Gaussian Graphical Models and Model Selection

Nicolas Verzelen

N° d'ordre: 9304

UNIVERSITE PARIS-SUD
FACULTE DES SCIENCES D'ORSAY

THESE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITE PARIS XI

Spécialité : Mathématiques

par

Nicolas Verzelen

Modèles graphiques gaussiens et sélection de modèles

Rapporteurs: Mme Florentina BUNEA
M. Jean-Philippe VERT

Soutenue le **17 Décembre 2008** devant la commission d'examen :

M.	Francis BACH	Examinateur
M.	Yannick BARAUD	Président du jury
Mme	Liliane BEL	Examinateur
M.	Xavier GUYON	Examinateur
M.	Pascal MASSART	Directeur de thèse
M.	Jean-Philippe VERT	Rapporteur

Remerciements

Mes premiers remerciements vont à Pascal. Lorsque j'étais en M2, tu m'as encouragé à réaliser mon stage dans l'Ohio sous la direction de Noel Cressie. Je remercie ce dernier de m'avoir initié aux statistiques spatiales. Je suis revenu à Orsay avec un germe de sujet de thèse. Pascal, tu as su m'aider à le faire évoluer et fructifier au cours de ces trois années. Au-delà des aspects purement scientifiques de ton encadrement, tu m'as appris que les discussions et échanges avec nos pairs sont certainement le plaisir essentiel en recherche.

Mon thème de recherche a beaucoup évolué au gré des rencontres ces dernières années. Fanny, ce fut un plaisir de travailler avec toi, qui m'a fait découvrir les applications en génomiques. Je pense que nous avons beaucoup appris mutuellement. Sylvie et Christophe, vos conseils et vos remarques ont eu une grande influence sur cette thèse et notre travail commun la poursuivra.

I was honoured that Florentina Bunea and Jean-Philippe Vert agreed to review my thesis. Francis Bach, Liliane Bel, Yannick Baraud et Xavier Guyon ont accepté de participer à mon jury, je leur en suis reconnaissant.

Lors de cours, séminaires, colloques, ou de simples pauses-cafés, j'ai pu dialoguer avec de nombreux statisticiens. Je remercie notamment Sylvain Arlot, Avner Bar-Hen, Lucien Birgé, Sthéphane Boucheron, Pierre Connault, Xavier Gendre, Mohammed Hebiri, Bertrand Michel, Patricia Reynaud-Bouret, Vincent Rivoirard, Gilles Stoltz, Laurent Zwald,...

Un grand merci à tous les relecteurs d'un ou plusieurs chapitres de cette thèse : Bérénice, Christophe, Dominique, Jean-Patrick, Martial, Martin et Sylvain. J'associe à mes remerciements Valérie Lavigne, qui m'a grandement facilité les tâches administratives.

Ce fut un plaisir de travailler à Orsay ; je remercie en particulier les doctorants du bureau 16. Enfin, merci à ceux (et surtout à celle) qui n'ont aucun lien scientifique avec cette thèse mais qui font que si je suis content d'aller au labo le matin, je le suis encore plus de le quitter le soir.

Avertissement

La présente thèse réunit des travaux sur les modèles graphiques gaussiens et la sélection de modèles. Si certains de ces travaux sont déjà publiés, d'autres sont récemment soumis ou sont encore en cours. Les chapitres 2 et 3 sont le fruit d'un travail commun avec Fanny Villers. Tous les autres sont issus d'un travail personnel.

Chaque chapitre peut être lu indépendamment des autres. Il en résulte quelques répétitions dans les concepts et résultats introduits. Cette thèse est organisée comme suit :

- Le Chapitre 1 est une présentation générale des connaissances actuelles en modélisation graphique et en sélection de modèles.
- Les Chapitres 2 et 3 traitent de tests d'adéquation du graphe d'un modèle graphique gaussien. Ils correspondent aux articles [VV08a] et [VV08b].
- Dans le Chapitre 4, nous nous intéressons à l'estimation adaptative dans un cadre de régression linéaire où le plan d'expérience est gaussien. Nous insistons également sur le lien existant entre ce cadre de régression et la modélisation graphique.
- Le Chapitre 5 traite d'estimation de covariance et de son lien avec les modèles graphiques dirigés.
- Dans les Chapitres 6 et 7, nous étudions l'estimation d'un champ gaussien stationnaire sur un réseau régulier. Nous y étudions une approche basée sur la modélisation graphique.

À l'exception du Chapitre 1, l'ensemble de cette thèse est rédigé en anglais.

Foreword

The present dissertation collects works on graphical models and model selection. The status of these works are heterogeneous (accepted for publication, to be submitted, or still in progress). Chapters 2 and 3 are joint works with Fanny Villers, while the other chapters are personal.

Each chapter can be read separately. This thesis is organized as follows.

- Chapter 1 is a general presentation (in french), of both the state-of-the-art and the contributions of this thesis.
- Chapters 2 and 3 deal with the problem of testing the graph of a Gaussian graphical model. They correspond to the articles [VV08a] and [VV08b].
- In Chapter 4, we consider the issue of adaptive estimation in a Gaussian design regression framework. Moreover, we emphasize the connection between this regression setting and Gaussian graphical models.
- Chapter 5 is about covariance estimation and directed graphical models.
- In Chapters 6 and 7, we study the estimation of a stationary Gaussian field using a graphical modelling approach.

Table des matières

Chapitre 1

Introduction

Dans de nombreux domaines, la compréhension de phénomènes complexes (climatologie, régulation génétique, etc...) repose à la fois sur de la modélisation et des expérimentations[1]. La récolte des données puis leur analyse statistique permet cet aller-retour entre modélisation et expérimentation. Aussi, est-il fondamental que scientifiques modélisateurs, expérimentateurs et statisticiens utilisent un langage commun afin de pouvoir facilement interagir.

En particulier, les protagonistes doivent s'accorder sur une modélisation probabiliste satisfaisante pour les uns comme pour les autres. Pour l'expérimentateur un « bon » modèle doit être simple à expliquer et être assez riche pour pouvoir prendre en compte les spécificités du phénomène étudié. En général, pour un statisticien un bon modèle doit être simple à *étudier* et se situer dans un cadre aussi abstrait que possible.

En ce sens, les modèles graphiques sont un bon candidat. Premièrement, leur représentation graphique les rend facilement interprétables. Deuxièmement, ce cadre est assez riche pour permettre la modélisation de phénomènes aussi différents que des systèmes de particules, des réseaux d'interactions entre gènes, des séries temporelles ou encore des systèmes de décision automatiques. Enfin, les modèles graphiques peuvent être étudiés dans un cadre maintenant bien formalisé par des décennies de recherche.

Le développement simultané de ces modèles dans des disciplines aussi différentes que la physique statistique, la statistique spatiale, l'apprentissage statistique, la psychologie s'est d'abord réalisé de façon relativement indépendante. Il a fallu attendre les années 90 pour que différents concepts soient unifiés. Il en résulte une terminologie changeante d'un domaine à l'autre. En statistique spatiale, on parle de champs de Markov ou de CAR[2]. En physique statistique, on utilise le terme champs de Gibbs tandis que le terme réseau bayésien provient de la théorie de l'apprentissage. Si ces notions sont légèrement différentes les unes des autres, toutes peuvent être replacées dans le cadre général de la modélisation graphique.

Cette thèse se situe dans le cadre de l'étude statistique des modèles graphiques gaussiens. L'estimation de la distribution d'un modèle graphique à structure connue est déjà bien traitée dans la littérature. Cependant, il n'en est pas de même lorsque la structure est inconnue. Dans cette thèse, nous proposons une étude *non-asymptotique* de diverses méthodes d'estimation et de test de la structure (chapitres 2, 3 et 4). La seconde moitié de la thèse (chapitres 5, 6 et 7) concerne l'estimation de la distribution d'un grand vecteur gaussien. Nous n'y utilisons plus les modèles graphiques comme un outil de modélisation mais plutôt comme une *bonne classe* d'approximation de la vraie distribution.

1.1 Modèles graphiques gaussiens

1.1.1 Définitions

Dans la suite du chapitre, Z désigne un vecteur gaussien de taille p, de moyenne nulle et de matrice de covariance Σ *inversible*. Nous introduisons maintenant les définitions et les propriétés des modèles graphiques nécessaires à la compréhension de cette thèse. Par souci de clarté, nous n'énoncerons les

[1] Dans certains domaines comme en écologie, il n'est pas possible de faire des expérimentations et il faut se contenter de *relevés de terrain*.

[2] Conditionally auto-regressive model

résultats que dans le cas gaussien avec covariance inversible. Le lecteur intéressé trouvera un traitement plus général dans le livre de Lauritzen [Lau96].

La notion centrale en modélisation graphique est l'*indépendance conditionnelle*. Pour tout ensemble $A \subset \{1, \dots, p\}$, on note Z_A l'ensemble $\{Z_i,\ i \in A\}$. Étant donnés trois sous-ensembles A, B et C de $\{1, \dots, p\}$, on dit que « Z_A est indépendant de Z_B conditionnellement à Z_C » et on note

$$Z_A \perp\!\!\!\perp Z_B | Z_C \ ,$$

si la densité $f_{Z_A,Z_B,Z_C}(z_A, z_B, z_C)$ se factorise sous la forme $f_{Z_A,Z_C}(z_A, z_C) f_{Z_B,Z_C}(z_B, z_C)$. On peut interpréter cette notion en terme d'information : « connaissant Z_C, Z_B n'apporte aucune information pour prédire Z_A ».

1.1.1.1 Modèles graphiques non-orientés

Soit $\mathcal{G} = (\Gamma, E)$ un graphe non-orienté fini. $\Gamma = \{1, \dots, p\}$ désigne l'ensemble des sommets tandis que E est l'ensemble des arêtes de $\mathcal{G}$. Dans la suite, on utilisera Γ lorsque l'on considère les variables (Z_i) comme étant non ordonnées. La notation $\{1, \dots, p\}$ sera réservée au cas où l'ordre des variables intervient. On écrit $i \leftrightarrow_{\mathcal{G}} j$ s'il existe une arête entre deux sommets i et j dans $\mathcal{G}$ et $i \nleftrightarrow_{\mathcal{G}} j$ s'il n'y a pas une telle arête. Pour tout sommet $i \in \Gamma$ on note $ne_{\mathcal{G}}(i)$ l'ensemble des voisins de i dans le graphe $\mathcal{G}$, i.e. : l'ensemble des sommets j qui sont reliés par une arête à i. Comme $\Gamma = \{1, \dots, p\}$, le vecteur Z est indexé par les sommets du graphe $\mathcal{G}$.

Définition 1.1. Le vecteur aléatoire Z satisfait la propriété de *Markov locale* par rapport à $\mathcal{G}$ si pour tout sommet i dans $\mathcal{G}$, Z_i est indépendant de ses non-voisins $Z_{\Gamma \setminus ne_{\mathcal{G}}(i) \cup \{i\}}$ conditionnellement à ses voisins $Z_{ne_{\mathcal{G}}(i)}$:

$$\forall i \in \Gamma, \qquad Z_i \perp\!\!\!\perp Z_{\Gamma \setminus ne_{\mathcal{G}}(i) \cup \{i\}} | Z_{ne_{\mathcal{G}}(i)} \ . \tag{1.1}$$

Dans ce cas, on dit que Z est un *modèle graphique gaussien* par rapport à $\mathcal{G}$.

Il en résulte que la propriété « Z est un modèle graphique gaussien par rapport au graphe $\mathcal{G}$ » est équivalente à la spécification d'un certain nombre d'indépendances conditionnelles liées à $\mathcal{G}$. Ainsi, supposer que Z est un modèle graphique gaussien par rapport au graphe vide (i.e. $E = \emptyset$) revient à supposer que les variables $(Z_i)_{i \in \Gamma}$ sont toutes indépendantes. Inversement, tout vecteur gaussien Z est un modèle graphique gaussien par rapport au graphe complet (i.e. les sommets sont deux à deux reliés). De plus, si $\mathcal{G}' = (\Gamma, E')$ est un sur-graphe de $\mathcal{G}$ (i.e. E' contient E) et si Z est un modèle graphique gaussien par rapport $\mathcal{G}$, alors Z est un modèle graphique gaussien par rapport à $\mathcal{G}'$. Ainsi, quelle que soit Σ la matrice de covariance inversible de Z, il existe un graphe $\mathcal{G}_{\min}$ minimal pour l'inclusion tel que Z est un modèle graphique gaussien par rapport à $\mathcal{G}_{\min}$. On verra dans la suite que ce graphe $\mathcal{G}_{\min}$ est unique.

Il existe d'autres propriétés de Markov sur des graphes que la propriété de Markov locale. Mentionnons la propriété de Markov globale qui nous sera utile dans la suite. Soient (A, B, C) trois sous-ensembles disjoints de Γ, on dit que C sépare A et B dans $\mathcal{G}$ si tout chemin du graphe qui part d'un élément de A pour aller à un élément de B passe par un élément de C.

Définition 1.2. Le vecteur Z satisfait la propriété de *Markov globale* par rapport à $\mathcal{G}$ si pour tout triplet (A, B, C) de sous-ensembles disjoints de Γ tel que C sépare A et B dans $\mathcal{G}$,

$$Z_A \perp\!\!\!\perp Z_B | Z_C \ . \tag{1.2}$$

La figure 1 donne un exemple de graphe et des propriétés de Markov associées. Elle est équivalente à la propriété de Markov locale dans notre cadre gaussien avec covariance inversible mais pas en général (cf. Lauritzen [Lau96] Sect. 3.1). Une conséquence directe de cette propriété de Markov globale est la suivante. Soit $\mathcal{G}$ un graphe. Si deux sommets i et j sont dans deux composantes connexes différentes de $\mathcal{G}$ et si Z est un modèle graphique par rapport à $\mathcal{G}$, alors Z_i et Z_j sont indépendants.

Les deux propositions suivantes sont à la base de la plupart des résultats de cette thèse. Elle énonce le lien entre graphe d'un modèle graphique et la matrice de précision $\Omega := \Sigma^{-1}$ du vecteur Z.

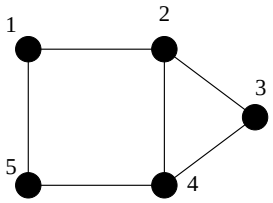


Figure 1 – D'après la propriété de Markov locale en 1, on a $Z_1 \perp\!\!\!\perp Z_{\{3,4\}} | Z_{\{2,5\}}$. La propriété de Markov globale implique aussi $Z_1 \perp\!\!\!\perp Z_3 | Z_{\{2,4\}}$.

Proposition 1.3. *Pour tout sommet $i \in \Gamma$, la distribution conditionnelle de Z_i par rapport aux variables restantes s'écrit*

$$Z_i = - \sum_{j \in \Gamma \setminus \{i\}} \Omega[i,j] \Omega[i,i]^{-1} Z_j + \epsilon_i \ ,$$

où ϵ_i suit une loi normale centrée de variance $\Omega[i,i]^{-1}$ et est indépendante de $Z_{\Gamma \setminus \{i\}}$.

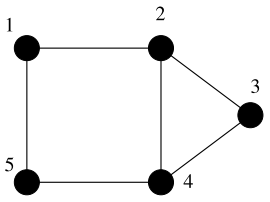
Proposition 1.4. *Supposons que Z est un modèle graphique gaussien par rapport à $\mathcal{G}$, alors*

$$\forall (i,j) \in \Gamma^2 \ , \quad (i \nleftrightarrow_{\mathcal{G}} j) \text{ et } i \neq j \quad \Longrightarrow \quad (\Omega[i,j] = 0) \ .$$

En d'autres termes, $\Omega[i,j] = 0$ pour tout couple (i,j) de sommets non reliés par une arête dans $\mathcal{G}$. Réciproquement, quel que soit la matrice de précision Ω inversible, Z est un modèle graphique gaussien par rapport au graphe $\mathcal{G}_\Omega = (\Gamma, E_\Omega)$ défini par

$$\forall (i,j) \in \Gamma^2 \ , \quad (i \leftrightarrow_{\mathcal{G}_\Omega} j) \quad \Longleftrightarrow \quad (i \neq j \text{ et } \Omega[i,j] \neq 0) \ .$$

La première proposition est un résultat connu d'analyse gaussienne multivariée (voir par exemple [Lau96] App.C). La seconde proposition est en fait une conséquence de la première. Il en ressort que le graphe minimal $\mathcal{G}_{\min}$ d'un vecteur Z est unique et entièrement caractérisé par l'emplacement des 0 dans la matrice de précision (i.e. $\mathcal{G}_{\min} = \mathcal{G}_\Omega$). Là réside tout l'intérêt de la modélisation graphique : on a d'une part une formulation en termes d'indépendances conditionnelles facilement interprétable à l'aide du graphe. D'autre part, il existe une formulation équivalente en termes de zéros de la matrice de précision, qui est elle propice à une étude statistique et probabiliste.



	1	2	3	4	5
1	*	*	0	0	*
2	*	*	*	*	0
3	0	*	*	*	0
4	0	*	*	*	*
5	*	0	0	*	*

Figure 2 – Lien entre graphe minimal de Z et zéros de sa matrice de précision.

1.1.1.2 Modèles graphiques orientés

Soit $\overrightarrow{\mathcal{G}} = (\Gamma, \overrightarrow{E})$ un graphe orienté fini. Les arêtes $(i,j) \in \overrightarrow{E}$ sont maintenant orientées. Pour tout couple (i,j) de sommets, j est un *parent* de i s'il existe une arête orientée qui va de j vers i. On le note $i \rightarrow_{\overrightarrow{\mathcal{G}}} j$. On suppose de plus que le graphe $\overrightarrow{\mathcal{G}}$ est acyclique, c'est à dire qu'il n'existe pas de chemin orienté (qui suit le sens des arêtes) dans $\overrightarrow{\mathcal{G}}$ qui réalise un cycle. Un sommet j est descendant de i s'il existe un chemin orienté allant de i vers j. Un sommet j est non-descendant de i s'il n'est pas un descendant de i. On note respectivement $pa_{\overrightarrow{\mathcal{G}}}(i)$ et $nd_{\overrightarrow{\mathcal{G}}}(i)$ l'ensemble des parents de i et l'ensemble des non-descendants de i.

Définition 1.5. Soit $\overrightarrow{\mathcal{G}}$ un graphe orienté fini. Le vecteur Z satisfait la propriété de *Markov locale orientée* par rapport à $\overrightarrow{\mathcal{G}}$ si pour tout sommet i dans $\overrightarrow{\mathcal{G}}$, Z_i est indépendant de ses non-descendants $Z_{nd_{\overrightarrow{\mathcal{G}}}(i)}$ conditionnellement à ses parents $Z_{pa_{\overrightarrow{\mathcal{G}}}(i)}$:

$$Z_i \perp\!\!\!\perp Z_{nd_{\overrightarrow{\mathcal{G}}}(i)} | Z_{pa_{\overrightarrow{\mathcal{G}}}(i)} \ . \tag{1.3}$$

On dit que Z est un *modèle graphique gaussien orienté* par rapport à $\overrightarrow{\mathcal{G}}$ s'il satisfait la propriété de Markov locale par rapport à $\overrightarrow{\mathcal{G}}$.

Exemple 1.6. *Considérons une série temporelle $(Z_i)_{i\in\mathbb{Z}}$ gaussienne stationnaire et autorégressive d'ordre 1 : il existe $a \in]-1;1[$ tel que pour tout indice i*

$$Z_{i+1} = aZ_i + \epsilon_i \ ,$$

où les variables ϵ_i sont indépendantes et ont une variance égale à $1-a^2$. Pour n'importe quel entier positif p, le vecteur $(Z_i)_{1\leq i\leq p}$ est un modèle graphique orienté par rapport au graphe qui relie tout sommet $i \in \{1; p-1\}$ à $i+1$. C'est également un modèle graphique orienté par rapport au graphe inverse (i.e. toutes les arêtes ont été retournées).

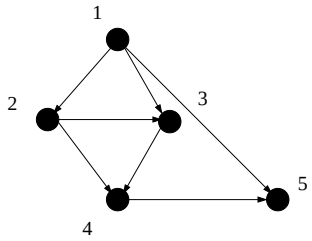


Figure 3 – Propriété de Markov locale orientée en 4 : $Z_4 \perp\!\!\!\perp Z_1 | Z_{\{2,3\}}$.

La figure 3 illustre la propriété de Markov locale orientée dans un cas particulier. Comme dans le cas non-orienté, la propriété « Z est un *modèle graphique gaussien orienté* par rapport à $\overrightarrow{\mathcal{G}}$ » est d'autant plus forte que le graphe $\overrightarrow{\mathcal{G}}$ est parcimonieux. Si $\overrightarrow{\mathcal{G}}'$ est un surgraphe de $\overrightarrow{\mathcal{G}}$ et si Z est un modèle graphique gaussien orienté par rapport à $\overrightarrow{\mathcal{G}}$, il l'est également par rapport $\overrightarrow{\mathcal{G}}'$. Si $\overrightarrow{\mathcal{G}}$ est vide, les variables (Z_i) sont indépendantes. Inversement, tout vecteur gaussien (de matrice de covariance inversible) est un modèle graphique orienté par rapport aux graphes orientés acycliques tels que tous les sommets sont deux à deux reliés. En effet, pour un tel graphe, les non-descendants d'un sommet sont tous des parents.

Comme le graphe $\overrightarrow{\mathcal{G}}$ est acyclique, on peut réaliser une numérotation des sommets compatible avec le sens des arêtes, i.e. qui satisfait la propriété :

$$\forall\ (i,j) \in \Gamma^2 \ , \quad (i \rightarrow_{\overrightarrow{\mathcal{G}}} j) \quad \Rightarrow \quad i \leq j \ .$$

Cette numérotation n'est toutefois pas unique. Un exemple d'une telle numérotation est donné sur la figure 3. Quitte à réordonner les variables, nous supposons que la numérotation naturelle des variables $(Z_i)_{1\leq i\leq p}$ par leur indice est compatible avec le graphe $\overrightarrow{\mathcal{G}}$. Dans la suite, (T,S) désigne l'unique couple de matrices qui satisfait

$$\Omega = T^* S^{-1} T \ ,$$

et tel que T est une matrice triangulaire inférieure de diagonale 1 et S est une matrice diagonale positive. T est appelé le *facteur de Cholesky* de la matrice Ω. Si les graphes des modèles graphiques non orientés sont caractérisés par les zéros de la matrice de précision, les propositions qui suivent lient le graphe d'un modèle graphique orienté avec les zéros du facteur de Cholesky T.

Proposition 1.7. *Pour tout sommet $i \in \{1,\dots p\}$, la distribution conditionnelle de Z_i par rapport à l'ensemble $Z_{<i}$ des variables $\{Z_j\}_{j<i}$ peut s'écrire*

$$Z_i = -\sum_{j=1}^{i-1} T[j,i] Z_j + \epsilon_i \ , \tag{1.4}$$

où ϵ_i est un variable gaussienne centrée de variance $S[i,i]$ indépendante de $Z_{<i}$. De plus, les variables $(\epsilon_i)_{1\leq i\leq p}$ sont indépendantes.

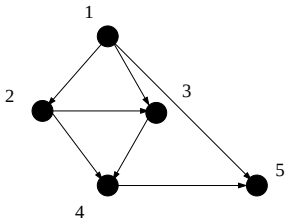
Proposition 1.8. *Supposons que Z est un modèle graphique gaussien par rapport au graphe orienté acyclique $\overrightarrow{\mathcal{G}}$ et que la numérotation des variables dans Z est compatible avec le graphe $\overrightarrow{\mathcal{G}}$, alors la matrice T satisfait :*

$$\forall\, i<j\in\{1,\dots,p\}^2,\quad i\nrightarrow_{\overrightarrow{\mathcal{G}}} j\ \Rightarrow\ T[j,i]=0\ .$$

Réciproquement, quelle que soit la matrice de précision Ω inversible, Z est un modèle graphique par rapport au graphe orienté acyclique $\overrightarrow{\mathcal{G}_T}$ défini par

$$(i\ \rightarrow_{\overrightarrow{\mathcal{G}_T}}\ j)\quad\Longleftrightarrow\quad[i<j\ \textit{et}\ T[j,i]\neq 0]\ .$$

Un exemple de cette dualité entre graphe et facteur de Cholesky est présenté sur la figure 4. Ayant numéroté les variables, il existe donc un unique graphe orienté acyclique $\overrightarrow{\mathcal{G}_T}$ minimal pour l'inclusion qui soit compatible avec la numérotation. De plus, ce graphe est entièrement défini par les zéros du facteur de Cholesky T. Cependant, il n'y a *pas* unicité du graphe orienté acyclique minimal si on n'impose pas de numérotation particulière.



	1	2	3	4	5
1	1	0	0	0	0
2	*	1	0	0	0
3	*	*	1	0	0
4	0	*	*	1	0
5	*	0	0	*	1

Figure 4 – Lien entre graphe du modèle graphique orienté et zéros du facteur de Cholesky.

Il est possible de simuler facilement un modèle graphique orienté à partir des distribution conditionnelles des variables Z_i : on simule d'abord Z_1, qui suit une distribution gaussienne centrée de variance $S[i,i]$. Pour tout $i\geq 2$, on simule Z_i en tirant la variable gaussienne ϵ_i de variance $S[i,i]$ et en utilisant la régression (1.4) :

$$Z_i=\sum_{j\in pa_{\overrightarrow{\mathcal{G}}}(i)} -T[j,i]Z_j+\epsilon_i\ .$$

Cette méthode de simulation illustre bien la notion de causalité sous-jacente aux modèles graphiques orientés : la variable Z_i est la somme d'une combinaison linéaire de ses parents et d'une *innovation* indépendante.

1.1.1.3 Liens entre modèles orientés et modèles non-orientés

Les formulations des modèles orientés et non-orientés sont légèrement différentes et chacune a son avantage selon le contexte. Nous verrons ainsi à la fin de cette sous-section que les modèles graphiques orientés sont plus facile à estimer. Pour pouvoir pleinement utiliser les avantages de ces deux types de modélisation, il est utile de connaître des transformations simples qui permettent de passer de l'un à l'autre.

Si Z est un modèle graphique gaussien orienté par rapport au graphe $\overrightarrow{\mathcal{G}}$ alors c'est également un modèle graphique gaussien par rapport au graphe moral $\mathcal{G}^m$ défini par :

$$i\leftrightarrow_{\mathcal{G}^m} j\quad\Leftrightarrow\quad\Big[i\rightarrow_{\overrightarrow{\mathcal{G}}} j\text{ ou }j\rightarrow_{\overrightarrow{\mathcal{G}}} i\text{ ou }\Big(\exists k\in\{1,\dots,p\},\ i\rightarrow_{\overrightarrow{\mathcal{G}}} k\text{ et }j\rightarrow_{\overrightarrow{\mathcal{G}}} k\Big)\Big]\ .$$

En d'autres termes, il y a une arête entre deux sommets i et j dans le graphe moral si i et j sont reliés dans le graphe orienté ou s'ils ont un enfant commun. Nous illustrons cette notion en figure 5.

Inversement, les modèles graphiques gaussiens par rapport à certains types de graphes $\mathcal{G}$ sont également des modèles graphiques orientés par rapport à un graphe $\overrightarrow{\mathcal{G}}_{ord}$ qui contient le même nombre

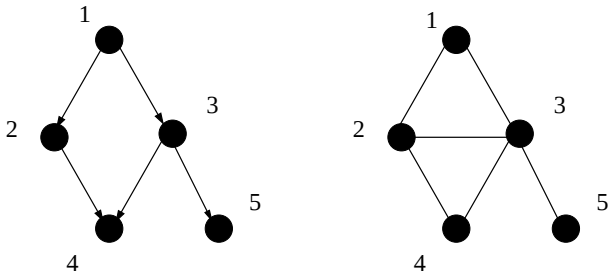


Figure 5 – Graphe orienté acyclique et graphe moral associé.

d'arêtes que le graphe initial. Pour cela, nous devons définir la notion de graphe *décomposable* (également appelé graphe *triangulaire*).

Soit $\mathcal{G}$ un graphe non-orienté et soit $i_1 i_2 \dots i_k$ un cycle de ce graphe. Une corde du cycle est une arête qui relie deux sommets non consécutifs du cycle. Le graphe $\mathcal{G}$ est dit *décomposable* si tout cycle de longueur supérieure ou égale à 4 contient au une moins corde. Un exemple de graphe décomposable et un exemple de graphe non décomposable sont représentés sur la figure 6.

Un ensemble A de sommets dans un graphe est dit *complet* si tous les éléments de A sont deux à deux connectés. On appelle numérotation *parfaite* des sommets d'un graphe $\mathcal{G}$ un ordre complet sur Γ qui satisfait

$$\forall\ j = 2, \dots, p : \quad ne_{\mathcal{G}}(j) \cap \{1, \dots, j-1\} \quad \text{est complet dans } \mathcal{G}\ .$$

Il y a équivalence entre la caractère décomposable d'un graphe $\mathcal{G}$ et l'existence d'une numérotation parfaite des sommets de $\mathcal{G}$. Si $\mathcal{G}$ est décomposable, la propriété de Markov locale par rapport à $\mathcal{G}$ implique la propriété de Markov locale orientée par rapport au graphe $\overrightarrow{\mathcal{G}}$ qui correspond au graphe $\mathcal{G}$ dont les arêtes ont été orientées pour être compatibles avec la numérotation parfaite (voir par exemple [Lau96]). Ainsi, un modèle graphique par rapport au graphe $\mathcal{G}$ est aussi un modèle graphique orienté par rapport à $\overrightarrow{\mathcal{G}}$. Un exemple d'une telle numérotation et du graphe orienté associé est donné sur la figure 6.

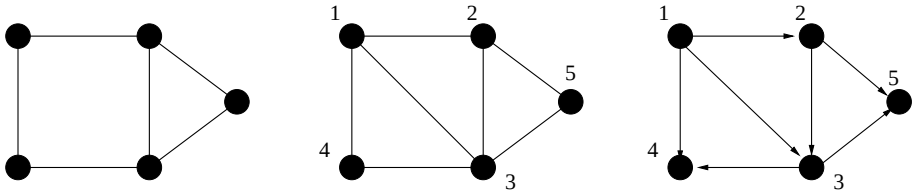


Figure 6 – Le premier graphe n'est pas décomposable. Le deuxième graphe est décomposable. Sa numérotation est parfaite ce qui permet de construire le graphe dirigé associé.

En résumé, les modèles graphiques par rapport à des graphes décomposables sont également des modèles graphiques orientés par rapport à un graphe de même taille. Si le graphe n'est pas décomposable, il est possible de se ramener à un modèle graphique orienté en rajoutant des arêtes au graphe initial pour le rendre décomposable. Cependant, cet ajout est algorithmiquement difficile[3].

Illustrons l'intérêt des transformations d'un modèle à un autre dans le cadre de l'*estimation par maximum de vraisemblance*. L'estimation de la matrice de covariance Σ lorsque le graphe est connu est bien étudiée depuis les années 90. On observe un n-échantillon du modèle graphique gaussien Z par rapport à un graphe $\overrightarrow{\mathcal{G}}$ orienté acyclique. L'estimateur du maximum de vraisemblance de Σ est presque sûrement défini si et seulement si $n > \max_{1 \le i \le p} |pa_{\overrightarrow{\mathcal{G}}}(i)|$. De plus, on en connaît une formule explicite (voir chapitre 5). Si Z est un modèle graphique par rapport à un graphe $\mathcal{G}$ non-orienté, il y a deux cas possibles. Soit $\mathcal{G}$ est décomposable et on peut se ramener à un graphe orienté. Soit $\mathcal{G}$ n'est pas décomposable et il n'existe pas de critère aussi simple pour caractériser l'existence du maximum de vraisemblance. Même lorsque celui-ci existe, on n'en connaît pas de formule explicite et il est nécessaire d'utiliser un algorithme itératif

[3] Ce problème est NP-dur.

pour le calculer.

Pour conclure cette rapide présentation des modèles graphiques, mentionnons qu'il existe d'autres façons de représenter des indépendances conditionnelles, parmi lesquelles les modèles graphiques chaînés qui utilisent des graphes partiellement orientés. On consultera [Lau96] pour plus de détails.

1.1.2 Applications et motivations

Nous présentons dans cette sous-section deux domaines d'application où interviennent des modèles graphiques gaussiens.

Puces à ADN et réseaux génétiques : un des enjeux majeurs en génomiques est la recherche de liens fonctionnels entre les gènes. On appelle réseau génétique un réseau dont les sommets représentent des gènes et les arêtes correspondent aux différentes régulations ayant lieu entre les gènes. L'émergence récente de nouvelles méthodes de mesures d'expression de gènes (ex : puces à ADN) a considérablement augmenté la quantité de données disponibles et permet maintenant l'étude statistique des réseaux génétiques. La particularité de ces données est que le nombre de répétitions n est beaucoup plus petit que le nombre p de gènes observés. Une hypothèse généralement faite est que les données provenant de la puce à ADN correspondent (après renormalisation) à un modèle graphique gaussien par rapport au réseau génétique. On peut alors utiliser les données de puces à ADN pour *inférer* le réseau génétique. Dans certains cas, les biologistes ont une bonne connaissance des interactions entre les gènes. L'objectif est alors de vérifier qu'aucune interaction n'a été oubliée.

Modèles spatiaux : en analyse d'image et en statistiques spatiales, les données sont parfois indexées par les sommets d'une grille régulière. Le champ aléatoire Z sous-jacent est généralement supposé stationnaire mais le nombre d'observations du champ peut valoir un. Il existe alors deux stratégies classiques pour estimer la distribution du champ : soit on considère les corrélations entre les variables, soit on considère la distribution conditionnelle d'une variable par rapport aux autres. Cette seconde approche vise à approcher la loi de Z par celle d'un modèle graphique gaussien stationnaire sur la grille. L'utilisation des modèles graphiques n'est pas tant motivée ici par leur interprétation en termes d'indépendances conditionelles que par leur « bonne » capacité d'approximation de la vrai distribution de Z.

1.1.3 Objectifs de la thèse

Formalisons maintenant les questions étudiées dans cette thèse. Soit Z un vecteur gaussien centré de taille p et de matrice de covariance Σ *inconnue* (mais inversible). On note Ω la matrice de précision et (T, S) les termes de la décomposition de Cholesky de Ω définis en Section 1.1.1.2. Supposons que l'on observe un n-échantillon de Z. Dans la suite, $\mathbf{Z}$ désigne la matrice $p \times n$ des observations de Z. Notre objectif est de construire et d'étudier des procédures statistiques qui répondent à l'un des trois problèmes suivants :

A.1. **Estimation du graphe** : Sachant que Z est un modèle graphique gaussien par rapport à un graphe minimal $\mathcal{G}$ inconnu (ou partiellement connu), on veut construire un estimateur $\widehat{\mathcal{G}}$ consistant de ce graphe. Par exemple, $\mathbf{Z}$ peut représenter des données de puces à ADN et l'objectif est d'inférer le réseau d'interactions géniques. On considère également le problème analogue dans le cas orienté. Dans ce cas, Z est un modèle graphique gaussien orienté par rapport à un graphe minimal orienté acyclique $\overrightarrow{\mathcal{G}}$ inconnu (ou partiellement connu) que l'on veut estimer. On verra en section 1.3.2 que le problème d'inférence d'un graphe orienté est mal posé si on ne connaît pas a priori une numérotation des variables compatible avec $\overrightarrow{\mathcal{G}}$. Dans une telle situation, il nous faudra légèrement reformuler la question.

A.2. **Test du graphe** : Étant donné un graphe non orienté $\mathcal{G}$ ou un graphe orienté acyclique $\overrightarrow{\mathcal{G}}$, on désire tester l'hypothèse : « Z est un modèle graphique gaussien par rapport à $\mathcal{G}$ (ou $\overrightarrow{\mathcal{G}}$) », contre une hypothèse alternative spécifiée ou non-paramétrique. Une application potentielle est la suivante : on a à notre disposition un modèle de réseau de régulations génétiques. En utilisant des données de puces à

ADN, on veut valider ce réseau ou vérifier que certaines interactions n'ont pas été oubliées.

A.3. **Estimation de la distribution de** Z : Cette fois-ci, nous ne supposons pas que Z est un modèle graphique par rapport à un graphe particulier. L'objectif est d'estimer la matrice de covariance Σ (ou de façon équivalente Ω). En toute généralité, il y a $p(p+1)/2$ paramètres à estimer. Si n est plus petit que p, il y a plus de paramètres à estimer que de données, et l'estimateur du maximum de vraisemblance de Σ n'est pas défini. De même, il est connu que la matrice de covariance empirique donne de mauvais résultats. Une approche classique dans cette situation consiste à supposer qu'il y a de la parcimonie en un certain sens : la matrice Σ, la matrice Ω ou encore le facteur de Cholesky T sont bien « approchés » par une matrice « creuse ». Nous nous intéressons en particulier aux approches basées sur la parcimonie de Ω ou de T. Grâce aux propositions 1.4 et 1.8, nous savons qu'il y a équivalence entre les propriétés « Ω ou T est approximativement creux » et « la distribution de Z est bien approchée par un modèle graphique gaussien de graphe parcimonieux ou un modèle graphique gaussien orienté de graphe parcimonieux et compatible avec la numérotation des variables ». Les modèles graphiques sont donc utilisés ici en tant que classes d'approximation de la matrice Σ. On veut donc construire une procédure statistique qui sélectionne un *bon* graphe et estime la distribution de Z parmi les modèles graphiques par rapport à ce graphe. Un *bon* graphe contient assez d'arêtes pour pouvoir bien approcher la distribution de Z, mais est assez parcimonieux pour que le nombre de paramètres à estimer dans Ω (ou T) reste assez petit. Cette approche se révèle utile en statistique spatiale pour estimer la distribution d'un champ gaussien stationnaire (Chapitre 6) ou pour réaliser de l'analyse linéaire discriminante d'IRM fonctionnelles. On expliquera plus précisément en section 1.4 ce qu'on entend par « bonne approximation » et dans quels cas il est plus pertinent de supposer que Ω ou T sont approximativement creux que de supposer que Σ est approximativement creux.

Dans notre cadre d'étude, une bonne procédure statistique doit satisfaire les qualités suivantes :

1. *Grande dimension.* La procédure est applicable et donne de bons résultats (en théorie) même lorsque la taille p du vecteur est beaucoup plus grande que le nombre d'observations n. Cette situation est en effet caractéristique des données génomiques ou des IRM fonctionnelles. Par exemple, les méthodes d'estimation reposant sur l'inversion de la matrice de covariance empirique ne sont définies que pour $n > p$, ce qui les rend inutilisable en grande dimension.
2. *Coût computationnel faible.* Le temps de calcul doit rester raisonnable même lorsque p ou n sont grands. Ainsi, pour le problème **A.1** d'estimation de graphes, il est irréaliste de vouloir considérer les $2^{p(p-1)/2}$ graphes possibles à p sommets lorsque p est plus grand que 10. Pour l'analyse de données de puces à ADN où p peut aller jusqu'à plusieurs milliers, une grande attention doit donc être portée à la complexité des algorithmes utilisés.
3. *Optimalité et adaptativité.* La procédure doit non seulement être consistante mais l'erreur d'estimation (ou de test) doit en plus converger vers 0 à la vitesse optimale. Des résultats de convergence non-asymptotiques sont préférés à des résultats asymptotiques ($n \to \infty$, p fixe) qui ne rendent pas forcément bien compte du comportement d'une procédure lorsque p est plus grand que n. Les vitesses optimales de convergence dépendent généralement de propriétés inconnues sur la cible. Par exemple un graphe $\mathcal{G}$ qui ne contient que très peu d'arêtes est plus facile à estimer qu'un graphe contenant beaucoup d'arêtes. Or, on ne connaît pas a priori le degré de parcimonie du graphe cible. Il est appréciable que la procédure statistique s'*adapte* à cette caractéristique, i.e. atteigne la vitesse optimale d'estimation quelle que soit la parcimonie et ceci sans la connaître a priori.
4. *Universalité.* On fait très peu (voire pas du tout) d'hypothèses sur la matrice Σ ou sur le graphe $\mathcal{G}$ pour que la procédure statistique converge ou atteigne la vitesse optimale de convergence.
5. *Flexibilité.* La procédure peut-elle s'adapter à des connaissance a priori ou au contexte ? Dans le cas des analyses de données de puces à ADN, les biologistes ont souvent une connaissance partielle du graphe d'interaction. Pour des données d'origine temporelle ou spatiale, on s'attend généralement à ce qu'il y ait plus de dépendances entre des variables proches dans le temps ou l'espace qu'entre des variables éloignées. Il est donc souhaitable que la procédure soit assez flexible pour intégrer ce type de connaissance a priori.
6. *Calibration.* La mise en oeuvre de la procédure dépend le moins possible (idéalement pas du tout) de

paramètres inconnus. Ainsi, pour l'estimation de la matrice Σ, il n'est pas très satisfaisant d'utiliser un critère pénalisé qui dépend de la plus grande valeur propre de Σ.

À notre connaissance, hormis pour la question **A.2**[4], il n'existe aucune procédure qui combine toutes ces qualités. En pratique, le choix d'une méthode par rapport à une autre dépend du contexte de l'étude et de l'importance relative accordée à chacun des critères précédents.

Pour chacune des trois questions évoquées, nous construisons dans cette thèse des procédures statistiques qui prennent en compte la grande dimension, valables sous peu d'hypothèses, flexibles et faciles à calibrer. Nous prenons un soin particulier à calculer les vitesses optimales de convergence et de test d'un point de vue *non-asymptotique*. Ainsi, nous pouvons caractériser les propriétés d'adaptativité des procédures introduites. Enfin, nous discutons leur coût computationnel qui, selon le cadre reste raisonnable ou devient prohibitif lorsque p grandit.

L'approche privilégiée dans cette thèse est d'étudier les propriétés générales de Z (problèmes **A.1**, **A.2**, ou **A.3**) sous l'angle des régressions conditionnelles. Grâce à la proposition 1.3, nous savons que pour $A \subset \Gamma \setminus \{i\}$ la distribution de Z_i conditionnellement à Z_A se décompose sous la forme :

$$Z_i = \sum_{j \in A} \theta_j Z_j + \epsilon \ ,$$

où $\theta \in \mathbb{R}^A$ et ϵ est indépendant de Z_A. Ainsi l'étude des distributions conditionnelles se ramène à l'étude de modèle de régression linéaire dit *à design gaussien*. Dans la prochaine section, nous explicitons le lien existant entre indépendances conditionnelles et support de la régression conditionnelle. De plus, nous expliquons en quoi chacun des problèmes **A.1**, **A.2**, et **A.3** trouve son pendant en régression conditionnelle. Puis, nous introduisons les concepts statistiques nécessaires à une étude « non-asymptotique » des méthodes d'estimation et de test pour cette régression. Dans la section 1.3, nous revenons aux problèmes **A.1** et **A.2** d'estimation et de test de graphes. Après une présentation des méthodes existantes, nous décrivons nos procédures basées sur les résultats obtenus en régression. Enfin, nous expliquons en section 1.4 dans quelle mesure l'étude du modèle de régression précédent est utile pour l'estimation de la distribution du vecteur gaussien Z (problème **A.3**). On évoquera en particulier le cas où les variables Z_i ont un ordre naturel et le cas où Z correspond à un champ stationnaire.

1.2 Régression linéaire à design gaussien

Dans cette section, nous considérons le modèle de régression suivant

$$Y = \sum_{i=1}^{p-1} \theta_i X_i + \epsilon \ , \tag{1.5}$$

où le vecteur réel $X = (X_i)_{1 \leq i \leq (p-1)}$ suit une loi jointe gaussienne de moyenne nulle et de matrice de covariance Λ *inconnue* mais néanmoins inversible. La variable ϵ est indépendante du vecteur X et suit une loi normale de moyenne nulle et de variance σ^2 *inconnue*. Nous avons mentionné en proposition 1.3 que cette régression correspond à la distribution de la variable Y conditionnellement à X. Il en ressort que σ^2 est la variance conditionnelle de Y par rapport à X et que l'espérance conditionnelle de Y par rapport à X est $\sum_{i=1}^{p-1} \theta_i X_i$.

$$\mathbb{E}(Y|X) = \sum_{i=1}^{p-1} \theta_i X_i \qquad \text{et} \qquad \mathrm{var}(Y|X) = \sigma^2 \ .$$

Explicitons maintenant la connexion entre indépendance conditionnelle et support de la régression. La preuve de ce lemme éclaire bien les différents concepts et nous la produisons donc ici.

Lemme 1.9. *Soit J un sous-ensemble de $\{1, \dots, p-1\}$. Il y a équivalence entre les deux assertions suivantes :*

1. *Y est indépendant de X_{J^c} conditionnellement à X_J*

[4] Voir le chapitre 3.

2. Le support de θ est inclus dans J.

Preuve: Supposons d'abord la première assertion réalisée. En conséquence, l'espérance de Y conditionnellement à X est égale à l'espérance de Y conditionnellement à X_J. Notons θ' l'unique vecteur à support dans J tel que $\mathbb{E}(Y|X_J) = \sum_{i\in J}\theta'_i X_i$ presque sûrement. Il en suit que l'identité $\sum_{i\in J}(\theta'_i - \theta_i)X_i = \sum_{i\in J^c}\theta_i X_i$ est réalisée presque sûrement. Comme la covariance du vecteur X est inversible, on conclut que θ_i vaut 0 pour tout indice i dans J^c.

Supposons maintenant la deuxième assertion réalisée. La distribution de Y conditionnellement à X_J est une loi normale de moyenne $\sum_{i\in J}\theta_i X_i$ et de variance σ^2. La distribution de Y conditionnellement à X est également une loi normale de moyenne $\sum_{i\in J}\theta_i X_i$ et de variance σ^2. Les lois de Y conditionnellement à X_J et conditionnellement à X sont identiques. En conséquence, Y est indépendant de X_{J^c} conditionnellement à X_J.

■

On dispose d'un n-échantillon du couple (Y, X). Dans la suite, je note $\mathbf{Y}$ le vecteur de taille n des observations de Y ainsi que $\mathbf{X}$ la matrice de taille $n \times (p-1)$ des observations du vecteur X. On considère alors les trois problèmes suivants :

B.1. Quel est le support de θ ? D'après le lemme 1.9, une reformulation équivalente est : quelles sont les indépendances conditionnelles de la variable Y par rapport aux variables $(X_i)_{1\leq i\leq p-1}$? Ainsi, l'inférence du support d'une régression conditionnelle permet d'inférer les voisins (ou les enfants) d'un sommet pour un modèle graphique. La question B.1 est donc un sous-problème de l'inférence du graphe **A.1** énoncé dans la précédente section. Notons m^*_{vrai} le support du vecteur θ. Notre objectif est alors de construire un estimateur $\widehat{m}$ de m^*_{vrai} tel que la probabilité $\mathbb{P}(\widehat{m} = m^*_{\text{vrai}})$ est aussi proche de 1 que possible.

B.2. Étant donné Θ_0 un ensemble de vecteurs de taille $p-1$, θ appartient-il à cet ensemble ? Si Θ_0 vaut S_J l'ensemble des vecteurs de taille $p-1$ dont le support vaut J, la question se reformule ainsi : Y est-il indépendant de X_{J^c} conditionnellement à X_J ? Cette question est reliée au test de graphe d'un modèle graphique (problème **A.2**). En termes statistiques, l'objectif est de construire une procédure de test T de l'hypothèse « $\theta \in \Theta_0$ » aussi puissante que possible.

B.3. Quelle est l'espérance conditionnelle de Y par rapport à X ? On verra l'intérêt d'un tel problème pour résoudre **A.3** en section 1.4. L'objectif ici est de construire un estimateur $\widehat{\theta}$ de θ qui permet de bien prédire la variable Y. Dans cette optique, il est naturel d'utiliser la fonction de perte des moindres carrés $l(.,.)$ pour caractériser la qualité d'un estimateur $\widehat{\theta}$:

$$l(\widehat{\theta}, \theta) := \mathbb{E}\left\{\left[X(\theta - \widehat{\theta})\right]^2\right\} = \mathbb{E}\left\{\left[\mathbb{E}_\theta(Y|X) - \mathbb{E}_{\widehat{\theta}}(Y|X)\right]^2\right\}\ , \tag{1.6}$$

où $\mathbb{E}_{\theta'}(Y|X)$ correspond à l'espérance conditionnelle de Y par rapport à X en utilisant le vecteur θ' dans la régression (1.5). Nous allons chercher à construire des estimateurs $\widehat{\theta}$ tels que le risque $\mathbb{E}[l(\widehat{\theta}, \theta)]$ est aussi petit que possible.

En section 1.2.1, nous montrons en quoi la théorie de la sélection de modèles permet de répondre aux questions **B.1** et **B.3**. Nous présentons ensuite des approches classiques qui permettent de traiter **B.1** et **B.3**. Enfin, nous décrivons rapidement la contribution de cette thèse. Dans la sous-section 1.2.2, nous évoquons la question **B.2** en lien avec la théorie de la décision. On s'attachera entre autres à montrer les similarités dans le traitement des tests d'hypothèses et de la sélection de modèles.

1.2.1 Sélection de modèles

Une méthode que nous allons privilégier dans cette thèse est la sélection de modèles, analysée d'un point de vue non-asymptotique. Nous présentons ici la problématique de la sélection de modèles dans le

cas particulier des modèles de régression linéaire (questions **B.1** et **B.3**). On s'intéressera dans la suite à d'autres cadres de sélection de modèles (en particulier pour l'estimation de covariance), mais les concepts introduits restent sensiblement les mêmes. Pour une présentation plus générale, on pourra lire le cours de Saint-Flour de Massart [Mas07] ou l'introduction de la thèse d'Arlot [Arl07].

1.2.1.1 Oracle, adaptivité, et consistance

Dans la suite, m désigne un sous-ensemble de $\{1,\dots,p-1\}$ qu'on appelle modèle. Nous notons S_m l'espace vectoriel des vecteurs de taille $p-1$ dont le support est inclus dans m.

Soit $\mathcal{M}$ une collection de modèles. Pour chacun de ces modèles $m \in \mathcal{M}$, on construit l'estimateur $\widehat{\theta}_m$ des moindres carrés. L'objectif de la sélection de modèles est de construire une procédure qui choisit un « bon » modèle $\widehat{m}$.

La notion de « bon » modèle et plus généralement la qualité d'une procédure de sélection de modèles se mesure selon deux points de vue : l'*efficacité* et l'*identification* (parfois appelée *consistance*). Dans le premier cas, on vise à obtenir un estimateur $\widehat{\theta}_{\widehat{m}}$ de risque $\mathbb{E}[l(\widehat{\theta}_{\widehat{m}},\theta)]$ petit. Dans le second cas, on suppose que m^*_{vrai}, le support de θ est inclus dans la collection $\mathcal{M}$ et on veut sélectionner ce modèle avec grande probabilité. La question **B.1** correspond à un point de vue identification tandis que **B.3** correspond à un point de vue efficacité.

Avant de préciser les notions qui sous-tendent l'analyse de procédures de sélection de modèles, mentionnons que le choix de la collection $\mathcal{M}$ dépend du cadre d'estimation ou des connaissances a priori. Supposons par exemple que les données sont temporelles. Il est alors légitime de penser qu'il existe un ordre dans les variables, de telle façon que la première est plus susceptible que la seconde d'être pertinente pour prédire Y, et ainsi de suite. Dans un tel cas, on prendra généralement une collection de modèles emboîtés $\{\emptyset,\{1\},\{1,2\},\dots..\}$. On parle alors de *sélection ordonnée*. Inversement, si on n'a aucune idée a priori sur le vrai vecteur θ, un choix raisonnable est la collection $\mathcal{M}^d_{co}$.qui contient tous les sous-ensembles de $\{1,\dots p-1\}$ de taille inféreure à d. Il s'agit d'un problème de *sélection complète*.

Efficacité. Construire une bonne procédure de sélection de modèles en termes d'efficacité revient à dire que le risque $\mathbb{E}_\theta[l(\widehat{\theta}_{\widehat{m}},\theta)]$ de l'estimateur sélectionné est faible. On peut notamment évaluer cette qualité de deux façons : les inégalités oracles et l'adaptativité.
Inégalités oracles : On appelle *oracle* le modèle :

$$m^* \in \arg\min_{m\in\mathcal{M}} \left\{ \mathbb{E}\left[l(\widehat{\theta}_m,\theta)\right]\right\} .$$

Remarquons que l'oracle dépend du vecteur θ, de la covariance Λ, et de la variance σ^2, qui sont tous les trois inconnus. En conséquence, nous n'avons pas accès à ce modèle. L'objectif d'une bonne sélection de modèles de ce point de vue est donc de trouver un modèle $\widehat{m}$ qui se comporte aussi bien (ou presque) que l'oracle. Cela peut se décrire sous la forme d'une *inégalité oracle* (non-asymptotique).

$$\mathbb{E}\left[l(\widehat{\theta}_{\widehat{m}},\theta)\right] \le L \inf_{m\in\mathcal{M}} \mathbb{E}\left[l(\widehat{\theta}_m,\theta)\right] = L\mathbb{E}\left[l(\widehat{\theta}_{m^*},\theta)\right],$$

où L est une constante numérique. En d'autres termes, le risque de l'estimateur sélectionné est comparable à celui de l'oracle. Il n'est pas toujours possible d'obtenir une inégalité oracle, en particulier lorsque la taille de la collection $\mathcal{M}$ est trop grande. Dans ce cas, on s'attachera à montrer des bornes du type :

$$\mathbb{E}\left[l(\widehat{\theta}_{\widehat{m}},\theta)\right] \le L \inf_{m\in\mathcal{M}} \left\{ \mathbb{E}\left[l(\widehat{\theta}_m,\theta)\right] + R(m,n)\right\} ,$$

où $R(m,n)\ge 0$ est un terme de reste aussi petit que possible. Par abus de langage, on appellera parfois ces bornes inégalités de type oracle.

Adaptativité : Une autre qualité recherchée pour un estimateur $\widehat{\theta}_{\widehat{m}}$ obtenu par sélection de modèles est l'adaptativité (voir par exemple Birgé et Massart [BM97]). Nous l'avons évoqué en section 1.1.3, mais nous le décrivons maintenant plus précisément. Supposons que le vrai vecteur θ appartient à une famille $\Theta = \cup_{s\in\mathcal{S}}\Theta_s$, où le paramètre inconnu s_0 tel que θ appartient à Θ_{s_0} représente une propriété de θ. Par exemple, s peut représenter la taille du support de θ. L'estimateur $\widehat{\theta}_{\widehat{m}}$ est dit *adaptatif au paramètre s* s'il n'utilise pas la connaissance de s_0 et reste aussi bon (en un sens défini dans le prochain paragraphe) que tout autre estimateur $\widehat{\theta}_{s_0}$ qui utiliserait la connaissance de s_0. Pour revenir à l'exemple précédent, on

dira que notre estimateur $\widehat{\theta}_{\widehat{m}}$ est adaptatif à la taille du support[5] de θ s'il n'utilise pas la connaissance de cette taille tout en étant aussi bon qu'un estimateur qui « connaît » cette taille a priori.

Dans cette thèse, on évaluera l'adaptativité d'un estimateur au sens du risque *minimax*. Si les inégalités oracles permettent de comparer la perte de l'estimateur $\widehat{\theta}_{\widehat{m}}$ à celle de l'oracle, l'approche minimax permet de se comparer à *tous* les estimateurs possibles. Rappelons la définition du risque minimax. Si on se donne Θ_s une collection de vecteurs, le risque minimax de cette collection est donné par

$$\mathcal{R}_{\min\,\max}(\Theta_s) := \inf_{\widehat{\theta}} \sup_{\theta\in\Theta_s} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] \ ,$$

où l'infimum est pris sur tous les estimateurs. Ce risque minimax mesure l'erreur dans le pire des cas dans la collection Θ_s. Un estimateur est dit minimax sur la collection Θ_s si son risque est uniformément majoré par $L\mathcal{R}_{\min\,\max}(\Theta_s)$ sur toute la collection Θ_s, où L est une constante numérique. Évaluer la qualité d'un estimateur au sens minimax n'est pas toujours pertinent du fait de la nature pessimiste du risque minimax. Considérons par exemple l'ensemble Θ de tous les vecteurs de taille $(p-1)$. On montrera dans le chapitre 4 que pour n assez grand le risque minimax vaut

$$\mathcal{R}_{\min\,\max}(\Theta) = L\sigma^2 \frac{p}{n} \ ,$$

où L est une constante numérique. Un estimateur dont le risque est de l'ordre de $\sigma^2 p/n$ pour tout vecteur θ est minimax sur Θ. Cependant, cet estimateur n'a que peu d'intérêt en pratique. Un bon estimateur doit en effet avoir un risque faible lorsque θ est « simple » à estimer, par exemple lorsque le vecteur θ est creux. On ne va donc pas tant s'intéresser à une propriété minimax en tant que telle qu'à de l'adaptativité au sens minimax : on dit que l'estimateur $\widehat{\theta}_{\widehat{m}}$ est *adaptatif au sens du minimax* si pour tout $s \in \mathcal{S}$,

$$\sup_{\theta\in\Theta_s} \mathbb{E}_\theta \left[l\left(\widehat{\theta}_{\widehat{m}}, \theta\right) \right] \leq L\mathcal{R}_{\min\,\max}(\Theta_s) \ ,$$

pour une constante numérique $L > 0$. En d'autres termes, l'estimateur $\widehat{\theta}_{\widehat{m}}$ atteint simultanément le risque minimax sur toutes les collections Θ_s pour $s \in \mathcal{S}$.

Identification. Si on se place dans le cadre de la question **B.1**, l'objectif est plutôt de déterminer le *vrai* modèle. Dans ce cas, on suppose que $\theta \in \cup_{m\in\mathcal{M}} S_m$ et on note m^*_{vrai} le *vrai modèle*, c'est à dire le modèle m le moins complexe tel que $\theta \in S_m$. En d'autres termes, m^*_{vrai} est le plus petit modèle m tel que le support de θ est inclus dans m. Ici, l'objectif est de construire une procédure de sélection de modèles telle que

$$\mathbb{P}\left(\widehat{m} = m^*_{\text{vrai}}\right) \to 1 \ ,$$

lorsque $n \to +\infty$. Reformulé en termes non-asymptotiques, l'objectif est de maximiser la probabilité $\mathbb{P}(\widehat{m} = m^*_{\text{vrai}})$.

Idéalement, on aimerait construire des procédures de sélection de modèles à la fois consistantes et de risque optimal. Cependant, ces deux objectifs sont parfois incompatibles. Yang [Yan05] a en effet montré dans un cadre légèrement différent du nôtre qu'un estimateur ne peut pas être simultanément minimax et consistant. Nous verrons néanmoins que les critères de prédiction (comme AIC) et de sélection de modèles (comme BIC) sont reliés, et qu'on peut modifier un bon critère de prédiction pour obtenir un bon critère d'identification.

1.2.1.2 Lien avec la régression à design fixe

S'il existe peu de résultats théoriques d'estimation dans le modèle (1.5), il y en a beaucoup plus pour la régression linéaire à *design fixe* à variance connue ou non. Supposons que l'on observe un vecteur $\mathbf{Y}$ de taille n et $\mathbf{X}$ une matrice de taille $n \times (p-1)$ de covariables telles que

$$\mathbf{Y} = \mathbf{X}\theta + \sigma\epsilon \ , \tag{1.7}$$

[5] sparsité en franglais

où le vecteur $\theta \in \mathbb{R}^{p-1}$ est inconnu, le terme de variance σ est positif et ϵ est un vecteur gaussien standard de taille n. La différence principale avec le modèle (1.5) décrit dans la sous-section précédente est que la matrice de design $\mathbf{X}$ est *considérée* comme *fixe*. Ainsi, lorsqu'on voudra évaluer la qualité d'un estimateur $\widehat{\theta}$ de θ, on utilisera la perte des moindres carrés $\|\mathbf{X}(\theta - \widehat{\theta})\|_n$ relative au design $\mathbf{X}$. Ici, $\|.\|_n$ correspond à la norme euclidienne sur $\mathbb{R}^n$ normalisée par $1/\sqrt{n}$. Cette approche ne signifie pas forcément que le plan d'expérience X est déterministe mais que la perte de l'estimateur se mesure par rapport à ce *design*. Dans le cas évoqué en **B.3**, la perte est au contraire intégrée sur la distribution (inconnue) de X. C'est pourquoi on parle dans un cas de régression à *design fixe* et dans l'autre de régression à *design aléatoire*. On distinguera deux cas selon que la variance σ^2 est connue ou non. Dans la suite de cette sous-section, nous rappelons quelques méthodes classiques en régression à design fixe pour l'estimation de θ ou l'estimation de son support.

Sélection de modèle par pénalisation à variance connue : Le principe de la sélection de modèles par pénalisation est de choisir un modèle $\widehat{m}$ qui minimise

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \{\| \mathbf{Y} - \mathbf{X}\widehat{\theta}_m\|_n^2 + \mathrm{pen}(m)\} \ , \tag{1.8}$$

où nous rappelons que $\widehat{\theta}_m$ est l'estimateur des moindres carrés. La fonction $pen : \mathcal{M} \to \mathbb{R}^+$ est appelée une *pénalité*.

Efficacité : Comme mentionné précédemment, l'objectif est de sélectionner un modèle $\widehat{m}$ qui réalise l'infimum des risques $\|\mathbf{X}(\widehat{\theta}_m - \theta)\|_n^2$. L'heuristique dite de *Mallows* est qu'une bonne pénalité doit permettre d'estimer sans biais le risque de l'estimateur $\widehat{\theta}_m$

$$pen_{\mathrm{Mallows}} := \frac{2\sigma^2|m|}{n} = \sigma^2 + \mathbb{E}\left[\|\mathbf{X}(\theta - \widehat{\theta}_m)\|_n^2\right] - \mathbb{E}\left[\|\mathbf{Y} - \mathbf{X}\widehat{\theta}_m\|_n^2\right] \ , \tag{1.9}$$

où $|m|$ désigne la taille du modèle m. Les espérances sont prises par rapport à $\mathbf{Y}$ tandis que $\mathbf{X}$ est considéré comme fixe. La deuxième égalité repose sur le calcul de l'espérance des deux termes en utilisant la définition d'un estimateur des moindres carrés. La formule (1.9) est la différence entre le risque de $\widehat{\theta}_m$ et l'espérance des moindres carrés résiduels pris en $\widehat{\theta}_m$. Le critère obtenu s'appelle la C_p de Mallows [Mal73]. C'est le même type de raisonnement qui a amené Akaike à introduire le critère AIC[6] [Aka73] qui pénalise non pas les moindres carrés résiduels comme en (1.8) mais la log-vraisemblance. De tels critères sont asymptotiquement optimaux en particulier lorsque la taille de la collection $\mathcal{M}$ reste raisonnable[7], comme cela a été prouvé par Shibata [Shi81] et Birgé et Massart [BM01]. En utilisant de façon systématique le phénomène de concentration de la mesure, Birgé et Massart [BM01] ont introduit des pénalités plus générales que celles de Mallows, et obtiennent ainsi un contrôle non-asymptotique du risque de l'estimateur $\widehat{\theta}_{\widehat{m}}$, et ce pour n'importe quelle collection de modèles. Dans le cas particulier de la sélection complète de variables, Massart [Mas07] Sect. 4.2 utilise la collection $\mathcal{M}_{\mathrm{co}}^{(p-1)\wedge n}$ qui contient tous les sous-ensembles de $\{1, \ldots, p-1\}$ de taille plus petite que $(p-1) \wedge n$, et préconise par exemple d'utiliser une pénalité du type :

$$\mathrm{pen}(m) = K\frac{\sigma^2|m|}{n}\left(1 + \sqrt{2 + \log\left(2(p-1)/|m|\right)}\right)^2 \ , \tag{1.10}$$

pour tout $m \in \mathcal{M}_{\mathrm{co}}^{(p-1)\wedge n}$. Il obtient ainsi une borne du type

$$\mathbb{E}\left[\|\mathbf{X}(\widehat{\theta}_{\widehat{m}} - \theta)\|_n^2\right] \leq L(K) \inf_{m \in \mathcal{M}_{\mathrm{co}}^{p-1}} \left\{\|\mathbf{X}(\theta_m - \theta)\|_n^2 + \frac{|m|\sigma^2}{n}\left(1 + \log\left(\frac{p-1}{|m|}\right)\right)\right\} \ . \tag{1.11}$$

Ici, θ_m désigne la projection de θ sur l'espace S_m par rapport à la norme $\|\mathbf{X}(. - .)\|_n$. La constante $L(K)$ ne dépend que de K. Le risque d'un estimateur paramétrique $\widehat{\theta}_m$ est $\|\mathbf{X}(\theta_m - \theta)\|_n^2 + \frac{|m|\sigma^2}{n}$. La borne précédente n'est donc pas une inégalité oracle du fait du terme en $\log((p-1)/|m|)$. Néanmoins, ce terme logarithmique est inévitable lorsqu'on considère un problème de sélection complète. On peut en fait déduire de la borne (1.11) que l'estimateur $\widehat{\theta}_{\widehat{m}}$ est adaptatif au sens minimax à la taille du support de θ (voir par exemple [Mas07] Sect.4.2).

[6] Akaike Information Criterion

[7] C'est le cas pour la sélection ordonnée de variables.

Identification : L'analyse de ces procédures de sélection de modèles du point de vue de l'identification a suscité moins de travaux. On a mentionné que les objectifs d'identification et d'optimalité en termes d'efficacité sont parfois incompatibles. De fait, il est connu que la pénalité C_p de Mallows ne permet pas d'obtenir une procédure consistante en sélection. En revanche la pénalité BIC[8] (Schwarz [Sch78]) qui peut s'écrire

$$pen_{BIC}(m) = \frac{\log(n)\sigma^2|m|}{n}$$

permet la consistance lorsque la collection de modèles n'est pas trop grande[9]. On peut également citer le critère MDL[10] (Rissanen [Ris78]).

Pour une analyse récente de tels critères, nous renvoyons le lecteur à Guyon et Yao [GY99]. Ces résultats suggèrent que surpénaliser est une bonne façon d'obtenir une procédure consistante. Dans le cas de la sélection complète, Abramovich *et al.* [ABDJ06] ont montré qu'une pénalité proche de (1.10) permettait de contrôler le FDR[11] lorsque le design est orthogonal. Ainsi, la procédure obtenue n'est pas consistante en sélection sauf si on se place dans l'asymptotique $p \to \infty$, $n \to \infty$ et la taille du support de θ est en $o(p)$. Toutefois, cette asymptotique est pertinente lorsque l'on considère de l'estimation en « grande dimension », i.e. p grand.

Avantages et inconvénients : La méthode de pénalisation introduite précédemment permet de construire des pénalités pour toute collection de modèles $\mathcal{M}$ initiale. Les procédures ainsi obtenues vérifient des inégalités oracles (lorsque c'est possible) non-asymptotiques. Pour certaines collections de modèles, il est possible de montrer des propriétés d'adaptativité au sens minimax (en particulier pour la sélection complète). En outre, ces pénalités peuvent être calibrées explicitement (tout au moins lorsque la variance σ^2 est connue). En ce qui concerne la consistance en termes de sélection, l'idée générale est de surpénaliser. Le défaut principal de ce type de procédure réside dans son coût computationnel. En effet, à moins que le design $\mathbf{X}$ soit orthogonal le temps de calcul du modèle $\widehat{m}$ est proportionnel à la taille de la collection $\mathcal{M}$. Si pour la sélection ordonnée le temps de calcul reste très faible, ce n'est pas le cas en sélection complète où la taille de la collection de modèles est telle que la procédure n'est applicable que pour p inférieur à 30.

Pénalisation à variance inconnue : Lorsque le terme de variance σ^2 est inconnu, on ne peut plus directement utiliser les procédures mentionnées précédemment, car les pénalités dépendent directement de σ^2. Une méthode souvent utilisée en pratique consiste à remplacer σ^2 par un estimateur $\widehat{\sigma}^2$. Cependant, l'estimation de σ^2 n'est pas évidente puisqu'on ne connaît pas a priori un bon modèle pour le vecteur θ. Birgé et Massart [BM07] proposent une méthode appelée *heuristique de pente* pour calibrer la pénalité dans une telle situation. Les critères AIC et BIC basés sur la vraisemblance pénalisée s'étendent lorsque la variance est inconnue. Plus récemment, Baraud, Giraud et Huet [BGH08] ont mené une étude générale des critères du type

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}} \left\{ \| \mathbf{Y} - \mathbf{X}\widehat{\theta}_m \|_n^2 \left(1 + \mathrm{pen}(m)\right)\right\} . \tag{1.12}$$

Remarquons que la pénalité est cette fois-ci multiplicative car $\|\mathbf{Y} - \mathbf{X}\widehat{\theta}_m\|_n^2$ fait office d'estimateur de la variance. Baraud *et al.* observent qu'avec de bons choix de pénalités on peut exprimer les critères classiques FPE[12] (Akaike [Aka70]), AIC et BIC sous la forme (1.12). Par ailleurs, ils étudient d'un point de vue non-asymptotique l'efficacité de chacun de ces critères en fonction de la taille de la collection de modèles. En particulier, ils prouvent que ces trois critères donnent de très mauvais résultats en sélection complète. Suite à cela, ils introduisent de nouvelles pénalités basées sur des quantiles de variables de Fisher et obtiennent des bornes de type oracle et de l'adaptativité au sens minimax sous des hypothèses légères sur la collection de modèles.

Finalement, la méthode générale de pénalisation de Baraud *et al.* présente les mêmes avantages et inconvénients que la procédure précédente hormis que l'on peut considérer le cas σ^2 inconnu.

[8] Bayesian Information Criterion
[9] Comme par exemple en sélection ordonnée.
[10] Minimum description length
[11] False discovery rate [BH95].
[12] Final Prediction error

Lasso : L'inflation de la taille des données à analyser dans de nombreux domaines a motivé le développement de procédures statistiques rapides et facilement implémentables. Parmi elles, la méthode des moindes carrés pénalisés par la norme l_1, baptisée lasso par Tibshirani [Tib96], est certainement l'une des plus populaires. Soit λ un paramètre positif dit de régularisation, l'estimateur $\widehat{\theta}^\lambda$ est défini dans notre cadre de régression par

$$\widehat{\theta}^\lambda := \arg \inf_{\theta \in \mathbb{R}^{p-1}} \|\mathbf{Y} - \sum_{i=1}^{p} \theta_i \mathbf{X}_i\|_n^2 + \lambda \|\theta\|_1 \ , \tag{1.13}$$

où $\|\theta\|_1$ désigne la norme de l_1 de θ. Comme le critère à minimiser est convexe en θ, on peut utiliser une procédure d'optimisation convexe pour calculer $\widehat{\theta}^\lambda$. Mieux, Hastie *et al.* [EHJT04] ont introduit l'algorithme LARS qui permet en un temps polynomial en n et p d'obtenir simultanément *tous* les estimateurs $\widehat{\theta}^\lambda$ avec $\lambda > 0$. Idéalement, le lasso permet à la fois d'obtenir un bon estimateur $\widehat{\theta}^\lambda$ de θ et de faire de la sélection de variables en sélectionnant le sous-ensemble $\widehat{m}^\lambda$ des variables qui correspond au support de $\widehat{\theta}^\lambda$.

Pour avoir un premier aperçu du comportement de l'estimateur lasso, supposons un instant que $p = n$ et que le design $\mathbf{X}$ vaut l'identité. Dans ce cas, l'estimateur lasso est équivalent à la méthode dite du *seuillage doux* :

$$\widehat{\theta}_i^\lambda = \begin{cases} Y_i - \frac{\lambda}{2} & \text{si} \quad Y_i \geq \frac{\lambda}{2} \ , \\ 0 & \text{si} \quad |Y_i| < \frac{\lambda}{2} \ , \\ Y_i + \frac{\lambda}{2} & \text{si} \quad Y_i \leq -\frac{\lambda}{2} \ . \end{cases}$$

Le seuillage doux permet de garder les plus grands coefficients Y_i et de réduire les autres à zéro. Contrairement au seuillage dur, l'estimateur est maintenant rétréci vers 0 de façon continue. Les propriétés d'adaptativité de cette méthode de seuillage ont été mises en évidence par Donoho et Johnstone [DJ94] dans le cadre de l'estimation fonctionnelle.

Ces dernières années, un grand nombre de papiers ont été consacrés au comportement de l'estimateur lasso lorsque le design n'est pas orthogonal (i.e. $\mathbf{X}^*\mathbf{X}$ n'est pas diagonal). La majorité de ces études tentent de répondre à l'une des deux questions suivantes :

1. Sous quelles conditions sur le design $\mathbf{X}$, l'estimateur $\widehat{\theta}^\lambda$ a-t-il des propriétés d'adaptativité au sens minimax ? [BTW07a, BRT08, CP08]
2. Sous quelles conditions sur le design $\mathbf{X}$, le modèle selectionné est-il consistent ? [MB06, ZY06, CP08]

Le cadre restreint de cette introduction ne nous permet pas de détailler ces résultats. Mentionnons néanmoins les résultats les plus significatifs. En particulier, Bickel *et al.* [BRT08] montrent que sous une condition sur les valeurs propres restreintes[13] du design (hypothèse $RE(s, c_0)$ dans leur article), le lasso atteint la vitesse minimax adaptative pour tout vecteur θ assez creux. Zhao and Yu [ZY06] ont quant à eux introduit la condition « irrepresentable condition » nécessaire et *presque* suffisante pour que le lasso soit consistant en sélection de modèles. Candès et Plan [CP08] montrent que si les corrélations (empiriques) entre les covariables sont inférieures à $1/\log p$, alors le lasso atteint la vitesse d'estimation optimale et est consistant en sélection de modèles dans une « majorité » de cas. De plus, Candès et Plan illustrent que leur condition sur la corrélation est presque nécessaire. Signalons que les procédures d'estimation par pénalisation l_1 ne se limitent pas aux modèles de régression linéaire, et que leur étude peut être étendue entre autres aux modèles linéaires généralisés [vdG08].

Avantages et inconvénients du lasso : L'estimateur lasso peut être calculé en un temps raisonnable même lorsque n et p sont grands. Il est consistant en sélection de variables et atteint des vitesses optimales d'estimation au sens de la perte $\|.\|_n$ sous des hypothèses précédemment citées. Cependant, cette méthode a trois inconvénients majeurs : premièrement, le lasso ne se comporte bien que sous certaines hypothèses sur le design ; deuxièmement, il considère uniquement la sélection complète de variables et ne permet pas d'intégrer une connaissance a priori sur la cible ; troisièmement, le choix du paramètre λ peut se révéler problématique. Si σ^2 est connu, Candès et Plan propose de prendre $\lambda = 2\sqrt{2\log(p-1)}\sigma$. Cependant, il n'existe pas encore de solution satisfaisante pour le choix de λ lorsque la variance σ^2 est inconnue. Les praticiens utilisent généralement une procédure de cross-validation pour sélectionner un bon

[13] Ce sont les valeurs des restrictions des sous-matrices de taille s d'une matrice donnée.

λ. Meinshausen et Bühlmann [MB06] ne préconisent pas cette méthode lorsque l'objectif est la sélection consistante de variables, car l'estimateur obtenu sélectionne généralement trop de variables.

Pour finir, mentionnons qu'il existe de nombreuses variantes du lasso qui permettent d'obtenir des résultats de consistance ou d'efficacité sous des hypothèses un peu moins restrictives. Citons par exemple l'adaptive lasso de Zou [Zou06] ou le bolasso de Bach [Bac08]. Cependant, ces procédures souffrent peu ou prou du même type d'inconvénients qu'énoncés précédemment.

Dantzig selector : Le Dantzig selector est une méthode d'estimation des modèles linéaires en grande dimension introduite par Candès et Tao [CT07]. Nous ne décrivons pas ici la définition de l'estimateur mais mentionnons que ses propriétés théoriques sont proches du lasso [BRT08].

Agrégation : Une méthode concurrente de la sélection de modèles est l'*agrégation* qui consiste non pas à choisir un estimateur parmi une famille comme on le fait en sélection de modèles mais à considérer des combinaisons d'estimateurs. Ce type de procédure n'est intéressante que dans un objectif d'efficacité. Mentionnons les résultats de Bunea *et al.* [BTW07b] et de Dalalyan et Tsybakov [DT08] dans un cadre de régression gaussienne assez général et ceux de Leung et Barron [LB06] en régression a design fixe. Giraud [Gir08b] considère des procédures analogues dans le cas où la variance σ^2 est inconnue. Nous ne reviendrons pas sur les méthodes d'agrégation dans la suite car il n'existe pas, à notre connaissance, de résultats dans les cadres étudiés dans cette thèse.

1.2.1.3 Retour au design aléatoire gaussien et contribution de la thèse

Dans la sous-section précédente, nous avons présenté quelques procédures d'estimation et de sélection de modèles dans le cadre de la régression linéaire avec un design considéré comme fixe (1.7). Ces méthodes sont de fait des bons candidats pour répondre aux questions **B.1** et **B.3** lorsque le design est gaussien. Nous n'avons pas mentionné jusqu'ici les procédure d'estimation et de sélection de modèles dans le cadre de la régression à design aléatoire (mais pas gaussien). En effet, la plupart des méthodes introduites spécifiquement dans ce cadre s'étendent difficilement à notre problème. Citons néanmoins les travaux de Baraud [Bar02a] et de Tsybakov [Tsy03] qui adoptent des points de vue assez proches du nôtre.

Évoquons maintenant l'utilisation du lasso dans le cadre du design gaussien. Meinshausen et Bühlmann [MB06] ont donné des conditions suffisantes sur le vecteur θ et la matrice Λ, pour que l'estimateur $\widehat{\theta}^{\lambda}$ construit avec un *bon* paramètre λ soit consistant en sélection de modèles. De plus, ils proposent un choix explicite du paramètre λ (mais qui n'est pas exactement celui de la théorie). À notre connaissance, il s'agit de l'unique résultat dans le cadre (1.7). Néanmoins, il est raisonnable de penser que l'estimateur lasso se comporte bien (en terme du risque), lorsque la matrice de covariance Λ est telle qu'avec grande probabilité le design empirique $\mathbf{X}$ vérifie une des conditions qui permettent au lasso de converger à vitesse optimale pour le modèle (1.7). En conséquence, l'estimateur lasso présente *a priori* les mêmes avantages et inconvénients que ceux énoncés précédemment.

Contribution de la thèse. Dans le chapitre 4 de cette thèse, nous étudions dans le cadre (1.5) une procédure de sélection de modèles proche de la méthode (1.12) introduite dans [BGH08]. Nous obtenons ainsi des inégalités non-asymptotiques de type oracle sans aucune hypothèse sur Λ, θ ou σ. La généralité de la procédure nous permet, selon le choix de la collection de modèles, de considérer de la sélection ordonnée, de la sélection complète, ou encore d'autres procédures. Dans le cas de la sélection ordonnée, nous obtenons également une inégalité oracle asympotique avec constante optimale. Par ailleurs, nous exhibons les vitesses minimax d'estimation en sélection ordonnée selon la décroissance du biais avec la dimension. Nous en déduisons que notre procédure pour l'estimation ordonnée est minimax adaptative à cette décroissance. En sélection complète, la procédure est minimax adaptative à la taille du support.

En ce qui concerne l'estimation du support de θ (problème **B.1**), nous ne donnons pas de résultat théorique. Néanmoins, les outils de concentration introduits permettent vraisemblablement de mener un raisonnement analogue à celui de Guyon et Yao [GY99]. Pour la sélection complète, les simulations numériques suggèrent que notre procèdure donne un bon estimateur du support de θ, au moins lorsque p est plus grand que n, et ce quelle que soit la matrice de covariance Λ.

En résumé, notre procédure souffre des mêmes avantages et inconvénients que les méthodes de pénalisation de Birgé et Massart [BM01] et Baraud *et al.* [BGH08] en design fixe. Ainsi la méthode est

très flexible et permet d'intégrer des connaissances à priori sur θ. On peut également traiter d'autres problèmes que la sélection ordonnée ou la sélection complète de variables. De plus, la procédure présente des propriétés non-asymptotiques d'oracle ou d'adaptativité au sens minimax, même lorsque le nombre de covariables p est plus élevé que n. Enfin, les pénalités introduites ne dépendent pas de paramètres inconnus. Cependant, le coût computationnel de la procédure est proportionnel à la taille de la collection $\mathcal{M}$, considérée et est donc très élevé en sélection complète.

1.2.2 Test d'hypothèse

L'objet de la théorie statistique des tests est de répondre à une question binaire telle que : « Y a-t-il ou non un lien significatif entre ces deux évènements ? ». Dans cette sous-section, nous redéfinissons les enjeux et quelques méthodes de la théorie des tests d'hypothèse dans le cas particulier du problème **B.2** qui nous intéresse. On se pose la question suivante. La distribution qui a généré les données ($\mathbf{Y}$,$\mathbf{X}$) vérifie-t-elle la propriété : « le support de θ est inclus dans Θ_0 » ?

Pour réaliser un test on formule deux hypothèses :

- une *hypothèse nulle* H_0. Elle correspond généralement au cas où il ne se passe rien de nouveau. Dans notre cas, l'hypothèse H_0 est : « le vecteur θ appartient à Θ_0 ». En pratique, l'ensemble Θ_0 sera souvent de la forme S_J où J est un sous-ensemble $\{1, \dots, p-1\}$. En d'autres termes, Θ_0 est l'ensemble des vecteurs de taille $p-1$ dont le support est inclus dans J.
- une *hypothèse alternative* H_1. Elle correspond à une découverte que l'on cherche à établir. Dans notre cas, on choisira différentes hypothèses alternatives selon le contexte. On peut par exemple choisir l'hypothèse alternative générale : « $\theta \in \mathbb{R}^{p-1} \setminus \Theta_0$ ». Si l'ensemble Θ_0 s'écrit sous la forme S_J, on peut également prendre comme hypothèse alternative « le support de θ contient entre 1 et k indices qui ne sont pas dans J ». Cette hypothèse correspond au cas où on suppose que le support de θ est parcimonieux. Dans la suite de ce chapitre, l'ensemble des vecteurs θ correspondant à l'hypothèse alternative est noté Θ_1.

Le choix du test dépend du choix de l'hypothèse nulle ainsi que de l'hypothèse alternative. Un *test statistique* est ici une application mesurable $T : \mathbb{R}^{np} \to \{0, 1\}$. Lorsque $T(\mathbf{Y}, \mathbf{X}) = 1$, on dit que *l'hypothèse nulle est rejetée*. Lorsque $T(\mathbf{Y}, \mathbf{X}) = 0$, on dit que *l'hypothèse nulle est acceptée*. On peut faire deux types d'erreur dans un test :

- Lorsqu'on rejette l'hypothèse nulle à tort, i.e. lorsque $T(\mathbf{Y}, \mathbf{X}) = 1$ alors que $\theta \in \Theta_0$, on parle d'erreur de *première espèce*.
- Lorsqu'on accepte l'hypothèse H_0 à tort (alors que H_1 est correcte), i.e. $T(\mathbf{Y}, \mathbf{X}) = 0$ alors que $\theta \in \Theta_1$, on parle d'erreur de *seconde espèce*.

Une approche communément adoptée consiste à contrôler la probabilité de l'erreur de première espèce :

$$\forall \theta \in \Theta_0 \ , \qquad \mathbb{P}_\theta \left[T(\mathbf{Y}, \mathbf{X}) = 1 \right] \leq \alpha \ , \tag{1.14}$$

pour un nombre $\alpha \in (0, 1)$. On appelle α le niveau du test. On cherche alors à choisir un test T,qui « minimise » la probabilité d'erreur de seconde espèce, parmi ceux qui satisfont (1.14). L'objectif est donc en premier lieu de garantir que l'erreur de première espèce est petite, i.e. qu'un rejet du test est valide.

La probabilité de l'erreur de seconde espèce dépend de $\theta \in \Theta_1$, de Λ et de σ^2. De même qu'en estimation il n'y a pas une unique façon d'envisager la minimisation du risque et la notion de « bon » estimateur, différentes approches permettent de considérer la minimisation de l'erreur.

Historiquement, la première d'entre elles revient à réaliser un contrôle uniforme de l'erreur de seconde espèce :

$$\forall \theta \in H_1 \ , \mathbb{P}_\theta \left[T(\mathbf{Y}, \mathbf{X}) = 0 \right] \leq 1 - \beta \ . \tag{1.15}$$

Lorsque (1.15) est réalisée, on dit que le test T est de puissance $\beta \in (0, 1)$. L'objectif est alors de construire un test T de niveau α prescrit et de puissance β maximale. Cependant, dans certains cas ce contrôle uniforme n'est pas possible ou non pertinent. Considérons par exemple le test d'hypothèse : $H_0 : \theta = 0$ contre $H_1 : \theta \neq 0$. Pour tout test T de niveau α, la puissance est inférieure à α. En effet, lorsque θ tend

vers 0, la distribution du test T tend au sens de la convergence faible vers sa loi sous l'hypothèse $\theta = 0$. Le contrôle uniforme de l'erreur ne permet donc pas de choisir une « bonne » procédure dans ce cas. Il faut donc utiliser d'autres critères pour comparer les procédures de test.

Test uniformément plus puissant : Dans certains cas, il est possible de définir un test uniformément plus puissant, i.e. un test T de niveau α tel que pour tout test T' de niveau α,

$$\forall \theta \in \Theta_1, \qquad \mathbb{P}_\theta(T = 1) \geq \mathbb{P}_\theta(T' = 1) \ .$$

Ainsi, la probabilité de l'erreur de seconde espèce est plus petite pour le test T, et ce uniformément pour tous les paramètres θ dans Θ_1. L'existence de tels tests est liée à la théorie de Neyman-Pearson (voir par exemple le livre de Lehman [Leh86]). Cependant, leur théorie ne s'applique pas dans de nombreuses situations (dont la nôtre), et il n'existe pas de test uniformément plus puissant.

Approche minimax et adaptativité : Pour simplifier les notations, nous supposons maintenant que $\Theta_0 = \{0_{p-1}\}$. Soient α un nombre appartenant à $]0;1[$ et β appartenant à $]0;1-\alpha[$. Typiquement, on choisit β petit. Soit T_α un test de niveau α de l'hypothèse « $\theta = 0$ » contre l'hypothèse « $\theta \in \Theta_1$ ». On peut mesurer la performance du test T_α en utilisant la quantité $\rho\left(\psi_\alpha, \Theta_1, \delta, \mathrm{var}(Y), \Sigma\right)$ définie par

$$\rho\left(T_\alpha, \Theta_1, \delta, \mathrm{var}(Y), \Lambda\right) \quad := \quad \inf\left\{\rho > 0, \ \inf\left\{\mathbb{P}_\theta(T_\alpha = 1), \ \theta \in \Theta_1 \text{ et } r_{s/n}(\theta) \geq \rho^2\right\} \geq 1 - \delta\right\},$$

où la quantité

$$r_{s/n}(\theta) := \frac{l(\theta, 0_{p-1})}{\mathrm{var}(Y) - l(\theta, 0_{p-1})} \tag{1.16}$$

correspond à $l(\theta, 0_{p-1})/\mathrm{var}(Y|X)$. De manière intuitive, on peut voir $r_{s/n}(\theta)$ comme la quantité d'information apportée par la variable X (i.e. le signal) divisée par la variance conditionnelle de Y (i.e. le bruit). Plus le rapport signal sur bruit est important, plus il devrait être « facile » de rejeter l'hypothèse H_0. Inversement, si le rapport signal/bruit est très faible, il est plus difficile de rejeter H_0 avec grande probabilité. En ce sens, la fonction $\rho(.)$ mesure la quantité $r_{s/n}(\theta)$ minimale pour que la probabilité de seconde espèce reste inférieure à δ. Pour attester de la qualité d'un test T_α, nous allons comparer la quantité $\rho\left(T_\alpha, \Theta_1, \delta, \mathrm{var}(Y), \Lambda\right)$ à

$$\rho_{\mathrm{minmax}}\left(\Theta_1, \alpha, \delta, \mathrm{var}(Y), \Lambda\right) := \inf_{T_\alpha} \rho\left(T_\alpha, \Theta_1, \delta, \mathrm{var}(Y), \Lambda\right) \ , \tag{1.17}$$

où l'infimum est pris sur tous les tests T_α de niveau α. Nous appelons la quantité $\rho_{\mathrm{minmax}}(.)$ la vitesse (α, δ)-minimax de test sur Θ_1[14].

Les vitesses minimax de tests ont été largement étudiées dans le cadre de la régression gaussienne à design fixe et orthogonal ou du bruit blanc gaussien. Dans ses trois articles fondateurs, Ingster [Ing93a, Ing93b, Ing93c] calcule des vitesses minimax asymptotiques de tests contre des alternatives non-paramètriques telles que des ellipsoïdes ou des espaces de fonctions höldériennes[15]. Spokoiny [Spo96] a montré que contrairement à l'estimation, il n'est pas possible d'obtenir une procédure adaptative au sens minimax à toute une collection d'ellipsoïdes (à moins de perdre un facteur logarithmique). Plus récemment, Baraud [Bar02b] a prouvé des résultats analogues dans un cadre non-asymptotique. De plus, il a calculé les vitesses minimax pour l'alternative : « le support de θ est de taille plus petite que k », où k est un entier entre 1 et n.

Dans ce même cadre de régression à design fixe et orthogonal mais à variance σ^2 inconnue, Baraud, Huet et Laurent [BHL03] ont introduit une procédure de test d'hypothèse linéaire basée sur des idées de sélection de modèles proches de celles développées précédemment. En utilisant des outils de concentration de la mesure, ils calculent le niveau et la probabilité de l'erreur de seconde espèce d'un point de vue non-asymptotique. Ils montrent ainsi que leur procédure atteint la vitesse minimax (à un facteur logarithmique près) de tests pour toute une collection d'ellipsoïdes. On trouvera dans leur article les

[14] On parle aussi parfois de vitesse minimax de séparation.
[15] Voir [Bar02b] pour une explication plus précise de ces résultats.

références à d'autres travaux adoptant le même point de vue.

Contribution de la thèse. Le chapitre 2 est le fruit d'une collaboration avec Fanny Villers. Nous proposons une procédure générale pour tester l'hypothèse $\theta \in S_J$ où J est un sous-ensemble de $\{1, \ldots, p-1\}$. Nous caractérisons de façon non-asymptotique ses probabilités d'erreur de première et seconde espèces. Puis, nous spécifions notre procédure selon l'alternative étudiée. Lorsque l'alternative est : « le vecteur θ contient au plus k composantes non-nulles en dehors de J », nous montrons que notre procédure atteint la vitesse minimax de test. De plus, cette vitesse minimax est elle-même identique à celle calculée par Baraud [Bar02b] en design fixe. Lorsque les covariables sont naturellement ordonnées (comme c'est le cas pour des données temporelles), nous introduisons deux procédures de tests qui atteignent la vitesse minimax (à facteur $\log(p)$ près) sur des ensembles analogues aux ellipsoïdes du design fixe.

Le coût computationnel de la procédure dépend de l'hypothèse alternative considérée. Mentionnons néanmoins qu'il reste faible même pour p grand lorsque les covariables sont supposées ordonnées ou lorsque l'hypothèse alternative est : « le support de θ contient exactement un élément qui n'est pas dans J. »

1.3 Estimation et tests de graphes

Revenons maintenant aux problèmes d'estimation et de tests de graphes pour des modèles graphiques (questions **A.1** et **A.2**). Nous allons d'abord considérer le cas des modèles graphiques non orientés puis le cas orientés. À chaque fois, nous présentons les méthodes classiques d'estimation et de test, puis nous détaillons la contribution de la thèse.

1.3.1 Estimation et tests de graphes non orientés

La majorité des méthodes d'estimations du graphe $\mathcal{G} = (\Gamma, E)$ sont basées sur l'une des approches suivantes. La question **A.2** du test d'adéquation d'un graphe a été peu traitée dans la littérature. Néanmoins, certaines méthodes d'estimation de graphe peuvent être reformulé sous la forme d'un test d'adéquation. Nous le préciserons lorsque c'est le cas.

Estimation bayésienne. Historiquement, les méthodes bayésiennes d'estimation et de sélection de modèles pour des modèles graphiques ont suscité un nombre important de recherches. En effet, le formalisme bayésien peut être reformulé dans le cadre des modèles graphiques[16]. Les procédures bayésiennes en modélisation graphique sont basées sur un principe commun : on part d'une mesure a priori $\pi(\mathcal{G}, \Omega)$ sur le graphe $\mathcal{G}$ du modèle graphique gaussien et sur la matrice de précision Ω, et on veut obtenir la distribution a posteriori $\pi(\mathcal{G}, \Omega|\mathbf{Z})$. Généralement, la mesure $\pi(\mathcal{G}, \Omega)$ satisfait la décomposition $\pi(\mathcal{G}, \Omega) = \pi(\Omega|\mathcal{G})\pi(\mathcal{G})$. En d'autres termes, on spécifie une loi a priori sur la matrice Ω pour chaque graphe $\mathcal{G}$, puis on spécifie une loi a priori à valeurs dans l'espace des graphes. Dans ce cas, la distribution a posteriori se décompose en $\pi(\mathcal{G}, \Omega|\mathbf{Z}) = \pi(\Omega|\mathbf{Z}, \mathcal{G})\pi(\mathcal{G}|\mathbf{Z})$. La construction de loi a priori $\pi(\Omega|\mathcal{G})$ satisfaisante n'est pas toujours chose aisée. Lorsque le graphe $\mathcal{G}$ est décomposable (voir section 1.1.1.3), Dawid et Lauritzen [DL93] ont introduit la distribution Hyper-Inverse Wishart. Celle-ci présente le double avantage d'être facilement simulable et surtout d'être conjuguée, i.e. on connaît une expression analytique de la distribution a posteriori $\pi(\Omega|\mathcal{G}, \mathbf{Z})$. Plus récemment, Roverato [Rov02] a défini des mesures Hyper-Inverse Wishart pour des graphes non décomposables. En revanche, ce nouvel a priori n'étant pas conjugué, le calcul de la loi a posteriori nécessite l'utilisation de méthode de type MCMC[17]. Letac et Massam [LM07] ont également généralisé les lois Hyper-Inverse Wishart et introduit de nouvelles distributions a priori pour les matrices de Wishart. La loi a priori sur l'espace des graphes est choisie pour charger uniformément tous les graphes, pour charger les graphes parcimonieux, ou alors pour charger uniquement les graphes décomposables. Dans le cas décomposable, Giudici et Green [GG99] ainsi que Wong et Carter [WCK03] ont développé des méthodes MCMC pour calculer le maximum a posteriori. Dellaportas *et al.* [DGR03] ont également introduit une procédure pour calculer le maximum a posteriori pour des lois a priori sur des graphes potentiellement non-décomposables.

[16] En particulier, il est possible de voir les modèles bayésiens hiérarchique comme un exemple de modèle graphique gaussien orienté.

[17] Markov Chain Monte Carlo.

L'avantage des méthodes bayésiennes est leur grande flexibilité puisqu'on peut facilement intégrer n'importe quelle connaissance a priori sur le graphe. Cependant, elles présentent un coût computationnel élevé. De fait, la taille des graphes considérés dans la littérature reste généralement faible (< 50). Considérer uniquement des graphes décomposables permet de réduire légérement le temps de calcul. Cependant, une telle procédure n'est plus consistante.

Estimation de la matrice de précision et tests multiples : Le principe général de ces méthodes est de construire un estimateur $\widehat{\Omega}$ de la matrice de précision, puis de sélectionner le graphe en utilisant une procédure de tests multiples basée sur les entrées de la matrice $\widehat{\Omega}$. Ainsi, Drton et Perlman [DP04] définissent $\widehat{\Omega}$ comme l'inverse de la matrice de covariance empirique. Ils proposent ensuite une procédures de tests multiple pour tester $\Omega[i,j] = 0$ pour tout $i \neq j$. Cette hypothèse est rejetée lorsque la statistique de test $\frac{\widehat{\Omega}[i,j]}{\sqrt{\widehat{\Omega}[i,i]\widehat{\Omega}[j,j]}}$ est « grande » en valeur absolue. Nous avons énoncé en Section 1.3.1 la bijection entre le graphe minimal $\mathcal{G}$ et les 0 de la matrice Ω. Drton et Perlman prouvent un contrôle asymptotique du FWER[18] qui correspond ici à la probabilité d'inclure une arête à tort. Cette procédure d'estimation est asymptotiquement consistante. De plus, leur méthode permet d'intégrer des connaissances a priori sur le graphe. On peut également l'étendre pour faire un test d'adéquation à un graphe donné (problème **A.2**). Cependant, leur procédure ne peut s'appliquer lorsque p est plus grand que n (pour inverser la matrice de covariance empirique). Pour résoudre ce défaut, Schäfer et Strimmer proposent d'appliquer une méthode de Bagging (Bootstrap aggregation voir Breiman [Bre96]) puis d'utiliser une pseudo-inversion pour mieux estimer Ω. Si leur procédure s'applique pour p et n arbitraires, il semble à la lueur des simulations réalisées par Villers *et al.* [VSBH08] que ses performances ne soient pas très concluantes.

Approximation du graphe : Wille et Bühlmann [WB06] proposent d'estimer le graphe d'indépendance du premier ordre noté $\mathcal{G}_{01}$ au lieu du graphe $\mathcal{G}$ du modèle graphique gaussien. Le graphe $\mathcal{G}_{01}$ est défini ainsi : Pour tout couple de sommets (a,b) dans Γ, on note $R_{ab\setminus\emptyset}$ la corrélation entre les variables Z_a et Z_b et pour tout $c \in \Gamma \setminus \{a,b\}$, on note $R_{ab\setminus c}$ la corrélation de Z_a et Z_b conditionnellement à Z_c. Il existe une arête entre a et b dans le graphe $\mathcal{G}_{01}$ si et seulement si $R_{ab\setminus\emptyset} \neq 0$ et $R_{ab\setminus c} \neq 0$ pour tout noeud $c \in \Gamma \setminus \{a,b\}$. Wille et Bühlmann proposent alors une procédure de tests multiples des hypothèses $\min_{c\in\Gamma\setminus\{a,b\}\cup\emptyset} |R_{ab\setminus c}| = 0$ pour tous les couples (a,b) dans Γ. Leur méthode sélectionne de façon consistante le graphe $\mathcal{G}_{01}$ dans l'asymptotique où $\log(p)/n$ tend vers 0. Cette procédure a plusieurs avantages : elle a un coût computationnel faible et s'applique même lorsque p est grand devant n. De plus, on peut éventuellement intégrer des connaissances a priori sur le graphe ou s'en servir pour construire une procédure de test d'adéquation (problème **A.2**). Cependant, la procédure estime le graphe $\mathcal{G}_{01}$ et non pas le graphe minimal du modèle graphique $\mathcal{G}$. Si $\mathcal{G}$ est acyclique, on a l'égalité $\mathcal{G} = \mathcal{G}_{01}$. Sous une hypothèse dite de « faithfullness » (voir [WB06]), Wille et Bühlmann montrent que $\mathcal{G} \subset \mathcal{G}_{01}$. Mais dans le cas général, ces deux graphes sont différents.

Kalisch et Bühlmann [KB07] ont introduit une procédure d'estimation du graphe basée sur une variante de la méthode *PC-Algorithm* (Spirtes *et al.* [SGS00]). Contrairement à la procédure précédente, on ne considère plus seulement les dépendances conditionnelles d'ordre 1, mais également les dépendances conditionnelles jusqu'à un ordre $\widehat{k}$ sélectionné par l'algorithme. Kalisch et Bühlmann prouvent que leur procédure est consistante pour la sélection de graphe dans une asymptotique où $p \gg n \gg deg(\mathcal{G})$ si la distribution Z satisfait l'hypothèse de « faithfulness » évoquée précédemment. La notation $deg(\mathcal{G})$ désigne le degré du graphe (i.e le nombre maximum de voisins). Le coût computationnel de la procédure est faible si le graphe sous-jacent est parcimonieux. Par contre, la procédure dépend d'un paramètre α dont le choix pratique reste un problème ouvert.

Estimation pénalisée de la matrice de précision : Récemment, plusieurs auteurs (Yuan et Lin [YL07], Barnerjee *et al.* [BEGd08], Friedman *et al.* [FHT08], Rothman *et al.* [RBLZ08]) ont proposé d'estimer la matrice de précision Ω en résolvant un problème de maximum de vraisemblance associé à une pénalité l_1 sur les entrées de la matrice de covariance. Plus précisément Banerjee *et al.* [BEGd08],

[18] Family Wise Error Rate. C'est la probabilité de faire une erreur de type 1 sur au moins un test.

proposent d'estimer Ω en minimisant le critère suivant :

$$\log(det(\Omega')) + tr(\overline{\mathbf{Z}^*\mathbf{Z}}\Omega') + \lambda \sum_{i=1}^{p}\sum_{j=1}^{p} |\Omega'[i,j]|$$

L'estimateur obtenu est noté $\widehat{\Omega}^\lambda$ et la procédure est appelée le *glasso*. L'estimateur $\widehat{G}^\lambda$ du graphe est construit en prenant pour arête les éléments non nul de $\widehat{\Omega}^\lambda$. L'algorithme proposé par Friedman permet de calculer rapidement $\widehat{\Omega}^\lambda$ et ce même pour p grand. Si Banerjee *et al.* contrôlent la probabilité de connecter (à tort) deux composantes connexes distrinctes du graphe, il n'existe pas à notre connaissance de résultat de consistance. Nous reparlerons de cette procédure d'estimation en Section 1.4.

Estimation et régression conditionnelle : Les deux méthodes que nous allons décrire reposent sur le lien entre indépendance conditionnelle et régression conditionnelle. Pour tout sommet a dans Γ, nous rappelons que la régression de Z_a conditionnellement à $Z_{\Gamma\setminus\{a\}}$ s'écrit

$$Z_a = \sum_{b\neq a} \theta_a[b] Z_b + \epsilon_a \tag{1.18}$$

où $\theta_a[b] = -\Omega[a,b]/\Omega[a,a]$ et ϵ_a est une variable gaussienne centrée indépendante des $(Z_b)_{b\neq a}$. Ainsi, il est équivalent d'estimer le voisinage de a et le support du vecteur (θ_a). On se ramène donc au problème **B.2** étudié dans la partie précédente.

Procédure lasso de Meinshausen et Bühlmann [MB06]. Les deux auteurs proposent d'estimer pour chacun des sommets $a \in \Gamma$ le vecteur (θ_a) de la régression conditionnelle en utilisant l'estimateur lasso de paramètre λ, noté $\widehat{\theta_a}^\lambda$. Ils en déduisent alors deux estimateurs du graphe $\mathcal{G}$: Pour le premier, on met une arête entre a et b si $\widehat{\theta_a}^\lambda[b]$ est différent de 0 ou si $\widehat{\theta_b}^\lambda[a]$ est différent de 0. Pour le second, on met une arête entre a et b si $\widehat{\theta_b}^\lambda[a]$ est différent de 0 et si $\widehat{\theta_a}^\lambda[b]$ est différent de 0. Comme énoncé dans la section précédente, Meinshausen et Bühlmann ont prouvé que sous certaines conditions qui assurent que le lasso est consistant en sélection de variables, leur méthode sélectionne de façon consistante le graphe $\mathcal{G}$. Leur résultat asymptotique correspond au cas où p, n, et le degré du graphe tendent vers l'infini avec $p \gg n \gg deg(\mathcal{G})$. Par ailleurs, les auteurs proposent un choix de λ donnant des bons résultats en pratique.

KGGM de Giraud [Gir08a]. Dans le même ordre d'idées, Giraud [Gir08a] estime le graphe $\mathcal{G}$ en utilisant simultanément les p problèmes de régression (1.18). Étant donné un collection $\mathcal{M}(\Gamma)$ de graphes non orientés de sommet Γ, il sélectionne le graphe $\widehat{\mathcal{G}}^K$ en minimisant le critère des moindres carrés simultanément pour chaque sommet $a \in \Gamma$

$$\widehat{\mathcal{G}}^K := \arg \min_{\mathcal{G}'\in\mathcal{M}(\Gamma)} \sum_{a\in\Gamma} (1 + q(K, |ne'_{\mathcal{G}}(a)|) \min_{\theta\in\mathbb{R}^{ne_{\mathcal{G}'}(a)}} \|\mathbf{Z}_a - \theta \mathbf{Z}_{ne_{\mathcal{G}'}(a)}\|_n^2 \ ,$$

où $q(K, |ne_{\mathcal{G}'}(a)|)$ qui dépend d'un paramètre $K > 1$ à choisir et de la taille $|ne_{\mathcal{G}'}(a)|$ du voisinage de a fait office de pénalité. En cela, KGGM repose sur une méthode de pénalisation très proche de la procédure de sélection de modèles étudiée par Baraud *et al.* [BGH08] en design fixe ou par V. dans le chapitre 4. Dans son travail simultané et indépendant du nôtre, Giraud obtient des bornes de type oracle sur l'efficacité de sa procédure.

Dans la Section 4.1.2, nous proposons une méthode proche de celle de Giraud. En quelques mots, nous proposons d'utiliser une approche analogue à celle de Meinshausen et Bühlmann mais en utilisant la méthode de sélection de modèles par pénalisation expliquée en Section 1.2.1.3. La différence principale avec la méthode de Giraud est que nous pouvons séparer notre critère à minimiser en p sous-critères à minimiser.

Ces deux méthodes sont très flexibles puisque le choix des collections de modèles est libre au statisticien. Si aucun résultat n'est établi en terme de consistance, les méthodes satisfont des inégalités non-asymptotiques de type oracle. Cependant, leur coût computationnel est très élevé, ce qui les rend incalculables pour p plus grand que 40. Néanmoins, dans un travail en cours en collaboration avec Giraud et Huet [GHV], nous combinons la méthode de Giraud avec des procédures statistiques rapides comme le lasso pour diminuer drastiquement le temps de calcul. Nous évoquons cette approche dans la discussion du chapitre 4.

Pour finir cette présentation (non exhaustive) de méthodes d'estimation du graphe, mentionnons le travail de Villers *et al.* [VSBH08]. Elles ont réalisé une étude comparative de certaines des procédures précédemment citées sur des données simulées ainsi que sur des données réelles. Il en ressort que la méthode de Meinshausen et Bühlmann donne en général de bons résultats. Néanmoins, les résultats diffèrent légèrement selon les exemples considérés et les valeurs relatives de n et p. Il semble que lorsque p est faible ($<$40), la méthode KGGM donne de meilleurs résultats.

Contributions de la thèse :

Estimation de graphe. Comme expliqué plus haut, nous introduisons et étudions dans le chapitre 4 une méthode d'estimation de graphe assez proche de KGGM.

Test d'adéquation de graphe. Le chapitre 3 de cette thèse est le fruit d'un travail commun avec Fanny Villers. Nous y introduisons une procédure de test de voisinage d'un modèle graphique gaussien, puis en déduisons une procédure de test d'adéquation du graphe (question **A.2**). Par test de voisinage d'un modèle graphique, nous entendons le problème suivant. Étant donné le graphe $\mathcal{G}$ et $a \in \Gamma$, nous voulons tester l'hypothèse : « Z satisfait la propriété de Markov local en a par rapport à $ne_{\mathcal{G}}(a)$ ». Cette étude est dans la droite ligne des résultats du chapitre 2 évoqués en section 1.2.2. Pour $J \subset \Gamma \setminus \{a\}$, nous avons énoncé dans le lemme 1.9 l'équivalence entre l'assertion « $Z_a \perp\!\!\!\perp Z_{\Gamma\setminus\{J\}\cup\{a\}} | Z_J$ » et « le support du vecteur θ_a de la régression conditionnelle de Z_a par rapport à $Z_{\Gamma\setminus\{a\}}$ est inclus dans J ». On se ramène donc au problème **B.3**, i.e. à tester l'hypothèse « le support θ_a est inclus dans $ne_{\mathcal{G}}(a)$ ». Nous définissons ainsi des procédures de test de voisinage qui sont des cas particuliers des procédures introduites dans le chapitre 2. Ces tests de voisinage héritent des propriétés décrites précédemment : ils sont peu coûteux en temps de calcul même lorsque p est grand, leur niveau est contrôlé et ils sont optimaux au sens minimax contre des voisinages alternatifs parcimonieux. Par ailleurs, leur définition est flexible et permet éventuellement de s'adapter à des choix spécifiques d'hypothèse alternative. Nous déduisons des tests de voisinage des tests d'adéquation de graphe en appliquant tout simplement une procédure de test de voisinage en chaque sommet a de Γ couplée à une méthode de Bonferroni. Enfin, nous présentons des illustrations numériques des performances sur des données simulées ainsi que sur des données réelles.

Nous avons mentionné précédemment que certaines méthodes d'estimation peuvent être déclinées sous la forme de tests de graphes (ex : Drton et Perlman [DP04] ou la procédure de Wille et Bühlmann [WB06]). Cependant, le problème d'estimation du graphe est intrinsèquement plus difficile. Les méthodes ainsi définies ne profitent pas de la « relative » simplicité du problème de test et ne permettent pas d'obtenir des résultats d'optimalité.

1.3.2 Estimation et tests de graphes orientés

Évoquons rapidement les méthodes d'estimation et de tests de graphes pour des modèles graphique orientés. Nous avons mentionné dans la section 1.1.1 qu'à un graphe dirigé acyclique on peut associer une numérotation (en général non unique) sur les sommets. La difficulté de l'estimation du graphe orienté $\overrightarrow{\mathcal{G}}$ est très différente si on connaît à l'avance une numérotation compatible avec $\overrightarrow{\mathcal{G}}$ ou non. Si on considère un test d'adéquation d'un graphe orienté acyclique $\overrightarrow{\mathcal{G}}$ aux données, la question ne se pose pas car on peut toujours utiliser une numérotation associée à ce graphe.

Numérotation connue. Lorsqu'une numérotation est connue, les méthodes d'estimation du graphe $\overrightarrow{\mathcal{G}}$ sont proches des méthodes utilisées dans le cas non-orienté. Ainsi, Consonni et Leucari [CL01] proposent une méthode bayésienne d'estimation utilisant une distribution a priori sur tous les graphes orientés acycliques compatibles avec la numérotation donnée. La méthode de Drton et Perlman [DP04] introduite dans la sous-section précédente se généralise à ce cadre-ci. Contrairement aux modèles graphiques non orientés pour lesquels le graphe était déduit du pattern de 0 de la matrice de précision Ω, le graphe $\overrightarrow{\mathcal{G}}$ est déduit du pattern de 0 du facteur T de Cholesky de Ω. Ainsi, au lieu d'utiliser l'estimateur *glasso*, Huang *et al.* proposent un estimateur de maximum de vraisemblance de T pénalisé par la norme l_1 de T. Globalement, ces méthodes présentent les mêmes avantages et inconvénients que leur pendant en estimation de graphe non-orienté.

Concernant le test d'adéquation des données au graphe $\overrightarrow{\mathcal{G}}$, on peut définir une procédure analogue à celle introduite dans la sous-section précédente. Les tests de voisinage sont maintenant remplacés par

les tests de l'hypothèse : « $Z_i \perp\!\!\!\perp Z_{<i\setminus pa_{\overrightarrow{\mathcal{G}}}(i)} | Z_{pa_{\overrightarrow{\mathcal{G}}}(i)}$ » où $Z_{<i\setminus pa_{\overrightarrow{\mathcal{G}}}(i)}$ désigne l'ensemble des Z_j tel que $j < i$ et $j \notin pa_{\overrightarrow{\mathcal{G}}}(i)$. Cette hypothèse s'exprime en termes de support des paramètres de la régression conditionnelle de Z_i par rapport $Z_{<i}$. Ainsi, on peut facilement étendre la méthodologie décrite pour les tests de graphes non orientés aux graphes orientés acycliques.

Numérotation inconnue. Lorsqu'on ne connaît pas a priori de numérotation liée au graphe orienté, l'estimation du graphe devient un problème non identifiable. En effet, si une distribution est génèrée par un graphe orienté acyclique $\overrightarrow{\mathcal{G}}$, alors il existe une classe d'équivalence de graphes orientés acycliques (minimaux au sens de l'inclusion) qui génèrent cette même distribution. On peut l'expliquer facilement de la façon suivante : La notion de causalité est sous-jacente à l'orientation d'un graphe d'un modèle graphique gaussien orienté. Or, en observant simultanément les $(Z_i)_{i\in\Gamma}$, on peut mettre en évidence des corrélation et pas des relations de causalité. L'objectif est donc ramené à l'estimation de la classe d'équivalence des DAG qui peuvent générer ces données (voir Chickering [Chi02]). On consultera [KB07] pour plus de détails sur ces classes d'équivalence ainsi que pour une courte revue des méthodes existantes.

1.4 Estimation de matrices de covariances

Nous revenons maintenant au problème **A.3** d'estimation de la distribution d'un vecteur gaussien. Nous considérons les deux situations suivantes :

- Le vecteur Z est indexé par un rectangle de $\mathbb{Z}^d$ et est stationnaire sur ce rectangle. La dimension $d = 1$ correspond à des séries temporelles. Nous apportons une attention particulière à la dimension 2, souvent considérée en analyse de données spatiales.
- Le vecteur Z n'est pas supposé stationnaire.

Entre ces deux cas, il existe des situations intermédiaires de stationnarité approchée ou de stationnarité sur des réseaux non-réguliers.

1.4.1 Cas non-stationnaire

Lorsque le nombre d'observations n est inférieur à p, nous avons mentionné que les performances de la matrice de covariance empirique sont mauvaises. Ceci a motivé l'introduction et la mise en oeuvre de procédures alternatives d'estimation de la matrice Σ (ou de façon équivalente de Ω).

L'estimation des matrices de covariance n'est généralement pas une fin en soi. C'est souvent une étape nécessaire avant l'application d'une méthode d'analyse de données comme l'*analyse en composantes principales* (ACP) ou l'*analyse linéaire discriminante*. En quelques mots, l'analyse en composantes principales vise à estimer les k plus grandes valeurs propres de la matrice Σ et leurs vecteurs propres associés. L'analyse discriminante linéaire est une technique de classification basée sur la comparaison entre les vraisemblances.

Contrairement aux modèles de régression linéaire évoqués en section 1.2, il n'y a pas une perte plus naturelle qu'une autre pour mesurer la qualité d'un estimateur $\widehat{\Sigma}$ (ou $\widehat{\Omega}$). Il est donc pertinent d'utiliser une perte liée à l'utilisation finale de l'estimateur de la matrice de covariance. El Karoui [EK08] montre qu'une bonne façon de mesurer la qualité de $\widehat{\Sigma}$ en vue d'une analyse en composante principale est de considérer la norme d'opérateur notée $\|.\|_o$. Comme l'analyse discriminante linéaire est basée sur la vraisemblance, un choix raisonnable dans ce cas est l'entropie de Kullback notée $\mathcal{K}(\Sigma;\widehat{\Sigma})$ entre la vraie distribution et la distribution estimée. Dans la littérature, on utilise parfois la perte de Frobenius $\|.\|_F^2$ définie comme la somme des carrés des entrées d'une matrice.

Dans la suite, nous décrivons les approches classiques pour l'estimation de Σ. Hormis pour des raisons computationnelles, le choix d'une approche plutôt qu'une autre est lié à des connaissances a priori sur la cible Σ ou à la procédure d'analyse de données appliquée.

Régularisation de la covariance Σ. Lorsqu'on veut estimer directement la matrice Σ, nombre de méthodes reviennent à régulariser la matrice Σ. Ainsi Ledoit et Wolf [LW04] proposent de remplacer la matrice de covariance empirique par une combinaison linéaire de celle-ci avec l'identité. Cependant, cette méthode de contraction n'estime pas de façon consistante les valeurs propres. En utilisant des résultats récents en théorie des matrices aléatoires, El Karoui [EK08] ainsi que Bickel et Levina [BL08a] étudient le seuillage de la matrice de covariance empirique. Ils obtiennent ainsi un estimateur consistant en

termes de norme d'opérateur. De plus, Bickel et Levina introduisent une méthode de rééchantillonnage qui sélectionne le paramètre du seuillage. Observons que ces procédures sont invariantes sous une permutation des variables.

Pour certaines applications (par exemple des données temporelles), il existe un ordre naturel sur les variables. Certaines procédures utilisent cette connaissance afin d'obtenir une convergence plus rapide. Entre autres, Furrer et Bentgsson [FB07] et Bickel et Levina [BL08b] introduisent et étudient des estimateurs de matrices quasi-diagonalles[19]. Plus précisément, ils seuillent toutes les composantes de $\widehat{\Sigma}$ à zéro sauf celles qui sont « proches » de la diagonale. Bickel et Levina montrent ainsi des résultats du même type que ceux obtenus dans le seuillage général, mais les vitesses sont cette fois plus rapides.

En résumé, les méthodes de régularisation de la covariance sont rapides à calculer et permettent d'obtenir des résultats de convergence en norme d'opérateur. Le choix pratique du paramètre de régularisation afin d'obtenir de l'adaptativité n'est pas complètement résolu.

Matrice de précision et modèles graphiques. Cette deuxième approche repose sur l'estimation de la matrice de précision Ω. L'idée est d'approcher la matrice Ω par une matrice $\widehat{\Omega}$ qui est « relativement » creuse. De façon équivalente, cela revient à estimer la distribution de Z dans une classe de modèles graphiques. Aussi, certaines méthodes évoquées en section 1.3.1 pour l'estimation de graphes non-orientés sont encore pertinentes ici. C'est le cas des procédures bayésiennes et des méthodes de vraisemblance pénalisée par la norme l_1 [YL07, BEGd08, RBLZ08]. Pour cette dernière méthode, Rothman *et al.* ont prouvé la convergence de leur estimateur $\widehat{\Omega}$ vers Ω en norme de Frobenius si Ω est creuse.

Décomposition de Cholesky et modèles graphiques orientés. Lorsqu'il existe un ordre naturel sur les variables, la régularisation est opérée via les facteurs de Cholesky T et S de la matrice Ω. On rappelle que $\Omega = T^* S^{-1} T$.

Pour des données temporelles, on s'attend à ce que les covariables pertinentes pour la régression de Z_i par rapport à $Z_{<i}$ sont ses proches prédécesseurs. Par la proposition 1.7, on s'attend donc à ce que la matrice T soit bien approchée par une matrice bande[20]. Dans cet esprit, Wu et Pourahmadi [WP03] estiment les k sous-diagonales de T par maximum de vraisemblance tandis que les autres composantes de la matrice valent 0. Le paramètre k est choisi grâce à une pénalité de type AIC. Cependant, ils ne prouvent aucun résultat général sur leur procédure. Bickel et Levina [BL08b] considèrent également le même type d'estimateur mais ils choisissent k grâce à une méthode de validation croisée. Connaissant un bon k, ils prouvent la convergence de leur estimateur en norme d'opérateur. Nous appelons ce problème l'*estimation par matrices bandes*.

Entre la régularisation de la matrice de précision et l'estimation de matrices bandes, il existe une troisième approche qui n'est pas invariante par permutation mais qui ne suppose pas non plus que T est approximativement une matrice bande. Elle consiste à approcher T par une matrice triangulaire inférieure creuse. Quand cette approche est-elle pertinente ? Rappelons d'abord qu'un modèle graphique orienté dont le graphe est compatible avec la numérotation a une matrice T creuse. Ainsi, si on pense que Z est un modèle graphique orienté, cette modélisation est intéressante. Plus généralement, nous pensons que cette approche donne de bons résultats, même si on n'a pas de connaissance a priori sur une bonne numérotation des variables. Certes, il n'y a pas invariance par permutation de la méthode, et il existe des matrices Ω creuses telles que T ne l'est pas. Un exemple est donné dans [RBLZ08] sect. 4. Néanmoins, de tels exemples sont pathologiques. De plus, on peut créer des exemples inverses où le facteur T est creux tandis que Ω ne l'est pas. Les capacités d'approximation des matrices T creuses sont légèrement différentes de celles des matrices de précision creuses, mais il n'est pas clair à l'heure actuelle qu'elles soient moins bonnes. Enfin, un aspect à ne pas négliger est que les procédures d'estimation basées sur T sont beaucoup plus rapides que leur pendant basé sur Ω (voir la section 5.2). Dans la suite, on appelle le problème d'estimation de T par une matrice creuse, l'*estimation complète du facteur de Cholesky*, par analogie avec l'estimation complète dans les problèmes de régression. Dans ce cadre Huang *et al.* [HLPL06] ont proposé une procédure d'estimation de T par maximum de vraisemblance pénalisée par la norme l_1 de T. Plus récemment, Lam et Fan [LF07] ont prouvé la consistance d'une telle méthode en norme de Frobenius si la cible T est creuse.

[19] Banded matrices et tapered matrices en anglais.
[20] T est nulle en dehors des termes diagonaux et des termes sous-diagonaux « proches » de la diagonale.

Contributions de la thèse. Dans le chapitre 5, nous introduisons une procédure générale de pénalisation du maximum de vraisemblance par la complexité pour l'estimation des matrices T et S. Nous pouvons ainsi traiter à la fois le problème d'estimation par matrices bandes et le problème de sélection complète du facteur de Cholesky T. Cette procédure satisfait une inégalité oracle non asymptotique par rapport à la perte de Kullback *sans* aucune hypothèse sur la matrice de précision Ω. Contrairement aux résultats évoqués précédemment, ces vitesses de convergence et inégalités oracles sont libres de toute dépendance cachée sur des paramètres comme la plus grande valeur propre de Σ.

Dans le cadre des matrices bandes, notre procédure est minimax adaptative à la vitesse de décroissance des termes lorsqu'on s'éloigne de la diagonale. Contrairement aux procédures l_1, nous calculons explicitement les constantes pour calibrer la pénalité. Enfin, la méthode est rapide à calculer.

Pour la sélection complète du facteur T, nous prouvons que la procédure est minimax adaptive à la parcimonie de la matrice T (d'un point de vue non-asymptotique). Nous obtenons également des résultats de convergence asymptotique en norme de Frobenius. La méthode est flexible et permet éventuellement d'intégrer des connaissances a priori sur T. Cependant, la procédure est dans ce cas très lourde en temps de calcul, ce qui la rend inapplicable pour p plus grand que 50.

1.4.2 Cas stationnnaire : Champs de Markov

Enfin dans la dernière partie de cette thèse (chapitres 6 et 7), nous nous intéressons à l'estimation de la distribution d'un champ Z stationnaire sur une grille rectangulaire notée $\mathcal{D}$. Ce type de question est souvent rencontré en statistique spatiale et en analyse d'image. Pour de telles applications, il arrive que le nombre n de répétitions soit égal à un. Aussi est-il crucial de bien utiliser la stationnarité du champ. Comme dans le cas non-stationnaire, on distingue les approches basées sur la covariance des approches basées sur la précision.

Par abus de notation, nous utiliserons dans cette sous-section, une notation matricielle pour Z. Ainsi, $Z[i,j]$ correspond à la valeur du champ en un point $(i,j) \in D$. Par convention, on suppose que le point $(0,0)$ appartient à $\mathcal{D}$.

Estimation de variogrammes : Sous l'hypothèse de stationnarité, la variance $\mathrm{var}(Z[i_1,j_1] - Z[i_2,j_2])$ ne dépend que de la différence $(i_1 - i_2, j_1 - j_2)$ entre les deux sommets. On peut donc définir une fonction appelée variogramme $\gamma(.)$ par

$$2\gamma(i,j) := \mathrm{var}(Z[i,j] - Z[0,0]) \qquad \forall (i,j) \in \mathcal{D} \ .$$

La connaissance du variogramme est équivalente à celle de la covariance du champ Z. Une approche classique en statistique spatiale revient à calculer un estimateur non-paramétrique du variogramme[21], puis à l'ajuster à un modèle paramétrique de variogramme choisi à l'avance. Il existe de nombreuses variations de cette approche et nous renvoyons le lecteur intéressé au chapitre 2 de Cressie [Cre93] pour une revue exhaustive. Les avantages de telles méthodes sont leur robustesse vis à vis de données non gaussiennes ainsi que leur grande flexibilité. Cependant, le choix du modèle de variogramme est potentiellement problématique. Un « mauvais » modèle peut entraîner une grande erreur d'estimation. La construction de procédures automatiques de sélection de variogrammes est difficile. Certes, des méthodes basées sur la validation croisée ont été proposées ([Cre93] sect. 2.6.4) mais la question reste encore ouverte. A l'opposé des ces méthodes paramétriques, Rosenblatt ([Ros85] ch.5) a développé une procédure non-paramétrique d'estimation de la covariance de Z. Si cette procédure est universellement consistante, elle n'atteint pas une vitesse paramétrique de convergence lorsque la vraie distribution appartient à un modèle paramétrique. En d'autres termes, pour l'estimation d'un covariance d'un champs stationnaire, la connexion n'a pas été faite entre estimation paramétrique et estimation non-paramétrique comme c'est le cas par exemple dans les modèles de régression.

Champs de Markov. Alternativement, on peut estimer la distribution du champ Z en se basant sur la distribution conditionnelle des variables du champ. Comme expliqué tout au long de ce chapitre, cette approche est reliée à l'estimation de la matrice de précision. Dans ce cas, le statisticien suppose

[21] appelé parfois *variogramme empirique*.

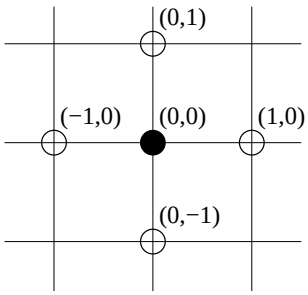


FIG. 1.1 – Exemple du voisinage aux 4 plus proches voisins.

généralement que Z est un modèle graphique gaussien par rapport à un certain graphe $\mathcal{G}$ *connu*. Comme le champ est supposé stationnaire, le graphe est également invariant par translation dans le rectangle et est donc défini de façon unique par le voisinage du point $(0,0)$ appartenant à $\mathcal{D}$. Les voisinages considérés sont généralement formés de points proches de $(0,0)$ selon l'intuition que les corrélations sont faibles pour des sommets éloignés sur le réseau.

Nous représentons par exemple sur la figure 1.1 le voisinage du point $(0,0)$ pour le graphe dit « aux quatre plus proches voisins ». Si le voisinage (ou de façon équivalente le graphe) est fixé, différentes procédures permettent d'estimer la distribution de Z. Citons le maximum de vraisemblance et le maximum de pseudo-vraisemblance. La pseudo-vraisemblance est le produit des distributions conditionnelles des $Z[i,j]$ par rapport aux autres variables.

$$p\mathcal{L}(Z) := \prod_{(i,j)\in\mathcal{D}} \mathcal{L}\left(Z[i,j] | \{Z[k,l](k,l) \in \mathcal{D} \setminus \{i,j\}\}\right) \ ,$$

où $\mathcal{L}(.|.)$ désigne la vraisemblance conditionnelle. Ce critère a été introduit pour des raisons numériques car la maximisation de la vraisemblance requiert l'optimisation du déterminant d'une grande matrice. Le comportement aysmptotique de ces estimateurs a été calculé dans une série de papiers par Besag (voir par exemple [BM75, Bes77, Guy87]). Cependant, il existe peu de résultats pour la *sélection* du voisinage. Guyon et Yao [GY99] donnent des conditions nécessaires et suffisantes pour qu'une procédure de sélection de modèles choisisse asymptotiquement le *vrai* voisinage. Rappelons que notre point de vue est légèrement différent. Nous ne supposons pas qu'il existe un *vrai* voisinage, mais nous cherchons à estimer au mieux la distribution de Z.

Contribution de cette thèse. Dans le chapitre 6 de cette thèse, nous introduisons une procédure de sélection de voisinage utilisant des estimateurs de pseudo-vraisemblance. Nous montrons qu'elle satisfait une inégalité de type oracle non-asymptotique. De plus, elle s'adapte d'un point de vue minimax à la parcimonie de la matrice Ω. Son temps de calcul reste faible même lorsque p est grand. Cependant, cette méthode présente deux inconvénients majeurs : elle n'est définie que pour des domaines $\mathcal{D}$ dont les conditions aux bords sont toriques. Par ailleurs, les pénalités introduites dépendent de paramètres inconnus comme la plus grande valeur propre de la matrice de covariance.

Le chapitre 7 résout ces deux problèmes, tout au moins d'un point de vue pratique. Birgé et Massart [BM07] ont introduit une méthode appelé « heuristique de pente » pour calibrer une pénalité connue à constante multiplicative près. Nous justifions l'utilisation de cette heuristique dans notre cadre de modèles spatiaux. De plus, nous étendons la procédure afin de considérer des conditions aux bords non toriques pour $\mathcal{D}$. Enfin, nous montrons sur des exemples numériques que notre procédure est parfois plus performantes que les procédures basées sur l'estimation paramétrique de variogramme, et ce même lorsque la covariance appartient à un modèle paramétrique de variogramme.

Chapter 2

Goodness-of-fit Tests for high-dimensional Gaussian linear models

Abstract. Let $(Y, (X_i)_{1\leq i\leq p})$ be a real zero mean Gaussian vector and V be a subset of $\{1,\dots,p\}$. Suppose we are given n i.i.d. replications of this vector. We propose a new test for testing that Y is independent of $(X_i)_{i\in\{1,\dots p\}\setminus V}$ conditionally to $(X_i)_{i\in V}$ against the general alternative that it is not. This procedure does not depend on any prior information on the covariance of X or the variance of Y and applies in a high-dimensional setting. It straightforwardly extends to test the neighbourhood of a Gaussian graphical model. The procedure is based on a model of Gaussian regression with random Gaussian covariates. We give non asymptotic properties of the test and we prove that it is rate optimal (up to a possible $\log(n)$ factor) over various classes of alternatives under some additional assumptions. Besides, it allows us to derive non asymptotic minimax rates of testing in this random design setting. Finally, we carry out a simulation study in order to evaluate the performance of our procedure.

2.1 Introduction

We consider the following regression model

$$Y = \sum_{i=1}^{p} \theta_i X_i + \epsilon \tag{2.1}$$

where θ is an unknown vector of $\mathbb{R}^p$. In the sequel, we note $\mathcal{I} := \{1,\dots,p\}$. The vector $X := (X_i)_{1\leq i\leq p}$ follows a real zero mean Gaussian distribution with non singular covariance matrix Σ and ϵ is a real zero mean Gaussian random variable independent of X. Straightforwardly, the variance of ϵ corresponds to the conditional variance of Y given X, $\mathrm{var}(Y|X)$.

The variable selection problem for this model in a high-dimensional setting has recently attracted a lot of attention. A large number of papers are now devoted to the design of new algorithms and estimators which are computationally feasible and are proven to converge; see for instance the works of Meinshausen and Bühlmann [MB06], Candès and Tao [CT07], Zhao and Yu [ZY06], Zou and Hastie [ZH05], Bühlmann and Kalisch [BK07], or Bunea *et al.* [BTW07a]. A common drawback of the previously mentioned estimation procedures is that they require restrictive conditions on the covariance matrix Σ in order to behave well. Our issue is the natural testing counterpart of this variable selection problem: we aim at defining a computationally feasible testing procedure that achieves an optimal rate for any covariance matrix Σ.

2.1.1 Presentation of the main results

We are given n i.i.d. replications of the vector (Y, X). Let us respectively note $\mathbf{Y}$ and $\mathbf{X}_i$ the vectors of the n observations of Y and X_i for any $i \in \mathcal{I}$. Let V be a subset of $\mathcal{I}$, then X_V refers to the set $\{X_i, i \in V\}$ and θ_V stands for the sequence $(\theta_i)_{i \in V}$. We first propose a collection of testing procedures T_α of the null hypothesis "$\theta_{\mathcal{I}\setminus V} = 0$" against the general alternative "$\theta_{\mathcal{I}\setminus V} \neq 0$". These procedures are based on the ideas of Baraud *et al.* [BHL03] in a random design. Their definition are very flexible as they require no prior knowledge of the covariance of X, the variance of ϵ, nor the variance of Y. Note that the property "$\theta_{\mathcal{I}\setminus V} = 0$" is equivalent to "$Y$ is independent of $X_{\mathcal{I}\setminus V}$ conditionally to X_V". Hence, it also permits to test conditional independences and applies for testing the graph of Gaussian graphical model (see below). Contrary to most approaches in this setting (e.g. Drton and Pearlman [DP07]), we are able to consider the difficult case of tests in a high-dimensional setting: the number of covariates p is possibly much larger than the number of observations n. Such situations arise in many statistical applications like in genomics or biomedical imaging. To our knowledge, the only testing procedures (e.g. [SS05]) that could handle high-dimensional alternatives lack of theoretical justifications. In this chapter, we exhibit some tests T_α that are both computationally amenable and optimal in the minimax sense.

From a theoretical perspective, we are able to control the Family Wise Error Rate (FWER) of our testing procedures T_α. Besides, we derive a general non asymptotic upper bound for their power. Contrary to the various rates of convergence obtained in the estimation setting (e.g. [MB06] or [CT07]), our upper bound holds for any covariance matrix Σ. Then, we derive from it non-asymptotic minimax rates of testing in the Gaussian random design framework. If the minimax rates are known for a long time in the fixed design Gaussian regression framework (e.g.[Bar02b]), they were unknown in our setting. For instance, if at most k components of θ are non-zero and if k is much smaller than p, we prove that the minimax rates of testing is of order $\frac{k \log(p)}{n}$ when the covariates X_i are independent. If the covariates are dependent, we derive faster minimax rates. To our knowledge, these are the first results for testing or estimation issues that illustrate minimax rates for dependent covariates. Afterwards, we show analogous results when k is large, or when the vector θ belongs to some ellipsoid or some collection of ellipsoids. For any of these alternatives, we exhibit some procedure T_α that achieves the optimal rate (at a possible $\log(n)$ factor). Finally, we illustrate the performance of the procedure on simulated examples.

2.1.2 Application to Gaussian Graphical Models (GGM)

Our work was originally motivated by the following question: let $(Z_j)_{j \in \mathcal{J}}$ be a random vector which follows a zero mean Gaussian distribution whose covariance matrix Σ' is non singular. We observe n i.i.d. replications of this vector Z and we are given a graph $\mathcal{G} = (\Gamma, E)$ where $\Gamma = \{1, \dots |\mathcal{J}|\}$ and E is a set of edges in $\Gamma \times \Gamma$. How can we test that Z is an undirected Gaussian graphical model (GGM) with respect to the graph $\mathcal{G}$?

The random vector Z is a GGM with respect to the graph $\mathcal{G} = (\Gamma, E)$ if for any couple (i, j) which is not contained in the edge set E, Z_i and Z_j are independent, given the remaining variables. See Lauritzen [Lau96] for definitions and main properties of GGM. Interest in these models has grown as they allow the description of dependence structure in high-dimensional data. As such, they are widely used in spatial statistics [Cre93, RH05] or probabilistic expert systems [CDLS99]. More recently, they have been applied to the analysis of microarray data. The challenge is to infer the network regulating the expression of the genes using only a small sample of data, see for instance Schäfer and Strimmer [SS05], Kishino and Waddell [KW00], or Wille *et al.* [WZV+04]. This issue has motivated the research for new estimation procedures to handle GGM in a high-dimensional setting.

It is beyond the scope of this chapter to give an exhaustive review of these. Many of these graph estimation methods are based on multiple testing procedures, see for instance Schäfer and Strimmer [SS05] or Wille and Bühlmann [WB06]. Other methods are based on variable selection for high-dimensional data we previously mentioned. For instance, Meinshausen and Bühlmann [MB06] proposed a computationally feasible model selection algorithm using Lasso penalization. Huang *et al.* [HLPL06] and Yuan and Lin [YL07] extend this method to infer directly the inverse covariance matrix Σ'^{-1} by minimizing the log-likehood penalized by the l^1 norm.

While the issue of graph and covariance estimation is extensively studied, few theoretical results are proved for the problem of hypothesis testing of GGM in a high-dimensional setting. We believe that this

issue is significant for two reasons: first, when considering a gene regulation network, the biologists often have a previous knowledge of the graph and may want to test if the microarray data match with their model. Second, when applying an estimation method in a high-dimensional setting, it could be useful to test the estimated graph as some of these methods reveal too conservative.

Admittedly, some of the previously mentioned estimation methods are based on multiple testing. However, as they are constructed for an estimation purpose, most of them do not take into account some previous knowledge about the graph. This is for instance the case for the approaches of Drton and Perlman [DP07] and Schäfer and Strimmer [SS05]. Some of the other existing procedures cannot be applied in a high-dimensional setting ($|\mathcal{J}| \geq n$). Finally, most of them lack theoretical justification in a non asymptotic way.

In Chapter 3, we define a test of graph based on the present work. It benefits the ability of handling high dimensional GGM and has minimax properties. Besides we show numerical evidence of its efficiency; see Chapter 3 for more details. In this article, we shall only present the idea underlying our approach.

For any $j \in \mathcal{J}$, we note $N(j)$ the set of neighbours of j in the graph $\mathcal{G}$. Testing that Z is a GGM with respect to $\mathcal{G}$ is equivalent to testing that the random variable Z_j conditionally to $(Z_l)_{l\in N(j)}$ is independent of $(Z_l)_{l\in\mathcal{J}\setminus(N(j)\cup\{j\})}$ for any $j \in \mathcal{J}$. As Z follows a Gaussian distribution, the distribution of Z_j conditionally to the other variables decomposes as follows:

$$Z_j = \sum_{k\in\mathcal{J}\setminus\{j\}} \theta_k Z_k + \epsilon_j,$$

where ϵ_j is normal and independent of $(Z_k)_{k\in\mathcal{J}\setminus\{j\}}$. Then, the statement of conditional independency is equivalent to $\theta_{\mathcal{J}\setminus\{j\}\cup N(j)} = 0$. This approach based on conditional regression is also used for estimation by Meinshausen and Bühlmann [MB06].

2.1.3 Organization of the chapter

In Section 2.2, we present the approach of our procedure and connect it with the fixed design framework. Besides, we define the notion of minimax rates of testing in this setting and gather the main notations. We define the testing procedures T_α in Section 2.3 and we non asymptotically characterise the set of vectors θ over which the test T_α is powerful. In Section 2.4 and 2.5, we apply our procedure to define tests and study their optimality for two different classes of alternatives. More precisely, in Section 2.4 we test $\theta = 0$ against the class of θ whose components equal 0, except at most k of them (k is supposed small). We define a test which under mild conditions achieves the minimax rate of testing. When the covariates are independent, it is interesting to note that the minimax rates exhibits the same ranges in our statistical model (2.1) and in fixed design regression model (2.2). In Section 2.5, we define two procedures that achieve the simultaneous minimax rates of testing over large classes of ellipsoids (to sometimes the price of a $\log(p)$ factor). Besides, we show that the problem of adaptation over classes of ellipsoids is impossible without a loss in efficiency. This was previously pointed out in [Spo96] in fixed design regression framework. The simulation studies are presented in Section 2.6. Finally, Sections 2.7, 2.8 and Appendix contain the proofs.

2.2 Description of the approach

2.2.1 Connection with tests in fixed design regression

Our work is directly inspired by the testing procedure of Baraud *et al.* [BHL03] in fixed design regression framework. Contrary to model (2.1), the problem of hypothesis testing in fixed design regression has been extensively studied. This is why we will use the results in this framework as a benchmark for the theoretical bounds in our model (2.1). Let us define this second regression model:

$$Y_i = f_i + \sigma\epsilon_i, \quad i \in \{1, \dots, N\}, \tag{2.2}$$

where f is an unknown vector of $\mathbb{R}^N$, σ some unknown positive number, and the ϵ_i's a sequence of i.i.d. standard Gaussian random variables. The problem at hand is testing that f belongs to a linear subspace

of $\mathbb{R}^N$ against the alternative that it does not. We refer to [BHL03] for a short review of non parametric tests in this framework. Besides, we are interested in the performance of the procedures from a minimax perspective. To our knowledge, there has been no results in model (2.1). However, there are numerous papers on this issue in the fixed design regression model. First, we refer to the seminal work of Ingster [Ing93a, Ing93b, Ing93c] who gives asymptotic minimax rates over non parametric alternatives. Our work is closely related to the results of Baraud [Bar02b] where he gives non asymptotic minimax rates of testing over ellipsoids or sparse signals. Throughout the chapter, we highlight the link between the minimax rates in fixed and in random design.

2.2.2 Principle of our testing procedure

Let us briefly describe the idea underlying our testing procedure. A formal definition will follow in Section 2.3.1.Let m be a subset of $\mathcal{I} \setminus V$. We respectively define S_V and $S_{V\cup m}$ as the linear subspaces of $\mathbb{R}^p$ such that $\theta_{\mathcal{I}\setminus V} = 0$, respectively $\theta_{\mathcal{I}\setminus(V\cup m)} = 0$. We note d and D_m for the cardinalities of V and m and N_m refers to $N_m = n - d - D_m$. If $N_m > 0$, we define the Fisher statistic ϕ_m by

$$\phi_m(\mathbf{Y}, \mathbf{X}) := \frac{N_m \|\Pi_{V\cup m}\mathbf{Y} - \Pi_V \mathbf{Y}\|_n^2}{D_m \|\mathbf{Y} - \Pi_{V\cup m}\mathbf{Y}\|_n^2} \ , \tag{2.3}$$

where Π_V refers to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_i)_{i\in V}$ and $\|.\|_n$ is the canonical norm in $\mathbb{R}^n$. We define the test statistic $\phi_{m,\alpha}(\mathbf{Y}, \mathbf{X})$ as

$$\phi_{m,\alpha}(\mathbf{Y}, \mathbf{X}) = \phi_m(\mathbf{Y}, \mathbf{X}) - \bar{F}^{-1}_{D_m,N_m}(\alpha) \ , \tag{2.4}$$

where $\bar{F}_{D_m,N_m}(u)$ denotes the probability for a Fisher variable with D and N degrees of freedom to be larger than u. Let us consider a finite collection $\mathcal{M}$ of non empty subsets of $\mathcal{I} \setminus V$ such that for each $m \in \mathcal{M}$, $N_m > 0$. Our testing procedure consists of doing a Fisher test for each $m \in \mathcal{M}$. We define $\{\alpha_m, m \in \mathcal{M}\}$ a suitable collection of numbers in $]0, 1[$ (which possibly depends on $\mathbf{X}$). For each $m \in \mathcal{M}$, we do the Fisher test ϕ_m of level α_m of:

$$H_0 : \theta \in S_V \quad \text{against the alternative} \quad H_{1,m} : \theta \in S_{V\cup m} \setminus S_V$$

and we decide to reject the null hypothesis if one of those Fisher tests does.

The main advantage of our procedure is that it is very flexible in the choices of the model $m \in \mathcal{M}$ and in the choices of the weights $\{\alpha_m\}$. Consequently, if we choose a suitable collection $\mathcal{M}$, the test is powerful over a large class of alternatives as shown in Sections 2.3.3, 2.4, and 2.5.

Finally, let us mention that our procedure easily extends to the case where the expectation of the random vector (Y, X) is unknown. Let $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ denote the projections of $\mathbf{X}$ and $\mathbf{Y}$ onto the unit vector $\mathbf{1}$. Then, one only has to apply the procedure to $(\mathbf{Y} - \overline{\mathbf{Y}}, \mathbf{X} - \overline{\mathbf{X}})$ and to replace d by $d+1$. The properties of the test remain unchanged and one can adapt all the proofs to the price of more technicalities.

2.2.3 Minimax rates of testing

In order to examine the quality of our tests, we will compare their performance with the minimax rates of testing. That is why we now define precisely what we mean by the (α, δ)-minimax rate of testing over a set Θ. We endow $\mathbb{R}^p$ with the Euclidean norm

$$\|\theta\|^2 := \theta^t \Sigma \theta = \mathrm{var}\left(\sum_{i=1}^p \theta_i X_i\right) \ . \tag{2.5}$$

As ϵ and X are independent, we derive from the definition of $\|.\|^2$ that $\mathrm{var}(Y) = \|\theta\|^2 + \mathrm{var}(Y|X)$. Let us remark that $\mathrm{var}(Y|X)$ does not depend on X. If we have $\|\theta\|$ vary, either the quantity $\mathrm{var}(Y)$ or $\mathrm{var}(Y|X)$ has to vary. In the sequel, we suppose that $\mathrm{var}(Y)$ is fixed. We briefly justify this choice in Section 2.4.2. Consequently, if $\|\theta\|^2$ is increasing, then $\mathrm{var}(Y|X)$ has to decrease so that the sum remains constant. Let α be a number in $]0; 1[$ and let δ be a number in $]0; 1-\alpha[$ (typically small). For a given

vector θ, matrix Σ and $\mathrm{var}(Y)$, we denote $\mathbb{P}_\theta$ the joint distribution of $(\mathbf{Y}, \mathbf{X})$. For the sake of simplicity, we do not emphasize the dependence of $\mathbb{P}_\theta$ on $\mathrm{var}(Y)$ or Σ. Let ψ_α be a test of level α of the hypothesis "$\theta = 0$" against the hypothesis "$\theta \in \Theta \setminus 0$". In our framework, it is natural to measure the performance of ψ_α using the quantity $\rho\left(\psi_\alpha, \Theta, \delta, \mathrm{var}(Y), \Sigma\right)$ defined by:

$$\rho\left(\psi_\alpha, \Theta, \delta, \mathrm{var}(Y), \Sigma\right) \quad := \quad \inf\left\{\rho > 0,\ \inf\left\{\mathbb{P}_\theta(\psi_\alpha = 1),\ \theta \in \Theta \text{ and } \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq \rho^2\right\} \geq 1 - \delta\right\},$$

where the quantity

$$r_{s/n}(\theta) := \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \tag{2.6}$$

appears naturally as it corresponds to the ratio $\|\theta\|^2/\mathrm{var}(Y|X)$ which is the quantity of information brought by X (i.e. the signal) over the conditional variance of Y (i.e. the noise). We aim at describing the quantity

$$\inf_{\psi_\alpha} \rho\left(\psi_\alpha, \Theta, \delta, \mathrm{var}(Y), \Sigma\right) := \rho\left(\Theta, \alpha, \delta, \mathrm{var}(Y), \Sigma\right), \tag{2.7}$$

where the infimum is taken over all the level-α tests ψ_α. We call this quantity the (α, δ)-minimax rate of testing over Θ.

A dual notion of this ρ function is the function β_Σ. For any $\Theta \subset \mathbb{R}^p$ and $\alpha \in]0, 1[$, we denote $\beta_\Sigma(\Theta)$ the quantity

$$\beta_\Sigma\left(\Theta\right) := \inf_{\psi_\alpha} \sup_{\theta \in \Theta} \mathbb{P}_\theta\left[\psi_\alpha = 0\right],$$

where the infimum is taken over all level-α tests ψ_α and where we recall that Σ refers to the covariance matrix of X.

2.2.4 Notations

Let recall the main notations that we shall use throughout the chapter. In the sequel, n stands for the number of independent observations, p is the number of covariates. Besides, X_V stands for the collection $(X_i)_{i \in V}$ of the covariates that correspond to the null hypothesis and d is the cardinality of the set V. The models m are subsets of $\mathcal{I} \subset V$ and we note D_m their cardinality. T_α stands for our testing procedure of level α. The statistics ϕ_m and the test $\phi_{m,\alpha}$ are respectively defined in (2.3) and (2.4). Finally, the norm $\|.\|$ is introduced in 2.5.

For $x, y \in \mathbb{R}$, we set

$$x \wedge y := \inf\{x, y\},\ x \vee y := \sup\{x, y\}.$$

For any $u \in \mathbb{R}$, $\bar{F}_{D,N}(u)$ denotes the probability for a Fisher variable with D and N degrees of freedom to be larger than u. In the sequel, L, L_1, L_2,... denote constants that may vary from line to line. The notation $L(.)$ specifies the dependency on some quantities. For the sake of simplicity, we only give the orders of magnitude in the results and we refer to the proofs for explicit constants.

2.3 The Testing procedure

2.3.1 Description of the procedure

Let us first fix some level $\alpha \in]0, 1[$. Throughout this chapter, we suppose that $n \geq d + 2$. Let us consider a finite collection $\mathcal{M}$ of non empty subsets of $\mathcal{I} \setminus V$ such that for all $m \in \mathcal{M}$, $1 \leq D_m \leq n - d - 1$. We introduce the following test of level α. We reject H_0: "$\theta \in S_V$" when the statistic

$$T_\alpha := \sup_{m \in \mathcal{M}} \left\{\phi_m(\mathbf{Y}, \mathbf{X}) - \bar{F}^{-1}_{D_m, N_m}(\alpha_m(\mathbf{X}))\right\} \tag{2.8}$$

is positive, where the collection of weights $\{\alpha_m(\mathbf{X}), m \in \mathcal{M}\}$ is chosen according to one of the two following procedures:

P_1 : The α_m 's do not depend on $\mathbf{X}$ and satisfy the equality :

$$\sum_{m\in\mathcal{M}} \alpha_m = \alpha \ . \tag{2.9}$$

P_2 : For all $m \in \mathcal{M}$, $\alpha_m(\mathbf{X}) = q_{\mathbf{X},\alpha}$, the α-quantile of the distribution of the random variable

$$\inf_{m\in\mathcal{M}} \bar{F}_{D_m,N_m}\left(\frac{\|\Pi_{V\cup m}(\boldsymbol{\epsilon}) - \Pi_V(\boldsymbol{\epsilon})\|_n^2/D_m}{\|\boldsymbol{\epsilon} - \Pi_{V\cup m}(\boldsymbol{\epsilon})\|_n^2/N_m}\right) \tag{2.10}$$

conditionally to $\mathbf{X}$.

Note that it is easy to compute the quantity $q_{\mathbf{X},\alpha}$. Let Z be a standard Gaussian random vector of size n independent of $\mathbf{X}$. As $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, the distribution of (2.10) conditionally to $\mathbf{X}$ is the same as the distribution of

$$\inf_{m\in\mathcal{M}} \bar{F}_{D_m,N_m}\left(\frac{\|\Pi_{V\cup m}(Z) - \Pi_V(Z)\|^2/D_m}{\|Z - \Pi_{V\cup m}(Z)\|^2/N_m}\right)$$

conditionally to $\mathbf{X}$. Hence, we can easily work out its quantile using Monte-Carlo method.

Clearly, the computational complexity of the procedure is linear with respect to the size of the collection of models $\mathcal{M}$ even when using Procedure P_2. Consequently, when we apply our procedure to high-dimensional data as in Section 2.6 or in Chapter 3, we favour collections $\mathcal{M}$ whose size is linear with respect to the number of covariates p.

2.3.2 Comparison of Procedures P_1 and P_2

We respectively refer to T_α^1 and T_α^2 for the tests (2.8) associated with Procedure P_1 and P_2. First, we are able to control the behavior of the test under the null hypothesis.

Proposition 2.1. *The test T_α^1 corresponds to a Bonferroni procedure and therefore satisfies*

$$\mathbb{P}_\theta(T_\alpha > 0) \leq \sum_{m\in\mathcal{M}} \alpha_m \leq \alpha,$$

whereas the test T_α^2 has the property to be of size exactly α:

$$\mathbb{P}_\theta(T_\alpha > 0) = \alpha.$$

The proof is given in Appendix. Besides, the test T_α^2 is more powerful than the corresponding test T_α^1 defined with weights $\alpha_m = \alpha/|\mathcal{M}|$.

Proposition 2.2. *For any parameter θ that does not belong to S_V, the procedure T_α^1 with weights $\alpha_m = \alpha/|\mathcal{M}|$ and the procedure T_α^2 satisfy*

$$\mathbb{P}_\theta\left(T_\alpha^2(\mathbf{X},\mathbf{Y}) > 0 \middle| \mathbf{X}\right) \geq \mathbb{P}_\theta\left(T_\alpha^1(\mathbf{X},\mathbf{Y}) > 0 \,|\, \mathbf{X}\right) \quad \mathbf{X}\ a.s.\ . \tag{2.11}$$

Again, the proof is given in Appendix. On the one hand, the choice of Procedure P_1 allows to avoid the computation of the quantile $q_{\mathbf{X},\alpha}$ and possibly permits to give a Bayesian flavor to the choice of the weights. On the other hand, Procedure P_2 is more powerful than the corresponding test with Procedure P_1. We will illustrate these considerations in Section 2.6. In sections 2.3.3, 2.4, and 2.5 we study the power and rates of testing of T_α with Procedure P_1.

2.3.3 Power of the Test

We aim at describing a set of vectors θ in $\mathbb{R}^p$ over which the test defined in Section 2.3 with Procedure P_1 is powerful. Since Procedure P_2 is more powerful than Procedure P_1 with $\alpha_m = \alpha/|\mathcal{M}|$, the test with Procedure P_2 will also be powerful on this set of θ.

Let α and δ be two numbers in $]0,1[$, and let $\{\alpha_m, m \in \mathcal{M}\}$ be weights such that $\sum_{m\in\mathcal{M}} \alpha_m \leq \alpha$. Let define Hypothesis $(H_\mathcal{M})$ as follows:

$(H_{\mathcal{M}})$ For all $m \in \mathcal{M}$, $\alpha_m \geq \exp(-N_m/10)$ and $\delta \geq \exp 2(-N_m/21)$.

For typical choices of the collections $\mathcal{M}$ and $\{\alpha_m, m \in \mathcal{M}\}$, these conditions are fulfilled as discussed in Sections 2.4 and 2.5. Let us now turn to the main result.

Theorem 2.3. *Let T_α be the test procedure defined by (2.8). We assume that $n > d + 2$ and that Assumption $(H_{\mathcal{M}})$ holds. Then, $\mathbb{P}_\theta(T_\alpha > 0) \geq 1 - \delta$ for all θ belonging to the set*

$$\mathcal{F}_{\mathcal{M}}(\delta) := \left\{ \theta \in \mathbb{R}^p, \exists m \in \mathcal{M} : \frac{\mathrm{var}(Y|X_V) - \mathrm{var}(Y|X_{V\cup m})}{\mathrm{var}(Y|X_{V\cup m})} \geq \Delta(m) \right\},$$

where

$$\Delta(m) := \frac{L_1\sqrt{D_m \log\left(\frac{2}{\alpha_m\delta}\right)}\left(1 + \sqrt{\frac{D_m}{N_m}}\right) + L_2\left(1 + 2\frac{D_m}{N_m}\right)\log\left(\frac{2}{\alpha_m\delta}\right)}{n - d}. \tag{2.12}$$

This result is similar to Theorem 1 in [BHL03] in fixed design regression framework and the same comment also holds: the test T_α under procedure P_1 has a power comparable to the best of the tests among the family $\{\phi_{m,\alpha}, m \in \mathcal{M}\}$. Indeed, let us assume for instance that $V = \{0\}$ and that the α_m are chosen to be equal to $\alpha/|\mathcal{M}|$. The test T_α defined by (2.8) is equivalent to doing several tests of $\theta = 0$ against $\theta \in S_m$ at level α_m for $m \in \mathcal{M}$ and it rejects the null hypothesis if one of those tests does. From Theorem 2.3, we know that under the hypothesis $H_{\mathcal{M}}$ this test has a power greater than $1 - \delta$ over the set of vectors θ belonging to $\bigcup_{m\in\mathcal{M}} \mathcal{F}'_m(\delta, \alpha_m)$ where $\mathcal{F}'_m(\delta, \alpha_m)$ is the set of vectors $\theta \in \mathbb{R}^p$ such that

$$\frac{\mathrm{var}(Y) - \mathrm{var}(Y|X_m)}{\mathrm{var}(Y|X_m)} \geq \frac{L(D_m, N_m)}{n}\left(\sqrt{D_m \log\left(\frac{2}{\alpha_m\delta}\right)} + \log\left(\frac{2}{\alpha_m\delta}\right)\right). \tag{2.13}$$

Besides, $L(D_m, N_m)$ behaves like a constant if the ratio D_m/N_m is bounded. Let us compare this result with the set of θ over which the Fisher test $\phi_{m,\alpha}$ at level α has a power greater than $1 - \delta$. Applying Theorem 2.3, we know that it contains $\mathcal{F}'_m(\delta, \alpha)$. Moreover, the following Proposition shows that it is not much larger than $\mathcal{F}'_m(\delta, \alpha)$:

Proposition 2.4. *Let $\delta \in]0, 1 - \alpha[$. If*

$$\frac{\mathrm{var}(Y) - \mathrm{var}(Y|X_m)}{\mathrm{var}(Y|X_m)} \leq L(\alpha, \delta)\frac{\sqrt{D_m}}{n},$$

then $\mathbb{P}_\theta(\phi_{m,\alpha} > 0) \leq 1 - \delta$

The proof is postponed to Section 2.8 and is based on a lower bound of the minimax rate of testing.

$\mathcal{F}'_m(\delta, \alpha)$ and $\mathcal{F}'_m(\delta, \alpha_m)$ defined in (2.13) differ from the fact that $\log(1/\alpha)$ is replaced by $\log(1/\alpha_m)$. For the main applications that we will study in Section 2.4, 2.5, and 2.6, the ratio $\log(1/\alpha_m)/\log(1/\alpha)$ is of order $\log(n)$, $\log\log n$, or $k\log(ep/k)$ where k is a "small" integer. Thus, for each $\delta \in]0, 1 - \alpha[$, the test based on T_α has a power greater than $1 - \delta$ over a class of vectors which is close to $\bigcup_{m\in\mathcal{M}} \mathcal{F}'_m(\delta, \alpha)$. It follows that for each $\theta \neq 0$ the power of this test under $\mathbb{P}_\theta$ is comparable to the best of the tests among the family $\{\phi_{m,\alpha}, m \in \mathcal{M}\}$.

In the next two sections, we use this theorem to establish rates of testing against different types of alternatives. First, we give an upper bound for the rate of testing $\theta = 0$ against a class of θ for which a lot of components are equal to 0. In Section 2.5, we study the rates of testing and simultaneous rates of testing $\theta = 0$ against classes of ellipsoids. For the sake of simplicity, we will only consider the case $V = \{0\}$. Nevertheless, the procedure T_α defined in (2.8) applies in the same way when one considers more complex null hypothesis and the rates of testing are unchanged except that we have to replace n by $n - d$ and $\mathrm{var}(Y)$ by $\mathrm{var}(Y|X_V)$.

2.4 Detecting non-zero coordinates

Let us fix an integer k between 1 and p. In this section, we are interested in testing $\theta = 0$ against the class of θ with a most k non-zero components. This typically corresponds to the situation encountered

when considering tests of neighborhood for large sparse graphs. As the graph is assumed to be sparse, only a small number of neighbors are missing under the alternative hypothesis.

For each pair of integers (k,p) with $k \leq p$, let $\mathcal{M}(k,p)$ be the class of all subsets of $\mathcal{I} = \{1,\dots,p\}$ of cardinality k. The set $\Theta[k,p]$ stands for the subset of vectors $\theta \in \mathbb{R}^p$, such that at most k coordinates of θ are non-zero.

First, we define a test T_α of the form (2.8) with Procedure P_1, and we derive an upper bound for the rate of testing of T_α against the alternative $\theta \in \Theta[k,p]$. Then, we show that this procedure is rate optimal when all the covariates are independent. Finally, we study the optimality of the test when $k=1$ for some examples of covariance matrix Σ.

2.4.1 Rate of testing of T_α

Proposition 2.5. *We consider the set of models $\mathcal{M} = \mathcal{M}(k,p)$. We use the test T_α under Procedure P_1 and we take the weights α_m all equal to $\alpha/|\mathcal{M}|$. Let us suppose that n satisfies:*

$$n \geq L\left[\log\left(\frac{2}{\alpha\delta}\right) + k\log\left(\frac{ep}{k}\right)\right]. \tag{2.14}$$

Let us set the quantity

$$\rho'^2_{k,p,n} := L(\alpha,\delta)\frac{k\log\left(\frac{ep}{k}\right)}{n}. \tag{2.15}$$

For any θ in $\Theta[k,p]$, such that $\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq \rho'^2_{k,p,n}$, $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1-\delta$.

We recall that the norm $\|.\|$ is defined in (2.5) and equals $\mathrm{var}(Y) - \mathrm{var}(Y|X)$. This proposition easily follows from Theorem 2.3 and its proof is given in Section 2.7. Note that the upper bound does not directly depend on the covariance matrix of the vector X. Besides, Hypothesis (2.14) corresponds to the minimal assumption needed for consistency and type-oracle inequalities in the estimation setting as pointed out by Wainwright ([Wai07] Th. 2) and Giraud ([Gir08a] Sect. 3.1). Hence, we conjecture that Hypothesis (2.14) is minimal so that Proposition 2.5 holds. We will further discuss the bound (2.15) after deriving lower bounds for the minimax rate of testing.

2.4.2 Minimax lower bounds for independent covariates

In the statistical framework considered here, the problem of giving minimax rates of testing under no prior knowledge of the covariance of X and of $\mathrm{var}(Y)$ is open. This is why we shall only derive lower bounds when $\mathrm{var}(Y)$ and the covariance matrix of X are known. In this section, we give non asymptotic lower bounds for the (α,δ)-minimax rate of testing over the set $\Theta[k,p]$ when the covariance matrix of X is the identity matrix (except Proposition 2.6). As these bounds coincide with the upper bound obtained in Section 2.4.1, this will show that our test T_α is rate optimal.

We first give a lower bound for the (α,δ)-minimax rate of detection of all p non-zero coordinates for any covariance matrix Σ.

Proposition 2.6. *Let us suppose that* $\mathrm{var}(Y)$ *is known. Let us set $\rho^2_{p,n}$ such that:*

$$\rho^2_{p,n} := L(\alpha,\delta)\frac{\sqrt{p}}{n}. \tag{2.16}$$

Then for all $\rho < \rho_{p,n}$,

$$\beta_\Sigma\left(\left\{\theta \in \Theta[p,p], \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = \rho^2\right\}\right) \geq \delta,$$

where we recall that Σ is the covariance matrix of X.

If $n \geq (1+\gamma)p$ for some $\gamma > 0$, Theorem 2.3 shows that the test $\phi_{\mathcal{I},\alpha}$ defined in (2.4) has power greater than δ over the vectors θ that satisfy

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq L(\gamma,\alpha,\delta)\frac{\sqrt{p}}{n}.$$

Hence, $\sqrt{p}/n$ is the minimax rate of testing $\Theta[p,p]$ at least when the number of observations is larger than the number of covariates. This is coherent with the minimax rate obtained in the fixed design framework (e.g. [Bar02b]). When p becomes larger we do not think that the lower bound given in Proposition 2.6 is still sharp. Note that this minimax rate of testing holds for any covariance matrix Σ contrary to Theorem 2.7.

We now turn to the lower bound for the (α,δ)-minimax rate of testing against $\theta \in \Theta[k,p]$.

Theorem 2.7. *Let us set $\rho^2_{k,p,n}$ such that*

$$\rho^2_{k,p,n} := L(\alpha,\delta)\frac{k}{n}\log\left(1+\frac{p}{k^2}+\sqrt{2\frac{p}{k^2}}\right). \tag{2.17}$$

We suppose that the covariance of X is the identity matrix I. Then, for all $\rho < \rho_{k,p,n}$,

$$\beta_I\left(\left\{\theta\in\Theta[k,p], \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2}=\rho^2\right\}\right) > \delta.$$

where the quantity $\mathrm{var}(Y)$ is known.

If $\alpha+\delta \le 53\%$, then one has

$$\rho^2_{k,p,n} \ge \frac{k}{2n}\log\left(1+\frac{p}{k^2}\vee\sqrt{\frac{p}{k^2}}\right).$$

This result implies the following lower bound for the minimax rate of testing

$$\rho\left(\Theta[k,p],\alpha,\delta,\mathrm{var}(Y),I\right)\right) \ge \rho^2_{k,p,n}.$$

The proof is given in Section 2.8. To the price of more technicalities, it is possible to prove that the lower bound still holds if the variables (X_i) are independent with known variances possibly different. Theorem 2.7 recovers approximately the lower bounds for the minimax rates of testing in signal detection framework obtained by Baraud [Bar02b]. The main difference lies in the fact that we suppose $\mathrm{var}(Y)$ known which in the signal detection framework translates in the fact that we would know the quantity $\|f\|^2+\sigma^2$.

We are now in position to compare the results of Proposition 2.5 and Theorem 2.7. Let distinguish between the values of k.

- When $k \le p^\gamma$ for some $\gamma < 1/2$, if n is large enough to satisfy the assumption of Proposition 2.5, the quantities $\rho^2_{k,p,n}$ and $\rho'^2_{k,p,n}$ are both of the order $\frac{k\log(p)}{n}$ times a constant (which depends on γ, α, and δ). This shows that the lower bound given in Theorem 2.7 is sharp. Additionally, in this case, the procedure T_α defined in Proposition 2.5 follows approximately the minimax rate of testing. We recall that our procedure T_α does not depend on the knowledge of $\mathrm{var}(Y)$ and $\mathrm{corr}(X)$. In applications, a small k typically corresponds to testing a Gaussian graphical model with respect to a graph $\mathcal{G}$, when the number of nodes is large and the graph is supposed to be sparse. When n does not satisfy the assumption of Proposition 2.5, we believe that our lower bound is not sharp anymore.

- When $\sqrt{p} \le k \le p$, the lower bound and the upper bound do not coincide anymore. Nevertheless, if $n \ge (1+\gamma)p$ for some $\gamma > 0$, Theorem 2.3 shows that the test $\phi_{\mathcal{I},\alpha}$ defined in (2.4) has power greater than δ over the vectors θ that satisfy

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \ge L(\gamma,\alpha,\delta)\frac{\sqrt{p}}{n}. \tag{2.18}$$

  This upper bound and the lower bound do not depend on k. Here again, the lower bound obtained in Theorem 2.7 is sharp and the test $\phi_{\mathcal{I},\alpha}$ defined previously is rate optimal. The fact that the rate of testing stabilizes around $\sqrt{p}/n$ for $k > \sqrt{p}$ also appears in signal detection and there is a discussion of this phenomenon in [Bar02b].

- When $k < \sqrt{p}$ and k is close to $\sqrt{p}$, the lower bound and the upper bound given by Proposition 2.5 differ from at most a $\log(p)$ factor. For instance, if k is of order $\sqrt{p}/\log p$, the lower bound in

Theorem 2.7 is of order $\sqrt{p}\log\log p/\log p$ and the upper bound is of order $\sqrt{p}$. We do not know if any of this bound is sharp and if the minimax rates of testing coincide when $\mathrm{var}(Y)$ is fixed and when it is not fixed.

All in all, the minimax rates of testing exhibit the same range of rates in our framework as in signal detection [Bar02b] when the covariates are independent. Moreover, this implies that the minimax rate of testing is slower when the $(X_i)_{i\in\mathcal{I}}$ are independent than for any other form of dependence. Indeed, the upper bounds obtained in Proposition 2.5 and in (2.18) do not depend on the covariance of X. Then, a natural question arises: is the test statistic T_α rate optimal for other correlation of X? We will partially answer this question when testing against the alternative $\theta \in \Theta[1,p]$.

2.4.3 Minimax rates for dependent covariates

In this section, we look for the minimax rate of testing $\theta = 0$ against $\theta \in \Theta[1,p]$ when the covariates X_i are no longer independent. We know that this rate is between the orders $\frac{1}{n}$, which is the minimax rate of testing when we know which coordinate is non-zero, and $\frac{\log(p)}{n}$, the minimax rate of testing for independent covariates.

Proposition 2.8. *Let us suppose that there exists a positive number c such that for any $i \neq j$,*

$$|corr(X_i, X_j)| \leq c$$

and that $\alpha + \delta \leq 53\%$. We define $\rho^2_{1,p,n,c}$ as

$$\rho^2_{1,p,n,c} := \frac{L}{n}\left(\log(p) \wedge \frac{1}{c}\right). \tag{2.19}$$

Then for any $\rho < \rho_{1,p,n,c}$,

$$\beta_\Sigma\left(\left\{\theta \in \Theta[1,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho^2\right\}\right) \geq \delta,$$

where Σ refers to the covariance matrix of X.

Remark: If the correlation between the covariates is smaller than $1/\log(p)$, then the minimax rate of testing is of the same order as in the independent case. If the correlation between the covariates is larger, we show in the following Proposition that under some additional assumption, the rate is faster.

Proposition 2.9. *Let us suppose that the correlation between X_i and X_j is exactly $c > 0$ for any $i \neq j$. Moreover, we assume that n satisfies the following condition:*

$$n \geq L\left[1 + \log\left(\frac{p}{\alpha\delta}\right)\right] \tag{2.20}$$

Let introduce the random variable $X_{p+1} := \frac{1}{p}\sum_{i=1}^p \frac{X_i}{\sqrt{var(X_i)}}$. If $\alpha < 60\%$ and $\delta < 60\%$ the test T_α defined by

$$T_\alpha = \left[\sup_{1\leq i\leq p} \phi_{\{i\},\alpha/(2p)}\right] \vee \phi_{\{p+1\},\alpha/2}$$

satisfies

$$\mathbb{P}_0\left(T_\alpha > 0\right) \leq \alpha \text{ and } \mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1 - \delta,$$

for any θ in $\Theta[1,p]$ such that

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq \frac{L(\alpha,\delta)}{n}\left(\log p \bigwedge \frac{1}{c}\right).$$

Consequently, when the correlation between X_i and X_j is a positive constant c, the minimax rate of testing is of order $\frac{\log(p)\wedge(1/c)}{n}$. When the correlation coefficient c is small, the minimax rate of testing coincides with the independent case, and when c is larger those rates differ. Therefore, the test T_α defined in Proposition 2.5 is not rate optimal when the correlation is known and is large. Indeed, when the correlation between the covariates is large, the tests statistics $\phi_{\{m\},\alpha_m}$ defining T_α are highly correlated.

The choice of the weights α_m in Procedure P_1 corresponds to a Bonferroni procedure, which is precisely known to behave bad when the tests are positively correlated.

This example illustrates the limits of Procedure P_1. However, it is not very realistic to suppose that the covariates have a constant correlation, for instance when one considers a GGM. Indeed, we expect that the correlation between two covariates is large if they are neighbors in the graph and smaller if they are far (w.r.t. the graph distance). This is why we derive lower bounds of the rate of testing for other kind of correlation matrices often used to model stationary processes.

Proposition 2.10. *Let $X_1, \dots, X_p$ form a stationary process on the one dimensional torus. More precisely, the correlation between X_i and X_j is a function of $|i-j|_p$ where $|.|_p$ refers to the toroidal distance defined by:*

$$|i-j|_p := (|i-j|) \wedge (p - |i-j|) \ .$$

$\Sigma_1(w)$ and $\Sigma_2(t)$ respectively refer to the correlation matrix of X such that

$$\begin{aligned} \mathrm{corr}(X_i, X_j) &= \exp(-w|i-j|_p) \text{ where } w > 0 \ , \\ \mathrm{corr}(X_i, X_j) &= (1+|i-j|_p)^{-t} \text{ where } t > 0 \ . \end{aligned}$$

Let us set $\rho^2_{1,p,n,\Sigma_1}(w)$ and $\rho^2_{1,p,n,\Sigma_2}(t)$ such that:

$$\begin{aligned} \rho^2_{1,p,n,\Sigma_1}(w) &:= \frac{1}{n}\log\left(1 + L(\alpha,\delta)p\frac{1-e^{-w}}{1+e^{-w}}\right) \\ \rho^2_{1,p,n,\Sigma_2}(t) &:= \begin{cases} \frac{1}{n}\log\left(1 + L(\alpha,\delta)\frac{p(t-1)}{t+1}\right) & \text{if} \quad t > 1 \\ \frac{1}{n}\log\left(1 + L(\alpha,\delta)\frac{p}{1+2\log(p-1)}\right) & \text{if} \quad t = 1 \\ \frac{1}{n}\log\left(1 + L(\alpha,\delta)p^t 2^{-t}(1-t)\right) & \text{if} \quad 0 < t < 1. \end{cases} \end{aligned}$$

Then, for any $\rho^2 < \rho^2_{1,p,n,\Sigma_1}(w)$,

$$\beta_{\Sigma_1(w)}\left(\left\{\theta \in \Theta[1,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho^2\right\}\right) \geq \delta,$$

and for any $\rho^2 < \rho^2_{1,p,n,\Sigma_2}(t)$,

$$\beta_{\Sigma_2(t)}\left(\left\{\theta \in \Theta[1,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho^2\right\}\right) \geq \delta.$$

If the range ω is larger than $1/p^\gamma$ or if the range t is larger than γ for some $\gamma < 1$, these lower bounds are of order $\frac{\log p}{n}$. As a consequence, for any of these correlation models the minimax rate of testing is of the same order as the minimax rate of testing for independent covariates. This means that our test T_α defined in Proposition 2.5 is rate-optimal for these correlations matrices. However, if ω is smaller than $1/p$ or if t is smaller than $1/\log(p)$, we recover the parametric rates $1/n$, which is achieved by the test $\phi_{\{p+1\},\alpha}$. This comes from the fact that the correlation $\mathrm{corr}(X_1, X_i)$ does not converge to zero for such choices of ω or t. We omit the details since the arguments are similar to the proof of Proposition 2.9.

To conclude, when $k \leq p^\gamma$ (for $\gamma \leq 1/2$), the test T_α defined in Proposition 2.5 is approximately (α,δ)-minimax against the alternative $\theta \in \Theta[k,p]$, when neither $\mathrm{var}(Y)$ nor the covariance matrix of X is fixed. Indeed, the rate of testing of T_α coincide (up to a constant) with the supremum of the minimax rates of testing on $\Theta[k,p]$ over all possible covariance matrices Σ:

$$\rho\left(\Theta[k,p], \alpha, \delta\right) := \sup_{\mathrm{var}(Y)>0, \Sigma>0} \rho\left(\Theta[k,p], \alpha, \delta, \mathrm{var}(Y), \Sigma\right),$$

where the supremum is taken over all positive $\mathrm{var}(Y)$ and every positive definite matrix Σ. When $k \geq \sqrt{p}$ and when $n \geq (1+\gamma)p$ (for $\gamma > 0$), the test defined in (2.18) has the same behavior.

However, our procedure does not adapt to Σ: for some correlation matrices (as shown for instance in Proposition 2.9), T_α with Procedure P_1 is not rate optimal. Nevertheless, we believe and this will be illustrated in Section 2.6 that Procedure P_2 slightly improves the power of the test when the covariates are correlated.

2.5 Rates of testing on “ellipsoids" and adaptation

In this section, we define tests T_α of the form (2.8) in order to test simultaneously $\theta = 0$ against θ belongs to some classes of ellipsoids. We will study their rates and show that they are optimal at sometimes the price of a $\log p$ factor.

For any non increasing sequence $(a_i)_{1\leq i\leq p+1}$ such that $a_1 = 1$ and $a_{p+1} = 0$ and any $R > 0$, we define the ellipsoid $\mathcal{E}_a(R)$ by

$$\mathcal{E}_a(R) := \left\{ \theta \in \mathbb{R}^p, \sum_{i=1}^{p} \frac{\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i})}{a_i^2} \leq R^2 \mathrm{var}(Y|X) \right\}, \tag{2.21}$$

where m_i refers to the set $\{1, \dots, i\}$ and $m_0 = \varnothing$.

Let us explain why we call this set an ellipsoid. Assume for instance that the (X_i) are independent identically distributed with variance one. In this case, the difference $\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i})$ equals $|\theta_i|^2$ and the definition of $\mathcal{E}_a(R)$ translates in

$$\mathcal{E}_a(R) = \left\{ \theta \in \mathbb{R}^p, \sum_{i=1}^{p} \frac{|\theta_i|^2}{a_i^2} \leq R^2 \mathrm{var}(Y|X) \right\}.$$

The main difference between this definition and the classical definition of an ellipsoid in the fixed design regression framework (as for instance in [Bar02b]) is the presence of the term $\mathrm{var}(Y|X)$. We added this quantity in order to be able to derive lower bounds of the minimax rate. If the X_i are not i.i.d. with unit variance, it is always possible to create a sequence X_i' of i.i.d. standard gaussian variables by orthogonalizing the X_i using Gram-Schmidt process. If we call θ' the vector in $\mathbb{R}^p$ such that $X\theta = X'\theta'$, it is straightforward to show that $\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i}) = |\theta_i'|^2$. We can then express $\mathcal{E}_a(R)$ using the coordinates of θ' as previously:

$$\mathcal{E}_a(R) = \left\{ \theta \in \mathbb{R}^p, \sum_{i=1}^{p} \frac{|\theta_i'|^2}{a_i^2} \leq R^2 \mathrm{var}(Y|X) \right\}.$$

The main advantage of Definition 2.21 is that it does not directly depend on the covariance of X. In the sequel we also consider the special case of ellipsoids with polynomial decay,

$$\mathcal{E}_s'(R) := \left\{ \theta \in \mathbb{R}^p, \sum_{i=1}^{p} \frac{\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i})}{i^{-2s}\mathrm{var}(Y|X)} \leq R^2 \right\}, \tag{2.22}$$

where $s > 0$ and $R > 0$. First, we define two tests procedures of the form (2.8) and evaluate their power respectively on the ellipsoids $\mathcal{E}_a(R)$ and on the ellipsoids $\mathcal{E}_s'(R)$. Then, we give some lower bounds for the (α, δ)-simultaneous minimax rates of testing. Extensions to more general l_p balls with $0 < p < 2$ are possible to the price of more technicalities by adapting the results of Section 4 in Baraud [Bar02b].

These alternatives correspond to the situation where we are given an order of relevance on the covariates that are not in the null hypothesis. This order could either be provided by a previous knowledge of the model or by a model selection algorithm such as LARS (least angle regression) introduced by Efron *et al.* [EHJT04]. We apply this last method to build a collection of models for our testing procedure (2.8) in Chapter 3.

2.5.1 Simultaneous Rates of testing of T_α over classes of ellipsoids

First, we define a procedure of the form (2.8) in order to test $\theta = 0$ against θ belongs to any of the ellipsoids $\mathcal{E}_a(R)$. For any $x > 0$, $[x]$ denotes the integer part of x.

We choose the class of models $\mathcal{M}$ and the weights α_m as follows:

- If $n < 2p$, we take the set $\mathcal{M}$ to be $\cup_{1\leq k\leq [n/2]} m_k$ and all the weights α_m are equal to $\alpha/|\mathcal{M}|$.
- If $n \geq 2p$, we take the set $\mathcal{M}$ to be $\cup_{1\leq k\leq p} m_k$. α_{m_p} equals $\alpha/2$ and for any k between 1 and $p-1$, α_{m_k} is chosen to be $\alpha/(2(p-1))$.

As previously, we bound the power of the tests T_α from a non-asymptotic point of view.

Proposition 2.11. *Let us assume that*

$$n \geq L\left[1+\log\left(\frac{1}{\alpha\delta}\right)\right]. \tag{2.23}$$

For any ellipsoid $\mathcal{E}_a(R)$, the test T_α defined by (2.8) with Procedure P_1 and with the class of models given just above satisfies

$$\mathbb{P}_0\left(T_\alpha \leq 0\right) \geq 1-\alpha,$$

and $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1-\delta$ for all $\theta \in \mathcal{E}_a(R)$ such that

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq L(\alpha,\delta)\log n \inf_{1\leq i\leq [n/2]}\left[a_{i+1}^2R^2+\frac{\sqrt{i}}{n}\right] \tag{2.24}$$

if $n<2p$, or

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq L(\alpha,\delta)\left\{\left[\log p \inf_{1\leq i\leq p-1}\left(a_{i+1}^2R^2+\frac{\sqrt{i}}{n}\right)\right]\bigwedge\frac{\sqrt{p}}{n}\right\} \tag{2.25}$$

if $n \geq 2p$.

All in all, for large values of n, the rate of testing is of order $\sup_{1\leq i\leq p}\left[a_i^2R^2 \wedge \frac{\sqrt{i\log(p)}}{n}\right]$. We show in the next subsection that the minimax rate of testing for an ellipsoid is of order:

$$\sup_{1\leq i\leq p}\left[a_i^2R^2\wedge\frac{\sqrt{i}}{n}\right].$$

Besides, we prove in Proposition 2.16 that a loss in $\sqrt{\log\log p}$ is unavoidable if one considers the simultaneous minimax rates of testing over a family of nested ellipsoids. Nevertheless, we do not know if the term $\sqrt{\log(p)}$ is optimal for testing simultaneously against all the ellipsoids $\mathcal{E}_a(R)$ for all sequences (a_i) and all $R>0$. When n is smaller than $2p$, we obtain comparable results except that we are unable to consider alternatives in large dimensions in the infimum (2.25).

We now turn to define a procedure of the form (2.8) in order to test simultaneously that $\theta=0$ against θ belongs to any of the $\mathcal{E}'_s(R)$. For this, we introduce the following collection of models $\mathcal{M}$ and weights α_m:

- If $n<2p$, we take the set $\mathcal{M}$ to be $\cup m_k$ where k belongs to $\left\{2^j, j\geq 0\right\}\cap\{1,\ldots,[n/2]\}$ and all the weights α_m are chosen to be $\alpha/|\mathcal{M}|$.
- If $n\geq 2p$, we take the set $\mathcal{M}$ to be $\cup m_k$ where k belongs to $\left(\left\{2^j, j\geq 0\right\}\cap\{1,\ldots,p\}\right)\cup\{p\}$, α_{m_p} equals $\alpha/2$ and for any k in the model between 1 and $p-1$, α_{m_k} is chosen to be $\alpha/(2(|\mathcal{M}|-1))$.

Proposition 2.12. *Let us assume that*

$$n \geq L\left[1+\log\left(\frac{1}{\alpha\delta}\right)\right] \tag{2.26}$$

and that $R^2 \geq \sqrt{\log\log n}/n$. For any $s>0$, the test procedure T_α defined by (2.8) with Procedure P_1 and with a class of models given just above satisfies:

$$\mathbb{P}_0\left(T_\alpha>0\right)\geq 1-\alpha,$$

and $\mathbb{P}_\theta\left(T_\alpha>0\right)\geq 1-\delta$ for any $\theta\in\mathcal{E}'_s(R)$ such that

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq L\left(\alpha,\delta\right)\left[R^{2/(1+4s)}\left(\frac{\sqrt{\log\log n}}{n}\right)^{4s/(1+4s)}+R^2\left(n/2\right)^{-2s}+\frac{\log\log n}{n}\right] \tag{2.27}$$

if $n<2p$ or

$$\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq L\left(\alpha,\delta\right)\left(\left[R^{2/(1+4s)}\left(\frac{\sqrt{\log\log p}}{n}\right)^{4s/(1+4s)}+\frac{\log\log p}{n}\right]\bigwedge\frac{\sqrt{p}}{n}\right) \tag{2.28}$$

if $n\geq 2p$.

Again, we retrieve similar results to those of Corollary 2 in [BHL03] in the fixed design regression framework. For $s > 1/4$ and $n < 2p$, the rate of testing is of order $\left(\frac{\sqrt{\log\log n}}{n}\right)^{4s/(1+4s)}$. We show in the next subsection that the logarithmic factor is due to the adaptive property of the test. If $s \leq 1/4$, the rate is of order n^{-2s}. When $n \geq 2p$, the rate is of order $\left(\frac{\sqrt{\log\log p}}{n}\right)^{4s/(1+4s)} \wedge \left(\frac{\sqrt{p}}{n}\right)$, and we mention at the end of the next subsection that it is optimal.

Here again, it is possible to define these tests with Procedure P_2 in order to improve the power of the test (see Section 2.6 for numerical results).

2.5.2 Minimax lower bounds

We first establish the (α, δ)-minimax rate of testing over an ellipsoid when the variance of Y and the covariance matrix of X are known.

Proposition 2.13. *Let us set the sequence $(a_i)_{1\leq i\leq p+1}$ and the positive number R. We introduce*

$$\rho^2_{a,n}(R) := \sup_{1\leq i\leq p} \left[\rho^2_{i,n} \wedge a_i^2 R^2\right], \tag{2.29}$$

where $\rho^2_{i,n}$ is defined by (2.16), then for any non singular covariance matrix Σ we have

$$\beta_\Sigma\left(\left\{\theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq \rho^2_{a,n}(R)\right\}\right) \geq \delta,$$

where the quantity var(Y) is fixed. If $\alpha + \delta \leq 47\%$ then

$$\rho^2_{a,n}(R) \geq \sup_{1\leq i\leq p}\left[\frac{\sqrt{i}}{n} \wedge a_i^2 R^2\right].$$

This lower bound is once more analogous to the one in the fixed design regression framework. Contrary to the lower bounds obtained in the previous section, it does not depend on the covariance of the covariates. We now look for an upper bound of the minimax rate of testing over a given ellipsoid. First, we need to define the quantity D^* as:

$$D^* := \inf\left\{1 \leq i \leq p, a_i^2 R^2 \leq \frac{\sqrt{i}}{n}\right\}$$

with the convention that $\inf \varnothing = p$.

Proposition 2.14. *Let us assume that $n \geq L\log\left[1 + \log\left(\frac{1}{\alpha\delta}\right)\right]$. If $R^2 > \frac{1}{n}$ and $D^* \leq n/2$, the test $\phi_{m_{D^*},\alpha}$ defined by (2.4) satisfies*

$$\mathbb{P}_0\left[\phi_{m_{D^*},\alpha} = 1\right] \leq \alpha \text{ and } \mathbb{P}_\theta\left[\phi_{m_{D^*},\alpha} = 0\right] \leq \delta$$

for all $\theta \in \mathcal{E}_a(R)$ such that

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq L(\alpha,\delta) \sup_{1\leq i\leq p}\left[\frac{\sqrt{i}}{n} \wedge a_i^2 R\right].$$

If $n \geq 2D^*$, the rates of testing on an ellipsoid are analogous to the rates on an ellipsoid in fixed design regression framework (see for instance [Bar02b]). If D^* is large and n is small, the bounds in Proposition 2.13 and 2.14 do not coincide. In this case, we do not know if this comes from the fact that the test in Proposition 2.14 does not depend on the knowledge of $\mathrm{var}(Y)$ or if one of the bounds in Proposition 2.13 and 2.14 is not sharp.

We are now interested in computing lower bounds for rates of testing simultaneously over a family of ellipsoids, in order to compare them with rates obtained in Section 2.5.1. First, we need a lower bound for the minimax simultaneous rate of testing over nested linear spaces. We recall that for any $D \in \{1, \ldots, p\}$, S_{m_D} stands for the linear spaces of vectors θ such that only their D first coordinates are possibly non-zero.

Proposition 2.15. *For $D \geq 2$, let us set*

$$\bar{\rho}^2_{D,n} := L(\alpha,\delta)\frac{\sqrt{\log\log(D+1)}\sqrt{D}}{n}. \tag{2.30}$$

Then, the following lower bound holds

$$\beta_I\left(\bigcup_{1\leq D\leq p}\left\{\theta\in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_D^2\right\}\right)\geq \delta,$$

if for all D between 1 and p, $r_D \leq \bar{\rho}_{D,n}$

Using this Proposition, it is possible to get a lower bound for the simultaneous rate of testing over a family of nested ellipsoids.

Proposition 2.16. *We fix a sequence $(a_i)_{1\leq i\leq p+1}$. For each $R>0$, let us set*

$$\bar{\rho}^2_{a,R,n} := \sup_{1\leq D\leq p}\left[\bar{\rho}^2_{D,n}\wedge(R^2a_D^2)\right]. \tag{2.31}$$

where $\bar{\rho}_{D,n}$ is given by (2.30). Then, for any non singular covariance matrix Σ of the vector X,

$$\beta_\Sigma\left(\bigcup_{R>0}\left\{\theta\in\mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2}\leq \bar{\rho}^2_{a,R,n}\right\}\right)\geq\delta.$$

This Proposition shows that the problem of adaptation is impossible in this setting: it is impossible to define a test which is simultaneously minimax over a class of nested ellipsoids (for $R>0$). This is also the case in fixed design as proved by [Spo96] for the case of Besov bodies. The loss of a term of the order $\sqrt{\log\log p}/n$ is unavoidable.

As a special case of Proposition 2.16, it is possible to compute a lower bound for the simultaneous minimax rate over $\mathcal{E}'_s(R)$ where R describes the positive numbers. After some calculation, we find a lower bound of order:

$$\left(\frac{\sqrt{\log\log p}}{n}\right)^{\frac{4s}{1+4s}}\bigwedge\frac{\sqrt{p\log\log p}}{n}.$$

This shows that the power of the test T_α obtained in (2.28) for $n\geq 2p$ is optimal when $R^2\geq\sqrt{\log\log n}/n$. However, when $n<2p$ and $s\leq 1/4$, we do not know if the rate n^{-2s} is optimal or not.

To conclude, when $n\geq 2p$ the test T_α defined in Proposition 2.12 achieves the simultaneous minimax rate over the classes of ellipsoids $\mathcal{E}'_s(R)$. On the other hand, the test T_α defined in Proposition 2.11 is not rate optimal simultaneously over all the ellipsoids $\mathcal{E}_a(R)$ and suffers a loss of a $\sqrt{\log p}$ factor even when $n\geq 2p$.

2.6 Simulations studies

The purpose of this simulation study is threefold. First, we illustrate the theoretical results established in previous sections. Second, we show that our procedure is easy to implement for different choices of collections $\mathcal{M}$ and is computationally feasible even when p is large. Our third purpose is to compare the efficiency of Procedures P_1 and P_2. Indeed, for a given collection $\mathcal{M}$, we know from Section 2.3.2 that the test (2.8) based on Procedure P_2 is more powerful than the corresponding test based on P_1. However, the computation of the quantity $q_{\mathbf{X},\alpha}$ is possibly time consuming and we therefore want to know if the benefit in power is worth the computational burden.

To our knowledge, when the number of covariates p is larger than the number of observations n there is no test with which we can compare our procedure.

2.6.1 Simulation experiments

We consider the regression model (2.1) with $\mathcal{I}=\{1,\dots,p\}$ and test the null hypothesis "$\theta=0$", which is equivalent to "Y is independent of X", at level $\alpha=5\%$. Let $(X_i)_{1\leqslant i\leqslant p}$ be a collection of p Gaussian variables with unit variance. The random variable is defined as follows: $Y=\sum_{i=1}^p\theta_iX_i+\varepsilon$ where ε is a zero mean gaussian variable with variance $1-\|\theta\|^2$ independent of X.

We consider two simulation experiments described below.

1. First simulation experiment: The correlation between X_i and X_j is a constant c for any $i \neq j$. Besides, in this experiment the parameter θ is chosen such that only one of its components is possibly non-zero. This corresponds to the situation considered in Section 2.4. First, the number of covariates p is fixed equal to 30 and the number of observations n is taken equal to 10 and 15. We choose for c three different values 0, 0.1, and 0.8, allowing thus to compare the procedure for independent, weakly and highly correlated covariates. We estimate the size of the test by taking $\theta_1 = 0$ and the power by taking for θ_1 the values 0.8 and 0.9. Theses choices of θ lead to a small and a large signal/noise ratio $r_{s/n}$ defined in (2.6) and equal in this experiment to $\theta_1^2/(1-\theta_1^2)$. Second, we examine the behavior of the tests when p increases and when the covariates are highly correlated: p equals 100 and 500, n equals 10 and 15, θ_1 is set to 0 and 0.8, and c is chosen to be 0.8.

2. Second simulation experiment: The covariates $(X_i)_{1\leqslant i\leqslant p}$ are independent. The number of covariates p equals 500 and the number of observations n equals 50 and 100. We set for any $i \in \{1,\dots,p\}$, $\theta_i = Ri^{-s}$. We estimate the size of the test by taking $R = 0$ and the power by taking for (R,s) the value $(0.2, 0.5)$, which corresponds to a slow decrease of the $(\theta_i)_{1\leqslant i\leqslant p}$. It was pointed out in the beginning of Section 2.5 that $|\theta_i|^2$ equals $\mathrm{var}(Y|X_{m_{i-1}}) - \mathrm{var}(Y|X_{m_i})$. Thus, $|\theta_i|^2$ represents the benefit in term of conditional variance brought by the variable X_i.

We use our testing procedure defined in (2.8) with different collections $\mathcal{M}$ and different choices for the weights $\{\alpha_m, m \in \mathcal{M}\}$.

The collections $\mathcal{M}$: we define three classes. Let us set $J_{n,p} = p \wedge [\frac{n}{2}]$, where $[x]$ denotes the integer part of x and let us define:

$$\begin{aligned}
\mathcal{M}^1 &:= \{\{i\}, 1 \leqslant i \leqslant p\} \\
\mathcal{M}^2 &:= \{m_k = \{1,2,\dots,k\}, 1 \leqslant k \leqslant J_{n,p}\} \\
\mathcal{M}^3 &:= \{m_k = \{1,2,\dots,k\}, k \in \{2^j, j \geq 0\} \cap \{1,\dots,J_{n,p}\}\}
\end{aligned}$$

We evaluate the performance of our testing procedure with $\mathcal{M} = \mathcal{M}^1$ in the first simulation experiment, and $\mathcal{M} = \mathcal{M}^2$ and $\mathcal{M}^3$ in the second simulation experiment. The cardinality of these three collections is smaller than p, and the computational complexity of the testing procedures is at most linear in p.

The collections $\{\alpha_m, m \in \mathcal{M}\}$: We consider Procedures P_1 and P_2 defined in Section 2.3. When we are using the procedure P_1, the α_m's equal $\alpha/|\mathcal{M}|$ where $|\mathcal{M}|$ denotes the cardinality of the collection $\mathcal{M}$. The quantity $q_{\mathbf{X},\alpha}$ that occurs in the procedure P_2 is computed by simulation. We use 1000 simulations for the estimation of $q_{\mathbf{X},\alpha}$. In the sequel we note $T_{\mathcal{M}^i,P_j}$ the test (2.8) with collection $\mathcal{M}^i$ and Procedure P_j.

In the first experiment, when p is large we also consider two other tests:

1. The test $\phi_{\{1\},\alpha}$ (2.4) of the hypothesis $\theta_1 = 0$ against the alternative $\theta_1 \neq 0$. This test corresponds to the single test when we know which coordinate is non-zero.

2. The test $\phi_{\{p+1\},\alpha}$ where $X_{p+1} := \frac{1}{p}\sum_{i=1}^p X_i$. Adapting the proof of Proposition 2.9, we know that this test is approximately minimax on $\Theta[1,p]$ if the correlation between the covariates is constant and large.

Contrary to our procedures, these two tests are based on the knowledge of $\mathrm{var}(X)$ (and eventually θ). We only use them as a benchmark to evaluate the performance of our procedure. We aim at showing that our test with Procedure P_2 is as powerful than $\phi_{\{p+1\},\alpha}$ and is close to the test $\phi_{\{1\},\alpha}$.

We estimate the size and the power of the testing procedures with 1000 simulations. For each simulation, we simulate the gaussian vector $(X_1,\dots,X_p)$ and then simulate the variable Y as described in the two simulation experiments.

2.6.2 Results of the simulation

The results of the first simulation experiment for $c = 0$ are given in Table 1. As expected, the power of the tests increases with the number of observations n and with the signal/noise ratio $r_{s/n}$. If the

Null hypothesis is true, $\theta_1 = 0$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.043	0.045
15	0.044	0.049

Null hypothesis is false

$\theta_1 = 0.8$, $r_{s/n} = 1.78$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.48	0.48
15	0.81	0.81

$\theta_1 = 0.9$, $r_{s/n} = 4.26$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.86	0.86
15	0.99	0.99

Table 1: First simulation study, independent case: $p = 30$, $c = 0$. Percentages of rejection and value of the signal/noise ratio $r_{s/n}$.

signal/noise ratio is large enough, we obtain powerful tests even if the number of covariates p is larger than the number of observations.

Null hypothesis is true, $\theta_1 = 0$

$c = 0$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.043	0.045
15	0.044	0.049

$c = 0.1$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.042	0.04
15	0.058	0.06

$c = 0.8$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.018	0.045
15	0.019	0.052

Null hypothesis is false, $\theta_1 = 0.8$

$c = 0$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.48	0.48
15	0.81	0.81

$c = 0.1$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.49	0.49
15	0.81	0.82

$c = 0.8$

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.64	0.77
15	0.89	0.94

Table 2: First simulation study, independent and dependent case. $p = 30$ Frequencies of rejection.

In Table 2 we present results of the first simulation experiment for $\theta_1 = 0.8$ when c varies.
Let us first compare the results for independent, weakly and highly correlated covariates when using Procedure P_1. The size and the power of the test for weakly correlated covariates are similar to the size and the power obtained in the independent case. Hence, we recover the remark following Proposition 2.8: when the correlation coefficient between the covariates is small, the minimax rate is of the same order as in the independent case. The test for highly correlated covariates is more powerful than the test for independent covariates, recovering thus the remark following Theorem 2.7: the worst case from a minimax rate perspective is the case where the covariates are independent. Let us now compare Procedures P_1 and P_2. In the case of independent or weakly correlated covariates, they give similar results. For highly correlated covariates, the power of $T_{\mathcal{M}^1,P_2}$ is much larger than the one of $T_{\mathcal{M}^1,P_1}$.

In Table 3 we present results of the multiple testing procedure and of the two tests $\phi_{\{1\},\alpha}$ and $\phi_{\{p+1\},\alpha}$ when $c = 0.8$ and the number of covariates p is large. For $p = 500$ and $n = 15$, one test takes less than one second with Procedure P_1 and less than 30 seconds with Procedure P_2. As expected, Procedure P_1 is too conservative when p increases. For $p = 100$, the power of the test based on Procedure P_1 is smaller than the power of the test $\phi_{\{p+1\},\alpha}$ and this difference increases when p is larger. The test based on Procedure P_2 is as powerful as $\phi_{\{p+1\},\alpha}$, and its power is close to the one of $\phi_{\{1\},\alpha}$. We recall that this last test is based on the knowledge of the non-zero component of θ contrary to ours. Besides, the test $\phi_{\{p+1\},\alpha}$ was shown in Proposition 2.9 to be optimal for this particular correlation setting. Hence, Procedure P_2 seems to achieve the optimal rate in this situation. Thus, we advise to use in practice Procedure P_2 if the number of covariates p is large, because Procedure P_1 becomes too conservative, especially if the covariates are correlated.

The results of the second simulation experiment are given in Table 4. As expected, Procedure P_2

Null hypothesis is true, $\theta_1 = 0$

	$p = 100$					$p = 500$			
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	$\phi_{\{1\},\alpha}$	$\phi_{\{p+1\},\alpha}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	$\phi_{\{1\},\alpha}$	$\phi_{\{p+1\},\alpha}$
10	0.01	0.056	0.051	0.045	10	0.009	0.044	0.040	0.040
15	0.016	0.053	0.047	0.053	15	0.011	0.040	0.042	0.034

Null hypothesis is false, $\theta_1 = 0.8$

	$p = 100$					$p = 500$			
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	$\phi_{\{1\},\alpha}$	$\phi_{\{p+1\},\alpha}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	$\phi_{\{1\},\alpha}$	$\phi_{\{p+1\},\alpha}$
10	0.60	0.77	0.91	0.79	10	0.52	0.76	0.91	0.77
15	0.85	0.92	0.99	0.92	15	0.77	0.94	0.99	0.94

Table 3: First simulation study, dependent case: $c = 0.8$. Frequencies of rejection.

Null hypothesis is true, $R = 0$

n	$T_{\mathcal{M}^2,P_1}$	$T_{\mathcal{M}^2,P_2}$	$T_{\mathcal{M}^3,P_1}$	$T_{\mathcal{M}^3,P_2}$
50	0.013	0.052	0.036	0.059
100	0.009	0.059	0.042	0.059

Null hypothesis is false, $R = 0.2, s = 0.5$

n	$T_{\mathcal{M}^2,P_1}$	$T_{\mathcal{M}^2,P_2}$	$T_{\mathcal{M}^3,P_1}$	$T_{\mathcal{M}^3,P_2}$
50	0.17	0.33	0.31	0.38
100	0.42	0.66	0.62	0.69

Table 4: Second simulation study. Frequencies of rejection.

improves the power of the test and the test $T_{\mathcal{M}^3,P_2}$ has the greatest power. In this setting, one should prefer the collection $\mathcal{M}^3$ to $\mathcal{M}^2$. This was previously pointed out in Section 2.5 from a theoretical point of view. Although $T_{\mathcal{M}^3,P_1}$ is conservative, it is a good compromise for practical issues: it is very easy and fast to implement and its performances are good.

2.7 Proofs of Theorem 2.3, Propositions 2.5, 2.9, 2.11, 2.12, and 2.14

Proof of Theorem 2.3. In a nutshell, we shall prove that conditionally to the design $\mathbf{X}$ the distribution of the test T_α is the same as the test introduced by Baraud *et al.* [BHL03]. Hence, we may apply their non asymptotic upper bound for the power.

Distribution of $\phi_m(\mathbf{Y}, \mathbf{X})$. First, we derive the distribution of the test statistic $\phi_m(\mathbf{Y}, \mathbf{X})$ under $\mathbb{P}_\theta$. The distribution of Y conditionally to the set of variables $(X_{V\cup m})$ is of the form

$$Y = \sum_{i \in V\cup m} \theta_i^{V\cup m} X_i + \epsilon^{V\cup m}, \tag{2.32}$$

where the vector $\theta^{V\cup m}$ is constant and $\epsilon^{V\cup m}$ is a zero mean Gaussian variable independent of $X_{V\cup m}$, whose variance is $\mathrm{var}(Y|X_{V\cup m})$. As a consequence, $\|\mathbf{Y} - \Pi_{V\cup m}\mathbf{Y}\|_n^2$ is exactly $\|\Pi_{(V\cup m)^\perp}\boldsymbol{\epsilon}^{V\cup m}\|_n^2$, where $\Pi_{(V\cup m)^\perp}$ denotes the orthogonal projection along the space generated by $(\mathbf{X}_i)_{i\in V\cup m}$.
Using the same decomposition of $\mathbf{Y}$ one simplifies the numerator of $\phi_m(\mathbf{Y}, \mathbf{X})$:

$$\|\Pi_{V\cup m}\mathbf{Y} - \Pi_V\mathbf{Y}\|_n^2 = \left\| \sum_{i\in V\cup m} \theta_i^{V\cup m}(\mathbf{X}_i - \Pi_V\mathbf{X}_i) + \Pi_{V^\perp\cap(V\cup m)}\boldsymbol{\epsilon}^{V\cup m} \right\|_n^2,$$

where $\Pi_{V^\perp \cap (V \cup m)}$ is the orthogonal projection onto the intersection between the space generated by $(\mathbf{X}_i)_{i \in V \cup m}$ and the orthogonal of the space generated by $(\mathbf{X}_i)_{i \in V}$.
For any $i \in m$, let us consider the conditional distribution of X_i with respect to X_V,

$$X_i = \sum_{j \in V} \theta_j^{V,i} X_j + \epsilon_i^V. \tag{2.33}$$

where $\theta_j^{V,i}$ are constants and ϵ_i^V is a zero-mean normal gaussian random variable whose variance is $\mathrm{var}(X_i|X_V)$ and which is independent of X_V. This enables us to express

$$\mathbf{X}_i - \Pi_V \mathbf{X}_i = \Pi_{V^\perp \cap (V \cup m)} \boldsymbol{\epsilon}_i^V, \quad \text{for all } i \in m \ .$$

Therefore, we decompose $\phi_m(\mathbf{Y}, \mathbf{X})$ in

$$\phi_m(\mathbf{Y}, \mathbf{X}) = \frac{N_m \| \Pi_{V^\perp \cap (V \cup m)} \left(\sum_{i \in m} \theta_i^{V \cup m} \boldsymbol{\epsilon}_i^V + \boldsymbol{\epsilon}^{V \cup m} \right) \|_n^2}{D_m \| \Pi_{(V \cup m)^\perp} \boldsymbol{\epsilon}^{V \cup m} \|_n^2}. \tag{2.34}$$

Let us define the random variable $Z_m^{(1)}$ and $Z_m^{(2)}$ where $Z_m^{(1)}$ refers to the numerator of (2.34) divided by N_m and $Z_m^{(2)}$ to the denominator divided by D_m. We now prove that $Z_m^{(1)}$ and $Z_m^{(2)}$ are independent.

The variables $(\boldsymbol{\epsilon}_j^V)_{j \in m}$ are $\sigma\left(\mathbf{X}_{V \cup m}\right)$-measurable as linear combinations of elements in $\mathbf{X}_{V \cup m}$. Moreover, $\boldsymbol{\epsilon}^{V \cup m}$ follows a zero mean normal distribution with covariance matrix $\mathrm{var}(Y|X_{V \cup m}) I_n$ and is independent of $\mathbf{X}_{V \cup m}$. As a consequence, conditionally to $\mathbf{X}_{V \cup m}$, $Z_m^{(1)}$ and $Z_m^{(2)}$ are independent by Cochran's Theorem as they correspond to projections onto two sets orthogonal from each other.

As $\boldsymbol{\epsilon}_j^V$ is a linear combination of the columns of $\mathbf{X}_{V \cup m}$, $Z_m^{(1)}$ follows a non-central χ^2 distribution conditionally to $\mathbf{X}_{V \cup m}$:

$$(Z_m^{(1)}|\mathbf{X}_{V \cup m}) \sim \mathrm{var}(Y|X_{V \cup m}) \chi^2 \left(\frac{\left\| \sum_{j \in m} \theta_j^{V \cup m} \Pi_{(V \cup m) \cap V^\perp} \boldsymbol{\epsilon}_j^V \right\|_n^2}{\mathrm{var}(Y|X_{V \cup m})}, D_m \right).$$

We denote $a_m^2(\mathbf{X}_{V \cup m}) := \frac{\| \sum_{j \in m} \theta_j^{V \cup m} \Pi_{(V \cup m) \cap V^\perp} \boldsymbol{\epsilon}_j^V \|_n^2}{\mathrm{var}(Y|X_{V \cup m})}$ this non-centrality parameter.

Power of T_α conditionally to $\mathbf{X}_{V \cup m}$. Conditionally to $\mathbf{X}_{V \cup m}$ our test statistic $\phi_m(\mathbf{Y}, \mathbf{X})$ is the same as that proposed by Baraud *et al* [BHL03] with $n - d$ data and $\sigma^2 = \mathrm{var}(Y|X_{V \cup m})$. Arguing as in their proof of Theorem 1, there exists some quantity $\bar{\Delta}_m(\delta)$ such that the procedure accepts the hypothesis with probability not larger than $\delta/2$ if $a_m^2(\mathbf{X}_{V \cup m}) > \bar{\Delta}_m(\delta)$:

$$\begin{aligned} \bar{\Delta}_m(\delta) := 2.5 \sqrt{1 + K_m^2(U)} \sqrt{D_m \log\left(\frac{4}{\alpha_m \delta}\right)} \left(1 + \sqrt{\frac{D_m}{N_m}}\right) + \\ 2.5 \left[k_m K_m(U) \vee 5\right] \log\left(\frac{4}{\alpha_m \delta}\right) \left(1 + \frac{2 D_m}{N_m}\right), \end{aligned} \tag{2.35}$$

where $U_m := \log(1/\alpha_m)$, $U := \log(2/\delta)$, $k_m := 2 \exp\left(4 U_m / N_m\right)$, and

$$K_m(u) := 1 + 2\sqrt{\frac{u}{N_m}} + 2 k_m \frac{u}{N_m}.$$

Consequently, we have

$$\mathbb{P}_\theta \left(T_\alpha \leq 0 | \mathbf{X}_{V \cup m}\right) \mathbf{1}\left\{a_m^2(\mathbf{X}_{V \cup m}) \geq \bar{\Delta}_m(\delta)\right\} \leq \delta/2. \tag{2.36}$$

Let derive the distribution of the non-central parameter $a_m(\mathbf{X}_{V \cup m})$. First, we simplify the projection term as $\boldsymbol{\epsilon}_j^V$ is a linear combinations of elements of $\mathbf{X}_{V \cup m}$.

$$\Pi_{(V \cup m) \cap V^\perp} \boldsymbol{\epsilon}_j^V = \Pi_{V \cup m} \boldsymbol{\epsilon}_j^V - \Pi_V \boldsymbol{\epsilon}_j^V = \Pi_{V^\perp} \boldsymbol{\epsilon}_j^V.$$

Let us define κ_m^2 as

$$\kappa_m^2 := \frac{\mathrm{var}\left(\sum_{j\in m}\theta_j^{V\cup m}\epsilon_j^V\right)}{\mathrm{var}(Y|X_{V\cup m})}.$$

As the variable $\sum_{j\in m}\theta_j^{V\cup m}\boldsymbol{\epsilon}_j^V$ is independent of $\mathbf{X}_V$, and as almost surely the dimension of the vector space generated by $\mathbf{X}_V$ is d, we get

$$\frac{\left\|\sum_{j\in m}\theta_j^{V\cup m}\Pi_{V^\perp}\boldsymbol{\epsilon}_j^V\right\|_n^2}{\mathrm{var}(Y|X_{V\cup m})} \sim \kappa_m^2\chi^2(n-d).$$

Hence, applying for instance Lemma 1 in [LM00], we get

$$\mathbb{P}_\theta\left[\frac{a_m^2(\mathbf{X}_{V\cup m})}{\kappa_m^2} \geq (n-d) - 2\sqrt{(n-d)U}\right] \leq \delta/2.$$

Let gather (2.36) with this last bound. If

$$\kappa_m^2 \geq \Delta'_m(\delta) := \frac{\bar{\Delta}_m(\delta)}{(n-d)\left(1-2\sqrt{\frac{U}{n-d}}\right)}\,, \tag{2.37}$$

then it holds that

$$\begin{aligned}
\mathbb{P}_\theta(T_\alpha \leq 0) \quad &\leq \quad \mathbb{P}_\theta\left(T_\alpha \leq 0, a_m^2(\mathbf{X}_{V\cup m}) > \bar{\Delta}_m(\delta)\right) + \mathbb{P}_\theta\left[a_m^2(\mathbf{X}_{V\cup m}) \leq \bar{\Delta}_m(\delta)\right] \\
&\leq \quad \mathbb{E}_\theta\left\{\mathbb{P}_\theta\left[T_\alpha \leq 0, a_m^2(\mathbf{X}_{V\cup m}) > \bar{\Delta}_m(\delta)|\mathbf{X}_{V\cup m}\right]\right\} + \\
&\qquad \mathbb{P}_\theta\left[\frac{a_m^2(\mathbf{X}_{V\cup m})}{\kappa_m^2} \geq (n-d) - 2\sqrt{(n-d)U}\right] \\
&\leq \delta.
\end{aligned}$$

Computation of κ_m^2. Let us now compute the quantity κ_m^2 in order to simplify Condition (2.37). Let first express $\mathrm{var}(Y|X_V)$ in terms of $\mathrm{var}(Y|X_{m\cup V})$ using the decomposition (2.32) of Y.

$$\begin{aligned}
\mathrm{var}(Y|X_V) \quad &= \quad \mathrm{var}\left(\sum_{j\in V\cup m}\theta_j^{V\cup m}X_j + \epsilon^{V\cup m}\,|X_V\right) \\
&= \quad \mathrm{var}\left(\sum_{j\in V\cup m}\theta_j^{V\cup m}X_j\,|X_V\right) + \mathrm{var}\left(\epsilon^{V\cup m}\,|X_V\right) \\
&= \quad \mathrm{var}\left(\sum_{j\in V\cup m}\theta_j^{V\cup m}X_j\,|X_V\right) + \mathrm{var}\left(Y\,|X_{V\cup m}\right), \qquad (2.38)
\end{aligned}$$

as $\epsilon^{V\cup m}$ is independent of $X_{V\cup m}$. Now using the definition of ϵ_j^V in (2.33), it turns out that

$$\begin{aligned}
\mathrm{var}\left(\sum_{j\in V\cup m}\theta_j^{V\cup m}X_j|X_V\right) \quad &= \quad \mathrm{var}\left(\sum_{j\in m}\theta_j^{V\cup m}X_j|X_V\right) \\
&= \quad \mathrm{var}\left(\sum_{j\in m}\theta_j^{V\cup m}\epsilon_j^V\,|X_V\right) \\
&= \quad \mathrm{var}\left(\sum_{j\in m}\theta_j^{V\cup m}\epsilon_j^V\right), \qquad (2.39)
\end{aligned}$$

as the $(\epsilon_j^V)_{j\in m}$ are independent of X_V. Gathering formulae (2.38) and (2.39), we get

$$\kappa_m^2 = \frac{\mathrm{var}(Y|X_V) - \mathrm{var}(Y|X_{V\cup m})}{\mathrm{var}(Y|X_{V\cup m})}. \tag{2.40}$$

Under Assumption $H_{\mathcal{M}}$, $U_m \leq N_m/10$ for all $m \in \mathcal{M}$ and $U \leq N_m/21$. Hence, the terms U/N_m, U_m/N_m, k_m, and $K_m(U)$ behave like constants and it follows from (2.37) that $\Delta'(m) \leq \Delta(m)$, which concludes the proof. □

Proof of Proposition 2.5. We first recall the classical upper bound for the binomial coefficient (see for instance (2.9) in [Mas07]).

$$\log|\mathcal{M}(k,p)| = \log\binom{p}{k} \leq k\log\left(\frac{ep}{k}\right).$$

As a consequence, $\log(1/\alpha_m) \leq \log(1/\alpha) + k\log\left(\frac{ep}{k}\right)$. Assumption (2.14) with $L = 21$ therefore implies Hypothesis $H_{\mathcal{M}}$. Hence, we are in position to apply the second result of Theorem 2.3. Moreover, the assumption on n implies that $n \geq 21k$ and D_m/N_m is thus smaller than $1/20$ for any model m in $\mathcal{M}(k,p)$. Formula (2.12) in Theorem 2.3 then translates into

$$\triangle(m) \leq \frac{(1+\sqrt{0.05})L_1\left(\sqrt{k^2\log\left(\frac{ep}{k}\right)} + \sqrt{k\log\left(\frac{2}{\alpha\delta}\right)}\right) + 1.1L_2\left(k\log\left(\frac{ep}{k}\right) + \log\left(\frac{2}{\alpha\delta}\right)\right)}{n},$$

and it follows that Proposition 2.5 holds.

□

Proof of Proposition 2.9. We fix the constant L in Hypothesis (2.20) to be $21\log(4e) \vee C_2\log(4)$ where the universal constant C_2 is defined later in the proof. This choice of constants allows the procedure $\left[\sup_{1\leq i\leq p}\phi_{\{i\},\alpha/(2p)}\right]$ to satisfy Hypothesis $H_{\mathcal{M}}$. An argument similar to the proof of Proposition 2.5 allows to show easily that there exists a universal constant C such that if we set

$$\rho_1'^2 \quad := \quad \frac{C\left(\log(p) + \log\left(\frac{4}{\alpha\delta}\right)\right)}{n} = \frac{C}{n}\log\left(\frac{4p}{\alpha\delta}\right), \tag{2.41}$$

then $\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq \rho_1'^2$ implies that $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1-\delta$. Here, the factor 4 in the logarithm comes from the fact that some weights α_m equal $\alpha/(2p)$.

Let ρ^2 and λ^2 be two positive numbers such that $\frac{\lambda^2}{\mathrm{var}(Y)-\lambda^2} = \rho^2$ and let $\theta \in \Theta[1,p]$ such that $\|\theta\|^2 = \lambda^2$. As $\mathrm{corr}(X_i, X_j) = c$ for any $i \neq j$, it follows that $\mathrm{var}(X_{p+1}) = c + \frac{1-c}{p}$ and $\mathrm{cov}(Y, X_{p+1})^2 = \|\theta\|^2\left[c + \frac{1-c}{p}\right]^2$.

$$\frac{\mathrm{var}(Y) - \mathrm{var}(Y|X_{p+1})}{\mathrm{var}(Y|X_{p+1})} = \frac{(c+(1-c)/p)\,\lambda^2}{\mathrm{var}(Y) - (c+(1-c)/p)\,\lambda^2}.$$

We now apply Theorem 2.3 to $\phi_{\{p+1\},\alpha/2}$ under $H_{\mathcal{M}}$. There exists a universal constant C_2 such that $\mathbb{P}_\theta\left(\phi_{\{p+1\},\alpha/2} > 0\right) \geq 1-\delta$ if

$$\frac{(c+(1-c)/p)\,\lambda^2}{\mathrm{var}(Y) - (c+(1-c)/p)\,\lambda^2} \geq \frac{C_2}{n}\log\left(\frac{4}{\alpha\delta}\right).$$

This last condition is implied by

$$\frac{c\lambda^2}{\mathrm{var}(Y) - c\lambda^2} \geq \frac{C_2}{n}\log\left(\frac{4}{\alpha\delta}\right),$$

which is equivalent to

$$\frac{\lambda^2}{\mathrm{var}(Y)} \geq \frac{C_2}{cn + cC_2\log\left(\frac{4}{\alpha\delta}\right)}\log\left(\frac{4}{\alpha\delta}\right). \tag{2.42}$$

Let us assume that $c \geq \log\left(\frac{4}{\alpha\delta}\right)/\log\left(\frac{4p}{\alpha\delta}\right)$. As $n \geq 2C_2\log\left(\frac{4p}{\alpha\delta}\right)$ (Hypothesis (2.20) and definition of L), $nc \geq 2C_2\log\left(\frac{4}{\alpha\delta}\right)$. As a consequence, Condition (2.42) is implied by:

$$\rho^2 \geq \frac{2C_2}{nc}\log\left(\frac{4}{\alpha\delta}\right). \tag{2.43}$$

Combining (2.41) and (2.43) allows to conclude that $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1-\delta$ if

$$\rho^2 \geq \frac{L}{n}\left(\log\left(\frac{4p}{\alpha\delta}\right) \bigwedge \frac{1}{c}\log\left(\frac{4}{\alpha\delta}\right)\right).$$

□

Proof of Proposition 2.11. We fix the constant L to $42\log(80)$ in Hypothesis (2.23). It follows that (2.23) implies

$$n \geq 42\left(\log\left(\frac{40}{\alpha}\right) \vee \log\left(\frac{2}{\delta}\right)\right). \tag{2.44}$$

First, we check that the test T_α satisfies Condition $H_{\mathcal{M}}$. As the dimension of each model is smaller than $n/2$, for any model m in $\mathcal{M}$, N_m is larger than $n/2$. Moreover, for any model m in $\mathcal{M}$, α_m is larger than $\alpha/(2|\mathcal{M}|)$ and $|\mathcal{M}|$ is smaller than $n/2$. As a consequence, the first condition of $H_{\mathcal{M}}$ is implied by the inequality

$$n \geq 20\log\left(\frac{n}{\alpha}\right). \tag{2.45}$$

Hypothesis (2.44) implies that $n/2 \geq 20\log\left(\frac{40}{\alpha}\right)$. Besides, for any $n > 0$ it holds that $n/2 \geq 20\log\left(\frac{n}{40}\right)$. Combining these two lower bounds enables to obtain (2.45). The second condition of $H_{\mathcal{M}}$ holds if $n \geq 42\log\left(\frac{2}{\delta}\right)$ which is a consequence of hypothesis (2.44).

Let first consider the case $n < 2p$ and let apply Theorem 2.3 under Hypothesis $H_{\mathcal{M}}$ to T_α. $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1-\delta$ for all $\theta \in \mathbb{R}^p$ such that

$$\exists i \in \{1,\dots,[n/2]\}, \frac{\mathrm{var}(Y) - \mathrm{var}\left(Y|X_{m_i}\right)}{\mathrm{var}\left(Y|X_{m_i}\right)} \geq C\frac{\sqrt{i\log\left(\frac{2[n/2]}{\alpha\delta}\right)} + \log\left(\frac{2[n/2]}{\alpha\delta}\right)}{n}, \tag{2.46}$$

where C is an universal constant. Let θ be an element of $\mathcal{E}_a(R)$ that satisfies

$$\|\theta\|^2 \geq (1+C)\left(\mathrm{var}(Y|X_{m_i}) - \mathrm{var}(Y|X)\right) + (1+C)\mathrm{var}(Y|X)\frac{\sqrt{i\log\left(\frac{n}{\alpha\delta}\right)} + \log\left(\frac{n}{\alpha\delta}\right)}{n},$$

for some $1 \leq i \leq [n/2]$. By Hypothesis (2.23), it holds that

$$\frac{\sqrt{i\log\left(\frac{n}{\alpha\delta}\right)} + \log\left(\frac{n}{\alpha\delta}\right)}{n} \leq 1\ ,$$

for any i between 1 and $[n/2]$. It is then straightforward to check that θ satisfies (2.46).
As θ belongs to the set $\mathcal{E}_a(R)$,

$$\begin{aligned}\mathrm{var}(Y|X_{m_i}) - \mathrm{var}(Y|X) &= a_{i+1}^2\mathrm{var}(Y|X)\sum_{j=i+1}^{p}\frac{\mathrm{var}(Y|X_{m_{j-1}}) - \mathrm{var}(Y|X_{m_j})}{a_{i+1}^2\mathrm{var}(Y|X)}\\ &\leq a_{i+1}^2\mathrm{var}(Y|X)R^2.\end{aligned}$$

Hence, if θ belongs to $\mathcal{E}_a(R)$ and satisfies

$$\|\theta\|^2 \geq (1+C)\mathrm{var}(Y|X)\left[\left(a_{i+1}^2R^2 + \frac{\sqrt{i\log\left(\frac{n}{\alpha\delta}\right)}}{n}\right) + \frac{1}{n}\log\left(\frac{n}{\alpha\delta}\right)\right],$$

then $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$. Gathering this condition for any i between 1 and $[n/2]$ allows to conclude that if θ satisfies

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq (1+C)\left[\inf_{1\leq i\leq[n/2]}\left(a_{i+1}^2R^2 + \frac{\sqrt{i\log\left(\frac{n}{\alpha\delta}\right)}}{n}\right) + \frac{1}{n}\log\left(\frac{n}{\alpha\delta}\right)\right],$$

then $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$.

Let us now turn to the case $n \geq 2p$. Let us consider T_α as the supremum of $p-1$ tests of level $\alpha/2(p-1)$ and one test of level $\alpha/2$. By considering the $p-1$ firsts tests, we obtain as in the previous case that $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$ if

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq (1+C)\left[\inf_{1\leq i\leq (p-1)}\left(a_{i+1}^2 R^2 + \frac{\sqrt{i\log\left(\frac{p}{\alpha\delta}\right)}}{n}\right) + \frac{1}{n}\log\left(\frac{p}{\alpha\delta}\right)\right].$$

On the other hand, using the last test statistic $\phi_{\mathcal{I},\alpha/2}$, $\mathbb{P}_\theta(T_\alpha \leq 0) \leq \delta$ if

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq C\frac{\sqrt{p\log\left(\frac{2}{\alpha\delta}\right)} + \log\left(\frac{2}{\alpha\delta}\right)}{n}.$$

Gathering these two conditions allows to prove (2.25). □

Proof of Proposition 2.12. The approach behind this proof is similar to the one for Proposition 2.11. We fix the constant L in Assumption 2.26, as in the previous proof. Hence, the collection of models $\mathcal{M}$ and the weights α_m satisfy hypothesis $H_{\mathcal{M}}$ as in the previous proof.

Let us give a sharper upper bound on $|\mathcal{M}|$:

$$|\mathcal{M}| \leq 1 + \log(n/2 \wedge p)/\log(2) \leq \log(n \wedge 2p)/\log(2). \tag{2.47}$$

We deduce from (2.47) that there exists a constant $L(\alpha,\delta)$ only depending on α and δ such that for all $m \in \mathcal{M}$,

$$\log\left(\frac{1}{\alpha_m \delta}\right) \leq L(\alpha,\delta)\log\log(n \wedge p).$$

First, let us consider the case $n < 2p$. We apply Theorem 2.3 under Assumption $H_{\mathcal{M}}$. As in the proof of Proposition 2.11, we obtain that $\mathbb{P}_\theta(T_\alpha > 0) \geq 1-\delta$ if

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq L(\alpha,\delta)\left[\inf_{i\in\{2^j, j\geq 0\}\cap\{1,\ldots,[n/2]\}}\left(R^2(i+1)^{-2s} + \frac{\sqrt{i\log\log n}}{n}\right) + \frac{\log\log n}{n}\right].$$

It is worth noting that $R^2 i^{-2s} \leq \frac{\sqrt{i\log\log n}}{n}$ if and only if

$$i \geq i^* = \left(\frac{R^2 n}{\sqrt{\log\log n}}\right)^{2/(1+4s)}.$$

Under the assumption on R, i^* is larger than one. Let us distinguish between two cases. If there exists i' in $\{2^j, j\geq 0\}\cap\{1,\ldots,[n/2]\}$ such that $i^* \leq i'$, one can take $i' \leq 2i^*$ and then

$$\begin{aligned}\inf_{i\in\{2^j, j\geq 0\}\cap\{1,\ldots,[n/2]\}}\left(R^2 i^{-2s} + \frac{\sqrt{i\log\log n}}{n}\right) &\leq 2\frac{\sqrt{i'\log\log n}}{n} \\ &\leq 2\sqrt{2}R^{2/(1+4s)}\left(\frac{\sqrt{\log\log n}}{n}\right)^{4s/(1+4s)}. \end{aligned}\tag{2.48}$$

Else, we take $i' \in \{2^j, j\geq 0\}\cap\{1,\ldots,[n/2]\}$ such that $n/4 \leq i' \leq n/2$. Since $i' \leq (i^* \wedge n/2)$ we obtain that

$$\inf_{i\in\{2^j, j\geq 0\}\cap\{1,\ldots,[n/2]\}}\left(R^2 i^{-2s} + \frac{\sqrt{i\log\log n}}{n}\right) \leq 2R^2 i'^{-2s} \leq 2R^2\left(\frac{n}{2}\right)^{-2s}. \tag{2.49}$$

Gathering inequalities (2.48) and (2.49) allows to prove (2.27).

We now turn to the case $n \geq 2p$. As in the proof of Proposition 2.11, we divide the proof into two parts: first we give an upper bound of the power for the $|\mathcal{M}|-1$ first tests which define T_α and then we give an upper bound for the last test $\phi_{\mathcal{I},\alpha/2}$. Combining these two inequalities allows us to prove (2.28). □

Proof of Proposition 2.14. We fix the constant L in the assumption as in the two previous proofs. We first note that the assumption on R^2 implies that $D^* \geq 2$. As N_m is larger than $n/2$, the $\phi_{m_{D^*}}$ test clearly satisfies Condition $H_{\mathcal{M}}$. As a consequence, we may apply Theorem 2.3. Hence, $\mathbb{P}_\theta(T^*_\alpha \leq 0) \leq \delta$ for any θ such that

$$\frac{\mathrm{var}(Y) - \mathrm{var}(Y|X_{m_{D^*}})}{\mathrm{var}(Y|X_{m_{D^*}})} \geq L(\alpha,\delta)\frac{\sqrt{D^*}}{n}. \tag{2.50}$$

Now, we use the same sketch as in the proof of Proposition 2.11. For any $\theta \in \mathcal{E}_a(R)$, Condition (2.50) is equivalent to:

$$\|\theta\|^2 \geq (\mathrm{var}(Y|X_{m_{D^*}}) - \mathrm{var}(Y|X))\left(1 + L(\alpha,\delta)\frac{\sqrt{D^*}}{n}\right) + \mathrm{var}(Y|X)L(\alpha,\delta)\frac{\sqrt{D^*}}{n}. \tag{2.51}$$

Moreover, as θ belongs to $\mathcal{E}_a(R)$,

$$\mathrm{var}(Y|X_{m_{D^*}}) - \mathrm{var}(Y|X) \leq a^2_{D^*+1}R^2\mathrm{var}(Y|X) \leq a^2_{D^*}\mathrm{var}(Y|X)R^2.$$

As $\sqrt{D^*}/n$ is smaller than one, Condition (2.51) is implied by

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq (1 + L(\alpha,\delta))\left(a^2_{D^*}R^2 + \frac{\sqrt{D^*}}{n}\right).$$

As $a^2_{D^*}R^2$ is smaller than $\frac{\sqrt{D^*}}{n}$ which is smaller $\sup_{1\leq i\leq p}\left[\frac{\sqrt{i}}{n} \wedge a_i^2R^2\right]$, it turns out that $\mathbb{P}_\theta(T^*_\alpha = 0) \leq \delta$ for any θ belonging to $\mathcal{E}_a(R)$ such that

$$\frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq 2(1 + L(\alpha,\delta)) \sup_{1\leq i\leq p}\left[\frac{\sqrt{i}}{n} \wedge a_i^2R^2\right].$$

□

2.8 Proofs of Theorem 2.7, Propositions 2.4, 2.6, 2.8, 2.10, 2.13, 2.15, and 2.16

Throughout this section, we shall use the notations $\eta := 2(1-\alpha-\delta)$ and $\mathcal{L}(\eta) := \frac{\log(1+2\eta^2)}{2}$.

Proof of Theorem 2.7. This proof follows the general method for obtaining lower bounds described in Section 7.1 in Baraud [Bar02b]. We first remind the reader of the main arguments of the approach applied to our model. Let ρ be some positive number and μ_ρ be some probability measure on

$$\Theta[k,p,\rho] := \left\{\theta \in \Theta[k,p], \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = \rho\right\}.$$

We define $\mathbb{P}_{\mu_\rho} = \int \mathbb{P}_\theta \mathrm{d}\mu_\rho(\theta)$ and Φ_α the set of level-α tests of the hypothesis "$\theta = 0$". Then,

$$\begin{aligned}
\beta_I(\Theta[k,p,\rho]) &\geq \inf_{\phi_\alpha\in\Phi_\alpha} \mathbb{P}_{\mu_\rho}[\phi_\alpha = 0] \\
&\geq 1 - \alpha - \sup_{A,\ \mathbb{P}_0(A)\leq\alpha} |\mathbb{P}_{\mu_\rho}(A) - \mathbb{P}_0(A)| \\
&\geq 1 - \alpha - \frac{1}{2}\|\mathbb{P}_{\mu_\rho} - \mathbb{P}_0\|_{TV},
\end{aligned} \tag{2.52}$$

where $\|\mathbb{P}_{\mu_\rho} - \mathbb{P}_0\|_{TV}$ denotes the total variation norm between the probabilities $\mathbb{P}_{\mu_\rho}$ and $\mathbb{P}_0$. If we suppose that $\mathbb{P}_{\mu_\rho}$ is absolutely continuous with respect to $\mathbb{P}_0$, we can upper bound the norm in total variation between these two probabilities as follows. We define

$$L_{\mu_\rho}(\mathbf{Y},\mathbf{X}) := \frac{\mathrm{d}\mathbb{P}_{\mu_\rho}}{\mathrm{d}\mathbb{P}_0}(\mathbf{Y},\mathbf{X}).$$

Then, we get the upper bound

$$\begin{aligned}\|\mathbb{P}_{\mu_\rho} - \mathbb{P}_0\|_{TV} &= \int |L_{\mu_\rho}(\mathbf{Y},\mathbf{X}) - 1| \mathrm{d}\mathbb{P}_0(\mathbf{Y},\mathbf{X}) \\ &\leq \left(\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right] - 1\right)^{1/2}.\end{aligned}$$

Thus, we deduce from (2.52) that

$$\beta_I(\Theta[k,p,\rho]) \geq 1 - \alpha - \frac{1}{2}\left(\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right] - 1\right)^{1/2}.$$

If we find a number $\rho^* = \rho^*(\eta)$ such that

$$\log\left(\mathbb{E}_0\left[L^2_{\mu_{\rho^*}}(\mathbf{Y},\mathbf{X})\right]\right) \leq \mathcal{L}(\eta), \tag{2.53}$$

then for any $\rho \leq \rho^*$,

$$\beta_I(\Theta[k,p,\rho]) \geq 1 - \alpha - \frac{\eta}{2} = \delta.$$

To apply this method, we first have to define a suitable prior μ_ρ on $\Theta[k,p,\rho]$. Let $\widehat{m}$ be some random variable uniformly distributed over $\mathcal{M}(k,p)$ and for each $m \in \mathcal{M}(k,p)$, let $\epsilon^m = (\epsilon^m_j)_{j\in m}$ be a sequence of independent Rademacher random variables. We assume that for all $m \in \mathcal{M}(k,p)$, ϵ^m and $\widehat{m}$ are independent. Let ρ be given and μ_ρ be the distribution of the random variable $\widehat{\theta} = \sum_{j\in\widehat{m}} \lambda \epsilon^{\widehat{m}}_j e_j$ where

$$\lambda^2 := \frac{\mathrm{var}(Y)\rho^2}{k(1+\rho^2)},$$

and where $(e_j)_{j\in\mathcal{I}}$ is the orthonormal family of vectors of $\mathbb{R}^p$ defined by

$$(e_j)_i = 1 \text{ if } i = j \text{ and } (e_i)_j = 0 \text{ otherwise.}$$

Straightforwardly, μ_ρ is supported by $\Theta[k,p,\rho]$. For any m in $\mathcal{M}(k,p)$ and any vector $(\zeta^m_j)_{j\in m}$ with values in $\{-1;1\}$, let $\mu_{m,\zeta^m,\rho}$ be the Dirac measure on $\sum_{j\in m}\lambda\zeta^m_j e_j$. For any m in $\mathcal{M}(k,p)$, $\mu_{m,\rho}$ denotes the distribution of the random variable $\sum_{j\in m}\lambda\zeta^m_j e_j$ where (ζ^m_j) is a sequence of independent Rademacher random variables. These definitions easily imply

$$L_{\mu_\rho}(\mathbf{Y},\mathbf{X}) = \frac{1}{\binom{p}{k}}\sum_{m\in\mathcal{M}(k,p)} L_{\mu_{m,\rho}}(\mathbf{Y},\mathbf{X}) = \frac{1}{2^k\binom{p}{k}}\sum_{m\in\mathcal{M}(k,p)}\sum_{\zeta^m\in\{-1,1\}^k} L_{\mu_{m,\zeta^m\rho}}(\mathbf{Y},\mathbf{X}).$$

We aim at bounding the quantity $\mathbb{E}_0(L^2_{\mu_\rho})$ and obtaining an inequality of the form (2.53). First, we work out $L_{\mu_{m,\zeta^m,\rho}}$:

$$\begin{aligned}L_{\mu_{m,\zeta^m,\rho}}(\mathbf{Y},\mathbf{X}) &= \left[\left(\frac{1}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right)^{n/2}\exp\left(-\frac{\|\mathbf{Y}\|_n^2}{2}\frac{\lambda^2 k}{\mathrm{var}(Y)(\mathrm{var}(Y)-\lambda^2 k)}\right.\right. \\ &+ \left.\left.\lambda\sum_{j\in m}\zeta^m_j\frac{<\mathbf{Y},\mathbf{X}_j>_n}{\mathrm{var}(Y)-\lambda^2 k} - \lambda^2\sum_{j,j'\in m}\zeta^m_j\zeta^m_{j'}\frac{<\mathbf{X}_j,\mathbf{X}_{j'}>_n}{2(\mathrm{var}(Y)-\lambda^2 k)}\right)\right],\end{aligned} \tag{2.54}$$

where $<.>_n$ refers to the canonical inner product in $\mathbb{R}^n$.

Let us fix m_1 and m_2 in $\mathcal{M}(k,p)$ and two vectors ζ^1 and ζ^2 respectively associated to m_1 and m_2. We aim at computing the quantity $\mathbb{E}_0\left(L_{\mu_{m_1,\zeta^1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2,\zeta^2,\rho}}(\mathbf{Y},\mathbf{X})\right)$. First, we decompose the set $m_1\cup m_2$ into four sets (which possibly are empty): $m_1\setminus m_2$, $m_2\setminus m_1$, m_3, and m_4, where m_3 and m_4 are defined by:

$$\begin{aligned}m_3 &:= \left\{j\in m_1\cap m_2 | \zeta^1_j = \zeta^2_j\right\} \\ m_4 &:= \left\{j\in m_1\cap m_2 | \zeta^1_j = -\zeta^2_j\right\}.\end{aligned}$$

For the sake of simplicity, we reorder the elements of $m_1 \cup m_2$ from 1 to $|m_1 \cup m_2|$ such that the first elements belong to $m_1 \setminus m_2$, then to $m_2 \setminus m_1$ and so on. Moreover, we define the vector $\zeta \in \mathbb{R}^{|m_1 \cup m_2|}$ such that $\zeta_j = \zeta_j^1$ if $j \in m_1$ and $\zeta_j = \zeta_j^2$ if $j \in m_2 \setminus m_1$. Using these notations, we compute the expectation of $L_{m_1,\zeta^1,\rho}(\mathbf{Y},\mathbf{X})L_{m_2,\zeta^2,\rho}(\mathbf{Y},\mathbf{X})$.

$$\mathbb{E}_0\left(L_{\mu_{m_1,\zeta^1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2,\zeta^2,\rho}}(\mathbf{Y},\mathbf{X})\right) = \left(\frac{1}{\mathrm{var}(Y)(1-\frac{\lambda^2 k}{\mathrm{var}(Y)})^2}\right)^{n/2} |A|^{-n/2}\ , \tag{2.55}$$

where $|.|$ refers to the determinant and A is a symmetric square matrix of size $|m_1 \cup m_2|+1$ such that:

$$A[1,j] := \begin{cases} \frac{\mathrm{var}(Y)+\lambda^2 k}{\mathrm{var}(Y)(\mathrm{var}(Y)-\lambda^2 k)} & \text{if} \quad j=1 \\ -\frac{\lambda\zeta_{j-1}}{\mathrm{var}(Y)-\lambda^2 k} & \text{if} \quad (j-1)\in m_1 \triangle m_2 \\ -2\frac{\lambda\zeta_{j-1}}{\mathrm{var}(Y)-\lambda^2 k} & \text{if} \quad (j-1)\in m_3 \\ 0 & \text{if} \quad (j-1)\in m_4\ , \end{cases}$$

where $m_1 \triangle m_2$ refers to $(m_1 \cup m_2) \setminus (m_1 \cap m_2)$. For any $i>1$ and $j>1$, A satisfies

$$A[i,j] := \begin{cases} \lambda^2\frac{\zeta_{i-1}\zeta_{j-1}}{\mathrm{var}(Y)-\lambda^2 k}+\delta_{i,j} & \text{if} \quad (i-1,j-1)\in (m_1\setminus m_2)\times m_1 \\ \lambda^2\frac{\zeta_{i-1}\zeta_{j-1}}{\mathrm{var}(Y)-\lambda^2 k}+\delta_{i,j} & \text{if} \quad (i-1,j-1)\in (m_2\setminus m_1)\times (m_2\setminus m_1 \cup m_3) \\ -\lambda^2\frac{\zeta_{i-1}\zeta_{j-1}}{\mathrm{var}(Y)-\lambda^2 k} & \text{if} \quad (i-1,j-1)\in (m_2\setminus m_1)\times m_4 \\ 2\lambda^2\frac{\zeta_{i-1}\zeta_{j-1}}{\mathrm{var}(Y)-\lambda^2 k}+\delta_{i,j} & \text{if} \quad (i-1,j-1)\in [m_3\times m_3]\cup[m_4\times m_4] \\ 0 & \text{else,} \end{cases},$$

where $\delta_{i,j}$ is the indicator function of $i=j$.

After some linear transformation on the lines of the matrix A, it is possible to express its determinant into

$$|A| = \frac{\mathrm{var}(Y)+\lambda^2 k}{\mathrm{var}(Y)(\mathrm{var}(Y)-\lambda^2 k)}\left|I_{|m_1\cup m_2|}+C\right|,$$

where $I_{|m_1\cup m_2|}$ is the identity matrix of size $|m_1 \cup m_2|$. C is a symmetric matrix of size $|m_1 \cup m_2|$ such that for any (i,j),

$$C[i,j] = \zeta_i\zeta_j D[i,j]$$

and D is a block symmetric matrix defined by

$$D := \begin{bmatrix} \frac{\lambda^4 k}{\mathrm{var}^2(Y)-\lambda^4k^2} & \frac{-\lambda^2\mathrm{var}(Y)}{\mathrm{var}^2(Y)-\lambda^4k^2} & \frac{-\lambda^2}{\mathrm{var}(Y)+\lambda^2k} & \frac{\lambda^2}{\mathrm{var}(Y)-\lambda^2k} \\ \frac{-\lambda^2\mathrm{var}(Y)}{\mathrm{var}^2(Y)-\lambda^4k^2} & \frac{\lambda^4 k}{\mathrm{var}^2(Y)-\lambda^4k^2} & \frac{-\lambda^2}{\mathrm{var}(Y)+\lambda^2k} & \frac{-\lambda^2}{\mathrm{var}(Y)-\lambda^2k} \\ \frac{-\lambda^2}{\mathrm{var}(Y)+\lambda^2k} & \frac{-\lambda^2}{\mathrm{var}(Y)+\lambda^2k} & \frac{-2\lambda^2}{\mathrm{var}(Y)+\lambda^2k} & 0 \\ \frac{\lambda^2}{\mathrm{var}(Y)-\lambda^2k} & \frac{-\lambda^2}{\mathrm{var}(Y)-\lambda^2k} & 0 & \frac{2\lambda^2}{\mathrm{var}(Y)-\lambda^2k} \end{bmatrix}.$$

Each block corresponds to one of the four previously defined subsets of $m_1 \cup m_2$ (i.e. $m_1 \setminus m_2$, $m_2 \setminus m_1$, m_3, and m_4). The matrix D is of rank at most four. By computing its non-zero eigenvalues, it is then straightforward to derive the determinant of A

$$|A| = \frac{\left[\mathrm{var}(Y)-\lambda^2(2|m_3|-|m_1\cap m_2|)\right]^2}{\mathrm{var}(Y)(\mathrm{var}(Y)-\lambda^2k)^2}.$$

Gathering this equality with (2.55) yields

$$\mathbb{E}_0\left(L_{\mu_{m_1,\zeta^1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2,\zeta^2,\rho}}(\mathbf{Y},\mathbf{X})\right) = \left[\frac{1}{1-\frac{\lambda^2(2|m_3|-|m_1\cap m_2|)}{\mathrm{var}(Y)}}\right]^n. \tag{2.56}$$

Then, we take the expectation with respect to ζ^1, ζ^2, m_1 and m_2. When m_1 and m_2 are fixed the expression (2.56) depends on ζ^1 and ζ^2 only towards the cardinality of m_3. As ζ^1 and ζ^2 correspond to

independent Rademacher variables, the random variable $2|m_3| - |m_1 \cap m_2|$ follows the distribution of Z, a sum of $|m_1 \cap m_2|$ independent Rademacher variables and

$$\mathbb{E}_0(L_{\mu_{m_1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2,\rho}}(\mathbf{Y},\mathbf{X})) = \mathbb{E}_0\left[\frac{1}{1-\frac{\lambda^2 Z}{\mathrm{var}(Y)}}\right]^n. \tag{2.57}$$

When Z is non-positive, this expression is smaller than one. Alternatively, when Z is non-negative:

$$\begin{aligned}\left[\frac{1}{1-\frac{\lambda^2 Z}{\mathrm{var}(Y)}}\right]^n &= \exp\left(n\log\left(\frac{1}{1-\frac{\lambda^2 Z}{\mathrm{var}(Y)}}\right)\right)\\ &\leq \exp\left[n\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 Z}{\mathrm{var}(Y)}}\right]\\ &\leq \exp\left[n\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right],\end{aligned}$$

as $\log(1+x) \leq x$ and as Z is smaller than k. We define an event $\mathbb{A}$ such that $\{Z > 0\} \subset \mathbb{A} \subset \{Z \geq 0\}$ and $\mathbb{P}(\mathbb{A}) = \frac{1}{2}$. This is always possible as the random variable Z is symmetric. As a consequence, on the event $\mathbb{A}^c$, the quantity (2.57) is smaller or equal to one. All in all, we bound (2.57) by:

$$\mathbb{E}_0(L_{\mu_{m_1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{m_2,\rho}}(\mathbf{Y},\mathbf{X})) \leq \frac{1}{2} + \mathbb{E}_0\left[\mathbf{1}_{\mathbb{A}}\exp\left[n\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right]\right], \tag{2.58}$$

where $\mathbf{1}_{\mathbb{A}}$ is the indicator function of the event $\mathbb{A}$. We now apply Hölder's inequality with a parameter $v \in]0;1]$, which will be fixed later.

$$\begin{aligned}\mathbb{E}_0\left[\mathbf{1}_{\mathbb{A}}\exp\left[n\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right]\right] &\leq \mathbb{P}(\mathbb{A})^{1-v}\left[\mathbb{E}_0\exp\left(\frac{n}{v}\frac{\frac{\lambda^2 Z}{\mathrm{var}(Y)}}{1-\frac{\lambda^2 k}{\mathrm{var}(Y)}}\right)\right]^v\\ &\leq \left(\frac{1}{2}\right)^{1-v}\left[\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)\right]^{|m_1\cap m_2|v}.\end{aligned} \tag{2.59}$$

Gathering inequalities (2.58) and (2.59) yields

$$\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right] \leq \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v}\frac{1}{\binom{p}{k}^2}\sum_{m_1,m_2\in\mathcal{M}(k,p)}\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^{|m_1\cap m_2|v}.$$

Following the approach of Baraud [Bar02b] in Section 7.2, we note that if m_1 and m_2 are taken uniformly and independently in $\mathcal{M}(k,p)$, then $|m_1 \cap m_2|$ is distributed as a Hypergeometric distribution with parameters p, k, and k/p. Thus, we derive that

$$\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right] \leq \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v}\mathbb{E}\left(\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^{vT}\right) \tag{2.60}$$

where T is a random variable distributed according to a Hypergeometric distribution with parameters p, k and k/p. We know from Aldous (p.173) [Ald85] that T has the same distribution as the random variable $\mathbb{E}(W|\mathcal{B}_p)$ where W is binomial random variable of parameters k, k/p and $\mathcal{B}_p$ some suitable σ-algebra. By a convexity argument, we then upper bound (2.60).

$$\begin{aligned}\mathbb{E}_0\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right] &\leq \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v}\mathbb{E}\left(\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^{vW}\right)\\ &= \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v}\left(1+\frac{k}{p}\left(\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^v - 1\right)\right)^k\\ &= \frac{1}{2} + \left(\frac{1}{2}\right)^{1-v}\exp\left[k\log\left(1+\frac{k}{p}\left(\cosh\left(\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)}\right)^v - 1\right)\right)\right].\end{aligned}$$

To get the upper bound on the total variation distance appearing in (2.52), we aim at constraining this last expression to be smaller than $1+\eta^2$. This is equivalent to the following inequality:

$$2^v \exp\left[k\log\left(1+\frac{k}{p}\left(\cosh\left(\frac{n\lambda^2 k}{vk(\mathrm{var}(Y)-\lambda^2 k)}\right)^v-1\right)\right)\right] \leq 1+2\eta^2 \ . \tag{2.61}$$

We now choose $v=\frac{\mathcal{L}(\eta)}{\log(2)}\wedge 1$. If v is strictly smaller than one, then (2.61) is equivalent to:

$$k\log\left[1+\frac{k}{p}\left(\cosh\left(\frac{n\lambda^2 k}{vk(\mathrm{var}(Y)-\lambda^2 k)}\right)^v-1\right)\right] \leq \frac{\log(1+2\eta^2)}{2}. \tag{2.62}$$

It is straightforward to show that this last inequality also implies (2.61) if v equals one. We now suppose that

$$\frac{n\lambda^2}{v(\mathrm{var}(Y)-\lambda^2 k)} \leq \log\left((1+u)^{\frac{1}{v}}+\sqrt{(1+u)^{\frac{2}{v}}-1}\right), \tag{2.63}$$

where $u=\frac{p\mathcal{L}(\eta)}{k^2}$. Using the classical equality $\cosh\left[\log(1+x+\sqrt{2x+x^2})\right]=1+x$ with $x=(1+u)^{\frac{1}{v}}-1$, we deduce that inequality (2.63) implies (2.62) because

$$\begin{aligned} k\log\left(1+\frac{k}{p}\left(\cosh\left(\frac{n\lambda^2 k}{vk(\mathrm{var}(Y)-\lambda^2 k)}\right)^v-1\right)\right) &\leq k\log\left(1+\frac{k}{p}u\right) \\ &\leq \frac{k^2}{p}u \leq \mathcal{L}(\eta) \ . \end{aligned}$$

For any $\beta\geq 1$ and any $x>0$, it holds that $(1+x)^\beta \geq 1+\beta x$. As $\frac{1}{v}\geq 1$, Condition (2.63) is implied by:

$$\frac{\lambda^2 k}{\mathrm{var}(Y)-\lambda^2 k} \leq \frac{kv}{n}\log\left(1+\frac{u}{v}+\sqrt{\frac{2u}{v}}\right).$$

One then combines the previous inequality with the definitions of u and v to obtain the upper bound

$$\frac{\lambda^2 k}{\mathrm{var}(Y)-\lambda^2 k} \leq \frac{k}{n}\left(\frac{\mathcal{L}(\eta)}{\log(2)}\wedge 1\right)\log\left(1+\frac{p(\log(2)\vee\mathcal{L}(\eta))}{k^2}+\sqrt{\frac{2p(\log(2)\vee\mathcal{L}(\eta))}{k^2}}\right) \ .$$

For any x positive and any u between 0 and 1, $\log(1+ux)\geq u\log(1+x)$. As a consequence, the previous inequality is implied by:

$$\begin{aligned} \frac{\lambda^2 k}{\mathrm{var}(Y)-\lambda^2 k} &\leq \frac{k}{n}\left(\frac{\mathcal{L}(\eta)}{\log(2)}\wedge 1\right)([\mathcal{L}(\eta)\vee\log(2)]\wedge 1)\log\left(1+\frac{p}{k^2}+\sqrt{\frac{2p}{k^2}}\right) \\ &= \frac{k}{n}(\mathcal{L}(\eta)\wedge 1)\log\left(1+\frac{p}{k^2}+\sqrt{\frac{2p}{k^2}}\right) \ . \end{aligned}$$

To resume, if we take ρ^2 smaller than (2.17), then

$$\beta_I\left(\Theta[k,p,\rho]\right)\geq\delta \ .$$

Besides, the lower bound is strict if ρ^2 is strictly smaller than (2.17). To prove the second part of the theorem, one has to observe that $\alpha+\delta\leq 53\%$ implies that $\mathcal{L}(\eta)\geq\frac{1}{2}$. □

Proof of Proposition 2.6. Let us first assume that the covariance matrix of X is the identity. We argue as in the proof of Theorem 2.7 taking $k=p$. The sketch of the proof remains unchanged except that we slightly modify the last part. Inequality (2.62) becomes

$$pv\log\left(\cosh\left(\frac{n\lambda^2 p}{vp(\mathrm{var}(Y)-\lambda^2 p)}\right)\right)\leq\mathcal{L}(\eta),$$

where we recall that $v=\frac{\mathcal{L}(\eta)}{\log 2}\wedge 1$. For all $x\in\mathbb{R}$, $\cosh(x)\leq\exp(x^2/2)$. Consequently, the previous inequality is implied by

$$\frac{\lambda^2 p}{\mathrm{var}(Y)-\lambda^2 p}\leq\sqrt{2v\mathcal{L}(\eta)}\frac{\sqrt{p}}{n},$$

and the result follows easily.

If we no longer assume that the covariance matrix Σ is the identity, we orthogonalize the sequence X_i thanks to Gram-Schmidt process. Applying the previous argument to this new sequence of covariates allows to conclude. □

Proof of Proposition 2.4. Let define the constant $L(\alpha, \delta)$ involved in the condition:

$$L(\alpha,\delta) := \sqrt{\log\left(1+8(1-\alpha-\delta)^2\right)}\left[1 \wedge \sqrt{\log\left(1+8(1-\alpha-\delta)^2\right)/(2\log 2)}\right]$$

Let us apply proposition 2.6. For any $\rho \leq L(\alpha,\delta)\frac{\sqrt{D_m}}{n}$ and any $\varsigma > 0$ there exists some $\theta \in S_m$ such that $\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = \rho^2$ and $\mathbb{P}_\theta(\phi_{m,\alpha} \leq 0) \geq \delta - \varsigma$. In the proof of Theorem 2.3, we have shown in (2.34) and following equalities that the distribution of the test statistic ϕ_m only depends on the quantity $\kappa_m^2 = \frac{\mathrm{var}(Y)-\mathrm{var}(Y|X_m)}{\mathrm{var}(Y|X_m)}$. Let θ' be an element of S_m such that $\kappa_m^2 = \rho^2$. The distribution of ϕ_m under $\mathbb{P}_{\theta'}$ is the same as its distribution under $\mathbb{P}_\theta$, and therefore

$$\mathbb{P}_{\theta'}\left(\phi_{m,\alpha} \leq 0\right) \geq \delta - \varsigma.$$

Letting ς go to 0 enables to conclude.

□

Proof of Proposition 2.8. This lower bound for dependent gaussian covariates is proved through the same approach as Theorem 2.7. We define the measure μ_ρ as in that proof. Under the hypothesis H_0, Y is independent of X. We note Σ the covariance matrix of X and $\mathbb{E}_{0,\Sigma}$ stands for the distribution of $(\mathbf{Y},\mathbf{X})$ under H_0 in order to emphasize the dependence on Σ.

First, one has to upper bound the quantity $\mathbb{E}_{0,\Sigma}\left[L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right]$. For the sake of simplicity, we make the hypothesis that every covariate X_j has variance 1. If this is not the case, we only have to rescale these variables. The quantity $\mathrm{corr}(i,j)$ refers to the correlation between X_i and X_j. As we only consider the case $k=1$, the set of models m in $\mathcal{M}(1,p)$ is in correspondence with the set $\{1,\dots,p\}$.

$$\mathbb{E}_{0,\Sigma}\left(L_{\mu_{i,\zeta^1,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{j,\zeta^2,\rho}}(\mathbf{Y},\mathbf{X})\right) = \left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-\mathrm{corr}(i,j)\lambda^2\zeta^1\zeta^2}\right)^n.$$

When i and j are fixed, we upper bound the expectation of this quantity with respect to ζ^1 and ζ^2 by

$$\mathbb{E}_{0,\Sigma}\left(L_{\mu_{i,\rho}}(\mathbf{Y},\mathbf{X})L_{\mu_{j,\rho}}(\mathbf{Y},\mathbf{X})\right) \leq \frac{1}{2} + \frac{1}{2}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-|\mathrm{corr}(i,j)|\lambda^2}\right)^n. \tag{2.64}$$

If $i \neq j$, $|\mathrm{corr}(i,j)|$ is smaller than c and if $i=j$, $\mathrm{corr}(i,j)$ is exactly one. As a consequence, taking the expectation of (2.64) with respect to i and j yields the upper bound

$$\mathbb{E}_{0,\Sigma}\left(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right) \leq \frac{1}{2} + \frac{1}{2}\left(\frac{1}{p}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-\lambda^2}\right)^n + \frac{p-1}{p}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-c\lambda^2}\right)^n\right). \tag{2.65}$$

Recall that we want to constrain this quantity (2.65) to be smaller than $1+\eta^2$. In particular, this holds if the two following inequalities hold:

$$\frac{1}{p}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-\lambda^2}\right)^n \leq \frac{1}{p} + \eta^2 \tag{2.66}$$

$$\frac{p-1}{p}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y)-c\lambda^2}\right)^n \leq \frac{p-1}{p} + \eta^2. \tag{2.67}$$

One then uses the inequality $\log(\frac{1}{1-x}) \leq \frac{x}{1-x}$ which holds for any positive x smaller than one. Condition (2.66) holds if

$$\frac{\lambda^2}{\mathrm{var}(Y)-\lambda^2} \leq \frac{1}{n}\log(1+p\eta^2), \tag{2.68}$$

whereas Condition (2.67) is implied by

$$\frac{c\lambda^2}{\mathrm{var}(Y) - c\lambda^2} \leq \frac{1}{n} \log\left(1 + \frac{p}{p-1}\eta^2\right) .$$

As c is smaller than one and $\frac{p}{p-1}$ is larger than 1, this last inequality holds if

$$\frac{\lambda^2}{\mathrm{var}(Y) - \lambda^2} \leq \frac{1}{nc} \log(1+\eta^2) . \tag{2.69}$$

Gathering conditions (2.68) and (2.69) allows to conclude and to obtain the desired lower bound (2.19).

□

Proof of Proposition 2.10. The sketch of the proof and the notations are analogous to the one in Proposition 2.8. The upper bound (2.64) still holds:

$$\mathbb{E}_{0,\Sigma}\left(L_{\mu_{i,\rho}}(\mathbf{Y},\mathbf{X}) L_{\mu_{j,\rho}}(\mathbf{Y},\mathbf{X})\right) \leq \frac{1}{2} + \frac{1}{2}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y) - |\mathrm{corr}(i,j)|\lambda^2}\right)^n .$$

Using the stationarity of the covariance function, we derive from (2.64) the following upper bound:

$$\mathbb{E}_{0,\Sigma}\left(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right) \leq \frac{1}{2} + \frac{1}{2p}\sum_{i=0}^{p-1}\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y) - \lambda^2|\mathrm{corr}(0,i)|}\right)^n ,$$

where $\mathrm{corr}(0,i)$ equals $\mathrm{corr}(X_1, X_{i+1})$. As previously, we want to constrain this quantity to be smaller than $1+\eta^2$. In particular, this is implied if for any i between 0 and $p-1$:

$$\left(\frac{\mathrm{var}(Y)}{\mathrm{var}(Y) - \lambda^2|\mathrm{corr}(i,0)|}\right)^n \leq 1 + \frac{2p\eta^2|\mathrm{corr}(i,0)|}{\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)|}.$$

Using the inequality $\log(1+u) \leq u$, it is straightforward to show that this previous inequality holds if

$$\frac{\lambda^2}{\mathrm{var}(Y) - \lambda^2|\mathrm{corr}(i,0)|} \leq \frac{1}{n|\mathrm{corr}(i,0)|} \log\left(1 + \frac{2p\eta^2|\mathrm{corr}(0,i)|}{\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)|}\right).$$

As $|\mathrm{corr}(i,0)|$ is smaller than one for any i between 0 and $p-1$, it follows that $\mathbb{E}_{0,\Sigma}\left(L^2_{\mu_\rho}(\mathbf{Y},\mathbf{X})\right)$ is smaller than $1+\eta^2$ if

$$\rho^2 \quad \leq \quad \bigwedge_{i=0}^{p-1} \frac{1}{n|\mathrm{corr}(i,0)|} \log\left(1 + \frac{2p\eta^2|\mathrm{corr}(0,i)|}{\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)|}\right).$$

We now apply the convexity inequality $\log(1+ux) \geq u\log(1+x)$ which holds for any positive x and any u between 0 and 1 to obtain the condition

$$\rho^2 \quad \leq \quad \frac{1}{n} \log\left(1 + \frac{2p\eta^2}{\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)|}\right). \tag{2.70}$$

It turns out we only have to upper bound the sum of $|\mathrm{corr}(i,0)|$ for the different types of correlation:

1. For $\mathrm{corr}(i,j) = \exp(-w|i-j|_p)$, the sum is clearly bounded by $1 + 2\frac{e^{-w}}{1-e^{-w}}$ and Condition (2.70) simplifies as
$$\rho^2 \leq \frac{1}{n}\log\left(1 + 2p\eta^2\frac{1-e^{-w}}{1+e^{-w}}\right).$$

2. if $\mathrm{corr}(i,j) = (1+|i-j|_p)^{-t}$ for t strictly larger than one, then $\sum_{i=0}^{p-1}|\mathrm{corr}(i,0)| \leq 1 + \frac{2}{t-1}$ and Condition (2.70) simplifies as
$$\rho^2 \leq \frac{1}{n}\log\left(1 + \frac{2p(t-1)\eta^2}{t+1}\right).$$

3. if $\mathrm{corr}(i,j) = (1+|i-j|_p)^{-1}$ then $\sum_{i=0}^{p-1} |\mathrm{corr}(i,0)| \leq 1+2\log(p-1)$ and Condition (2.70) simplifies as
$$\rho^2 \leq \frac{1}{n}\log\left(1+\frac{2p\eta^2}{1+2\log(p-1)}\right).$$

4. if $\mathrm{corr}(i,j) = (1+|i-j|_p)^{-t}$ for $0<t<1$, then
$$\sum_{i=0}^{p-1} |\mathrm{corr}(i,0)| \leq 1+\frac{2}{1-t}\left[\left(\frac{p}{2}\right)^{1-t}-1\right] \leq \frac{2}{1-t}\left(\frac{p}{2}\right)^{1-t}$$
and Condition (2.70) simplifies as
$$\rho^2 \leq \frac{1}{n}\log\left(1+p^t 2^{1-t}(1-t)\eta^2\right).$$

□

Proof of Proposition 2.13. For each dimension D between 1 and p, we define $r_D^2 = \rho_{D,n}^2 \wedge a_D^2 R^2$. Let us fix some $D \in \{1,\dots,p\}$. Since $r_D^2 \leq a_D^2$ and since the a_j's are non increasing,
$$\sum_{j=1}^{D} \frac{\mathrm{var}(Y|X_{m_{j-1}}) - \mathrm{var}(Y|X_{m_j})}{a_j^2} \leq \mathrm{var}(Y|X)R^2,$$
for all $\theta \in S_{m_D}$ such that $\frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_D^2$. Indeed, $\|\theta\|^2 = \sum_{j=1}^{D} \mathrm{var}(Y|X_{m_{j-1}}) - \mathrm{var}(Y|X_{m_j})$ and $\mathrm{var}(Y) - \|\theta\|^2 = \mathrm{var}(Y|X)$. As a consequence,
$$\left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_D^2\right\} \subset \left\{\theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq r_D^2\right\}.$$
Since $r_D \leq \rho_{D,n}$, we deduce from Proposition 2.6 that
$$\beta_\Sigma\left(\left\{\theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} \geq r_D^2\right\}\right) \geq \delta.$$
The first result of Proposition 2.13 follows by gathering these lower bounds for all D between 1 and p.

Moreover, $\rho_{i,n}^2$ is defined in Proposition 2.6 as $\rho_{i,n}^2 = \sqrt{2}\left[\sqrt{\mathcal{L}(\eta)} \wedge \frac{\mathcal{L}(\eta)}{\sqrt{\log 2}}\right]\frac{\sqrt{i}}{n}$. If $\alpha+\delta \leq 47\%$, it is straightforward to show that $\rho_{i,n}^2 \geq \frac{\sqrt{i}}{n}$.

□

Proof of Proposition 2.15. We first need the following Lemma.

Lemma 2.17. *We consider $(I_j)_{j\in\mathcal{J}}$ a partition of $\mathcal{I}$. For each $j \in \mathcal{J}$ let $p(j) = |I_j|$. For any $j \in \mathcal{J}$, we define Θ_j as the set of $\theta \in \mathbb{R}^p$ such that their support is included in I_j. For any sequence of positive weights k_j such that*
$$\sum_{j\in\mathcal{J}} k_j = 1,$$
it holds that
$$\beta_I\left(\bigcup_{j\in\mathcal{J}}\left\{\theta \in \Theta_j, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r_j^2\right\}\right) \geq \delta,$$
if for all $j \in \mathcal{J}$, $r_j \leq \rho_{p(j),n}(\eta/\sqrt{k_j})$, where the function $\rho_{p(j),n}$ is defined by (2.16).

For all $j \geq 0$ such that $2^{j+1}-1 \in \mathcal{I}$ (i.e. for all $j \leq J$ where $J = \log(p+1)/\log(2) - 1$), let $\bar{S}_j$ be the linear span of the e_k's for $k \in \{2^j,\dots,2^{j+1}-1\}$. Then, $\dim(\bar{S}_j) = 2^j$ and $\bar{S}_j \subset S_{m_D}$ for $D = D(j) = 2^{j+1}-1$. It is straightforward to show that
$$\bigcup_{j=0}^{J} \bar{S}_j[r_{D(j)}] \subset \bigcup_{j=0}^{J} S_{m_{D(j)}}[r_{D(j)}] \subset \bigcup_{D=1}^{p} S_{m_D}[r_D],$$

where $\bar{S}_j[r_D(j)] := \left\{\theta \in \bar{S}_j, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r^2_{D(j)}\right\}$ and $S_{m_D}[r_D] := \left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r^2_{D}\right\}$.

Let choose $\mathcal{J} = \{1, \dots, J\}$. For any $j \in \mathcal{J}$, we define $I_j = \left\{2^j, 2^j+1, \dots, 2^{j+1}-1\right\}$. Applying Lemma 2.17 with $k_j := [(j+1)R(p)]^{-1}$ where $R(p) := \sum_{k=0}^{J} 1/(k+1)$ we get

$$\beta_I\left(\bigcup_{D=1}^{p}\left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r^2_D\right\}\right) \geq \delta \ ,$$

if for all those $D = D(j)$

$$r^2_D \leq \sqrt{\log(1+2\eta^2/k_j)}\left(1 \wedge \frac{\sqrt{\log(1+2\eta^2/k_j)}}{\sqrt{2\log 2}}\right)\frac{\sqrt{D}}{n}.$$

For $D = D(j)$, this last quantity is lower bounded by

$$\sqrt{\log(1+2\eta^2/k_j)}\left(1 \wedge \frac{\sqrt{\log(1+2\eta^2/k_j)}}{\sqrt{2\log 2}}\right)\frac{\sqrt{D}}{n} \geq \sqrt{\log(1+2\eta^2(j+1)R(p))}\left(1 \wedge \frac{\sqrt{\log(1+2\eta^2)}}{\sqrt{2\log 2}}\right)\frac{2^{j/2}}{n} \ . \tag{2.71}$$

It remains to check that (2.71) is larger than $\bar{\rho}_{D(j),n}$. Using $j+1 = \log(D+1)/\log(2) \geq \log(D+1)$, we get $2^{j/2} \geq \sqrt{D/2}$. Thanks to the convexity inequality $\log(1+ux) \geq u\log(1+x)$, which holds for any $x > 0$ and any $u \in]0,1]$, we obtain

$$\begin{aligned}
\sqrt{\log(1+2\eta^2(j+1)R(p))}2^{j/2} &\geq \sqrt{D/2}\left(\eta\sqrt{2R(p)} \wedge 1\right)\sqrt{\log\left[1+\log(D+1)\right]} \\
&\geq \left((\eta\sqrt{2}) \wedge 1\right)\sqrt{\log\log(D+1)}\sqrt{D/2}, \\
&\geq \frac{1}{\sqrt{2}}\left(1 \wedge \sqrt{\log(1+2\eta^2)}\right)\sqrt{\log\log(D+1)}\sqrt{D} \ ,
\end{aligned}$$

as $R(p)$ is larger than one for any $p \geq 1$. All in all, we get the lower bound

$$\begin{aligned}
&\sqrt{\log(1+2\eta^2(j+1)^2R(p))}\left(1 \wedge \frac{\sqrt{\log(1+2\eta^2)}}{\sqrt{2\log 2}}\right)\frac{2^{j/2}}{n} \\
&\qquad \geq \frac{1}{2\sqrt{\log(2)}}\left(1 \wedge \log(1+2\eta^2)\right)\sqrt{\log\log(D+1)}\frac{\sqrt{D}}{n} = \bar{\rho}^2_{D,n} \ .
\end{aligned}$$

Thus, if for all $1 \leq D \leq p$, r^2_D is smaller than $\bar{\rho}^2_{D,n}$, it holds that

$$\beta_I\left(\bigcup_{D=1}^{p}\left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r^2_D\right\}\right) \geq \delta \ .$$

□

Proof of Lemma 2.17. Using a similar approach to the proof of Theorem 2.7, we know that for each $r_j \leq \tilde{\rho}_j(\eta/\sqrt{k_j})$ there exists some measure μ_j over

$$\Theta_j[r_j] := \left\{\theta \in \Theta_j, \frac{\|\theta\|^2}{\mathrm{var}(Y)-\|\theta\|^2} = r^2_j\right\}$$

such that

$$\mathbb{E}_0\left[L^2_{\mu_j}(Y,X)\right] \leq 1+\eta^2/k_j \ . \tag{2.72}$$

We now define a probability measure $\mu = \sum_{j\in\mathcal{J}} k_j\mu_j$ over $\bigcup_{j\in\mathcal{J}}\Theta_j[r_j]$. L_{μ_j} refers to the density of $\mathbb{P}_{\mu_j}$ with respect to $\mathbb{P}_0$. Thus,

$$L_\mu(Y) = \frac{\mathrm{d}\mathbb{P}_\mu}{\mathrm{d}\mathbb{P}_0}(\mathbf{Y},\mathbf{X}) = \sum_{j\in\mathcal{J}} k_j L_{\mu_j}(\mathbf{Y},\mathbf{X}) \ ,$$

and

$$\mathbb{E}_0\left[L_\mu^2(\mathbf{Y},\mathbf{X})\right] = \sum_{j,j'\in\mathcal{J}} k_j k_{j'} \mathbb{E}_0\left[L_{\mu_j}(\mathbf{Y},\mathbf{X}) L_{\mu_{j'}}(\mathbf{Y},\mathbf{X})\right] .$$

Using expression (2.56), it is straightforward to show that if $j \neq j'$, then

$$\mathbb{E}_0\left[L_{\mu_j}(\mathbf{Y},\mathbf{X}) L_{\mu_{j'}}(\mathbf{Y},\mathbf{X})\right] = 1.$$

This follows from the fact that the sets Θ_j and $\Theta_{j'}$ are orthogonal with respect to the inner product (2.5). Thus,

$$\mathbb{E}_0\left[L_\mu(\mathbf{Y},\mathbf{X})\right] = 1 + \sum_{j\in\mathcal{J}} k_j^2 \left(\mathbb{E}_0\left[L_{\mu_j}^2(\mathbf{Y},\mathbf{X})\right] - 1\right) \leq 1 + \eta^2$$

thanks to (2.72). Using the argument (2.53) as in the proof of Theorem 2.7 allows to conclude. □

Proof of Proposition 2.16. First of all, we only have to consider the case where the covariance matrix of X is the identity. If this is not the case, one only has to apply Gram-Schmidt process to X and thus obtain a vector X' and a new basis for $\mathbb{R}^p$ which is orthonormal. We refer to the beginning of Section 2.5 for more details.

Like the previous bounds for ellipsoids, we adapt the approach of Section 6 in Baraud [Bar02b]. We use the same notations as in proof of Proposition 2.13. Let $D^*(R) \in \{1,\dots,p\}$ an integer which achieves the supremum of $\bar{\rho}_D^2 \wedge (R^2 a_D^2) = \bar{r}_D^2$. As in proof of Proposition 2.13, for any $R > 0$,

$$\left\{\theta \in S_{m_{D^*(R)}}, \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = r_{D^*(R)}^2\right\} \subset \left\{\theta \in \mathcal{E}_a(R), \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq r_{D^*(R)}^2\right\} .$$

When R varies, $D^*(R)$ describes $\{1,\dots,p\}$. Thus, we obtain

$$\begin{aligned}\bigcup_{1\leq D\leq p} \left\{\theta \in S_{m_D}, \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = r_D^2\right\} &= \bigcup_{R>0} \left\{\theta \in S_{m_{D^*(R)}}, \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} = r_{D^*(R)}^2\right\} \\ &\subset \bigcup_{R>0} \left\{\theta \in \mathcal{E}_a(R) \frac{\|\theta\|^2}{\mathrm{var}(Y) - \|\theta\|^2} \geq r_{D^*(R)}^2\right\} ,\end{aligned}$$

and the result follows from proposition 2.15. □

2.9 Appendix

Proof of Proposition 2.1. The test associated with Procedure P_1 corresponds to a Bonferroni procedure. Hence, we prove that its size is less than α by arguing as follows: let θ be an element of S_V (defined in Section 2.2.2),

$$\mathbb{P}_\theta(T_\alpha > 0) \leq \sum_{m\in\mathcal{M}} \mathbb{P}_\theta\left(\phi_m(\mathbf{Y},\mathbf{X}) - \bar{F}_{D_m,N_m}^{-1}(\alpha_m) > 0\right),$$

where $\phi_m(\mathbf{Y},\mathbf{X})$ is defined in (2.3). The test is rejected if for some model m, $\phi_m(\mathbf{Y},\mathbf{X})$ is larger than $\bar{F}_{D_m,N_m}^{-1}(\alpha_m)$. As θ belongs to S_V, $\Pi_{V\cup m}\mathbf{Y} - \Pi_V\mathbf{Y} = \Pi_{V\cup m}\boldsymbol{\epsilon} - \Pi_V\boldsymbol{\epsilon}$ and $\mathbf{Y} - \Pi_{V\cup m}\mathbf{Y} = \boldsymbol{\epsilon} - \Pi_{V\cup m}\boldsymbol{\epsilon}$. Then, the quantity $\phi_m(\mathbf{Y},\mathbf{X})$ is equal to

$$\phi_m(\mathbf{Y},\mathbf{X}) = \frac{N_m \|\Pi_{V\cup m}\boldsymbol{\epsilon} - \Pi_V\boldsymbol{\epsilon}\|_n^2}{D_m \|\boldsymbol{\epsilon} - \Pi_{V\cup m}\boldsymbol{\epsilon}\|_n^2}.$$

Because $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, the distribution of $\phi_m(\mathbf{Y},\mathbf{X})$ conditionally to $\mathbf{X}$ is a Fisher distribution with D_m and N_m degrees of freedom. As a consequence, $\phi_{m,\alpha_m}(\mathbf{Y},\mathbf{X})$ is a Fisher test with D_m and N_m degrees of freedom. It follows that:

$$\mathbb{P}_\theta(T_\alpha > 0) \leq \sum_{m\in\mathcal{M}} \alpha_m \leq \alpha.$$

The test associated with Procedure P_2 has the property to be of size exactly α. More precisely, for any $\theta \in S_V$, we have that

$$\mathbb{P}_\theta(T_\alpha > 0|\mathbf{X}) = \alpha \quad \mathbf{X} \text{ a.s. .}$$

The result follows from the fact that $q_{\mathbf{X},\alpha}$ satisfies

$$\mathbb{P}_\theta\left(\sup_{m\in\mathcal{M}}\left\{\frac{N_m\|\Pi_{V\cup m}(\boldsymbol{\epsilon})-\Pi_V(\boldsymbol{\epsilon})\|_n^2}{D_m\|\boldsymbol{\epsilon}-\Pi_{V\cup m}(\boldsymbol{\epsilon})\|_n^2} - \bar{F}^{-1}_{D_m,N_m}\left(q_{\mathbf{X},\alpha}\right)\right\} > 0\middle|\mathbf{X}\right) = \alpha,$$

and that for any $\theta \in S_V$, $\Pi_{V\cup m}\mathbf{Y} - \Pi_V\mathbf{Y} = \Pi_{V\cup m}\boldsymbol{\epsilon} - \Pi_V\boldsymbol{\epsilon}$ and $\mathbf{Y} - \Pi_{V\cup m}\mathbf{Y} = \boldsymbol{\epsilon} - \Pi_{V\cup m}\boldsymbol{\epsilon}$.

□

Proof of Proposition 2.2. Let come back to the definitions of T^1_α and T^2_α:

$$\begin{aligned} T^1_\alpha(\mathbf{X},\mathbf{Y}) &= \sup_{m\in\mathcal{M}}\left\{\phi_m(\mathbf{Y},\mathbf{X}) - \bar{F}^{-1}_{D_m,N_m}(\alpha/|\mathcal{M}|)\right\} \\ T^2_\alpha(\mathbf{X},\mathbf{Y}) &= \sup_{m\in\mathcal{M}}\left\{\phi_m(\mathbf{Y},\mathbf{X}) - \bar{F}^{-1}_{D_m,N_m}(q_{\mathbf{X},\alpha})\right\} \end{aligned}$$

Conditionally on $\mathbf{X}$, the size of T^1_α is smaller than α, whereas the size T^2_α is exactly α. As a consequence $q_{\mathbf{X},\alpha} \geq \alpha/|\mathcal{M}|$ as the statistics T^1_α and T^2_α differ only through these quantities. Thus, $T^2_\alpha(\mathbf{X},\mathbf{Y}) \geq T^1_\alpha(\mathbf{X},\mathbf{Y})$, $(\mathbf{X},\mathbf{Y})$ almost surely and the result (2.11) follows. □

Chapter 3

Tests for Gaussian graphical models

Abstract. Gaussian graphical models are promising tools for analyzing genetic networks. In many applications, biologists have some knowledge of the genetic network and may want to assess the quality of their model thanks to gene expression data. This is why one introduces a novel procedure for testing the neighborhoods of a Gaussian graphical model. It is based on the connection between local Markov property and conditional regression of a Gaussian random variable. Adapting recent results on tests for high-dimensional Gaussian linear models, one proves that the testing procedure inherits appealing theoretical properties. Besides, it applies and is computationally feasible in a high-dimensional setting: the number of nodes may be much larger than the number of observations. A large part of the study is deserved to illustrate and discuss applications to simulated data and to biological data.

3.1 Introduction

Biological processes regulating the expression of the genes lead to complex high-dimensional systems. Thus, inferring these underlying networks recently became an arising issue in systems biology. More precisely, the challenge at hand is to use gene expression data coming from microarray experiments to estimate or to test the network. In this regard, mathematical tools were developed to provide a suitable framework for modeling complex dependence structures. Among these, Gaussian graphical models (GGMs) (see Lauritzen [Lau96], Edwards [Edw00]) have gained a lot of attention and have already been applied in several works (see [KW00], [TH02], [WYR03], [WZV$^+$04], [SS05]). However, the number of genes p will typically exceed by far the number n of the samples given by the microarray experiments. In this high-dimensional setting, estimating or assessing a GGM raises difficult statistical and computational issues. For instance, most of the methodologies based on asymptotic statistics do not apply anymore.

In recent years, the problem of graph estimation for massive data sets became a hot spot in statistics. Most of the emerging methods fall in two categories. On the one hand, some are based on multiple testing procedures, see for instance Schäfer and Strimmer [SS05] or Wille and Bühlmann [WB06]. On the other hand, other methods are based on variable selection for high-dimensional data. We mention the seminal work of Meinshausen and Bühlmann [MB06] who proposed a computationally feasible model selection algorithm using Lasso penalization (see [Tib96]). Huang *et al.* [HLPL06] and Yuan and Lin [YL07] extend this method to infer directly the graph by minimizing the log-likelihood penalized by the l_1 norm.

In contrast, there are not many results about the problem of hypothesis testing in a high-dimensional setting. We believe that this issue is significant for two reasons: First, when considering a gene regulation network, the biologists often have a previous knowledge of the graph and may want to test if the microarray data match with their model. Second, when applying an estimation method in a high-dimensional setting, it could be useful to test the estimated graph as some of these methods reveal too conservative. Admittedly, some of the previously mentioned estimation methods are based on multiple testing. However, as they are constructed for an estimation purpose, most of them do not take into account some previous knowledge about the graph. This is for instance the case for the approaches of Drton and Perlman

[DP07] and Schäfer and Strimmer [SS05]. Some of the other existing procedures cannot be applied in a high-dimensional setting (e.g. Drton and Perlman [DP08]). Finally, most of them lack of theoretical justifications in a non asymptotic way. This is why we propose a testing procedure to assess whether some connections are missing in a graph. The procedure starts from a minimal graph, minimal in the sense that all edges are assumed to be relevant : typically this graph is provided by biologists thanks to their prior knowledge. The aim of the procedure is to test if microarray data match with this minimal graph or if there are missing edge. The interest of this test is first for biologists to assess the quality of their knowledge. Second, when the test is rejected, it suggests potential connections between genes that steer biologists towards new experimentations.

Let us precise our objective: consider $X = (X_1, \dots, X_p)^t$ a random vector distributed as a multivariate Gaussian $\mathcal{N}(0, \Sigma)$. Throughout this chapter, we assume that the matrix Σ is non-singular. The conditional independence structure of this distribution can be represented by an undirected graph $\mathcal{G} = (\Gamma, E)$ where $\Gamma = \{1, \dots, p\}$ is the set of nodes and E the set of edges. There is an edge between nodes a and b if and only if the random variables X_a and X_b are conditionally dependent given all remaining variables $X_{-\{a,b\}} = \{X_i, i \in \Gamma \setminus \{a, b\}\}$. The random vector X is then said to be a Gaussian graphical model with respect to the graph $\mathcal{G}$. Given a node $a \in \Gamma$, we define its neighborhood $ne(a)$ as the set of nodes $b \in \Gamma \setminus \{a\}$ such that $(a, b) \in E$. We say that X follows the local Markov property at node a with respect to the graph $\mathcal{G}$ if X_a is independent from $\{X_i, i \in \Gamma \setminus (ne(a) \cup \{a\})\}$ given $\{X_i, i \in ne(a)\}$. Lauritzen [Lau96] shows that X is a Gaussian graphical model with respect to $\mathcal{G}$ if and only if it follows the local Markov property at each node $a \in \Gamma$.

Suppose we are given a n-sample of the vector X and an undirected graph $\mathcal{G} = (\Gamma, E)$. In the present chapter, we construct testing procedures of the hypothesis "X follows the local Markov property at the node a with respect to the graph $\mathcal{G}$" against the hypothesis that it does not. In the following, we refer to such tests as *test of neighborhood*. We deduce testing procedures of the hypothesis "X is a Gaussian graphical model with respect to the graph $\mathcal{G}$" against the hypothesis that it is not. We call these tests *tests of graph*. Our test of neighborhood applies and is computationally feasible in a high-dimensional setting as long as the graph $\mathcal{G}$ is sparse. Besides, it inherits the appealing theoretical properties shown in a Chapter 2: we are able to compute non asymptotic bounds of its power and we show its optimality in the minimax sense.

In Section 3.2.1.1 we highlight the connection between tests of neighborhood and tests in Gaussian linear regression in a random Gaussian design. Thus, we construct procedures based on tests of linear hypothesis in this regression framework introduced in Chapter 2. They are feasible in a high-dimensional setting and we control exactly their family-wise error rate. Then, we exhibit non asymptotic results on their power in Section 3.2.2. Finally, we apply our procedures to simulated data in Section 3.3 and to real data sets in Section 3.4. In the sequel, we denote $\overline{ne}(a) := ne(a) \cup \{a\}$ for any node $a \in \Gamma$.

3.2 Description of the testing procedures

3.2.1 Test of neighborhood

3.2.1.1 Connection with conditional Gaussian regression

In this part, we highlight the connection between the local Markov property and conditional regression of a Gaussian random variable. We define precisely the testing procedure in the next part, following the approach introduced in Chapter 2.

Let $\mathcal{G} = (\Gamma, E)$ be an undirected graph and $a \in \Gamma$ be a node of this graph. We want to test the hypothesis "X_a is independent from $X_{\Gamma \setminus \overline{ne}(a)}$ conditionally to $X_{ne(a)}$" against the general alternative that it is not. This hypothesis corresponds to the local Markov property defined in Lauritzen [Lau96] of X at the node a. In order to perform this test, we use a different characterization of conditional independence.

Let us consider the conditional distribution of X_a given all remaining variables $X_{-a} = \{X_b, b \in \Gamma \setminus \{a\}\}$. Using standard Gaussian properties (see for instance [Lau96] appendix C), we know that this conditional distribution is a Gaussian distribution whose mean is a linear combination of elements in X_{-a}

and whose variance does not depend on X_{-a}. Hence, we can decompose X_a as:

$$X_a = \sum_{b \in \Gamma \setminus a} \theta_b^a X_b + \epsilon_a, \tag{3.1}$$

where θ^a is a vector of coefficients in $\mathbb{R}^{p-1}$ and ϵ_a is a zero mean Gaussian random variable independent from X_{-a} whose variance equals the conditional variance of X_a given X_{-a}, $\mathrm{var}(X_a|X_{-a})$. The vector θ^a is determined by the inverse covariance matrix K of X (see [Edw00]). More precisely, $\theta_b^a = -K[a,b]/K[a,a]$ for any $b \neq a$ and $\mathrm{var}(X_a|X_{-a}) = 1/K[a,a]$. As a consequence, the set of non-zero coefficients of θ^a corresponds to the non zero-components of the a-th row of K. Equivalently, there is an edge between the nodes a and b in the graph if the quantity $K[a,b]$ is not zero. For any set $V \subset \Gamma \setminus \{a\}$, θ_V^a denotes the sequence $(\theta_b^a)_{b \in V}$.

Testing the null-hypothesis "X_a is independent from $X_{\Gamma \setminus \overline{ne}(a)}$ conditionally to $X_{ne(a)}$" against the general alternative is therefore equivalent to testing the null-hypothesis $H_{0,a}$: "$\theta^a_{\Gamma \setminus \overline{ne}(a)} = 0$" against the general alternative $H_{1,a}$: "$\theta^a_{\Gamma \setminus \overline{ne}(a)} \neq 0$". Consequently, the test of neighborhood amounts to goodness-of-fit tests for Gaussian regression with random Gaussian covariates as considered in Chapter 2.

3.2.1.2 Description of the procedure

In this part, we adapt the test introduced in Chapter 2 to our statistical context. We are given n observations of the vector $X = (X_1, \ldots, X_p)^t$. For any $a \in \Gamma$, let us note $\mathbf{X}_a$ the n-vector of observations of X_a and $\mathbf{X}_{-a}$ the set of vectors $\mathbf{X}_b$ where b belongs to $\Gamma \setminus \{a\}$. The joint distribution of (X_a, X_{-a}) is uniquely defined by the vector θ^a, the covariance matrix of X_{-a} denoted Σ_{-a}, and $\mathrm{var}(X_a|X_{-a})$ the conditional variance of X_a. In the sequel, $\mathbb{P}_{\theta^a}$ refers to the joint distribution of $(\mathbf{X}_a, \mathbf{X}_{-a})$. For the sake of simplicity, we do not emphasize the dependency of $\mathbb{P}_{\theta^a}$ on Σ_{-a} and $\mathrm{var}(X_a|X_{-a})$.

Let us first fix some level $\alpha \in]0,1[$ and let m be a subset of $\Gamma \setminus \overline{ne}(a)$. In the sequel d_a and D_m denote the cardinalities of $ne(a)$ and m, and we define N_m as $n - d_a - D_m$. We assume that $n \geq d_a + 2$. We define the Fisher statistic ϕ_m by

$$\phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) := \frac{N_m \|\Pi_{ne(a)\cup m}\mathbf{X}_a - \Pi_{ne(a)}\mathbf{X}_a\|_n^2}{D_m \|\mathbf{X}_a - \Pi_{ne(a)\cup m}\mathbf{X}_a\|_n^2}, \tag{3.2}$$

where $\|.\|_n$ is the canonical norm in $\mathbb{R}^n$, and $\Pi_{ne(a)}$ and $\Pi_{ne(a)\cup m}$ respectively refer to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_b)_{b \in ne(a)}$ and to the orthogonal projection onto the space generated by the vectors $(\mathbf{X}_b)_{b \in ne(a)\cup m}$. Then, ϕ_m corresponds to the statistic of the Fisher test of the null hypothesis

$$\begin{aligned} &H_{0,a} : \theta_{\Gamma \setminus \overline{ne}(a)} = 0 \text{ against the alternative} \\ &H_{1,a,m} : \theta_{\Gamma \setminus \overline{ne}(a)} \neq 0 \text{ and } \theta_{\Gamma \setminus (\overline{ne}(a)\cup m)} = 0. \end{aligned} \tag{3.3}$$

In the sequel, $\Pi_{ne(a)^\perp}$ stands for the orthogonal projection along the space generated by $(\mathbf{X}_b)$ with b belonging to $ne(a)$. Let us consider a finite collection $\mathcal{M}_a$ of non empty subsets of $\Gamma \setminus \overline{ne}(a)$. For all $m \in \mathcal{M}_a$, the cardinality D_m must be smaller than $n - d_a$. We define $\{\alpha_m, m \in \mathcal{M}_a\}$ a suitable collection of numbers in $]0,1[$ (which possibly depend on $\mathbf{X}_{-a}$). Our testing procedure consists in doing for each $m \in \mathcal{M}_a$ the Fisher test based on the statistic ϕ_m defined in Equation (3.2) at level α_m and rejecting the null hypothesis $H_{0,a}$ if one of those tests does. More precisely, we define the test T_α as

$$T_\alpha := \sup_{m \in \mathcal{M}_a} \left\{ \phi_m(\mathbf{X}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m, N_m}(\alpha_m(\mathbf{X}_{-a})) \right\}, \tag{3.4}$$

where for any $u \in \mathbb{R}$, $\bar{F}_{D,N}(u)$ denotes the probability for a Fisher variable with D and N degrees of freedom to be larger than u. We therefore reject the null hypothesis when T_α is positive. The main difference between this procedure and the one defined in Chapter 2 lies in the fact that we now deal with possibly random collection of models.

In order to ensure that the level T_α is less than α, the collection of weights $\{\alpha_m(\mathbf{X}_{-a}), m \in \mathcal{M}_a\}$ in $]0,1[$ must satisfy the property: for all $\theta \in \mathbb{R}^{p-1}$ such that $\theta_{\Gamma \setminus \overline{ne}(a)} = 0$, then $\mathbb{P}_\theta(T_\alpha > 0) \leq \alpha$. We choose the collection $\{\alpha_m(\mathbf{X}_{-a}), m \in \mathcal{M}_a\}$ in accordance with one of the two following procedures:

- P_1 : The α_m 's do not depend on $\mathbf{X}_{-a}$ and satisfy the equality :

$$\sum_{m \in \mathcal{M}_a} \alpha_m = \alpha. \tag{3.5}$$

- P_2 : For all $m \in \mathcal{M}_a$, $\alpha_m(\mathbf{X}_{-a}) = q_{\mathbf{X}_{-a},\alpha}$, where $q_{\mathbf{X}_{-a},\alpha}$ is defined conditionally to $\mathbf{X}_{-a}$ as the α-quantile of the distribution of the random variable

$$\inf_{m \in \mathcal{M}_a} \bar{F}_{D_m,N_m}\left(\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})\right). \tag{3.6}$$

  Note that this last distribution does not depend on the variance of $\boldsymbol{\epsilon}_a$ and thus we can work out $q_{\mathbf{X}_{-a},\alpha}$ using Monte-Carlo method.

3.2.1.3 Comparison of Procedures P_1 and P_2

If the collection of models is not random, one can either use Procedure P_1 or P_2. In Section 2.3.2, we show that the test T_α with Procedure P_1 has a size less than α, whereas the size of T_α with Procedure P_2 is exactly α. We deduce from this fact that the test T_α with procedure P_2 is more powerful than the corresponding test defined with Procedure P_1 with weights $\alpha_m = \alpha/|\mathcal{M}_a|$.

On the one hand the choice of Procedure P_1 allows to avoid the computation of the quantile $q_{\mathbf{X}_{-a},\alpha}$ and possibly permits to give a Bayesian flavor to the choice of the weights. On the other hand, Procedure P_1 becomes too conservative when the collection of models $\mathcal{M}_a$ is large. This is often the case when the number p of nodes in the graph is large. That is why we advise to use Procedure P_2 when considering large graphs. We compare both Procedures in practice in Section 2.6 and 3.3.

3.2.1.4 Collection of models $\mathcal{M}_a$

The main advantage of our procedure is that it is very flexible in the choices of the models $m \in \mathcal{M}_a$. If we choose suitable collections $\mathcal{M}_a$, the test is powerful over a large class of alternatives as shown in Chapter 2 for non random collections. In this part, we propose two relevant classes of models $\mathcal{M}_a^1$ and $\mathcal{M}_a^2$ for our issue of test of neighborhood.

The collection $\mathcal{M}_a^1$ is defined as $\mathcal{M}_a^1 := \{\{b\}, b \in \Gamma \setminus \overline{ne}(a)\}$ and consists in taking each node in $\Gamma \setminus \overline{ne}(a)$ in turn. In Section 3.2.2, we present theoretical results of the power of T_α with collection $\mathcal{M}_a^1$ and Procedure P_1. This collection presents the advantage to be relatively small compared to other possible collections and the obtained procedure is consequently computationally attractive.

We have shown in Chapter 2, and this will be illustrated again in Section 3.3, that if there are several non-zero coefficients in $\theta^a_{\Gamma \setminus \overline{ne}(a)}$, considering models of larger dimensions can improve the performance of the test. For instance, if we are given an order on the nodes and if the vector θ^a belongs to an ellipsoid relative to this order, one should choose the collection of nested models defined by this order (see Section 2.5). There is not such an order in our context as we do not know in principle which nodes are more relevant to test. That is why we propose to use the LARS (least angle regression) algorithm introduced by [EHJT04]. This model selection algorithm provides an order of relevance of the covariates in linear regression. Besides, one of its main advantage lies in its computationally attractiveness. The collection of models $\mathcal{M}_a^2$ is built as follows. We first choose an integer J which corresponds to the maximal size of the models we want to consider. We advise to take J smaller than $n/2$. Then, we apply the LARS algorithm to the response $\Pi_{ne(a)^\perp}\mathbf{X}_a$ with the set of covariates $\Pi_{ne(a)^\perp}\mathbf{X}_b$ where $b \in \Gamma \setminus \overline{ne}(a)$ and we obtain the sequence $s_{LARS} = (j_1, \ldots, j_J)$. Finally we define the collection $\mathcal{M}_a^2$ as:

$$\mathcal{M}_a^2 := \{\{j_1, \ldots, j_k\}, 1 \leq k \leq J\}.$$

As the collection of models $\mathcal{M}_a^2$ given by the LARS algorithm now depends on the data, we need do define a new procedure to handle random collections.

Suppose we are given a random collection of models $\mathcal{M}_a$ which only depends on

$$\Psi(\mathbf{X}_a, \mathbf{X}_{-a}) := \left(\frac{\Pi_{ne(a)^\perp}\mathbf{X}_a}{\|\Pi_{ne(a)^\perp}\mathbf{X}_a\|_n}, \mathbf{X}_{-a}\right), \tag{3.7}$$

then we shall use the test statistic (3.4) with weights given by the procedure P_3 defined as follows:

- P_3: For all $m \in \mathcal{M}_a\left[\Psi\left(\mathbf{X}_a, \mathbf{X}_{-a}\right)\right]$, $\alpha_m(\mathbf{X}_{-a}) = q'_{\mathbf{X}_{-a},\alpha}$, the α-quantile of the distribution of the random variable
$$\inf_{m \in \mathcal{M}_a[\Psi(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})]} \bar{F}_{D_m, N_m}\left(\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})\right), \tag{3.8}$$
conditionally to $\mathbf{X}_{-a}$. As for the procedure P_2, the distribution of (3.8) does not depend on the variance of $\boldsymbol{\epsilon}_a$ and thus we are able to compute $q'_{\mathbf{X}_{-a},\alpha}$ using Monte-Carlo method.

Clearly, if the collection of models is not random, Procedures P_2 and P_3 lead to the same weights. As with Procedure P_2, the size of T_α with Procedure P_3 is exactly α. More Precisely, for any $\theta^a \in \mathbb{R}^{p-1}$ such that $\theta^a_{\Gamma \setminus \overline{ne}(a)} = 0$, we have that
$$\mathbb{P}_{\theta^a}\left(T_\alpha | \mathbf{X}_{-a}\right) = \alpha \ \ \mathbf{X}_{-a} \text{ a.s. } .$$

The result follows from the fact that $q'_{\mathbf{X}_{-a},\alpha}$ satisfies
$$\mathbb{P}_{\theta^a}\left(\sup_{m \in \mathcal{M}_a[\Psi(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a})]} \left\{\phi_m(\boldsymbol{\epsilon}_a, \mathbf{X}_{-a}) - \bar{F}^{-1}_{D_m, N_m}\left(q'_{\mathbf{X}_{-a},\alpha}\right)\right\} > 0 \middle| \mathbf{X}_{-a}\right) = \alpha,$$
and for any $\theta^a \in \mathbb{R}^{p-1}$ such that $\theta_{\Gamma \setminus \overline{ne}(a)} = 0$,
$$\Pi_{ne(a) \cup m} \mathbf{X}_a - \Pi_{ne(a)} \mathbf{X}_a = \Pi_{ne(a) \cup m} \boldsymbol{\epsilon}_a - \Pi_{ne(a)} \boldsymbol{\epsilon}_a \ ,$$
and
$$\mathbf{X}_a - \Pi_{ne(a) \cup m} \mathbf{X}_a = \boldsymbol{\epsilon}_a - \Pi_{ne(a) \cup m} \boldsymbol{\epsilon}_a \ .$$
As the sequence of relevant variables given by the LARS algorithm does not depend on the norm of the response, the collection $\mathcal{M}^2_a$ only depends on $\Psi(\mathbf{X}_a, \mathbf{X}_{-a})$ and thus we are able to apply Procedure P_3.

The size of these two collections $\mathcal{M}^1_a$ and $\mathcal{M}^2_a$ is smaller than the number of nodes p. Consequently, the computational complexity of our procedure is at most linear with respect to p when considering the collection $\mathcal{M}^1_a$ and is of the same order as the complexity of the LARS algorithm when considering $\mathcal{M}^2_a$.

3.2.2 Properties of the test of neighborhood with collection $\mathcal{M}^1_a$

For the convenience of the reader, we recall in this part some of the theoretical results established in Chapter 2. First, we give a proposition which characterizes the set of vectors θ^a over which the test T_α with the collection $\mathcal{M}^1_a$ and weights $\alpha_m = \alpha / |\mathcal{M}^1_a|$ is powerful. We shall then discuss the optimality of this test.

Proposition 3.1. *Let us assume that n satisfies:*
$$n - d_a - 1 \geq \left[10 \log\left(\frac{p - d_a - 1}{\alpha}\right) \vee 21 \log\left(1/\delta\right)\right].$$
Let us set the quantity
$$\rho^2_{n-d_a, p-d_a} := \frac{C_1}{n - d_a} \log\left(\frac{p - d_a - 1}{\alpha\delta}\right), \tag{3.9}$$
where C_1 is a universal constant. For any θ^a in $\mathbb{R}^{\Gamma \setminus \{a\}}$, $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1 - \delta$ if there exists $b \in \Gamma \setminus \overline{ne}(a)$ such that
$$\frac{\mathrm{var}_{\theta^a}(X_a | X_{ne(a)}) - \mathrm{var}_{\theta^a}(X_a | X_{ne(a) \cup \{b\}})}{\mathrm{var}_{\theta^a}(X_a | X_{ne(a) \cup \{b\}})} \geq \rho^2_{n-d_a, p-d_a}. \tag{3.10}$$

This proposition is a straightforward corollary of Theorem 2.3 in Chapter 2. One interprets the quantity appearing in (3.10) as follows: the quotient of conditional variances measures the ratio of the quantity of information brought by X_i for the prediction of X_a to the part of X_a not explained by $X_{ne(a) \cup \{i\}}$. In other words, the test T_α has a power larger than δ for vectors θ^a such that there exists a node $i \in \Gamma \setminus \overline{ne}(a)$ which improves enough the prediction of X_a.

This test is optimal in the minimax sense if we test against the alternative "$\theta^a_{\Gamma \setminus \overline{ne}(a)}$ has only one non-zero component" and if the covariates are independent (see Section 2.4.2). The condition of independence for covariates is unrealistic in a Gaussian graphical context, but it is nevertheless relevant as the independent case is an important benchmark from the minimax point of view (see Section 2.4.2 for more details). When the covariates are correlated we know from a simulation study (see Section 2.6) that using Procedure P_2 slightly improves the power of the test T_α.

3.2.3 Test of graph

From the test of neighborhood we define a procedure to test a graph. More precisely, we test the null hypothesis H_0 that "X is a Gaussian graphical model with respect to $\mathcal{G}$" against the alternative that it is not. Let $\{\alpha_a, a \in \Gamma\}$ be a collection of numbers in $]0,1[$. For each node $a \in \Gamma$, we test at level α_a the neighborhood of the node a with one of the procedures explained in Section 3.2.1.2. We decide to reject the null hypothesis H_0 as soon as one of the test $T^a_{\alpha_a}$ is rejected. We obtain a test of level α of the graph $\mathcal{G}$ if we take $\{\alpha_a, a \in \Gamma\}$ such that $\sum_{a\in\Gamma} \alpha_a = \alpha$. In the sequel we choose $\alpha_a = \alpha/p$ for each $a \in \Gamma$.

This procedure corresponds to a Bonferroni choice of the weights. As a consequence, if the number p of nodes is very large, our test may suffer a loss of its size. This restricts ourselves to consider tests of graph only for relatively small graphs, or for subgraphs of a large graph. Let us recall that when we apply the test of neighborhood to one node, the number p of nodes can be arbitrary large without any loss in the size of the test, provided that we use Procedure P_2 or P_3.

3.3 Simulations

In this section we present two simulation studies. First, we study the test of graph when the number of nodes is small. On the one hand we compare the efficiency of Procedures P_1 and P_2 and on the other hand we show the influence of the percentage of edges in the graph on the power of the test. Second, we study the test of neighborhood when p is large, illustrating the power of our procedure in a high-dimensional setting. Besides, we compare the efficiency of the tests based on the collections of models $\mathcal{M}^1_a$ and $\mathcal{M}^2_a$ defined in Section 3.2.1.4.

3.3.1 Simulation of a GGM

3.3.1.1 Simulation of a graph

In our simulations we use two different methods to generate random graphs. The first one allows to control the number of nodes p and the percentages of edges η in the graph. It consists in choosing uniformly and independently the positions of the $\eta \times p(p-1)/2$ edges. We use this method in the simulation experiment on the test of graph, with different values of η to measure the influence of the percentage of edges on the test.

However, the vertices of real-world networks are often structured in clusters, i.e groups of proteins functionally related, with different connectivity properties. That is why Daudin *et al.* [PDR08] proposed a model called ERMG for Erdös-Rényi Mixtures for Graphs, which describes the way edges connect nodes, accounting for some groups of nodes, and some preferential connections between the groups. The ERMG model assumes that the nodes are spread into Q clusters with probabilities $\{p_1, \ldots, p_Q\}$. We are given a connectivity matrix C of size $Q \times Q$ which specifies the probability of connection between two nodes according to the clusters they belong to. More precisely, the probability that two nodes belonging to the clusters i and j share an edge equals $C[i,j]$. We use this method to generate a graph in the simulation experiment on the test of neighborhood, with the following parameters provided by Daudin *et al.* [PDR08]: $p = 199$ nodes, $Q = 7$ clusters, the probabilities $(p_1, \ldots, p_Q)$ and the connectivity matrix C equal:

$$(p_1, \ldots, p_Q) = \begin{pmatrix} 0.038 & 0.052 & 0.060 & 0.082 & 0.083 & 0.125 & 0.560 \end{pmatrix}, \tag{3.11}$$

$$C = \begin{pmatrix} 0.999 & 0.319 & 1e-06 & 0.116 & 1e-06 & 1e-06 & 0.007 \\ 0.319 & 0.869 & 1e-06 & 1e-06 & 0.140 & 0.004 & 0.002 \\ 1e-06 & 1e-06 & 0.467 & 0.0155 & 0.005 & 0.014 & 0.004 \\ 0.116 & 1e-06 & 0.016 & 0.216 & 1e-06 & 0.017 & 0.005 \\ 1e-06 & 0.140 & 0.005 & 1e-06 & 0.229 & 1e-06 & 0.004 \\ 1e-06 & 0.004 & 0.014 & 0.017 & 1e-06 & 0.239 & 0.013 \\ 0.007 & 0.002 & 0.004 & 0.005 & 0.0041 & 0.0129 & 0.0163 \end{pmatrix}. \tag{3.12}$$

Using these parameters, the percentage of edges η in the graph equals 2.5%.

3.3.1.2 Simulation of the data

Given a graph we generate random vectors whose conditional independence structure is represented by the graph.

First, we generate the partial correlation matrix Π as follows : to a graph with p nodes we associate a symmetric $p \times p$ matrix U such that for any $(i,j) \in \{1,\dots,p\}^2$, $U[i,j]$ is drawn from the uniform distribution between -1 and 1 if there is an edge between the nodes i and j and $U[i,j]$ is set to 0 in the other case. We then compute column-wise sums of the absolute values of the matrix U entries, and set the corresponding diagonal element equal to this sum plus a small constant. This ensures that the resulting matrix is diagonally dominant and thus positive definite. Finally, we standardize the matrix so that the diagonal entries all equal 1 to obtain the simulated partial correlation matrix Π.

Second, we simulate data of the sample size n. We generate n independent samples from the multivariate normal distribution with mean zero, unit variance, and correlation structure associated to the partial correlation matrix Π. In the sequel, we note $\mathbf{X}$ the $n \times p$ associated data matrix.

3.3.2 Simulation setup

3.3.2.1 Simulation study of the test of graph

We evaluate the performance of the test of graph, first with simulations on randomly generated graphs, and secondly on a network coming from the data base KEGG.

1. First simulation experiment: We estimate the level and the power of the test of graph with 1000 simulations. For fixed parameters (p,η,n), we generate 1000 graphs by using the first method described in Section 3.3.1.1 and 1000 data matrices as described in Section 3.3.1.2. Let $\mathcal{G}^s$ and $\mathbf{X}^s$ for $s = 1,\dots,1000$ denote the graphs and the data matrices for the 1000 simulations. For each simulation s, we test the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}^s$". We thus estimate the level of the test by dividing the number of simulations for which we reject the null hypothesis by 1000. Let q be a number in $]0,1[$. For each simulation s, let $\mathcal{G}^s_{-q}$ be the graph built from the graph $\mathcal{G}^s$ in which we delete randomly $q\frac{p(p-1)}{2}\eta$ edges. For each simulation s, we test the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}^s_{-q}$". We estimate the power of the test by dividing the number of simulations for which we reject the null hypotheses by 1000.

   The number of variables p is set to 15, whereas the number of observations n is taken equal to 10, 15 and 30 to study the effect of the sample size. We examine the influence of the percentage of edges in the graph, by taking $\eta = 0.1$ and 0.15. Besides, we show the effect of the percentage q of missing edges on the power, by presenting the results for q equal to 10%, 40% and 100%.

2. Second simulation experiment: This simulation is based on the cell cycle of yeast (Saccharomyces cerevisiae). This experiment aims at showing the performance of our procedure with simulations on a real biological network. The graph corresponding to the cell cycle of yeast is available in the data base KEGG from the following website: `http://www.genome.jp/kegg/pathway/sce/sce04111.html`. We focus on a part of this pathway involving 16 proteins and 18 interactions. The graph, denoted in the sequel $\mathcal{G}_{cellcycle}$ is shown in Figure 3.1. We estimate the level and the power of the test by simulating 1000 data matrix $(\mathbf{X}^s)_{s=1,\dots,1000}$ from the graph $\mathcal{G}_{cellcycle}$ as described in Section 3.3.1.2. We first estimate the level of the test by testing for each simulation s, the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_{cellcycle}$". Then, we delete the three edges involving the protein complex $SCF\ Cdc4$ in $\mathcal{G}_{cellcycle}$ in order to define the graph $\mathcal{G}^{-Cdc4}_{cellcycle}$. This protein complex $SCF\ Cdc4$ participates in cell death. We estimate the power of the test by testing for each simulation s the null hypothesis "$\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}^{-Cdc4}_{cellcycle}$. In other words we evaluate the ability of our procedure to detect the link of the protein complex $SCF\ Cdc4$ with the cell cycle.

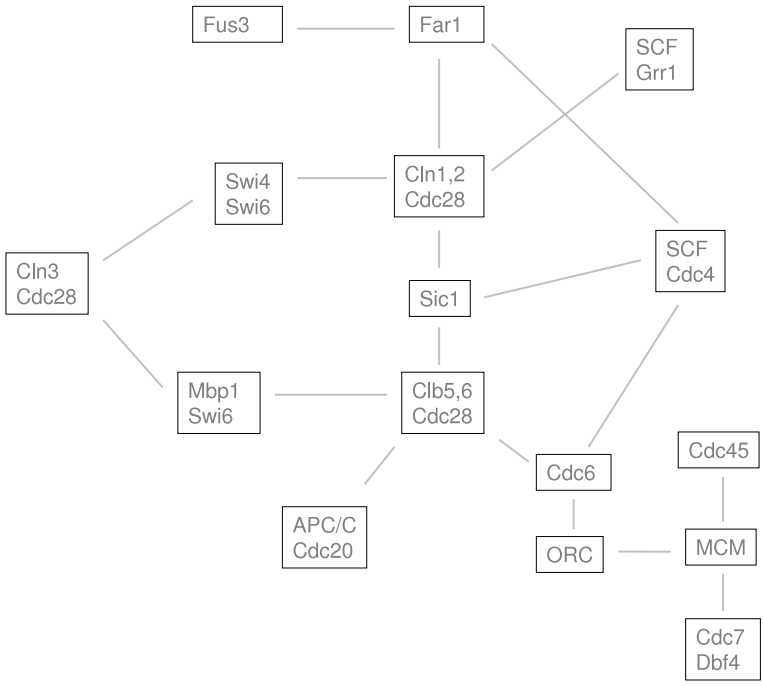


Figure 3.1: $\mathcal{G}_{cellcycle}$

3.3.2.2 Simulation study of the test of neighborhood

We first simulate a graph $\mathcal{G}$ according to the ERMG model described in Section 3.3.1.1 with $p = 199$ nodes, $Q = 7$ clusters, and the parameters $(p_1, \ldots, p_Q)$ and the matrix C defined in Equations (3.11) and (3.12). We then focus on a node a of this graph, chosen such that it has several neighbors. In our simulation this node has 6 neighbors. Let us denote $ne(a)$ its neighborhood given by the graph $\mathcal{G}$. We simulate 1000 data matrix as described in Section 3.3.1.2 from the graph $\mathcal{G}$ and estimate the level of the test by testing the null hypothesis that the node a has no other neighbor than the set $ne(a)$, and the power by testing the null hypothesis that the node a has no neighbor. We present results when the sample size n is equal to 50, 100, and 200.

3.3.2.3 Collections of models $\mathcal{M}_a$ and collections $\{\alpha_m, m \in \mathcal{M}_a\}$

For each node a, we use the testing procedure defined in (3.4) with different collections $\mathcal{M}_a$ and different choices of the weights $\{\alpha_m, m \in \mathcal{M}_a\}$. Let us recall that $ne(a)$ denotes the neighborhood of the node a under the null hypothesis and α_a the level of the test of neighborhood for the node a. For the test of graph we choose $\alpha_a = \alpha/p$ and for the test of neighborhood α_a equals α.

The collections $\mathcal{M}_a$: we consider the two collections defined in Section 3.2.1.4.

$$\mathcal{M}_a^1 = \{\{b\}, b \in \Gamma \setminus \overline{ne}(a)\}$$

and

$$\mathcal{M}_a^2 = \{\{j_1, \ldots, j_k\}, 1 \leq k \leq J\}.$$

where $S_{Lars}\left[\Psi\left(\mathbf{X}_a, \mathbf{X}_{-a}\right)\right] = \{j_1, j_2, \ldots, j_J\}$ is the sequence given by the LARS algorithm for the prediction of $\Pi_{ne(a)^\perp}\mathbf{X}_a$ with the set of covariates $\Pi_{ne(a)^\perp}\mathbf{X}_b$ where $b \in \Gamma \setminus \overline{ne}(a)$. The maximum number of steps J is taken equal to 10. We evaluate the performance of our testing procedure with $\mathcal{M}_a^1$ in the simulation experiment on the test of graph, and we compare collections $\mathcal{M}_a^1$ and $\mathcal{M}_a^2$ in the simulation experiment on the test of neighborhood. Indeed, in the second simulation experiment p and thus the collection $\mathcal{M}_a^1$ are large. It is therefore interesting to compare their respective computational cost.

The collection $\{\alpha_m, m \in \mathcal{M}_a\}$: When we consider the collection of models $\mathcal{M}_a^1$ we use either Procedure P_1 or Procedure P_2 defined in Section 3.2.1.2. For Procedure P_1 the α_m's are taken equal to

$\alpha_a/|\mathcal{M}_a|$. The quantity $q_{\mathbf{X}_{-a},\alpha_a}$ occurring in Procedure P_2 is evaluated by simulation. Let Z be a standard Gaussian random vector of size n independent from $\mathbf{X}_{-a}$. As $\boldsymbol{\epsilon}_a$ is independent from $\mathbf{X}_{-a}$, the distribution of (3.6) conditionally to $\mathbf{X}_{-a}$ is the same as the distribution of

$$\inf_{m\subset\mathcal{M}_a} \bar{F}_{D_m,N_m}\frac{\|\Pi_{ne(a)\cup m}(Z)-\Pi_{ne(a)}(Z)\|^2/D_m}{\|Z-\Pi_{ne(a)\cup m}(Z)\|^2/N_m},$$

conditionally to $\mathbf{X}_{-a}$. Consequently, we estimate the quantile $q_{\mathbf{X}_{-a},\alpha_a}$ by a Monte-Carlo method with 1000 samples. When we use the collection $\mathcal{M}_a^2$ we apply Procedure P_3. The quantile $q'_{\mathbf{X}_{-a},\alpha_a}$ is again computed by a Monte-Carlo method with 1000 simulations.. The difference with the simulation of $q_{\mathbf{X}_{-a},\alpha_a}$ lies in the fact that the collection $\mathcal{M}_a^2$ is random and depends on ϵ_a. For each simulation, let Z be a standard Gaussian random vector of size n independent from $\mathbf{X}_{-a}$. We apply the LARS algorithm for the prediction of $\Pi_{ne(a)^\perp}Z$ with the set of covariates $\Pi_{ne(a)^\perp}\mathbf{X}_b$ where $b\in\Gamma_{-a}\setminus ne(a)$. We obtain the sequence $S_{Lars}\left[\Psi\left(Z,\mathbf{X}_{-a}\right)\right]$ which leads to the collection of models $\mathcal{M}_a^2\left[\Psi\left(Z,\mathbf{X}_{-a}\right)\right]$. The Ψ function is defined in (3.7). As $\boldsymbol{\epsilon}_a$ is independent from $\mathbf{X}_{-a}$, the distribution of (3.8) conditionally to $\mathbf{X}_{-a}$ is the same as the distribution of

$$\inf_{m\in\mathcal{M}_a[\Psi(Z,\mathbf{X}_{-a})]} \bar{F}_{D_m,N_m}\left(\frac{\|\Pi_{ne(a)\cup m}Z-\Pi_{ne(a)}Z\|_n^2/D_m}{\|Z-\Pi_{ne(a)\cup m}Z\|_n^2/N_m}\right),$$

conditionally to $\mathbf{X}_{-a}$ and we therefore estimate the quantile $q'_{\mathbf{X}_{-a},\alpha_a}$. In the sequel, we note $T_{\mathcal{M}_a^i,P_j}$ the test (3.4) with collection $\mathcal{M}_a^i$ and Procedure P_j.

3.3.3 The results

In Table 3.1 and 3.2 we present results of the first simulation experiment on the test of graph respectively for $\eta=0.1$ and $\eta=0.15$. As expected, the power of the tests increases with the number of observations n. Besides, the power of the tests increases also with the percentage of missing edges q, the tests being indeed more powerful when the graphs under the null and the alternative hypotheses are more different. As expected, the tests based on Procedure P_2 are more powerful than the corresponding tests based on Procedure P_1. However because p is small, the difference between the two procedures is not really significant. Nevertheless, Procedure P_1 may become too conservative when p is large. As expected, its implementation is faster: for $p=15$ and $n=10$ a single simulation using Procedure P_1 takes approximately a tenth of a second whereas a single simulation using Procedure P_2 takes approximately 9 seconds. For p small, Procedure P_1 is therefore a good compromise in practice, Procedure P_2 being rather recommended when considering large graphs. Let us now compare the influence of η on the power of the test. When the percentage of edges η in the graph increases, the tests are less powerful. It is especially significant for $q=10\%$. In fact, when η increases the average number of neighbors for each node increases as well. In practice, the test of neighborhood is less powerful for a node which already has several neighbors under the null hypothesis. Consequently, the issue of testing the graph is more difficult when η is large.

In Table 3.3 we give the results of the second experiment for the test of graph. The percentage of edges in the graph $\mathcal{G}_{cellcycle}$ equals 15%, whereas the ratio of missing edges is $q=1/6$ as we delete 3 edges among 18 in $\mathcal{G}_{cellcycle}$. In fact, as q is between 10% and 40% the powers of the tests in this setting are comparable to the results in Table 3.2. For $n=20$ observations the test is powerful and detects the relation between the protein complex $SCF\ Cdc4$ and the cell cycle with large probability. Even when n is smaller than p, the test detects the relation with a moderate probability.

In Table 3.4 we give the results of the experiment on the test of neighborhood. For $n=50$ and 100 the test is more powerful when using the collection of models $\mathcal{M}_a^1$ whereas when n is larger both procedures exhibit a comparable power. This comes from the fact that the test with collection $\mathcal{M}_a^2$ is performed in two steps: first, the selection of the relevant covariates using LARS and second, the test (3.4) itself. When n is small, LARS makes mistakes and possibly selects irrelevant covariates. In this case, the collection of models is bad and the test seldom rejects. When n is large, LARS often selects the relevant variables and the test $T_{\mathcal{M}^2,P_3}$ therefore takes advantage of exploiting models of several dimensions. However, its performances are not much better than the ones of $T_{\mathcal{M}^1,P_2}$ even when n is large. Let us now compare the computational efficiency of these two procedures. For $p=200$ and $n=100$ a single simulation using collection $\mathcal{M}_a^1$ is almost three times longer than using collection $\mathcal{M}_a^2$. It seems natural to exploit

Table 3.1: Test of graph, first simulation. $\eta = 0.1$. Estimated levels and powers. The nominal level is $\alpha = 5\%$. The standard deviation of these estimators equals 0.007.

Estimated levels

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.028	0.046
15	0.035	0.061
30	0.033	0.054

Estimated powers

$q = 10\%$			$q = 40\%$			$q = 100\%$		
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.73	0.75	10	0.94	0.94	10	0.99	0.99
15	0.83	0.84	15	0.97	0.98	15	1	1
30	0.95	0.95	30	1	1	30	1	1

Table 3.2: Test of graph, first simulation. $\eta = 0.15$. Estimated levels and powers. The nominal level is $\alpha = 5\%$. The standard deviation of these estimators equals 0.007.

Estimated levels

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.031	0.050
15	0.044	0.053
30	0.041	0.058

Estimated powers

$q = 10\%$			$q = 40\%$			$q = 100\%$		
n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$	n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.28	0.32	10	0.70	0.72	10	0.90	0.91
15	0.44	0.46	15	0.87	0.88	15	0.99	0.99
30	0.73	0.75	30	0.99	0.99	30	1	1

Table 3.3: Test of graph, second simulation experiment. Estimated levels and powers. The nominal level is $\alpha = 5\%$. The standard deviation of these estimators equals 0.007.

Estimated levels

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.040	0.055
20	0.046	0.063
30	0.040	0.058

Estimated powers

n	$T_{\mathcal{M}^1,P_1}$	$T_{\mathcal{M}^1,P_2}$
10	0.43	0.46
20	0.76	0.79
30	0.89	0.90

model of several dimensions especially when we consider the test of neighborhood for a node which has several missing neighbors. However, the LARS algorithm does not really improve the performance of the procedure. Nevertheless, using collection $\mathcal{M}_a^2$ is computationally more attractive than using collection $\mathcal{M}_a^1$.

Table 3.4: Test of neighborhood for the simulation experiment described in Section 3.3.2.2. Estimated levels and powers. The nominal level is $\alpha = 5\%$. The standard deviation of these estimators equals 0.007.

Estimated levels

n	$T_{\mathcal{M}^1,P_2}$	$T_{\mathcal{M}^2,P_3}$
50	0.056	0.052
100	0.044	0.054
200	0.041	0.043

Estimated powers

n	$T_{\mathcal{M}^1,P_2}$	$T_{\mathcal{M}^2,P_3}$
50	0.19	0.15
100	0.47	0.41
200	0.85	0.86

3.4 Application to biological data

In this section, we apply the test of graph to the multivariate flow cytometry data produced by Sachs *et al.* [SPP+05]. These data concern a human T cell signaling pathway whose deregulation may lead to carcenogenesis. Therefore, this pathway was extensively studied in the literature and a network involving 11 proteins and 16 interactions was conventionally accepted (see [SPP+05]). See Figure 3.2 for a representation of this network. The data from Sachs consist of quantitative amounts of these 11 proteins, simultaneously measured from single cells under perturbation conditions. In the sequel, we focus on one general perturbation (anti-CD3/CD28 + ICAM-2) that overall stimulates the cellular signaling network. In this condition the quantities of the 11 proteins are measured in 902 cells. Let denote D this data set constituted of $p = 11$ variables and $n = 902$ observations. Contrary to most of postgenomic data, flow cytometry data provide a large sample of observations that allow us to measure the influence of the sample size on the power. From this data set we infer the network using three methods and we apply our test of graph as a tool to validate these estimations. As such abundance of data is rarely available in postgenomic data, we secondly carry out a simulation study to determine the influence of the number of observations on the test. From the empirical covariance matrix obtained with the whole data set D, we generate data of different sample sizes and we evaluate the performance of the test with respect to the sample size.

We use the methods proposed by Drton and Perlman [DP08], Wille and Bühlmann [WB06], and Meinshausen and Bühlmann [MB06] to infer the network. Let us briefly describe them. The SINful approach introduced by Drton and Perlman is a model selection algorithm based on multiple testing. For any couple of nodes they perform a test of existence of an edge between these two nodes and select the graph by computing the simultaneous p-values of these tests. This method assumes that the number of observations n is larger than the number of variables p. The two other methods have been recently proposed to deal with the usual fact in genomics of p large and n small. Wille and Bühlmann [WB06] estimate a lower-order conditional independence graph instead of the concentration graph, while Meinshausen and Bühlmann [MB06] estimate the neighborhood of any node with the Lasso method. We represent the three estimated graphs in Figure 3.3.

Let us define the graph $\mathcal{G}_\cap$ as the intersection of the graph estimated by these three methods and of the graph with the connections well-established in the literature. This graph $\mathcal{G}_\cap$ is represented in Figure 3.4. We test with our procedure the null hypothesis $H_{\mathcal{G}_\cap}$: "the data set D follows the distribution of a Gaussian graphical model with respect to the graph $\mathcal{G}_\cap$". We use for each node a of the graph the collection of models $\mathcal{M}_a^1$ defined in Section 3.2.1.4 and the procedure P_1. As p is small, the difference between Procedure P_2 and P_1 is indeed not significant and the implementation of P_1 is faster. If we apply our procedure at level $\alpha = 5\%$, we reject the null hypothesis $H_{\mathcal{G}_\cap}$. In fact the p-value of the test is smaller than 10^{-10}. As our procedure consists in testing the neighborhood of each node, it is interesting to look

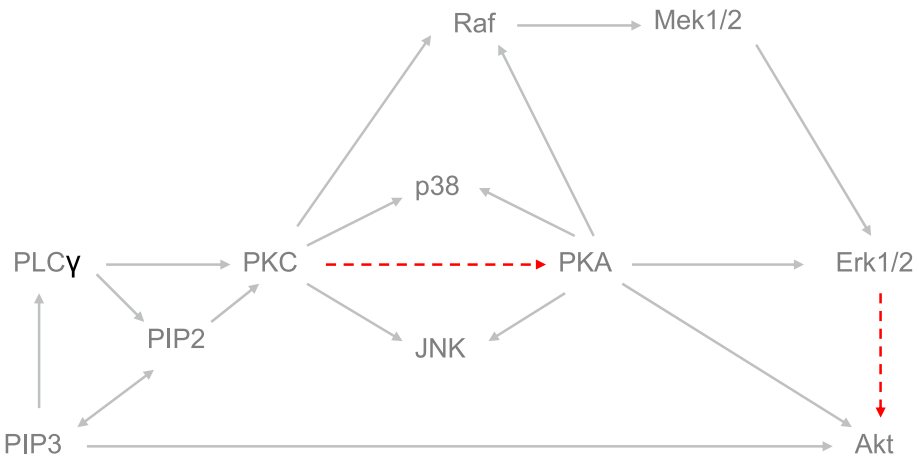


Figure 3.2: Classic signaling network of the human T cell pathway. The connections well-established in the literature are in grey and the connections cited at least once in the literature are represented by red dotted lines.

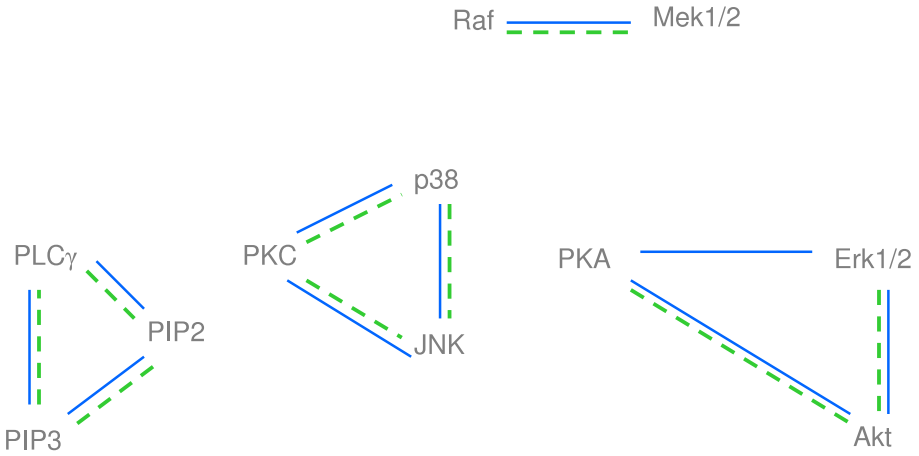


Figure 3.3: Inferred graphs. The graphs estimated with the methods of Drton and Perlman and Wille and Bühlmann are identical and represented in blue. The graph estimated with the method of Meinshausen and Bühlmann is in green dotted line

for the nodes for which the test of neighborhood is rejected. For any of these rejected neighborhood tests, we then look for the alternatives leading to this rejection. In Table 3.5 we enumerate the nodes for which the test of neighborhood is rejected and the alternatives which lead to this decision.

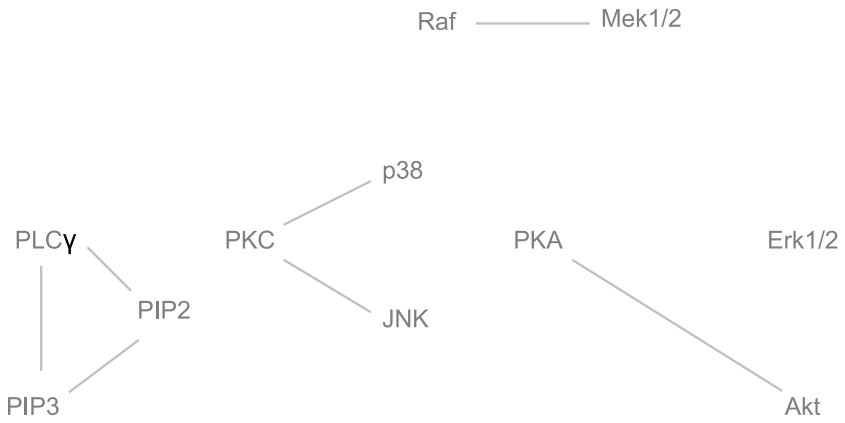


Figure 3.4: Graph $\mathcal{G}_\cap$

Table 3.5: Rejection of $H_{\mathcal{G}_\cap}$.

Rejection of the neighborhood of	because of node(s)
node	
Erk1/2	Akt, PKA
Akt	Erk1/2
PKA	Erk1/2
p38	JNK
JNK	p38

As the connection $PKA - Erk1/2$ is well-established and the connection $Erk1/2 - Akt$ is cited at least once in the literature, we decide to add those two edges in the graph $\mathcal{G}_\cap$, defining thus a new graph $\mathcal{G}_2$ shown in Figure 3.5. The test of the null hypothesis $H_{\mathcal{G}_2}$ at level $\alpha = 5\%$: "the data set D follows the distribution of a Gaussian graphical model with respect to the graph $\mathcal{G}_2$" is rejected, the p-value of the test being smaller than 10^{-10}. The reason is that the tests concerning respectively nodes $p38$ and JNK are rejected when we consider in the alternative respectively nodes JNK and $p38$.

We therefore define a new graph $\mathcal{G}_T$ by adding the connection $p38 - JNK$, even if this connection is not well-established in the literature. Let us note that the graph $\mathcal{G}_T$ is the same as the network inferred by Sachs *et al.* [SPP+05] with approximately the same data set by using a Bayesian approach. We apply our test of graph and we accept the hypothesis that the data set D is a Gaussian graphical model with respect to the graph $\mathcal{G}_T$ at the level $\alpha = 5\%$. In fact, the p-value of the test equals 8%. As n is large we use the result of the test with confidence and assume that the graph $\mathcal{G}_T$ (Figure 3.6) represents the conditional independence structure of the data set D.

We now carry out a simulation study from this data set to determine the influence of the number of observations n on the power of our procedure. >From the empirical covariance matrix obtained with the data set D, we generate 1000 simulated data $(\mathbf{X}^s)_{s=1,\ldots,1000}$ of different sample sizes n whose conditional independence structure is represented by the graph $\mathcal{G}_T$. First, we estimate the level of the test for different values of n by testing for each simulation that $\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_T$. Second, we delete the two edges involving protein PKC in $\mathcal{G}_T$ in order to define $\mathcal{G}_T^-$. We estimate the power of the test for different values of n by testing for each simulation that $\mathbf{X}^s$ is a Gaussian graphical model with respect to the graph $\mathcal{G}_T^-$.

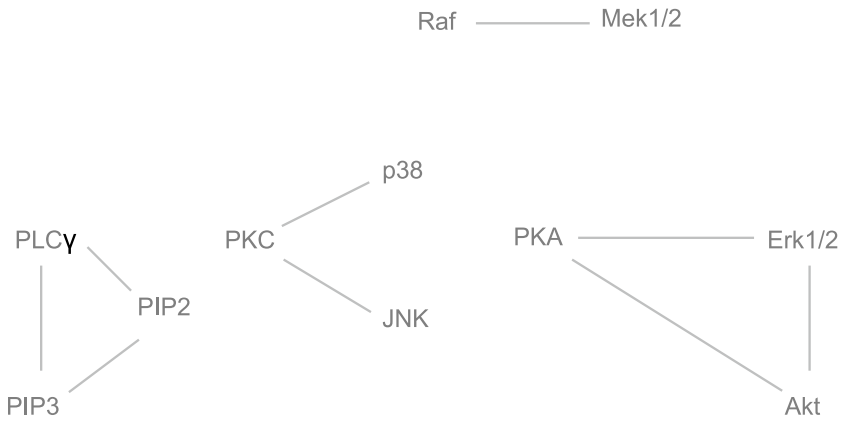


Figure 3.5: Graph $\mathcal{G}_2$

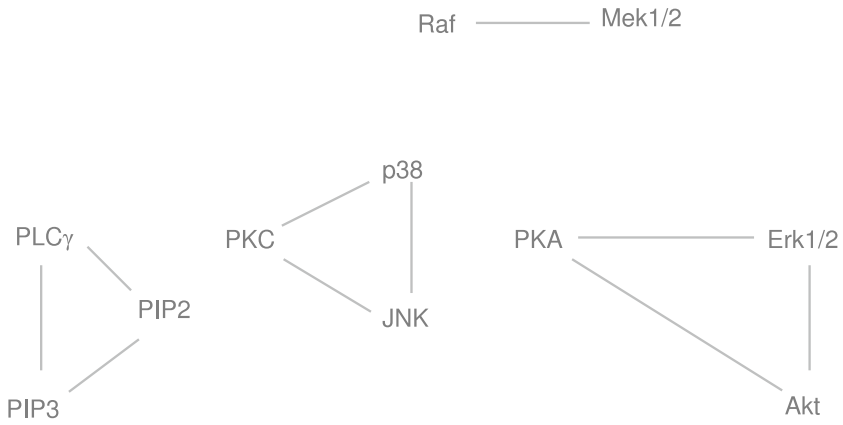


Figure 3.6: Graph $\mathcal{G}_T$

Table 3.6: Sachs data. Estimated levels and powers

Estimated levels

n	$T_{\mathcal{M}^1,P_1}$
10	0.032
15	0.036
20	0.033

Estimated powers

n	$T_{\mathcal{M}^1,P_1}$
10	0.49
15	0.86
20	0.97

The results of the simulation study from the selected Sachs' data are presented in Table 3.6. We recall that the graph involves $p = 11$ proteins and we take for the sample size n the values 10, 15, and 20. As expected, the power of the test increases with the number of observations n. However, the number of observations do not have to be very large to obtain a powerful test. For $n = 15$ observations the test is able to recover that the protein PKC is not independent from the proteins $p38$ and JNK with large probability.

3.5 Conclusion

In this chapter, we propose a multiple testing procedure to assess whether some connections are missing in a minimal graph derived from experimental knowledge. Besides, when the procedure is rejected the different p-values of the tests suggest potential connections between genes/proteins that steer biologists towards new experimentations.

Our procedure is feasible in a high-dimensional setting. Hence, we advise it to analyze microarray data for which the number of genes p typically exceeds the number of samples. Of course, when p becomes very large the power of the procedure decreases but this is intrinsic to the statistical problem.

Chapter 4

High-dimensional Gaussian model selection on a Gaussian design

Abstract. We consider the problem of estimating the conditional mean of a real Gaussian variable $Y = \sum_{i=1}^{p} \theta_i X_i + \epsilon$ where the vector of the covariates $(X_i)_{1 \leq i \leq p}$ follows a joint Gaussian distribution. This issue often occurs when one aims at estimating the graph or the distribution of a Gaussian graphical model. We introduce a general model selection procedure which is based on the minimization of a penalized least squares type criterion. It handles a variety of problems such as ordered and complete variable selection, allows to incorporate some prior knowledge on the model and applies when the number of covariates p is larger than the number of observations n. Moreover, it is shown to achieve a non-asymptotic oracle inequality independently of the correlation structure of the covariates. We also exhibit various minimax rates of estimation in the considered framework and hence derive adaptivity properties of our procedure.

4.1 Introduction

4.1.1 Regression model

We consider the following regression model

$$Y \quad = \quad X\theta + \epsilon \ , \tag{4.1}$$

where θ is an unknown vector of $\mathbb{R}^p$. The row vector $X := (X_i)_{1 \leq i \leq p}$ follows a real zero mean Gaussian distribution with non singular covariance matrix Σ and ϵ is a real zero mean Gaussian random variable independent of X with variance σ^2. The variance of ϵ corresponds to the conditional variance of Y given X, $\mathrm{var}(Y|X)$. In the sequel, the parameters θ, Σ, and σ^2 are considered as unknown.

Suppose we are given n i.i.d. replications of the vector (Y, X). We respectively write $\mathbf{Y}$ and $\mathbf{X}$ for the vector of n observations of Y and the $n \times p$ matrix of observations of X. In the present work, we propose a new procedure to estimate the vector θ, when the matrix Σ and the variance σ^2 are both unknown. This corresponds to estimating the conditional expectation of the variable Y given the random vector X. Besides, we want to handle the difficult case of high-dimensional data, i.e. the number of covariates p is possibly much larger than n. This estimation problem is equivalent to building a suitable predictor of Y given the covariates $(X_i)_{1 \leq i \leq p}$. Classically, we shall use the mean-squared prediction error to assess the quality of our estimation. For any $(\theta_1, \theta_2) \in \mathbb{R}^p$, it is defined by

$$l(\theta_1, \theta_2) := \mathbb{E}\left[(X\theta_1 - X\theta_2)^2 \right] \ . \tag{4.2}$$

4.1.2 Applications to Gaussian graphical models (GGM)

Estimation in the regression model (4.1) is mainly motivated by the study of Gaussian graphical models (GGM). Let Z be a Gaussian random vector indexed by the elements of a finite set Γ. The vector Z is

a GGM with respect to an undirected graph $\mathcal{G} = (\Gamma, E)$ if for any couple (i, j) which is not contained in the edge set E, Z_i and Z_j are independent, given the remaining variables. See Lauritzen [Lau96] for definitions and main properties of GGM. Estimating the neighborhood of a given point $i \in \Gamma$ is equivalent to estimating the support of the regression of Z_i with respect to the covariates $(Z_j)_{j\in\Gamma\setminus\{i\}}$. Meinshausen and Bühlmann [MB06] have taken this point of view in order to estimate the graph of a GGM. Similarly, we can apply the model selection procedure we shall introduce in this chapter to estimate the support of the regression and therefore the graph $\mathcal{G}$ of a GGM.

Interest in these models has grown since they allow the description of dependence structure of high-dimensional data. As such, they are widely used in spatial statistics [Cre93, RH05] or probabilistic expert systems [CDLS99]. More recently, they have been applied to the analysis of microarray data. The challenge is to infer the network regulating the expression of the genes using only a small sample of data, see for instance Schäfer and Strimmer [SS05], or Wille *et al.* [WZV$^+$04].

This has motivated the search for new estimation procedures to handle the linear regression model (4.1) with Gaussian random design. Finally, let us mention that the model (4.1) is also of interest when estimating the distribution of directed graphical models or more generally the joint distribution of a large Gaussian random vector. Estimating the joint distribution of a Gaussian vector $(Z_i)_{1\leq i\leq p}$ indeed amounts to estimating the conditional expectations and variance of Z_i given $(Z_j)_{1\leq j\leq i-1}$ for any $1 \leq i \leq p$.

4.1.3 General oracle inequalities

Estimation of high-dimensional Gaussian linear models has now attracted a lot of attention. Various procedures have been proposed to perform the estimation of θ when $p > n$. The challenge at hand it to design estimators that are both computationally feasible and are proved to be efficient. The Lasso estimator has been introduced by Tibshirani [Tib96]. Meinshausen and Bühlmann [MB06] have shown that this estimator is consistent under a neighborhood stability condition. These convergence results were refined in the works of Zhao and Yu [ZY06], Bunea *et al.* [BTW07a], Bickel *et al.* [BRT08], or Candès and Plan [CP08] in a slightly different framework. Candès and Tao [CT07] have also introduced the Dantzig-selector procedure which performs similarly as l_1 penalization methods. In the more specific context of GGM, Bühlmann and Kalisch [BK07] have analyzed the PC algorithm and have proven its consistency when the GGM follows a faithfulness assumption. All these methods share an attractive computational efficiency and most of them are proven to converge at the optimal rate when the covariates are nearly independent. However, they also share two main drawbacks. First, the l_1 estimators are known to behave poorly when the covariates are highly correlated and even for some covariance structures with small correlation (see e.g. [CP08]). Similarly, the PC algorithm is not consistent if the faithfulness assumption is not fulfilled. Second, these procedures do not allow to integrate some biological or physical prior knowledge. Let us provide two examples. Biologists sometimes have a strong preconception of the underlying biological network thanks to previous experimentations. For instance, Sachs *et al.* [SPP$^+$05]) have produced multivariate flow cytometry data in order to study a human T cell signaling pathway. Since this pathway has important medical implications, it was already extensively studied and a network is conventionally accepted (see [SPP$^+$05]). For this particular example, it could be more interesting to check whether some interactions were forgotten or some unnecessary interactions were added in the model than performing a complete graph estimation. Moreover, the covariates have in some situations a temporal or spatial interpretation. In such a case, it is natural to introduce an *order* between the covariates, by assuming that a covariate which is *close* (in space or time) to the response Y is more likely to be significant. Hence, an ordered variable selection method is here possibly more relevant than the complete variable selection methods previously mentioned.

Let us emphasize the main differences of our estimation setting with related studies in the litterature. Birgé and Massart [BM01] consider model selection in a fixed design setting with known variance. Bunea *et al.* [BTW07b] also suppose that the variance is known. Yet, they consider a random design setting, but they assume that the regression functions are bounded (Assumption A.2 in their paper) which is not the case here. Moreover, they obtain risk bounds with respect to the empirical norm $\|\mathbf{X}(\widehat{\theta}-\theta)\|_n^2$ and not the integrated loss $l(.,.)$. Here, $\|.\|_n$ refers to the canonical norm in $\mathbb{R}^n$ reweighted by $\sqrt{n}$. As mentioned earlier, our objective is to infer the conditional expectation of Y given X. Hence, it is more significant to assess the risk with respect to the loss $l(.,.)$. Baraud *et al.* [BGH08] consider fixed design regression

but do not assume that the variance is known.

Our objective is twofold. First, we introduce a general model selection procedure that is very flexible and allows to integrate any prior knowledge on the regression. We prove non-asymptotic oracle inequalities that hold without any assumption on the correlation structure between the covariates. Second, we obtain non-asymptotic rates of estimation for our model (4.1) that help us to derive adaptive properties for our criterion.

In the sequel, a *model* m stands for a subset of $\{1,\dots,p\}$. We note d_m the size of m whereas the linear space S_m refers to the set of vectors $\theta \in \mathbb{R}^p$ whose support is included in m. If d_m is smaller than n, then we define $\widehat{\theta}_m$ as the least-square estimator of θ over S_m. In the sequel, Π_m stands for the projection of $\mathbb{R}^n$ into the space generated by $(\mathbf{X}_i)_{i\in m}$. Hence, we have the relation $\mathbf{X}\widehat{\theta}_m = \Pi_m \mathbf{Y}$. Since the covariance matrix Σ is non singular, observe that almost surely the rank of Π_m is d_m. Given a collection $\mathcal{M}$ of models, our purpose is to select a model $\widehat{m} \in \mathcal{M}$ that exhibits a risk as small as possible with respect to the prediction loss function $l(.,.)$ defined in (4.2). The model m^* that minimizes the risks $\mathbb{E}[l(\widehat{\theta}_m, \theta)]$ over the whole collection $\mathcal{M}$ is called an oracle. Hence, we want to perform as well as the oracle $\widehat{\theta}_{m^*}$. However, we do not have access to m^* as it requires the knowledge of the true vector θ. A classical method to estimate a *good* model $\widehat{m}$ is achieved through *penalization* with respect to the complexity of models. In the sequel, we shall select the model $\widehat{m}$ as

$$\widehat{m} := \arg\min_{m\in\mathcal{M}} \mathrm{Crit}(m) := \arg\min_{m\in\mathcal{M}} \|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2 \left[1 + \mathrm{pen}(m)\right] \ , \tag{4.3}$$

where pen(.) is a positive function defined on $\mathcal{M}$. Besides, we recall that $\|.\|_n$ refers to the canonical norm in $\mathbb{R}^n$ reweighted by $\sqrt{n}$. Observe that $Crit(m)$ is the sum of the least-square error $\|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2$ and a penalty term pen(m) rescaled by the least-square error in order to come up with the fact that the conditional variance σ^2 is unknown. We precise in Section 4.2 the heuristics underlying this model selection criterion. Baraud *et al.* [BGH08] have extensively studied this penalization method in the fixed design Gaussian regression framework with unknown variance. In their introduction, they explain how one may retrieve classical criteria like AIC [Aka73], BIC [Sch78], and FPE [Aka70] by choosing a suitable penalty function pen(.).

This model selection procedure is really flexible through the choice of the collection $\mathcal{M}$ and of the penalty function pen(.). Indeed, we may perform complete variable selection by taking the collection of subsets of $\{1,\dots,p\}$ whose is smaller than some integer d. Otherwise, by taking a nested collection of models, one performs ordered variable selection. We give more details in Sections 4.2 and 4.3. If one has some prior idea on the true model m, then one could only consider the collection of models that are close in some sense to m. Moreover, one may also give a Bayesian flavor to the penalty function pen(.) and hence specify some prior knowledge on the model.

First, we state a non-asymptotic oracle inequality when the complexity of the collection $\mathcal{M}$ is small and for penalty functions pen(m) that are larger than $Kd_m/(n-d_m)$ with $K > 1$. Then, we prove that the FPE criterion of Akaike [Aka70] which corresponds to the choice $K = 2$ achieves an asymptotic exact oracle inequality for the special case of ordered variable selection. For the sake of completeness, we prove that choosing K smaller than one yields terrible performances.

In Section 4.3.2, we consider general collection of models $\mathcal{M}$. By introducing new penalties that take into account the complexity of $\mathcal{M}$ as in [BM07], we are able to state a non-asymptotic oracle inequality. In particular, we consider the problem of complete variable selection.

Interestingly, these rates of convergence do not depend on the covariance matrix Σ of the covariates, whereas known results on the Lasso or the Dantzig selector rely on some assumptions on Σ, as discussed in Section 4.3.2. We illustrate in Section 4.5 on simulated examples that for some covariance matrices Σ the Lasso performs poorly whereas our methods still behaves well. Besides, our penalization method does not require the knowledge of the conditional variance σ^2. In contrast, the Lasso and the Dantzig selector are constructed for known variance. Since σ^2 is unknown, one either has to estimate it or has to use a cross-validation method in order to calibrating the penalty. In both cases, there is some room for

improvements for the practical calibration of these estimators.

However, our model selection procedure suffers a computational cost that depends linearly on the size of the collection $\mathcal{M}$. For instance, the complete variable selection problem is NP-hard. This makes it intractable when p becomes too large (i.e. more than 50). In contrast, our criterion applies for arbitrary p when considering ordered variable selection since the size of $\mathcal{M}$ is linear with n. We shall mention in the discussion some possible extensions that we hope can cope with the computational issues.

In a simultaneous and independent work to ours, Giraud [Gir08a] applies an analogous procedure to estimate the graph of a GGM. Using slightly different techniques, he obtains non-asymptotic results that are complementary to ours. However, he needs to threshold his estimators in order to study them. Moreover, he does not consider the case of nested collections of models as we do in Section 4.2. Finally, he does not derive minimax rates of estimation.

4.1.4 Minimax rates of estimation

In order to assess the optimality of our procedure, we investigate in Section 4.4 the minimax rates of estimation for ordered and complete variable selection. For ordered variable selection, we compute the minimax rate of estimation over ellipsoids which is analogous to the rate obtained in the fixed design framework. We derive that our penalized estimator is adaptive to the collection of ellipsoids and to the covariance matrix Σ. For complete variable selection, we prove that the minimax rates of estimator of vectors θ with at most k non-zero components is of order $\frac{k \log p}{n}$ when the covariates are independent. This is again coherent with the situation observed in the fixed design setting. Then, the estimator $\widetilde{\theta}$ defined for complete variable selection problem is shown to be adaptive to any sparse vector θ. Moreover, it seems that the minimax rates may become faster when the matrix Σ is far from identity. We investigate this phenomenon in Section 4.4.2. All these minimax rates of estimation are, to our knowledge, new in the Gaussian random design regression. Tsybakov [Tsy03] has derived minimax rates of estimation in a general random design regression setup, but his results do not apply in our setting as explained in Section 4.4.2.

4.1.5 Organization of the chapter and some notations

In Section 4.2, we precise our estimation procedure and explain the heuristics underlying the penalization method. The main results are stated in Section 4.3. In Section 4.4, we derive the different minimax rates of estimation and assess the adaptivity of the penalized estimator $\widehat{\theta}_{\widehat{m}}$. We perform a simulation study and compare the behaviour of our estimator with Lasso and adaptive Lasso in Section 4.5 . Section 4.6 contains a final discussion and some extensions, whereas the proofs are postponed to Section 4.7.

Throughout the chapter, $\|.\|_n^2$ stands for the square of the canonical norm in $\mathbb{R}^n$ reweighted by n. For any vector Z of size n, we recall that $\Pi_m Z$ denotes the orthogonal projection of Z onto the space generated by $(\mathbf{X}_i)_{i\in m}$. The notation X_m stands for $(X_i)_{i\in m}$ and $\mathbf{X}_m$ represents the $n \times d_m$ matrix of the n observations of X_m. For the sake of simplicity, we write $\widetilde{\theta}$ for the penalized estimator $\widehat{\theta}_{\widehat{m}}$. For any $x > 0$, $\lfloor x \rfloor$ is the largest integer smaller than x and $\lceil x \rceil$ is the smallest integer larger than x. Finally, L, L_1, L_2,... denote universal constants that may vary from line to line. The notation $L(.)$ specifies the dependency on some quantities.

4.2 Estimation procedure

In this section, we detail our model selection procedure and provide an heuristic explanation. For any vector θ' in $\mathbb{R}^p$, we define the mean-squared error $\gamma(.)$ and its empirical counterpart $\gamma_n(.)$ as

$$\gamma(\theta') := \mathbb{E}_\theta \left[\left(Y - X\theta' \right)^2 \right] \quad \text{and} \quad \gamma_n(\theta') := \|\mathbf{Y} - \mathbf{X}\theta'\|_n^2 \ . \tag{4.4}$$

The function $\gamma(.)$ is closely connected to the loss function $l(.,.)$ through the relation $l(\beta, \theta) = \gamma(\beta) - \gamma(\theta)$.

Given a model m of size strictly smaller than n, we refer to θ_m as the unique minimizer of $\gamma(.)$ over the subset S_m. It then follows that $\mathbb{E}\left(Y | X_m\right) = \sum_{i\in m} \theta_i X_i$ and $\gamma(\theta_m)$ is the conditional variance of Y

given X_m. As for it, the least squares estimator $\widehat{\theta}_m$ is the minimizer of $\gamma_n(.)$ over the space S_m.

$$\widehat{\theta}_m := \arg\min_{\theta' \in S_m} \gamma_n(\theta') \text{ a.s. .}$$

It is almost surely uniquely defined since Σ is assumed to be non-singular and since $d_m < n$. Besides $\gamma_n(\widehat{\theta}_m)$ equals $\|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2$. Let us derive two simple properties of $\widehat{\theta}_m$ that will give us some hints to perform model selection.

Lemma 4.1. *For any model m whose dimension is smaller than $n-1$, the expected mean-squared error of $\widehat{\theta}_m$ and the expected least squares of $\widehat{\theta}_m$ respectively equal*

$$\mathbb{E}\left[\gamma(\widehat{\theta}_m)\right] = \left[l(\theta_m, \theta) + \sigma^2\right]\left(1 + \frac{d_m}{n - d_m - 1}\right) , \tag{4.5}$$

$$\mathbb{E}\left[\gamma_n(\widehat{\theta}_m)\right] = \left[l(\theta_m, \theta) + \sigma^2\right]\left(1 - \frac{d_m}{n}\right) . \tag{4.6}$$

The proof is postponed to the Appendix. From Equation (4.5), we derive a bias variance decomposition of the risk of the estimator $\widehat{\theta}_m$:

$$\mathbb{E}\left[l(\widehat{\theta}_m, \theta)\right] = l(\theta_m, \theta) + \left[\sigma^2 + l(\theta_m, \theta)\right] \frac{d_m}{n - d_m - 1} .$$

Hence, $\widehat{\theta}_m$ converges to θ_m in probability when n converges to infinity. Contrary to the fixed design regression framework, the variance term $\left[\sigma^2 + l(\theta_m, \theta)\right] \frac{d_m}{n-d_m-1}$ depends on the bias term $l(\theta_m, \theta)$. Besides, this variance term does not necessarily increase when the dimension of the model increases.

Let us now explain the idea underlying our model selection procedure. We aim at choosing a model $\widehat{m}$ that nearly minimizes the mean-squared error $\gamma(\widehat{\theta}_m)$. Since we do not have access to $\gamma(\widehat{\theta}_m)$ nor to the bias $l(\theta_m, \theta)$, we perform an unbiased estimation of the risk as done by Mallows [Mal73] in the fixed design framework.

$$\begin{aligned}
\gamma\left(\widehat{\theta}_m\right) &\approx \gamma_n\left(\widehat{\theta}_m\right) + \mathbb{E}\left[\gamma_n\left(\widehat{\theta}_m\right) - \gamma\left(\widehat{\theta}_m\right)\right] \\
&\approx \gamma_n\left(\widehat{\theta}_m\right) + \mathbb{E}\left[\gamma_n\left(\widehat{\theta}_m\right)\right] \frac{d_m}{n - d_m}\left[2 + \frac{d_m + 1}{n - d_m - 1}\right] \\
&\approx \gamma_n\left(\widehat{\theta}_m\right)\left[1 + \frac{d_m}{n - d_m}\left(2 + \frac{d_m + 1}{n - d_m - 1}\right)\right] .
\end{aligned} \tag{4.7}$$

By Lemma 4.1, these approximations are in fact equalities in expectation. Since the last expression only depends on the data, we may compute its minimizer over the collection $\mathcal{M}$. This approximation is effective and minimizing (4.7) provides a good estimator $\widetilde{\theta}$ when the size of the collection $\mathcal{M}$ is moderate as stated in Theorem 4.2. We recall that $\|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2$ equals $\gamma_n(\widehat{\theta}_m)$. Hence, our previous heuristics would lead to a choice of penalty $\text{pen}(m) = \frac{d_m}{n-d_m}\left(2 + \frac{d_m+1}{n-d_m-1}\right)$ in our criterion (4.3), whereas FPE criterion corresponds to $\text{pen}(m) = \frac{2d_m}{n-d_m}$. These two penalties are equivalent when the dimension d_m is small in front of n. In Theorem 4.2, we explain why these criteria allow to derive approximate oracle inequalities when there is a small number of models. However, when the size of the collections $\mathcal{M}$ increases, we need to design other penalties that take into account the complexity of the collection $\mathcal{M}$ (see Section 4.3.2).

4.3 Oracle inequalities

4.3.1 A small number of models

In this section, we restrict ourselves to the situation where the collection of models $\mathcal{M}$ only contains a small number of models as defined in [BM07] Sect 3.1.2.

($\mathbb{H}_{Pol}$): for each $d \geq 1$ the number of models $m \in \mathcal{M}$ such that $d_m = d$ grows at most polynomially with respect to d. In other words, there exists α and β such that for any $d \geq 1$, $\text{Card}\left(\{m \in \mathcal{M},\ d_m = d\}\right) \leq$

αd^β.

($\mathbb{H}_\eta$): The dimension d_m of every model m in $\mathcal{M}$ is smaller than ηn. Moreover, the number of observations n is larger than $6/(1-\eta)$.

Assumption ($\mathbb{H}_{Pol}$) states that there is at most a polynomial number of models with a given dimension. It includes in particular the problem of ordered variable selection, on which we will focus in this section. Let us introduce the collection of models relevant for this issue. For any positive number i smaller or equal to p, we define the model $m_i := \{1, \dots, i\}$ and the nested collection $\mathcal{M}_i := \{m_0, m_1, \dots m_i\}$. Here, m_0 refers to the empty model. Any collection $\mathcal{M}_i$ satisfies ($\mathbb{H}_{Pol}$) with $\beta = 0$ and $\alpha = 1$.

Theorem 4.2. *Let η be any positive number smaller than one. Assume that the collection $\mathcal{M}$ satisfies ($\mathbb{H}_{Pol}$) and ($\mathbb{H}_\eta$). If the penalty pen(.) is lower bounded as follows*

$$pen(m) \geq K \frac{d_m}{n - d_m} \text{ for all } m \in \mathcal{M} \text{ and some } K > 1 \ , \tag{4.8}$$

then

$$\mathbb{E}\left[l(\widetilde{\theta}, \theta)\right] \leq L(K, \eta) \inf_{m \in \mathcal{M}} \left[l(\theta_m, \theta) + \frac{n - d_m}{n} pen(m) \left[\sigma^2 + l(\theta_m, \theta)\right] \right] + \tau_n \ , \tag{4.9}$$

where the error term τ_n is defined as

$$\tau_n = \tau_n \left[var(Y), K, \eta, \alpha, \beta\right] := L_1(K, \eta, \alpha, \beta) \left[\frac{\sigma^2}{n} + n^{3+\beta} var(Y) \exp\left[-n L_2(K, \eta)\right] \right] \ ,$$

and $L_2(K, \eta)$ is positive.

The theorem applies for any n, any p and there is no hidden dependency on n or p in the constants. Besides, observe that the theorem does not depend at all on the covariance matrix Σ between the covariates. If we choose the penalty $\text{pen}(m) = K \frac{d_m}{n - d_m}$, we obtain an approximate oracle inequality.

$$\mathbb{E}\left[l(\widetilde{\theta}, \theta)\right] \leq L(K, \eta) \inf_{m \in \mathcal{M}} \mathbb{E}\left[l(\widehat{\theta}_m, \theta)\right] + \tau_n \left[\text{var}(Y), K, \eta, \alpha, \beta\right] \ ,$$

thanks to Lemma 4.1. The term in $n^{3+\beta}\text{var}(Y)\exp[-nL_2(K,\eta)]$ converges exponentially fast to 0 when n goes to infinity and is therefore considered as negligible. One interesting feature of this oracle inequality is that it allows to consider models of dimensions as close to n as we want providing that n is large enough. This will not be possible in the next section when handling more complex collections of models.

If we have stated that $\widetilde{\theta}$ performs almost as well as the oracle model, one may wonder whether it is possible to perform exactly as well as the oracle. In the next proposition, we shall prove that under additional assumption the estimator $\widetilde{\theta}$ with $K = 2$ follows an asymptotic exact oracle inequality. We state the result for the problem of ordered variable selection. Let us assume for a moment that the set of covariates is infinite, i.e. $p = +\infty$.

Definition 4.3. Let s and R be two positive numbers. We define the so-called ellipsoid $\mathcal{E}'_s(R)$ as

$$\mathcal{E}'_s(R) := \left\{ (\theta_i)_{i \geq 0}, \quad \sum_{i=1}^{+\infty} \frac{l\left(\theta_{m_{i-1}}, \theta_{m_i}\right)}{i^{-s}} \leq R^2 \sigma^2 \right\} \ .$$

In Section 4.4.1, we explain why we call this set $\mathcal{E}'_s(R)$ an ellipsoid.

Proposition 4.4. *Assume there exists s, s', and R such that $\theta \in \mathcal{E}'_s(R)$ and such that for any positive numbers R', $\theta \notin \mathcal{E}'_{s'}(R')$. We consider the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ and the penalty $pen(m) = 2\frac{d_m}{n - d_m}$. Then, the estimator $\widetilde{\theta}$ satisfies an asymptotic exact oracle inequality, for n going to infinity*

$$\frac{l(\widetilde{\theta}, \theta)}{\inf_{m \in \mathcal{M}_{\lfloor n/2 \rfloor}} l(\widehat{\theta}_m, \theta)} \to 1 \qquad p.s. \ .$$

Admittedly, we make n go to the infinity in this proposition but we are still in a high dimensional setting since $p = +\infty$ and since the size of the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ goes to infinity with n. In fact, the assumptions of Proposition 4.4 may be weakened. Looking closely at the proof, one observes that the result holds if θ does not belong to any model of the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ and if the dimension of the oracle model m^* is small before n. Notice that it is classical to assume that the bias is non-zero for every model m for proving the asymptotic optimality of Mallows' C_p (*cf.* Shibata [Shi81] and Birgé and Massart [BM07]). Moreover, the choice of the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ is arbitrary and one can extend it to many collections that satisfy $(\mathbb{H}_{Pol})$ and $(\mathbb{H}_\eta)$. As mentioned in Section 4.2, the penalty $\mathrm{pen}(m) = 2\frac{d_m}{n-d_m}$ corresponds to the FPE model selection procedure. In conclusion, the choice of the FPE criterion turns out to be asymptotically optimal when the complexity of $\mathcal{M}$ is small.

We now underline that the condition $K > 1$ in Theorem 4.2 is almost necessary. Indeed, choosing K smaller than one yields terrible statistical performances.

Proposition 4.5. *Suppose that p is larger than $n/2$. Let us consider the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ and assume that for some $\nu > 0$,*

$$pen(m) = (1-\nu)\frac{d_m}{n-d_m} \ , \tag{4.10}$$

for any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$. Then given $\delta \in (0,1)$, there exists some $n_0(\nu,\delta)$ only depending on ν and δ such that for $n \geq n_0(\nu,\delta)$,

$$\mathbb{P}_\theta\left[d_{\widehat{m}} \geq \frac{n}{4}\right] \geq 1-\delta \quad \textit{and} \quad \mathbb{E}\left[l(\widetilde{\theta},\theta)\right] \geq l(\theta_{m_{\lfloor n/2 \rfloor}},\theta) + L(\delta,\nu)\sigma^2 \ .$$

If one chooses a too small penalty, then the dimension $d_{\widehat{m}}$ of the selected model is huge and the penalized estimator $\widetilde{\theta}$ performs poorly. The hypothesis $p \geq n/2$ is needed for defining the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$. Once again, the choice of the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ is rather arbitrary and the result of Proposition 4.5 still holds for collections $\mathcal{M}$ which satisfy $(\mathbb{H}_{Pol})$ and $(\mathbb{H}_\eta)$ and contain at least one model of large dimension. Theorem 4.2 and Proposition 4.5 tell us that $\frac{d_m}{n-d_m}$ is the minimal penalty.

In practice, we advise to choose K between 2 and 3. Admittedly, $K = 2$ is asymptotically optimal by Proposition 4.4. Nevertheless, we have observed on simulations that $K = 3$ gives slightly better results when n is small. For ordered variable selection, we suggest to take the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$.

4.3.2 A general model selection theorem

In this section, we study the performance of the penalized estimator $\widetilde{\theta}$ for general collections $\mathcal{M}$. Classically, we need to penalize stronger the models m, incorporating the complexity of the collection. As a special case, we shall consider the problem of complete variable selection. This is why we define the collections $\mathcal{M}_p^d$ that consist of all subsets of $\{1,\dots,p\}$ of size less or equal to d.

Definition 4.6. Given a collection $\mathcal{M}$, we define the function $H(.)$ by

$$H(d) := \frac{1}{d}\log\left[\mathrm{Card}\left(\{m \in \mathcal{M},\, d_m = d\}\right)\right] \ ,$$

for any integer $d \geq 1$.

This function measures the complexity of the collection $\mathcal{M}$. For the collection $\mathcal{M}_p^d$, $H(k)$ is upper bounded by $\log(ep/k)$ for any $k \leq d$ (see Eq.(4.10) in [Mas07]). Contrary to the situation encountered in ordered variable selection, we are not able to consider models of arbitrary dimensions and we shall do the following assumption.

$(\mathbb{H}_{K,\eta})$: Given $K > 1$ and $\eta > 0$, the collection $\mathcal{M}$ and the number η satisfy

$$\forall m \in \mathcal{M}, \quad \frac{\left[1+\sqrt{2H(d_m)}\right]^2 d_m}{n-d_m} \leq \eta < \eta(K) \ , \tag{4.11}$$

where $\eta(K)$ is defined as $\eta(K) := [1 - 2(3/(K+2))^{1/6}]^2 \bigvee [1 - (3/K+2)^{1/6}]^2/4$.

The function $\eta(K)$ is positive and increases when K is larger than one. Besides, $\eta(K)$ converges to one when K converges to infinity. We do not claim that the expression of $\eta(K)$ is optimal. We are more interested in its behavior when K is large.

Theorem 4.7. *Let $K > 1$ and let $\eta < \eta(K)$. Assume that n is larger than some quantity $n_0(K)$ only depending on K and the collection $\mathcal{M}$ satisfies $(\mathbb{H}_{K,\eta})$. If the penalty pen(.) is lower bounded as follows*

$$pen(m) \geq K \frac{d_m}{n - d_m} \left(1 + \sqrt{2H(d_m)}\right)^2 \quad \text{for any } m \in \mathcal{M} \ , \tag{4.12}$$

then

$$\mathbb{E}\left[l(\widetilde{\theta}, \theta)\right] \leq L(K,\eta) \inf_{m \in \mathcal{M}} \left\{ l(\theta_m, \theta) + \frac{n - d_m}{n} pen(m) \left[\sigma^2 + l(\theta_m, \theta)\right] \right\} + \tau_n \ , \tag{4.13}$$

where τ_n is defined as

$$\tau_n = \tau_n \left[var(Y), K, \eta\right] := \sigma^2 \frac{L_1(K,\eta)}{n} + L_2(K,\eta) n^{5/2} var(Y) \exp\left[-n L_3(K,\eta)\right] \ ,$$

and $L_3(K,\eta)$ is positive.

This theorem provides an oracle type inequality of the same type as the one obtained in the Gaussian sequential framework by Birgé and Massart [BM01]. The risk of the penalized estimator $\widetilde{\theta}$ almost achieves the infimum of the risks plus a penalty term depending on the function $H(.)$. As in Theorem 4.2, the error term $\tau_n \left[\mathrm{var}(Y), K, \eta\right]$ depends on θ but this part goes exponentially fast to 0 with n.

Comments:

- As for Theorem 4.2, the result holds for arbitrary large p as long as n is larger than the quantity $n_0(K)$ (independent of p). There is no hidden dependency on p except in the complexity function $H(.)$ and Assumption $\mathbb{H}_{K,\eta}$ that we shall discuss for the particular case of complete variable selection. Moreover, one may easily check Assumption $\mathbb{H}_{K,\eta}$ since it only depends on the collection $\mathcal{M}$ and not on some unknown quantity.

- This result (as well as of Theorem 4.2) does not depend at all on the covariance matrix Σ between the covariates.

- The penalty introduced in this theorem only depends on the collection $\mathcal{M}$ and a number $K > 1$. Hence, performing the procedure does not require any knowledge on σ^2, Σ, or θ. We give hints at the end of the section for choosing the constant K.

Let us now restate Theorem 4.7 for the particular issue of complete variable selection. Consider $K > 1$, $\eta < \eta(K)$ and $d > 1$ such that $\mathcal{M}_p^d$ satisfies Assumption $(\mathbb{H}_{K,\eta})$. If we take for any model $m \in \mathcal{M}_p^d$ the penalty term

$$\mathrm{pen}(m) = K \frac{d_m}{n - d_m} \left[1 + \sqrt{2 \log\left(\frac{ep}{d_m}\right)}\right]^2 \ , \tag{4.14}$$

then we get

$$\mathbb{E}\left[l(\widetilde{\theta}, \theta)\right] \leq L(K,\eta) \inf_{m \in \mathcal{M}_p^d} \left\{ l(\theta_m, \theta) + \frac{d_m}{n} \log\left(\frac{ep}{d_m}\right) \sigma^2 \right\} + \tau_n \left[\mathrm{var}(Y), K, \eta\right] \ .$$

We shall prove in Section 4.4.2, that the term $\log(p/d_m)$ is unavoidable and that the obtained estimator is optimal from a minimax point of view. If the true parameter θ belongs to some unknown model m, then the rates of estimation of $\tilde{\theta}$ is of the order $\frac{d_m}{n} \log(p/d_m)\sigma^2$. Let us compare our result with other procedures.

- The oracle type inequalities look similar to the ones obtained by Birgé and Massart [BM01], Bunea *et al.* [BTW07b] and Baraud *et al.* [BGH08]. However, Birgé and Massart and Bunea *et al.* assume that the variance σ^2 is known. Moreover, Birgé and Massart and Baraud *et al.* only consider a fixed design setting. Yet, Bunea *et al.* allow the design to be random, but they assume that the regression functions are bounded (Assumption A.2 in their paper) which is not the case here. Moreover, they only get risk bounds with respect to the empirical norm $\|.\|_n$ and not the integrated loss $l(.,.)$.

- As mentioned previously, our oracle inequality holds for any covariance matrix Σ. In contrast, Lasso and Dantzig selector estimators have been shown to satisfy oracle inequalities under assumptions on the empirical design $\mathbf{X}$. In [CT07], Candès and Tao indeed assume that the singular values of $\mathbf{X}$ restricted to any subset of size proportional to the sparsity of θ are bounded away from zero. Bickel *et al.* [BRT08] introduce an extension of this condition prove both for the Lasso and the Dantzig selector. In a recent work [CP08], Candès and Plan state that if the empirical correlation between the covariates is smaller than $L(\log p)^{-1}$,then the Lasso follows an oracle inequality in a majority of cases. Their condition is in fact almost necessary. On the one hand, they give examples of some low correlated situations, where the Lasso performs poorly. On the other hand, they prove that the Lasso fails to work well if the correlation between the covariates if larger than $L(\log p)^{-1}$. Yet, Candès and Plan consider the loss function $\|\mathbf{X}\widehat{\theta} - \mathbf{X}\theta\|_n^2$, whereas we use the *integrated* loss $l(\widehat{\theta}, \theta)$, but this does not really change the impact of their result. We refer to their paper for further details. The main point is that for some correlation structures, our procedure still works well, whereas the Lasso and the Dantzig selector procedures perfom poorly. In many problems such as GGM estimation, the correlation between the covariates may be high and even the relaxed assumptions of Candès and Plan may not be fulfilled. In Section 4.5, we illustrate this phenomenon by comparing our procedure with the Lasso on numerical examples for independent and highly correlated covariates.

- Suppose that the covariates are independent and that θ belongs to some model m, the rates of convergence of the Lasso is then of the order $\frac{d_m}{n}\log(p)\sigma^2$, whereas ours is $\frac{d_m}{n}\log(p/d_m)\sigma^2$. Consider the case where p, and d_m are of the same order whereas n is large. Our model selection procedure therefore outperforms the Lasso by a $\log(p)$ factor even if the covariates are independent.

- Let us restate Assumption $(\mathbb{H}_{K,\eta})$ for the particular collection $\mathcal{M}_p^d$. Given some $K > 1$ and some $\eta < \eta(K)$, the collection $\mathcal{M}_p^d$ satisfies $(\mathbb{H}_{K,\eta})$ if

$$d \leq \eta \frac{n}{1 + \left[1 + \sqrt{2(1 + \log(p/d))}\right]^2} \,. \tag{4.15}$$

  If p is much larger than n, the dimension d of the largest model has to be be smaller than the order $\eta\frac{n}{2\log(p)}$. Candès and Plan state a similar condition for the lasso. We believe that this condition is unimprovable. Indeed, Wainwright states in Th.2 of [Wai07] a result going in this sense: it is impossible to estimate reliably the support of a k-sparse vector θ if n is smaller than the order $k\log(p/k)$. If $\log(p)$ is larger than n, then we cannot apply Theorem 4.7. This ultra-high dimensional setting is also not handled by the theory for the Lasso and the Dantzig selector. Finally, if p is of the same order as n, then Condition (4.15) is satisfied for dimensions d of the same order as n. Hence, our method works well even when the sparsity is of the same order as n, which is not the case for the Lasso or the Dantzig selector.

Let us discuss the practical choice of d and K for complete variable selection. From numerical studies, we advise to take $d \leq \frac{n}{2.5\left[2+\log\left(\frac{p}{n}\vee 1\right)\right]} \wedge p$ even if this quantity is slightly larger than what is ensured by the theory. The practical choice of K depends on the aim of the study. If one aims at minimizing the risk, $K = 1.1$ gives rather good result. A larger K like 1.5 or 2 allows to obtain a more conservative procedure and consequently a lower FDR. We compare these values of K on simulated examples in Section 4.5.

4.4 Minimax lower bounds and Adaptivity

Throughout this section, we emphasize the dependency of the expectations $\mathbb{E}(.)$ and the probabilities $\mathbb{P}(.)$ on θ by writing $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$. We have stated in Section 4.3 that the penalized estimator $\widetilde{\theta}$ performs almost as well as the best of the estimators $\widehat{\theta}_m$. We now want to compare the risk of $\widetilde{\theta}$ with the risk of any other possible estimator estimator $\widehat{\theta}$. There is no hope to make a pointwise comparison with an arbitrary estimator. Therefore, we classically consider the maximal risk over some suitable subsets Θ of $\mathbb{R}^p$. The *minimax risk* over the set Θ is given by $\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta[l(\widehat{\theta},\theta)]$, where the infimum is taken over all possible estimators $\widehat{\theta}$ of θ. Then, the estimator $\tilde{\theta}$ is said to be *approximately minimax* with respect to the set Θ if the ratio

$$\frac{\sup_{\theta\in\Theta} \mathbb{E}_\theta\left[l\left(\tilde{\theta},\theta\right)\right]}{\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right]}$$

is smaller than a constant that does not depend on σ^2, n, or p. The minimax rates of estimation were extensively studied in the fixed design Gaussian regression framework and we refer for instance to [BM01] for a detailed discussion. In this section, we apply a classical methodology known as Fano's Lemma in order to derive minimax rates of estimation for ordered and complete variable selection. Then, we deduce adaptive properties of the penalized estimator $\widetilde{\theta}$.

4.4.1 Adaptivity with respect to ellipsoids

In this section, we prove that the estimator $\widetilde{\theta}$ introduced in Section 4.3.1 to perform ordered variable selection is adaptive to a large class of ellipsoids.

Definition 4.8. For any non increasing sequence $(a_i)_{1\leq i\leq p+1}$ such that $a_1 = 1$ and $a_{p+1} = 0$ and any $R > 0$, we define the ellipsoid $\mathcal{E}_a(R)$ by

$$\mathcal{E}_a(R) := \left\{\theta\in\mathbb{R}^p, \sum_{i=1}^{p} \frac{l\left(\theta_{m_{i-1}},\theta_{m_i}\right)}{a_i^2} \leq R^2\right\} .$$

This definition is very similar to the notion of ellipsoids introduced in Chapter 2. Let us explain why we call this set an ellipsoid. Assume for one moment that the $(X_i)_{1\leq i\leq p}$ are independent identically distributed with variance one. In this case, the term $l\left(\theta_{m_{i-1}},\theta_{m_i}\right)$ equals θ_i^2 and the definition of $\mathcal{E}_a(R)$ translates in

$$\mathcal{E}_a(R) = \left\{\theta\in\mathbb{R}^p, \sum_{i=1}^{p} \frac{\theta_i^2}{a_i^2} \leq R^2\right\} ,$$

which precisely corresponds to a *classical* definition of an ellipsoid. If the $(X_i)_{1\leq i\leq p}$ are not i.i.d. with unit variance, it is always possible to create a sequence X_i' of i.i.d. standard Gaussian variables by orthonormalizing the X_i using Gram-Schmidt process. If we call θ' the vector in $\mathbb{R}^p$ such that $X\theta = X'\theta'$, then it holds that $l\left(\theta_{m_{i-1}},\theta_{m_i}\right) = \theta_i'^2$. Then, we can express $\mathcal{E}_a(R)$ using the coordinates of θ' as previously:

$$\mathcal{E}_a(R) = \left\{\theta\in\mathbb{R}^p, \sum_{i=1}^{p} \frac{\theta_i'^2}{a_i^2} \leq R^2\right\} .$$

The main advantage of this definition is that it does not directly depend on the covariance of $(X_i)_{1\leq i\leq p}$.

Proposition 4.9. *For any sequence $(a_i)_{1\leq i\leq p}$ and any positive number R, the minimax rate of estimation over the ellipsoid $\mathcal{E}_a(R)$ is lower bounded by*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathcal{E}_a(R)} \mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right] \geq L \sup_{1\leq i\leq p}\left[a_i^2R^2 \wedge \frac{\sigma^2 i}{n}\right] . \tag{4.16}$$

This result is analogous to the lower bounds obtained in the fixed design regression framework (see e.g. [Mas07] Th. 4.9). Hence, the estimator $\widetilde{\theta}$ built in Section 4.3.1 is adaptive to a large class of ellipsoids.

Corrolary 4.10. *Assume that n is larger than 12. We consider the penalized estimator $\widetilde{\theta}$ with the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ and the penalty $pen(m) = K\frac{d_m}{n-d_m}$. Let $\mathcal{E}_a(R)$ be an ellipsoid whose radius R satisfies $\frac{\sigma^2}{n} \leq R^2 \leq \sigma^2 n^\beta$ for some $\beta > 0$. Then, $\widetilde{\theta}$ is approximately minimax on $\mathcal{E}_a(R)$*

$$\sup_{\theta \in \mathcal{E}_a(R)} l(\widetilde{\theta}, \theta) \leq L(K, \beta) \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}_a(R)} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] ,$$

if either $n \geq 2p$ or $a^2_{\lfloor n/2 \rfloor + 1} R^2 \leq \sigma^2/2$.

In the fixed design framework, one may build adaptive estimators to any ellipsoid satisfying $R^2 \geq \sigma^2/n$ so that the ellipsoid is not degenerate (see e.g. [Mas07] Sect. 4.3.3). In our setting, when p is small the estimator $\widetilde{\theta}$ is adaptive to all the ellipsoids that have a moderate radius $\sigma^2/n \leq R^2 \leq n^\beta$. The technical condition $R^2 \leq n^\beta$ is not really restrictive. It comes from the term $n^3 l(0_p, \theta) \exp(-nL(K))$ in Theorem 4.2 which goes exponentially fast to 0 with n. When p is larger, $\widetilde{\theta}$ is adaptive to the ellipsoids that also satisfies $a^2_{\lfloor n/2 \rfloor + 1} R^2 \leq \sigma^2/2$. In other words, we require that the ellipsoid is well approximated by the space $S_{m \lfloor n/2 \rfloor}$ of vectors θ whose support is included in $\{1, \ldots, \lfloor n/2 \rfloor\}$. If this condition is not fulfilled, the estimator $\widetilde{\theta}$ is not proved to be minimax on $\mathcal{E}_a(R)$. For such situations, we believe on the one hand that the estimator $\widetilde{\theta}$ should be refined and on the other hand that our lower bounds are not sharp. Finally, the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$ may be replaced by any $\mathcal{M}_{\lfloor n\eta \rfloor}$ in Corollary 4.10.

Since the methods used for minimax lower bounds and the oracle inequalities are analogous to the ones in the Gaussian sequence framework, one may also adapt in our setting the arguments developed in [Mas07] Sect. 4.3.5 to derive minimax rates of estimation over other sets such Besov bodies. However, this is not really relevant for the regression model (4.1).

4.4.2 Adaptivity with respect to sparsity

Our aim is now to analyze the minimax risk for the complete variable selection problem. Let us fix an integer k between 1 and p. We are interested in estimating the vector θ within the class of vectors with a most k non-zero components. This typically corresponds to the situation encountered in graphical modeling when estimating the neighborhoods of large sparse graphs. As the graph is assumed to be sparse, only a small number of components of θ are non-zero.

In the sequel, the set $\Theta[k, p]$ stands for the subset of vectors $\theta \in \mathbb{R}^p$, such that at most k coordinates of θ are non-zero. For any $r > 0$, we denote $\Theta[k, p](r)$ the subset of $\Theta[k, p]$ such that any component of θ is smaller than r in absolute value.

First, we derive a lower bound for the minimax rates of estimation when the covariates are independent. Then, we prove the estimator $\widetilde{\theta}$ defined with some collection $\mathcal{M}_p^d$ and the penalty (4.14) is adaptive to any sparse vector θ. Finally, we investigate the minimax rates of estimation for correlated covariates.

Proposition 4.11. *Assume that the covariates X_i are independent and have a unit variance. For any $k \leq p$ and any radius $r > 0$,*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] \geq Lk \left[r^2 \wedge \sigma^2 \frac{1 + \log\left(\frac{p}{k}\right)}{n} \right] . \tag{4.17}$$

Thanks to Theorem 4.7, we derive the minimax rate of estimation over $\Theta[k, p]$.

Corrolary 4.12. *Consider $K > 0$, $\beta > 0$, and $\eta < \eta(K)$. Assume that $n \geq n_0(K)$ and that the covariates X_i are independent and have a unit variance. Let d be a positive integer such that $\mathcal{M}_p^d$ satisfies $(\mathbb{H}_{K,\eta})$. The penalized estimator $\widetilde{\theta}$ defined with the collection $\mathcal{M}_p^d$ and the penalty (4.14) is adaptive minimax over the sets $\Theta[k, p](n^\beta)$*

$$\sup_{\theta \in \Theta[k,p]} \mathbb{E}_\theta \left[l(\widetilde{\theta}, \theta) \right] \leq L(K, \beta, \eta) \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](n^\beta)} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] ,$$

for any k smaller than d.

Hence, the minimax rates of estimation over $\Theta[k,p](n^\beta)$ is of order $k\frac{\log\left(\frac{ep}{k}\right)}{n}$, which is similar to the rates obtained in the fixed design regression framework. As in previous Section, we restrict ourselves to a radius r in $\Theta[k,p](r)$ smaller than n^β because of the term $\tau_n(\mathrm{var}(Y),K,\eta)$ which depends on $l(0_p,\theta)$ but goes exponentially fast to 0 when n goes to infinity. Let us interpret Corollary 4.12 with regard to Condition (4.15). If p is of the same order as n, the estimator $\widetilde{\theta}$ is simultaneously minimax over all sets $\Theta[k,p](n^\beta)$ when k is smaller than a constant times n. If p is much larger than n, the estimator $\widetilde{\theta}$ is simultaneously minimax over all sets $\Theta[k,p](n^\beta)$ with k smaller than $Ln/\log(p)$. We conjecture that the minimax rate of estimation is larger than $k\log(p/k)/n$ when k becomes larger than $n/\log p$. Let us mention that Tsybakov [Tsy03] has proved general minimax lower bounds for aggregation in Gaussian random design regression. However, his result does not apply in our Gaussian design setting setting since he assumes that the density of the covariates X_i is lower bounded by a constant μ_0.

We have proved that the estimator $\widetilde{\theta}$ is adaptive to an unknown sparsity when the covariates are independent. The performance of $\widetilde{\theta}$ exhibited in Theorem 4.7 do not depend on the covariance matrix Σ. Hence, the minimax rates of estimation on $\Theta[k,p]$ is smaller or equal to the order $k\log(p/k)/n$ for any dependence between the covariance. One may then wonder whether the minimax rate of estimation over $\Theta[k,p]$ is not faster when the covariates are correlated. We are unable to derive the minimax rates for a general covariance matrix Σ. This is why we restrict ourselves to particular examples of correlation structures. Let us first consider a pathological situation: Assume that $X_1,\dots,X_k$ are independent and that $X_{k+1},\dots,X_p$ are all equal to X_1. Admittedly, the covariance matrix Σ is henceforth non invertible. In the discussion, we mention that Theorems 4.2 and 4.7 easily extend when Σ is non-invertible if we take into account that the estimators $\widehat{\theta}_m$ and $\widehat{m}$ are non-necessarily uniquely defined. We may derive from Lemma 4.1 that the estimator $\widehat{\theta}_{\{1,\dots,k\}}$ achieves the rate k/n over $\theta[k,p](n^\beta)$. Conversely, the parametric rate k/n is optimal. However, the estimator $\widetilde{\theta}$ defined with the collection $\mathcal{M}_p^k$ and penalty (4.14) only achieves the rate $k\log(p/k)/n$. Hence, $\widetilde{\theta}$ is not minimax over $\Theta[k,p]$ for this particular covariance matrix and the minimax rate is degenerate. This emergence of faster rates for correlation covariates also occurs for testing problems in the model (4.1) as stated in Section 2.4.3. This is why we provide sufficient conditions on Σ so that the minimax rate of estimation is still of the same order as in the independent case. In the following proposition, $\|.\|$ refers to the canonical norm in $\mathbb{R}^p$.

Proposition 4.13. *Let Ψ denote the correlation matrix of the covariates $(X_i)_{1\le i\le p}$. Let k be a positive number smaller $p/2$ and let $\delta>0$. Assume that*

$$(1-\delta)^2\|\theta\|^2\ ,\le \theta^*\Psi\theta\le(1+\delta)^2\|\theta\|^2\ , \tag{4.18}$$

for all $\theta\in\mathbb{R}^p$ with at most $2k$ non-zero components. Then, the minimax rate of estimation over $\Theta[k,p](r)$ is lower bounded as follows

$$\inf_{\widehat{\theta}}\sup_{\theta\in\Theta[k,p](r)}\mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right]\ge L(1-\delta)^2k\left[r^2\wedge\sigma^2\frac{1+\log\left(\frac{p}{k}\right)}{(1+\delta)^2n}\right]\ .$$

Assumption (4.18) corresponds to the δ-Restricted Isometry Property of order $2k$ introduced by Candès and Tao [CT05]. Under such a condition, the minimax rates of estimation is the same as the one in the independent case up to a constant depending on δ and the estimator $\widetilde{\theta}$ defined in Corollary 4.12 is still approximately minimax over such sets $\Theta[k,p]$.

However, the δ-Restricted Isometry Property is quite restrictive and seems not to be necessary so that the minimax rate of estimation stays of the order $k\log(p/k)/n$. Besides, in many situations this condition is not fulfilled. Assume for instance that the random vector X is a Gaussian Graphical model with respect to a given sparse graph. We expect that the correlation between two covariates is large if they are neighbors in the graph and small if they are far-off (w.r.t. the graph distance). This is why we derive lower bounds on the rate of estimation for correlation matrices often used to model stationary processes.

Proposition 4.14. *Let $X_1,\dots,X_p$ form a stationary process on the one dimensional torus. More precisely, the correlation between X_i and X_j is a function of $|i-j|_p$ where $|.|_p$ refers to the toroidal distance defined by:*

$$|i-j|_p:=(|i-j|)\wedge(p-|i-j|)\ .$$

$\Psi_1(\omega)$ and $\Psi_2(t)$ respectively refer to the correlation matrix of X such that

$$\begin{aligned}\mathrm{corr}(X_i, X_j) &:= \exp(-\omega|i-j|_p) \text{ where } \omega > 0,\\ \mathrm{corr}(X_i, X_j) &:= (1+|i-j|_p)^{-t} \text{ where } t > 0.\end{aligned}$$

Then, the minimax rates of estimation are lower bounded as follows

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta[k,p]} \mathbb{E}_{\theta,\Psi_1(\omega)}\left[l(\widehat{\theta},\theta)\right] \geq L\frac{k\sigma^2}{n}\left[1+\log\left(\frac{\lfloor p\lceil \log(4k)/\omega\rceil^{-1}\rfloor}{k}\right)\right] ,$$

if k is smaller than $p/\lceil\log(4k)/\omega\rceil$ and

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta[k,p]} \mathbb{E}_{\theta,\Psi_2(t)}\left[l(\widehat{\theta},\theta)\right] \geq L\frac{k\sigma^2}{n}\left[1+\log\left(\frac{\lfloor p\lceil (4k)^{\frac{1}{t}}-1\rceil^{-1}\rfloor}{k}\right)\right] ;$$

if k is smaller than $p/\lceil(4k)^{\frac{1}{t}}-1\rceil$.

If the range ω is larger than $1/p^\gamma$ or if the range t is larger than γ for some $\gamma < 1$, the lower bounds are of order $\sigma^2\frac{k}{n}(1+\log p/k)$. As a consequence, for any of these correlation models the minimax rate of estimation is of the same order as the minimax rate of estimation for independent covariates. This means that the estimator $\widetilde{\theta}$ defined in Proposition 4.12 is rate-optimal for these correlations matrices.

In conclusion, the estimator $\widetilde{\theta}$ defined in Corollary 4.12 may not be adaptive to the covariance matrix Σ but rather achieves the minimax rate over all covariance matrices Σ:

$$\sup_{\Sigma\geq 0}\sup_{\theta\in\Theta[k,p](n^\beta)} \mathbb{E}_\theta\left[l(\widetilde{\theta},\theta)\right] \leq L(K,\beta,\eta)\inf_{\widehat{\theta}}\sup_{\Sigma\geq 0}\sup_{\theta\in\Theta[k,p](n^\beta)} \mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right] .$$

Nevertheless, the result makes sense if one considers GGMs since the resulting covariance matrices are typically far from being independent.

4.5 Numerical study

In this section, we carry out a small simulation study to evaluate the performance of our estimator $\widetilde{\theta}$. As pointed out earlier, an interesting feature of our criterion lies in its flexibility. However, we restrict ourselves here to the variable selection problem. Indeed, it allows to assess the efficiency of our procedure with having regard to the Lasso [Tib96] and adaptive Lasso proposed by Zou [Zou06]. Even if these two procedures assume that the conditional variance σ^2 is known, they give good results in practice and the comparison with our method is of interest. The calculations are made with *R* `www.r-project.org/`.

4.5.1 Simulation scheme

We consider the regression model (4.1) with $p = 20$, and $\sigma^2 = 1$. The number of observations n equal 15, 20, and 30. We perform two simulation experiments.

1. First simulation experiment: The covariance matrix Σ_1 is the identity matrix. This corresponds to the situation where the covariates are all independent. The vector θ_1 has all its components to zero except the three first ones, which respectively equal 2, 1, and 0.5.

2. Second simulation experiment: Let A be the $p\times p$ matrix whose lines $(a_1,\ldots,a_p)$ are respectively defined by

$$\begin{aligned}a_1 &:= (1,-1,0,\ldots,0)/\sqrt{2}\\ a_2 &:= (-1,1.2,0,\ldots,0)/\sqrt{1+1.2^2}\\ a_3 &:= (1/\sqrt{2},1/\sqrt{2},1/p,\ldots,1/p)/\sqrt{1/2+(p-2)/p^2} ,\end{aligned}$$

   and for $4\leq j\leq p$, a_j corresponds to the j^{th} canonical vector of $\mathbb{R}^p$. Then, we take the covariance matrix $\Sigma_2 = A^*A$ and the vector $\theta_2^* = (40,40,0,\ldots,0)$. This choice of parameters derives from the simulation experiments of [BGH08]. Observe that the two first covariates are highly correlated.

For each sample we estimate θ with our procedure, the Lasso and the adaptive Lasso. For our procedure we use the collection $\mathcal{M}_p^3$ for $n = 15$, $\mathcal{M}_p^4$ for $n = 20$ and, $\mathcal{M}_p^5$ for $n = 30$. The choice of smaller collections for $n = 15$ and 20 is due to Condition (4.15). We take the penalty (4.14) with $K = 1.1$ 1.5, and 2. For the Lasso and adaptive Lasso procedures, we first normalize the covariates $(\mathbf{X}_i)$. Here, $2\sqrt{\log p}\sigma$ would be a good choice for the parameter λ of the Lasso. However, we do not have access to σ. Hence, we use an estimation of the variance $\widehat{\text{var}}(Y)$ which is a (possibly inaccurate) upper bound of σ^2. This is why we choose the parameter λ of the Lasso between $0.3 \times 2\sqrt{\log p \widehat{\text{var}}(Y)}$ and $2\sqrt{\log p \widehat{\text{var}}(Y)}$ by leave-one-out cross-validation. The number 0.3 is rather arbitrary. In practice, the performances of the Lasso do not really depend on this number as soon it is neither too small nor close to one. For the adaptive Lasso procedure, the parameters γ and λ are also estimated thanks to leave-one-out cross-validation: γ can take three values $(0.5, 1, 2)$ and the values of λ vary between $0.3 \times 2\sqrt{\log p \widehat{\text{var}}(Y)}$ and $2\sqrt{\log(p) \widehat{\text{var}}(Y)}$.

We evaluate the risk ratio

$$\text{ratio.Risk} = \frac{\mathbb{E}\left[l(\widehat{\theta}, \theta)\right]}{\inf_{m \in \mathcal{M}_p^5} \mathbb{E}\left[l(\widehat{\theta}_m, \theta)\right]}$$

as well as the power and the FDR on the basis of 1000 simulations.

4.5.2 Results

	$n = 15$			$n = 20$		
Estimator	ratio.Risk	Power	FDR	ratio.Risk	Power	FDR
$K = 1.1$	4.8 ± 0.4	0.67 ± 0.02	0.23 ± 0.02	4.8 ± 0.3	0.77 ± 0.01	0.28 ± 0.02
$K = 1.5$	5.7 ± 0.4	0.62 ± 0.02	0.20 ± 0.01	5.3 ± 0.4	0.74 ± 0.02	0.25 ± 0.01
$K = 2$	7.3 ± 0.5	0.54 ± 0.02	0.17 ± 0.01	6.6 ± 0.5	0.68 ± 0.02	0.21 ± 0.01
Lasso	5.8 ± 0.2	0.64 ± 0.01	0.29 ± 0.02	6.0 ± 0.2	0.74 ± 0.01	0.23 ± 0.01
A. Lasso	4.8 ± 0.3	0.64 ± 0.02	0.30 ± 0.02	4.7 ± 0.4	0.75 ± 0.02	0.30 ± 0.01

	$n = 30$		
Estimator	ratio.Risk	Power	FDR
$K = 1.1$	4.2 ± 0.3	0.87 ± 0.01	0.23 ± 0.02
$K = 1.5$	4.1 ± 0.2	0.84 ± 0.01	0.19 ± 0.01
$K = 2$	4.3 ± 0.2	0.81 ± 0.01	0.14 ± 0.01
Lasso	6.6 ± 0.2	0.83 ± 0.01	0.18 ± 0.01
A. Lasso	4.3 ± 0.5	0.86 ± 0.02	0.26 ± 0.01

Table 5: Our procedure with $K = 1.1$, 1.5, and 2 and Lasso and adaptive Lasso procedures: Estimation and 95% confidence interval of Risk ratio (ratio.Risk), Power and FDR when $p = 20$, $\Sigma = \Sigma_2$, $\theta = \theta_2$, and $n = 15$, 20, and 30.

The results of the first simulation experiment are given in Table 5. We observe that the five estimators perform more or less similarly as expected by the theory. The results of the second simulation study are reported in Table 6. Clearly, the Lasso and adaptive Lasso procedures are not consistent in this situation since the power is close to 0 and the FDR is close to one. Consequently, the risk ratio is quite large and the adaptive Lasso even seems unstable. In contrast, our method exhibits a large power and a reasonable FDR.

In the two studies, choosing a larger K reduces the power of the estimator but also decreases the FDR. It seems that the choice $K = 1.1$ yields a good risk ratio, whereas $K = 2$ gives a better control of the FDR. Contrary to the parameter λ for the lasso, we do not need an *ad-hoc* method such as cross-validation to calibrate K. The second example is certainly quite pathological but it illustrates that our estimator $\widetilde{\theta}$ performs well even when the Lasso does not provide an accurate estimation. The good behavior of our method illustrates the strength of Theorem 4.7 that does not depend on the correlation of the explanatory variables.

	$n = 15$			$n = 20$		
Estimator	ratio.Risk	Power	FDR	ratio.Risk	Power	FDR
$K = 1.1$	5.3 ± 0.4	0.77 ± 0.03	0.41 ± 0.02	6.4 ± 0.5	0.87 ± 0.02	0.39 ± 0.02
$K = 1.5$	5.3 ± 0.4	0.76 ± 0.03	0.41 ± 0.02	5.9 ± 0.5	0.87 ± 0.02	0.36 ± 0.02
$K = 2$	5.5 ± 0.5	0.75 ± 0.03	0.40 ± 0.02	5.5 ± 0.5	0.86 ± 0.02	0.33 ± 0.02
Lasso	13.5 ± 0.3	0.02 ± 0.01	0.99 ± 0.01	16.7 ± 0.3	0.02 ± 0.01	0.98 ± 0.01
A. Lasso	15.0 ± 1.2	0.02 ± 0.01	0.90 ± 0.02	20.5 ± 1.8	0.04 ± 0.01	0.89 ± 0.02

	$n = 30$		
Estimator	ratio.Risk	Power	FDR
$K = 1.1$	4.5 ± 0.3	0.96 ± 0.02	0.24 ± 0.02
$K = 1.5$	3.9 ± 0.3	0.95 ± 0.01	0.19 ± 0.02
$K = 2$	3.5 ± 0.3	0.94 ± 0.01	0.16 ± 0.02
Lasso	22.0 ± 0.3	0.02 ± 0.01	0.99 ± 0.01
A. Lasso	31.8 ± 3.0	0.04 ± 0.01	0.88 ± 0.02

Table 6: Our procedure with $K = 1.1$, 1.5, and 2 and Lasso and adaptive Lasso procedures: Estimation and 95% confidence interval of Risk ratio (ratio.Risk), Power and FDR when $p = 20$, $\Sigma = \Sigma_1$, $\theta = \theta_1$, and $n = 15$, 20, and 30.

4.6 Discussion and concluding remarks

Until now, we have assumed that the covariance matrix Σ of the covariates is non-singular. If Σ is singular, the estimators $\widehat{\theta}_m$ and the model $\widehat{m}$ are not necessarily uniquely defined. However, upon defining $\widehat{\theta}_m$ as *one* of the minimizers of $\gamma_n(\theta')$ over S_m, one may readily extend the oracle inequalities stated in Theorem 4.2 and 4.7.

Let us recall the main features of our method. We have defined a model selection criterion that satisfies oracle inequalities regardless of the correlation between the covariates and regardless of the collection of models. Hence, the estimator $\widetilde{\theta}$ achieves nice adaptive properties for ordered variable selection or for complete variable selection. Besides, one can easily combine this method with prior knowledge on the model by choosing a proper collection $\mathcal{M}$ or by modulating the penalty pen(.). Moreover, we may easily calibrate the penalty even when σ^2 is unknown, whereas the Lasso-type procedures require a cross-validation strategy to choose the parameter λ. The compensation for these nice properties is a computational cost that depends linearly on the size of $\mathcal{M}$. Hence, the complete variable selection problem is NP-hard. This makes it intractable when p becomes too large (i.e. more than 50). In contrast, our criterion applies for arbitrary p when considering ordered variable selection since the size of $\mathcal{M}$ is linear with n. In situations where one has a good prior knowledge on the true model, the collection $\mathcal{M}$ is then not too large and our criterion is also fastly calculable even for large p.

For complete variable selection, Lasso-type procedures are computationally feasible even when p is large and achieve oracle inequalities under assumptions on the covariance structure. However, there are both theoretical and practical problems with these estimators. On the one hand, they are known to perform poorly for some covariance structures. On the other hand, there is some room for improvement in the practical calibration of the lasso, especially when σ^2 is unknown. In a future work, we would like to combine the strength of our method with these computationally fast algorithms. The problem at hand is to design a fast data-driven method that picks a subcollection $\widehat{\mathcal{M}}$ of reasonable size. Afterwards, one applies our procedure to $\widehat{\mathcal{M}}$ instead of $\mathcal{M}$. A direction that needs further investigation is taking for $\widehat{M}$ all the subsets of the regularization path given by the lasso.

4.7 Proofs

4.7.1 Some notations and probabilistic tools

First, let us define the random variable ϵ_m by

$$Y = X\theta_m + \epsilon_m + \epsilon \ a.s. \ . \tag{4.19}$$

By definition of θ_m, ϵ_m follows a normal distribution and is independent of ϵ and of X_m. Hence, the variance of ϵ_m equals $l(\theta_m, \theta)$. The vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_m$ refer to the n samples of ϵ and ϵ_m. For any model m and any vector Z of size n, $\Pi_m^{\perp} Z$ stands for $Z - \Pi_m Z$. For any subset m of $\{1, \ldots, p\}$, Σ_m denotes the covariance matrix of the vector X_m^*. Moreover, we define the row vector $Z_m := X_m \sqrt{\Sigma_m^{-1}}$ in order to deal with standard Gaussian vectors. Similarly to the matrix $\mathbf{X}_m$, the $n \times d_m$ matrix $\mathbf{Z}_m$ stands for the n observations of Z_m. The notation $\langle ., . \rangle_n$ refers to the empirical inner product associated with the norm $\|.\|_n$. Lastly, $\varphi_{\max}(A)$ denotes the largest eigenvalue (in absolute value) of a symmetric square matrix A.

We shall extensively use the explicit expression of $\widehat{\theta}_m$:

$$\mathbf{X}\widehat{\theta}_m = \mathbf{X}_m (\mathbf{X}_m^* \mathbf{X}_m)^{-1} \mathbf{X}_m^* \mathbf{Y} \ . \tag{4.20}$$

Let us state a first lemma that gives the expressions of $\gamma_n(widehat\theta_m)$, $\gamma(\widehat{\theta}_m)$, and the loss $l(\widehat{\theta}_m, \theta_m)$.

Lemma 4.15. *For any model m of size smaller than n,*

$$\gamma_n\left(\widehat{\theta}_m\right) = \|\Pi_m^{\perp}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)\|_n^2 \ , \tag{4.21}$$

$$\gamma\left(\widehat{\theta}_m\right) = \sigma^2 + l(\theta_m, \theta) + l(\widehat{\theta}_m, \theta_m) \ , \tag{4.22}$$

$$l(\widehat{\theta}_m, \theta_m) = (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})^* \mathbf{Z}_{\widehat{m}} (\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-2} \mathbf{Z}_{\widehat{m}}^* (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}}) \ . \tag{4.23}$$

The proof is postponed to the Appendix.

We now introduce the main probabilistic tools used throughout the proofs. First, we need to bound the deviations of χ^2 random variables.

Lemma 4.16. *For any integer $d > 0$ and any positive number x,*

$$\mathbb{P}\left(\chi^2(d) \leq d - 2\sqrt{dx}\right) \leq \exp(-x) \ ,$$

$$\mathbb{P}\left(\chi^2(d) \geq d + 2\sqrt{dx} + 2x\right) \leq \exp(-x) \ .$$

These bounds are classical and are shown by applying Laplace method. We refer to Lemma 1 in [LM00] for more details. Moreover, we state a refined bound for the lower deviations of a χ^2 distribution.

Lemma 4.17. *For any integer $d > 0$ and any positive number x,*

$$\mathbb{P}\left[\chi^2(d) \leq d\left[\left(1 - \delta_d - \sqrt{\frac{2x}{d}}\right) \vee 0\right]^2\right] \leq \exp(-x) \ ,$$

$$\text{where } \delta_d := \sqrt{\frac{\pi}{2d}} + \exp(-d/16) \ . \tag{4.24}$$

The proof is postponed the Appendix. Finally, we shall bound the largest eigenvalue of standard Wishart matrices and standard inverse Wishart matrices. The following deviation inequality is taken from Theorem 2.13 in [DS01].

Lemma 4.18. *Let Z^*Z be a standard Wishart matrix of parameters (n, d) with $n > d$. For any positive number x,*

$$\mathbb{P}\left\{\varphi_{\max}\left[(Z^*Z)^{-1}\right] \geq \left[n\left(1 - \sqrt{\frac{d}{n}} - x\right)^2\right]^{-1}\right\} \leq \exp(-nx^2/2) \ ,$$

and

$$\mathbb{P}\left[\varphi_{\max}\left(Z^*Z\right) \leq n\left(1+\sqrt{\frac{d}{n}}+x\right)^2\right] \quad \leq \quad \exp(-nx^2/2)\ .$$

4.7.2 Proof of Theorem 4.2

Proof of Theorem 4.2. For the sake of simplicity we divide the main steps of the proof in several lemmas. First, let us fix a model m in the collection $\mathcal{M}$. By definition of $\widehat{m}$, we know that

$$\gamma_n(\widetilde{\theta})\left[1+\mathrm{pen}(\widehat{m})\right] \leq \gamma_n(\theta_m)\left[1+\mathrm{pen}(m)\right]\ .$$

Subtracting $\gamma(\theta)$ to both sides of this inequality yields

$$l(\widetilde{\theta},\theta) \leq l(\theta_m,\theta)+\gamma_n(\theta_m)\mathrm{pen}(m)+\overline{\gamma}_n(\theta_m)-\gamma_n(\widetilde{\theta})\mathrm{pen}(\widehat{m})-\overline{\gamma}_n(\widetilde{\theta})\ , \tag{4.25}$$

where $\overline{\gamma}_n(.) := \gamma_n(.)-\gamma(.)$. The proof is based on the concentration of the term $-\overline{\gamma}_n(\widetilde{\theta})$. More precisely, we shall prove that with overwhelming probability this quantity is of the same order as the penalty term $\gamma_n(\widetilde{\theta})\mathrm{pen}(\widehat{m})$.

Let κ_1 and κ_2 be two positive numbers smaller than one that we shall fix later. For any model $m' \in \mathcal{M}$, we introduce the random variables $A_{m'}$ an $B_{m'}$ as

$$\begin{aligned} A_{m'} \quad &:= \quad \kappa_1+1-\frac{\|\Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\|_n^2}{l(\theta_{m'},\theta)}+\kappa_2 n\varphi_{\max}\left[\left(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'}\right)^{-1}\right]\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\ &- \quad K\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi_{m'^{\perp}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2}\ , \qquad (4.26) \\ B_{m'} \quad &:= \quad \kappa_1^{-1}\frac{\langle\Pi^{\perp}_{m'}\boldsymbol{\epsilon},\Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)}+\frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{\sigma^2}+\kappa_2 n\varphi_{\max}\left[\left(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'}\right)^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\ &- \quad K\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi^{\perp}_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2}\ . \qquad (4.27) \end{aligned}$$

We recall that the notations ϵ_m, Z_m, $\langle .,. \rangle_n$, and $\varphi_{\max}(.)$ are defined in Section 4.7.1. We may upper bound the expression $-\overline{\gamma}_n(\widetilde{\theta})-\gamma_n(\widetilde{\theta})\mathrm{pen}(\widehat{m})$ with respect to $A_{\widehat{m}}$ and $B_{\widehat{m}}$ as follows.

Lemma 4.19. *Almost surely, it holds that*

$$-\overline{\gamma}_n(\widetilde{\theta})-\gamma_n(\widetilde{\theta})pen(\widehat{m})-\sigma^2+\|\boldsymbol{\epsilon}\|_n^2 \leq l(\widetilde{\theta},\theta)\left[A_{\widehat{m}}\vee(1-\kappa_2)\right]+\sigma^2 B_{\widehat{m}}\ . \tag{4.28}$$

Let us set the constants

$$\kappa_1 := \frac{1}{4} \quad \text{and} \quad \kappa_2 := \frac{(K-1)(1-\sqrt{\eta})^2}{16}\wedge 1\ . \tag{4.29}$$

We do not claim that this choice is optimal, but we are not really concerned about the constants for this result. The core of this proof consists in showing that with overwhelming probability the variable $A_{\widehat{m}}$ is smaller than 1 and $B_{\widehat{m}}$ is smaller than a constant over n.

Lemma 4.20. *The event Ω_1 defined as*

$$\Omega_1 := \left\{A_{\widehat{m}} \leq \frac{7}{8}\right\}\bigcap\left\{\kappa_2 n\varphi_{max}\left[\left(\mathbf{Z}^*_{\widehat{m}}\mathbf{Z}_{\widehat{m}}\right)^{-1}\right] \leq \frac{K-1}{4}\right\}$$

satisfies $\mathbb{P}(\Omega_1^c) \leq L\mathrm{Card}(\mathcal{M})\exp\left[-nL'(K,\eta)\right]$, where $L'(K,\eta)$ is positive.

Lemma 4.21. *There exists an event Ω_2 of probability larger than $1-\exp(-nL)$ with $L>0$ such that*

$$\mathbb{E}\left[B_{\widehat{m}}\mathbf{1}_{\Omega_1\cap\Omega_2}\right] \leq \frac{L(K,\eta,\alpha,\beta)}{n}\ .$$

Gathering the upper bound (4.25) and Lemma 4.19, 4.20, and 4.21, we conclude that

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1\cap\Omega_2}\left(\kappa_2\wedge\frac{1}{8}\right)\right] &\leq l(\theta_m,\theta)+\mathbb{E}\left[\gamma_n(\theta_m)\mathrm{pen}(m)\right]\\ &+ \sigma^2\frac{L(K,\eta,\alpha,\beta)}{n}+\mathbb{E}\left[\mathbf{1}_{\Omega_1\cap\Omega_2}\left(\overline{\gamma}_n\left(\theta_m\right)+\sigma^2-\|\boldsymbol{\epsilon}\|_n^2\right)\right]\ .\end{aligned}$$

As the expectation of the random variable $\overline{\gamma}_n\left(\theta_m\right)+\sigma^2-\|\boldsymbol{\epsilon}\|_n^2$ is zero, it holds that

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\Omega_1\cap\Omega_2}\left(\overline{\gamma}_n\left(\theta_m\right)+\sigma^2-\|\boldsymbol{\epsilon}\|_n^2\right)\right] &= \mathbb{E}\left[\mathbf{1}_{\Omega_1^c\cup\Omega_2^c}\left(\overline{\gamma}_n\left(\theta_m\right)+\sigma^2-\|\boldsymbol{\epsilon}\|_n^2\right)\right]\\ &\leq \sqrt{\mathbb{P}(\Omega_1^c)+\mathbb{P}(\Omega_2^c)}\left[\sqrt{\mathbb{E}\left[\|\boldsymbol{\epsilon}_m\|_n^2-l(\theta_m,\theta)\right]^2}+2\sqrt{\mathbb{E}\left[\langle\boldsymbol{\epsilon},\boldsymbol{\epsilon}_m\rangle_n^2\right]}\right]\\ &\leq \sqrt{\mathbb{P}(\Omega_1^c)+\mathbb{P}(\Omega_2^c)}\sqrt{\frac{2}{n}}\left[l(\theta_m,\theta)+\sigma\sqrt{2l(\theta_m,\theta)}\right]\ .\end{aligned}$$

The probabilities $\mathbb{P}(\Omega_1^c)$ and $\mathbb{P}(\Omega_2^c)$ converge to 0 at an exponential rate with respect to n. Hence, by taking the infimum over all the models $m\in\mathcal{M}$, we obtain

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1\cap\Omega_2}\right] &\leq L(K,\eta)\inf_{m\in\mathcal{M}}\left[l(\theta_m,\theta)+\left(\sigma^2+l(\theta_m,\theta)\right)\mathrm{pen}(m)\right]+L_2(K,\eta,\alpha,\beta)\frac{\sigma^2}{n}+\\ &+ L_3(K,\eta)\sqrt{\frac{\mathrm{Card}(\mathcal{M})}{n}}\left[\sigma^2+l(0_p,\theta)\right]\exp\left[-nL_4(K,\eta)\right]\ , \qquad (4.30)\end{aligned}$$

with $L_4(K,\eta)>0$. In order to conclude, we need to control the loss of the estimator $\widetilde{\theta}$ on the event of small probability $\Omega_1^c\cup\Omega_2^c$. Thanks to the following lemma, we may upper bound the r-th risk of the estimators $\widehat{\theta}_m$.

Proposition 4.22. *For any model m and any integer $r\geq 2$ such that $n-d_m-2r+1>0$,*

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^r\right]^{\frac{1}{r}}\leq Lrd_mn\left[\sigma^2+l(\theta_m,\theta)\right]\ .$$

The proof is postponed to Section 4.7.4. We derive from this bound a strong control on $\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1^c\cup\Omega_2^c}\right]$.

Lemma 4.23.

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1^c\cup\Omega_2^c}\right]\leq L(K,\eta)n^2\,Card(\mathcal{M})\,var(Y)\exp\left[-nL'(K,\eta)\right]\ , \qquad (4.31)$$

where $L'(K,\eta)$ is positive.

By Assumptions ($\mathbb{H}_{Pol}$) and ($\mathbb{H}_\eta$), the cardinality of the collection of $\mathcal{M}$ is smaller than $\alpha n^{1+\beta}$. We gather the upper bounds (4.30) and (4.31) and so we conclude. □

Proof of Lemma 4.19. Thanks to Lemma 4.15, we decompose $\overline{\gamma}_n(\widetilde{\theta})$ as

$$\overline{\gamma}_n(\widetilde{\theta})=\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2-\sigma^2-l(\theta_{\widehat{m}},\theta)-(1-\kappa_2)l(\widetilde{\theta},\theta_{\widehat{m}})-\kappa_2(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})^*\mathbf{Z}_{\widehat{m}}(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-2}\mathbf{Z}_{\widehat{m}}^*(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\ .$$

Since $2ab\leq\kappa_1a^2+\kappa_1^{-1}b^2$ for any $\kappa_1>0$, it holds that

$$\begin{aligned}-\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2+\|\boldsymbol{\epsilon}\|_n^2 &= \|\Pi_{\widehat{m}}\boldsymbol{\epsilon}\|_n^2-\|\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2-2\langle\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon},\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\rangle_n\\ &\leq \sigma^2\left[\kappa_1+\frac{\|\Pi_{\widehat{m}}\boldsymbol{\epsilon}\|_n^2}{\sigma^2}\right]+l(\theta_{\widehat{m}},\theta)\left[\frac{\|\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{l(\theta_{\widehat{m}},\theta)}+\kappa_1^{-1}\frac{\langle\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon},\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\rangle_n^2}{\sigma^2l(\theta_{\widehat{m}},\theta)}\right]\ .\end{aligned}$$

Besides, we upper bound Expression (4.23) of $l(\widetilde{\theta},\theta_{\widehat{m}})$ using the largest eigenvalue of $(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}$.

$$\begin{aligned}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})^*\mathbf{Z}_{\widehat{m}}(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-2}\mathbf{Z}_{\widehat{m}}^*(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}}) &\leq \varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right](\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})^*\mathbf{Z}_{\widehat{m}}(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\mathbf{Z}_{\widehat{m}}^*(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\\ &\leq \left[\sigma^2+l(\theta_{\widehat{m}},\theta)\right]n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]\frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2+l(\theta_{\widehat{m}},\theta)}\ . \qquad (4.32)\end{aligned}$$

Thanks to Assumption (4.8), we upper bound the penalty terms as follows:

$$-\gamma_n(\widetilde{\theta})\mathrm{pen}(\widehat{m}) \leq \left[\sigma^2 + l(\theta_{\widehat{m}},\theta)\right] \frac{\|\Pi^{\perp}_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2 + l(\theta_{\widehat{m}},\theta)} K \frac{d_{\widehat{m}}}{n-d_{\widehat{m}}} \ .$$

By gathering the four last identities, we get

$$-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})\mathrm{pen}(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \leq l(\widetilde{\theta},\theta)\left[A_{\widehat{m}} \vee (1-\kappa_2)\right] + \sigma^2 B_{\widehat{m}} \ ,$$

since $l(\widetilde{\theta},\theta)$ decomposes into the sum $l(\widetilde{\theta},\theta_{\widehat{m}}) + l(\theta_{\widehat{m}},\theta)$. □

Proof of Lemma 4.20. We recall that for any model $m \in \mathcal{M}$,

$$\begin{aligned} A_m &:= \frac{5}{4} + \frac{\|\Pi^{\perp}_m \boldsymbol{\epsilon}_m\|_n^2}{l(\theta_m,\theta)} + \kappa_2 n \varphi_{\max}\left[(\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right] \frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} \\ &- K\frac{d_m}{n-d_m}\frac{\|\Pi^{\perp}_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} \ . \end{aligned}$$

In order to control the variable $A_{\widehat{m}}$, we shall simultaneously bound the deviations of the four random variables involved in any variable A_m.

Since $\mathbf{X}_m$ is independent of $\boldsymbol{\epsilon}_m/\sqrt{l(\theta_m,\theta)}$ and since $\boldsymbol{\epsilon}_m/\sqrt{l(\theta_m,\theta)}$ is a standard Gaussian vector of size n, the random variable $\|\Pi^{\perp}_m\boldsymbol{\epsilon}_m\|_n^2/l(\theta_m,\theta)$ follows a χ^2 distribution with $n-d_m$ degrees of freedom conditionally on $\mathbf{X}_m$. As this distribution does not depend on $\mathbf{X}_m$, $\|\Pi^{\perp}_m\boldsymbol{\epsilon}_m\|_n^2/l(\theta_m,\theta)$ follows a χ^2 distribution with $n-d_m$ degrees of freedom. Similarly, the random variables $\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2/[l(\theta_m,\theta)+\sigma^2]$ and $\|\Pi^{\perp}_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2/[l(\theta_m,\theta)+\sigma^2]$ follow χ^2 distributions with respectively d_m and $n-d_m$ degrees of freedom. Besides, the matrix $(\mathbf{Z}_m^*\mathbf{Z}_m)$ follows a standard Wishart distribution with parameters (n,d_m).

Let x be a positive number we shall fix later. By Lemma 4.16 and 4.18, there exists an event Ω_1' of large probability

$$P(\Omega_1^{'c}) \leq 4\exp(-nx)\mathrm{Card}(\mathcal{M}) \ ,$$

such that for conditionally on Ω_1',

$$\frac{\|\Pi^{\perp}_m\boldsymbol{\epsilon}_m\|_n^2}{l(\theta_m,\theta)} \geq \frac{n-d_m}{n} - 2\sqrt{\frac{(n-d_m)x}{n}} \ , \tag{4.33}$$

$$\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} \leq \frac{d_m}{n} + 2\sqrt{\frac{d_m x}{n}} + 2x \ , \tag{4.34}$$

$$\frac{\|\Pi^{\perp}_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} \geq \frac{n-d_m}{n} - 2\sqrt{\frac{(n-d_m)x}{n}} \ , \tag{4.35}$$

$$\varphi_{\max}\left[(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right] \leq \left\{ n\left[\left(1-\sqrt{\frac{d_m}{n}}-\sqrt{2x}\right)\vee 0\right]^2\right\}^{-1} \ , \tag{4.36}$$

for every model $m \in \mathcal{M}$. Let us prove that for a suitable choice of the number x, $A_{\widehat{m}}\mathbf{1}_{\Omega_1'}$ is smaller than $7/8$. First, we constrain $n\kappa_2\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]$ to be smaller than $\frac{K-1}{4}$ on the event Ω_1'. By (4.36), it holds that

$$n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq \left[\left(1-\sqrt{\eta}-\sqrt{2x}\right)\vee 0\right]^{-2} \ .$$

Constraining x to be smaller than $\frac{(1-\sqrt{\eta})^2}{8}$ ensures that the largest eigenvalue of $(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}$ satisfies

$$n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq \frac{4}{\left(1-\sqrt{\eta}\right)^2} \ .$$

By definition (4.29) of κ_2, it follows that $n\kappa_2\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq (K-1)/4$. Applying inequality $2ab \leq \delta a^2 + \delta^{-1}b^2$ to the bounds (4.33), (4.34), and (4.35) yields

$$\begin{aligned} -\frac{\|\Pi^{\perp}_{\widehat{m}}\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{l(\theta_{\widehat{m}},\theta)} &\leq -\frac{1}{2} + \frac{d_{\widehat{m}}}{2n} + 2x \\ \kappa_2 n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]\frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2+l(\theta_{\widehat{m}},\theta)} &\leq \frac{K-1}{2}\left[\frac{d_{\widehat{m}}}{n}+\frac{3x}{2}\right] \\ -K\frac{d_{\widehat{m}}}{n-d\widehat{m}}\frac{\|\Pi^{\perp}_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2+l(\theta_{\widehat{m}},\theta)} &\leq -K\frac{d_{\widehat{m}}}{2n} + x\frac{2K\eta}{1-\eta} \ . \end{aligned}$$

Gathering these three inequalities, we get

$$A_{\widehat{m}}\mathbf{1}_{\Omega_1'} \leq \frac{3}{4} + x\left[2 + \frac{3(K-1)}{4} + 2K\frac{\eta}{1-\eta}\right] .$$

If we set x to

$$x := \left[8\left(2 + \frac{3(K-1)}{4} + 2K\frac{\eta}{1-\eta}\right)\right]^{-1} \wedge \frac{\left(1-\sqrt{\eta}\right)^2}{8} ,$$

then $A_{\widehat{m}}\mathbf{1}_{\Omega_1'}$ is smaller than $\frac{7}{8}$ and the result follows. □

Proof of Lemma 4.21. We shall simultaneously bound the deviations of the random variables involved in the definition of B_m for all models $m \in \mathcal{M}$. Let us first define the random variable E_m as

$$E_m \quad := \quad \kappa_1^{-1}\frac{\langle \Pi_m^\perp \boldsymbol{\epsilon}, \Pi_m^\perp \boldsymbol{\epsilon}_m\rangle_n^2}{\sigma^2 l(\theta_m,\theta)} + \frac{\|\Pi_m\boldsymbol{\epsilon}\|_n^2}{\sigma^2} .$$

Factorizing by the norm of $\boldsymbol{\epsilon}$, we get

$$E_m \quad \leq \quad \kappa_1^{-1}\frac{\|\boldsymbol{\epsilon}\|_n^2}{\sigma^2}\frac{\langle \frac{\Pi_m^\perp\boldsymbol{\epsilon}}{\|\Pi_m^\perp\boldsymbol{\epsilon}\|_n}, \Pi_m^\perp\boldsymbol{\epsilon}_m\rangle_n^2}{l(\theta_m,\theta)} + \frac{\|\Pi_m\boldsymbol{\epsilon}\|_n^2}{\sigma^2} . \tag{4.37}$$

The variable $\frac{\|\boldsymbol{\epsilon}\|_n^2}{\sigma^2}$ follows a χ^2 distribution with n degrees of freedom divided by n. By Lemma 4.16 there exists an event Ω_2 of probability larger than $1-\exp(n/8)$ such that $\frac{\|\boldsymbol{\epsilon}\|_n^2}{\sigma^2}$ is smaller than 2. As $\kappa_1^{-1} = 4$, we obtain

$$E_m\mathbf{1}_{\Omega_2} \quad \leq \quad 8\frac{\langle \frac{\Pi_m^\perp\boldsymbol{\epsilon}}{\|\Pi_m^\perp\boldsymbol{\epsilon}\|_n}, \Pi_m^\perp\boldsymbol{\epsilon}_m\rangle_n^2}{l(\theta_m,\theta)} + \frac{\|\Pi_m\boldsymbol{\epsilon}\|_n^2}{\sigma^2} .$$

Since $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}_m$, and $\mathbf{X}_m$ are independent, it holds that conditionally on $\mathbf{X}_m$ and $\boldsymbol{\epsilon}$,

$$n\frac{\langle \frac{\Pi_m^\perp\boldsymbol{\epsilon}}{\|\Pi_m^\perp\boldsymbol{\epsilon}\|_n}, \Pi_m^\perp\boldsymbol{\epsilon}_m\rangle_n^2}{l(\theta_m,\theta)} \sim \chi^2(1) .$$

Since the distribution depends neither on $\mathbf{X}_m$ nor on $\boldsymbol{\epsilon}$, this random variable follows a χ^2 distribution with 1 degree of freedom. Besides, it is independent of the variable $\frac{\|\Pi_m.\boldsymbol{\epsilon}\|_n^2}{\sigma^2}$. Arguing as previously, we work out the distribution

$$\frac{\|\Pi_m\boldsymbol{\epsilon}\|_n^2}{\sigma^2} \sim \frac{\chi^2(d_m)}{n} .$$

Consequently, the variable $E_m\mathbf{1}_{\Omega_2}$ is upper bounded by a random variable that follows the distribution of

$$\frac{8}{n}T_1 + \frac{1}{n}T_2 ,$$

where T_1 and T_2 are two independent χ^2 distribution with respectively 1 and d_m degrees of freedom. Moreover, the random variables $n\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$ and $n\frac{\|\Pi_m^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$ respectively follow a χ^2 distribution with d_m and $n-d_m$ degrees of freedom.

Let us bound the deviations of the random variables $E_m\mathbf{1}_{\Omega_2}$, $\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$, and $\frac{\|\Pi_m^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$ for any model $m \in \mathcal{M}$. We apply Lemma 1 in [LM00] for $E_m\mathbf{1}_{\Omega_2}$ and Lemma 4.16 for the two remaining random variables. Hence, for any $x > 0$, there exists an event $\mathbb{F}(x)$ of large probability

$$\begin{aligned}\mathbb{P}\left[\mathbb{F}(x)^c\right] \quad &\leq \quad e^{-x}\left(\sum_{m\in\mathcal{M}} e^{-\xi_1 d_m} + e^{-\xi_2 d_m} + e^{-\xi_3 d_m}\right) \\ &\leq \quad e^{-x}\left[3 + \alpha\sum_{d=1}^{+\infty} d^\beta(e^{-\xi_1 d} + e^{-\xi_2 d} + e^{-\xi_3 d})\right] ,\end{aligned}$$

such that conditionally on $\mathbb{F}(x)$,

$$\left\{\begin{array}{rcl} E_m \mathbf{1}_{\Omega_2} & \leq & \frac{d_m+8}{n} + \frac{2}{n}\sqrt{\left[d_m+8^2\right](\xi_1 d_m + x)} + 16\frac{\xi_1 d_m + x}{n} \\ \frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} & \leq & \frac{1}{n}\left(d_m + 2\sqrt{d_m\left[d_m\xi_2 + x\right]} + 2\left(d_m\xi_2 + x\right)\right) \\ -\frac{Kd_m}{n-d_m}\frac{\|(\Pi_m^{\perp}\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} & \leq & -\frac{Kd_m}{n(n-d_m)}\left(n - d_m - 2\sqrt{(n-d_m)(\xi_3 d_m + x)}\right) \end{array}\right. ,$$

for all models $m \in \mathcal{M}$. We shall fix later the positive constants ξ_1, ξ_2, and ξ_3. Let us apply extensively the inequality $2ab \leq \tau a^2 + \tau^{-1}b^2$. Hence, conditionally on $\mathbb{F}(x)$, the model $\widehat{m}$ satisfies

$$\left\{\begin{array}{rcl} E_{\widehat{m}} \mathbf{1}_{\Omega_2} & \leq & \frac{d_{\widehat{m}}}{n}\left[1 + 2\sqrt{\xi_1} + 17\xi_1 + \tau_1\right] + \frac{x}{n}\left[17 + \tau_1^{-1}\right] + \frac{72}{n} \\ \frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{l(\theta_{\widehat{m}},\theta)+\sigma^2} & \leq & \frac{d_{\widehat{m}}}{n}\left[1 + 2\sqrt{\xi_2} + 2\xi_2 + \tau_2\right] + \frac{x}{n}\left[2 + \tau_2^{-1}\right] \\ -\frac{Kd_{\widehat{m}}}{n-d_{\widehat{m}}}\frac{\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2+l(\theta_{\widehat{m}},\theta)} & \leq & -K\frac{d_{\widehat{m}}}{n}\left[1 - 2\sqrt{\xi_3\frac{d_{\widehat{m}}}{n-d_{\widehat{m}}}} - \tau_3\right] + K\frac{x}{n}\tau_3^{-1}\frac{d_{\widehat{m}}}{n-d_{\widehat{m}}} \end{array}\right. .$$

By Lemma 4.20, we know that conditionally on Ω_1, $\kappa_2 n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-1}\right]$ is smaller than $\frac{K-1}{4}$. By assumption $(\mathbb{H}_\eta)$, the ratio $\frac{d_{\widehat{m}}}{n-d_{\widehat{m}}}$ is smaller than $\frac{\eta}{1-\eta}$. Gathering these inequalities we upper bound $B_{\widehat{m}}$ on the event $\Omega_1 \cap \Omega_2 \cap \mathbb{F}(x)$,

$$B_{\widehat{m}} \leq \frac{d_{\widehat{m}}}{n}U + \frac{x}{n}V + \frac{72}{n} ,$$

where U and V are defined as

$$\begin{aligned} U &:= 1 + 2\sqrt{\xi_1} + 17\xi_1 + \tau_1 + \frac{K-1}{4}\left[1 + 2\sqrt{\xi_2} + 2\xi_2 + \tau_2\right] - K\left[1 - 2\sqrt{\xi_3}\sqrt{\frac{\eta}{1-\eta}} - \tau_3\right] \\ V &:= 17 + \tau_1^{-1} + \frac{K-1}{4}\left[2 + \tau_2^{-1}\right] + K\tau_3^{-1}\frac{\eta}{1-\eta} . \end{aligned}$$

Looking closely at U, one observes that it is the sum of the quantity $-\frac{3(K-1)}{4}$ and an expression that we can make arbitrary small by choosing the positive constants ξ_1, ξ_2, ξ_3, τ_1, τ_2, and τ_3 small enough. Consequently, there exists a suitable choice of these constants only depending on K and η that constrains the quantity U to be non positive. It follows that for any $x > 0$, with probability larger than $1 - e^{-x}L(K,\eta,\alpha,\beta)$,

$$B_{\widehat{m}}\mathbf{1}_{\Omega_1\cap\Omega_2} \leq \frac{x}{n}L(K,\eta) + \frac{L'(K,\eta)}{n} .$$

Integrating this upper bound for any $x > 0$, we conclude

$$\mathbb{E}\left[B_{\widehat{m}}\mathbf{1}_{\Omega_1\cap\Omega_2}\right] \leq \frac{L(K,\eta,\alpha,\beta)}{n} .$$

□

Proof of Lemma 4.23. We perform a very crude upper bound by controlling the sum of the risk of every estimator $\widehat{\theta}_m$.

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] \leq \sqrt{\mathbb{P}(\Omega_1^c) + \mathbb{P}(\Omega_2^c)}\sqrt{\sum_{m\in\mathcal{M}}\mathbb{E}\left[l(\widehat{\theta}_m,\theta)^2\right]} .$$

As for any model $m \in \mathcal{M}$, $l(\widehat{\theta}_m,\theta) = l(\theta_m,\theta) + l(\widehat{\theta}_m,\theta_m)$, it follows that

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta)^2\right] \leq 2\left\{l(\theta_m,\theta)^2 + \mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^2\right]\right\} .$$

For any model $m \in \mathcal{M}$, it holds that $n - d_m - 3 \geq (1-\eta)n - 3$, which is positive by assumption $(\mathbb{H}_\eta)$. Hence, we may apply Lemma 4.22 with $r = 2$ to all models $m \in \mathcal{M}$:

$$\begin{aligned} \mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^2\right] &\leq L\left[d_m n(\sigma^2 + l(\theta_m,\theta))\right]^2 \\ &\leq Ln^4\mathrm{var}(Y)^2 , \end{aligned}$$

since for any model m, $\sigma^2 + l(\theta_m,\theta) \leq \mathrm{var}(Y)$. By summing this bound for all models $m \in \mathcal{M}$ and applying Lemma 4.20 and 4.21, we get

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1^c\cup\Omega_2^c}\right] \leq n^2\mathrm{Card}(\mathcal{M})L(K,\eta)\mathrm{var}(Y)\exp\left[-nL'(K,\eta)\right] ,$$

where $L'(K,\eta)$ is positive.

□

4.7.3 Proof of Theorem 4.7

Proof of Theorem 4.7. This proof follows the same approach as the one of Theorem 4.2. We shall only emphasize the differences with this previous proof. The bound (4.25) still holds. Let us respectively define the three constants κ_1, κ_2 and $\nu(K)$ as

$$\begin{aligned}\kappa_1 &:= \frac{\sqrt{\frac{3}{K+2}}}{1-\sqrt{\eta}-\nu(K)}\ , \qquad \kappa_2 := \frac{(K-1)\left[1-\sqrt{\eta}\right]^2\left[1-\sqrt{\eta}-\nu(K)\right]^2}{16} \wedge 1\ ,\\ \nu(K) &:= \left(\frac{3}{K+2}\right)^{1/6} \wedge \frac{1-\left(\frac{3}{K+2}\right)^{1/6}}{2}\ .\end{aligned}$$

We also introduce the random variables $A_{m'}$ and $B_{m'}$ for any model $m' \in \mathcal{M}$.

$$\begin{aligned}A_{m'} &:= \kappa_1 + 1 - \frac{\|\Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\|_n^2}{l(\theta_{m'},\theta)} + \kappa_2 n\varphi_{\max}\left[(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2}\\ &- K\left[1+\sqrt{2H(d'_m)}\right]^2\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi_{m'^{\perp}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2}\ ,\\ B_{m'} &:= \kappa_1^{-1}\frac{\langle\Pi^{\perp}_{m'}\boldsymbol{\epsilon},\Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)} + \frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{\sigma^2} + \kappa_2 n\varphi_{\max}\left[(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2}\\ &- K\frac{d_{m'}}{n-d_{m'}}\left[1+\sqrt{2H(d'_m)}\right]^2\frac{\|\Pi^{\perp}_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2}\ .\end{aligned}$$

The bound given in Lemma 4.19 clearly extends to

$$-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})\mathrm{pen}(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \le l(\widetilde{\theta},\theta)\left[A_{\widehat{m}} \vee (1-\kappa_2)\right] + \sigma^2 B_{\widehat{m}}\ .$$

As previously, we control the variable $A_{\widehat{m}}$ on an event of large probability Ω_1 and take the expectation of $B_{\widehat{m}}$ on an event of large probability $\Omega_1 \cap \Omega_2$.

Lemma 4.24. *Let Ω_1 be the event*

$$\Omega_1 := \{A_{\widehat{m}} \le s(K,\eta)\} \bigcap \left\{\kappa_2 n\varphi_{max}\left[(\mathbf{Z}^*_{\widehat{m}}\mathbf{Z}_{\widehat{m}})^{-1}\right] \le \frac{(K-1)\left(1-\sqrt{\eta}-\nu(K)\right)^2}{4}\right\}\ ,$$

where $s(K,\eta)$ is a function smaller than one. Then, $\mathbb{P}\left(\Omega_1^c\right) \le L(K)n\exp\left[-nL'(K,\eta)\right]$ with $L'(K,\eta) > 0$.

The function $s(K,\eta)$ is given explicitly in the proof of Lemma 4.24

Lemma 4.25. *Let us assume that n is larger than some quantities $n_0(K)$. Then, there exists an event Ω_2 of probability larger than $1-\exp\left[-nL(K,\eta)\right]$ where $L(K,\eta) > 0$ such that*

$$\mathbb{E}\left[B_{\widehat{m}}\mathbf{1}_{\Omega_1\cap\Omega_2}\right] \le \frac{L(K,\eta)}{n}\ .$$

Gathering inequalities (4.25), (4.28), Lemma 4.24 and 4.25, we obtain as on the previous proof that

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)\mathbf{1}_{\Omega_1\cap\Omega_2}\right] &\le L(K,\eta)\inf_{m\in\mathcal{M}}\left[l(\theta_m,\theta) + \left(\sigma^2 + l(\theta_m,\theta)\right)\mathrm{pen}(m)\right] +\\ &+ L'(K,\eta)\left[\frac{\sigma^2}{n} + \left(\sigma^2 + l(0_p,\theta)\right)n\exp\left[-nL''(K,\eta)\right]\right]\ . \qquad (4.38)\end{aligned}$$

Afterwards, we control the loss of the estimator $\widetilde{\theta}$ on the event of small probability $\Omega_1^c \cup \Omega_2^c$.

Lemma 4.26. *If n is larger than some quantity $n_0(K)$,*

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] \le n^{5/2}\left(\sigma^2 + l(0_p,\theta)\right)L(K,\eta)\exp\left[-nL'(K,\eta)\right]\ ,$$

where $L(K,\eta)$ is positive.

Gathering this last bound with (4.38) enables to conclude. □

Proof of Lemma 4.24. This proof is analogous to the proof of Lemma 4.20, except that we shall change the weights in the concentration inequalities in order to take into account the complexity of the collection of models. Let x be a positive number we shall fix later. Applying Lemma 4.16, Lemma 4.17, and Lemma 4.18 ensures that there exists an event Ω_1' such that

$$P(\Omega_1^{'c}) \leq 4\exp(-nx) \sum_{m\in\mathcal{M}} \exp\left[-d_m H(d_m)\right] \ ,$$

and for all models $m \in \mathcal{M}$,

$$\frac{\|\Pi_m^\perp \boldsymbol{\epsilon}_m\|_n^2}{l(\theta_m,\theta)} \geq \frac{n-d_m}{n}\left[\left(1-\delta_{n-d_m}-\sqrt{\frac{2d_m H(d_m)}{n-d_m}}-\sqrt{\frac{2xn}{n-d_m}}\right)\vee 0\right]^2 , \quad (4.39)$$

$$\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} \leq \frac{2d_m}{n}\left[1+\sqrt{H(d_m)}+H(d_m)\right]+3x \ , \quad (4.40)$$

$$\frac{\|\Pi_m^\perp(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)} \geq \frac{n-d_m}{n}\left[\left(1-\delta_{n-d_m}-\sqrt{\frac{2d_m H(d_m)}{n-d_m}}-\sqrt{\frac{2xn}{n-d_m}}\right)\vee 0\right]^2 , \quad (4.41)$$

$$n\varphi_{\max}\left[(\mathbf{Z}_m^*\mathbf{Z}_m)^{-1}\right] \leq \left[\left(1-\left(1+\sqrt{2H(d_m)}\right)\sqrt{\frac{d_m}{n}}-\sqrt{2x}\right)\vee 0\right]^{-2} .$$

We recall that δ_d is defined in (4.24). Besides, it holds that

$$\mathbb{P}(\Omega_1^{'c}) \leq 4\exp[-nx]\sum_{d=0}^{n} \mathrm{Card}\left[\{m\in\mathcal{M},\ d_m=d\}\right]\exp[-dH(d)] \leq 4n\exp[-nx] \ .$$

By Assumption $(\mathbb{H}_{K,\eta})$, the expression $\left(1+\sqrt{2H(d_m)}\right)\sqrt{\frac{d_m}{n}}$ is bounded by $\sqrt{\eta}$. Hence, conditionally on Ω_1',

$$n\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right] \leq \left[\left(1-\sqrt{\eta}-\sqrt{2x}\right)\vee 0\right]^{-2} ,$$

Constraining x to be smaller than $\frac{\left(1-\sqrt{\eta}\right)^2}{8}$ ensures that

$$n\kappa_2\varphi_{\max}\left[(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]\mathbf{1}_{\Omega_1'} \leq \frac{(K-1)(1-\sqrt{\eta}-\nu(K))^2}{4} \ .$$

By assumption $(\mathbb{H}_{K,\eta})$, the dimension of any model $m\in\mathcal{M}$ is smaller than $n/2$. If n is larger than some quantities only depending on K, then $\delta_{n/2}$ is smaller than $\nu(K)$. Let us assume first that this is the case. We recall that $\nu(K)$ is defined at the beginning of the proof of Theorem 4.7. Since $\nu(K)\leq 1-\sqrt{\eta}$, inequality (4.39) becomes

$$\frac{\|\Pi_m^\perp \boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{l(\theta_{\widehat{m}},\theta)} \geq \left(1-\frac{d_{\widehat{m}}}{n}\right)\left[1-\nu(K)-\sqrt{\eta}\right]^2-2\sqrt{2x} \ .$$

Bounding analogously the remaining terms of $A_{\widehat{m}}$, we get

$$A_{\widehat{m}} \leq \kappa_1+1-\left[1-\sqrt{\eta}-\delta_{n/2}\right]^2+\frac{d_{\widehat{m}}}{n}(1-\sqrt{\eta}-\delta_{n/2})^2U_1+\sqrt{x}U_2+xU_3 \ ,$$

where U_1, U_2, and U_3 are respectively defined as

$$\begin{cases} U_1 := -K\left[1+\sqrt{2H(d_{\widehat{m}})}\right]^2+1+(K-1)/2\left[1+\sqrt{H(d_{\widehat{m}})}\right]^2\leq 0 \\ U_2 := 2\sqrt{2}\left[1+K\eta\right] \\ U_3 := \frac{3}{4}(K-1)\left[1-\sqrt{\eta}-\nu(K)\right]^2 \ . \end{cases}$$

Since U_1 is non-positive, we obtain an upper bound of $A_{\widehat{m}}$ that does not depend anymore on $\widehat{m}$. By assumption $(\mathbb{H}_{K,\eta})$, we know that $\eta<(1-\nu(K)-(\frac{3}{K+2})^{1/6})^2$. Hence, coming back to the definition of κ_1 allows to prove that κ_1 is strictly smaller than $[1-\sqrt{\eta}-\nu(K)]^2$. Setting

$$x := \left[\frac{\left[1-\sqrt{\eta}-\nu(K)\right]^2-\kappa_1}{4U_2}\right]^2 \wedge \frac{\left[1-\sqrt{\eta}-\nu(K)\right]^2-\kappa_1}{4U_3} \wedge \frac{\left(1-\sqrt{\eta}\right)^2}{8} \ ,$$

we get

$$A_{\widehat{m}} \leq 1 - \frac{1}{2}\left[(1-\sqrt{\eta}-\nu(K))^2 - \kappa_1\right] < 1 \ ,$$

on the event Ω_1'.

In order to take into account the case $\delta_{n/2} \geq \nu(K)$, we only have to choose a large constant $L(K)$ in the upper bound of $\mathbb{P}(\Omega_1^c)$. □

Proof of Lemma 4.25. Once again, the sketch of the proof closely follows the proof of Lemma 4.25. Let us consider the random variables E_m defined as

$$E_m := \kappa_1^{-1} \frac{\langle \Pi_{m'}^{\perp}\boldsymbol{\epsilon}, \Pi_{m'}^{\perp}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)} + \frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{\sigma^2} \ .$$

Since $\|\boldsymbol{\epsilon}\|_n^2/\sigma^2$ follows a χ^2 distribution with n degrees of freedom, there exists an event Ω_2 of probability larger than $1-\exp[-nL(K)]$such that $\|\boldsymbol{\epsilon}\|_n^2/\sigma^2$ is smaller than $\kappa_1^{-1} = \sqrt{(K+2)/3}[1-\sqrt{\eta}-\nu(K)]$ on Ω_2. The constant $L(K)$in the exponential is positive. We shall simultaneously upper bound the deviations of the random variables E_m, $\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2}$, and $\frac{\|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2+l(\theta_m,\theta)}$. Let ξ be some positive constant that we shall fix later. For any $x>0$, we define an event $\mathbb{F}(x)$ such that conditionally on $\mathbb{F}(x)\cap\Omega_2$,

$$\left\{\begin{array}{ccc} E_m & \leq & \frac{d_m+\kappa_1^{-2}}{n} + \frac{2}{n}\sqrt{\left[d_m+\kappa_1^{-4}\right]\left[d_m(\xi+H(d_m))+x\right]} \\ & + & 2\kappa_1^{-2}\frac{\xi(d_m+H(d_m))+x}{n} \\ \frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} & \leq & \frac{1}{n}\left[d_m + 2\sqrt{d_m\left[d_m(\frac{1}{16}+H(d_m))+x\right]} + 2\left[d_m(\frac{1}{16}+H(d_m))+x\right]\right] \\ \frac{\|\Pi_m^{\perp}\epsilon_m+\epsilon\|_n^2}{\sigma^2+l(\theta_m,\theta)} & \geq & \frac{n-d_m}{n}\left[\left(1-\delta_{n-d_m}-\sqrt{\frac{d_m(1+2H(d_m))}{n-d_m}}-\sqrt{\frac{2x}{n-d_m}}\right)\vee 0\right]^2 \ , \end{array}\right.$$

for any model $m\in\mathcal{M}$. Then, the probability of $\mathbb{F}(x)$ satisfies

$$\begin{aligned} \mathbb{P}\left[\mathbb{F}(x)^c\right] &\leq e^{-x}\left[\sum_{m\in\mathcal{M}} \exp\left[-H(d_m)\right]\left(e^{-\xi d_m} + e^{-\frac{d_m}{16}} + e^{-\frac{d_m}{2}}\right)\right] \\ &\leq e^{-x}\left(\frac{1}{1-e^{-\xi}} + \frac{1}{1-e^{-1/16}} + \frac{1}{1-e^{-1/2}}\right) \ . \end{aligned}$$

Let us expand the three deviation bounds thanks to the inequality $2ab \leq \tau a^2 + \tau^{-1}b^2$:

$$\begin{aligned} E_m &\leq \frac{d_m}{n}\left[1+2\sqrt{\xi}+2\kappa_1^{-2}\xi+\tau_1\xi+\tau_2\right] + \frac{x}{n}\left[2\kappa_1^{-2}+\tau_2^{-1}+\tau_1\right] \\ &+ \frac{\kappa_1^{-2}}{n}\left[1+\tau_1^{-1}\kappa_1^{-2}\right] + \frac{d_m H(d_m)}{n}\left[2\kappa_1^{-2}+\tau_1\right] + 2\frac{d_m\sqrt{H(d_m)}}{n} \\ &\leq \frac{d_m}{n}\left(1+\sqrt{2H(d_m)}\right)^2\left[\kappa_1^{-2}+2\sqrt{\xi}+2\kappa_1^{-2}\xi+\tau_1\xi+\tau_2\right] \\ &+ \frac{x}{n}\left[2\kappa_1^{-2}+\tau_2^{-1}+\tau_1\right] + \frac{\kappa_1^{-2}}{n}\left[1+\tau_1^{-1}\kappa_1^{-2}\right] \ . \end{aligned}$$

Similarly, we get

$$\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} \leq 2\frac{d_m}{n}\left[1+\sqrt{2H(d_m)}\right]^2 + 5\frac{x}{n} \ .$$

If n is larger than some quantity $n_0(K)$, then $\delta_{n/2}$ is smaller than $\nu(K)$. Applying Assumption $(\mathbb{H}_{K,\eta})$, we get

$$\begin{aligned} -K\frac{d_m}{n-d_m}&\left(1+\sqrt{2H(d_m)}\right)^2\frac{\|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{l(\theta_m,\theta)+\sigma^2} \\ &\leq -K\frac{d_m}{n}\left(1+\sqrt{2H(d_m)}\right)^2\left[\left(1-\sqrt{\eta}-\nu(K)-\sqrt{\frac{2x}{n-d_m}}\right)\vee 0\right]^2 \\ &\leq -K\frac{d_m}{n}\left(1+\sqrt{2H(d_m)}\right)^2\left[(1-\sqrt{\eta}-\nu(K))^2-\tau_3\right] + 2K\eta\tau_3^{-1}\frac{x}{n} \ . \end{aligned}$$

Let us combine these three bounds with the definitions of B_m, κ_1, and κ_2. Hence, Conditionally to the event $\Omega_1 \cap \Omega_2 \cap \mathbb{F}(x)$,

$$B_{\widehat{m}} \leq \frac{d_{\widehat{m}}}{n}\left[1+\sqrt{2H(\widehat{m})}\right]^2 U_1 + \frac{x}{n}U_2 + \frac{L(K,\eta)}{n}U_3 \ , \tag{4.42}$$

where

$$\left\{\begin{array}{lcl} U_1 & := & -\frac{K-1}{6}\left(1-\sqrt{\eta}-\nu(K)\right)^2 + K\tau_3 + 2\sqrt{\xi} + 2\kappa_1^{-2}\xi + \tau_1\xi + \tau_2 \ , \\ U_2 & := & \tau_2^{-1} + \tau_1 + L(K,\eta)(1+\tau_3^{-1}) \ , \\ U_3 & := & 1+\tau_1^{-1} \ . \end{array}\right.$$

Since $K>1$, there exists a suitable choice of the constants ξ, τ_1, and τ_2, only depending on K and η that constrains U_1 to be non positive. Hence, conditionally on the event $\Omega_1 \cap \Omega_2 \cap \mathbb{F}(x)$,

$$B_{\widehat{m}} \leq \frac{L(K,\eta)}{n} + L'(K,\eta)\frac{x}{n} \ .$$

Since $\mathbb{P}\left[\mathbb{F}(x)^c\right] \leq e^{-x}L(K,\eta)$, we conclude by integrating the last expression with respect to x.

□

Proof of Lemma 4.26. As in the ordered selection case, we apply Cauchy-Schwarz inequality

$$\mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] \quad \leq \quad \sqrt{\mathbb{P}(\Omega_1^c)+\mathbb{P}(\Omega_2^c)}\sqrt{\mathbb{E}\left[l(\widetilde{\theta},\theta)^2\right]} \ .$$

However, there are too many models to bound efficiently the risk of $\widetilde{\theta}$ by the sum of the risks of the estimators $\widehat{\theta}_m$. This is why we use here Hölder's inequality

$$\begin{aligned} \mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] &\leq L(K)\sqrt{n}\exp\left[-nL(K,\eta)\right]\sqrt{\mathbb{E}\left[\sum_{m\in\mathcal{M}}\mathbf{1}_{m=\widehat{m}}l(\widehat{\theta}_m,\theta)^2\right]} \\ &\leq L(K)\sqrt{n}\exp\left[-nL(K,\eta)\right]\sqrt{\sum_{m\in\mathcal{M}}\mathbb{P}\left(m=\widehat{m}\right)^{1/u}\mathbb{E}\left[l(\widehat{\theta}_m,\theta)^{2v}\right]^{1/v}}, \end{aligned} \tag{4.43}$$

where $v := \left\lfloor \frac{n}{8} \right\rfloor$, and $u =: \frac{v}{v-1}$. We assume here that n is larger than 8. For any model $m \in \mathcal{M}$, the loss $l(\widehat{\theta}_m,\theta)$ decomposes into the sum $l(\theta_m,\theta)+l(\widehat{\theta}_m,\theta_m)$. Hence,we obtain the following upper bound by applying Minkowski's inequality

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta)^{2v}\right]^{1/2v} \leq l(\theta_m,\theta) + \mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^{2v}\right]^{1/2v} \leq \mathrm{var}(Y) + \mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^{2v}\right]^{1/2v} \ . \tag{4.44}$$

We shall upper bound this last term thanks to Lemma 4.22. Since v is smaller than $n/8$ and since d_m is smaller than $n/2$, it follows that for any model $m \in \mathcal{M}$, $n-d_m-4v+1$ is positive and

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^{2v}\right]^{1/2v} \leq 2vLnd_m\sqrt{1+\frac{2}{n}}\left(\sigma^2+l(\theta_m,\theta)\right) \ ,$$

for any model $m \in \mathcal{M}$. Since $d_m \leq n$ and since $\sigma^2+l(\theta_m,\theta) \leq \mathrm{var}(Y)$, we obtain

$$\mathbb{E}\left[l(\widehat{\theta}_m,\theta_m)^{2v}\right]^{1/2v} \leq 2vLn^2\mathrm{var}(Y)\sqrt{1+\frac{2}{n}} \ . \tag{4.45}$$

Gathering upper bounds (4.43), (4.44), and (4.45) we get

$$\begin{aligned} \mathbb{E}\left[l(\widetilde{\theta},\theta)1_{\Omega_1^c\cup\Omega_2^c}\right] &\leq L(K)\sqrt{n}\exp\left[-nL'(K,\eta)\right] \\ &\times \left[\mathrm{var}(Y)+2vLn^2\sqrt{1+\frac{2}{n}}\mathrm{var}(Y)\right]\sqrt{\sum_{m\in\mathcal{M}}\mathbb{P}\left(m=\widehat{m}\right)^{1/u}} \ . \end{aligned}$$

Since the sum over $m \in \mathcal{M}$ of $\mathbb{P}(m = \widehat{m})$ is one, the last term of the previous expression is maximized when every $\mathbb{P}(m = \widehat{m})$ equals $\frac{1}{\mathrm{Card}(\mathcal{M})}$. Hence,

$$\mathbb{E}\left[l(\widetilde{\theta}, \theta) 1_{\Omega_1^c \cup \Omega_2^c}\right] \quad \leq \quad n^{5/2} \mathrm{var}(Y) L(K, \eta) \mathrm{Card}(\mathcal{M})^{1/(2v)} \exp\left[-nL'(K, \eta)\right] \ ,$$

where $L'(K, \eta)$ is positive. Let us first bound the cardinality of the collection $\mathcal{M}$. We recall that the dimension of any model $m \in \mathcal{M}$ is assumed to be smaller than $n/2$ by $(\mathbb{H}_{K,\eta})$. Besides, for any $d \in \{1, \ldots, n/2\}$, there are less than $\exp(dH(d))$ models of dimension d. Hence,

$$\log(\mathcal{M}) \leq \log(n) + \sup_{d=1,\ldots,n/2} dH(d) \ .$$

By assumption $(\mathbb{H}_{K,\eta})$, $dH(d)$ is smaller than $n/2$. Thus, $\log(\mathcal{M}) \leq \log(n) + n/2$ and it follows that $\mathrm{Card}(\mathcal{M})^{1/(2v)}$ is smaller than an universal constant providing that n is larger than 8. All in all, we get

$$\mathbb{E}\left[l(\widetilde{\theta}, \theta) 1_{\Omega_1^c \cup \Omega_2^c}\right] \quad \leq \quad n^{5/2} \mathrm{var}(Y) L(K, \eta) \exp\left[-nL'(K, \eta)\right] \ ,$$

where $L'(K, \eta)$ is positive. □

4.7.4 Proof of Proposition 4.22

Proof of Proposition 4.22. Let m be a subset of $\{1, \ldots, p\}$. Thanks to (4.23), we know that

$$l(\widehat{\theta}_m, \theta_m) = (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^* \mathbf{Z}_m \left(\mathbf{Z}_m^* \mathbf{Z}_m\right)^{-2} \mathbf{Z}_m^* (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m) \ .$$

Applying Cauchy-Schwarz inequality, we decompose the r-th loss of $\widehat{\theta}_m$ in two terms

$$\begin{aligned} \mathbb{E}\left[l(\widehat{\theta}_m, \theta_m)^r\right]^{\frac{1}{r}} \quad &\leq \quad \mathbb{E}\left[\left\|(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^*\right\|_F^r \left\|\mathbf{Z}_m \left(\mathbf{Z}_m^* \mathbf{Z}_m\right)^{-2} \mathbf{Z}_m^*\right\|_F^r\right]^{\frac{1}{r}} \\ &\leq \quad \mathbb{E}\left[\left\|(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^*\right\|_F^r\right]^{\frac{1}{r}} \mathbb{E}\left\{tr\left[\left(\mathbf{Z}_m^* \mathbf{Z}_m\right)^{-2}\right]^{\frac{r}{2}}\right\}^{\frac{1}{r}} , \end{aligned} \tag{4.46}$$

by independence of $\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}_m$, and $\mathbf{Z}_m$. Here, $\|.\|_F$ stands for the Frobenius norm in the space of square matrices. We shall successively upper bound the two terms involved in (4.46).

$$\left\|(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^*\right\|_F^r \quad = \quad \left[\sum_{1 \leq i,j \leq n} (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)[i]^2 (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)[j]^2\right]^{r/2} \ .$$

This last expression corresponds to the $L_{r/2}$ norm of a Gaussian chaos of order 4. By Theorem 3.2.10 in [dlPG99], such chaos satisfy a Khintchine-Kahane type inequality:

Lemma 4.27. *For all $d \in \mathbb{N}$ there exists a constant $L_d \in (0, \infty)$ such that, if X is a Gaussian chaos of order d with values in any normed space F with norm $\|.\|$ and if $1 < s < q < \infty$, then*

$$\left(\mathbb{E}\|X\|^q\right)^{\frac{1}{q}} \leq L_d \left(\frac{q-1}{s-1}\right)^{d/2} \mathbb{E}\left[\|X\|^s\right]^{\frac{1}{s}} \ .$$

Let us assume that r is larger than four. Applying the last lemma with $d = 4$, $q = r/2$, and $s = 2$ yields

$$\mathbb{E}\left[\left\|(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^*\right\|_F^r\right]^{\frac{2}{r}} \leq L_4 (r/2 - 1)^2 \mathbb{E}\left[\left\|(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^*\right\|_F^4\right]^{\frac{1}{2}} .$$

By standard Gaussian properties, we compute the fourth moment of this chaos and obtain

$$\mathbb{E}\left[\left\|(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^*\right\|_F^4\right]^{\frac{1}{2}} \quad \leq \quad Ln^2 \left[\sigma^2 + l(\theta_m, \theta)\right]^2 \ .$$

Hence, we get the upper bound

$$\mathbb{E}\left[\left\|(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)^*\right\|_F^r\right]^{\frac{1}{r}} \leq L(r-1)n \left[\sigma^2 + l(\theta_m, \theta)\right] \ . \tag{4.47}$$

Straightforward computations allow to extend this bound to $r = 2$ and $r = 3$.

Let us turn to bounding the second term of (4.46). Since the eigenvalues of the matrix $(\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}$ are almost surely non-negative, it follows that

$$tr\left[(\mathbf{Z}_m^* \mathbf{Z}_m)^{-2}\right] \leq tr\left[(\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right]^2 .$$

Consequently, we shall upper bound the r-th moment of the trace of an inverse standard Wishart matrix. For any couple of matrices A and B respectively of size $p_1 \times q_1$ and $p_2 \times q_2$, we define the Kronecker product matrix $A \otimes B$ as the matrix of size $p_1 p_2 \times q_1 q_2$ that satisfies:

$$A \otimes B[i_2 + p_2(i_1 - 1); j_2 + q_2(j_1 - 1)] := A[i_1; j_1] B[i_2; j_2] , \quad \text{for any } \begin{cases} 1 \leq i_1 \leq p_1 \\ 1 \leq i_2 \leq p_2 \\ 1 \leq j_1 \leq q_1 \\ 1 \leq j_2 \leq q_2 \end{cases} .$$

For any matrix A, $\otimes^k A$ refers to the k-th power of A with respect to the Kronecker product. Since $tr(A)^k = tr\left(\otimes^k A\right)$ for any square matrix A, we obtain

$$\begin{aligned} \mathbb{E}\left[tr(\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right]^k &= \mathbb{E}\left[tr\left(\otimes^k (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right)\right] \\ &= tr\left[\mathbb{E}\left(\otimes^k (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right)\right] \\ &\leq \sqrt{d_m^k} \left\|\mathbb{E}\left[\otimes^k (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right]\right\|_F , \end{aligned}$$

thanks to Cauchy-Schwarz inequality. In Equation (4.2) of [vR88], Von Rosen has characterized recursively the expectation of $\otimes^k (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}$ as long as $n - d_m - 2k - 1$ is positive:

$$\mathrm{vec}\left(\mathbb{E}\left[\otimes^{k+1} (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right]\right) = A(n, d_m, k)^{-1} \mathrm{vec}\left(\mathbb{E}\left[\otimes^k (\mathbf{Z}_m'^* \mathbf{Z}_m)^{-1}\right] \otimes I\right) , \tag{4.48}$$

where 'vec' refers to the vectorized version of the matrix. See Section 2 of [vR88] for more details about this definition. $A(n, d_m, k)$ is a symmetric matrix of size $d_m^{k+1} \times d_m^{k+1}$ which only depends on n, d_m, and k and is known to be diagonally dominant. More precisely, any diagonal element of $A(n, d_m, k)$ is greater or equal to one plus the corresponding row sums of the absolute values of the off-diagonal elements. Hence, the matrix A is invertible and its smallest eigenvalue is larger or equal to one. Consequently, $\varphi_{\max}\left(A^{-1}\right)$ is smaller or equal to one. It then follows from (4.48) that

$$\begin{aligned} \left\|\mathbb{E}\left[\otimes^{k+1} (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right]\right\|_F &= \left\|\mathrm{vec}\left(\mathbb{E}\left[\otimes^{k+1} (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right]\right)\right\|_F \\ &\leq \varphi_{\max}(A^{-1}) \left\|\mathrm{vec}\left(\mathbb{E}\left[\otimes^k (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right] \otimes I\right)\right\|_F \\ &\leq \sqrt{d_m} \left\|\mathbb{E}\left[\otimes^k (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right]\right\|_F . \end{aligned}$$

By induction, we obtain

$$\mathbb{E}\left[tr(\mathbf{Z}_m^* \mathbf{Z}_m)^{-1}\right]^r \leq d_m^r , \tag{4.49}$$

if $n - d_m - 2r + 1 > 0$. Combining upper bounds (4.47) and (4.49) enables to conclude

$$\mathbb{E}\left[l(\widehat{\theta}_m, \theta_m)^r\right]^{\frac{1}{r}} \leq Lr d_m n(\sigma^2 + l(\theta_m, \theta)) .$$

□

4.7.5 Proof of Proposition 4.4

Proof of Proposition 4.4. We shall prove that there exists a sequence $\tau(n)$ going to 0 such that with probability larger than $1 - L(s, R)\frac{\log n}{n^2}$,

$$\frac{l(\widetilde{\theta}, \theta)}{\inf_{m \in \mathcal{M}_{\lfloor n/2 \rfloor}}} \leq 1 + \tau(n) .$$

Then, we conclude by applying Borel-Cantelli Lemma.

Let m_* be the model that minimizes the loss function $l(\widehat{\theta}_m, \theta)$:

$$m_* = \arg \inf_{m \in \mathcal{M}_{\lfloor n/2 \rfloor}} l(\widehat{\theta}_m, \theta) \ .$$

It is almost surely uniquely defined. Contrary to the oracle m^*, the model m_* is random. By definition of $\widehat{m}$, we derive that

$$l(\widetilde{\theta}, \theta) \leq l(\widehat{\theta}_{m_*}, \theta) + \gamma_n(\widehat{\theta}_{m^*})\text{pen}(m_*) + \overline{\gamma}_n(\widehat{\theta}_{m_*}) - \gamma_n(\widetilde{\theta})\text{pen}(\widehat{m}) - \overline{\gamma}_n(\widetilde{\theta}) \ , \tag{4.50}$$

where $\overline{\gamma}_n$ is defined in the proof of Theorem 4.2. The proof divides in two parts. First, we state that conditionally to an event Ω_1 of large probability, the dimensions of $\widehat{m}$ and of m^* are moderate. Afterwards, we prove that conditionally to another event of large probability $\Omega_1 \cap \Omega_2 \cap \Omega_3$, the ratio $l(\widetilde{\theta}, \theta)/l(\widehat{\theta}_{m^*}, \theta)$ is close to one.

Lemma 4.28. *Let us define the event Ω_1 as:*

$$\Omega_1 := \left\{ \log^2(n) < d_{m_*} < \frac{n}{\log n} \quad \textit{and} \quad \log^2(n) < d_{\widehat{m}} < \frac{n}{\log n} \right\} \ .$$

The event Ω_1 is achieved with large probability: $\mathbb{P}(\Omega_1) \geq 1 - \frac{L(R,s)}{n^2}$.

Lemma 4.29. *There exists an event Ω_2 of probability larger than $1 - L\frac{\log n}{n}$ such that*

$$\left[-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta}) pen(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \right] \mathbf{1}_{\Omega_1 \cap \Omega_2} \leq l(\widetilde{\theta}, \theta)\tau_1(n),$$

where $\tau_1(n)$ is a positive sequence converging to zero when n goes to infinity.

Lemma 4.30. *There exists an event Ω_3 of probability larger than $1 - L\frac{\log n}{n}$ such that*

$$\left[\overline{\gamma}_n(\widetilde{\theta}) + \gamma_n(\widehat{\theta}_m) pen(\widehat{m}) + \sigma^2 - \|\boldsymbol{\epsilon}\|_n^2 \right] \mathbf{1}_{\Omega_1 \cap \Omega_3} \leq l\left(\widehat{\theta}_{m_*}, \theta\right) \tau_2(n),$$

where $\tau_2(n)$ is a positive sequence converging to zero when n goes to infinity.

Gathering these three lemma, we derive from the upper bound (4.50) the inequality

$$\frac{l(\widetilde{\theta}, \theta)}{l(\theta_{m_*}, \theta)} \mathbf{1}_{\Omega_1 \cap \Omega_2 \cap \Omega_3} \leq \frac{1 + \tau_2(n)}{1 - \tau_1(n)} \ ,$$

which allows to conclude.

□

Proof of Lemma 4.28. Let us consider the model $m_{R,s}$ defined by $d_{m_{R,s}} := \lfloor (nR^2)^{\frac{1}{1+s}} \rfloor$. If n is larger than some quantity $L(R,s)$, then $d_{m_{R,s}}$ is smaller than $n/2$ and $m_{R,s}$ therefore belongs to the collection $\mathcal{M}_{\lfloor n/2 \rfloor}$. We shall prove that outside an event of small probability, the loss $l(\widehat{\theta}_{m_{R,s}}, \theta)$ is smaller than the loss $l(\widehat{\theta}_m, \theta)$ of all models $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$ whose dimension is smaller than $\log^2(n)$ or larger than $\frac{n}{\log n}$. Hence, the model m_* satisfies $\log^2(n) < d_{m_*} < \frac{n}{\log n}$ with large probability.

First, we need to upper bound the loss $l(\widehat{\theta}_{m_{R,s}}, \theta)$. Since $l(\widehat{\theta}_{m_{R,s}}, \theta) = l(\theta_{m_{R,s}}, \theta) + l(\widehat{\theta}_{m_{R,s}}, \theta_{m_{R,s}})$, it comes to upper bounding both the bias term and the variance term. Since θ belongs to $\mathcal{E}'_s(R)$,

$$\begin{aligned} l\left(\theta_{m_{R,s}}, \theta\right) &= \sum_{i > d_{m_{R,s}}}^{+\infty} l(\theta_{m_{i-1}}, \theta_{m_i}) \\ &\leq (d_{m_i} + 1)^{-s} \sum_{i > d_{m_{R,s}}}^{+\infty} \frac{l(\theta_{m_{i-1}}, \theta_{m_i})}{i^{-s}} \leq \sigma^2 \left(\frac{R^2}{n^s} \right)^{\frac{1}{1+s}} \ . \end{aligned} \tag{4.51}$$

Then, we bound the variance term $l(\widehat{\theta}_{m_{R,s}}, \theta_{m_{R,s}})$ thanks to (4.32) as in the proof of Lemma 4.19.

$$l\left(\widehat{\theta}_{m_{R,s}}, \theta_{m_{R,s}}\right) \leq \left[\sigma^2 + l\left(\theta_{m_{R,s}}, \theta\right)\right] \varphi_{\max}\left[n(\mathbf{Z}^*_{m_{R,s}}\mathbf{Z}_{m_{R,s}})^{-1}\right] \frac{\left\|\Pi_{m_{R,s}}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{m_{R,s}})\right\|_n^2}{\sigma^2 + l(\theta_{m_{R,s}}, \theta)} .$$

The two random variables involved in this last expression respectively follow the distribution of an inverse Wishart matrix with parameters $(n, d_{m_{R,s}})$ and a χ^2 distribution with $d_{m_{R,s}}$ degrees of freedom. Thanks to Lemma 4.16 and 4.18, we prove that outside an event of probability smaller than $L(R,s)\exp[-L'(R,s)n^{\frac{1}{1+s}}]$ with $L'(R,s) > 0$,

$$l\left(\widehat{\theta}_{m_{R,s}}, \theta_{m_{R,s}}\right) \leq 4\left[\sigma^2 + l\left(\theta_{m_{R,s}}, \theta\right)\right] \frac{d_{m_{R,s}}}{n} ,$$

if n is large enough. Gathering this last upper bound with (4.51) yields

$$l\left(\widehat{\theta}_{m_{R,s}}, \theta\right) \leq \sigma^2 \left[5\frac{R^{\frac{2}{1+s}}}{n^{\frac{s}{1+s}}} + 4\left(\frac{R^{\frac{2}{1+s}}}{n^{\frac{s}{1+s}}}\right)^2\right] \leq \sigma^2 \frac{C(R,s)}{n^{\frac{s}{1+s}}} \tag{4.52}$$

where $C(R,s)$ is a constant that only depends on R and s.

Let us prove that the bias term of any model of dimension smaller than $\log^2(n)$ is larger than (4.52) if n is large enough. Obviously, we only have to consider the model of dimension $\lfloor \log^2(n) \rfloor$. Assume that there exists an infinite increasing sequence of integers u_n satisfying:

$$\sum_{i > \log^2(u_n)} l\left(\theta_{m_{i-1}}, \theta_{m_i}\right) \leq \frac{C(R,s)}{(u_{n+1})^{\frac{s}{1+s}}} . \tag{4.53}$$

Then, the sequence (v_n) defined by $v_n := \log^2(u_n)$ satisfies

$$\sum_{i > v_n} l\left(\theta_{m_{i-1}}, \theta_{m_i}\right) \leq C(R,s)\exp\left[-\sqrt{v_{n+1}}\frac{s}{1+s}\right] .$$

Let us consider a subsequence of (v_n) such that $\lfloor v_n \rfloor$ is strictly increasing. For the sake of simplicity we still call it v_n. It follows that

$$\begin{aligned}
\sum_{i=\lfloor v_0 \rfloor + 1}^{+\infty} \frac{l\left(\theta_{m_{i-1}}, \theta_{m_i}\right)}{i^{-s'}} &= \sum_{n=0}^{+\infty} \sum_{i=\lfloor v_n \rfloor + 1}^{\lfloor v_{n+1} \rfloor} \frac{l\left(\theta_{m_{i-1}}, \theta_{m_i}\right)}{i^{-s'}} \\
&\leq C(R,s) \sum_{n=0}^{+\infty} \lfloor v_{n+1} \rfloor^{s'} \exp\left[-\sqrt{\lfloor v_{n+1} \rfloor}\frac{s}{1+s}\right] \leq \infty ,
\end{aligned}$$

and θ therefore belongs to some ellipsoid $\mathcal{E}_{s'}(R')$. This contradicts the assumption θ does not belong to any ellipsoid $\mathcal{E}_{s'}(R')$. As a consequence, there only exists a finite sequence of integers u_n that satisfy Condition (4.53). For n large enough, the bias term of any model of dimension less than $\log^2(n)$ is therefore larger than the loss $l(\widehat{\theta}_{m_{R,s}}, \theta)$ with overwhelming probability.

Let us turn to the models of dimension larger than $n/\log n$. We shall prove that with large probability, for any model m of dimension larger than $n/\log n$, the variance term $l(\widehat{\theta}_m, \theta_m)$ is larger than the order $\sigma^2/\log n$. For any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$,

$$l\left(\widehat{\theta}_m, \theta_m\right) \geq \frac{n\sigma^2}{\varphi_{\max}\left(\mathbf{Z}^*_m \mathbf{Z}_m\right)} \frac{\|\Pi_m(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)\|_n^2}{\sigma^2 + l(\theta_m, \theta)} .$$

The two random variables involved in this expression respectively follow a Wishart distribution with parameters (n, d_m) and a χ^2 distribution with d_m degrees of freedom rescaled by n. Again, we apply Lemma 4.16 and 4.18 to control the deviations of these random variables. Hence, outside an event of probability smaller than $L(\xi)\exp[-n\xi/\log n]$,

$$l\left(\widehat{\theta}_m, \theta_m\right) \geq \sigma^2 \left(1 + \sqrt{\frac{d_m}{n}} + \sqrt{2\xi\frac{d_m}{n}}\right)^{-2} \frac{d_m}{n}\left(1 - 2\sqrt{\xi}\right) ,$$

for any model m of dimension larger than $n/\log n$. For any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$, the ratio d_m/n is smaller than $1/2$. As a consequence, we get

$$l\left(\widehat{\theta}_m, \theta_m\right) \geq \frac{\sigma^2}{\log n}\left(1-2\sqrt{\xi}\right)\left(1+\sqrt{1/2}+\sqrt{\xi}\right)^{-2} .$$

Choosing for instance $\xi = 1/16$ ensures that for n large enough the loss $l(\widehat{\theta}_m, \theta_m)$ is larger than $l(\widehat{\theta}_{m_{R,s}}, \theta)$ for every model m of dimension larger than $n/\log n$ outside an event of probability smaller than $L_1 \exp[-L_2 n/\log n] + L_3(R,s)\exp[-L_4(R,s)n^{1/(1+s)}]$ with $L_4(R,s) > 0$.

Let us now turn to the selected model $\widehat{m}$. We shall prove that outside an event of small probability,

$$\gamma_n\left(\widehat{\theta}_{m_{R,s}}\right)[1+\mathrm{pen}(m_{R,s})] \leq \gamma_n\left(\widehat{\theta}_m\right)[1+\mathrm{pen}(m)] , \tag{4.54}$$

for all models m of dimension smaller than $\log^2 n$ or larger than $n/\log n$. We first consider the models of dimension smaller than $\log^2(n)$. For any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$, $\gamma_n(\widehat{\theta}_m)[1+\mathrm{pen}(m)]$ follows a χ^2 distribution with $n-d_m$ degrees of freedom times $[\sigma^2 + l(\theta_m,\theta)][1+2d_m/(n-d_m)]/n$. Again, we apply Lemma 4.16. Hence, with probability larger than $1-e/[n^2(e-1)]$, the following upper bound holds for any model m of dimension smaller than $\log^2(n)$.

$$\begin{aligned}
\gamma_n\left(\widehat{\theta}_m\right)[1+\mathrm{pen}(m)] &\geq \sigma^2\left[1+\frac{l(\theta_m,\theta)}{\sigma^2}\right]\left(1+2\frac{d_m}{n-d_m}\right)\left[\frac{n-d_m}{n}-2\frac{\sqrt{(n-d_m)(d_m+2\log(n))}}{n}\right] \\
&\geq \sigma^2\left[1+\frac{l(\theta_m,\theta)}{\sigma^2}\right]\left(1+\frac{d_m}{n}\right)\left[1-2\sqrt{\frac{d_m+2\log(n)}{n-d_m}}\right] \\
&\geq \sigma^2\left[1+\frac{l(\theta_m,\theta)}{\sigma^2}\right]\left[1-4\frac{\log n}{\sqrt{n}}\right] ,
\end{aligned}$$

for n large enough. Besides, outside an event of probability smaller than $\frac{1}{n^2}$,

$$\begin{aligned}
\gamma_n\left(\widehat{\theta}_{m_{R,s}}\right)[1+\mathrm{pen}(m_{R,s})] &\leq \sigma^2\left[1+\frac{l(\theta_{m_{R,s}},\theta)}{\sigma^2}\right]\left(1+2\frac{d_{m_{R,s}}}{n-d_{m_{R,s}}}\right)\times \\
&\qquad \left[\frac{n-d_{m_{R,s}}}{n}+2\frac{\sqrt{(n-d_{m_{R,s}})2\log n}}{n}+4\frac{\log n}{n}\right] \\
&\leq \sigma^2\left[1+\frac{l(\theta_{m_{R,s}},\theta)}{\sigma^2}\right]\left(1+\frac{d_{m_{R,s}}}{n}\right)\left[1+2\frac{\sqrt{2\log n}}{\sqrt{n-d_{m_{R,s}}}}+4\frac{\log n}{n-d_{m_{R,s}}}\right] .
\end{aligned}$$

For n large enough, $d_{m_{R,s}}$ is smaller than $\frac{n}{2}$, and the last upper bound becomes:

$$\gamma_n\left(\widehat{\theta}_{m_{R,s}}\right)[1+\mathrm{pen}(m_{R,s})] \leq \sigma^2\left[1+\frac{C(R,s)}{n^{\frac{s}{1+s}}}\right]^2\left(1+10\frac{\log(n)}{\sqrt{n}}\right) .$$

Hence, $\gamma_n\left(\widehat{\theta}_{m_{R,s}}\right)[1+\mathrm{pen}(m_{R,s})] \leq \gamma_n\left(\widehat{\theta}_m\right)[1+\mathrm{pen}(m)]$ if

$$\frac{l(\theta_{m_{\lfloor \log^2 n \rfloor}},\theta)}{\sigma^2} \geq 3\frac{C(R,s)}{n^{\frac{s}{1+s}}}\times\frac{1+10\log(n)/\sqrt{n}}{1-4\log(n)/\sqrt{n}}+14\frac{\log(n)}{\sqrt{n}} .$$

As previously, this inequality always holds except for a finite number of n, since θ does not belong to any ellipsoid $\mathcal{E}_{s'}(R')$. Thus, outside an event of probability smaller than $\frac{L}{n^2}$, $d_{\widehat{m}}$ is larger than $\log^2 n$.

Let us now turn to the models of large dimension. Inequality (4.54) holds if the quantity

$$\|\epsilon\|_n^2\left(\frac{2d_{m_{R,s}}}{n-d_{m_{R,s}}}-\frac{2d_m}{n-d_m}\right)+\|\Pi_m\epsilon\|_n^2\left(1+\frac{2d_m}{n-d_m}\right)+\langle\Pi^{\perp}_{m_{R,s}}\epsilon_{m_{R,s}},\Pi^{\perp}_{m_{R,s}}\epsilon+2\epsilon_{m_{R,s}}\rangle_n\left(1+\frac{2d_{m_{R,s}}}{n-d_{m_{R,s}}}\right) \tag{4.55}$$

is non-positive. The three following bounds hold outside an event of probability smaller than $\frac{L(\xi)}{n^2}$:

$$\begin{aligned}
\|\boldsymbol{\epsilon}\|_n^2 &\geq 1 - 4\frac{\sqrt{\log n}}{\sqrt{n}} \ , \\
\|\Pi_m \boldsymbol{\epsilon}\|_n^2 &\leq (1+\xi)\frac{d_m}{n}, \ \text{ for all models } m \text{ of dimension } d_m > \frac{n}{\log n} \ , \\
\langle \Pi^{\perp}_{m_{R,s}} \boldsymbol{\epsilon}_{m_{R,s}}, \Pi^{\perp}_{m_{R,s}} \boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}_{m_{R,s}} \rangle_n &\leq l(\theta_{m_{R,s}}, \theta) \left[\frac{n - d_{m_{R,s}}}{n} + 4\frac{\sqrt{(n-d_{m_{R,s}})\log n}}{n} + \frac{4\log n}{n} \right] \\
&+ 4\sqrt{l(\theta_{m_{R,s}}, \theta)}\sigma \frac{\sqrt{(n-d_{m_{R,s}})\log n}}{n} \ .
\end{aligned}$$

Gathering these three inequalities we upper bound (4.55) by

$$\sigma^2 \frac{d_m}{n - d_m} \left[-2 + 8\sqrt{\frac{\log n}{n}} + (1+\xi)\left(\frac{n+d_m}{n}\right) \right] + 2\sigma^2 \frac{d_{m_{R,s}}}{n - d_{m_{R,s}}} + \\
+\sigma^2 L \left(1 + \frac{d_{m_{R,s}}}{n}\right) \left(\frac{l(\theta_{m_{R,s}}, \theta)}{\sigma^2} + \frac{\sqrt{l(\theta_{m_{R,s}}, \theta)}}{\sigma} \right) \left(1 + \sqrt{\frac{\log n}{n - d_{m_{R,s}}}} \right) \ .$$

The dimension of any model $m \in \mathcal{M}_{\lfloor n/2 \rfloor}$ is assumed to be smaller than $n/2$ and the dimensions of the models m considered are larger than $\frac{n}{\log n}$. For ξ small enough and n large enough, the previous expression is therefore upper bounded by

$$\sigma^2 \frac{2}{\log n} \left[\frac{3}{2}(1+\xi) - 2 + 8\sqrt{\frac{\log n}{n}} \right] + L\sigma^2 \left[\frac{R^{\frac{2}{1+s}}}{n^{\frac{s}{1+s}}} + \frac{R^{\frac{1}{1+s}}}{n^{\frac{a}{2(1+a)}}} \right] \ . \tag{4.56}$$

For n large enough, this last quantity is clearly non-positive.

All in all, we have proved that for n large enough outside an event of probability smaller than $\frac{L(R,s)}{n^2}$, it holds that

$$\log^2(n) < d_{m_*} < \frac{n}{\log n} \quad \text{and} \quad \log^2(n) < d_{\widehat{m}} < \frac{n}{\log n} \ .$$

□

Proof of Lemma 4.29. Arguing as in the proof of Theorem 4.2, we upper bound

$$-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})\mathrm{pen}(\widehat{m}) + \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2 \leq l(\theta_{\widehat{m}}, \theta) A_{\widehat{m}} + \sigma^2 B_{\widehat{m}} + (1 - \kappa_2(n)) l(\widetilde{\theta}, \theta_{\widehat{m}}) \ , \tag{4.57}$$

where $A_{\widehat{m}}$ and $B_{\widehat{m}}$ are respectively defined in (4.26) and in (4.27). We will fix the quantities $\kappa_1(n)$ and $\kappa_2(n)$ later. Besides, we define and bound the quantity $E_{\widehat{m}}$ as in (4.37).

Applying Lemma 4.16 and Lemma 4.18 and arguing as in the proofs of Lemma 4.20 and Lemma 4.21, there exists an event Ω_2 of large probability

$$\mathbb{P}(\Omega_1^c) \leq \exp\left[-n/8\right] + 5 \sum_{d=\log^2(n)}^{\frac{n}{\log n}} \exp\left[-\frac{2d}{\log n} \right] \leq \exp\left[-n/8\right] + \frac{5\log n}{2n^2(1 - 1/\log n)} \ ,$$

and such that conditionally on $\Omega_1 \cap \Omega_2$,

$$\begin{aligned}
\frac{\|\Pi^{\perp}_{\widehat{m}} \boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{l(\theta_{\widehat{m}}, \theta)} &\geq \frac{n - d_{\widehat{m}}}{n} - 2\frac{\sqrt{2(n-d_{\widehat{m}})d_{\widehat{m}}/\log n}}{n}, \\
\frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2 + l(\theta_{\widehat{m}}, \theta)} &\leq \frac{d_{\widehat{m}}}{n} + \frac{2\sqrt{2}d_{\widehat{m}}}{n\sqrt{\log n}} + 4\frac{d_{\widehat{m}}}{n\log n}, \\
\frac{\|\Pi^{\perp}_{\widehat{m}}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\sigma^2 + l(\theta_{\widehat{m}}, \theta)} &\geq \frac{n - d_{\widehat{m}}}{n} - 2\frac{\sqrt{2(n-d_{\widehat{m}})d_{\widehat{m}}/\log n}}{n} \\
\varphi_{\max}\left[(\mathbf{Z}^*_{\widehat{m}} \mathbf{Z}_{\widehat{m}})^{-1} \right] &\leq n^{-1} \left(1 - \left(1 + \sqrt{\frac{4}{\log n}} \right) \sqrt{\frac{d_{\widehat{m}}}{n}} \right)^{-2} \\
\|\boldsymbol{\epsilon}\|_n^2 &\leq 2 \\
E_{\widehat{m}} &\leq \frac{d_{\widehat{m}} + 2\kappa_1^{-1}(n)}{n} + \frac{2}{n}\sqrt{\left[d_{\widehat{m}} + \left(2\kappa_1^{-1}(n)\right)^2 \right] \frac{2d_{\widehat{m}}}{\log n}} + 8\kappa_1^{-1}(n)\frac{d_{\widehat{m}}}{n\log n} \ .
\end{aligned}$$

Gathering these six upper bounds, we are able to upper bound $A_{\widehat{m}}$ and $B_{\widehat{m}}$,

$$\begin{aligned}
A_{\widehat{m}} &\leq \kappa_1(n) + L_1\sqrt{\frac{d_{\widehat{m}}}{n\log n}} + \frac{d_{\widehat{m}}}{n}\left[-1 + L_2\sqrt{\frac{d_{\widehat{m}}}{(n-d_{\widehat{m}})\log n}} + \kappa_2(n)\frac{1+L_3/\sqrt{\log(n)}}{\left[1-(1+\sqrt{\frac{4}{\log n}})\sqrt{\frac{d_{\widehat{m}}}{n}})\right]^2}\right] , \\
B_{\widehat{m}} &\leq \frac{d_{\widehat{m}}}{n}\left[-1 + L_1\sqrt{\frac{d_{\widehat{m}}}{(n-d_{\widehat{m}})\log n}} + \kappa_2(n)\frac{1+L_2/\sqrt{\log(n)}}{\left[1-(1+\sqrt{\frac{4}{\log n}})\sqrt{\frac{d_{\widehat{m}}}{n}})\right]^2}\right] \\
&+ L_3\frac{d_{\widehat{m}}}{n}\left[\frac{\kappa_1^{-1}(n)}{d_{\widehat{m}}} + \frac{\kappa_1^{-1}(n)}{\log n} + \frac{1}{\sqrt{\log(n)}} + \frac{\kappa_1^{-1}(n)}{\sqrt{\log(n)d_{\widehat{m}}}}\right] .
\end{aligned}$$

Conditionally to the event Ω_1, the dimension of $\widehat{m}$ is moderate. Setting κ_1 to $\frac{1}{\log n}$, we get

$$\begin{aligned}
A_{\widehat{m}} &\leq \frac{L_1}{\log n} + \frac{d_{\widehat{m}}}{n}\left[-1 + \frac{L_2}{\log n} + \kappa_2(n)\frac{1+\frac{L_3}{\sqrt{\log n}}}{\left[1-\frac{L_4}{\sqrt{\log(n)}}\right]^2}\right] , \\
B_{\widehat{m}} &\leq \frac{d_{\widehat{m}}}{n}\left[-1 + \frac{L_1}{\log n} + \kappa_2(n)\frac{1+\frac{L_2}{\sqrt{\log n}}}{\left[1-\frac{L_3}{\sqrt{\log(n)}}\right]^2} + \frac{L_4}{\sqrt{\log n}}\right] .
\end{aligned}$$

Hence, there exists a sequence $\kappa_2(n)$ converging to one such that conditionally on $\Omega_1 \cap \Omega_2$, $B_{\widehat{m}}$ is non-positive and $A_{\widehat{m}}$ is bounded by $\frac{L}{\log n}$ when n is large enough. Coming back to the inequality (4.57) yields

$$\left[-\overline{\gamma}_n(\widetilde{\theta}) - \gamma_n(\widetilde{\theta})\mathrm{pen}(\widehat{m}) - \sigma^2 + \|\boldsymbol{\epsilon}\|_n^2\right]\mathbf{1}_{\Omega_1\cap\Omega_2} \leq l(\widetilde{\theta},\theta)\left[\frac{L}{\log n} \vee (1-\kappa_2(n))\right] ,$$

which concludes the proof. □

Proof of Lemma 4.30. We follow a similar approach to the previous proof.

$$\overline{\gamma}_n(\widehat{\theta}_{m_*}) + \gamma_n(\widehat{\theta}_{m_*})\mathrm{pen}(m_*) + \sigma^2 - \|\boldsymbol{\epsilon}\|_n^2 \leq C_{m_*}l(\theta_{m_*},\theta) + D_{m_*}\sigma^2 + \kappa_2(n)l(\widehat{\theta}_{m_*},\theta_{m_*}), \quad (4.58)$$

where for any model $m' \in \mathcal{M}_{\lfloor n/2\rfloor}$, $C_{m'}$ and $D_{m'}$ are respectively defined as

$$\begin{aligned}
C_{m'} &= \kappa_1(n) + \frac{\|\Pi_{m'}^{\perp}\boldsymbol{\epsilon}_{m'}\|_n^2}{l(\theta_{m'},\theta)} - 1 + 2\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi_{m'}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\
&- (1+\kappa_2(n))\frac{n}{\varphi_{\max}(\mathbf{Z}_{m'}^*\mathbf{Z}_{m'})}\frac{\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} \\
D_{m'} &= \kappa_1^{-1}(n)\frac{\langle\Pi_{m'}^{\perp}\boldsymbol{\epsilon},\Pi_{m'}^{\perp}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{\sigma^2 l(\theta_{m'},\theta)} - \frac{\|\Pi_{m'}\epsilon\|_n^2}{\sigma^2} \\
&- (1+\kappa_2(n))\frac{n}{\varphi_{\max}(\mathbf{Z}_{m'}^*\mathbf{Z}_{m'})}\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} + 2\frac{d_{m'}}{n-d_{m'}}\frac{\|\Pi_{m'}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(\theta_{m'},\theta)+\sigma^2} .
\end{aligned}$$

We fix $\kappa_1(n) = 1/\log n$ whereas $\kappa_2(n)$ will be fixed later. Arguing as in the proof of Lemma 4.29, there exists an event Ω_3 of large probability

$$\mathbb{P}(\Omega_3^c) \leq \exp\left[-n/8\right] + 5\sum_{d=\log^2(n)}^{\frac{n}{\log n}}\exp\left[-\frac{2d}{\log n}\right] \leq \exp\left[-n/8\right] + \frac{5\log n}{2n^2(1-1/\log(n))} ,$$

such that conditionally on $\Omega_1 \cap \Omega_3$, the two following bounds hold:

$$\begin{aligned} C_{m_*} &\leq \frac{L_1}{\log n} + \frac{d_{m_*}}{n}\left[1 + \frac{L_2}{\log n} - (1+\kappa_2(n))\frac{1 + L_3\sqrt{\frac{2}{\log n}}}{\left[1+\frac{L_4}{\sqrt{\log n}}\right]^2}\right] , \\ D_{m_*} &\leq \frac{d_{m_*}}{n}\left[1 + \frac{L_1}{\log n} + \frac{L_2}{\sqrt{\log n}} - (1+\kappa_2(n))\frac{1 + L_3\sqrt{\frac{2}{\log n}}}{\left[1+\frac{L_4}{\sqrt{\log n}}\right]^2}\right] , \end{aligned}$$

if n is large. The main difference with the proof of Lemma 4.29 lies in the fact that we now control the largest eigenvalue of $\mathbf{Z}_m^*\mathbf{Z}_m$ thanks to the second result of Lemma 4.18. There exists a sequence $\kappa_2(n)$ converging to 0 such that conditionally on $\Omega_1 \cap \Omega_3$, D_{m_*} is non-positive and C_{m_*} is bounded by $\frac{L}{\log n}$ when n is large. Coming back to (4.57) yields

$$\left[\overline{\gamma}_n(\widehat{\theta}_{m_*}) + \mathrm{pen}(m_*) + \sigma^2 - \|\epsilon\|_n^2\right]\mathbf{1}_{\Omega_1\cup\Omega_3} \leq l(\widehat{\theta}_{m_*},\theta)\left[\frac{L}{\log n} \vee \kappa_2(n)\right] ,$$

which concludes the proof. □

4.7.6 Proof of Proposition 4.5

Proof of Proposition 4.5. The approach is similar to the proof of Proposition 1 in [BM07]. For any model $m \in \mathcal{M}_{\lfloor n/2\rfloor}$, let us define

$$\Delta(m, m_{\lfloor n/2\rfloor}) := \gamma_n\left(\widehat{\theta}_{m_{\lfloor n/2\rfloor}}\right)\left[1+\mathrm{pen}(m_{\lfloor n/2\rfloor})\right] - \gamma_n\left(\widehat{\theta}_m\right)\left[1+\mathrm{pen}(m)\right] .$$

We shall prove that with large probability the quantity $\Delta(m, m_{\lfloor n/2\rfloor})$ is negative for any model m of dimension smaller than $n/4$. Hence, with large probability $d_{\widehat{m}}$ will be larger than $n/4$. Let us fix a model m of dimension smaller than $n/4$.

First, we use Expression (4.21) to lower bound $\gamma_n(\widehat{\theta}_m)$.

$$\begin{aligned} \gamma_n\left(\widehat{\theta}_m\right) &= \|\Pi_m^\perp\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\|_n^2 + \|\Pi_m^\perp\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\|_n^2 + 2\langle\Pi_m^\perp\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right), \Pi_m^\perp\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\rangle_n \\ &\geq \|\Pi_m^\perp\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\|_n^2 - \left\langle \Pi_m^\perp\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right), \frac{\Pi_m^\perp\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)}{\|\Pi_m^\perp\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\|_n}\right\rangle_n^2 , \end{aligned}$$

since $2ab \geq -a^2 - b^2$ for any number a and b. Hence, we may upper bound $\Delta(m, m_{\lfloor n/2\rfloor})$ by

$$\begin{aligned} \Delta(m, m_{\lfloor n/2\rfloor}) &\leq \|\Pi^\perp_{m_{\lfloor n/2\rfloor}}\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\|_n^2\left[\mathrm{pen}(m_{\lfloor n/2\rfloor}) - \mathrm{pen}(m)\right] \\ &- \left\|[\Pi_m^\perp - \Pi^\perp_{m_{\lfloor n/2\rfloor}}]\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\right\|_n^2\left[1+\mathrm{pen}(m)\right] \\ &+ \left\langle \Pi_m^\perp\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right), \frac{\Pi_m^\perp\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)}{\|\Pi_m^\perp\left(\boldsymbol{\epsilon}_m - \boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\|_n}\right\rangle_n^2\left[1+\mathrm{pen}(m)\right] . \end{aligned} \tag{4.59}$$

Arguing as the proof of Lemma 4.1, we observe that $\|\Pi^\perp_{m_{\lfloor n/2\rfloor}}\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\|_n^2$ follows a χ^2 distribution with $n - \lfloor n/2\rfloor$ degrees of freedom times $[\sigma^2 + l(\theta_{m_{\lfloor n/2\rfloor}})]/n$. Analogously, the random variable $\|[\Pi_m^\perp - \Pi^\perp_{m_{\lfloor n/2\rfloor}}]\left(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m_{\lfloor n/2\rfloor}}\right)\|_n^2$ follows a χ^2 distribution with $(d_{m_{\lfloor n/2\rfloor}} - d_m)$ degrees of freedom times $[\sigma^2 + l(\theta_{m_{\lfloor n/2\rfloor}})]/n$. Let us turn to the distribution of the third term. Coming back to the definition of ϵ_m, we observe that

$$\epsilon_m - \epsilon_{m_{\lfloor n/2\rfloor}} = Y - X\theta_m - (Y - X\theta_{m_{\lfloor n/2\rfloor}}) = X(\theta_m - \theta_{m_{\lfloor n/2\rfloor}}) .$$

Hence, $\epsilon_m - \epsilon_{m_{\lfloor n/2\rfloor}}$ is both independent of X_m and of $\epsilon + \epsilon_{m_{\lfloor n/2\rfloor}}$. Consequently, by conditioning and unconditioning, we conclude that the random variable defined in (4.59) follows a χ^2 distribution with 1 degree of freedom times $[\sigma^2 + l(\theta_{m_{\lfloor n/2\rfloor}})]/n$.

Once again, we apply Lemma 4.16 and the classical deviation bound $\mathbb{P}\left(|\mathcal{N}(0,1)| \geq \sqrt{2x}\right) \leq 2e^{-x}$. Let x be some positive number smaller than one that we shall fix later. There exists an event Ω_x of

probability larger than $1-\exp(-nx/2)-3\exp(-(n/4-1)x)\frac{1}{1-e^{-x}}$ such for any model of dimension smaller than $n/4$,

$$\begin{aligned}\frac{\Delta(m,m_{\lfloor n/2\rfloor})}{\sigma^2+l(\theta_{m_{\lfloor n/2\rfloor}})} &\leq \left(\frac{n-\lfloor n/2\rfloor}{n}\right)\left(1+2\sqrt{x}+2x\right)\left(\mathrm{pen}(m_{\lfloor n/2\rfloor})-\mathrm{pen}(m)\right)\\ &- \frac{\lfloor n/2\rfloor-d_m}{n}(1-2\sqrt{x}-2x)(1+\mathrm{pen}(m))\ .\end{aligned}$$

We now replace the penalty terms by their values thanks to Assumption (4.10). Conditionally to Ω_x, we obtain that

$$\frac{\Delta(m,m_{\lfloor n/2\rfloor})}{\sigma^2+l(\theta_{m_{\lfloor n/2\rfloor}})} \leq \frac{\lfloor n/2\rfloor-d_m}{n}\left\{4(1-\nu)(\sqrt{x}+x)\left[1+\frac{d_m}{n-d_m}\right]-\nu(1-2\sqrt{x}-2x)\right\}\ .$$

Since the dimension of the model m is smaller than $n/4$, $\frac{d_m}{n-d_m}$ is smaller than $1/3$. Hence, the last upper bound becomes

$$\frac{\Delta(m,m_{\lfloor n/2\rfloor})}{\sigma^2+l(\theta_{m_{\lfloor n/2\rfloor}})} \leq \frac{\lfloor n/2\rfloor-d_m}{n}\left\{\frac{16}{3}(1-\nu)(\sqrt{x}+x)-\nu(1-2\sqrt{x}-2x)\right\}\ .$$

There exists some $x(\nu)$ such that conditionally on $\Omega_{x(\nu)}$, $\Delta(m,m_{\lfloor n/2\rfloor})$ is negative for any model m of dimension smaller than $n/4$. Since $\mathbb{P}(\Omega^c_{x(\nu)})$ goes exponentially fast with ν to 0, there exists some $n_0(\nu,\delta)$ such that for any n larger than $n_0(\nu,\delta)$, $\mathbb{P}(\Omega^c_{x(\nu)})$ is smaller than δ. We have proved that with probability larger than $1-\delta$, the dimension of $\widehat{m}$ is larger than $n/4$.

Let us simultaneously lower bound the loss $l(\widehat{\theta}_m,\theta_m)$ for every model $m\in\mathcal{M}$ of dimension larger than $n/4$. Thanks to (4.23), we stochastically lower bound $l(\widehat{\theta}_m,\theta_m)$

$$\begin{aligned}l(\widehat{\theta}_m,\theta_m) &\geq n\varphi_{\max}\left(\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-1}\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2\\ &\succeq \varphi_{\max}\left(n\mathbf{Z}_m^*\mathbf{Z}_m\right)^{-1}\|\Pi_m\boldsymbol{\epsilon}\|_n^2,\end{aligned}$$

where $\mathbf{Z}_m^*\mathbf{Z}_m$ follows a standard wishart distribution with parameters (n,d_m). Applying Lemma 4.16 and Lemma 4.18 in order to simultaneously lower bound the loss $l(\widehat{\theta}_m,\theta_m)$, we find an event Ω' of probability larger than $1-\frac{2\exp(-n/4)}{1-e^{-1/16}}$, such that

$$l(\widehat{\theta}_m,\theta_m)\mathbf{1}_{\Omega'} \geq \left(1+\sqrt{\frac{d_m}{n}}+\sqrt{\frac{2d_m}{16n}}\right)^{-2}\frac{d_m}{2n}\sigma^2 \geq \frac{d_m}{8n}\sigma^2\ ,$$

for any model $m\in\mathcal{M}$ of dimension larger than $n/4$. On the event $\Omega_{x(\nu)}$, the dimension $d_{\widehat{m}}$ is larger than $n/4$. As a consequence, $l(\widetilde{\theta},\theta_{\widehat{m}})\mathbf{1}_{\Omega'\cap\Omega_{x(\nu)}}\geq\frac{\sigma^2}{32}$. All in all, we obtain

$$\begin{aligned}\mathbb{E}\left[l(\widetilde{\theta},\theta)\right] &\geq l(\theta_{m_{\lfloor n/2\rfloor}},\theta)+\mathbb{E}\left[\mathbf{1}_{\Omega'\cap\Omega_{x(\nu)}}l(\widetilde{\theta},\theta_{\widehat{m}})\right]\\ &\geq l(\theta_{m_{\lfloor n/2\rfloor}},\theta)+\left[1-\mathbb{P}(\Omega^c_{x(\nu)})-\mathbb{P}(\Omega'^c)\right]\frac{\sigma^2}{32}\\ &\geq l(\theta_{m_{\lfloor n/2\rfloor}},\theta)+L(\delta,\nu)\sigma^2\ ,\end{aligned}$$

if n is larger than some $n_0(\nu,\delta)$.

□

4.7.7 Proofs of the minimax lower bounds

All these minimax lower bounds are based on Birgé's version of Fano's Lemma [Bir05].

Lemma 4.31. (Birgé's Lemma) *Let (Θ,d) be some pseudo-metric space and $\{\mathbb{P}_\theta,\theta\in\Theta\}$ be some statistical model. Let κ denote some absolute constant smaller than one. Then for any estimator $\widehat{\theta}$ and any finite subset Θ_1 of Θ, setting $\delta=min_{\theta,\theta'\in\Theta_1,\theta\neq\theta'}d(\theta,\theta')$, provided that $\max_{\theta,\theta'\in\Theta_1}\mathcal{K}(\mathbb{P}_\theta,\mathbb{P}_{\theta'})\leq\kappa\log|\Theta_1|$, the following lower bound holds for every $p\geq1$,*

$$\sup_{\theta\in\Theta_1}\mathbb{E}_\theta[d^p(\widehat{\theta},\theta)]\geq 2^{-p}\delta^p(1-\kappa)\ .$$

First, we compute the Kullback-Leibler divergence between the distribution $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$.

$$\mathcal{K}\left(\mathbb{P}_\theta;\mathbb{P}_{\theta'}\right) = \mathcal{K}\left(\mathbb{P}_\theta(X);\mathbb{P}_{\theta'}(X)\right) + \mathbb{E}_\theta\left[\mathcal{K}\left(\mathbb{P}_\theta(Y|X);\mathbb{P}_{\theta'}(Y|X)\right)\,|\,X\right]$$

The two marginal distributions $\mathbb{P}_\theta(X)$ and $\mathbb{P}_{\theta'}(X)$ are equal. The conditional distributions $\mathbb{P}_\theta(Y|X)$ and $\mathbb{P}_{\theta'}(Y|X)$ are Gaussian with variance σ^2 and with mean respectively equal to $X\theta$ and $X\theta'$. Hence, the conditional Kullback-Leibler divergence equals

$$\mathcal{K}\left(\mathbb{P}_\theta(Y|X);\mathbb{P}_{\theta'}(Y|X)\right) = \frac{\left[X(\theta-\theta')\right]^2}{2\sigma^2}\ .$$

Reintegrating with respect to X yields

$$\mathcal{K}\left(\mathbb{P}_\theta;\mathbb{P}_{\theta'}\right) = \frac{l(\theta',\theta)}{2\sigma^2} \text{ and } \mathcal{K}\left(\mathbb{P}_\theta^{\otimes n};\mathbb{P}_{\theta'}^{\otimes n}\right) = n\frac{l(\theta',\theta)}{2\sigma^2}\ . \tag{4.60}$$

Proof of Proposition 4.9. First, we need a lower bound of the minimax rate of estimation on a subspace of dimension D.

Lemma 4.32. *Let D be some positive number smaller than p and r be some arbitrary positive number. Let S_D be the set of vectors in $\mathbb{R}^p$ whose support in included in $\{1,\dots,D\}$. Then, for any estimator $\widehat{\theta}$ of θ,*

$$\sup_{\theta\in S_D,\, l(0_p,\theta)\le Dr^2} \mathbb{E}_\theta\left[l(\widehat{\theta},\theta)\right] \ge LD\left[r^2\wedge\frac{\sigma^2}{n}\right]\ . \tag{4.61}$$

Let us fix some $D\in\{1,\dots,p\}$. Consider the set $\Theta_D := \left\{\theta\in S_D, l(0_p,\theta)\le a_D^2R^2\right\}$. Since the a_j's are non increasing, it holds that

$$\sum_{i=1}^{p}\frac{l(\theta_{m_{i-1}},\theta_{m_i})}{a_i^2} \le \sum_{i=1}^{D}\frac{l(\theta_{m_{i-1}},\theta_{m_i})}{a_D^2} \le \frac{l(0_p,\theta)}{a_D^2} \le R^2\ ,$$

for any $\theta\in\Theta_D$. Hence Θ_D is included in $\mathcal{E}_a(R)$. Applying Lemma 4.32, we get

$$\begin{aligned}\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}_a(R)} &\ge LD\left[\frac{a_D^2R^2}{D}\wedge\frac{\sigma^2}{n}\right]\\ &\ge L\left[a_D^2R^2\wedge\frac{D\sigma^2}{n}\right]\ .\end{aligned}$$

Taking the supremum over D in $\{1,\dots,p\}$ enables to conclude.

□

Proof of Lemma 4.32. Let us assume first that $\Sigma = I_p$. Consider the hypercube $\mathcal{C}_D(r) := \{0,r\}^D\times\{0\}^{p-D}$. Thanks to (4.60), we upper bound the Kullback-Leibler divergence between the distributions $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$

$$\mathcal{K}\left(\mathbb{P}_\theta^{\otimes n};\mathbb{P}_{\theta'}^{\otimes n}\right) \le \frac{nDr^2}{2\sigma^2}\ ,$$

where θ and θ' belong to $\mathcal{C}_D(r)$. Then, we apply Varshamov-Gilbert's lemma (e.g. Lemma 4.7 in [Mas07]) to the set $\mathcal{C}_D(r)$.

Lemma 4.33 (Varshamov-Gilbert's lemma)**.** *Let $\{0,1\}^D$ be equipped with Hamming distance d_H. There exists some subset Θ of $\{0,1\}^D$ with the following properties*

$$d_H(\theta,\theta') > D/4 \text{ for every } (\theta,\theta')\in\Theta^2 \text{ with } \theta\ne\theta' \text{ and } \log|\Theta|\ge D/8\ .$$

Combining Lemma 4.31 with the set Θ defined in the last lemma yields

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{C}_D(r)}\mathbb{E}_\theta\left[d_H(\widehat{\theta},\theta)\right] \ge \frac{D}{16}\ ,$$

provided that $\frac{nDr^2}{2\sigma^2} \leq D/16$. Coming back to the loss function $l(.,.)$ yields

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{C}_D(r)} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] \geq LDr^2 \ ,$$

if $r^2 \leq L\frac{\sigma^2}{n}$. Finally, we get

$$\inf_{\widehat{\theta}} \sup_{\theta \in S_D,\ l(0_p,\theta) \leq Dr^2} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] \geq LD \left[r^2 \wedge \frac{\sigma^2}{n} \right] \ .$$

If we no longer assume that the covariance matrix Σ is the identity, we orthogonalize the sequence X_i thanks to Gram-Schmidt process. Applying the previous argument to this new sequence of covariates allows to conclude. □

Proof of Corollary 4.10. This result follows from the upper bound on the risk of $\widetilde{\theta}$ in Theorem 4.2 and the minimax lower bound of Proposition 4.9. Let $\mathcal{E}_a(R)$ an ellipsoid satisfying $\frac{\sigma^2}{n} \leq R^2 \leq \sigma^2 n^\beta$, then $l(0_p, \theta)$ is smaller than $\sigma^2 n^\beta$. By Theorem 4.2, the estimator $\widetilde{\theta}$ defined with the collection $\mathcal{M}_{\lfloor n/2 \rfloor \wedge p}$ and $\mathrm{pen}(m) = K\frac{d_m}{n-d_m}$ satisfies

$$\begin{aligned} \mathbb{E}_\theta \left[l(\widetilde{\theta}, \theta) \right] &\leq L(K) \inf_{1 \leq i \leq \lfloor n/2 \rfloor \wedge p} \left\{ l(\theta_{m_i}, \theta) + K\frac{i}{n-i}[\sigma^2 + l(\theta_{m_i}, \theta)] \right\} + L(K,\beta)\frac{\sigma^2}{n} \\ &\leq L(K,\beta) \inf_{1 \leq i \leq \lfloor n/2 \rfloor \wedge p} \left[l(\theta_{m_i}, \theta) + \frac{i}{n}\sigma^2 \right] \ . \end{aligned}$$

If θ belongs to $\mathcal{E}_a(R)$, then

$$l(\theta_{m_i}, \theta) \leq a_{i+1}^2 \sum_{j=i+1}^{p} \frac{l(\theta_{m_j}, \theta_{m_{j-1}})}{a_j^2} \leq R^2 a_{i+1}^2 \ ,$$

since the (a_i)'s are increasing. It follows that

$$\mathbb{E}_\theta \left[l(\widetilde{\theta}, \theta) \right] \leq L(K,\beta) \inf_{1 \leq i \leq \lfloor n/2 \rfloor \wedge p} \left[R^2 a_{i+1}^2 + \frac{i}{n}\sigma^2 \right] \ . \tag{4.62}$$

Let us define $i^* := \sup \left\{ 1 \leq i \leq p \,,\, R^2 a_i^2 \geq \frac{\sigma^2 i}{n} \right\}$, with the convention $\sup \varnothing = 0$. Since $R^2 \geq \sigma^2/n$, i^* is larger or equal to one. By Proposition 4.9, the minimax rates of estimation is lower bounded as follows

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}_a(R)} \mathbb{E}_\theta \left[l(\widehat{\theta}, \theta) \right] \geq L \left[a_{i^*+1}^2 R^2 \vee \frac{\sigma^2 i^*}{n} \right] \geq L \left[a_{i^*+1}^2 R^2 + \frac{\sigma^2 i^*}{n} \right] \ .$$

If either $p \leq 2n$ or $a_{\lfloor n/2 \rfloor + 1}^2 R^2 \leq \sigma^2/2$, then i^* is smaller or equal to $\lfloor n/2 \rfloor \wedge p$ and we obtain thanks to (4.62) that

$$\begin{aligned} \mathbb{E}_\theta \left[l(\widetilde{\theta}, \theta) \right] &\leq L(K,\beta) \left[a_{i^*+1}^2 R^2 + \frac{\sigma^2 i^*}{n} \right] \\ &\leq L(K,\beta) \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}_a(R)} \mathbb{E} \left[l(\widehat{\theta}, \theta) \right] \ . \end{aligned}$$

□

Proof of Proposition 4.11. First, we use (4.60) to upper bound the Kullback-Leibler divergence between the distributions corresponding to parameters θ and θ' in the set $\Theta[k,p](r)$

$$\mathcal{K}\left(\mathbb{P}_\theta^{\otimes n}; \mathbb{P}_{\theta'}^{\otimes n} \right) \leq \frac{nkr^2}{2\sigma^2} \ ,$$

since the covariates are i.i.d standard Gaussian variables. Let us state a combinatorial argument due to Birgé and Massart [BM98].

Lemma 4.34. *Let $\{0,1\}^p$ be equipped with Hamming distance d_H and given $1 \leq k \leq p/4$, define $\{0,1\}^p_k := \{x \in \{0,1\}^p : d_H(0,x) = k\}$. There exists some subset Θ of $\{0,1\}^p_k$ with the following properties*

$$d_H(\theta,\theta') > k/8 \text{ for every } (\theta,\theta') \in \Theta^2 \text{ with } \theta \neq \theta' \text{ and } \log|\Theta| \geq k/5 \log\left(\frac{p}{k}\right) .$$

Suppose that k is smaller than $p/4$. Applying Lemma 4.31 with Hamming distance d_H and the set $r\Theta$ introduced in Lemma 4.34 yields

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[d_H\left(\widehat{\theta},\theta\right)\right] \geq \frac{k}{16} , \qquad \text{provided that} \qquad \frac{nkr^2}{2\sigma^2} \leq \frac{k}{10}\log\left(\frac{p}{k}\right) . \tag{4.63}$$

Since the covariates X_i are independent and of variance 1, the lower bound (4.63) is equivalent to

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l\left(\widehat{\theta},\theta\right)\right] \geq \frac{kr^2}{16}.$$

All in all, we obtain

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l\left(\widehat{\theta},\theta\right)\right] \geq Lk\left(r^2 \wedge \frac{\log\left(\frac{p}{k}\right)}{n}\sigma^2\right) .$$

Since p/k is larger than 4, we obtain the desired lower bound by changing the constant L:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l\left(\widehat{\theta},\theta\right)\right] \geq Lk\left(r^2 \wedge \frac{1+\log\left(\frac{p}{k}\right)}{n}\sigma^2\right) .$$

If p/k is smaller than 4, we know from the proof of Lemma 4.32, that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{C}_k(r)} \mathbb{E}_\theta \left[l\left(\widehat{\theta},\theta\right)\right] \geq Lk\left(r^2 \wedge \frac{\sigma^2}{n}\right) .$$

We conclude by observing that $\log(p/k)$ is smaller than $\log(4)$ and that $\mathcal{C}_k(r)$ is included in $\Theta[k,p](r)$. □

Proof of Proposition 4.13. Assume first the covariates (X_i) have a unit variance. If this is not the case, then one only has to rescale them. By Condition (4.18), the Kullback-Leibler divergence between the distributions corresponding to parameters θ and θ' in the set $\Theta[k,p](r)$ satisfies

$$\mathcal{K}\left(\mathbb{P}_\theta^{\otimes n}; \mathbb{P}_{\theta'}^{\otimes n}\right) \leq (1+\delta)^2 \frac{nkr^2}{2\sigma^2},$$

We recall that $\|.\|$ refers to the canonical norm in $\mathbb{R}^p$. Arguing as in the proof of Proposition 4.11, we lower bound the risk of any estimator $\widehat{\theta}$ with the loss function $\|.\|$,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[\|\widehat{\theta}-\theta\|^2\right] \geq Lk\left(r^2 \wedge \frac{1+\log\left(\frac{p}{k}\right)}{(1+\delta)^2 n}\sigma^2\right) ,$$

Applying again Assumption (4.18) allows to obtain the desired lower bound on the risk

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p](r)} \mathbb{E}_\theta \left[l(\widehat{\theta},\theta)\right] \geq Lk(1-\delta)^2\left(r^2 \wedge \frac{1+\log\left(\frac{p}{k}\right)}{(1+\delta)^2 n}\sigma^2\right) .$$

□

Proof of Proposition 4.14. We first consider the correlation matrix $\Psi_1(\omega)$. Let us pick a maximal subset $\Phi \subset \{1,\dots p\}$ of points that are $\lceil \log(4k)/\omega \rceil$ spaced with respect to the toroidal distance. Hence, the cardinality of Φ is $\lfloor p\lceil \log(4k)/\omega\rceil^{-1}\rfloor$. Assume that k is smaller than this quantity. We call C the correlation matrix of the points that belong to Φ. Obviously, for any $(i,j) \in \Phi^2$, it holds that $|C(i,j)| \leq 1/(4k)$ if $i \neq j$. Hence, any submatrix of C with size $2k$ is diagonally dominant and the sum of the absolute value of its non-diagonal elements is smaller than $1/2$. The matrix C therefore follows a $1/2$-Restricted Isometry Property of size $2k$. Consequently, we may apply Proposition 4.13 with the subset of covariates Φ and the result follows. The second case is handled similarly. □

4.8 Appendix

Proof of Lemma 4.15. We recall that $\gamma_n(\widehat{\theta}_m) = \|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2$. Thanks to the definition (4.19) of ϵ and ϵ_m , we obtain the first result. Let us turn to the mean squared error $\gamma(\widehat{\theta}_m)$. By definition,

$$\begin{aligned}\gamma(\widehat{\theta}_m) &= \mathbb{E}_{Y,X}\left[Y - X\widehat{\theta}_m\right]^2 = \mathbb{E}_{\epsilon,\epsilon_m,X_m}\left[\epsilon + \epsilon_m + X\left(\theta_m - \widehat{\theta}_m\right)\right]^2 \\ &= \left[\sigma^2 + l(\theta_m,\theta)\right] + \mathbb{E}_{X_m}\left[X\left(\theta_m - \widehat{\theta}_m\right)\right]^2 ,\end{aligned}$$

since $X\left(\theta_m - \widehat{\theta}_m\right)$ only depends on X_m and $\epsilon + \epsilon_m$ is independent of X_m. Hence, we get

$$\gamma(\widehat{\theta}_m) = \left[\sigma^2 + l(\theta_m,\theta)\right] + l(\widehat{\theta}_m,\theta_m) .$$

By definition, we derive that

$$l(\widehat{\theta}_m,\theta_m) = \mathbb{E}_{X_m}\left[X\left(\theta_m - \widehat{\theta}_m\right)\right]^2 = \left(\theta_m - \widehat{\theta}_m\right)^* \Sigma \left(\theta_m - \widehat{\theta}_m\right) .$$

Since $\widehat{\theta}_m$ is least-square of θ_m in the model (4.19), it follows that

$$l(\widehat{\theta}_m,\theta_m) = (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})^* \mathbf{X}_{\widehat{m}} (\mathbf{X}_{\widehat{m}}^* \mathbf{X}_{\widehat{m}})^{-1} \Sigma_m (\mathbf{X}_{\widehat{m}}^* \mathbf{X}_{\widehat{m}})^{-1} \mathbf{X}_{\widehat{m}}^* (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}}) .$$

We replace $\mathbf{X}_m$ by $\mathbf{Z}_m \sqrt{\Sigma_m}$ and therefore obtain

$$l(\widehat{\theta}_m,\theta_m) = (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})^* \mathbf{Z}_{\widehat{m}} (\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-2} \mathbf{Z}_{\widehat{m}}^* (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}}) .$$

□

Proof of Lemma 4.1. Thanks to Equation (4.21), we know that $\gamma_n(\widehat{\theta}_m) = \|\Pi_m^{\perp}(\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m)\|_n^2$. The variance of $\epsilon + \epsilon_m$ is $\sigma^2 + l(\theta_m,\theta)$. Since $\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m$ is independent of $\mathbf{X}_m$, $\gamma_n(\widehat{\theta}_m)$ follows a χ^2 distribution with $n - d_m$ degrees of freedom times $\frac{\sigma^2 + l(\theta_m,\theta)}{n}$ and the result follows.

Let us turn to the expectation of $\gamma(\widehat{\theta}_m)$. By (4.22), $\gamma(\widehat{\theta}_m)$ equals

$$\gamma\left(\widehat{\theta}_m\right) = \sigma^2 + l(\theta_m,\theta) + (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}})^* \mathbf{Z}_{\widehat{m}} (\mathbf{Z}_{\widehat{m}}^* \mathbf{Z}_{\widehat{m}})^{-2} \mathbf{Z}_{\widehat{m}}^* (\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_{\widehat{m}}) ,$$

following the arguments of the proof of Lemma 4.15. Since $\epsilon + \epsilon_m$ and X_m are independent, one may integrate with respect to $\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_m$

$$\mathbb{E}\left[\gamma(\widehat{\theta}_m)\right] = \left[\sigma^2 + l(\theta_m,\theta)\right]\left\{1 + \mathbb{E}\left[tr\left(\mathbf{Z}_m^* \mathbf{Z}_m\right)^{-1}\right)\right]\right\} ,$$

where the last term it the expectation of the trace of an inverse standard Wishart matrix of parameters (n, d_m). Thanks to [vR88], we know that it equals $\frac{d_m}{n - d_m - 1}$. □

Proof of Lemma 4.17. The random variable $\sqrt{\chi^2(d)}$ may be interpreted as a Lipschitz function with constant 1 on $\mathbb{R}^d$ equipped with the standard Gaussian measure. Hence, we may apply the Gaussian concentration theorem (see e.g. [Mas07] Th. 3.4). For any $x > 0$,

$$\mathbb{P}\left(\sqrt{\chi^2(d)} \leq \mathbb{E}\left[\sqrt{\chi^2(d)}\right] - \sqrt{2x}\right) \leq \exp(-x) . \tag{4.64}$$

In order to conclude, we need to lower bound $\mathbb{E}\left[\sqrt{\chi^2(d)}\right]$. Let us introduce the variable $Z := 1 - \sqrt{\frac{\chi^2(d)}{d}}$. By definition, Z is smaller or equal to one. Hence, we upper bound $\mathbb{E}(Z)$ as

$$\mathbb{E}(Z) \leq \int_0^1 \mathbb{P}(Z \geq t)dt \leq \int_0^{\sqrt{\frac{1}{8}}} \mathbb{P}(Z \geq t)dt + \mathbb{P}(Z \geq \sqrt{\frac{1}{8}}) .$$

Let us upper bound $\mathbb{P}(Z \geq t)$ for any $0 \leq t \leq \sqrt{\frac{1}{8}}$ by applying Lemma 4.16

$$\begin{aligned}\mathbb{P}(Z \geq t) &\leq \mathbb{P}\left(\chi^2(d) \leq d\left[1 - t\right]^2\right) \\ &\leq \mathbb{P}\left(\chi^2(d) \leq d - 2\sqrt{d}\sqrt{dt^2/2}\right) \leq \exp\left(-\frac{dt^2}{2}\right) ,\end{aligned}$$

since $t \leq 2 - \sqrt{2}$. Gathering this upper bound with the previous inequality yields

$$\begin{aligned} \mathbb{E}(Z) &\leq \exp\left(-\frac{d}{16}\right) + \int_0^{+\infty} \exp\left(-\frac{dt^2}{2}\right) dt \\ &\leq \exp\left(-\frac{d}{16}\right) + \sqrt{\frac{\pi}{2d}}\,. \end{aligned}$$

Thus, we obtain $\mathbb{E}\left(\sqrt{\chi^2(d)}\right) \geq \sqrt{d} - \sqrt{d}\exp(-d/16) - \sqrt{\pi/2}$. Combining this lower bound with (4.64) allows to conclude. □

Chapter 5

Adaptive estimation of covariance matrices via Cholesky decomposition

Abstract. This chapter considers the estimation of a covariance matrix of p variables from n observations. We introduce a novel model selection procedure that is based on the Cholesky decomposition of the inverse of the covariance. Two different settings are investigated. In the first one, the variables are assumed to have a natural ordering and the Cholesky factor T is then supposed to be approximately banded, i.e. the entries of T that are far from the diagonal are close to 0. In the second setting, the variables are not supposed anymore to be ordered, but the matrix is assumed to be approximately sparse. This hypothesis is closely connected to the notion of ordered Gaussian graphical model. In both settings, we are able to prove non-asymptotic oracle-type inequalities with respect to the Kullback-Leibler entropy. Moreover, we derive various minimax rates of estimation that are achieved by our procedure. Other types of loss functions are also considered.

5.1 Introduction

The problem of estimating large covariance matrices has recently attracted a lot of attention. On the one hand, there is an inflation of high-dimensional data in many scientific areas: gene arrays, functional magnetic resonance imaging (fMRI), image classification, and climate studies. On the other hand, many data analysis tools require an estimation of the covariance matrix Σ. This is for instance the case for principal component analysis (PCA), for linear discriminant analysis (LDA), or for establishing independences or conditional independences between the variables. It is known for a long time that the simplest estimator, the sample covariance matrix performs poorly when the size of the vector p is larger than the number of observations n (see for instance Johnstone [Joh01]).

Depending on the objectives of the analysis and on the applications, different approaches are used for estimating high-dimensional covariance matrices. Indeed, if one wants to perform PCA or to establish independences between the covariates, then it is advised to estimate directly the covariance matrix Σ. In contrast, performing LDA further relies on the inverse of the covariance matrix. In the sequel, we call this matrix the precision matrix and note it Ω. Sparse precision matrices are also of interest because of their connection with graphical models and conditional independence. The pattern of zero in Ω indeed corresponds to the graph structure of the distribution (see for instance Lauritzen [Lau96] Sect.5.1.3).

When directly estimating the covariance matrix Σ, most of the methods amount to regularizing the empirical covariance matrix. Let us mention the work of Ledoit and Wolf [LW04] who propose to replace the sample covariance with its linear combination with the identity matrix. However, these shrinkage methods are known to provide an inconsistent estimation of the eigenvectors [JL04]. Applying recent results on random matrix theory, El Karoui [EK08] and Bickel and Levina [BL08a] have studied thresholding estimators of Σ. The resulting estimator is sparse and is proved (for instance [BL08a]) to be consistent with respect to the operator norm under mild conditions as long as $\log(p)/n$ goes to 0.

These results are particularly of interest for performing PCA since they imply a consistent estimation of the eigenvalues and the eigenvectors. Observe that all these methods are invariant under permutation of the variables. Yet, in many applications (for instance times series, spectroscopy, climate data), there exists a natural ordering in the data. In such a case, one may use other procedures and obtain faster rates of convergence. Among other, Furrer and Bentgsson [FB07] and Bickel and Levina [BL08b] use banded or tapering estimators. Again, the consistency of such estimators is proved. Let us mention that all these methods share an attractive computational cost. We refer to the introduction of [BL08a] for a more complete review.

The estimation procedures of the precision matrix Ω fall into three categories depending whether there exists an ordering on the variables and to what extent this ordering is important. If there is not such an ordering, d'Aspremont *et al.* [BEGd08] and Yuan and Lin [YL07] have adopted a penalized likelihood approach by applying a l_1 penalty to the entries of the precision matrix. It has also been discussed by Rothman *et al.* [RBLZ08] and Friedman *et al.* [FHT08] and extended by Fan and Lam *et al.* [LF07] to other penalization methods. These estimators are known to converge with respect to the Frobenius norm (for instance [RBLZ08]) when the underlying precision matrix is sparse enough.

When there is a natural ordering on the covariates, the regularization is introduced via the Cholesky decomposition:

$$\Omega = T^* S^{-1} T \ ,$$

where T is a lower triangular matrix with one in the diagonal and S is a diagonal matrix with positive entries. The elements of the i-th row may be interpreted as regression coefficient of i-th component given its predecessors. This will be further explained in Section 5.1.1. For time series or spectroscopy data, it is more likely that the relevant covariates for this regression of the i-th component are its closest predecessors. In other word, it is expected that the matrix T is approximately banded. With this in mind, Wu and Pourahmadi [WP03] introduce a k-banded estimator of the matrix T by smoothing along the first k subdiagonals and setting the rest to 0. The choice of k is made by applying AIC (Akaike [Aka73]). They prove element-wise consistency of their estimator but did not provide any high-dimensional result with respect to a loss function such as Kullback or Frobenius. Bickel and Levina [BL08b] also consider k-banded estimator of T and are able to prove rates of convergence in the matrix operator norm. Moreover, they introduce a cross-validation approach for choosing a suitable k, but they do not prove that the selection method achieves adaptiveness. More recently, Levina *et al.* [LRZ08] propose a new banding procedure based on a nested lasso penalty. Unlike the previous methods, they allow the number $k = k_j$ used for banding to depend on the line j of T. They do not state any theoretical result, but they exhibit numerical evidence of its efficiency. In the sequel, we call the issue of estimating Ω by banding the matrix T the *banding problem*.

Between the first approach based on precision matrix regularization and the second one which relies on banding the Cholesky factor, there exists a third one which is not permutation invariant, but does not assume that the matrix T is approximately banded. It consists in approximating T by a sparse lower triangular matrix. When is it interesting to adopt this approach? If we consider a directed graphical model whose graph is sparse and compatible with the ordering of the variables, then the Cholesky factor T is sparse, since its pattern of zero is related to the directed acyclic graph (DAG) of the directed graphical model associated to this ordering (see for instance [Lau96]). More generally, it may be worth using this strategy even if one does not have a clue on the precision matrix Ω. Admittedly, it is not completely satisfying to apply a method that depends on the ordering of the variables when we do not know a *good* ordering. There are indeed examples of sparse precision matrices Ω such that for a *bad* ordering, the Cholesky factor is not sparse at all (see [RBLZ08] Sect.4). Nevertheless, we see two reasons for advocating an approach based on sparse Cholesky factor estimation. On the one hand, there also exists examples of sparse Cholesky factor T such that the precision matrix Ω is not sparse at all. Consider for instance a matrix T which is zero except on the diagonal and on the last line. Our point is that sparse precision matrices and sparse Cholesky factors have different approximation capacities, but it remains still unclear which one should be favored. On the other hand, most of procedures based on the estimation of T are computationally faster than their counterpart based on the estimation of Ω. This is due to the decomposition of the likelihood into p independent terms explained in Section 5.2.

In the sequel, we call the issue of estimating T in the class of sparse lower triangular matrices the *complete graph selection* problem by analogy to the complete variable estimation problem in regression

problems. In this setting, Huang *et al.* [HLPL06] proposed adding an l_1 penalty on the elements of T. More recently, Lam and Fan [LF07] have extended their method to other types of penalty and have proved its consistency in the Frobenius norm if the matrix T is exactly sparse. To finish, let us mention that Wagaman and Levina [WE08] have developed a data-driven method based on the isomap algorithm for picking a suitable ordering on the variables.

In this chapter, we consider both the banding problem and the complete graph selection problem. We introduce a general l_0 penalization method based on maximum likelihood for estimating the matrices T and S. We exhibit a non-asymptotic oracle inequality with respect to the Kullback loss *without* any assumption on the target Ω. For the adaptive banding issue, our method is also shown to achieve simultaneously the minimax rates of estimation over various sets that we call ellipsoids. We also compute asymptotic rates of convergence in the Frobenius norm. Contrary to the l_1 penalization methods, we explicitly provide the constant for tuning the penalty. Finally, the method is shown to be computationally efficient.

For complete graph selection, we prove that our estimator is minimax adaptive to the unknown sparsity of the matrix T from a non-asymptotic point of view. We also provide asymptotic rates of convergence with respect to the Frobenius norm. Moreover, our method is flexible and allows to integrate some prior knowledge on the graph. However, the obtained procedure is computationally intensive which makes it infeasible for p larger than 20.

Since data analysis methods like LDA are based on likelihood we find it more relevant to obtain rates of convergence with respect to the Kullback-Leibler loss than Frobenius rates of convergence. Moreover, considering Kullback loss allows us to obtain rates of convergence which are free of hidden dependency on parameter such as the largest eigenvalue of Σ. In this sense, we argue that this loss function is more natural for the statistical problem we consider.

The chapter is organized as follows. In Section 5.2, we describe the procedure and provide an algorithm for computing more efficiently the matrix $\widetilde{\Omega}$. In Section 5.3, we provide a bias-variance decomposition for the Kullback risk of the estimator $\widehat{\Omega}_m$. Moreover, we state the main result of the chapter, namely a general non-asymptotic oracle type inequality for the risk of $\widetilde{\Omega}$. In Section 5.4, we specify our result to the problem of adaptive banding. Moreover, we prove that our so-defined estimator is minimax adaptive to the decay of the off-diagonal coefficients of the matrix T. Asymptotic rates of convergence with respect to the Frobenius norm are also provided. In Section 5.5, we investigate the complete graph selection issue. We first derive a non-asymptotic oracle inequality and then derive that our procedure is minimax adaptive to the unknown sparsity of the Cholesky factor T. As previously, we provide asymptotic rates of convergence with respect to the Frobenius loss function. We make a few concluding remarks in Section 5.6, while the proofs are postponed to Section 5.7.

5.1.1 Notations

In this subsection, we briefly describe the approach underlying our model selection procedure and introduce the main notations.

We consider the estimation of the vector $X = (X_i)_{1\leq i\leq p}$ of size p which follows a centered normal distribution with covariance matrix Σ. We are given n independent observations of the vector X. In the sequel, we note $\mathbf{X}$ the $n \times p$ matrix of the observations. Moreover, for any $1 \leq i \leq p$ and any subset A of $\{1, \ldots, p-1\}$, $\mathbf{X}_i$ and $\mathbf{X}_A$ respectively refer to the vector of the n observations of X_i and to the $n \times |A|$ matrix of the observations of $(X_i)_{i\in A}$.

We recall that precision matrix Ω uniquely decomposes as $\Omega = T^* S T$ where T is a lower triangular matrix with unit diagonal and S is a diagonal matrix. Let us first emphasize the connection between the modified Cholesky factor T and conditional regressions. In the sequel, for any i between 1 and $p-1$we note t_i the vector of size $i-1$ made of the $i-1$-th first elements of the ith-line of T. By convention t_1 is the vector of null size. Besides, we note s_i the diagonal element of i-th diagonal element of the matrix S. Let us define the vector $\epsilon = (\epsilon_i)_{1\leq i\leq p}$ of size p as $\epsilon := TX$. By standard Gaussian properties, the covariance matrix of ϵ is S. Since the diagonal of T is one, it follows that for any $1 \leq i \leq p$

$$X[i] = \sum_{j=1}^{i-1} -t_i[j]X[j] + \epsilon_i \ , \tag{5.1}$$

where $\mathrm{var}(\epsilon_i) = s_i$ and the $(\epsilon_i)_{1\leq i\leq p}$ are independent.

In order to assess the performances, we use the Kullback divergence. In the sequel, $\mathcal{K}(\Omega;\Omega')$ stands for the Kullback divergence between the centered normal distribution with covariance Ω^{-1} and the centered normal distribution with covariance Ω'^{-1}. We shall also sometimes assess the performance of our procedure using the Frobenius norm and the l_2 operator norm. This is why we respectively define $\|A\|_F^2 := \sum_{i,j} A[i,j]^2$ and $\|A\|$ as the Frobenius norm and the l_2 operator norm of the matrix A. Finally, $\varphi_{\max}(\Omega)$ stands for the largest eigenvalue of Ω.

5.2 Description of the procedure

In this section, we explain why the computation of the parametric estimators $\widehat{\Omega}_m$ and the penalized estimator $\widetilde{\Omega}$ may be divided in p independent sub-problems.

In the sequel, for any i between 1 and p, m_i stands for a subset of $\{1,\ldots,i-1\}$. By convention, $m_1 = \emptyset$. Besides, we call any set m of the form $m = m_1 \times m_2 \times \ldots \times m_p$ a model. For any i between 1 and p, $\mathcal{M}_i$ refers to a collection subsets of $\{1,\ldots,i-1\}$ and we call $\mathcal{M} := \mathcal{M}_1 \times \ldots \times \mathcal{M}_p$ a collection of models. The choice of the collection $\mathcal{M}$ depends on the estimation problem we consider. We specify the collection for the banding problem and the complete graph selection problem in Sections 5.4 and 5.5.

For any model $m \in \mathcal{M}$, we define $\mathcal{T}_m$ as the affine space of lower triangular triangular matrices T such for any i between 1 and p, the support of t_i is included in m_i. We note $Diag(p)$ the set of all diagonal matrices with positive entries on the diagonal. The matrices $\widehat{T}_m$ and $\widehat{S}_m$ are then defined as the maximum likelihood estimators of T and S

$$\left(\widehat{T}_m, \widehat{S}_m\right) = \arg \min_{T'\in\mathcal{T}_m,\ S'\in \mathrm{Diag}(p)} \mathcal{L}_n(T,S) := \frac{1}{2} tr\left[T^* S^{-1} T \overline{\mathbf{X}^*\mathbf{X}}\right] + \frac{1}{2}\log|S| \tag{5.2}$$

In fact, $\mathcal{L}_n(T,S)$ stands for the opposite of the log-likelihood. Hence, the estimated precision matrix is $\widehat{\Omega}_m = \widehat{T}_m^* \widehat{S}_m^{-1} \widehat{T}_m$.

Our objective is to select a model $\widehat{m}$ among the collection of models $\mathcal{M}$. We achieve it through penalization. For any $1 \leq i \leq p$, $pen_i : \mathcal{M}_i \to \mathbb{R}^+$ is a positive function that we shall explicitly define later. The penalty function $pen : \mathcal{M} \to \mathbb{R}^+$ is defined as $\mathrm{pen}(m) = \sum_{i=1}^p pen_i(m_i)$. Then, we select a model $\widehat{m}$ that minimizes the following criterion

$$\widehat{m} := \arg\min_{m\in\mathcal{M}} 2\mathcal{L}_n(\widehat{T}_m, \widehat{S}_m) + \mathrm{pen}(m) = \arg\min_{m\in\mathcal{M}} tr\left[\widehat{\Omega}_m \overline{\mathbf{X}^*\mathbf{X}}\right] - \log|\widehat{\Omega}_m| + \mathrm{pen}(m)$$

For short, we write $\widetilde{\Omega} := \widehat{\Omega}_{\widehat{m}}$, $\widetilde{S} := \widehat{S}_{\widehat{m}}$, and $\widetilde{T} = \widehat{T}_{\widehat{m}}$.

As mentioned earlier, the idea underlying the use of the matrices T and S lies in the regression models (5.1). Indeed, these regressions naturally appear when deriving the negative log-likelihood (5.2):

$$2\mathcal{L}_n(T',S') = \sum_{i=1}^p s_i'^{-1} \|\mathbf{X}_i + \mathbf{X}_{<i}(t_i')^*\|_n^2 + \log(s_i') \ ,$$

where $\|.\|_n$ stands for the canonical norm in $\mathbb{R}^n$ divided by $\sqrt{n}$. By definition of $\widehat{T}_m$ and $\widehat{S}_m$, we easily derive that the i-th row vector $\widehat{t}_{i,m_i}$ of $\widehat{T}_m$ and the i-th diagonal element $\widehat{s}_{i,m_i}$ of $\widehat{S}_m$ respectively equal

$$\widehat{t}_{i,m_i} = \arg\min_{\mathrm{supp}(t_i')\subset m_i} \|\mathbf{X}_i + \mathbf{X}_{<i}(t_i')^*\|_n^2 \quad \text{and} \quad \widehat{s}_{i,m_i}^2 = \|\mathbf{X}_i + \mathbf{X}_{<i}\widehat{t}_{i,m_i}^*\|_n^2 \ , \tag{5.3}$$

for any $1 \leq i \leq p$. Here, $\mathrm{supp}(t_i')$ stands for the support of t_i'. Hence, the row vector $\widehat{t}_{i,m_i}$ is the least square estimator of t_i in the regression model (5.1) and $\widehat{s}_{i,m_i}$ is the empirical conditional variance of X_i given X_{m_i}. There are two main consequences: first, Expression (5.3) emphasizes the connection between covariance estimation and linear regression in a Gaussian design. Second, it highly simplifies the computational cost of our procedure. Indeed, the log-likelihood $-\mathcal{L}_n(\widehat{T}_m, \widehat{S}_m)$ now writes

$$\mathcal{L}_n\left(\widehat{T}_m, \widehat{S}_m\right) = \frac{1}{2}\sum_{i=1}^p \left[\log\left(\widehat{s}_{i,m_i}\right) + 1\right] \ .$$

and it follows that $\widehat{m}_i = \arg\min_{m_i \in \mathcal{M}_i} \log(\widehat{s}_{i,m_i}) + pen_i(m_i)$. This is why we suggest to compute $\widehat{m}$ and $\widehat{\Omega}$ as follows. Assume we are given a collection $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_p)$ and a penalty function $(pen_1(.), \ldots, pen_p(.))$.

Algorithm 5.1. *Computation of $\widehat{m}$ and $\widetilde{\Omega}$.*

1. *For i going from 1 to p,*
   - *Compute $\widehat{s}_{i,m_i}$ for each model $m_i \in \mathcal{M}_i$.*
   - *Take $\widehat{m}_i = \arg\min_{m_i \in \mathcal{M}_i} \log(\widehat{s}_{i,m_i}) + pen(m_i)$.*
2. *Set $\widehat{m} = (\widehat{m}_1, \ldots, \widehat{m}_p)$ and built $(\widetilde{T}, \widetilde{S})$ by gathering the estimators $(\widehat{t}_{i,\widehat{m}_i}, \widehat{s}_{i,\widehat{m}_i})$.*
3. *Take $\widetilde{\Omega} = \widetilde{T}\widetilde{S}^{-1}\widetilde{T}$.*

In order to select $\widehat{m}$, one needs to compute all $\widehat{s}_{i,m_i}$ for any $i \in \{1, \ldots, p\}$ and any model $m_i \in \mathcal{M}_i$. Hence, the complexity of the procedure is proportional to $\sum_{i=1}^{p} |\mathcal{M}_i|$. We further discuss computational issue in Section 5.6.

5.3 Risk analysis

In this section, we first provide a bias-variance decomposition for the Kullback risk of the parametric estimator $\widehat{\Omega}$. Afterwards, we state a general non-asymptotic risk bound for $\widetilde{\Omega}$.

5.3.1 Parametric estimation

Let m be model in $\mathcal{M}$. Let us define the matrix Ω_m as the best approximation of Ω that corresponds to the model m. The matrices T_m and S_m are defined as the minimizers in $\mathcal{T}_m$ and $Diag(p)$ of the Kullback loss with Ω

$$(T_m, S_m) := \arg \min_{T' \in \mathcal{T}_m,\ S' \in \mathrm{Diag}(p)} \mathcal{K}\left(\Omega; T'^{*} S'^{-1} T'\right)$$

We note $\Omega_m = T_m^* S_m^{-1} T_m$.

We define the conditional Kullback-Leibler divergence of the distribution of X_i given $X_{<i}$ by

$$\mathcal{K}(t_i, s_i; t_i', s_i') := \mathbb{E}\left\{\mathcal{K}\left[\mathbb{P}_{t_i,s_i}(X_i|X_{<i}); \mathbb{P}_{t_i',s_i'}(X_i|X_{<i})\right]\right\} \ , \tag{5.4}$$

where $\mathbb{P}_{t_i,s_i}(X_i|X_{<i})$ stands for the conditional distribution of X_i given $X_{<i}$ with parameters (t_i, s_i). Applying the chain rule, we obtain that $\mathcal{K}(\Omega; \Omega') = \sum_{i=1}^{p} \mathcal{K}(t_i, s_i; t_i', s_i')$. Consequently, we analyse the Kullback risk $\mathbb{E}[\mathcal{K}(\Omega; \widehat{\Omega}_m)]$ by controlling each conditional risk $\mathbb{E}\left[\mathcal{K}(t_i, s_i; \widehat{t}_{i,m_i}, \widehat{s}_{i,m_i})\right]$. Let us define t_{i,m_i} and s_{i,m_i} as the *projections* of (t_i, s_i) on the space associated to the model m_i with respect to the Kullback divergence $\mathcal{K}(t_i, s_i; ., .)$. In other words, t_{i,m_i} and s_{i,m_i} satisfy

$$t_{i,m_i} = \arg \min_{\mathrm{supp}(t_i') \subset m_i} \mathbb{E}\left[\left(X_i + X_{<i}(t_i')^*\right)^2\right] \quad \text{and} \quad s_{i,m_i} = \mathrm{var}\left(X_i|X_{<i}\right) \ .$$

Applying the chain rule, we check that t_{i,m_i} corresponds to $(i-1)$-th first elements of the i-th line of T_m and s_{i,m_i} is the i-th diagonal element of S_m. Thanks to the previous property, we derive a bias-variance decomposition for the Kullback risk $\mathbb{E}\left[\mathcal{K}(t_i, s_i; \widehat{s}_{i,m_i}, \widehat{s}_{i,m_i})\right]$.

Proposition 5.2. *Assume that $|m_i|$ is smaller than $n-2$. The Kullback risk of $(\widehat{t}_{i,m_i}, \widehat{s}_{i,m_i})$ decomposes as follows*

$$\mathbb{E}\left[\mathcal{K}\left(t_i, s_i; \widehat{t}_{i,m_i}, \widehat{s}_{i,m_i}\right)\right] = \mathcal{K}\left(t_i, s_i; t_{i,m_i}, s_{i,m_i}\right) + R_{n,|m_i|} \ , \tag{5.5}$$

where $R_{n,d}$ is defined as

$$R_{n,d} := \frac{d+1}{n-d-2} + \frac{d(d+1)}{2(n-d-1)(n-d-2)} + \frac{1}{2}\left[\Psi\left(n-d\right) + \log\left(1 - \frac{d}{n}\right)\right] \ ,$$

and $\Psi(n-d) := \mathbb{E}\left[\log\left(\frac{\chi^2(n-d)}{n-d}\right)\right]$. *Besides,* $R_{n,d}$ *is bounded as follows*

$$\frac{d+1}{2(n-d-2)} \leq R_{n,d} \leq \frac{d+1}{n-d-2} + \frac{1}{2}\left[\frac{d+1}{n-d-2}\right]^2$$
$$\text{and } R_{n,d} = \frac{d+1}{2(n-d-2)} + \mathcal{O}\left(\frac{d+1}{n}\right)^2 .$$

An explicit expression of $R_{n,d}$ is provided in the proof. Applying the chain rule, we then derive a bias-variance decomposition for the maximum likelihood estimator $\widehat{\Omega}_m$.

Corrolary 5.3. *Let* $m = (m_1, \ldots, m_p)$ *be a model such that the size* $|m_i|$ *of each submodel is smaller than* $n-2$. *Then, the Kullback risk of the maximum likelihood estimator* $\widehat{\Omega}_m$ *decomposes into*

$$\mathbb{E}\left[\mathcal{K}\left(\Omega; \widehat{\Omega}_m\right)\right] = \mathcal{K}\left(\Omega; \Omega_m\right) + \sum_{i=1}^{p} R_{n,|m_i|} .$$

If the size $|m_i|$ of each submodels is small with respect to n, the variance term is of the order $\sum_{i=1}^{p} \frac{|m_i+1|}{2(n-|m_i|-2)}$. For other loss functions such as the Frobenius norm or the l_2 operator norm between Ω and $\widehat{\Omega}_m$, there is no such bias-variance decomposition with a variance term that does not depend on the target.

5.3.2 Main result

In this subsection, we state a general non-asymptotic oracle inequality for the Kullback-Leibler risk of the estimator $\widetilde{\Omega}$. We first need to define a penalty function pen(.) that takes into account the complexity of the collection of models. Given a collection $\mathcal{M}$, we introduce the functions $H_i(.)$ as a measure of complexity.

Definition 5.4. For any integer i between 2 and p, the function $H_i(.)$ is defined as

$$H_i(d) := \frac{1}{d}\log\left[\text{Card}\left(\{m \in \mathcal{M}_i,\ |m_i| = d\}\right)\right] ,$$

where d is any integer larger or equal to 1. Besides, $H_i(0)$ is set to 0 for any i between 1 and p.

These functions are analogous to the complexity measures introduced in [BM07] Sect.1.3 or in Section 4.3.2. We shall obtain an oracle inequality under the following assumption.

$(\mathbb{H}_{K,\eta})$: Given $K > 1$ and $\eta > 0$, the collection $\mathcal{M}$ and the number η satisfy

$$\forall\ 2 \leq i \leq p\ , \forall m_i \in \mathcal{M}_i\ , \quad \frac{\left[1+\sqrt{2H_i(|m_i|)}\right]^2 |m_i|}{n-|m_i|} \leq \eta < \eta(K)\ , \tag{5.6}$$

where $\eta(K)$ is defined as in Equation (4.11) in Chapter 4. The function $\eta(.)$ is positive and increases to one. This condition requires that the size of the collection is not too large. Assumption $(\mathbb{H}_{K,\eta})$ is similar to the assumption made in Section 4.3.2 for obtaining an oracle inequality in the linear regression with Gaussian design framework. We further discuss $(\mathbb{H}_{K,\eta})$ in Sections 5.4 and 5.5 when considering the particular problems of ordered and complete variable selection.

Theorem 5.5. *Let* $K > 1$ *and let* $\eta < \eta(K)$. *Assume that* n *is larger than some quantity* $n_0(K)$ *only depending on* K *and that the collection* $\mathcal{M}$ *satisfies* $(\mathbb{H}_{K,\eta})$. *If the penalty pen*(.) *is lower bounded as follows*

$$pen_i(m_i) \geq K\frac{|m_i|}{n-|m_i|}\left(1+\sqrt{2H_i(|m_i|)}\right)^2 \quad \text{for any } 1 \leq i \leq p \text{ and any } m_i \in \mathcal{M}_i\ , \tag{5.7}$$

then the risk of $\widetilde{\Omega}$ *is upper bounded by*

$$\mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}\right)\right] \leq L(K,\eta) \inf_{m\in\mathcal{M}}\left[\mathcal{K}\left(\Omega; \Omega_m\right) + pen(m) + \frac{p}{n}\right] + \tau_n\ , \tag{5.8}$$

where τ_n *is defined as*

$$\tau_n = \tau\left(\Omega, K, \eta, n, p\right) := L(K,\eta) n^{5/2}\left[p + \mathcal{K}(\Omega; I_p)\right] \exp\left[-nL_2(K,\eta)\right] \ ,$$

and $L_2(K,\eta)$ *is positive. Here,* I_p *stands for the identity matrix of size* p*.*

Comments:

- This theorem tells us $\widetilde{\Omega}$ performs almost as well as the best trade-off between the bias term $\mathcal{K}(\Omega;\Omega_m)$ and the penalty term $\mathrm{pen}(m)$. Observe that the term p/n is unavoidable since it is of the same order as the variance term for the null model by Corollary 5.3. The error term τ_n is considered as negligible since converges exponentially fast to 0 with n.
- The result is non-asymptotic and holds for arbitrary large p as longs n is larger than the quantity $n_0(K)$ (independent of p). There is no hidden dependency on p except in the complexity functions $H_i(.)$ and Assumption $(\mathbb{H}_{K,\eta})$ that we shall discuss for particular cases in Sections 5.4.1 and 5.5.1.
- Observe that we are not performing any assumption on the true precision matrix Ω except that it is invertible. In particular, we do not assume that it is sparse and we give a rate of convergence that only depends on a bias variance trade-off. Besides, there is no hidden constant that depends on Ω (except for τ_n).
- Finally, the penalty introduced in this theorem only depends on the collection $\mathcal{M}$ and on a number $K > 1$. As Condition (5.7) is only a lower bound, one may give a Bayesian flavor to the penalty by integrating some prior knowledge. Moreover, one can choose the parameter K depending on how conservative one wants the procedure to be. We further discuss the practical choice of K in Sections 5.4 and 5.5. In any case, the main point is that we do not need any additional method to calibrate the penalty.

In Section 5.4 and 5.5, we apply the method for performing adaptive banding and complete estimation of the Cholesky factor.

5.4 Adaptive banding

5.4.1 Oracle inequalities

Let us fix $K > 1$, η smaller than $\eta(K)$ and let us choose an integer d smaller than $n\frac{\eta}{1+\eta}$. Here, d stands for the largest dimension of the models m_i. For any $2 \leq i \leq p$, we consider the ordered collections

$$\mathcal{M}^d_{i,\mathrm{ord}} := \{\emptyset, \{1\}, \{1,2\}, \ldots, \{1 \wedge (i-d), \ldots, i-1\}\} \ ,$$

and $\mathcal{M}^d_{1,\mathrm{ord}} := \{\emptyset\}$. For any $1 \leq i \leq p$ and any model m_i in $\mathcal{M}^d_{i,\mathrm{ord}}$ we fix the penalty

$$pen_i(m_i) = K\frac{|m_i|}{n - |m_i|} \ . \tag{5.9}$$

We write $\widetilde{\Omega}^d_{\mathrm{ord}}$ for the estimator $\widetilde{\Omega}$ defined with the collection $\mathcal{M}^d_{\mathrm{ord}}$ and the penalty (5.9).

Corrolary 5.6. *If* n *is larger than some quantity* $n_0(K)$*, then*

$$\mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{\mathrm{ord}}\right)\right] \leq L(K,\eta) \inf_{m \in \mathcal{M}^d_{\mathrm{ord}}} \mathbb{E}\left[\mathcal{K}\left(\Omega; \widehat{\Omega}_m\right)\right] + \tau_n\left(\Omega, K, \eta, n, p\right) \ . \tag{5.10}$$

Comments:

- The term τ_n is defined in Theorem 5.5 and is considered as negligible since it converges to 0 exponentially fast towards 0. Hence, the penalized estimator $\widetilde{\Omega}$ achieves an oracle inequality *without* any assumption on the target Ω.
- This oracle inequality is non-asymptotic and holds any p and any n larger than $n_0(K)$. Moreover, by choosing a constant K large enough, one can consider a maximal dimension of model d up to the order of n, because $\eta(K)$ converges to one when K increases.

Let us briefly discuss the practical choice of the parameters K and d. Setting K to 2 gives a criterion close to $AICc$ (see for instance [MT98]). Besides, it is justified theoretically by Proposition 4.4. A choice of $K = 3$ may be advised if one wants a more conservative procedure. We have stated Corollary 5.6 for collection $\mathcal{M}_i$ of size smaller than $\frac{\eta}{1+\eta}n$. In practice, taking the size $n/2$ yields rather good results even if it is not completely ensured by the theory.

The computational cost of our method is rather small since the complexity is the same as p times the complexity of an ordered variable selection in a classical regression framework. From numerical comparisons, it seems to be slightly faster than the methods of Bickel and Levina [BL08b] and Levina *et al.* [LRZ08] which require cross-validation type strategies.

5.4.2 Adaptiveness with respect to ellipsoids

We now state that the estimator $\widetilde{\Omega}^d_{\rm ord}$ is simultaneously minimax over a large class of sets that we call ellipsoids.

Definition 5.7. Let $(a_i)_{1\le i\le p-1}$ be non-increasing sequence of positive numbers such that $a_1 = 1$ and let R be a positive number. Then, the set $\mathcal{E}(a, R, p)$ is made of all the non-singular matrices $\Omega = T^* S^{-1} T$ where S is in $Diag(p)$ and T is a lower triangular matrix with unit diagonal that satisfies the following property

$$\sum_{j=1}^{i-1} \frac{T[i, i-j]^2}{a_j^2} \le R^2 \ , \qquad \forall\ 2 \le i \le p \ . \tag{5.11}$$

By convention, we set $a_p = 0$. The sequence (a_i) measures the rate of decay of each line of T when one moves away the diagonal. Observe that in this definition, every line of T decreases the same rate. To the price of more technicity, we may also consider different rates of decay for each line of T.

In order to compute the minimax rates of estimation over ellipsoids, we first need to consider a lower bound over the sets $\mathcal{T}_{\rm ord}[k_1, \dots, k_p, r]$ and $\mathcal{U}_{\rm ord}[k_1, \dots, k_p, r]$.

Definition 5.8. Let $(k_1, \dots, k_p)$ be an element of $\mathcal{M}^\infty_{ord}$ and let r be a positive number. We respectively define the sets $\mathcal{T}_{\rm ord}[k_1, \dots, k_p, r]$ and $\mathcal{U}_{\rm ord}[k_1, \dots, k_p, r]$ as

$$\mathcal{T}_{\rm ord}[k_1, \dots, k_p, r] := \left\{ T \in Trig(p) \text{ s.t. } \forall 2 \le i \le p, T[i,j] = \begin{array}{ccc} 0 & \text{if} & 1 \le j \le i - k_i - 1 \\ 0 \text{ or } r & \text{if} & i - k_i \le j \le i-1 \end{array} \right\} , \tag{5.12}$$

$$\mathcal{U}_{\rm ord}[k_1, \dots, k_p, r] := \left\{ T^* S^{-1} T \ , \ \ T \in \mathcal{T}_{\rm ord}[k_1, \dots, k_p, r] \text{ and } S \in Diag(p) \right\} \ . \tag{5.13}$$

The set $\mathcal{T}_{\rm ord}[k_1, \dots, k_p, r]$ contains lower triangular matrices with unit diagonal such that for each line i between 2 and p, the support of the vector $(T[i,j])_{1\le j\le i-1}$ is included in $\{i - k_i, i - k_i + 1, \dots, i-1\}$. We are able to lower bound the minimax rates of estimation over $\mathcal{U}_{ord}[(k_1, \dots, k_p), r]$.

Proposition 5.9. *$k := 1 \vee \max_{1\le i\le p} k_i$ is smaller than $\sqrt{n}$.*

$$\inf_{\widehat{\Omega}} \sup_{\Omega \in \mathcal{U}_{\rm ord}[(k_1,\dots,k_p),r]} \mathbb{E}\left[\mathcal{K}\left(\Omega; \widehat{\Omega}\right)\right] \ge L \left[\sum_{i=2}^{p} k_i + p \right] \left(r^2 \wedge \frac{1}{n} \right)$$

These minimax rates of estimation are not really surprising, since they correspond to the minimax rates of estimation of p different parametric regression problems whose minimax rates is known to be of the order $k_i(r^2 \wedge 1/n$. We refer for instance to [Mas07] Prop. 4.8. Moreover, the term p/n is due to the diagonal matrices S in $\Omega = T^* S^{-1} T$. We believe that the assumption k is smaller than $\sqrt{n}$ is not necessary but we do not know how to remove it.

We shall restrict ourselves to covariance matrices with eigenvalues that lie in a compact when considering the ellipsoid $\mathcal{E}(a, R, p)$

$$\mathcal{B}_{\rm op}(\gamma) := \left\{ \varphi_{\min}(\Omega) \ge \frac{1}{\gamma} \text{ and } \varphi_{\max}(\Omega) \le \gamma \right\} \ . \tag{5.14}$$

Proposition 5.10. *For any ellipsoid $\mathcal{E}(a,R,p)$, the minimax rates of estimation is lower bounded by*

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{E}(a,R,p)} \mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] \geq Lp \sup_{k=1,\dots,\lfloor\sqrt{n}\rfloor} \left(R^2 a_k^2 \wedge \frac{k+1}{n}\right) . \tag{5.15}$$

Let us consider the estimator $\widetilde{\Omega}^d_{ord}$ defined in Section 5.4.1 with $d=\lfloor n\frac{\eta}{1+\eta}\rfloor$ and the penalty (5.9). We also fix $\gamma>0$. If the sequence (a_i) and R also satisfy $R^2\geq\frac{1}{n}$ and $a^2_{\lfloor\sqrt{n}\rfloor\wedge p}\leq\frac{1}{R^2\sqrt{n}}$, then

$$\sup_{\Omega\in\mathcal{E}(a,R,p)\cap\mathcal{B}_{op}(\gamma)} \mathbb{E}\left[\mathcal{K}\left(\Omega;\widetilde{\Omega}^d_{\mathrm{Co}}\right)\right] \leq L(K,\eta,\beta,\gamma)\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{E}(a,R,p)\cap\mathcal{B}_{op}(\gamma)} \mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] , \tag{5.16}$$

if n is larger than $n_0(K)$

Observe that the minimax rates of convergence over $\mathcal{E}(a,R,p)$ in the lower bound (5.15) is similar to the one obtained for classical ellipsoids in the Gaussian fixed design regression setting (see for instance [Mas07] Th. 4.9). We conclude from the second result that our estimator $\widetilde{\Omega}^d_{\mathrm{ord}}$ is minimax adaptive to the ellipsoids that are not degenerate (i.e. $R^2\geq 1/n$) and whose rates (a_i) does not converge too slowly towards zero (i.e. $a^2_{\lfloor\sqrt{n}\rfloor\wedge p}\leq\frac{1}{R^2\sqrt{n}}$). Note that all the sequences (a_i) such that $a_i^2\leq R^2/i$ satisfy the last assumption. However, the estimator $\widetilde{\Omega}^d_{\mathrm{ord}}$ is not adaptive to the parameter γ since the constant L in (5.16) depends on γ. This is not really surprising. Indeed, the oracle inequality (5.10) is expressed in terms of the Kullback loss while the ellipsoids are defined in terms of the entries of T. If we would have considered the minimax rates of estimation over sets analogous to $\mathcal{E}(a,R,p)$ but defined in terms of the decay of the Kullback bias, then we would have obtained minimax adaptiveness without any condition on the eigenvalues.

We are also able to prove asymptotic rates of convergence and asymptotic minimax properties with respect to the Frobenius loss function. For any $s>0$, we define the ellipsoid $\mathcal{E}'(s,p,R)$ as the ellipsoid $\mathcal{E}(a,R,p)$ with the sequence $(a_i)_{1\leq i\leq p-1}:=i^{-s}$.

Corrolary 5.11. *If $\sum_{i=1}^{p_n}k_i+p_n=o(n)$ and $k:=1\vee\max_{1\leq i\leq p}k_i$ is smaller than $\sqrt{n}$ then Uniformly over the set $\mathcal{U}_{\mathrm{ord}}[(k_1,\dots,k_{p_n}),+\infty]\cap\mathcal{B}_{op}(\gamma)$,*

$$\|\Omega-\widetilde{\Omega}^d_{\mathrm{ord}}\|_F^2=\mathcal{O}_P\left(\frac{\sum_{i=1}^{p_n}k_i+p_n}{n}\right) \tag{5.17}$$

If $s>1/2$, then uniformly over the set $\mathcal{E}'(s,R,p_n)\cap\mathcal{B}_{op}(\gamma)$, the estimator $\widetilde{\Omega}^d_{\mathrm{ord}}$ satisfies

$$\|\Omega-\widetilde{\Omega}^d_{\mathrm{ord}}\|_F^2=\mathcal{O}_P\left[p_n\left(\left(\frac{R}{n^s}\right)^{\frac{2}{2s+1}}\wedge\frac{p_n}{n}\right)\right] . \tag{5.18}$$

Moreover, these two rates are optimal from a minimax point of view.

The estimator $\widetilde{\Omega}^d_{\mathrm{ord}}$ achieves the minimax rates of estimation over special cases of ellipsoids. However, all these results depend on γ and are of *asymptotic* nature.

5.5 Complete graph selection

We now turn to the complete Cholesky factor estimation problem. First, we adapt the model selection procedure to this setting. Then, we derive an oracle inequality for the Kullback loss. Afterwards, we state that the procedure is minimax adaptive to the unknown sparsity both with respect to the Kullback entropy and the Frobenius norm.

5.5.1 Oracle inequalities

Again, we fix K and η smaller than $\eta(K)$. Let us take a maximal dimension d that satisfies

$$d\leq\eta L\frac{n}{1+[\log(p/d)\vee 0]} , \tag{5.19}$$

where L is a numerical constant that may be explicitly derived from Equation (4.15) in Chapter 4. We consider the collections of models $\mathcal{M}^d_{i,\mathrm{co}}$ that contain all the subsets of $\{1,\dots,i-1\}$ of size smaller or equal to d. In the sequel, $\widetilde{\Omega}^d_{\mathrm{co}}$ corresponds to selected estimator with the collection $\mathcal{M}^d_{\mathrm{co}}$ and the penalty defined below.

Corrolary 5.12. *For any $2 \leq i \leq p$ and any model m_i in $\mathcal{M}^d_{i,co}$ we fix the penalty*

$$pen(m_i) = K \frac{|m_i|}{n - |m_i|} \left\{ 1 + \sqrt{2 \left[1 + \log\left(\frac{i-1}{|m_i|} \right) \right]} \right\}^2 .$$

If n is larger than some quantity $n_0(K)$, then $\widetilde{\Omega}^d_{co}$ satisfies

$$\begin{aligned} \mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{\text{co}} \right) \right] &\leq L(K,\eta) \inf_{m \in \mathcal{M}^d_{\text{co}}} \left\{ \mathcal{K}\left(\Omega; \Omega_m\right) + \sum_{i=2}^{p} \frac{|m_i|}{n - |m_i|} \left[1 + \log\left(\frac{i-1}{|m_i|} \right) \right] + \frac{p}{n} \right\} + \\ &+ \tau_n \ , \end{aligned} \tag{5.20}$$

where τ_n is defined in Theorem 5.5.

Hence, we get an oracle inequality up to logarithms factors, but we prove in Section 5.5.2 that these terms $\log[(i-1)/|m_i|]$ are in fact unavoidable. For the sake of simplicity, we may straightforwardly derive from (5.20) the less sharp but more readable upper bound

$$\mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{\text{co}} \right) \right] \leq L(K,\eta) \inf_{m \in \mathcal{M}^d_{\text{co}}} \left\{ \mathcal{K}\left(\Omega; \Omega_m\right) + \frac{p + |m| \log p}{n} \right\} + \tau_n\left(\Omega, K, \eta, n, p\right) \ ,$$

where $|m| := \sum_{i=1}^{p} |m_i|$.

Again, we do not perform any assumption on the target Ω and the obtained upper bound is non-asymptotic. By Condition (5.19), we may consider dimension d up to the order $\frac{n}{\log(p/n \vee 1)}$. If p is much larger than n, the maximal dimension has to be smaller than the order $n/\log(p)$. This is not really surprising since it is also the case for linear regression with Gaussian design as stated in Section 4.3.2. There is no precise results that proves that this $n/\log(p)$ bound is optimal but we believe that it is unimprovable. If p is of the same order as n, it is possible to consider dimensions up to the same order as p.

In practice, we set the maximal dimension to $\frac{n}{2.5[2+log(p/(n \wedge p))]}$. Concerning the choice of K, we advise to use the value 1.1, if one aims at minimizing risk. If one assumes that there exists a true and wants to retrieve, a larger value of K like 1.5 or 2 is advised in order to decrease the FDR.

5.5.2 Adaptiveness to unknown sparsity

In this section, we state that the estimator $\widetilde{\Omega}^d_{\text{co}}$ achieves simultaneously the minimax rates of estimation for sparsity of the matrix T. In the sequel, $\mathcal{U}_1[k,p]$ stands for the set of positive square matrices $\Omega = T^* S^{-1} T$ of size p such that its Cholesky factor T contains at most k non-zero off-diagonal coefficients on each line. The set $\mathcal{U}_1[k,p]$ contains the precision matrices of the directed Gaussian graphical models whose underlying directed acyclic graph $\overrightarrow{\mathcal{G}}$ satisfies the two following properties:

- It is compatible with the ordering on the variables.
- Each node of $\overrightarrow{\mathcal{G}}$ has at most k parents.

We shall also consider the set $\mathcal{U}_2[k,p]$ that contains positive square matrices whose whose Cholesky factor is k-sparse (i.e. contains at most k non-zero elements). Hence, the set $\mathcal{U}_2[k,p]$ corresponds to the precision matrices of the directed Gaussian graphical models whose underlying directed acyclic graph $\overrightarrow{\mathcal{G}}$ is compatible with the ordering on the variables and has at most k edges. In contrast to the previous situation, we say the underlying Cholesky factors T are ultra-sparse.

For deriving the minimax rates of estimation, we shall restrict ourselves to precision matrices whose Kullback divergence with the identity is not too large. This is why we define

$$\mathcal{B}_{\mathcal{K}}(r) := \{\Omega \text{ s.t. } \mathcal{K}(\Omega; I_p) \leq pr\} \ ,$$

for any positive number $r > 0$.

Proposition 5.13. *Let k and p be two positive integers such that $k \leq p$. Assume that $n \geq Lk^2[1 + \log(p/k)]$, where L is some universal constant exhibited in the proof. Then, the minimax rates of estimation over the set $\mathcal{U}_1[k,p]$ is lower bounded as follows*

$$\inf_{\widehat{\Omega}} \sup_{\Omega \in \mathcal{U}_1[k,p]} \mathcal{K}\left(\Omega; \widehat{\Omega}\right) \geq Lkp \frac{1 + \log\left(\frac{p}{k}\right)}{n} \ . \tag{5.21}$$

Consider $K > 1$, $\beta > 1$, and $\eta < \eta(K)$. Assume that $n \geq n_0(K)$ and choose a positive integer d that satisfies Condition (5.19). The penalized estimator $\widetilde{\Omega}^d_{co}$ defined in Corollary 5.12 is adaptive minimax over the sets $\mathcal{U}_1[k,p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)$

$$\sup_{\Omega \in \mathcal{U}_1[k,p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)} \mathbb{E}_\Omega \left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{co}\right)\right] \leq L(K,\beta,\eta) \inf_{\widehat{\Omega}} \sup_{\Omega \in \mathcal{U}_1[k,p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)} \mathbb{E}_\Omega \left[\mathcal{K}\left(\Omega; \widehat{\Omega}\right)\right] \ ,$$

for all positive integers k smaller than d that also satisfy $n \geq k^2(1 + \log(p/k))$.

The minimax rates of estimation over $\mathcal{U}_1[k,p]$ is of order $kp\frac{1+\log\left(\frac{p}{k}\right)}{n}$. We do not think that the condition $n \geq Lk^2[1 + \log(p/k)]$ is necessary but we do not know how to remove it. The technical condition $\mathcal{K}(\Omega; I_p) \leq pn^\beta$ is not really restrictive. It comes from the term $n^{5/2}\mathcal{K}(\Omega; I_p)\exp[-nL(K,\eta)]$ in Theorem 5.5 which goes exponentially fast to 0 with n as long as $\mathcal{K}(\Omega, I_p)/p$ is grows polynomially with respect to n. In conclusion, our estimator $\widetilde{\Omega}^d_{\text{co}}$ is adaptive to the sparsity of its Cholesky factor T. Equivalently, it is minimax adaptive for estimating the distribution of a sparse directed Gaussian graphical model whose underlying graph is unknown. Let us turn to the minimax rate of estimation over ultra-sparse matrices.

Proposition 5.14. *Let k and p be two positive integers such that $k \leq p$. Assume that n is larger than some universal constant exhibited in the proof. Then, for any $\beta > 1$, the minimax rates of estimation over the set $\mathcal{U}_2[k,p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)$ is lower bounded as follows*

$$\inf_{\widehat{\Omega}} \sup_{\Omega \in \mathcal{U}_2[k,p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)} \mathcal{K}\left(\Omega; \widehat{\Omega}\right) \geq L \frac{p + k\log(p)}{n} \ . \tag{5.22}$$

Consider $K > 1$ and $\eta < \eta(K)$. Assume that $n \geq n_0(K)$ and choose d be a positive integer that satisfies Condition (5.19). The penalized estimator $\widetilde{\Omega}^d_{co}$ defined in Corollary 5.12 is adaptive minimax over the sets $\mathcal{U}_2[k,p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)$

$$\sup_{\Omega \in \mathcal{U}_2[k,p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)} \mathbb{E}_\Omega \left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{co}\right)\right] \leq L(K,\beta,\eta) \inf_{\widehat{\Omega}} \sup_{\Omega \in \mathcal{U}_2[k,p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)} \mathbb{E}_\Omega \left[\mathcal{K}\left(\Omega; \widehat{\Omega}\right)\right] \ ,$$

for all positive k smaller than d.

The estimator $\widetilde{\Omega}^d_{\text{co}}$ is therefore adaptive minimax to the ultra-sparse Cholesky factor in $\mathcal{U}_2[k,p]$.

We may also consider the rates of convergence with respect to the Frobenius norm or the operator norm in the spirit of the results of Lam and Fan [LF07]. We recall that $\|.\|_F$ and $\|.\|$ respectively refer to the Frobenius norm and the operator norm in the space of matrices. We also recall that the set $\mathcal{B}_{\text{op}}(\gamma)$ is defined in (5.14).

Corrolary 5.15. *Let $K > 1$, $\eta < \eta(K)$, $\gamma > 2$, and let d be the largest integer that satisfies (5.19). If $p_n k_n[1 + \log(\frac{p_n}{k_n})] = o(n)$,*

$$\|\Omega - \widetilde{\Omega}^d_{\text{co}}\|_F^2 = \mathcal{O}_P\left(k_n\left[1 + \log\left(\frac{p_n}{k_n}\right)\right]\frac{p_n}{n}\right) \quad \textit{and} \quad \|\Omega - \widetilde{\Omega}^d_{\text{co}}\| = \mathcal{O}_P\left(\sqrt{k_n\left[1 + \log\left(\frac{p_n}{k_n}\right)\right]\frac{p_n}{n}}\right) \tag{5.23}$$

uniformly on $\mathcal{U}_1[k_n, p_n] \cap \mathcal{B}_{op}[\gamma]$. Moreover, the Frobenius rate of convergence is optimal from a minimax point of view.

Hence, if the Cholesky factor T has at most k non-zero off-diagonal entries on each line, our estimator achieves the rate $k\log\left(\frac{p}{k}\right)\frac{p}{n}$ with respect to the Frobenius loss. Observe that these results are of asymptotic nature and that p_n has to be much smaller than n. Besides, the upper bounds on the rates highly

depend on the largest eigenvalue $\varphi_{\max}(\Omega)$. This is why we have restricted ourselves to precision matrices whose eigenvalues lie in the compact $[1/\gamma;\gamma]$. Nevertheless, to our knowledge all results in this setting suffer the same drawbacks. See for instance Th.11 of Lam and Fan [LF07].

The estimator $\widetilde{\Omega}^d_{\text{co}}$ is therefore asymptotically minimax adaptive to the sets $\mathcal{U}_1[k,p] \cap \mathcal{B}_{op}(\gamma)$ with respect to the Frobenius norm. However, we do not think that the rates of convergence with respect to the operator norm are sharp.

Corrolary 5.16. *Let $K > 1$, $\eta < \eta(K)$, $\gamma > 2$, and take the largest maximal dimension d that satisfies (5.19). If $p_n + k_n \log(p_n) = o(n)$, then*

$$\|\Omega - \widetilde{\Omega}^d_{\text{co}}\|_F^2 = \mathcal{O}_P\left(\frac{p_n + k_n \log(p_n)}{n}\right) \quad \textit{and} \quad \|\Omega - \widetilde{\Omega}^d_{\text{co}}\| = \mathcal{O}_P\left(\sqrt{\frac{p_n + k_n \log(p_n)}{n}}\right) , \tag{5.24}$$

uniformly on $\mathcal{U}_2[k_n, p_n] \cap \mathcal{B}_{op}[\gamma]$. Moreover, the Frobenius rate of convergence is optimal from a minimax point of view.

Again, we point out that $\widetilde{\Omega}^d_{\text{co}}$ is asymptotically adaptive minimax to the sets $\mathcal{U}_2[k_n, p_n]$. Moreover, these rates are coherent with the ones obtained by Lam and Fan in Sect.4 of [LF07]. Again, we do not think that the operator norm rates of convergence are optimal.

5.6 Discussion

- In this chapter, we have considered a general penalized maximum likelihood procedure for estimating the distribution of a Gaussian vector. We have obtained a non-asymptotic oracle inequality with respect to the Kullback loss. We advocate the use of this loss in order to obtain rates of convergence that are free of parameters like the largest eigenvalue. In contrast, the Frobenius rates of estimation are only asymptotic and depend in a complex form on the target Ω. Moreover, the Frobenius loss is less interpretable in terms of distribution than the Kullback. It would be of interest to study further the performances of our procedure with respect to the operator norm. Even if it is certainly not possible to obtain oracle inequalities as in Corollary 5.6, rates of convergence and minimax rates of convergence for the operator norm are of interest in many applications such as PCA as advocated for instance by Bickel and Levina [BL08b]. In their article, they also used maximum likelihood estimator but with a slightly different collection of models. Hence, we believe that our penalization method may inherit some of the nice features they proved for their procedure.

- For banded matrices, our method achieves an oracle inequality and is adaptive to the decay in the Cholesky factor T. We are able to derive corresponding asymptotic results for the Frobenius loss function. Moreover, our estimator is computationally competitive with the other existing procedures. Finally, we provide explicitly our penalty and it is therefore not needed to tune the method as done for instance in [LRZ08] or [LF07]. In a future work, we would like to study its performances with respect to the operator norm and prove similar results to the ones of [BL08b] along with the corresponding minimax lower bounds.

- In the complete graph estimation, we have derived that our procedure satisfies an oracle type inequality and achieves adaptiveness to the unknown sparsity of the Cholesky factor T. Again, asymptotic results with respect to the Frobenius loss function are provided. As in the banded case, we provide an explicit penalty. However, contrary to the l_1 penalization methods, our procedure is computationally feasible only for small p. In a future work, we would like to lower the computational cost of our procedure. In short, the idea would be to apply a computationally amenable procedure in order to reduce the collection $\mathcal{M}^d_{co}$ to some smaller collection $\widehat{\mathcal{M}^d_{co}}$. Finally, it should also be of interest to investigate further the difference of the approximation capacities of sparse Cholesky decomposition and approximation capacities of sparse precision matrices.

5.7 Proofs

5.7.1 Some notations and probabilistic tools

First, we introduce the prediction contrasts $l_i(.,.)$. Consider i be an integer between 2 and p and let (t,t') be two row vectors in $\mathbb{R}^{i-1}$ then the contrast $l_i(t,t')$ is defined as

$$l_i(t,t') := \mathrm{var}\left[\sum_{j=1}^{i-1}(t[j]-t'[j])X[j]\right] . \tag{5.25}$$

Let i be an integer larger than one and smaller than p and let m_i be a model in $\mathcal{M}_i$. We define the random variable ϵ_{m_i} by

$$X[i] = \sum_{j\in m} -t_{i,m_i}[j]X[j] + \epsilon_{m_i} + \epsilon_i \qquad \text{a.s. .} \tag{5.26}$$

By definition of t_{i,m_i}, the variable ϵ_{m_i} is independent of ϵ and of X_{m_i}. Besides, its variance equals $l_i(t_{i,m_i},t_i)$. If follows from the definition of s_{i,m_i} that $s_{i,m_i} = l_i(t_{i,m_i},t_i)+s_i$. The vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_m$ refer to the n samples of ϵ and ϵ_m. For any model m and any vector Z of size n, $\Pi_m Z$ refers to the projection of Z onto the subspace generated by $(\mathbf{X}_i)_{i\in m}$ whereas $\Pi_m^{\perp} Z$ stands for $Z-\Pi_m Z$. For any subset m of $\{1,\dots,p\}$, Σ_m denotes the covariance matrix of the vector X_m^*. Moreover, we define the row vector $Z_m := X_m\sqrt{\Sigma_m^{-1}}$ in order to deal with standard Gaussian vectors. Similarly to the matrix $\mathbf{X}_m$, the $n\times d_m$ matrix $\mathbf{Z}_m$ stands for the n observations of Z_m. We note $d_H(.,.)$ the Hamming distance between two vectors. The Hamming distance between two matrices of size p is defined as the Hamming distance between the two associated vectors of size p^2. It is also noted $d_H(.,.)$.

Lemma 5.17.

$$\mathcal{K}\left(t_i,s_i;t_i',s_i'\right) = \frac{1}{2}\left[\log\frac{s_i'}{s_i} + \frac{s_i}{s_i'} - 1 + \frac{l_i(t_i,t_i')}{s_i'}\right] . \tag{5.27}$$

The estimators $\widehat{t}_{i,m_i}$ and $\widehat{s}_{i,m_i}$ expressed as follows

$$\mathbf{X}_{<i}\widehat{t}_{i,m_i}^* = \mathbf{X}_{m_i}(\mathbf{X}_{m_i}^*\mathbf{X}_m)^{-1}\mathbf{X}_{m_i}^*\mathbf{X}_i , \tag{5.28}$$

$$\widehat{s}_{i,m_i} = \|\Pi_{m_i}^{\perp}\mathbf{X}_i\|_n^2 = \|\Pi_{m_i}^{\perp}(\boldsymbol{\epsilon}_{i,m_i}+\boldsymbol{\epsilon}_i)\|_n^2 . \tag{5.29}$$

Lemma 5.18. *Let V be a χ^2 random variable with $N>2$ degrees of freedom and let p be some positive integer such that $N>2p$, then*

$$\mathbb{E}\left[\frac{1}{V^p}\right] = \frac{1}{(N-2)\dots(N-2p)} .$$

We refer to Lemma 5 in [BGH08] for the proof of slightly more general version of this lemma.

5.7.2 Proof of Proposition 5.2

Proof of Proposition 5.2. First, we decompose the Kullback-Leibler divergence into a bias term and a variance term thanks to Expression (5.27).

$$\mathbb{E}\left[2\mathcal{K}\left(t_i,s_i;\widehat{t}_{i,m_i},\widehat{s}_{i,m_i}\right)\right] = \mathbb{E}\left[\log\frac{\widehat{s}_{i,m_i}}{s_i} + \frac{s_i^2 + l_i(\widehat{t}_{i,m_i},t_i)}{\widehat{s}_{i,m_i}} - 1\right]$$

By definition, $\widehat{t}_{i,m_i}$ is the least squares estimator of t_i over the set of vectors of size $i-1$ whose support is included in m_i and t_{i,m_i} is the best predictor of X_i given X_{m_i}. Hence, the prediction error $l_i(\widehat{t}_{i,m_i},t_i)+s_i$ equals $l_i(\widehat{t}_{i,m_i},t_{i,m_i})+s_{i,m_i}$ and it follows that

$$\begin{aligned}\mathbb{E}\left[2\mathcal{K}\left(t_i,s_i;\widehat{t}_{i,m_i},\widehat{s}_{i,m_i}\right)\right] &= 2\mathcal{K}\left(t_i,s_i;t_{i,m_i},s_{i,m_i}\right) + \\ &+ \mathbb{E}\left[\log\frac{\widehat{s}_{i,m_i}}{s_{i,m_i}} + \frac{l_i\left(\widehat{t}_{i,m_i},t_{i,m_i}\right)}{\widehat{s}_{i,m_i}} + \left(\frac{s_{i,m_i}}{\widehat{s}_{i,m_i}} - 1\right)\right] .\end{aligned} \tag{5.30}$$

Let us compute the expectation of these three last terms. Notice that $\widehat{s}_{i,m_i} = \|\Pi^{\perp}_{m_i}\mathbf{X}_i\|_n^2$ follows the distribution of a χ^2 distribution with $n-|m_i|$ degrees of freedom times $s_{i,m_i}/n$.

$$\mathbb{E}\left[\frac{s_{i,m_i}}{\widehat{s}_{i,m_i}} - 1\right] = \mathbb{E}\left[\frac{n}{\chi^2(n-|m_i|)} - 1\right] = \frac{|m_i|+2}{n-|m_i|-2}\ , \tag{5.31}$$

by Lemma 5.18. Similarly, we compute the expectation of the logarithm as follows:

$$\begin{aligned}\mathbb{E}\left[\log\frac{\widehat{s}_{i,m_i}}{s_{i,m_i}}\right] &= \mathbb{E}\left[\log\left(\frac{\chi^2(n-|m_i|)}{n}\right)\right] = \mathbb{E}\left[\log\left(\frac{\chi^2(n-|m_i|)}{n-|m_i|}\right)\right] + \log\left(\frac{n-|m_i|}{n}\right)\\ &= \Psi(n-|m_i|) + \log\left(\frac{n-|m_i|}{n}\right)\ ,\end{aligned} \tag{5.32}$$

by definition of the function $\Psi(.)$. The last term $\frac{l_i(\widehat{t}_{i,m_i},t_{i,m_i})}{\widehat{s}_{i,m_i}}$ is slightly more difficult to handle. Let us first write $l_i(\widehat{t}_{i,m_i},t_{i,m_i})$:

$$\begin{aligned}l_i(\widehat{t}_{i,m_i},t_{i,m_i}) &= (t_{i,m_i}-\widehat{t}_{i,m_i})^*\Sigma_{m_i}(t_{i,m_i}-\widehat{t}_{i,m_i})\\ &= (\mathbf{X}_i+\mathbf{X}_{m_i}t^*_{i,m_i})^*\mathbf{X}_{m_i}(\mathbf{X}^*_{m_i}\mathbf{X}_{m_i})^{-1}\Sigma_{m_i}(\mathbf{X}^*_{m_i}\mathbf{X}_{m_i})^{-1}\mathbf{X}^*_{m_i}(\mathbf{X}_i+\mathbf{X}_{m_i}t^*_{i,m_i})\\ &= (\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_{i,m_i})^*\mathbf{X}_{m_i}(\mathbf{X}^*_{m_i}\mathbf{X}_{m_i})^{-1}\Sigma_{m_i}(\mathbf{X}^*_{m_i}\mathbf{X}_{m_i})^{-1}\mathbf{X}^*_{m_i}(\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_{i,m_i})\ ,\end{aligned}$$

By definition of ϵ_{i,m_i}. Observe that $\epsilon_i+\epsilon_{i,m_i}$ is independent of X_{m_i}. Hence, conditionally to $\mathbf{X}_{m_i}$, $l_i(\widehat{t}_{i,m_i},t_{i,m_i})$ only depends on $\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_{i,m_i}$ through its orthogonal projection onto the space generated by $(\mathbf{X}_j)_{j\in m_i}$. Meanwhile, $\widehat{s}_{i,m_i} = \|\Pi^{\perp}_{m_i}(\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_{i,m_i})\|_n^2$ is the orthogonal projection of $(\boldsymbol{\epsilon}_i+\boldsymbol{\epsilon}_{i,m_i})$ along the same subspace. Thus, $l_i(\widehat{t}_{i,m_i},t_{i,m_i})$ and $\widehat{s}_{i,m_i}$ are independent conditionally to $\mathbf{X}_{m_i}$. Moreover, $\widehat{s}_{i,m_i}$ is independent of $\mathbf{X}_{i,m_i}$. Hence, $l_i(\widehat{t}_{i,m_i},t_{i,m_i})$ and $\widehat{s}_{i,m_i}$ are independent. Following the proof of Lemma 4.1, we observe that $\mathbb{E}[l_i(\widehat{t}_{i,m_i},t_{i,m_i})]$ is the expectation of the trace of an inverse Wishart $Wish^{-1}(|m_i|,n)$ times s_{i,m_i}. We then obtain that

$$\mathbb{E}\left[\frac{l_i\left(\widehat{t}_{i,m_i},t_{i,m_i}\right)}{\widehat{s}_{i,m_i}}\right] = \mathbb{E}\left[\frac{Wish^{-1}(|m_i|,n)}{\chi^2(n-|m_i|)/n}\right] = \frac{n|m_i|}{(n-|m_i|-1)(n-|m_i|-2)}\ , \tag{5.33}$$

since $\mathbb{E}\left[Wish^{-1}(|m_i|,n)\right] = |m_i|/(n-|m_i|-1)$ by Von Rosen [vR88]. Gathering identities (5.31), (5.32), and (5.33) with (5.30) yields the first result (5.5).

Let us now compute the function $\Psi(.)$.

Lemma 5.19. *For any d larger than 3,*

$$-\frac{1}{d-2} \le \Psi(d) \le 0 \quad \textit{and} \quad \Psi(d) = -\frac{1}{d} + \mathcal{O}\left(\frac{1}{d^2}\right)\ .$$

The proof is given in the Appendix. Since $\log(1-d/n)$ is negative, we obtain the first upper bound on $R_{n,d}$. For any positive number x, $\log(1+x) \le x$ and consequently $\log(1-x)$ is smaller than $-x/(1-x)$ for any positive number x smaller than one. It then follows that $\Psi(n-d)+\log(1-d/n) \ge -\frac{d+1}{n-d-2}$ and $R_{n,d} \ge \frac{d+1}{2(n-d-2)}$. Analogously, we obtain the expansion of $R_{n,d}$ when d/n goes to 0 thanks to Lemma 5.19 and the Taylor expansion of log. □

5.7.3 Proof of Theorem 5.5

Proof of Theorem 5.5. This result is based on a Kullback oracle inequality for all the estimators $(\widetilde{t}_i,\widetilde{s}_i)$ with $1 \le i \le p$. Let i be an integer between 1 and p.

Assumption $(\mathbb{H}^i_{K,\eta})$: Given $K>1$ and $\eta>0$, the collection $\mathcal{M}$ and the number η satisfy

$$\forall m_i \in \mathcal{M}_i\ , \quad \frac{\left[1+\sqrt{2H_i(|m_i|)}\right]^2|m_i|}{n-|m_i|} \le \eta < \eta(K)\ , \tag{5.34}$$

where we recall that $\eta(K)$ is defined in Equation 4.11 in Chapter 4.

Proposition 5.20. *Let $K > 1$ and $\eta < \eta(K)$. Assume that n is larger than some quantity $n_0(K)$ and that the collection $\mathcal{M}_i$ satisfies the condition $(\mathbb{H}^i_{K,\eta})$ and that the penalty function is such that*

$$pen_i(m) \geq K\frac{|m|}{n-|m|}\left(1+\sqrt{2H_i(|m|)}\right)^2 \text{ for any } m \in \mathcal{M}_i \text{ and some } K > 1 \ . \tag{5.35}$$

Then, the penalized estimator $(\widetilde{t}_i, \widetilde{s}_i)$ satisfies

$$\mathbb{E}\left[\mathcal{K}\left(t_i, s_i; \widetilde{t}_i, \widetilde{s}_i\right)\right] \leq L(K,\eta) \inf_{m_i \in \mathcal{M}_i}\left[\mathbb{E}\left[\mathcal{K}\left(t_i, s_i; \widehat{t}_{i,m}, \widehat{s}_{i,m}\right)\right] + pen_i(m)\right] + \tau_n\left[t_i, s_i, K, \eta\right] \ .$$

The small term $\tau_n(t_i, s_i, K, \eta)$ is defined as

$$\tau_n\left[t_i, s_i, K, \eta\right] := \frac{L(K)}{n} + L'(K,\eta)n^{5/2}\left[1+\mathcal{K}\left(t_i, s_i; 0, 1\right)\right]\exp\left[-nL(K,\eta)\right] \ ,$$

where 0 stands here for the null vector of size $i-1$.

Then, we apply p times this property and use the chain rule as in Section 5.3.1 to conclude. □

Proof of Proposition 5.20. The proof of this theorem is mainly inspired by ideas introduced in the proofs of Theorem 3 in [BGH08] and of Theorem 4.7. The case $i = 1$ is a consequence of Proposition 5.2 since $|\mathcal{M}_1| = 1$. In the sequel, we assume that i is larger than one. For the sake of clarity, we forget the subscripts i in the remainder of the proof.

Let us introduce some new notations. First, $\langle ., . \rangle_n$ is the natural inner product in $\mathbb{R}^n$ associated to the norm $\|.\|_n$. Let m be any model in the collection $\mathcal{M}$. We define the row vector $Z_m := X_m\sqrt{\Sigma_m^{-1}}$ in order to deal with standard Gaussian vectors. Similarly to the matrix $\mathbf{X}_m$, the $n \times d_m$ matrix $\mathbf{Z}_m$ stands for the n observations of Z_m.

We shall use the constants κ_1, κ_2, and $\nu(K)$ as defined in the proof of Theorem 4.7. We provide their expression for completeness although it not really of interest.

$$\begin{aligned}
\kappa_1 &:= \frac{\sqrt{\frac{3}{K+2}}}{1-\sqrt{\eta}-\nu(K)} \ , \qquad \kappa_2 := \frac{(K-1)\left[1-\sqrt{\eta}\right]^2\left[1-\sqrt{\eta}-\nu(K)\right]^2}{16} \wedge 1 \ , \\
\nu(K) &:= \left(\frac{3}{K+2}\right)^{1/6} \wedge \frac{1-\left(\frac{3}{K+2}\right)^{1/6}}{2} \ .
\end{aligned}$$

Besides, we introduce the positive constant κ_0 as the largest number that satisfies

$$\kappa_0 \leq 1 - \frac{2}{K+1} \text{ and } \frac{K+2}{3} \leq (1+\kappa_0)\frac{K+1.5}{2.5} \ .$$

For clarity, the proof is splitted into five lemmas.

Lemma 5.21.

$$\begin{aligned}
2(1-\kappa_0)\mathcal{K}\left[t, s; \widetilde{t}, \widetilde{s}\right] &\leq 2\mathcal{K}\left[t, s; \widehat{t}_m, \widehat{s}_m\right] + (1-\kappa_0)pen(m) + \frac{l(\widetilde{t}, t)}{\widetilde{s}}\left[R_1(\widehat{m}) \vee (1-\kappa_2)(1-\kappa_0)\right] + \\
&+ R_2(m) + \frac{s}{\widetilde{s}}R_3(\widehat{m}) + R_4(m, \widehat{m}) \ ,
\end{aligned}$$

where for all model $m' \in \mathcal{M}$,

$$\begin{aligned}
R_1(m') &:= \kappa_1 + 1 - \kappa_0 - \frac{\|\Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\|_n^2}{l(t_{m'},t)} + \kappa_2(1-\kappa_0)\varphi_{max}\left[n(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(t_{m'},t)+s} \ , \\
&- K(1-\kappa_0)\left[1+\sqrt{2H(|m'|)}\right]^2 \frac{|m'|}{n-|m'|}\frac{\|\Pi_{m'^{\perp}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(t_{m'},t)+s} \\
R_2(m) &:= 2\frac{\langle \Pi^{\perp}_m\boldsymbol{\epsilon}, \Pi^{\perp}_m\boldsymbol{\epsilon}_m\rangle_n}{\widehat{s}_m} + \frac{\|\Pi^{\perp}_m\boldsymbol{\epsilon}_m\|_n^2 - l(t_m,t)}{\widehat{s}_m} \ , \\
R_3(m') &:= \kappa_1^{-1}\frac{\langle \Pi^{\perp}_{m'}\boldsymbol{\epsilon}, \Pi^{\perp}_{m'}\boldsymbol{\epsilon}_{m'}\rangle_n^2}{sl(t_{m'},t)} + \kappa_2(1-\kappa_0)\varphi_{max}\left[n(\mathbf{Z}^*_{m'}\mathbf{Z}_{m'})^{-1}\right]\frac{\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(t_{m'},t)+s} \\
&+ \frac{\|\Pi_{m'}\boldsymbol{\epsilon}\|_n^2}{s} - K(1-\kappa_0)\left[1+\sqrt{2H(|m'|)}\right]^2\frac{|m'|}{n-|m'|}\frac{\|\Pi^{\perp}_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2}{l(t_{m'},t)+s} \ , \\
R_4(m,m') &:= \left(\|\boldsymbol{\epsilon}\|_n^2 - s(1-\kappa_0)\right)\left[\frac{1}{\widehat{s}_m} - \frac{1}{\widehat{s}_{m'}}\right] \ .
\end{aligned}$$

Lemma 5.22. *Assume that* n *is larger than some quantity* $n_0(K)$. *There exists an event* Ω_1 *of probability larger than* $1 - L(K)n\exp\left[-nL'(K,\eta)\right]$ *such that*

$$R_1(\widehat{m})\mathbf{1}_{\Omega_1} \leq v(K,\eta)(1-\kappa_0) \qquad \textit{and} \qquad \mathbb{E}\left[\frac{s}{\widetilde{s}}R_3(\widehat{m})\mathbf{1}_{\Omega_1}\right] \leq \frac{L(K,\eta)}{n},$$

where $v(K,\eta)$ *is a positive constant (strictly) smaller than* 1.

Since $l(\widetilde{t},t)/\widetilde{s}$ is smaller than $2\mathcal{K}\left[t,s;\widetilde{t},\widetilde{s}\right]$, it follows that

$$2\mathbb{E}\left[\mathcal{K}\left(t,s;\widetilde{t},\widetilde{s}\right)\mathbf{1}_{\Omega_1}\right] \leq L(K,\eta)\left\{2\mathbb{E}\left[\mathcal{K}\left(t,s;\widehat{t}_m,\widehat{s}_m\right)\right] + \mathrm{pen}(m) + \mathbb{E}\left[(R_2(m)+R_4(m,\widehat{m}))\mathbf{1}_{\Omega_1}\right]\right\} \ .$$

Lemma 5.23. *Suppose that* n *is larger than some quantity* $n_0(K)$. *Considering the event* Ω_1 *defined in Lemma 5.22, we bound* $R_2(m)$ *by*

$$\mathbb{E}\left[R_2(m)\mathbf{1}_{\Omega_1}\right] \leq L(K)\sqrt{n}\exp\left[-nL(K,\eta)\right] \ .$$

Lemma 5.24. *Suppose that* n *is larger than some quantity* $n_0(K)$. *Considering the event* Ω_1 *defined in Lemma 5.22, then*

$$\mathbb{E}\left[R_4(m,\widehat{m})\mathbf{1}_{\Omega_1}\right] \leq Lpen(m) + n\exp\left[-nL(K)\right] \ .$$

Gathering these two lemma, we control the Kullback risk of $(\widetilde{t},\widetilde{s})$ on the event Ω_1

$$\begin{aligned}
2\mathbb{E}\left[\mathcal{K}\left(t,s;\widetilde{t},\widetilde{s}\right)\mathbf{1}_{\Omega_1}\right] &\leq L(K,\eta)\left\{2\mathbb{E}\left[\mathcal{K}\left[t,s;\widehat{t}_m,\widehat{s}_m\right]\right] + \mathrm{pen}(m)\right\} \\
&+ \frac{L(K)}{n} + (n+L)\exp\left[-nL(K)\right] \ . \qquad (5.36)
\end{aligned}$$

To conclude, we need to control the Kullback risk of the estimator $(\widetilde{t},\widetilde{s})$ on the event Ω_1^c.

Lemma 5.25.

$$\mathbb{E}\left[\mathcal{K}\left(t,s;\widetilde{t},\widetilde{s}\right)\mathbf{1}_{\Omega_1^c}\right] \leq L(K,\eta)n^{5/2}\left[1+\mathcal{K}(t,s;0,1)\right]\exp\left[-nL(K)\right] \ .$$

Combining (5.36) and Lemma 5.25 allows to conclude

$$\begin{aligned}
\mathbb{E}\left[\mathcal{K}\left(t,s;\widetilde{t},\widetilde{s}\right)\right] &\leq L(K,\eta)\left[\mathbb{E}\left[\mathcal{K}\left(t,s;\widehat{t}_m,\widehat{s}_m\right)\right] + \mathrm{pen}(m)\right] + \frac{L(K)}{n} \\
&+ L(K,\eta)n^{5/2}\left[1+\mathcal{K}(t,s;0,1)\right]\exp\left[-nL(K)\right] \ .
\end{aligned}$$

□

Proof of Lemma 5.21. Using expression (5.27) of $\mathcal{K}\left[t,s;\widehat{t}_m,\widehat{s}_m\right]$, we derive

$$\begin{aligned}
2(1-\kappa_0)\mathcal{K}\left(t,s;\widetilde{t},\widetilde{s}\right) &= 2\mathcal{K}\left[t,s;\widehat{t}_m,\widehat{s}_m\right] + (1-\kappa_0)\log\left(\frac{\widetilde{s}}{\widehat{s}_m}\right) + (1-\kappa_0)\frac{s+l(\widetilde{t},t)}{\widetilde{s}} \\
&- \frac{s+l(\widehat{t}_m,t)}{\widehat{s}_m} + \kappa_0 + \kappa_0\log\left(\frac{s}{\widehat{s}_m}\right) \ .
\end{aligned}$$

By definition of $\widehat{m}$, $\log\left(\frac{\widetilde{s}}{\widehat{s}_m}\right) \leq \mathrm{pen}(m) - \mathrm{pen}(\widehat{m})$. Hence,

$$\begin{aligned}
2(1-\kappa_0)\mathcal{K}\left(t,s;\widetilde{t},\widetilde{s}\right) &\leq 2\mathcal{K}\left(t,s;\widehat{t}_m,\widehat{s}_m\right) + (1-\kappa_0)\left[\mathrm{pen}(m) - \mathrm{pen}(\widehat{m})\right] \\
&+ \kappa_0\frac{s}{\widehat{s}_m} + \kappa_0\left[-\frac{s}{\widehat{s}_m} + 1 + \log\left(\frac{s}{\widehat{s}_m}\right)\right] - \frac{s + l(\widehat{t}_m,t) - \|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\widehat{s}_m} + \\
&+ \frac{l(\widetilde{t},t)(1-\kappa_0) + s(1-\kappa_0) - \|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\widetilde{s}} ,
\end{aligned}$$

since $\widehat{s}_m = \|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2$ and $\widetilde{s} = \|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2$. As the function $x - \log x - 1$ is positive, the term $\left[-\frac{s}{\widehat{s}_m} + 1 + \log\left(\frac{s}{\widehat{s}_m}\right)\right]$ is negative. Since t_m is the best predictor of X_i conditionally to X_m, It follows that $l(\widetilde{t},t) = l(\widetilde{t},t_{\widehat{m}}) + l(t_{\widehat{m}},t)$. Hence,

$$\kappa_0\frac{s}{\widehat{s}_m} - \frac{s + l(\widehat{t}_m,t) - \|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2}{\widehat{s}_m} \leq -(1-\kappa_0)\frac{s}{\widehat{s}_m} + \frac{\|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2 - l(t_m,t)}{\widehat{s}_m} .$$

In the proof of Lemma 4.19, we state that $l(\widehat{t}_{m'},t_{m'}) \leq \varphi_{\max}\left[n(\mathbf{Z}_{m'}^*\mathbf{Z}_{m'})^{-1}\right]\|\Pi_{m'}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{m'})\|_n^2$. This yields

$$\begin{aligned}
\frac{[l(\widetilde{t},t)+s](1-\kappa_0)}{\widetilde{s}} &\leq (1-\kappa_0)\frac{s + l(t_{\widehat{m}},t) + \kappa_2\varphi_{\max}\left[n(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\widetilde{s}} \\
&+ (1-\kappa_0)(1-\kappa_2)\frac{l(\widetilde{t},t_{\widehat{m}})}{\widetilde{s}} .
\end{aligned}$$

Let us gather all these bounds

$$\begin{aligned}
2(1-\kappa_0)\mathcal{K}\left[t,s;\widetilde{t},\widetilde{s}\right] &\leq 2\mathcal{K}\left[t,s;\widehat{t}_m,\widehat{s}_m\right] + (1-\kappa_0)\left[\mathrm{pen}(m) - \mathrm{pen}(\widehat{m})\right] \\
&+ (1-\kappa_0)\frac{l(t_{\widehat{m}},t) + (1-\kappa_2)l(\widetilde{t},t_{\widehat{m}}) + \kappa_2\varphi_{\max}\left[n(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\widetilde{s}} \\
&- \frac{\|\Pi_{\widehat{m}}^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\widetilde{s}} + s(1-\kappa_0)\left(\frac{1}{\widetilde{s}} - \frac{1}{\widehat{s}_m}\right) + \frac{\|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2 - l(t_m,t)}{\widehat{s}_m} \\
&\leq 2\mathcal{K}\left[t,s;\widehat{t}_m,\widehat{s}_m\right] + (1-\kappa_0)\left[\mathrm{pen}(m) - \mathrm{pen}(\widehat{m})\right] \\
&+ (1-\kappa_0)\frac{l(t_{\widehat{m}},t) + (1-\kappa_2)l(\widetilde{t},t_{\widehat{m}}) + \kappa_2\varphi_{\max}\left[n(\mathbf{Z}_{\widehat{m}}^*\mathbf{Z}_{\widehat{m}})^{-1}\right]\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{\widetilde{s}} \\
&+ \left(\|\boldsymbol{\epsilon}\|_n^2 - s(1-\kappa_0)\right)\left(\frac{1}{\widehat{s}_m} - \frac{1}{\widetilde{s}}\right) + \frac{\|\Pi_{\widehat{m}}\boldsymbol{\epsilon}\|_n^2}{\widetilde{s}} + 2\frac{\langle\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon},\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\rangle_n}{\widetilde{s}} \\
&- \frac{\|\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\|_n^2}{\widetilde{s}} + 2\frac{\langle\Pi_m^{\perp}\boldsymbol{\epsilon},\Pi_m^{\perp}\boldsymbol{\epsilon}_m\rangle_n}{\widehat{s}_m} + \frac{\|\Pi_m^{\perp}\boldsymbol{\epsilon}_m\|_n^2 - l(t_m,t)}{\widehat{s}_m} .
\end{aligned}$$

In order to conclude, we use Condition (5.35) on $\mathrm{pen}(\widehat{m})$, we write any $\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2$ as $[s+l(t_{\widehat{m}},t)]\frac{\|\Pi_{\widehat{m}}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_{\widehat{m}})\|_n^2}{s+l(t_{\widehat{m}},t)}$, and apply the inequality

$$2\frac{\langle\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon},\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\rangle_n}{\widetilde{s}} \leq \kappa_1\frac{l(t_{\widehat{m}},t)}{\widetilde{s}} + \kappa_1^{-1}\frac{s}{\widetilde{s}}\frac{\langle\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon},\Pi_{\widehat{m}}^{\perp}\boldsymbol{\epsilon}_{\widehat{m}}\rangle_n^2}{sl(t_{\widehat{m}},t)} .$$

Hence, we conclude that

$$\begin{aligned}
2(1-\kappa_0)\mathcal{K}\left[t,s;\widetilde{t},\widetilde{s}\right] &\leq 2\mathcal{K}\left[t,s;\widehat{t}_m,\widehat{s}_m\right] + (1-\kappa_0)\mathrm{pen}(m) \\
&+ \frac{l(\widetilde{t},t)}{\widetilde{s}}\left[R_1(\widehat{m}) \vee (1-\kappa_2)(1-\kappa_0)\right] + R_2(m) + \frac{s}{\widetilde{s}}R_3(\widehat{m}) + R_4(m,\widehat{m}) .
\end{aligned}$$

□

Proof of Lemma 5.22. We argue exactly as in the proofs of Lemma 4.24 and 4.25. Assume that n is larger than $n_0(K)$, where $n_0(K)$ is defined in Lemma 4.25. Let us define the event Ω_1 as the intersection of the event Ω_1' introduced in 4.24, Ω_2 introduced in Lemma 4.25 and Ω_3 defined as:

$$\Omega_3 := \left\{\frac{\widehat{s}_m}{l(t_m,t)+s} \geq \frac{1}{2}\left(1 - \delta_0(K) - \sqrt{\eta}\right)^2\right\} .$$

Hence, $\mathbb{P}[\Omega_1^c] \leq L(K)n\exp[-nL'(K,\eta)]$. As we have chosen κ_0 small enough, the upper bounds made through Lemma concentrationthrm1complete easily adapt to our situations and it follows that $R_1(\widehat{m})\mathbf{1}_{\Omega_1} \leq s(K,\eta)$. As κ_0 is small enough, we are also able to follow the computations of Lemma majorationesperanceconditionellecomplete.Hence, for any $x>0$

$$R_3(\widehat{m})\mathbf{1}_{\Omega_1} \leq \frac{L(K,\eta)(1+x)}{n} ,$$

with probability larger than $1-\exp(-x)$. Integrating with respect to x and noting that $s/\widetilde{s}$ is smaller than $L(K,\eta)$ on Ω_1 allows to conclude.

□

Proof of Lemma 5.23. The random variable $\frac{\langle \Pi_m^\perp \boldsymbol{\epsilon}, \Pi_m^\perp \boldsymbol{\epsilon}_m\rangle_n}{\widehat{s}_m}$ equals $\frac{\langle \Pi_m^\perp \boldsymbol{\epsilon}, \Pi_m^\perp \boldsymbol{\epsilon}_m\rangle_n}{\|\Pi_m^\perp \boldsymbol{\epsilon}+\Pi_m^\perp \boldsymbol{\epsilon}_m\|_n^2}$. Since X_m, $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_m$ are independent, this random variable follows the same distribution as $T(U,V) =: \langle\sqrt{s}U, \sqrt{l(t_m,t)}V\rangle/\|\sqrt{s}U+\sqrt{l(t_m,t)}V\|^2$, where U and V follow standard normal distributions of size $n-|m|$. Let us condition the random variable $T(U,V)$ by $U=u\neq 0$. Observe that if $T(u,v)>0$ then $T(u,-v)<0$ and $|T(u,-v)|\geq|T(u,v)|$. Since the distribution of V is symmetric, it follows that the expectation of T is non-positive.

$$\begin{aligned}
\mathbb{E}[R_2(m)\mathbf{1}_{\Omega_1}] &= \mathbb{E}\left[2\frac{\langle \Pi_m^\perp \boldsymbol{\epsilon}, \Pi_m^\perp \boldsymbol{\epsilon}_m\rangle_n}{\widehat{s}_m}\mathbf{1}_{\Omega_1}\right] + \mathbb{E}\left[1-\frac{l(t_m,t)}{\widehat{s}_m}\mathbf{1}_{\Omega_1}\right] \\
&\leq -\mathbb{E}\left[2\frac{\langle \Pi_m^\perp \boldsymbol{\epsilon}, \Pi_m^\perp \boldsymbol{\epsilon}_m\rangle_n}{\widehat{s}_m}\mathbf{1}_{\Omega_1^c}\right] + \mathbb{E}\left[1-\frac{l(t_m,t)}{\widehat{s}_m}\right] - \mathbb{E}\left[\left(1-\frac{l(t_m,t)}{\widehat{s}_m}\right)\mathbf{1}_{\Omega_1^c}\right] .
\end{aligned}$$

We compute the expectation of the square inverse of a χ^2 random variable applying Lemma 5.18. Observe that the expectation of the second term is negative. Hence,

$$\begin{aligned}
\mathbb{E}[R_2(m)\mathbf{1}_{\Omega_1}] &\leq 2\sqrt{\mathbb{P}(\Omega_1^c)}\left[\mathbb{E}\left(\langle \Pi_m^\perp \boldsymbol{\epsilon}, \Pi_m^\perp \boldsymbol{\epsilon}_m\rangle_n^4\right)\mathbb{E}\left(\frac{1}{\widehat{s}_m^4}\right)\right]^{1/4} + \sqrt{\mathbb{P}(\Omega_1^c)}\sqrt{\mathbb{E}\left[1-\frac{l(t_m,t)}{\widehat{s}_m}\right]^2} \\
&\leq \sqrt{\mathbb{P}(\Omega_1^c)}\left[2\left(\frac{3(n-|m|)(n-|m|+2)}{(n-|m|-2)\dots(n-|m|-8)}\right)^{1/4} + \sqrt{\frac{|m|^2+6|m|+8}{(n-|m|-2)(n-|m|-4)}}\right] \\
&\leq L(K)\sqrt{n}\exp[-nL'(K,\eta)] ,
\end{aligned}$$

since by Assumption $\mathbb{H}_{K,\eta}$, $|m|$ is smaller than $n/2$ and since n is larger than 17. □

Proof of Lemma 5.24. We bound the quantity $R_4(m,\widehat{m})$ using the same arguments as in the proof of Theorem 3 in [BGH08]. We first split this quantity into a sum of two terms:

$$\begin{aligned}
R_4(m,\widehat{m}) &= \left(\|\boldsymbol{\epsilon}\|_n^2 - s(1-\kappa_0)\right)_+\left[\frac{1}{\widehat{s}_m}-\frac{1}{\widetilde{s}}\right] + \left(s(1-\kappa_0)-\|\boldsymbol{\epsilon}\|_n^2\right)_+\left[-\frac{1}{\widehat{s}_m}+\frac{1}{\widetilde{s}}\right] \\
&\leq R_{4,1}(m,\widehat{m}) + R_{4,2}(\widehat{m}) ,
\end{aligned}$$

where $R_{4,1}(m,\widehat{m})$ and $R_{4,2}(\widehat{m})$ are respectively defined as

$$\begin{aligned}
R_{4,1}(m,\widehat{m}) &:= \left(\|\boldsymbol{\epsilon}\|_n^2 - s(1-\kappa_0)\right)_+\left[\frac{1}{\widehat{s}_m}-\frac{1}{\widetilde{s}}\right] \\
R_{4,2}(\widehat{m}) &:= \left(s(1-\kappa_0)-\|\boldsymbol{\epsilon}\|_n^2\right)_+\frac{1}{\widetilde{s}}.
\end{aligned}$$

By definition, we know that $\log\left(\frac{\widetilde{s}}{\widehat{s}_m}\right)$ is smaller than $\mathrm{pen}(m)-\mathrm{pen}(\widehat{m})$.

$$\begin{aligned}
R_{4,1}(m,\widehat{m}) &\leq \left(\|\boldsymbol{\epsilon}\|_n^2 - s(1-\kappa_0)\right)_+\frac{1}{\widehat{s}_m}\log\left(\frac{\widetilde{s}}{\widehat{s}_m}\right) \\
&\leq \left(\|\boldsymbol{\epsilon}\|_n^2 - s(1-\kappa_0)\right)_+\frac{1}{\widehat{s}_m}\mathrm{pen}(m) .
\end{aligned}$$

Applying Cauchy-Schwarz inequality yields

$$\begin{aligned}
\mathbb{E}\left[R_{4,1}(m,\widehat{m})\mathbf{1}_{\Omega_1}\right] &\leq \mathbb{E}\left[\frac{\left(\|\boldsymbol{\epsilon}\|_n^2 - s(1-\kappa_0)\right)_+}{\widehat{s}_m}\right]\mathrm{pen}(m) \\
&\leq \sqrt{\mathbb{E}\left[\left(\|\boldsymbol{\epsilon}\|_n^2 - s(1-\kappa_0)\right)^2\right]\mathbb{E}\left[\frac{1}{\widehat{s}_m^2}\right]}\mathrm{pen}(m) \\
&\leq \frac{s}{s_m}\sqrt{\left[\kappa_1^2 + \frac{2}{n}\right]\frac{n^2}{(n-|m|-2)(n-|m|-4)}}\mathrm{pen}(m) \\
&\leq L\mathrm{pen}(m)\ ,
\end{aligned}$$

since $|m| \leq n/2$ by Assumption $\mathbb{H}_{K,\eta}$ and since $n \geq 17$. Let us turn to $R_{4,2}(\widehat{m})$. We apply Hölder's inequality with $v := \lfloor \frac{n}{8} \rfloor$ and $u = \frac{v}{v-1}$.

$$\begin{aligned}
\mathbb{E}\left[R_{4,2}(\widehat{m})\mathbf{1}_{\Omega_1}\right] &\leq \mathbb{E}\left[\left(s(1-\kappa_0) - \|\boldsymbol{\epsilon}\|_n^2\right)_+ \frac{1}{\widetilde{s}}\right] \\
&\leq \mathbb{E}\left[\mathbf{1}_{s(1-\kappa_0)\geq\|\boldsymbol{\epsilon}\|_n^2}\frac{s}{\widetilde{s}}\right] \\
&\leq \left[\mathbb{P}\left[\|\boldsymbol{\epsilon}\|_n^2 \leq s(1-\kappa_0)\right]\right]^{1/u}\left[\mathbb{E}\left(\frac{s}{\widetilde{s}}\right)^v\right]^{1/v} \\
&\leq \left[\mathbb{P}\left[\|\boldsymbol{\epsilon}\|_n^2 \leq s(1-\kappa_0)\right]\right]^{1/u}\left[\sum_{m\in\mathcal{M}}\mathbb{E}\left(\frac{s}{\widehat{s}_m}\right)^v\right]^{1/v}\ .
\end{aligned}$$

Since v is smaller than $n/8$ and since $|m|$ is smaller than $n/2$ it follows that $n - |m| - 2v$ is larger than $n/4$. Hence, we may apply Lemma 5.18 to any model $m \in \mathcal{M}$.

$$\begin{aligned}
\mathbb{E}\left[R_{4,2}(\widehat{m})\mathbf{1}_{\Omega_1}\right] &\leq \exp\left[-n\frac{\kappa_0^2}{4u}\right]\left[\sum_{m\in\mathcal{M}}\frac{n^v}{(n-|m|-2)\ldots..(n-|m|-2v)}\right]^{1/v} \\
&\leq \frac{n}{n-d_{\max}-2v}\exp\left[-n\frac{\kappa_0^2}{4u}\right]|\mathcal{M}|^{1/v} \\
&\leq n\exp\left[-n\frac{\kappa_0^2}{4u}\right]|\mathcal{M}|^{1/v}\ .
\end{aligned}$$

Let us bound the cardinality of the collection $\mathcal{M}$. We recall that the dimension of any model $m \in \mathcal{M}$ is assumed to be smaller than $n/2$ by $(\mathbb{H}_{K,\eta})$. Besides, for any $d \in \{1,\ldots,n/2\}$, there are less than $\exp(dH(d))$ models of dimension d. Hence,

$$\log(\mathcal{M}) \leq \log(n) + \sup_{d=1,\ldots,n/2} dH(d)\ .$$

By assumption $(\mathbb{H}_{K,\eta})$, $dH(d)$ is smaller than $n/2$. Thus, $\log(\mathcal{M}) \leq \log(n) + n/2$ and it follows that $|(\mathcal{M})|^{1/v}$ is smaller than an universal constant providing that n is larger than 8. All in all, we get

$$\mathbb{E}\left[R_{4,2}(\widehat{m})\mathbf{1}_{\Omega_1}\right] \leq Ln\exp\left[-n\frac{\kappa_0^2}{4u}\right]$$

□

Proof of Lemma 5.25. For any $x > 0$, the following inequality holds

$$x - 1 - \log(x) \leq \frac{9}{64}\left(x - \frac{1}{x}\right)^2\ .$$

This statement is easy to establish by studying the derivative of the associated function. Hence, we upper bound the Kullback divergence

$$\begin{aligned}
\mathcal{K}\left[t,s;\widehat{t}_m,\widehat{s}_m\right] &= \frac{s}{\widehat{s}_m} + 1 - \log\left(\frac{s}{\widehat{s}_m}\right) + \frac{l(\widehat{t}_m,t)}{\widehat{s}_m} \\
&\leq \frac{9}{64}\left[\frac{s^2}{\widehat{s}_m^2} + \frac{\widehat{s}_m^2}{s^2}\right] + \frac{l(t_m,t)}{\widehat{s}_m} + \frac{l(\widehat{t}_m,t_m)}{\widehat{s}_m}\ .
\end{aligned}$$

Thanks to Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\mathcal{K}\left(t,s;\widetilde{t},\widetilde{s}\right)\mathbf{1}_{\Omega_1^c}\right] &\leq \mathbb{E}\left[\left(\frac{9}{64}\left[\frac{s^2}{\widetilde{s}^2}+\frac{\widetilde{s}^2}{s^2}\right]+\frac{l(t_m,t)}{\widetilde{s}}+\frac{l(\widetilde{t},t_{\widehat{m}})}{\widetilde{s}}\right)\mathbf{1}_{\Omega_1^c}\right] \\ &\leq L\sqrt{\mathbb{P}\left[\Omega_1^c\right]}\sqrt{\sum_{m\in\mathcal{M}}\mathbb{E}\left[\mathbf{1}_{m=\widehat{m}}\left(\frac{s^4}{\widehat{s}_m^4}+\frac{\widehat{s}_m^4}{s^4}+\frac{l(t_m,t)^2}{\widehat{s}_m^2}+\frac{l^2(\widehat{t}_m,t_m}{\widehat{s}_m^2}\right)\right]}\ .\end{aligned}$$

As in the proof of Lemma 5.24, we apply Hölder's inequality with $v=\lfloor\frac{n}{16}\rfloor$ and $u=\frac{v}{v-1}$. Again, we check that for any model $m\in\mathcal{M}$, $n-|m|-8v\geq 1$.

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}_{m=\widehat{m}}\left(\frac{s^4}{\widehat{s}_m^4}+\frac{\widehat{s}_m^4}{s^4}+\frac{l(t_m,t)^2}{\widehat{s}_m^2}+\frac{l^2(\widehat{t}_m,t_m}{\widehat{s}_m^2}\right)\right] \\ &\quad\leq \mathbb{P}\left[m=\widehat{m}\right]^{\frac{1}{u}}\left[\mathbb{E}\left(\frac{s^{4v}}{\widehat{s}_m^{4v}}\right)^{\frac{1}{v}}+\mathbb{E}\left(\frac{\widehat{s}_m^{4v}}{s^{4v}}\right)^{\frac{1}{v}}+\mathbb{E}\left(\frac{l(t_m,t)^{4v}}{\widehat{s}_m^{2v}}\right)^{\frac{1}{v}}+\mathbb{E}\left(\frac{l(\widehat{t}_m,t_m)^{2v}}{\widehat{s}_m^{2v}}\right)^{\frac{1}{v}}\right]\ .\end{aligned}$$

We bound the first two terms applying Lemma 5.18 or computing the v-th moment of χ^2 random variable.

$$\begin{aligned}\mathbb{E}\left[\frac{s^{4v}}{\widehat{s}_m^{4v}}\right]^{\frac{1}{v}} &\leq \frac{n^4}{(n-|m|-8v)^4}\ , \\ \mathbb{E}\left[\frac{\widehat{s}_m^{4v}}{s^{4v}}\right]^{\frac{1}{v}} &= \left(\frac{(n-|m|)(n-|m|+2)\dots(n-|m|+2(4v-1))(s_m)^{4v}}{(ns)^{4v}}\right)^{\frac{1}{v}}\leq\frac{(n-|m|+8v)^4\left(s+l(0,t)\right)^4}{n^4s^4}\ .\end{aligned}$$

As $\|\Pi_m^{\perp}(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2$ is independent of the couple $(\|\Pi_m(\boldsymbol{\epsilon}+\boldsymbol{\epsilon}_m)\|_n^2,\mathbf{X_m})$, the random variables $\widehat{s}_m$ and $l(\widehat{t}_m,t_m)$ are independent. We bound the the l^{2v}-risk of $l(\widehat{t}_m,t)$ thanks to Proposition 4.22.

$$\begin{aligned}\mathbb{E}\left(\frac{l(\widehat{t}_m,t_m)^{2v}}{\widehat{s}_m^{2v}}\right)^{\frac{1}{v}} &= \left(\mathbb{E}\left[l(\widehat{t}_m,t_m)^{2v}\right]\mathbb{E}\left[\frac{1}{\widehat{s}_m^{2v}}\right]\right)^{\frac{1}{v}} \\ &\leq \frac{(Lv|m|ns_m)^2n^2}{(n-|m|-4v)^2(s_m)}\leq(Lv|m|n)^2\frac{n^2}{(n-|m|-4v)^2}\ .\end{aligned}$$

Combining these upper bounds and noting that $n-|m|-8v\geq 1$ and $|m|\leq n/2$ yields

$$\begin{aligned}\mathbb{E}\left[\mathcal{K}\left(t,s;\widetilde{t},\widetilde{s}\right)\mathbf{1}_{\Omega_1^c}\right] &\leq \left[\frac{2n^2}{(n-|m|-8v)^2}+\frac{Lv|m|n^2}{n-|m|-4v}+\frac{(n-|m|+8v)^2}{n^2}\left(1+\frac{l(0,t)}{s}\right)^2\right]\times \\ &\times\ L\sqrt{\mathbb{P}\left[\Omega_1^c\right]}|\mathcal{M}|^{\frac{1}{2v}} \\ &\leq L(K,\eta)n^{5/2}\left[1+\frac{l(0,t)}{s}\right]\exp\left[-nL(K,\eta)\right]\ ,\end{aligned}$$

since $|\mathcal{M}|^{1/2v}$ is smaller than than an universal constant as explained in the proof of Lemma 5.24. Finally observe that $\frac{l(0,t)}{s}$ is smaller than $\mathcal{K}(t,s;0,1)$.

□

5.7.4 Proofs of the minimax bounds

We first state two lemma we shall often apply in these proofs. The first one is known as Varshanov-Gilbert's lemma, whereas the second one is a modified version of Birgé's lemma for covariance estimation.

Lemma 5.26 (Varshanov-Gilbert's lemma)**.** *Let $\{0,1\}^D$ be equipped with Hamming distance d_H. There exists some subset Θ of $\{0,1\}^D$ with the following properties*

$$d_H(\theta,\theta')>D/4 \text{ for every } (\theta,\theta')\in\Theta^2 \text{ with } \theta\neq\theta' \text{ and } \log|\Theta|\geq D/8\ .$$

In the sequel, we note $\|t\|_{l_2}$ the l_2 norm of a vector t.

Lemma 5.27. *Let A be a subset of $\{1,\dots,p\}$. For any positive matrices Ω and Ω', we define the function $d(\Omega,\Omega')$ as*

$$d(\Omega,\Omega') := \sum_{i\in A} \log\left[1+\frac{\|t_i-t_i'\|_{l_2}^2}{4}\right] + \sum_{i\in A^c} \frac{s_i}{s_i'} + \log\left(\frac{s_i}{s_i'}\right) - 1 \ . \tag{5.37}$$

Let Υ be a subset of square matrices of size p which satisfies the following assumptions:

1. *For all $\Omega\in\Upsilon$, $\phi_{max}(\Omega)\le 2$ and $\phi_{min}(\Omega)\ge 1/2$.*
2. *There exists $(\mathbf{s}_1,\mathbf{s}_2)\in[1;2]^2$ such that for any $\Omega\in\Upsilon$ and any $1\le i\le p$, $s_i\in\{\mathbf{s}_1,\mathbf{s}_2\}$.*

Setting $\delta = min_{\Omega,\Omega'\in\Upsilon,\Omega\neq\Omega'} d(\Omega,\Omega')$, provided that $\max_{\Omega,\Omega'\in\Upsilon}\mathcal{K}(\mathbb{P}_\Omega^{\otimes n},\mathbb{P}_{\Omega'}^{\otimes n}) \le \kappa_1\log|\Upsilon|$, the following lower bound holds ,

$$\inf_{\widehat{\Omega}}\sup_{\Omega\in\Upsilon}\mathbb{E}_\Omega[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)] \ge \kappa_2\delta \ ,$$

where κ_1 and κ_2 are two numerical constants.

Proof of Lemma 5.27. This lemma is mainly based on an application of Birgé's version of Fano's lemma [Bir05]. We provide a statement of the result that is taken from [Mas07] Sect. 2.4.

Lemma 5.28. *Let $(\mathbb{P}_i)_{0\le i\le N}$ be some family of probability distributions and $(A_i)_{0\le i\le N}$ be some family of disjoint events. Let $a=\min_{0\le i\le N}\mathbb{P}_i(A_i)$, then*

$$a \le \kappa\vee\left(\frac{\max_{1\le i\le N}\mathcal{K}\left(\mathbb{P}_i;\mathbb{P}_0\right)}{\log(1+N)}\right) \ .$$

Let $\widehat{\Omega}$ be an estimator of Ω. We define an estimator $\widetilde{\Omega}$ taking its values in Υ such that

$$d(\widetilde{\Omega},\widehat{\Omega}) = \min_{\Omega'\in\Upsilon} d(\Omega',\widehat{\Omega}) \ .$$

We define $(\widetilde{T},\widetilde{S})$ and $(\widehat{T},\widehat{S})$ the Cholesky decompositions of $\widetilde{\Omega}$ and $\widehat{\Omega}$. Let i be an integer between 1 and p. By the triangle inequality,

$$\frac{\|t_i-\widetilde{t}_i\|_{l_2}^2}{4} \le 2\left[\frac{\|t_i-\widehat{t}_i\|_{l_2}^2}{4}+\frac{\|\widehat{t}_i-\widetilde{t}_i\|_{l_2}^2}{4}\right] \ .$$

For any positive number a and b, $\log(1+a+b)\le log(1+a)+\log(1+b)$. Moreover, for any positive number a, $\log(1+2a)\le 2\log(1+a)$ because the log function is concave. Hence, we get

$$\log\left[1+\frac{\|t_i-\widetilde{t}_i\|_{l_2}^2}{4}\right] \le 2\log\left[1+\frac{\|t_i-\widehat{t}_i\|_{l_2}^2}{4}\right] + 2\log\left[1+\frac{\|\widehat{t}_i-\widetilde{t}_i\|_{l_2}^2}{4}\right] \ . \tag{5.38}$$

Let us define the function f as $f(x):=x-\log(x)-1$ for any positive number x and let us prove that there exists some numerical constant L such that

$$f\left(\frac{s_i}{\widetilde{s}_i}\right) \le L\left[f\left(\frac{s_i}{\widehat{s}_i}\right)+f\left(\frac{\widetilde{s}_i}{\widehat{s}_i}\right)\right] \ . \tag{5.39}$$

If $s_i=\widetilde{s}_i$, this inequality holds for any positive L since $f(1)=0$ and f is non negative. If $s_i\neq\widetilde{s}_i$, there are two possibilities: either $s_i=\mathbf{s}_1$ and $\widetilde{s}_i=\mathbf{s}_2$ or $s_i=\mathbf{s}_2$ and $\widetilde{s}_i=\mathbf{s}_1$. By deriving $f(\frac{\mathbf{s}_1}{x})+f(\frac{\mathbf{s}_2}{x})$, one observes that this sum is minimized for $x=\frac{\mathbf{s}_1+\mathbf{s}_2}{2}$ and that this minimum equals $f\left[2/\left(1+\frac{\mathbf{s}_1}{\mathbf{s}_2}\right)\right]+f\left[2/\left(1+\frac{\mathbf{s}_2}{\mathbf{s}_1}\right)\right]$. Hence, we obtain that

$$f\left(\frac{s_i}{\widetilde{s}_i}\right) \le \frac{f\left(\frac{\mathbf{s}_1}{\mathbf{s}_2}\right)\vee f\left(\frac{\mathbf{s}_2}{\mathbf{s}_1}\right)}{f\left[2/\left(1+\frac{\mathbf{s}_1}{\mathbf{s}_2}\right)\right]+f\left[2/\left(1+\frac{\mathbf{s}_2}{\mathbf{s}_1}\right)\right]}\left[f\left(\frac{s_i}{\widehat{s}_i}\right)+f\left(\frac{\widetilde{s}_i}{\widehat{s}_i}\right)\right] \ .$$

Since $\mathbf{s}_1$ and $\mathbf{s}_2$ lie between one and two, it follows that

$$f\left(\frac{s_i}{\widetilde{s}_i}\right) \leq \sup_{1/2\leq x\leq 2} \frac{f(x)}{f\left[2/(1+x)\right]+f\left[2/(1+\frac{1}{x})\right]} \left[f\left(\frac{s_i}{\widehat{s}_i}\right)+f\left(\frac{\widetilde{s}_i}{\widehat{s}_i}\right)\right] . \tag{5.40}$$

The ratio $f(x)/(f\left[2/(1+x)\right]+f\left[2/(1+\frac{1}{x})\right])$ is positive and continuous on $[1/2;1[$ and $]1;2]$. By studying the Taylor series of $f(x)$ at x equals one, we observe that $f(x)=(x-1)^2/2+o[(x-1)^2]$, $f(2/(1+x))=(x-1)^2/8+o[(x-1)^2]$, and $f(2/(1+1/x))=(x-1)^2/8+o[(x-1)^2]$. Hence, there exists a continuation of the ratio $f(x)/(2f(\sqrt{x}))$ around one. The supremum (5.40) is therefore finite and the upper bound (5.39) holds.

Combining the upper bounds (5.38) and (5.39) with the definition of $\widetilde{\Omega}$ yields

$$\begin{aligned} d(\Omega,\widetilde{\Omega}) &\leq 2\sum_{i\in A}\left[\log\left[1+\frac{\|t_i-\widehat{t}_i\|_{l_2}^2}{4}\right]+\log\left[1+\frac{\|\widehat{t}_i-\widetilde{t}_i\|_{l_2}^2}{4}\right]\right]+L\sum_{i\in A^c}\left[f\left(\frac{s_i}{\widehat{s}_i}\right)+f\left(\frac{s_i}{\widetilde{s}_i}\right)\right] \\ &\leq L\left[d(\Omega,\widehat{\Omega})+d(\widetilde{\Omega},\widehat{\Omega})\right] \leq Ld(\Omega,\widehat{\Omega}) . \end{aligned}$$

Hence, one may lower bound the risk of $\widehat{\Omega}$ as follows

$$\sup_{\Omega\in\Upsilon}\mathbb{E}_\Omega\left[d\left(\Omega,\widehat{\Omega}\right)\right] \geq L^{-1}\delta\sup_{\Omega\in\Upsilon}\mathbb{P}_\Omega\left[\Omega\neq\widetilde{\Omega}\right] = L^{-1}\delta\left(1-\min_{\Omega\in\Upsilon}\mathbb{P}_\Omega\left[\Omega=\widehat{\Omega}\right]\right) .$$

Applying Lemma 5.28, we conclude that

$$\inf_{\widehat{\Omega}}\sup_{\Omega\in\Upsilon}\mathbb{E}_\Omega\left[d\left(\Omega,\widehat{\Omega}\right)\right] \geq L^{-1}(1-\kappa)\delta , \tag{5.41}$$

if $\max_{\Omega,\Omega'\in\Upsilon}\mathcal{K}(\mathbb{P}_\Omega^{\otimes n},\mathbb{P}_{\Omega'}^{\otimes n})\leq\kappa\log|\Upsilon|$.

Let us now express this minimax lower bound in term of Kullback divergence. Thanks to the chain rule and Lemma 5.21, the Kullback divergence between two positive matrices Ω and Ω' decomposes as

$$\mathcal{K}(\Omega;\Omega') = \sum_{i=1}^p \frac{1}{2}\left[\log\frac{s_i'}{s_i}+\frac{s_i}{s_i'}-1+\frac{l_i(t_i,t_i')}{s_i'}\right] .$$

Straightforward computations allow to prove that the function $\log\frac{s_i'}{s_i}+\frac{s_i}{s_i'}-1+\frac{l_i(t_i,t_i')}{s_i'}$ is minimized with respect to s_i' when $s_i'=s_i+l_i(t_i,t_i')$. This leads to the lower bound

$$\log\frac{s_i'}{s_i}+\frac{s_i}{s_i'}-1+\frac{l_i(t_i,t_i')}{s_i'} \geq \log\left(1+\frac{l_i(t_i,t_i')}{s_i}\right) .$$

By Definition (5.25) of $l_i(.,.)$ the quantity $l_i(t_i,t_i')$ is lower bounded by $\left[\phi_{\max}(\Omega)\right]^{-1}\|t_i-t_i'\|_{l_2}^2$. By assumption, $\left[\phi_{\max}(\Omega)\right]^{-1}$ is larger than $1/2$ for any $\Omega\in\Upsilon$. Moreover, s_i is smaller than 2. We conclude that for any $\Omega\in\Upsilon$ and any positive matrix Ω', the following lower bound holds

$$2\mathcal{K}(\Omega;\Omega') \geq \sum_{i\in A}\log\left(1+\frac{\|t_i-t_i'\|_{l_2}^2}{4}\right)+\sum_{i\in A^c}\frac{s_i}{s_i'}-\log\left(\frac{s_i}{s_i'}\right)-1 = d(\Omega,\Omega') .$$

We conclude by gathering this last bound with (5.41). □

5.7.4.1 Adaptive banding

Proof of Proposition 5.9. Let r be a positive number. Let $\mathcal{T}'_{\text{ord}}[k_1,\ldots,k_p,r]$ be a maximal subset of $\mathcal{T}_{\text{ord}}[k_1,\ldots,k_p,r]$ which satisfies the property: "for any two different elements T and T' of $\mathcal{T}'_{\text{ord}}[k_1,\ldots,k_p,r]$, the Hamming distance $d_H(T,T')$ is larger than $\sum_{1\leq i\leq p}k_i/4$".

By Lemma 5.26, this set satisfies $\log\left[\text{Card}\left(\mathcal{T}'_{\text{ord}}[k_1,\ldots,k_p,r]\right)\right]\geq\sum_{2\leq i\leq p}k_i/8$. Let T be a matrix in $\mathcal{T}'_{\text{ord}}[k_1,\ldots,k_p,r]$. Standard computations allow to prove that the diagonal elements of T^*T lie between

1 and $1 + kr^2$. Besides, the sum of the absolute values of the off-diagonal elements on each line is upper bounded as follows. Let i be an integer between 1 and p.

$$\begin{aligned}\sum_{j\neq i} |T^*T[i,j]| = \sum_{l=1}^{p}\sum_{j\neq i} |T[l,i]T[l,j]| &\leq \sum_{j\neq i} T[i,j] + \sum_{j\neq i} T[j,i] + \sum_{j\neq i}\sum_{j\neq l\neq i} T[l,i]T[l,j] \\ &\leq 2kr + k^2r^2 \ .\end{aligned}$$

If r is smaller than $1/(8k)$, the matrices T^*T are diagonally dominant and their eigenvalues lie between 5/8 and 1.3. Let us define the subset A of $\{1,\ldots,p\}$ as $A := \{i,\ k_i > 0\}$. We introduce the subset $\mathcal{S}[A,p,r]$ as

$$\mathcal{S}[A,p,r] := \{S \in \mathrm{Diag}(p),\ S[i,i] = 1 \text{ if } i \in A \text{ and } S[i,i] = 1 \text{ or } 1+r \text{ if } i \in A^c\} \ .$$

Applying again Lemma 5.26, we define a subset $\mathcal{S}'[A,p,r]$ of $\mathcal{S}[A,p,r]$ such that $\log\left[\mathrm{Card}\left(\mathcal{S}'[A,p,r]\right)\right] \geq \log\left[\mathrm{Card}(A^c)\right]/8$ and such that its elements are $\mathrm{Card}(A^c)/4$-separated with respect to the Hamming distance. If r is smaller than 0.5, then the eigenvalues of any matrix in $\mathcal{S}'[A,p,r]$ are between 1 and 1.5. Finally, we define the set $\mathcal{U}'_{\mathrm{ord}}[k_1,\ldots,k_p,r]$ as

$$\mathcal{U}'_{\mathrm{ord}}[k_1,\ldots,k_p,r] := \{T^*ST,\ T \in \mathcal{T}'_{\mathrm{ord}}[k_1,\ldots,k_p,r] \text{ and } S \in \mathcal{S}'[A,p,r]\} \ .$$

We thus easily lower bound its cardinality

$$\log\left[\mathrm{Card}\left(\mathcal{U}'_{\mathrm{ord}}[k_1,\ldots,k_p,r]\right)\right] \geq \left(\mathrm{Card}(A^c) + \sum_{i=1}^{p} k_i\right)/8 \geq \left(p + \sum_{i=1}^{p} k_i\right)/16 \ .$$

Moreover, if $r \leq \frac{1}{8k}$, the eigenvalues of any matrix in this set are between 1/2 and 2. Let us upper bound the Kullback entropy between any two elements $\Omega = T^*S^{-1}T$ and $\Omega' = T'^*S'^{-1}T'$ of $\mathcal{U}'_{\mathrm{ord}}[k_1,\ldots,k_p,r]$.

$$2\mathcal{K}\left(\Omega;\Omega'\right) = \sum_{i\in A} \frac{l_i\left(t_i,t'_i\right)}{s'_i} + \sum_{i\in A^c} \frac{s_i}{s'_i} + \log\left(\frac{s_i}{s'_i}\right) - 1 \ .$$

If $i \in A$, then $s'_i = 1$. Besides, $l_i(t_i,t'_i) \leq [\phi_{\min}(\Omega)]^{-1}\|t_i - t'_i\|_{l_2}^2 \leq 2k_ir^2$. Recalling that the function f is defined by $f(x) = x - \log x - 1$ and that $r \leq 1/8$, straightforward computations lead to $f\left(\frac{s'_i}{s_i}\right) \leq Lr^2$. Hence, for any $(\Omega_1,\Omega_2) \in \mathcal{U}'_{\mathrm{ord}}[k_1,\ldots,k_p,r]$, it holds that $\mathcal{K}\left(\mathbb{P}_{\Omega_1}^{\otimes n};\mathbb{P}_{\Omega_2}^{\otimes n}\right) \leq L\left(p + \sum_{i=2}^{p} k_i\right)r^2$. Moreover, it holds that $f(1+r) \geq Lr^2$ and $f((1+r)^{-1}) \geq Lr^2$ since $f(1+x) = x^2/2 + o(x^2)$ and $r \leq 1/8$. If $\Omega_1 \neq \Omega_2$, then $d(\Omega_1,\Omega_2)$ is lower bounded as follows

$$\begin{aligned} d(\Omega_1,\Omega_2) &\geq \sum_{i\in A} \log\left(1 + \frac{k_ir^2}{16}\right) + \frac{\mathrm{Card}\left(A^c\right)}{4}\left[f(1+r) \wedge f(1/(1+r))\right] \\ &\geq L\left[\sum_{i\in A} k_i + \mathrm{Card}\left(A^c\right)\right]r^2 \geq L\left[\sum_{i=1}^{p} k_i + p\right]r^2 \ , \end{aligned}$$

since r is smaller than $1/8$ and $k_ir^2/16$ is smaller than $1/64$. Hence, as long as $r \leq L\frac{1}{\sqrt{n}} \wedge \frac{1}{8k} \wedge \frac{1}{2}$, one may apply Lemma 5.27.

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{U}_{ord}[k_1,\ldots,k_p,r]} \mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] \geq L\left[\sum_{i=2}^{p} k_i + p\right]\left(r^2 \wedge \frac{1}{n} \wedge \frac{1}{k^2}\right) \ .$$

By assumption, $1/k^2$ is larger than $1/n$ and the result follows.

□

Proof of Proposition 5.10. The lower bound (5.15) is a consequence of Proposition 5.9. Let k be a positive integer smaller than $\lfloor\sqrt{n}\rfloor \wedge (p-1)$. Let r be a positive number smaller than $\sqrt{\frac{a_k^2R^2}{k}}$. We consider the set $\mathcal{U}_{\mathrm{ord}}[0,1,\ldots,k-1,k,\ldots,k,r]$. Let (T,S) refer to the Cholesky decomposition of a matrix Ω belonging to this set. By definition of $\mathcal{U}_{\mathrm{ord}}$,

$$\sum_{j=1}^{i-1} \frac{t[i,i-j]^2}{a_j^2} = \sum_{j=1}^{k} \frac{t[i,i-j]^2}{a_j^2} \leq a_k^2kr^2 \leq R^2 \ .$$

Hence, the set $\mathcal{U}_{\rm ord}[0,1,\ldots,k-1,k,\ldots,k,r]$ is included in $\mathcal{E}(a,R,p)$. By Proposition 5.9, we obtain the minimax lower bound.

$$\begin{aligned}\inf_{\widehat{\Omega}}\sup_{\Omega\in\mathcal{E}(a,R,p)}\mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] &\geq Lp(k+1)\left(\frac{a_k^2R^2}{k}\wedge\frac{1}{n}\right)\\ &\geq Lp\left(a_k^2R^2\wedge\frac{k+1}{n}\right)\ .\end{aligned}$$

Similarly if $k=0$, the set $\mathcal{U}_{\rm ord}[0,\ldots,0,+\infty]$ is included in $\mathcal{E}(a,R,p)$ and the minimax rates of estimation over $\mathcal{E}(a,R,p)$ is lower bounded by $Lp(a_0R^2\wedge\frac{1}{n})$ with the convention $a_0=+\infty$. Taking the infimum for all non-positive integers smaller than $\lfloor\sqrt{n}\rfloor\wedge(p-1)$ yields the first result.

Let us now turn to the second part of the proposition. Let γ be some number larger than two. The previous lower bound still holds if Ω also belongs to $\mathcal{B}_{\rm op}(\gamma)$. Indeed, the matrices Ω considered in the proof of Proposition 5.9 for bounding the minimax rates of estimation over set of the type $\mathcal{U}_{\rm ord}[0,1,\ldots,k-1,k,\ldots,k,r]$ have their eigenvalues between $1/2$ and 2. Hence, we get:

$$\inf_{\widehat{\Omega}}\sup_{\Omega\in\mathcal{E}(a,R,p)\cap\mathcal{B}_{\rm op}(\gamma)}\mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] \geq Lp\sup_{k=0,\ldots,\lfloor\sqrt{n}\rfloor\wedge(p-1)}\left(a_k^2R^2\wedge\frac{k+1}{n}\right)\ . \tag{5.42}$$

Let k be a non-negative integer smaller or equal to $d\wedge(p-1)$, where d is the maximal dimension of the models defining the estimator $\widetilde{\Omega}^d_{\rm ord}$. We consider the model

$$m=(\emptyset,\{1\},\ldots,\{i-1,\ldots,i-k\},\ldots,\{p-1,\ldots,p-k\})$$

in $\mathcal{M}^d_{\rm ord}$ which corresponds to the k-th banded matrices T. By Corollary 5.6, the risk of $\widetilde{\Omega}^d_{\rm ord}$ is upper bounded by

$$\mathbb{E}\left[\mathcal{K}\left(\Omega;\widetilde{\Omega}^d_{\rm ord}\right)\right]\leq L(K,\eta)\left[\frac{p(k+1)}{n}+\mathcal{K}\left(\Omega;\Omega_m\right)\right]+\tau_n(\Omega,K,\eta)\ . \tag{5.43}$$

Let us upper bound the bias term $\mathcal{K}\left(\Omega;\Omega_m\right)$. By Equation (5.27), it decomposes as

$$\begin{aligned}2\mathcal{K}\left(\Omega;\Omega_m\right) &= \sum_{i=1}^p\frac{s_i}{s_{i,m_i}}-\log\left(\frac{s_i}{s_{i,m_i}}\right)-1+\frac{l_i(t_i,t_{i,m_i})}{s_{i,m_i}}\\ &= \sum_{i=1}^p\log\left(1+\frac{l_i(t_i,t_{i,m_i})}{s_i}\right)\ \leq\sum_{i=1}^p\frac{l_i(t_i,t_{i,m_i})}{s_i}\ ,\end{aligned}$$

since we have mentioned in the proof of Lemma 5.27 that $s_{i,m_i}=s_i+l_i(t_i,t_{i,m_i})$.

Let i be an integer between $k+2$ and p (if there exists one). We define t'_{i,m_i} as the projection of t_i with respect to the canonical norm in $\mathbb{R}^{i-1}$. Since Ω belongs to $\mathcal{B}_{\rm op}(\gamma)$, it follows that s_i is larger than $1/\gamma$ and that the largest eigenvalue of Ω^{-1} is smaller than γ. Hence, we obtain that

$$\begin{aligned}\frac{l_i(t_i,t_{i,m_i})}{s_i}\leq\gamma l_i(t_i,t'_{i,m_i}) &\leq \gamma^2\left[\sum_{j=1}^{i-1}\left(t_i[i-j]-t'_{i,m_i}[i-j]\right])^2\right]\\ &= \gamma^2\left[\sum_{j=k+1}^{i-1}t_i[i-j]^2\right]\leq\gamma^2a_{k+1}^2R^2\ .\end{aligned}$$

If i is smaller or equal to $k+2$, then $l_i(t_i,t_{i,m_i})=0$. Combining this upper bound with (5.43), we get

$$\mathbb{E}\left[\mathcal{K}\left(\Omega;\widetilde{\Omega}^d_{\rm ord}\right)\right]\leq L(K,\eta,\gamma)p\left[\frac{(k+1)}{n}+a_{k+1}R^2\right]+\tau_n(\Omega,K,\eta)\ .$$

Since Ω belong to $\mathcal{B}_{\rm op}(\gamma)$

$$\begin{aligned}2\mathcal{K}(\Omega;I_p)/p &= 1/p\sum_{i=1}^p\varphi_i(\Omega)-\log\left(\varphi_i(\Omega)\right)-1\\ &\leq [\varphi_{\min}(\Omega)-\log\left(\varphi_{\min}(\Omega)\right)-1]\vee[\varphi_{\max}(\Omega)-\log\left(\varphi_{\max}(\Omega)\right)-1]\leq L(\gamma)\ .\end{aligned}$$

Hence, the term $\tau_n(\Omega, K, \eta)$ is smaller than some $L(K, \eta, \gamma)\frac{p}{n}$. For n larger than some universal constant, the largest dimension d in the model collection that defines $\widetilde{\Omega}^d_{\rm ord}$ is larger than $\lfloor\sqrt{n}\rfloor$. Taking the infimum over k in $0, \ldots, \lfloor\sqrt{n}\rfloor \wedge (p-1)$, we conclude that

$$\mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{\rm ord}\right)\right] \leq L(K, \eta, \gamma, \beta)p \inf_{k=1,\ldots,\lfloor\sqrt{n}\rfloor\wedge(p-1)} \left(a^2_{k+1}R^2 + \frac{k+1}{n}\right) .$$

Let us define $d^* := \sup\left\{d' \geq 0 \text{ s.t. } \frac{d'+1}{n} \leq a_{d'}R^2\right\}$. By assumption, d^* is smaller or equal to $\lfloor\sqrt{n}\rfloor$. Hence,

$$\begin{aligned}\mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{\rm ord}\right)\right] &\leq L(K, \eta, \gamma, \beta)p\left(a^2_{d^*+1}R^2 + \frac{d^*+1}{n}\right) \\ &\leq L(K, \eta, \gamma, \beta) \inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{E}(a,R,p)\cap\mathcal{B}_{\rm op}(\gamma)} \mathbb{E}\left[\mathcal{K}\left(\Omega; \widehat{\Omega}\right)\right] ,\end{aligned}$$

thanks to Equation (5.42).

□

5.7.4.2 Complete graph selection

Let $T \in \mathcal{T}_1[k, p, r]$ be the set of lower triangular matrices of size p with a unit diagonal and such that each line contains at most k non-zero off-diagonal entries. These entries are also smaller than r in absolute value. We first provide a minimax lower bound on the minimax risk over $T \in \mathcal{T}_1[k, p, r]$.

Proposition 5.29. *Let k and p be two positive integers such that $k \leq p$. Assume that $n \geq Lk^2[1 + \log(p/k)]$, where L is some universal constant exhibited in the proof. Then, for any $r > 0$, the minimax rates of estimation over the set $\mathcal{T}_1[k, p, r]$ is lower bounded as follows*

$$\inf_{\widehat{\Omega}} \sup_{T\in\mathcal{T}_1[k,p,r]} \mathcal{K}\left(T^*T; \widehat{\Omega}\right) \geq Lkp\left[r^2 \wedge \frac{1 + \log\left(\frac{p}{k}\right)}{n}\right] . \tag{5.44}$$

We believe that the condition $n \geq Lk^2[1 + \log(p/k)]$ is essentially technical but we do not know how to remove it. Thanks to Corollary 5.12, we may easily derive the minimax rates of estimation over the sets $\mathcal{U}_1[k, p]$. Let us first provide the proof of Proposition 5.29 and then derive the proof of Proposition 5.13.

Proof of Proposition 5.29. Assume first that k is a power of 2, that $2k$ divides p and that $\log(p/k)$ is larger than 19. Let us consider the set $\mathcal{T}^{(1)}_1[k, p, r]$ of lower triangular square matrices T of size p such that:

- the diagonal of T is made of 1,
- the lower left submatrix of T of size $p/2$ contains exactly k entries that equal r on each line and on each column,
- every other entry of T is zero.

Clearly, $\mathcal{T}^{(1)}_1[k, p, r]$ is in one to one correspondence with the set $\Theta[k, p/2]$ of binary square matrices of size $p/2$ that contain exactly k non-zero coefficients on each line and each column.

Let T be a matrix that belongs to $\mathcal{T}^{(1)}_1[k, p, r]$. We claim that as long as r is smaller than $1/8k$, the eigenvalues of T^*T are between $1/2$ and 2. Indeed, the diagonal elements of T^*T are all between 1 and $1 + kr^2$. Besides, the sum of the off-diagonal elements is upper bounded by

$$\begin{aligned}\sum_{j\neq i} |T^*T[i, j]| = \sum_{l=1}^{p}\sum_{j\neq i} |T[l, i]T[l, j]| &\leq \sum_{j\neq i} T[i, j] + \sum_{j\neq i} T[j, i] + \sum_{j\neq i}\sum_{j\neq l\neq i} T[l, i]T[l, j] \\ &\leq 2kr + k^2r^2 .\end{aligned}$$

Hence, if $r \leq 1/8k$, the matrix T is diagonally dominant and the sum of the off diagonal terms is smaller than $3/8$ whereas the diagonal term is between 1 and $1 + 1/8$.

Let T and T' be two elements of $\mathcal{T}_1^{(1)}[k,p,r]$. Let us upper bound the Kullback entropy between the corresponding precision matrices.

$$2\mathcal{K}\left(T^*T; T'^*T'\right) = \sum_{i=1}^{p} l_i(t_i, t_i') \leq \sum_{i=p/2+1}^{p} \varphi_{\max}(T^*T)\|t_i - t_i'\|_{l_2}^2 \leq 4kpr^2 \ . \tag{5.45}$$

Lemma 5.30. *Assume that* $\log(p/k)$ *is larger than* 19. *Let* $\Theta[k,p/2]$ *be equipped with Hamming distance* d_H. *There exists some subset* $\Theta'[k,p/2]$ *of* $\Theta[k,p/2]$ *with the following properties*

$$d_H(\theta,\theta') > pk/4 \text{ for every } (\theta,\theta') \in \Theta'^2 \text{ with } \theta \neq \theta' \text{ and } \log\left[Card(\Theta')\right] \geq kp/20 \log\left(\frac{p}{k}\right) \ . \tag{5.46}$$

The proof of this lemma is postponed to the end of this subsection. By Lemma 5.30, there exists some subset $\mathcal{T}_1^{(2)}[k,p,r]$ of $\mathcal{T}_1^{(1)}[k,p,r]$ such that $d_H(T,T') \geq pk/8$ for every $(T,T') \in \mathcal{T}_1^{(2)}[k,p,r]$ with $T \neq T'$ and

$$\log\left[\mathrm{Card}\left(\mathcal{T}_1^{(2)}[k,p,r]\right)\right] \geq kp/20 \log\left(\frac{p}{k}\right) \ . \tag{5.47}$$

Let us define $A = \{1, \dots p\}$ and let us consider the function $d(.,.)$ defined in Lemma 5.27. Observe that $2kr^2 \leq 1/32$ since $r \leq 1/(8k)$. By the mean value theorem, we obtain that $\log(1 + x/4) \geq \frac{x}{8}$ for any positive number x smaller than $2kr^2$. Hence, we get

$$d(T^*T, T'^*T') = \sum_{i=1}^{p} \log\left[1 + \frac{d_H(t_i,t_i')r^2}{4}\right] \geq \sum_{i=1}^{p} \frac{d_H(t_i,t_i')r^2}{8} \tag{5.48}$$

$$\geq d_H(T,T')\frac{r^2}{8} \geq \frac{pkr^2}{32} \ , \tag{5.49}$$

for any $T \neq T'$ in $\mathcal{T}_1^{(2)}[k,p,r]$. We are now in position to apply Lemma 5.27 to $\mathcal{T}_1^{(2)}[k,p,r]$ with the bounds (5.45), (5.47), and (5.48).

$$\inf_{\widehat{\Omega}} \sup_{T \in \mathcal{T}_1^{(1)}[k,p,r]} \mathbb{E}\left[\mathcal{K}\left(T^*T; \widehat{\Omega}\right)\right] \geq \frac{\kappa_2}{64} pkr^2 \ ,$$

as long as $4kpnr^2 \leq \kappa_1 kp/20 \log\left(\frac{p}{k}\right)$ and $r \leq \frac{1}{8k}$. This yields

$$\begin{aligned} \inf_{\widehat{\Omega}} \sup_{T \in \mathcal{T}_1^{(1)}[k,p,r]} \mathbb{E}\left[\mathcal{K}\left(T^*T; \widehat{\Omega}\right)\right] &\geq Lpk\left[r^2 \wedge \frac{\log\left(\frac{p}{k}\right)}{n}\right] \\ &\geq Lpk\left[r^2 \wedge \frac{1 + \log\left(\frac{p}{k}\right)}{n}\right] \ , \end{aligned}$$

since $n \geq k^2[1 + \log(p/k)]$ and $\log(p/k)$ is assumed to be larger than 19.

We now turn to the case where k is not a power of 2 or $2k$ does not divide p. We only assume that $\log(p/k)$ is larger than $19 + \log(2)$. Let us define $k' := 2^{\lfloor \log_2 k \rfloor}$ and p' as the largest integer that is divided by $2k'$ and is smaller than p. Here $\log_2$ refers to the function $\log(.)/\log(2)$. It follows from this definition and the assumption that k' is between $k/2$ and k and p' is larger than $p/2$. Thus, $\log(p'/k')$ is larger than $\log(p/2k) \geq 19$. Let $\mathcal{T}_1^{(1)}[k',p',p,r]$ denote the set of lower triangular matrices T such that the diagonal elements of T equal 1, the lower left submatrix of T of size $p'/2$ contains exactly k' entries that equal r on each line and on each column and such that every other entry of T is zero. Arguing as in the first case, we obtain that

$$\begin{aligned} \inf_{\widehat{\Omega}} \sup_{T \in \mathcal{T}_1[k,p,r]} \mathbb{E}\left[\mathcal{K}\left(T^*T; \widehat{\Omega}\right)\right] &\geq Lp'k'\left[r^2 \wedge \frac{1 + \log\left(\frac{p'}{k'}\right)}{n}\right] \\ &\geq Lpk\left[r^2 \wedge \frac{1 + \log\left(\frac{p}{k}\right)}{n}\right] \ . \end{aligned}$$

Finally, we consider the situation where the ratio $\log(p/k)$ is smaller than $19 + \log(2)$. Observe that the set $\mathcal{T}_{\rm ord}[(0, 1, \ldots, k-1, k, \ldots, k), r]$ is included in $\mathcal{T}_1[k, p, r]$. If we choose the set A to be $\{1, \ldots, p\}$, then a slight modification in the proof of Proposition 5.9 allows to show the minimax lower bound:

$$\inf_{\widehat{\Omega}} \sup_{T \in \mathcal{T}_{\rm ord}[(0,\ldots,k,\ldots,k),r]} \mathcal{K}\left[T^*T; \widehat{\Omega}\right] \geq Lkp\left[r^2 \wedge \frac{1}{n}\right] ,$$

as long as $k \leq \sqrt{n}$. Hence, it follows that

$$\inf_{\widehat{\Omega}} \sup_{T \in \mathcal{T}_1[k,p,r]} \mathcal{K}\left[T^*T; \widehat{\Omega}\right] \quad \geq \quad Lkp\left[r^2 \wedge \frac{1}{n}\right] \geq Lkp\left[r^2 \wedge \frac{1 + \log\left(\frac{p}{k}\right)}{n}\right] ,$$

since $\log(p/k)$ is smaller than $19 + \log(2)$.

□

Proof of Proposition 5.13. We derive from the proof of Proposition 5.29 a minimax lower bound over $\mathcal{U}_1[k, p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)$. First, we consider the case where k is a power of 2, $2k$ divides p and $\log(p/k)$ is larger than 19. Let us set $r^2 = \left(\frac{1+\log(p/k)}{n}\right) \wedge \left(\frac{1}{8k}\right)^2$. In the previous proof, we have shown that

$$\inf_{\widehat{\Omega}} \inf_{T \in \mathcal{T}_1^{(1)}[k,p,r]} \mathbb{E}\left[\mathcal{K}\left(T^*T; \widehat{\Omega}\right)\right] \geq Lkp\left(r^2 \wedge \frac{1 + \log(p/k)}{n}\right) \geq Lkp\frac{1 + \log(p/k)}{n} ,$$

since $n \geq k^2(1+\log(p/k)$. Moreover, we have mentioned that for any matrix T in $\mathcal{T}_1^{(1)}[k, p, r]$, $\varphi_{\min}(T^*T) \geq 1/2$. Let us now upper bound the Kullback divergence with the identity matrix.

$$\begin{aligned} \mathcal{K}\left(T^*T; I_p\right) &\leq \frac{1}{2}\sum_{i=2}^{p} l_i(t_i, 0_{i-1}) \leq \frac{\varphi_{\min}(T^*T)}{2}\|T - I_p\|_F^2 \\ &\leq kpr^2 \leq p \leq pn^\beta . \end{aligned}$$

Hence, the set $\left\{T^*T,\, T \in \mathcal{T}_1^{(1)}[k, p, r]\right\}$ is included in $\mathcal{U}_1[k, p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)$ and the lower bound follows.

The case where k is not a power of 2 or $2k$ does not divide p is handled similarly if one uses the set $\mathcal{T}_1^{(1)}[k', p', p, r]$ defined in the proof of Proposition 5.29. Finally, one uses the set $\mathcal{T}_{\rm ord}[(0, \ldots, k, \ldots, k), r]$ if $\log(p/k) \leq 19 + \log(2)$.

Let us turn to the upper bound on the risk. By Proposition 5.12, the estimator $\widetilde{\Omega}_{\rm co}^d$ satisfies

$$\begin{aligned} \mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}_{\rm co}^d\right)\right] &\leq L(K, \eta)pk\frac{1 + \log\left(\frac{p}{k}\right)}{n} + L(K, \eta)pn^{5/2}\left[1 + \mathcal{K}(\Omega; I_p)\right]\exp\left[-nL(K, \eta)\right] \\ &\leq L(K, \eta, \beta)pk\frac{1 + \log\left(\frac{p}{k}\right)}{n} , \end{aligned}$$

for any $\Omega \in \mathcal{U}_1[k, p] \cap \mathcal{B}_{\mathcal{K}}(n^\beta)$. We conclude by gathering the lower and the upper bounds.

□

Proof of Proposition 5.14. This proof follows the same sketch as the proof of Proposition 5.29. Let k be an integer smaller than $p/2$. Observe that it is sufficient to prove the result (5.22) for all k smaller than $p/2$ since this lower bound is up to a numerical constant. Assume first that $\log(p)$ is larger than 21.

Let us define the set $A := \{p - k + 1, \ldots, p\}$. We need to build a suitable subset of $\mathcal{U}_2[k, p]$ that is well separated with respect to the function $d(.,.)$ introduced in Lemma 5.27. Let r_1 and r_2 be two positive numbers respectively smaller than $1/4$ and $1/8$. We shall fix them later.

Let us first introduce the set $\mathcal{S}[A^c, p, r_1]$ of diagonal matrices S such that $S[i, i] = 1$ if $i \in A$ and $S[i, i]$ is either 1 or $1 + r_1$ if $i \in A^c$. The cardinality of this set is $2^{\mathrm{Card}(A^c)}$. Applying Lemma 5.26, there exists a subset $\mathcal{S}'[A^c, p, r_1]$ that satisfies $\log\left[\mathrm{Card}\left(\mathcal{S}'[A^c, p, r_1]\right)\right] \geq \mathrm{Card}(A^c)/8$ and such that any two elements of $\mathcal{S}'[A^c, p, r_1]$ are $\mathrm{Card}(A^c)/4$ separated with respect to the Hamming distance d_H.

Let us now introduce the set $\mathcal{T}_2^{(1)}[k,p,r_2]$ of lower triangular matrices T of size p with unit diagonal and such that the $k \times \lfloor p/2 \rfloor$ lower left submatrix of T contains exactly one r_2 on each line and at most one r_2 on each column. Every other entries of T is zero.

Let us consider the set $\mathcal{T}_2^{(1)}[k,p,r]$ of lower triangular square matrices T of size p such that: of size p such that:

- the diagonal of T is made of 1,
- the lower left submatrix of T of size $k \times \lfloor p/2 \rfloor$ contains exactly one entry that equals r_2 on each line and at most one on each column.
- every other entry of T is zero.

Lemma 5.31. *Assume that* $\log p \geq 21$. *There exists some subset* $\mathcal{T}_2^{(2)}[k,p,r_2]$ *of* $\mathcal{T}_2^{(1)}[k,p,r_2]$ *such that the Hamming distance between any two different elements of* $\mathcal{T}_2^{(2)}[k,p,r_2]$ *is larger than* $k/2$ *and such that* $\log\left[Card\left(\mathcal{T}_2^{(2)}[k,p,r_2]\right)\right] \geq k\log(p)/10$.

We now define the subset $\mathcal{U}_2'[k,p,r_1,r_2]$ of $\mathcal{U}_2[k,p]$ as

$$\mathcal{U}_2'[k,p,r_1,r_2] := \left\{\Omega = T^* S^{-1} T \ , \quad T \in \mathcal{T}_2^{(2)}[k,p,r_2] \text{ and } S \in \mathcal{S}'[A^c,p,r_1]\right\} \ .$$

In order to apply Lemma 5.27, we need to bound the eigenvalues of the matrices in $\mathcal{U}_2'[k,p,r_1,r_2]$, lower bound the function $d(.,.)$ defined in (5.37), upper bound the Kullback divergence between elements of $\mathcal{U}_2'[k,p,r_1,r_2]$ and lower bound the cardinality.of $\mathcal{U}_2'[k,p,r_1,r_2]$.

1. Let us first bound the smallest and the largest eigenvalues of the matrices Ω in this set. Let T and S correspond to the Cholesky decomposition of Ω. Straightforward computations allow to prove that each diagonal element of T^*T is between 1 and $1+r_2^2$ and the sum of the absolute value of the off-diagonal elements of T^*T on each line is smaller than $2r_2$. Hence, T^*T is diagonally dominant and $\varphi_{\max}(T^*T) \leq (1+r_2)^2$ and $\varphi_{\min}(T^*T) \geq 1-2r_2$. Since r_2 is constrained to be smaller than $1/8$ than the eigenvalues of T^*T are between $3/4$ and $3/2$. The eigenvalues of S are between 1 and $5/4$, because r_1 is constrained to be smaller than $1/4$. The eigenvalues of Ω are bounded using the eigenvalues of T and S: $\varphi_{\max}(\Omega) \leq \varphi_{\max}(T^*T)\varphi_{\max}(S^{-1})$ and $\varphi_{\min}(\Omega) \leq \varphi_{\min}(T^*T)\varphi_{\min}(S^{-1})$. Hence, we conclude that the eigenvalues of Ω are between $1/2$ and 2.

2. Let us now lower bound $d(\Omega,\Omega')$ if $\Omega \neq \Omega'$. The quantity $d_H(t_i,t_i')r_2^2/4$ is smaller than 2. Hence, by the mean value theorem $\log(1+d_H(t_i,t_i')r_2^2/4)$ is larger than $d_H(t_i,t_i')r_2^2/4$. By definition of the sets $\mathcal{T}_2^{(2)}[k,p,r_2]$ and $\mathcal{S}'[A^c,p,r_1]$, we get

$$\begin{aligned} d(\Omega,\Omega') &= \sum_{i=1}^{p-k} f\left(\frac{s_i}{s_i'}\right) + \sum_{i=p-k+1}^{k} \log\left(1+\frac{d_H(t_i,t_i')r_2^2}{4}\right) \\ &\geq \frac{p-k}{4}\left[f(1+r_1) \wedge f(1/(1+r_1))\right] + \frac{d_H(T,T')r_2^2}{8} \geq L\left[(p-k)r_1^2 + k\log(p)r_2^2\right] \\ &\geq L\left[pr_1^2 + k\log(p)r_2^2\right] \ , \end{aligned}$$

   since k is assumed to be smaller than $p/2$.

3. Let us upper bound the Kullback divergence between two element Ω and Ω' in $\mathcal{U}_2'[k,p,r_1,r_2]$

$$\begin{aligned} 2\mathcal{K}\left(\Omega;\Omega'\right) &= \sum_{i=1}^{p} \frac{s_i}{s_i'} - \log\left(\frac{s_i}{s_i'}\right) - 1 + \frac{l_i(t_i,t_i')}{s_i'} \\ &= \sum_{i=1}^{p-k} \frac{s_i}{s_i'} - \log\left(\frac{s_i}{s_i'}\right) - 1 + \sum_{i=p-k+1}^{k} l_i(t_i,t_i') \ . \end{aligned}$$

   Since the smallest eigenvalue of Ω is smaller than $1/2$, it follows that $l_i(t_i.t_i')$ is smaller than $2\|t_i - t_i'\|_{l_2}^2$ which is smaller than $4r_2^2$ by definition of $\mathcal{T}_2^{(2)}[k,p,r_2]$. Let us recall that the function f defined as $f(x) = x - 1 - \log(x)$ is positive and equivalent to $(x-1)^2$ when x is close to one.

Since r_1 is smaller than $1/4$, there exists some numerical constant L such that $f(\frac{s_i}{s_i'}) \leq L r_1^2$. All in all, we obtained the upper bound

$$\mathcal{K}(\Omega;\Omega') \leq L\left[(p-k)r_1^2 + kr_2^2\right] \leq L\left[pr_1^2 + kr_2^2\right].$$

4. Finally, we lower bound the cardinality of $\mathcal{U}_2'[k,p,R_1,R_2]$.

$$\log\left[\mathrm{Card}\left(\mathcal{U}_2'[k,p,r_1,r_2]\right)\right] \geq \frac{p-k}{8} + \frac{k\log p}{8} \geq L\left[p + k\log(p)\right] .$$

By Lemma 5.27,

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{U}'[k,p,r_1,r_2]} \mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] \geq L\left[pr_1^2 + k\log(p)r_2^2\right] ,$$

provided that $r_1 \leq 1/4$, $r_2 \leq 1/8$, and $n\left[pr_1^2 + kr_2^2\right] \leq L_1\left[p + k\log(p)\right]$. Choosing $r_1^2 = \frac{1}{16} \wedge \frac{(L_1\wedge 1)}{n}$ and $r_2^2 = \frac{(L_1\wedge 1)\log(p)}{n} \wedge \frac{1}{64}$ since n is assumed to be larger than $\log(p)$.

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{U}_2'[k,p,r_1,r_2]} \mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] \geq L\frac{p+k\log p}{n} .$$

Let us now prove that the set $\mathcal{U}_2'[k,p,r_1,r_2]$ is included in $\mathcal{B}_{\mathcal{K}}(1)$.

$$\mathcal{K}(\Omega;I_p) \quad\leq\quad \left[\varphi_{\min}(\Omega)\right]^{-1}\frac{kr_2^2}{2} + \frac{p-k}{2}f(1+r_1) \leq \frac{k\log p}{n} + p - k \leq p ,$$

since $f(5/4) \leq 1$ and $n \geq \log(p)$. Hence, the following minimax lower bound also holds

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{U}_2[k,p]\cap\mathcal{B}_{\mathcal{K}}(n^\beta)} \mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] \geq L\frac{p+k\log p}{n} .$$

If $\log p$ is smaller than 21, we consider the set $\mathcal{U}_{\mathrm{ord}}[0,\ldots,0,1,\ldots,1,r]$ where there are $p-k$ times 0 and k times 1. It is included in $\mathcal{U}_2[k,p]$ and by Proposition 5.9, we conclude that

$$\inf_{\widehat{\Omega}} \sup_{\Omega\in\mathcal{U}_2[k,p]} \mathbb{E}\left[\mathcal{K}\left(\Omega;\widehat{\Omega}\right)\right] \geq L\frac{p+k}{n} \geq L\frac{p+k\log p}{n} ,$$

since $\log p$ is smaller than 21. Besides, one can prove that the set $\mathcal{U}_{\mathrm{ord}}[0,\ldots,0,1,\ldots,1,r]$ is included in $\mathcal{B}_{\mathcal{K}}(1)$ if r is smaller than $\frac{1}{\sqrt{n}} \wedge \frac{1}{8}$. Hence, the same minimax lower bound holds on $\mathcal{U}_2[k,p]\cap\mathcal{B}_{\mathcal{K}}(n^\beta)$ □

Proof of Lemma 5.30. Let $\Theta'[k,p/2]$ be a maximal subset of $\Theta[k,p/2]$ which is $pk/4$-separated with respect to the Hamming distance. Then, the closed Hamming balls $\mathcal{B}_H(x,pk/4)$ centered at the elements of $\Theta'[k,p/2]$ and with radius $0.5k\frac{p}{2}$ are covering $\Theta[k,p/2]$. Hence,

$$\mathrm{Card}\left(\Theta[k,p/2]\right) \leq \sum_{x\in\Theta'[k,p/2]} \mathrm{Card}\left(\mathcal{B}_H(x,0.5k\frac{p}{2})\right) .$$

The balls $\mathcal{B}_H(x,k\frac{p}{4})$ can also be considered as subsets of the set $\{0,1\}_{kp/2}^{(p/2)^2}$ of binary sequences of size $(p/2)^2$ with exactly $kp/2$ non-zero coefficients. In the proof of Lemma 4.10 in [Mas07], Massart shows that if $p \geq 8k$

$$\mathrm{Card}\left(\mathcal{B}_H(x,0.5k\frac{p}{2})\right) \leq \binom{(p/2)^2}{kp/2}^{-1}\left(\frac{p}{2k}\right)^{\rho k(p/2)} ,$$

where $\rho \geq 0.23$. Since $\log(p/k)$ is assumed to be larger than 19, we can apply this result. If follows that

$$\log\left[\mathrm{Card}(\Theta'[k,p/2])\right] \geq \frac{\rho}{2}kp\log\left(\frac{p}{2k}\right) + \log\left[\mathrm{Card}\left(\Theta[k,p/2]\right)\binom{(p/2)^2}{kp/2}^{-1}\right] . \tag{5.50}$$

Let us now lower bound the cardinality of $\Theta[k,p/2]$. Observe that $\mathrm{Card}\left(\Theta[2k,2p]\right) \geq \left(\Theta[k,p]\right)^4$. Let us indeed cut the square matrix of size $2p$ into four square matrices of size p. Then, any combination of any

four elements of $\Theta[k,p]$ yields an unique element of $\Theta[2k,2p]$. Since $k=2^s$ for some integer $s>0$ and since $2k$ divides p, one concludes by straightforward induction that

$$\log\left(\mathrm{Card}(\Theta[k,p/2])\right) \geq k^2 \log\left(\mathrm{Card}(\Theta[1,p/(2k)])\right) \ .$$

Moreover, $\Theta[1,p/2k]$ is in correspondence with the set of permutations of $p/2k$ elements. Thus,

$$\log\left(\mathrm{Card}(\Theta[1,p/(2k)])\right) \geq \left(\frac{p}{2k}\right)! \geq \frac{p}{2k}\log\left(\frac{p}{2ek}\right) \ ,$$

since $a! \geq (a/e)^a$ for any positive integer a. If follows that $\log\left(\mathrm{Card}(\Theta[k,p/2])\right) \geq \frac{pk}{2}\log\left(\frac{p}{2ek}\right)$. In contrast, $\log\binom{(p/2)^2}{kp/2}$ is upper bounded by $\frac{kp}{2}\log[pe/(2k)]$ since $\binom{a}{b} \leq (ae/b)^b$ for any positive integers a and b. Gathering these bounds with (5.50) yields

$$\log\left[\mathrm{Card}(\Theta'[k,p/2])\right] \geq \frac{pk}{2}\left[\rho\log\left(\frac{p}{k}\right) - \rho\log 2 - 2\right] \geq \frac{\rho}{4}kp\log\left(\frac{p}{k}\right) \ ,$$

since $\log(p/k)$ is assumed to be larger than 19 which is larger than $2\log 2 + 4/\rho$.

□

Proof of Lemma 5.31. The set $\mathcal{T}_2^{(1)}[k,p,r_2]$ is in one to one correspondence with the set $\Theta_2[k,\lfloor p/2\rfloor]$ of binary matrices of size $k\times\lfloor p/2\rfloor$ with exactly one non-zero entry on each line and at most one on each column. The proof is then quite similar to the proof of Lemma 5.30. Let $\Theta_2'[k,\lfloor p/2\rfloor]$ be a maximal subset of $\Theta_2[k,\lfloor p/2\rfloor]$ such that the Hamming distance between any two different elements of $\Theta_2'[k,\lfloor p/2\rfloor]$ is larger than $\frac{k}{2}$. Then, the set $\mathcal{T}_2^{(2)}[k,p,r_2]$ is the subset $\mathcal{T}_2^{(1)}[k,p,r_2]$ which naturally in correspondence with $\Theta_2'[k,\lfloor p/2\rfloor]$.

Let us lower bound the cardinality of $\Theta_2'[k,\lfloor p/2\rfloor]$. Since the closed Hamming balls with radius $r/2$ and centered at the elements of $\Theta_2'[k,\lfloor p/2\rfloor]$ cover $\Theta_2[k,\lfloor p/2\rfloor]$, we get

$$\mathrm{Card}\left(\Theta_2[k,\lfloor p/2\rfloor]\right) \geq \sum_{x\in\Theta_2'[k,\lfloor p/2\rfloor]} \mathrm{Card}\left(\mathcal{B}_H(x,k/2)\right) \ .$$

One can consider these balls as subsets of the set $\{0,1\}_k^{k\lfloor p/2\rfloor}$of binary sequence of size $k\lfloor p/2\rfloor$ with exactly k non-zero coefficients. We use the same lower bound for the Hamming balls as in the previous proof:

$$\mathrm{Card}\left(\mathcal{B}_H(x,k/2)\right) \leq \binom{k\lfloor p/2\rfloor}{k}^{-1} (\lfloor p/2\rfloor)^{\rho k} \ ,$$

if $p\geq 8$. We recall that $\rho\geq 0.23$. We can apply this result since $\log(p)\geq 21$. It follows that

$$\log\left[\mathrm{Card}\left(\Theta_2'[k,\lfloor p/2\rfloor]\right)\right] \geq \rho k\log\left(\lfloor p/2\rfloor\right) + \log\left[\mathrm{Card}\left(\Theta_2[k,\lfloor p/2\rfloor]\right)\right] - \log\binom{k\lfloor p/2\rfloor}{k} \ .$$

Observe that the cardinality of $\Theta_2[k,\lfloor p/2\rfloor]$ is $\frac{\lfloor p/2\rfloor!}{(\lfloor p/2\rfloor-k)!}$. For any positive integers a and b, it holds that $\binom{a}{b} \geq (a/b)^b$ and $a! \geq (a/e)^a$. Hence, we obtain

$$\log\left[\mathrm{Card}\left(\Theta_2[k,\lfloor p/2\rfloor]\right)\right] \geq \log\binom{\lfloor p/2\rfloor}{k}\log(k!) \geq k\log\left(\frac{\lfloor p/2\rfloor}{e}\right) \ .$$

Let us combine the previous bounds and and let us apply the inequality $\binom{a}{b} \leq (ae/b)^b$ which holds for any positive integer a and b. Hence, we get

$$\log\left[\mathrm{Card}\left(\Theta_2'[k,\lfloor p/2\rfloor]\right)\right] \geq \rho k\log\left(\lfloor p/2\rfloor\right) - 2k \geq k\left[\rho\log(p) - \rho\log(4) - 2\right] \geq \frac{\rho}{2}\log(p) \ ,$$

Since $\log(p)$ is larger than 21.

□

5.7.5 Proof of the corollaries

Proof of Corollary 5.6. The functions $H_i(.)$ equal 0 for all the collections $\mathcal{M}^d_{i,\mathrm{ord}}$. Hence, the collections $\mathcal{M}^d_{\mathrm{ord}}$ satisfies $\mathbb{H}_{K,\eta}$. We conclude by gathering Proposition 5.2 and Theorem 5.5. □

Proof of Corollary 5.12. For any $2 \leq i \leq p$ and any $1 \leq k \leq (i-1) \wedge d$, $H_i(k)$ is smaller than $1 + \log((i-1)/k)$. Hence, the collection $\mathcal{M}^d_{\mathrm{co}}$ satisfies Assumption $(\mathbb{H}_{K,\eta})$. The result then derives from Theorem 5.5. □

Proof of Corollary 5.11. For any symmetric matrix A, we denote $\{\varphi_i(A)\}_{1 \leq i \leq p_n}$ the set of its eigenvalues. Since $x - \log x - 1$ is equivalent to $(x-1)^2$ when x goes to one, the Kullback-Leibler divergence $\mathcal{K}(\Omega; \Omega')$ decomposes as

$$\begin{aligned}
\mathcal{K}(\Omega; \Omega') &= \frac{1}{2}\left[tr\left(\Omega'\Sigma\right) - \log\left(|\Omega'\Sigma|\right) - p_n\right] \\
&= \frac{1}{2}\sum_{i=1}^{p_n}\left\{\varphi_i\left(\sqrt{\Sigma}\Omega'\sqrt{\Sigma}\right) - \log\left[\varphi_i\left(\sqrt{\Sigma}\Omega'\sqrt{\Sigma}\right)\right] - 1\right\} \\
&= \frac{1}{4}\sum_{i=1}^{p_n}\left[\varphi_i\left(\sqrt{\Sigma}\Omega'\sqrt{\Sigma}\right) - 1\right]^2 + o\left[\mathcal{K}(\Omega; \Omega')\right] \\
&= \frac{1}{4}\sum_{i=1}^{p_n}\varphi_i^2\left(\sqrt{\Sigma}\Omega'\sqrt{\Sigma} - I_{p_n}\right) + o\left[\mathcal{K}(\Omega; \Omega')\right] \ ,
\end{aligned}$$

when $\mathcal{K}(\Omega; \Omega')$ is close to 0. This last sum corresponds to the Frobenius norm of $\sqrt{\Sigma}\Omega'\sqrt{\Sigma} - I_{p_n}$. Hence, we get

$$\|\sqrt{\Sigma}\Omega'\sqrt{\Sigma} - I_{p_n}\|_F^2 = 4\left[\mathcal{K}(\Omega; \Omega')\right] + o\left[\mathcal{K}(\Omega; \Omega')\right] \ , \tag{5.51}$$

when $\mathcal{K}(\Omega; \Omega')$ is close to 0. Let us come back to the Frobenius distance between Ω' and Ω,

$$\begin{aligned}
\|\Omega' - \Omega\|_F^2 &= tr\left[\sqrt{\Omega}\left(\sqrt{\Sigma}\Omega'\sqrt{\Sigma} - I_{p_n}\right)\Omega\left(\sqrt{\Sigma}\Omega'\sqrt{\Sigma} - I_{p_n}\right)\sqrt{\Omega}\right] \\
&\leq \varphi^2_{\max}(\Omega)\,\|\sqrt{\Sigma}\Omega'\sqrt{\Sigma} - I_{p_n}\|_F^2 \ .
\end{aligned}$$

Gathering this upper bound with the preceding result yields

$$\|\Omega' - \Omega\|_F^2 \leq 4\varphi^2_{\max}(\Omega)\left[\mathcal{K}(\Omega; \Omega') + o\left(\mathcal{K}(\Omega; \Omega')\right)\right] \ , \tag{5.52}$$

when $\mathcal{K}(\Omega; \Omega')$ is close to 0. By Corollary 5.6, the risk of $\widetilde{\Omega}^d_{\mathrm{ord}}$ on $\mathcal{U}_{\mathrm{ord}}[k_1, , \ldots, k_p, +\infty] \cap \mathcal{B}_{\mathrm{op}}(\gamma)$ is upper bounded

$$\mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{\mathrm{ord}}\right)\right] \leq L(K, \eta)\frac{p_n + \sum_{i=1}^{p_n} k_i}{n} + \tau_n(K, \eta, \Omega) \ .$$

The Kullback divergence $\mathcal{K}(\Omega; I_{p_n})/p_n$ is upper bounded by $[\varphi_{\max}(\Omega) \vee (\log[1/\varphi_{\min}(\Omega)] - 1)] \leq L(\gamma)$. Hence, the term $\tau_n(K, \eta, \Omega)$ is upper bounded by $L(K, \eta, \gamma)\frac{p_n}{n}$. We conclude that

$$\mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{\mathrm{ord}}\right)\right] \leq L(K, \eta, \gamma)\frac{p_n + \sum_{i=1}^{p_n} k_i}{n} \ .$$

Gathering this upper bound with (5.52) yields the first result. By Proposition 5.10, we know that

$$\mathbb{E}\left[\mathcal{K}\left(\Omega; \widetilde{\Omega}^d_{\mathrm{ord}}\right)\right] \leq L(K, \eta, \gamma)p_n\left(R^{\frac{2}{2s+1}}n^{-\frac{2s}{2s+1}} \wedge \frac{p_n}{n}\right) \ .$$

We prove the second result using this last bound and (5.52).

The corresponding minimax lower bounds are proved as Propositions 5.9 and 5.10. Indeed, we consider again the set $\mathcal{U}'_{\mathrm{ord}}[k_1, \ldots, k_{p_n}, r]$ defined in the proof of proposition 5.9 with $r \leq \frac{1}{8k} \wedge \frac{1}{n}$. We recall that it belongs to $\mathcal{B}_{\mathrm{op}}(2)$. For any two matrices $\Omega_1 \neq \Omega_2$ in this set,

$$\mathcal{K}(\Omega_1; \Omega_2) \geq 2d(\Omega_1, \Omega_2) \geq L\left[\sum_{i=1}^{p_n} k_i + p_n\right] r^2 \ ,$$

where $d(.,.)$ is introduced in Lemma 5.27. The second lower bound is given at the end of the proof of Proposition 5.9. We also have stated the converse upper bound

$$\mathcal{K}(\Omega_1;\Omega_2) \leq L\left[\sum_{i=1}^{p_n} k_i + p_n\right] r^2 \ .$$

Arguing as previously, we connect the Frobenius distance between Ω_1 and Ω_2 with the Kullback entropy.

$$\begin{aligned}\|\Omega_1-\Omega_2\|_F^2 &\geq \varphi_{\min}^2(\Omega_1)\,\|\sqrt{\Omega_1}^{-1}\Omega_2\sqrt{\Omega_1}^{-1}-I_{p_n}\|_F^2\\ &\geq 4\varphi_{\min}^2(\Omega_1)\,\mathcal{K}(\Omega_1;\Omega_2)+o\left[\mathcal{K}(\Omega_1;\Omega_2)\right]\\ &\geq \mathcal{K}(\Omega_1;\Omega_2)+o\left[\mathcal{K}(\Omega_1;\Omega_2)\right]\ ,\end{aligned}$$

because $\varphi_{\min}(\Omega_1)$ is larger than $1/2$. Since r^2 is assumed to be smaller than $1/n$ and since $\sum_{i=1}^{p_n} k_i + p_n = o(n)$, $\mathcal{K}(\Omega_1;\Omega_2)$ goes to 0 when n goes to infinity. Hence, for n sufficiently large,

$$\|\Omega_1-\Omega_2\|_F^2 \geq \frac{1}{2}\mathcal{K}(\Omega_1;\Omega_2) \geq L\left[\sum_{i=1}^{p_n} k_i + p_n\right] r^2 \ .$$

Applying suitably Lemma 5.28 yields

$$\inf_{\widehat{\theta}}\sup_{\Omega\in\mathcal{U}_{\mathrm{ord}}[k_1,\ldots,k_{p_n},r]\cap\mathcal{B}_{\mathrm{op}}(\gamma)} \mathbb{E}\left[\|\Omega-\widehat{\Omega}\|_F^2\right] \geq L\left[\sum_{i=1}^{p_n} k_i + p_n\right]\left(r^2\wedge\frac{1}{n}\right) \ ,$$

as long as n is large enough. This proves the first minimax lower bound.

Let us define $k_n := (R^2 n)^{1/(2s+1)} \wedge (p-1)$ and $r_n = 1/(8\sqrt{n})$. Since $s > 1/4$, k_n is smaller than $\lfloor\sqrt{n}\rfloor$ for n large enough. We straightforwardly check as in the proof of Proposition 5.10 that $\mathcal{U}_{\mathrm{ord}}[0,1,\ldots,k_n,\ldots k_n,r_n]$ is included in $\mathcal{E}'[s,p_n,R]\cap\mathcal{B}_{\mathrm{op}}(2)$. Thanks to the last minimax lower bound, we then conclude that

$$\inf_{\widehat{\theta}}\sup_{\Omega\in\mathcal{E}'[s,p_n,R]\cap\mathcal{B}_{\mathrm{op}}(2)} \mathbb{E}\left[\|\Omega-\widehat{\Omega}\|_F^2\right] \geq L\frac{p_n k_n}{n} \geq Lp_n\left(\left(\frac{R}{n^s}\right)^{\frac{2}{2s+1}}\wedge\frac{p_n-1}{n}\right),$$

for n large enough.

□

Proof of Corollary 5.15. From the previous proof, we know that for any estimator $\widetilde{\Omega}$ such that $\mathcal{K}\left(\Omega;\widetilde{\Omega}\right) = o_P(1)$, $\|\widetilde{\Omega}-\Omega\|_F^2 = \mathcal{O}_P\left[\mathcal{K}\left(\Omega;\widetilde{\Omega}\right)\right]$. Let us apply Corollary 5.12

$$\mathbb{E}\left[\mathcal{K}\left(\Omega;\widetilde{\Omega}\right)\right] \leq L(K,\eta)\frac{(k_n+1)\log p_n}{n} + L(K,\eta)n^{5/2}\left[p+\mathcal{K}(\Omega;I_{p_n})\right]\exp\left[-nL(K,\eta)\right] \ .$$

The Kullback divergence $\mathcal{K}(\Omega;I_{p_n})$ is upper bounded by $p_n\left[\varphi_{\max}(\Omega)\vee(\log[1/\varphi_{\min}(\Omega)]-1)\right]$. Hence, we get

$$\mathbb{E}\left[\mathcal{K}\left(\Omega;\widetilde{\Omega}\right)\right] \leq L(K,\eta,\gamma)\frac{p_n+k_n\log p_n}{n}[1+o(1)] \ .$$

Gathering this last upper bound with (5.51) yields the first result. Since the Frobenius norm dominates the operator norm, the second result follows.

The corresponding asymptotic minimax lower bound is proved as in Proposition 5.13 using again the fact the Frobenius distance $\|\Omega-\widehat{\Omega}\|_F^2$ is lower bounded by the Kullback divergence $\mathcal{K}(\Omega;\widehat{\Omega})$ times a constant if $\mathcal{K}(\Omega;\widehat{\Omega})$ is small enough..

□

Proof of Corollary 5.16. The proof is analogous to the previous one. The idea is to combine the results obtained in terms of Kullback divergence with the inequalities that compare Kullback divergence and Frobenius norm (proof of Corollary 5.11). □

5.8 Appendix

Proof of Lemma 5.19. Let d be a positive integer larger than one. By Jensen's inequality, we first notice that $\Psi(d)$ is non-positive. Using the density of a $\chi^2(d)$ distribution, we obtain

$$\Psi(d) = \int_0^{+\infty} \frac{\log(t) e^{-t} t^{d/2-1}}{2^{d/2}\Gamma(n/2)} dt - \log(d) := I_d - \log(d) \ ,$$

where $\Gamma(.)$ stands for the Gamma function. Let us exhibit a recurrence relation for I_d applying integration by parts:

$$\begin{aligned} I_d &= \int_0^{+\infty} \frac{\log(2t) e^{-t} t^{d/2-1}}{\Gamma(d/2)} dt = 0 + \int_0^{+\infty} e^{-t} t^{d/2-2} \frac{1 + \log(2t)(d/2-1)}{\Gamma(d/2)} \\ &= \frac{1}{d/2-1} + I_{d-2} \ . \end{aligned}$$

Hence, we only have to work out I_1 and I_2 in order to compute I_d.

$$\begin{aligned} I_2 &= \log(2) + \Gamma'(1)/\Gamma(1) = \log(2) - \gamma \ , \\ I_1 &= \log(2) + \Gamma'(1/2)/\Gamma(1/2) = -\log(2) - \gamma \ , \end{aligned}$$

where γ is the Euler constant. For any positive integer d, we therefore derive that

$$\begin{aligned} \Psi(2d) &= \sum_{i=1}^{d-1} \frac{1}{i} - \gamma - \log(d) \ , \\ \Psi(2d+1) &= \sum_{i=1}^{d} \frac{2}{2i-1} - \gamma - 2\log(2) - \log(d+1/2) \ . \end{aligned}$$

Using the asymptotic expansion of the harmonic series yields $\Psi(2d) = \frac{-1}{2d} + \mathcal{O}\left(\frac{1}{(2d)^2}\right)$. Let us note $H(d)$ the d-th partial sum of harmonic series. Straightforwards computations lead to

$$\begin{aligned} \Psi(2d+1) &= 2H(2d) - H(d) - \gamma - 2\log(2) - \log(d+1/2) \\ &= \mathcal{O}\left(\frac{1}{d^2}\right) + \log\left(\frac{d}{d+1/2}\right) = \frac{-1}{2d} + \mathcal{O}\left(\frac{1}{(2d)^2}\right) \ . \end{aligned}$$

Thus, we obtain the asymptotic expansion $\Psi(d) = -\frac{1}{d} + \mathcal{O}\left(\frac{1}{d^2}\right)$. Let us turn to the lower bound. From now on, we assume that $d \geq 3$. We define the sequence v_d as $v_d = \Psi(d) + 1/(d-2)$. We know that v_d converges to 0 when d goes to infinity. Let us prove that the subsequences $(v_{2d})_{d>1}$ and $(v_{2d+1})_{d\geq 1}$ are decreasing. Since $\log(1-x) \leq -x - \frac{x^2}{2}$ for any $0 \leq x < 1$,

$$\begin{aligned} v_{2d+2} - v_{2d} &= \frac{3}{2d} - \frac{1}{2d-2} + \log\left(1 - \frac{1}{d+1}\right) \\ &\leq \frac{1}{d} - \frac{1}{2d(d-1)} - \frac{1}{d+1} - \frac{1}{2(d+1)^2} \\ &\leq \frac{1}{2d(d+1)^2} - \frac{1}{d(d+1)(d-1)} < 0 \ . \end{aligned}$$

Analogously, we compute

$$\begin{aligned} v_{2d+1} - v_{2d-1} &= \frac{3}{3d-1} - \frac{1}{2d-3} + \log\left(1 - \frac{2}{2d+1}\right) \\ &\leq \frac{4}{(2d-1)(2d+1)^2} - \frac{8}{(2d-3)(2d-1)(2d+1)} < 0 \ . \end{aligned}$$

We conclude that v_d is non-negative for any $d \geq 3$. It follows that $\Psi(d) \geq -\frac{1}{d-2}$.

□

Chapter 6

Adaptive estimation of stationary Gaussian fields

Abstract. We study the non-parametric covariance estimation of a stationary Gaussian field X. In the time series setting, some procedures like AIC are proved to achieve optimal model selection among autoregressive models. However, there exists no such equivalent results of adaptivity in a spatial setting. By considering collections of Gaussian Markov random fields (GMRF) as approximation sets for the distribution of X, we introduce a novel model selection procedure for spatial fields. Given a neighborhood m, this procedure first amounts to computing a covariance estimator of X within the GMRFs of neighborhood m. We then select neighborhood a $\widehat{m}$ among a given collection of neighborhoods by applying a penalization strategy. The so-defined method satisfies a nonasymptotic oracle type inequality. If X is a GMRF, the procedure is also minimax adaptive to the sparsity of its neighborhood. More generally, the procedure is adaptive to the rate of approximation of the true distribution by GMRFs with growing neighborhoods.

6.1 Introduction

In this chapter, we study the estimation of the distribution of a stationary Gaussian field $X = (X_{[i,j]})_{(i,j)\in\Lambda}$ indexed by the nodes of a square lattice Λ of size $p \times p$. This problem is often encountered in spatial statistics or in image analysis.

Various estimation methods have been proposed to handle this question. Most of them fall into two categories. On the one hand, one may consider direct covariance estimation. A popular approach amounts to first computing an empirical variogram and then fitting a suitable parametric variogram model such as the exponential or Matérn model. It is beyond the scope of this chapter to do an exhaustive review of these methods and we refer to [Cre93] Ch.2 for more details. Some procedures also apply to non-regular lattices. However, a bad choice of the variogram model may lead to poor results. The issue of variogram model selection has not been completely solved yet, although some procedures based on cross-validation have been proposed. See [Cre93] Sect.2.6.4 for a discussion. Alternatively, Rosenblatt [Ros85] Ch.5 has developed a non-parametric estimator of the spectral density of the field X. This procedure is shown to be universally consistent, but it fails to achieve the optimal rate of convergence when the true distribution belongs to one parametric model.

On the other hand, a second approach to the problem amounts to considering the conditional distribution at one node given the remaining nodes. This point of view is closely connected to the notion of *Gaussian Markov Random field* (GMRF). Let $\mathcal{G}$ be a graph whose vertex set is Λ. The field X satisfies the local Markov property with respect to $\mathcal{G}$ if it satisfies the following property: for any node $(i,j) \in \Lambda$, conditionally to the set of variables $X_{[k,l]}$ such that (k,l) is a neighbor of (i,j) in $\mathcal{G}$, $X_{[i,j]}$ is independent

from all the remaining variables. The field X is said to be a GMRF with respect to the graph $\mathcal{G}$ if it fulfills the local Markov property with respect to $\mathcal{G}$. GMRFs are also sometimes called Gaussian graphical models. A huge literature develops around this subject since Gaussian graphical models are promising tools to analyze complex high-dimensional systems involved for instance in postgenomic data. See [Lau96] and [Edw00] for introductions to Gaussian graphical models and Markov properties. In the sequel, we assume that the node $(0,0)$ belongs to Λ. Since we assume here that the field X is stationary, defining a graph $\mathcal{G}$ is equivalent to defining the neighborhood m of the node $(0,0)$. Indeed, the neighborhood of any node $(i,j) \in \Lambda$ is the transposition of m by (i,j). In the sequel, we call m *the neighborhood* of a GMRF. If the neighborhood is empty, then the Markov property states that the components of X are all independent. Alternatively, any zero-mean Gaussian stationary field is a GMRF with respect to the complete neighborhood (i.e. containing all the nodes except $(0,0)$). Let us mention the idea underlying our approach: using the same data, we select a *suitable* neighborhood and estimate the distribution of X in the space of stationary GMRFs with respect to this neighborhood.

Numerous papers have been devoted to parametric estimation for stationary GMRFs with a known neighborhood. The authors have derived their asymptotic properties of such estimators (see [BM75, Bes77, Guy95]). If the field X is assumed to be a GMRF with respect to a *known* neighborhood in all these works, the issue of neighborhood selection has been less studied. Besag and Kooperberg [BK95], Rue and Tjelmeland [RT02], and Cressie and Verzelen [CV08] have tackled the problem of *approximating* the distribution of a Gaussian field by a GMRF, but this requires the knowledge of the true distribution. Guyon and Yao have stated in [GY99] necessary conditions and sufficient conditions for a model selection procedure to choose asymptotically the true neighborhood of a GMRF with probability one. Our point of view is slightly different: we do not assume that the field X is a GMRF with respect to a sparse neighborhood and do not aim at estimating the true neighborhood, we rather want to select a neighborhood that allows to estimate *well* the distribution of X.

Our problem on a two-dimensional field has a natural one-dimensional counterpart in time series analysis. It is indeed known that an auto-regressive process (AR) of order p is also a GMRF with $2p$ nearest neighbors and reciprocally (see [Guy95] Sect. 1.3). In this one-dimensional setting, our issue reformulates as follows: how can we select the order of an AR to estimate well the distribution of a time series? It is known that order selection by minimization of criteria like AICC, AIC or FPE satisfy asymptotically oracle inequalities (Shibata [Shi80] and Hurvich and Tsai [HT89]). We refer to Brockwell and Davis [BD91] and McQuarrie and Tsai [MT98] for detailed discussions. However, one cannot readily extend these results to a spatial setting because of computational and theoretical difficulties.

6.1.1 Conditional regression

Let us now precise the notations and present the ideas underlying our approach. In the sequel, Λ stands for the toroidal lattice of size $p \times p$. We consider the random field $X = (X_{[i,j]})_{1 \le i,j \le p}$ indexed by the nodes of Λ. Besides, X^v refers to the vectorialized version of X with the convention $X_{[i,j]} = X^v{}_{[(i-1)\times p + j]}$ for any $1 \le i,j \le p$. Using this new notation amounts to "forgetting" the spatial structure of X and allows to get into a more classical statistical framework. For the sake of simplicity, the components of X are defined modulo p in the remainder of the chapter.

Throughout this chapter, we assume the field X is centered. In practice, the statistician has to first subtract some parametric form of the mean value. Hence, the vector X^v follows a zero-mean Gaussian distribution $\mathcal{N}(0,\Sigma)$, where the $p^2 \times p^2$ matrix Σ is non singular but unknown. Besides, we suppose that the field X is stationary on the torus Λ. More precisely, for any $r > 0$, any $(i,j) \in \{1,\dots,p\}^2$, and any $(k_1,l_1),\dots,(k_r,l_r) \in \{1,\dots,p\}^{2r}$, it holds that

$$(X_{[k_1,l_1]},\dots,X_{[k_r,l_r]}) \sim (X_{[k_1+i,l_1+j]},\dots,X_{[k_r+i,l_r+j]}) \ .$$

We observe $n \ge 1$ i.i.d. replications of the vector X^v. In the sequel, $\mathbf{X^v}$ denotes the $p^2 \times n$ matrix of the n observations of X^v. For any $1 \le i \le n$, the $p \times p$ matrix $\mathbf{X}_i$ stands for the i-th observation of the field X. All these notations are recalled in Table 7 in Section 6.1.4. In practice, the number of observations n often equals one. Our goal is to estimate the matrix Σ.

We sometimes assume that the field X is isotropic. Let G be the group of vector isometries of the unit square. For any node $(i,j) \in \Lambda$ and any isometry $g \in G$, $g.(i,j)$ stands for the image of (i,j) in Λ under the action of g. We say that X is isotropic on Λ if for any $r > 0$, $g \in G$, and $(k_1,l_1),\dots,(k_r,l_r) \in \{1,\dots,p\}^{2r}$,

$$(X_{[k_1,l_1]},\dots,X_{[k_r,l_r]}) \sim (X_{[g.(k_1,l_1)]},\dots,X_{[g.(k_r,l_r)]}) \ .$$

As mentioned earlier, we aim at estimating the distribution of the field X through a conditional distribution approach. By standard Gaussian derivations (see for instance [Lau96] App.C), there exists a unique $p \times p$ matrix θ such that $\theta[0,0] = 0$ and

$$X_{[0,0]} = \sum_{(i,j)\in\Lambda\setminus\{(0,0)\}} \theta_{[i,j]} X_{[i,j]} + \epsilon_{[0,0]} \ , \tag{6.1}$$

where the random variable $\epsilon_{[0,0]}$ follows a zero-mean normal distribution and is independent from the covariates $(X_{[i,j]})_{(i,j)\in\Lambda\setminus\{(0,0)\}}$. Equation (6.1) describes the conditional distribution of $X_{[0,0]}$ given the remaining variables. Since the field X is stationary, the matrix θ also satisfies $\theta_{[i,j]} = \theta_{[-i,-j]}$ for any $(i,j) \in \Lambda$. Let us note σ^2 the conditional variance of $X_{[0,0]}$ and I_{p^2} the identity matrix of size p^2. The matrix θ is closely related to the covariance matrix Σ of X^v through the following property:

$$\Sigma = \sigma^2 \left[I_{p^2} - C(\theta)\right]^{-1} \ , \tag{6.2}$$

where the $p^2 \times p^2$ matrix $C(\theta)$ is defined as $C(\theta)_{[i_1(p-1)+j_1,i_2(p-1)+j_2]} := \theta_{[i_2-i_1,j_2-j_1]}$ for any $1 \le i_1,i_2,j_1,j_2 \le p$. The matrix $(I_{p^2} - C(\theta))$ is called the partial correlation matrix of the field X. The so-defined matrix $C(\theta)$ is symmetric block circulant with $p \times p$ blocks as stated below. We refer to [RH05] Sect.2.6 or the book of Gray [Gra06] for definitions and main properties on circulant and block circulant matrices.

Lemma 6.1. *Let θ be a square matrix of size p such that*

$$\text{for any } 1 \le i,j \le p,\ \theta_{[i,j]} = \theta_{[-i,-j]}, \tag{6.3}$$

then the matrix $C(\theta)$ is symmetric block circulant with $p \times p$ blocks. Conversely, if B is a $p^2 \times p^2$ symmetric block circulant matrix with $p \times p$ blocks, then there exists a square matrix θ of size p satisfying (6.3) and such that $B = C(\theta)$.

A proof is given in the appendix. In conclusion, estimating the matrix Σ/σ^2 amounts to estimating the matrix $C(\theta)$, which is also equivalent to estimating the $p \times p$ matrix θ. This is why, we shall focus on the estimation of the matrix θ.

Let us precise the set of possible values for θ. In the sequel, Θ denote the vector space of the $p \times p$ matrices that satisfy $\theta_{[0,0]} = 0$ and $\theta_{[i,j]} = \theta_{[-i,-j]}$, for any $(i,j) \in \Lambda$. A matrix $\theta \in \Theta$ corresponds to the distribution of a stationary Gaussian field if and only if the $p^2 \times p^2$ matrix $(I_{p^2} - C(\theta))$ is positive definite. This is why we define the convex subset Θ^+ of Θ by

$$\Theta^+ := \left\{\theta \in \Theta \ \text{ s.t. } \left(I_{p^2} - C(\theta)\right) \text{ is positive definite}\right\} \ . \tag{6.4}$$

The set of covariance matrices of stationary Gaussian fields on Λ with unit conditional variance is therefore in one to one correspondence with the set Θ^+. Let us define the corresponding set Θ^{iso} and $\Theta^{+,\text{iso}}$ for isotropic Gaussian fields.

$$\Theta^{\text{iso}} := \{\theta \in \Theta \ , \ \ \theta_{[i,j]} = \theta_{[g.(i,j)]} \ , \forall (i,j) \in \Lambda, \ \forall g \in G\} \quad \text{and } \Theta^{+,\text{iso}} := \Theta^+ \cap \Theta^{\text{iso}} \ . \tag{6.5}$$

6.1.2 Model selection

The issue of covariance estimation may be reformulated as a problem of conditional regression defined in Equation (6.1). However, the set Θ^+ of admissible parameters for the estimation is huge. The dimension of Θ is indeed of the same order as p^2 whereas we only observe p^2 non-independent data if n equals one. In order to avoid the curse of dimensionality, it is natural to assume that the target θ is approximately *sparse*.

It is indeed likely that the coefficients $\theta_{[i,j]}$ are *close* to zero for the nodes (i,j) which are *far* from the origin $(0,0)$. By Equation (6.1), this means that $X_{[0,0]}$ is *well* predicted by the covariates $X_{[i,j]}$ whose corresponding nodes (i,j) are close to the origin. We do not want to perform a restrictive assumption on the true distribution. In general, we aim at adapting to the *sparsity* of the matrix θ.

In the sequel, m refers to a subset of $\Lambda \setminus \{0,0\}$. We call it a model. By Equation (6.1), the property "X is a GMRF with respect to the neighborhood m" is equivalent to "the support of θ is included in m". We are given a nested collection $\mathcal{M}$ of models. For any of these models $m \in \mathcal{M}$, we compute $\widehat{\theta}_{m,\rho_1}$ the Conditional least squares estimator (CLS) of θ for the model m by maximizing the pseudolikelihood over a subset of matrices θ whose support is included in m. These estimators as well as their dependency on the quantity ρ_1 are defined in Section 6.2.

The model m that minimizes the risk of $\widehat{\theta}_{m,\rho_1}$ over the collection $\mathcal{M}$ is called an oracle and is noted m^*. In practice, this model is unknown and we have to estimate it. The art of model selection is to pick a model $m \in \mathcal{M}$ that is large enough to enable a good approximation of θ but is small enough so that the variance of $\widehat{\theta}_{m,\rho_1}$ is small. Let us reformulate the approach in terms of GMRFs: given a collection $\mathcal{M}$ of neighborhoods, we compute an estimator of θ in the set of GMRFs with neighborhood m, for any $m \in \mathcal{M}$. Our purpose is to select a suitable neighborhood $\widehat{m}$ so that the estimator $\widehat{\theta}_{\widehat{m}}$ has a risk as small as possible.

A classical method to estimate a *good model* $\widehat{m}$ is achieved through *penalization* with respect to the size of the models. In the following expression, $\gamma_{n,p}(.)$ stands for the CLS empirical contrast that we shall define in Section 6.2. We recall that it is closely connected to the pseudolikelihood. We select a model $\widehat{m}$ by minimizing the criterion

$$\widehat{m} = \arg\min_{m \in \mathcal{M}} \left[\gamma_{n,p}(\widehat{\theta}_{m,\rho_1}) + \mathrm{pen}(m) \right] . \tag{6.6}$$

where pen(.) denotes a positive function defined on $\mathcal{M}$. In this chapter, we prove that under a suitable choice of the penalty function pen(.), the risk of the estimator $\widehat{\theta}_{\widehat{m}}$ is as small as possible.

6.1.3 Risk bounds and adaptation

We shall assess our procedure using two different loss functions. First, we introduce the loss function $l(.,.)$ that measures how well we estimate the conditional distribution (6.1) of the field. For any $\theta_1, \theta_2 \in \Theta$, the distance $l(\theta_1, \theta_2)$ is defined by

$$l(\theta_1, \theta_2) := \frac{1}{p^2} tr\left[(C(\theta_1) - C(\theta_2)) \Sigma (C(\theta_1) - C(\theta_2)) \right] . \tag{6.7}$$

Let us reformulate $l(\theta_1, \theta_2)$ in terms of conditional expectation

$$l(\theta_1, \theta_2) = \mathbb{E}_\theta \left\{ \left[\mathbb{E}_{\theta_1} \left(X_{[0,0]} | X_{\Lambda \setminus \{0,0\}} \right) - \mathbb{E}_{\theta_2} \left(X_{[0,0]} | X_{\Lambda \setminus \{0,0\}} \right) \right]^2 \right\} ,$$

where $\mathbb{E}_\theta(.)$ stands for the expectation with respect to the distribution of X^v, $\mathcal{N}(0, \sigma^2 (I_{p^2} - C(\theta))^{-1})$. Hence, $l(\widehat{\theta}, \theta)$ corresponds the mean squared prediction loss which is often used in the random design regression framework or in time series analysis [HT89]. Moreover, the loss function $l(\widehat{\theta}, \theta)$ is also connected to the notion of kriging error. The kriging predictor of $X_{[0,0]}$ is defined as the best linear combination of the covariates $(X_{[k,l]})_{(k,l) \in \Lambda \setminus \{(0,0\}}$ for predicting the value $X_{[0,0]}$. By Equation (6.1), this predictor is exactly $\sum_{(k,l) \in \Lambda \setminus \{(0,0\}} \theta_{[k,l]} X_{[k,l]}$ and the mean squared prediction error is σ^2. If we do not know θ but we are given an estimator $\widehat{\theta}$, then the corresponding kriging predictor $\sum_{(k,l) \in \Lambda \setminus \{(0,0\}} \widehat{\theta}_{[k,l]} X_{[k,l]}$ has a mean squared prediction error equal to $\sigma^2 + l(\widehat{\theta}, \theta)$. Kriging is a key concept in spatial statistics and it is therefore interesting to consider a loss function that measures the kriging performances when one estimates θ. We refer to Cressie [Cre93] Ch.2 for a comprehensive introduction on kriging and related notions.

We shall also assess our results thanks to the Frobenius distance noted $\|.\|_F$ and defined by $\|A\|_F^2 := \sum_{1 \le i,j \le p} A_{[i,j]}{}^2$. Observe that the Frobenius distance $\|\theta_1 - \theta_2\|_F^2$ also equals the Frobenius distance

between the partial correlation matrices $(I_{p^2} - C(\theta_1))$ and $(I_{p^2} - C(\theta_2))$ (up to a factor p^2)

$$\|\theta_1 - \theta_2\|_F^2 = \frac{1}{p^2}\| \left(I_{p^2} - C(\theta_1)\right) - \left(I_{p^2} - C(\theta_2)\right) \|_F^2 \ , \tag{6.8}$$

Our aim is then to define a suitable penalty function pen(.) in (6.6) so that the estimator $\widehat{\theta}_{\widehat{m},\rho_1}$ performs almost as well as the oracle estimator $\widehat{\theta}_{m^*,\rho_1}$. For any model $m \in \mathcal{M}$, we define θ_{m,ρ_1} as the matrix which minimizes the loss $l(\theta', \theta)$ over the sets of matrices θ' corresponding to model m. The loss $l(\theta_{m,\rho_1}, \theta)$ is called the *bias*. Our main result is stated in Section 6.3. We provide a condition on the penalty function pen(.), so that the selected estimator satisfies a risk bound of the form

$$\mathbb{E}_\theta \left[l\left(\widehat{\theta}_{\widehat{m},\rho_1}, \theta\right)\right] \leq L \inf_{m\in\mathcal{M}} \left[l(\theta_{m,\rho_1}, \theta) + \varphi_{\max}(\Sigma)\frac{\mathrm{Card}(m)}{np^2}\right] \ , \tag{6.9}$$

where $\varphi_{\max}(\Sigma)$ is the largest eigenvalue of Σ. Contrary to most results in a spatial setting, this upper bound on the risk is nonasymptotic and holds in a general setting. The term $\varphi_{\max}(\Sigma)\frac{\mathrm{Card}(m)}{np^2}$ grows linearly with the size of m and goes to 0 with n and p. In Section 6.4, we prove that the variance term of a model m is of the same order as $\varphi_{\max}(\Sigma)\frac{\mathrm{Card}(m)}{np^2}$. Hence, the bound (6.9) tells us that the risk of $\widehat{\theta}_{\widehat{m},\rho_1}$ is smaller than a quantity which is the same order as the risk $\mathbb{E}_\theta[l(\widehat{\theta}_{m^*,\rho_1}, \theta)]$ of the oracle m^*. We say that the selected estimator achieves an *oracle-type inequality*.

In Section 6.4, we bound the asymptotic expectations $\mathbb{E}[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ and connect them to the variance terms in Bound (6.9). As a consequence, we prove that under mild assumptions on the target θ, the upper bound (6.9) is optimal from the asymptotic point of view (up to a multiplicative numerical constant). We discuss the assumptions in Section 6.5. In Section 6.6, we compute nonasymptotic minimax lower bounds with respect to the loss functions $l(.,.)$ and $\|.\|_F^2$. We then derive that under mild assumptions, our estimator $\widehat{\theta}_{\widehat{m},\rho_1}$ is minimax adaptive to the sparsity of θ and minimax adaptive to the decay of θ.

To our knowledge, these are the first oracle-type inequalities in a spatial setting. The computation of the minimax rates of convergence is also new. Moreover, most of our results are nonasymptotic. Although we have considered a square on the two-dimensional lattice, our method straightforwardly extends to any d-dimensional toroidal rectangle with $d \geq 1$. In the one-dimensional setting, we retrieve a oracle-type inequality that is close to the work of Shibata [Shi80]. Yet, he has stated an asymptotic oracle inequality for the estimation of autoregressive processes. In contrast, our result applies on a torus and is only optimal up to constants but it is nonasympotic and most of all applies for higher dimensional lattices. In Section 6.7, we further discuss the advantages and the weak points of our method. Moreover, we mention the extensions made in a subsequent chapter. All the proofs are postponed to Section 6.8 and to the appendix.

6.1.4 Some notations

Throughout this chapter, $L, L_1, L_2, \ldots$ denote constants that may vary from line to line. The notation $L(.)$ specifies the dependency on some quantities. For any matrix A, $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$ respectively refer the largest eigenvalue and the smallest eigenvalues of A. We recall that $\|A\|_F$ is the Frobenius norm of A. For any matrix θ of size p, $\|\theta\|_1$ stands for the sum of of the absolute values of the components of θ, we call it its l_1 norm. In the sequel, 0_p is the square matrix of size p whose indices are 0. Given $\rho > 0$, the ball $\mathcal{B}_1(0_p; \rho)$ is defined as the set of square matrices of size p whose l_1 norm is smaller than ρ. Finally, Table 7 gathers the notations involving X.

X	Matrix of size $p \times p$	Random field
X^v	Vector of length p^2	Vectorialized version of X
$\mathbf{X^v}$	Matrix of size $p^2 \times n$	Observations of X^v
$\mathbf{X}_i$	Matrix of size $p \times p$	i-th observation of the field X

Table 7: Notations for the random field and the data.

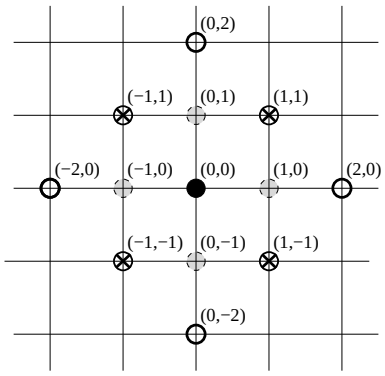


Figure 6.1: *Examples of models. The four gray nodes refer to m_1. The model m_2 also contains the nodes with a cross whereas m_3 contains all the nodes except $(0,0)$.*

6.2 Model selection procedure

In this section, we formally define our model selection procedure.

6.2.1 Collection of models

For any node (i,j) belonging to the lattice Λ, let us define the toroidal norm by

$$|(i,j)|_t^2 := [i \wedge (p-i)]^2 + [j \wedge (p-j)]^2$$

We aim at selecting a "good" neighborhood for the GMRF. Since X corresponds to some "spatial" process, it is natural to assume that nodes that are close to $(0,0)$ are more likely to be significant. This is why we restrict ourselves in the sequel to the collection $\mathcal{M}_1$ of neighborhoods.

Definition 6.2. A subset $m \subset \Lambda \setminus \{(0,0)\}$ belongs to $\mathcal{M}_1$ if there exists a number $r_m > 1$ such that

$$m = \{(i,j) \in \Lambda \setminus \{(0,0)\} \quad \text{s.t.} \quad |(i,j)|_t \leq r_m\} \ . \tag{6.10}$$

The collection $\mathcal{M}_1$ is totally ordered with respect to the inclusion and we therefore order our models $m_0 \subset m_1 \subset \ldots \subset m_i \ldots$. For instance, m_0 corresponds to the empty neighborhood whereas m_1 stands for the neighborhood of size 4. See Figure 6.1 for other examples.

For any model $m \in \mathcal{M}_1$, we define the vector space Θ_m as the subset of the elements of Θ whose support is included in m. We recall that Θ is defined in Section 6.1.1. Similarly Θ_m^{iso} is the subset of Θ^{iso} whose support is included in m. The dimensions of Θ_m and Θ_m^{iso} are respectively noted d_m and d_m^{iso}. Since we aim at estimating the positive matrix $(I_{p^2} - C(\theta))$, we shall consider the convex subsets of Θ_m^+ and $\Theta_m^{+,\text{iso}}$ that correspond to non-negative precision matrices.

$$\Theta_m^+ := \Theta_m \cap \Theta^+ \qquad \text{and} \qquad \Theta_m^{+,\text{iso}} := \Theta_m^{\text{iso}} \cap \Theta^{+,\text{iso}} \ . \tag{6.11}$$

For instance, the set $\Theta_{m_1}^+$ is in one to one correspondence with the sets of GMRFs whose neighborhood is made of the four nearest neighbors. Similarly, $\Theta_{m_1}^+$ is in one to one correspondence with the GMRFs with eight nearest neighbors. In our estimation procedure, we shall restrict ourselves to precision matrices whose largest eigenvalue is upper bounded by a constant. This is why we define the subsets Θ_{m_2,ρ_1}^+ and $\Theta_{m,\rho_1}^{+,\text{iso}}$ for any $\rho_1 \geq 2$.

$$\Theta_{m,\rho_1}^+ := \left\{\theta \in \Theta_m^+ \ , \varphi_{\max}\left[I_{p^2} - C(\theta)\right] < \rho_1\right\} \tag{6.12}$$

$$\Theta_{m,\rho_1}^{+,\text{iso}} := \left\{\theta \in \Theta_m^{+,\text{iso}} \ , \varphi_{\max}\left[I_{p^2} - C(\theta)\right] < \rho_1\right\} \ . \tag{6.13}$$

Finally, we need a generating family of the spaces Θ_m and Θ_m^{iso}. For any node $(i,j) \in \Lambda \setminus \{(0,0)\}$, let us define the $p \times p$ matrix $\Psi_{i,j}$ as

$$\Psi_{i,j}[k,l] := \begin{cases} 1 & \text{if} \quad (k,l) = (i,j) \text{ or } (k,l) = -(i,j) \\ 0 & \text{otherwise .} \end{cases} \tag{6.14}$$

Hence, Θ_m is generated by the matrices $\Psi_{i,j}$ for which (i,j) belongs to m. Similarly, for any $(i,j) \in \Lambda \setminus \{(0,0)\}$, let us define the matrix $\Psi^{\text{iso}}_{i,j}$ by

$$\Psi^{\text{iso}}_{i,j}[k,l] := \begin{cases} 1 & \text{if} \quad \exists g \in G,\ (k,l) = g.(i,j) \\ 0 & \text{otherwise .} \end{cases} \tag{6.15}$$

6.2.2 Estimation by Conditional Least Squares (CLS)

Let us turn to the conditional least squares estimator. For any $\theta' \in \Theta^+$, the criterion $\gamma_{n,p}(\theta')$ is defined by

$$\gamma_{n,p}(\theta') \quad := \quad \frac{1}{np^2} \sum_{i=1}^{n} \sum_{1 \le j_1, j_2 \le p} \left(\mathbf{X}_i[j_1,j_2] - \sum_{(l_1,l_2) \in \Lambda \setminus \{(0,0)\}} \theta'[l_1,l_2] \mathbf{X}_i[j_1+l_1, j_2+l_2] \right)^2 . \tag{6.16}$$

In a nutshell, $\gamma_{n,p}(\theta')$ is a least squares criterion that allows to perform the simultaneous linear regression of all $\mathbf{X}_i[j_1,j_2]$ with respect to the covariates $(\mathbf{X}_i[l_1,l_2])_{(l_1,l_2) \neq (j_1,j_2)}$. The advantage of this criterion is that it does not require the computation of a determinant of a huge matrix as for the likelihood. We shall often use an alternative expression of $\gamma_{n,p}(\theta')$ in terms of the factor $C(\theta')$ and the empirical covariance matrix $\overline{\mathbf{X}^{\mathbf{v}} \mathbf{X}^{\mathbf{v}*}}$:

$$\gamma_{n,p}(\theta') \quad = \quad \frac{1}{p^2} tr \left[(I_{p^2} - C(\theta')) \overline{\mathbf{X}^{\mathbf{v}} \mathbf{X}^{\mathbf{v}*}} (I_{p^2} - C(\theta')) \right] . \tag{6.17}$$

One proves the equivalence between these two expressions by coming back to the definition of $C(\theta')$. Let $\rho_1 > 2$ be fixed. For any model $m \in \mathcal{M}$, we compute the CLS estimators $\widehat{\theta}_{m,\rho_1}$ and $\widehat{\theta}^{\text{iso}}_{m,\rho_1}$ by minimizing the criterion $\gamma_{n,p}(.)$ as follows

$$\widehat{\theta}_{m,\rho_1} := \arg \min_{\theta' \in \overline{\Theta^+_{m,\rho_1}}} \gamma_{n,p}(\theta') \qquad \text{and} \qquad \widehat{\theta}^{\text{iso}}_{m,\rho_1} := \arg \min_{\theta' \in \overline{\Theta^{+,\text{iso}}_{m,\rho_1}}} \gamma_{n,p}(\theta') , \tag{6.18}$$

where $\overline{A}$ stands for the closure of the set A. The existence and the uniqueness of $\widehat{\theta}_{m,\rho_1}$ and $\widehat{\theta}^{\text{iso}}_{m,\rho_1}$ are ensured by the following lemma.

Lemma 6.3. *For any $\theta \in \Theta^+$, $\gamma_{n,p}(.)$ is almost surely strictly convex on $\overline{\Theta^+}$.*

The proof is postponed to the appendix. We discuss the dependency of $\widehat{\theta}_{m,\rho_1}$ on the parameter ρ_1 in Section 6.5. For stationary Gaussian fields, minimizing the CLS criterion $\gamma_{n,p}(.)$ over a set Θ^+_{m,ρ_1} is equivalent to minimizing the product of the conditional likelihoods $\left(X_{[i,j]} | X_{-\{i,j\}}\right)$, called *Conditional Pseudo-Likelihood* (CPL):

$$p\mathcal{L}_n(\theta', \mathbf{X}^{\mathbf{v}}) := \prod_{\substack{1 \le i \le n, \\ (j_1,j_2) \in \Lambda}} \mathcal{L}_{n,\theta'} \left(\mathbf{X}_i[j_1,j_2] | (\mathbf{X}_i)_{-\{j_1,j_2\}} \right) = \left(\sqrt{2\pi}\sigma \right)^{-np^2} \exp\left(-\frac{1}{2} \frac{np^2 \gamma_{n,p}(\theta')}{\sigma^2} \right) ,$$

where we recall that σ^2 refers to the conditional variance of any $X_{[i,j]}$. In fact, CLS estimators were first introduced by Besag [Bes75] who call them pseudolikelihood estimators since they minimize the CPL.

Let us define the function $\gamma(.)$ as an infinite sampled version of the CLS criterion $\gamma_{n,p}(.)$:

$$\gamma(\theta') := \mathbb{E}_\theta \left[\gamma_{n,p}(\theta') \right] = \mathbb{E}_\theta \left[\left(X_{[0,0]} - \sum_{(i,j) \neq (0,0)} \theta'[i,j] X_{[i,j]} \right)^2 \right] , \tag{6.19}$$

for any $\theta', \theta \in \Theta^+$. The function $\gamma(\theta')$ measures the prediction error of $X_{[0,0]}$ if one uses $\sum_{(i,j) \neq (0,0)} \theta'[i,j] X_{[i,j]}$ as a predictor. Moreover, it is a special case of the CMLS criterion introduced by Cressie and Verzelen in (Eq.10) of [CV08] to approximate a Gaussian field by a GMRF. Hence, one may interpret the CLS criterion as a finite sampled version of their approximation method. Observe that the function $\gamma(.)$ is minimized over Θ^+ at the point θ and that $\gamma(\theta) = \text{var}_\theta \left(X_{[0,0]} \left| X_{-\{0,0\}} \right. \right) = \sigma^2$. Moreover, the difference

$\gamma(\theta') - \gamma(\theta)$ equals the loss $l(\theta', \theta)$ defined by (6.7).

For any model $m \in \mathcal{M}$, we introduce the projections θ_{m,ρ_1} and $\theta^{\text{iso}}_{m,\rho_1}$ as the best approximation of θ in $\overline{\Theta^+_{m,\rho_1}}$ and $\overline{\Theta^{+,\text{iso}}_{m,\rho_1}}$.

$$\theta_{m,\rho_1} := \arg \min_{\theta' \in \overline{\Theta^+_{m,\rho_1}}} l(\theta', \theta) \qquad \text{and} \qquad \theta^{\text{iso}}_{m,\rho_1} := \arg \min_{\theta' \in \overline{\Theta^{+,\text{iso}}_{m,\rho_1}}} l(\theta', \theta) \ . \tag{6.20}$$

Since $\gamma(.)$ is strictly convex on Θ^+, the matrices θ_{m,ρ_1} and $\theta^{\text{iso}}_{m,\rho_1}$ are uniquely defined. By its definition (6.7), one may interpret $l(.,.)$ as an inner product on the space Θ; therefore, the orthogonal projection of θ onto the convex closed set $\overline{\Theta^+_{m,\rho_1}}$ (resp. $\overline{\Theta^{+,\text{iso}}_{m,\rho_1}}$) with respect to $l(.,.)$ is θ_{m,ρ_1} (resp. $\theta^{\text{iso}}_{m,\rho_1}$). It then follows from a property of orthogonal projections that the loss of $\widehat{\theta}_{m,\rho_1}$ is upper bounded by

$$l(\widehat{\theta}_{m,\rho_1}, \theta) \leq l(\theta_{m,\rho_1}, \theta) + l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1}) \ . \tag{6.21}$$

The first term $l(\theta_{m,\rho_1}, \theta)$ accounts for the bias, whereas the second term $l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})$ is a variance term. Observe that $\theta \in \Theta^+_m$ does not necessarily imply that the bias $l(\theta_{m,\rho_1}, \theta)$ is null because in general $\overline{\Theta^+_m} \neq \overline{\Theta^+_{m,\rho_1}}$. This will be the case only if θ satisfies the following hypothesis.

$$(\mathbb{H}_1): \qquad \varphi_{\max}(I_{p^2} - C(\theta)) < \rho_1 \ . \tag{6.22}$$

Assumption $(\mathbb{H}_1)$ is necessary to ensure the existence of a model $m \in \mathcal{M}$ such that the bias is zero (i.e. $\theta_{m,\rho_1} = \theta$). By identity (6.2), one observes that $(\mathbb{H}_1)$ is equivalent to a lower bound on the smallest eigenvalue of Σ, i.e. $\varphi_{\min}(\Sigma) \leq \frac{\sigma^2}{\rho_1}$. We further discuss $(\mathbb{H}_1)$ in Section 6.5.

For the sake of completeness, we recall the penalization criterion introduced in (6.6). Given a subcollection of models $\mathcal{M} \subset \mathcal{M}_1$ and a positive function $\text{pen} : \mathcal{M} \to \mathbb{R}^+$ that we call a penalty, we select a model as follows

$$\widehat{m} := \arg \min_{m \in \mathcal{M}} \left[\gamma_{n,p}\left(\widehat{\theta}_{m,\rho_1}\right) \right] + \text{pen}(m) \quad \text{and} \quad \widehat{m}^{\text{iso}} := \arg \min_{m \in \mathcal{M}} \left[\gamma_{n,p}\left(\widehat{\theta}^{\text{iso}}_{m,\rho_1}\right) \right] + \text{pen}(m) \ .$$

Observe that $\widehat{m}$ and $\widehat{m}^{\text{iso}}$ depend on ρ_1. For the sake clarity, we do not emphasize this dependency in the notation. In the sequel, we write $\widetilde{\theta}_{\rho_1}$ and $\widetilde{\theta}^{\text{iso}}_{\rho_1}$ for $\widehat{\theta}_{\widehat{m},\rho_1}$ and $\widehat{\theta}^{\text{iso},\rho_1}_{\widehat{m}^{\text{iso}}}$.

6.3 Main Result

We now provide a nonasymptotic upper bound for the risk of the estimators $\widetilde{\theta}_{\rho_1}$ and $\widetilde{\theta}^{\text{iso}}_{\rho_1}$. Let us recall that Σ stands for the covariance matrix of X^v.

Theorem 6.4. *Let K be a positive number larger than a universal constant K_0 and let $\mathcal{M}$ be a subcollection of $\mathcal{M}_1$. If for every model $m \in \mathcal{M}$,*

$$pen(m) \geq K\rho_1^2 \varphi_{max}(\Sigma) \frac{d_m}{np^2} \ , \tag{6.23}$$

then for any $\theta \in \Theta^+$, the estimator $\widetilde{\theta}_{\rho_1}$ satisfies

$$\mathbb{E}_\theta \left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right) \right] \leq L_1(K) \inf_{m \in \mathcal{M}} \left[l(\theta_{m,\rho_1}, \theta) + pen(m) \right] + L_2(K) \frac{\rho_1^2 \varphi_{max}(\Sigma)}{np^2} \ , \tag{6.24}$$

A similar bound holds if one replaces $\widetilde{\theta}_{\rho_1}$ by $\widetilde{\theta}^{\text{iso}}_{\rho_1}$, Θ^+ by $\Theta^{+,iso}$, θ_{m,ρ_1} by θ^{iso}_m, and d_m by d^{iso}_m.

The proof is postponed to Section 6.8.2. It is based on a novel concentration inequality for suprema of Gaussian chaos stated and proved in Section 6.8.1. The constant K_0 is made explicit in the proof. Observe that the theorem holds for any n, any p and that we have not performed any assumption on the target $\theta \in \Theta^+$ (resp. $\Theta^{+,\text{iso}}$). If the collection $\mathcal{M}$ does not contain the empty model, one gets the more readable upper bound

$$\mathbb{E}_\theta \left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right) \right] \leq L(K) \inf_{m \in \mathcal{M}} \left[l(\theta_{m,\rho_1}, \theta) + \text{pen}(m) \right] \ .$$

This theorem tells us that $\widetilde{\theta}_{\rho_1}$ essentially performs as well as the best trade-off between the bias term $l(\theta_{m,\rho_1},\theta)$ and $\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}$ that plays the role of a variance. Here are some additional comments.

Comments:

1. Consider the special case where the target θ belongs to some parametric set Θ_m^+ with $m \in \mathcal{M}$. Suppose that the hypothesis $(\mathbb{H}_1)$ defined in (6.22) is fulfilled. Choosing a penalty $\text{pen}(m) = K\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}$, we get

$$\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right] \leq L(K)\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}\ . \tag{6.25}$$

   We shall prove in Section 6.4.2 and 6.6.1 that this rate is optimal both from an asymptotic oracle and a minimax point of view. We have mentioned in Section 6.2.2 that $(\mathbb{H}_1)$ is necessary for the bound (6.25) to hold. If ρ_1 is chosen large enough, then Assumption $(\mathbb{H}_1)$ is fulfilled. We do not have access to this minimal ρ_1 that ensures $(\mathbb{H}_1)$, since it requires the knowledge of θ. Nevertheless, we argue in Section 6.5 that "moderate" values for ρ_1 ensure Assumption $(\mathbb{H}_1)$ when the model m is small.

2. We have mentioned in the introduction that our objective was to obtain oracle inequalities of the form

$$\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right] \leq L(K)\inf_{m\in\mathcal{M}}\mathbb{E}\left[l\left(\widehat{\theta}_{m,\rho_1},\theta\right)\right] = L(K)\mathbb{E}\left[\left(\widehat{\theta}_{m^*,\rho_1},\theta\right)\right]\ .$$

   This is why we want to compare the sum $l(\theta_{m,\rho_1},\theta)+\text{pen}(m)$ with $\mathbb{E}[l(\widehat{\theta}_{m,\rho_1},\theta)]$. First, we provide in Section 6.4.1 a sufficient condition so that the risk $\mathbb{E}[l(\widehat{\theta}_{m,\rho_1},\theta)]$ decomposes exactly as the sum $l(\theta_{m,\rho_1},\theta)+\mathbb{E}[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$. Moreover, we compute in Section 6.4.2 the asymptotic variance term $\mathbb{E}[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$ and compare it with the penalty term $\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}$. We shall then derive oracle type inequalities and discuss the dependency of the different bounds on $\varphi_{\max}(\Sigma)$.

3. Condition (6.23) gives a lower bound on the penalty function pen(.) so that the result holds. Choosing a proper penalty term according to (6.23) therefore requires an upper bound on the largest eigenvalue of Σ. However, such a bound is seldom known in practice. We shall mention in Section 6.7 a practical method to calibrate the penalty.

A bound similar to (6.24) holds for the Frobenius distance between the partial correlation matrices $(I_{p^2}-C(\theta))$ and $(I_{p^2}-C(\widetilde{\theta}_{\rho_1}))$.

Corrolary 6.5. *Assume the same as in Theorem 6.4, except that there is equality in (6.23). Then,*

$$\begin{aligned}\mathbb{E}_\theta\left[\|C(\widetilde{\theta}_{\rho_1})-C\left(\theta\right)\|_F^2\right] &\leq L_1\left(K\right)\frac{\varphi_{\max}(\Sigma)}{\varphi_{\min}(\Sigma)}\inf_{m\in\mathcal{M}}\left[\|C(\theta_{m,\rho_1})-C(\theta)\|_F^2+\frac{K\rho_1^2 d_m}{n}\right]\\ &+ L_2(K)\frac{\varphi_{\max}(\Sigma)}{\varphi_{\min}(\Sigma)}\frac{\rho_1^2}{n}\ .\end{aligned} \tag{6.26}$$

A similar result holds for isotropic GMRFs.

Proof of Corollary 6.5. This is a consequence of Theorem 6.4. By definition (6.7) of the loss function $l(.,.)$, the two following bounds hold

$$\begin{aligned}p^2 l(\theta_1,\theta_2) &\geq \varphi_{\min}(\Sigma)\|C(\theta_1)-C(\theta_2)\|_F^2\\ p^2 l(\theta_1,\theta_2) &\leq \varphi_{\max}(\Sigma)\|C(\theta_1)-C(\theta_2)\|_F^2\ .\end{aligned}$$

Gathering these bounds with (6.24) yields the result. □

The same comments as for Theorem (6.4) hold. We may express this Corollary 6.5 in terms of the risk $\mathbb{E}(\|\widetilde{\theta}_{\rho_1}-\theta\|_F^2)$, since $\|C(\theta_1)-C(\theta_2)\|_F^2 = p^2\|\theta_1-\theta_2\|_F^2$:

$$\begin{aligned}\mathbb{E}_\theta\left[\|\widetilde{\theta}_{\rho_1}-\theta\|_F^2\right] &\leq L_1\left(K\right)\frac{\varphi_{max}(\Sigma)}{\varphi_{min}(\Sigma)}\inf_{m\in\mathcal{M}}\left[\|\theta_{m,\rho_1}-\theta\|_F^2+\frac{K\rho_1^2 d_m}{np^2}\right]\\ &+ L_2(K)\frac{\varphi_{max}(\Sigma)}{\varphi_{min}(\Sigma)}\frac{\rho_1^2}{np^2}\end{aligned}$$

6.4 Parametric risk and asymptotic oracle inequalities

In this section, we study the risk of the parametric estimators $\widehat{\theta}_{m,\rho_1}$ in order to assess the optimality of Theorem 6.4.

6.4.1 Bias-variance decomposition

The properties of the parametric estimator $\widehat{\theta}_{m,\rho_1}$ and of the projection θ_{m,ρ_1} differ slightly whether θ_{m,ρ_1} belongs to the open set Θ^+_{m,ρ_1} or to its border. Observe that Hypothesis $(\mathbb{H}_1)$ defined in (6.22) does not necessarily imply that the projection θ_{m,ρ_1} belongs to Θ^+_m. This is why we introduce the condition $(\mathbb{H}_2)$.

$$(\mathbb{H}_2): \qquad \theta \in \mathcal{B}_1(0_p, 1) \qquad \Longleftrightarrow \qquad \|\theta\|_1 < 1. \tag{6.27}$$

The condition $\|\theta\|_1 < 1$ is equivalent to $[I_{p^2} - C(\theta)]$ is strictly diagonally dominant. Condition $(\mathbb{H}_2)$ implies that the largest eigenvalue of $(I_{p^2} - C(\theta))$ is smaller than 2 and therefore that $(\mathbb{H}_1)$ is fulfilled since ρ_1 is supposed larger than 2. We further discuss this assumption in Section 6.5.

Lemma 6.6. *Let $\theta \in \Theta^+$ such that $(\mathbb{H}_2)$ holds and let $m \in \mathcal{M}_1$. Then, the minimum of $\gamma(.)$ over Θ_m is achieved in $\Theta^+_{m,2}$. This implies that*

$$\theta_{m,\rho_1} = \arg\min_{\theta' \in \Theta_m} \gamma(\theta') \qquad \text{and} \qquad \gamma(\theta_{m,\rho_1}) = var_\theta\left(X_{[0,0]}|X_m\right) \ .$$

Besides, $\|\theta_{m,\rho_1}\|_1 \leq \|\theta\|_1$. The same results holds for θ^{iso}_{m,ρ_1} if θ in $\Theta^{+,\mathrm{iso}}$.

The purpose of this property is threefold. First, we derive that Assumption $(\mathbb{H}_2)$ ensures that θ_{m,ρ_1} belongs Θ^+_{m,ρ_1} and that the smallest eigenvalue of $(I_{p^2} - C(\theta_{m,\rho_1}))$ is larger than $1 - \|\theta\|_1$. Second, it allows to express the projection θ_{m,ρ_1} in terms of conditional expectation (Corollary 6.7). Finally, we deduce a bias-variance decomposition of the estimator $\widehat{\theta}_{m,\rho_1}$ (Corollary 6.8). In other words, the equality holds in (6.21).

Corrolary 6.7. *Let $\theta \in \Theta^+$ such that $(\mathbb{H}_2)$ holds and let $m \in \mathcal{M}_1$. The projection θ_{m,ρ_1} is uniquely defined by the equation*

$$\mathbb{E}_\theta\left(X_{[0,0]}|X_m\right) = \sum_{(i,j)\in m} \theta_{m,\rho_1[i,j]} X_{[i,j]} \ ,$$

and $\theta_{m,\rho_1[i,j]} = 0$ for any $(i,j) \notin m$. Similarly, if $\theta \in \Theta^{+,\mathrm{iso}}$ satisfies $(\mathbb{H}_2)$, then $\theta^{\mathrm{iso}}_{m,\rho_1}$ is uniquely defined by the equation

$$\mathbb{E}_\theta\left(X_{[0,0]}|X_m\right) = \sum_{(i,j)\in m} \theta^{\mathrm{iso}}_{m,\rho_1[i,j]} X_{[i,j]} \ ,$$

and $\theta^{\mathrm{iso}}_{m,\rho_1[i,j]} = 0$ for any $(i,j) \notin m$.

Consequently, $\sum_{1\leq i,j\leq p} \theta_{m,\rho_1[i,j]} X_{[i,j]}$ is the best linear predictor of $X_{[0,0]}$ given the covariates $X_{[i,j]}$ with $(i,j) \in m$. This is precisely the definition of the kriging parameters (see Cressie [Cre93] Ch.3 for an introduction). Hence, the matrix θ_{m,ρ_1} corresponds to the kriging parameters of $X_{[0,0]}$ with kriging neighborhood's range of r_m. The distance r_m is introduced in Definition 6.2 and stands for the radius of m.

Corrolary 6.8. *Let $\theta \in \Theta^+$ such that $(\mathbb{H}_2)$ holds and let $m \in \mathcal{M}_1$. The loss of $\widehat{\theta}_{m,\rho_1}$ decomposes as $l(\widehat{\theta}_{m,\rho_1}, \theta) = l(\theta_{m,\rho_1}, \theta) + l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})$. If θ belongs to $\Theta^{+,\mathrm{iso}}_m$ and $(\mathbb{H}_2)$ holds, then we also have the decomposition $l(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}, \theta) = l(\theta^{\mathrm{iso}}_{m,\rho_1}, \theta) + l(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}, \theta_{m,\rho_1})$.*

If θ does not satisfy Assumption $(\mathbb{H}_2)$, then θ_{m,ρ_1} does not necessarily belong to Θ^+_{m,ρ_1} and there may not be such a bias variance decomposition.

6.4.2 Asymptotic risk

In this section, we evaluate the risk of each estimator $\widehat{\theta}_{m,\rho_1}$ and use it as a benchmark to assess the result of Theorem 6.4. We have mentioned in Corollary 6.8 that under $(\mathbb{H}_2)$ the risk $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta)]$ decomposes into the sum of the bias $l(\theta_{m,\rho_1},\theta)$ and a variance term $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$. If this last quantity is of the same order as the penalty $\mathrm{pen}(m)$ introduced in (6.23), then Theorem 6.4 yields an oracle inequality. However, we are unable to express this variance term $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$ in a simple form. This is why we restrict ourselves to study the risks when n tends to infinity. Nevertheless, these results give us some hints to appreciate the strength and the weaknesses of Theorem 6.4 and the upper bound (6.25).

In the following proposition, we adapt a result of Guyon [Guy95] Sect.4.3.2 to obtain an asymptotic expression of the risk $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$. We first need to introduce some new notations. For any model m in the collection $\mathcal{M}_1\setminus\{\emptyset\}$, we fix a sequence $(i_k,j_k)_{k=1\dots d_m}$ of integers such that $(\Psi_{i_1,j_1},\dots,\Psi_{i_{d_m},j_{d_m}})$ is a basis of the space Θ_m. Then, $\chi_{m[0,0]}$ stands for the random vector of size d_m that contains the neighbors of $X_{[0,0]}$

$$\chi_{m[0,0]}{}^* := \left[tr\left(\Psi_{i_1,j_1}X^v\right),\dots,tr\left(\Psi_{i_{d_m},j_{d_m}}X^v\right)\right] .$$

Besides, for any $\theta\in\Theta^+$, we define the matrices V, W and IL_m as

$$\left\{\begin{array}{rcl} V & := & \mathrm{cov}_\theta(\chi_{m[0,0]}) \\ W_{[k,l]} & := & \frac{1}{p^2}tr\left[C\left(\Psi_{i_k,j_k}\right)\left[I_{p^2}-C(\theta_{m,\rho_1})\right]^2\left[I_{p^2}-C(\theta)\right]^{-2}C\left(\Psi_{i_l,j_l}\right)\right], \text{ for any } k=1,\dots,d_m \\ IL_m & := & \mathrm{Diag}\left(\|\Psi_{i_k,j_k}\|_F^2,\ k=1,\dots,d_m\right) , \end{array}\right.$$

where for any vector u, $\mathrm{Diag}(u)$ is the diagonal matrix whose diagonal elements are the components of u. We also define the corresponding quantities $\chi^{\mathrm{iso}}_{m[0,0]}$, V^{iso}, W^{iso}, and IL^{iso}_m in order to consider the isotropic estimator $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$.

Proposition 6.9. *Let m be a model in $\mathcal{M}_1\setminus\{\emptyset\}$ and let θ be an element of Θ^+_m that satisfies $(\mathbb{H}_1)$. Then, $\widehat{\theta}_{m,\rho_1}$ converges to θ in probability and*

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}_{m,\rho_1},\theta\right)\right] = 2\sigma^4 tr\left[IL_mV^{-1}\right] . \tag{6.28}$$

Let θ in Θ^+ such that $(\mathbb{H}_2)$ is fulfilled. Then, $\widehat{\theta}_{m,\rho_1}$ converges to θ_{m,ρ_1} in probability and

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1}\right)\right] = 2\sigma^4 tr(WV^{-1}) . \tag{6.29}$$

Both results still hold for the estimator $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$ if θ belongs to $\Theta^{+,\mathrm{iso}}$ and if one replace V, W, and IL_m by V^{iso}, W^{iso}, and IL^{iso}_m.

In the first case, Assumption $(\mathbb{H}_1)$ ensures that $\theta\in\Theta^+_{m,\rho_1}$, whereas Assumption $(\mathbb{H}_2)$ ensures that $\theta_{m,\rho_1}\in\Theta^+_{m,\rho_1}$. The proof is based on the extension of Guyon's approach in the toroidal framework.

The expressions (6.28) and (6.29) are not easily interpretable in the present form. This is why we first derive (6.28) when θ is zero. Observe that it is equivalent to the independence of the $(X_{[i,j]})_{(i,j)\in\Lambda}$.

Example 6.10. *Assume that θ is zero. Then, for any model $m\in\mathcal{M}_1$, the asymptotic risks of $\widehat{\theta}_{m,\rho_1}$ and $\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1}$ satisfy*

$$\lim_{n\to+\infty} np^2\mathbb{E}_{0_p}\left[l\left(\widehat{\theta}_{m,\rho_1},0_p\right)\right] = 2\sigma^2 d_m \text{ and } \lim_{n\to+\infty} np^2\mathbb{E}_{0_p}\left[l\left(\widehat{\theta}^{\mathrm{iso}}_{m,\rho_1},0_p\right)\right] = 2\sigma^2 d^{\mathrm{iso}}_m ,$$

where we recall that d^{iso}_m is the dimension of the space Θ^{iso}_m.

Proof. Since the components of X are independent, the matrix V equals $\sigma^2 IL_m$. We conclude by applying Proposition 6.9 □

Therefore, when the variables $X_{[i,j]}$ are independent, the asymptotic risk of $\widehat{\theta}_{m,\rho_1}$ equals, up to a factor 2, the variance term of the least squares estimator in the fixed design Gaussian regression framework. This quantity is of the same order as the penalty introduced in Section 6.3. When the matrix θ is non zero, we can lower bound the limits (6.28) and (6.29).

Corrolary 6.11. *Let m be a model in $\mathcal{M}_1$ and let $\theta \in \Theta_m^+$ that satisfies $(\mathbb{H}_1)$. Then, the variance term is asymptotically lower bounded as follows*

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}_{m,\rho_1},\theta\right)\right] \geq L\sigma^2\varphi_{min}\left[I_{p^2}-C(\theta)\right]d_m = L\sigma^4\frac{d_m}{\varphi_{max}(\Sigma)}\ , \tag{6.30}$$

where L is a universal constant. Let $\theta \in \Theta^+$ that satisfies $(\mathbb{H}_2)$. For any model $m \in \mathcal{M}_1$,

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1}\right)\right] \geq L\sigma^2\left(1-\|\theta\|_1\right)^3 d_m\ , \tag{6.31}$$

Again, analogous lower bounds hold for $\widehat{\theta}^{\text{iso}}_{m,\rho_1}$ when θ belongs to $\Theta^{\text{iso},+}$. This corollary states that asymptotically with respect to n the variance term of $\widehat{\theta}_{m,\rho_1}$ is larger than the order $\frac{d_m}{np^2}$. This expression is not really surprising since d_m stands for the dimension of the model m and np^2 corresponds to the number of data observed. Let define $R_{\theta,\infty}(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1}) := \lim_{n\to+\infty} np^2\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$ as the asymptotic variance term for $\widehat{\theta}_{m,\rho_1}$ rescaled by the number np^2 of observations.

The first part of the corollary (6.30) states that from an asymptotic point of view the upper bound (6.25) is optimal. By Theorem 6.4, if we choose $\text{pen}(m) = K\rho_1^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}$, then it holds that

$$\mathbb{E}\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right] \leq L\left(K,\rho_1,\varphi_{\min}\left[I_{p^2}-C(\theta)\right]\right)\frac{R_{\theta,\infty}(\widehat{\theta}_{m,\rho_1},\theta)}{np^2},$$

for any model $m \in \mathcal{M}\setminus\emptyset$ and any $\theta \in \Theta_m^+$ that satisfies $(\mathbb{H}_1)$. This property holds for any n and any p. Hence, $\widetilde{\theta}_{\rho_1}$ performs as well as the parametric estimator $\widehat{\theta}_{m,\rho_1}$ if the support of θ belongs to some unknown model m and if θ satisfies $(\mathbb{H}_1)$.

If we assume that $\|\theta\|_1 < 1$ (Hypothesis $(\mathbb{H}_2)$), we are able to derive a stronger result.

Proposition 6.12. *Considering $K \geq K_0$, $\rho_1 \geq 2$, $\eta < 1$ and a collection $\mathcal{M} \subset \mathcal{M}_1\setminus\emptyset$, we define the estimator $\widetilde{\theta}_{\rho_1}$ with the penalty $\text{pen}(m) = K\rho_1^2\frac{d_m}{np^2(1-\eta)}$. Then, the risk of $\widetilde{\theta}_{\rho_1}$ is upper bounded by*

$$\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right] \leq L(K,\rho_1,\eta)\inf_{m\in\mathcal{M}}\left\{l\left(\theta_{m,\rho_1},\theta\right)+\frac{R_{\theta,\infty}\left(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1}\right)}{np^2}\right\}\ , \tag{6.32}$$

for any $\theta \in \Theta^+ \cap \mathcal{B}_1(0_p,\eta)$.

Observe that this property holds for any n and any p. If the matrix θ is strictly diagonally dominant, we therefore obtain an upper bound similar to an oracle inequality, except that the variance term $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})]$ has been replaced by its asymptotic counterpart $R_{\theta,\infty}(\widehat{\theta}_{m,\rho_1},\theta_{m,\rho_1})/(np^2)$. However, this inequality is not valid uniformly over any $\eta < 1$: when η converges to one, the constant $L(K,\rho_1,\eta)$ tends to infinity. Indeed, if $\|\theta\|_1$ converges to one, the lower bound (6.31) on the variance term can behave like $(1-\|\theta\|_1)^3 d_m/(np^2)$ for some matrices θ whereas the penalty term $d_m/[np^2(1-\|\theta\|_1)]$ tends to infinity.

In the remaining part of the section, we illustrate that the constant $L(K,\eta,\rho_1)$ has to go to infinity when η goes to one. Let us consider the model m_1. It consists of GMRFs with 4-nearest neighbors.

Example 6.13. *Let θ be a non zero element of $\Theta^{\text{iso}}_{m_1}$, then the asymptotic risk of $\widehat{\theta}^{\text{iso}}_{m_1,\rho_1}$ simplifies as*

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}^{\text{iso}}_{m_1,\rho_1},\theta\right)\right] = 2\frac{\sigma^4\theta_{[1,0]}}{cov(X_{[1,0]},X_{[0,0]})}\ . \tag{6.33}$$

If we let the size p of the network tend to infinity and $\theta_{[1,0]}$ go to $1/4$, the risk is equivalent to

$$\lim_{p\to+\infty}\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}^{\text{iso}}_{m_1,\rho_1},\theta\right)\right] \underset{\theta_{[1,0]}\to 1/4}{\sim} \frac{16\sigma^2(1-4\theta_{[1,0]})}{\log(16)}\ .$$

If follows from the second result that the lower bound (6.30) is sharp since in this particular case $\varphi_{\min}(I_{p^2} - C(\theta)) = \sigma^2(1 - 4\theta_{[1,0]})$. When $\theta_{[1,0]}$ tends to $1/4$, then $\|\theta\|_1$ tends to one and $\mathbb{E}_\theta[l(\widehat{\theta}^{iso}_{m_1,\rho_1}, \theta)]$ behaves like $\sigma^2(1 - \|\theta\|_1)d^{\text{iso}}_{m_1}/(np^2)$ whereas the penalty pen(m_1) given in Theorem 6.4 has to be larger than $\sigma^2 d^{\text{iso}}_{m_1}/[np^2(1 - \|\theta\|_1)]$. Hence, the variance term and the penalty pen(.) are not necessarily of the same order when $\|\theta\|_1$ tends to one. Theorem 6.4 cannot lead to an oracle inequality of the type (6.32), which is valid uniformly on $\eta < 1$.

Example 6.14. *Let α be a positive number smaller than $1/4$. For any integer p which is divisible by 4, we define the $p \times p$ matrix $\theta^{(p)}$ by*

$$\begin{cases} {\theta^{(p)}}_{[p/4,p/4]} = {\theta^{(p)}}_{[-p/4,p/4]} = {\theta^{(p)}}_{[p/4,-p/4]} = {\theta^{(p)}}_{[-p/4,-p/4]} & := \quad \alpha \\ {\theta^{(p)}}_{[i,j]} & := \quad 0 \quad \textit{else} \ . \end{cases}$$

Then, the variance term is asymptotically lower bounded as follows

$$\lim_{p\to+\infty} \lim_{n\to+\infty} np^2 \mathbb{E}_{\theta^{(p)}} \left[l\left(\widehat{{\theta^{(p)}}}^{\text{iso}}_{m_1,\rho_1}, [\theta^{(p)}]^{\text{iso}}_{m_1,\rho_1} \right) \right] \geq \frac{L\sigma^2}{1 - 4\alpha} \ .$$

Consequently, this variance is of order $\sigma^2 \frac{d^{\text{iso}}_m}{np^2(1-\|\theta\|_1)} = \varphi_{\max}(\Sigma)\frac{d^{\text{iso}}_m}{np^2}$ when $\|\theta\|_1$ goes to one. The penalty pen(m) introduced in Proposition 6.12 is therefore a sharp upper bound of the variance terms.

On the one hand, we take a penalty pen(m) larger than $\sigma^2 \frac{d_m}{np^2(1-\|\theta\|_1)}$, whereas in some cases the variance of $\widehat{\theta}_{m,\rho_1}$ is of order $\sigma^2(1 - \|\theta\|_1)\frac{d_m}{np^2}$. The bound (6.32) cannot therefore hold uniformly over any $\eta < 1$. We think that it is intrinsic to the penalization strategy.

6.5 Comments on the assumptions

In this section, we discuss the dependency of the estimators $\widehat{\theta}_{m,\rho_1}$ on ρ_1 as well as Assumptions $(\mathbb{H}_1)$ and $(\mathbb{H}_2)$.

Dependency of $\widehat{\theta}_{m,\rho_1}$ on ρ_1. We recall that the estimator $\widehat{\theta}_{m,\rho_1}$ is defined in (6.18) as the minimizer of the CLS empirical contrast $\gamma_{n,p}(.)$ over Θ^+_{m,ρ_1}. It may seem restrictive to perform the minimization over the set Θ^+_{m,ρ_1} instead of Θ^+_m. Nevertheless, we advocate that it is not the case, at least for small models. Let us indeed define

$$\rho(m) := \sup_{\theta\in\Theta^+_m} \varphi_{\max}\left[I_{p^2} - C(\theta)\right] \qquad \text{and} \qquad \rho^{iso}(m) := \sup_{\theta\in\Theta^{+,iso}_m} \varphi_{\max}\left[I_{p^2} - C(\theta)\right] \ .$$

The quantities $\rho(m)$ and $\rho^{iso}(m)$ are finite since Θ^+_m is bounded. If one takes ρ_1 larger than $\rho(m)$ (resp. $\rho^{\text{iso}}(m)$), then the set Θ^+_{m,ρ_1} (resp. $\Theta^{+,\text{iso}}_{m,\rho_1}$) is exactly Θ^+_m (resp. $\Theta^{+,\text{iso}}_m$). We illustrate in Table 8 that $\rho(m)$ and $\rho^{\text{iso}}(m)$ are small, when the model m is small. Consequently, choosing a moderate value for ρ_1 is not really restrictive for small models. However, when the size of the model m increases, the sets Θ^+_{m,ρ_1} and Θ^+_m become different for moderate values of ρ_1.

d_m	2	4	6	10
$\rho(m)$	2.0	4.0	5.0	6.8
d^{iso}_m	1	2	3	4
$\rho^{\text{iso}}(m)$	2.0	4.0	5.0	6.8

Table 8: Approximate computation of $\rho(m)$ and $\rho^{\text{iso}}(m)$ for the four smallest models with $p = 50$.

Assumption $(\mathbb{H}_1)$ defined in (6.22) states that the largest eigenvalue of $(I_{p^2} - C(\theta))$ is smaller than ρ_1. We have illustrated in Table 8 that if the support of θ belongs to a small model m, then the maximal absolute value of $(I_{p^2} - C(\theta))$ is small. Hence, Assumption $(\mathbb{H}_1)$ is ensured for "moderate" values of ρ_1 as soon as the support of θ belongs to some small model. In practice, we do not know in advance if a given choice of rho_1 ensure $(\mathbb{H}_1)$. We discuss an extension of the procedure which has not this drawback in Section 6.7.

If θ is not sparse but approximately sparse it is likely that the largest eigenvalue of θ remain moderate.

Assumption $(\mathbb{H}_2)$ defined in (6.27) states that $\theta \in \mathcal{B}_1(0_p, 1)$ or equivalently that the matrix $(I_{p^2} - C(\theta))$ is diagonally dominant. Rue and Held prove in [RH05] Sect.2.7 that $\Theta^+_{m_1}$ is included in $\mathcal{B}_1(0_p, 1)$. They also point that a small part of $\Theta^+_{m_2}$ does not belong to $\mathcal{B}_1(0_p, 1)$. In fact, Assumption $(\mathbb{H}_2)$ becomes more and more restrictive if the support of θ becomes larger. Nevertheless, Assumption $(\mathbb{H}_2)$ is also quite common in the literature (as for instance in [Guy95]).

If one looks closely at our proofs involving Assumptions $(\mathbb{H}_2)$, one realizes that this assumptions is only made to ensure the following facts.

1. The *projection* θ_{m,ρ_1} belongs to the open set Θ^+_{m,ρ_1} for any model $m \in \mathcal{M}$ (Corollary 6.8).
2. The smallest eigenvalue of $(I_{p^2} - C(\theta_{m,\rho_1}))$ is lower bounded by some positive number ρ_2,.uniformly over all models $m \in \mathcal{M}$.

From empirical observations, Assumption these two last facts seem far more restrictive than $(\mathbb{H}_2)$. We used $(\mathbb{H}_2)$ in the statement of the results, because we did not find a weaker but simple condition that ensure facts 1 and 2.

6.6 Minimax rates

In Theorem 6.4 and Proposition 6.12 we have shown that under mild assumptions on θ the estimator $\widetilde{\theta}_{\rho_1}$ behaves almost as well as the best estimator among the family $\{\widehat{\theta}_{m,\rho_1},\, m \in \mathcal{M}\}$. We now compare the risk of $\widetilde{\theta}_{\rho_1}$ with the risk of any other possible estimator $\widehat{\theta}$. This includes comparison with maximum likelihood methods. There is no hope to make a pointwise comparison with an arbitrary estimator. Therefore, we classically consider the maximal risk over some suitable subsets $\mathcal{T}$ of Θ^+. The *minimax risk* over the set $\mathcal{T}$ is given by $\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{T}} \mathbb{E}_\theta[l(\widehat{\theta}, \theta)]$, where the infimum is taken over all possible estimators $\widehat{\theta}$ of θ. Then, the estimator $\widetilde{\theta}_{\rho_1}$ is said to be *approximately minimax* with respect to the set $\mathcal{T}$ if the ratio

$$\frac{\sup_{\theta \in \mathcal{T}} \mathbb{E}_\theta \left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right]}{\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{T}} \mathbb{E}_\theta \left[l\left(\widehat{\theta}, \theta\right)\right]}$$

is smaller than a constant that does not depend on σ^2, n or p. An estimator is said to be *adaptive* to a collection $(\mathcal{T}_i)_{i \in \mathcal{I}}$ if it is simultaneously minimax over each $\mathcal{T}_i$. The problem of designing adaptive estimation procedures is in general difficult. It has been extensively studied in the fixed design Gaussian regression framework. See for instance [BM01] for a detailed discussion. In the sequel, we adapt some of their ideas to the GMRF framework.

We prove in Section 6.6.1 that the estimator $\widetilde{\theta}_{\rho_1}$ is adaptive to the unknown sparsity of the matrix θ. Moreover, it is also adaptive if we consider the Frobenius distance between partial correlation matrices. In Section 6.6.2, we show that $\widetilde{\theta}_{\rho_1}$ is also adaptive to the rates of decay of the bias.

We need to restrain ourselves to set of matrices θ such that the largest eigenvalue of the covariance matrix Σ is uniformly bounded. This is why we define

$$\forall \rho_2 > 1\ , \quad \mathcal{U}(\rho_2) := \left\{ \theta \in \Theta, \varphi_{\min}\left(I_{p^2} - C(\theta)\right) \geq \frac{1}{\rho_2} \right\}\ . \tag{6.34}$$

Observe that $\theta \in \mathcal{U}(\rho_2)$ is exactly equivalent to $\varphi_{\max}(\Sigma) \leq \sigma^2 \rho_2$ since $\Sigma = \sigma^2(I_{p^2} - C(\theta))$.

6.6.1 Adapting to unknown sparsity

In this subsection, we prove that under mild assumptions the penalized estimator $\widetilde{\theta}_{\rho_1}$ is adaptive to the unknown sparsity of θ. We first lower bound the minimax rate of convergence on given hypercubes.

Definition 6.15. Let m be a model in the collection $\mathcal{M}_1 \setminus \emptyset$. We consider $\left(\Psi_{i_1,j_1}, \ldots, \Psi_{i_{d_m},j_{d_m}}\right)$ a basis of the space Θ_m defined by (6.14). For any $\theta' \in \Theta^+_m$, the hypercube $\mathcal{C}_m(\theta', r)$ is defined as

$$\mathcal{C}_m\left(\theta', r\right) := \left\{ \theta' + \sum_{k=1}^{d_m} \Psi_{i_k,j_k} \phi_k,\ \phi \in \{0,1\}^{d_m} \right\}\ ,$$

if the positive number r is small enough so that $\mathcal{C}_m(\theta', r) \subset \Theta^+$. For any $\theta' \in \Theta_m^{+,iso}$, we analogously define the hypercubes $\mathcal{C}_m^{iso}(\theta', r)$ using a basis $\left(\Psi_{i_1,j_1}^{iso}, \ldots, \Psi_{i_{d_m},j_{d_m}}^{iso}\right)$.

Proposition 6.16. *Let m be a model in $\mathcal{M}_1 \setminus \emptyset$ whose dimension d_m is smaller than $p\sqrt{n}$. Then, for any estimator $\widehat{\theta}$,*

$$\sup_{\theta \in \Theta_m^+} \mathbb{E}_\theta \left[l\left(\widehat{\theta}, \theta\right)\right] \geq \sup_{\theta \in \Theta_{m,2}^+} \mathbb{E}_\theta \left[l\left(\widehat{\theta}, \theta\right)\right] \geq L\sigma^2 \frac{d_m}{np^2} \ . \tag{6.35}$$

Let θ' be an element of Θ_m^+ that satisfies $(\mathbb{H}_2)$. For any estimator $\widehat{\theta}$ of θ,

$$\sup_{\theta \in \mathrm{Co}\left[\mathcal{C}_m\left(\theta', \frac{1-\|\theta'\|_1}{\sqrt{np^2}}\right)\right]} \mathbb{E}_\theta \left[l\left(\widehat{\theta}, \theta\right)\right] \geq L\sigma^2 \varphi_{\min}^2 \left[I_{p^2} - C(\theta')\right] \frac{d_m}{np^2} \ , \tag{6.36}$$

where $\mathrm{Co}\left[\mathcal{C}_m\left(\theta', r\right)\right]$ denotes the convex hull of $\mathcal{C}_m\left(\theta', r\right)$.

An analogous result holds for isotropic hypercubes. The first bound (6.35) means that for any estimator $\widehat{\theta}$, the supremum of the risks $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ over Θ_m^+ is larger than $\sigma^2 \frac{d_m}{np^2}$ (up to some numerical constant). This rate $\sigma^2 \frac{d_m}{np^2}$ is achieved by the CLS estimator by Theorem 6.4.

The second lower bound (6.36) is of independent interest. It implies that in a small neighborhood of θ' the risk $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ is larger than $\sigma^2\varphi_{\min}^2[I_{p^2} - C(\theta')]d_m/np^2$. This confirms the lower bound (6.30) of Proposition 6.11 in a nonasymptotic way. Indeed, these two expressions match up to a factor $\varphi_{\min}[I_{p^2} - C(\theta')]$. This difference comes from the fact that the lower bound (6.36) holds for any estimator $\widehat{\theta}$. Bound (6.36) is sharp in the sense that the maximum likelihood estimator $\widehat{\theta}_{m_1}^{\mathrm{iso},mle}$ of isotropic GMRF in m_1 exhibits an asymptotic risk of order $\sigma^2\varphi_{\min}^2[I_{p^2} - C(\theta)]/(np^2)$ for the parameter θ studied in Example 6.13. It is shown using the methodology introduced in the proof of Example 6.13. We now state that $\widetilde{\theta}_\rho$ is adaptive to the sparsity of m.

Corrolary 6.17. *Considering $K \geq K_0$, $\rho_1 \geq 2$, $\rho_2 > 2$ and a collection $\mathcal{M} \subset \mathcal{M}_1$, we define the estimator $\widetilde{\theta}_{\rho_1}$ with the penalty $pen(m) = K\sigma^2\rho_1^2\rho_2\frac{d_m}{np^2}$. For any non empty model m,*

$$\sup_{\theta \in \Theta_{m,\rho_1}^+ \cap \mathcal{U}(\rho_2)} \mathbb{E}_\theta \left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right] \leq L\left(K, \rho_1, \rho_2\right) \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_{m,\rho_1}^+ \cap \mathcal{U}(\rho_2)} \mathbb{E}\left[l\left(\widehat{\theta}, \theta\right)\right] \ , \tag{6.37}$$

where $\mathcal{U}(\rho_2)$ is defined in (6.34).

A similar result holds for $\widetilde{\theta}_{\rho_1}^{\mathrm{iso}}$ and $\Theta_{m,\rho_1}^{+,\mathrm{iso}}$. Corollary 6.17 is nonasymptotic and applies for any n and any p. If θ belongs to some model m, then the optimal risk from a minimax point of view is of order $\frac{d_m}{np^2}$. In practice, we do not know the true model m. Nevertheless, the procedure simultaneously achieves the minimax rates for all supports m possible. This means that $\widetilde{\theta}_{\rho_1}$ reaches this minimax rate $\frac{d_m}{np^2}$ *without* knowing in advance the true model m.

The procedure is not adaptive to the smallest and the largest eigenvalue of $(I_{p^2} - C(\theta))$ which correspond to ρ_1 and ρ_2. Indeed, the constant $L\left(K, \rho_1, \rho_2\right)$ depends on ρ_1 and ρ_2. We are not aware of any other covariance estimation procedure which is really adaptive the smallest and the largest eigenvalue of the matrix.

Finally, $\widetilde{\theta}_{\rho_1}$ exhibits the same adaptive properties with respect to the Frobenius norm.

Corrolary 6.18. *Under the same assumptions as Corollary 6.17,*

$$\sup_{\theta \in \Theta_{m,\rho_1}^+ \cap \mathcal{U}(\rho_2)} \mathbb{E}_\theta \left[\|C(\widetilde{\theta}_{\rho_1}) - C(\theta)\|_F^2\right] \leq L\left(K, \rho_1, \rho_2\right) \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_{m,\rho_1}^+ \cap \mathcal{U}(\rho_2)} \mathbb{E}\left[\|C(\widehat{\theta}) - C(\theta)\|_F^2\right] \ . \tag{6.38}$$

Proof of Corollary 6.18. As in the proof of Corollary 6.5, we observe that

$$\|C(\theta_1) - C(\theta_2)\|_F \geq \frac{p^2\rho_1}{\sigma^2} l(\theta_1, \theta_2) \ ,$$

if θ satisfies Assumption $(\mathbb{H}_1)$. We conclude by applying Proposition 6.16 and Corollary 6.5. □

6.6.2 Adapting to the decay of the bias

In this section, we prove that the estimator $\widetilde{\theta}_{\rho_1}$ is adaptive to a range of sets that we call *pseudo-ellipsoids*.

Definition 6.19 (Pseudo-ellipsoids)**.** Let $(a_j)_{1\leq j\leq \mathrm{Card}(\mathcal{M}_1)}$ be a non-increasing sequence of positive numbers. Then, $\theta\in\Theta^+$ belongs to the *pseudo-ellipsoid* $\mathcal{E}(a)$ if and only if

$$\sum_{i=1}^{\mathrm{Card}(\mathcal{M}_1)} \frac{var_\theta\left(X_{[0,0]}|X_{\mathcal{N}(m_{i-1})}\right)-var_\theta\left(X_{[0,0]}|X_{\mathcal{N}(m_i)}\right)}{a_i^2} \quad\leq\quad 1\ . \tag{6.39}$$

Condition (6.39) measures how fast $\mathrm{var}_\theta(X_{[0,0]}|X_{\mathcal{N}(m_i)})$ tends to $\mathrm{var}_\theta(X_{[0,0]}|X_{\Lambda\setminus\{(0,0)\}})$. Suppose that Assumption $(\mathbb{H}_2)$ defined in (6.27) is fulfilled. By Corollary 6.7, $\mathrm{var}_\theta\left(X_{[0,0]}|X_{\mathcal{N}(m_i)}\right)$ is the sum of $l(\theta_{m_i},\theta)$ and σ^2 and Condition (6.39) is equivalent to

$$\sum_{i=1}^{\mathrm{Card}(\mathcal{M}_1)} \frac{l\left(\theta_{m_{i-1}},\theta\right)-l(\theta_{m_i},\theta)}{a_i^2}\leq 1\ . \tag{6.40}$$

Hence, the sequence (a_i) gives some condition on the *rate of decay* of the bias when the dimension of the model increases. These sets $\mathcal{E}(a)$ are not true ellipsoids. Nevertheless, one may consider them as counterparts of the classical ellipsoids studied in the fixed design Gaussian regression framework (see for instance [Mas07] Sect.4.3).

To prove adaptivity, we shall need the equivalence between Conditions (6.39) and (6.40). This equivalence holds if $\mathrm{var}_\theta\left(X_{[0,0]}|X_{\mathcal{N}(m_i)}\right)$ decomposes as $l(\theta_{m_i},\theta)+\sigma^2$, for any model $m\in\mathcal{M}_1$. As mentioned earlier, Assumption $(\mathbb{H}_2)$ is sufficient (but not necessary) for this property to hold. This is why we restrict ourselves to study sets of the type $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)$. We shall also perform the following assumption on the ellipsoids $\mathcal{E}(a)$

$$(\mathbb{H}_a):\qquad a_i^2\leq\frac{\sigma^2}{d_{m_i}},\ \text{for any } 1\leq i\leq|\mathcal{M}_1|\ .$$

It essentially means that the sequence (a_i) converges fast enough towards 0. For instance, all the sequences $a_i=\sigma(d_{m_i})^{-s}$ with $s\geq 1/2$ satisfy $(\mathbb{H}_a)$.

Proposition 6.20. *Under Assumption $(\mathbb{H}_a)$, the minimax rate of estimation on $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(2)$ is lower bounded by*

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(2)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right]\geq L\sup_{1\leq i\leq Card(\mathcal{M}_1)}\left(a_i^2\wedge\sigma^2\frac{d_{m_i}}{np^2}\right)\ . \tag{6.41}$$

This lower bound is analogous to the minimax rate of estimation for ellipsoids in the Gaussian sequence model. Gathering Theorem 6.4 and Proposition 6.20 enables to derive adaptive properties for $\widetilde{\theta}_{\rho_1}$.

Proposition 6.21. *Considering $K\geq K_0$, $\rho_1\geq 2$, $\rho_2>2$ and the collection $\mathcal{M}_1$, we define the estimator $\widetilde{\theta}_{\rho_1}$ with the penalty $pen(m)=K\sigma^2\rho_1^2\rho_2\frac{d_m}{np^2}$. For any ellipsoid $\mathcal{E}(a)$ that satisfies $(\mathbb{H}_a)$ and such that $a_1^2\geq\frac{1}{np^2}$, the estimator $\widetilde{\theta}_{\rho_1}$ is minimax over the set $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)$:*

$$\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)}\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right]\leq L(K,\rho_1,\rho_2)\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)}\mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right]\ . \tag{6.42}$$

Let us first illustrate this result. We have mentioned earlier, that Assumption $(\mathbb{H}_a)$ is satisfied for all sequences $a_i=\sigma(d_{m_i})^{-s}$ with $s\geq 1/2$. We note $\mathcal{E}'(s)$ such a pseudo-ellipsoid. By Propositions 6.20 and 6.21, the minimax rate over *one* pseudo ellipsoid $\mathcal{E}'(s)$ is $\sigma^2(np^2)^{-2s/(1+2s)}$. The larger s is, the faster the minimax rates is. The estimator $\widetilde{\theta}_{\rho_1}$ achieves simultaneously the rate $\sigma^2(np^2)^{-2s/(1+2s)}$ for all $s\geq 1/2$. Consequently, $\widetilde{\theta}_{\rho_1}$ is adaptive to the rate s of decay of the bias: it achieves the optimal rates without knowing s in advance.

Let us further comment Proposition 6.21. By (6.42), the estimator $\widetilde{\theta}_{\rho_1}$ is adaptive over $\mathcal{E}(a)\cap\mathcal{B}_1(0_p,1)\cap\mathcal{U}(\rho_2)$ for all sequences (a) such that $(\mathbb{H}_a)$ is satisfied and such that $a_1^2\geq\frac{1}{np^2}$. Again, the result applies for any n and any p. The condition $a_1^2\geq 1/(np^2)$ is classical. It ensures that the pseudo-ellipsoid $\mathcal{E}(a)$ is not degenerate, i.e. that the minimax rates of estimation is not smaller than $\sigma^2/(np^2)$. We have explained earlier that we restricts ourselves to parameters θ in $\mathcal{B}_1(0_p,1)$ only because this enforces the equivalence between (6.39) and (6.40). In contrast, the hypothesis $\varphi_{\max}(\Sigma)\leq\sigma^2\rho_2$ is really necessary because we fail to be adaptive to ρ_2.

Corrolary 6.22. *Under Assumption* $(\mathbb{H}_a)$*, the minimax rate of estimation over* $\mathcal{E}(a) \cap \mathcal{U}(2) \cap \mathcal{B}_1(0_p, 1)$ *is lower bounded by*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p,1) \cap \mathcal{U}(2)} \mathbb{E}_\theta \left[\|C(\widehat{\theta}) - C(\theta)\|_F^2 \right] \geq L \sup_{1 \leq i \leq Card(\mathcal{M}_1)} \left(a_i^2 p^2 \wedge \frac{d_{m_i}}{n} \right) .$$

Under the same assumptions as Corollary 6.21,

$$\sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p,1) \cap \mathcal{U}(\rho_2)} \mathbb{E}_\theta \left[\|C(\widehat{\theta}) - C(\theta)\|_F^2 \right] \leq L(K, \rho_1, \rho_2) \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p,1) \cap \mathcal{U}(\rho_2)} \mathbb{E}_\theta \left[\|C(\widehat{\theta}) - C(\theta)\|_F^2 \right] .$$

Proof of Corollary 6.22. As in the proof of Corollary 6.5, we observe that

$$\|C(\theta_1) - C(\theta_2)\|_F \geq p^2 [\varphi_{\max}(\Sigma)]^{-1} l(\theta_1, \theta_2) \geq \frac{p^2}{\rho_2 \sigma^2} l(\theta_1, \theta_2) ,$$

$$\|C(\theta_1) - C(\theta_2)\|_F \leq p^2 [\varphi_{\min}(\Sigma)]^{-1} l(\theta_1, \theta_2) \leq p^2 \frac{\varphi_{\max}[I_{p^2} - C(\theta)]}{\sigma^2} l(\theta_1, \theta_2) \leq \frac{\rho_2 p^2}{\sigma^2} l(\theta_1, \theta_2) ,$$

if $\theta \in \mathcal{B}_1(0_p, 1) \cap \mathcal{B}_{\mathrm{op}}(\rho_2)$. We conclude by applying Proposition 6.20 and Corollary 6.21. □

Again, $\widetilde{\theta}_{\rho_1}$ satisfies the same minimax properties with respect to the Frobenius norm. All these properties easily extend to isotropic fields if one defines the corresponding sets $\mathcal{E}^{\mathrm{iso}}(a) \cap \mathcal{B}_1(0_p, 1) \cap \mathcal{U}(\rho_2)$ of isotropic GMRFs.

6.7 Discussion

6.7.1 Comparison with maximum likelihood estimation

Let us first compare the computational cost the CLS estimation method and the maximum likelihood estimator (MLE). For toroidal lattices, fast algorithms based on two-dimensional Fast-Fourier transformation (see for instance [RT02]) allow to compute the MLE as fast as the CLS estimator. More details on the computation of the CLS estimators for toroidal lattices are given in Section 7.2.3. When the lattice is not a torus, the MLE becomes intractable because it involves the optimization of a determinant of size p^2. In contrast, the CLS criterion $\gamma_{n,p}(.)$ defined in (6.16) is a quadratic function of θ. Consequently, CLS estimators are still computationally amenable. Since we extend our model selection to non-toroidal lattices in Chapter 7, it is preferable to use CLS estimators.

Let us compare the risk of CLS estimators and MLE. Given a small dimensional model m, the risk of the *parametric* CLS estimator and the *parametric* MLE have been compared from an asymptotic point of view ([Guy95] Sect.4.3). It is generally accepted (see for instance Cressie [Cre93] Sect. 7.3.1) and that *parametric* CLS estimators are almost as efficient as parametric MLE for the major part of the parameter spaces Θ_m^+. We have non-asymptotically assessed this statement in Proposition 6.16 by minimax arguments. Nevertheless, for some parameters θ that are close to the border of Θ_m^+, Kashyap and Chellappa [KC83] have pointed out that CLS estimators are less efficient than MLE. If we have proved nonasymptotic bounds for CLS based model selection method, we are not aware of any result for model selection procedures based on MLE.

6.7.2 Concluding remarks

We have developed a model selection procedure for choosing the neighborhood of a GMRF. In Theorem 6.4, we have proven a nonasymptotic upper bound for the risk of the estimator $\widetilde{\theta}_{\rho_1}$ with respect to the prediction error $l(.,.)$. Under Assumption $(\mathbb{H}_1)$, this bound is shown to be optimal from an asymptotic point of view if the support of θ belongs to one of the models in the collection. If Assumption $(\mathbb{H}_2)$ is fulfilled, we are able to obtain an oracle type inequality for $\widetilde{\theta}_{\rho_1}$. Moreover, $\widetilde{\theta}_{\rho_1}$ is minimax adaptive to the sparsity of θ under $(\mathbb{H}_1)$. Finally, it simultaneously achieves the minimax rates of estimation over a large class of sets $\mathcal{E}(a)$ if $(\mathbb{H}_2)$ holds. Some of these properties still hold if we use the Frobenius loss function. The case of isotropic Gaussian fields is handled similarly.

However, in the oracle inequality (6.32) and in the minimax bounds (6.37) and (6.42), we either perform an assumption on the l_1 norm of θ or on the smallest eigenvalue of $(I_{p^2} - C(\theta))$. When $\|\theta\|_1$ tends to one or $\varphi_{\min}[I_{p^2} - C(\theta)]$ tends to 0, there is a distortion between the upper bound $\mathbb{E}_\theta[l(\widetilde{\theta}_{\rho_1}, \theta)]$ provided by Theorem 6.4 and the lower bounds given by Proposition 6.11 or Proposition 6.16. This limitation seems intrinsic to our penalization method which is linear with respect to the dimension, whereas the asymptotic variance term $\mathbb{E}_\theta[l(\widehat{\theta}_{m,\rho_1}, \theta)]$ depends in a complex way on the dimension of the model m and on the target θ. In our opinion, achieving adaptivity with respect to the smallest eigenvalue of $(I_{p^2} - C(\theta))$ (or equivalently the largest value of Σ) would require a different penalization technique. Nevertheless, we are not aware of any procedure in a covariance estimation setting that is adaptive to the largest eigenvalues of Σ.

So far, we have provided an estimation procedure for $(I_{p^2} - C(\theta)) = \sigma^2\Sigma^{-1}$. If we aim at estimating the precision matrix Σ^{-1}, we also have to take into account the quantity σ^2. It is natural to estimate it by $\widetilde{\sigma}^2 := \gamma_{n,p^2}(\widetilde{\theta}_{\rho_1})$ as done for instance by Guyon in [Guy95] Sect.4.3 in the parametric setting. Then, we obtain the estimate $\widetilde{\Sigma^{-1}} := \widetilde{\sigma}^2(I_{p^2} - C(\widetilde{\theta}_{\rho_1})$. It is of interest to study the adaptive properties of this estimator with respect to loss functions such as the Frobenius or operator norm as done in [RBLZ08] in the non-stationary setting. Nevertheless, let us mention that the matrix $\widetilde{\Sigma^{-1}}$ is not necessarily invertible since the estimator $\widetilde{\theta}_{\rho_1}$ belongs to the closure of Θ^+.

The choice of the quantity ρ_1 is problematic. On the one hand, ρ_1 should be large enough so that Assumption $(\mathbb{H}_1)$ is fulfilled. On the other hand, a large value of ρ_1 yields worse bounds in Theorem 6.4. Moreover, the largest eigenvalue of $(I_{p^2} - C(\theta))$ is unknown in practice, which makes more difficult the choice of ρ_1. We see two possible answers to this issue:

- First, moderate values of ρ_1 are sufficient to enforce $(\mathbb{H}_1)$ if the target θ is sparse as illustrated in Table 8.

- Second, we believe that the bounds for the risk are pessimistic with respect to ρ_1. A future direction of research is to derive risk bounds for $\widetilde{\theta}_{\rho_1}$ with $\rho_1 = +\infty$. In Chapter 7, we illustrate that such a procedure gives rather good results in practice.

In Theorem 6.4, we only provide a lower bound of the penalty so that the procedure performs well. However, this bound depends on the largest eigenvalue of Σ which is seldom known in practice and we did not give any advice for choosing a "reasonable" constant K in practice. This is why we define in Chapter 7 a data-driven method based on the *slope heuristics* of Birgé and Massart [BM07] for calibrating the penalty. We also provide numerical evidence of its performances on simulated data.

We mentioned in the introduction that the toroidal assumption for the lattice is somewhat artificial in several applications. Nevertheless, we needed to neglect the edge effects in order to derive non asymptotic properties for $\widetilde{\theta}_{\rho_1}$ as in Theorem 6.4. In practice, it is often more realistic to suppose that we observe a small window of a Gaussian field defined on the whole plane $\mathbb{Z}^2$. The previous nonasymptotic properties do not extend to this new setting. Nevertheless, Lakshman and Derin have shown in [LD93] that there is no phase transition within the valid parameter space for GMRFs defined on the plane $\mathbb{Z}^2$. In short, this implies that the distribution of a field observed in a fixed window of a GMRF does not asymptotically depend on the bound condition. Therefore, it is reasonable to think that our estimation procedure performs well if it was adapted to this new setting. In Chapter 7, we describe this extension and we provide numerical evidence of its performances.

6.8 Proofs

6.8.1 A concentration inequality

In this section, we prove a new concentration inequality for suprema of Gaussian chaos of order 2. It will be useful for proving Theorem 6.4.

Proposition 6.23. *Let F be a compact set of symmetric matrices of size r, $(Y^1, \dots, Y^n)$ be a n-sample of a standard Gaussian vector of size r, and Z be the random variable defined by*

$$Z := \sup_{R \in F} tr\left[R(\overline{YY^*} - I_r)\right] \ .$$

Then

$$\mathbb{P}(Z \geq \mathbb{E}(Z) + t) \leq \exp\left[-\left(\frac{t^2}{L_1 \mathbb{E}(W)} \bigwedge \frac{t}{L_2 B}\right)\right], \tag{6.43}$$

where the quantities B and W are such that

$$\begin{aligned} B &:= \frac{2}{n} \sup_{R \in F} \varphi_{max}(R) \\ W &:= \frac{4}{n} \sup_{R \in F} tr(R\overline{YY^*}R') \ . \end{aligned}$$

Proof of Proposition 6.23. The main argument of this proof is to transfer a deviation inequality for suprema of Rademacher chaos of order 2 to suprema of Gaussian Chaos. Talagrand [Tal96] has first given in Theorem 1.2 a concentration inequality for such suprema of Rademacher chaos. Boucheron *et al.* [BBLM05] have recovered the upper bound applying a new methodology based on the entropy method. We shall adapt their proof to consider non-necessarily homogeneous chaos of order 2.

First, we recall the notations introduced in [BBLM05]. Let N be a positive integer. Then, $\mathcal{I}_N$ stands for the family of subsets of $\{1, \dots, N\}$ of size less than 2. Let $\mathcal{T}$ be a set of vectors indexed by $\mathcal{I}_N$. In the sequel, $\mathcal{T}$ is assumed to be a compact subset of $\mathbb{R}^{(N(N+1)/2)+1}$. The following lemma states a slightly modified version of the upper bound in remark 7 in [BBLM05].

Lemma 6.24. *Let T be a suprema of Rademacher chaos indexed by $\mathcal{I}_N$ of the form*

$$T := \sup_{t \in \mathcal{T}} \left| \sum_{\{i,j\}} U_i U_j t_{\{i,j\}} + \sum_{i=1}^{N} t_{\{i\}} + t_{\varnothing} \right| \ ,$$

where $U_1, \dots, U_N$ are independent Rademacher random variables. Then for any $x > 0$,

$$\mathbb{P}\left\{T \geq \mathbb{E}[T] + x\right\} \leq 4 \exp\left(-\frac{x^2}{L_1 \mathbb{E}[D]^2} \wedge \frac{x}{L_2 E}\right) \ , \tag{6.44}$$

where D and E are defined by:

$$\begin{aligned} D &:= \sup_{t \in \mathcal{T}} \sup_{\alpha : \|\alpha\|_2 \leq 1} \left| \sum_{i=1}^{N} U_i \sum_{j \neq i} \alpha_j t_{\{i,j\}} \right| \ , \\ E &:= \sup_{t \in \mathcal{T}} \sup_{\alpha^{(1)}, \alpha^{(2)}, \|\alpha^{(1)}\|_2 \leq 1 \ \|\alpha^{(2)}\| \leq 1} \left| \sum_{i=1}^{N} \sum_{j \neq i} t_{\{i,j\}} \alpha_i^{(1)} \alpha_j^{(2)} \right| \ . \end{aligned}$$

Contrary to the original result of [BBLM05], the chaos are not assumed to be homogeneous. Besides, the $t_{\{i\}}$ are redundant with $t_{\varnothing}$. In fact, we introduced this family in order to emphasize the connection with Gaussian chaos in the next result.

A suitable application of the central limit theorem enables to obtain a corresponding bound for Gaussian chaos of order 2.

Lemma 6.25. *Let T be a supremum of Gaussian chaos of order 2.*

$$T := \sup_{t \in \mathcal{T}} \left| \sum_{\{i,j\}} t_{\{i,j\}} Y_i Y_j + \sum_i t_i Y_i^2 + t_{\varnothing} \right| \ , \tag{6.45}$$

where $Y_1, \dots, Y_N$ are independent standard Gaussian random variable. Then, for any $x > 0$,

$$\mathbb{P}\left\{T \geq \mathbb{E}[T] + x\right\} \leq \exp\left(-\frac{x^2}{\mathbb{E}[D]^2 L_1} \wedge \frac{x}{E L_2}\right) \ , \tag{6.46}$$

where

$$
\begin{aligned}
D &:= \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^N \|\alpha\|_2\leq 1} \sum_{i,j} Y_i(1+\delta_{i,j})\alpha_j t_{\{i,j\}} \ , \\
E &:= \sup_{t\in\mathcal{T}} \sup_{\alpha_1,\ \|\alpha_1\|_2\leq 1} \sup_{\alpha_2,\ \|\alpha_2\|_2\leq 1} \sum_{i,j} \alpha_{1,i}\alpha_{2,j} t_{\{i,j\}}(1+\delta_{i,j}) \ .
\end{aligned}
$$

The proof of this Lemma is postponed to the end of this section. To conclude, we derive the result of Proposition 6.23 from this last lemma. For any matrix $R \in F$, we define the vector $t^R \in \mathbb{R}^{nr(nr+1)/2+1}$ indexed by $\mathcal{I}_{nr}$ as follows

$$
t^R_{\{(i,k),(j,l)\}} := \delta_{k,l}(2-\delta_{i,j})\frac{R_{[i,j]}}{n}, \qquad t^R_{\{(i,k)\}} := \frac{R_{[i,i]}}{n}, \quad \text{and } t^R_{\varnothing} := -tr(R) \ ,
$$

where $\delta_{i,j}$ is the indicator function of $i=j$. In order to apply Lemma 6.25 with $N = nr$ and $\mathcal{T} = \left\{t^R | R \in F\right\}$, we have to work out the quantities D and E.

$$
\begin{aligned}
D &= \sup_{t^R\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^{nr},\ \|\alpha\|_2\leq 1} \left\{ \sum_{i=1}^r\sum_{k=1}^n Y_{[i,k]} \sum_{j=1}^r\sum_{l=1}^n t^{R,k,l}_{ij}(1+\delta_{i,j}\delta_{k,l})\alpha^l_j \right\} \\
&= \sup_{R\in F} \sup_{\alpha\in\mathbb{R}^{nr},\ \|\alpha\|_2\leq 1} 2\left\{ \sum_{i=1}^r\sum_{k=1}^n Y_{[i,k]} \sum_{j=1}^r \frac{R_{[i,j]}\alpha^k_j}{n} \right\} \\
&= \sup_{R\in F} \sup_{\alpha\in\mathbb{R}^{nr},\ \|\alpha\|_2\leq 1} \frac{2}{n}\left\{ \sum_{k=1}^n\sum_{j=1}^r \alpha^k_j \left(\sum_{i=1}^r Y_{[i,k]}R_{[i,j]} \right) \right\} \ .
\end{aligned}
$$

Applying Cauchy-Schwarz identity yields

$$
\begin{aligned}
D^2 &= \frac{4}{n^2}\sup_{R\in F}\left\{ \sum_{k=1}^n\sum_{j=1}^r \left(\sum_{i=1}^r Y_{[i,k]}R_{[i,j]} \right)^2 \right\} \\
&= \frac{4}{n}\sup_{R\in F} tr(R\overline{YY^*}R^*) \ . \qquad (6.47)
\end{aligned}
$$

Let us now turn the constant E

$$
\begin{aligned}
E &= \sup_{t^R\in\mathcal{T}} \sup_{\substack{\alpha_1,\alpha_2\in\mathbb{R}^{nr} \\ \|\alpha_1\|_2\leq 1, \|\alpha_2\|_2\leq 1}} \sum_{1\leq i,j\leq r}\sum_{1\leq k,l\leq n} (1+\delta_{ij}\delta_{k,l}) t^{R,kl}_{i,j}\alpha^k_{1,i}\alpha^l_{2,j} \\
&= \sup_{R\in F} \sup_{\substack{\alpha_1,\alpha_2\in\mathbb{R}^{nr} \\ \|\alpha_1\|_2\leq 1, \|\alpha_2\|_2\leq 1}} \frac{2}{n}\sum_{1\leq i,j\leq r}\sum_{1\leq k\leq n} R_{[i,j]}\alpha^k_{1,i}\alpha^k_{2,j} \ .
\end{aligned}
$$

From this last expression, it follows that E is a supremum of L_2 operator norms

$$
E = \frac{2}{n}\sup_{R\in F}\varphi_{\max}\left(Diag^{(n)}(R)\right) \ ,
$$

where $Diag^{(n)}(R)$ is the $(nr\times nr)$ block diagonal matrix such that each diagonal block is made of the matrix R. Since the largest eigenvalue of $Diag^{(n)}(R)$ is exactly the largest eigenvalue of R, we get

$$
E = \frac{2}{n}\sup_{R\in F}\varphi_{\max}(R) \ . \qquad (6.48)
$$

Applying Proposition 6.25 and gathering identities (6.47) and (6.48) yields

$$
\mathbb{P}(Z \geq \mathbb{E}(Z)+t) \leq \exp\left[-\left(\frac{t^2}{L_1\mathbb{E}(V)} \bigwedge \frac{t}{L_2 B} \right) \right] \ ,
$$

where $B = E$ and $V = D^2$. □

Proof of Lemma 6.24. This result is an extension of Corollary 4 in [BBLM05]. We shall closely follow the sketch of their proof adapting a few arguments. First, we upper bound the moments of $(T - \mathbb{E}(T))_+$. Then, we derive the deviation inequality from it. Here, $x_+ = \max(x, 0)$.

Lemma 6.26. *For all real numbers $q \geq 2$,*

$$\|(T - \mathbb{E}(T))_+\|_q \leq \sqrt{Lq}\mathbb{E}(D) + LqE \ , \tag{6.49}$$

where $\|T\|_q^q$ stands for the q-th moment of the random variable T. The quantities D and E are defined in Lemma 6.24.

By Lemma 6.26, for any $t \geq 0$ and any $q \geq 2$,

$$\begin{aligned} \mathbb{P}\left(T \geq \mathbb{E}(T) + t\right) &\leq \frac{\mathbb{E}\left[(T - \mathbb{E}(T))_+^q\right]}{t^q} \\ &\leq \left(\frac{\sqrt{Lq}\mathbb{E}(D) + LqE}{t}\right)^q \ . \end{aligned}$$

The right-hand side is at most 2^{-q} if $\sqrt{Lq}\mathbb{E}(D) \leq t/4$ and $LqE \leq t/4$. Let us set

$$q_0 := \frac{t^2}{16L\mathbb{E}(D)^2} \wedge \frac{t}{4LE} \ .$$

If $q_0 \geq 2$, then $\mathbb{P}\left(T \geq \mathbb{E}(T) + t\right) \leq 2^{-q_0}$. On the other hand if $q_0 < 2$, then $4 \times 2^{-q_0} \geq 1$. It follows that

$$\mathbb{P}\left(T \geq \mathbb{E}(T) + t\right) \leq 4\exp\left(-\frac{\log(2)}{4L}\left[\frac{t^2}{4\mathbb{E}(D)^2} \wedge \frac{t}{E}\right]\right) \ .$$

□

Proof of Lemma 6.26. This result is based on the entropy method developed in [BBLM05]. Let $f : \mathbb{R}^N \to \mathbb{R}$ be a measurable function such that $T = f(U_1, \ldots, U_N)$. In the sequel, $U_1', \ldots, U_N'$ denote independent copies of $U_1, \ldots, U_N$. The random variable T_i' and V^+ are defined by

$$\begin{aligned} T_i' &:= f(U_1, \ldots, U_{i-1}, U_i', U_{i+1}, \ldots, U_N) \ , \\ V^+ &:= \mathbb{E}\left[\sum_{i=1}^N (T - T_i')_+^2 | U_1^N\right] \ , \end{aligned}$$

where U_1^N refers to the set $\{U_1, \ldots, U_N\}$. Theorem 2 in [BBLM05] states that for any real $q \geq 2$,

$$\|(T - \mathbb{E}(T))_+\|_q \leq \sqrt{Lq}\|\sqrt{V^+}\|_q \ . \tag{6.50}$$

To conclude, we only have bound the moments of $\sqrt{V^+}$. By definition,

$$T = \sup_{t \in \mathcal{T}} \left|\sum_{\{i,j\}} U_i U_j t_{\{i,j\}} + \sum_{i=1}^N t_{\{i\}} + t_\varnothing\right| \ .$$

Since the set $\mathcal{T}$ is compact, this supremum is achieved almost surely at an element t^0 of $\mathcal{T}$. For any $1 \leq i \leq N$,

$$\begin{aligned} (T - T_i')_+^2 &\leq \left(\left|\sum_{\{k,l\}} U_k U_l t^0_{\{k,l\}} + \sum_{k=1}^N t^0_{\{k\}} + t^0_\varnothing\right| - \left|\sum_{\{k,l\}, k\neq i, l\neq i} U_i U_j t^0_{\{k,l\}} + \sum_{k\neq i} U_i' U_k t^0_{\{k,i\}} + \sum_{k=1}^N t^0_{\{k\}} + t^0_\varnothing\right|\right)_+^2 \\ &\leq \left((U_i - U_i')\left|\sum_{j\neq i} U_j t^0\{i,j\}\right|\right)^2 \ . \end{aligned}$$

Gathering this bound for any i between 1 and N, we get

$$\begin{aligned} V^+ &\leq \sum_{i=1}^N \mathbb{E}\left[\left((U_i - U_i')\left|\sum_{j\neq i} U_j t^0\{i,j\}\right|\right)^2 \middle| U_1^N\right] \\ &\leq 2\sum_{i=1}^N \left[\sum_{j\neq i} U_j t^0\{i,j\}\right]^2 \\ &\leq 2 \sup_{\alpha\in\mathbb{R}^N,\ \|\alpha\|_2\leq 1} \left[\sum_{i=1}^N \alpha_i \left(\sum_{j\neq i} t^0_{\{i,j\}} U_j\right)\right]^2 \\ &\leq 2 \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^N,\ \|\alpha\|_2\leq 1} \sum_{i=1}^N \left[U_i \sum_{j\neq i} \alpha_j t_{\{i,j\}}\right]^2 = 2D^2 \ . \end{aligned}$$

Combining this last bound with (6.50) yields

$$\begin{aligned} \|(T - \mathbb{E}(T))_+\|_q &\leq \sqrt{Lq}\sqrt{2}\|D\|_q \\ &\leq \sqrt{Lq}\left[\mathbb{E}(D) + \|(D - \mathbb{E}(D))_+\|_q\right] \ . \end{aligned} \tag{6.51}$$

Since the random variable D defined in Lemma 6.24 is a measurable function f_2 of the variables $U_1, \ldots, U_N$, we apply again Theorem 2 in [BBLM05].

$$\|(D - \mathbb{E}(D))_+\|_q \leq \sqrt{Lq}\|\sqrt{V_2^+}\|_q,$$

where V_2^+ is defined by

$$V_2^+ := \mathbb{E}\left[\sum_{i=1}^N (D - D_i')_+^2 \middle| U_i^N\right],$$

and $D_i' := f_2(U_1, \ldots, U_{i-1}, U_i', U_{i+1}, \ldots, U_N)$. As previously, the supremum in D is achieved at some random parameter (t^0, α^0). We therefore upper bound V_2^+ as previously.

$$\begin{aligned} V_2^+ &\leq \sum_{i=1}^N \mathbb{E}\left[\left((U_i - U_i')\left(\sum_{j\neq i} \alpha_j^0 t^0_{\{i,j\}}\right)\right)^2 \middle| U_1^N\right] \\ &\leq 2\sum_{i=1}^N \left(\sum_{j\neq i} \alpha_j^0 t^0_{\{i,j\}}\right)^2 \\ &\leq 2 \sup_{\alpha^{(2)}\in\mathbb{R}^N, \|\alpha\|_2\leq 1} \left(\sum_{i=1}^N \alpha_j^{(2)} \sum_{j\neq i} \alpha_i^0 t_{\{i,j\}}\right)^2 = 2E^2 \ . \end{aligned}$$

Gathering this upper bound with (6.51) yields

$$\|(T - \mathbb{E}(T))_+\|_q \leq \sqrt{Lq}\mathbb{E}(D) + LqE \ .$$

□

Proof of Lemma 6.25. We shall apply the central limit theorem in order to transfer results for Rademacher chaos to Gaussian chaos. Let f be the unique function satisfying $T = f(y_1, \ldots, y_N)$ for any $(y_1, \ldots, y_N) \in \mathbb{R}^N$. As the set $\mathcal{T}$ is compact, the function f is known to be continuous. Let $(U_i^{(j)})_{1\leq i\leq N, j\geq 0}$ an i.i.d. family of Rademacher variables. For any integer $n > 0$, the random variables $Y^{(n)}$ and $T^{(n)}$ are defined by

$$\begin{aligned} Y^{(n)} &:= \left(\sum_{j=1}^n \frac{U_1^{(j)}}{\sqrt{n}}, \ldots, \sum_{j=1}^n \frac{U_N^{(j)}}{\sqrt{n}}\right) \ , \\ T^{(n)} &:= f\left(Y^{(n)}\right) \ . \end{aligned}$$

Clearly, $T^{(n)}$ is a supremum of Rademacher chaos of order 2 with nN variables and a constant term. By the central limit theorem, $T^{(n)}$ converges in distribution towards T as n tends to infinity. Consequently, deviation inequalities for the variables $T^{(n)}$ transfer to T as long as the quantities $\mathbb{E}\left[D^{(n)}\right]$, $E^{(n)}$, and $\mathbb{E}(T^{(n)})$ converge.

We first prove that the sequence $T^{(n)}$ converges in expectation towards T. As $T^{(n)}$ converges in distribution, it is sufficient to show that the sequence $T^{(n)}$ is asymptotically uniformly integrable. The set $\mathcal{T}$ is compact, thus there exists a positive number t_∞ such that

$$\begin{aligned} T^{(n)} &\leq t_\infty \left[\sum_{i,j} |Y_i^{(n)} Y_j^{(n)}| + 1 \right] \\ &\leq t_\infty \left[1 + (N+1)/2 \sum_{i=1}^{N} \left(Y_i^{(n)}\right)^2 \right] . \end{aligned}$$

It follows that

$$\left(T^{(n)}\right)^2 \leq t_\infty^2 \left(\frac{N+1}{2}\right)^2 \frac{N+2}{2} \left[1 + \sum_{i=1}^{N} \left(Y_i^{(n)}\right)^4 \right] . \tag{6.52}$$

The sequence $Y_i^{(n)}$ does not only converge in distribution to a standard normal distribution but also in moments (see for instance [Bil95] p.391). It follows that $\overline{\lim}\mathbb{E}\left[\left(T^{(n)}\right)^2\right] \leq \infty$ and the sequence $f\left(Y^{(n)}\right)$ is asymptotically uniformly integrable. As a consequence,

$$\lim_{n\to\infty} \mathbb{E}\left[T^{(n)}\right] = \mathbb{E}[T] .$$

Let us turn to the limit of $\mathbb{E}\left[D^{(n)}\right]$. As the variable $T^{(n)}$ equals

$$T^{(n)} = \sup_{t\in\mathcal{T}} \left| \sum_{\{i,j\}} t_{\{i,j\}} \sum_{1\leq k,l\leq n} \frac{U_i^{(k)} U_j^{(l)}}{n} + \sum_i t_i \sum_{1\leq k\leq n} \frac{U_i^{(k)}}{\sqrt{n}} \sum_{l\neq k} \frac{U_i^{(l)}}{\sqrt{n}} + t_\varnothing + \sum_i t_i \right| ,$$

it follows that

$$\begin{aligned} D^{(n)} &= \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^{nN},\ \|\alpha\|_2\leq 1} \left| \sum_{1\leq i\leq N} \sum_{1\leq k\leq n} U_i^{(k)} \left\{ \sum_{j\neq i} \frac{t_{\{i,j\}}}{n} \sum_{1\leq l\leq n} \alpha_j^{(l)} + 2 \sum_{l\neq k} 2\frac{t_{\{i\}}}{n} \alpha_i^{(l)} \right\} \right| \\ &\leq \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^{nN},\ \|\alpha\|_2\leq 1} \left\{ \sum_i \frac{U_i^{(k)}}{\sqrt{n}} \sum_j (1+\delta_{i,j}) t_{\{i,j\}} \frac{\sum_{1\leq l\leq n} \alpha_j^{(l)}}{\sqrt{n}} \right\} + A^{(n)} , \end{aligned} \tag{6.53}$$

where the random variable $A^{(n)}$ is defined by

$$A^{(n)} := \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^{nN},\ \|\alpha\|_2\leq 1} \sum_{i=1}^{N} \sum_{j=1}^{n} t_{\{i\}} \frac{U_i^{(j)}}{n} \alpha_i^j .$$

Straightforwardly, one upper bounds $A^{(n)}$ by $\frac{t_\infty}{n}\sqrt{\sum_{i=1}^N \sum_{j=1}^n \left(U_i^{(j)}\right)^2}$ and its expectation satisfies

$$\mathbb{E}\left(\left|A^{(n)}\right|\right) \leq t_\infty \sqrt{\frac{N}{n}} ,$$

which goes to 0 when n goes to infinity. Thus, we only have to upper bound the expectation of the first term in (6.53). Clearly, the supremum is achieved only when for all $1 \leq j \leq N$, the sequence $(\alpha_j^{(l)})_{1\leq l\leq n}$ is constant. In such a case, the sequence $(\alpha_j^{(1)})_{1\leq j\leq N}$ satisfies $\|\alpha^{(1)}\|_2 \leq 1/\sqrt{n}$. it follows that

$$\mathbb{E}\left[D^{(n)}\right] = \mathbb{E}\left\{ \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^N \|\alpha\|_2\leq 1} \mathbb{E}\left[\sum_i Y_i^{(n)} \sum_j (1+\delta_{i,j})\alpha_j \right] \right\} + \mathcal{O}(\frac{1}{\sqrt{n}}) .$$

Let g be the function defined by $g(y_1,\dots,y_N) = \sup_{t\in\mathcal{T}} \sup_{\alpha\in\mathbb{R}^N \|\alpha\|_2\leq 1} \left[\sum_i y_i \sum_j (1+\delta_{i,j})\alpha_j\right]$, for any $(y_1,\dots,y_N)\in\mathbb{R}^N$. The function $g(.)$ is measurable and continuous as the supremum is taken over a compact set. As a consequence, $g(Y^{(n)})$ converges in distribution towards $g(Y)$. As previously, the sequence is asymptotically uniformly integrable since its moment of order 2 is uniformly upper bounded. It follows that $\lim \mathbb{E}\left[D^{(n)}\right] = \mathbb{E}[D]$.

Third, we compute the limit of $E^{(n)}$. By definition,

$$\begin{aligned} E^{(n)} &= \sup_{t\in\mathcal{T}} \sup_{\alpha_1,\alpha_2\in\mathbb{R}^{nN},\ \|\alpha_1\|_2\leq 1,\|\alpha_2\|_2\leq 1} \sum_{i=1}^N \sum_{k=1}^n \alpha_{1,i}^k \left[\sum_{j\neq i}\sum_{l=1}^n \alpha_{2,j}^{(l)} \frac{t_{\{i,j\}}}{n} + 2\sum_{l\neq k} \alpha_{2,i}^{(l)} \frac{t_{\{i\}}}{n}\right] \\ &= \sup_{t\in\mathcal{T}} \sup_{\alpha_1,\alpha_2,\ \|\alpha_1\|_2\leq 1,\|\alpha_2\|_2\leq 1} \sum_{i=1}^N \sum_{j=1}^N (1+\delta_{i,j}) \frac{t_{\{i,j\}}}{n} \left[\sum_{k=1}^n \sum_{l=1}^n \alpha_{1,i}^{(k)} \alpha_{2,j}^{(l)}\right] + \mathcal{O}(\frac{1}{n}) \ . \end{aligned}$$

As for the computation of $D^{(n)}$, the supremum is achieved when the sequences $(\alpha_{1,i}^k)_{1\leq k\leq n}$ and $(\alpha_{2,j}^l)_{1\leq l\leq n}$ are constant for any $i\in\{1,\dots,N\}$. Thus, we only have to consider the supremum over the vectors α_1 and α_2 in $\mathbb{R}^N$.

$$E^{(n)} = \sup_{t\in\mathcal{T}} \sup_{\alpha_1,\alpha_2\in\mathbb{R}^N \|\alpha_i\|_2\leq 1} \sum_{i=1}^N \sum_{j=1}^N (1+\delta_{ij}) t_{i,j} \alpha_{1,i}\alpha_{2,j} + \mathcal{O}\left(\frac{1}{n}\right).$$

It follows that $E^{(n)}$ converges towards E when n tends to infinity.

The random variable $T^{(n)} - \mathbb{E}(T^{(n)})$ converges in distribution towards $T - \mathbb{E}(T)$. By Lemma 6.24 ,

$$\mathbb{P}(T - \mathbb{E}(T) \geq x) \leq \underline{\lim}\ \exp\left(-\frac{x^2}{\mathbb{E}[D^{(n)}]^2 L_1} \wedge \frac{x}{E^{(n)} L_2}\right) \ ,$$

for any $x>0$. Combining this upper bound with the convergence of the sequences $D^{(n)}$ and $E^{(n)}$ allows to conclude. □

6.8.2 Proof of Theorem 6.4

Proof of Theorem 6.4. We only consider the case of anisotropic estimators. The proofs and lemma are analogous for isotropic estimators. We first fix a model $m\in\mathcal{M}$. By definition, the model $\widehat{m}$ satisfies

$$\gamma_{n,p}(\widetilde{\theta}_{\rho_1}) + \mathrm{pen}(\widehat{m}) \leq \gamma_{n,p}(\theta_{m,\rho_1}) + \mathrm{pen}(m) \ .$$

For any $\theta'\in\Theta^+$, $\overline{\gamma}_{n,p}(\theta')$ stands for the difference between $\gamma_{n,p}(\theta')$ and its expectation $\gamma(\theta')$. Then, the previous inequality turns into

$$\gamma(\widetilde{\theta}_{\rho_1}) \leq \gamma(\theta_{m,\rho_1}) + \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) + \mathrm{pen}(m) - \mathrm{pen}(\widehat{m}) \ .$$

Subtracting the quantity $\gamma(\theta)$ to both sides of this inequality yields

$$l(\widetilde{\theta}_{\rho_1},\theta) \leq l(\theta_{m,\rho_1},\theta) + \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) + \mathrm{pen}(m) - \mathrm{pen}(\widehat{m}) \ . \tag{6.54}$$

The proof is based on the control of the random variable $\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1})$.

Lemma 6.27. *For any positive number α, ξ, and $\delta>1$ the event Ω_ξ defined by*

$$\Omega_\xi = \left\{ \begin{array}{r} \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}\left(\widetilde{\theta}_{\rho_1}\right) \leq \frac{1}{\sqrt{\delta}} l\left(\widetilde{\theta}_{\rho_1},\theta\right) + \frac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1},\theta) \\ + \frac{K_0\delta^2\rho_1^2\varphi_{max}(\Sigma)}{np^2}\left[(1+\alpha/2)(d_m + d_{\widehat{m}}) + \frac{\xi^2}{\delta-1}\right] \end{array} \right\} \ ,$$

satisfies

$$\mathbb{P}(\Omega_\xi^c) \leq \exp\left\{-L_1\xi\left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n}\right]\right\} \sum_{m'\in\mathcal{M}} \exp\left\{-L_2\sqrt{d_{m'}}\left(\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \frac{\alpha^2}{1+\alpha/2}\right)\right\} \ .$$

A similar lemma holds in the isotropic case. In particular, we choose $\alpha = \frac{K-K_0}{K_0}$ and $\delta = \sqrt{\frac{1+\alpha}{1+\alpha/2}}$. Lemma 6.27 implies that on the event Ω_ξ,

$$\begin{aligned}\overline{\gamma}_{n,p}\left(\theta_{m,\rho_1}\right) - \overline{\gamma}_{n,p}\left(\widetilde{\theta}_{\rho_1}\right) &\leq \frac{1}{\sqrt{\delta(\alpha)}} l\left(\widetilde{\theta}_{\rho_1}, \theta\right) + \frac{\sqrt{\delta(\alpha)}}{\sqrt{\delta(\alpha)}-1} l\left(\theta_{m,\rho_1}, \theta\right) + \mathrm{pen}(m) \\ &+ \mathrm{pen}(\widehat{m}) + \frac{K_0 \xi^2 \delta(\alpha)^2 \rho_1^2 \varphi_{\max}(\Sigma)}{np^2\left(\delta(\alpha)-1\right)} .\end{aligned}$$

Thus, gathering this bound with inequality (6.54) yields

$$\frac{\delta(\alpha)^{1/2}-1}{\delta(\alpha)^{1/2}} l(\widetilde{\theta}_{\rho_1}, \theta) \leq \left[1 + \delta(\alpha)^{-1/2}(\delta(\alpha)^{1/2}-1)^{-1}\right] l(\theta_{m,\rho_1}, \theta) + 2\mathrm{pen}(m) + \frac{K_0 \xi^2 \rho_1^2 \varphi_{\max}(\Sigma)\delta(\alpha)^2}{np^2(\delta(\alpha)-1)} ,$$

with probability larger than $1 - \mathbb{P}\left(\Omega_\xi\right)$. Integrating this inequality with respect to $\xi > 0$ leads to

$$\begin{aligned}\frac{\delta(\alpha)^{1/2}-1}{\delta(\alpha)^{1/2}} \mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right] \leq \left[1 + \delta(\alpha)^{-1/2}\left(\delta(\alpha)^{1/2}-1\right)^{-1}\right] l\left(\theta_{m,\rho_1}, \theta\right) + \\ 2\mathrm{pen}(m) + \frac{\delta(\alpha)^2 L(\alpha)}{\left(\delta(\alpha)-1\right)\left[\frac{\alpha^2}{1+\alpha/2} \wedge n\right]} \frac{\rho_1^2 \varphi_{\max}(\Sigma)}{np^2} . \qquad (6.55)\end{aligned}$$

We upper bound $[(\alpha^2/(1+\alpha/2)) \wedge n]^{-1}$ by.$[(\alpha^2/(1+\alpha/2)) \wedge 1]^{-1}$. Since $\alpha = \frac{K-K_0}{K_0}$, it follows that

$$\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right] \leq L_1(K)\left[l\left(\theta_{m,\rho_1}, \theta\right) + \mathrm{pen}(m)\right] + L_2\left(K\right) \frac{\rho_1^2 \varphi_{\max}(\Sigma)}{np^2} ,$$

Taking the infimum over the models $m \in \mathcal{M}$ allows to conclude. □

Proof of Lemma 6.27. Throughout this proof, it is more convenient to express the quantities $\overline{\gamma}_{n,p}(.)$ and $l(.)$ in terms of covariance and precision matrices. Thanks to Equation (6.19), we also provide a matricial expression for $\gamma(.)$:

$$\gamma(\theta') = \frac{1}{p^2} tr\left[\left(I - C(\theta')\right)\Sigma\left(I - C(\theta')\right)\right] . \qquad (6.56)$$

Gathering identities (6.56) and (6.17), we get

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) = \frac{1}{p^2} tr\left[\left(\left[I_{p^2} - C(\theta_{m,\rho_1})\right]^2 - \left[I_{p^2} - C(\widetilde{\theta}_{\rho_1})\right]^2\right)\left(\overline{\mathbf{X}^{\mathbf{v}}\mathbf{X}^{\mathbf{v}*}} - \Sigma\right)\right] .$$

Since the matrices Σ, $(I_{p^2} - C(\theta_{m,\rho_1}))$, and $(I_{p^2} - C(\widetilde{\theta}_{\rho_1}))$ correspond to covariance or precision matrices of stationary fields on the two dimensional torus, they are symmetric block circulant. By Lemma 6.37, they are jointly diagonalizable in the same orthogonal basis. In the sequel, P stands for an orthogonal matrix associated to this basis. Then, the matrices $C(\theta_{m,\rho_1})$, $C(\widetilde{\theta}_{\rho_1})$, and Σ respectively decompose in

$$C(\theta_{m,\rho_1}) = P^* D(\theta_{m,\rho_1}) P,\ C(\widetilde{\theta}_{\rho_1}) = P^* D(\widetilde{\theta}_{\rho_1}) P,\ \Sigma = P^* D_\Sigma P,$$

where the matrices $D(\theta_{m,\rho_1})$, $D(\widetilde{\theta}_{\rho_1})$, and D_Σ are diagonal. Let the $p^2 \times n$ matrix $\mathbf{Y}$ be defined by $\mathbf{Y} := \sqrt{\Sigma^{-1}}\mathbf{X}^{\mathbf{v}}$. Clearly, the components of $\mathbf{Y}$ follow independent standard normal distributions. Gathering these new notations, we get

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) = \frac{1}{p^2} tr\left[\left(\left[I_{p^2} - D(\theta_{m,\rho_1})\right]^2 - \left[I_{p^2} - D(\widetilde{\theta}_{\rho_1})\right]^2\right) D_\Sigma \left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right] . \qquad (6.57)$$

Except $\overline{\mathbf{Y}\mathbf{Y}^*}$ all the matrices in this last expression are diagonal and we may therefore commute them in the trace.

Let $< ., . >_{\mathcal{H}}$ and $< ., . >_{\mathcal{H}'}$ be two inner products in the space of square matrices of size p^2 respectively defined by

$$< A, B >_{\mathcal{H}} := \frac{tr(A^* \Sigma B)}{p^2} \text{ and } < A, B >_{\mathcal{H}'} := \frac{tr(A^* D_\Sigma B)}{p^2} .$$

This first inner product is related to the loss function $l(.,.)$ through the identity

$$l\left(\theta',\theta\right) = \|C(\theta') - C(\theta)\|_{\mathcal{H}}^2 \ .$$

Besides, these two inner products clearly satisfy $\|C(\theta')\|_{\mathcal{H}} = \|D(\theta')\|_{\mathcal{H}'}$ for any $\theta' \in \Theta^+$. Gathering these new notations, we may upper bound (6.57) by

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) \quad - \quad \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \leq \|[I_{p^2} - D(\theta_{m,\rho_1})]^2 - [I_{p^2} - D(\widetilde{\theta}_{\rho_1})]^2\|_{\mathcal{H}'} \times \sup_{\substack{\theta_1 \in \Theta_m,\, \theta_2 \in \Theta_{\widehat{m}}, \\ \|[I_{p^2} - D(\theta_1)]^2 - [I_{p^2} - D(\theta_2)]^2\|_{\mathcal{H}'} \leq 1}} \left\langle \left[I_{p^2} - D(\theta_1)\right]^2 - \left[I_{p^2} - D(\theta_2)\right]^2, \left[\overline{\mathbf{YY}^*} - I_{p^2}\right] \right\rangle_{\mathcal{H}'} \quad (6.58)$$

The first term in this product is easily bounded as these matrices are diagonal.

$$\begin{aligned} \| \left[I_{p^2} - D(\theta_{m,\rho_1})\right]^2 - \left[I_{p^2} - D(\widetilde{\theta}_{\rho_1})\right]^2 \|_{\mathcal{H}'} &= tr\left[\left([I_{p^2} - D(\theta_{m,\rho_1})]^2 - [I_{p^2} - D(\widetilde{\theta}_{\rho_1})]^2\right)^2 \frac{D_\Sigma}{p^2}\right]^{\frac{1}{2}} \\ &= tr\left[\left[D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\right]^2 \frac{D_\Sigma}{p^2}\left[2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\right]^2\right]^{1/2} \\ &\leq \varphi_{\max}\left[2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\right] \|D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\|_{\mathcal{H}'} \ . \quad (6.59) \end{aligned}$$

Since θ_{m,ρ_1} and $\widetilde{\theta}_{\rho_1}$ respectively belong to Θ^+_{m,ρ_1} and $\Theta^+_{\widehat{m},\rho_1}$, the largest eigenvalues of the matrices $I_{p^2} - C(\theta_{m,\rho_1})$ and $I_{p^2} - C(\widetilde{\theta}_{\rho_1})$ are smaller than ρ_1. Hence, we get

$$\varphi_{\max}\left[2I_{p^2} - D(\theta_{m,\rho_1}) - D(\widetilde{\theta}_{\rho_1})\right] = \varphi_{\max}\left[I_{p^2} - C(\theta_{m,\rho_1})\right] + \varphi_{\max}\left[I_{p^2} - C(\widetilde{\theta}_{\rho_1})\right] \leq 2\rho_1 \ .$$

Let us turn to the second term in (6.58). First, we embed the set of matrices over which the supremum is taken in a ball of a vector space. For any model $m' \in \mathcal{M}$, let $U_{m'}$ be the space generated by the matrices $D(\theta')^2$ and $D(\theta')$ for $\theta' \in \Theta_{m'}$. In the sequel, we note $d_{m'^2}$ the dimension of $U_{m'}$. The space $U_{m,m'}$ is defined as the sum of U_m and $U_{m'}$ whereas d_{m^2,m'^2} stands for its dimension. Finally, we note $\mathcal{B}^{\mathcal{H}'}_{m^2,m'^2}$ the unit ball of $U_{m,m'}$ with respect to the inner product $<|>_{\mathcal{H}'}$. Gathering these notations, we get

$$\sup_{\substack{R = [I - D(\theta_1)]^2 - \left[I_{p^2} - D(\theta_2)\right]^2, \\ \theta_1 \in \Theta_m, \theta_2 \in \Theta_{\widehat{m}} \text{ and } \|R\|_{\mathcal{H}'} \leq 1}} \left\langle R, \overline{\mathbf{YY}^*} - I_{p^2}\right\rangle_{\mathcal{H}'} \leq \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,\widehat{m}^2}} \frac{1}{p^2} tr\left[RD_\Sigma\left(\overline{\mathbf{YY}^*} - I_{p^2}\right)\right] \ .$$

Applying the classical inequality $ab \leq \delta a^2 + \delta^{-1}b^2/4$ and gathering inequalities (6.58) and (6.59) yields

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \leq \delta^{-1}\|C(\theta_{m,\rho_1}) - C(\widetilde{\theta}_{\rho_1})\|^2_{\mathcal{H}} + \rho_1^2\delta \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,\widehat{m}^2}} \frac{1}{p^2} tr^2\left[RD_\Sigma\left(\overline{\mathbf{YY}^*} - I_{p^2}\right)\right] \ . \quad (6.60)$$

For any model $m' \in \mathcal{M}$, we define the random variable $Z_{m'}$ as

$$Z_{m'} := \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,m'^2}} \frac{1}{p^2} tr\left[RD_\Sigma\left(\overline{\mathbf{YY}^*} - I_{p^2}\right)\right] \ .$$

The variables $Z_{m'}$ turn out to be suprema of Gaussian chaos of order 2. In order to bound $Z_{\widehat{m}}$, we simultaneously control the deviations of $Z_{m'}$ for any model $m' \in \mathcal{M}$ thanks to the following lemma.

Lemma 6.28. *For any positive numbers α and ξ and any model $m' \in \mathcal{M}$,*

$$\mathbb{P}\left(Z_{m'} \geq \sqrt{\frac{2\varphi_{max}(\Sigma)}{n}}\left\{\sqrt{1+\alpha/2}\sqrt{d_{m^2,m'^2}} + \xi\right\}\right) \leq \exp\left\{-L_2\sqrt{d_{m'}}\left(\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \frac{\alpha^2}{1+\alpha/2}\right) - L_1\xi\left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n}\right]\right\} \ .$$

This result is a consequence from a general concentration inequality for suprema Gaussian chaos of order 2 stated in Proposition 6.23. Its proof is postponed to the end of the section. Let us fix the positive numbers α and ξ. Applying Lemma 6.28 to any model $m' \in \mathcal{M}$, the event Ω'_ξ defined by

$$\Omega'_\xi = \left\{ Z_{\widehat{m}} \leq \sqrt{\frac{2\varphi_{\max}(\Sigma)}{n}} \left[\sqrt{1+\alpha/2} \sqrt{d_{m^2,\widehat{m}^2}} + \xi \right] \right\}$$

satisfies

$$\mathbb{P}(\Omega'^c_\xi) \leq \exp\left\{ -L_1 \xi \left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n} \right] \right\} \sum_{m' \in \mathcal{M}} \exp\left\{ -L_2 \sqrt{d_{m'}} \left(\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \frac{\alpha^2}{1+\alpha/2} \right) \right\} .$$

From inequality (6.60), it follows that

$$\overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) \leq \delta^{-1} \|C(\theta_{m,\rho_1}) - C(\widetilde{\theta}_{\rho_1})\|^2_{\mathcal{H}} + \frac{2\delta\rho_1^2 \varphi_{\max}(\Sigma)}{np^2} \left\{ \sqrt{1+\alpha/2} \sqrt{d_{m^2,\widehat{m}^2}} + \xi \right\}^2 ,$$

conditionally to Ω'_ξ. By triangle inequality,

$$\|C(\theta_{m,\rho_1}) - C(\widetilde{\theta}_{\rho_1})\|_{\mathcal{H}} \leq \|C(\theta_{m,\rho_1}) - C(\theta)\|_{\mathcal{H}} + \|C(\widetilde{\theta}_{\rho_1}) - C(\theta)\|_{\mathcal{H}} .$$

We recall that the loss function $l(\theta', \theta)$ equals $\|C(\theta') - C(\theta)\|^2_{\mathcal{H}}$. We apply twice the inequality $(a+b)^2 \leq (1+\beta)a^2 + (1+\beta^{-1})b^2$. Setting the first β to $\sqrt{\delta} - 1$, it follows that

$$\begin{aligned} \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) &\leq \frac{1}{\sqrt{\delta}} l(\widetilde{\theta}_{\rho_1}, \theta) + \frac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1}, \theta) \\ &+ \frac{2\delta\rho_1^2 \varphi_{\max}(\Sigma)}{np^2} \left[d_{m^2,\widehat{m}^2}(1+\beta)(1+\alpha/2) + \xi^2(1+\beta^{-1}) \right] . \end{aligned}$$

By definition of $U_{m,\widehat{m}}$, its dimension $d_{m^2,\widehat{m}^2}$ is bounded by $d_{m^2} + d_{\widehat{m}^2}$. Choosing $\beta = \delta - 1$ yields

$$\begin{aligned} \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) &\leq \frac{1}{\sqrt{\delta}} l(\widetilde{\theta}_{\rho_1}, \theta) + \frac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1}, \theta) \\ &+ \frac{2\delta^2\rho_1^2 \varphi_{\max}(\Sigma)}{np^2} \left[d_{m^2}(1+\alpha/2) + d_{\widehat{m}^2}(1+\alpha/2) \right] + \frac{8\xi^2 \varphi_{\max}(\Sigma)\delta^2}{np^2(\delta-1)} \quad (6.61) \end{aligned}$$

To conclude, we need to compare the dimension $d_{m'^2}$ of the space $U_{m'}$ with $d_{m'}$.

Lemma 6.29. *For any model $m \in \mathcal{M}$, it holds that*

$$d_{m^2} \leq L d_m ,$$

where L is a numerical constant between 4 and 5.48.

The proof is postponed to the end of this section. Defining the universal constant $K_0 := 2L$, we derive from (6.61) that

$$\begin{aligned} \overline{\gamma}_{n,p}(\theta_{m,\rho_1}) - \overline{\gamma}_{n,p}(\widetilde{\theta}_{\rho_1}) &\leq \frac{1}{\sqrt{\delta}} l(\widetilde{\theta}_{\rho_1}, \theta) + \frac{\sqrt{\delta}}{\sqrt{\delta}-1} l(\theta_{m,\rho_1}, \theta) \\ &+ \frac{K_0\delta^2\rho_1^2 \varphi_{\max}(\Sigma)}{np^2} \left[d_m(1+\alpha/2) + d_{\widehat{m}}(1+\alpha/2) + \frac{\xi^2}{\delta-1} \right] , \end{aligned}$$

with probability larger than $\mathbb{P}\left(\Omega'_\xi\right)$. The isotropic case is analogous if we replace d_m by d_m^{iso}. □

Proof of Lemma 6.28. We only consider here the anisotropic case, since the isotropic case is analogous. This result is based on the deviation inequality for suprema of Gaussian chaos of order 2 stated in Proposition 6.23. For any model m' belonging to $\mathcal{M}$, we shall upper bound the quantities $\mathbb{E}(Z_{m'})$, $B_{m'}$, and $\mathbb{E}(W_{m'})$ defined in (6.43).

1. Let us first consider the expectation of $Z_{m'}$. Let $U'_{m,m'}$ be the new vector space defined by

$$U'_{m,m'} := U_{m,m'} \frac{\sqrt{D_\Sigma}}{p} \ ,$$

where $U_{m,m'}$ is introduced in the proof of Lemma 6.27. This new space allows to handle the computation with the canonical inner product in the space of matrices. Let $\mathcal{B}^{(2)}_{m^2,m'^2}$ be the unit ball of $U'_{m,m'}$ with respect to the canonical inner product. If R belongs to $U_{m,m'}$, then $\|R\|_{\mathcal{H}'} = \|R\frac{\sqrt{D_\Sigma}}{p}\|_F$, where $\|.\|_F$ stands for the Frobenius norm.

$$\begin{aligned} Z_{m'} &= \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,m'^2}} \frac{1}{p^2} tr\left[RD_\Sigma(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2})\right] \\ &= \sup_{R \in \mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R\frac{\sqrt{D_\Sigma}}{p}\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right] \qquad (6.62) \\ &= \|\Pi_{U'_{m,m'}} \frac{\sqrt{D_\Sigma}}{p}\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\|_F \ , \end{aligned}$$

where $\Pi_{U'_{m,m'}}$ refers to the orthogonal projection with respect to the canonical inner product onto the space $U'_{m,m'}$. Let $F_1, \ldots, F_{d_{m^2,m'^2}}$ denote an orthonormal basis of $U'_{m,m'}$.

$$\begin{aligned} \mathbb{E}(Z^2_{m'}) &= \sum_{i=1}^{d_{m^2,m'^2}} \mathbb{E}\left[tr^2\left(F_i\sqrt{\frac{D_\Sigma}{p^2}}\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right)\right] \\ &= \sum_{i=1}^{d_{m^2,m'^2}} \mathbb{E}\left[\sum_{j=1}^{p^2} F_i[j,j]\frac{\sqrt{D_\Sigma[j,j]}}{p}(\overline{\mathbf{Y}\mathbf{Y}^*}[j,j] - 1)\right]^2 \\ &= \sum_{i=1}^{d_{m^2,m'^2}} \frac{2}{np^2} tr(F_i D_\Sigma F_i) \\ &\leq \sum_{i=1}^{d_{m^2,m'^2}} \frac{2\varphi_{\max}(D_\Sigma)}{np^2} = \frac{2d_{m^2,m'^2}\varphi_{\max}(\Sigma)}{np^2} \ . \end{aligned}$$

Applying Cauchy-Schwarz inequality, it follows that

$$\mathbb{E}(Z_{m'}) \leq \sqrt{\frac{2d_{m^2,m'^2}\varphi_{\max}(\Sigma)}{np^2}} \ . \qquad (6.63)$$

2. Using the identity (6.62), the quantity $B_{m'}$ equals

$$B_{m'} = \frac{2}{n} \sup_{R \in \mathcal{B}^{(2)}_{m^2,m'^2}} \varphi_{\max}\left(R\frac{\sqrt{D_\Sigma}}{p}\right) \ .$$

As the operator norm is under-multiplicative and as it dominates the Frobenius norm, we get the following bound

$$B_{m'} \leq \frac{2\sqrt{\varphi_{\max}(\Sigma)}}{np} \ . \qquad (6.64)$$

3. Let us turn to bounding the quantity $\mathbb{E}(W_{m'})$. Again, by introducing the ball $\mathcal{B}^{(2)}_{m^2,m'^2}$, we get

$$\begin{aligned} W_{m'} &= \frac{4}{n} \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,m'^2}} \frac{1}{p^2} tr\left[R\overline{\mathbf{Y}\mathbf{Y}^*}D_\Sigma R\right] \\ &\leq \frac{4\varphi_{\max}(\Sigma)}{np^2} \sup_{R \in \mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R\overline{\mathbf{Y}\mathbf{Y}^*}R\right] \\ &\leq \frac{4\varphi_{\max}(\Sigma)}{np^2}\left(1 + \sup_{R \in \mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)R\right]\right) \ . \end{aligned}$$

Let $F_1, \dots F_{d_{m^2,m'^2}}$ an orthonormal basis of $U'_{m,m'}$ and let λ be a vector in $\mathbb{R}^{d_{m^2,m'^2}}$. We write $\|\lambda\|_2$ for its L_2 norm.

$$\begin{aligned}\mathbb{E}\left(\sup_{R\in\mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R\left(\overline{\mathbf{YY}^*}-I_{p^2}\right)R\right]^2\right) &= \mathbb{E}\left(\sup_{\|\lambda\|_2\le 1}\sum_{i,j=1}^{d_{m^2,m'^2}}\lambda_i\lambda_j tr\left[F_iF_j(\overline{\mathbf{YY}^*}/n-I_{p^2})\right]\right)^2\\ &\le \sum_{i,j=1}^{d_{m^2,m'^2}}\mathbb{E}\left(tr\left[F_iF_j(\mathbf{YY}^*/n-I_{p^2})\right]^2\right)\ .\end{aligned}$$

The second inequality is a consequence of Cauchy-Schwarz inequality in $\mathbb{R}^{(d_{m^2,m'^2})^2}$ since the l_2 norm of the vector $(\lambda_i\lambda_j)_{1\le i,j\le d_{m^2,m'^2}} \in \mathbb{R}^{d^2_{m^2,m'^2}}$ is bounded by 1. Since the matrices F_i are diagonal, we get

$$\mathbb{E}\left(\sup_{R\in\mathcal{B}^{(2)}_{m^2,m'^2}} tr\left[R(\mathbf{YY}^*/n-I)R\right]^2\right)\le\frac{2}{n}\sum_{i,j=1}^{d_{m^2,m'^2}}\|F_iF_j\|_2^2\ .$$

It remains to bound the norm of the products F_iF_j for any i,j between 1 and d_{m^2,m'^2}.

$$\sum_{i,j=1}^{d_{m^2,m'^2}}\|F_iF_j\|_2^2 = \sum_{i,j=1}^{d_{m^2,m'^2}}\sum_{k=1}^{p^2}F_{i[k,k]}{}^2F_{j[k,k]}{}^2=\sum_{k=1}^{p^2}\left(\sum_{i=1}^{d_{m^2,m'^2}}F_{i[k,k]}{}^2\right)^2\ .$$

For any $k\in\{1,\dots,p^2\}$, $\sum_{i=1}^{d_{m^2,m'^2}}F_{i[k,k]}{}^2\le 1$ since $\left(F_1,\dots,F_{d_{m^2,m'^2}}\right)$ form an orthonormal family. Hence, we get

$$\sum_{i,j=1}^{d_{m^2,m'^2}}\|F_iF_j\|_2^2 \le \sum_{k=1}^{p^2}\sum_{i=1}^{d_{m^2,m'^2}}F_{i[k,k]}{}^2=d_{m^2,m'^2}\ .$$

All in all, we have proved that

$$\mathbb{E}(W_{m'})\le\frac{4\varphi_{\max}(\Sigma)}{np^2}\left[1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right]. \tag{6.65}$$

Gathering these three bounds and applying Proposition 6.23 allows to obtain the following deviation inequality:

$$\begin{aligned}&\mathbb{P}\left(Z_{m'}\ge\sqrt{\tfrac{2\varphi_{\max}(\Sigma)}{n}}\left\{\sqrt{1+\alpha/2}\sqrt{d_{m^2,m'^2}}+\xi\right\}\right)\\ &\le\exp\left\{-\left[\frac{\left[\left(\sqrt{1+\alpha/2}-1\right)\sqrt{d_{m^2,m'^2}}+\xi\right]^2}{2L_1\left(1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right)}\bigwedge\frac{\sqrt{n}\left[\left(\sqrt{1+\alpha/2}-1\right)\sqrt{d_{m^2,m'^2}}+\xi\right]}{\sqrt{2}L_2}\right]\right\}\\ &\le\exp\left\{-\left[\frac{\left[\sqrt{1+\alpha/2}-1\right]^2d_{m^2,m'^2}}{2L_1\left(1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right)}\bigwedge\frac{\sqrt{n}\left(\sqrt{1+\alpha/2}-1\right)\sqrt{d_{m^2,m'^2}}}{\sqrt{2}L_2}\right]-\left[\frac{\xi\left[\sqrt{1+\alpha/2}-1\right]\sqrt{d_{m^2,m'^2}}}{L_1\left[1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right]}\bigwedge\frac{\sqrt{n}\xi}{\sqrt{2}L_2}\right]\right\}.\end{aligned}$$

As n and d_{m^2,m'^2} are larger than one, there exists a universal constant L'_2 such that

$$\left[\frac{(\sqrt{1+\alpha/2}-1)^2d_{m^2,m'^2}}{2L_1\left(1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right)}\bigwedge\frac{\sqrt{n}(\sqrt{1+\alpha/2}-1)\sqrt{d_{m^2,m'^2}}}{\sqrt{2}L_2}\right]\ge 4L'_2\sqrt{d_{m^2,m'^2}}\left[\left(\sqrt{1+\alpha/2}-1\right)^2\wedge\left(\sqrt{1+\alpha/2}-1\right)\right].$$

Since the vector space $U_{m,m'}$ contains all the matrices $D(\theta')$ with θ' belonging to m', d_{m^2,m'^2} is larger than $d_{m'}$. Besides, by concavity of the square root function, it holds that $\sqrt{1+\alpha/2}-1\ge\frac{\alpha}{4\sqrt{1+\alpha/2}}$.

Setting $L_1' := \frac{1}{4L_1(1+\sqrt{2})} \wedge \frac{1}{\sqrt{2}L_2}$ and arguing as previously leads to

$$\left[\frac{\xi(\sqrt{1+\alpha/2}-1)\sqrt{d_{m^2,m'^2}}}{L_1\left(1+\sqrt{\frac{2d_{m^2,m'^2}}{n}}\right)} \bigwedge \frac{\sqrt{n}\xi}{\sqrt{2}L_2} \right] \geq L_1'\xi \left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n} \right] .$$

Gathering these two inequalities allows us to conclude that

$$\mathbb{P}\left(Z_{m'} \geq \sqrt{\frac{2\varphi_{\max}(\Sigma)}{n}} \left\{ \sqrt{(1+\alpha/2)\, d_{m^2,m'^2}} + \xi \right\} \right) \leq \exp\left\{ -L_2'\sqrt{d_{m'}} \left(\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \frac{\alpha^2}{1+\alpha/2} \right) - L_1'\xi \left[\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \sqrt{n} \right] \right\} .$$

□

Proof of Lemma 6.29. The approach falls in two parts. First, we relate the dimensions d_m and d_{m^2} to the number of nodes of the torus Λ that are closer than r_m or $2r_m$ to the origin $(0,0)$. We recall that the quantity r_m is introduced in Definition 6.2. Second, we compute a nonasymptotic upper bound of the number of points in $\mathbb{Z}^2$ that lie in the disc of radius r. This second step is quite tedious and will only give the main arguments.

Let m be a model of the collection $\mathcal{M}_1$. By definition, m is the set of points lying in the disc of radius r_m centered on $(0,0)$. Hence,

$$\Theta_m = vect\left\{ \Psi_{i,j},\ (i,j) \in m \right\} ,$$

where the matrices $\Psi_{i,j}$ are defined by (6.14). As $\Psi_{i,j} = \Psi_{-i,-j}$, the dimension d_m of Θ_m is exactly the number of orbits of m under the action of the central symmetry s.

As d_{m^2} is defined as the dimension of the space U_m, it also corresponds to the dimension of the space

$$vect\left\{ C(\theta), \theta \in \Theta_m \right\} + vect\left\{ C(\theta)^2, \theta \in \Theta_m \right\} , \tag{6.66}$$

which is clearly in one to one correspondence with U_m. Straightforward computations lead to the following identity:

$$C(\Psi_{i_1,j_1})C(\Psi_{i_2,j_2}) = C(\Psi_{i_1+i_2,j_1+j_2})\left[1 + s_{i_1+i_2,j_1+j_2}\right] + C(\Psi_{i_1-i_2,j_1-j_2})\left[1 + s_{i_1-i_2,j_1-j_2}\right]) ,$$

where $s_{x,y}$ is the indicator function of $x = -x$ and $y = -y$ in the torus Λ. Combining this property with the definition of Θ_m, we embed the space (6.66) in the space

$$vect\left\{ C\left(\Psi_{i_1+i_2,j_1+j_2}\right),\ (i_1,j_1),(i_2,j_2) \in m \cup \{(0,0)\} \right\} ,$$

and this last space is in one to one correspondence with

$$vect\left\{ \Psi_{i_1+i_2,j_1+j_2},\ (i_1,j_1),(i_2,j_2) \in m \cup \{(0,0)\} \right\} . \tag{6.67}$$

In the sequel, $\mathcal{N}(m)$ stands for the set $\{(i_1+i_2, j_1+j_2),\ (i_1,j_1),(i_2,j_2) \in m \cup \{(0,0)\}\}$. Thus, the dimension d_{m^2} is smaller or equal to the number of orbits of $\mathcal{N}(m)$ under the action of the symmetry s.

To conclude, we have to compare the number of orbits in m and the number of orbits in $\mathcal{N}(m)$. We distinguish two cases depending whether $2r_m + 1 \leq p$ or $2r_m + 1 > p$. First, we assume that $2r_m + 1 \leq p$. For such values the disc of radius r_m centered on the points $(0,0)$ in not overlapping itself on the torus except on a set of null Lebesgue measure. In the sequel, $\lfloor x \rfloor$ refers to the largest integer smaller than x. We represent the orbit space of m as in Figure 6.2. To any of these points, we associate a square of size 1. If we add $2 + 2\lfloor r_m \rfloor$ squares to the d_m first squares, we remark that the half disc centered on $(0,0)$ and with length r_m is contained in the reunion of these squares. Then, we get

$$d_m + 2 + 2\lfloor r_m \rfloor \geq \frac{\pi r_m^2}{2} . \tag{6.68}$$

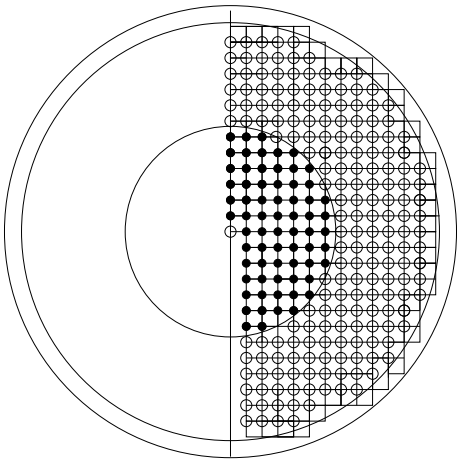


Figure 6.2: *The black dots represent the orbit space of m and the white dots represent the remaining points of the orbit space of $\mathcal{N}(m)$.*

The points in $\mathcal{N}(m)$ are closer than $2r_m$ from the origin. Consequently, all the squares associated to representants of $\mathcal{N}(m)$ are included in the disc of radius $2r_m + \sqrt{2}$.

$$d_{m^2} + 2 + 2\lfloor 2r_m \rfloor \leq \frac{\pi}{2}\left\{2r_m + \sqrt{2}\right\}^2 .$$

Combining these two inequalities, we are able to upper bound d_{m^2}

$$\begin{aligned} 2 + 2\lfloor 2r_m \rfloor + d_{m^2} &\leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 (d_m + 1 + 2\lfloor r_m \rfloor) , \\ d_{m^2} &\leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 d_m + 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 (1 + 2\lfloor r_m \rfloor) . \end{aligned}$$

Applying again inequality (6.68), we upper bound r_m:

$$r_m \leq \frac{2}{\pi}\left[1 + \sqrt{1 + \frac{\pi}{2}(1 + d_m)}\right] .$$

Gathering these two last bounds yields

$$d_{m^2} \leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 \left[1 + \frac{1}{d_m}\left(1 + \frac{4}{\pi}\left[1 + \sqrt{1 + \frac{\pi}{2}(1 + d_m)}\right]\right)\right] d_m .$$

This upper bound is equivalent to $4d_m$, when d_m goes to infinity. Computing the ratio d_{m^2}/d_m for every model m of small dimension allows to conclude.

Let us turn to the case $2r_m + 1 > p$. Suppose that p is larger or equal to 9. The lower bound (6.68) does not necessarily hold anymore. Indeed, the disc is overlapping with itself because of toroidal effects. Nevertheless, we obtain a similar lower bound by replacing r_m by $(p-1)/2$:

$$d_m + 2 + 2\lfloor \frac{p-1}{2} \rfloor \geq \frac{\pi(p-1)^2}{8}.$$

The number of orbits of Λ under the action of the symmetry s is $\frac{p^2+1}{2}$ if p is odd and $\frac{(p+1)^2-1}{2}$ if p is even. It follows that $d_{m^2} \leq \frac{(p+1)^2-1}{2}$. Gathering these two bounds, we get

$$\frac{d_{m^2}}{d_m} \leq \frac{(p+1)^2}{\pi(p-1)^2/4 - 2(p+1)} .$$

This last quantity is smaller than 4 for any $p \geq 9$. An exhaustive computation of the ratios when $p < 9$ allows to conclude.

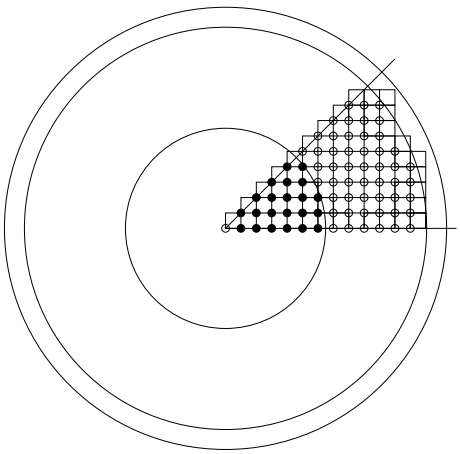


Figure 6.3: *The black dots represent the orbit space of m under the action of G and the white dots represent the remaining points of the orbit space of $\mathcal{N}^{iso}(m)$.*

Let us turn to the isotropic case. Arguing as previously, we observe that the dimension $d_m^{\rm iso}$ is the number of orbits of the set m under the action of the group G introduced in Definition 6.5 whereas d_{m^2} is smaller or equal to the number of orbits of $\mathcal{N}^{\rm iso}(m)$ under the action of G. As for anisotropic models, we choose represent these orbits on the torus and associate squares of size 1 (see Figure 6.3). Assuming that $r_m < (p-1)/2$, we bound d_m and d_{m^2}.

$$\begin{aligned} d_m + 1 &\geq \frac{1}{8}\pi r_m^2 + \frac{1}{2}\lfloor \frac{\sqrt{2}r_m}{2} \rfloor \ , \\ d_{m^2} &\leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 \frac{1}{8}\pi r_m^2 + \frac{1}{2}\lfloor \sqrt{2}r_m \rfloor \ . \end{aligned}$$

Gathering these two inequalities, we get

$$d_{m^2} \leq 4\left\{1 + \frac{\sqrt{2}}{2r_m}\right\}^2 d_m \ .$$

As a consequence, d_{m^2} is smaller than $4d_m$ when d_m goes to infinity. As previously, computing the ratio d_{m^2}/d_m for models m of small dimension allows to conclude. The case $r_m > (p-1)/2$ is handled as for the anisotropic case. □

6.8.3 Proofs of the minimax results

Let us first prove a minimax lower bound on hypercubes $\mathcal{C}_m(\theta', r)$. We recall that these hypercubes are introduced in Definition 6.15.

Lemma 6.30. *Let m be a model in $\mathcal{M}_1$ that satisfies $d_m \leq \sqrt{n}p$ and let θ' be a matrix in $\Theta_m \cap \mathcal{B}_1(0_p, 1)$. Then, for any positive number r such that $(1 - \|\theta'\|_1 - 2rd_m)$ is positive,*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathrm{Co}[\mathcal{C}_m(\theta', r)]} \mathbb{E}_\theta \left[l\left(\widehat{\theta}, \theta\right) \right] \geq L\sigma^2 \left(r \wedge \frac{1 - \|\theta'\|_1}{\sqrt{np^2}} \right)^2 d_m \ ,$$

where $\mathrm{Co}\left[\mathcal{C}_m(\theta', r)\right]$ denotes the convex hull of $C_m(\theta', r)$. Similarly, let m be a model in $\mathcal{M}_1$ such $d_m^{\rm iso} \leq \sqrt{n}p$ and let θ' be a matrix in $\Theta_m^{\rm iso} \cap \mathcal{B}_1(0_p, 1)$. Then, for any positive number r such that $(1 - \|\theta'\|_1 - 8rd_m^{\rm iso})$ is positive,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathrm{Co}[\mathcal{C}_m^{\rm iso}(\theta', r)]} \mathbb{E}_\theta \left[l\left(\widehat{\theta}, \theta\right) \right] \geq L\sigma^2 \left(r \wedge \frac{1 - \|\theta'\|_1}{\sqrt{np^2}} \right)^2 d_m^{\rm iso} \ .$$

Proof of Proposition 6.16. The first result derives from Lemma 6.30 applied to the hypercube $\mathcal{C}_m(0_p, \frac{1}{\sqrt{np^2}})$. We prove the second result using the same lemma with $\mathcal{C}_m(\theta', \frac{1-\|\theta\|_1}{\sqrt{np}})$. □

Proof of Lemma 6.30. This lower bound is based on an application of Fano's approach. See [Yu97] for a review of this method and comparisons with Le Cam's and Assouad's Lemma. The proof follows three main steps: First, we upper bound the Kullback-Leibler entropy between distributions corresponding to θ_1 and θ_2 in the hypercube. Second, we find a set of points in the hypercube well separated with respect to the Hamming distance. Finally, we conclude by applying Birgé's version of Fano's lemma.

Lemma 6.31. *The Kullback-Leibler entropy between two mean zero-Gaussian vectors of size p^2 with precision matrices $\left[I_{p^2} - C(\theta_1)\right]/\sigma^2$ and $\left[I_{p^2} - C(\theta_2)\right]/\sigma^2$ equals*

$$\mathcal{K}(\theta_1, \theta_2) = 1/2 \left[\log\left(\frac{|I_{p^2} - C(\theta_1)|}{|I_{p^2} - C(\theta_2)|} \right) + tr\left(\left[I_{p^2} - C(\theta_2)\right] \left[I_{p^2} - C(\theta_1)\right]^{-1} \right) - p^2 \right] ,$$

where for any square matrix A, $|A|$ refers to the determinant of A.

This statement is classical and its proof is omitted. The matrices $\left[I_{p^2} - C(\theta_1)\right]$ and $\left[I_{p^2} - C(\theta_2)\right]$ are diagonalizable in the same basis since they are symmetric block circulant (Lemma 6.37). Transforming vectors of size p^2 into $p \times p$ matrices, we respectively define λ_1 and λ_2 as the $p \times p$ matrices of eigenvalues of $\left[I_{p^2} - C(\theta_1)\right]$ and $\left[I_{p^2} - C(\theta_2)\right]$. It follows that

$$\mathcal{K}(\theta_1, \theta_2) = 1/2 \sum_{1 \leq i,j \leq p} \left(\frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} - \log\left(\frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} \right) - 1 \right) .$$

For any $x > 0$, the following inequality holds

$$x - 1 - \log(x) \leq \frac{9}{64} \left(x - \frac{1}{x} \right)^2 .$$

It is easy to establish by studying the derivative of corresponding functions. As a consequence,

$$\begin{aligned} \frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} - \log\left(\frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} \right) - 1 &\leq \frac{9}{64} \left(\frac{\lambda_{2[i,j]}}{\lambda_{1[i,j]}} - \frac{\lambda_{1[i,j]}}{\lambda_{2[i,j]}} \right)^2 \\ &\leq \frac{9}{64} \left(\frac{1}{\lambda_{1[i,j]}} + \frac{1}{\lambda_{2[i,j]}} \right)^2 \left(\lambda_{1[i,j]} - \lambda_{2[i,j]} \right)^2 . \end{aligned} \tag{6.69}$$

Let us first consider the anisotropic case. Let m be a model in $\mathcal{M}_1$ and let θ' belong $\Theta_m \cap \mathcal{B}_1(0_p, 1)$. We also consider a positive radius r such that $(1 - \|\theta'\|_1 - 2rd_m)$ is positive. For any θ_1, θ_2 in $\mathcal{C}_m(\theta', r)$ the matrices $(I_{p^2} - C(\theta_1))$ and $(I_{p^2} - C(\theta_2))$ are diagonally dominant and their eigenvalues $\lambda_{1[i,j]}$ and $\lambda_{2[i,j]}$ are larger than $1 - \|\theta'\|_1 - 2rd_m$.

$$\begin{aligned} \mathcal{K}(\theta_1, \theta_2) &\leq \frac{9}{16(1 - \|\theta'\|_1 - 2rd_m)^2} \sum_{1 \leq i,j \leq p} \left(\lambda_{1[i,j]} - \lambda_{2[i,j]} \right)^2 \\ &\leq \frac{9}{16(1 - \|\theta'\|_1 - 2rd_m)^2} \|C(\theta_1) - C(\theta_2)\|_F^2 \\ &\leq \frac{9 d_m r^2 p^2}{8(1 - \|\theta'\|_1 - 2rd_m)^2} . \end{aligned} \tag{6.70}$$

We recall that $\|.\|_F$ refers to the Frobenius norm in the space of matrices.

Let us state Birgé's version of Fano's lemma [Bir05] and a combinatorial argument known under the name of Varshamov-Gilbert's lemma. These two lemma are taken from [Mas07] and respectively correspond to Corollary 2.18 and Lemma 4.7.

Lemma 6.32. (Birgé's lemma) *Let (S,d) be some pseudo-metric space and $\{\mathbb{P}_s, s \in S\}$ be some statistical model. Let κ denote some absolute constant smaller than one. Then for any estimator $\widehat{s}$ and any finite subset T of S, setting $\delta = min_{s,t\in T, s\neq t} d(s,t)$, provided that $\max_{s,t\in T} \mathcal{K}(\mathbb{P}_s, \mathbb{P}_t) \leq \kappa \log|T|$, the following lower bound holds for every $p \geq 1$,*

$$\sup_{s\in S} \mathbb{E}_s[d^p(s,\widehat{s})] \geq 2^{-p}\delta^p(1-\kappa) \ .$$

Lemma 6.33. (Varshamov-Gilbert's lemma) *Let $\{0,1\}^d$ be equipped with Hamming distance d_H. There exists some subset Φ of $\{0,1\}^d$ with the following properties*

$$d_H(\phi,\phi') > d/4 \text{ for every } (\phi,\phi') \in \Phi^2 \text{ with } \phi \neq \phi' \text{ and } \log|\Phi| \geq \frac{d}{8} \ .$$

Applying Lemma 6.32 with Hamming distance d_H and the set Φ introduced in Lemma 6.33 yields

$$\sup_{\theta\in\mathcal{C}_m(\theta',r)} \mathbb{E}_\theta\left[d_H\left(\widehat{\theta},\theta\right)\right] \geq \frac{d_m}{8}(1-\kappa) \ , \tag{6.71}$$

provided that

$$\frac{9d_m r^2 p^2 n}{8(1-\|\theta'\|_1 - 2rd_m)^2} \leq \frac{\kappa d_m}{8} \ . \tag{6.72}$$

Let us express (6.71) in terms of the Frobenius $\|.\|_F$ norm.

$$\sup_{\theta\in\mathcal{C}_m(\theta',r)} \mathbb{E}_\theta\left[\|C(\widehat{\theta}) - C(\theta)\|_F^2\right] \geq \frac{d_m r^2 p^2}{4}(1-\kappa) \ .$$

Since for every θ in the hypercube, $\sigma^{-2}(I_{p^2} - C(\theta))$ is diagonally dominant, its largest eigenvalue is smaller than $2\sigma^{-2}$. The loss function $l(\widehat{\theta},\theta)$ equals $\frac{\sigma^2}{p^2} tr\{[C(\widehat{\theta}) - C(\theta)](I - C(\theta))^{-1}[C(\widehat{\theta}) - C(\theta)]\}$. It follows that

$$\sup_{\theta\in\mathcal{C}_m(\theta',r)} \mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right] \geq \sigma^2\frac{d_m r^2}{8}(1-\kappa) \ . \tag{6.73}$$

Condition (6.72) is equivalent to $\frac{r^2}{(1-\|\theta'\|_1 - 2rd_m)^2} \leq \frac{\kappa}{9p^2n}$. If we assume that

$$r^2 \leq \frac{\kappa\left(1-\|\theta'\|_1\right)^2}{18p^2n} \ , \tag{6.74}$$

then $1 - \|\theta'\|_1 - 2rd_m \geq (1-\|\theta'\|_1)(1 - 2d_m\sqrt{\frac{\kappa}{18np^2}})$. This last quantity is larger than $(1-\|\theta'\|_1)/\sqrt{2}$ if d_m is smaller than $\frac{3}{2}(\sqrt{2}-1)\sqrt{np^2/\kappa}$. Gathering inequality (6.73) and condition (6.74), we get the lower bound

$$\begin{aligned}
\inf_{\widehat{\theta}} \sup_{\theta\in \mathrm{Co}[\mathcal{C}_m(\theta',r)]} \mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right] &\geq \inf_{\widehat{\theta}} \sup_{\theta\in\mathcal{C}_m\left[\theta', r\wedge(1-\|\theta'\|_1)\sqrt{\frac{\kappa}{18p^2n}}\right]} \mathbb{E}_\theta\left[l\left(\widehat{\theta},\theta\right)\right] \\
&\geq L\left(r^2 \wedge \frac{\left(1-\|\theta'\|_1\right)^2}{np^2}\right) d_m\sigma^2 \ .
\end{aligned}$$

One handles models of dimension d_m between $\frac{3}{2}(\sqrt{2}-1)\sqrt{np^2/\kappa}$ and $\sqrt{n}p$ by changing the constant L in the last lower bound.

Let us turn to sets of isotropic GMRFs. The proof is similar to the non-isotropic case, except for a few arguments. Let m belongs to the collection $\mathcal{M}_1$ and let θ' be an element of $\Theta_m^{\mathrm{iso}} \cap \mathcal{B}_1(0_p,1)$. Let r be such that $1 - \|\theta'\|_1 - 8d_m^{iso}$ is positive. If θ_1 and θ_2 belong to the hypercube $\mathcal{C}_m^{\mathrm{iso}}(\theta',r)$, then

$$\mathcal{K}(\theta_1,\theta_2) \leq \frac{9d_m r^2 p^2}{2(1-\|\theta'\|_1 - 8rd_m^{\mathrm{iso}})^2} \ .$$

Applying Lemma 6.32 and 6.33, it follows that

$$\inf_{\widehat{\theta}} \sup_{\theta \in C_m^{\mathrm{iso}}(\theta', r)} \mathbb{E}_\theta \left[d_H \left(\widehat{\theta}, \theta \right) \right] \geq \frac{d_m^{\mathrm{iso}}}{8} (1 - \kappa) \ ,$$

provided that $\frac{9 d_m r^2 p^2 n}{2(1 - \|\theta'\|_1 - 8 r d_m^{\mathrm{iso}})^2} \leq \frac{\kappa d_m^{\mathrm{iso}}}{8}$. As a consequence,

$$\inf_{\widehat{\theta}} \sup_{\theta \in C_m^{\mathrm{iso}}(\theta', r)} \mathbb{E}_\theta \left[l \left(\widehat{\theta}, \theta \right) \right] \geq \frac{d_m^{\mathrm{iso}} r^2}{8} (1 - \kappa) \ ,$$

if $\frac{r^2}{(1 - \|\theta'\|_1 - 8 r d_m^{\mathrm{iso}})^2} \leq \frac{\kappa}{36 p^2 n}$. We conclude by arguing as in the isotropic case. □

Proof of Proposition 6.20. First, observe that the set $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)$ is included in $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1) \cap \mathcal{U}(2)$. We then derive minimax lower bounds on $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)$ from the lower bounds on hypercubes.

Let m_i be a model in $\mathcal{M}_1$ such that d_m is smaller than $\sqrt{n}p$. Let us look for positive numbers r such that the hypercube $[\mathcal{C}_{m_i}(0_p, r)]$ is included in the set $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)$.

Lemma 6.34. *Let m be a model in $\mathcal{M}_1$ and r be a positive number smaller than $1/(4d_m)$. For any $\theta \in \mathrm{Co}\left[\mathcal{C}_m(0_p, r)\right]$,*

$$\mathrm{var}_\theta \left(X_{[0,0]} \right) \leq \sigma^2 \left(1 + 16 d_m r^2 \right) \ .$$

If we choose

$$r \leq \frac{a_i}{16 \sigma \sqrt{d_{m_i}}} \ ,$$

then $2 r d_{m_i}$ is smaller than $1/8$ by assumption $(\mathbb{H}_a)$. Applying Lemma 6.34, we then derive that $\mathrm{var}_\theta \left(X_{[0,0]} \right) \leq \sigma^2 + a_i^2$. Hence, we get the upper bound $\sum_{j=1}^{i} \left[\mathrm{var} \left(X_{[0,0]} | X_{m_{j-1}} \right) - \mathrm{var} \left(X_{[0,0]} | X_{m_j} \right) \right] \leq a_i^2$ and it follows that

$$\sum_{j=1}^{\mathrm{Card}(\mathcal{M}_1)} \frac{\mathrm{var} \left(X_{[0,0]} | X_{m_{k-1}} \right) - \mathrm{var} \left(X_{[0,0]} | X_{m_j} \right)}{a_j^2} \leq 1 \ ,$$

since the sequence $(a_j)_{1 \leq j \leq \mathrm{Card}(\mathcal{M}_1)}$ is non increasing. Consequently, $\mathrm{Co}\left[\mathcal{C}_m(0_p, r)\right]$ is a subset of $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)$. By Lemma 6.30, we get

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)} \mathbb{E}_\theta \left[l \left(\widehat{\theta}, \theta \right) \right] &\geq L \sigma^2 \left(\frac{a_i^2}{16 \sigma^2} \wedge \frac{d_{m_i}}{n p^2} \right) \\ &\geq L \left(a_i^2 \wedge \frac{\sigma^2 d_{m_i}}{n p^2} \right) \ . \end{aligned} \tag{6.75}$$

Considering all models $m \in \mathcal{M}_1$ such that $d_m \leq \sqrt{n}p$ yields

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1/2)} \mathbb{E}_\theta \left[l \left(\widehat{\theta}, \theta \right) \right] \geq L \sup_{i \leq \mathrm{Card}(\mathcal{M}_1),\ d_{m_i} \leq \sqrt{n}p} \left(a_i^2 \wedge \frac{\sigma^2 d_{m_i}}{n p^2} \right) \ . \tag{6.76}$$

If the maximal dimension $d_{m_{\mathrm{Card}(\mathcal{M}_1)}}$ is smaller than $\sqrt{n}p$, the proof is finished. In the opposite case, we need to show that the supremum (6.41) over all models $m \in \mathcal{M}_1$ is achieved at some model m of dimension less than $\sqrt{n}p$.

Lemma 6.35. *For any integer $1 \leq i \leq Card(\mathcal{M}_1) - 1$, the ratio $d_{m_{i+1}}/d_{m_i}$ is less than 2.*

Let i' be the largest integer such that $d_{m_{i'}} \leq \sqrt{n}p$. Since i' is smaller than $\mathrm{Card}(\mathcal{M}_1)$, we know from Lemma 6.35 that $\sqrt{n}p/2 \leq d_{m_{i'}} \leq \sqrt{n}p$. By assumption $(\mathbb{H}_a)$, $a_{i'}^2$ is smaller than $\frac{\sigma^2}{d_{m_{i'}}}$. Gathering these bounds yields

$$a_{i'}^2 \leq \frac{\sigma^2}{d_{m_{i'}}} \leq \frac{4 d_{m_{i'}} \sigma^2}{n p^2} \ .$$

Since the sequence $(a_i)_{1 \leq i \leq \mathrm{Card}(\mathcal{M}_1)}$ is non increasing, the supremum (6.41) over all models in $\mathcal{M}_1$ is either achieved for some $i \leq i'$ or is smaller than $4 \left(a_{i'}^2 \wedge \frac{\sigma^2 d_{m_{i'}}}{n p^2} \right)$. □

Proof of lemma 6.34. Let m be a model in $\mathcal{M}_1$, r be a positive number smaller than $\frac{1}{4d_m}$, and θ be an element of the convex hull of $\mathcal{C}_m(0_p, r)$. The covariance matrix of the vector X^v is $\Sigma = \sigma^2 \left[I - C(\theta)\right]^{-1}$. Since the field X is stationary, $\mathrm{var}_\theta(X_{[0,0]})$ equals any diagonal element of Σ. In particular, $\mathrm{var}_\theta(X_{[0,0]})$ corresponds to the mean of the eigenvalues of Σ. The matrix $[I - C(\theta)]$ is block circulant. As in the proof of Lemma 6.71, we note λ the $p \times p$ matrix of the eigenvalues of $(I_{p^2} - C(\theta))$. By Lemma 6.37,

$$\lambda_{[i,j]} = 1 + \sum_{(k,l)\in\Lambda} \theta_{[k,l]} \cos\left[2\pi\left(\frac{ik}{p} + \frac{jl}{p}\right)\right] ,$$

for any $1 \leq i, j \leq p$. Since θ belongs to the convex hull of $\mathcal{C}_m(0_p, r)$, $\theta_{[k,l]}$ is zero if $(k,l) \notin m$ and $|\theta_{[k,l]}| \leq r$ if $(k,l) \in m$. Thus $\sum_{(k,l)\in\Lambda} |\theta_{[k,l]}|$ is smaller than $1/2$. Applying Taylor-Lagrange inequality, we get

$$\frac{1}{1+x} \leq 1 - x + \frac{x^2}{(1-|x|)^3} ,$$

for any x between -1 and 1. It follows that

$$\lambda_{[i,j]}^{-1} \leq 1 - \sum_{k,l\in\Lambda} \theta_{[k,l]} \cos\left[2\pi\left(\frac{ik}{p} + \frac{jl}{p}\right)\right] + 8\left\{\sum_{k,l\in\Lambda} \theta_{[k,l]} \cos\left[2\pi\left(\frac{ik}{p} + \frac{jl}{p}\right)\right]\right\}^2 . \tag{6.77}$$

Summing this inequality for all $(i,j) \in \{1,\ldots,p\}^2$, the first order term turns out to be $tr[C(\theta)]/p^2$ which is zero whereas the second term equals $8tr[C(\theta)^2]/p^2$. Since there are less than $2d_m$ non-zero terms on each line of the matrix $C(\theta)$, its Frobenius norm is smaller than $2d_m p^2 r^2$. Consequently, we obtain

$$\mathrm{var}_\theta\left(X_{[0,0]}\right) \leq \sigma^2\left(1 + 16 d_m r^2\right) .$$

□

Proof of Lemma 6.35. This property seems straightforward but the proof is a bit tedious. Let i be a positive integer smaller than $\mathrm{Card}(\mathcal{M}_1)$. By definition of the radius r_m in Equation (6.10), the model m_{i+1} is the set of nodes in $\Lambda \setminus \{(0,0)\}$ at a distance smaller or equal to $r_{m_{i+1}}$ from $(0,0)$, whereas the model m_i only contains the points in $\Lambda \setminus \{(0,0)\}$ at a distance strictly smaller than $r_{m_{i+1}}$ from the origin.

Let us first assume that $2r_{m_{i+1}} \leq p$. In such a case, the disc centered on $(0,0)$ with radius $r_{m_{i+1}}$ does not overlap with itself on the torus Λ. To any node in the neighborhood m_{i+1} and to the node $(0,0)$, we associate the square of size 1 centered on it. All these squares do not overlap and are included in the disc of radius $r_{m_{i+1}} + \sqrt{2}/2$. Hence, we get the upper bound $2d_{m_{i+1}} + 1 \leq \pi(r_{m_{i+1}} + \sqrt{2}/2)^2$. Similarly, the disc of radius $r_{m_{i+1}} - \sqrt{2}/2$ is included in the union of the squares associated to the nodes $m_i \cup \{0,0\}$. It follows that $2d_{m_i} + 1$ is larger or equal to $\pi\left(r_{m_{i+1}} - \sqrt{2}/2\right)^2$. Gathering these two inequalities, we obtain

$$\frac{d_{m_{i+1}}}{d_{m_i}} \leq \frac{\left(r_{m_{i+1}} + \frac{\sqrt{2}}{2}\right)^2 - 1}{\left(r_{m_{i+1}} - \frac{\sqrt{2}}{2}\right)^2 - 1} ,$$

if $r_{m_{i+1}}$ is larger than $1 + \sqrt{2}/2$. If $r_{m_{i+1}}$ larger than 5, this upper bound is smaller than two. An exhaustive computation for models of small dimension allows to conclude.

If $2r_{m_{i+1}} \geq p$ and $2r_{m_i} < p$, then the preceding lower bound of d_{m_i} and the preceding upper bound of $d_{m_{i+1}}$ still hold. Finally, let us assume that $2r_{m_i} \geq p$. Arguing as previously, we conclude that $2d_{m_i} + 1 \geq \pi(p/2 - \sqrt{2}/2)^2$. The largest dimension of a model $m \in \mathcal{M}_1$ is $(p^2-1)/2$ if p is odd and $((p+1)^2 - 3)/2$ if p is even. Thus, $d_{m_{i+1}} \leq \frac{(p+1)^2-3}{2}$. Gathering these two bounds yields

$$\frac{d_{m_{i+1}}}{d_{m_i}} \leq \frac{(p+1)^2 - 3}{\left(\frac{p}{2} - \frac{\sqrt{2}}{2}\right)^2} ,$$

which is smaller than 2 if p is larger than 10. Exhaustive computations for small p allow to conclude.

□

Proof of Corollary 6.17. Observe that $\mathrm{Co}[\mathcal{C}_m(0_p, 1/(4d_m)]$ is included in $\Theta_m \cap \mathcal{B}_1(0_p, 1/2)$. This last set is itself included in $\Theta^+_{m,\rho_1} \cap \mathcal{U}(\rho_2)$. Applying Lemma 6.30, we get the following minimax lower bound

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta^+_{m,\rho_1} \cap \mathcal{U}(\rho_2)} \mathbb{E}\left[l\left(\widehat{\theta}, \theta\right)\right] \geq L\sigma^2 \frac{d_m}{np^2} \ ,$$

since the dimension d_m is smaller than np^2. Applying Theorem 6.4, we derive that

$$\begin{aligned} \sup_{\theta \in \Theta^+_{m,\rho_1} \cap \mathcal{U}(\rho_2)} \mathbb{E}\left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right] &\leq L(K)\sigma^2\rho_1^2\rho_2 \frac{d_m}{np^2} + L_2(K)\frac{\rho_1^2}{np^2} \sup_{\theta \in \Theta^+_{m,\rho_1} \cap \mathcal{U}(\rho_2)} \varphi_{\max}(\Sigma) \\ &\leq L(K, \rho_1, \rho_2)\sigma^2 \frac{d_m}{np^2}. \end{aligned}$$

We conclude by combining the two different bounds. □

Proof of Proposition 6.21. This result derives from the upper bound of the risk of $\widetilde{\theta}_{\rho_1}$ stated in Theorem 6.4 and the minimax lower bound stated in Proposition 6.20.

Let $\mathcal{E}(a)$ be a pseudo-ellipsoid that satisfies Assumption $(\mathbb{H}_a)$ and such that $a_1^2 \geq \frac{1}{np^2}$. For any θ in $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1) \cap \mathcal{U}(\rho_2)$, the penalty term satisfies $\mathrm{pen}(m) = K\sigma^2\rho_1^2\rho_2 d_m/np^2$ is larger than $Kd_m\varphi_{\max}(\Sigma)/np^2$. Applying Theorem6.4, we upper bound the risk $\widetilde{\theta}_{\rho_1}$

$$\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right] \leq L_1(K) \inf_{m \in \mathcal{M}_1} \left[l(\theta_{m,\rho_1}, \theta) + \mathrm{pen}(m)\right] + L_2(K)\rho_2 \frac{\sigma^2}{np^2} \ ,$$

for any $\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1) \cap \mathcal{U}(\rho_2)$. It follows that

$$\sup_{\theta \in} \mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right] \leq L(K) \inf_{m \in \mathcal{M}_1 \ , d_m > 0} \left[l(\theta_{m,\rho_1}, \theta) + \rho_1^2\rho_2\sigma^2 \frac{d_m}{np^2}\right] \ .$$

Let i be a positive integer smaller or equal than $\mathrm{Card}(\mathcal{M}_1)$. We know from Section 6.4.1 that the bias $l(\theta_{m_i}, \theta)$ of the model m_i equals $\mathrm{var}(X_{[0,0]}|X_{m_i}) - \sigma^2$. Since θ belongs to the set $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1)$, the bias term is smaller or equal to a^2_{i+1} with the convention $a^2_{\mathrm{Card}(\mathcal{M}_1)+1} = 0$. Hence , the previous upper bound becomes

$$\begin{aligned} \mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right] &\leq L(K) \inf_{1 \leq i \leq \mathrm{Card}(\mathcal{M}_1)} \left[a^2_{i+1} + \rho_1^2\rho_2\sigma^2 \frac{d_{m_i}}{np^2}\right] \\ &\leq L(K, \rho_1, \rho_2) \inf_{1 \leq i \leq \mathrm{Card}(\mathcal{M}_1)} \left[a^2_{i+1} + \frac{\sigma^2 d_{m_i}}{np^2}\right] \ . \end{aligned} \tag{6.78}$$

Applying Proposition 6.20 to the set $\mathcal{E}(a) \cap \mathcal{B}_1(0_p, 1) \cap \mathcal{U}(2)$, we get

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p,1) \cap \mathcal{U}(\rho_2)} \mathbb{E}_\theta\left[l\left(\widehat{\theta}, \theta\right)\right] &\geq \inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p,1) \cap \mathcal{U}(2)} \mathbb{E}_\theta\left[l\left(\widehat{\theta}, \theta\right)\right] \\ &\geq L \sup_{1 \leq i \leq \mathrm{Card}(\mathcal{M}_1)} \left(a_i^2 \wedge \sigma^2 \frac{d_{m_i}}{np^2}\right) \ . \end{aligned}$$

Let us define i^* by

$$i^* := \sup\left\{1 \leq i \leq \mathrm{Card}(\mathcal{M}_1) \, , \, a_i^2 \geq \frac{\sigma^2 d_{m_i}}{np^2}\right\} \ ,$$

with the convention $\sup \varnothing = 0$. Since $a_1^2 \geq \sigma^2/np^2$, i^* is larger or equal to one. It follows that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{E}(a) \cap \mathcal{B}_1(0_p, \eta)} \mathbb{E}_\theta\left[l\left(\widehat{\theta}, \theta\right)\right] \geq L_2\left(a^2_{i^*+1} \vee \frac{\sigma^2 d_{m_{i^*}}}{np^2}\right) \ .$$

Meanwhile, the upper bound (6.78) on the risk of $\widetilde{\theta}_{\rho_1}$ becomes

$$\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1}, \theta\right)\right] \leq L(K, \rho_1, \rho_2)\left(a^2_{i^*+1} + \frac{\sigma^2 d_{m_{i^*}}}{np^2}\right) \leq 2L(K, \rho_1, \rho_2)\left(a^2_{i^*+1} \vee \frac{\sigma^2 d_{m_{i^*}}}{np^2}\right) \ ,$$

which allows to conclude. □

6.8.4 Proofs of the asymptotic risk bounds

Proof of Proposition 6.9. This result is closely related to Proposition 4.11 in [Guy95]. In fact, we extend his proof to stationary fields on a torus. In the sequel, we shall only consider non-isotropic GMRFs, the isotropic case being similar. Let us fix a model m in the collection $\mathcal{M}_1$ and let us assume $(\mathbb{H}_1)$.

We define the $d_m \times p^2$ matrix χ_m^v as

$$(\chi_m^v)^* := \left([C(\Psi_{i_k,j_k})X^v]\,,\ k=1,\dots,d_m\right)\ .$$

For any $(i,j) \in \{1,\dots,p\}^2$, the $(i-1)p+j$-th row of χ_m^v corresponds to the list of covariates used when performing the regression of $X_{[i,j]}$ with respect to its neighbours in the model m. Contrary to the previous proofs, we need to express the $n \times p^2$ matrix $\mathbf{X^v}$ in terms of a vector. This is why we define the vector $\mathbf{X^V}$ of size np^2 as

$$\mathbf{X^V}_{[p^2(j-1)+p(i_1-1)+i_2]} := \mathbf{X}^j{}_{[i_1,i_2]}\ ,$$

for any $(i_1,i_2) \in \{1,\dots,p\}^2$ and any $j \le n$. Similarly, let $\boldsymbol{\chi}_m^V$ be the $d_m \times np^2$ matrix defined as

$$\boldsymbol{\chi}_m^{\mathbf{V}}{}_{[k,p^2(j-1)+p(i_1-1)+i_2]} := \boldsymbol{\chi}_m^j{}_{[p(i_1-1)+i_2]}\ ,$$

for any $(i_1,i_2) \in \{1,\dots,p\}^2$ and any $j \le n$.

We are not able to work out directly the asymptotic risk of $\widehat{\theta}_{m,\rho_1}$. This is why we introduce a new estimator $\check{\theta}_m$ whose asymptotic distribution is easier to derive. Afterwards, we shall prove that $\check{\theta}_m$ and $\widehat{\theta}_{m,\rho_1}$ have the same asymptotic distribution. Let us respectively define the estimators $\check{a}_m$ in $\mathbb{R}^{d_m}$ and $\check{\theta}_m$ as

$$\begin{aligned} \check{a}_m &:= \left(\left(\boldsymbol{\chi}_m^{\mathbf{V}}\right)^* \boldsymbol{\chi}_m^{\mathbf{V}}\right)^{-1} \boldsymbol{\chi}_m^{\mathbf{V}}\mathbf{X^V} \\ \check{\theta}_m &:= \sum_{k=1}^{d_m} \check{a}_{m[k]}\Psi_{i_k,j_k}\ , \end{aligned} \tag{6.79}$$

where we recall that $(\Psi_{i_1,j_1},\dots,\Psi_{i_{d_m},j_{d_m}})$ is a basis of Θ_m. Obviously, $\check{\theta}_m$ is a Conditional least squares estimator since it minimizes the expression (6.16) of $\gamma_{n,p}(.)$ over the whole space Θ_m. Consequently, $\check{\theta}_m$ coincides with $\widehat{\theta}_{m,\rho_1}$ if $\check{\theta}_m$ belongs to Θ^+_{m,ρ_1}.

For the second result, we assume that Assumption $(\mathbb{H}_2)$ holds. Applying Corollary 6.7, we know that for any $(k,l) \in \Lambda$, $X_{[k,l]}$ decomposes as

$$X_{[k,l]} = \sum_{(i,j)\in m} \theta_{m,\rho_1[i,j]} X_{[k+i,l+j]} + \epsilon_{m[k,l]}\ , \tag{6.80}$$

where $\epsilon_{m[k,l]}$ is independent from $\{X_{[k+i,l+j]},\ (i,j) \in m\}$. For the first result, the same decomposition holds since θ is assumed to belong to Θ^+_{m,ρ_1} and θ_{m,ρ_1} therefore equals θ.

Let $a_m \in \mathbb{R}^{d_m}$ be the unique vector such that $\theta_{m,\rho_1} = \sum_{k=1}^{d_m} a_{m[k]}\Psi_{i_k,j_k}$. Then, the previous decomposition becomes

$$X^v = a_m^* \chi_m^v + \epsilon_m^v\ .$$

Gathering this last identity with (6.79) yields

$$\check{a}_m - a_m = \left(\frac{1}{np^2}(\boldsymbol{\chi}_m^{\mathbf{V}})^*\boldsymbol{\chi}_m^{\mathbf{V}}\right)^{-1}\left(\frac{1}{np^2}\boldsymbol{\chi}_m^{\mathbf{V}}\boldsymbol{\epsilon}_m^{\mathbf{V}}\right)\ ,$$

where the vector $\boldsymbol{\epsilon}_m^{\mathbf{V}}$ of size np^2 corresponds to the n observations of the vector ϵ_m^v. When n goes to the infinity, $\frac{1}{np^2}\left(\boldsymbol{\chi}_m^{\mathbf{V}}\right)^*\boldsymbol{\chi}_m^{\mathbf{V}}$ converges almost surely to the covariance matrix V by the law of large numbers. By definition, the variable $\epsilon_{m[i,j]}$ is independent from the $(i-1)p+j$th row of $\chi^v_{m[i,j]}$. It follows that $\mathbb{E}_\theta\left(\boldsymbol{\chi}_m^{\mathbf{V}}\boldsymbol{\epsilon}^{\mathbf{V}}\right) = 0$. Applying again the law of large numbers we conclude that $\check{a}_m$ converges almost surely towards a_m and that $\check{\theta}_m$ converges almost surely towards θ_{m,ρ_1}. Besides, the central limit theorem states that the random vector $\frac{1}{\sqrt{np}}\boldsymbol{\chi}_m^{\mathbf{V}}\boldsymbol{\epsilon}^{\mathbf{V}}$ converges in distribution towards a zero mean Gaussian vector whose

covariance matrix equals $\frac{1}{p^2}\mathrm{var}_\theta\left(\chi_m^v\epsilon_m^v\right)$. By decomposition (6.80), $\epsilon_m^v = (I - C(\theta_{m,\rho_1}))X^v$ while the k-th row of χ_m^v equals $[C(\Psi_{i_k,j_k})X^v]^*$. Thus, for any $1 \leq k,l \leq d_m$,

$$\frac{1}{p^2}\mathrm{var}_\theta\left(\chi_m^v\epsilon_m^v\right)_{[k,l]} = \frac{1}{p^2}\mathrm{cov}_\theta\left[(X^v)^* C(\Psi_{i_k,j_k})\left[I - C(\theta_{m,\rho_1})\right]X^v, (X^v)^* C(\Psi_{i_l,j_l})\left[I - C(\theta_{m,\rho_1})\right]X^v\right] .$$

As the covariance matrix of X^v is $\sigma^2\left[I - C(\theta)\right]^{-1}$, we obtain by standard Gaussian properties

$$\frac{1}{p^2}\mathrm{var}_\theta\left(\chi_m^v\epsilon_m^v\right)_{[k,l]} = \\ \frac{2\sigma^4}{p^2}\mathrm{cov}_\theta\left[\left[I - C(\theta)\right]^{-1} C(\Psi_{i_k,j_k})\left[I - C(\theta_{m,\rho_1})\right]\left[I - C(\theta)\right]^{-1} C(\Psi_{i_l,j_l})\left[I - C(\theta_{m,\rho_1})\right]\right] . \quad (6.81)$$

By Lemma 6.37, all these matrices are diagonalizable in the same basis and therefore commute with each other. We conclude that $\frac{1}{p^2}\mathrm{var}_\theta\left(\chi_m^v\epsilon_m^v\right) = 2\sigma^4 W$ and

$$\sqrt{n}p\left(\check{a}_m - a_m\right) \to \mathcal{N}\left(0, V^{-1}WV^{-1}\right) .$$

As $\widehat{\theta}_{m,\rho_1}$ belongs to Θ^+_{m,ρ_1}, there exists a unique vector $\widehat{a}_m \in \mathbb{R}^{d_m}$ such that $\widehat{\theta}_{m,\rho_1} = \sum_{k=1}^{d_m}\widehat{a}_{m[k]}\Psi_{i_k,j_k}$. The matrix θ_{m,ρ_1} belongs to the open set Θ^+_{m,ρ_1} for the two cases of the propositions. Indeed, θ_{m,ρ_1} equals θ in the first situation. In the second situation, this is due to the fact that θ satisfies $(\mathbb{H}_2)$ and to Lemma 6.6.

Since $\check{\theta}_m$ converges almost surely to θ_{m,ρ_1}, the matrix $\check{\theta}_m$ belongs to m with probability going to one when n goes to infinity. If follows that the estimators $\check{a}_m$ and $\widehat{a}_m$ coincide with probability going to one. By Slutsky's Lemma, we obtain that

$$\sqrt{n}p\left(\widehat{a}_m - a_m\right) \to \mathcal{N}\left(0, V^{-1}WV^{-1}\right) .$$

Let us express the risk of $\widehat{\theta}_{m,\rho_1}$ with respect to the distribution of $\widehat{a}_m$.

$$l\left(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1}\right) = \mathbb{E}_\theta\left[\sum_{k=1}^{d_m}\left(\widehat{a}_{m[k]} - a_{m[k]}\right)tr\left(\Psi_{i_k,j_k}X\right)\right]^2 = tr\left[V\left(\widehat{a}_m - a_m\right)^*\left(\widehat{a}_m - a_m\right)\right] . \quad (6.82)$$

By Portmanteau's Lemma, $np^2 l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})$ converges in distribution towards a random variable whose expectation is $tr\left(WV^{-1}\right)$. In order to conclude, it remains to prove that the sequence $\left[np^2 l\left(\widehat{\theta}_{m,\rho_1}, \theta\right)\right]_{n\geq 1}$ is asymptotically uniformly integrable.

Let us consider a model selection procedure with the collection $\mathcal{M} = \{m\}$ and a penalty term satisfying the assumptions of Theorem 6.4. Arguing as in the proof of this theorem, we derive from identity (6.55) the following property. For any $\xi > 0$, with probability larger than $1 - L_1\exp\left[-L_2\xi\right]$,

$$np^2 l\left(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1}\right) \leq L_3 d_m \varphi_{\max}(\Sigma) + L_4\xi^2\varphi_{\max}(\Sigma) .$$

This clearly implies that the sequence $[np^2 l(\widehat{\theta}_{m,\rho_1}, \theta_{m,\rho_1})]_{n\geq 1}$ is asymptotically uniformly integrable and the first part of the result follows.

For the first result of the proposition, we have stated that θ equals Θ_m. As a consequence,

$$\lim_{n\to+\infty}\mathbb{E}_\theta\left[l\left(\widehat{\theta}_{m,\rho_1}, \theta\right)\right] = 2\sigma^4 tr\left[WV^{-1}\right] .$$

Besides, the term $W_{[k,l]}$ here equals $tr\left[C(\Psi_{i_k,j_k})C(\Psi_{i_l,j_l})\right]$. This last quantity is zero if $k \neq l$ and equals $\|C(\Psi_{i_k,j_k})\|_F^2$ if $k = l$. □

Proof of Corollary 6.11. For the sake of simplicity, we assume that for any node $(i,j) \in m$, the nodes (i,j) and $(-i,-j)$ are different in Λ. If this is not the case, we only have to slightly modify the proof in

order to take account that $\|\Psi_{i,j}\|_F^2$ may equal one. The matrix V is the covariance of the vector of size d_m

$$\left(X_{i_1,j_1} + X_{-i_1,-j_1}, \ldots, X_{i_{d_m},j_{d_m}} + X_{-i_{d_m},-j_{d_m}}\right) . \tag{6.83}$$

Since the matrix Σ of X^v is positive, V is also positive. Moreover, its largest eigenvalue is larger than $2\varphi_{\max}(\Sigma)$.

Let us assume first the θ belongs to Θ_m^+ and that Assumption $(\mathbb{H}_1)$ is fulfilled. By the first result of Proposition 6.9,

$$\lim_{n\to+\infty} np^2\mathbb{E}\left[l\left(\widehat{\theta}_{m,\rho_1},\theta\right)\right] = 2\sigma^4 tr\left[IL_m V^{-1}\right] \geq \frac{\sigma^4}{\varphi_{\max}(\Sigma)} tr[IL_m] = 2\sigma^4 \frac{d_m}{\varphi_{\max}(\Sigma)} ,$$

which corresponds to the first lower bound (6.30).

Let us turn to the second result. We now assume that θ satisfies Assumption $(\mathbb{H}_2)$. By the identity (6.28) of Proposition 6.9, we only have to lower bound the quantity $tr\left[VW^{-1}\right]$.

$$tr\left[V^{-1}W\right] \quad \geq \quad \varphi_{\max}(V)^{-1} tr\left[W\right] \geq \frac{1}{2\varphi_{\max}(\Sigma)} tr[W] .$$

Since the matrix $\Sigma^{-1} = \sigma^{-2}\left[I_{p^2} - C(\theta)\right]$ is diagonally dominant, its smallest eigenvalue is larger than $\sigma^{-2}(1-\|\theta\|_1)$. The matrix $\left[I_{p^2} - C(\theta_{m,\rho_1})\right]^2\left[I_{p^2} - C(\theta)\right]^{-2}$ is symmetric positive. It follows that W is also symmetric positive definite. Hence, we get

$$tr\left[V^{-1}W\right] \quad \geq \quad \frac{\sigma^{-2}}{2}\left[1-\|\theta\|_1\right]\sum_{k=1}^{d_m}\frac{1}{p^2} tr\left[C(\Psi_{i_k,j_k})^2\left[I_{p^2} - C(\theta_{m,\rho_1})\right]^2\left[I_{p^2} - C(\theta)\right]^{-2}\right] . \tag{6.84}$$

The largest eigenvalue of $\left[I_{p^2} - C(\theta)\right]$ is smaller than 2 and the smallest eigenvalue of $\left[I_{p^2} - C(\theta_{m,\rho_1})\right]$ is larger than $1 - \|\theta_{m,\rho_1}\|_1$. By Lemma 6.37, these two matrices are jointly diagonalizable and the smallest eigenvalue of $\left[I_{p^2} - C(\theta_{m,\rho_1})\right]^2\left[I_{p^2} - C(\theta)\right]^{-2}$ is therefore larger than $(1-\|\theta_{m,\rho_1}\|_1)^2/4$. Gathering this lower bound with (6.84) yields

$$tr\left[V^{-1}W\right] \geq \frac{d_m\sigma^{-2}}{2}\left[1-\|\theta\|_1\right]\left[1-\|\theta_{m,\rho_1}\|_1\right]^2 .$$

Lemma 6.6 states that $\|\theta_{m,\rho_1}\|_1 \leq \|\theta\|_1$. Combining these two lower bounds enables to conclude.

□

Proof of Proposition 6.12. As θ belongs to $\Theta^+ \cap \mathcal{B}_1(0_p,\eta)$, the largest eigenvalue of Σ is smaller than $\frac{\sigma^2}{1-\eta}$. Applying Theorem 6.4, we get

$$\begin{aligned}\mathbb{E}_\theta\left[l\left(\widetilde{\theta}_{\rho_1},\theta\right)\right] &\leq L(K)\inf_{m\in\mathcal{M}}\left[l(\theta_{m,\rho_1},\theta) + K\frac{\sigma^2}{np^2(1-\eta)}\right] \\ &\leq L(K,\eta)\inf_{m\in\mathcal{M}}\left[l(\theta_{m,\rho_1},\theta) + K\frac{\sigma^2}{np^2}(1-\eta)^3\right] .\end{aligned}$$

Gathering this bound with the result of Corollary 6.11 enable us to conclude. □

Proof of Example 6.13.

Lemma 6.36. *For any θ is the space $\Theta_{m_1}^{+,\mathrm{iso}}$, the asymptotic variance term of $\widehat{\theta}_{m_1,\rho_1}^{\mathrm{iso}}$ equals*

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}_{m_1,\rho_1}^{\mathrm{iso}},\theta\right)\right] = 2\sigma^4\frac{tr\left(H^2\right)}{tr\left(H^2\Sigma\right)} .$$

If θ belongs to $\Theta^{+,\mathrm{iso}}$ and also satisfies $(\mathbb{H}_2)$, then

$$\lim_{n\to+\infty} np^2\mathbb{E}_\theta\left[l\left(\widehat{\theta}_{m_1,\rho_1}^{\mathrm{iso}},\theta_{m_1,\rho_1}^{\mathrm{iso}}\right)\right] = 2\frac{tr\left\{\left[(I-\theta_{m_1,\rho_1}^{\mathrm{iso}}[1,0]H)H\Sigma\right]^2\right\}}{tr(H^2\Sigma)} , \tag{6.85}$$

where the $p^2\times p^2$ matrix H is defined as $H := C\left(\Psi_{1,0}^{\mathrm{iso}}\right)$.

Proof of Lemma 6.36. Apply Proposition 6.9 noting that $V = tr[H\Sigma H]/p^2$ and

$$W = \frac{tr\left\{\left[(I - \theta_{m_1^{\text{iso}}[1,0]}H)H\Sigma\right]^2\right\}}{\sigma^4 p^2} .$$

To prove the second result, we observe that $\Theta_{m_1}^{+,\text{iso}}$ equals $\Theta_{m_1,2}^{+,\text{iso}}$. It is stated for instance in Table 8. □

Since the matrix θ belongs to $\Theta_{m_1}^{+,\text{iso}}$, we may apply the second result of Lemma 6.36. Straightforward computations lead to $tr(H^2) = \|C\left(\Psi_{1,0}^{\text{iso}}\right)\|_F^2 = 4p^2$ and

$$tr(H^2\Sigma) = 4p^2\left[\text{var}(X_{[0,0]}) + 2\text{cov}_\theta(X_{[0,0]}, X_{[1,1]}) + \text{cov}_\theta\left(X_{[0,0]}, X_{[2,0]}\right)\right] .$$

Since the field X is an isotropic GMRF with four nearest neighbors,

$$X_{[0,0]} = \theta_{[1,0]}\left(X_{[1,0]} + X_{[-1,0]} + X_{[0,1]} + X_{[0,-1]}\right) + \epsilon_{[0,0]} ,$$

where $\epsilon_{[0,0]}$ is independent from every variable $X_{[i,j]}$ with $(i,j) \neq 0$. Multiplying this identity by $X_{[1,0]}$ and taking the expectation yields

$$\text{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right) = \theta_{[1,0]}\left[\text{var}\left(X_{[0,0]}\right) + 2\text{cov}_\theta(X_{[0,0]}, X_{[1,1]}) + \text{cov}_\theta\left(X_{[0,0]}, X_{[2,0]}\right)\right] .$$

Hence, we obtain $tr\left(H^2\Sigma\right) = 4\text{cov}_\theta(X_{[0,0]}, X_{[1,0]})/\theta_{[1,0]}$ and

$$\frac{tr(H^2)}{tr(H^2\Sigma)} = \frac{\theta_{[1,0]}}{\text{cov}_\theta(X_{[0,0]}, X_{[1,0]})} ,$$

which concludes the first part of the proof.

This second part is based on the spectral representation of the field X and follows arguments which come back to Moran [Mor73]. We shall compute the limit of $\text{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right)$ when the size of Λ goes to infinity. As the field X is stationary on Λ, we may diagonalize its covariance matrix Σ applying Lemma 6.37. We note D_Σ the corresponding diagonal matrix defined by

$$D_{\Sigma[(i-1)p+j,(i-1)p+j]} = \sum_{k=1}^{p}\sum_{l=1}^{p}\text{cov}_\theta\left(X_{[0,0]}, X_{[k,l]}\right)\cos\left[2\pi\left(\frac{ki}{p} + \frac{lj}{p}\right)\right] ,$$

for any $1 \leq i,j \leq p$. Straightforwardly, we express $\text{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right)$ as a linear combination of the eigenvalues

$$\text{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right) = \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\cos\left(2\pi\frac{i}{p}\right) D_{\Sigma[(i-1)p+j,(i-1)p+j]} .$$

Applying Lemma 6.37 to the matrix Σ^{-1} and noting that $\theta \in \Theta^{\text{iso},+}$ allows to get another expression of the eigenvalues of Σ

$$D_{\Sigma[(i-1)p+j,(i-1)p+j]} = \frac{\sigma^2}{1 - 2\theta_{[1,0]}\left[\cos\left(\frac{2\pi i}{p}\right) + \cos\left(\frac{2\pi j}{p}\right)\right]} .$$

We then combine these expression. By symmetry between i and j we get

$$\text{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right) = \frac{\sigma^2}{2p^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\cos\left(2\pi\frac{i}{p}\right) + \cos\left(2\pi\frac{j}{p}\right)}{1 - 2\theta_{[1,0]}\left[\cos\left(2\pi\frac{i}{p}\right) + \cos\left(2\pi\frac{j}{p}\right)\right]} .$$

If we let p go to infinity, this sum converges to the following integral

$$\begin{aligned}\lim_{p\to+\infty}\text{cov}_\theta\left(X_{[0,0]}, X_{[1,0]}\right) &= \frac{\sigma^2}{2}\int_0^1\int_0^1\frac{\cos(2\pi x) + \cos(2\pi y)}{1 - 2\theta_{[1,0]}\left(\cos(2\pi x) + \cos(2\pi y)\right)}dxdy \\ &= \frac{\sigma^2}{2\theta_{[1,0]}}\left[-1 + \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{2\pi}\frac{1}{1 - 2\theta_{[1,0]}\left[\cos(x) + \cos(y)\right]}dxdy\right] .\end{aligned}$$

This last elliptic integral is asymptotically equivalent to $\frac{\log 16}{4(1-4\theta_{[1,0]})}$ when $\theta_{[1,0]} \to \frac{1}{4}$ as observed for instance by Moran [Mor73]. We conclude by substituting this limit in expression (6.33). □

Proof of Example 6.14. First, we compute $[\theta^{(p)}]^{\text{iso}}_{m_1}[1,0]$. By Lemma 6.6, it minimizes the function $\gamma(.)$ defined in (6.19) over the whole space $\Theta_{m_1^{\text{iso}}}$. We therefore obtain

$$[\theta^{(p)}]^{\text{iso}}_{m_1}[1,0] = \frac{tr\left[\Sigma H\right]}{tr\left[\Sigma H^2\right]} \ .$$

Once again, we apply Lemma 6.37 to simultaneously diagonalize the matrices H and Σ^{-1}. As previously, we note D_Σ the corresponding diagonal matrix of Σ.

$$\begin{aligned} D_{\Sigma[(i-1)p+j,(i-1)p+j]} &= \frac{\sigma^2}{1-2\alpha\left[\cos\left(2\pi\left(\frac{pi}{4p}+\frac{pj}{4p}\right)\right)+\cos\left(2\pi\left(\frac{-pi}{4p}+\frac{pj}{4p}\right)\right)\right]} \\ &= \frac{\sigma^2}{1-4\alpha\cos\left(\pi\frac{i}{2}\right)\cos\left(\pi\frac{j}{2}\right)} \ . \end{aligned}$$

Analogously, we compute the diagonal matrix $D\left(\Psi^{\text{iso}}_{1,0}\right)$

$$D\left(\Psi^{\text{iso}}_{1,0}\right)_{[(i-1)p+j,(i-1)p+j]} = 2\left[\cos\left(2\pi\frac{i}{p}\right)+\cos\left(2\pi\frac{j}{p}\right)\right] \ .$$

Combining these two last expressions, we obtain

$$tr(H\Sigma) = \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^2\frac{2\left[\cos\left(2\pi\frac{i}{p}\right)+\cos\left(2\pi\frac{j}{p}\right)\right]}{1-4\alpha\cos\left(\pi\frac{i}{2}\right)\cos\left(\pi\frac{j}{2}\right)} \ .$$

Let us split this sum in 16 parts depending on the congruence of i and j modulo 4. As each if of these 16 sums is shown to be zero, we conclude that $tr(H\Sigma) = [\theta^{(p)}]^{\text{iso}}_{m_1}[1,0] = 0$. By Lemma 6.36, the asymptotic risk of $\widehat{\theta^{(p)}}^{\text{iso},\rho_1}_{m_1}$ therefore equals

$$\lim_{n\to+\infty} np^2\mathbb{E}_{\theta^{(p)}}\left[l\left(\widehat{\theta^{(p)}}^{\text{iso},\rho_1}_{m_1},[\theta^{(p)}]^{\text{iso}}_{m_1}\right)\right] = \frac{tr(H^4\Sigma^2)}{tr(H^2\Sigma)} \ .$$

First, we lower bound the numerator

$$tr(H^4\Sigma^2) = \sigma^4\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\left\{2\left[\cos\left(2\pi\frac{i}{p}\right)+\cos\left(2\pi\frac{j}{p}\right)\right]\right\}^4}{\left\{1-4\alpha\cos\left(\pi\frac{i}{2}\right)\cos\left(\pi\frac{j}{2}\right)\right\}^2} \ .$$

As each term of this sum is non-negative, we may only consider the coefficients i and j which are congruent to 0 modulo 4.

$$tr(H^4\Sigma^2) \geq \sigma^4\sum_{i=0}^{p/4-1}\sum_{j=0}^{p/4-1}\frac{16\left[\cos\left(2\pi\frac{i}{p/4}\right)+\cos\left(2\pi\frac{j}{p/4}\right)\right]^4}{(1-4\alpha)^2} \ .$$

If we let go p to infinity, we get the lower bound

$$\lim_{p\to+\infty}\frac{tr(H^4\Sigma^2)}{p^2} \geq \frac{\sigma^4}{(1-4\alpha)^2}\int_0^1\int_0^1\left[\cos(2\pi x)+\cos(2\pi y)\right]^4 dxdy \ .$$

Similarly, we upper bound $tr(H^2\Sigma)$ and let p go to infinity

$$\lim_{p\to+\infty}\frac{tr(H^2\Sigma)}{p^2} \leq \frac{4\sigma^2}{1-4\alpha}\int_0^1\int_0^1\left[\cos(2\pi x)+\cos(2\pi y)\right]^2 dxdy \ .$$

Combining these two bounds allows to conclude

$$\lim_{p\to+\infty}\lim_{n\to+\infty} np^2R_{\theta^{(p)}}\left(\widehat{\theta^{(p)}}^{\text{iso},\rho_1}_{m_1},[\theta^{(p)}]^{\text{iso}}_{m_1}\right) \geq \frac{L\sigma^2}{1-4\alpha} \ .$$

□

Appendix

Lemma 6.37. *There exists an orthogonal matrix P which simultaneously diagonalizes every $p^2 \times p^2$ symmetric block circulant matrices with $p \times p$ blocks. Conversely, if θ is a square matrix of size p which satisfies (6.3), then the matrix $D(\theta) = PC(\theta)P^*$ is diagonal and satisfies*

$$D(\theta)_{[(i-1)p+j,(i-1)p+j]} = \sum_{k=1}^{p}\sum_{l=1}^{p} \theta_{[k,l]} \cos\left(2\pi(ki/p + lj/p)\right) \tag{6.86}$$

for any $1 \leq i, j \leq p$.

It is proved as in [RH05] Sect.2.6.2 to the price of a slight modification in order to take into account the fact that P has is orthogonal and not unitary. The difference comes from the fact that contrary to Rue and Held we also assume that $C(\theta)$ is symmetric.

This lemma states that all symmetric block circulant matrices are simultaneously diagonalizable. Moreover, Expression (6.86) explicitly provides the eigenvalues of the $C(\theta)$ as the two-dimensional discrete Fourier transform of the $p \times p$ matrix θ.

Proof of Lemma 6.1. Let θ be a $p \times p$ matrix that satisfies condition (6.3). For any $1 \leq i_1, i_2 \leq p$, we define the $p \times p$ submatrix C_{i_1,i_2} as

$$C_{i_1,i_2}[j_1,j_2] := C(\theta)_{[(i_1-1)p+j_1,(i_2-1)p+j_2]} \ ,$$

for any $1 \leq j_1, j_2 \leq p$. For the sake of simplicity, the subscripts (i_1, i_2) are taken modulo p. By definition of $C(\theta)$, it holds that $C_{i_1,i_2} = C_{0,i_2-i_1}$ for any $1 \leq i_1, i_2 \leq p$. Besides, the matrices $C_{0,i}$ are circulant for any $1 \leq i \leq p$. In short, the matrix $C(\theta)$ is of the form

$$C(\theta) = \begin{pmatrix} C_{0,1} & C_{0,2} & \cdots & C_{0,p} \\ \vdots & \vdots & \vdots & \vdots \\ C_{0,p} & C_{0,1} & \cdots & C_{0,p-1} \end{pmatrix},$$

where the matrices $C_{0,i}$ are circulant. Let (i_1, i_2, j_1, j_2) be in $\{1, \ldots, p\}^4$. By definition,

$$C(\theta)_{[(i_1-1)p+j_1,(i_2-1)p+j_2]} = \theta_{[i_2-i_1,j_2-j_1]} \ .$$

Since the matrix θ satisfies condition (6.3), $\theta_{[i_2-i_1,j_2-j_1]} = \theta_{[i_1-i_2,j_1-j_2]}$. As a consequence, $C(\theta)_{[(i_1-1)p+j_1,(i_2-1)p+j_2]} = C(\theta)_{[(i_2-1)p+j_2,(i_1-1)p+j_1]}$ and $C(\theta)$ is therefore symmetric.

Conversely, let B be a $p^2 \times p^2$ symmetric block circulant matrix. Let us define the matrix θ of size p by

$$\theta_{[i,j]} := B_{[1,(i-1)p+j]} \ ,$$

for any $1 \leq i, j \leq p$. Since the matrix B is block circulant, it follows that $C(\theta) = B$. By definition, $\theta_{[i,j]} = C(\theta)_{[1,(i-1)p+j]}$ and $\theta_{[-i,-j]} = C(\theta)_{[(i-1)p+j,1]}$ for any integers $1 \leq i, j \leq p$. Since the matrix B is symmetric, we conclude that $\theta_{[i,j]} = \theta_{[-i,-j]}$. □

Proof of Lemma 6.3. For any $\theta' \in \Theta^+$, $\gamma_{n,p}(\theta')$ is defined as

$$\gamma_{n,p}(\theta') = \frac{1}{p^2} tr\left[(I_{p^2} - C(\theta'))\overline{\mathbf{X}^{\mathbf{v}}\mathbf{X}^{\mathbf{v}*}}(I_{p^2} - C(\theta'))\right] \ .$$

Applying Lemma 6.37, there exists an orthogonal matrix P that simultaneously diagonalizes Σ and any matrix $C(\theta')$. Let us define $\mathbf{Y}^i := \sqrt{\Sigma}^{-1}\mathbf{X}_i$ and $D_\Sigma := P\Sigma P^*$. Gathering these new notations yields

$$\gamma_{n,p}(\theta') = \frac{1}{p^2} tr\left[\left(I_{p^2} - D(\theta')\right) D_\Sigma \overline{\mathbf{Y}\mathbf{Y}^*} \left(I_{p^2} - D(\theta')\right)\right] \ ,$$

where the vectors $\mathbf{Y}^i$ are independent standard Gaussian random vectors. Except $\overline{\mathbf{Y}\mathbf{Y}^*}$, every matrix involved in this last expression is diagonal. Besides, the diagonal matrix D_Σ is positive since Σ is non-singular. Thus, $tr\left[(I_{p^2} - D(\theta'))D_\Sigma \overline{\mathbf{Y}\mathbf{Y}^*}(I_{p^2} - D(\theta'))\right]$ is almost surely a positive quadratic form on the vector space generated by I_{p^2} and $D(\Theta^+)$. Since the function $D(.)$ is injective and linear on Θ^+, it follows that $\gamma_{n,p}(.)$ is almost surely strictly convex on Θ^+. □

Proof of Lemma 6.6 and Corollary 6.7. The proof only uses the stationarity of the field X on Λ and the l_1 norm of θ. However, the computations are a bit cumbersome. Let θ be an element of Θ^+. By standard Gaussian properties, the expectation of $X_{[0,0]}$ given the remaining covariates is

$$\mathbb{E}_\theta\left(X_{[0,0]}|X_{-\{0,0\}}\right) = \sum_{(i,j)\in\Lambda\setminus(0,0)} \theta_{[i,j]}X_{[i,j]} \ .$$

By assumption $(\mathbb{H}_2)$, the l_1 norm of θ is smaller than one. We shall prove by backward induction that for any subset A of $\Lambda\setminus\{(0,0)\}$ the matrix θ^A uniquely defined by

$$\mathbb{E}_\theta\left(X_{[0,0]}|X_A\right) = \sum_{(i,j)\in A} {\theta^A}_{[i,j]}X_{[i,j]} \text{ and } {\theta^A}_{[i,j]} = 0 \text{ for any } (i,j)\notin A$$

satisfies $\|\theta^A\|_1 \leq \|\theta\|_1$. The property is clearly true if $A = \Lambda\setminus\{(0,0)\}$. Suppose we have proved it for any set of cardinality q larger than one. Let A be a subset of $\Lambda\setminus\{(0,0)\}$ of cardinality $q-1$ and (i,j) be an element of $\Lambda\setminus(A\cup\{(0,0)\})$. Let us derive the expectation of $X_{[0,0]}$ conditionally to X_A from the expectation of $X_{[0,0]}$ conditionally to $X_{A\cup\{(i,j)\}}$.

$$\begin{aligned}\mathbb{E}_\theta\left(X_{[0,0]}|X_A\right) &= \mathbb{E}_\theta\left[\mathbb{E}(X_{[0,0]}|X_A)|X_{A\cup\{(i,j)\}}\right] \\ &= \sum_{(k,l)\in A} {\theta^{A\cup\{(i,j)\}}}_{[k,l]}X_{[k,l]} + {\theta^{A\cup\{(i,j)\}}}_{[i,j]}\mathbb{E}_\theta\left[X_{[i,j]}|X_A\right] \ . \end{aligned} \tag{6.87}$$

Let us take the conditional expectation of $X_{[i,j]}$ with respect to $X_{A\cup\{(0,0)\}}$. Since the field X is stationary on Λ and by the induction hypothesis, the unique matrix $\theta_{(i,j)}^{A\cup\{(0,0)\}}$ defined by

$$\mathbb{E}_\theta\left(X_{[i,j]}|X_{A\cup\{(0,0)\}}\right) = \sum_{(k,l)\in A\cup\{(0,0)\}} {\theta_{(i,j)}^{A\cup\{(0,0)\}}}_{[k,l]}X_{[k,l]}$$

and ${\theta_{(i,j)}^{A\cup\{(0,0)\}}}_{[k,l]} = 0$ for any $(k,l)\notin A\cup\{(0,0)\}$ satisfies $\|\theta_{(i,j)}^{A\cup\{(0,0)\}}\|_1 \leq \|\theta\|_1$. Taking the expectation conditionally to X_A of this previous expression leads to

$$\mathbb{E}_\theta\left(X_{[i,j]}|X_A\right) = \sum_{(k,l)\in A} {\theta_{(i,j)}^{A\cup\{(0,0)\}}}_{[k,l]}X_{[k,l]} + {\theta_{(i,j)}^{A\cup\{(0,0)\}}}_{[0,0]}\mathbb{E}\left(X_{[0,0]}|X_A\right) \ . \tag{6.88}$$

Gathering identities (6.87) and (6.88) yields

$$\mathbb{E}_\theta\left(X_{[0,0]}|X_A\right) = \sum_{(k,l)\in A} \frac{{\theta^{A\cup\{i,j\}}}_{[k,l]} + {\theta^{A\cup\{(i,j)\}}}_{[i,j]}{\theta_{(i,j)}^{A\cup\{0,0\}}}_{[k,l]}}{1 - {\theta^{A\cup\{(i,j)\}}}_{[i,j]}{\theta_{(i,j)}^{A\cup\{0,0\}}}_{[0,0]}}X_{[k,l]} \ ,$$

since $\left|{\theta^{A\cup\{i,j\}}}_{[i,j]}{\theta^{i,j}_{A\cup\{0,0\}}}_{[0,0]}\right| < 1$. Then, we upper bound the l_1 norm of θ^A using that $\|\theta^{A\cup\{(i,j)\}}\|_1$ and

$\|\theta^{A\cup\{(0,0)\}}_{(i,j)}\|_1$ are smaller or equal to $\|\theta\|_1$.

$$\begin{aligned}
\|\theta^A\|_1 &\leq \frac{1}{1-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|}\left(\sum_{(k,l)\in A}\left|\theta^{A\cup\{j+1\}}{}_{[k,l]}\right|+\sum_{(k,l)\in A}\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{0,0\}}_{(i,j)}{}_{[k,l]}\right|\right)\\
&\leq \frac{\|\theta\|_1+\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\right|\left(\sum_{(k,l)\in A\cup\{(0,0)\}}\left|\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[k,l]}\right|-1-\left|\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|\right)}{1-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{(0,0)\}}_{i,j}{}_{[0,0]}\right|}\\
&\leq \frac{\|\theta\|_1\left(1+\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\right|\right)-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\right|\left(1+\left|\theta^{A\cup\{(0,0)\}}_{i,j}{}_{[0,0]}\right|\right)}{1-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|}\\
&\leq \|\theta\|_1+\frac{\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\right|\left(\|\theta\|_1-1\right)\left(1+\left|\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|\right)}{1-\left|\theta^{A\cup\{(i,j)\}}{}_{[i,j]}\theta^{A\cup\{(0,0)\}}_{(i,j)}{}_{[0,0]}\right|}\ .
\end{aligned}$$

Since $\|\theta\|_1$ is smaller than one, it follows that $\|\theta^A\|_1\leq\|\theta\|_1$.

Let m be a model in the collection $\mathcal{M}_1$. Since m stands for a set of neighbors of $(0,0)$, we may define θ^m as above. It follows that $\|\theta^m\|_1\leq\|\theta\|_1$. Since the field X is stationary on the torus, X follows the same distribution as the field X^s defined by $X^s{}_{[i,j]}=X_{[-i,-j]}$. By uniqueness of θ^m, we obtain that $\theta^m{}_{[i,j]}=\theta^m{}_{[-i,-j]}$. Thus, θ^m belongs to the space Θ_m. Moreover, θ^m minimizes the function $\gamma(.)$ on Θ_m. Since the l_1 norm of θ^m is smaller than one, θ^m belongs to $\Theta^+_{m,2}$. The matrices θ^m and θ_{m,ρ_1} are therefore equal, which concludes the proof in the non-isotropic case.

Let us now turn to the isotropic case. Let θ belong to $\Theta^{\mathrm{iso},+}$ and let m be a model in $\mathcal{M}_1$. As previously, the matrix θ^m satisfies $\|\theta^m\|_1\leq\|\theta\|_1$. Since the distribution of X is invariant under the action of the group G, θ^m belongs to Θ^{iso}_m. Since $\|\theta^m\|_1\leq\|\theta\|_1$, θ^m lies in $\Theta^{+,\mathrm{iso}}_{m,2}$

It follows that $\theta^m=\theta_m$. □

Proof of Corollary 6.8. Let θ be a matrix in Θ^+ such that $(\mathbb{H}_2)$ holds and let m be a model in $\mathcal{M}_1$. We decompose $\gamma(\widehat{\theta}_{m,\rho_1})$ using the conditional expectation of $X_{[0,0]}$ given X_m.

$$\begin{aligned}
\gamma(\widehat{\theta}_{m,\rho_1}) &= \mathbb{E}_\theta\left[X_{[0,0]}-\sum_{(i,j)\in m}\widehat{\theta}_{m,\rho_1}{}_{[i,j]}X_{[i,j]}\right]^2\\
&= \mathbb{E}_\theta\left[X_{[0,0]}-\mathbb{E}_\theta\left(X_{[0,0]}\,|X_m\right)\right]^2+\mathbb{E}_\theta\left[\mathbb{E}_\theta\left(X_{[0,0]}\,|X_m\right)-\sum_{(i,j)\in m}\widehat{\theta}_{m,\rho_1}{}_{[i,j]}X_{[i,j]}\right]^2.
\end{aligned}$$

By Corollary 6.7, we know that

$$\mathbb{E}_\theta\left(X_{[0,0]}\,|X_m\right)=\sum_{(i,j)\in m}\theta_{m,\rho_1}{}_{[i,j]}X_{[i,j]}\ .$$

Combining these two last identities yields

$$\gamma(\widehat{\theta}_{m,\rho_1}) = \gamma(\theta_{m,\rho_1})+\mathbb{E}_\theta\left[\sum_{(i,j)\in\Lambda\setminus\{(0,0)\}}\left(\theta_{m,\rho_1}-\widehat{\theta}_{m,\rho_1}\right)_{[i,j]}X_{[i,j]}\right]^2.$$

Subtracting $\gamma(\theta)$, we obtain the first result. The proof is analogous in the isotropic case. □

Chapter 7

Data-driven neighborhood selection of a Gaussian field

Abstract. We study the non-parametric covariance estimation of a stationary Gaussian field X observed on a lattice. In a Chapter 6, we have introduced a model selection procedure to tackle this issue. It amounts to selecting a neighborhood $\widehat{m}$ and estimating the covariance of X in the space of Gaussian Markov random fields (GMRFs) with neighborhood $\widehat{m}$. This strategy is shown to satisfy oracle inequalities as well as minimax adaptive properties. However, it suffers several drawbacks which makes it difficult to apply in practice. First, the penalty depends on some unknown quantities. Second, the procedure is only defined for toroidal lattices, which it unrealistic for many applications. In this chapter, we slightly modify the procedure to handle these problems. On the one hand, we propose a data-driven algorithm for tuning the penalty function. On the other hand, we extend the procedure to non-toroidal lattices. We then study the numerical performances of this new method on simulated examples. This suggests that Gaussian Markov random field selection is sometimes a good alternative to variogram estimation.

7.1 Introduction

In this chapter, we study the estimation of the distribution of a stationary Gaussian field $(X_{[i,j]})_{(i,j)\in\Lambda}$ indexed by the nodes of a rectangular lattice Λ of size $p_1 \times p_2$. This problem is often encountered in spatial statistics or in image analysis. Most statistical procedures amount to first subtracting a parametric form of the mean value. Henceforth, we assume that the field X is centered. Given a n-sample of the field X, the challenge is to infer the covariance of the field. In practice, the number n of observations often equals one. Different methods have been proposed to tackle this problem.

A first approach amounts to computing an empirical variogram and then fitting a suitable parametric variogram model such as the exponential or Matérn model (see [Cre93] Ch.2 for more details). It such methods perform well if one has chosen a *good* variogram model, it is not adaptive to any form of correlation.

A second approach amounts to estimating the covariance of X in a Gaussian Markov random field (GMRF) setting. Let us fix a node $(0,0)$ at the center of Λ and let m be a subset of $\Lambda \setminus \{(0,0)\}$. The field X is a GMRF with respect to the neighborhood m if conditionally to $(X_{[k,l]})_{(k,l)\in m}$, the variable $X_{[0,0]}$ is independent from all the remaining variables in Λ. We refer to Rue and Held [RH05] for a comprehensive introduction on GMRFs. If we know that X is a GMRF with respect to the neighborhood m, then we can estimate the covariance by applying likelihood or pseudolikelihood maximization. These parametric procedures are well understood, at least from an asymptotic point of view (see for instance [Guy95] Sect.4). However, we do not know in practice which neighborhood m we should choose. For instance, choosing the empty neighborhood amounts to assuming that all the components of X are independent. Alternatively, if we choose the complete neighborhood, which contains all the nodes of Λ except $(0,0)$,

then the number of parameters is huge and estimation performances are poor.

In this chapter, we tackle the problem of neighborhood selection from a practical point of view. The purpose is to define a data-driven procedure that picks a suitable neighborhood $\widehat{m}$ and then estimates the distribution of X in the space of GMRFs with neighborhood $\widehat{m}$. This procedure should not require any knowledge on the distribution of X. Moreover, it has to achieve a fast rate of convergence if the distribution of X is a GMRF with respect to a small neighborhood. In statistical terms, the procedure has to be non-parametric and adaptive.

Besag and Kooperberg [BK95], Rue and Tjelmeland [RT02], and Cressie and Verzelen [CV08] have considered the problem of *approximating* the distribution of a Gaussian field by a GMRF, but this requires the knowledge of the true distribution. Guyon and Yao have stated in [GY99] necessary conditions and sufficient conditions for a model selection procedure to choose asymptotically the true neighborhood of a GMRF with probability one. Our point of view is slightly different: we do not assume that the field X is a GMRF with respect to a sparse neighborhood and do not aim at estimating the true neighborhood, we rather want to select a neighborhood that allows to estimate *well* the distribution of X. The distinction between these two points of view has been nicely described in the first chapter of MacQuarrie and Tsai [MT98].

In Chapter 6, we have introduced a neighborhood selection procedure based on pseudolikelihood maximization and penalization. Under mild assumptions, the procedure achieves optimal neighborhood selection. More precisely, it satisfies an oracle inequality and it is minimax adaptive to the sparsity of the neighborhood. To our knowledge, these are the first results of neighborhood selection in this spatial setting.

If the procedure exhibits appealing theoretical properties, it suffers several drawbacks from a practical perspective. First, the method constrains the largest eigenvalue of the estimated covariance to be smaller than some parameter ρ. In practice, it is difficult to choose ρ since we do not know the largest eigenvalue of the true covariance. Second, the penalty function pen(.) introduced in Section 6.3 depends on the largest eigenvalue of the covariance of the field X. Hence, we need a practical method for tuning the penalty. Third, in Chapter 6 we have only defined the procedure when the lattice Λ is a square torus.

Our contribution is twofold. On the one hand, we propose practical versions of our neighborhood selection procedure that overcome the previously-mentioned drawbacks:

- The procedure is extended to rectangular lattices.
- We do not constrain anymore the largest eigenvalue of the covariance.
- We provide an algorithm based on the so-called *slope heuristics* of Birgé and Massart [BM07] for tuning the penalty. Theoretical justifications for its use are also given.
- Finally, we extend the procedure to the case where the lattice Λ is not a torus.

On the other hand, we illustrate the performances of this new procedure on numerical examples. When Λ is a torus, we compare it with likelihood-based methods like AIC [Aka73] and BIC [Sch78], even if they were not studied in this setting. When Λ is not toroidal, likelihood methods become intractable and compare our procedure with variogram-based methods.

The chapter is organized as follows. In Section 7.2, we define a new version of the estimation procedure of Chapter 6 that does not require anymore the choice of the constant ρ. We discuss its computational complexity. In Section 7.3, we connect this new procedure to the method introduced in Chapter 6 and we recall some theoretical results. We provide an algorithm for tuning the penalty in practice in Section 7.4. In Section 7.5, we extend our procedure for handling non-toroidal lattices. This simulation studies are provided in Section 7.6, while the proofs are postponed to Section 7.7.

Let us introduce some notations. In the sequel, X^v refers to the vectorialized version of X with the convention $X_{[i,j]} = X^v_{[(i-1)\times p_1 + j]}$ for any $1 \leq i \leq p_1$ and $1 \leq j \leq p_2$. Using this new notation amounts to "forgetting" the spatial structure of X and allows to get into a more classical statistical framework. We

note $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ the n observations of the field X. The matrix Σ stands for the covariance matrix of X^v. For any matrix A, $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$ respectively refer the largest eigenvalue and the smallest eigenvalues of A. Finally, I_r denotes the identity matrix of size r.

7.2 Neighborhood selection on a torus

In this section, we introduce the main concepts and notations for GMRFs on a torus. Afterwards, we describe our procedure based on pseudolikelihood maximization. Finally, we discuss some computational aspects. Throughout this section and the two following sections, the lattice Λ is assumed to be toroidal. Consequently, the components of the matrices X are taken modulo p_1 and p_2.

7.2.1 GMRFs on the torus

The notion of conditional distribution is underlying the definition of GMRFs. By standard Gaussian derivations (see for instance [Lau96] App.C), there exists a unique $p \times p$ matrix θ such that $\theta_{[0,0]} = 0$ and

$$X_{[0,0]} = \sum_{(i,j)\in\Lambda\setminus\{(0,0)\}} \theta_{[i,j]} X_{[i,j]} + \epsilon_{[0,0]} \ , \tag{7.1}$$

where the random variable $\epsilon_{[0,0]}$ follows a zero-mean normal distribution and is independent from the covariates $(X_{[i,j]})_{(i,j)\in\Lambda\setminus\{(0,0)\}}$. In the sequel, we note σ^2 the variance of $\epsilon_{[0,0]}$ and we call it the conditional variance of $X_{[0,0]}$.

Equation (7.1) describes the conditional distribution of $X_{[0,0]}$ given the remaining variables. By stationarity of the field X, it holds that that $\theta_{[i,j]} = \theta_{[-i,-j]}$. The covariance matrix Σ is closely related to θ through the following equation:

$$\Sigma = \sigma^2 \left[I_{p_1 p_2} - C(\theta) \right]^{-1} \ , \tag{7.2}$$

where the $p^2 \times p^2$ matrix $C(\theta)$ is defined as $C(\theta)_{[i_1(p-1)+j_1, i_2(p-1)+j_2]} := \theta_{[i_2-i_1, j_2-j_1]}$ for any $1 \leq i_1, i_2, j_1, j_2 \leq p$. The matrix $(I_{p_1 p_2} - C(\theta))$ is called the partial correlation matrix of the field X. The so-defined matrix $C(\theta)$ is symmetric block circulant with $p \times p$ blocks. We refer to [RH05] Sect.2.6 or the book of Gray [Gra06] for definitions and main properties on circulant and block circulant matrices.

There are two main consequences. First, estimating the $p_1 \times p_2$ matrix θ amounts to estimating the covariance matrix Σ up to a multiplicative constant. We shall therefore focus on θ. Second, by Equation (7.1), the field X is a GMRF whose neighborhood is support of θ. The adaptive estimation of the distribution of X by neighborhood selection therefore reformulates as an adaptive estimation problem of the matrix θ.via support selection.

Let us now precise the set of possible values for θ. The set Θ denotes the vector space of the $p_1 \times p_2$ matrices that satisfy $\theta_{[0,0]} = 0$ and $\theta_{[i,j]} = \theta_{[-i,-j]}$, for any $(i,j) \in \Lambda$. Hence, a matrix $\theta \in \Theta$ corresponds to the distribution of a stationary Gaussian field if and only if the $p_1 p_2 \times p_1 p_2$ matrix $(I_{p_1 p_2} - C(\theta))$ is positive definite. This is why we define the convex subset Θ^+ of Θ by

$$\Theta^+ := \{\theta \in \Theta \text{ s.t. } [I_{p_1 p_2} - C(\theta)] \text{ is positive definite}\} \ . \tag{7.3}$$

The set of covariance matrices of stationary Gaussian fields on Λ with unit conditional variance is in one to one correspondence with the set Θ^+. We sometimes assume that the field X is isotropic. The corresponding sets Θ^{iso} and $\Theta^{+,\text{iso}}$ for isotropic fields are introduced as:

$$\Theta^{\text{iso}} := \{\theta \in \Theta \ , \ \theta_{[i,j]} = \theta_{[-i,j]} = \theta_{[j,i]} \ , \forall (i,j) \in \Lambda\} \quad \text{and } \Theta^{+,\text{iso}} := \Theta^+ \cap \Theta^{\text{iso}} \ .$$

Finally, we consider the toroidal norm $|(i,j)|_t$ is defined by

$$|(i,j)|_t^2 := [i \wedge (p_1 - i)]^2 + [j \wedge (p_2 - j)]^2 \ ,$$

for any node $(i,j) \in \Lambda$.

7.2.2 Description of the procedure

In the sequel, a model m stands for a subset of $\Lambda \setminus \{(0,0)\}$. It is also called a neighborhood. For the sake of simplicity, we shall only use the collection of models $\mathcal{M}_1$ defined below.

Definition 7.1. A subset $m \subset \Lambda \setminus \{(0,0)\}$ belongs to $\mathcal{M}_1$ if and only if there exists a number $r_m > 1$ such that

$$m = \{(i,j) \in \Lambda \setminus \{(0,0)\} \quad \text{s.t.} \quad |(i,j)|_t \leq r_m\} \ . \tag{7.4}$$

In other words, the neighborhoods m in $\mathcal{M}_1$ are sets of nodes lying in a disc centered at $(0,0)$. Obviously, $\mathcal{M}_1$ is totally ordered with respect to the inclusion. Consequently, we order the models $m_0 \subset m_1 \subset \ldots \subset m_i \ldots$. For instance, m_0 corresponds to the empty neighborhood, m_1 stands for the neighborhood of size 4, and m_2 refers to the neighborhood with 8 neighbours. See Figure 6.1 in Chapter 6 for other examples.

For any model $m \in \mathcal{M}_1$, the vector space Θ_m is the subset of matrices Θ whose support is included in m. Similarly Θ_m^{iso} is the subset of Θ^{iso} whose support is included in m. The dimensions of Θ_m and Θ_m^{iso} are respectively noted d_m and d_m^{iso}. Since we aim at estimating the positive matrix $(I_{p_1 p_2} - C(\theta))$, we also consider the convex subsets of Θ_m^+ and $\Theta_m^{+,\text{iso}}$ which correspond to non-negative matrices precision matrices.

$$\Theta_m^+ := \Theta_m \cap \Theta^+ \qquad \text{and} \qquad \Theta_m^{+,\text{iso}} := \Theta_m^{\text{iso}} \cap \Theta^{+,\text{iso}} \ . \tag{7.5}$$

For any $\theta' \in \Theta^+$, the conditional least-squares (CLS) criterion ([Guy95]) $\gamma_{n,p_1,p_2}(\theta')$ is defined by

$$\gamma_{n,p_1,p_2}(\theta') \quad := \quad \frac{1}{np_1p_2} \sum_{i=1}^{n} \sum_{(j_1,j_2)\in\Lambda} \left(\mathbf{X}_{i[j_1,j_2]} - \sum_{(l_1,l_2)\in\Lambda\setminus\{(0,0)\}} \theta'_{[l_1,l_2]} \mathbf{X}_{i[j_1+l_1,j_2+l_2]} \right)^2 \ . \tag{7.6}$$

The function $\gamma_{n,p_1,p_2}(.)$ is a least squares criterion that allows us to perform the simultaneous linear regression of all $\mathbf{X}_{i[j_1,j_2]}$ with respect to the covariates $(\mathbf{X}_{i[l_1,l_2]})_{(l_1,l_2)\neq(k_1,k_2)}$. The advantage of this criterion is that it does not require the computation of a determinant as for the likelihood. We explain in Section 6.2 its connection with the pseudolikelihood introduced by [Bes75]. For any model $m \in \mathcal{M}_1$, the estimators are defined as the unique minimizers of $\gamma_{n,p_1,p_2}(.)$ on the sets Θ_m^+ and $\Theta_m^{+,\text{iso}}$.

$$\widehat{\theta}_m := \arg \min_{\theta' \in \overline{\Theta_m^+}} \gamma_{n,p_1,p_2}(\theta') \qquad \text{and} \qquad \widehat{\theta}_m^{\text{iso}} := \arg \min_{\theta' \in \overline{\Theta_m^{+,\text{iso}}}} \gamma_{n,p_1,p_2}(\theta') \ , \tag{7.7}$$

where $\overline{A}$ stands for the closure of A. Let us mention that the estimator $\widehat{\theta}_m$ corresponds to the estimator $\widehat{\theta}_{m,\rho_1}$ defined in Section 6.2.2 with $\rho_1 = +\infty$. We further discuss the connection between $\widehat{\theta}_m$ and $\widehat{\theta}_{m,\rho_1}$ in Section 7.3.

Given a subcollection of models $\mathcal{M}$ of $\mathcal{M}_1$ and a positive function pen : $\mathcal{M} \to \mathbb{R}^+$ called a penalty, we select a model as follows:

$$\widehat{m} := \arg \min_{m\in\mathcal{M}} \left[\gamma_{n,p_1,p_2}\left(\widehat{\theta}_m\right) + \text{pen}(m) \right] \quad \text{and} \quad \widehat{m}^{\text{iso}} := \arg \min_{m\in\mathcal{M}} \left[\gamma_{n,p_1,p_2}\left(\widehat{\theta}_m^{\text{iso}}\right) + \text{pen}(m) \right] \ . \tag{7.8}$$

For short, we write $\widetilde{\theta}$ and $\widetilde{\theta}^{\text{iso}}$ for $\widehat{\theta}_{\widehat{m}}$ and $\widehat{\theta}^{\text{iso}}_{\widehat{m}^{\text{iso}}}$. We discuss the choice of the penalty function in Section 7.4.

7.2.3 Computational aspects

Since the lattice Λ is a torus, the computation of the estimators $\widehat{\theta}_m$ is performed efficiently thanks to the following lemma.

Lemma 7.2. *For any $p \times p$ matrix A and for any $1 \leq i \leq p_1$ and $1 \leq j \leq p_2$, $\lambda_{[i,j]}(A)$ is the (i,j)-th term of two-dimensional discrete Fourier transform of the matrix A:*

$$\lambda_{[i,j]}(A) \quad := \quad \sum_{k=1}^{p_1} \sum_{l=1}^{p_2} A_{[i,j]} \exp\left[2\iota\pi \left(\frac{ki}{p_1} + \frac{jl}{p_2} \right) \right] \ , \tag{7.9}$$

where $\iota^2 = -1$. The conditional least-squares criterion $\gamma_{n,p_1,p_2}(\theta')$ simplifies as

$$\gamma_{n,p_1,p_2}(\theta') = \frac{1}{np_1^2p_2^2}\left\{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\left[1-\lambda_{[i,j]}(\theta)\right]^2\left[\sum_{k=1}^{n}\lambda_{[i,j]}\left(\mathbf{X}_k\right)\overline{\lambda_{[i,j]}\left(\mathbf{X}_k\right)}\right]\right\} .$$

A proof is given in Section 7.7. Optimization of $\gamma_{n,p_1,p_2}(.)$ over the set Θ_m^+ is performed fastly using the Fast Fourier transform (FFT). Nevertheless, this is not the privilege of CLS estimators, since maximum likelihood estimators are also computed fastly by FFT when Λ is a torus.

In Section 7.5, we shall mention that the computation of the CLS estimators $\widehat{\theta}_m$ remains quite easy when Λ is not a torus whereas likelihood maximization becomes intractable.

7.3 Theoretical results

Throughout this section, Λ is assumed to be square lattice and we note p its size. Let us mention that the restriction to square lattices made in Chapter 6 allows to simplify the proofs but is not necessary so that the theoretical results hold. In this section, we first recall the procedure introduced in Chapter 6 and we emphasize the differences with the one defined in the previous section. We also mention a result of optimality stated in the previous chapter. This will provide some insights for calibrating the penalty pen(.) in Section 7.4.

Given $\rho > 2$ be a positive constant, we define the subsets $\Theta_{m,\rho}^+$ and $\Theta_{m,\rho}^{+,\mathrm{iso}}$ by

$$\begin{aligned}\Theta_{m,\rho}^+ &:= \left\{\theta\in\Theta_m^+ \ , \ \varphi_{\max}\left[I_{p_1p_2}-C(\theta)\right]<\rho\right\} \\ \Theta_{m,\rho}^{+,\mathrm{iso}} &:= \left\{\theta\in\Theta_m^{+,\mathrm{iso}} \ , \ \varphi_{\max}\left[I_{p_1p_2}-C(\theta)\right]<\rho\right\} .\end{aligned} \tag{7.10}$$

Then, the corresponding estimators $\widehat{\theta}_{m,\rho}$ and $\widehat{\theta}_{m,\rho}^{\mathrm{iso}}$ are defined as in (7.7), except that we now consider $\Theta_{m,\rho}^+$ instead of Θ_m^+.

$$\widehat{\theta}_{m,\rho} := \arg\min_{\theta'\in\overline{\Theta_{m,\rho}^+}}\gamma_{n,p,p}(\theta') \qquad \text{and} \qquad \widehat{\theta}_{m,\rho}^{\mathrm{iso}} := \arg\min_{\theta'\in\overline{\Theta_{m,\rho}^{+,\mathrm{iso}}}}\gamma_{n,p,p}(\theta') .$$

Given a subcollection $\mathcal{M}$ of $\mathcal{M}_1$ and a penalty function pen(.), we select the models $\widehat{m}_\rho$ and $\widehat{m}_\rho^{\mathrm{iso}}$ as in (7.8) except that we use $\widehat{\theta}_{m,\rho}$ and $\widehat{\theta}_{m,\rho}^{\mathrm{iso}}$ instead of $\widehat{\theta}_m$ and $\widehat{\theta}_m^{\mathrm{iso}}$. We also note $\widetilde{\theta}_\rho$ and $\widetilde{\theta}_\rho^{\mathrm{iso}}$ for $\widehat{\theta}_{\widehat{m}_\rho,\rho}$ and $\widehat{\theta}_{\widehat{m}_\rho^{\mathrm{iso}},\rho}^{\mathrm{iso}}$.

The only difference between the estimators $\widetilde{\theta}$ and $\widetilde{\theta}_\rho$ is that the largest eigenvalue of the precision matrix $(I_{p^2}-C(\widetilde{\theta}))$ is restricted to be smaller than ρ. We make this restriction in Chapter 6 to facilitate the analysis.

In order to assess the performance of the penalized estimator $\widetilde{\theta}_\rho$ and $\widetilde{\theta}_\rho^{\mathrm{iso}}$, we use the prediction loss function $l(\theta_1,\theta_2)$ defined by

$$l(\theta_1,\theta_2) := \frac{1}{p^2}tr\left[(C(\theta_1)-C(\theta_2))\Sigma(C(\theta_1)-C(\theta_2))\right] . \tag{7.11}$$

As explained in Section 6.1.3, the loss $l(\theta_1,\theta_2)$ expresses in terms of conditional expectation

$$l(\theta_1,\theta_2) = \mathbb{E}_\theta\left\{\left[\mathbb{E}_{\theta_1}\left(X_{[0,0]}|X_{\Lambda\setminus\{0,0\}}\right)-\mathbb{E}_{\theta_2}\left(X_{[0,0]}|X_{\Lambda\setminus\{0,0\}}\right)\right]^2\right\} , \tag{7.12}$$

where $\mathbb{E}_\theta(.)$ stands for the expectation with respect to the distribution $\mathcal{N}(0,\sigma^2(I_{p_1p_2}-C(\theta))^{-1})$. Hence, $l(\widehat{\theta},\theta)$ corresponds the mean squared prediction loss of $X_{[0,0]}$ given the other covariates. For any neighborhood $m\in\mathcal{M}$, we define the *projection* $\theta_{m,\rho}$ as the closest element of θ in $\Theta_{m,\rho}^+$ with respect to the loss $l(.,.)$.

$$\theta_{m,\rho} := \arg\min_{\theta'\in\overline{\Theta_{m,\rho}^+}} l(\theta',\theta) \qquad \text{and} \qquad \theta_{m,\rho}^{\mathrm{iso}} := \arg\min_{\theta'\in\overline{\Theta_{m,\rho}^{+,\mathrm{iso}}}} l(\theta',\theta) .$$

We call the loss $l(\theta_{m,\rho},\theta)$ the bias of the set $\Theta_{m,\rho}^+$. This implies that the $\widehat{\theta}_{m,\rho}$ cannot perform better than this loss.

Theorem 7.3. *Let $\rho > 2$, K be a positive number larger than an universal constant K_0 and $\mathcal{M}$ be a subcollection of $\mathcal{M}_1$. If every model $m \in \mathcal{M}$, it holds that*

$$pen(m) \geq K\rho^2 \varphi_{max}(\Sigma) \frac{d_m + 1}{np^2} \ , \tag{7.13}$$

then for any $\theta \in \Theta^+$, the estimator $\widetilde{\theta}_\rho$ satisfies

$$\mathbb{E}_\theta[l(\widetilde{\theta}_\rho, \theta)] \leq L(K) \inf_{m \in \mathcal{M}} [l(\theta_{m,\rho}, \theta) + pen(m)] \ , \tag{7.14}$$

where $L(K)$ only depends on K. A similar bound holds if one replaces $\widetilde{\theta}_\rho$ by $\widetilde{\theta}^{\text{iso}}_\rho$, Θ^+ by $\Theta^{+,iso}$, $\theta_{m,\rho}$ by $\theta^{\text{iso}}_{m,\rho}$, and d_m by d^{iso}_m.

This nonasymptotic bound is provided in a slightly different version in Chapter 6. It states that $\widetilde{\theta}_\rho$ achieves a trade-off between the bias and a variance term if the penalty is suitable chosen. In Theorem 7.3, we use the penalty $K\rho^2\varphi_{\max}(\Sigma)\frac{d_m+1}{np^2}$ instead of the penalty $K\rho^2\varphi_{\max}(\Sigma)\frac{d_m}{np^2}$ stated in Chapter 6. This makes the bound (7.14) simpler. Observe that these two penalties yield the same model selection since they only differ by a constant. Let us further discuss two points.

- In this chapter, we use the estimator $\widetilde{\theta}$ rather than $\widetilde{\theta}_\rho$. Given a collection of models $\mathcal{M}$, there exists some finite $\rho > 2$, such that these two estimators coincide. Take for instance $\rho = \sup_{m \in \mathcal{M}} \sup_{\theta \in \Theta^+_m} \varphi_{\max}(I_{p_1 p_2} - C(\theta))$. Admittedly, the so-obtained ρ may be large, especially if there are large models in $\mathcal{M}$. The upper bound (7.14) on the risk therefore becomes worse. Nevertheless, we do not think that the dependency of (7.14) on ρ is sharp. Indeed , we illustrate in Section 7.6 that the risk of $\widetilde{\theta}$ exhibits good statistical performances..
- Theorem 7.3 provides a suitable form of the penalty for obtaining oracle inequalities. However, this penalty depends on $\varphi_{\max}(\Sigma)$ which is not known in practice. This is why we develop a data-driven penalization method in the next section.

7.4 Slope Heuristics

Let us introduce a data-driven method for calibrating the penalty function pen(.). It is based on the so-called *slope heuristic* introduced by Birgé and Massart [BM07] in the fixed design Gaussian regression framework (see also [Mas07] Sect.8.5.2). This heuristic relies on the notion of minimal penalty. In short, assume that one knows that a good penalty has a form $\text{pen}(m) = NF(d_m)$ (where d_m is the dimension of the model and N is a tuning parameter). Let us define $\widehat{m}(N)$ the selected model as a function of N. There exists a quantity $\widehat{N}_{\min}$ satisfying the following property: If $N > \widehat{N}_{\min}$, the dimension of the selected model $d_{\widehat{m}(N)}$ is reasonable and if $N < \widehat{N}_{\min}$, the dimension of the selected model is huge. The function $pen_{\min}(.) := \widehat{N}_{\min} F(.)$ is called the minimal penalty. In fact, a *dimension jump* occurs for $d_{\widehat{m}(N)}$ at the point $\widehat{N}_{\min}$. Thus, the quantity $\widehat{N}_{\min}$ is clearly observable for real data sets. In their Gaussian framework, Birgé and Massart have shown that twice the minimal penalty is nearly the optimal penalty. In other words, the model $\widehat{m} := \widehat{m}(2\widehat{N}_{\min})$ yields an efficient estimator.

The slope heuristic method has been successfully applied for multiple change-point detection [Leb05]. Applications are also being developed in other frameworks such as mixture models [MM08], clustering [BCM08], estimation of oil reserves [Lep02], and genomic [Vil07].

If this method was originally introduced for fixed design Gaussian regression, Arlot and Massart [AM08] have proved more recently that a similar phenomenon occurs in the heteroscedastic random-design case. In the GMRF setting, we are only able to partially justify this heuristic. For the sake of simplicity, let us assume in the next proposition that the lattice Λ is a square of size p.

Proposition 7.4. *Consider $\rho > 2$, and $\eta < 1$ and suppose that p is larger than some numerical constant p_0. Let m' be the largest model in $\mathcal{M}_1$ that satisfies $d_{m'} \leq \sqrt{np^2}$.*

For any model $m \in \mathcal{M}_1$, we assume that

$$pen(m') - pen(m) \leq K_1(1-\eta)\sigma^2 \left\{ \varphi_{min}\left(I_{p^2} - C(\theta)\right) \wedge \left[\rho - \varphi_{max}\left(I_{p^2} - C(\theta)\right)\right] \right\} \frac{d_{m'} - d_m}{np^2} \ , \tag{7.15}$$

where K_1 is a universal constant defined in the proof. Then, for any $\theta \in \Theta^+_{m',\rho}$, it holds that

$$\mathbb{P}\left\{ d_{\widehat{m}_\rho} > L\left[\sqrt{np^2} \wedge p^2\right]\right\} \geq \frac{1}{2} \ ,$$

where L only depends on η, ρ, $\varphi_{min}\left(I_{p^2} - C(\theta)\right)$, and $\varphi_{max}\left(I_{p^2} - C(\theta)\right)$.

The proof is postponed to Section 7.7. Let us define

$$N_1 := K_1\sigma^2 \left\{\varphi_{\min}\left(I_{p_1p_2} - C(\theta)\right) \wedge \left[\rho - \varphi_{\max}\left(I_{p_1p_2} - C(\theta)\right)\right]\right\} \ ,$$

and let us consider penalty functions $\mathrm{pen}(m) = N\frac{d_m}{np_1p_2}$ for some $N > 0$. The proposition states that if N is smaller than N_1, then the procedure selects a model of huge dimension with large probability, i.e $d_{\widehat{m}(N)}$ is huge. Alternatively, let us define

$$N_2 := K_0 \frac{\sigma^2\rho^2}{\varphi_{\min}\left(I_{p_1p_2} - C(\theta)\right)} \frac{d_m}{np_1p_2} \ ,$$

where the numerical constant K_0 is introduced in Theorem 6.4. By this theorem, choosing $N > N_2$ ensures that the risk of $\widetilde{\theta}_\rho$ achieves a type-oracle inequality and the dimension $d_{\widehat{m}_\rho(N)}$ is reasonable. The quantities N_1 and N_2 are different especially when the eigenvalues of $(I_{p_1p_2} - C(\theta))$ are far from 1. Since we do not know the behavior of the selected model $\widehat{m}_\rho(N)$ when N is between N_1 and N_2, we are not able to really prove a dimension jump as the fixed design Gaussian regression framework. Besides, we have mentioned in the preceding section that we are more interested to the estimator $\widetilde{\theta}$ than $\widetilde{\theta}_\rho$. Nevertheless, we clearly observe in simulation studies a dimension jump for some N between N_1 and N_2 even if we use the estimators $\widehat{\theta}_m$ instead of $\widehat{\theta}_{m,\rho}$. This lets us think that the slope heuristic is still valid in the GMRF framework.

Algorithm 7.5. *(Data-driven penalization with slope heuristic). Let $\mathcal{M}$ be a subcollection of $\mathcal{M}_1$.*

1. *Compute the selected model $\widehat{m}(N)$ as a function of $N > 0$*

$$\widehat{m}(N) \in \arg\min_{m\in\mathcal{M}} \left\{ \gamma_{n,p_1,p_2}\left(\widehat{\theta}_m\right) + N\frac{d_m}{np_1p_2}\right\} \ .$$

2. *Find $\widehat{N}_{\min} > 0$ such that the jump $d_{\widehat{m}\left(\left[\widehat{N}_{\min}\right]_-\right)} - d_{\widehat{m}\left(\left[\widehat{N}_{\min}\right]_+\right)}$ is maximal.*

3. *Select the model $\widehat{m} = \widehat{m}(2\widehat{N}_{\min})$.*

The difference $f(x_-) - f(x_+)$ measures the discontinuity of a function f at the point x. Step 2 may need to introduce huge models in the collection $\mathcal{M}$ all the other ones being considered as "reasonably small". As the function $\widehat{m}(.)$ is piecewise linear with at most $\mathrm{Card}(\mathcal{M})$ jumps, so that steps 1-2 have a complexity $\mathcal{O}\left(\mathrm{Card}(\mathcal{M})\right)^2$. We refer to App.A.1 of [AM08] for more details on the computational aspects of steps 1 and 2. Let us mention that there are other ways of estimating $\widehat{N}_{\min}$ than choosing the largest jump as described in [AM08] App.A.2. Finally, the methodology described in this section straightforwardly extends to the case of isotropic GMRFs estimation by replacing $\widehat{m}(N)$ by $\widehat{m}^{\mathrm{iso}}(N)$ and d_m by d_m^{iso}.

Algorithm 7.6. *(Data-driven penalization for isotropic models). Let $\mathcal{M}$ be a subcollection of $\mathcal{M}_1$.*

1. *Compute the selected model $\widehat{m}^{\mathrm{iso}}(N)$ as a function of $N > 0$*

$$\widehat{m}^{\mathrm{iso}}(N) \in \arg\min_{m\in\mathcal{M}} \left\{ \gamma_{n,p_1,p_2}\left(\widehat{\theta}^{\mathrm{iso}}_m\right) + N\frac{d^{\mathrm{iso}}_m}{np_1p_2}\right\} \ .$$

2. *Find $\widehat{N}^{\mathrm{iso}}_{\min} > 0$ such that the jump $d_{\widehat{m}^{\mathrm{iso}}\left(\left[\widehat{N}^{\mathrm{iso}}_{\min}\right]_-\right)} - d_{\widehat{m}^{\mathrm{iso}}\left(\left[\widehat{N}^{\mathrm{iso}}_{\min}\right]_+\right)}$ is maximal.*

3. *Select the model $\widehat{m}^{\mathrm{iso}} = \widehat{m}^{\mathrm{iso}}(2\widehat{N}^{\mathrm{iso}}_{\min})$.*

In conclusion, the model selection procedure described in Algorithm 7.5 is completely data-driven and does not require any prior knowledge on the matrix Σ. Moreover, its computational burden remains small. We illustrate its efficiency in Section 7.6.

7.5 Extension to non-toroidal lattices

It is often artificial to consider the field X as stationary on a torus. However, we needed this hypothesis for deriving nonasymptotic properties of the estimator $\widetilde{\theta}$ in Chapter 6. In many applications, it is more realistic to assume that we observe a small window of a Gaussian field defined on the plane $\mathbb{Z}^2$. If we are unable to prove nonasymptotic risk bounds in this new setting. Nevertheless, Lakshman and Derin have shown in [LD93] that there is no phase transition within the valid parameter space for GMRFs defined on the plane $\mathbb{Z}^2$. Let us briefly explain what this means: consider a GMRF defined on a square lattice of size p, but only observed on a square lattice of size p'. The absence of phase transition implies the distribution of this field observed on this fixed window of size p' does not asymptotically depend on the bound conditions when p goes to infinity. Consequently, it is reasonable to think that our estimation procedure still performs well to the price of slight modifications. In the sequel, we assume that the field X is defined on $\mathbb{Z}^2$, but the data $\mathbf{X}$ still correspond to n independent observations of the field X on the window Λ of size $p_1 \times p_2$. The conditional distribution of $X_{[0,0]}$ given the remaining covariates now decomposes as

$$X_{[0,0]} = \sum_{(i,j)\in\mathbb{Z}^2\setminus\{(0,0)\}} \theta_{[i,j]} X_{[i,j]} + \epsilon_{[0,0]} \ , \tag{7.16}$$

where $\theta_{[.,.]}$ is an "infinite" matrix defined on $\mathbb{Z}^2$ and where $\epsilon_{[0,0]}$ is a centered Gaussian variable of variance σ^2 independent of $(X_{[i,j]})_{(i,j)\in\Lambda\setminus\{(0,0)\}}$. The distribution of the field X is uniquely defined by the function θ and positive number σ^2. The set $\Theta^{+,\infty}$ of valid parameter for θ is now defined using the spectral density function. We refer to Rue and Held [RH05] Sect.2.7 for more details.

Definition 7.7. A function $\theta : \mathbb{Z}^2 \to \mathbb{R}$ belongs to the set $\Theta^{+,\infty}$ if and only if it satisfies the three following conditions:

1. $\theta_{[0,0]} = 0$.
2. For any $(i,j) \in \mathbb{Z}^2$, $\theta_{[i,j]} = \theta_{[-i,-j]}$.
3. For any $(\omega_1, \omega_2) \in [0, 2\pi)^2$, $1 - \sum_{(i,j)\in\mathbb{Z}^2} \theta_{[i,j]} \cos(i\omega_1 + j\omega_2) > 0$.

Similarly, we define the set $\Theta^{+,\infty,\text{iso}}$ for the isotropic GMRFs on the lattices. As done in Section 7.2 for toroidal lattices, we now introduce the parametric parameter sets. For any model $m \in \mathcal{M}_1$, the set $\Theta_m^{+,\infty}$ refers to the subset of matrices θ in $\Theta^{+,\infty}$ whose support is included in m. Analogously, we define the parametric set $\Theta_m^{+,\infty,\text{iso}}$ corresponding to isotropic GMRFs.

We cannot directly extend the CLS empirical contrast $\gamma_{n,p_1,p_2}(.)$ defined in (7.6) in this new setting because we have to take the edge effect into account. Indeed, if we want to compute the conditional regression of $\mathbf{X}_{i[j_1,j_2]}$, we have to observe *all* its neighbors with respect to m, i.e. $\{\mathbf{X}_{i[j_1+l_1,j_2+l_2]},\ (l_1,l_2) \in m\}$. In this regard, we define the set Λ_m for any model $m \in \mathcal{M}_1$.

$$\Lambda_m := \{(i_1,i_2) \in \Lambda\,,\ (m + (i_1,i_2)) \subset \Lambda\} \ ,$$

where $(m + (i,j))$ denotes the set m of nodes translated by (i,j). For instance, if we consider the model m_1 with four nearest neighbors, the edge effect size is one and Λ_m contains all the nodes that do not lie on the border. The model m_3 with 12 nearest neighbors yields an edge effect of size 2 and Λ_m contains all the nodes in Λ, except those which are at a (euclidean) distance strictly smaller than 2 from the border.

For any model $m \in \mathcal{M}_1$ and any θ' belonging to $\Theta_m^{+,\infty}$, we define $\gamma_{n,p_1,p_2}^{\Lambda_m}(.)$ as an analogous of $\gamma_{n,p_1,p_2}(.)$ except that it only relies on the conditional regression of the nodes in Λ_m.

$$\gamma_{n,p_1,p_2}^{\Lambda_m}(\theta') \ := \ \frac{1}{n\text{Card}(\Lambda_m)} \sum_{i=1}^{n} \sum_{(j_1,j_2)\in\Lambda_m} \left(\mathbf{X}_{i[j_1,j_2]} - \sum_{(l_1,l_2)\in m} \theta'_{[l_1,l_2]} \mathbf{X}_{i[j_1+l_1,j_2+l_2]} \right)^2 .$$

Then a *CLS* estimators $\widehat{\theta}_m^{\Lambda_m}$ and $\widehat{\theta}_m^{\Lambda_m,\text{iso}}$ are defined by

$$\widehat{\theta}_m^{\Lambda_m} \in \arg\min_{\theta'\in\Theta_m^{+,\infty}} \gamma_{n,p_1,p_2}^{\Lambda_m}(\theta') \qquad \text{and} \qquad \widehat{\theta}_m^{\Lambda_m,\text{iso}} \in \arg\min_{\theta'\in\Theta_m^{+,\infty,\text{iso}}} \gamma_{n,p_1,p_2}^{\Lambda_m}(\theta') \ .$$

Contrary to $\widehat{\theta}_m$, the estimator $\widehat{\theta}_m^{\Lambda_m}$ is not necessarily unique especially if the size of Λ_m is smaller than d_m. Let us mention that it is quite classical in the literature to remove nodes if we want to take edge effects or missing data into account (see e.g. [Guy95] Sect.4.3). We cannot use anymore Fast Fourier transform for computing the parametric estimator. Nevertheless, $\widehat{\theta}_m^{\Lambda_m}$ is still computationally amenable, since it minimizes a quadratic function on the closed convex set $\Theta_m^{+,\infty}$.

When considering toroidal networks, we have applied a linear penalization of the order $d_m/(n\mathrm{Card}(\Lambda))$, i.e. the dimension of the model divided by the total number of nodes observed. Since $\gamma_{n,p_1,p_2}^{\Lambda_m}(\theta')$ only evaluates the conditional regression of the nodes in the subset Λ_m, we shall penalize our model by the quantity $\frac{d_m}{n\mathrm{Card}(\Lambda_m)}$. In the sequel, $n\mathrm{Card}(\Lambda_m)$ is called the *effective number* of observations.

We now provide a data-driven model selection procedure for choosing the neighborhood. It is based on the slope heuristic developed in the previous section. Suppose we are given a subcollection $\mathcal{M}$ of $\mathcal{M}_1$. Then, the penalized estimator $\widehat{\theta}_{\widehat{m}}^{\Lambda_{\widehat{m}}}$ is computed as follows.

Algorithm 7.8. *(Data-driven penalization for non-toroidal lattice).*

1. *Compute the selected model $\widehat{m}(L)$ as a function of $L > 0$*

$$\widehat{m}(L) \in \arg\min_{m\in\mathcal{M}} \left\{ \gamma_{n,p_1,p_2}^{\Lambda_m}(\widehat{\theta}_m^{\Lambda_m}) + L\frac{d_m}{n\,Card(\Lambda_m)} \right\} .$$

2. *Find $\widehat{L}_{\min} > 0$ such that the jump $d_{\widehat{m}\left([\widehat{L}_{\min}]_-\right)} - d_{\widehat{m}\left([\widehat{L}_{\min}]_+\right)}$ is maximal.*

3. *Select the model $\widehat{m} = \widehat{m}(2\widehat{L}_{\min})$.*

This procedure straightforwardly extends to the case of isotropic GMRFs estimation by replacing $\widehat{m}(N)$ by $\widehat{m}^{\mathrm{iso}}(N)$ and d_m by d_m^{iso}. As for Algorithm 7.5, it is advised to introduce huge models in the collection $\mathcal{M}$ in order to better detect the dimension jump. However, when the dimension of the models increases the size of Λ_m decreases and the estimator $\widehat{\theta}_m^{\Lambda_m}$ may become unreliable. The method therefore requires a reasonable number of data. In practice, Λ should not contain less than 100 nodes.

7.6 Simulation study

In the first simulation experiment, we compare the efficiency of our procedure with penalized maximum likelihood methods for estimating the parameters of a GMRF on a torus. In the second experiment, we study the estimation of a Gaussian field observed on a rectangles. The calculations are made with *R* `www.r-project.org`.

7.6.1 Isotropic GMRF on a torus

First, we consider X an isotropic GMRF on the torus Λ of size $p = p_1 = p_2 = 20$. There are therefore 400 points in the lattice. The number of observations n equals one and the conditional variance σ^2 is one. We introduce a radius $r := \sqrt{17}$. Then, for any number $\phi > 0$, we define the $p \times p$ matrix θ^ϕ as:

$$\left\{ \begin{array}{lcll} \theta^\phi{}_{[0,0]} & := & 0 , & \\ \theta^\phi{}_{[i,j]} & := & \phi & \text{if } |(i,j)|_t \leq r \text{ and } (i,j) \neq (0,0) , \\ \theta^\phi{}_{[i,j]} & := & 0 & \text{if } |(i,j)|_t > r . \end{array} \right.$$

In practice, we set ϕ to 0, 0.0125, 0.015, and 0.0175. Observe that these choices constrain $\|\theta^\phi\|_1 < 1$. The matrix θ^ϕ therefore belongs to the set $\Theta_{m_{10}}^{+,\mathrm{iso}}$ of dimension 10 introduced in Definition 7.1.

In the sequel, we shall compare the efficiency of three model selection procedures. For each of them, we use the collection $\mathcal{M} := \{m_0, m_1, \ldots, m_{20}\}$ whose maximal dimension $d_{m_{20}}^{\mathrm{iso}}$ is 21. The estimator $\widetilde{\theta}^{\mathrm{iso}}$ is built using the CLS model selection procedure introduced in Algorithm 7.6. The two other procedures are based on likelihood maximization. In this regard, we first define the parametric maximum likelihood estimator $\widehat{\theta}_m^{\mathrm{mle}}$ for any model $m \in \mathcal{M}$.

$$\left(\widehat{\theta}_m^{\mathrm{mle}}, \widehat{\sigma}_m^{\mathrm{mle}}\right) := \arg\min_{\theta'\in\Theta_m^{+,\mathrm{iso}},\sigma'} -\mathcal{L}_p(\theta', \sigma', \mathbf{X}) ,$$

where $\mathcal{L}_p(\theta', \mathbf{X})$ stands for the log-likelihood at the parameter θ'. We then select a model m applying either an AIC-type criterion [Aka73] or a BIC-type criterion [Sch78]:

$$\begin{aligned}\widehat{m}^{\mathrm{AIC}} &:= \arg\min_{m\in\mathcal{M}} \left\{-2\mathcal{L}_p(\widehat{\theta}^{\mathrm{mle}}_m, \widehat{\sigma}^{\mathrm{mle}}_m, \mathbf{X}) + 2d^{\mathrm{iso}}_m\right\} ,\\ \widehat{m}^{\mathrm{BIC}} &:= \arg\min_{m\in\mathcal{M}} \left\{-2\mathcal{L}_p(\widehat{\theta}^{\mathrm{mle}}_m, \widehat{\sigma}^{\mathrm{mle}}_m, \mathbf{X}) + \log(p^2)d^{\mathrm{iso}}_m\right\} .\end{aligned}$$

For short, we write $\widehat{\theta}^{\mathrm{AIC}}$ and $\widehat{\theta}^{\mathrm{BIC}}$ for the two obtained estimators $\widehat{\theta}^{\mathrm{mle}}_{\widehat{m}^{\mathrm{AIC}}}$ and $\widehat{\theta}^{\mathrm{mle}}_{\widehat{m}^{\mathrm{BIC}}}$. Although AIC and BIC procedures are not justified in this setting, we still apply them as they are widely used in lots of framework. Moreover, their computation is performed efficiently by using Fast-Fourier transform described in Section 7.2.3.

The experiments are repeated $N = 1000$ times. The Gaussian field is simulated by using the Fast Fourier transform. The quality of the estimations is assessed thanks to the prediction loss function $l(.,.)$ defined in (7.11). For any ϕ and any of these three estimators, we evaluate the risks $\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}^{\mathrm{AIC}}, \theta^\phi)]$, $\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}^{\mathrm{BIC}}, \theta^\phi)]$, and $\mathbb{E}_{\theta^\phi}[l(\widetilde{\theta}, \theta^\phi)]$ as well as the corresponding empirical 95% confidences intervals by Monte-Carlo method. Thanks to the N experiments, we also estimate the risk of $\widehat{\theta}_m$ for each model $m \in \mathcal{M}$. It then allows to evaluate the oracle risks $\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}_{m^*}, \theta^\phi)]$. and the risk ratios $\mathbb{E}_{\theta^\phi}[l(\widetilde{\theta}, \theta^\phi)]/\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}_{m^*}, \theta^\phi)]$ The results are given in Table 9.

$\phi \times 10^2$	0	1.25	1.5	1.75
$\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}^{\mathrm{AIC}}, \theta^\phi)] \times 10^2$	1.2 ± 0.2	3.1 ± 0.2	4.3 ± 0.2	6.4 ± 0.2
$\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}^{\mathrm{BIC}}, \theta^\phi)] \times 10^2$	0.01 ± 0.01	1.9 ± 0.1	3.7 ± 0.1	9.7 ± 0.3
$\mathbb{E}_{\theta^\phi}[l(\widetilde{\theta}, \theta^\phi)] \times 10^2$	1.6 ± 0.2	3.2 ± 0.2	4.2 ± 0.1	7.2 ± 0.3
$\mathbb{E}_{\theta^\phi}[l(\widetilde{\theta}, \theta^\phi)]/\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}_{m^*}, \theta^\phi)]$	$+\infty$	1.9 ± 0.7	1.3 ± 0.2	1.5 ± 0.3

Table 9: First simulation study. Estimates and 95% confidence intervals of the risks $\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}^{\mathrm{AIC}}, \theta^\phi)]$, $\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}^{\mathrm{BIC}}, \theta^\phi)]$, and $\mathbb{E}_{\theta^\phi}[l(\widetilde{\theta}, \theta^\phi)]$ and of the ratio $\mathbb{E}_{\theta^\phi}[l(\widetilde{\theta}, \theta^\phi)]/\mathbb{E}_{\theta^\phi}[l(\widehat{\theta}_{m^*}, \theta^\phi)]$.

The BIC criterion outperforms the other procedures when $\phi = 0$, 0.0125, or 0.015 but behaves bad for a large ϕ. Indeed, the BIC criterion has a tendency to overpenalize the models. For the two first values of ϕ the oracle model in $\mathcal{M}$ is m_0. Hence, overpenalizing increases the performance of estimation in this case. However, when ϕ increases, the dimension of the oracle model is larger and BIC therefore selects too small models.

In contrast, AIC and the CLS estimator exhibit similar behaviors. If we forget the case $a = 0$ for which the oracle risk is 0, the risk of $\widetilde{\theta}$ is close to the risk of the oracle model. Hence, the model selection procedure for $\widetilde{\theta}$ is almost optimal.

In conclusion, $\widetilde{\theta}$ or $\widehat{\theta}^{\mathrm{AIC}}$ both exhibit good performances for estimating the distribution of a regular Gaussian field on a torus. The strength of our model selection procedure lies in the fact it easily generalizes to non-toroidal lattices as illustrated in the next section.

7.6.2 Isotropic Gaussian fields on $\mathbb{Z}^2$

We now consider X an isotropic Gaussian field defined on $\mathbb{Z}^2$ but only observed on a square Λ of size $p = p_1 = p_2 = 20$. This corresponds to the situation described in Section 7.5. The variance of $X_{[0,0]}$ is set to one and the distribution of the field is therefore uniquely defined by its correlation function $\rho(k, l) := \mathrm{corr}(X_{[k,l]}, X_{[0,0]})$. Again, the number of replications n is chosen to be one. In the experiment, we use four classical correlation functions: exponential, spherical, circular, and Matérn (e.g. [Cre93]

Sect.2.3.1 and [Mat86]).

$$\begin{aligned}
\text{Exponential:}\quad \rho(k,l) &= \exp\left(-\frac{d(k,l)}{r}\right)\\
\text{Circular:}\quad \rho(k,l) &= \begin{cases} 1-\frac{2}{\pi}\left[\frac{d(k,l)}{r}\sqrt{1-\left(\frac{d(k,l)}{r}\right)^2}+\sin^{-1}\left(\sqrt{\frac{d(k,l)}{r}}\right)\right] & \text{if} \quad d(k,l)\le r\\ 0 & \text{else}\end{cases}\\
\text{Spherical:}\quad \rho(k,l) &= \begin{cases} 1-1.5\frac{d(k,l)}{r}+0.5\left(\frac{d(k,l)}{r}\right)^3 & \text{if} \quad d(k,l)\le r\\ 0 & \text{else}\end{cases}\\
\text{Matérn:}\quad \rho(k,l) &= \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{d(k,l)}{r}\right)^{\kappa}\mathcal{K}_\kappa\left(\frac{d(k,l)}{r}\right),
\end{aligned}$$

where $d(k,l)$ denotes the euclidean distance from (k,l) to $(0,0)$ and $\mathcal{K}_\kappa(.)$ is the modified Bessel function of order κ. In a nutshell, the parameter r represents the range of correlation, whereas κ may be regarded as a smoothness parameter for the Matérn function. In this simulation experiment, we set r to 3. When considering the Matérn model, we take κ equal to 0.05, 0.25, 0.5 and 1.

We compute the estimator $\widetilde{\theta}$ based on Algorithm 7.8 with the collection $\mathcal{M} := \{m \in \mathcal{M}_1^{\text{iso}}\,,\, d_m \le 18\}$. Since the lattice Λ is not a torus, methods based on likelihood maximization exhibit a prohibitive computational burden. Consequently, we do not use MLE in this experiment. We shall compare the efficiency of $\widetilde{\theta}$ with a variogram-based estimation method.

Let us recall that the matrix θ corresponds to the coefficients of the conditional regression of $X_{[0,0]}$ given $\{X_{[i,j]},\ (i,j) \in \Lambda\backslash\{(0,0)\}\}$ (Equation (7.1)). Hence, the linear combination $\sum_{(i,j)\in\Lambda\backslash\{(0,0)\}} \theta_{[i,j]}X_{[i,j]}$ is the kriging predictor of $X_{[0,0]}$ given the remaining variables. A natural method to estimate θ in this spatial setting amounts to estimating the variogram of the observed Gaussian field and then performing ordinary kriging at the node $(0,0)$ that lies in the center of Λ. More precisely, we first estimate the empirical variogram by applying the modulus estimator of Hawkes and Cressie (e.g. [Cre93] Eq.(2.2.8)) to the observed field of 400 points. Afterwards, we fit this empirical variogram to a variogram model using the reweighted least-squares suggested by Cressie [Cre85]. This procedure therefore requires the choice of a particular variogram model. In this simulation study, we choose *the model* that has generated the data. Observe that this method is *not* adaptive since it requires the knowledge of the variogram model. In practice, we use Library *geoR* [RJ01] implemented in *R* to estimate the parameters r, $\text{var}(X_{[0,0]})$ and eventually κ of the variogram model. Then, we compute the estimator $\widehat{\theta}^K$ by performing ordinary kriging at the center node of Λ. For each of these estimations, we assume that the variogram model is known. For computational reasons, we use a kriging neighborhood of size 11×11 that contains 120 points. Previous simulations have indicated that this neighborhood choice does not decrease the precision of the estimation.

We assess the performances of the procedures using the loss $l(.,.)$. Even, if the $l(.,.)$ is defined in (7.11) for a torus, the alternative definition (7.12) clearly extends to this non toroidal setting. Consequently, the loss $l(\widehat{\theta},\theta)$ measures the difference between the prediction error of $X_{[0,0]}$ when using $\sum_{(i,j)\in\Lambda\backslash\{(0,0)\}} \widehat{\theta}_{[i,j]}X_{[i,j]}$ and the prediction error of $X_{[0,0]}$ when using the best predictor $\mathbb{E}[X_{[0,0]}|(X_{[i,j]})_{(i,j)\in\Lambda\backslash\{(0,0)\}}]$. In other words, $l(\widehat{\theta},\theta)$ is the difference of the kriging error made with the estimated parameters $\widehat{\theta}$ and the kriging error made with the true parameter θ.

For any of the four correlation models previously mentioned, we evaluate the risks $\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]$ and $\mathbb{E}_\theta[l(\widehat{\theta}^K,\theta)]$ by Monte-Carlo and also compare the risk of $\widetilde{\theta}$ with the oracle $\widehat{\theta}_{m^*}$. As in Section 7.6.1, we evaluate the oracle risk $\mathbb{E}[l(\widehat{\theta}_{m^*},\theta)]$ by taking the minimum of the evaluations of the risks $\mathbb{E}[l(\widehat{\theta}_m,\theta)]$ over all models $m \in \mathcal{M}$. Results of the simulation experiment are given in Table 10 and 11.

Observe that none of these fields are GMRFs. Here, the GMRF models should only be viewed as a collection of approximation sets of the true distribution. This simulation experiment is in the spirit of the study of Rue and Tjelmeland [RT02]. However, there are some major differences. Contrary to them, we perform estimation and not only approximation. Moreover, our lattice is not a torus. Finally, we use our prediction loss $l(.,.)$ to assess the performance, whereas they compare the obtained correlation functions.

In Table 10 and 11, the ratio $\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]/\mathbb{E}_\theta[l(\widehat{\theta}_{m^*},\theta)]$ stays close to one. Hence, the model selection is almost optimal from an efficiency point of view. Apart from the exponential model, the estimator $\widetilde{\theta}$

Model	Exponential	Circular	Spherical
$\mathbb{E}_\theta[l(\widehat{\theta}^K,\theta]\times 10^2$	0.08 ± 0.03	8.8 ± 0.5	3.0 ± 0.2
$\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]\times 10^2$	0.70 ± 0.09	6.2 ± 0.3	3.0 ± 0.2
$\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]/\mathbb{E}_\theta[l(\widehat{\theta}_{m^*},\theta)]$	1.4 ± 0.2	1.1 ± 0.1	1.6 ± 0.3

Table 10: Estimates and 95% confidence intervals of the risks $\mathbb{E}_\theta[l(\widehat{\theta}^K,\theta)]$ and $\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]$ and of the ratio $\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]/\mathbb{E}_\theta[l(\widehat{\theta}_{m^*},\theta)]$ for the exponential, circular and spherical models.

κ	0.05	0.25	0.5	1
$\mathbb{E}_\theta[l(\widehat{\theta}^K,\theta)]\times 10^2$	13.0 ± 1.0	9.9 ± 1.0	9.0 ± 1.70	2.7 ± 0.4
$\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]\times 10^2$	5.2 ± 0.8	1.6 ± 0.3	0.71 ± 0.09	0.37 ± 0.06
$\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]/\mathbb{E}_\theta[l(\widehat{\theta}_{m^*},\theta)]$	2.5 ± 0.7	3.6 ± 1.4	1.4 ± 0.2	2.3 ± 0.5

Table 11: Estimates and 95% confidence intervals of the risks $\mathbb{E}_\theta[l(\widehat{\theta}^K,\theta)]$ and $\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]$ and of the ratio $\mathbb{E}_\theta[l(\widetilde{\theta},\theta)]/\mathbb{E}_\theta[l(\widehat{\theta}_{m^*},\theta)]$ for Matérn model.

outperforms the estimator $\widehat{\theta}^K$ based on geostatistical methods. This is particularly striking for Matérn correlation model since in that case the computation of $\widehat{\theta}^K$ requires the estimation of the additional parameter κ. Indeed, let us recall that the exponential model and the Matérn model with $\kappa = 0.5$ are exactly equivalent. Hence, the risk of $\widehat{\theta}^K$ is 100 times higher when κ has to be estimated than when κ is known to equal 0.5.

Moreover, the kriging estimator $\widehat{\theta}^K$ requires the knowledge or the choice of a correlation model. However, there is no reason for real-world data to follow exactly the distribution of one of these correlation models. In contrast, our method does not rely on any previous knowledge on the distribution of the field. Consequently, the estimator $\widetilde{\theta}$ may even more outperform the variogram-type estimation methods if the true distribution is far from one of the classical variogram models. However, our estimated partial correlation matrix $(I_{p_1p_2} - C(\widetilde{\theta}))$ is possibly singular and we may therefore not necessarily be able to derive from it an estimation of the correlation matrix.

7.7 Proofs

Let us introduce some notations that shall be used throughout the proofs. For any $1 \le k \le n$, the vector $\mathbf{X}_k^v$ denotes the vectorialized version of the k-th sample of X. Moreover, $\mathbf{X^v}$ is the matrix of size $p^2 \times n$ of the n realisations of the vector $\mathbf{X}_k^v$. Throughout these proofs, L, L_1, L_2 denote constants that may vary from line to line. The notation $L(.)$ specifies the dependency on some quantities. Finally, the $\gamma(.)$ function stands for an infinite sampled version of the CLS criterion $\gamma_{n,p_1,p_2}(.)$: $\gamma(.) := \mathbb{E}[\gamma_{n,p_1,p_2}(.)]$.

7.7.1 Proof of Lemma 7.2

Let us provide an alternative expression of $\gamma_{n,p_1,p_2}(\theta')$ in term of the factor $C(\theta')$ and the empirical covariance matrix $\overline{\mathbf{X^v}\mathbf{X^{v*}}}$.

$$\gamma_{n,p_1,p_2}(\theta') \quad = \quad \frac{1}{np_1p_2} tr\left[(I_{p_1p_2} - C(\theta'))\overline{\mathbf{X^v}\mathbf{X^{v*}}}(I_{p_1p_2} - C(\theta'))\right] \ . \tag{7.17}$$

This is justified in Section 6.2.2.

Lemma 7.9. *There exists an orthogonal matrix P which simultaneously diagonalizes every $p^2 \times p^2$ symmetric block circulant matrices with $p \times p$ blocks. Let θ be a matrix of size $p_1 \times p_2$ such that $C(\theta)$ is symmetric. The matrix $D(\theta) = P^*C(\theta)P$ is diagonal and satisfies*

$$D(\theta)_{[(i-1)p_1+j,(i-1)p_1+j]} = \sum_{k=1}^{p_1}\sum_{l=1}^{p_2} \theta_{[k,l]} \cos\left[2\pi(ki/p_1 + lj/p_2)\right] \ , \tag{7.18}$$

for any $1 \leq i \leq p_1$ *and* $1 \leq j \leq p_2$.

It is proved as in [RH05] Sect.2.6.2 to the price of a slight modification in order to take into account the fact that P is orthogonal and not unitary. The difference comes from the fact that contrary to Rue and Held we also assume that $C(\theta)$ is symmetric. This lemma states that all symmetric block circulant matrices are simultaneously diagonalizable. Observe that for any $1 \leq i \leq p_1$ and $1 \leq j \leq p_2$, it holds that $D(\theta)_{[(i-1)p_1+j,(i-1)p_1+j]} = \lambda_{[i,j]}(\theta)$ since $\theta_{[k,l]} = \theta_{[p_1-k,p_2-l]}$. Hence, Expression (7.17) becomes

$$\gamma_{n,p_1,p_2}(\theta') \quad = \quad \frac{1}{np_1p_2}\left\{\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\left[1-\lambda_{[i,j]}(\theta)\right]^2\left[\sum_{k=1}^{n}\left[P^*\mathbf{X}_k^v(\mathbf{X}_k^v)^*P\right]_{[(i-1)p_1+j,(i-1)p_1+j]}\right]\right\} ,$$

where $\mathbf{X}_k^v$ is the vectorialized version of the k-th observation of the field X. Straightforward computations allow us to prove that the quantities

$$\left(P^*\mathbf{X}_k^v(\mathbf{X}_k^v)^*P\right)_{[(i-1)p_1+j,(i-1)p_1+j]} + \left(P^*\mathbf{X}_k^v(\mathbf{X}_k^v)^*P\right)_{[(p_1-i-1)p_1+p_2-j,(p_1-i-1)p_1+p_2-j]}$$

and

$$\frac{1}{\sqrt{p_1p_2}}\lambda_{[i,j]}(\mathbf{X}_k^v)\overline{\lambda_{[i,j]}(\mathbf{X}_k^v)} + \frac{1}{\sqrt{p_1p_2}}\lambda_{[p_1-i,p_2-j]}(\mathbf{X}_k^v)\overline{\lambda_{[p_1-i,p_2-j]}(\mathbf{X}_k^v)}$$

are equal for any $1 \leq i \leq p_1$ and $1 \leq j \leq p_2$. Here, the entries of the matrix $\lambda(.)$ are taken modulo p_1 and p_2 and the entries of $[P^*\mathbf{X}_k^v(\mathbf{X}_k^v)^*P]$ are taken modulo p_1p_2. The result follows.

7.7.2 Proof of Proposition 7.4

Proof of Proposition 7.4. We only consider the anisotropic case, since the proof for isotropic estimation is analogous. For any model $m \in \mathcal{M}_1$, we define

$$\Delta(m,m') := \gamma_{n,p,p}\left(\widehat{\theta}_{m,\rho}\right) + \mathrm{pen}(m) - \gamma_{n,p,p}\left(\widehat{\theta}_{m',\rho}\right) - \mathrm{pen}(m') .$$

We aim at showing that with large probability, the quantity $\Delta(m,m')$ is positive for all small dimensional model m. Hence, we would conclude the dimension of $\widehat{m}$ is large. In this regard, we bound the deviations of the differences

$$\begin{aligned}\gamma_{n,p,p}\left(\widehat{\theta}_{m,\rho}\right) - \gamma_{n,p,p}\left(\widehat{\theta}_{m',\rho}\right) \quad &= \quad \left[\gamma_{n,p,p}\left(\widehat{\theta}_{m,\rho}\right) - \gamma_{n,p,p}\left(\theta_{m,\rho}\right)\right] + \left[\gamma_{n,p,p}\left(\theta_{m,\rho}\right) - \gamma_{n,p,p}\left(\theta\right)\right] \\ &+ \quad \left[\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}\left(\widehat{\theta}_{m',\rho}\right)\right] .\end{aligned}$$

Lemma 7.10. *Let* K_2 *be some universal constant that we shall define in the proof. With probability larger than* $3/4$,

$$\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}(\theta_{m,\rho}) \leq \frac{K_2}{2}\rho^2\varphi_{max}(\Sigma)\frac{d_m \vee 1}{np^2} \quad \textit{and} \quad \gamma_{n,p,p}\left(\theta_{m,\rho}\right) - \gamma_{n,p,p}\left(\widehat{\theta}_{m,\rho}\right) \leq \frac{K_2}{2}\rho^2\varphi_{max}(\Sigma)\frac{d_m}{np^2}$$

for all model $m \in \mathcal{M}_1$.

Lemma 7.11. *Assume that* p *is larger than some numerical constant* p_0. *With probability larger than* $3/4$, *it holds that*

$$\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}(\widehat{\theta}_{m',\rho}) \geq K_3\sigma^2\left[\varphi_{min}\left[I_{p^2} - C(\theta)\right] \wedge \left[\rho - \varphi_{max}\left(I_{p^2} - C(\theta)\right)\right]\right]\frac{d_{m'}}{np^2} ,$$

where K_3 *is a universal constant defined in the proof.*

Let us define K_1 to be exactly K_3. Gathering these two Lemma with Assumption (7.15), there exists an event Ω of probability larger than $1/2$ such that

$$\Delta(m,m') \geq \frac{\sigma^2}{np^2}\left\{K_1\eta d_{m'}\left[\varphi_{\min}\left(I_{p^2} - C(\theta)\right) \wedge \left[\rho - \varphi_{\max}\left(I_{p^2} - C(\theta)\right)\right]\right] - K_2(d_m \vee 1)\frac{\rho^2}{\varphi_{\min}\left[I_{p^2} - C(\theta)\right]}\right\} ,$$

for all model $m \in \mathcal{M}_1$. Thus, conditionally to the event Ω, $\Delta(m, m')$ is positive for all models $m \in \mathcal{M}_1$ that satisfy

$$\frac{d_m \vee 1}{d_{m'}} \leq \frac{K_3 \eta}{K_2 \rho^2} \varphi_{\min}\left(I_{p^2} - C(\theta)\right) \left\{\varphi_{\min}\left(I_{p^2} - C(\theta)\right) \wedge \left[\rho - \varphi_{\max}\left(I_{p^2} - C(\theta)\right)\right]\right\} \ .$$

By Lemma 6.35, the dimension $d_{m'}$ is larger than $\frac{1}{2}\left[\sqrt{np^2} \wedge \left(p^2 - 1\right)\right]$. We conclude that

$$d_{\widehat{m}_\rho} \vee 1 \geq \left[\sqrt{np^2} \wedge p^2 - 1\right] \frac{K_3 \eta}{K_2 \rho^2} \varphi_{\min}\left(I_{p^2} - C(\theta)\right) \left\{\varphi_{\min}\left(I_{p^2} - C(\theta)\right) \wedge \left[\rho - \varphi_{\max}\left(I_{p^2} - C(\theta)\right)\right]\right\} \ ,$$

with probability larger than $1/2$. □

Proof of Lemma 7.10. In the sequel, $\overline{\gamma}_{n,p}(.)$ denotes the difference $\gamma_{n,p,p}(.)$ -$\gamma(.)$. Given a model m, we consider the difference

$$\gamma_{n,p,p}\left(\theta\right) - \gamma_{n,p,p}\left(\theta_{m,\rho}\right) \quad = \quad \overline{\gamma}_{n,p}\left(\theta\right) - \overline{\gamma}_{n,p}\left(\theta_{m,\rho}\right) - l(\theta_{m,\rho}, \theta) \ .$$

Upper bounding the difference of $\gamma_{n,p,p}$ therefore amounts to bounding the difference of $\overline{\gamma}_{n,p}$. By definition of $\gamma_{n,p,p}$ and γ, it expresses

$$\overline{\gamma}_{n,p}\left(\theta\right) - \overline{\gamma}_{n,p}\left(\theta_{m,\rho}\right) \quad = \quad \frac{1}{p^2} tr\left\{\left[(I_{p^2} - C(\theta))^2 - (I_{p^2} - C(\theta_{m,\rho}))^2\right]\left(\overline{\mathbf{X}^{\mathbf{v}}\mathbf{X}^{\mathbf{v}*}} - \Sigma\right)\right\}$$

The matrices Σ, $(I_{p^2} - C(\theta))$, and $I_{p^2} - C(\theta_{m,\rho})$ are symmetric block circulant. By Lemma 7.9, they are jointly diagonalizable in the same orthogonal basis. If we note P an orthogonal matrix associated to this basis, then $C(\theta_{m,\rho})$, $C(\theta)$ and Σ respectively decompose in

$$C(\theta_{m,\rho}) = P^* D(\theta_{m,\rho}) P \ , \ C(\theta) = P^* D(\theta) P \ \text{ and } \Sigma = P^* D(\Sigma) P \ ,$$

where the matrices $D(\theta_{m,\rho})$, $D(\theta)$, and $D(\Sigma)$ are diagonal.

$$\overline{\gamma}_{n,p}\left(\theta\right) - \overline{\gamma}_{n,p}\left(\theta_{m,\rho}\right) \quad = \quad \frac{1}{p^2} tr\left\{D(\theta_{m,\rho}) - D(\theta)\right)\left(2I_{p^2} - D(\theta) - D(\theta_{m,\rho}) D_\Sigma\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right\} \quad (7.19)$$

where the matrix $\mathbf{Y}$ is defined as $P\sqrt{\Sigma^{-1}}\mathbf{X}^v P^*$. Its components follow independent standard Gaussian distributions. Since the matrices involved in (7.19) are diagonal, this last expression is a linear combination of centered χ^2 random variables. We shall apply the following Lemma to bound its deviations.

Lemma 7.12. *Let $(Y_1, \ldots, Y_D)$ be i.i.d. standard Gaussian variables. Let $a_1, \ldots, a_D$ be fixed numbers. We set*

$$\|a\|_\infty := \sup_{i=1,\ldots,D} |a_i|, \quad \|a\|_2^2 := \sum_{i=1}^D a_i^2$$

Let T be the random variable defined by

$$T := \sum_{i=1}^D a_i \left(Y_i^2 - 1\right) \ .$$

Then, the following deviation inequality holds for any positive x

$$\mathbb{P}\left[T \geq 2\|a\|_2 \sqrt{x} + 2\|a\|_\infty x\right] \leq e^{-x} \ .$$

This result is very close to Lemma 1 of Laurent and Massart in [LM00]. The only difference lies in the fact that they constrain the coefficients a_i to be non-negative. Nevertheless, their proof easily extends to our situation. Let us define the vector ϕ of size $n \times p^2$ as

$$a^i[j] := \frac{D_{\Sigma[i,i]}\left(D(\theta_{m,\rho})_{[i,i]} - D(\theta)_{[i,i]}\right)\left(2 - D(\theta[i,i] - D(\theta_{m,\rho})[i,i]\right)}{np^2} \ ,$$

for any $1 \leq i \leq n$ and any $1 \leq j \leq p^2$. Since the matrices $I - C(\theta)$ and $I - C(\theta_{m,\rho})$ belong to the set Θ_ρ^+, their largest eigenvalue is smaller than ρ. By Definition (7.11) of the loss function $l(.,.)$, $\|a\|_2 \leq 2\rho\sqrt{\frac{\varphi_{\max}(\Sigma) l(\theta_{m,\rho},\theta)}{np^2}}$ and $\|a\|_\infty \leq 4\rho^2 \frac{\varphi_{\max}(\Sigma)}{np^2}$. By Applying Lemma 7.12 to Expression (7.19), we conclude that

$$\mathbb{P}\left[\overline{\gamma}_{n,p}(\theta) - \overline{\gamma}_{n,p}(\theta_{m,\rho}) \geq l(\theta_{m,\rho},\theta) + 12\rho^2 \frac{\varphi_{\max}(\Sigma)}{np^2} x\right] \leq e^{-x} ,$$

for any $x > 0$. Consequently, for any $K > 0$, the difference of $\gamma_{n,p,p}$ satisfies

$$\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}(\theta_{m,\rho}) \leq \frac{K}{2}\rho^2 \varphi_{\max}(\Sigma) \frac{d_m \vee 1}{np^2} ,$$

simultaneously for all models $m \in \mathcal{M}_1$ with probability larger than $1 - \sum_{m \in \mathcal{M}_1 \setminus \emptyset} e^{-K(d_m \vee 1)/24}$. If K is chosen large enough the previous upper bound holds on an event of probability larger than $7/8$. We call K_2' such a value.

Let us now turn to the second part of the result. As previously, we decompose the difference of empirical contrasts

$$\gamma_{n,p,p}(\theta_{m,\rho}) - \gamma_{n,p,p}\left(\widehat{\theta}_{m,\rho}\right) = \overline{\gamma}_{n,p}(\theta_{m,\rho}) - \overline{\gamma}_{n,p}\left(\widehat{\theta}_{m,\rho}\right) - l\left(\widehat{\theta}_{m,\rho}, \theta_{m,\rho}\right)$$

Arguing as in the proof of Theorem 6.4, we obtain an upper bound analogous to Equation (6.60)

$$\overline{\gamma}_{n,p}(\theta_{m,\rho}) - \overline{\gamma}_{n,p}\left(\widehat{\theta}_{m,\rho}\right) \leq l(\widehat{\theta}_{m,\rho}, \theta_{m,\rho}) + \rho^2 \left\{ \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,m^2}} \frac{1}{p^2} tr\left[RD_\Sigma\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right] \right\}^2$$

The set $\mathcal{B}^{\mathcal{H}'}_{m^2,m^2}$ is defined in the proof of Lemma 6.27 but we won't really need it in what follows. Coming back to a difference of $\gamma_{n,p,p}$, we get

$$\gamma_{n,p,p}(\theta_{m,\rho}) - \gamma_{n,p,p}\left(\widehat{\theta}_{m,\rho}\right) \leq \rho^2 \left\{ \sup_{R \in \mathcal{B}^{\mathcal{H}'}_{m^2,m^2}} \frac{1}{p^2} tr\left[RD_\Sigma\left(\overline{\mathbf{Y}\mathbf{Y}^*} - I_{p^2}\right)\right] \right\}^2 .$$

We consecutively apply Lemma 6.28 and 6.29 to bound the deviation of this supremum. We therefore obtain that for any positive number α,

$$\gamma_{n,p,p}(\theta_{m,\rho}) - \gamma_{n,p,p}\left(\widehat{\theta}_{m,\rho}\right) \leq L_1(1+\alpha/2)\rho^2 \varphi_{\max}(\Sigma) \frac{d_m}{np^2} . \tag{7.20}$$

with probability larger than $1 - \exp\left[-L_2\sqrt{d_m}\left(\frac{\alpha}{\sqrt{1+\alpha/2}} \wedge \frac{\alpha^2}{1+\alpha/2}\right)\right]$. Hence, there exists some numerical constant α_0 such that the upper bound (7.20) with $\alpha = \alpha_0$ holds simultaneously for all models $m \in \mathcal{M}_1 \setminus \emptyset$ with probability larger than $7/8$. Choosing K_2 to be the supremum of K_2' and $2L_1(1+\alpha_0/2)$ allows to conclude.

□

Proof of Lemma 7.11. Thanks to the definition (7.17) of $\gamma_{n,p,p}(.)$ we obtain

$$\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}(\widehat{\theta}_{m',\rho}) = \frac{1}{p^2} \sup_{\theta' \in \Theta^+_{m',\rho}} tr\left[\left(C(\theta') - C(\theta)\right)\left(2I_{p^2} - C(\theta) - C(\theta')\right)\Sigma\overline{\mathbf{Z}\mathbf{Z}^*}\right] ,$$

where the $p^2 \times n$ matrix $\mathbf{Z}$ is defined by $\mathbf{Z} := \sqrt{\Sigma}^{-1}\mathbf{X}^{\mathbf{v}}$. We recall that the matrices Σ, $C(\theta)$ and $C(\theta')$ commute since they are jointly diagonalizable by Lemma 7.9. Let $\left(\Theta^+_{m',\rho} - \theta\right)$ be the set $\Theta^+_{m',\rho}$ translated by θ. Since $C(\theta) + C(\theta') = C(\theta + \theta')$, we lower bound the difference of $\gamma_{n,p,p}(.)$ as follows

$$\begin{aligned}
\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}(\widehat{\theta}_{m',\rho}) &= \frac{1}{p^2} \sup_{\theta' \in \left(\Theta^+_{m',\rho} - \theta\right)} 2\sigma^2 tr\left[C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right] - tr\left[C(\theta')^2\Sigma\overline{\mathbf{Z}\mathbf{Z}^*}\right] \\
&\geq \frac{\sigma^2}{p^2} \sup_{\theta' \in \left(\Theta^+_{m',\rho} - \theta\right)} \left\{2tr\left[C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right] - \varphi_{\min}^{-1}\left[I_{p^2} - C(\theta)\right] tr\left[C(\theta')^2\overline{\mathbf{Z}\mathbf{Z}^*}\right]\right\} .
\end{aligned}$$

Let us consider $\Psi_{i_1,j_1},\ldots,\Psi_{i_{d_{m'}},j_{d_{m'}}}$ a basis of the space $\Theta_{m'}$ defined in Equation (6.14). Let α be a positive number that we shall define later. We then define θ' as

$$\theta' := \varphi_{\min}\left[I_{p^2} - C(\theta)\right]\frac{\alpha}{p^2}\sum_{k=1}^{d_{m'}} tr\left[C\left(\Psi_{i_k,j_k}\right)\overline{\mathbf{Z}\mathbf{Z}^*}\right]\Psi_{i_k,j_k} \ .$$

Since θ is assumed to belong to $\Theta^+_{m',\rho}$, the parameter θ' belongs to $\left(\Theta^+_{m',\rho} - \theta\right)$ if $\varphi_{\max}[C(\theta')] \leq \varphi_{\min}\left(I_{p^2} - C(\theta)\right)$ and $\varphi_{\min}[C(\theta')] \geq -\rho + \varphi_{\max}\left(I_{p^2} - C(\theta)\right)$. The largest eigenvalue of $C(\theta')$ is smaller than $\|\theta'\|_1$ whereas its smallest eigenvalue is larger than $-\|\theta'\|_1$. Let us upper bound the l_1 norm of θ':

$$\begin{aligned}\|\theta'\|_1 &= 2\varphi_{\min}\left[I_{p^2} - C(\theta)\right]\frac{\alpha}{p^2}\sum_{k=1}^{d_{m'}}\left|tr\left[C\left(\Psi_{i_k,j_k}\right)\overline{\mathbf{Z}\mathbf{Z}^*}\right]\right| \\ &\leq 2\sqrt{\frac{\alpha}{p^2}\varphi_{\min}\left[I_{p^2} - C(\theta)\right]d_{m'}tr\left[C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right]} \ . \end{aligned} \tag{7.21}$$

Hence, θ' belongs to $\left(\Theta^+_{m',\rho} - \theta\right)$ if

$$\|\theta'\|_1 \leq \varphi_{\min}\left(I_{p^2} - C(\theta)\right) \wedge \left[\rho - \varphi_{\max}\left(I_{p^2} - C(\theta)\right)\right] \ . \tag{7.22}$$

Thus, we get the lower bound

$$\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}(\widehat{\theta}_{m',\rho}) \geq \frac{\sigma^2}{p^2}\left\{2tr\left[C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right] - \varphi^{-1}_{\min}\left[I_{p^2} - C(\theta)\right]tr\left[C(\theta')^2\overline{\mathbf{Z}\mathbf{Z}^*}\right]\right\} \ , \tag{7.23}$$

if Condition (7.22) is satisfied.

Let us now bound the deviations of the two random variables involved in (7.21) and (7.23) applying Markov's and Tchebychev's inequality. For the sake of simplicity, we assume that $d_{m'}$ is smaller than $(p^2-2p)/2$. In such a case, all the nodes in m' are different from their symmetric in Λ. We omit the proof for $d_{m'}$ larger than $(p^2-2p)/2$ because the approach is analogous but the computations are slightly more involved. Straightforwardly, we get

$$\mathbb{E}\left[tr\left(C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right)\right] = 4\alpha\varphi_{\min}\left[I_{p^2} - C(\theta)\right]\frac{d_{m'}}{n} \ ,$$

since the neighborhood m' only contains points (i,j) whose symmetric $(-i,-j)$ is different. A cumbersome but pedestrian computation leads to the upper bound

$$\mathrm{var}\left[tr\left(C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right)\right] \leq L_1\alpha^2\varphi^2_{\min}\left[I_{p^2} - C(\theta)\right]\frac{d_{m'}}{n^2} \ ,$$

where L_1 is a numerical constant. Similarly, we upper bound the expectation of $tr\left[C(\theta')^2\overline{\mathbf{Z}\mathbf{Z}^*}\right]$

$$\mathbb{E}\left[tr\left(C(\theta')^2\overline{\mathbf{Z}\mathbf{Z}^*}\right)\right] \leq L_2\alpha^2\varphi^2_{\min}\left[I_{p^2} - C(\theta)\right]\frac{d_{m'}}{n} \ .$$

Let us respectively apply Tchebychev's inequality and Markov's inequality to the variables $tr\left[C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right]$ and $tr\left[C(\theta')^2\overline{\mathbf{Z}\mathbf{Z}^*}\right]$. Hence, there exists an event Ω of probability larger than $3/4$ such that

$$\begin{aligned}&2tr\left[C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right] - \varphi^{-1}_{\min}\left[I_{p^2} - C(\theta)\right]tr\left[C(\theta')^2\overline{\mathbf{Z}\mathbf{Z}^*}\right] \geq \\ &\qquad \varphi_{\min}\left[I_{p^2} - C(\theta)\right]\frac{d_{m'}}{n}\left\{8\alpha\left(1 - \sqrt{\frac{L'_1}{d_{m'}}}\right) - \alpha^2 L'_2\right\}\end{aligned}$$

and

$$tr\left[C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right] \leq 4\alpha\varphi_{\min}\left[I_{p^2} - C(\theta)\right]\frac{d_{m'}}{n}\left(1 + \sqrt{\frac{L'_1}{d_{m'}}}\right) \ .$$

In the sequel, we assume that p is larger than some universal constant p_0, which ensures the dimension $d_{m'}$ to be larger than $4L'_1$. Gathering (7.21) with the upper bound on $tr\left[C(\theta')\overline{\mathbf{Z}\mathbf{Z}^*}\right]$ yields

$$\|\theta'\|_1 \leq 2\sqrt{2}\alpha\varphi_{\min}\left[I_{p^2} - C(\theta)\right]\frac{d_{m'}}{\sqrt{np^2}} \leq 2\sqrt{2}\alpha\varphi_{\min}\left[I_{p^2} - C(\theta)\right] \ ,$$

since $d_{m'} \leq p\sqrt{n}$. If $2\sqrt{2}\alpha$ is smaller than $1 \wedge \left[\left[\rho - \varphi_{\max}\left(I_{p^2} - C(\theta)\right)\right]\varphi_{\min}^{-1}\left[I_{p^2} - C(\theta)\right]\right]$, then Condition (7.22) is fulfilled on the event Ω and it follows from (7.23) that

$$\mathbb{P}\left\{\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}(\widehat{\theta}_{m',\rho}) \geq 4\sigma^2\varphi_{\min}\left[I_{p^2} - C(\theta)\right]\frac{d_{m'}}{np^2}\left[\alpha - \alpha^2 L_2'/4\right]\right\} \geq \frac{3}{4}\ .$$

Choosing $\alpha = \frac{2}{L_2'} \wedge \frac{\sqrt{2}}{4} \wedge \sqrt{2}\frac{\rho - \varphi_{\max}\left(I_{p^2} - C(\theta)\right)}{4\varphi_{\min}\left[I_{p^2} - C(\theta)\right]}$, we get

$$\mathbb{P}\left\{\gamma_{n,p,p}(\theta) - \gamma_{n,p,p}(\widehat{\theta}_{m',\rho}) \geq K_3\sigma^2\left[\varphi_{\min}\left[I_{p^2} - C(\theta)\right] \wedge \left[\rho - \varphi_{\max}\left(I_{p^2} - C(\theta)\right)\right]\right]\frac{d_{m'}}{np^2}\right\} \geq \frac{3}{4},$$

where K_3 is an universal constant. □

Bibliographie

[ABDJ06] Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2) :584–653, 2006.

[Aka70] Hirotugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22 :203–217, 1970.

[Aka73] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[Ald85] David J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.

[AM08] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res. (accepted)*, 2008.

[Arl07] Sylvain Arlot. *Resampling and Model selection.* PhD thesis, University Paris XI, December 2007.

[Bac08] Francis Bach. model consistent lasso estimation through the bootstrap. In *Twenty-fifth International Conference on Machine Learning (ICML)*, 2008.

[Bar02a] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6 :127–146 (electronic), 2002.

[Bar02b] Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5) :577–606, 2002.

[BBLM05] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2) :514–560, 2005.

[BCM08] Jean-Patrick Baudry, Gilles Celeux, and Jean-Michel Marin. Selecting models focussing the modeller's purpose. In *Compstat 2008 : Proceedings in Computational Statistics*. Springer-Verlag, 2008.

[BD91] Peter J. Brockwell and Richard A. Davis. *Time series : theory and methods*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1991.

[BEGd08] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9 :485–516, 2008.

[Bes75] Julian E. Besag. Statistical Analysis of Non-Lattice Data. *The Statistician*, 24(3) :179–195, 1975.

[Bes77] Julian E. Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64(3) :616–618, 1977.

[BGH08] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *Ann. Statist. (to appear)*, 2008.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1) :289–300, 1995.

[BHL03] Yannick Baraud, Sylvie Huet, and BÃ©atrice Laurent. Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*, 31(1) :225–251, 2003.

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[Bir05] Lucien Birgé. A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory*, 51(4) :1611–1615, 2005.

[BK95] Julian E. Besag and Charles Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4) :733–746, 1995.

[BK07] Peter Bühlmann and Markus Kalisch. Variable selection for high-dimensional models : partial faithful distributions, strong assocations and the PC-algorithm. Technical report, *Seminar für Statistik, ETH Zürich*, 2007.

[BL08a] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Statist. (to appear)*, 2008.

[BL08b] Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1) :199–227, 2008.

[BM75] Julian E. Besag and Patrick A. P. Moran. On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, 62(3) :555–562, 1975.

[BM97] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM98] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3) :329–375, 1998.

[BM01] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268, 2001.

[BM07] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[Bre96] Leo Breiman. Bagging predictors. *Machine Learning*, 24 :123–140, 1996.

[BRT08] Peter Bickel, Ya'acov Ritov, and Alexandre Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist. (to appear)*, 2008.

[BTW07a] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1 :169–194 (electronic), 2007.

[BTW07b] Florentina Bunea, Alexandre Tsybakov, and Martin Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4) :1674–1697, 2007.

[CDLS99] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic networks and expert systems*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 1999.

[Chi02] David Maxwell Chickering. Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.*, 2(3) :445–498, 2002.

[CL01] Guido Consonni and Valentina Leucari. Model determination for directed acyclic graphs. *The Statistician*, 50(3) :243–256, 2001.

[CP08] Emmanuel J. Candès and Yaniv Plan. Near-ideal model selection by l_1 minimization. *Ann. Statist. (to appear)*, 2008.

[Cre85] Noel A. C. Cressie. Fitting variogram models by weighted least squares. *Mathematical Geology*, 17 :563–586, 1985.

[Cre93] Noel A. C. Cressie. *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1993. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

[CT05] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12) :4203–4215, 2005.

[CT07] Emmanuel J. Candès and Terence Tao. The Dantzig selector : statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6) :2313–2351, 2007.

[CV08] Noel A. C. Cressie and Nicolas Verzelen. Conditional-mean least-squares of Gaussian Markov random fields to Gaussian fields. *Comput. Statist. Data Analysis*, 52(5) :2794–2807, 2008.

[DGR03] Petros Dellaportas, Paolo Giudici, and Gareth Roberts. Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā*, 65(1) :43–55, 2003.

[DJ94] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.

[DL93] A. Philip Dawid and Steffen L. Lauritzen. Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21(3) :1272–1317, 1993.

[dlPG99] Víctor H. de la Peña and Evarist Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

[DP04] Mathias Drton and Michael D. Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91(3) :591–602, 2004.

[DP07] Mathias Drton and Michael D. Perlman. Mutiple testing and error control in Gaussian graphical model selection. *Statist. Sci.*, 22(3) :430–449, 2007.

[DP08] Mathias Drton and Michael D. Perlman. A SINful approach to Gaussian graphical model selection. *J. Statist. Plann. Inference*, 138(4) :1179–1200, 2008.

[DS01] Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001.

[DT08] Arnak Dalalyan and Alexandre Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning*, 72(1) :39–61, 2008.

[Edw00] David Edwards. *Introduction to graphical modelling*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2000.

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2) :407–499, 2004. With discussion, and a rejoinder by the authors.

[EK08] Noureddine El Karoui. Operator norm consistent estimation of large dimensional covariance matrices. *Ann. Statist. to appear*, 2008.

[FB07] Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.*, 98(2) :227–255, 2007.

[FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2008.

[GG99] Paolo Giudici and Peter J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4) :785–801, 1999.

[GHV] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. A practical procedure for estimating gaussian graphs. In preparation.

[Gir08a] Christophe Giraud. Estimation of Gaussian graphs by model selection. *Electron. J. Stat.*, 2 :542–563, 2008.

[Gir08b] Christophe Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4) :1089–1107, 2008.

[Gra06] Robert M. Gray. *Toeplitz and Circulant Matrices : A Review*. Now Publishers, Norwell, Massachusetts, rev. edition, 2006.

[Guy87] Xavier Guyon. Estimation d'un champ par pseudo-vraisemblance conditionnelle : étude asymptotique et application au cas markovien. In *Spatial processes and spatial time series analysis (Brussels, 1985)*, volume 11 of *Travaux Rech.*, pages 15–62. Publ. Fac. Univ. Saint-Louis, Brussels, 1987.

[Guy95] Xavier Guyon. *Random fields on a network*. Probability and its Applications (New York). Springer-Verlag, New York, 1995. Modeling, statistics, and applications, Translated from the 1992 French original by Carenne Ludeña.

[GY99] Xavier Guyon and Jian-feng Yao. On the underfitting and overfitting sets of models chosen by order selection criteria. *J. Multivariate Anal.*, 70(2) :221–249, 1999.

[HLPL06] Jianhua Z. Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1) :85–98, 2006.

[HT89] Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2) :297–307, 1989.

[Ing93a] Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.*, 2(2) :85–114, 1993.

[Ing93b] Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Math. Methods Statist.*, 2(3) :171–189, 1993.

[Ing93c] Yu. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Math. Methods Statist.*, 2(4) :249–268, 1993.

[JL04] Iain M. Johnstone and Arthur Lu. Sparse principal components analysis. Technical report, Stanford university, 2004.

[Joh01] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2) :295–327, 2001.

[KB07] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8 :613–636, 2007.

[KC83] Rangasami Kashyap and Ramalingam Chellapa. Estimation and choice of neighbors in spatial-interaction models of images. *IEEE Trans. Inform. Theory*, 29(1) :60–72, 1983.

[KW00] Hirohisa Kishino and Peter J. Waddell. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Informatics*, 11 :83–95, 2000.

[Lau96] Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[LB06] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8) :3396–3410, 2006.

[LD93] Sridhar Lakshmanan and Haluk Derin. Valid parameter space for 2-D Gaussian Markov random fields. *IEEE Trans. Inform. Theory*, 39(2) :703–709, 1993.

[Leb05] Émilie Lebarbier. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal processing*, 85(4) :717–736, 2005.

[Leh86] Erich L. Lehmann. *Testing statistical hypotheses*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, second edition, 1986.

[Lep02] Vincent Lepez. *Some estimation problems related to oil reserves*. PhD thesis, University Paris XI, 2002.

[LF07] Clifford Lam and Jianqinq Fan. Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation. Technical report, Princeton University, 2007.

[LM00] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5) :1302–1338, 2000.

[LM07] Gérard Letac and Hélène Massam. Wishart distributions for decomposable graphs. *Ann. Statist.*, 35(3) :1278–1323, 2007.

[LRZ08] Elizaveta Levina, Adam J. Rothman, and Ji Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Stat.*, 2(1) :245–263, 2008.

[LW04] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2) :365–411, 2004.

[Mal73] Colin L. Mallows. Some comments on c_p. *Technometrics*, 15 :661–675, 1973.

[Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[Mat86] Bertil Matérn. *Spatial variation*, volume 36 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, second edition, 1986. With a Swedish summary.

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3) :1436–1462, 2006.

[MM08] Cathy Maugis and Bertrand Michel. Slope heuristics for variable selection and clustering via gaussian mixtures. Technical Report RR-6550, INRIA, 2008.

[Mor73] Patrick A. P. Moran. A Gaussian Markovian process on a square lattice. *J. Appl. Probability*, 10 :54–62, 1973.

[MT98] Allan D. R. McQuarrie and Chih-Ling Tsai. *Regression and time series model selection*. World Scientific Publishing Co. Inc., River Edge, NJ, 1998.

[PDR08] Franck Picard, Jean-Jacques Daudin, and Stéphane Robin. A mixture model for random graphs. *Stat. Comput.*, 18(2) :173–183, 2008.

[RBLZ08] Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2 :494–515, 2008.

[RH05] Havard Rue and Leonhard Held. *Gaussian Markov random fields*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, 2005. Theory and applications.

[Ris78] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14 :465–471, 1978.

[RJ01] Paulo J. Jr. Ribeiro and Diggle Peter J. geoR : A package for geostatistical analysis. *R-NEWS*, 1(2) :1609–3631, 2001.

[Ros85] Murray Rosenblatt. *Stationary sequences and random fields*. Birkhäuser Boston Inc., Boston, MA, 1985.

[Rov02] Alberto Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.*, 29(3) :391–411, 2002.

[RT02] Håvard Rue and Håkon Tjelmeland. Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, 29(1) :31–49, 2002.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[SGS00] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

[Shi80] Ritei Shibata. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.*, 8(1) :147–164, 1980.

[Shi81] Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1) :45–54, 1981.

[Spo96] Vladimir G. Spokoiny. Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6) :2477–2498, 1996.

[SPP+05] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721) :523–529, 2005.

[SS05] Julian Schäfer and Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association network. *Bioinformatics*, 21 :754–764, 2005.

[Tal96] Michel Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3) :505–563, 1996.

[TH02] H. To and K. Horimoto. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modelling. *Bioinformatics*, 18 :287–297, 2002.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1) :267–288, 1996.

[Tsy03] Alexandre Tsybakov. Optimal rates of aggregation. In *Proceedings of the 16th Annual Conference on Learning Theory*, volume 2777, pages 303–313. Springer-Verlag, 2003.

[vdG08] Sara van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2) :614–645, 2008.

[Vil07] Fanny Villers. *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, University Paris XI, December 2007.

[vR88] Dietrich von Rosen. Moments for the inverted Wishart distribution. *Scand. J. Statist.*, 15(2) :97–109, 1988.

[VSBH08] Fanny Villers, Brigitte Schaeffer, Caroline Bertin, and Sylvie Huet. Assessing the validity domains of graphical gaussian models in order to infer relationships among components of complex biological systems. *Stat. Appl. Genet. Mol. Biol.*, 7(2), 2008. Article 14.

[VV08a] Nicolas Verzelen and Fanny Villers. Goodness-of-fit tests for high-dimensional gaussian linear models. *Ann. Statist. (to appear)*, 2008.

[VV08b] Nicolas Verzelen and Fanny Villers. Tests for gaussian graphical models. *Comput. Statist. Data Analysis*, 2008.

[Wai07] Martin J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. Technical Report 725, Department of Statistics, UC Berkeley, 2007.

[WB06] Anja Wille and Peter Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.*, 5 :Art. 1, 34 pp. (electronic), 2006.

[WCK03] Frederick Wong, Christopher K. Carter, and Robert Kohn. Efficient estimation of covariance selection models. *Biometrika*, 90(4) :809–830, 2003.

[WE08] Amy Wagaman and Levina Elizaveta. Discovering sparse covariance structures with the isomap. *J. Comput. Graph. Statist. (to appear)*, 2008.

[WP03] Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4) :831–844, 2003.

[WYR03] Xintao Wu, Yong Ye, and Subramanian Kalpathi R. Interactive analysis of gene interactions using graphical Gaussian model. In *ACM SIGKDD Workshop on Data Mining in Bioinformatics*, volume 3, pages 63–69, 2003.

[WZV+04] Anja Wille, P. Zimmermann, E. Vranova, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann. Sparse graphical Gaussian modelling of the isoprenoid gene network in *arabidopsis thaliana*. *Genome Biology*, **5**(11), 2004.

[Yan05] Yuhong Yang. Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation. *Biometrika*, 92(4) :937–950, 2005.

[YL07] Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1) :19–35, 2007.

[Yu97] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2) :301–320, 2005.

[Zou06] Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476) :1418–1429, 2006.

[ZY06] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7 :2541–2563, 2006.

Modèles graphiques gaussiens et Sélection de modèles

Résumé : Cette thèse s'inscrit dans les domaines de la statistique non-paramétrique, de la théorie statistique de l'apprentissage et des statistiques spatiales. Son objet est la compréhension et la mise en oeuvre de méthodes d'estimation et de décision pour des modèles graphiques gaussiens. Ces outils probabilistes rencontrent un succès grandissant pour la modélisation de systêmes complexes dans des domaines aussi différents que la génomique ou l'analyse spatiale. L'inflation récente de la taille des données analysées rend maintenant nécessaire la construction de procédures statistiques valables en « grande dimension », c'est à dire lorsque le nombre de variables est potentiellement plus grand que le nombre d'observations. Trois problèmes généraux sont considérés dans cette thèse : le test d'adéquation d'un graphe à un modèle graphique gaussien, l'estimation du graphe d'un modèle graphique gaussien et l'estimation de la covariance d'un modèle graphique gaussien, ou plus généralement d'un vecteur gaussien. Suite à cela, nous étudions l'estimation de la covariance d'un champ gaussien stationnaire sur un réseau, sous l'angle de la modélisation graphique.

En utilisant le lien entre modèles graphiques et régression linéaire à plan d'expérience gaussien, nous développons une approche basée sur des techniques de sélection de modèles. Les procédures ainsi introduites sont analysés d'un point de vue non-asymptotique. Nous prouvons notamment des inégalités oracles et des propriétés d'adaptation au sens minimax valables en grande dimension. Les performances pratiques des méthodes statistiques sont illustrées sur des données simulées ainsi que sur des données réelles.

Mots-clés : Modèles graphiques, statistique spatiales, sélection de modèles, régression linéaire, vitesse minimax, adaptation, pénalisation, tests mutiples, champs de Markov, pseudo-vraisemblance.

Gaussian Graphical Models and Model Selection

Abstract : This work is linked to the theories of non-parametric statistics, statistical learning, and spatial statistics. Its goal is to provide and study statistical procedures for Gaussian graphical models. Graphical models have emerged as useful tools for modelling complex systems in many fields such as genomics or spatial analysis. The recent availability of a huge amount of data challenges us with new issues : the number of variables under study is possibly much larger than the sample size. This motivates the search for methods that remain valid in a high-dimensional setting. In this setting, three main issues are considered : the goodness of fit test of the graph of a Gaussian graphical model, the graph estimation of a Gaussian graphical model, and the covariance estimation of a Gaussian graphical model or more generally of a Gaussian vector. Furthermore, we use graphical models to study the covariance estimation of a stationary Gaussian field on a lattice.

Our approach is based on the connection between Gaussian graphical model and linear regression with Gaussian design. This connection motivates the use of model selection techniques by penalization. The procedures introduced to analyze each of the four previous issues satisfy non-asymptotic oracle inequalities and are adaptive in the minimax sense. All these results still hold in a high-dimensional setting. The practical efficiency of the procedures is assessed on simulated and real-world data.

Keywords : Graphical models, spatial statistics, model selection, linear regression, minimax rate, adaptivity, penalization, multiple testing, Markov random field, pseudo-likelihood.

AMS Classification : 62H20, 62J05, 62H11, 62M40, 62G08, 62H15, 62C20.

N° d'impression 2880
4ème trimestre 2008