Statistiques discrètes et Statistiques bayésiennes en grande dimension

Dominique Bontemps

N° d'ordre : 10073

THÈSE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS-SUD XI

Spécialité: Mathématiques

par

Dominique BONTEMPS

Statistiques discrètes et Statistiques bayésiennes en grande dimension

Soutenue le 2 décembre 2010 devant la Commission d'examen:

M.	Olivier CATONI	(Rapporteur)
Mme	Elisabeth GASSIAT	(Directrice de thèse)
M.	Pascal MASSART	
Mme	Judith ROUSSEAU	(Présidente du jury)
M.	Harry VAN ZANTEN	(Rapporteur)

Thèse préparée au
Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8628), Bât. 425
Université Paris-Sud 11
91 405 Orsay CEDEX

Résumé

Dans cette thèse de doctorat, nous présentons les travaux que nous avons effectués dans trois directions reliées : la compression de données en alphabet infini, les statistiques bayésiennes en dimension infinie, et les mélanges de distributions discrètes multivariées.

Dans le cadre de la compression de données sans perte, nous nous sommes intéressé à des classes de sources stationnaires sans mémoire sur un alphabet infini, définies par une condition d'enveloppe à décroissance exponentielle sur les distributions marginales. Un équivalent de la redondance minimax de ces classes a été obtenue. Un algorithme approximativement minimax ainsi que des a-priori approximativement les moins favorables, basés sur l'a-priori de Jeffreys en alphabet fini, ont en outre été proposés.

Le deuxième type de travaux porte sur la normalité asymptotique des distributions a-posteriori (théorèmes de Bernstein-von Mises) dans différents cadres non-paramétriques et semi-paramétriques.
Tout d'abord, dans un cadre de régression gaussienne lorsque le nombre de régresseurs augmente avec la taille de l'échantillon. Les théorèmes non-paramétriques portent sur les coefficients de régression, tandis que les théorèmes semi-paramétriques portent sur des fonctionnelles de la fonction de régression. Dans nos applications au modèle de suites gaussiennes et à la régression de fonctions appartenant à des classe de Sobolev ou de régularité C^α, nous obtenons simultanément le théorème de Bernstein-von Mises et la vitesse d'estimation fréquentiste minimax. L'adaptativité est atteinte pour l'estimation de fonctionnelles dans ces applications.
Par ailleurs nous présentons également un théorème de Bernstein-von Mises non-paramétrique pour des modèles exponentiels de dimension croissante.

Enfin, le dernier volet de ce travail porte sur l'estimation du nombre de composantes et des variables pertinentes dans des modèles de mélange de lois multinomiales multivariées, dans une optique de classification non supervisée. Ce type de modèles est utilisé par exemple pour traiter des données génotypiques. Un critère du maximum de vraisemblance pénalisé est proposé, et une inégalité oracle non-asymptotique est obtenue. Le critère retenu en pratique comporte une calibration grâce à l'heuristique de pente. Ses performances sont meilleurs que celles des critères classiques **BIC** et **AIC** sur des données simulées. L'ensemble des procédures est implémenté dans un logiciel librement accessible.

Mots-clefs : Alphabet infini dénombrable, A-priori bayésien le moins favorable, Codage universel, Compression adaptative, Compression de données sans perte, Redondance minimax, Estimation adaptative, Modèles exponentiels, Normalité asymptotique a-posteriori, Paramètre de la valeur moyenne, Statistiques bayésiennes non-paramétriques, Statistiques bayésiennes semi-paramétriques, Théorème de Bernstein-von Mises, Biostatistiques, Génotypes multilocus, Heuristique de pente, Mélange de multinomiales multivariées, Modèles à classes latentes, Sélection de modèle, Sélection de variables, Vraissemblance pénalisée.

Discrete Statistics and Bayesian Statistics in large dimension

Abstract

In this PhD thesis we present the results we obtained in three linked fields: data compression for infinite alphabets; infinite-dimensinal Bayesian Statistics; multivariate multinomial mixture models.

The first point deals with the problem of universal lossless coding on a countable infinite alphabet. It focuses on some classes of stationary memoryless sources defined by an envelope condition on the marginal distribution, namely exponentially decreasing envelope classes. An equivalent of the minimax redundancy of such classes is obtained. Then an approximately maximin prior distribution is provided and an adaptive algorithm is proposed, whose maximum redundancy is equivalent to the minimax redundancy.

The next works deal with the asymptotic normality of a-posteriori distributions (Bernstein-von Mises theorems) in several nonparametric and semiparametric frameworks.
First, in Gaussian linear regression models when the number of regressors increases with the sample size. Two kinds of Bernstein-von Mises Theorems are obtained in this framework: nonparametric theorems for the parameter itself, and semiparametric theorems for functionals of the parameter. We apply them to the Gaussian sequence model and to the regression of functions in Sobolev and C^α classes, in which we get the minimax convergence rates. Adaptivity is reached for the Bayesian estimators of functionals in our applications.
We also get a nonparametric Bernstein-von Mises theorem for increasing-dimensional exponential models.

In the last part of our work we consider the problem of estimating the number of components and the relevant variables in a multivariate multinomial mixture, in order to perform an unsupervised classification. Such models arise in particular when dealing with multilocus genotypic data. A new penalized maximum likelihood criterion is proposed, and a non-asymptotic oracle inequality is obtained. The criterion used in practice needs a calibration thanks to the slope heuristics, in an automatic data-driven procedure. Using simulated data, we found that this procedure improves the performances of the selection procedure with respect to classical criteria such as **BIC** and **AIC**. The procedures are implemented in a free-of-charge software.

Keywords: Adaptive compression, Infinite countable alphabets, Less favorable Bayes prior, Lossless data compression, Minimax redundancy, Universal coding, Adaptive estimation, Bernstein-von Mises Theorem, Nonparametric Bayesian Statistics, Posterior asymptotic normality, Semiparametric Bayesian Statistics, Biostatistics, Latent class models, Model selection, Multivariate multinomial mixtures, Multilocus genotypes, Penalized Likelihood, Slope heuristics, Variables selection.

Remerciements

Cette thèse de doctorat n'aurait pu arriver à son terme sans le soutien de nombreuses personnes, que je tiens à remercier ici.

Tout d'abord, merci à toi, Elisabeth, pour ton travail pendant ces trois ans. Tu as su être disponible et de bon conseil à tout moment. Aux moments plus difficiles, tu as été à l'écoute, tu m'as soutenu, tu m'as encouragé. Tu m'as guidé vers de nouvelles questions et m'as procuré l'occasion de collaborations enrichissantes. T'avoir eu comme directrice est une des meilleures choses qui me soit arrivées.

Special thanks to my referees, Olivier Catoni and Harry van Zanten, for having accepted to report my thesis in a relatively short time. Merci aussi aux autres membres du jury, Judith Rousseau et Pascal Massart.

Pascal, je tiens également à te remercier pour la qualité de tes cours de M2, ainsi que pour ta disponibilité à mes questions mathématiques à plusieurs occasions ces dernières années — tu te souviens peut-être de ce mail que je t'avais envoyé un soir pendant les vacances. Merci également pour les conseils d'orientation qu'avec Wendelin Werner vous nous dispensiez en M2. C'est grâce à vous que j'ai choisi cette voie et je ne le regrette pas !

Au cours de ma thèse j'ai également été accompagné par plusieurs personnes. Une place spéciale revient à Wilson Toussile. Notre travail et nos discussions ont joué un grand rôle pendant ces trois ans, et notre amitié m'est précieuse.
Nos discussions, nos groupes de travail m'ont nourri : merci à Aurélien, à Ismaël, à Judith, à Vincent, à Stéphane. Je pense également aux anciens doctorants d'Orsay dont les discussions m'ont enrichi : Nicolas, Nathalie, Jean-Patrick, Bertrand, Cathy, et bien d'autres, merci à vous.

Un travail de thèse ne se fait pas dans l'isolement, et je dois beaucoup au groupe que nous formons à Orsay. Je pense bien sûr aux responsables de l'école doctorale, aux organisateurs du séminaire des doctorants : merci en particulier à Pierre Pansu, à Valérie Lavigne pour sa grande disponibilité, à Robin, à Ramla. Mais je pense surtout à tous les doctorants, à la bonne ambiance que j'ai connu parmi vous : merci Pierre, Thi Thu, Camille, Abed, Oana, Bernardo, Emmanuel, Hatem... Un remerciement spécial à Sébastien pour avoir accepté de relire dans l'urgence des parties de mon manuscrit.

Enfin, je tiens à remercier tous les membres de ma famille pour m'avoir accompagné ces dernières années.

Table des matières

Chapitre 1

Introduction

Dans cette thèse de doctorat, nous présentons les travaux que nous avons effectués dans trois directions reliées : la compression de données en alphabet infini, les statistiques bayésiennes en dimension infinie, et les mélanges de distributions discrètes multivariées. La présente introduction est divisée en trois parties qui reprennent ces différents points. L'étude des processus à valeurs dans des espaces discrets de grande taille apparaît comme un fil directeur au long de notre travail.

Tout d'abord, nous resituons dans la section 1.1 la problématique de la compression de données sans perte, et le cas des alphabets infinis dénombrables. Tout cela est l'occasion de faire apparaître les relations entre théorie de l'information et estimation statistique aussi bien fréquentiste que bayésienne. Nos résultats sur les classes enveloppe en alphabet infini sont replacés dans ce cadre.

Si la théorie de l'information est un cadre idéal pour mettre en relation méthodes bayésiennes et résultats fréquentistes, le théorème de Bernstein-von Mises permet à son tour d'approfondir les propriétés fréquentistes des estimateurs bayésiens. C'est l'objet de la section 1.2, où nous présentons les problématiques liées à la normalité asymptotique des distributions bayésiennes a posteriori. Nous y présentons aussi les développements que nous avons obtenus dans ce domaine. Enfin, notre paragraphe 1.2.3 illustre les relations entre théorie de l'information et théorème de Berstein-von Mises. Ce dernier apparaît en effet comme un outil important pour la compréhension des phénomènes qui sous-tendent la compression de données, et pour l'obtention de bornes fines sur la redondance des algorithmes de codage.

Notre voyage à travers la théorie de l'information a également illustré l'intérêt de modélisations fines des espaces discrets en grande dimension, et des mélanges de lois discrètes. Dans la section 1.3 nous considérons d'autres mélanges de lois discrètes, dans une perspective de classification non-supervisée pour des données discrètes multivariées. Notre travail dans ce domaine a été également l'occasion de mettre en œuvre des méthodes de calcul liées à l'entropie métrique, que nous avions déjà utilisées pour la compression de données en alphabet infini.

1.1 Compression de données sans perte

1.1.1 Du codage sans perte aux probabilités

La théorie de la compression de données a été introduite par Shannon dans son célèbre article [98]. Outre la présentation que nous en faisons ci-après, on trouvera des exposés de qualité dans les livres de synthèse [29] et [30], ainsi que dans la thèse de doctorat de Garivier [47] et les notes de cours de Gassiat [49].

Soit $\mathcal{X}$ un ensemble fini ou dénombrable, qu'on appelle alphabet. On notera $\mathcal{X}^*$ l'ensemble des suites finies d'éléments de $\mathcal{X}$:

$$\mathcal{X}^* = \bigcup_{n=1}^{\infty} \mathcal{X}^n.$$

Les éléments de $\mathcal{X}$ sont appelés lettres ou symboles, et les éléments de $\mathcal{X}^*$ mots. On note $\cdot$ l'opération de concaténation entre les mots.

On désire encoder les mots de $\mathcal{X}^*$ en une suite de bits 0 ou 1. Un code sans perte f est une application injective de $\mathcal{X}^*$ dans $\{0,1\}^*$. Le but de la compression de données sans perte est de trouver des codes sans perte qui minimisent le taux de compression, c'est-à-dire le rapport de la longueur du mot de code sur la longueur initiale

$$\frac{l[f(x)]}{l(x)}$$

où $l(x) = n$ si $x \in \mathcal{X}^n$ et $l(y) = m$ si $y \in \{0,1\}^m$.

Un code est dit uniquement décodable s'il y a une unique manière de décomposer une concaténation de mots de codes :

$$f(w_1) \cdot f(w_2) \cdots f(w_n) = f(w'_1) \cdot f(w'_2) \cdots f(w'_m) \\ \Longrightarrow n = m, w_1 = w'_1, \ldots, w_n = w'_n.$$

Un cas particulier de codes uniquement décodables sont les codes préfixes, dans lesquels aucun mot n'est préfixe d'un autre mot :

$$f(w) \cdot y = f(w') \Longrightarrow y = \emptyset, w = w'.$$

Les codes préfixes sont aussi instantanément décodables : quand on arrive à la fin d'un mot de code, on le sait.

Les résultats qui suivent fondent les relations entre compression de données et statistiques. L'inégalité de Kraft [71] a d'abord été montrée pour des codes préfixes, puis étendue aux codes uniquement décodables par McMillan [79] :

Proposition 1. *Si f est un code uniquement décodable de $A \subset \mathcal{X}^*$ dans $\{0,1\}^*$, alors*

$$\sum_{w \in A} 2^{-l[f(w)]} \leq 1.$$

À toute longueur de code $l \circ f$ on peut donc associer une sous-probabilité $q_f(\cdot) = 2^{-l[f(\cdot)]}$ sur A. Réciproquement on peut construire des codes préfixes dont la longueur correspond à une sous-probabilité arbitraire :

Proposition 2. *1. Si λ est une fonction d'une partie A de $\mathcal{X}^*$ à valeurs entières qui vérifie*

$$\sum_{w\in A} 2^{-\lambda(w)} \leq 1,$$

alors il existe un code préfixe f de A dans $\{0,1\}^*$ dont λ est la longueur.

2. Plus généralement, si q est une sous-probabilité sur $A \subset \mathcal{X}^*$, il existe un code préfixe f de A dans $\{0,1\}^*$ tel que

$$\forall w \in A,\ l[f(w)] \leq q(w) + 1.$$

Remarquons que si q est une sous-probabilité, on peut construire une probabilité q' en augmentant le poids de certains éléments. Alors le code construit sur q' sera uniformément moins long que celui construit sur q. Si on ne veut pas passer par q', on peut aussi ajouter un point externe auquel on attribue la masse manquante.

Les codes de Shannon et de Huffman [64] sont des exemples de codes vérifiant la proposition 2. Le codage arithmétique [86] quant à lui réalise une inégalité très légèrement affaiblie

$$\forall w \in A,\ l[f(w)] \leq q(w) + 2.$$

Mais il présente surtout l'avantage d'une grande efficacité algorithmique, en particulier pour encoder séquentiellement les lettres d'un mot long ; la perte d'un maximum de 2 bits est considérée comme négligeable lorsque le mot à encoder devient long.

Ainsi donc, les propositions 1 et 2 établissent une quasi-équivalence entre code et loi de probabilité. Le codage arithmétique en particulier sert de support à de très nombreux résultats mathématiques en théorie de l'information, en permettant d'implémenter efficacement des techniques d'estimation statistique sous forme d'algorithmes de compression de données.

1.1.2 Universalité faible

Une source sur $\mathcal{X}$ est une distribution de probabilités $\boldsymbol{P}$ sur l'ensemble $\mathcal{X}^{\mathbb{N}}$ des mots infinis d'éléments de $\mathcal{X}$. La suite de lettres émise par une source est une variable aléatoire $\boldsymbol{X} = (X_n)_{n\geq 1}$, dont les valeurs sont indiquées par les caractères minuscules $\boldsymbol{x} = (x_n)_{n\geq 1}$. Si $\boldsymbol{P}$ est la loi de $\boldsymbol{X}$, P désigne la distribution de X_1 et P^n celle de $X_{1:n} = (X_1, \ldots, X_n)$. La source $\boldsymbol{P}$ est dite stationnaire sans mémoire si $X_1, X_2, \ldots$ sont indépendantes et identiquement distribuées.

L'inégalité de Shannon fournit une borne inférieure pour la longueur moyenne d'un code lorsque le message à encoder est produit par une source aléatoire $\boldsymbol{P}$:

Théorème 1. *Pour $n \geq 1$, soit P^n une loi de probabilité sur $\mathcal{X}^n$, et f un code uniquement décodable de $\mathcal{X}^n$ dans $\{0,1\}^*$. Alors*

$$E_{P^n}\left[l(f(X_{1:n}))\right] \geq H(P^n) = E_{P^n}\left[-\log_2 P^n(X_{1:n})\right].$$

$H(P^n)$ est appelée entropie de Shannon. Par la suite on utilisera log pour désigner le logarithme en base 2, et on conservera ln pour désigner le logarithme naturel.

Si on considère une source dans son ensemble on appelle taux d'entropie d'un processus de loi $\boldsymbol{P}$ la limite, si elle existe,

$$H_*(\boldsymbol{P}) = \lim_{n\to\infty} \frac{1}{n} H(P^n).$$

Le taux d'entropie existe en particulier lorsque la source $\boldsymbol{P}$ est stationnaire et que la première marginale P admet une entropie finie :

$$H_*(\boldsymbol{P}) = \inf_{n\to\infty} \frac{1}{n} H(P^n) = \lim_{n\to\infty} H(X_n|X_{1:n-1})$$

où l'entropie conditionnelle $H(Y|Z) = E\left[H\left(P_{Y|Z}\right)\right]$ est l'espérance de l'entropie de la loi conditionnelle de Y sachant Z.

Sous des hypothèses très faibles (source stationnaire ergodique), on a convergence presque sûre de $-\log P^n(X_{1:n})$ vers le taux d'entropie $H_*(\boldsymbol{P})$. Ce théorème a été démontré par Shannon [98] pour les sources stationnaires sans mémoire et les sources markoviennes ; Shannon l'avait également énoncé pour les sources stationnaires ergodiques, mais la preuve n'en a été faite que par McMillan [78] (pour la convergence $\mathbb{L}^1$) et par Breiman [18] (pour la convergence p.s.).

Théorème 2. *Si $\boldsymbol{P}$ est un processus stationnaire ergodique à valeurs dans un ensemble $\mathcal{X}$ fini ou dénombrable et de taux d'entropie $H_*(\boldsymbol{P})$ fini, $-\log P^n(X_{1:n})$ converge $\boldsymbol{P}$-p.s. vers $H_*(\boldsymbol{P})$*

À partir du théorème de Shannon-Breiman-McMillan et de l'inégalité de Kraft-McMillan, on obtient un pendant presque sûr à l'inégalité de Shannon :

Théorème 3. *Soit $\boldsymbol{P}$ un processus stationnaire ergodique à valeurs dans un ensemble $\mathcal{X}$ fini ou dénombrable et de taux d'entropie $H_*(\boldsymbol{P})$ fini. Si f est une suite de codes uniquement décodables sur $\mathcal{X}^n$, alors $\boldsymbol{P}$-p.s.*

$$\liminf_{n\to\infty} \frac{l\left[f\left(X_{1:n}\right)\right)]}{n} \geq H_*(\boldsymbol{P}).$$

De la proposition 2 on peut facilement déduire que l'inégalité de Shannon est fine, en considérant un code construit sur la loi de codage P^n.

Théorème 4. *Si P^n une loi de probabilité sur $\mathcal{X}^n$, il existe un code préfixe f de $\mathcal{X}^n$ dans $\{0,1\}^*$ tel que*

$$E_{P^n}\left[l(f(X_{1:n}))\right] \leq H(P^n) + 1.$$

Malheureusement construire ce code optimal nécessite de connaître exactement la loi de la source, ce qui n'est pas le cas en général. On peut dès lors se demander combien de bits un code quelconque utilisera en trop vis-à-vis de ce code optimal inconnu. C'est ce qu'on appelle la redondance.

Définition 1. *Soit $n \geq 1$.*

1. *La redondance (ou redondance moyenne) d'ordre n d'une loi de probabilité Q^n définie sur $\mathcal{X}^n$ par rapport à P^n est définie par*

$$R_n(Q^n, P^n) = E_{P^n}\left[-\log \frac{Q^n(X_{1:n})}{P^n(X_{1:n})}\right] = D(P^n, Q^n).$$

2. *Le regret (ou redondance ponctuelle) d'ordre n d'une loi de probabilité Q^n définie sur $\mathcal{X}^n$ par rapport à P^n est définie par*

$$R_n^*(Q^n, P^n) = \sup_{x_{1:n} \in \mathcal{X}^n} \{-\log Q^n(x_{1:n}) + \log P^n(x_{1:n})\}.$$

On reconnait dans la formule de la redondance la divergence de Kullback-Leibler, calculée en base 2. Aussi bien $R_n(Q^n, P^n)$ que $R_n^*(Q^n, P^n)$ s'intéressent à la différence entre la longueur de code de Q^n et celle de P^n, mais le regret considère le pire des cas alors que la redondance est la différence en moyenne entre les longueurs de code.

À partir de la notion de redondance, on peut maintenant s'intéresser à des codes qui compressent efficacement toutes les sources dans une famille de processus aléatoires. Cela nous mène dans un premier temps à la notion d'universalité faible.

Définition 2. *Soit Λ une famille de processus aléatoires à valeurs dans $\mathcal{X}$.*

1. *La classe Λ est faiblement universelle s'il existe une suite de probabilités de codage $(Q^n)_{n\geq 1}$ dont la redondance moyenne par symbole tend vers 0 pour toute source de Λ :*

$$\sup_{\boldsymbol{P} \in \Lambda} \lim_{n\to\infty} \frac{1}{n} R_n(Q^n, P^n) = 0.$$

   $(Q^n)_{n\geq 1}$ est alors appelée code faiblement universel.

2. *Une suite $\rho(n)$ est une vitesse faible de codage pour Λ si $\rho(n) = o(n)$ et s'il existe une suite de probabilités de codage $(Q^n)_{n\geq 1}$ telle que pour toute source $\boldsymbol{P} \in \Lambda$, il existe une constante $K(\boldsymbol{P})$ telle que*

$$R_n(Q^n, P^n) \leq K(\boldsymbol{P})\rho(n).$$

Pour les alphabets finis on sait que la classe des processus stationnaires ergodiques est faiblement universelle ; des codes comme ceux de Lempel-Ziv [124, 125] réalisent la définition et sont dits universels. En revanche Shields [101] a prouvé que la classe des processus stationnaires ergodiques est trop grande pour admettre une vitesse faible.

Pour les alphabets infinis la situation est bien différente : même la classe des processus stationnaires sans mémoire n'y est pas faiblement universelle. Kieffer [67] a démontré une condition nécessaire et suffisante, simplifiée par [60, 61], pour qu'une classe de sources stationnaires soit faiblement universelle :

Théorème 5. *Une classe Λ de sources stationnaires sur un alphabet dénombrable $\mathcal{X}$ est faiblement universelle si et seulement si d'une part toutes les sources admettent des marginales d'entropie finie*

$$\forall \boldsymbol{P} \in \Lambda,\ H(P) < \infty,$$

et si d'autre part il existe une probabilité Q sur $\mathcal{X}$ telle que

$$\forall \boldsymbol{P} \in \Lambda,\ D(P, Q) < \infty.$$

En particulier l'existence d'un codage faiblement universel ne dépend pas de la structure de dépendance des lois de Λ.

1.1.3 Universalité forte

La notion d'universalité forte apparaît lorsqu'on s'intéresse à des algorithmes universels qui atteignent leur vitesse de codage uniformément sur une classe. Cette notion est reliée à l'approche minimax, qui cherche la stratégie de codage qui se comporte le mieux pour la pire des sources de la classe.

Définition 3. *1. La classe Λ est fortement universelle s'il existe une suite de probabilités de codage $(Q^n)_{n\geq 1}$ dont la redondance moyenne par symbole tend vers 0 uniformément sur Λ :*

$$\lim_{n\to\infty} \sup_{\boldsymbol{P}\in\Lambda} \frac{1}{n} R_n(Q^n, P^n) = 0.$$

2. La redondance minimax d'ordre n de la classe Λ est définie par

$$R_n(\Lambda) = \inf_{Q^n} \sup_{\boldsymbol{P}\in\Lambda} R_n(Q^n; P^n).$$

La redondance minimax renormalisée $R_n(\Lambda)/n$ apparaît donc comme la vitesse forte minimale d'une classe fortement universelle.

De nombreux auteurs (par exemple [22, 40, 42, 120]) approfondissent quant à eux l'approche trajectorielle plus contraignante du regret minimax.

Définition 4. *Le regret minimax d'ordre n de la classe Λ est définie par*

$$R_n^*(\Lambda) = \inf_{Q^n} \sup_{\boldsymbol{P}\in\Lambda} R_n^*(Q^n; P^n).$$

Le regret minimax peut servir à étudier la redondance minimax grâce à l'inégalité $R_n(\Lambda) \leq R_n^*(\Lambda)$. Mais il présente aussi l'avantage que lorsqu'il est fini, on connait explicitement le code qui l'atteint : c'est la distribution *Normalized Maximum Likelihood* (NML) de Shtarkov [102].

Définition 5. *Soit Λ une classe de sources, et $n \geq 1$ fixé. Étant donné $x_{1:n} \in \mathcal{X}^n$, on note $\widehat{P}_{x_{1:n}} = \sup_{\boldsymbol{P}\in\Lambda} P^n(x_{1:n})$ le maximum de vraisemblance en $x_{1:n}$. Si $\sum_{x_{1:n}\in\mathcal{X}^n} \widehat{P}_{x_{1:n}} < \infty$, le maximum de vraisemblance normalisé est la probabilité sur $\mathcal{X}^n$ définie par*

$$NML_n(x_{1:n}) = \frac{\widehat{P}_{x_{1:n}}}{\sum_{y_{1:n}\in\mathcal{X}^n} \widehat{P}_{y_{1:n}}}.$$

On vérifie en outre facilement qu'alors le regret minimax n'est autre que

$$R_n^*(\Lambda) = \log \sum_{x_{1:n}\in\mathcal{X}^n} \sup_{\boldsymbol{P}\in\Lambda} P^n(x_{1:n}). \tag{1.1}$$

Cependant $(NML_n)_n$ n'est pas une suite consistante de lois de probabilité : NML_{n-1} n'est pas la marginale de NML_n pour les $n-1$ premières coordonnées.

Dans une toute autre direction, l'approche bayésienne s'oppose à l'approche minimax en ce que, au lieu de regarder le pire des cas, elle pondère chaque source selon un a priori. Les classes de sources sont munies de la topologie de la convergence faible, et de la tribu borélienne associée.

Définition 6. *Soit Λ une classe de sources, et μ une distribution de probabilité sur Λ. La redondance bayésienne de la classe Λ par rapport à l'a priori μ est définie par*

$$R_{n,\mu}(\Lambda) = \inf_{Q^n} \int_\Lambda R_n(Q^n; P^n)\, \mathrm{d}\mu(\boldsymbol{P}).$$

Une stratégie de codage Q^n qui atteint cette valeur est appelé stratégie de Bayes.

Il n'y a en réalité qu'une unique stratégie de Bayes, c'est le mélange bayésien

$$M_{n,\mu}(x_{1:n}) = \int_\Lambda P^n(x_{1:n})\, \mathrm{d}\mu(\boldsymbol{P}). \tag{1.2}$$

Dans le jeu bayésien $\boldsymbol{P}$ est elle-même une variable aléatoire de loi μ, tandis que $M_{n,\mu}$ est la loi marginale de $X_{1:n}$. La redondance bayésienne peut alors s'écrire comme l'information mutuelle entre $\boldsymbol{P}$ et $X_{1:n}$

$$I(\mu, X_{1:n}) = D\left(P_{(\boldsymbol{P},X_{1:n})}, \mu \otimes M_{n,\mu}\right), \tag{1.3}$$

où $P_{(\boldsymbol{P},X_{1:n})}$ désigne la loi du couple $(\boldsymbol{P}, X_{1:n})$ lorsque $\boldsymbol{P}$ est tiré selon μ et la loi conditionnelle de $X_{1:n}$ sachant $\boldsymbol{P}$ est $\boldsymbol{P}$ lui-même.

Dans l'ignorance d'un a priori pertinent sur Λ, on s'intéresse à l'a priori le moins favorable, appelé aussi a priori le moins informatif ou a priori de référence : c'est l'approche maximin. La redondance maximin est définie par

$$\sup_\mu R_{n,\mu}(\Lambda).$$

Le fait que la perte maximin soit un minorant de la perte minimax est une propriété mathématique très générale. Il suffit dès lors d'exhiber un a priori quelconque pour obtenir une minoration de la redondance minimax. Ce qui est remarquable lorsqu'on prend la divergence de Kullback-Leibler comme fonction de perte, c'est que perte maximin et perte minimax sont égales. Le théorème qui suit peut être trouvé dans Gallager [46] et dans Davisson et Leon-Garcia [33]. Haussler [62] a montré qu'il s'étend à toutes les classes de processus stationnaires ergodiques sur un espace métrique séparable et complet.

Théorème 6. *Soit Λ une classe de sources, et $n \geq 1$. Alors*

$$R_n(\Lambda) = \sup_\mu R_{n,\mu}(\Lambda),$$

où le supremum est pris sur toutes les mesures de probabilité sur Λ. Si cette redondance est finie, la borne minimax est en outre atteinte par un mélange de Bayes M_{n,μ_n^*}.

L'a priori le moins favorable a été identifié par Clarke et Barron [27] comme étant asymptotiquement l'a priori de Jeffreys dans le cadre très large des classes paramétriques régulières de dimension finie. Pour la classe des sources stationnaires sans mémoire en alphabet fini de cardinal k, l'a priori de Jeffreys coïncide avec la distribution de Dirichlet $D(1/2, 1/2, \ldots, 1/2)$ sur le simplex de $\mathbb{R}^k$; le code correspondant est aussi appelé mélange de Krichevsky-Trofimov, du nom des auteurs [72] qui ont introduit en compression de données les a priori de Dirichlet et ont étudié en premier leur propriétés. Xie et Barron [119, 120] ont montré que la stratégie bayésienne associée à l'a priori

de Jeffreys était asymptotiquement maximin sur la classe des sources sans mémoire en alphabet fini, mais qu'elle n'est pas asymptotiquement minimax — bien qu'une légère modification le soit. Ces outils nous serviront de base lorsque nous nous intéresserons à des classes de sources stationnaires sans mémoire en alphabet infini.

En passant par la redondance bayésienne, de nombreuses approximations de la redondance minimax ont pu être proposées pour différentes classes de sources en alphabet fini, et des algorithmes de compression ont été proposés. En particulier pour la classe des sources stationnaires sans mémoire divers papiers [7, 21, 40, 72, 103, 119, 120] sont arrivés à des approximations de plus en plus précises de la redondance minimax

$$R_n(\Lambda) = \frac{k-1}{2}\log\frac{n}{2\pi e} + \log\frac{\Gamma(1/2)^k}{\Gamma(k/2)} + o(1),$$

où k est le cardinal (fixé) de l'alphabet et $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,\mathrm{d}t$.
La classe de toutes les sources stationnaires sans mémoire sur un alphabet de cardinal k est un cas particulier de classe paramétrique de dimension $k-1$ sur un alphabet éventuellement plus grand ; pour ce dernier type de classes de sources, des approximations égales au premier ordre de la redondance minimax ont été obtenus sous diverses conditions par Clarke et Barron [6, 26, 27]. Dans [4, 32, 72, 118] d'autres résultats semblables sont encore disponibles pour les classes de chaînes de Markov de différents ordres, et les classes à arbres de contexte fini. Enfin Rissanen [85] avait obtenu une minoration valable pour une classe paramétrique Λ suffisamment régulière de dimension k sur un alphabet arbitraire :

$$\limsup_{n\to\infty} \frac{1}{\log n} R_n(\Lambda) \geq \frac{k}{2}.$$

Un phénomène remarquable dans toutes ces familles paramétriques, c'est que la redondance minimax et le regret minimax sont équivalentes. De fait, plusieurs de ces résultats sont obtenus d'une part en minorant la redondance bayésienne d'un a priori bien choisi, et d'autre part en majorant le regret minimax grâce à son expression explicite (1.1). On obtient alors un encadrement du type

$$\frac{d}{2}\log n + O(1) \leq R_n(\Lambda) \leq R_n^*(\Lambda) \leq \frac{d}{2}\log n + O(1)$$

où d est la dimension de la classe considérée (voir [85, 116, 119]).

Un cas particulier est celui des classes de sources stationnaires sans mémoire de distribution monotone sur un alphabet fini de taille variable k, étudié par Shamir [96]. La redondance minimax est alors équivalente à $\frac{1}{2}\log(n/k^3)$ si $k = o(n^{1/3})$, et en $O(n^{1/3+\epsilon})$ si $k = O(n)$.

Les approximations de la redondance minimax deviennent moins précises lorsqu'on s'intéresse à des classes de sources non-paramétriques, même en alphabet fini. Un exemple typique est celui de la classe des processus de renouvellement. Pour cette classe de sources Csiszár et Shields [31] ont obtenu des minorations et des majorations de la forme

$$c\sqrt{n} \leq R_n(\Lambda) \leq R_n^*(\Lambda) \leq C\sqrt{n}$$

avec c et C des constantes. Flajolet et Szpankowski [43] ont amélioré en

$$R_n^*(\Lambda) = \frac{2}{\ln 2}\sqrt{\left(\frac{\pi^2}{6} - 1\right) n} + O(\log n).$$

Pour les processus de renouvellement stationnaire markoviens d'ordre r, [31] fournit l'encadrement

$$cn^{\frac{r+1}{r+2}} \leq R_n(\Lambda) \leq R_n^*(\Lambda) \leq Cn^{\frac{r+1}{r+2}}.$$

Dans ces situations on ne sait pas si redondance minimax et regret minimax coïncident au premier ordre.

1.1.4 Classes enveloppe en alphabet infini

L'utilisation d'alphabets infinis permet de s'intéresser à diverses situations où la taille de l'alphabet est inconnue, ou bien grande vis-à-vis de la longueur des messages qui doivent être encodés. Ces situations peuvent par exemple se rencontrer en linguistique (avec l'alphabet qui correspond aux mots d'une langue donnée), à la compression sans perte pour des codecs multimédia, etc. Par simplicité nous considérerons ici que l'alphabet est l'ensemble des entiers strictement positifs $\mathcal{X} = \mathbb{N}_*$.

La difficulté des alphabets infinis vient de ce que les classes de sources usuelles deviennent trop grandes pour admettre des codes même faiblement universels. La condition de Kieffer (théorème 5) illustre ce problème. Une difficulté à souligner est qu'une probabilité sur un espace discret dénombrable n'a pas nécessairement une entropie finie ; dès lors il n'existe aucun code dont la longueur de code moyenne est finie !

Une première façon de contourner le problème est le codage par motif, introduit par Äberg, Shtarkov et Smeets [1], puis objet de nombreuses études par Shamir, Orlitsky, Santhanam et divers auteurs [37, 65, 81–83, 91–95, 97]. Cette approche consiste à ne transmettre que la structure du message, c'est-à-dire les répétitions en omettant les valeurs des symboles.

Si cependant on désire conserver la notion usuelle de redondance et encoder le message dans son ensemble, il convient d'étendre la notion de dimension pour conserver un contrôle sur la richesse des classes non-paramétriques. Remarquons en effet que les résultats cités précédemment font apparaître une dépendance linéaire de la redondance minimax en la dimension paramétrique des classes de sources considérées. L'entropie métrique pour la distance de Hellinger se révèle être la notion pertinente dans ce cadre, et Haussler et Opper [63] ainsi que Barron et Yang [9] l'ont parallèlement utilisée avec succès pour caractériser la redondance de classes non-paramétriques de sources. L'entropie métrique se révèle de fait un outil très puissant dans différents domaines des statistiques, et nous aurons plusieurs fois l'occasion de la croiser au long de cette thèse. Dans le livre de Massart [75] on pourra en trouver différentes applications, et de nombreuses références.

Pour des classes de sources stationnaires sans mémoire d'entropie finie, Boucheron, Garivier et Gassiat [16] ont relié l'existence d'un regret minimax fini à la présence d'une enveloppe intégrable sur les premières marginales de la classe :

$$R_n^*(\Lambda) < \infty \Longleftrightarrow \sum_{k\geq 1} \sup_{\boldsymbol{P}\in\Lambda} P(k) < \infty.$$

La preuve repose sur la sous-additivité de la redondance minimax et du regret minimax pour de telles classes de sources.

À partir de là il devient naturel de considérer des classes enveloppe de sources sans mémoire, qui rassemblent toutes les sources dont les marginales d'ordre 1 sont dominées par une fonction f :

$$\Lambda_f = \{\boldsymbol{P} \ : \ \boldsymbol{P} \text{ est stationnaire et sans mémoire et } \forall k \geq 1, \ P(k) \leq f(k)\}.$$

[16] a montré que pour les classes enveloppe, la redondance minimax et le regret minimax sont finis simultanément :

$$R_n(\Lambda_f) < \infty \Longleftrightarrow R_n^*(\Lambda_f) < \infty \Longleftrightarrow \sum_{k\geq 1} f(k) < \infty.$$

Nous nous intéressons particulièrement à deux types de classes enveloppe, définis par la décroissance exponentielle ou polynomiale de l'enveloppe.

Définition 7. *1. Soit $\alpha > 0$ et $C > 0$ tels que $C\sum_{k\geq 1} k^{-\alpha} \geq 2^\alpha$. La classe enveloppe à décroissance polynomiale $\Lambda_{C\cdot^{-\alpha}}$ est la classe des sources stationnaires sans mémoire définie par la fonction enveloppe $k \mapsto Ck^{-\alpha}$.*

2. Soit $\alpha > 1$ et $C > e^{2\alpha}$ des réels. La classe enveloppe à décroissance exponentielle $\Lambda_{Ce^{-\alpha\cdot}}$ est la classe des sources stationnaires sans mémoire définie par la fonction enveloppe $k \mapsto Ce^{-\alpha k}$.

Ces conditions d'enveloppe imposent en particulier des restrictions sur les queues des distributions des éléments de la classe. Il faut remarquer que les classes enveloppe imposent une borne supérieure aux probabilités des symboles, mais pas de bornes inférieures. En utilisant des bornes sur le regret minimax qui font intervenir la suite des restes de la série $\sum_{k\geq 1} f(k)$, [16] obtient les encadrements suivants

$$A_{C,\alpha} n^{1/\alpha} \leq R_n\left(\Lambda_{C\cdot^{-\alpha}}\right) \leq R_n^*\left(\Lambda_{C\cdot^{-\alpha}}\right) \leq \left(\frac{2Cn}{\alpha - 1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1) \tag{1.4}$$

$$\frac{1+o(1)}{8\alpha \log e} \log^2 n \leq R_n\left(\Lambda_{Ce^{-\alpha\cdot}}\right) \leq R_n^*\left(\Lambda_{Ce^{-\alpha\cdot}}\right) \leq \frac{1}{2\alpha \log e} \log^2 n + O(1)$$

où $A_{C,\alpha}$ est une constante explicite dépendant de C et α.

Ces mêmes auteurs proposent également un algorithme séquentiel linéaire en temps, dont la redondance approche la redondance minimax des différentes classes enveloppe à des facteurs logarithmiques près. Son principe de fonctionnement est l'encodage séparé des symboles grands, en utilisant un code d'Elias [41]. La partie principale du message est la suite des symboles censurés, à laquelle est appliqué un codage du type mélange de Krichevsky-Trofimov. Le point délicat de cet algorithme est la détermination du seuil de censure à utiliser, qui doit être choisi en fonction du paramètre α des différentes classes enveloppe (cependant une version adaptative est aussi proposée).

Dans le chapitre 2 nous améliorons l'approximation de la redondance minimax des classes enveloppe à décroissance exponentielle, en obtenant un équivalent

$$R_n\left(\Lambda_{Ce^{-\alpha\cdot}}\right) \sim \frac{1}{4\alpha \log e} \log^2 n. \tag{1.5}$$

Pour ce faire nous nous appuyons sur les méthodes d'entropie métrique de Haussler et Opper [63]. D'autre part nous sommes en mesure de proposer des a priori approximativement maximin, en adaptant l'a priori de Jeffreys pour les alphabets finis. Enfin

nous améliorons l'algorithme de [16] en un algorithme adaptatif, pour lequel le seuil de censure est choisi automatiquement, et dont nous sommes capables de montrer qu'il atteint la redondance minimax à un équivalent près pour toutes les classes enveloppe à décroissance exponentielle. Considérés ensemble, ces deux derniers points constituent une deuxième preuve indépendante de la relation (1.5).
Une version abrégée [15] du chapitre 2 a été accepté pour publication à *IEEE Transactions on Information Theory*. Après la soumission de cet article, nous avons découvert une façon algorithmiquement très simplifiée de mettre en œuvre le principe du codage par censure qui est appliqué aussi bien dans [16] que dans [15] ; cette simplification fait l'objet de l'annexe 2.D du chapitre 2.

Perspectives. Nous aimerions étendre à d'autres classes de sources le travail que nous avons effectué pour les classes enveloppe à décroissance exponentielle, en considérant en particulier les classes à décroissance polynomiale. Les méthodes utilisées dans le premier cas ne nous permettent pour le moment pas de trouver un équivalent de la redondance minimax des classes polynomiales ; elles font même moins bien que l'encadrement (1.4). Elles ont cependant une marge d'amélioration, et on peut en particulier espérer que des a-priori construits plus finement à partir des distributions de Dirichlet permettront d'obtenir des minorations améliorées de la redondance maximin.

D'autre part, nous aimerions passer à des structures de dépendance différentes, en particulier à des dépendances markoviennes. Pour garder des classes de sources d'une complexité raisonnable, qui admettent des vitesses de compression, nous pensons à des situations où les lettres se regroupent en un nombre fini de clusters : les probabilités de transition ne dépendraient alors du ou des lettres précédentes que par l'intermédiaire du cluster auquel la lettre appartient. En outre ces probabilités de transition vérifieraient eux-même une condition d'enveloppe.

Nous désirons également approfondir l'aspect expérimental et les connexions avec des domaines plus appliqués. Nous voudrions être en mesure de mieux connaître les types de sources en alphabet grand que l'on y rencontre.

Enfin, le désir d'une compréhension plus fine des classes de sources en dimension infinie a été l'une des motivations qui nous ont poussés vers l'étude du théorème de Bernstein-von Mises dans des situations non-paramétriques. Nous revenons dans le paragraphe 1.2.3 de la prochaine section sur les relations profondes qui existent entre redondance et théorème de Bernstein-von Mises.

1.2 Le théorème de Bernstein-von Mises

1.2.1 Modèles statistiques et paradigme bayésien

Dans un cadre statistique très général, on considère un espace mesurable $(\Omega, \mathcal{A})$, une mesure de probabilité P_0 inconnue sur $(\Omega, \mathcal{A})$, et une observation $\boldsymbol{X}$ tirée selon la loi P_0. Le rôle du statisticien sera d'estimer, à partir de $\boldsymbol{X}$, une certaine caractéristique $g(P_0)$ de la loi inconnue. Le statisticien fréquentiste mesurera la performance d'un estimateur

$\widehat{g}$, mesurable par rapport à $\boldsymbol{X}$, par un risque de la forme

$$R_{P_0}(\widehat{g}) = E_{P_0}\left[d(\widehat{g}, g(P_0))\right]$$

où $d(\cdot,\cdot)$ est ce qu'on appelle la fonction de perte.

Dans le but d'estimer la quantité d'intérêt $g(P_0)$, le statisticien établit un modèle statistique $(P_\theta)_{\theta\in\Theta}$ à partir de diverses hypothèses venant de la réalité physique qui est modélisée, ou de contraintes mathématiques. Plusieurs situations peuvent se rencontrer.

- Le modèle paramétrique : Θ est un domaine de $\mathbb{R}^k$. Diverses hypothèses de régularité sur la dépendance de P_θ vis-à-vis de θ seront faites selon les situations. Supposons en particulier qu'il existe une mesure ν dominante commune au modèle, et que le modèle est *différentiable en moyenne quadratique* pour tout $\theta \in \Theta$. Cela signifie qu'il existe une fonction score $\dot{l_\theta}(\cdot)$ à valeurs dans $\mathbb{R}^k$ et de carré intégrable sous P_θ, telle que

$$E_{P_\theta}\left[\left(\sqrt{\frac{\mathrm{d}P_{\theta+h}(\boldsymbol{X})}{\mathrm{d}P_\theta(\boldsymbol{X})}} - 1 - \frac{1}{2}h^T\dot{l_\theta}(\boldsymbol{X})\right)^2\right] = o\left(\|h\|^2\right),$$

  où $\mathrm{d}P_\theta$ désigne la densité de P_θ vis-à-vis de ν. L'information de Fisher $I(\theta)$ est alors définie comme la matrice $E_{P_\theta}\left[\dot{l_\theta}(\boldsymbol{X})\dot{l_\theta}^T(\boldsymbol{X})\right]$.
  La démarche fréquentiste usuelle pour estimer $g(P_0)$ dans ce cadre sera d'exhiber dans un premier temps un estimateur $\widehat{\theta}$ de sorte que $P_{\widehat{\theta}}$ approxime au mieux P_0 (selon un critère à définir) ; dans un deuxième temps on choisira $g\left(P_{\widehat{\theta}}\right)$ comme estimateur plug-in de $g(P_0)$.
  Parmi les estimateurs fréquemment considérés, il convient de citer en particulier l'estimateur du maximum de vraisemblance (MLE, pour *Maximum Likelihood Estimator*), défini par

$$\widehat{\theta}^{MLE} = \arg\max_{\theta\in\Theta} \mathrm{d}P_\theta(\boldsymbol{X})$$

  si cette quantité existe.
  Enfin, le modèle $(P_\theta)_{\theta\in\Theta}$ est dit mal spécifié si $P_0 \notin \{P_\theta : \theta \in \Theta\}$. Un modèle mal spécifié peut en particulier présenter l'avantage d'être de dimension bien plus petite. Cela produit des estimateurs $g\left(P_{\widehat{\theta}}\right)$ dont la variance est bien plus faible, ce qui peut compenser l'augmentation inévitable du biais $E\left[g\left(P_{\widehat{\theta}}\right)\right] - g(P_0)$ — c'est ce qu'on appelle l'équilibre biais-variance. Ainsi on réduira en particulier le risque quadratique

$$E\left[\left(g\left(P_{\widehat{\theta}}\right) - g(P_0)\right)^2\right] = E\left[\left(g\left(P_{\widehat{\theta}}\right) - E\left[g\left(P_{\widehat{\theta}}\right)\right]\right)^2\right] + \left(E\left[g\left(P_{\widehat{\theta}}\right)\right] - g(P_0)\right)^2.$$

- Le modèle non-paramétrique : Θ est un espace infini-dimensionnel, et on suppose généralement que le modèle est bien spécifié. Dans ce cadre le statisticien développera généralement des méthodes qui visent à estimer directement $g(P_0)$ sans passer par une estimation du paramètre θ.
  Le modèle semi-paramétrique est un cas particulier du modèle non-paramétrique, dans lequel la quantité d'intérêt $g(P_\theta)$ est à valeurs dans un espace paramétrique. Cela s'oppose aux situations non-paramétriques où la quantité d'intérêt est elle-même non-paramétrique (par exemple la loi P_0 elle-même). Un cas particulier

de modèles semi-paramétriques est celui où θ est en fait un couple (η, f), avec η le paramètre d'intérêt fini-dimensionnel, et f un paramètre de nuisance infini-dimensionnel.

– Les modèles de type crible. Le terme crible (*sieve* en anglais) peut désigner un ensemble fini qui sert à approximer un ensemble plus grand — on parle alors de crible fini. Nous utilisons ici un second sens du mot crible, pour désigner des familles de modèles paramétriques dont la dimension n'est pas bornée. L'idée qui les sous-tend est qu'il faut s'adapter à la complexité de la vraie loi P_0, en évitant les modèles de dimension trop grande. Dans une optique d'approximation, chacun des modèles peut être mal-spécifié : lorsque la dimension du modèle augmente la distance à P_0 diminue (en divergence de Kullback-Leibler ou en distance de Hellinger par exemple). Typiquement la dimension k_n du modèle sélectionné croîtra avec la taille n de l'observation $\boldsymbol{X}$. Du fait de la dimension croissante les résultats associés aux modèles crible rentrent dans la famille des résultats non-paramétriques.

Signalons également les modèles exponentiels, pour lesquels la densité P_θ des observations est proportionnelle à une expression exponentielle de la forme $\exp\{\theta^T t(\boldsymbol{X})\}$ (pour la paramétrisation en paramètre dit naturel ; des changements de paramètres sont possibles). Ces modèles seront paramétriques ou non-paramétrique selon que θ est de dimension finie ou dans $\ell^2(\mathbb{R})$.

Le statisticien bayésien quant à lui reprend ces différents modèles, mais y ajoute une nouvelle couche d'aléa. Munissons Θ de la topologie de la convergence faible et de la tribu borélienne associée. Alors le statisticien bayésien proposera une distribution de probabilité μ sur Θ, appelée distribution *a priori*, qui cherchera à intégrer les informations dont il dispose sur les valeurs attendues du paramètre. La mesure de la performance d'un estimateur ne sera plus le risque calculé pour un choix fixé de la loi P_0. Ce sera désormais le risque bayésien

$$R_\mu(\widehat{g}) = \int_\Theta R_P(\widehat{g})\,\mathrm{d}\mu(P).$$

Le MLE n'est pas pertinent dans ce cadre, on s'intéressera plutôt à des estimateurs du type $E_\mu[\theta|\boldsymbol{X}]$. Plus généralement on considèrera la distribution *a posteriori*, qui est la loi conditionnelle définie par la formule de Bayes

$$\mathrm{d}\mu(\theta|\boldsymbol{X}) = \frac{\mathrm{d}P_\theta(\boldsymbol{X})\,\mathrm{d}\mu(\theta)}{\int_\Theta \mathrm{d}P_\nu(\boldsymbol{X})\,\mathrm{d}\mu(\nu)}.$$

Pour le bayésien le choix de la distribution a priori s'avèrera particulièrement critique. Ce choix peut être guidé par des considérations mathématiques (a priori conjugués par exemple, pour lesquels la distribution a posteriori est explicitement connue et du même type). La recherche des a priori les moins informatifs ou les moins favorables, associés à une approche maximin, aura son intérêt propre — nous l'avons illustré dans le paragraphe 1.1.3 dans le cadre de la compression de données.

1.2.2 Approche fréquentiste des méthodes bayésiennes : de la consistance à la normalité asymptotique

Ces dernières années ont vu de nombreux développements de la théorie bayésienne non-paramétrique fréquentiste. Cette théorie étudie les propriétés fréquentistes des méthodes bayésiennes non-paramétriques. On peut en particulier distinguer plusieurs types des résultats, qui forment trois étapes d'une même démarche. Les premiers portent sur la consistance des distributions a posteriori vers la vraie loi P_0. Les suivants étudient les vitesses de convergence, et sont particulièrement nombreux. Les derniers établissent la normalité asymptotique de la distribution a posteriori autour de l'estimateur du maximum de vraisemblance, sous des hypothèses adéquates : ce sont les théorèmes de Bernstein-von Mises ou théorèmes de la limite centrale bayésiens.

La consistance et les vitesses de convergence. Si P et Q sont deux mesures de probabilité de densités respectives $\mathrm{d}P$ et $\mathrm{d}Q$ par rapport à une même mesure ν, la distance de Hellinger entre P et Q est définie par

$$\mathbf{h}^2(P,Q) = \int \left(\sqrt{\mathrm{d}P(x)} - \sqrt{\mathrm{d}Q(x)}\right)^2 \mathrm{d}\nu(x).$$

La consistance de la distribution a posteriori pour la distance de Hellinger s'écrit

$$\forall \varepsilon > 0,\ \mu\left(\mathbf{h}(P_\theta, P_0) > \varepsilon | \boldsymbol{X}\right) \to 0 \text{ lorsque } n \to \infty.$$

Une telle consistance a été démontrée ou infirmée sur des exemples non-paramétriques précis par Diaconis et Freedman [38,39]. Des résultats génériques ont ensuite été obtenus dans [8,54]. Les hypothèses de ces résultats sont doubles : l'a priori doit charger un voisinage de P_0 au sens de la divergence de Kullback-Leibler ; la richesse du modèle doit être contrôlée en termes d'entropie métrique pour la distance de Hellinger. Cette dernière hypothèse permet la construction de cribles adaptés et de tests uniformément consistants. Walker [113, 114] a proposé des méthodes alternatives pour établir la consistance de l'a posteriori.

Ces mêmes méthodes d'entropie métrique ont été raffinées et ont permis l'établissement de vitesses de convergence dans de nombreuses situations. Une suite ε_n qui converge vers 0 est une vitesse de convergence en distance de Hellinger pour la distribution a posteriori si, pour un réel $M > 0$ suffisamment grand,

$$\mu\left(\mathbf{h}(P_\theta, P_0) > M\varepsilon_n | \boldsymbol{X}\right) \to 0 \text{ lorsque } n \to \infty.$$

Ghosal, Ghosh et van der Vaart [55] ont obtenu des résultats généraux de vitesse de convergence lorsque l'observation $\boldsymbol{X}$ est un vecteur $(X_1, \ldots, X_n)$ dont les coordonnées X_i sont indépendantes et identiquement distribuées, avec plusieurs applications, en particulier aux modèles log-splines qui rentrent dans la famille des modèles exponentiels de dimension croissante (la log-densité est approximée par des splines). Shen et Wasserman [100] ont obtenu le même type de résultats dans un travail indépendant ; les modèles crible sont aussi considérés parmi leurs applications. Ghosal et van der Vaart [58] considèrent des observations qui ne sont pas indépendantes et identiquement distribuées, avec diverses applications non-paramétriques ou en dimension croissante. Enfin Kleijn

et van der Vaart [70] s'intéressent à l'influence de la mis-spécification dans un cadre assez général.

Parmi les a priori qui se sont révélés d'une grande efficacité pour l'obtention de vitesses performantes de convergence a posteriori, il convient de signaler les processus gaussiens [20, 35, 107–109]. Ghosal et van der Vaart [57, 59] utilisent des mélanges de Dirichlet de lois gaussiennes comme a priori, pour estimer dans un premier temps des densités de la forme mélange de lois gaussiennes, et dans un second temps des densités régulières ; les vitesses de convergence obtenues correspondent ou sont proches des vitesses fréquentistes minimax d'estimation sur ces classes. [56, 123] considèrent des a priori hiérarchiques : dans une situation où une collection de modèles est disponible, un a priori est choisi pour chaque modèle, et à un niveau supérieur un a priori porte sur l'index du modèle. Cela permet en particulier de passer des modèles crible à des modèles non-paramétriques. Scricciolo [89] considère des modèles exponentiels de dimension croissante, et obtient les vitesses minimax d'estimation de la log-densité sur des classes de Sobolev périodiques. [90] obtient également les vitesses de convergence non-paramétriques pour l'estimation de densité par des histogrammes.

Concernant les méthodes de preuve, Walker et al. [115] proposent des variations techniques pour démontrer leurs résultats non-paramétriques. Xing [121, 122] obtient ses résultats non-paramétriques en utilisant l'α-entropie de Hausdorff.

La normalité asymptotique a posteriori. Dans cette thèse nous nous intéressons plus particulièrement à l'étape suivante dans la théorie : l'obtention de théorèmes de Bernstein-von Mises. Ce type de théorème établit, sous des hypothèses adéquates, que la distribution a posteriori est asymptotiquement gaussienne et centrée sur l'estimateur MLE, avec une variance égale à la variance asymptotique fréquentiste du MLE. Généralement la distance en variation totale (ou distance $\mathbb{L}^1$) est utilisée pour mesurer la convergence. Pour gérer les modèles mal spécifiés, notons P_* la projection de P_0 sur le modèle pour la divergence de Kullback-Leibler, et θ_* le paramètre associé. On obtient typiquement des résultats du type

$$E_{P_0} \left\| \mu(\theta|\boldsymbol{X}) - \mathcal{N}\left(\widehat{\theta}^{MLE}, I(\theta_*)^{-1}\right) \right\|_{\mathrm{TV}} \to 0,$$

où $I(\theta_*)$ est la matrice de l'information de Fisher calculée en θ_*. Une première difficulté qui surgit est que l'estimateur MLE n'est pas défini dans les modèles non-paramétriques, et dans ces modèles on peut s'interroger sur ce que devrait être le bon centrage pour un théorème de Bernstein-von Mises. De fait nous n'avons pas connaissance de tels théorèmes dans des cadres purement non-paramétriques. Par rapport à la consistance, les théorèmes de Bernstein-von Mises nécessitent en outre que l'a priori soit suffisamment plat sur des voisinages de P_0 ou de sa projection sur le modèle, et ce point pose également problème en non-paramétrique.

En dimension finie (fixée), le théorème de Bernstein-von Mises est désormais un résultat bien connu. Sa preuve est disponible dans des livres de synthèse, tel [106]. Dans un cadre non-paramétrique on dispose de contre-exemples, tels celui proposés par Freedman [45]. Les résultats positifs disponibles sont nettement moins nombreux que ceux portant sur les vitesses de convergence ; ils concernent des modèles crible (de dimension croissante) et des modèles semi-paramétriques.

En dimension croissante, signalons le papier précurseur de Ghosal [52] sur les modèles de régression, où les erreurs sont indépendantes identiquement distribuées mais pas nécessairement gaussiennes ; les hypothèses portent en particulier sur la distribution des erreurs et sur la croissance du nombre k_n de régresseurs, qui doit vérifier $k_n^4 \ln k_n = o(n)$. Ghosal [53] s'est aussi intéressé très tôt aux modèles exponentiels de dimension croissante. Les conditions utilisées pour établir la normalité asymptotique de la distribution a posteriori portent classiquement sur la masse de l'a priori autour du vrai paramètre et sur la régularité de l'a priori en ce point, mais également sur la croissance de la dimension, sur les valeurs propres de la matrice de l'information de Fisher, ainsi que sur des moments d'ordre 4 des lois du modèle autour de P_0. Boucheron et Gassiat [17] obtiennent un théorème de Bernstein-von Mises pour les probabilités discrètes à valeurs dans un ensemble dénombrable, auxquels les résultats de Ghosal [53] ne pouvaient pas s'appliquer. Ils développent en outre des théorèmes de Bernstein-von Mises semi-paramétriques en distance de Levy-Prokhorov pour des fonctionnelles de ces lois de probabilité, en particulier les entropies de Rényi d'ordre α [29]. Clarke et Ghosal [25] reviennent sur les modèles exponentiels et s'intéressent à des a priori de référence. Le théorème de Bernstein-von Mises leur permet d'obtenir un développement asymptotique de l'information mutuelle de Shannon pour ces a priori. Ce dernier point nous intéresse particulièrement dans une optique de compression de données, nous y reviendrons dans le paragraphe 1.2.3.

Parmi les théorèmes de Bernstein-von Mises semi-paramétriques, signalons tout d'abord les modèles spécifiques étudiés par Kim et Lee [68, 69] : le modèle de censure à droite et le modèle de Cox de hasard proportionnel. Castillo [19] obtient des résultats semi-paramétriques génériques avec des a priori de type processus gaussiens, sur des modèles semi-paramétriques où le paramètre se décompose en un paramètre d'intérêt et un paramètre de nuisance. La situation diffère selon qu'il y a ou non perte d'information. Ce travail clarifie un article de Shen [99] qui étudie des modèles semi-paramétriques génériques. Rivoirard et Rousseau [87] utilisent des modèles de dimension croissante pour obtenir un théorème de Bernstein-von Mises sur les fonctionnelles linéaires de la densité des observations. Ils atteignent en outre la vitesse fréquentiste minimax d'estimation pour les densités dans des classes de régularité spécifique avec un choix déterministe (non-adaptatif) de la dimension du modèle crible. Pour obtenir un résultat adaptatif ils considèrent des a priori hiérarchiques, mais le théorème de Bernstein-von Mises n'est pas démontré dans ce cas : ils fournissent même un contre-exemple où la limite asymptotique de la distribution a posteriori est un mélange de gaussiennes avec des centres différents.

Résultats obtenus. Dans cette thèse nous nous sommes intéressés au théorème de Bernstein-von Mises dans deux situations différentes : la régression linéaire gaussienne avec un nombre croissant de régresseurs, et les modèles exponentiels de dimension croissante.

Le chapitre 3 est consacré à la régression linéaire gaussienne. Nous y obtenons des théorèmes de Bernstein-von Mises non-paramétriques (sur le paramètre de dimension croissante), et des théorèmes de Bernstein-von Mises semi-paramétriques. Ces résultats ont été soumis sous forme d'article.

Les résultats non-paramétriques couvrent le cas d'un a priori gaussien spécifiquement

adapté à la paramétrisation, et le cas plus général d'a priori réguliers. Nous améliorons, en particulier en termes de croissance de la dimension, les résultats initiaux de Ghosal [52] dans la direction des erreurs gaussiennes indépendantes de même variance σ_n^2 connue. En outre nos modèles peuvent être mal-spécifiés ; en d'autres termes nous n'exigeons pas que P_0 appartienne à l'un des modèles de notre crible. Nos hypothèses portent sur deux choses : le poids de l'a priori sur un voisinage d'une projection orthogonale de P_0 sur le modèle ; l'a priori doit être suffisamment plat sur ce voisinage. Ces théorèmes sont appliqués aux fonctions de régression qui appartiennent à des classes de Sobolev périodiques et à des classes de régularité $C^\alpha[0,1]$, ainsi qu'au modèle de suite gaussienne. Dans ces différentes situations nous obtenons en outre la vitesse de convergence fréquentiste minimax avec des choix adaptés (non-adaptatifs) de la dimension k_n en fonction de n. Nous illustrons également des situations où la distribution a posteriori concentre à la vitesse minimax mais pour lesquels le théorème de Bernstein-von Mises n'est pas vérifié.

Nos théorèmes de Bernstein-von Mises semi-paramétriques couvrent les fonctionnelles linéaires et non-linéaires du paramètre. Le cas linéaire est immédiat à partir des théorèmes non-paramétriques et ne nécessite aucune hypothèse supplémentaire. Lorsque la fonction de régression f appartient à une classe de Sobolev périodique, nous sommes en mesure de proposer des estimateurs bayésiens adaptatifs pour les fonctionnelles linéaires de f ainsi que pour la norme $\mathbb{L}^2$ de f : en plus de la normalité asymptotique a posteriori, ces estimateurs atteignent la vitesse d'estimation minimax sur chacune de ces classes sans nécessiter de connaître à l'avance la classe exacte de f.

L'annexe A présente un travail encore en cours consacré aux modèles exponentiels de dimension croissante. Nous sommes en mesure d'améliorer les conditions du papier précurseur de Ghosal [53] dans une direction différente de [25]. En particulier nous relâchons les hypothèses (liées entre elles) sur la croissance de la dimension, les valeurs propres de l'information de Fisher, et les moments d'ordre 3 et 4 des lois du modèle au voisinage de la projection P_* de P_0 pour la divergence de Kullback-Leibler. Cela nous permet de retrouver les résultats de Boucheron et Gassiat [17] comme un cas particulier. Dans le cas général, notre théorème fait intervenir une espérance sous P_* au lieu d'une espérance sous P_0, ce qui nous limite pour le moment aux modèles de dimension croissante mais bien spécifiés — c'était déjà le cas dans [25, 53]. La limitation principale que nous rencontrons, c'est que les conditions sur les valeurs propres de l'information de Fisher restent difficiles à vérifier en pratique.

Perspectives. Aussi bien dans le chapitre 3 que dans le chapitre A, nous pensons qu'une réécriture minutieuse permettra de faire apparaître des vitesses non-asymptotiques explicites de la convergence de la distribution a posteriori vers sa limite asymptotique. Ces vitesses seraient valables uniformément sur des classes d'a priori, et cela nous serait utile dans un cadre de compression de données, pour se diriger vers des développements asymptotiques ou au-moins des encadrements de la redondance maximin. Le cheminement pour y arriver est l'objet du prochain paragraphe 1.2.3.

Concernant la régression non gaussienne, nous souhaiterions améliorer les résultats de Ghosal [52] en continuant sur la lancée d'une adaptation fine des méthodes de van der Vaart [106]. Sur la distribution des erreurs par exemple, l'idéal serait de se ramener à des conditions de différentiabilité en moyenne quadratique.

Pour les modèles exponentiels, nous envisageons dans un premier temps de poursuivre la recherche dans deux directions : la gestion des modèles mal spécifiés, et des méthodes efficaces pour vérifier dans les situations pratiques les conditions portant sur la matrice de Fisher. Ces deux points pourraient ouvrir la porte à de nombreuses applications. Dans une autre direction, les fonctionnelles du paramètre sont de première importance. Il nous faudra adapter aux modèles exponentiels généraux les outils utilisés par Boucheron et Gassiat [17] ou dans le chapitre 3.

1.2.3 Liens avec la compression de données

Les relations entre le théorème de Bernstein-von Mises et la théorie de l'information ont été mises en valeur dès 1990, dans l'article de Clarke et Barron [26]. Ces auteurs partent de la normalité asymptotique de la distribution a posteriori sur des classes de sources paramétriques régulières de dimension d fixée, et en déduisent un développement asymptotique de la redondance du mélange bayésien $M_{n,\mu}$ (voir (1.2)) :

$$R_n\left(M_{n,\mu}, P_\theta^n\right) = \frac{d}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\det I(\theta) + \log\frac{1}{\mathrm{d}\mu(\theta)} + o(1) \tag{1.6}$$

où $I(\theta)$ est la matrice de l'information de Fisher du modèle, $\mathrm{d}\mu$ est la densité de μ par rapport à la mesure de Lebesgue, et θ est un point intérieur du modèle.

La redondance du mélange bayésien peut en effet se réécrire sous la forme

$$\begin{aligned} R_n\left(M_{n,\mu}, P_\theta^n\right) &= E_{P_\theta^n}\left[-\log\frac{M_{n,\mu}(X_{1:n})}{P_\theta^n(X_{1:n})}\right] \\ &= E_{P_\theta^n}\left[\log\frac{\mathrm{d}\mu(\theta|X_{1:n})}{\mathrm{d}\mu(\theta)}\right]. \end{aligned}$$

La formule (1.6) découlera alors de l'approximation de $\mathrm{d}\mu(\theta|X_{1:n})$ par la densité gaussienne $\mathrm{d}\mathcal{N}\left(\widehat{\theta}^{MLE}, I(\theta)^{-1}\right)(\theta)$. Cependant l'utilisation de résultats de plus bas niveau vis-à-vis du théorème de Bernstein-von Mises, tels la concentration de la mesure a posteriori et les développements LAN de la log-vraisemblance (*Local Asymptotic Normality*, voir [106]), permet de réduire le nombre d'hypothèses nécessaires à l'obtention de développements du type (1.6).

L'étape suivante dans cette démarche est l'obtention de développements asymptotiques de l'information mutuelle de Shannon (1.3) pour des a priori bien choisis. L'information mutuelle n'est en effet rien d'autre que la redondance bayésienne

$$I(\mu, X_{1:n}) = \int_\Theta R_n\left(M_{n,\mu}, P_\theta^n\right)\,\mathrm{d}\mu(\theta).$$

Cela mène, sous des conditions adéquates, à un développement du type

$$I(\mu, X_{1:n}) = \frac{d}{2}\log\frac{n}{2\pi e} + \int_\Theta \log\frac{\sqrt{\det I(\theta)}}{\mathrm{d}\mu(\theta)}\,\mathrm{d}\mu(\theta) + o(1)$$

qui est asymptotiquement maximisé par l'a priori de Jeffreys $\mathrm{d}\mu(\theta) \propto \sqrt{\det I(\theta)}$: c'est le résultat obtenu par Clarke et Barron [27] pour les modèles paramétriques. Clarke et Ghosal [25] ont récemment effectué un travail similaire pour des modèles exponentiels de dimension croissante, en se basant sur le travail de Ghosal [53] pour la normalité asymptotique de la distribution a posteriori.

1.3 Modèles de mélange discrets

1.3.1 Une classification non-supervisée à base de modèle, avec sélection de variables

Motivation et modèles existants. Notre motivation initiale pour nous intéresser aux modèles de mélange discrets était le traitement de données génotypiques. Supposons que pour une population de n individus d'une espèce donnée, un certain nombre L de marqueurs génétiques soient relevés. Ces marqueurs, ou loci, seront les variables discrètes de notre étude statistique.
Le plupart des espèces vivantes étudiées sont diploïdes, c'est-à-dire que leur génome comporte deux exemplaires de chaque gène. En conséquence, ce qui est observé pour chaque individu i et chaque marqueur génétique l est un ensemble de deux valeurs ou allèles $\{x_i^{l,1}, x_i^{l,2}\}$. Ces deux allèles sont éventuellement égaux, et on ne peut pas leur attribuer d'ordre du fait qu'on ne sait pas distinguer l'origine des deux chromosomes d'une même paire.
Dans un certain nombre de situations, les biologistes pensent que la population globale des individus est en réalité divisée en plusieurs sous-populations. Ils sont donc demandeurs de méthodes statistiques de classification non supervisée, pour retrouver le nombre K de sous-populations différentes, caractériser ces sous-populations par leurs fréquences alléliques, et enfin attribuer chaque individu à une des K sous-populations.

Nous nous intéressons plus particulièrement aux méthodes de classification non supervisée à base de modèle (*model-based clustering* en anglais). Une telle classification fait particulièrement sens : l'identification des sous-populations en tant que composantes d'un mélange de distributions ouvre la voie à une interprétation biologique en termes de caractéristiques génomiques. Parmi les modèles utilisés dans ce domaine, les modèles les plus courants sont ceux considérés par Pritchard et al. [84], dont le logiciel `Structure` est par ailleurs très largement utilisé et fait référence. Ces modèles sont au nombre de deux.
Tout d'abord, un modèle qui coïnciderait avec le modèle à classe latente standard, à ceci près que les variables sont en fait des ensembles à deux éléments comme décrit plus haut. On peut le décrire ainsi :

- Les individus sont indépendants et identiquement distribués.
- À chaque individu i on attribue un label z_i, qui représente la sous-population à laquelle il appartient. z_i est une variable aléatoire distribuée selon une loi multinomiale $\pi = (\pi_1, \ldots, \pi_K)$.
- Sachant la population z_i, les différentes variables $x_i^l, 1 \leq l \leq L$, correspondant aux différents loci sont indépendantes. C'est ce que les biologistes appellent équilibre de liaison. Du fait de la recombinaison génétique entre chromosomes et à l'intérieur de chaque paire de chromosomes à chaque génération, cette hypothèse est d'autant plus justifiée que les loci sont éloignés dans le génome.
- Sachant la population z_i, les deux allèles $x_i^{l,1}$ et $x_i^{l,2}$ du même locus sont indépendantes. Cette hypothèse, spécifique aux données génotypiques diploïdes, est ce que les biologistes appellent équilibre de Hardy-Weinberg. Cela est justifié dans la mesure où les parents de chaque individu sont supposés tirés aléatoirement dans la sous-population à laquelle l'individu appartient.

En notant $\alpha_{k,l,j}$ la fréquence de l'allèle j au locus l dans la sous-population k, la vrai-

semblance de l'individu i devient dans ce modèle

$$P(x_i) = \sum_{k=1}^{K} \pi_k \prod_{l=1}^{L} \left(2 - \mathbb{1}_{x_i^{l,1}=x_i^{l,2}}\right) \alpha_{k,l,x_i^{l,1}} \alpha_{k,l,x_i^{l,2}}.$$

Le second modèle est privilégié par Pritchard et ses co-auteurs. La différence avec celui que nous venons d'exposer est que l'on n'attribue plus à chaque individu i le label z_i d'une unique classe, mais plutôt une mesure de probabilité $q^{(i)} = (q_1^{(i)}, \ldots, q_K^{(i)})$ sur la famille des sous-populations, où $q_k^{(i)}$ représente la proportion du génome de l'individu i qui provient de la sous-population k. Une famille de variables indépendantes $z_l^{(i,a)}$ est tirée selon $q^{(i)}$, qui représentent pour $1 \leq i \leq n$, $1 \leq l \leq L$ et $a \in \{1, 2\}$ la population d'où est originaire l'allèle a au locus l pour l'individu i.
Un tel modèle semble évidemment pertinent au niveau de l'interprétation biologique. En revanche le nombre de paramètres est très important. Les méthodes bayésiennes sont dès lors inévitables, et ce sont d'ailleurs souvent elles qui sont mises en œuvre dans de nombreux travaux du domaine (voir par exemple [23, 28]).

À notre connaissance, la sélection de variables a été introduite pour le traitement des données génotypiques par Toussile et Gassiat [104]. Cela permet d'améliorer les performances de la classification en éliminant les variables dont le bruit perturbe l'estimation statistique. Cela ouvre surtout la voie à une interprétation pratique : identifier les gènes qui différencient les populations peut être une information très utile du point de vue biologique.
Le modèle correspondant n'est plus défini par le seul nombre K de sous-populations, mais également par le sous-ensemble S des loci pertinents pour la classification. Les variables qui ne sont pas dans S sont désormais distribuées à l'identique sur les différentes sous-populations, et les fréquences alléliques pour ces variables sont notées $\beta_{l,j}$ (au lieu de $\alpha_{k,l,j}$). On se place dans la situation plus simple où à chaque individu est associé le label z_i d'une unique classe. La vraisemblance de l'individu i devient donc

$$P(x_i) = \sum_{k=1}^{K} \pi_k \prod_{l \in S} \left(2 - \mathbb{1}_{x_i^{l,1}=x_i^{l,2}}\right) \alpha_{k,l,x_i^{l,1}} \alpha_{k,l,x_i^{l,2}} \prod_{l \notin S} \left(2 - \mathbb{1}_{x_i^{l,1}=x_i^{l,2}}\right) \beta_{l,x_i^{l,1}} \beta_{l,x_i^{l,2}}.$$

La méthode statistique choisie par Toussile et Gassiat [104] est l'estimation par maximum de vraisemblance pénalisé. Cette méthode estime simultanément le nombre de composantes K et la collection S des variables pertinentes ; elle rentre dans la famille des méthodes de sélection de modèle en deux temps — estimation dans chaque modèle puis choix du modèle — à la différence de critères pénalisés comme **LASSO** ou **ICL** dans lesquels la pénalité ne dépend pas uniquement du modèle, mais de l'estimateur dans son ensemble. La sélection de modèle en deux temps peut se conceptualiser de la façon suivante :

1. On dispose d'une collection de modèles $(\mathcal{M}_{(K,S)})_{(K,S)\in\mathbb{M}}$.
2. Une fonction de constraste empirique sert à mesurer la performance des estimateurs. Son rôle est d'imiter le risque statistique auquel on n'a pas accès, éventuellement translaté d'une constante. Pour la sélection par maximum de vraisemblance pénalisé, on utilise le contraste de la log-vraisemblance :

$$\gamma_n(P) = -\frac{1}{n} \sum_{i=1}^{n} \ln P(x_i).$$

3. Dans chaque modèle $\mathcal{M}_{(K,S)}$, un estimateur $\widehat{P}_{(K,S)}$ est choisi, généralement en minimisant le contraste empirique sur le modèle. Dans notre cas, nous prenons l'estimateur MLE
$$\widehat{P}_{(K,S)} = \underset{P \in \mathcal{M}_{(K,S)}}{\arg\min}\, \gamma_n\left(P\right).$$
En pratique une version approchée du MLE est calculée grâce à l'algorithme `EM`.
4. Pour chaque modèle une pénalité $\mathbf{pen}_n(K,S)$ est choisie, afin de contrebalancer le risque de sur-adaptation (*overfit*) de l'estimateur aux données, que le contraste ne prend pas en compte. Un modèle est alors sélectionné de façon à minimiser le contraste pénalisé, ou critère :
$$\left(\widehat{K},\widehat{S}\right) = \underset{(K,S)\in\mathbb{M}}{\arg\min}\left[\gamma_n\left(\widehat{P}_{(K,S)}\right) + \mathbf{pen}_n(K,S)\right].$$
Parmi les critères classiques en sélection de modèles, il convient de signaler **AIC** (*Akaike's Information Criterion* [2]) et **BIC** (*Bayesian Information Criterion* [88]). **AIC** est caractérisé par la pénalité
$$\mathbf{pen_{AIC}}(K,S) = D_{(K,S)},$$
avec $D_{(K,S)}$ la dimension du modèle $\mathcal{M}_{(K,S)}$. **AIC** provient d'un développement asymptotique de la divergence de Kullback-Leibler, dans une optique d'efficacité asyptotique et d'estimation de densité. Quant à **BIC**, il est caractérisé par la pénalité
$$\mathbf{pen_{BIC}}(K,S) = \frac{\ln n}{2} D_{(K,S)}.$$
BIC provient d'un développement asymptotique de la vraissemblance bayésienne complétée pour l'a-priori de référence, dans une optique d'identification du vrai modèle. Nous renvoyons à la thèse de doctorat de Baudry [10] pour une discussion sur ces différents critères.
5. L'estimateur choisi à la fin de la procédure est $\widehat{P}_{\left(\widehat{K},\widehat{S}\right)}$.

Dans le cas des modèles introduits par Toussile et Gassiat [104], ces auteurs démontrent la consistance asymptotique de la procédure de sélection de modèle pour des pénalités du type **BIC**. Une fois le modèle choisi et les probabilités alléliques estimées, les individus sont classées dans les différentes sous-populations par la règle du maximum a posteriori (MAP) : on attribue à l'individu i l'étiquette $\widehat{z}_i$ qui maximise la probabilité conditionnelle $\widehat{P}_{\left(\widehat{K},\widehat{S}\right)}(z|x_i)$, calculée sur l'estimateur sélectionné.
Dans de tels modèles l'algorithme `EM` admet des formules explicites à chaque itération (voir [11] pour des détails algorithmiques autour de `EM`). En revanche la famille de modèles à considérer grandit rapidement (à une vitesse exponentielle en L), et son exploration se fait par un algorithme backward-stepwise adapté de ce que Maugis et Michel [76] font pour les mélanges finis de gaussiennes multivariées. Le bénéfice apporté par la sélection de variables est illustrée expérimentalement dans l'article [104] en particulier pour l'estimation de K. Les simulations illustrent aussi un phénomène attendu sur de tels modèles, qui est que **BIC** a tendance à surpénaliser et choisira longtemps (lorsque n grandit) des modèles de dimension trop petite, tandis que **AIC** sous-pénalise et aura tendance à sélectionner un nombre trop grand de sous-populations. Cela provient en particulier des approximations asymptotiques qui sous-tendent **BIC**. D'où la recherche de critères non-asymptotiques.

Modèle retenu et résultats. Le chapitre 4 présente un travail effectué en collaboration avec Wilson Toussile, qui prolonge le travail initié par Toussile et Gassiat. Son contenu fait l'objet d'un article soumis.

Nous considérons deux collections de modèles parallèles, construits sur les mêmes principes pour deux situations distinctes, et pour lesquels nous obtenons des résultats semblables. La première collection de modèles est la même que celle de Toussile et Gassiat [104]. La seconde regroupe des modèles à classe latente, avec sélection de variable comme dans la première collection, mais adaptés à des données discrètes multivariées habituelles : la variable x_i^l est simplement une variable multinomiale à valeurs dans un ensemble fini.

Comme Toussile et Gassiat, nous procédons par maximum de vraisemblance pénalisé. À la différence de ces auteurs, nous ne privilégions plus la consistance de la procédure de sélection de modèles, mais nous considérons une approche oracle. Cette approche accepte explicitement que les modèles soient mal spécifiés. Cela apporte une justification supplémentaire à la modélisation choisie : en particulier le fait que les probabilités alléliques ne puissent pas être exactement identiques entre les sous-populations n'est pas gênant. L'objectif n'est plus de choisir le vrai modèle, mais le modèle qui minimise le risque d'estimation mesuré en distance de Hellinger entre la distribution estimée et la vraie distribution des données. Ce risque ne mesure donc pas directement la qualité de la classification — nous y revenons dans le paragraphe 1.3.1.

En utilisant la théorie de Massart [75] basée sur l'entropie métrique, nous établissons une inégalité oracle non-asymptotique

$$E_{P_0}\left[\mathbf{h}^2\left(P_0, \widehat{P}_{(\widehat{K}_n,\widehat{S}_n)}\right)\right] \leq C\left(\inf_{(K,S)\in\mathbb{M}}\left(\mathbf{KL}\left(P_0, \mathcal{M}_{(K,S)}\right) + \mathbf{pen}_n(K,S)\right) + \rho + \frac{(3/4)^L}{n}\right)$$

où ρ mesure l'erreur permise lors du calcul de l'estimateur MLE

$$\gamma_n\left(\widehat{P}_{(K,S)}\right) \leq \inf_{P\in\mathcal{M}_{(K,S)}} \gamma_n(P) + \rho,$$

$\mathbf{KL}\left(P_0, \mathcal{M}_{(K,S)}\right)$ est l'infimum de la divergence de Kullback-Leibler entre P_0 et les distributions du modèle $\mathcal{M}_{(K,S)}$, et C est une constante absolue (supérieure à 1). La pénalité est soumise à une minoration

$$\mathbf{pen}_n(K,S) \geq \kappa\left(5 + \sqrt{\max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right)}\right)^2 \frac{D_{(K,S)}}{n}$$

où κ est une autre constante absolue supérieure à 1. L'inégalité oracle apporte une certaine validation théorique de la procédure de sélection de modèle, en permettant de relier le risque de l'estimateur sélectionné au risque du meilleur estimateur inconnu de la collection. Elle présente également l'avantage d'admettre explicitement que l'estimateur choisi dans chaque modèle n'atteigne pas exactement le maximum de vraisemblance.

En revanche la pénalité de l'inégalité oracle est trop conservative et a tendance à surpénaliser la complexité des modèles. La pénalité retenue en pratique n'est donc pas exactement celle-là, même si elle en est inspirée. Elle est de la forme désormais assez répandue $\lambda D_{(K,S)}$, où $D_{(K,S)}$ est la dimension du modèle courant et λ est un paramètre calibré automatiquement sur les données en utilisant l'heuristique de pente due à Birgé

et Massart [13]. Des simulations sont présentées, qui illustrent la bonne performance de la procédure, tant pour la consistance de la sélection de modèle que pour le choix d'un modèle proche de l'oracle, en comparaison aux critères classiques **BIC** et **AIC**.

L'ensemble de ces procédures a été implémenté dans la version 2 du logiciel `MixMoGenD`, dont la version 1 a été proposée par Toussile et Gassiat [104], et qui est librement téléchargeable. Nous revenons sur les aspects algorithmiques dans le paragraphe 1.3.2.

Nous tenons à signaler le travail récent de Biernacki, Celeux et Govaert [12] qui s'intéressent au modèle à classe latente standard (sans sélection de variables) ; grâce à des formules explicites de la vraisemblance intégrée des données complétées pour l'a priori de Jeffreys, ces auteurs développent des critères bayésiens non-asymptotiques de sélection de modèles, qui évitent l'approximation **BIC**. Il serait sans doute intéressant de les étendre aux modèles que nous considérons.

Perspectives. La sélection de variables permet de réduire la dimension des modèles considérés. Malgré cela la dimension reste grande : elle croît linéairement en le nombre de sous-populations, linéairement en le cardinal de S, et linéairement en le nombre moyen d'états possible pour les différentes variables. Dans le même temps, la taille des échantillons disponibles pour certains projets biologiques reste modeste. Il est fréquent que la dimension des modèles soit nettement supérieure au nombre d'individus (même si les tableaux de données sont plus grands du fait qu'il y a plusieurs variables). En particulier, nous avons commencé à nous intéresser à des façons moins coûteuses en nombre de paramètres de représenter les lois de probabilité multinomiales — par exemple avec des histogrammes. Nous regardons aussi des modèles qui font intervenir des sur-clusters, qui regrouperaient les sous-populations de manière différente pour chaque variable. Parmi les bénéfices attendus, ces sur-clusters donnent lieu à une interprétation pratique. Notre idée peut s'exprimer ainsi : la probabilité d'un état donné pour une variable donnée ne dépendrait du cluster auquel l'individu appartient que par l'intermédiaire du sur-cluster associé à la variable. La difficulté dans ces nouvelles collections de modèles est d'en concevoir une exploration adaptée ; l'exploration de la collection de modèles utilisée dans le chapitre 4 est déjà relativement coûteuse.
Nous souhaitons rapprocher davantage le critère de sélection de modèle que nous retenons en pratique de la théorie illustrée par l'inégalité oracle. Cela pourrait se faire par le biais de justifications théoriques de l'heuristique de pente, pour nos modèles ou de façon générale.
Parmi les critères de vraisemblance pénalisés, il paraît naturel de s'intéresser au bénéfice qu'apporteraient des pénalités ℓ^1 (de type **LASSO**) vis-à-vis de la procédure actuelle de sélection de modèle par pénalité ℓ^0 (proportionnelle à la dimension du modèle). Cependant une pénalité ℓ^1 pertinente pour des mélanges de lois multinomiales multivariées reste à définir.
Nous voudrions sortir des critères liés à l'estimation de densité pour avoir un critère plus proche de la classification, et obtenir des bornes non asymptotiques pour les procédures associées. On peut penser au critère **ICL** et à ses dérivés, mais ils favorisent des clusters bien séparés plutôt que la détection des composantes d'un mélange (voir [10] par exemple). Cela ne serait pas forcément un avantage pour l'interprétation biologique.
Enfin, nous désirons intégrer la gestion des données manquantes, qui sont fréquentes en biologie pour ce type de données.

1.3.2 Aspects algorithmiques

Parallèlement à la recherche mathématique exposée au chapitre 4, nous avons collaboré avec Wilson Toussile à l'amélioration et l'extension du logiciel `MixMoGenD`. Une petite partie de ces améliorations se trouve déjà dans la version du logiciel qui accompagne la version définitive de l'article [104].
Le logiciel est développé en C++ mais les toutes premières versions de développement avaient été écrites en C. Le premier travail, invisible au niveau des fonctionnalités, a donc été d'achever la conversion à la programmation orientée objet et à une gestion rigoureuse des allocations dynamiques de la mémoire.

On peut distinguer trois niveaux de l'algorithme :

1. la recherche de l'estimateur MLE sur les modèles individuels, au moyen de l'algorithme `EM` ;
2. l'exploration de la famille des modèles disponibles par une procédure backward-stepwise ;
3. la sélection du modèle et donc de l'estimateur final en maximisant un critère de maximum de vraisemblance pénalisé sur les modèles explorés. Cette dernière étape comporte une calibration de la pénalité par heuristique de pente, si on a choisi le critère de sélection que nous proposons.

Concernant l'algorithme `EM`, nous utilisons la méthode proposée par [11], qui consiste à faire des lancements brefs de `EM` à partir de points de départ aléatoires, puis à choisir le plus prometteur. Le nombre de points de départ et le nombre d'itérations pour chacun sont paramétrables en option. L'algorithme `SEM` (Stochastic EM) est également proposé en option. Toutefois les versions grand public de `MixMoGenD` désactivent ces différentes options pour ne pas impacter la facilité d'utilisation.
Pour le tirage aléatoire des points de départs de l'algorithme `EM`, nous avons tenté de tirer les probabilités des différents états de chaque variable selon une distribution de Dirichlet $D(1/2, \ldots, 1/2)$. Les résultats de l'algorithme `EM` étaient très mauvais ! En revanche une distribution uniforme sur le simplex a produit de bons résultats : avec 10 points de départs nous n'avons constaté aucun problème sur toutes les simulations présentées paragraphe 4.4.3 et sur un certain nombre d'autres simulations semblables que nous avons effectuées.

Dans un premier temps nous reprenions l'algorithme backward-stepwise utilisé par Maugis et Michel [76] pour des mélanges de gaussiennes multivariées. La procédure explore, pour une valeur K fixée du nombre de sous-populations, les sous-ensembles S de variables les plus pertinents. On part de $S = \{1, \ldots, L\}$ plein, et on essaie successivement d'enlever puis de rajouter des variables, à condition que cela améliore le critère de sélection de modèles, jusqu'à ce qu'on ne puisse plus. Nous avons constaté que l'exploration pouvait s'interrompre sur des ensembles S relativement grands, à cause de phénomènes de minima locaux, alors que l'ensemble S qui minimise le critère est beaucoup plus petit. Nous avons alors modifié l'algorithme de sorte que la descente soit forcée dans les situations où auparavant l'algorithme s'interrompait. Sur toutes les simulations effectuées, cela nous a permis de passer réellement par les modèles les plus pertinents. Accessoirement, cela nous a permis d'ajouter en option une étape forward-stepwise, sur le même principe que l'algorithme backward-stepwise, mais en partant de $S = \emptyset$.

Concernant la procédure de calibration de la pénalité, une première difficulté était

liée à l'exploration des modèles : l'algorithme backward-stepwise utilise le critère de sélection de modèles pour choisir son chemin d'exploration. De l'autre côté, le critère que nous proposons n'est calibré qu'*après* la procédure d'exploration. Nous avons résolu le problème en utilisant une grille exponentielle de valeurs du paramètre à calibrer, de manière à couvrir un éventail de pénalités qui contienne et dépasse les pénalités associées à **AIC** et **BIC**. L'algorithme backward-stepwise est alors lancé pour chacune des valeurs de la grille. Dans de nombreux cas cela n'est pas aussi coûteux que l'on pourrait le craindre, parce que les chemins d'exploration coïncident sur des parties importantes de leur trajet. La calibration de la pénalité quant à elle fait intervenir une détection de rupture, que nous avons rendue plus robuste en utilisant une fenêtre glissante (voir paragraphe 4.4.1).

À côté du code stable qui est compilé dans les versions téléchargeables du logiciel `MixMoGenD`, nous développons une version expérimentale. Celle-ci fait appel à des méthodes de programmation évoluées, en particulier au polymorphisme. Nous avons par exemple mutualisé et sécurisé du code qui intervenait dans plusieurs contextes, grâce à l'utilisation de méthodes génériques (*template methods*).
L'utilisation de l'héritage et de classes virtuelles, y compris de classes abstraites, ainsi que de classes génériques (*template classes*), nous a permis de développer la modularité du logiciel. En particulier les nouvelles collections de modèles présentées à la fin du paragraphe 1.3.1 ont pu être intégrées au logiciel expérimental, et des premières simulations ont été effectuées. Cela ouvre aussi la porte à la gestion de données mélangeant différents types de variables discrètes, ou bien des variables discrètes et des variables continues.

Nous projetons de proposer des versions de `MixMoGenD` sous forme de paquet `R`. Cela est déjà possible en réalité, mais l'intégration à `R` reste à améliorer. Enfin, nous projetons à moyen terme de publier le code stabilisé du logiciel `MixMoGenD` sous licence GPL.

Chapitre 2

Codage universel en alphabet infini : Enveloppes à décroissance exponentielle

Universal Coding on Infinite Alphabets: Exponentially Decreasing Envelopes

Abstract

This chapter deals with the problem of universal lossless coding on a countable infinite alphabet. It focuses on some classes of stationary memoryless sources defined by an envelope condition on the marginal distribution, namely exponentially decreasing envelope classes with exponent α.

The minimax redundancy of exponentially decreasing envelope classes is proved to be equivalent to $\frac{1}{4\alpha \log e} \log^2 n$. Then an approximately maximin prior distribution is provided. At last, an adaptive algorithm is proposed, whose maximum redundancy is equivalent to the minimax redundancy. A recent simplification of this algorithm is available in appendix.

Keywords: Adaptive compression, Bayes mixture, Data compression, Infinite countable alphabets, Redundancy, Universal coding.

Sommaire

2.1 Introduction

Compression of data is broadly used in our daily life: from the movies we watch to the office documents we produce. In this article, we are interested in lossless data compression on an unknown alphabet. This has applications in areas such as language modeling or lossless multimedia codecs.

First, we present briefly the problematics of data compression. More details are available in general textbooks, like [29]. Then we make a short review of preceding results, in which we situate the topic of this article, exponentially decreasing envelope classes, and we announce our results.

2.1.1 Lossless data compression

Consider a finite or countably infinite alphabet $\mathcal{X}$. A source on $\mathcal{X}$ is a probability distribution $\boldsymbol{P}$, on the set $\mathcal{X}^{\mathbb{N}}$ of infinite sequences of symbols from $\mathcal{X}$. Its marginal distributions are denoted by P^n, $n \geq 1$ (for $n = 1$, we only note P). The scope of lossless data compression is to encode a sequence of symbols $X_{1:n}$, generated according to P^n, into a sequence of bits as small as possible. The algorithm has to be uniquely decodable.

The binary entropy $H(P^n) = E_{P^n}[-\log_2 P^n(X_{1:n})]$ is known to be a lower bound for the expected codelength of $X_{1:n}$. From now on, log denotes the logarithm taken to base 2, while ln is used to denote the natural logarithm. Since arithmetic coding based on P^n encodes a message $x_{1:n}$ with $\lceil -\log P^n(x_{1:n}) \rceil + 1$ bits, this lower bound can be achieved within two bits. Then, the expected redundancy measures the mean number of extra bits, in addition to the entropy, a coding strategy uses to encode X^n. In the sequel, we use the word *redundancy* instead of *expected redundancy*.

Furthermore, together with Kraft-McMillan inequality, arithmetic coding provides an almost perfect correspondence between coding algorithms and probability distributions on $\mathcal{X}^n$. In this setting, if an algorithm is associated to the probability distribution Q^n, its expected redundancy reduces to the Kullbach-Liebler divergence between P^n and Q^n

$$D(P^n; Q^n) = E_{P^n}\left[\log \frac{P^n(X_{1:n})}{Q^n(X_{1:n})}\right].$$

We call this quantity (expected) redundancy of the distribution Q^n (with respect to P^n).

Unfortunately, the true statistics of the source are not known in general, but P^n is supposed to belong to some large class Λ of sources (for instance, the class of all stationary memoryless sources, or the class of Markov sources). In this paper, the maximum redundancy

$$R_n(Q^n; \Lambda) = \sup_{\boldsymbol{P} \in \Lambda} R_n(Q^n; P^n)$$

measures how well a coding probability Q^n behave on an entire class Λ. With this point of view, the best coding probability is a *minimax* coding probability, that achieves the *minimax redundancy*

$$R_n(\Lambda) = \inf_{Q^n} R_n(Q^n; \Lambda).$$

Another way to measure the ability of a class of sources to be efficiently encoded is the *Bayes redundancy*

$$R_{n,\mu}(\Lambda) = \inf_{Q^n} \int_\Lambda R_n(Q^n; P^n)\, \mathrm{d}\mu(\boldsymbol{P})$$

where μ is a prior distribution on Λ endowed with the topology of weak convergence and the Borel σ-field. Only one coding strategy achieves the Bayes redundancy: the Bayes mixture

$$M_{n,\mu}(x_{1:n}) = \int_\Lambda P^n(x_{1:n})\, \mathrm{d}\mu(\boldsymbol{P}).$$

When Λ is a class of stationary memoryless sources on the set $\mathcal{X} = \mathbb{N}_* = \mathbb{N}\backslash\{0\}$, there is a natural parametrization of Λ by $P_{\boldsymbol{\theta}}(j) = \theta_j$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots) \in \Theta_\Lambda$. Θ_Λ is then a subset of

$$\Theta = \left\{ \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots) \in [0,1]^{\mathbb{N}} : \sum_{i \geq 1} \theta_i = 1 \right\}$$

and it is endowed with the topology of pointwise convergence. In this case we write μ as a prior on Θ_Λ.

Minimax redundancy and Bayes redundancy are linked by an important relation [33, 46]; it is written here in the context of stationary memoryless sources on a finite or countably infinite alphabet, but Haussler [62] has shown that it can be generalized for all classes of stationary ergodic processes on a complete separable metric space.

Theorem 1. *Let Λ be a class of stationary memoryless sources, such that the parameter set Θ_Λ is a measurable subset of Θ. Let $n \geq 1$. Then*

$$R_n(\Lambda) = \sup_\mu R_{n,\mu}(\Lambda),$$

where the supremum is taken over all (Borel) probability measures on Θ_Λ.

The quantity $\sup_\mu R_{n,\mu}(\Lambda)$ is called *maximin redundancy*. A prior whose Bayes redundancy corresponds to the maximin redundancy is said to be maximin, or least favorable.
Theorem 1 says that maximin redundancy and minimax redundancy are the same. It provides a tool to calculate the minimax redundancy.

Before speaking about known results, let us make mention of other two notions.
With an asymptotic point of view, a sequence of coding probabilities $(Q_n)_{n \geq 1}$ is said to be weakly universal if the per-symbol redundancy tends to 0 on Λ:

$$\sup_{\boldsymbol{P} \in \Lambda} \lim_{n \to \infty} \frac{1}{n} D(P^n; Q^n) = 0.$$

Instead of the expected redundancy, many authors consider individual sequences. In this case, the *minimax regret*

$$R_n^*(\Lambda) = \inf_{Q^n} \sup_{\boldsymbol{P} \in \Lambda} \sup_{x_{1:n} \in \mathcal{X}^n} \log \frac{P^n(x_{1:n})}{Q^n(x_{1:n})}$$

plays the role that the minimax redundancy plays with the expected redundancy.

2.1.2 Exponentially decreasing envelope classes

In the case of a finite alphabet of size k, many classes of sources have been studied in the literature, for which estimates of the redundancy have been provided. In particular we have the class of all stationary memoryless sources (see [7, 21, 40, 72, 119, 120], and references therein), whose minimax redundancy is

$$\frac{k-1}{2}\log\frac{n}{2\pi e}+\log\frac{\Gamma(1/2)^k}{\Gamma(k/2)}+o(1).$$

This last class can be seen as a particular case of a $(k-1)$-dimensional class of stationary memoryless sources on a (possibly) bigger alphabet, for which we have a similar result under certain conditions (see [6, 26, 27]). Similar results are still available for classes of Markov processes and finite memory tree sources on a finite alphabet (see [4,32,72,118]), and for k-dimensional classes of even non stationary memoryless sources on an arbitrary alphabet (see [85]).
The results become less precise when one considers infinite dimensional classes on a finite alphabet. A typical example is the class of renewal processes, for which we do not have an equivalent of the expected redundancy, but we know that it is lower and upper bounded by a constant times $\sqrt{n}$ (see [31, 43]).
Eventually, it is well known that the class of stationary ergodic sources on a finite alphabet is weakly universal (see [29]). However, Shields [101] showed that this class does not admit non-trivial universal redundancy rates.

In the case of a countably infinite alphabet, the situation is significantly different. Even the class of all stationary memoryless sources is not weakly universal (see [61,67]). Kieffer characterized weakly universal classes in [67] (see also [60, 61]):

Proposition 3. *A class Λ of stationary sources on $\mathbb{N}_*$ is weakly universal if and only if there exists a probability distribution Q on $\mathbb{N}_*$ such that for every $\boldsymbol{P}\in\Lambda$, $D(P;Q)<\infty$.*

In the literature, we find two main ways to deal with infinite alphabets. The first one [24,48,50,65,81–83,91,97] separates the message into two parts: a description of the symbols appearing in the message, and the *pattern* they form. Then the compression of patterns is studied.

A second approach [16,41,44,60,66] studies collections of sources satisfying Kieffer's condition, and proposes compression algorithms for these classes. A result from [16] indicates us such a way:

Proposition 4. *Let Λ be a class of stationary memoryless sources over $\mathbb{N}_*$. Let the envelope function f be defined by $f(x)=\sup_{\boldsymbol{P}\in\Lambda}P(x)$. Then the minimax regret satisfies*

$$R_n^*(\Lambda)<\infty\Leftrightarrow\sum_{x\in\mathbb{N}_*}f(x)<\infty.$$

It is therefore quite natural to consider classes of stationary memoryless sources with envelope conditions on the marginal distribution. In this article we study specific classes of stationary memoryless sources introduced by [16], and called *exponentially decreasing envelope classes*.

Definition 1. *Let C and α be positive numbers satisfying $C > e^{2\alpha}$. The exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$ is the class of sources defined by*

$$\Lambda_{Ce^{-\alpha\cdot}} = \{\boldsymbol{P} : \forall k \geq 1,\ P(k) \leq Ce^{-\alpha k} \text{ and } \boldsymbol{P} \text{ is stationary and memoryless.}\}$$

The first condition addresses mainly the queue of the distribution of X_1; it means that great numbers must be rare enough. It does not mean that the distribution is geometrical: if C is big enough, many other distributions are possible. Furthermore we will see that the exact value of C does not change significantly the minimax redundancy, unlike α.

Since in this paper we are going to only talk about exponentially decreasing envelope classes, we simplify the notations $R_n(Q^n; \Lambda_{Ce^{-\alpha\cdot}})$, $R_n(\Lambda_{Ce^{-\alpha\cdot}})$, and $R_{n,\mu}(\Lambda_{Ce^{-\alpha\cdot}})$ into $R_n(Q^n; C, \alpha)$, $R_n(C, \alpha)$, and $R_{n,\mu}(C, \alpha)$ respectively. The subset of Θ corresponding to $\Lambda_{Ce^{-\alpha\cdot}}$ is denoted by

$$\Theta_{C,\alpha} = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots) \in [0,1]^{\mathbb{N}} : \sum_{i\geq 1} \theta_i = 1 \text{ and } \forall i \geq 1,\ \theta_i \leq Ce^{-\alpha i}\}. \tag{2.1}$$

We present two main results about these classes.

In Section 2.2 we calculate the minimax redundancy of exponentially decreasing envelope classes, and we find that it is equivalent to $\frac{1}{4\alpha \log e} \log^2 n$ as n tends to the infinity. This rate is interesting for two main reasons. Up to our knowledge, exponentially decreasing envelope classes are the first family of classes on an infinite alphabet for which an equivalent of the minimax redundancy is known. Then, even the rate is new: until now only rates in $\log n$ or $\sqrt{n}$ have been obtained.

In Section 2.3 we are concerned with the problem of finding a maximin Bayes prior. We construct a sequence of Bayes priors whose Bayes redundancy is equivalent to the maximin redundancy, as the length n of the message tends to the infinity.

Once the minimax redundancy of a class of sources is known, we are interested in finding a minimax coding algorithm. Section 2.4 proposes a new adaptive coding algorithm `ACcode`, and we show that its maximum redundancy is equivalent to the minimax redundancy of exponentially decreasing envelope classes.

Considered together, these two results (of Section 2.3 and Section 2.4) provide a proof of our approximation of the minimax redundancy independent on the one given in Section 2.2.

The Appendix contains some proofs and some auxiliary results used in the main analysis. Eventually, a recent simplification of our algorithm `ACcode` is proposed in Appendix 2.D.

2.2 Minimax redundancy

In this section we state our main result. Theorem 2 below gives an equivalent of the minimax redundancy of exponentially decreasing envelope classes. To get it, we use a

result due to Haussler and Opper [63].

Theorem 2. *Let C and α be positive numbers such that $C > e^{2\alpha}$. The minimax redundancy of the exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha \cdot}}$ satisfies*

$$R_n(C,\alpha) \underset{n\to\infty}{\sim} \frac{1}{4\alpha \log e} \log^2 n.$$

Theorem 2 improves on a preceding result of [16, Theorem 7]. In that article the following bounds of the minimax redundancy of exponentially decreasing envelope classes are given:

$$\begin{aligned} &\frac{1}{8\alpha \log e} \log^2 n \,(1+o(1)) \\ &\quad \leq R_n(C,\alpha) \\ &\quad \leq \frac{1}{2\alpha \log e} \log^2 n + O(1). \end{aligned}$$

In subsection 2.2.1 we outline the work done in [63], and then we use it in subsection 2.2.2 to prove Theorem 2. Eventually, we discuss in subsection 2.2.3 the adaptation of this method to other envelope classes.

2.2.1 From the metric entropy to the minimax redundancy

To study the redundancy of a class of sources, [63] considers the Hellinger distance between the first marginal distributions of each source. Bounds on the minimax redundancy are provided in terms of the metric entropy of the set of the first marginal distributions, with respect to the Hellinger distance. As a consequence, that method can be applied only to stationary memoryless sources. However it is very efficient in the case of exponentially decreasing envelope classes.

First, we need to define the Hellinger distance and the metric entropy. In the case of sources on a countably infinite alphabet, the Hellinger distance can be defined in the following way:

Definition 2. *Let P and Q two probability distributions on $\mathbb{N}_*$. Then the Hellinger distance between P and Q is defined by*

$$h(P,Q) = \sqrt{\sum_{k\geq 1} \left(\sqrt{P(k)} - \sqrt{Q(k)}\right)^2}.$$

A related metric can be defined on the parameter set Θ:

$$d(\boldsymbol{\theta},\boldsymbol{\theta}') = h(P_{\boldsymbol{\theta}}, P_{\boldsymbol{\theta}'}) = \sqrt{\sum_{k\geq 1} \left(\sqrt{\theta_k} - \sqrt{\theta'_k}\right)^2}.$$

From a metric we can define the *metric entropy*. We consider first some numbers.

Definition 3. *Let S be a subset of Θ, and ϵ be a positive number.*

1. *We denote by* $\mathcal{D}_\epsilon(S,d)$ *the cardinality of the smallest finite partition of* S *with sets of diameter at most* ϵ*, or we set* $\mathcal{D}_\epsilon(S,d)=\infty$ *if no such finite partition exists.*
2. *The metric entropy of* (S,d) *is defined by*

$$\mathcal{H}_\epsilon(S,d)=\ln \mathcal{D}_\epsilon(S,d).\ ^{1}$$

3. *An* ϵ-cover *of* S *is a subset* $A\subset S$ *such that, for all* x *in* S*, there is an element* y *of* A *with* $d(x,y)<\epsilon$*. The* covering number $\mathcal{N}_\epsilon(S,d)$ *is the cardinality of the smallest finite* ϵ*-cover of* S*, or we define* $\mathcal{N}_\epsilon(S,d)=\infty$ *if no finite* ϵ*-cover exists.*
4. *An* ϵ-separated subset *of* S *is a subset* $A\subset S$ *such that, for all distinct* x*,* y *in* A*,* $d(x,y)>\epsilon$*. The* packing number $\mathcal{M}_\epsilon(S,d)$ *is the cardinality of the largest finite* ϵ*-separated subset of* S*, or we define* $\mathcal{M}_\epsilon(S,d)=\infty$ *if arbitrary large* ϵ*-separated subsets exist.*

The following lemma explains how these numbers are linked. It is a classical result that can be found for instance in [110].

Lemma 1. *Let* S *be a subset of* Θ*. For all* $\epsilon>0$*,*

$$\mathcal{M}_{2\epsilon}(S,d)\le\mathcal{D}_{2\epsilon}(S,d)\le\mathcal{N}_\epsilon(S,d)\le\mathcal{M}_\epsilon(S,d).$$

Lemma 1 enables us to choose the most convenient number to calculate the metric entropy.

From the metric entropy one can define the notion of metric dimension, which generalizes the classical notion of dimension. But the metric entropy lets us know in some way how dense the elements are in a set, even infinite dimensional.

Another quantity that [63] uses is the *minimax risk for the* $(1+\lambda)$*-affinity*

$$R_\lambda(\Lambda)=\inf_Q\sup_{\boldsymbol{\theta}\in\Theta_\lambda}\sum_{k\ge1}P_{\boldsymbol{\theta}}(k)^{1+\lambda}Q(k)^{-\lambda},$$

defined for all $\lambda>0$.
More precisions about the $(1+\lambda)$-affinity are given in [63]. See also [14] for a special regard payed to envelope classes.

In the case of an envelope class Λ_f defined by an integrable envelope function f, it is easy to see that $R_\lambda(\Lambda_f)<\infty$ for all $\lambda>0$. Indeed the choice

$$Q(k)=\frac{f(k)}{\sum_{l\ge1}f(l)}$$

leads to the relation

$$R_\lambda(\Lambda_f)\le\left(\sum_{k\ge1}f(k)\right)^\lambda.$$

We can now write a slightly modified version [2] of Theorem 5 of [63] in the context of data compression on an infinite alphabet.

1. We follow [63] in this definition of the metric entropy. Several authors use a slightly different definition, based on the covering number or the packing number.

2. The separation of the upper and lower bounds have no effect on the proof given by Haussler and Opper. A complete justification is available in [14].

Theorem 3. *Let Λ be a class of stationary memoryless sources on $\mathbb{N}_*$, such that the parameter set Θ_Λ is a measurable subset of Θ. Assume that there exists $\lambda > 0$ such that $R_\lambda(\Lambda) < \infty$. Let $h(x)$ be a continuous, non-decreasing function defined on the positive reals such that, for all $\gamma \geq 0$ and $C > 0$,*

1. $$\lim_{x\to\infty} \frac{h\left(Cx\left(h(x)\right)^\gamma\right)}{h(x)} = 1$$

   and

2. $$\lim_{x\to\infty} \frac{h\left(Cx(\ln x)^\gamma\right)}{h(x)} = 1.$$

Then

1. *If*
   $$\mathcal{H}_\epsilon(\Theta_\Lambda, d) \underset{\epsilon\to 0}{\sim} h\left(\frac{1}{\epsilon}\right),$$
   then
   $$R_n(\Lambda) \underset{n\to\infty}{\sim} (\log e)\, h(\sqrt{n}).^{3}$$
2. *If, for some $\alpha > 0$ and $c > 0$,*
   $$\liminf_{\epsilon\to 0} \frac{\mathcal{H}_\epsilon(\Theta_\Lambda, d)}{(1/\epsilon)^\alpha h(1/\epsilon)} \geq c,$$
   then
   $$\liminf_{n\to\infty} \frac{R_n(\Lambda)}{n^{\alpha/(\alpha+2)}\left[h(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)}} > 0.$$
3. *If, for some $\alpha > 0$ and $C > 0$,*
   $$\limsup_{\epsilon\to 0} \frac{\mathcal{H}_\epsilon(\Theta_\Lambda, d)}{(1/\epsilon)^\alpha h(1/\epsilon)} \leq C,$$
   then
   $$\limsup_{n\to\infty} \frac{R_n(\Lambda)}{(n\ln n)^{\alpha/(\alpha+2)}\left[h(n^{1/(\alpha+2)})\right]^{2/(\alpha+2)}} < \infty.$$

The conditions concerning the function h mean that h cannot grow too fast. For instance, h can grow like $C(\ln x)^\eta$, with $\eta \geq 0$.

The first case in the theorem is the one we use for exponentially decreasing envelope classes. In this case, the fast decreasing envelope produces a "not too big" metric entropy. Theorem 3 gives us an equivalent of the minimax redundancy of the class of sources when n goes to the infinity. This turns out very useful, as it improves a preceding result of [16]. However it is only an asymptotic result, without any convergence speed.

The second and the third items correspond to bigger classes of sources. In these cases the result is a bit less interesting: it gives a speed for the growth of the redundancy, but without the associated constant factor. Furthermore there is a gap of $(\ln n)^{\alpha/(\alpha+2)}$ between the lower bound of point 2 and the upper bound of point 3. However it allows us to retrieve more or less a result of [16] for another type of envelope classes.

We now develop these applications.

3. The $(\log e)$ factor comes from the use of the logarithm taken to base 2, in the definition of R_n.

2.2.2 The minimax redundancy of exponentially decreasing envelope classes

We want to apply Theorem 3 in order to prove Theorem 2, and therefore we have to calculate the metric entropy of exponentially decreasing envelope classes:

Proposition 5. *Let C and α be positive numbers such that $C > e^{2\alpha}$. The metric entropy of the parameter set $\Theta_{C,\alpha}$ satisfies*

$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) = (1 + o(1))\frac{1}{\alpha}\ln^2(1/\epsilon),$$

where $o(1)$ is a function $g(\epsilon)$ such that $g(\epsilon) \to 0$ as $\epsilon \to 0$.

Proof of Theorem 2. Just apply Theorem 3, with $h(x) = \frac{1}{\alpha}\ln^2(x)$, to get the result. □

We make first some general considerations about the entropy of envelope classes, and then prove Proposition 5. Let Λ_f be the envelope class defined by the integrable envelope function f. Let Θ_f be the corresponding parameter set

$$\Theta_f = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots) \in [0,1]^{\mathbb{N}} : \sum_{i\geq 1} \theta_i = 1 \text{ and } \forall i \geq 1,\ \theta_i \leq f(i)\}.$$

The function $\boldsymbol{\theta} \mapsto (\sqrt{\theta_1}, \sqrt{\theta_2}, \ldots)$ is an isometry between the metric space (Θ_f, d) and the subset $A_f \cap \{\|x\| = 1\}$ of ℓ^2, equipped with the classical euclidean norm $\|\cdot\|$, where A_f is defined by

$$A_f = \{(x_k)_{k\in\mathbb{N}^*} \in \ell^2 :\ \forall k \in \mathbb{N}^*,\ 0 \leq x_k \leq \sqrt{f(k)}\}. \tag{2.2}$$

The metric entropy of (Θ_f, d) can be calculated in this space.

Next we truncate some coordinates, to work in a finite dimensional space instead of ℓ^2. Together with an adequate use of Lemma 1, this helps us to obtain upper and lower bounds of the metric entropy of (Θ_f, d). The outlines of the proofs of the next lemmas can be found in Appendix 2.A. We start with the upper bound.

Lemma 2. *Let Λ_f be the envelope class defined by the integrable envelope function f, and let ϵ be a positive number. Let N_ϵ denote the integer*

$$N_\epsilon = \inf\left\{n \geq 1 :\ \sum_{k\geq n+1} f(k) \leq \frac{\epsilon^2}{16}\right\}.$$

For $U \in \mathbb{R}^N$ and $a > 0$, let $B_{\mathbb{R}^N}(U, a)$ denote the ball in $\mathbb{R}^n$ with center U and radius a. Then

$$\mathcal{H}_\epsilon(\Theta_f, d) \leq N_\epsilon \ln(1/\epsilon) + 3N_\epsilon \ln 2 + A(N_\epsilon) + B(\epsilon),$$

where

$$A(N) = -\ln \mathrm{Vol}\,(B_{\mathbb{R}^N}(0,1)) = \ln \frac{\Gamma(\frac{N}{2}+1)}{\pi^{\frac{N}{2}}}$$

and

$$B(\epsilon) = \sum_{k=1}^{N_\epsilon} \ln\left(\sqrt{f(k)} + \frac{\epsilon}{4}\right).$$

Furthermore

$$A(N_\epsilon) \underset{\epsilon\to 0}{\sim} \frac{N_\epsilon}{2} \ln N_\epsilon.$$

Note that

$$-N_\epsilon \ln(1/\epsilon) - 2N_\epsilon \ln 2 \leq B(\epsilon) \leq \frac{\epsilon}{4} N_\epsilon.$$

These bounds on $B(\epsilon)$ show that $B(\epsilon)$ tends to decrease the upper bound, while $A(N_\epsilon)$ contributes to its growth. If $\ln N_\epsilon$ behaves like $\ln(1/\epsilon)$ up to a constant factor, then the upper bound given in Lemma 2 corresponds to a constant times $N_\epsilon \ln N_\epsilon$, and we are concerned with the point 3 of Theorem 3.

Next we state a lower bound on the metric entropy. In this case too, we want to truncate some coordinates to bring ourselves to a smaller finite dimensional space. This time we truncate the first coordinates. Let us consider the number

$$l_f = \min\{l \geq 0 : \sum_{k\geq l+1} f(k) \leq 1\}.$$

Lemma 3. *Let Λ_f be the envelope class defined by an integrable envelope function f, which satisfies*

$$\sum_{k\geq 1} f(k) \geq 2.$$

Let $\epsilon > 0$ be a positive number, and let $m \geq 1$ be an integer. Then

$$\mathcal{H}_\epsilon(\Theta_f, d) \geq \frac{1}{2} \sum_{k=l_f+1}^{l_f+m} \ln f(k) + m \ln\left(\frac{1}{\epsilon}\right) + A(m),$$

where $A(m)$ is defined as in Lemma 2:

$$A(m) = -\ln \mathrm{Vol}\left(B_{\mathbb{R}^m}(0,1)\right) \underset{m\to\infty}{\sim} \frac{m}{2} \ln m.$$

Let us now apply these two results to the exponentially decreasing envelope classes.

Proof of Proposition 5. We first apply Lemma 2 and obtain

$$N_\epsilon \leq \frac{2}{\alpha} \ln(1/\epsilon) + \frac{1}{\alpha} \ln \frac{16C}{1-e^{-\alpha}} \underset{\epsilon\to 0}{\sim} \frac{2}{\alpha} \ln(1/\epsilon). \tag{2.3}$$

Therefore we can upper bound $B(\epsilon)$:

$$\begin{aligned} B(\epsilon) &\leq \sum_{k=1}^{N_\epsilon} \ln\left(\frac{\epsilon}{4} + \sqrt{C} e^{-\frac{\alpha}{2}k}\right) \\ &\leq \int_0^{N_\epsilon} \left[\frac{\ln C}{2} - \frac{\alpha}{2}x + \ln\left(1 + \frac{\epsilon\, e^{\frac{\alpha}{2}x}}{4\sqrt{C}}\right)\right] \mathrm{d}x. \end{aligned}$$

On the other hand (2.3) gives

$$\frac{\epsilon\, e^{\frac{\alpha}{2}N_\epsilon}}{4\sqrt{C}} \leq \frac{1}{\sqrt{1-e^{-\alpha}}},$$

and then

$$\begin{aligned} B(\epsilon) &\leq -\frac{\alpha}{4}N_\epsilon^2 + \left(\frac{\ln C}{2} + \frac{1}{\sqrt{1-e^{-\alpha}}}\right) N_\epsilon \\ &\sim -\frac{\alpha}{4}N_\epsilon^2 \sim -\frac{1}{\alpha}\ln^2(1/\epsilon). \end{aligned}$$

We have also

$$A(N_\epsilon) \sim \frac{2}{\alpha}\ln(1/\epsilon)\cdot\ln\ln(1/\epsilon) = o(\ln^2(1/\epsilon))$$

and, gathering all these results, we get

$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) \leq (1+o(1))\frac{1}{\alpha}\ln^2(1/\epsilon). \tag{2.4}$$

Consider now Lemma 3. Note that the exponentially decreasing envelopes satisfy the condition $\sum_{k\geq 1} f(k) \geq 2$. Indeed this envelope is

$$f(k) = \min(1, Ce^{-\alpha k}),$$

and the condition $C > e^{2\alpha}$ entails that $f(1) = f(2) = 1$.

Since $\sum_{k\geq l_f+1} f(k) \leq 1$, $Ce^{-\alpha k} \leq 1$ for all $k \geq l_f + 1$. Therefore

$$\ln f(k) = -\alpha k + \ln C$$

for all $k \geq l_f + 1$, and

$$\begin{aligned} \frac{1}{2}\sum_{k=l_f+1}^{l_f+m} \ln f(k) &= \frac{m}{2}\ln C - \frac{\alpha}{2}\sum_{k=l_f+1}^{l_f+m} k \\ &= -(1+o(1))\frac{\alpha}{4}m^2. \end{aligned}$$

From Lemma 3 we obtain

$$\mathcal{H}_\epsilon(\Theta_f, d) \geq m\ln\left(\frac{1}{\epsilon}\right) + (1+o(1))\frac{m}{2}\ln m - (1+o(1))\frac{\alpha}{4}m^2.$$

With the choice $m = \left\lfloor \frac{2}{\alpha}\ln\left(\frac{1}{\varepsilon}\right)\right\rfloor$, the term $\frac{m}{2}\ln m$ becomes negligible and we get the following lower bound:

$$\mathcal{H}_\epsilon(\Theta_{C,\alpha}, d) \geq (1+o(1))\frac{1}{\alpha}\ln^2(1/\epsilon). \tag{2.5}$$

Note that the lower bound (2.5) is the same as the upper bound (2.4). Therefore this concludes the proof of Proposition 5. □

2.2.3 What about other envelope classes?

In [16] the redundancy of another type of envelope classes is also studied. The *power-law envelope class* $\Lambda_{C\cdot^{-\alpha}}$ is defined, for $C > 1$ and $\alpha > 1$, by the envelope function $f_{\alpha,C}(x) = \min(1, \frac{C}{x^\alpha})$. The bounds obtained in [16, Theorem 6] are

$$\begin{aligned} A(\alpha) n^{1/\alpha} \log\lfloor C\zeta(\alpha)\rfloor & \\ \leq \mathbb{R}_n(\Lambda_{C\cdot^{-\alpha}}) & \\ \leq \left(\frac{2Cn}{\alpha-1}\right)^{1/\alpha} (\log n)^{1-1/\alpha} + O(1), & \end{aligned} \tag{2.6}$$

where

$$A(\alpha) = \frac{1}{\alpha}\int_1^\infty \frac{1 - e^{-1/(\zeta(\alpha)u)}}{u^{1-1/\alpha}}\mathrm{d}u,$$

and ζ denotes the classical function $\zeta(\alpha) = \sum_{k\geq 1} \frac{1}{k^\alpha}$, for $\alpha > 1$.

If one adapts the calculus made earlier to the power-law envelope classes, one can get the following upper and lower bounds:
There are two (calculable) constants $K_1, K_2 > 0$ such that, for all $\epsilon > 0$,

$$K_1 \left(\frac{1}{\epsilon}\right)^{\frac{2}{\alpha-1}} \leq \mathcal{H}_\epsilon \leq K_2(1+o(1)) \left(\frac{1}{\epsilon}\right)^{\frac{2}{\alpha-1}} \ln\left(\frac{1}{\epsilon}\right).$$

Unfortunately this formula leaves a gap between the lower bound and the upper bound. The application of Theorem 3 makes the gap worse. Indeed the polynomial part $\left(\frac{1}{\epsilon}\right)^{\frac{2}{\alpha-1}}$ of the metric entropy causes an additional gap of $\log^{1/\alpha} n$. In practice the bounds are the following:
There are two (unknown) constants $C, c > 0$ such that, for all $n \geq 1$,

$$c(1+o(1))n^{1/\alpha} \leq R_n(\Lambda_{C\cdot^{-\alpha}}) \leq C(1+o(1))n^{1/\alpha}\log n. \tag{2.7}$$

These inequalities improve in no way the result of [16]. May a better calculation of the metric entropy improve either their lower bound or their upper bound? Anyway the metric entropy of power-law envelope classes is "too big" to efficiently apply Theorem 3: it does not leave the hope for an equivalence, as for exponentially decreasing envelope classes. To summarize, the strategy based on the metric entropy and Theorem 3 turns out efficient for "small" classes of sources.

2.3 Dirichlet's prior

Theorem 1 gives a way to calculate a lower bound of the minimax redundancy of a class of sources. Indeed the minimax redundancy is lower bounded by the Bayes redundancy of any prior. In this context, the choice of an appropriate prior is a relevant matter.

In the context of coding on a finite alphabet, the Bayes strategy using *Jeffreys' prior* plays a significant role. In the important case of the class of all stationary memoryless

sources on a finite alphabet, the Jeffrey's prior is the Dirichlet$(1/2, 1/2, \ldots, 1/2)$ prior. [27] proves that "Jeffrey's prior is asymptotically least favorable" under some conditions. [119] goes further and shows that Dirichlet's prior is asymptotically maximin but not asymptotically minimax. Then [120] proposes an asymptotically minimax modification of the Dirichlet prior.

In this section we construct a sequence of priors μ_k as the Dirichlet prior on a finite set of coordinates, supported by the envelope class and normalized. With an appropriate choice of k depending on n, priors μ_k are "almost" asymptotically least favorable for the exponentially decreasing envelope classes: Theorem 4 below states that their Bayes redundancy is equivalent, as n tends to the infinity, to the minimax redundancy.

Let us now go on in the definition of priors μ_k. First of all, we need to properly define Dirichlet's prior on the class of all stationary memoryless sources on a finite alphabet.

An stationary memoryless source $\boldsymbol{P}$ on the alphabet $\{1, 2, \ldots, k\}$ is characterized by the statistics of its first marginal distribution $P(i)$, $1 \leq i \leq k$. The class of all stationary memoryless sources on this alphabet can be parameterized as $(\boldsymbol{P_\theta})$, where the parameter $\boldsymbol{\theta}$ is an element of the simplex of $\mathbb{R}^k$

$$\mathbb{S}_k = \{(\theta_1, \theta_2, \ldots, \theta_k) \in [0,1]^k : \sum_{1 \leq i \leq k} \theta_i = 1\},$$

and $P_{\boldsymbol{\theta}}(i) = \theta_i$.

An equivalent notation of the simplex is obtained by setting $\theta_1 = 1 - \sum_{2 \leq i \leq k} \theta_i$, with $(\theta_2, \ldots, \theta_k)$ an element of the set

$$\mathbb{S}'_k = \{(\theta_2, \ldots, \theta_k) \in [0,1]^{k-1} : \sum_{2 \leq i \leq k} \theta_i \leq 1\}.$$

This makes easier to define the Lebesgue measure $\mathrm{d}\boldsymbol{\theta}$ on the simplex of $\mathbb{R}^k$, by restriction of the Lebesgue measure on $[0,1]^{k-1}$.

If we consider the sequence $X_{1:n}$ of the first n symbols produced by a source $\boldsymbol{P_\theta}$, let T_i denote the number of occurrences of symbol i

$$T_i = \sum_{j=1}^{n} \mathbb{1}_{X_j = a_i} \quad \text{for all } 1 \leq i \leq k,$$

where

$$\mathbb{1}_{X_j = a_i} = \begin{cases} 1 & \text{if } X_j = a_i, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the probability of the sequence $X_{1:n}$ under the distribution $\boldsymbol{P_\theta}$ is

$$P^n_{\boldsymbol{\theta}}(X_{1:n}) = \theta_1^{T_1} \theta_2^{T_2} \cdots \theta_k^{T_k}. \tag{2.8}$$

Dirichlet's prior has a density proportional to $\theta_1^{-1/2} \theta_2^{-1/2} \cdots \theta_k^{-1/2}$ with respect to the Lebesgue measure. The associated Bayes mixture, also called *Krichevsky-Trofimov*

mixture, is

$$KT_k(X_{1:n}) = \frac{\int_{\mathbb{S}'_k} \theta_1^{T_1-1/2}\theta_2^{T_2-1/2}\cdots\theta_k^{T_k-1/2}\mathrm{d}\boldsymbol{\theta}}{\int_{\mathbb{S}'_k} \theta_1^{-1/2}\theta_2^{-1/2}\cdots\theta_k^{-1/2}\mathrm{d}\boldsymbol{\theta}} = \frac{D_k(T_1+\frac{1}{2},\cdots,T_k+\frac{1}{2})}{D_k(\frac{1}{2},\cdots,\frac{1}{2})}, \tag{2.9}$$

where $D_k(\cdot,\ldots,\cdot)$ denotes the Dirichlet integrals

$$D_k(\lambda_1,\ldots,\lambda_k) = \int_{\mathbb{S}'_k} \theta_1^{\lambda_1-1}\theta_2^{\lambda_2-1}\cdots\theta_k^{\lambda_k-1}\mathrm{d}\boldsymbol{\theta} = \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)\cdots\Gamma(\lambda_k)}{\Gamma\left(\sum_{i=1}^k \lambda_i\right)}. \tag{2.10}$$

Another classical definition of Krichevsky-Trofimov mixtures gives the conditional probabilities: if $x_{1:n}$ is a message on the alphabet $\{1,\ldots,k\}$, then, for all $0 \le i \le n-1$ and for all $1 \le j \le k$,

$$KT_k(X_{i+1} = j | X_{1:i} = x_{1:i}) = \frac{n_i^j + \frac{1}{2}}{i + \frac{k}{2}}, \tag{2.11}$$

where n_i^j is the number of occurrences of symbol j in $x_{1:i}$.

Let f be an integrable envelope function, and Λ_f be the associated envelope class. Choose any fixed $m \in \mathbb{N}_*$ such that

$$\sum_{i \ge m} f(i) < 1.$$

For $k \ge m+1$, let Θ_k denote the subset of Θ_f defined by

$$\Theta_k = \Big\{(\theta_1, 0, \ldots, 0, \theta_m, \ldots, \theta_k, 0, \ldots) : \theta_1 = 1 - \sum_{i=m}^k \theta_i \text{ and } \forall m \le i \le k,\ 0 \le \theta_i \le f(i)\Big\}.$$

The presence of zeros between θ_1 and θ_m is motivated by computational reasons.

Let μ_k denote the prior on Θ_k which is proportional to the Dirichlet prior μ on the simplex $\mathbb{S}_{k-m+2}$, using the coordinates $(\theta_1, \theta_m, \ldots, \theta_k)$:

$$\begin{aligned}
&\mathrm{d}\mu_k(\theta_1, 0, \ldots, 0, \theta_m, \ldots, \theta_k, 0, \ldots) \\
&\quad = \frac{\theta_1^{-1/2}\theta_m^{-1/2}\cdots\theta_k^{-1/2}\,\mathrm{d}\boldsymbol{\theta}}{\int_{\Theta_k} \theta_1^{-1/2}\theta_m^{-1/2}\cdots\theta_k^{-1/2}\,\mathrm{d}\boldsymbol{\theta}} \\
&\quad = \frac{\int_{\mathbb{S}_{k-m+2}} \theta_1^{-1/2}\theta_m^{-1/2}\cdots\theta_k^{-1/2}\,\mathrm{d}\boldsymbol{\theta}}{\int_{\Theta_k} \theta_1^{-1/2}\theta_m^{-1/2}\cdots\theta_k^{-1/2}\,\mathrm{d}\boldsymbol{\theta}} \mathrm{d}\mu(\theta_1, \theta_m, \ldots, \theta_k).
\end{aligned}$$

In this formula, $\mathrm{d}\boldsymbol{\theta}$ is the Lebesgue measure on the simplex $\mathbb{S}_{k-m+2}$ indexed by $(\theta_1, \theta_m, \ldots, \theta_k)$, and Θ_k is identified with its projection on the simplex. Similarly, let $\widetilde{KT}_{k-m+2}$ denote the Bayes mixture associated to the Dirichlet prior μ indexed by $(\theta_1, \theta_m, \ldots, \theta_k)$.

Theorem 4. *Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Let $k_n = \left\lfloor \frac{1}{\alpha \log e} \log n \right\rfloor$. Then the sequence of priors μ_{k_n} verifies*

$$R_{n,\mu_{k_n}}(C,\alpha) \geq (1+o(1))\frac{1}{4\alpha \log e}\log^2 n,$$

where $o(1)$ is a function $g(n)$ such that $g(n) \to 0$ as $n \to \infty$.

Additionally, Theorem 4 enables us to retrieve in an independent way the lower bound of the minimax redundancy obtained in the section 2.2. Theorem 4 in proved in Appendix 2.B.

What about other envelope classes? For power-law envelope classes the choice $k_n = \left\lfloor \frac{n^{-1/\alpha}}{e} \right\rfloor$ gives

$$R_n(\Lambda_{C\cdot^{-\alpha}}) \geq R_{n,\mu_{k_n}}(\Lambda_{C\cdot^{-\alpha}}) \geq (1+o(1))\frac{\alpha}{2e}n^{1/\alpha}.$$

This result is similar to those presented in (2.6) and (2.7), and that is good. However it does not permit to fill the gap between this lower bound and the upper bound given in (2.6).

2.4 AutoCensuring Code

This section presents a new algorithm called AutoCensuring Code (`ACcode`). It is in fact a modification of the Censuring Code proposed by Boucheron, Garivier and Gassiat in [16]. We keep the idea that big symbols are very few, and must be encoded differently, with an Elias code. Smaller symbols are encoded by arithmetic coding based on Krichevsky-Trofimov mixtures, which are known to be effective for finite alphabets. Our innovation is a data-driven cutoff $M_i = \sup_{1\leq k\leq i} X_k$ used to encode X_{i+1}: with this choice we do not need to know the exact parameters of the exponentially decreasing envelope.

`ACcode` is a prefix code on the set of all finite length messages, and it works on line. Its maximum redundancy on an exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$ is equivalent to the minimax redundancy of this class of sources. Furthermore `ACcode` is adaptive, as the same algorithm satisfies this property with all exponentially decreasing envelope classes. This is formulated in the following theorem, proved in Appendix 2.C. Let $\texttt{ACcode}(x_{1:n})$ denote the binary string produced by `ACcode` when it encodes the message $x_{1:n}$, and let $l(\cdot)$ denote the length of a string.

Theorem 5. *For any positive numbers C and α satisfying $C > e^{2\alpha}$,*

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}[l(\texttt{ACcode}(X_{1:n})) - H(P^n)] \underset{n\to\infty}{\sim} R_n(C,\alpha).$$

The difference between the redundancy of `ACcode` and the minimax redundancy is not necessarily bounded: there may exist codes whose redundancy is smaller than the redundancy of `ACcode`, but with a benefit asymptotically negligible with respect to $\log^2 n$.

Additionally, Theorem 5 enables us to retrieve the upper bound of the minimax redundancy obtained in the section 2.2.

Let us now define `ACcode`. Let $n \geq 1$ be some positive integer, and let $x_{1:n} = x_1 x_2 \dots x_n$ be a string from $\mathbb{N}_*^n$ to be encoded. We define the sequence of maxima

$$m_0 = 0 \text{ and } m_i = \sup_{1 \leq k \leq i} x_k, \text{ for all } 1 \leq i \leq n.$$

The sequence $(m_i)_{1 \leq i \leq n}$ is non-decreasing, piecewise constant. For $1 \leq i \leq n$, let $n_i^0 = \sum_{j=1}^{i} \mathbb{1}_{m_j > m_{j-1}}$ be the number of plateaus between 1 and i. For $1 \leq k \leq n_n^0$, let $\widetilde{m}_k$ be the k^{th} new maximum:

$$m_i = \widetilde{m}_{n_i^0}. \tag{2.12}$$

We define also $\widetilde{m}_0 = 0$. Let string $\widetilde{\boldsymbol{m}}$ be the sequence $(\widetilde{m}_1 - \widetilde{m}_0 + 1), \dots, (\widetilde{m}_{n_n^0} - \widetilde{m}_{n_n^0 - 1} + 1), 1$. $\widetilde{\boldsymbol{m}}$ is encoded into a binary string `C2` by applying Elias penultimate code (see [41]) to each number in $\widetilde{\boldsymbol{m}}$. It is a prefix code which uses $l_E(x)$ bits to encode a positive integer x, with

$$\begin{aligned} &l_E(1) = 1, \\ &l_E(x) = 1 + \lfloor \log x \rfloor + 2 \lfloor \log \lfloor \log x \rfloor + 1 \rfloor \quad \text{if } x \geq 2. \end{aligned} \tag{2.13}$$

Meanwhile the sequence of censored symbols is encoded using side information from $\widetilde{\boldsymbol{m}}$. Consider the censored sequence $\widetilde{x}_{1:n} = \widetilde{x}_1 \widetilde{x}_2 \dots \widetilde{x}_n$ defined by

$$\widetilde{x}_i = x_i \mathbb{1}_{x_i \leq m_{i-1}} = \begin{cases} x_i & \text{if } x_i \leq m_{i-1}, \\ 0 & \text{otherwise.} \end{cases}$$

All symbols greater than m_{i-1} are encoded together: they are replaced by the extra symbol 0, and this extra symbol is encoded instead. 0 has a special use in our setting: it makes the decoder to know when m_i changes, and that the new value has to be read in `C2`. We add at the end of $\widetilde{x}_{1:n}$ an additional 0, which acts as a termination signal together with the last 1 in $\widetilde{\boldsymbol{m}}$. This makes our code to be prefix on the set of all finite length messages (whatever n).

Therefore we produce the binary string `C1` by arithmetic coding of $\widetilde{x}_{1:n}0$. The conditional coding probabilities are defined by

$$\begin{aligned} Q_{i+1}(\widetilde{X}_{i+1} = j | X_{1:i} = x_{1:i}) &= \frac{n_i^j + \frac{1}{2}}{i + \frac{m_i + 1}{2}} \quad \text{if} \quad 1 \leq j \leq m_i, \\ Q_{i+1}(\widetilde{X}_{i+1} = 0 | X_{1:i} = x_{1:i}) &= \frac{1/2}{i + \frac{m_i + 1}{2}}, \end{aligned}$$

where for $j \geq 1$ and $i \geq 0$, n_i^j is the number of occurrences of symbol j in $x_{1:i}$ (with convention $n_0^j = 0$ for all $j \geq 1$).

If $i \leq n - 1$, the event $\{\widetilde{X}_{i+1} = 0\}$ is equal to $\{X_{i+1} > M_i\}$. If $x_{i+1} = j > m_i$, then $n_i^j = 0$, and we still have

$$Q_{i+1}(\widetilde{X}_{i+1} = 0 | X_{1:i} = x_{1:i}) = \frac{n_i^j + \frac{1}{2}}{i + \frac{m_i + 1}{2}}.$$

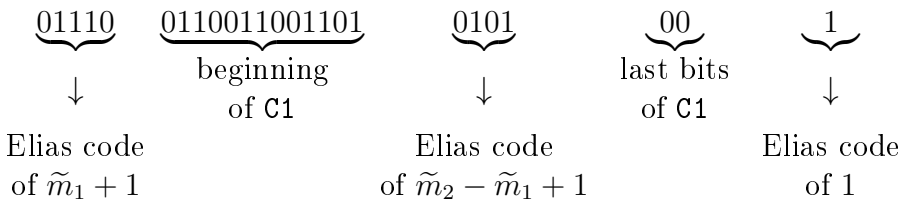

Figure 2.1: Example of `ACcode`

In the sequel we note the coding probability used to encode the entire string $\widetilde{x}_{1:n}0$ by

$$Q^{n+1}(\widetilde{x}_{1:n}0) = Q_{n+1}(0|x_{1:n}) \prod_{i=0}^{n-1} Q_{i+1}(\widetilde{x}_{i+1}|x_{1:i}).$$

A remark we can do is that the symbol 0 is always considered as new: when $x_{i+1} > m_i$, we encode 0 but we increment the counter $n_i^{x_{i+1}}$. (This choice has been made to simplify the calculation of the redundancy of `ACcode`, but we suspect that changing this behavior could improve the performances.)

Now we have defined `C1` and `C2`, we have to describe how they are transmitted. To keep our code on line, we overlap these two strings in the following way.

Arithmetic code needs a certain amount of bits, say l_i, to send the first i symbol of $\widetilde{x}_{1:n}$. Unfortunately, l_i depends on whether $i = n + 1$ or not. In previous case $l_{n+1} = \lceil -\log Q^{n+1}(\widetilde{x}_{1:n}0) \rceil + 1$, and in later one l_i depends on the following symbols and has to be computed.
`ACcode` begins with `C2`, by the transmission of the Elias code of $\widetilde{m}_1 + 1$. Then the transmission of `C1` is initiated. Suppose that $\widetilde{x}_i = 0$ and $n_i^0 = k$. As soon as l_i bits of `C1` have been sent, the `ACcode` algorithm sends the Elias code of $\widetilde{m}_k - \widetilde{m}_{k-1} + 1$. Then `C1` is transmitted again, from the next bit.

To decode the i^{th} symbol in `C1`, the knowledge of the current maximum m_{i-1} is needed; it is obtained from the beginning of the string `C2`. The decoder also needs the counters $(n_{i-1}^j)_{j \geq 1}$, which can be computed from the first $i - 1$ decoded symbols. As soon as l_i bits of `C1` have been received, $\widetilde{x}_i$ can be decoded. When the decoder meets a 0 at the i^{th} position, he knows that the next bits are the Elias code of the next symbol in $\widetilde{\boldsymbol{m}}$, and deduces m_i via (2.12). Since the Elias code is prefix, the decoder knows when he receives `C1` again. Then the $(i+1)^{\text{th}}$ symbol can be processed.

Fig. 2.1 shows an illustration of the transmission process. In this example, the initial message is $x_{1:4} = 5, 3, 2, 7$. Then the message encoded in `C1` is $\widetilde{x}_{1:4}0 = 0, 3, 2, 0, 0$. 13 bits are needed to transmit the second 0, and 15 bits for the last one. In `C2` we transmit $\widetilde{\boldsymbol{m}} = 6, 3, 1$.

In the previous example, exact calculations have been performed, but this is not sensible for a practical implementation of arithmetic coding. Some rule is needed to set the precision in calculus, and it must be used by both coder and decoder. To avoid a too big extra redundancy caused by approximations, precision can grow as n grows. For instance, calculations can be made in memory with a further precision of $2\lceil \log i \rceil$ bits, in addition to the $\lceil -\log Q^i(\widetilde{x}_{1:i}) \rceil + 1$ bits needed to encode $x_{1:i}$; this insures that the extra redundancy is bounded.

Acknowledgment

I wish thank E. Gassiat for her helpful advice, for the many ideas in this paper she suggested me, and for her constant availability to my questions.

Thanks also to A. Garivier and to S. Boucheron for the useful discussions we had. I don't forget the ideas I got from my classmates, especially R. Imekraz.

2.A Metric entropy of exponentially decreasing envelope classes

We give here the proofs of the lemmas we stated in subsection 2.2.2.

Proof of Lemma 2. N_ϵ denotes the threshold from which we want to truncate the coordinates. If $y = (y_k)_{k\geq 1}$ is an element of A_f, its truncated version is $\widetilde{y} = (y_k \mathbb{1}_{k\leq N_\epsilon})_{k\geq 1}$. Then

$$\|y - \widetilde{y}\| = \sqrt{\sum_{k\geq N_\epsilon+1} y_k^2} \leq \sqrt{\sum_{k\geq N_\epsilon+1} f(k)} \leq \frac{\epsilon}{4}.$$

Suppose now that S is an $\epsilon/4$-cover of $\{y \in A_f : \forall n \geq N_\epsilon, y_n = 0\}$. Let z denote an element of A_f. Then it exists some $y \in S$ such that $\|\widetilde{z} - y\| \leq \epsilon/4$. Thus $\|z - y\| \leq \epsilon/2$, and S is an $\epsilon/2$-cover of A_f. This leads to

$$\begin{aligned}
\mathcal{D}_\epsilon(\Theta_f, d) &\leq \mathcal{N}_{\epsilon/2}(A_f, \|\cdot\|_{\ell^2}) \\
&\leq \mathcal{N}_{\epsilon/4}\left(\prod_{1\leq k\leq N_\epsilon} \left[0, \sqrt{f(k)}\right], \|\cdot\|_{\mathbb{R}^{N_\epsilon}}\right) \\
&\leq \mathcal{M}_{\epsilon/4}\left(\prod_{1\leq k\leq N_\epsilon} \left[0, \sqrt{f(k)}\right], \|\cdot\|_{\mathbb{R}^{N_\epsilon}}\right) \\
&\leq \frac{\mathrm{Vol}\left(\prod_{1\leq k\leq N_\epsilon} \left[-\frac{\epsilon}{8}, \sqrt{f(k)} + \frac{\epsilon}{8}\right]\right)}{\mathrm{Vol}\left(B_{\mathbb{R}^{N_\epsilon}}(0, \frac{\epsilon}{8})\right)} \\
&\leq \left(\frac{\epsilon}{8}\right)^{-N_\epsilon} \frac{\Gamma(\frac{N_\epsilon}{2} + 1)}{\pi^{N_\epsilon/2}} \prod_{k=1}^{N_\epsilon} \left(\sqrt{f(k)} + \frac{\epsilon}{4}\right).
\end{aligned}$$

A first consequence of that calculus is that $\mathcal{D}_\epsilon(\Theta_f, d)$ is finite for all $\epsilon > 0$. The first assertion of Lemma 2 is then obtained by applying the logarithm function.

The rest of Lemma 2 follows from the Feller bounds, in their version proposed by [117, ch. XII]: For all $x > 0$, there exists $\beta \in [0, 1]$ such that

$$\Gamma(x) = \sqrt{2\pi}\, x^{x-1/2} e^{-x} e^{\frac{\beta}{12x}}. \tag{2.14}$$

Therefore

$$\begin{aligned} A(N) &= -\frac{N}{2}\ln\pi + \frac{N+1}{2}\ln\left(\frac{N}{2}+1\right) - \left(\frac{N}{2}+1\right) \\ &\quad + \frac{\ln(2\pi)}{2} + \frac{\beta}{12\left(\frac{N}{2}+1\right)} \\ &\sim \frac{N}{2}\ln N. \end{aligned}$$

□

Proof of Lemma 3. Let $m \geq 1$ be an integer. We project the set $A_f \cap \{\|x\| = 1\}$ over the m-dimensional space

$$E_m = \{0\}^{l_f} \times \mathbb{R}^m \times \{0\}^{\{k:k\geq l_f+m+1\}}$$

generated by the coordinates from $l_f + 1$ to $l_f + m$. The resulting set is isomorphic to the rectangle $\prod_{k=l_f+1}^{l_f+m}\left[0, \sqrt{f(k)}\right]$. This leads to

$$\begin{aligned} \mathcal{D}_\epsilon(\Theta_f, d) &\geq \mathcal{N}_\epsilon\left(\prod_{k=l_f+1}^{l_f+m}\left[0,\sqrt{f(k)}\right], \|\cdot\|_{\mathbb{R}^m}\right) \\ &\geq \frac{\mathrm{Vol}\left(\prod_{k=l_f+1}^{l_f+m}\left[0,\sqrt{f(k)}\right]\right)}{\mathrm{Vol}\left(B_{\mathbb{R}^m}(0,\epsilon)\right)}. \end{aligned}$$

It only remains to apply the logarithm function. □

2.B Proof of Theorem 4

We use the following, which is a modification of Proposition 1 in [119].

Proposition 6. *Let $\boldsymbol{\theta}$ be an element of the simplex $\mathbb{S}_k$, and KT_k the Krichevsky-Trofimov mixture. Then*

$$D(P^n_{\boldsymbol{\theta}}; KT_k) \geq \frac{k-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma(1/2)^k}{\Gamma(k/2)} - \frac{5k}{3}\log e.$$

Proof of Proposition 6. Let $E_{\boldsymbol{\theta}}$ denote the expected value under the distribution $\boldsymbol{P}_{\boldsymbol{\theta}}$. The calculus is made using natural logarithm:

$$\begin{aligned} &(\ln 2)\, D(P^n_{\boldsymbol{\theta}}; KT_k) \\ &\quad = E_{\boldsymbol{\theta}} \ln\frac{P^n_{\boldsymbol{\theta}}(X_{1:n})}{KT_k(X_{1:n})} \\ &\quad = E_{\boldsymbol{\theta}}\sum_{i=1}^k T_i \ln\theta_i - E_{\boldsymbol{\theta}}\ln\frac{D_k(T_1+\frac{1}{2},\cdots,T_k+\frac{1}{2})}{D_k(\frac{1}{2},\cdots,\frac{1}{2})} \\ &\quad = \ln\frac{\Gamma(1/2)^k}{\Gamma(k/2)} + \overbrace{\sum_{i=1}^k n\theta_i\ln\theta_i - E_{\boldsymbol{\theta}}\ln\frac{\prod_{i=1}^k\Gamma(T_i+\frac{1}{2})}{\Gamma(n+\frac{k}{2})}}^{\text{(A)}}. \end{aligned} \tag{2.15}$$

The second line comes from (2.8) and (2.9), and the third from (2.10). We use now the Feller bounds given in (2.14), with β_0 denoting the coefficient corresponding to the formula of $\Gamma(n+\frac{k}{2})$, and β_i being the coefficient in the formula of $\Gamma(T_i+\frac{1}{2})$.

$$\begin{aligned}(\mathrm{A}) &= \sum_{i=1}^{k} n\theta_i \ln\theta_i - E_{\boldsymbol{\theta}} \ln \frac{\prod_{i=1}^{k}\left(\sqrt{2\pi}(T_i+\frac{1}{2})^{T_i}\right)}{\sqrt{2\pi}(n+\frac{k}{2})^{n+(k-1)/2}} \\ &\quad - \sum_{i=1}^{k} E_{\boldsymbol{\theta}} \frac{\beta_i}{12(T_i+\frac{1}{2})} + \frac{\beta_0}{12(n+\frac{k}{2})} \\ &\geq -\frac{k-1}{2}\ln 2\pi - \frac{k}{6} \\ &\quad + \overbrace{\sum_{i=1}^{k}\left(n\theta_i\ln\theta_i - E_{\theta_i}T_i\ln\left(T_i+\frac{1}{2}\right)\right)}^{(\mathrm{B})} \\ &\quad + \overbrace{\left(n+\frac{k-1}{2}\right)\ln\left(n+\frac{k}{2}\right)}^{(\mathrm{C})}.\end{aligned}$$

Now we lower bound separately (B) and (C). For the later,

$$\begin{aligned}(\mathrm{C}) &= \left(n+\frac{k-1}{2}\right)\ln n + \left(n+\frac{k-1}{2}\right)\ln\left(1+\frac{k}{2n}\right) \\ &\geq n\ln n + \frac{k-1}{2}\ln n.\end{aligned}$$

On the other hand,

$$\begin{aligned}(\mathrm{B}) &= -n\ln n + \sum_{i=1}^{k} \overbrace{(n\theta_i\ln n\theta_i - E_{\theta_i}T_i\ln T_i)}^{(\mathrm{B}_i)} \\ &\quad - \sum_{i=1}^{k} \overbrace{E_{\theta_i}T_i\ln\left(1+\frac{1}{2T_i}\right)}^{\leq 1/2}.\end{aligned}$$

From the relation $\ln t \leq t-1$, with $t = \frac{T_i}{n\theta_i}$, we get

$$\begin{aligned}\ln T_i - \ln n\theta_i &\leq \frac{T_i - n\theta_i}{n\theta_i} \\ T_i\ln T_i - T_i\ln n\theta_i &\leq \frac{(T_i-n\theta_i)^2}{n\theta_i} + (T_i - n\theta_i) \\ E_{\theta_i}T_i\ln T_i - n\theta_i\ln n\theta_i &\leq \frac{\operatorname{Var} T_i}{n\theta_i} = 1-\theta_i.\end{aligned}$$

As a consequence,

$$\begin{aligned}(\mathrm{B}_i) &\geq -(1-\theta_i) \geq -1, \\ (\mathrm{B}) &\geq -n\ln n - \frac{3k}{2},\end{aligned}$$

and

$$(\mathrm{A}) \geq \frac{k-1}{2} \ln \frac{n}{2\pi} - \frac{5k}{3}.$$

All that remains is to collect the different elements of (2.15) to get the announced result. □

Proof of Theorem 4. To simplify the notations, let us define

$$\begin{aligned} C(k) &= \int_{\Theta_k} \theta_1^{-1/2} \theta_m^{-1/2} \cdots \theta_k^{-1/2} \, \mathrm{d}\boldsymbol{\theta}, \\ D(k) &= \int_{\mathbb{S}_{k-m+2}} \theta_1^{-1/2} \theta_m^{-1/2} \cdots \theta_k^{-1/2} \, \mathrm{d}\boldsymbol{\theta} \\ &= \frac{\Gamma(1/2)^{k-m+2}}{\Gamma(\frac{k-m+2}{2})}. \end{aligned}$$

In pratice, prior μ_k is supported by the alphabet $A_k = \{1, m, m+1, \ldots, k\}$. The corresponding Bayes strategy doesn't encode messages with other symbols. As a consequence, it is far from being asymptotically minimax!

Let $x_{1:n}$ be an element of A_k^n. Then

$$\begin{aligned} M_{n,\mu_k}(x_{1:n}) &= \frac{D(k)}{C(k)} \int_{\Theta_k} P_{\boldsymbol{\theta}}(x_{1:n}) \, \mathrm{d}\mu(\theta_1, \theta_m, \ldots, \theta_k) \\ &\leq \frac{D(k)}{C(k)} \widetilde{KT}_{k-m+2}(x_{1:n}). \end{aligned}$$

Therefore, if $\boldsymbol{\theta}$ is an element of Θ_k, Proposition 6 entails

$$\begin{aligned} D(P_{\boldsymbol{\theta}}^n; M_{n,\mu_k}) &\geq D(P_{\boldsymbol{\theta}}^n; \widetilde{KT}_{k-m+2}) + \log C(k) - \log D(k) \\ &\geq \frac{k-m+1}{2} \log \frac{n}{2\pi} + \log C(k) \\ &\quad - \frac{5(k-m+2)}{3} \log e. \end{aligned}$$

Consequently,

$$\begin{aligned} R_{n,\mu_k}(\Lambda_f) &= \int_{\Theta_k} D(P_{\boldsymbol{\theta}}^n; M_{n,\mu_k}) \, \mathrm{d}\mu_k(\boldsymbol{\theta}) \\ &\geq \log C(k) + \frac{k}{2} \log n \\ &\quad - \frac{10 \log e + 3 \log 2\pi}{6} k - \frac{m-1}{2} \log n \\ &\quad + \left(\frac{m-1}{2} \log 2\pi + \frac{5(m-2)}{3} \log e \right). \end{aligned} \tag{2.16}$$

Now, let us calculate $C(k)$. First note that the choice of m made before is such that all values of $(\theta_m, \ldots, \theta_k)$ in the rectangle $[0, f(m)] \times \cdots \times [0, f(k)]$ are possible. It allows

us to write the integrals over Θ_k as integrals over that rectangle.

$$\begin{aligned} C(k) &= \int_{\Theta_k} \frac{\mathrm{d}\theta_m \cdots \mathrm{d}\theta_k}{\sqrt{\theta_1 \theta_m \cdots \theta_k}} \\ &\geq \prod_{i=m}^{k} \int_0^{f(i)} \frac{\mathrm{d}\theta_i}{\sqrt{\theta_i}} \\ &= \prod_{i=m}^{k} 2\sqrt{f(i)}. \end{aligned}$$

At this point we need to specify f. In the case of the exponentially decreasing envelope class $\Lambda_{Ce^{-\alpha\cdot}}$, $f(i) = \min(1, Ce^{-\alpha i})$. Since $\sum_{i \geq m} f(i) < 1$, $f(i) = Ce^{-\alpha i} < 1$ for all $i \geq m$. Thus

$$C(k) \geq \prod_{i=m}^{k} 2\sqrt{C} e^{-\frac{\alpha}{2} i},$$

and

$$\log C(k) \geq (k - m + 1) \left(1 + \frac{\log C}{2}\right) - \frac{\alpha \log e}{2} \sum_{i=m}^{k} i.$$

If we plug it in (2.16), we get $R_{n,\mu_k}(C, \alpha) \geq g(n, k)$, where

$$\begin{aligned} g(n,k) = -\frac{\alpha \log e}{4} k^2 + \frac{k}{2} \log n + \Bigg[1 + \frac{\log C}{2} - \frac{\alpha \log e}{4} \\ - \frac{5 \log e}{3} - \frac{\log 2\pi}{2}\Bigg] k - \frac{m-1}{2} \log n \\ + \Bigg[\frac{\alpha(m^2 + m - 1) \log e}{4} - (m-1)\left(1 + \frac{\log C}{2}\right) \\ + \frac{m-1}{2} \log 2\pi + \frac{5(m-2) \log e}{3}\Bigg]. \end{aligned}$$

With the choice $k_n = \left\lfloor \frac{1}{\alpha \log e} \log n \right\rfloor$, only the first two terms matter, and

$$g(n, k_n) \sim \frac{1}{4\alpha \log e} \log^2 n.$$

This achieves the proof of Theorem 4. □

2.C Redundancy of `ACcode`

2.C.1 Moments of M_n

We first need a lemma which contains several useful results about the moments of M_n.

Lemma 4. *Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Then, for all $n \geq 1$,*

1.
$$\sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_P[M_n] \leq \frac{1}{\alpha}\left(\ln n + \ln \frac{C}{1-e^{-\alpha}} + 1\right).$$

2.
$$\sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_P\left[M_n \mathbb{1}_{M_n > \frac{1}{\alpha}\ln\frac{Cn^2}{1-e^{-\alpha}}}\right] = O\left(\frac{\ln n}{n}\right).$$

3.
$$\sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_P[M_n \ln M_n] = o(\ln^2 n).$$

Proof. Let F denote the distribution function associated with P. For $t \geq 0$, we have

$$\begin{aligned} P(X_1 > t) &= \sum_{k \geq \lfloor t \rfloor + 1} P(k) \\ &\leq \frac{C}{1-e^{-\alpha}} e^{-\alpha(\lfloor t \rfloor + 1)} \\ &\leq e^{-\alpha(t-\beta)}, \end{aligned}$$

where $\beta = \frac{1}{\alpha}\ln\frac{C}{1-e^{-\alpha}}$. Therefore $F(t) \geq G(t)$ for all $t \in \mathbb{R}$, where

$$G(t) = \mathbb{1}_{t \geq \beta}\left(1 - e^{-\alpha(t-\beta)}\right).$$

G is the distribution function of a random variable $\beta + Y$, where Y follows the exponential distribution with parameter α.

Let $U_1, \ldots, U_n$ be n iid random variables following the uniform distribution on $[0,1]$. For $1 \leq i \leq n$, let us define

$$\begin{aligned} X'_i &= F^{-1}(U_i) \\ Y_i &= G^{-1}(U_i) - \beta, \end{aligned}$$

where F^{-1} and G^{-1} denote the pseudo-inverses of F and G:

$$\forall t \in [0,1], \quad F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}.$$

Then the n-dimensional vector $X'_{1:n} = (X'_1, \ldots, X'_n)$ has the same distribution as $X_{1:n}$, and the maxima $M'_n = \sup_{1 \leq i \leq n} X'_i$ and M_n follow the same distribution.

On the other hand, the relation $F \geq G$ entails $X'_i \leq \beta + Y_i$, for all $1 \leq i \leq n$. As the consequence, if $M''_n = \sup_{1 \leq i \leq n} Y_i$ denotes the maximum of all Y_i, we have $M'_n \leq \beta + M''_n$. Since the random variables Y_i are independent, the probability distribution of M''_n is easy to calculate. Indeed for all $t > 0$,

$$\begin{aligned} P(M''_n \leq t) &= P(\forall\ 1 \leq i \leq n,\ Y_i \leq t) \\ &= \left(1 - e^{-\alpha t}\right)^n. \end{aligned}$$

We can write down the density function of M''_n:

$$f(t) = \begin{cases} n\,\alpha\, e^{-\alpha t}(1-e^{-\alpha t})^{n-1} & \text{if } t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Now we look for an upper bound of $E[M_n]$ by taking advantage of the knowledge of that distribution:

$$\begin{aligned} E[M_n] &= E[M'_n] \\ &\leq E[\beta + M''_n] \\ &= \beta + \int_0^\infty t\, n\, \alpha\, e^{-\alpha t}(1-e^{-\alpha t})^{n-1}\, \mathrm{d}t \\ &= \beta + \int_0^\infty \left(1 - (1-e^{-\alpha t})^n\right)\, \mathrm{d}t \end{aligned}$$

integrating by parts. Use now the change of variables

$$\begin{cases} u = 1 - e^{-\alpha t} \\ t = \frac{-\ln(1-u)}{\alpha} \end{cases}$$

$$\begin{aligned} E[M_n] &\leq \beta + \frac{1}{\alpha}\int_0^1 \frac{1-u^n}{1-u}\mathrm{d}u \\ &\leq \frac{1}{\alpha}\left(\ln n + 1 + \ln\frac{C}{1-e^{-\alpha}}\right). \end{aligned}$$

Since the upper bound does not depend on P, that achieves the proof of the point 1. We can handle the point 2 in the same way. For all $t > 0$, we have

$$\begin{aligned} E\left[M_n \mathbb{1}_{M_n > \beta + t}\right] &\leq E\left[(\beta + M''_n)\mathbb{1}_{M''_n > t}\right] \\ &\leq \int_t^\infty (\beta + u) n\, \alpha\, e^{-\alpha u}\, \mathrm{d}u \\ &= n e^{-\alpha t}\left(t + \frac{1}{\alpha} + \beta\right). \end{aligned}$$

With $t = \frac{2}{\alpha}\ln n$, we get the second point of Lemma 4.

The third item is similar. Since the function $x \mapsto x \ln x$ is increasing on $[1, +\infty)$ and $1 \leq M'_n \leq \beta + M''_n$, we have

$$\begin{aligned} &E\left[M_n \ln M_n\right] \\ &\quad\leq E\left[(\beta + M''_n)\ln(\beta + M''_n)\right] \\ &\quad= E\left[\mathbb{1}_{M''_n \leq \beta}(\beta + M''_n)\ln(\beta + M''_n)\right] \\ &\qquad+ E\left[\mathbb{1}_{M''_n > \beta}\mathbb{1}_{M''_n \leq \frac{2}{\alpha}\ln n}(\beta + M''_n)\ln(\beta + M''_n)\right] \\ &\qquad+ E\left[\mathbb{1}_{M''_n > \beta}\mathbb{1}_{M''_n > \frac{2}{\alpha}\ln n}(\beta + M''_n)\ln(\beta + M''_n)\right] \\ &\quad\leq 2\beta\ln(2\beta) + \frac{4}{\alpha}(\ln n)\ln\left(\frac{4}{\alpha}\ln n\right) \\ &\qquad+ E\left[2M''_n \ln(2M''_n)\mathbb{1}_{M''_n > \frac{2}{\alpha}\ln n}\right] \\ &\quad\leq 2\beta\ln(2\beta) + \left(\frac{4}{\alpha}\ln\frac{4}{\alpha}\right)\ln n + \frac{4}{\alpha}(\ln n)(\ln\ln n) \\ &\qquad+ E\left[4M''^2_n \mathbb{1}_{M''_n > \frac{2}{\alpha}\ln n}\right]. \end{aligned}$$

Let us define

$$\gamma(n) = 2\beta \ln(2\beta) + \left(\frac{4}{\alpha} \ln \frac{4}{\alpha}\right) \ln n + \frac{4}{\alpha}(\ln n)(\ln \ln n).$$

Note that $\gamma(n) = o(\ln^2 n)$. Then

$$\begin{aligned} E\left[M_n \ln M_n\right] &\leq \gamma(n) + \int_{\frac{2}{\alpha}\ln n}^{\infty} 4u^2 n\, \alpha\, e^{-\alpha u}\, \mathrm{d}u \\ &= \gamma(n) + \frac{4ne^{-2\ln n}}{\alpha^2}(4\ln^2 n + 4\ln n + 2). \end{aligned}$$

Taking the supremum over P, we get

$$\begin{aligned} \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_P\left[M_n \ln M_n\right] &\leq \gamma(n) + \frac{16\ln^2 n + 16 \ln n + 8}{\alpha^2\, n} \\ &= o(\ln^2 n). \end{aligned}$$

□

2.C.2 Contribution of C1

Proposition 7. *Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Then*

$$\begin{aligned} &\sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_{P^n}[-\log Q^n(\widetilde{X}_{1:n}) - H(P^n)] \\ &\qquad \leq (1 + o(1)) \frac{1}{4\alpha \log e} \log^2 n. \end{aligned}$$

Proof. We give here the sketch of the proof, and we delay the proofs of (2.17), (2.18), (2.19), and (2.20).

Here we deal with the quantity

$$(\mathrm{A}) = \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_{P^n}[-\log Q^n(\widetilde{X}_{1:n}) - H(P^n)].$$

that corresponds to the contribution of C1. As we saw in Section 2.4, the coding probability Q^n is based on Krichevsky-Trofimov mixtures. For $k \geq 1$, let KT_k denote the usual Krichevsky-Trofimov mixture on the alphabet $\{1, \ldots, k\}$, whose conditional probabilities are, for all $0 \leq i \leq n-1$ and for all $1 \leq j \leq k$,

$$KT_k(X_{i+1} = j | X_{1:i} = x_{1:i}) = \frac{n_i^j + \frac{1}{2}}{i + \frac{k}{2}}.$$

Let us choose $k = m_n + 1$. In this case, there is a simple relation between KT_{m_n+1} and Q^n. For any sequence of n positive integers $x_{1:n} \in \mathbb{N}_*^n$,

$$\begin{aligned} &Q_{i+1}(\widetilde{X}_{i+1} = \widetilde{x}_{i+1} | X_{1:i} = x_{1:i}) \\ &\qquad = \frac{2i + 1 + m_n}{2i + 1 + m_i} KT_{m_n+1}(X_{i+1} = x_{i+1} | X_{1:i} = x_{1:i}). \end{aligned}$$

As a consequence, we can link the redundancy of Q^n to the redundancy of KT_{m_n+1}:

$$\log Q^n(\widetilde{X}_{1:n}) = \log KT_{M_n+1}(X_{1:n}) + \sum_{i=0}^{n-1} \log \frac{2i+1+M_n}{2i+1+M_i}$$

and therefore

$$(\mathrm{A}) = \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \left(\overbrace{E_{P^n}\big[-\log KT_{M_n+1}(X_{1:n}) - H(P^n)\big]}^{(\mathrm{A}_1)} \right.$$
$$\left. \overbrace{- E_{P^n}\left[\sum_{i=0}^{n-1} \log \frac{2i+1+M_n}{2i+1+M_i} \right]}^{(\mathrm{A}_2)} \right).$$

Note that (A_2) corresponds to the gain in redundancy of Q^n with respect to KT_{M_n+1}. It illustrates the benefit of taking M_i instead of M_n as cutoff to encode X_{i+1}.

On the one hand, we have

$$(\mathrm{A}_1) \leq \frac{E[M_n]}{2} \log n + E[\log(M_n+1)]. \tag{2.17}$$

Since $E[\log M_n] \leq E[M_n]$, Lemma 4 entails

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E[\log(M_n+1)] = o(\log^2 n).$$

Applying Lemma 4 again, we see that (A_1) produces a redundancy equivalent to $\frac{1}{2\alpha \log e} \log^2 n$, which is twice bigger than the minimax redundancy obtained in Theorem 2. So, we will hope the corrective term (A_2) to be about $\frac{1}{4\alpha \log e} \log^2 n$.
To deal with (A_2), we use the concavity of the log function, and we group the terms in the sum, M_n by M_n. Let $m = \left\lfloor \frac{n-1}{M_n} \right\rfloor$ be the number of bundles.
To simplify the expression, we also neglect few terms at the beginning of the sum. Let $(h_n)_{n\geq 1}$ be a non-decreasing sequence of positive integers, such that $h_n \to \infty$ as $n \to \infty$, and let us define $\lambda_n = 2h_n \log\left(1 + \frac{1}{2h_n}\right)$. Then

$$(\mathrm{A}_2) \geq \lambda_n E_{P^n}\left[\sum_{k=h_n+1}^{m} \frac{M_n - M_{kM_n}}{2(k+1)} \right]. \tag{2.18}$$

It is easy to check that the function $x \mapsto x \log\left(1 + \frac{1}{x}\right)$ is non-decreasing, and tends to

$\log e$ when x tends to the infinity; therefore (λ_n) tends to $\log e$. We can write now

$$\begin{aligned}(\mathrm{A}) &\leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \left(\frac{E[M_n]}{2}\log n \right. \\ &\qquad \left. - \lambda_n E\left[\sum_{k=h_n+1}^{m} \frac{M_n - M_{kM_n}}{2(k+1)}\right]\right) + o(\log^2 n) \\ &\leq \overbrace{\frac{1}{2}\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}\left[M_n \log n - \lambda_n M_n \sum_{k=h_n+1}^{m} \frac{1}{k+1}\right]}^{(\mathrm{A}_3)} \\ &\qquad + \frac{\lambda_n}{2}\overbrace{\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}\left[\sum_{k=h_n+1}^{m} \frac{M_{kM_n}}{k+1}\right]}^{(\mathrm{A}_4)} + o(\log^2 n)\end{aligned}$$

Let us choose $h_n = \max\{1, \lfloor \ln n - 2 \rfloor\}$. Then

$$(\mathrm{A}_3) = o(\log^2 n) \tag{2.19}$$

$$(\mathrm{A}_4) \leq \frac{\log^2 n}{2\alpha \log^2 e} + o(\log^2 n). \tag{2.20}$$

Therefore we have

$$(\mathrm{A}) \leq (1 + o(1))\frac{1}{4\alpha \log e}\log^2 n$$

which concludes the proof of Proposition 7. □

Proof of (2.17). Let

$$\widehat{P}^n(x_{1:n}) = \sup_{P^n} P^n(x_{1:n}) = \prod_{j\in\{x_1,\ldots,x_n\}} \left(\frac{n_j^n}{n}\right)^{n_j^n}$$

be the maximum likelihood of the string $x_{1:n}$ over all iid distribution on $\mathbb{N}^n$. Then

$$\begin{aligned}(\mathrm{A}_1) &\leq E_{P^n}\left[\log \frac{\widehat{P}^n(X_{1:n})}{KT_{M_n+1}(X_{1:n})}\right] \\ &\leq E_{P^n}\left[\sup_{x_{1:n}\in\{1,\ldots,M_n+1\}} \log \frac{\widehat{P}^n(x_{1:n})}{KT_{M_n+1}(x_{1:n})}\right]\end{aligned}$$

Now we can apply a result from Catoni [21, prop 1.4.1]:

Lemma 5. *For all $k \geq 1$ and for all $x_{1:n} \in \{1,\ldots,k\}^n$,*

$$-\log KT_k(x_{1:n}) + \log \widehat{P}^n(x_{1:n}) \leq \frac{k-1}{2}\log n + \log k.$$

□

Proof of (2.18). We group the terms in (A_2), M_n by M_n:

$$(\mathrm{A}_2) \geq E_{P^n}\left[\sum_{k=1}^{m-1}\sum_{i=kM_n+1}^{(k+1)M_n} \log\left(1+\frac{M_n-M_i}{2i+M_i+1}\right)\right].$$

From the relation $M_k \leq M_{k'}$ for all $k' \geq k \geq 1$, we can infer, for all $i \geq kM_n$,

$$\frac{M_n-M_i}{2i+M_i+1} \leq \frac{M_n}{2kM_n} = \frac{1}{2k}$$

Since log is a concave function, we have $\log(1+x) \geq \frac{x\log(1+a)}{a}$ for all $a>0$ and $0 \leq x \leq a$. Consequently, if we choose $a = \frac{1}{2k}$,

$$\begin{aligned}(\mathrm{A}_2) &\geq E_{P^n}\left[\sum_{k=1}^{m-1}\sum_{i=kM_n+1}^{(k+1)M_n} 2k\log\left(1+\frac{1}{2k}\right)\frac{M_n-M_i}{2i+M_i+1}\right]\\ &\geq E_{P^n}\left[\sum_{k=h_n+1}^{m}\lambda_n\frac{M_n-M_{kM_n}}{2k+2}\right].\end{aligned}$$

□

Proof of (2.19). We have

$$\begin{aligned}(\mathrm{A}_3) = \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}}\Bigg[&\sum_{j\geq 1} P^n(M_n=j)\\ &\times\left(j\log n - \lambda_n j\sum_{k=h_n+1}^{\lfloor\frac{n-1}{j}\rfloor}\frac{1}{k+1}\right)\Bigg].\end{aligned}$$

Then we plug in $h_n = \lfloor \ln n - 2\rfloor$. For n large enough, $h_n \geq 1$, and we have

$$\begin{aligned}j\sum_{k=h_n+1}^{\lfloor\frac{n-1}{j}\rfloor}\frac{1}{k+1} &\geq j\int_{\ln n-1}^{\lfloor\frac{n-1}{j}\rfloor+1}\frac{\mathrm{d}x}{x+1}\\ &= j\left(\ln\left(\left\lfloor\frac{n-1}{j}\right\rfloor+2\right)-\ln(\ln n)\right)\\ &\geq j\ln(n-1)-j\ln j-j\ln(\ln n),\end{aligned}$$

and therefore

$$\begin{aligned}(\mathrm{A}_3) \leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}}\big[&(\log e-\lambda_n)E[M_n]\ln n\\ &+\lambda_n E[M_n]\ln\frac{n}{n-1}\\ &+\lambda_n E[M_n\ln M_n]+\lambda_n E[M_n]\ln(\ln n)\big].\end{aligned}$$

Then, if we use Lemma 4 and the fact that λ_n tends to $\log e$, we get (2.19). □

Proof of (2.20). We want to commute the expected value and the sum in (A_4). To do it, we need to get rid of m. We can note that the condition $k \leq m = \lfloor \frac{n-1}{M_n} \rfloor$ entails $kM_n \leq n-1$. Consequently, for n big enough,

$$\begin{aligned}(A_4) &\leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}\left[\sum_{k=3}^{m} \frac{M_{kM_n}}{k+1}\right] \\ &\leq \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}\left[\sum_{k=3}^{n-1} \frac{M_{kM_n}\mathbb{1}_{kM_n\leq n-1}}{k+1}\right] \\ &\leq \sum_{k=3}^{n-1} \frac{\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}[M_{kM_n}\mathbb{1}_{M_n\leq l_n}]}{k+1} \\ &\quad + \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}[M_n\mathbb{1}_{M_n>l_n}]\sum_{k=3}^{n-1}\frac{1}{k+1},\end{aligned}$$

where $l_n = \left\lfloor \frac{1}{\alpha}\left(2\ln n + \ln \frac{C}{1-e^{-\alpha}}\right)\right\rfloor$. We can now plug in the results of Lemma 4:

$$\begin{aligned}(A_4) &\leq \sum_{k=3}^{n-1} \frac{\ln(kl_n) + 1 + \ln\frac{C}{1-e^{-\alpha}}}{(k+1)\alpha} + o(1)\sum_{k=3}^{n-1}\frac{1}{k+1} \\ &\leq \frac{1}{\alpha}\sum_{k=3}^{n-1}\frac{\ln k}{k+1} + \frac{1}{\alpha}(\ln l_n + O(1))\sum_{k=3}^{n-1}\frac{1}{k+1}.\end{aligned}$$

Note that $l_n = O(\ln n)$, and consequently $\ln l_n = O(\ln\ln n)$. So

$$\begin{aligned}(A_4) &\leq \frac{1}{\alpha}\int_3^n \frac{\ln x}{x}\,\mathrm{d}x + O(\ln\ln n)\int_3^n \frac{\mathrm{d}x}{x} \\ &\leq \frac{1}{2\alpha}\ln^2 n + o(\ln^2 n).\end{aligned}$$

□

2.C.3 Contribution of C2

Proposition 8. *Let C and α be positive numbers satisfying $C > e^{2\alpha}$. Then*

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}[l(\texttt{C2})] \leq o(\log^2 n).$$

Proof. Like in the previous subsection, we give first the sketch of the proof, and we delay several technical lemmas.

$$\begin{aligned}&\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}[l(\texttt{C2})] \\ &\qquad\qquad \leq 1 + \sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \sum_{i=1}^{n} E_{P^n}\left[\mathbb{1}_{X_i>M_{i-1}} l_E(X_i+1)\right].\end{aligned}$$

We deal with this sum thanks to the following lemma:

Lemma 6. *Let g be a positive and non-decreasing function on $[1,\infty)$. Let $(K_n)_{n\geq 1}$ be a non-decreasing sequence of positive integers. Then, for all $n \geq 1$,*

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} \sum_{i=1}^{n} E_{P^n}\left[\mathbb{1}_{X_i>M_{i-1}} g(X_i)\right] \leq (1+o(1))g(K_n)\frac{1}{\alpha}\ln n + C\,n\int_{K_n}^{\infty} g(x+1)e^{-\alpha x}\,\mathrm{d}x.$$

To apply Lemma 6, we extend the definition of l_E on $[1,\infty)$ by

$$l_E(x) = \begin{cases} 1 & \text{if } x \in [1,2), \\ 1 + \lfloor \log x \rfloor + 2\lfloor \log \lfloor \log x \rfloor + 1 \rfloor & \text{if } x \geq 2. \end{cases}$$

We get

$$\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}} E_{P^n}[l(\mathtt{C2})] \leq (1+o(1))\frac{l_E(K_n+1)}{\alpha}\ln n + C\,n\int_{K_n}^{\infty} l_E(x+2)e^{-\alpha x}\,\mathrm{d}x$$

Then we can choose $K_n = \max\{1, \lfloor \frac{1}{\alpha}\ln n \rfloor\}$. This entails

$$l_E(K_n+1) \underset{n\to\infty}{\sim} \log K_n \sim \log\log n = o(\log n),$$

and therefore

$$\frac{1}{\alpha} l_E(K_n+1)\ln n = o(\log^2 n).$$

The remaining term is treated by Lemma 7, which achieves the proof of Proposition 8:

Lemma 7. *Let $\alpha > 0$ be a real number, and let $K_n = \max\{1, \lfloor \frac{1}{\alpha}\ln n \rfloor\}$. Then*

$$n\int_{K_n}^{\infty} l_E(x+2)e^{-\alpha x}\,\mathrm{d}x = o(\log n).$$

□

Proof of Lemma 6. Let $\boldsymbol{P}$ be an element of $\Lambda_{Ce^{-\alpha\cdot}}$. Let us define, for all $k \geq 0$,

$$\bar{p}(k) = P(X_1 > k) = \sum_{j\geq k+1} P(k),$$

and

$$(\mathrm{B}_1) = \sum_{i=1}^{n} E_{P^n}\left[\mathbb{1}_{X_i>M_{i-1}} g(X_i)\right].$$

Note that, for all $1 \leq i \leq n$, X_i and M_{i-1} are independent random variables, and

$$\begin{aligned} P^n(M_i \leq k) &= P^n(\forall\ 1 \leq j \leq i,\ X_j \leq k) \\ &= (1-\bar{p}(k))^i. \end{aligned}$$

Then we can write

$$\begin{aligned}(\mathrm{B}_1) &= \sum_{i=1}^{n}\sum_{k\geq 0} P^n(M_{i-1}=k)\sum_{m\geq k+1} P(m)g(m)\\ &= \sum_{m\geq 1} P(m)g(m)\sum_{i=1}^{n}\sum_{k=0}^{m-1} P(M_{i-1}=k)\\ &= P(1)g(1)+\sum_{m\geq 2} P(m)g(m)\sum_{i=1}^{n}(1-\bar{p}(m-1))^{i-1}\\ &= \sum_{m\geq 1} P(m)g(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)}.\end{aligned}$$

If we take $g(x)=1$ for all x, we get

$$\begin{aligned}\sum_{m\geq 1} P(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)} &= E\left[\sum_{i=1}^{n}\mathbb{1}_{X_i>M_{i-1}}\right]\\ &\leq E[M_n].\end{aligned}$$

In the general case, we can split the sum at K_n, and we get

$$\begin{aligned}(\mathrm{B}_1) &= \sum_{m=1}^{K_n} P(m)g(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)}\\ &\quad+\sum_{m\geq K_n+1} P(m)g(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)}\\ &\leq g(K_n)\sum_{m\geq 1} P(m)\frac{1-(1-\bar{p}(m-1))^n}{\bar{p}(m-1)}\\ &\quad+\sum_{m\geq K_n+1} nP(m)g(m)\\ &\leq g(K_n)E[M_n]+C\,n\sum_{m\geq K_n+1} g(m)e^{-\alpha m}.\end{aligned}$$

At this point, we can take the supremum over all sources $\boldsymbol{P}$ in $\Lambda_{Ce^{-\alpha\cdot}}$:

$$\begin{aligned}&\sup_{\boldsymbol{P}\in\Lambda_{Ce^{-\alpha\cdot}}}\sum_{i=1}^{n} E_{P^n}\left[\mathbb{1}_{X_i>M_{i-1}}g(X_i)\right]\\ &\quad\leq (1+o(1))g(K_n)\frac{1}{\alpha}\ln n + C\,n\int_{K_n}^{\infty} g(x+1)e^{-\alpha x}\,\mathrm{d}x.\end{aligned}$$

□

Proof of Lemma 7.

$$
\begin{aligned}
& n \int_{K_n}^{\infty} l_E(x+2) e^{-\alpha x} \, \mathrm{d}x \\
& \quad \leq n \int_{K_n}^{\infty} \left(\log(x+2) + 2 \log\log(x+3) + 1 \right) e^{-\alpha x} \, \mathrm{d}x \\
& \quad \leq n e^{-\alpha K_n} \log(K_n + 3) \\
& \qquad \int_{K_n+3}^{\infty} \frac{\log x + 2 \log\log x + 1}{\log(K_n+3)} e^{-\alpha(x - K_n - 3)} \, \mathrm{d}x \\
& \quad \leq e^{\alpha} \log(K_n+3) \left(\sup_{x \geq K_n+3} \frac{\log x + 2\log\log x + 1}{\log x} \right) \\
& \qquad \int_{K_n+3}^{\infty} \frac{\log x}{\log(K_n+3)} e^{-\alpha(x-K_n-3)} \, \mathrm{d}x \\
& \quad = O(\log K_n) \int_0^{\infty} \left(1 + \frac{\log\left(1 + \frac{x}{K_n+3}\right)}{\log(K_n+3)} \right) e^{-\alpha x} \, \mathrm{d}x \\
& \quad = o(\log n).
\end{aligned}
$$

The supremum is correctly defined and bounded, because the function

$$x \mapsto \frac{\log x + 2\log\log x + 1}{\log x}$$

is continuous and tends to 1 as x tends to the infinity. □

2.C.4 Proof of Theorem 5

The message sent by the `ACcode` algorithm is compound of two strings `C1` and `C2`. `C1` corresponds to the part of the message encoded by the arithmetic code, with coding probability Q^{n+1}. The arithmetic code encodes a message $\widetilde{x}_{1:n}0$ with $\lceil -\log Q^{n+1}(\widetilde{x}_{1:n}0) \rceil + 1$ bits. We have

$$
\begin{aligned}
E_{P^n}[-\log Q_{n+1}(0|X_{1:n})] &= E_{P^n}[\log(M_n + 1 + 2n)] \\
&\leq \log(2n) + \frac{E_{P^n}[M_n + 1]}{2n} \\
&= O(\log n)
\end{aligned}
$$

thanks to Lemma 4. Therefore the redundancy of `ACcode` can be upper bounded, for all $n \geq 2$, by

$$
\begin{aligned}
& \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_{P^n}[l(\mathtt{C1}) + l(\mathtt{C1})] - H(P^n)] \\
& \qquad \leq \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_{P^n}[-\log Q^n(\widetilde{X}_{1:n}) - H(P^n)] \\
& \qquad\quad + \sup_{\boldsymbol{P} \in \Lambda_{Ce^{-\alpha \cdot}}} E_{P^n}[l(\mathtt{C2})] + O(\log n).
\end{aligned}
$$

We conclude thanks to Propositions 7 and 8.

2.D Simplification de ACcode

Le contenu du présent chapitre a fait l'objet d'un article [15] accepté à *IEEE Transactions on Information Theory*. Après sa soumission à cette revue, nous avons découvert une façon algorithmiquement très simplifiée de mettre en œuvre le principe du codage par censure. Cette simplification s'applique aussi bien à `ACcode` qu'à l'algorithme `CensuringCode` proposé par Boucheron, Garivier et Gassiat [16].

Dans les deux cas, le signal encodé est composé de deux parties. Un premier mot de code `C1` est le résultat du codage arithmétique d'une suite censurée $\widetilde{X}_{1:n}$, dans laquelle le symbole 0 indique la présence d'un symbole censuré. Le second mot de code `C2` est formé par la concaténation des codes Elias [41] des symboles censurés.

Nous proposons de ne former qu'un unique mot de code par codage arithmétique. L'idée de censure, qui nous fait traiter différemment les symboles petits et les symboles grands, sera directement transcrite dans la probabilité de codage. En particulier, la probabilité de codage des symboles non-censurés ne changerait pas ; la probabilité des symboles censurés serait le produit entre d'une part la probabilité de codage anciennement attribuée au symbole 0, et d'autre part une quantité qui imiterait la longueur du code Elias.

Pour fixer les choses, posons

$$C = \left(\sum_{k=1}^{\infty} \frac{1}{k \log^2(k+1)} \right)^{-1},$$

et considérons s la loi de probabilité sur $\mathbb{N}_*$ définie par

$$s(k) = \frac{C}{k \log^2(k+1)}.$$

Ainsi $-\log s(k) = \log k + 2 \log\log(k+1) - \log C$, ce qui correspond à la longueur du code Elias à un $O(1)$ près.
Dans notre algorithme simplifié, le codeur encode directement le message $X_{1:n}$ par codage arithmétique, en utilisant les probabilités conditionnelles

$$Q_{i+1}(X_{i+1} = j | X_{1:i} = x_{1:i}) = \begin{cases} \dfrac{n_i^j + 1/2}{i + \frac{m_i+1}{2}} & \text{si} \quad 1 \le j \le m_i, \\ \dfrac{1/2}{i + \frac{m_i+1}{2}} s(j - m_i) & \text{sinon.} \end{cases}$$

L'analyse de la longueur de code qui a été faite pour `ACcode` reste valide — le calcul en est même simplifié. On vérifie d'ailleurs facilement que la longueur moyenne du nouveau code est celle de `ACcode` à un $O(\log n)$ près. Remarquons au passage que le choix particulier que nous avons fait pour s pourra éventuellement être amélioré (nous avons choisi cette formule pour sa simplicité). Une alternative pourrait simplement être de définir s explicitement à partir de la longueur du code Elias (2.13).

Ainsi ce nouvel algorithme n'apporte que des avantages : une longueur de code comparable et une implémentation grandement simplifiée.

Chapitre 3

Le théorème de Bernstein-von Mises pour la régression gaussienne sous un nombre croissant de régresseurs

Bernstein-von Mises Theorems for Gaussian regression with increasing number of regressors

Abstract

This chapter brings a contribution to the Bayesian theory of nonparametric and semiparametric estimation. We are interested in the asymptotic normality of the posterior distribution in Gaussian linear regression models when the number of regressors increases with the sample size. Two kinds of Bernstein-von Mises Theorems are obtained in this framework: nonparametric theorems for the parameter itself, and semiparametric theorems for functionals of the parameter. We apply them to the Gaussian sequence model and to the regression of functions in Sobolev and C^α classes, in which we get the minimax convergence rates. Adaptivity is reached for the Bayesian estimators of functionals in our applications.

Keywords: Nonparametric Bayesian Statistics, Semiparametric Bayesian Statistics, Bernstein-von Mises Theorem, posterior asymptotic normality, adaptive estimation.

Sommaire

3.1 Introduction

To estimate a parameter of interest in a statistical model, a Bayesian puts a prior distribution on it and looks at the posterior distribution, given the observations. A Bernstein-von Mises Theorem is a result stating that under adequate conditions the posterior distribution is asymptotically normal, centered at the maximum likelihood estimator (MLE) of the model used, with a variance equal to the asymptotic frequentist variance of the MLE.

Such an asymptotic posterior normality is important because it allows to construct approximate credible regions, based on the posterior distribution, which keep good frequentist properties. In particular it is difficult to build frequentist confidence regions in complex models, while the Monte-Carlo Markov chain algorithms (MCMC) make more feasible the construction of Bayesian confidence regions — however Bernstein-von Mises Theorems are difficult to derive in complex models.

For parametric models, the Bernstein-von Mises Theorem is a well-known result, for which we refer to [106]. In nonparametric models (where the parameter space is infinite-dimensional or growing), and semiparametric models (when the parameter of interest is a finite-dimensional functional of the complete infinite-dimensional parameter), there are still relatively few asymptotic normality results. [45] gives negative results, and we recall some positive ones below. However many recent papers deal with the convergence rate of posterior distributions in various settings, which is linked with the model complexity: we refer to [55, 100] as early representatives of this school.
Nonparametric Bernstein-von Mises Theorems have been developed for models based on a sieve approximation, where the dimension of the parameter grows with the sample size. In particular two situations have been studied: regression models in [52]; exponential models in [53], [25], and [17] (this last one deals with the discrete case, when the observations follow some unknown infinite multinomial distribution).
In semiparametric frameworks the asymptotic normality has been obtained in several situations. [69] and [68] study the nonparametric right-censoring model and the proportional hazard model. [19] obtains Bernstein-von Mises Theorems for Gaussian process priors, in the semiparametric framework where the unknown quantity is (θ, f), with θ the parameter of interest and f an infinite-dimensional nuisance parameter. It clarifies a preceding paper [99], which considers also the more general framework where the quantity of interest is a finite-dimensional function $g(f)$ of the infinite-dimensional parameter f of the model. [87] obtains the Bernstein-von Mises Theorem for linear functionals of the density of the observations, in the context of a sieve approximation; they achieve also the frequentist minimax estimation rate for densities in specific regularity classes with a deterministic (non-adaptive) value of the cutoff k_n.

In the current paper we obtain nonparametric and semiparametric Bernstein-von Mises Theorems in a Gaussian regression framework with an increasing number of regressors.
Our nonparametric results cover the case of a specific Gaussian prior, and the case of more generic smooth priors. They are said nonparametric because we use sieve priors and the dimension of the parameter grows. These results improve on the preceding ones by [52] which did not suppose the normality of the errors but imposed other conditions, in particular on the growth rate of the number of regressors. We apply them to the pe-

riodic Sobolev classes and to regularity classes $C^\alpha[0,1]$ in the context of the regression model (using respectively trigonometric polynomials and splines as regressors), as well as to the Gaussian sequence model. In all these situations we get the asymptotic normality of the posterior in addition to the minimax convergence rates, with appropriate (non-adaptive) choices of the prior. We also show that for some priors known to reach this convergence rate, the Bernstein-von Mises Theorem does not hold.
We derive also semiparametric Bernstein-von Mises Theorems for linear and nonlinear functionals of the parameter. The linear case is an immediate corollary of the nonparametric theorems and do not need any additional condition. We apply these results to the periodic Sobolev classes to estimate a linear functional and the L^2 norm of the regression function f if enough smoothness is present, and in both cases we are able to build an adaptive Bayesian estimator which achieves the minimax convergence rate whatever the unknown parameter of the class is, in addition to the asymptotic normality.

The paper is organized as follows. We present the framework in section 3.2. Section 3.3 states the nonparametric Bernstein-von Mises Theorems, for Gaussian or non-Gaussian priors. In section 3.4 we expound the semiparametric Bernstein-von Mises Theorems for linear and non-linear functionals of the parameter. Then we consider in section 3.5 applications to the Gaussian sequence model, and to the regression of a function in a Sobolev and $C^\alpha[0,1]$ class. In section 3.6 the nonparametric and semiparametric Bernstein-von Mises Theorems are proved. Eventually the Appendix contains various technical tools used in the main analysis.

3.2 Framework

We consider a Gaussian linear regression framework. For any $n \geq 1$, our observation $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ is a Gaussian random vector

$$Y = F + \varepsilon \tag{3.1}$$

where the vector of errors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \sim \mathcal{N}(0, \sigma_n^2 I_n)$ is centered normal and the mean vector F belongs to $\mathbb{R}^n$. The observations Y_i and the variance σ_n^2 of the errors may depend on n, but σ_n^2 is known. Fix a (sequence of) mean vector(s) F_0. We denote by P_{F_0} the probability distribution of a random variable following $\mathcal{N}(F_0, \sigma_n^2 I_n)$, and E the associated expectation.

Let $\phi_1, \ldots, \phi_{k_n}$ a collection of k_n linearly independent regressors in $\mathbb{R}^n$, where $k_n \leq n$ grows with n. We gather these regressors in the $n \times k_n$-matrix Φ of rank k_n, and we denote $\langle \phi \rangle$ their linear span. $\langle \phi \rangle$ is the misspecified model in which the Bernstein-von Mises Theorems will be stated. It can be parametrized as $\langle \phi \rangle = \left\{ \Phi\theta : \theta = (\theta_1, \ldots, \theta_{k_n}) \in \mathbb{R}^{k_n} \right\}$. We denote by P_θ the probability distribution of a random variable following $\mathcal{N}(\Phi\theta, \sigma_n^2 I_n)$, and E_θ the associated expectation.

As examples, we present three different frameworks, each one with its own collection of regressors. In section 3.5 the Bernstein-von Mises Theorems are applied to each one of these frameworks.

1. **The Gaussian sequence model.**
   Our first application concerns the Gaussian sequence model, which is also equivalent to the white noise model (see [75, ch. 4] for instance). We consider the infinite

dimensional setting

$$Y_j = \theta_j^0 + \frac{1}{\sqrt{n}}\xi_j, \qquad j \geq 1 \tag{3.2}$$

where the random variables $\xi_j, j \geq 1$ are independant and have distribution $\mathcal{N}(0,1)$. Projecting on the first k_n coordinates with $k_n \leq n$, we retrieve our model (3.1) with $\theta_0 = (\theta_j^0)_{1\leq j\leq k_n}$, $\sigma_n = 1/\sqrt{n}$, and $\Phi^T\Phi = I_{k_n}$.

2. **Regression of a function in a Sobolev class.**
Let $f : [0,1] \to \mathbb{R}$ be a function in $\mathbb{L}^2([0,1])$. We observe realizations of random variables

$$Y_i = f(i/n) + \varepsilon_i \tag{3.3}$$

for $1 \leq i \leq n$, where the errors ε_i are iid $\mathcal{N}(0,\sigma_n^2)$ and σ_n does not depend on n. We denote by $(\varphi_j)_{j\geq 1}$ the Fourier basis

$$\begin{aligned} \varphi_1 &\equiv 1 \\ \varphi_{2m}(x) &= \sqrt{2}\,\cos(2\pi m x) \qquad \forall m \geq 1 \\ \varphi_{2m+1}(x) &= \sqrt{2}\,\sin(2\pi m x) \qquad \forall m \geq 1 \end{aligned} \tag{3.4}$$

For the regression on Fourier's basis we choose a regular design $x_i = i/n$ for $1 \leq i \leq n$. This gives the collection of regressors $\phi_j = (\varphi_j(i/n))_{1\leq i\leq n}$, $1 \leq j \leq k_n$. In practice we suppose that f belongs to one of the Sobolev classes:

Definition 4. *Let $\alpha > 0$ and $L > 0$. Let $(\varphi_j)_{j\geq 1}$ denote the Fourier basis (3.4). We define the Sobolev class $\mathcal{W}(\alpha, L)$ as the collection of all functions $f = \sum_{j=1}^{\infty}\theta_j\varphi_j$ in $\mathbb{L}^2([0,1])$ such that $\theta = (\theta_j)_{j\geq 1}$ is an element of the ellipsoid of $\ell^2(\mathbb{N})$*

$$\Theta(\alpha, L) = \left\{\theta \in \ell^2(\mathbb{N}) : \sum_{j=1}^{\infty} a_j^2\theta_j^2 \leq \frac{L^2}{\pi^{2\alpha}}\right\}$$

where

$$a_j = \begin{cases} j^\alpha & \text{if } j \text{ is even;} \\ (j-1)^\alpha & \text{if } j \text{ is odd.} \end{cases} \tag{3.5}$$

3. **Regression of a function in $C^\alpha[0,1]$.**
Fix a regularity $\alpha > 0$, and consider a function $f \in C^\alpha[0,1]$. This means that f is α_0 times continuously differentiable with $\|f\|_\alpha < \infty$, α_0 being the greatest integer less than α and the seminorm being defined by

$$\|f\|_\alpha = \sup_{x\neq x'} \frac{\left|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(x')\right|}{|x - x'|^{\alpha-\alpha_0}}.$$

Consider a design $\left(x_i^{(n)}\right)_{n\geq 1, 1\leq i\leq n}$, not necessarily uniform. Here F_0 is the vector $\left(f\left(x_i^{(n)}\right)\right)_{1\leq i\leq n}$. Once again we suppose that $\sigma_n = \sigma$ does not depend on n.

Fix an integer $q \geq \alpha$, and let $K = k_n + 1 - q$. Partition the interval $(0,\ 1]$ into K subintervals $((j-1)/K,\ j/K]$ for $1 \leq j \leq K$. We want to perform the regression of f in the space of splines of order q defined on that partition, and use the B-splines basis $(B_j)_{1\leq j\leq k_n}$ (see [34] for instance). Our collection of regressors is $\phi_j = \left(B_j\left(x_i^{(n)}\right)\right)_{1\leq i\leq n}$, for $1 \leq j \leq k_n$.

For any value of $n \geq 1$, let W be a prior distribution on F, with support included in $\langle \phi \rangle$. Equivalently, W is induced by a probability distribution $\widetilde{W}$ on θ by the application $\theta \mapsto \Phi\theta$. P^W denotes the marginal distribution of Y under prior W, and $W(\mathrm{d}G(F)|Y)$ denotes the posterior distribution of a functional $G(F)$. Note that everything depends on n — W for instance is a distribution on $\mathbb{R}^n$ — even if we do not use n as index to simplify our notations.
W is a sieve prior. Such priors are specially well adapted for increasing dimension frameworks; they also make clear the relations between the parametric and nonparametric results. On the other hand the question of the choice of the cutoff k_n arises.
The exact parametrization by θ and the corresponding collection of regressors $\phi_1, \ldots, \phi_{k_n}$ are somehow arbitrary: what matters is the posterior distribution of F and this depends on $\langle \phi \rangle$, which is characterized by the matrix $\Sigma = \Phi(\Phi^T\Phi)^{-1}\Phi^T$ of the orthogonal projection onto $\langle \phi \rangle$. In practice it is difficult to dissociate $\langle \phi \rangle$ and the collection $\phi_1, \ldots, \phi_{k_n}$, but we have chosen to emphasize W and F over $\widetilde{W}$ and θ.

In the model $\langle \phi \rangle$, the MLE of F_0 is the orthogonal projection $Y_{\langle \phi \rangle}$ of Y; so $Y_{\langle \phi \rangle} = \Sigma Y$. We set $\theta_Y = (\Phi^T\Phi)^{-1}\Phi^T Y$ its associated parameter. Let also $F_{\langle \phi \rangle} = \Phi\theta_0$ be the projection of F_0 on $\langle \phi \rangle$, with $\theta_0 = (\Phi^T\Phi)^{-1}\Phi^T F_0$. Even if $\langle \phi \rangle$ contains the support of the prior distribution W, we do not suppose that F_0 belongs to $\langle \phi \rangle$, and this improves on some previous results. θ_0 has not to be seen as some "true" parameter.

Although the MLE is naturally defined *in the sieve* $\langle \phi \rangle$, it heavily depends on the choice of $\langle \phi \rangle$. Therefore the Bernstein-von Mises Theorems we establish depends on the choice of the sieve the prior distribution is built on. This is true in particular in a maybe more veiled way for our semiparametric results, in which the centering point is a plug-in estimator based on the MLE defined on $\langle \phi \rangle$. In nonparametric models constructed on an infinite dimensional parameter, there is no definition of a MLE; what should be the natural centering for a Bernstein-von Mises Theorem in such situations is not clear.

To conclude this section, the following immediate frequentist result gives the distribution of $Y_{\langle \phi \rangle}$ under P_{F_0}:

$$Y_{\langle \phi \rangle} \sim \mathcal{N}\left(F_{\langle \phi \rangle}, \sigma_n^2 \Sigma\right).$$

3.3 Nonparametric Bernstein-von Mises Theorems

The proofs of our nonparametric results are delayed to section 3.6.

3.3.1 With Gaussian priors

We consider here a centered, normal prior distribution W which is isotropic on $\langle \phi \rangle$, so that $W = \mathcal{N}(0, \tau_n^2 \Sigma)$ for some sequence τ_n. Essentially the only assumption needed in this case is that the prior becomes flat enough as n grows. $\|Q - Q'\|_{\mathrm{TV}}$ denotes the total variation norm between two probability distributions Q and Q'.

Theorem 6. *Assume that* $\sigma_n = o(\tau_n)$, $\|F_0\| = o(\tau_n^2/\sigma_n)$ *and* $k_n = o(\tau_n^4/\sigma_n^4)$. *Then*

$$E \left\| W(\mathrm{d}F|Y) - \mathcal{N}\left(Y_{\langle \phi \rangle}, \sigma_n^2 \Sigma\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

Since the support of W is included in $\langle\phi\rangle$, we can equivalently state

$$E\left\|\widetilde{W}(\mathrm{d}\theta|Y) - \mathcal{N}\left(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

Theorem 6 does not deal with the modeling bias introduced by taking a prior restricted to $\langle\phi\rangle$. This is an important question in nonparametric statistics, and k_n has to be chosen in order to achieve the bias-variance tradeoff. In most cases this bias has already been studied in frequentist papers on sieve approximation.

As an example, let us consider an usual regression framework with $F_0 = (f(x_i))_{1\leq i\leq n}$, where f is some function and $(x_i)_{1\leq i\leq n}$ some design. If σ_n does not depend on n, both conditions $\|F_0\| = o(\tau_n^2/\sigma_n)$ and $k_n = o(\tau_n^4/\sigma_n^4)$ are verified for instance if f is bounded and $n^{1/4} = o(\tau_n)$. These conditions can be read in the other way: τ_n^4 must be large enough with respect to $\|F_0\|$ and k_n.

3.3.2 With smooth priors

We consider now more general priors. We get an abstract result, but with powerful applications.

Theorem 7. *Suppose that W is induced by a distribution on θ admitting a density $w(\theta)$ with respect to Lebesgue measure. If there exists a sequence $(M_n)_{n\geq 1}$ such that*

1. $\sup_{\|\Phi h\|^2\leq\sigma_n^2 M_n, \|\Phi g\|^2\leq\sigma_n^2 M_n} \frac{w(\theta_0+h)}{w(\theta_0+g)} \to 1$ *as* $n\to\infty$.
2. $k_n \ln k_n = o(M_n)$
3. $\max\left(0, \ln\left(\frac{\sqrt{\det(\Phi^T\Phi)}}{\sigma_n^{k_n} w(\theta_0)}\right)\right) = o(M_n)$

Then

$$E\left\|W(\mathrm{d}F|Y) - \mathcal{N}\left(Y_{\langle\phi\rangle}, \sigma_n^2\Sigma\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } n\to\infty.$$

Since the support of W is included in $\langle\phi\rangle$, we can equivalently state

$$E\left\|\widetilde{W}(\mathrm{d}\theta|Y) - \mathcal{N}\left(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } n\to\infty.$$

With Condition 1 we ask for a sufficiently flat prior in a given neighborhood of θ_0. By Conditions 2 and 3 we insure that this neighborhood has enough prior weight. This kind of assumptions is quite common in the literature dealing with the concentration of posterior distributions. These assumptions are needed together in order to get the Gaussian shape of the posterior distribution. Several of our applications illustrate that priors known to induce the posterior minimax convergence rate may not be flat enough to get the Gaussian shape with the asymptotic variance $\sigma_n^2\Sigma$.

Our main applications, to the Gaussian sequence model, and to the regression model using trigonometric polynomials and splines, are developed in section 3.5. We now present two remarks about the parametric case and the comparison with the pioneer work of Ghosal [52].

The parametric case. Consider the regression of a function f defined on $[0,1]$, with a fixed number k of regressors. Set a design $(x_i^{(n)})_{n\geq 1, 1\leq i\leq n}$, with $x_i^{(n)} \in [(i-1)/n, i/n]$ for any $n \geq 1$, and $F_0 = \left(f(x_i^{(n)})\right)_{1\leq i\leq n}$. Choose a finite number of piecewise continuous and linearly independent regressors $(\varphi_j)_{1\leq j\leq k}$ on $[0,1]$, and set $\phi_j = \left(\varphi_j\left(x_i^{(n)}\right)\right)_{1\leq i\leq n}$ for $1 \leq j \leq k$. f, $k_n = k$, $\sigma_n = \sigma$, and W do not depend on n.
We would like to compare Theorem 7 with the usual Bernstein-von Mises Theorem for parametric models, applied to such a regression framework. In that setting, let us suppose that w is continuous and positive, and that f is bounded. Then Condition 1 becomes $M_n = o(n)$, while Condition 3 reduces to $\ln n = o(M_n)$. Clearly, there exist such sequences $(M_n)_{n\geq 1}$, and Theorem 7 applies.
The rescaling by $\sqrt{n}$ of the Bernstein-von Mises Theorem for parametric models is here hidden in the asymptotic posterior variance $\sigma^2(\Phi^T\Phi)^{-1}$ of the parameter θ. Indeed, $(1/n)\ \Phi^T\Phi$ is a Riemann sum, and converges towards the Gramian matrix of the collection $(\varphi_j)_{1\leq j\leq k}$ in $\mathbb{L}^2([0,1])$.

Proof. We have $\|\Phi\theta_0\| \leq \|F_0\| \leq \sqrt{n}\|f\|_\infty$, and $\|\theta_0\|^2 \leq \left\|(\Phi^T\Phi)^{-1}\right\| \|\Phi\theta_0\|^2 \leq \left\|n(\Phi^T\Phi)^{-1}\right\| \|f\|_\infty^2$. $(1/n)\ \Phi^T\Phi$ converges towards the Gramian matrix of the collection $(\varphi_j)_{1\leq j\leq k}$ in $\mathbb{L}^2([0,1])$, and its smallest eigenvalue is lower bounded for n large enough. Therefore θ_0 is bounded, and we can consider it lies in some compact set on which w is uniformly continuous and lower bounded by a positive constant. The rest follows. □

Comparison with Ghosal's conditions. The Bernstein-von Mises Theorem in a regression setting when the number of parameters goes to infinity has been first studied by Ghosal [52] as an early step in the development of frequentist nonparametric Bayesian theory. In his paper the errors ε_i are not supposed to be Gaussian. Under the Gaussianity assumption, we get improved results. In particular our condition for the prior smoothness is simpler, and the growth rate of the dimension k_n is much less constrained.

- [52] does not admit a modeling bias between F_0 and $\Phi\theta_0$. In the present work the normality of the errors permits to take $F_0 \neq \Phi\theta_0$ without any cost, as it appears in the core of the proof (Lemma 14).
- In [52] σ_n is constant, which does not allow the application to the Gaussian sequence model.
- At last, [52] restricts the growth of the dimension k_n to $k_n^4 \ln k_n = o(n)$ (see below). It is then not possible to obtain the applications to the Gaussian sequence model or to the regression model for Sobolev or C^α classes.

Let $\delta_n^2 = \|(\Phi^T\Phi)^{-1}\|$ be the operator norm of $(\Phi^T\Phi)^{-1}$ for the ℓ^2 metric, and let η_n^2 be the maximal value on the diagonal of Σ. With our notations, the remaining assumptions of [52] become

(A3) There exists $\eta_0 > 0$ such that $w(\theta_0) > \eta_0^{k_n}$. Moreover

$$|\ln w(\theta) - \ln w(\theta_0)| \leq L_n(C)\|\theta - \theta_0\|, \tag{3.6}$$

whenever $\|\theta - \theta_0\| \leq C\delta_n k_n \sqrt{\ln k_n}$, where the Lipschitz constant $L_n(C)$ is subject to some growth restriction (see assumption A4).

(A4)

$$\forall C > 0, L_n(C)\delta_n k_n \sqrt{\ln k_n} \to 0 \quad \text{and} \quad \eta_n k_n^{3/2} \sqrt{\ln k_n} \to 0. \tag{3.7}$$

Further the design satisfies a condition on the trace of $\Phi^T\Phi$:

$$\mathrm{Tr}(\Phi^T\Phi) = O(nk_n). \tag{3.8}$$

Since Σ is an orthogonal projection matrix on a k_n-dimensional space, $\mathrm{Tr}(\Sigma) = k_n$ and $\eta_n^2 \geq k_n/n$. Consequently the last part of (3.7) entails $k_n^4 \ln k_n = o(n)$.

If we add the normality of the errors and a slight technical condition $\ln n = o(k_n \ln k_n)$, these assumptions entail ours. Indeed, set $M_n = C^2 k_n^2 \ln k_n$ for some arbitrary value of C. Our condition 2 is immediate. Condition 1 is got from (3.6) and the first part of (3.7). The beginning of (A3) entails $-\ln w(\theta_0) = O(k_n) = o(M_n)$. Using the concavity of the ln function and (3.8), we get $\ln\det(\Phi^T\Phi) \leq k_n \ln \mathrm{Tr}(\Phi^T\Phi) - k_n \ln k_n = O(k_n \ln n) = o(M_n)$. Therefore our condition 3 holds.

3.4 Semiparametric Bernstein-von Mises Theorems

We consider two kinds of functionals of F: linear and non-linear ones. These results can be easily adapted to functionals of θ, using the maps $\theta \mapsto \Phi\theta$ and $F \mapsto (\Phi^T\Phi)^{-1}\Phi^T F$.

3.4.1 The linear case

For linear functionals of F, we have the following corollary:

Corollary 1. *Let $p \geq 1$ fixed, and G be a $\mathbb{R}^p \times \mathbb{R}^n$-matrix. Suppose that the conditions of either Theorem 6 or Theorem 7 are verified. Then*

$$E \left\| W(\mathrm{d}(GF)|Y) - \mathcal{N}\left(GY_{\langle\phi\rangle}, \sigma_n^2 G\Sigma G^T\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

Further, the distribution of $GY_{\langle\phi\rangle}$ is $\mathcal{N}\left(GF_{\langle\phi\rangle}, \sigma_n^2 G\Sigma G^T\right)$.

Corollary 1 is just a linear transform of the preceding Theorems, and of the distribution of $Y_{\langle\phi\rangle}$.

An example of application is given in subsection 3.5.2, in the context of the regression on Fourier's basis.

3.4.2 The nonlinear case

Let $p \geq 1$ fixed, and $G : \mathbb{R}^n \mapsto \mathbb{R}^p$ be a twice continuously differentiable function. For $F \in \mathbb{R}^n$, let $\dot{G}_F$ denote the Jacobian matrix of G at F, and $D_F^2 G(\cdot,\cdot)$ the second derivative of G, as a bilinear function on $\mathbb{R}^n$. For any $F \in \langle\phi\rangle$ and $a > 0$, let

$$B_F(a) = \sup_{h \in \langle\phi\rangle : \|h\|^2 \leq \sigma_n^2 a} \sup_{0 \leq t \leq 1} \left\| D_{F+th}^2 G(h,h) \right\|. \tag{3.9}$$

where $\|\cdot\|$ denotes the Euclidean norm of $\mathbb{R}^p$.

We also consider the following nonnegative symmetric matrix

$$\Gamma_F = \sigma_n^2 \dot{G}_F \Sigma \dot{G}_F^T. \tag{3.10}$$

In the following, $\|\Gamma_F^{-1}\|$ denotes the Euclidean operator norm of Γ_F^{-1}, which is also the inverse of the smallest eigenvalue of Γ_F.

Let $\mathcal{I}$ be the collection of all intervals in $\mathbb{R}$, and for any $I \in \mathcal{I}$, let $\psi(I) = P(Z \in I)$, where Z is a $\mathcal{N}(0,1)$ random variable.

Theorem 8. *Let $G : \mathbb{R}^n \mapsto \mathbb{R}^p$ be a twice continuously differentiable function, and let Γ_F be as just defined. Suppose that $\Gamma_{F_{\langle\phi\rangle}}$ is nonsingular, and that there exists a sequence $(M_n)_{n\geq 1}$ such that $k_n = o(M_n)$ and*

$$B^2_{F_{\langle\phi\rangle}}(M_n) = o\left(\left\|\Gamma^{-1}_{F_{\langle\phi\rangle}}\right\|^{-1}\right). \tag{3.11}$$

Suppose further that the conditions of either Theorem 6 or Theorem 7 are verified. Then, for any $b \in \mathbb{R}^p$,

$$E\left[\sup_{I\in\mathcal{I}}\left|W\left(\frac{b^T\left(G(F) - G(Y_{\langle\phi\rangle})\right)}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}} \in I \,\middle|\, Y\right) - \psi(I)\right|\right] \to 0 \text{ as } n \to \infty.$$

Under the same conditions,

$$\sup_{I\in\mathcal{I}}\left|P\left(\frac{b^T\left(G(Y_{\langle\phi\rangle}) - G(F_{\langle\phi\rangle})\right)}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}} \in I\right) - \psi(I)\right| \to 0 \text{ as } n \to \infty. \tag{3.12}$$

$\sup_{I\in\mathcal{I}}|Q(I) - Q'(I)|$ is the Levy-Prokhorov distance between two distributions Q and Q' on $\mathbb{R}$. The Levy-Prokhorov distance metricizes the convergence in distribution. So, when $p = 1$ (3.12) says that the Levy-Prokhorov distance between the distribution of $\frac{b^T\left(G(Y_{\langle\phi\rangle}) - G(F_{\langle\phi\rangle})\right)}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}}$ and $\mathcal{N}(0,1)$ goes to 0 in mean.

An application of Theorem 8 is given in subsection 3.5.2, in the context of the regression on Fourier's basis. The proof is delayed to subsection 3.6.3.

3.5 Applications

We present now the three applications announced in section 3.2. The models studied and the collections of regressors used have been defined there.

3.5.1 The Gaussian sequence model

We consider the model (3.2). Here the MLE is only the projection $\theta_Y = (Y_j)_{1\leq j\leq k_n}$.

The nonparametric case corresponds to the estimation of θ^0. Under the assumption that θ^0 is in some regularity class, we obtain a Bernstein-von Mises Theorem with the

posterior convergence rate already obtained in previous works. On the contrary, for some priors known to achieve this rate, the centering point and the asymptotic variance of the posterior distribution do not fit with the ones expected in a Bernstein-von Mises Theorem. We also look at the semiparametric estimation of the squared ℓ^2 norm of θ^0.

The nonparametric estimation of θ^0

Proposition 9. *Suppose that* $\sum_{j=1}^{k_n}(\theta_j^0)^2$ *is bounded. This is verified in particular when* θ^0 *is an element of* $\ell^2(\mathbb{N})$ *non depending on* n. *With a prior* $\widetilde{W} = \mathcal{N}(0, \tau_n^2 I_{k_n})$ *such that* $n^{-1/4} = o(\tau_n)$, *we have whatever* $k_n \leq n$,

$$E\left\|\widetilde{W}(\mathrm{d}\theta|Y) - \mathcal{N}\left(\theta_Y, \frac{1}{n}I_{k_n}\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

and the convergence rate of θ towards θ_0 *is* $\sqrt{\frac{k_n}{n}}$: *for every* $\lambda_n \to \infty$,

$$E\left[\widetilde{W}\left(\|\theta - \theta_0\| \geq \lambda_n\sqrt{\frac{k_n}{n}}\middle| Y\right)\right] \to 0.$$

Proof. The beginning is an immediate corollary of Theorem 6. For the convergence rate, let $\lambda_n \to \infty$. Since $\theta_Y - \theta_0 \sim \mathcal{N}\left(0, \frac{1}{n}I_{k_n}\right)$,

$$P\left(\|\theta_Y - \theta_0\| \geq \frac{\lambda_n}{2}\sqrt{\frac{k_n}{n}}\right) \to 0.$$

In the same way

$$\begin{aligned} E\left[\widetilde{W}\left(\|\theta - \theta_Y\| \geq \frac{\lambda_n}{2}\sqrt{\frac{k_n}{n}}\right)\right] &\leq E\left\|\widetilde{W}(\mathrm{d}\theta|Y) - \mathcal{N}\left(\theta_Y, \frac{1}{n}I_{k_n}\right)\right\|_{\mathrm{TV}} \\ &\quad + \mathcal{N}\left(0, \frac{1}{n}I_{k_n}\right)\left(\left\{h:\ \|h\| \leq \frac{\lambda_n}{2}\sqrt{\frac{k_n}{n}}\right\}\right) \\ &\to 0. \end{aligned}$$

Therefore

$$E\left[\widetilde{W}\left(\|\theta - \theta_0\| \geq \lambda_n\sqrt{\frac{k_n}{n}}\right)\right] \to 0.$$

□

However in such a general setting we have no information about the bias between θ^0 and its projection θ_0. Several authors add the assumption that the true parameter belongs to a Sobolev class of regularity $\alpha > 0$, defined by the relation $\sum_{j=1}^{\infty}|\theta_j^0|^2 j^{2\alpha} < \infty$. In this setting we show that for some priors the induced posterior may achieve the nonparametric convergence rate but with a centering point and a variance different from what is expected in the Bernstein-von Mises Theorem. Then we exhibit priors for which both the Bernstein-von Mises Theorem and the nonparametric convergence rate hold.

From now on, we suppose that $\sum_{j=1}^{\infty} |\theta_j^0|^2 j^{2\alpha} < \infty$. In this setting [58, §7.6] considers a prior $\widetilde{W}$ such that θ_1, θ_2,... are independent, and θ_j is normally distributed with variance σ_{j,k_n}^2. Further, the variances are supposed to verify

$$c/k_n \leq \min\{\sigma_{j,k_n}^2 j^{2\alpha} : 1 \leq j \leq k_n\} \leq C/k_n \tag{3.13}$$

for some positive constants c and C. Suppose that $\alpha \geq 1/2$ and there exists constants C_1 and C_2 such that $C_1 n^{1/(1+2\alpha)} \leq k_n \leq C_2 n^{1/(1+2\alpha)}$. Then [58, Theorem 11] proved that the posterior converges at the rate $n^{-\alpha/(1+2\alpha)}$.

In order to get $n^{-1}I_{k_n}$ as asymptotic variance, we need more stringent conditions on k_n, or a flatter prior. As a counterexample consider, for $k_n \approx n^{1/(1+2\alpha)}$, the following choices of σ_{j,k_n}:

$$\sigma_{j,k_n}^2 = \begin{cases} k_n^{-1} & \text{if } 1 \leq j \leq k_n/2, \\ 2^{2\alpha}/n & \text{if } j > k_n/2. \end{cases}$$

Then $\min\{\sigma_{j,k_n}^2 j^{2\alpha} : 1 \leq j \leq k_n\} \approx k_n^{-1}$, and [58, Theorem 11] applies.
In this case we can perform an explicit calculus of the posterior distribution, similar to the one made in the proof of Theorem 6. The coordinates are independent, and

$$\widetilde{W}(\mathrm{d}\theta_j|Y) = \mathcal{N}\left(\frac{\sigma_{j,k_n}^2}{\sigma_n^2 + \sigma_{j,k_n}^2} Y_j, \frac{\sigma_n^2 \sigma_{j,k_n}^2}{\sigma_n^2 + \sigma_{j,k_n}^2}\right).$$

For $j > k_n/2$, $\frac{\sigma_{j,k_n}^2}{\sigma_n^2 + \sigma_{j,k_n}^2} = \frac{4^\alpha}{1+4^\alpha}$, and therefore $\left\| \widetilde{W}(\mathrm{d}\theta_j|Y) - \mathcal{N}(Y_j, \sigma_n^2) \right\|_{\mathrm{TV}}$ is bounded away from 0.

On the contrary with an isotropic prior, flat in all directions, we obtain the centering point and the asymptotic variance we expected, and the same convergence rate as previously.

Proposition 10. *Suppose that θ^0 belongs to the Sobolev class of regularity $\alpha > 0$. Choose a prior $\widetilde{W} = \mathcal{N}(0, \tau_n^2 I_{k_n})$ such that $n^{-1/4} = o(\tau_n)$, which insures the asymptotic normality of the posterior distribution as in Proposition 9.*
If further $k_n \approx n^{1/(1+2\alpha)}$, then the convergence rate of θ towards θ_0 and towards θ^0 is $n^{-\alpha/(1+2\alpha)}$: for every $\lambda_n \to \infty$,

$$E\left[\widetilde{W}\left(\|\theta - \theta^0\| \geq \lambda_n n^{-\alpha/(1+2\alpha)} \middle| Y\right)\right] \to 0.$$

Proof. We consider θ and θ_0 as elements of $\ell^2(\mathbb{N})$ by setting $\theta_j = \theta_{0,j} = 0$ for $j \geq k_n + 1$. The convergence rate towards θ_0 has already been established in Proposition 9. Since $\theta_{0,j} = \theta_j^0$ for $1 \leq j \leq k_n$, $\|\theta^0 - \theta_0\| \leq k_n^{-\alpha} \sqrt{\sum_{j=k_n+1}^{\infty} (\theta_j^0)^2 j^{2\alpha}} = O(k_n^{-\alpha})$. Therefore the convergence rate of θ towards θ^0 is also $n^{-\alpha/(1+2\alpha)}$. □

Semiparametric theorem for the ℓ^2 norm of θ^0

We still consider the same prior distribution as before, but now we look at the posterior distribution of $\|\theta\|^2$. To get the asymptotic normality with variance $n^{-1/2}$, we just need $k_n = o(\sqrt{n})$. To control the bias term we need $\alpha > 1/2$, and in this case we get an adaptive Bayesian estimator.

Proposition 11. *Let $\alpha > 1/2$ and suppose that θ^0 belongs to the Sobolev class of regularity α. Choose a prior $\widetilde{W} = \mathcal{N}(0, \tau_n^2 I_{k_n})$ such that $n^{-1/4} = o(\tau_n)$. Then, for any choice of k_n such that $k_n = o(\sqrt{n})$ and $\sqrt{n} = o(k_n^{2\alpha})$,*

$$E\left[\sup_{I\in\mathcal{I}}\left|\widetilde{W}\left(\frac{\sqrt{n}\left(\|\theta\|^2 - \|\theta_Y\|^2\right)}{2\|\theta^0\|} \in I \middle| Y\right) - \psi(I)\right|\right] \to 0 \text{ as } n \to \infty$$

and $\dfrac{\sqrt{n}\left(\|\theta_Y\|^2 - \|\theta_0\|^2\right)}{2\|\theta^0\|} \to \mathcal{N}(0,1)$ *in distribution, as $n \to \infty$. Further, the bias is negligible with respect to the square root of the variance:*

$$\frac{\sqrt{n}\left(\|\theta_0\|^2 - \|\theta^0\|^2\right)}{2\|\theta^0\|} = o(1).$$

In particular the choice $k_n = \sqrt{n/\ln n}$ is adaptive in α.

Proof. The conditions of Theorem 6 are fulfilled, as in Proposition 9.

Here $G(\theta) = \theta^T\theta$, $\dot{G}_\theta = 2\theta^T$ and $\ddot{G}_\theta = 2I_{k_n}$. Therefore $B_{\theta_0}(M_n) = 2M_n/n$, while $\Gamma_{\theta_0} = 4\|\theta_0\|^2/n$.

Let us choose $(M_n)_{n\geq 1}$ such that $k_n = o(M_n)$ and $M_n = o(\sqrt{n})$. Such sequences exist and fulfill the conditions of Theorem 8.

Since $\|\theta_0\|^2 \to \|\theta^0\|^2$, we can substitute the variance Γ_{θ_0} by $4\|\theta^0\|^2/n$ and get the two asymptotic normality results.

Eventually $\|\theta^0\|^2 - \|\theta_0\|^2 = \|\theta^0 - \theta_0\|^2 = O\left(k_n^{-2\alpha}\right)$, as in the proof of Proposition 10. If $\sqrt{n} = o(k_n^{2\alpha})$, we get $\sqrt{n}\left(\|\theta_0\|^2 - \|\theta^0\|^2\right) = o(1)$. □

3.5.2 Regression on Fourier's basis

Now we consider the regression model (3.3) with a function f in a Sobolev class $\mathcal{W}(\alpha, L)$, and use Fourier's basis (3.4). For any $\theta \in \mathbb{R}^{k_n}$, we define $f_\theta = \sum_{j=1}^{k_n} \theta_j \varphi_j$. We also denote by $\theta^0 \in \ell^2(\mathbb{N})$ the sequence of Fourier's coefficients of f: $f = \sum_{j=1}^{\infty} \theta_j^0 \varphi_j$.

The following useful Lemma about our collection of regressors can be found for instance in [105] (we slightly modified it to take into account the case n even):

Lemma 8. *Suppose either that n is odd and $k_n \leq n$, or n is even and $k_n \leq n-1$. Consider the collection $(\phi_j)_{1\leq j\leq k_n}$ defined before, and Φ the associated matrix. Then*

$$\Phi^T\Phi = nI_{k_n}.$$

This makes the regression on Fourier's basis very close to the Gaussian sequence model, and the result we obtain are similar.

We consider first the nonparametric estimation of f in a Sobolev class, for which we get a Bernstein-von Mises Theorem and the frequentist minimax $n^{-\alpha/(1+2\alpha)}$ posterior convergence rate for the L^2 norm.

Then we consider two semiparametric settings: the estimation of a linear functional of f, and the estimation of the L^2 norm of f. We get the adaptive $\sqrt{n}$ convergence rate for any $\alpha > 1/2$.

Nonparametric Bernstein-von Mises Theorem in Sobolev classes

Proposition 12. *Suppose that f belongs to some Sobolev class $\mathcal{W}(\alpha, L)$ for $L > 0$ and $\alpha > 1/2$. Let $k_n \approx n^{1/(1+2\alpha)}$ and $\widetilde{W} = \mathcal{N}(0, \gamma_n I_{k_n})$ be the prior on θ, for a sequence $(\gamma_n)_{n\geq 1}$ such that $1/\sqrt{n} = o(\gamma_n)$. Then*

$$E\left\|\widetilde{W}(\mathrm{d}\theta|Y) - \mathcal{N}\left(\theta_Y, \frac{\sigma^2}{n} I_{k_n}\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty$$

and the convergence rate relative to the euclidean norm for f_θ is $n^{-\alpha/(1+2\alpha)}$: for every $\lambda_n \to \infty$,

$$E\left[\widetilde{W}\left(\|f_\theta - f\| \geq \lambda_n n^{-\alpha/(1+2\alpha)} \middle| Y\right)\right] \to 0.$$

Proof. The conditions of Theorem 6 are fulfilled: with $\tau_n^2 = n\gamma_n$, we have $n = o(\tau_n^4)$. The first assertion follows.

Because of the orthogonal nature of Fourier's basis, $\|f_\theta - f\| = \|\theta - \theta^0\|$ in $\ell^2(\mathbb{N})$. We use the decomposition $\|\theta - \theta^0\|^2 \leq \|\theta - \theta_0\|^2 + \|\theta_0 - \theta^0\|^2$. In the same way as in the proof of Proposition 9, for any $\lambda_n \to \infty$,

$$E\left[\widetilde{W}\left(\|\theta - \theta_0\| \geq \lambda_n \sqrt{\frac{k_n}{n}}\right)\right] \to 0.$$

Going back to Definition 4, we have

$$\|\theta_0 - \theta^0\|^2 = \sum_{j=k_n+1}^{\infty} (\theta_j^0)^2 \leq k_n^{-2\alpha} \sum_{j=k_n+1}^{\infty} a_j^{2\alpha} (\theta_j^0)^2 = O(k_n^{-2\alpha}).$$

This permits to get

$$E\left[\widetilde{W}\left(\|\theta - \theta^0\| \geq \lambda_n n^{-\alpha/(1+2\alpha)} \middle| Y\right)\right] \to 0.$$

□

Linear functionals of f

Let $g : [0,1] \to \mathbb{R}$ be a function in $\mathbb{L}^2([0,1])$. We want to estimate $\mathcal{F}(f) = \int_0^1 fg$, and we approximate it by

$$\frac{1}{n}\sum_{i=1}^{n} g(i/n) f(i/n) = GF_0$$

where $G = (g(i/n)/n)_{1\leq i\leq n}^T$. The plug-in MLE estimator of GF_0 in the misspecified model $\langle\phi\rangle$ is $GY_{\langle\phi\rangle}$. More generally, we consider the functional $F \mapsto GF$.

The following result is adaptive, in the sense that the same choice $k_n = \lfloor n/\ln n \rfloor$ entails the convergence rate $n^{-1/2}$ for all values of $\alpha > 1/2$.

Proposition 13. *Suppose f is bounded, and let W be the prior induced by the $\mathcal{N}(0, \gamma_n\ I_{k_n})$ distribution on θ, for a sequence $(\gamma_n)_{n\geq 1}$ such that $1/\sqrt{n} = o(\gamma_n)$. Then*

1. $$E\left\|W(\mathrm{d}(GF)|Y) - \mathcal{N}\left(GY_{\langle\phi\rangle}, \sigma^2 G\Sigma G^T\right)\right\|_{\mathrm{TV}} \to 0$$
   and the distribution of $GY_{\langle\phi\rangle}$ *is* $\mathcal{N}\left(GF_{\langle\phi\rangle}, \sigma^2 G\Sigma G^T\right)$.
2. *Suppose further that* f *and* g *belong to some Sobolev class* $\mathcal{W}(\alpha, L)$ *for* $L > 0$ *and* $\alpha > 1/2$. *Then* $G\Sigma G^T \sim \frac{1}{n}\int_0^1 g^2$,
   $$E\left\|W\left(\mathrm{d}\frac{\sqrt{n}(GF - GY_{\langle\phi\rangle})}{\sigma\sqrt{\int_0^1 g^2}}\middle| Y\right) - \mathcal{N}(0,1)\right\|_{\mathrm{TV}} \to 0,$$
   and $\dfrac{\sqrt{n}(GY_{\langle\phi\rangle} - GF_{\langle\phi\rangle})}{\sigma\sqrt{\int_0^1 g^2}} \to \mathcal{N}(0,1)$ *in distribution, as* $n \to \infty$.
3. *Suppose that* f *and* g *belong to some Sobolev class* $\mathcal{W}(\alpha, L)$ *for* $L > 0$ *and* $\alpha > 1/2$, *and suppose further that* k_n *is large enough so that* $n = o(k_n^{2\alpha})$. *Then the bias is negligible with respect to the square root of the variance:*
   $$\frac{\sqrt{n}\left(GF_{\langle\phi\rangle} - \mathcal{F}(f)\right)}{\sigma\sqrt{\int_0^1 g^2}} = o(1).$$

Before the proof we give two lemmas, proved in Appendix 3.B, about the error terms of the approximation of a Sobolev class by a sieve build on Fourier's basis, and of the approximation of an integral by a Riemann sum.

Lemma 9. *Let* $\alpha > 1/2$ *and* $L > 0$. *We suppose* n *odd or* $k_n < n$. *If* $f \in \mathcal{W}(\alpha, L)$,
$$\|F_0 - F_{\langle\phi\rangle}\| \leq (1 + o(1))\frac{\sqrt{2}L}{\pi^\alpha}\frac{\sqrt{n}}{k_n^\alpha}.$$
Further, $\|F_0\| \sim \sqrt{n\int_0^1 f^2}$ *and* $\|F_0 - F_{\langle\phi\rangle}\| = O(k_n^{-\alpha}\|F_0\|)$.

Lemma 10. *Let two functions* $f \in \mathcal{W}(\alpha, L)$ *and* $g \in \mathcal{W}(\alpha', L')$ *for some* $\alpha, \alpha' > 1/2$ *and two positive numbers* L *and* L'. *Then*
$$\left|\frac{1}{n}\sum_{i=1}^n f(i/n)g(i/n) - \int_0^1 fg\right| = O\left(n^{-\inf(\alpha,\alpha')}\right).$$

Proof of Proposition 13.
1. The first assertion is just Corollary 1. The conditions of Theorem 6 are fulfilled, as in the proof of Proposition 12.
2. If $g \in \mathcal{W}(\alpha, L)$ for $L > 0$ and $\alpha > 1/2$, $G\Sigma G^T = \|\Sigma G^T\|^2 \sim \|G^T\|^2$ by Lemma 9. In the meantime $\|G^T\|^2 = \frac{1}{n^2}\sum_{i=1}^n g^2(x_i) \sim \frac{1}{n}\int_0^1 g^2$ by Lemma 10. So $G\Sigma G^T \sim \frac{1}{n}\int_0^1 g^2$, and the variance in the formulas of Corollary 1 can be substituted with $\frac{1}{n}\int_0^1 g^2$.
3. We decompose the bias into two terms, $|GF_0 - \mathcal{F}(f)|$ and $|GF_{\langle\phi\rangle} - GF_0|$, and show that both are $o(n^{-1/2})$. The first term is controlled by Lemma 10. For the last one, $|GF_{\langle\phi\rangle} - GF_0| \leq \|G^T\|\,\|F_{\langle\phi\rangle} - F_0\|$. $\|G^T\| = O(n^{-1/2})$, $\|F_{\langle\phi\rangle} - F_0\| = O(k_n^{-\alpha}\|F_0\|)$ by Lemma 9, and $\|F_0\| = O(\sqrt{n})$. We conclude thanks to the assumption $n = o(k_n^{2\alpha})$.

□

L^2 norm of f

Suppose that we want to estimate $\mathcal{F}(f) = \int_0^1 f^2$. We can consider the plug-in MLE estimator

$$G(Y_{\langle\phi\rangle}) = \frac{1}{n}\left\|Y_{\langle\phi\rangle}\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{k_n}\theta_{Y,j}\varphi_j(i/n)\right)^2.$$

More generally we define, for any $F \in \mathbb{R}^n$,

$$G(F) = \frac{1}{n}\|F\|^2. \tag{3.14}$$

With a Gaussian prior, we obtain the following result, which is also adaptive: the same $k_n = \lfloor\sqrt{n}/\ln n\rfloor$ is suitable whatever $\alpha > 1/2$.

Proposition 14. *Let $G(F) = \|F\|^2/n$. Suppose that $f \in \mathcal{W}(\alpha, L)$ for some $L > 0$ and $\alpha > 1/2$. Let W be the prior induced by the $\mathcal{N}(0, \gamma_n\, I_{k_n})$ distribution on θ, for a sequence $(\gamma_n)_{n\geq 1}$ such that $1/\sqrt{n} = o(\gamma_n)$. The sequence $(k_n)_{n\geq 1}$ can be chosen such that $k_n = o(\sqrt{n})$ and $\sqrt{n} = o(k_n^{2\alpha})$, and with such a choice,*

$$E\left[\sup_{I\in\mathcal{I}}\left|W\left(\frac{\sqrt{n}\left(G(F) - G(Y_{\langle\phi\rangle})\right)}{2\sigma\sqrt{\mathcal{F}(f)}} \in I \,\middle|\, Y\right) - \psi(I)\right|\right] \to 0 \text{ as } n \to \infty$$

and $\dfrac{\sqrt{n}\left(G(Y_{\langle\phi\rangle}) - G(F_{\langle\phi\rangle})\right)}{2\sigma\sqrt{\mathcal{F}(f)}} \to \mathcal{N}(0,1)$ *in distribution, as $n \to \infty$. Further, the bias is negligible with respect to the square root of the variance:*

$$\frac{\sqrt{n}\left(G(F_{\langle\phi\rangle}) - \mathcal{F}(f)\right)}{2\sigma\sqrt{\mathcal{F}(f)}} = o(1).$$

A similar corollary can be stated for a non-Gaussian prior.

Proof. First, let us note that the conditions of Theorem 6 are fulfilled, as in the proof of Proposition 12. Lemma 17 in Appendix 3.B insures that f is bounded.

In this setting $\dot{G}_F = (2/n)\,F^T$ and $D_F^2 G(h,h) = (2/n)\,\|h\|^2$ for any $F \in \mathbb{R}^n$ and any $h \in \mathbb{R}^n$. Therefore $B_F(a) = 2\sigma^2 a/n$, and $\Gamma_F = 4(\sigma^2/n^2)\,\|F\|^2$. By Lemma 9, $\|F_{\langle\phi\rangle}\|^2 \sim \|F_0\|^2 \sim n\,\mathcal{F}(f)$. Thus $\Gamma_{F_{\langle\phi\rangle}} = 4(1 + o(1))\mathcal{F}(f)/n$.

Let us choose $(M_n)_{n\geq 1}$ such that $k_n = o(M_n)$ and $M_n = o(\sqrt{n})$. Such sequences exist and fulfill the conditions of Theorem 8. We can substitute the variance $\Gamma_{F_{\langle\phi\rangle}}$ by $4\mathcal{F}(f)/n$ and get the two asymptotic normality results.

Let us now consider the bias term.

$$\mathcal{F}(f) - G(F_{\langle\phi\rangle}) \leq \frac{\|F_0\|^2 - \|F_{\langle\phi\rangle}\|^2}{n} + \left(\int_0^1 f^2 - \frac{1}{n}\sum_{i=1}^{n} f^2(i/n)\right)$$

We use Lemma 9 to control $\|F_0\|^2 - \|F_{\langle\phi\rangle}\|^2$, and Lemma 10 for the other term:

$$\left|\mathcal{F}(f) - G(F_{\langle\phi\rangle})\right| = O\left(k_n^{-2\alpha}\right) + O\left(n^{-\alpha}\right).$$

This is a $o(1/\sqrt{n})$ under the assumptions of Corollary 14. □

3.5.3 Regression on splines

Here we consider the regression model for functions in $C^\alpha[0,1]$ with $\alpha > 0$, using splines. The problem has been set in section 3.2. We first develop further the framework and the assumptions used here, and recall the previous result of [58, §7.7.1] which obtains the posterior concentration at the frequentist minimax rate. Then we present two Bernstein-von Mises Theorems: the first one with the same prior as [58] but a stronger condition on k_n (or equivalently on α); the second one with a flatter prior, for which we retrieve the minimax convergence rate in addition to the asymptotic Gaussianity of the posterior distribution.

For any $\theta \in \mathbb{R}^{k_n}$, we define $f_\theta = \sum_{j=1}^{k_n} \theta_j B_j$. The B-splines basis has the following approximation property: for any $\alpha > 0$, there exist $C_\alpha > 0$ such that, if $f \in C^\alpha[0,1]$, there exists $\theta^\infty \in \mathbb{R}^{k_n}$ verifying

$$\|f - f_{\theta^\infty}\|_\infty \leq C_\alpha k_n^{-\alpha} \|f\|_\alpha. \tag{3.15}$$

We need the design $\left(x_i^{(n)}\right)_{n\geq 1, 1\leq i\leq n}$ to be sufficiently regular but, as stressed in [58], the spacial separation property of B-splines permits to express the precise condition in terms of the covariance matrix $\Phi^T\Phi$. We suppose that there exist positive constants C_1 and C_2 such that, as n increases, whatever $\theta \in \mathbb{R}^{k_n}$,

$$C_1 \frac{n}{k_n} \|\theta\|^2 \leq \theta^T \Phi^T \Phi \theta \leq C_2 \frac{n}{k_n} \|\theta\|^2. \tag{3.16}$$

A norm $\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n |f(x_i)|^2}$ is associated to the design. Note that $\sqrt{n}\|f_\theta\|_n = \|\Phi\theta\|$ if $\theta \in \mathbb{R}^{k_n}$. Under condition (3.16) we have a relation between $\|\cdot\|_n$ and the euclidean norm on the parameter space: for every θ_1 and θ_2

$$C_1\|\theta_1 - \theta_2\| \leq \sqrt{k_n}\, \|f_{\theta_1} - f_{\theta_2}\|_n \leq C_2 \|\theta_1 - \theta_2\|.$$

With these conditions [58, Theorem 12] gets the posterior concentration at the minimax rate. Take $\alpha \geq 1/2$, let $\widetilde{W} = \mathcal{N}(0, I_{k_n})$ be the prior on the spline coefficients, and suppose there exists constants C_3 and C_4 such that $C_3 n^{1/(1+2\alpha)} \leq k_n \leq C_4 n^{1/(1+2\alpha)}$. Then the posterior concentrates at the minimax rate $n^{-\alpha/(1+2\alpha)}$ relative to $\|\cdot\|_n$: for every $\lambda_n \to \infty$,

$$E\left[\widetilde{W}\left(\|f_\theta - f\|_n \geq \lambda_n n^{-\alpha/(1+2\alpha)} \middle| Y\right)\right] \to 0.$$

This is equivalent to a convergence rate $n^{\frac{1-2\alpha}{2(1+2\alpha)}}$ relative to the euclidean norm for θ:

$$E\left[\widetilde{W}\left(\|\theta - \theta_0\| \geq \lambda_n n^{\frac{1-2\alpha}{2(1+2\alpha)}} \middle| Y\right)\right] \to 0.$$

Indeed (3.15) and the projection property entail

$$\|f_{\theta_0} - f\|_n \leq \|f_{\theta^\infty} - f\|_n \leq \|f_{\theta^\infty} - f\|_\infty \leq C_\alpha \|f\|_\alpha k_n^{-\alpha}.$$

With modified assumptions we get also the Bernstein-von Mises Theorem in two different settings. First, with the same prior as [58]:

Proposition 15. *Assume that f is bounded, $k_n = o\left(\left(\frac{n}{\ln n}\right)^{1/3}\right)$, and (3.16) holds. Let $\widetilde{W} = \mathcal{N}(0, I_{k_n})$ be the prior on the spline coefficients. Then*

$$E\left\|\widetilde{W}(\mathrm{d}\theta|Y) - \mathcal{N}\left(\theta_Y, \sigma^2(\Phi^T\Phi)^{-1}\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty$$

and the convergence rate relative to the euclidean norm for θ is $\frac{k_n}{\sqrt{n}}$.

We need $\alpha > 1$ to get the Gaussian shape with the same convergence rate as in [58]. The conditions of Proposition 15 are verified in particular if there exists constants C_3 and C_4 such that $C_3 n^{1/(1+2\alpha)} \le k_n \le C_4 n^{1/(1+2\alpha)}$. In this case the convergence rate for θ is $n^{\frac{1-2\alpha}{2(1+2\alpha)}}$.

Proof. We apply Theorem 7. We can choose M_n such that $k_n \ln n = o(M_n)$ and $M_n = o\left(\frac{n}{k_n^2}\right)$. Assumption 2 is then trivially verified.

From (3.16) we get $\left\|\Phi^T\Phi\right\| \le C_2\frac{n}{k_n}$ and $\left\|(\Phi^T\Phi)^{-1}\right\| \le C_1^{-1}\frac{k_n}{n}$. We have also $\ln\det(\Phi^T\Phi) \le k_n \ln C_2 + k_n \ln\left(\frac{n}{k_n}\right) = O(k_n \ln n) = o(M_n)$. Since $\theta_0 = \Phi(\Phi^T\Phi)^{-1}F_0$,

$$\|\theta_0\|^2 \le \frac{k_n}{C_1 n}\|F_0\|^2 \le \frac{\|f\|_\infty}{C_1}k_n.$$

Therefore $-\ln w(\theta_0) = O(1) + \frac{1}{2}\|\theta_0\|^2 = O(k_n) = o(M_n)$, and assumption 3 holds.

Let $h \in \mathbb{R}^{k_n}$ such that $\|\Phi h\|^2 \le \sigma^2 M_n$. We have $\|h\|^2 \le \left\|(\Phi^T\Phi)^{-1}\right\|\|\Phi h\|^2 \le \frac{\sigma^2 k_n M_n}{C_1 n} = o\left(k_n^{-1}\right)$. Therefore

$$\sup_{\|\Phi h\|^2 \le \sigma^2 M_n}\left|\ln\frac{w(\theta_0+h)}{w(\theta_0)}\right| \le \sup_{\|\Phi h\|^2 \le \sigma^2 M_n}\frac{\|h\|^2 + 2\|h\|\|\theta_0\|}{2} = o(1) \tag{3.17}$$

and assumption 1 follows.

Let us now prove the convergence rate. Let $\lambda_n \to \infty$. Then

$$P\left(\|\theta_Y - \theta_0\| \ge \frac{\lambda_n k_n}{2\sqrt{n}}\right) \le P\left(\|\Phi(\theta_Y - \theta_0)\|^2 \ge \frac{C_1\lambda_n^2 k_n}{4}\right) \to 0$$

since $\|\Phi(\theta_Y - \theta_0)\|^2 \sim \sigma^2\chi^2(k_n)$. In the same way

$$\begin{aligned} E\left[\widetilde{W}\left(\|\theta - \theta_Y\| \ge \frac{\lambda_n k_n}{2\sqrt{n}}\right)\right] &\le E\left\|\widetilde{W}(\mathrm{d}\theta|Y) - \mathcal{N}\left(\theta_Y, \sigma^2(\Phi^T\Phi)^{-1}\right)\right\|_{\mathrm{TV}} \\ &\quad + \mathcal{N}\left(0, \sigma^2(\Phi^T\Phi)^{-1}\right)\left(\left\{h:\ \|h\| \le \frac{\lambda_n k_n}{2\sqrt{n}}\right\}\right) \\ &\to 0. \end{aligned}$$

Therefore

$$E\left[\widetilde{W}\left(\|\theta - \theta_0\| \ge \frac{\lambda_n k_n}{\sqrt{n}}\right)\right] \to 0.$$

□

The situation is similar to the one we encountered with the Gaussian sequence model. To get the Bernstein-von Mises Theorem with the same convergence rate as [58] for $\alpha \leq 1$, we need a flatter prior:

Proposition 16. *Assume that f is bounded and (3.16) holds. Let $\widetilde{W} = \mathcal{N}(0, \tau_n^2 I_{k_n})$ be the prior on the spline coefficients, with the sequence τ_n verifying*

$$\frac{k_n^2 \ln n}{n} = o(\tau_n^2) \qquad \textit{and} \qquad \frac{k_n^3 \ln n}{n} = o(\tau_n^4).$$

Then

$$E \left\| \widetilde{W}(\mathrm{d}\theta|Y) - \mathcal{N}\left(\theta_Y, \sigma^2(\Phi^T\Phi)^{-1}\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty$$

and the convergence rate relative to the euclidean norm for θ is $\frac{k_n}{\sqrt{n}}$.

When $\alpha > 0$ and k_n is of order $n^{1/(1+2\alpha)}$, the conditions reduce to $n^{\frac{2-2\alpha}{1+2\alpha}} \ln n = o(\tau_n^4)$. So we retrieve the convergence rate of [58] in addition to the Gaussian shape with the same k_n, even for $\alpha \leq 1$, but with a different prior.

Proof. The proof is essentially the same as for Proposition 15. M_n can be chosen such as $k_n \ln n = o(M_n)$, $M_n = o\left(\frac{n\tau_n^2}{k_n}\right)$, and $M_n = o\left(\frac{n\tau_n^4}{k_n^2}\right)$. These last two conditions are the ones needed to obtain the same upper bounds as in (3.17). □

3.6 Proofs

3.6.1 Proof of Theorem 6

In the present setting all distributions are explicit and admit densities with respect to the corresponding Lebesgue measure. We decompose any $y \in \mathbb{R}^n$ in two orthogonal components $y = \Phi\theta_y + y'$, with $\Phi^T y' = 0$. Then

$$\begin{aligned}
\mathrm{d}P_\theta(y) &= c_1 \exp\left\{ -\frac{1}{2\sigma_n^2}\left(\|\Phi\theta\|^2 + \|\Phi\theta_y\|^2 + \|y'\|^2 - 2\theta^T\Phi^T\Phi\theta_y \right) \right\} \\
\mathrm{d}\widetilde{W}(\theta) &= c_2 \exp\left\{ -\frac{1}{2\tau_n^2}\|\Phi\theta\|^2 \right\} \\
\mathrm{d}P_\theta(y)\,\mathrm{d}\widetilde{W}(\theta) &= c_1 c_2 \exp\left\{ -\frac{\sigma_n^2+\tau_n^2}{2\sigma_n^2\tau_n^2} \left\| \Phi\left(\theta - \frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\theta_y\right) \right\|^2 \right. \\
&\qquad \left. -\frac{1}{2(\sigma_n^2+\tau_n^2)}\|\Phi\theta_y\|^2 - \frac{1}{2\sigma_n^2}\|y'\|^2 \right\}
\end{aligned}$$

where $c_1 = (2\pi)^{-n/2}\sigma_n^{-n}$ and $c2 = (2\pi)^{-k_n/2}\tau_n^{-k_n}\det(\Phi^T\Phi)^{-1}$.

Using the Bayes rule, we get the density of $\widetilde{W}(\mathrm{d}\theta|Y)$, in which we recognize the normal distribution

$$\widetilde{W}(\mathrm{d}\theta|Y) = \mathcal{N}\left(\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\theta_Y, \frac{\sigma_n^2\tau_n^2}{\sigma_n^2+\tau_n^2}(\Phi^T\Phi)^{-1} \right). \tag{3.18}$$

At that point, we have got an exact expression of $\widetilde{W}(\mathrm{d}\theta|Y)$, but nor the centering nor the variance correspond to the limit distribution given in Theorem 6. Therefore we make use of the triangle inequality, with intermediate distribution $Q = \mathcal{N}\left(\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}\right)$. We first deal with the change in the variance.

Let $\alpha_n = \frac{\tau_n}{\sigma_n}\sqrt{\ln\left(1+\frac{\sigma_n^2}{\tau_n^2}\right)}$, and f and g be respectively the density functions of $\mathcal{N}(0, I_{k_n})$ and $\mathcal{N}\left(0, \frac{\tau_n^2}{\sigma_n^2+\tau_n^2}I_{k_n}\right)$. Let U be a random variable following the chi-square distribution with k_n degrees of freedom $\chi^2(k_n)$. Then

$$\begin{aligned}
\left\|\widetilde{W}(\mathrm{d}\theta|Y) - Q\right\|_{\mathrm{TV}} &= \left\|\mathcal{N}(0, I_{k_n}) - \mathcal{N}\left(0, \frac{\tau_n^2}{\sigma_n^2+\tau_n^2}I_{k_n}\right)\right\|_{\mathrm{TV}} \\
&= \int_{\mathbb{R}^{k_n}} (g-f)_+ = \int_{\|x\|\le\sqrt{k_n}\alpha_n} g(x) - f(x)\,\mathrm{d}^n x \\
&= P(U \le k_n\alpha_n^2) - P\left(U \le \frac{\sigma_n^2+\tau_n^2}{\tau_n^2}k_n\alpha_n^2\right) \\
&= P\left(\alpha_n^2 \le \frac{U}{k_n} \le \frac{\sigma_n^2+\tau_n^2}{\tau_n^2}\alpha_n^2\right).
\end{aligned}$$

As n goes to infinity, $\frac{U}{k_n}$ converges towards $\mathcal{N}(0,1)$ in distribution. Since $\sigma_n = o(\tau_n)$, both $\frac{\sigma_n^2+\tau_n^2}{\tau_n^2}$ and α_n go to 1. As a consequence, $\left\|\widetilde{W}(\mathrm{d}\theta|Y) - Q\right\|_{\mathrm{TV}}$ goes to zero as n goes to infinity.

Let us now deal with the centering term.

Lemma 11. *Let U be a standard normal random variable, let $k \ge 1$, and let $Z \in \mathbb{R}^k$. Then*

$$\|\mathcal{N}(0, I_k) - \mathcal{N}(Z, I_k)\|_{\mathrm{TV}} = P(|U| \le \|Z\|/2) \le \|Z\|/\sqrt{2\pi}.$$

Proof. Let g be the density of $\mathcal{N}(0, I_k)$. Then

$$\begin{aligned}
\|\mathcal{N}(0, I_k) - \mathcal{N}(Z, I_k)\|_{\mathrm{TV}} &= \int_{\mathbb{R}^k} (g(x) - g(x-Z))_+\,\mathrm{d}^k x \\
&= \int_{\{2x^TZ \le \|Z\|^2\}} (g(x) - g(x-Z))\,\mathrm{d}^k x \\
&= P(U \le \|Z\|/2) - P(U + \|Z\| \le \|Z\|/2) \\
&\le \|Z\|/\sqrt{2\pi}.
\end{aligned}$$

The last line comes from the density of $\mathcal{N}(0,1)$ being bounded by $1/\sqrt{2\pi}$. □

Let $\sqrt{\Phi^T\Phi}$ be a square root of the matrix $\Phi^T\Phi$. Then

$$\begin{aligned}
\left\|\mathcal{N}\left(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}\right) - Q\right\|_{\mathrm{TV}} &= \left\|\mathcal{N}(0, I_{k_n}) - \mathcal{N}\left(\frac{\sigma_n}{\tau_n^2+\sigma_n^2}\sqrt{\Phi^T\Phi}\theta_Y, I_{k_n}\right)\right\|_{\mathrm{TV}} \\
&\le \frac{1}{\sqrt{2\pi}}\frac{\sigma_n}{(\tau_n^2+\sigma_n^2)}\|\Phi\theta_Y\| \\
&\le \frac{1}{\sqrt{2\pi}}\frac{\sigma_n}{(\tau_n^2+\sigma_n^2)}\left(\|F_0\| + \sqrt{\varepsilon^T\Sigma\varepsilon}\right).
\end{aligned}$$

$\varepsilon^T \Sigma \varepsilon$ is a random variable following $\sigma_n^2 \chi^2(k_n)$ distribution. Therefore

$$E \left\| \mathcal{N}\left(\theta_Y, \sigma_n^2 (\Phi^T \Phi)^{-1}\right) - Q \right\|_{\mathrm{TV}} \leq \frac{1}{\sqrt{2\pi}} \frac{\sigma_n}{\tau_n^2 + \sigma_n^2} \left(\|F_0\| + \sigma_n \sqrt{k_n} \right).$$

This goes to zero under the assumptions of Theorem 6.

To conclude the proof, let us just note that we deduce the results on $W(\mathrm{d}F|Y)$ from the ones on $\widetilde{W}(\mathrm{d}\theta|Y)$, by the linear relation $F = \Phi\theta$.

3.6.2 Proof of Theorem 7.

We make the proof for $\widetilde{W}(\mathrm{d}\theta|Y)$. Then the result for $W(\mathrm{d}F|Y)$ is immediate. Our method is adapted from [17].

For $M > 0$, consider the ellipsoid

$$\mathcal{E}_{\theta_0,\Phi}(M) = \left\{ \theta \in \mathbb{R}^{k_n} : (\theta - \theta_0)^T \Phi^T \Phi (\theta - \theta_0) \leq \sigma_n^2 M \right\}. \tag{3.19}$$

To any probability measure P on $\mathbb{R}^{k_n}$, we associate the probability

$$P^M = \frac{P\left(\cdot \cap \mathcal{E}_{\theta_0,\Phi}(M)\right)}{P\left(\mathcal{E}_{\theta_0,\Phi}(M)\right)} \tag{3.20}$$

with support in $\mathcal{E}_{\theta_0,\Phi}(M)$. It can be easily checked that

$$\left\| P - P^M \right\|_{\mathrm{TV}} = P\left(\mathcal{E}^c_{\theta_0,\Phi}(M)\right). \tag{3.21}$$

Then the calculus is divided in three parts, M_n being used as a threshold to truncate the queues of the probability distributions. Gathered, these lemmas give Theorem 7.

Lemma 12. *If $k_n < 4M_n$, then*

$$E \left\| \mathcal{N}\left(\theta_Y, \sigma_n^2 (\Phi^T \Phi)^{-1}\right) - \mathcal{N}^{M_n}\left(\theta_Y, \sigma_n^2 (\Phi^T \Phi)^{-1}\right) \right\|_{\mathrm{TV}} \leq 2 e^{-\frac{\left(\sqrt{M_n} - 2\sqrt{k_n}\right)^2}{8}}.$$

If $k_n = o(M_n)$, for n large enough, this bound can be replaced by $\exp(-M_n/9)$.

Proof. Two cases occur, depending on whether θ_Y is near or far from θ_0:

$$\begin{aligned} &\left\| \mathcal{N}\left(\theta_Y, \sigma_n^2 (\Phi^T \Phi)^{-1}\right) - \mathcal{N}^{M_n}\left(\theta_Y, \sigma_n^2 (\Phi^T \Phi)^{-1}\right) \right\|_{\mathrm{TV}} \\ &\qquad = \mathcal{N}\left(\theta_Y, \sigma_n^2 (\Phi^T \Phi)^{-1}\right)\left(\mathcal{E}^c_{\theta_0,\Phi}(M_n)\right) \\ &\qquad \leq \mathbb{1}_{(\theta_Y - \theta_0)^T \Phi^T \Phi (\theta_Y - \theta_0) > \sigma_n^2 M_n/4} \\ &\qquad\quad + \mathcal{N}\left(\theta_0, \sigma_n^2 (\Phi^T \Phi)^{-1}\right)\left(\mathcal{E}^c_{\theta_0,\Phi}(M_n/4)\right) \end{aligned}$$

Let U a random variable following the $\chi^2(k_n)$ distribution. Taking the expectation in the last line we get

$$E \left\| \mathcal{N}\left(\theta_Y, \sigma_n^2 (\Phi^T \Phi)^{-1}\right) - \mathcal{N}^{M_n}\left(\theta_Y, \sigma_n^2 (\Phi^T \Phi)^{-1}\right) \right\|_{\mathrm{TV}} \leq 2P(U > M_n/4).$$

To conclude we use Cirelson's inequality [75]:

$$P(\sqrt{U} > \sqrt{k_n} + \sqrt{2x}) \leq \exp(-x) \tag{3.22}$$

□

Lemma 13. *If* $\sup_{\|\Phi h\|^2 \le \sigma_n^2 M_n, \|\Phi g\|^2 \le \sigma_n^2 M_n} \frac{w(\theta_0+h)}{w(\theta_0+g)} \to 1$ *as* $n \to \infty$, *then*

$$E\left\|\widetilde{W}^{M_n}(\mathrm{d}\theta|Y) - \mathcal{N}^{M_n}\left(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty.$$

Proof. Let us first note that, for every θ and τ in $\mathbb{R}^{k_n}$, for every $Y \in \mathbb{R}^n$,

$$\frac{\mathrm{d}P_\theta(Y)}{\mathrm{d}P_\tau(Y)} = \exp\left\{\frac{-\|\Phi\theta\|^2 + \|\Phi\tau\|^2 - 2Y^T\Phi(\tau-\theta)}{2\sigma_n^2}\right\} = \frac{\mathrm{d}\mathcal{N}\left(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}\right)(\theta)}{\mathrm{d}\mathcal{N}\left(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}\right)(\tau)}. \tag{3.23}$$

In the following we mainly use the convexity of $x \mapsto (1-x)_+$. We abbreviate $\mathcal{N}^{M_n}\left(\theta_Y, \sigma_n^2(\Phi^T\Phi)^{-1}\right)$ into $\mathcal{N}^{M_n}$. Then

$$\begin{aligned}
&\left\|\widetilde{W}^{M_n}(\mathrm{d}\theta|Y) - \mathcal{N}^{M_n}\right\|_{\mathrm{TV}} \\
&\qquad = \int\left(1 - \frac{\mathrm{d}\mathcal{N}^{M_n}(\theta)}{\mathrm{d}\widetilde{W}^{M_n}(\theta|Y)}\right)_+ \mathrm{d}\widetilde{W}^{M_n}(\theta|Y) \\
&\qquad = \int\left(1 - \frac{\mathrm{d}\mathcal{N}^{M_n}(\theta)\int \frac{w(\tau)}{\mathrm{d}\mathcal{N}^{M_n}(\tau)}\,\mathrm{d}P_\tau(Y)\,\mathrm{d}\mathcal{N}^{M_n}(\tau)}{w(\theta)\,\mathrm{d}P_\theta(Y)}\right)_+ \mathrm{d}\widetilde{W}^{M_n}(\theta|Y) \\
&\qquad \le \int\int\left(1 - \frac{w(\tau)\,\mathrm{d}\mathcal{N}^{M_n}(\theta)\,\mathrm{d}P_\tau(Y)}{w(\theta)\,\mathrm{d}\mathcal{N}^{M_n}(\tau)\,\mathrm{d}P_\theta(Y)}\right)_+ \mathrm{d}\mathcal{N}^{M_n}(\tau)\,\mathrm{d}\widetilde{W}^{M_n}(\theta|Y) \\
&\qquad = \int\int\left(1 - \frac{w(\tau)}{w(\theta)}\right)_+ \mathrm{d}\mathcal{N}^{M_n}(\tau)\,\mathrm{d}\widetilde{W}^{M_n}(\theta|Y) \\
&\qquad \le 1 - \inf_{\|\Phi h\|^2 \le \sigma_n^2 M_n, \|\Phi g\|^2 \le \sigma_n^2 M_n} \frac{w(\theta_0+h)}{w(\theta_0+g)}.
\end{aligned}$$

□

Proposition 17 (Posterior concentration)**.** *Suppose that Condition 1, Condition 2, and Condition 3 of Theorem 7 hold. Then*

$$\begin{aligned}
E\left\|\widetilde{W}(\mathrm{d}\theta|Y) - \widetilde{W}^{M_n}(\mathrm{d}\theta|Y)\right\|_{\mathrm{TV}} &= E\left[\widetilde{W}\left(\mathcal{E}^C_{\theta_0,\Phi}(M_n)\middle|\, Y\right)\right] \\
&\to 0 \text{ as } n \to \infty.
\end{aligned}$$

Proposition 17 is proved in Appendix 3.A, using the important following Lemma.

Lemma 14. *Let* $a \in \mathbb{R}^n$ *such that* $\Phi^T a = 0$. *Then, for any* $y \in \mathbb{R}^n$, $W(\cdot|Y=y) = W(\cdot|Y=y+a)$.

Lemma 14 states that the distribution $W(\cdot|Y)$ is invariant by any translation of Y orthogonal to $\langle\phi\rangle$. As a consequence, proving Proposition 17 in the case $F_0 = \Phi\theta_0$ is enough.

Proof. We decompose any $y \in \mathbb{R}^n$ in two orthogonal components $y = \Phi\theta_y + y'$, with $\Phi^T y' = 0$. The density of P_θ is equal to

$$\begin{aligned}\mathrm{d}P_\theta(y) &= \frac{1}{\left(\sigma_n\sqrt{2\pi}\right)^n}\exp\left\{-\frac{\|\Phi\theta_y + y' - \Phi\theta\|^2}{2\sigma_n^2}\right\} \\ &= \frac{1}{\left(\sigma_n\sqrt{2\pi}\right)^n}\exp\left\{-\frac{\|y'\|^2}{2\sigma_n^2}\right\}\exp\left\{-\frac{\|\Phi\theta_y - \Phi\theta\|^2}{2\sigma_n^2}\right\}.\end{aligned}$$

On the same way,

$$\begin{aligned}\mathrm{d}P_\theta(y+a) &= \frac{1}{\left(\sigma_n\sqrt{2\pi}\right)^n}\exp\left\{-\frac{\|y'+a\|^2}{2\sigma_n^2}\right\}\exp\left\{-\frac{\|\Phi\theta_y - \Phi\theta\|^2}{2\sigma_n^2}\right\} \\ &= \exp\left\{-\frac{\|y'+a\|^2 - \|y'\|^2}{2\sigma_n^2}\right\}\mathrm{d}P_\theta(y).\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{d}P^W(y+a) &= \int \mathrm{d}P_\theta(y+a)w(\theta)\,\mathrm{d}\theta \\ &= \exp\left\{-\frac{\|y'+a\|^2 - \|y'\|^2}{2\sigma_n^2}\right\}\mathrm{d}P^W(y)\end{aligned}$$

and

$$\begin{aligned}\widetilde{W}(\mathrm{d}\theta|Y = y+a) &= \frac{\mathrm{d}P_\theta(y+a)w(\theta)\,\mathrm{d}\theta}{\mathrm{d}P^W(y+a)} \\ &= \widetilde{W}(\mathrm{d}\theta|Y = y).\end{aligned}$$

□

3.6.3 Proof of Theorem 8.

Let us consider the following Taylor expansion:

$$\begin{aligned}G(F) - G(Y_{\langle\phi\rangle}) &= \dot{G}_{F_{\langle\phi\rangle}}(F - Y_{\langle\phi\rangle}) \\ &\quad + \frac{1}{2}\int_0^1 (1-t)D^2_{F_{\langle\phi\rangle}+t\,(F-F_{\langle\phi\rangle})}G(F - F_{\langle\phi\rangle}, F - F_{\langle\phi\rangle})\,\mathrm{d}t \\ &\quad - \frac{1}{2}\int_0^1 (1-t)D^2_{F_{\langle\phi\rangle}+t\,(Y_{\langle\phi\rangle}-F_{\langle\phi\rangle})}G(Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}, Y_{\langle\phi\rangle} - F_{\langle\phi\rangle})\,\mathrm{d}t.\end{aligned}$$

Suppose that $F \in \langle\phi\rangle$, $\|F - F_{\langle\phi\rangle}\|^2 \leq \sigma_n^2 M_n$, and $\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 \leq \sigma_n^2 M_n$. Then, for any $b \in \mathbb{R}^p$,

$$\left|b^T\left(G(F) - G(Y_{\langle\phi\rangle}) - \dot{G}_{F_{\langle\phi\rangle}}(F - Y_{\langle\phi\rangle})\right)\right| \leq \|b\|B_{F_{\langle\phi\rangle}}(M_n).$$

On the other hand, $\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b} \geq \sqrt{\left\|\Gamma^{-1}_{F_{\langle\phi\rangle}}\right\|^{-1}}\|b\|$. Moreover

$$\left\| W\left(\mathrm{d}\frac{b^T\dot{G}_{F_{\langle\phi\rangle}}(F - Y_{\langle\phi\rangle})}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}} \middle| Y \right) - \mathcal{N}(0,1) \right\|_{\mathrm{TV}} \leq \left\| W(\mathrm{d}F|Y) - \mathcal{N}\left(Y_{\langle\phi\rangle}, \sigma_n^2\Sigma\right) \right\|_{\mathrm{TV}}.$$

Let $\eta_n = \sqrt{\left\|\Gamma^{-1}_{F_{\langle\phi\rangle}}\right\|} B_{F_{\langle\phi\rangle}}(M_n)$, which tends to 0 by hypothesis. Let also

$$I_{\eta_n} = \{x \in \mathbb{R} : \exists x' \in I, |x - x'| \leq \eta_n\}.$$

Note that $\psi\left(I_{\eta_n}\right) \leq \psi(I) + \sqrt{\frac{2}{\pi}}\eta_n$.

Gathering all this information, we can get the upper bound

$$\begin{aligned} W\left(\frac{b^T\left(G(F) - G(Y_{\langle\phi\rangle})\right)}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}} \in I \middle| Y \right) \leq \psi(I) + \sqrt{\frac{2}{\pi}}\eta_n \\ + \left\| W(\mathrm{d}F|Y) - \mathcal{N}\left(Y_{\langle\phi\rangle}, \sigma_n^2\Sigma\right) \right\|_{\mathrm{TV}} \\ + \mathbb{1}_{\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n} \\ + W\left(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n \middle| Y \right). \end{aligned}$$

A lower bound is obtained in the same way. Taking the expectation,

$$\begin{aligned} E\left| W\left(\frac{b^T\left(G(F) - G(Y_{\langle\phi\rangle})\right)}{\sqrt{b^T\Gamma_{F_{\langle\phi\rangle}}b}} \in I \middle| Y \right) - \psi(I) \right| \\ \leq o(1) + P(\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n) \\ + E\left[W\left(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n \middle| Y \right) \right]. \end{aligned}$$

But $\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2$ follows the $\sigma_n^2\chi^2(k_n)$ distribution, and since $k_n = o(M_n)$,

$$P(\|Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n) = o(1).$$

To conclude the proof of the Bayesian part of Theorem 8, we use the following:

Lemma 15. *Suppose that the conditions of either Theorem 6 or Theorem 7 are verified. Then*

$$E\left[W\left(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n \middle| Y \right) \right] \to 0 \text{ as } n \to \infty.$$

Proof. For smooth priors, this is an immediate corollary of Proposition 17. Let us suppose we are under the conditions of Theorem 6.

Let Z be a $\mathcal{N}\left(0, \frac{\sigma_n^2\tau_n^2}{\sigma_n^2+\tau_n^2}(\Phi^T\Phi)^{-1}\right)$ random vector in $\mathbb{R}^n$ independent on Y, and U a

random variable following $\chi^2(k_n)$. Using (3.18), we get

$$\begin{aligned} W\left(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n \middle| Y\right) & \\ &= P\left(\left\|Z + \frac{\tau_n^2}{\sigma_n^2+\tau_n^2} Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\right\|^2 > \sigma_n^2 M_n\right) \\ &\leq P\left(\|Z\| > \sigma_n\sqrt{M_n} - \left\|\frac{\tau_n^2}{\sigma_n^2+\tau_n^2} Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\right\|\right) \\ &\leq \begin{cases} 1 & \text{if } \left\|\frac{\tau_n^2}{\sigma_n^2+\tau_n^2} Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\right\| > \frac{2\sigma_n\sqrt{M_n}}{3} \\ P\left(U > \frac{\sigma_n^2+\tau_n^2}{\tau_n^2}\frac{M_n}{9}\right) & \text{otherwise.} \end{cases} \end{aligned}$$

Since $k_n = o(M_n)$, $P(U > M_n/9) = o(1)$. On the other hand,

$$\begin{aligned} \left\|\frac{\tau_n^2}{\sigma_n^2+\tau_n^2} Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}\right\| &= \left\|\Sigma\left(\frac{\tau_n^2}{\sigma_n^2+\tau_n^2}\varepsilon + \frac{\sigma_n^2}{\sigma_n^2+\tau_n^2} F_0\right)\right\| \\ &\leq \|\Sigma\varepsilon\| + \frac{\sigma_n}{\sqrt{\sigma_n^2+\tau_n^2}}\|F_0\| \end{aligned}$$

Since $\|F_0\| = o(\tau_n^2/\sigma_n)$, $\frac{\sigma_n^2\|F_0\|^2}{\sigma_n^2+\tau_n^2} = o(1) < \frac{M_n}{9}$ for n large enough. $\|\Sigma\varepsilon\|^2$ is a $\chi^2(k_n)$ variable. Therefore, for n large enough,

$$E\left[W\left(\|F - F_{\langle\phi\rangle}\|^2 > \sigma_n^2 M_n \middle| Y\right)\right] \leq 2P(U > M_n/9) = o(1).$$

□

The frequentist assertion (3.12) is proved in a similar way from Taylor's expansion

$$\begin{aligned} G(Y_{\langle\phi\rangle}) - G(F_{\langle\phi\rangle}) = \dot{G}_{F_{\langle\phi\rangle}}(Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}) & \\ + \frac{1}{2}\int_0^1 (1-t) D^2_{F_{\langle\phi\rangle}+t\,(Y_{\langle\phi\rangle}-F_{\langle\phi\rangle})} G(Y_{\langle\phi\rangle} - F_{\langle\phi\rangle}, Y_{\langle\phi\rangle} - F_{\langle\phi\rangle})\,\mathrm{d}t. & \end{aligned}$$

Acknowledgment

The author would like to thank I. Castillo and E. Gassiat for valuable discussions and suggestions.

3.A Posterior Consistency

Here we prove Proposition 17. Lemma 14 allows us to suppose $F_0 = \Phi\theta_0$. Let U a random variable following the $\chi^2(k_n)$ distribution. Proceeding as in [17, 106], we introduce a test

$$T_n = \mathbb{1}_{(\theta_Y-\theta_0)^T\Phi^T\Phi\,(\theta_Y-\theta_0) > \sigma_n^2 M_n/4}. \tag{3.24}$$

Note that $ET_n = P(U > M_n/4) = o(1)$. Then

$$E\left[\widetilde{W}\left(\mathcal{E}^C_{\theta_0,\Phi}(M_n)\middle| Y\right)\right] \leq ET_n + E\left[(1-T_n)\widetilde{W}\left(\mathcal{E}^C_{\theta_0,\Phi}(M_n)\middle| Y\right)\right].$$

Next, let $(r_n)_{n\geq 1}$ be a sequence of positive numbers such that r_n goes to 0 and $-\ln(r_n) = o(M_n/k_n)$ as n goes to infinity. We replace the distribution P_{θ_0} by the mixture distribution $P^W_{\theta_0,r_n}$ with density

$$\mathrm{d}P^W_{\theta_0,r_n}(y) = \int_{\mathcal{E}_{\theta_0,\Phi}(r_n)} \mathrm{d}P_\theta(y)\widetilde{W}^{r_n}(\mathrm{d}\theta). \tag{3.25}$$

where $\widetilde{W}^{r_n}$ is the rescaled restriction of $\widetilde{W}$ to $\mathcal{E}_{\theta_0,\Phi}(r_n)$, as in (3.20). The following Lemma illustrates the link between P_{θ_0} and $P^W_{\theta_0,r_n}$.

Lemma 16. *Using the preceding notations,*

$$\left\|P^W_{\theta_0,r_n} - P_{\theta_0}\right\|_{\mathrm{TV}} \leq \sqrt{\frac{r_n}{2\pi}} = o(1).$$

Proof. We use convexity, and Lemma 11 since $P_\theta = \mathcal{N}(\Phi\theta, \sigma_n^2 I_n)$:

$$\left\|P^W_{\theta_0,r_n} - P_{\theta_0}\right\|_{\mathrm{TV}} \leq \sup_{\|\Phi h\|^2 \leq \sigma_n^2 r_n} \|P_{\theta_0+h} - P_{\theta_0}\|_{\mathrm{TV}} \leq \frac{\sqrt{r_n}}{\sqrt{2\pi}}.$$

□

At that point, the Bayes rule and the Fubini Theorem give

$$\begin{aligned}
&E^W_{\theta_0,r_n}\left[(1-T_n)\widetilde{W}\left(\mathcal{E}^C_{\theta_0,\Phi}(M_n)\middle|\, Y\right)\right] \\
&= \frac{1}{\widetilde{W}(\mathcal{E}(r_n))}\int_{\mathcal{E}(r_n)}\left(\int_{\mathbb{R}^n}\left[(1-T_n)\int_{\mathcal{E}^C(M_n)}\frac{\mathrm{d}P_\tau(Y)w(\tau)\,\mathrm{d}\tau}{\int_{\mathbb{R}^{k_n}}\mathrm{d}P_\eta(Y)w(\eta)\,\mathrm{d}\eta}\right]\mathrm{d}P_\theta(Y)\right)w(\theta)\,\mathrm{d}\theta \\
&= \frac{1}{\widetilde{W}(\mathcal{E}(r_n))}\int_{\mathcal{E}^C(M_n)} E_\tau\left[(1-T_n)\widetilde{W}(\mathcal{E}(r_n)|Y)\right]w(\tau)\,\mathrm{d}\tau \\
&\leq \frac{1}{\widetilde{W}\left(\mathcal{E}_{\theta_0,\Phi}(r_n)\right)}\sup_{\|\Phi h\|^2>\sigma_n^2 M_n} E_{\theta_0+h}(1-T_n) \\
&\leq \frac{1}{\widetilde{W}\left(\mathcal{E}_{\theta_0,\Phi}(r_n)\right)}\sup_{\|\Phi h\|^2>\sigma_n^2 M_n} P_{\theta_0+h}\left(\|\Phi\left(\theta_Y-\theta_0-h\right)\|^2 > \sigma_n^2 M_n/4\right) \\
&= \frac{P(U>M_n/4)}{\widetilde{W}\left(\mathcal{E}_{\theta_0,\Phi}(r_n)\right)}.
\end{aligned}$$

Let $B_k(0,1)$ be the unit ball in $\mathbb{R}^k$. We make use of the following relation (see for instance [15, Lemma 2])

$$-\ln \mathrm{Vol}\left(B_k(0,1)\right) = \ln\frac{\Gamma(1+k/2)}{\pi^{k/2}} \underset{k\to\infty}{\sim} \frac{k}{2}\ln k$$

together with a control on the volume of the ellipsoid $\mathcal{E}_{\theta_0,\Phi}(r_n)$

$$\widetilde{W}\left(\mathcal{E}_{\theta_0,\Phi}(r_n)\right) \geq \left(\inf_{\|\Phi h\|^2\leq\sigma_n^2 r_n}\frac{w(\theta_0+h)}{w(\theta_0)}\right)\frac{\sigma_n^{k_n}w(\theta_0)}{\sqrt{\det(\Phi^T\Phi)}}r_n^{k_n/2}\,\mathrm{Vol}\left(B_{k_n}(0,1)\right).$$

Next we can use Cirelson's inequality (3.22) as in Lemma 12 and get, for n large enough,

$$\begin{aligned}\ln\left(E^W_{\theta_0,r_n}\left[(1-T_n)\widetilde{W}\left(\mathcal{E}_{\theta_0,\Phi}(M_n)|\,Y\right)\right]\right)&\\ \leq \ln\left(\frac{\sqrt{\det(\Phi^T\Phi)}}{\sigma_n^{k_n}w(\theta_0)}\right)-\frac{M_n}{9}-\frac{k_n}{2}\ln(r_n)-\ln\mathrm{Vol}\left(B_{k_n}(0,1)\right)+o(1)&\\ \sim -\frac{M_n}{9}&\end{aligned}$$

which goes to minus infinity as n goes to infinity.

3.B Sobolev classes

We begin with a simple lemma, then we prove Lemma 9 and Lemma 10

Lemma 17. *Let $\alpha > 1/2$, $L > 0$, and $\theta \in \Theta(\alpha, L)$. Then*

$$\sum_{j=1}^{\infty}|\theta_j| < \infty.$$

As a consequence, f is the uniform limit of the series $\sum_{j=1}^{\infty}\theta_j\varphi_j$ and f is continuous.

Proof of Lemma 17. We have a simple control on the sum of the coefficients

$$\sum_{j=2}^{\infty}|\theta_j| \leq \sqrt{\sum_{j\geq 2}a_j^{-2}}\sqrt{\sum_{j\geq 2}a_j^2\theta_j^2} \leq \frac{L}{\pi^\alpha}\sqrt{\sum_{j\geq 1}j^{-2\alpha}} < \infty.$$

Since all functions φ_j are continuous and bounded by $\sqrt{2}$, the other points follow. □

Proof of Lemma 9. $F_{\langle\phi\rangle}$ is the orthogonal projection of F_0 on the convex span of the first k_n vectors of the orthogonal basis $(\phi_j)_{1\leq j\leq n}$ of $\mathbb{R}^n$. So

$$\|F_0 - F_{\langle\phi\rangle}\|^2 = \sum_{j=k_n+1}^{n}(F_0^T\phi_j)^2 = n\sum_{j=k_n+1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}f(i/n)\varphi_j(i/n)\right)^2.$$

Following [105], we set $\zeta_j = \frac{1}{n}\sum_{i=1}^{n}f(i/n)\varphi_j(i/n) - \theta_j^0$ for $1 \leq j \leq n$. Then

$$\|F_0 - F_{\langle\phi\rangle}\|^2 = n\sum_{j=k_n+1}^{n}\left(\zeta_j+\theta_j^0\right)^2 \leq 2n\left(\sum_{j=1}^{n}\zeta_j^2 + \sum_{j=k_n+1}^{n}(\theta_j^0)^2\right).$$

Using Lemma 8, for any $1 \leq j \leq n$,

$$\begin{aligned}\zeta_j &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{m=1}^{\infty}\theta_m^0\varphi_m(i/n)\right)\varphi_j(i/n) - \theta_j^0\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{m=n+1}^{\infty}\theta_m^0\varphi_m(i/n)\right)\varphi_j(i/n).\end{aligned}$$

So, using Lemma 8 again,

$$\sum_{j=1}^{n} \zeta_j^2 \leq \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{m=n+1}^{\infty} \theta_m^0 \varphi_m(i/n) \right)^2 .$$

We recognize a Riemann sum of the function $\left(\sum_{m=n+1}^{\infty} \theta_m^0 \varphi_m\right)^2$, which is continuous according to Lemma 17. Therefore

$$\sum_{j=1}^{n} \zeta_j^2 \leq (1+o(1)) \int_0^1 \left(\sum_{m=n+1}^{\infty} \theta_m^0 \varphi_m \right)^2 = \sum_{m=n+1}^{\infty} (\theta_m^0)^2$$

and

$$\begin{aligned} \|F_0 - F_{\langle\phi\rangle}\|^2 &\leq (2n+o(n)) \sum_{m=k_n+1}^{\infty} (\theta_m^0)^2 \\ &\leq \frac{2n+o(n)}{a_{k_n+1}^2} \sum_{m=k_n+1}^{\infty} a_m^2 (\theta_m^0)^2 \\ &\leq \frac{2L^2}{\pi^{2\alpha}} \frac{n+o(n)}{k_n^{2\alpha}}. \end{aligned}$$

On the other hand, f is continuous, and $(1/n)\,\|F_0\|^2$ is a Riemann sum of f^2. Therefore $(1/n)\,\|F_0\|^2$ goes to $\int_0^1 f^2$ as n goes to infinity. □

Proof of Lemma 10. Let $(\theta'_j)_{j\geq 1}$ the Fourier coefficients of g. As in the previous proof, we set $\zeta_j = \frac{1}{n}\sum_{i=1}^{n} f(i/n)\varphi_j(i/n) - \theta_j^0$ and $\zeta'_j = \frac{1}{n}\sum_{i=1}^{n} g(i/n)\varphi_j(i/n) - \theta'_j$ for $1 \leq j \leq n$. We have $F_0 = \sum_{j=1}^{n} (\zeta_j + \theta_j^0)\phi_j$, so

$$\frac{1}{n}\sum_{i=1}^{n} f(i/n)g(i/n) = \sum_{j=1}^{n} (\zeta_j + \theta_j^0)(\zeta'_j + \theta'_j).$$

In the meantime

$$\int_0^1 fg = \sum_{j=1}^{\infty} \theta_j^0 \theta'_j.$$

So

$$\begin{aligned} \left| \frac{1}{n}\sum_{i=1}^{n} f(i/n)g(i/n) - \int_0^1 fg \right| &= \left| \sum_{j=1}^{n} \zeta_j\zeta'_j + \sum_{j=1}^{n} \zeta_j\theta'_j + \sum_{j=1}^{n} \theta_j^0\zeta'_j - \sum_{j=n+1}^{\infty} \theta_j^0\theta'_j \right| \\ &\leq \sqrt{\sum_{j=1}^{n} \zeta_j^2} \sqrt{\sum_{j=1}^{n} \zeta_j'^2} + \sqrt{\sum_{j=1}^{n} \zeta_j^2} \sqrt{\sum_{j=1}^{n} \theta_j'^2} \\ &\quad + \sqrt{\sum_{j=1}^{n} \zeta_j'^2} \sqrt{\sum_{j=1}^{n} (\theta_j^0)^2} + \sqrt{\sum_{j=n+1}^{\infty} (\theta_j^0)^2} \sqrt{\sum_{j=n+1}^{\infty} \theta_j'^2}. \end{aligned}$$

As in the proof of Lemma 9, we have

$$\sum_{j=1}^{n} \zeta_j^2 \sim \sum_{j=n+1}^{\infty} (\theta_j^0)^2 \leq \frac{L^2}{\pi^{2\alpha} n^{2\alpha}}$$

and on the other hand,

$$\sum_{j=1}^{n}(\theta_j^0)^2 \leq \int_0^1 f^2 = \|f\|^2.$$

Thus

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} f(i/n)g(i/n) - \int_0^1 fg\right| &\leq (1+o(1))\left(\frac{LL'}{\pi^{\alpha+\alpha'}n^{\alpha+\alpha'}} + \frac{L'\|f\|}{\pi^{\alpha'}n^{\alpha'}} + \frac{L\|g\|}{\pi^{\alpha}n^{\alpha}}\right)\\ &= O\left(n^{-\inf(\alpha,\alpha')}\right).\end{aligned}$$

□

Chapitre 4

Mélange de lois multinomiales multivariées : un critère pénalisé pour la sélection de variable et le clustering

Multivariate multinomial mixtures: a data-driven penalized criterion for variable selection and clustering

Abstract

We consider the problem of estimating the number of components and the relevant variables in a multivariate multinomial mixture. This kind of models arise in particular when dealing with multilocus genotypic data. A new penalized maximum likelihood criterion is proposed, and a non-asymptotic oracle inequality is obtained. Further, under weak assumptions on the true probability underlying the observations, the selected model is asymptotically consistent. On a practical aspect, the shape of our proposed penalty function is defined up to a multiplicative parameter which is calibrated thanks to the slope heuristics, in an automatic data-driven procedure. Using simulated data, we found that this procedure improves the performances of the selection procedure with respect to classical criteria such as **BIC** and **AIC**. The new criterion gives an answer to the question "Which criterion for which sample size?".

Keywords: Biostatistics, Model selection, Multilocus genotypic data, Multivariate multinomial mixture, Latent class model, Penalized Likelihood, Population genetics, Slope heuristics, Variables selection.

Sommaire

4.1 Introduction

This article is concerned with the unsupervised classification on categorical multivariate data. The model-based clustering, which uses finite mixture models, is an intuitive and rigorous framework for the unsupervised classification. However there is no clear consensus on the way to gather individuals in general: on the basis of well separated clusters, or on the basis of the components of the mixture distribution? We refer to [10] for a general discussion on this topic. Finite mixture models are specially adapted when each class is supposed to be characterized by a set of parameters, for instance in population genetics: in this case the populations that the biologists look for are characterized by their allelic frequencies and a genetic equilibrium; this corresponds to the notion of population as a reproduction unit, or a group of individuals sharing the same genetic structure. Finite mixture models are also known in the literature as the latent class models.

The observations are n independent realizations of a random vector, whose number L of coordinates (variables) may be large. The individuals of the sample are clustered into a certain unknown number K of populations on the basis of the frequencies of apparition of the possible states of each variable. It may happen that only a subset S of the variables are relevant for clustering purposes, and the others are just noise. Thus, in addition to the number K of populations and the frequencies of the different states, we are also interested in the subset S, which may have significance in the interpretation of the results.

A number of clustering methods for categorical multivariate data have been proposed in recent years in the context of genomics (see [23, 28, 84]). But the problem of variable selection for clustering using such data was first addressed in [104], where the question is regarded as a model selection problem in a density estimation framework. First the components of a finite mixture distribution are identified, then the individuals are clustered into these components using the Maximum A Posteriori (MAP) method. Using simulated data, that article shows that the variable selection procedure based on the Bayesian Information Criterion (**BIC**) significantly improves clustering and prediction capacities in our framework. It also gives a theoretical consistency result: when the true density P_0 underlying the observations belongs to one of the competing models, then there exists a smallest model $\mathcal{M}_{(K_0,\, S_0)}$ containing P_0; further, the **BIC** type criteria select $\mathcal{M}_{(K_0,\, S_0)}$ with probability tending to one as the sample size n goes to infinity. This consistency approach requires large sample sizes which may be difficult to obtain. However the knowledge of the true model, aside the frequencies of the states, is an important information for the interpretation of the results.

In the present paper we adopt an oracle approach. We do not aim at choosing the true model underlying the data, even if our procedure performs well also for that. The criteria are rather designed to minimize some risk function of the estimated density with respect to the true density. In this context simpler models can be preferred to $\mathcal{M}_{(K_0,\, S_0)}$, in which too many parameters can entail estimators which overfit the data. Actually there is no need to assume that P_0 belongs to one of the competing models $\mathcal{M}_{(K,S)}$.

BIC relies on a strong asymptotic assumption, and can thus require large sample sizes to reach its asymptotic behavior; practically **BIC** is known to overpenalize, and therefore selects too small models for small or medium values of n (see [80]). On the

contrary Akaike's Information Criterion (**AIC**) is known to underpenalize, and selects too large models for large and medium values of n. We would like a criterion which gathers the virtues of both **AIC** and **BIC**, and performs well for different values of n.

In this article, we propose a non asymptotic penalized criterion based on the metric entropy theory of Massart (in particular [75]). It leads to a non asymptotic oracle inequality, which compares the risk of the selected estimator to the risk of the estimator associated with the unknown best model (see Theorem 9 below). There exists a large literature on model selection via penalization from a non asymptotic perspective. This literature is still in development with the appearance of sophisticated tools of probability such as concentration and deviations inequalities (see [75] and the references therein). In mixture models the non asymptotic approach is very recent, the first related work being [77] for the Gaussian mixture model.

However, the obtained penalty function presents drawbacks: it depends on a multiplicative constant for which sharp upper bounds are not available, and it leads in practice to an overpenalization — even worse than **BIC**. Therefore our theoretical result mainly suggests the shape of the penalty function:

$$\mathbf{pen}_n(m) = \lambda D_m/n,$$

where D_m is the dimension of model m, and λ an unknown parameter depending on the sample size and the complexity of the collection of models under competition, which has to be calibrated. A calibration of λ with the so-called slope heuristics has been proposed in [13] in such a case. We propose a modified version based on a sliding window of this calibration method. The resulting criterion does not require an ad-hoc choice of the penalty parameters and adapts automatically to the data. Although the full theoretical validation of slope heuristics is provided only in the Gaussian homoscedastic and heteroscedastic regression frameworks [3,13], they have been implemented in several other frameworks (see [74, 76, 111, 112] for applications in density estimation, genomics, etc.). The simulations performed in Subsection 4.4.3 illustrate that our criterion behaves well with respect to more classical criteria as **BIC** and **AIC**, both to estimate the density, even when n is relatively small, and to retrieve the true model. It can be seen as a representative of the family of the General Information Criteria (see for instance [5] whose criterion is less intuitive but presents some analogy with the slope heuristics).

The paper is organized as follows. Section 4.2 is devoted to the presentation of the mixture models framework and to the model selection paradigm. In Section 4.3 we state and prove our main result, the oracle inequality. Section 4.4 is devoted to the practical aspect of our procedure which has been implemented in the stand alone software `MixMoGenD` (Mixture Model using Genotypic Data) (see [104]). Results on simulated experiments are also presented: we compare our proposed criterion to classical **BIC** and **AIC**, in both points of view of the selection of the true model and of density estimation. Eventually, the Appendices contain several technical results used in the main analysis.

4.2 Model and methods

4.2.1 Framework

We suppose we deal with independent and identically distributed (iid) realizations of a multivariate random vector $X = (X^l)_{1 \leq l \leq L}$. We consider two main settings:

1. Each X^l is a multinomial variable taking values in $\{1, \ldots, A_l\}$.
2. Each X^l consists in a (non ordered) set $\left\{X^{l,1}, X^{l,2}\right\}$ of two (that may be equal) qualitative variables taking their values in the same set $\{1, \ldots, A_l\}$.

All along this article, these two settings will be referred to as Case 1 and Case 2. In both cases, the numbers A_l of allowed states are supposed to be known, and to verify $A_l \geq 2$.

The first case is a usual latent class model with various applications (psychomatrics, marketing, credit scoring, genomics, etc.), while the last one is more specific to genotypic data. In this context $X = (X^l)_{1 \leq l \leq L}$ represents the genotype of an individual at L loci of its DNA. Case 1 corresponds to haploid organisms, with a single representative of each chromosome; at any locus l a single allele X^l is measured. Case 2 corresponds to diploid organisms, with two representatives of each chromosome; at any locus l, two alleles $X^{l,1}$ and $X^{l,2}$ are observed together.

We consider a model-based clustering, which means that the sample is a finite mixture of an unknown number K of populations (clusters), each being characterized by a set of frequencies of the states. Let denote by Z the (unobserved) population an individual comes from. Variable Z takes its values in the set $\{1, \ldots, K\}$ of the labels of the different clusters. Its distribution is given by the vector $\pi = (\pi_k)_{1 \leq k \leq K}$, where $\pi_k = P(Z = k)$. Conditionally to Z, the variables $X^1, \ldots, X^L$ are supposed to be independent. In Case 2, the states $X^{l,1}$ and $X^{l,2}$ for the l^{th} variable are also supposed to be independent conditionally to Z. The preceding two assumptions are what biologists respectively call *Linkage Equilibrium* (LE) and *Hardy-Weinberg Equilibrium* (HWE). According to these assumptions, the probability distribution of a genotype $x = \left(x^l\right)_{1 \leq l \leq L}$ in a population k is given in the following equations

$$\begin{aligned} P\left(x \middle| \, Z = k\right) &= \prod_{l=1}^{L} P\left(x^l | Z = k\right) \\ \text{Case 1: } P\left(x^l | Z = k\right) &= \alpha_{k,l,x^l} \\ \text{Case 2: } P\left(x^l | Z = k\right) &= \left(2 - \mathbb{1}_{x^{l,1} = x^{l,2}}\right) \alpha_{k,l,x^{l,1}} \alpha_{k,l,x^{l,2}} \end{aligned} \tag{4.1}$$

where $\alpha_{k,l,j}$ is the probability of state j associated to variable X^l in population k. The mixing proportions π_k and the probabilities $\alpha_{k,l,j}$ will be treated as parameters.

In the context of genomics, Hardy-Weinberg and linkage equilibria are based on several simplifying assumptions that can seem unrealistic; however they have still proven to be useful in describing many population genetic attributes and serve as a base model in the development of more realistic models of microevolution. Further, the choice of estimators derived from the maximum likelihood estimator (MLE) responds to the wish of biologists to group the sample into clusters minimizing the Hardy-Weinberg and linkage desequilibria, and this brings some robustness to our modeling (see [73] and references therein).

Going deeper, the oracle approach emphasizes that we should often prefer simplified and misspecified models. This introduces a modeling bias in order to get more robust estimators and classifiers, and at the end we get a smaller estimation error. This legitimizes also the following simplification.

It may happen that the structure of interest is contained in only a subset S of the L available variables, the others been useless or even harmful to detect a reasonable clustering into statistically different populations. For the variables in S, the frequencies of the states in at least two populations are different: we will call them clustering variables. For the other variables, the states are supposed to be equally distributed across the clusters. This approximation is theoretically justified by the oracle heuristics, which is able to take advantage of the misspecification; the simulations performed in [104] illustrate its benefits.
We denote by $\beta_{l,j}$ the frequency of state j associated to variable X^l in the whole population:

$$\beta_{l,j} = \alpha_{1,l,j} = \cdots = \alpha_{k,l,j} \cdots = \alpha_{K,l,j} \text{ for any } l \notin S \text{ and } 1 \leq j \leq A_l.$$

Obviously, $S = \emptyset$ if $K = 1$, otherwise S belongs to $\mathcal{P}^*(L)$, the set of all non empty subsets of $\{1, \ldots, L\}$.

Summarizing all these assumptions, we can write down the likelihood of an observation $x = \left(x^l\right)_{1 \leq l \leq L}$:

$$\begin{aligned}
\text{Case 1: } P_{(K,S,\theta)}(x) &= \left[\sum_{k=1}^{K} \pi_k \prod_{l \in S} \alpha_{k,l,x^l}\right] \times \prod_{l \notin S} \beta_{l,x^l} \\
\text{Case 2: } P_{(K,S,\theta)}(x) &= \left[\sum_{k=1}^{K} \pi_k \prod_{l \in S} \left(2 - \mathbb{1}_{x^{l,1}=x^{l,2}}\right) \alpha_{k,l,x^{l,1}} \times \alpha_{k,l,x^{l,2}}\right] \\
&\quad \times \prod_{l \notin S} \left(2 - \mathbb{1}_{x^{l,1}=x^{l,2}}\right) \beta_{l,x^{l,1}} \beta_{l,x^{l,2}}
\end{aligned} \tag{4.2}$$

where $\theta = (\pi, \alpha, \beta)$ is a multidimensional parameter, with

$$\begin{aligned}
\alpha &= (\alpha_{k,l,j})_{1 \leq k \leq K;\ l \in S;\ 1 \leq j \leq A_l} \\
\beta &= (\beta_{l,j})_{l \notin S;\ 1 \leq j \leq A_l}.
\end{aligned}$$

For a given K and S, $\theta = \theta_{(K,S)}$ ranges in the set

$$\Theta_{(K,S)} = \mathbb{S}_{K-1} \times \left[\prod_{l \in S} \mathbb{S}_{A_l - 1}\right]^K \times \prod_{l \notin S} \mathbb{S}_{A_l - 1}, \tag{4.3}$$

where $\mathbb{S}_{r-1} = \left\{p = (p_1,\ p_2, \ldots,\ p_r) \in [0,\ 1]^r : \sum_{j=1}^{r} p_j = 1\right\}$ is the $(r-1)$-dimensional simplex.

Then we consider the collection of all parametric models

$$\mathcal{M}_{(K,S)} = \left\{P_{(K,S,\theta)} : \theta \in \Theta_{(K,S)}\right\} \tag{4.4}$$

with $(K, S) \in \mathbb{M} := \{(1, \emptyset)\} \cup (\mathbb{N} \backslash \{0, 1\}) \times \mathcal{P}^*(L)$. To alleviate notations, we will often use the single index $m \in \mathbb{M}$ instead of (K, S).

Each model $\mathcal{M}_{(K,S)}$ corresponds to a particular structure situation with K clusters and a subset S of clustering variables. Inferring K and S becomes a model selection problem in a density estimation framework. It also leads to a data clustering, via the estimation $\widehat{\theta}$ of the parameter $\theta_{(K,S)}$ and the prediction of the class z of an observation x by the MAP method:

$$\widehat{z} = \underset{1\leq k\leq K}{\arg\max}\, P_{(K,S,\widehat{\theta})}\left(Z = k | X = x\right).$$

4.2.2 Model selection via penalization

A common method to solve model selection problems consists in the minimization of a penalized maximum likelihood criterion. In each model $\mathcal{M}_{(K,S)}$, consider the maximum likelihood estimator (MLE) $\widehat{P}_{(K,S)} = P_{(K,S,\widehat{\theta})}$, which minimizes the log-likelihood contrast

$$\gamma_n\left(P\right) = -\frac{1}{n}\sum_{i=1}^{n} \ln P\left(X_i\right) \tag{4.5}$$

where X_i describes the individual i in the sample. Then a data driven selected model $\mathcal{M}_{\left(\widehat{K}_n,\widehat{S}_n\right)}$ is chosen, where $\left(\widehat{K}_n, \widehat{S}_n\right)$ minimizes a penalized maximum likelihood criterion of the form

$$\mathbf{crit}(K,S) = \gamma_n\big(\widehat{P}_{(K,S)}\big) + \mathbf{pen}_n(K,S),$$

where $\mathbf{pen}_n : \mathbb{M} \to \mathbb{R}_+$ is the penalty function. Eventually the selected estimator is $\widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}$.

The penalty function is designed to avoid overfit problems. Classical penalties, such as the ones used in **AIC** and **BIC** criteria, are based on the dimension of the model. In the following, we will refer to the number of free parameters

$$D_{(K,S)} = K - 1 + K\sum_{l\in S}\left(A_l - 1\right) + \sum_{l\notin S}\left(A_l - 1\right) \tag{4.6}$$

as the dimension of the model $\mathcal{M}_{(K,S)}$. The penalty functions of **AIC** and **BIC** are respectively defined by

$$\begin{aligned} \mathbf{pen_{AIC}}\left(m\right) &= \frac{1}{n}\cdot D_m; \\ \mathbf{pen_{BIC}}\left(m\right) &= \frac{\ln n}{2n}\cdot D_m. \end{aligned} \tag{4.7}$$

Our work is centered on the MLE estimator $\widehat{P}_{(K,\ S)}$, but this last one presents a drawback. For the sake of density estimation, we would like to use the Kullback-Leibler divergence **KL** as a risk function to measure the quality of an estimator. Unfortunately, when an state is not present in the sample, the MLE estimator assigns to it a zero probability. As a consequence, the Kullback risk $E_{P_0}\left[\mathbf{KL}\left(P_0,\ \widehat{P}_{(K,\ S)}\right)\right]$ is infinite.

The Hellinger distance offers an alternative to the Kullback-Leibler divergence. Let us consider two probability distribution P and Q, admitting respectively s and t as

density functions with respect to a common σ-finite measure μ. We call Hellinger distance between P and Q the quantity $\mathbf{h}(P,Q)$ defined by

$$\mathbf{h}(P,Q)^2 = \int \left(\sqrt{s(x)} - \sqrt{t(x)}\right)^2 d\mu(x). \tag{4.8}$$

Let $(K^*,\ S^*)$ be a minimizer in $(K,\ S)$ of the Hellinger risk of the MLE estimator

$$R_{(K,\ S)} = E_{P_0}\left[\mathbf{h}^2\left(P_0,\ \widehat{P}_{(K,\ S)}\right)\right]. \tag{4.9}$$

The density $\widehat{P}_{(K^*,\ S^*)}$ is called oracle for the Hellinger risk. It is not an estimator, since it depends on the true density P_0. However it can be used as a benchmark to quantify the quality of our model selection procedure: in the simulation performed in paragraph 4.4.3, we compare the Hellinger risk of the selected estimator $\widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}$ to the oracle risk.

4.3 New criteria and non asymptotic risk bounds

4.3.1 Main result

Our main theorem provides an oracle inequality for both Case 1 and Case 2. It links the Hellinger risk of the selected estimator to the Kullback-Leibler divergence **KL** between the true density and each model in the models collection. Unlike **KL** which is not a metric, the Hellinger distance **h** permits to take advantage of the metric properties (metric entropy) of the models.

Theorem 9. *We consider the collection $\mathbb{M}$ of models defined above, and a corresponding collection of ρ-MLEs $\left(\widehat{P}_{(K,S)}\right)_{(K,S)\in\mathbb{M}}$, which means that for every $(K,S)\in\mathbb{M}$*

$$\gamma_n\left(\widehat{P}_{(K,S)}\right) \leq \inf_{Q\in\mathcal{M}_{(K,S)}} \gamma_n(Q) + \rho.$$

Let $A_{\max} = \sup_{1\leq l\leq L} A_l$, and let ξ be defined by $\xi = \dfrac{4\sqrt{A_{\max}}\sqrt{L}}{2^{L+1}-1}$ in Case 1 and $\xi = \dfrac{4\sqrt{A_{\max}}\sqrt{L}}{2(1+3\sqrt{2})^L - 1}$ in Case 2. Assume that $\xi < 1$ or $n > \xi^2 K$.

There exists absolute constants κ and C such that whenever

$$\mathbf{pen}_n(K,S) \geq \kappa\left(5 + \sqrt{\max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right)}\right)^2 \frac{D_{(K,S)}}{n} \tag{4.10}$$

for every $(K,S)\in\mathbb{M}$, then the model $\mathcal{M}_{\left(\widehat{K}_n,\widehat{S}_n\right)}$ where $\left(\widehat{K}_n,\widehat{S}_n\right)$ minimizes

$$\mathbf{crit}(K,S) = \gamma_n\left(\widehat{P}_{(K,S)}\right) + \mathbf{pen}_n(K,S)$$

over $\mathbb{M}$ *exists and moreover, whatever the underlying probability* P_0*,*

$$E_{P_0}\left[\mathbf{h}^2\left(P_0, \widehat{P}_{(\widehat{K}_n,\widehat{S}_n)}\right)\right] \leq C\left(\inf_{(K,S)\in\mathbb{M}}\left(\mathbf{KL}\left(P_0, \mathcal{M}_{(K,S)}\right) + \mathbf{pen}_n(K,S)\right) + \rho + \frac{(3/4)^L}{n}\right)$$

where, for every $(K,S)\in\mathbb{M}$, $\mathbf{KL}\left(P_0, \mathcal{M}_{(K,S)}\right) = \inf_{Q\in\mathcal{M}_{(K,S)}} \mathbf{KL}(P_0, Q)$.

The condition $\xi < 1$ is used in the proof to avoid more complicated calculations. In practice, ξ is very likely to be smaller than 1 for L not too small.

Note that as soon as $n \geq 2L$, (4.10) is simplified in the following way

$$\mathbf{pen}_n(K,S) \geq \kappa\left(5 + \sqrt{\frac{1}{2}\ln n + \frac{1}{2}\ln L}\right)^2 \frac{D_{(K,S)}}{n}.$$

The leading term for large n is $\kappa\frac{\ln n}{2}\frac{D_{(K,S)}}{n}$, which is a multiple of the penalty function of **BIC**. As a consequence, we can apply Theorem 2 from [104]: when the underlying distribution P_0 belongs to one of the competing models, the smallest model (K_0, S_0) containing P_0 is selected with probability tending to 1 as n goes to infinity.

Such a penalty is not surprising in our context: it is in fact very similar to the one obtained in [77] in a Gaussian mixture framework.

Sharp estimates of κ are not available. Theorem 9 is too conservative in practice, and leads to an over-penalized criterion which is outperformed by smaller penalties. So it is mainly used to suggest the shape of the penalty function

$$\mathbf{pen}_n(K,S) = \lambda\frac{D_{(K,S)}}{n} \tag{4.11}$$

where λ is a parameter to be chosen depending on n and the collection $\mathbb{M}$ — but not on (K,S). Slope heuristics [3, 13] can be used in practice to calibrate λ: this is done in Section 4.4, where we use change-point detection [74] in relation to slope heuristics.

Since $\mathbf{h}^2$ is upper bounded by 2, the non-asymptotic feature of Theorem 9 is interesting when n is large enough with respect to $D_{(K,S)}$. However, even with small values of n, the simulations performed in Subsection 4.4.3 show that the penalized criterion calibrated using the slope heuristics keep good behaviors.

4.3.2 A general tool for model selection

Theorem 9 is obtained from [75, Theorem 7.11]. This last result deals with model selection problems by proposing penalty functions related to geometrical properties of the models, namely metric entropy with bracketing for Hellinger distance.

The framework here is the following. We consider some measurable space $(A, \mathcal{A})$, and μ a σ-finite positive measure on A. A collection of models $(\mathcal{M}_m)_{m\in\mathbb{M}}$ is given, where each model $\mathcal{M}_m$ is a set of probability density functions s with respect to μ.

The following relation permits us to extend the definition of $\mathbf{h}$ to positive functions s or t whose integral is finite but not necessary 1. Denoting $\sqrt{s}$ the function defined by $\sqrt{s}(x) = \sqrt{s(x)}$, and by $\|\cdot\|_2$ the usual norm in $\mathbb{L}^2(\mu)$, then

$$\mathbf{h}(s,t) = \|\sqrt{s} - \sqrt{t}\|_2.$$

Let us now recall the definition of metric entropy with bracketing. Consider some collection F of measurable functions on A, and d one of the following metrics on F: $\mathbf{h}$, $\|\cdot\|_1$, or $\|\cdot\|_2$. A bracket $[l,u]$ is the collection of all measurable functions f such that $l \leq f \leq u$. Its d-diameter is the distance $d(u,l)$. Then, for every positive number ε, we denote by $N_{[.]}(\varepsilon, F, d)$ the minimal number of brackets with d-diameter not larger than ε which are needed to cover F. The d-entropy with bracketing of F is defined as the logarithm of $N_{[.]}(\varepsilon, F, d)$, and is denoted by $H_{[.]}(\varepsilon, F, d)$.

We assume that for each model $\mathcal{M}_m$ the square entropy with bracketing $\sqrt{H_{[.]}(\varepsilon, \mathcal{M}_m, \mathbf{h})}$ is integrable at 0. Let us consider some function ϕ_m on $\mathbb{R}_+$ with the following properties

(I). ϕ_m is nondecreasing, $x \mapsto \phi_m(x)/x$ is non-increasing on $(0, +\infty)$ and for every $\sigma \in \mathbb{R}_+$ and every $u \in \mathcal{M}_m$

$$\int_0^\sigma \sqrt{H_{[.]}\left(x, S_m(u,\sigma), \mathbf{h}\right)}\mathrm{d}x \leq \phi_m(\sigma),$$

where $S_m(u,\sigma) = \left\{t \in \mathcal{M}_m : \|\sqrt{t} - \sqrt{u}\|_2 \leq \sigma\right\}$.

(I) is verified in particular with $\phi_m(\sigma) = \int_0^\sigma \sqrt{H_{[.]}\left(x, \mathcal{M}_m, \mathbf{h}\right)}\mathrm{d}x$.

In order to avoid measurability problems, we suppose that for each $m \in \mathbb{M}$, the following separability condition is verified for $\mathcal{M}_m$:

(M). There exists some countable subset $\mathcal{M}'_m$ of $\mathcal{M}_m$ and a set $A' \subset A$ with $\mu(A') = \mu(A)$ such that for every $t \in \mathcal{M}_m$, there exists some sequence $(t_k)_{k\geq 1}$ of elements of $\mathcal{M}'_m$ such that for every $x \in A'$, $\ln(t_k(x))$ tends to $\ln(t(x))$ as k tends to infinity.

Theorem 10. *Let $X_1, \ldots, X_n$ be iid random variables with unknown density s with respect to some positive measure μ. Let $\{\mathcal{M}_m\}_{m\in\mathbb{M}}$ be some at most countable collection of models, each fulfilling* (M). *We consider a corresponding collection of ρ-MLEs $(\widehat{s}_m)_{m\in\mathbb{M}}$. Let $\{x_m\}_{m\in\mathbb{M}}$ be some family of nonnegative numbers such that*

$$\sum_{m\in\mathbb{M}} e^{-x_m} = \Sigma < \infty,$$

and for every $m \in \mathbb{M}$ considering ϕ_m with property (i) define σ_m as the unique positive solution of the equation

$$\phi_m(\sigma) = \sqrt{n}\sigma^2. \tag{4.12}$$

Let $\mathbf{pen}_n : \mathbb{M} \to \mathbb{R}_+$ and consider the penalized log-likelihood criterion

$$\mathbf{crit}(m) = \gamma_n\left(\widehat{s}_m\right) + \mathbf{pen}_n(m).$$

Then, there exists some absolute constants κ and C such that whenever

$$\mathbf{pen}_n(m) \geq \kappa\left(\sigma_m^2 + \frac{x_m}{n}\right) \text{ for every } m \in \mathbb{M},$$

some random variable $\widehat{m}$ *minimizing* **crit** *over* $\mathbb{M}$ *exists and moreover, whatever the density* s

$$E_s\left[\mathbf{h}^2\left(s,\widehat{s}_{\widehat{m}}\right)\right] \leq C\left(\inf_{m\in\mathbb{M}}\left(\mathbf{KL}\left(s,\mathcal{M}_m\right)+\mathbf{pen}_n(m)\right)+\rho+\frac{\Sigma}{n}\right).$$

In Theorem 10, σ_m^2 has the role of a variance term of $\widehat{s}_m$, while the weights x_m take into account the number of models m having the same dimension.

4.3.3 Proof of Theorem 9

In order to apply Theorem 10, we need to compute the metric entropy with bracketing of each model $\mathcal{M}_{(K,S)}$. This is done in the following result, which is proved in Appendix 4.A.

Proposition 18 (Bracketing entropy of a model)**.** *Let* $\eta:\mathbb{R}_+\to\mathbb{R}_+$ *be the increasing convex function defined by*

$$\begin{aligned}&\textit{Case 1: } \eta(\varepsilon)=(1+\varepsilon)^{L+1}-1,\\ &\textit{Case 2: } \eta(\varepsilon)=(1+\varepsilon)\left(1+\sqrt{2}\,\varepsilon\,(2+\varepsilon)\right)^{L}-1.\end{aligned}$$

For any choice of K *and* S, $\mathcal{M}_{(K,S)}$ *fulfills* (M)*. For any* $\varepsilon\in(0,1)$,

$$H_{[\cdot]}\left(\eta(\varepsilon),\mathcal{M}_{(K,S)},\mathbf{h}\right)\leq D_{(K,S)}\ln\left(\frac{1}{\varepsilon}\right)+C_{(K,S)},$$

where

$$\begin{aligned}C_{(K,S)}=\frac{1}{2}\Bigg(&\ln(2\pi e)D_{(K,S)}+\ln(4\pi e)\left(\mathbb{1}_{K\geq 2}+L+(K-1)|S|\right)\\&+\mathbb{1}_{K\geq 2}\ln(K+1)+\sum_{l=1}^{L}\ln(A_l+1)+(K-1)\sum_{l\in S}\ln(A_l+1).\Bigg)\end{aligned}\tag{4.13}$$

$C_{(K,S)}$ is a technical quantity measuring the complexity of a model $\mathcal{M}_{(K,S)}$.

In the next step we establish an expression for ϕ_m. All following results are proved in Appendix 4.B.

Proposition 19. *For any choice of* $m=(K,S)$*, the function* ϕ_m *defined on* $(0,\eta(1)]$ *by*

$$\phi_m(\sigma)=\left(2\sqrt{\ln 2}\,\sqrt{D_{(K,S)}}+\sqrt{C_{(K,S)}-D_{(K,S)}\ln\eta^{-1}(\sigma)}\right)\sigma$$

fulfills (I).

We do not define ϕ_m for σ bigger than $\eta(1)$, to avoid more complicated expressions. This is why a condition on ξ appears in the following lemma:

Lemma 18. *Let* $A_{\max} = \sup_{1\leq l\leq L} A_l$, $\xi = \dfrac{4\sqrt{A_{\max}}\sqrt{L}}{2^{L+1}-1}$ *in Case 1, and* $\xi = \dfrac{4\sqrt{A_{\max}}\sqrt{L}}{2(1+3\sqrt{2})^L-1}$ *in Case 2. Then, for all* $n \geq 1$ *if* $\xi < 1$, *and for* $n > \xi^2 K$ *otherwise, the solution* σ_m *of (4.12) verifies* $\sigma_m < \eta(1)$.

From Proposition 19 we can deduce an upper bound for σ_m, with a similar reasoning to [77]. First, $\sigma_m \leq \eta(1)$ entails $\eta^{-1}(\sigma_m) \leq 1$, and we obtain the lower bound $\sigma_m \geq \widetilde{\sigma}_m$, where

$$\widetilde{\sigma}_m = \frac{1}{\sqrt{n}}\left(2\sqrt{\ln 2}\,\sqrt{D_m} + \sqrt{C_m}\right). \tag{4.14}$$

This can be used to get an upper bound

$$\sigma_m \leq \frac{1}{\sqrt{n}}\left(2\sqrt{\ln 2}\,\sqrt{D_m} + \sqrt{C_m - D_m \ln \eta^{-1}(\widetilde{\sigma}_m)}\right). \tag{4.15}$$

Let us now choose the weights x_m. If we take something bigger than $n\sigma_m^2$, this will change the shape of the penalty in Theorem 10. We define

$$x_m = (\ln 2) D_m.$$

The following Lemma shows that this choice is suitable.

Lemma 19. *For any model* $\mathcal{M}_m$, *with* $m \in \mathbb{M}$ *as above, let us set* $x_m = (\ln 2)D_m$. *Then*

$$\sum_{m\in\mathbb{M}} e^{-x_m} \leq (3/4)^L.$$

To express the penalty function we have to lower bound $\eta^{-1}(\widetilde{\sigma}_m)$. This is done in the following Lemma.

Lemma 20. *Using the preceding notations,*

$$\sigma_m^2 + \frac{x_m}{n} \leq \frac{D_{(K,S)}}{n}\left(5 + \sqrt{\max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right)}\right)^2.$$

This ends the proof of Theorem 9.

4.4 In practice

In real datasets the numbers A_l of possible states at each variable X^l are not necessarily known. The numbers $\widehat{A}_l$ of observed states can be used instead. In fact, the MLE estimator select a density with null weight on non-observed states. Then, in each model $\mathcal{M}_{(K,S)}$, an approximated MLE estimator can be computed thanks to the Expectation-Maximization (EM) algorithm (see [36]).

The other two points that have to be done before reaching the final estimator $\widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}$ are the choice of the penalty function, and the sub-collection of models on which the EM algorithm will be used. These two points are discussed in Subsections 4.4.1 and 4.4.2. Then simulations are presented in Subsection 4.4.3.

4.4.1 Slope heuristics and Dimension jump

Theorem 9 suggests to take a penalty function of the shape (4.11), defined modulo a multiplicative parameter λ which has to be calibrated. Slope heuristics, as presented in [3,13], provide a practical method to find an optimal penalty $\mathbf{pen}_{\text{opt}}(m) = \lambda_{\text{opt}} D_m/n$. These heuristics are based on the conjecture that there exists a minimal penalty $\mathbf{pen}_{\text{min}}(m) = \lambda_{\text{min}} D_m/n$ required for the model selection procedure to work: when the penalty is smaller that $\mathbf{pen}_{\text{min}}$, the selected model is one of the most complex models, and the risk of the selected estimator is large. On the contrary, when the penalty is larger than $\mathbf{pen}_{\text{min}}$, the complexity of the selected model is much smaller. Then the optimal penalty is close to twice the minimal penalty:

$$\mathbf{pen}_{\text{opt}}(m) \approx 2\lambda_{\text{min}} D_m/n.$$

The name "slope heuristics" comes from λ_{min} being the slope of the linear regression $\gamma_n\left(\widehat{P}_m\right) \sim D_m/n$ for a certain sub-collection of the most competing models m. For example, on the left panel of Figure 4.1 below, a slope is visible for the models containing the true model $\mathcal{M}_{(K_0,\, S_0)}$. Even if this example is favorable and mainly here for illustration purposes, it shows that the slope heuristics are sensible with the modelings of the present work.

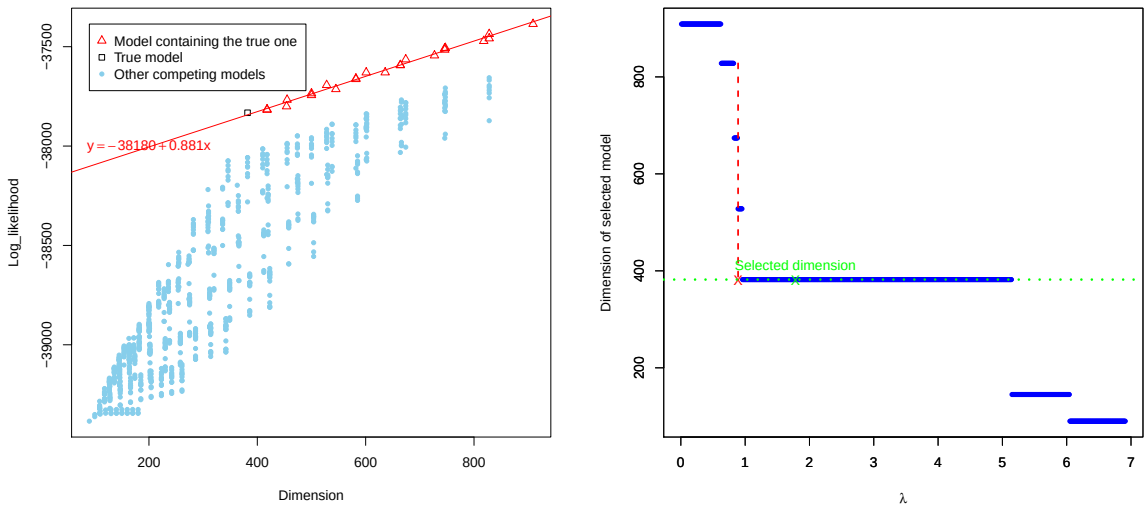

Figure 4.1: Two ways to compute the slope, on a simulated sample of 1000 individuals, with 8 clustering variables among 10, and 5 populations. Models have been explored via the modified backward-stepwise described in subsection 4.4.2, the number K of clusters varying from 1 to 10. The size of the sliding window is 0.15.

Instead of estimating λ_{min} by linear regression, another method is jump detection. Suppose we have at hand a reasonable grid $\lambda_1 < \ldots < \lambda_r$ of candidate values of λ_{min}, and a sub-collection $\mathbb{M}_{explored}$ of the most competitive models. Each λ_i leads to a selected model $\widehat{m}_i$ with dimension $D_{\widehat{m}_i}$. If you plot $D_{\widehat{m}_i}$ as a function of λ_i, λ_{min} is expected to lie at the position of the biggest jump. However, the right panel of Figure 4.1 illustrates

an important point: in that example the biggest jump is at $\lambda \approx 5.1$, but $\lambda_{\min}$ is around 0.9, which corresponds to several successive jumps. We propose an improved version of the dimension jump method of [3], based on a sliding window: we consider at a time all jumps in an window of $h \geq 1$ following intervals in the grid. Algorithm 1 below describes the procedure.

Algorithm 1 Calibration of Penalty$\left(\mathbb{M}_{explored}, (\lambda_i)_{i=1,\dots,r}, h\right)$

for $i = 1$ to n_λ **do**

$\widehat{m}_i \leftarrow \underset{m \in \mathbb{M}_{explored}}{\arg\min} \left\{ \gamma_n\left(\widehat{P}_m\right) + \lambda_i D_m / n \right\}$

end for

$i_{end} \leftarrow \min \underset{i \in \{h+1,\dots,r\}}{\arg\max} \left\{ D_{\widehat{m}_{i-h}} - D_{\widehat{m}_i} \right\}$

$i_{init} \leftarrow \max \left\{ j \in [i_{end} - h, i_{end} - 1], D_{\widehat{m}_j} - D_{\widehat{m}_{i_{end}}} = D_{\widehat{m}_{i_{end}-h}} - D_{\widehat{m}_{i_{end}}} \right\}$

$\widehat{\lambda}_{\min} \leftarrow \dfrac{\lambda_{i_{init}} + \lambda_{i_{end}}}{2}$

return $\widehat{\lambda}_{\min}$

4.4.2 Sub-collection of models for calibration

For a given maximum value $K_{\max}$ of the number of clusters, the number of models under competition is equal to $1 + (K_{\max} - 1) * \left(2^L - 1\right)$. Since this number is huge in most situations, it is very painful to consider all competing models for calibration of the parameter λ. On the other hand, we need enough models to ensure that there is a clear jump in the sequence of selected dimension. We consider the modified backward-stepwise algorithm proposed in [104], which explores of cardinalities of S. It enables to gather the most competitive models among all possible S for a given number K of clusters and a given penalty function $\mathbf{pen}_n$. It gives also the choice to add a complementary exploration step based on a similarly modified forward strategy. We will refer to this algorithm as *explorer* $(K,\ \mathbf{pen}_n)$.

Since we do not know the final penalty during the exploration step, we consider a reasonable grid $1/2 = \lambda_1 < \ldots < \lambda_r = \ln n$ containing both penalty functions associated to **AIC** and **BIC** (4.7). To each value λ_i of the grid is associated a penalty function $\mathbf{pen}_{\lambda_i}$. We launch *explorer* $\left(K,\ \mathbf{pen}_{\lambda_i}\right)$ for all values of K in $\{1, \ldots,\ K_{\max}\}$ and for all values of λ_i of the above grid, and we gather the explored models in $\mathbb{M}_{explored}$. This sub-collection seemly contains the most competitive models and it is then used to calibrate λ.

4.4.3 Numerical experiments

Our proposed procedure with a data-driven calibration of the penalty function has been implemented for Case 2 in the software `MixMoGenD` (Mixture Model using Genotypic Data), which already proposed a selection procedure based on asymptotic criteria **BIC** and **AIC** (see [104]). Here, we conduct numerical experiments on simulated datasets

for performances assessment of the new non asymptotic criterion with respect to **BIC** and **AIC**.

We present two experiments, both in Case 2. The first one considers the consistency of the selected model: we study how the procedure retrieves the main features of the true model as the number of individuals in the datasets increases. In the second one, we are rather interested in a validation of the model selection procedure from the oracle point of view: we compare the Hellinger risk of the selected estimator to the oracle risk.

Consistency performances

In this experiment we consider a setting with $L = 10$ variables of 10 states each. We chose a parameter with $K_0 = 5$ populations of equal probability. The frequencies of the states have been chosen such that the genetic differentiation between the populations is decreasing with the variables rank. In the first 6 variables, the populations are more separated. In the following 2 variables, the populations are very poorly differentiated. In the last 2 variables, the states follow the same uniform distribution in all populations. The whole parameter is available at `http://www.math.u-psud.fr/~toussile/`.

We considered different values n of the sample size in [50, 900] and for each of them, 10 datasets have been simulated. The results are summarized in Figure 4.2 and Table 4.4.3. The left panel in Figure 4.2 gives the proportion of selecting the subset $\widehat{S}_n$ of clustering variables containing the first 6 variables, which are the most genetically differentiated variables. The right panel gives the proportion of selected models with $\widehat{K}_n = K_0$. Table 4.4.3 gives further details on the selected number of components $\widehat{K}_n$.

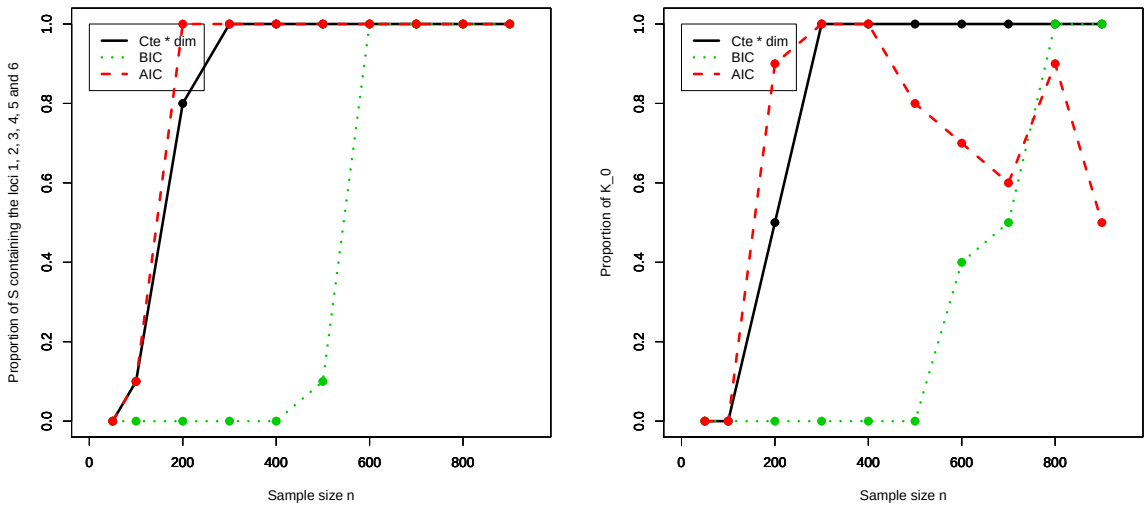

Figure 4.2: The figure in the left panel gives the proportion of selected models with $\widehat{S}_n \supseteq \{1, \ldots, 6\}$, and the one in the right gives the proportion of selected models with $\widehat{K}_n = K_0$, versus the sample size.

In this experiment, **AIC** seems to be the best criterion for variable selection; however the different between **AIC** and the new criterion is not significant. It also appears that

		$\widehat{K}_n$							
n	**crit**	1	2	3	4	5	6	7	8
50	Cte*dim	0	10	0	0	0	0	0	0
	AIC	0	10	0	0	0	0	0	0
	BIC	10	0	0	0	0	0	0	0
100	Cte*dim	0	2	8	0	0	0	0	0
	AIC	0	0	9	0	0	0	0	0
	BIC	10	0	0	0	0	0	0	0
200	Cte*dim	0	0	3	2	5	0	0	0
	AIC	0	0	0	1	9	0	0	0
	BIC	10	0	0	0	0	0	0	0
300	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	10	0	0	0
	BIC	0	9	1	0	0	0	0	0
400	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	10	0	0	0
	BIC	0	9	1	0	0	0	0	0
500	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	8	1	1	0
	BIC	0	5	4	1	0	0	0	0
600	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	7	3	0	0
	BIC	0	0	0	0	5	5	0	0
700	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	6	2	2	0
	BIC	0	0	0	0	10	0	0	0
800	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	9	1	0	0
	BIC	0	0	0	0	10	0	0	0

Table 4.1: Selection of the number of populations using different criteria: AIC, BIC and Cte*dim (for the criterion we propose, with automatic data-driven calibration of the penalty function).

AIC estimates the number of clusters better than the other criteria for small sample sizes (around $n = 100$ and $n = 200$), but it overestimates this number from $n = 500$. On the contrary, the new criterion perfectly estimates the number of clusters for sample sizes ≥ 300. **BIC** performs poorly for both variables selection and classification on datasets with small sizes. As expected, the data-driven calibration of the penalty function improves globally the performances of the selection procedure, and it gives thus an answer to the question "Which penalty for which sample size?".

It may happen that the results obtained on small sample sizes change a little from one run to another. In fact, the EM algorithm can miss the global maximum on such sample sizes, in particular in models of higher dimension. In our experiments, it is probably the case with some datasets of size $n \leq 300$, when the number of free parameters in the simulated model is ≥ 310.

Oracle performances of the estimator

Since the new criterion is designed in an oracle perspective, it is interesting to compare the associated estimator to the oracle for Hellinger risk. Recall that the oracle is the estimator associated to the model indexed by the minimizer (K^*, S^*) of the risk

$E\left[\mathbf{h}^2\left(P_0,\ \widehat{P}_{(K,\ S)}\right)\right]$ over the collection of models $\mathbb{M}$.

In this experiment, we consider simulated datasets with reduced variability in order to reduce the computation time. The parameter underlying the data admits $L = 6$ variables, 3 states for each variable, and $K_0 = 3$ populations with equal probability. The frequencies of the states have been chosen in such a way that the genetic differentiation between the population is significant on the first 3 variables, very small on the 4^{th} and 5^{th} variables, while the states of the 6^{th} variable follow the uniform distribution in all populations. Thus the true model is defined by $K_0 = 3$ and $S_0 = \{1,\ 2,\ 3,\ 4,\ 5\}$. The whole parameter is available at `http://www.math.u-psud.fr/~toussile/`.

We estimated the oracle using a Monte Carlo procedure on 100 simulated datasets of size 500 each, and got $\widehat{K^*} = 3$ and $\widehat{S^*} = \{1,\ 2,\ 3,\ 4\}$. The results we obtained are summarized in Figure 4.3 and Table 4.2.

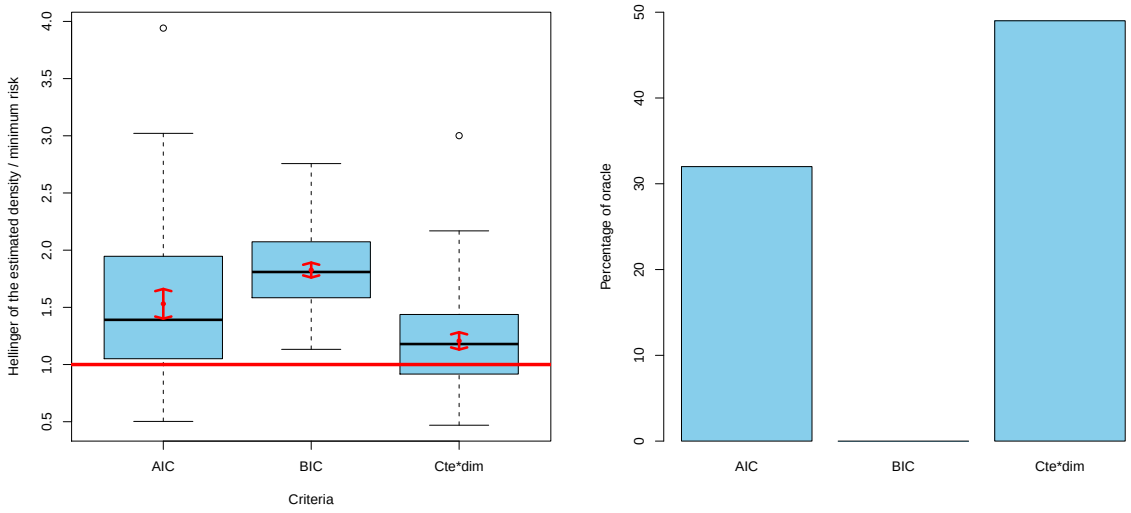

Figure 4.3: The left panel gives the boxplots, means and their 95% confident intervals, for $\dfrac{\mathbf{h}^2\left(P_0,\ \widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}\right)}{\mathbf{h}^2\left(P_0,\ \widehat{P}_{(\widehat{K^*},\ \widehat{S^*})}\right)}$; the right panel gives the percentages of selection of the estimated oracle $\left(\widehat{K^*},\ \widehat{S^*}\right)$; three criteria have been used: **AIC**, **BIC**, and **Cte*Dim** which denotes the new criterion with data-driven calibration of the penalty function.

	AIC	**BIC**
AIC	-	$< 5.40e-05$
Cte*Dim	$< 2.02e-05$	$< 2.20e-16$

Table 4.2: The p-values of pairwise student tests comparing the means of the $\mathbf{h}^2\left(P_0,\ \widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}\right)$. The alternative hypothesis is that the mean of the Hellinger distance associated to the criterion in the first column is less than the one associated to the criterion in the first line.

The worst behavior comes from **BIC** and it is not a surprise for two main reasons.

First **BIC** is designed to find the true model which is different to the oracle in our experiments. Second, it is based on asymptotic approximation and therefore requires large samples. In contrary, compared to **AIC** and **BIC**, the new criterion with data-driven calibration of the penalty function is significantly the best in the sense of Hellinger risk and the capacity of selecting the oracle. Recall that both **AIC** and the new criterion are designed to find the oracle (see Table 4.2). But like **BIC**, **AIC** is based on asymptotic approximations. So the advantage of the new criterion over **AIC** is probably that it is designed in a non asymptotic perspective.

4.5 Conclusion

In this paper, we have considered a model selection via penalization, which performs simultaneously a variables selection and a detection of the number of populations, in the specific framework of multivariate multinomial mixture. This leads to a clustering in a second time. Our main result provides an oracle inequality, under the condition of some lower bound on the penalty function. The weakness of such a result is that the associated penalized criterion is not directly usable. Nevertheless, it suggests a shape of the penalty function which is of the form $\mathbf{pen}_n(m) = \lambda D_m/n$, where $\lambda = \lambda(n,\ \mathbb{M})$ is a parameter which depends on the data and the collection of the competing models. In practice λ is calibrated via the slope heuristics.

In the simulated experiments we conducted, the new criterion with penalty calibration shows good behaviors for density estimation as well as for the selection of the true model. It also performs well both when the number of individuals is large and when it is reasonnably small. This gives an answer to the question "Which criterion for with sample size?"

In the modeling we considered, the model dimension grows rapidly. In real experiments the number of individuals can be small, so other modeling with reduced dimension may be needed. We currently work on models which cluster the populations differently for each variable, as well as models which allocate the same probability to several states.

Acknowledgment

The authors gratefully acknowledge the comments and advice of Elisabeth Gassiat, Pascal Massart, and Gilles Celeux. Many thanks also to Nathalie Akakpo, Nicolas Verzelen, and Cathy Maugis, for the useful discussion we had.

4.A Metric entropy with bracketing

We first state several results about the entropy with bracketing, which will be used to prove Proposition 18. They are mainly adapted from [51], but several are improved or writen here in a more general form. These lemmas can be seen as a toolbox to calculate the metric entropy with bracketing of complex models from the metric entropy of simpler elements.

We consider a measurable space $(A, \mathcal{A})$, and μ a σ-finite positive measure on A. We consider a model $\mathcal{M}$, which is a set of probability density functions with respect to μ. All functions considered in the following will be positive functions in $\mathbb{L}^1(\mu)$.

Lemma 21. *Let $\varepsilon > 0$. Let $[l, u]$ be a bracket in $\mathbb{L}^1(\mu)$, with* **h***-diameter less than ε, and containing s, a probability density function with respect to μ. Then*

$$\int l \, \mathrm{d}\mu \leq 1 \leq \int u \, \mathrm{d}\mu \leq (1+\varepsilon)^2.$$

Proof. First two inequalities are immediate, from $l \leq s \leq u$. For the last one, we use triangle inequality in $\mathbb{L}^2(\mu)$, and the definition of **h**:

$$\begin{aligned}\int u \, \mathrm{d}\mu &= \int \left(\sqrt{l} + \left(\sqrt{u} - \sqrt{l}\right)\right)^2 \mathrm{d}\mu \\ &\leq \left(\sqrt{\int l \, \mathrm{d}\mu} + \mathbf{h}(u, l)\right)^2 \\ &\leq (1+\varepsilon)^2.\end{aligned}$$

□

Lemma 22 (Bracketing entropy of product densities)**.** *Let $n \geq 2$, and consider a collection $(A_i, \mathcal{A}_i, \mu_i)_{1 \leq i \leq n}$ of measured space. For any $1 \leq i \leq n$, let $\mathcal{M}_i$ be a collection of probability density functions on A_i fulfilling* (M)*. Consider the product model*

$$\mathcal{M} = \left\{ s = \otimes_{i=1}^n s_i ; \forall 1 \leq i \leq n, s_i \in \mathcal{M}_i \right\}.$$

$\mathcal{M}$ contains density functions on $A = \prod_{i=1}^n A_i$ with respect to $\mu = \otimes_{i=1}^n \mu_i$.

$\mathcal{M}$ fulfills (M) *and, for any sequence of positive numbers $(\delta_i)_{1 \leq i \leq n}$, if $\varepsilon \geq \prod_{i=1}^n (1 + \delta_i) - 1$ then*

$$H_{[\cdot]}(\varepsilon, \mathcal{M}, \mathbf{h}) \leq \sum_{i=1}^n H_{[\cdot]}(\delta_i, \mathcal{M}_i, \mathbf{h}).$$

Proof. Let us consider some $s = \otimes_{i=1}^n s_i$ in $\mathcal{M}$. For $1 \leq i \leq n$, let $\mathcal{M}_i'$, A_i' and a sequence $(t_{i,k})_{k \geq 1}$ be such as needed for $\mathcal{M}_i$ to verify (M). Then, with the choice $t_k = \otimes_{i=1}^n t_{i,k}$ and $A' = \prod_{i=1}^n A_i'$, (M) is true for $\mathcal{M}$ too.

Let $\delta > 0$. For any $1 \leq i \leq n$, let $[l_i, u_i]$ a bracket containing s_i, with **h**-diameter less than δ_i. Let us set $l = \otimes_{i=1}^n l_i$, and $u = \otimes_{i=1}^n u_i$. Then s belongs to bracket $[l, u]$. We can compute its **h**-diameter:

$$\begin{aligned}\mathbf{h}(l, u) &= \sqrt{\int_A \left(\sum_{j=1}^n \left(\prod_{i=1}^{j-1} \sqrt{l_i} \prod_{i=j}^n \sqrt{u_i} - \prod_{i=1}^{j} \sqrt{l_i} \prod_{i=j+1}^n \sqrt{u_i} \right) \right)^2 \mathrm{d}\mu} \\ &\leq \sum_{j=1}^n \prod_{i=1}^{j-1} \sqrt{\int_{A_i} l_i \, \mathrm{d}\mu_i} \prod_{i=j+1}^n \sqrt{\int_{A_i} u_i \, \mathrm{d}\mu_i} \; \mathbf{h}(l_j, u_j) \\ &\leq \sum_{j=1}^n \delta_j \prod_{i=j+1}^n (1 + \delta_i) = \prod_{j=1}^n (1 + \delta_j) - 1\end{aligned}$$

thanks to triangle inequality and Lemma 21 (empty products equal 1).

Let $\varepsilon \geq \prod_{i=1}^{n}(1+\delta_i)-1$. For any $1 \leq i \leq n$ consider a minimal covering of $\mathcal{M}_i$ with brackets of $\mathbf{h}$-diameter less than δ_i. With the previous process we can build a covering of $\mathcal{M}$ with brackets of $\mathbf{h}$-diameter less than ε. So the minimal cardinality of such a covering verifies

$$N_{[\cdot]}(\varepsilon, \mathcal{M}, \mathbf{h}) \leq \prod_{i=1}^{n} N_{[\cdot]}(\delta_i, \mathcal{M}_i, \mathbf{h}).$$

□

Lemma 23 (Bracketing entropy of mixture densities)**.** *Let $n \geq 2$, and for any $1 \leq i \leq n$, let $\mathcal{M}_i$ be a set of probability density functions, all on the same measured space $(A, \mathcal{A}, \mu)$ and fulfilling* (M)*. Let us consider the set of all mixture densities*

$$\mathcal{M} = \left\{ \sum_{i=1}^{n} \pi_i s_i : \pi = (\pi_i)_{1 \leq i \leq n} \in \mathbb{S}_{n-1}; \forall 1 \leq i \leq n, s_i \in \mathcal{M}_i \right\}.$$

Then $\mathcal{M}$ fulfills (M)*, and for any $\delta > 0$, $\eta > 0$, and $\varepsilon \geq \delta + \eta + \delta\eta$,*

$$H_{[\cdot]}(\varepsilon, \mathcal{M}, \mathbf{h}) \leq H_{[\cdot]}(\delta, \mathbb{S}_{n-1}, \mathbf{h}) + \sum_{i=1}^{n} H_{[\cdot]}(\eta, \mathcal{M}_i, \mathbf{h}).$$

Proof. First, let us note that $\mathbb{S}_{n-1}$ is separable for its usual topology. Then, checking that $\mathcal{M}$ fulfills (M) is easy, and we do not explicit it.

We do not develop either the proof of the last relation, because it is exactly the same as in [51, proof of Theorem 2]. Let us just say that at the end we get, using our Lemma 21 instead of [51, Lemma 3],

$$\begin{aligned} \mathbf{h}^2(l,u) &\leq \eta^2 (1+\delta)^2 + \delta^2 + 2\eta\,\delta\,(1+\delta) \\ &\leq \varepsilon^2. \end{aligned}$$

□

Next result is just Lemma 2 from [51]:

Lemma 24 (Bracketing entropy of the simplex)**.** *Let $n \geq 2$ be an integer. Let μ be the counting measure on $\{1, \ldots, n\}$. We identify any probability on $\{1, \ldots, n\}$ with its density $s \in \mathbb{S}_{n-1}$ with respect to μ. Then, if $0 < \delta \leq 1$,*

$$H_{[\cdot]}(\delta, \mathbb{S}_{n-1}, \mathbf{h}) \leq (n-1)\ln\left(\frac{1}{\delta}\right) + \frac{\ln 2 + \ln(n+1) + n\ln(2\pi e)}{2}.$$

To deal with Case 2, we also need the metric entropy of the collection of all Hardy-Weinberg genotype distributions for a given variable.

Lemma 25 (Bracketing entropy of Hardy-Weinberg genotype distributions)**.** *Suppose that, for some variable l, there exist $A_l \geq 2$ different states. Let Ω_l be the collection of all genotype distributions following Hardy-Weinberg model (4.1). Then Ω_l fulfills* (M)*, and for any $\delta > 0$ and $\varepsilon \geq \sqrt{2}\,\delta\,(2+\delta)$,*

$$H_{[\cdot]}(\varepsilon, \Omega_l, \mathbf{h}) \leq H_{[\cdot]}(\delta, \mathbb{S}_{A_l-1}, \mathbf{h}).$$

Proof. (4.1) permits to associate a parameter $\alpha = (\alpha_1, \dots, \alpha_{A_l}) \in \mathbb{S}_{A_l-1}$ to any density in Ω_l. More generally, for any $\alpha \in [0,1]^{A_l}$, we define a function

$$d_\alpha(x) = (2 - \mathbb{1}_{x_1=x_2})\, \alpha_{x_1} \alpha_{x_2}$$

on the set of all genotypes $x = \{x^1, x^2\}$ on A_l states. Consider some $\delta > 0$ and $d_\alpha \in \Omega_l$. Let $[l, u]$ be some bracket containing α, with $\mathbf{h}$-diameter less than δ. Then d_α belongs to the bracket $[d_l, d_u]$. Let us calculate its diameter.

$$\begin{aligned}
\mathbf{h}^2(d_l, d_u) &= \sum_{a=1}^{A_l} (u_a - l_a)^2 + \sum_{1 \le a < b \le A_l} \left(\sqrt{2u_a u_b} - \sqrt{2l_a l_b}\right)^2 \\
&\le 2 \sum_{a=1}^{A_l} \sum_{b=1}^{A_l} \left(\sqrt{u_a u_b} - \sqrt{u_a l_b} + \sqrt{u_a l_b} - \sqrt{l_a l_b}\right)^2 \\
&\le 2 \left(\sqrt{\sum_{a=1}^{A_l} u_a \sum_{b=1}^{A_l} \left(\sqrt{u_b} - \sqrt{l_b}\right)^2} + \sqrt{\sum_{a=1}^{A_l} \left(\sqrt{u_a} - \sqrt{l_a}\right)^2 \sum_{b=1}^{A_l} l_b} \right)^2 \\
&\le 2 \left((1+\delta)\, \delta + \delta \right)^2
\end{aligned}$$

using Lemma 21. So $\mathbf{h}(d_l, d_u) \le \sqrt{2}\, \delta\, (2 + \delta)$.

Let $(\alpha^{(k)})_{k \ge 1}$ a sequence of elements of $\mathbb{S}_{A_l-1} \cap \mathbb{Q}^{A_l}$, which tends to α for the usual topology as k tends to infinity. Then, for any genotype $x = \{x^1, x^2\}$, $\ln d_{\alpha^{(k)}}(x)$ tends to $\ln d_\alpha(x)$. Therefore Ω_l fulfills (M). □

Proof of Proposition 18. We build the proof for Case 2. For Case 1 everything is similar, with a simplification: we directly have $\mathbb{S}_{A_l-1}$ instead of Ω_l.

Using (4.2) we see that a probability $P_{(K,S)}\big(\,\cdot\,|\theta\big)$ is the product of a mixture density corresponding to the variables in S, and a product density in $\bigotimes_{l \notin S} \Omega_l$ for the other variables. Let us call $\mathcal{M}$ the collection of all mixtures of K densities in $\bigotimes_{l \in S} \Omega_l$.

We first deal with the non clustering variables. Using Lemma 22 and Lemma 25, $\bigotimes_{l \notin S} \Omega_l$ fulfills (M). For any $\varepsilon \in (0,1)$,

$$\begin{aligned}
H_{[\cdot]}\left((1 + 2\sqrt{2}\varepsilon + \sqrt{2}\varepsilon^2)^{L-|S|} - 1, \bigotimes_{l \notin S} \Omega_l, \mathbf{h} \right) &\le \sum_{l \notin S} H_{[\cdot]}\left(2\sqrt{2}\varepsilon + \sqrt{2}\varepsilon^2, \Omega_l, \mathbf{h}\right) \\
&\le \sum_{l \notin S} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right).
\end{aligned}$$

On the same way

$$H_{[\cdot]}\left((1 + 2\sqrt{2}\varepsilon + \sqrt{2}\varepsilon^2)^{|S|} - 1, \bigotimes_{l \in S} \Omega_l, \mathbf{h} \right) \le \sum_{l \in S} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right).$$

We can apply Lemma 23, and get that $\mathcal{M}$ fulfills (M) and

$$\begin{aligned}
H_{[\cdot]}\Big((1 + 2\sqrt{2}\varepsilon + \sqrt{2}\varepsilon^2)^{|S|}(1+\varepsilon) - 1, \mathcal{M}, \mathbf{h} \Big) & \\
\le \mathbb{1}_{K \ge 2} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{K-1}, \mathbf{h}\right) + K \sum_{l \in S} H_{[\cdot]}\left(\varepsilon, \mathbb{S}_{A_l-1}, \mathbf{h}\right).&
\end{aligned}$$

Lemma 22 again, applied to $\mathcal{M}$ and $\bigotimes_{l\notin S}\Omega_l$, gives that $\mathcal{M}_{(K,S)}$ fulfills (M), and for any $\varepsilon\in(0,1)$,

$$H_{[\cdot]}\left(\eta(\varepsilon),\mathcal{M}_{(K,S)},\mathbf{h}\right) \leq \mathbb{1}_{K\geq 2}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{K-1},\mathbf{h}\right)+K\sum_{l\in S}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{A_l-1},\mathbf{h}\right)+\sum_{l\notin S}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{A_l-1},\mathbf{h}\right).$$

At this point, it only remains to use Lemma 24 and to compute the constants. □

4.B Establishing the penalty

First, we need to establish some properties of function η.

Lemma 26 (Properties of function η)**.** *We consider the function η defined in Proposition 18, from $\mathbb{R}_+$ into $\mathbb{R}_+$. η is nonnegative, increasing and convex. $\eta(0)=0$, and $\eta'(0)=L+1$ in Case 1 while $\eta'(0)=2\sqrt{2}\,L+1$ in Case 2.*

Proof. The proof in Case 1 is immediate, so we develop only Case 2.

Setting $u(x)=1+2\sqrt{2}\,x+\sqrt{2}\,x^2$, we can write $\eta(x)=(1+x)\,u(x)^L-1$. Then, calculus gives

$$\eta'(x)=(2L+1)\,u(x)^L+2L\,(\sqrt{2}-1)\,u(x)^{L-1}.$$

Since u is positive on $(0,+\infty)$, η is increasing. But $\eta(0)=0$, so η is nonnegative on $\mathbb{R}_+$. We also have $\eta'(0)=2\sqrt{2}\,L+1$. Next,

$$\eta''(x)=2\sqrt{2}\,(1+x)\,\left((2L^2+L)\,u(x)^{L-1}+2L\,(L-1)\,(\sqrt{2}-1)\,u(x)^{L-2}\right)$$

which is positive on $\mathbb{R}_+$. □

Proof of Proposition 19. Let $0<\sigma\leq\eta(1)$, and $\delta=\eta^{-1}(\sigma)$. Then, for any $u\in\mathcal{M}_m$,

$$\begin{aligned}
&\int_0^\sigma \sqrt{H_{[\cdot]}\left(x,\mathcal{M}_m(u,\sigma),\mathbf{h}\right)}\mathrm{d}x\\
&\quad\leq\sum_{j=1}^{\infty}\int_{\eta(2^{-j}\delta)}^{\eta(2^{-j+1}\delta)}\sqrt{H_{[\cdot]}\left(x,\mathcal{M}_m,\mathbf{h}\right)}\mathrm{d}x\\
&\quad\leq\sum_{j=1}^{\infty}\left(\eta\left(2^{-j+1}\delta\right)-\eta\left(2^{-j}\delta\right)\right)\sqrt{C_m-D_m\ln\delta+D_m j\ln 2}\\
&\quad\leq\eta(\delta)\sqrt{C_m-D_m\ln\delta}\\
&\qquad+\sqrt{D_m\ln 2}\sum_{j=1}^{\infty}\sqrt{j}\,\left(\eta\left(2^{-j+1}\delta\right)-\eta\left(2^{-j}\delta\right)\right).
\end{aligned}$$

We deal with the last term of this sum in the following way:

$$\begin{aligned}\sum_{j=1}^{\infty}\sqrt{j}\,\left(\eta\left(2^{-j+1}\delta\right)-\eta\left(2^{-j}\delta\right)\right) &\leq \sum_{j=1}^{\infty} j\,\left(\eta\left(2^{-j+1}\delta\right)-\eta\left(2^{-j}\delta\right)\right)\\ &= \sum_{k=1}^{\infty}\eta\left(2^{-k+1}\delta\right)\\ &\leq \sum_{k=1}^{\infty}2^{-k+1}\eta(\delta)=2\sigma.\end{aligned}$$

So

$$\int_0^{\sigma}\sqrt{H_{[\cdot]}\left(x,\mathcal{M}_m(u,\sigma),\mathbf{h}\right)}\mathrm{d}x\leq\phi_m(\sigma).$$

Since η is increasing, $\phi_m(x)/x$ is decreasing. To check that ϕ_m is nondecreasing, it is enough to prove that function $f(x)=x\sqrt{b-\ln\eta^{-1}(x)}$ is nondecreasing on $(0,\eta(1)]$, where $b=\frac{C_m}{D_m}$. From (4.13), we get $C_m>\frac{\ln(2\pi e)}{2}D_m>D_m$, so $b>1$. Calculus gives

$$f'(x)=\sqrt{b-\ln\eta^{-1}(x)}-\frac{x}{2\eta^{-1}(x)\,\eta'\left(\eta^{-1}(x)\right)\sqrt{b-\ln\eta^{-1}(x)}}.$$

Let $y\in(0,1]$. η is convex on $(0,1]$, and that entails $\frac{\eta(y)}{y\,\eta'(y)}\leq 1$. Thus

$$\sqrt{b-\ln y}\,f'\left(\eta(y)\right)\geq b-\ln y-1/2>0.$$

□

Proof of Lemma 18. Since $\phi_m(x)/x$ is non-increasing, for any $\sigma>0$ such that $\sqrt{n}\,\sigma^2>\phi_m(\sigma)$, $\sigma>\sigma_m$. So, we look for situations such that $\sqrt{n}>\dfrac{\phi_m(\eta(1))}{\eta^2(1)}$.

For all $1\leq l\leq L$, $A_l\geq 2$. Since $\frac{1}{2}\ln(1+x)\leq x-1$ for $x\geq 2$, we get the following bounds

$$\frac{1+\ln(2\pi)}{2}D_m\leq C_m\leq\left(2+\ln(2\pi)+\frac{\ln 2}{2}\right)D_m. \tag{4.16}$$

Therefore

$$\frac{\phi_m(\eta(1))}{\eta^2(1)}<\frac{4\sqrt{D_m}}{\eta(1)}$$

On the other hand, we have

$$D_m\leq K\,L\,A_{\max}.$$

So, since $\phi_m(x)/x^2$ is decreasing, $\sigma_m<\eta(1)$ as soon as $n>\xi^2K$. This is true when $\xi<1$, since $K\leq n$: the number of clusters is not bigger than the number of individuals. □

Proof of Lemma 19. We define $\delta=1/2$, from which $e^{-x_m}=\delta^{D_m}$. If we consider the collection $\mathbb{M}$, we can discern two cases: $K=1$ and $S=\emptyset$, or $K\geq 2$ and $S\neq\emptyset$. So,

using (4.6),

$$\begin{aligned}\sum_{m\in\mathbb{M}} e^{-x_m} &= \delta^{\sum_{l=1}^{L}(A_l-1)}\left(1+\sum_{S\neq\emptyset}\sum_{K\geq 2}\left(\delta^{1+\sum_{l\in S}(A_l-1)}\right)^{K-1}\right)\\ &= \delta^{\sum_{l=1}^{L}(A_l-1)}\left(1+\sum_{S\neq\emptyset}\frac{\delta^{1+\sum_{l\in S}(A_l-1)}}{1-\delta^{1+\sum_{l\in S}(A_l-1)}}\right)\\ &\leq \delta^{L}\left(1+\frac{\delta}{1-\delta}\sum_{S\neq\emptyset}\delta^{|S|}\right)\\ &= \delta^{L}(1+\delta)^{L}.\end{aligned}$$

□

Proof of Lemma 20. η^{-1} in concave and nondecreasing, $\eta(0)=0$, so for any $0\leq x\leq \eta(1)$,

$$\eta^{-1}(x)\geq \frac{\eta^{-1}(2)}{2}\,\min(x,2).$$

On the other hand (4.14) and (4.16) entail

$$\widetilde{\sigma}_m \geq C_1\sqrt{\frac{D_m}{n}} \geq C_1\sqrt{\frac{L}{n}} \tag{4.17}$$

where $C_1 = 2\sqrt{\ln 2}+\sqrt{\frac{1+\ln(2\pi)}{2}} > 2\sqrt{2}$. Therefore

$$-\ln\eta^{-1}\left(\widetilde{\sigma}_m\right) \leq -\ln\left(\frac{\eta^{-1}(2)}{2}\right) - \ln 2 + \max\left(0, \frac{1}{2}\left(\ln n - \ln L - \ln 2\right)\right).$$

Consider Case 1. Since η is a convex function and $\eta'(0)=L+1$,

$$\eta^{-1}(2)\leq \frac{2}{L+1}.$$

Now,

$$\eta\left(\frac{2}{L+1}\right) = \left(1+\frac{2}{L+1}\right)^{L+1} - 1 \leq e^2-1.$$

Then

$$\frac{\eta^{-1}(2)}{2} \geq \frac{2/(L+1)}{\eta\left(2/(L+1)\right)} \geq \frac{2}{(e^2-1)(L+1)}.$$

Therefore

$$-\ln\left(\frac{\eta^{-1}(2)}{2}\right) \leq \ln(e^2-1) - \ln 2 + \ln L + \ln(3/2)$$

and

$$-\ln\eta^{-1}\left(\widetilde{\sigma}_m\right) \leq \ln(e^2-1) - \frac{7}{2}\ln 2 + \ln 3 + \max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right).$$

Using now (4.15), we get

$$\begin{aligned}\sigma_m^2 + \frac{x_m}{n} &\le \frac{D_m}{n}\left(\frac{1}{2} + \left(2\sqrt{\ln 2} + \sqrt{2 + \ln(2\pi) + \frac{\ln 2}{2} - \ln \eta^{-1}\left(\widetilde{\sigma}_m\right)}\right)^2\right) \\ &\le \frac{D_m}{n}\left(\frac{1}{\sqrt{2}} + 2\sqrt{\ln 2} + \sqrt{2 + \ln(2\pi) - 3\ln 2 + \ln 3 + \ln(e^2 - 1)}\right. \\ &\qquad \left. + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2 \\ &\le \frac{D_m}{n}\left(5 + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2 .\end{aligned}$$

Next, consider Case 2, and follow the same method. Then

$$\eta^{-1}(2) \le \frac{1}{\sqrt{2}\, L}$$

and

$$\eta\left(\frac{1}{\sqrt{2}\, L}\right) \le 2\exp\left(2 + \frac{1}{\sqrt{2}}\right).$$

This leads to

$$-\ln \eta^{-1}(x) \le 2 + \frac{1}{\sqrt{2}} + \frac{3\ln 2}{2} + \ln L - \ln \min(x, 2)$$

and

$$-\ln \eta^{-1}\left(\widetilde{\sigma}_m\right) \le 2 + \frac{1}{\sqrt{2}} + \max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right).$$

Now we obtain

$$\begin{aligned}\sigma_m^2 + \frac{x_m}{n} &\le \frac{D_m}{n}\left(\frac{1}{\sqrt{2}} + 2\sqrt{\ln 2} + \sqrt{4 + \ln(2\pi) + \frac{\sqrt{2} + \ln 2}{2}}\right. \\ &\qquad \left. + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2 \\ &\le \frac{D_m}{n}\left(5 + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2 .\end{aligned}$$

□

Annexe A

Un théorème de Bernstein-von Mises non-paramétrique pour les modèles exponentiels

A non-parametric Bernstein-Von Mises Theorem for exponential models

Abstract

In this chapter we obtain a non-parametric Bernstein-von Mises Theorem for increasing-dimensional exponential models. These results improve on the pioneer paper of Ghosal [53], which allows us to retrieve the results of Boucheron and Gassiat [17] as a special case. The work is still in progress in several directions: misspecified models, tools to check in practice the conditions on the Fisher information matrix, and application to functionals of the parameter.

Keywords: Nonparametric Bayesian Statistics, Bernstein-von Mises Theorem, exponential models, mean parameter.

Sommaire

A.1 Background

Let $\mathcal{X}$ be some measurable set (the sample space). Let $X_{1:n} = (X_1, \ldots, X_n)$ be a vector of independent variables identically distributed according to some probability distribution P_0 on $\mathcal{X}$.

A.1.1 Exponential Model and Mean Value Parameter

Let μ be a σ-finite measure on $\mathcal{X}$. Let $(k_n)_{n\geq 1}$ be some non-decreasing sequence of positive integers, with $p_n \to \infty$ as $n \to \infty$.

For each value $k = k_n$, let $t^k = (t_1^k, \ldots, t_k^k) : \mathcal{X} \to \mathbb{R}^k$ be a sequence of bounded functions such that, if $\mathbf{1}$ denotes the constant function with value 1 on $\mathcal{X}$, then $\mathbf{1}$, t_1^k, $\ldots$, t_k^k are linearly independent functions on the support of μ.

Let f^k be some positive, measurable function, which has a finite integral with respect to μ. For each value $k = k_n$, let

$$\psi^k(\theta) = \ln \int_{\mathcal{X}} \exp\left(\theta^T t^k(x)\right) f(x)\, \mathrm{d}\mu(x)$$

be defined on $\mathbb{R}^k$ (it is the natural parameter space, since t^k is bounded).

For simplicity, we will use the notation t instead of t^{k_n}, ψ instead of ψ^{k_n}, and so on. The same will be done with incoming notations.

We consider the exponential model $\Lambda_{\mathtt{can}} = \{P_{\theta,\mathtt{can}}\}_{\theta\in\Theta}$, where Θ is a connected open subset of $\mathbb{R}^{k_n}$, and $P_{\theta,\mathtt{can}}$ is the probability distribution on $\mathcal{X}$ admitting

$$f_{\theta,\mathtt{can}}(x) = f(x) \exp\left(\theta^T t(x) - \psi(\theta)\right)$$

as density function with respect to μ, where the exponent T denote the transposition operator. $E_{\theta,\mathtt{can}}[g(X)]$ and $E_{\theta,\mathtt{can}}[g(X_{1:n})]$ respectively denote the mean values of $g(X)$ and $g(X_{1:n})$ when the distribution of X is $P_{\theta,\mathtt{can}}$, and $X_{1:n}$ is a n-sample of the distribution $P_{\theta,\mathtt{can}}$. We also use the notation $P_{\theta,\mathtt{can}}(B)$ with the same meaning.

θ is called *canonical parameter* (or natural parameter) of the exponential family $\Lambda_{\mathtt{can}}$. The index "$\mathtt{can}$" in $P_{\theta,\mathtt{can}}$ emphasizes this point.

ψ is known to be analytic on $\mathbb{R}^{k_n}$ (see for instance [106, p. 38]), and

$$\dot{\psi}(\theta) = E_{\theta,\mathtt{can}}\, t(X)$$
$$\ddot{\psi}(\theta) = \mathrm{Cov}_{\theta,\mathtt{can}}\, t(X)$$

$\ddot{\psi}(\theta)$ is also the Fisher information matrix of the model $\Lambda_{\mathtt{can}}$. The fact that $\mathbf{1}$, t_1^k, $\ldots$, t_k^k are linearly independent functions on the support of μ entails that $\ddot{\psi}(\theta)$ is positive for any θ in $\mathbb{R}^{k_n}$, and consequently $\dot{\psi}$ is injective.

Let $\Pi = \dot{\psi}(\Theta)$ the image of Θ by $\dot{\psi}$. In the sequel, we consider the substitution $\pi = \dot{\psi}(\theta) = E_{\theta,\mathtt{can}}\, t(X)$. π is called *mean value parameter*. Equivalently, let $Q = \dot{\psi}^{-1}$ the inverse function; thus $\theta = Q(\pi)$. We now write P_π instead of $P_{\theta,\mathtt{can}}$, and so on. $\phi(\pi)$ will be used to denote $\psi \circ Q(\pi)$. The density of P_π is

$$f_\pi(x) = f(x) \exp\left(Q(\pi)^T t(x) - \phi(\pi)\right).$$

The Fisher information matrix of the new model $\Lambda = \{P_\pi\}_{\pi\in\Pi}$ is

$$I(\pi) = \ddot{\psi}^{-1}(\theta).$$

A.1.2 Maximum likelihood estimator

For any value of k_n, consider a parameter $\pi_* \in \Pi$, and suppose $X_{1:n}$ to be a a vector of independent variables identically distributed according to P_{π_*}. In other words, $P_0 = P_{\pi_*}$. Let $\theta_* = Q(\pi_*)$.

Consider the empirical mean

$$\overline{t(X)} = \frac{1}{n}\sum_{i=1}^{n} t(X_i).$$

$\overline{t(X)}$ will also be denoted by $\widehat{\pi}_n$. $\widehat{\pi}_n$ can be seen as an estimator of π_*: indeed, $\widehat{\pi}_n$ is linked to the usual maximum likelihood estimator (MLE) $\widehat{\theta}_n$ by the relation $\widehat{\theta}_n = Q(\widehat{\pi}_n)$. $P_{\widehat{\pi}_n}$ is expected to approximate P_{π_*}.
However $\widehat{\pi}_n$ may be outside Π: in this case the MLE does not exist.

In the sequel, we denote by Δ_n the normalized difference

$$\Delta_n = \sqrt{n}(\widehat{\pi}_n - \pi_*).$$

Then the central limit theorem gives $\Delta_n \to \mathcal{N}\left(0, \ddot{\psi}(\theta_*)\right)$ in distribution.

A.1.3 Prior and posterior distributions

Let $W = W_{k_n}$ be a probability distribution on Π, admitting $w(\cdot)$ as a density function with respect to the Lebesgue measure.

We consider the Bayesian setting where W is a prior distribution on the parameter π, and the conditional distribution of $X_{1:n}$ given π is the n-fold product measure $P_\pi^{\otimes n}$ (given π, $X_1, \ldots, X_n$ are iid and follow P_π).

The posterior distribution, or conditional distribution of π given $X_{1:n}$, is denoted by $P(\cdot|X_{1:n})$.

In the sequel, we denote by H_n the normalized difference

$$H_n = \sqrt{n}(\pi - \pi_*).$$

Its conditional (posterior) distribution given $X_{1:n}$ is denoted by $P_{H_n|X_{1:n}}$.

A Bernstein-Von Mises Theorem is a theorem stating the convergence (in distribution, in total variation, etc.) of the posterior distribution $P_{H_n|X_{1:n}}$ towards a Gaussian distribution.

A.2 A non-parametric Bernstein-Von Mises Theorem

A.2.1 Assumptions

Let $(M_n)_{n\geq 1}$ denote some unbounded, non-decreasing sequence on $(0,\infty)$. M_n will appear as a tool in our calculus.

Assumption 1.

$$k_n = o(M_n).$$

For $h \in \mathbb{R}^{k_n}$, let $\sigma_n^2(h) = h^T I(\pi_*)h$. For $M > 0$, let us consider the set $\mathcal{E}(M)$ defined by

$$\mathcal{E}_{n,\pi_*}(M) = \left\{h \in \mathbb{R}^{k_n} : \sigma_n^2(h) \leq M\right\}$$

We only consider interior points of the model:

Assumption 2. *For n large enough, $\mathcal{E}_{n,\pi_*}(M_n) \subset \Pi$.*

Our assumptions concern the model, which has to be smooth enough, and the prior distribution, which has to charge a neighborhood of π_* and to be flat enough on it.

Model hypotheses

For any $\pi \in \Pi$, let J_π be a square root of $\ddot{\psi}(\theta)$, *i.e.* a matrix such that $J_\pi J_\pi^T = \ddot{\psi}(\theta)$. Then, $(J_\pi^T)^{-1}$ is a square root of $I(\pi)$. Let $h = \sqrt{n}(\pi - \pi_*)$. For a $k \times k$-matrix A, let $\|A\|$ denote the operator norm associated with the usual euclidean norm $\|\cdot\|$ of $\mathbb{R}^k$. We need to control the deformation matrix $J_{\pi_*}^{-1}\ddot{\psi}(\theta)(J_{\pi_*}^T)^{-1}$ as a function of $\sigma_n(h)$:

Assumption 3. *There exists a constant C and a sequence $\varepsilon_n \to 0$ such that*

$$\sup_{h\in\mathcal{E}(M_n)} \|J_{\pi_*}^T I(\pi) J_{\pi_*}\| \leq 1 + \varepsilon_n$$

and, whenever $\sigma_n^2(h) \geq M_n$,

$$\left\| J_{\pi_*}^{-1}\ddot{\psi}(\theta)(J_{\pi_*}^T)^{-1} \right\| \leq C\frac{\sigma_n^2(h)}{M_n},$$

where I_{k_n} denotes the identity matrix of dimension $k_n \times k_n$.

Since two square roots of a positive definite matrix are orthogonal multiples of each other, $\|J_{\pi_*}^{-1}\ddot{\psi}(\theta)(J_{\pi_*}^T)^{-1}\|$ is correctly defined.

Next, we need to control the moments of t. In this paper, we only consider exponential models where t is bounded, with the following condition

Assumption 4.

$$M_n \|I(\pi_*)\| \sup_{x\in\mathcal{X}} \|t(x)\|^2 = o(n).$$

In the application to multinomial distributions paragraph A.4, we have $\sup_{x\in\mathcal{X}} \|t(x)\| = 1$, and Assumption 4 entails $n\|I(\pi_*)\|^{-1} \to \infty$, which corresponds to Conditions 2.1 and 3.1 of [17].

Our last assumptions concern the moments of order 3 and 4 of the distributions of the exponential model in a neighborhood of π_*.

Assumption 5. *For $h \in \mathcal{E}(M)$ and $u \in (0,1)$, let $\pi_u = \pi_* + \frac{u}{\sqrt{n}}h$. We suppose*

$$\sup_{h\in\mathcal{E}(M_n)} \sup_{u\in(0,1)} \left\| E_{\pi_u}\left[J_{\pi_u}^{-1}(t(X)-\pi_u)\left((t(X)-\pi_u)^T I(\pi_u)h\right)^2\right]\right\| = o\left(\sqrt{\frac{n}{M_n}}\right).$$

Assumption 6. *For $h \in \mathcal{E}(M)$ and $u \in (0,1)$, let $\pi_u = \pi_* + \frac{u}{\sqrt{n}}h$. We suppose*

$$\sup_{h\in\mathcal{E}(M_n)} \sup_{u\in(0,1)} E_{\pi_u}\left[\left((t(X)-\pi_u)^T I(\pi_u)h\right)^4\right] = o(n)$$

Apart the growth condition, these assumptions differ from the ones of the precursor paper of Ghosal [53] mainly in the fact that we avoid the simpler but non-tight moments $\sup_{\|a\|=1} E_{\pi_u}\left[\left|a^T J_{\pi_u}^{-1}(t(X)-\pi_u)\right|^4\right]$:

Prior hypotheses

Assumption 7 (Prior Smoothness)**.**

$$\sup_{h,g\in\mathcal{E}(M_n)} \frac{w(\pi_* + \frac{h}{\sqrt{n}})}{w(\pi_* + \frac{g}{\sqrt{n}})} \to 1 \quad as \quad n\to\infty.$$

Assumption 8 (Prior concentration)**.**

$$2\ln w(\pi_*) - k_n \ln n - \ln\det I(\pi_*) = o(M_n).$$

As an example, Boucheron and Gassiat [17, Proposition 3.14] study the Dirichlet priors $D(\beta,\dots,\beta)$ with $\beta > 0$ in the case of multinomial distributions: they fulfill the Prior hypotheses, under a bit stronger assumption on the growth rate $k_n \ln n = o(M_n)$, together with $\ln\det I(\pi_*) = o(M_n)$.

A.2.2 Main result

Let $\mathcal{N}(\Delta_n, \ddot{\psi}(\theta_*))$ denote the normal distribution on $\mathbb{R}^{k_n}$ with mean Δ_n and covariance matrix $\ddot{\psi}(\theta_*)$. Let $\|\cdot\|_{\mathrm{TV}}$ denote the total variation norm.

Theorem 11. *Suppose that $(k_n)_{n\geq1}$ and $(\pi_*^{k_n})_{n\geq1}$ are such that, for some sequence $(M_n)_{n\geq1}$, Assumptions 1, 2, 3, 4, 5, and 6, 7, and 8 hold. Then*

$$E_{\pi_*}\left\|\mathcal{N}\big(\Delta_n, \ddot{\psi}(\theta_*)\big) - P_{H_n|X_{1:n}}\right\|_{\mathrm{TV}} \to 0 \quad as \quad n\to\infty.$$

Assumptions 5 and 6 can be deduced from a stronger version of Assumption 4:

Assumption 9.

$$M_n^3 \|I(\pi_*)\| \sup_{x\in\mathcal{X}} \|t(x)\|^2 = o(n).$$

Lemma 27. *Suppose Assumptions 2, 3, and 9 hold. Then Assumption 5 is satisfied.*

Proof. Let us denote (A) the quantity

$$\text{(A)} = \left\| E_{\pi_u}\left[J_{\pi_u}^{-1}(t(X)-\pi_u)\left((t(X)-\pi_u)^T I(\pi_u)h\right)^2\right]\right\|.$$

$$\begin{aligned}\text{(A)} &\leq 2\|J_{\pi_u}^{-1}\| \sup_{x\in\mathcal{X}}\|t(x)\| E_{\pi_u}\left[\left((t(X)-\pi_u)^T I(\pi_u)h\right)^2\right] \\ &\leq 2\sqrt{\|J_{\pi_*}^T I(\pi_u) J_{\pi_*}\| \, \|I(\pi_*)\|} \sup_{x\in\mathcal{X}}\|t(x)\| \; h^T I(\pi_u) h\end{aligned}$$

Therefore

$$\begin{aligned}\sup_{h\in\mathcal{E}(M_n)} \sup_{u\in(0,1)} \text{(A)} &\leq 2M_n \sup_{h\in\mathcal{E}(M_n)} \|J_{\pi_*}^T I(\pi) J_{\pi_*}\|^{3/2} \sqrt{\|I(\pi_*)\|} \sup_{x\in\mathcal{X}}\|t(x)\| \\ &= o\left(\sqrt{\frac{n}{M_n}}\right)\end{aligned}$$

thanks to Assumptions 3 and 9. □

Lemma 28. *Suppose that Assumptions 2 and 3 hold. Suppose further that*

$$M_n^2 \|I(\pi_*)\| \sup_{x\in\mathcal{X}} \|t(x)\|^2 = o(n).$$

Then Assumption 6 holds.

Proof of Lemma 28. With the same arguments as in the proof of Lemma 27, we get the upper bound

$$\begin{aligned}&\sup_{h\in\mathcal{E}(M_n)} \sup_{u\in(0,1)} E_{\pi_u}\left[\left((t(X)-\pi_u)^T I(\pi_u)h\right)^4\right] \\ &\quad\leq 4M_n^2 \sup_{h\in\mathcal{E}(M_n)} \|J_{\pi_*}^T I(\pi) J_{\pi_*}\|^3 \|I(\pi_*)\| \sup_{x\in\mathcal{X}}\|t(x)\|^2 \\ &\quad= o(n).\end{aligned}$$

□

With these lemmas, we get a simpler but slightly less powerful Bernstein-von Mises Theorem:

Theorem 12. *Suppose that $(k_n)_{n\geq 1}$ and $(\pi_*^{k_n})_{n\geq 1}$ are such that, for some sequence $(M_n)_{n\geq 1}$, Assumptions 1, 2, 3, 7, 8, and 9 hold. Then*

$$E_{\pi_*}\left\| \mathcal{N}\left(\Delta_n, \ddot{\psi}(\theta_*)\right) - P_{H_n|X_{1:n}}\right\|_{\text{TV}} \to 0 \quad as \quad n\to\infty.$$

A.3 Proof of Theorem 11

A.3.1 Sketch of proof

For any probability distribution P on $\mathbb{R}^{k_n}$, let P^M be the probability distribution defined by

$$P^M(B) = \frac{P(B \cap \mathcal{E}(M))}{P(\mathcal{E}(M))}.$$

$$\begin{aligned}\left\|\mathcal{N}\big(\Delta_n, \ddot{\psi}(\theta_*)\big) - P_{H_n|X_{1:n}}\right\|_{\mathrm{TV}} \leq{}& \left\|\mathcal{N}\big(\Delta_n, \ddot{\psi}(\theta_*)\big) - \mathcal{N}^{M_n}\big(\Delta_n, \ddot{\psi}(\theta_*)\big)\right\|_{\mathrm{TV}} \\ &+ \left\|\mathcal{N}^{M_n}\big(\Delta_n, \ddot{\psi}(\theta_*)\big) - P^{M_n}_{H_n|X_{1:n}}\right\|_{\mathrm{TV}} \\ &+ \left\|P^{M_n}_{H_n|X_{1:n}} - P_{H_n|X_{1:n}}\right\|_{\mathrm{TV}}\end{aligned}$$

We deal with each part separately with the following Propositions.

Proposition 20. *Assume Assumptions 1, 2, 3, 4, 5, 6, and 7 hold. Then*

$$E_{\pi_*}\left\|\mathcal{N}^{M_n}\big(\Delta_n, \ddot{\psi}(\theta_*)\big) - P^{M_n}_{H_n|X_{1:n}}\right\|_{\mathrm{TV}} \to 0 \quad \textit{as} \quad n \to \infty.$$

Proposition 21. *Assume Assumptions 1, 2, 3, 5, 6, 7, and 8 hold. Then*

$$E_{\pi_*}\left\|P^{M_n}_{H_n|X_{1:n}} - P_{H_n|X_{1:n}}\right\|_{\mathrm{TV}} \to 0 \quad \textit{as} \quad n \to \infty.$$

Proposition 22. *Assume Assumptions 1 and 2 hold. Suppose further that*

$$M_n^{-1}\|I(\pi_*)\| \sup_{x \in \mathcal{X}} \|t(x)\|^2 = o(n).$$

Then

$$E_{\pi_*}\left\|\mathcal{N}\big(\Delta_n, \ddot{\psi}(\theta_*)\big) - \mathcal{N}^{M_n}\big(\Delta_n, \ddot{\psi}(\theta_*)\big)\right\|_{\mathrm{TV}} \to 0 \quad \textit{as} \quad n \to \infty.$$

Proposition 20 and Proposition 21 are proved respectively in Subsection A.3.2 and Subsection A.3.3. Proposition 22 is the same as [17, Proposition 3.11]; its proof relies on Cirelson's Inequality and Lemma 30 below.

A.3.2 Truncated distributions

In order to prove Proposition 20, let us introduce the simpler notation

$$(\mathrm{A}) = \left\|\mathcal{N}^{M_n}\big(\Delta_n, \ddot{\psi}(\theta_*)\big) - P^{M_n}_{H_n|X_{1:n}}\right\|_{\mathrm{TV}}.$$

The proof relies on Taylor expansions of log-likelihood ratios. Consider $\pi_* \in \Pi$, and h such that $\left[\pi_*, \pi_* + \frac{h}{\sqrt{n}}\right] \subset \Pi$. Let π_u denote $\pi_* + u\frac{h}{\sqrt{n}}$. Let us define

$$A_n(h) = \sqrt{n}\, J_{\pi_*}^T \left(Q(\pi_1) - Q(\pi_0) - \left.\frac{\mathrm{d}Q(\pi_u)}{\mathrm{d}u}\right|_{u=0} \right) \tag{A.1}$$

$$\begin{aligned} B_n(h) = n\,\pi_*^T \left(Q(\pi_1) - Q(\pi_0) - \left.\frac{\mathrm{d}Q(\pi_u)}{\mathrm{d}u}\right|_{u=0} - \frac{1}{2}\left.\frac{\mathrm{d}^2Q(\pi_u)}{\mathrm{d}u^2}\right|_{u=0} \right) \\ - n\left(\phi(\pi_1) - \phi(\pi_0) - \left.\frac{\mathrm{d}\phi(\pi_u)}{\mathrm{d}u}\right|_{u=0} - \frac{1}{2}\left.\frac{\mathrm{d}^2\phi(\pi_u)}{\mathrm{d}u^2}\right|_{u=0} \right). \end{aligned} \tag{A.2}$$

Eventually, let $P_{n,h}$ be a notation for $P^{\otimes n}_{\pi_* + \frac{h}{\sqrt{n}}}$. We prove in Appendix A.A

$$\ln \frac{\mathrm{d}P_{n,h}}{\mathrm{d}P_{n,0}}(X_{1:n}) = \Delta_n^T I(\pi_*)h - \frac{1}{2}h^T I(\pi_*)h + \Delta_n^T (J_{\pi_*}^T)^{-1} A_n(h) + B_n(h). \tag{A.3}$$

Then, the same calculus as [17, Appendix D], which relies on the Jensen inequality, gives

$$\begin{aligned} (\mathrm{A}) \leq \iint \left(1 - \frac{w\left(\pi_* + \frac{g}{\sqrt{n}}\right)}{w\left(\pi_* + \frac{h}{\sqrt{n}}\right)} \exp\left\{ \Delta_n^T (J_{\pi_*}^T)^{-1} \big(A_n(g) - A_n(h) \big) \right\} \right. \\ \left. \cdot \exp\big\{ B_n(g) - B_n(h) \big\} \right)_+ \mathrm{d}\mathcal{N}^{M_n}(g)\, \mathrm{d}P^{M_n}_{H_n|X_{1:n}}(h) \end{aligned} \tag{A.4}$$

where $(x)_+ = \max(0, x)$, and $\mathcal{N}^{M_n}$ is an abbreviation for $\mathcal{N}^{M_n}\big(\Delta_n, \ddot{\psi}(\theta_*)\big)$.

Using the relation

$$\forall x \geq 0, \quad (1 - x\, e^a)_+ \leq |a|\, x + (1 - x)_+ \tag{A.5}$$

we can write

$$(\mathrm{A}) \leq \left(1 - \frac{1}{C}\right) + 2C \left(\|J_{\pi_*}^{-1}\Delta_n\| \sup_{h \in \mathcal{E}(M_n)} \|A_n(h)\| + \sup_{h \in \mathcal{E}(M_n)} |B_n(h)| \right)$$

where

$$C = \sup_{g,h \in \mathcal{E}(M_n)} \frac{w\left(\pi_* + \frac{g}{\sqrt{n}}\right)}{w\left(\pi_* + \frac{h}{\sqrt{n}}\right)}$$

converges to 1 if Assumption 7 holds. Notice that everything is deterministic, but $\|J_{\pi_*}^{-1}\Delta_n\|$. Taking the expectation, we have

$$E_{\pi_*} \|J_{\pi_*}^{-1}\Delta_n\| \leq \sqrt{E_{\pi_*}\left[\|J_{\pi_*}^{-1}\Delta_n\|^2\right]} = \sqrt{k_n} \tag{A.6}$$

Then, we need the following result, proved in Appendix A.B.

Proposition 23. *Assume Assumptions 1, 2, 3, 5, and 6 hold. Then*

$$\sqrt{k_n} \sup_{h \in \mathcal{E}(M_n)} \|A_n(h)\| + \sup_{h \in \mathcal{E}(M_n)} |B_n(h)| \to 0 \quad \textit{as} \quad n \to \infty.$$

A.3.3 Posterior Concentration

The first step is given by the following immediate Lemma:

Lemma 29. *Let P denote a probability distribution on some space $(\Omega, \mathcal{F})$. Let A denote an event with non-null P-probability and let P^A denote the conditional probability given A, that is $P^A(B) = P(A \cap B)/P(A)$, then*

$$\|P^A - P\| = P(A^c).$$

Here A^c denote the complementary set $\Omega \setminus A$. Then, we adapt the argument of [17, 106] to upper bound $E_{\pi_*}\left[P\big(H_n^T I(\pi_*) H_n > M_n | X_{1:n}\big)\right]$.

Let us consider a sequence of test $\phi_n = \mathbb{1}_{\Delta_n^T I(\pi_*)\Delta_n > M_n/4}$. Then, we get from Lemma 29

$$E_{\pi_*}\left\|P^{M_n}_{H_n|X_{1:n}} - P_{H_n|X_{1:n}}\right\|_{\mathrm{TV}} \leq E_{\pi_*}\,\phi_n + E_{\pi_*}\left[(1-\phi_n)P\big(H_n^T I(\pi_*) H_n > M_n | X_{1:n}\big)\right]$$

For the first part, we use the following Lemma, proved in Appendix A.C.

Lemma 30. *Assume Assumption 1 holds, and suppose*

$$M_n^{-1}\|I(\pi_*)\| \sup_{x\in\mathcal{X}} \|t(x)\|^2 = o(n).$$

Then

$$P_{\pi_*}\big(\Delta_n^T I(\pi_*)\Delta_n > M_n/4\big) \to 0 \quad as \quad n \to \infty.$$

Then, let us choose a sequence r_n of positive numbers such that $r_n \to 0$ and $-\ln r_n = o(M_n/k_n)$. Assumption 1 insures that such a choice is possible. Let us define the probability distribution $P_{n,\mathcal{E}(r_n)}$ on $\mathcal{X}^n$ as the mixture of $P_{n,h}$ when the prior is conditioned on the ellipsoid $\mathcal{E}(r_n)$:

$$P_{n,\mathcal{E}(r_n)}(B) = \frac{\int_{\mathcal{E}(r_n)} \mathrm{d}h\, n^{-k_n/2} w\left(\pi_* + \frac{h}{\sqrt{n}}\right) P_{n,h}(B)}{W\left(\pi_* + \frac{1}{\sqrt{n}}\mathcal{E}(r_n)\right)}. \tag{A.7}$$

For commodity in long integrals, we put the integration variable at the beginning.

The link between $P_{\pi_*}^{\otimes n}$ and $P_{n,\mathcal{E}(r_n)}$ is given by

$$\forall X : \Omega \to [0,1], \quad |E_P\, X - E_Q\, X| \leq \|P - Q\| \tag{A.8}$$

and the following Lemma, proved in Appendix A.D.

Lemma 31. *Assume Assumptions 1, 2, 3, 5, and 6 hold. Then*

$$\left\|P_{\pi_*}^{\otimes n} - P_{n,\mathcal{E}(r_n)}\right\|_{\mathrm{TV}} \to 0 \quad as \quad n \to \infty.$$

This lemma plays the same role as the contiguity argument of [106]. It can be adapted to get an explicit convergence speed, if needed.

Thus, it is enough if the following quantity converges to 0:

$$(\mathrm{B}) = E_{n,\mathcal{E}(r_n)}\left[(1-\phi_n)P\big(H_n^T I(\pi_*)H_n > M_n | X_{1:n}\big)\right].$$

Using a Fubini integration, we get

$$\begin{aligned}(\mathrm{B}) &= \frac{1}{W\left(\pi_* + \frac{1}{\sqrt{n}}\mathcal{E}(r_n)\right)} \int_{\mathcal{E}(M_n)^c} \mathrm{d}h\, n^{-k_n/2} w\left(\pi_* + \frac{h}{\sqrt{n}}\right) \\ &\quad \cdot E_{n,h}\left[(1-\phi_n)P\big(H_n^T I(\pi_*)H_n \le r_n | X_{1:n}\big)\right] \\ &\le \frac{1}{W\left(\pi_* + \frac{1}{\sqrt{n}}\mathcal{E}(r_n)\right)} \sup_{h\in\mathcal{E}(M_n)^c} E_{n,h}(1-\phi_n)\end{aligned}$$

We conclude the proof of Proposition 21 thanks to the following two lemmas:

Lemma 32. *Assume Assumptions 2, 7, and 8 hold. Then*

$$-\ln W\left(\pi_* + \frac{1}{\sqrt{n}}\mathcal{E}(r_n)\right) = o(M_n).$$

Lemma 33. *Assume Assumptions 3 and 4 hold. Then, there exists a constant $c > 0$ such that, for n large enough,*

$$\sup_{h\in\mathcal{E}(M_n)^c} E_{n,h}(1-\phi_n) \le \exp\left(-c\, M_n\right).$$

Proof of Lemma 32. For n large enough, $r_n \le M_n$. By Assumptions 2 and 7,

$$\begin{aligned}W\left(\pi_* + \frac{1}{\sqrt{n}}\mathcal{E}(r_n)\right) &= (1+o(1))\, w(\pi_*)\, \mathrm{Vol}\left(\frac{1}{\sqrt{n}}\mathcal{E}(r_n)\right) \\ &= (1+o(1))\, n^{-k_n/2} w(\pi_*)\, r_n^{k_n/2} \left(\det I(\pi_*)\right)^{-1/2} \\ &\quad \cdot \mathrm{Vol}\left(\left\{x\in\mathbb{R}^{k_n} : \|x\| \le 1\right\}\right) \\ &\ge (1+o(1))\, n^{-k_n/2} r_n^{k_n/2} \left(\det I(\pi_*)\right)^{-1/2} w(\pi_*).\end{aligned}$$

We conclude using Assumption 8 and the fact that $-k_n \ln r_n = o(M_n)$. □

Proof of Lemma 33. For $\sigma_n(h) > \sqrt{M_n}$, let $\pi = \pi_* + \frac{h}{\sqrt{n}}$. Then

$$\begin{aligned}\sqrt{\Delta_n^T I(\pi_*)\Delta_n} &= \sup_{a:\|a\|=1} a^T J_{\pi_*}^{-1}\Delta_n \\ &\ge \frac{1}{\sqrt{n}}\sum_{i=1}^n a_*^T J_{\pi_*}^{-1}(t(X_i)-\pi) + \sigma_n(h)\end{aligned}$$

where $a_* = \frac{J_{\pi_*}^{-1}h}{\sigma_n(h)}$. Let $Y_i = -\frac{1}{\sqrt{n}} a_*^T J_{\pi_*}^{-1}(t(X_i)-\pi)$, for $1 \le i \le n$. Then

$$E_{n,h}\left(1-\phi_n\right) \le P_{n,h}\left(\sum_{i=1}^n Y_i \ge \sigma_n(h) - \frac{\sqrt{M_n}}{2}\right).$$

Using Bennett's inequality [75, (2.16)], we get

$$E_{n,h}\left(1-\phi_n\right) \leq \exp\left(-\frac{(\sigma_n(h)-\sqrt{M_n}/2)^2}{2v+2b(\sigma_n(h)-\sqrt{M_n}/2)/3}\right)$$

where

$$v = \mathrm{Var}_\pi(a_*^T J_{\pi_*}^{-1} t(X)) \leq \|J_{\pi_*}^{-1}\ddot{\psi}(\theta)(J_{\pi_*}^T)^{-1}\|$$
$$b = 2\sqrt{\frac{\|I(\pi_*)\|}{n}}\sup_{x\in\mathcal{X}}\|t(x)\|.$$

Assumption 3 and the fact that $\sigma_n(h) > \sqrt{M_n}$ entails

$$v \leq \frac{\sigma_n^2(h)}{M_n}\,O(1).$$

Using Assumption 4, we get

$$E_{n,h}\left(1-\phi_n\right) \leq \exp\left\{-\frac{M_n}{O(1)}\right\}$$

where $O(1)$ denotes a positive bounded function of n, non depending on h. □

A.4 Application to multinomial distributions

The special case of multinomial distributions has been studied in [17, 53]. Here, we follow the approach of Boucheron and Gassiat [17], where we aim at approximating multinomial distributions on $\mathbb{N}_* = \mathbb{N}\backslash\{0\}$ by distribution lying in increasing-dimensional exponential models.

Let $\mu = \sum_{m\geq 1}\delta_{\{m\}}$ be the counting measure on $\mathbb{N}_*$. Let Q be some probability distribution on $\mathbb{N}_*$. For any $k \geq 1$, let Q_k be defined by

$$\begin{array}{ll} Q_k(j) = 0 & \text{if } 1 \leq j \leq k \\ Q_k(j) = Q(j-k) & \text{if } j \geq k+1. \end{array}$$

Let us choose $f^k(j) = \mathbb{1}_{1\leq j\leq k} + Q_k(j)$. We define $t_j = \mathbb{1}_{\{j\}}$, $\Theta = \mathbb{R}^k$, and

$$\Pi = \left\{\pi \in \mathbb{R}^k : \forall 1 \leq j \leq k, \quad \pi(j) > 0 \quad \text{and} \quad \sum_{j=1}^k \pi(j) < 1\right\}.$$

Then, for any $\pi \in \Pi$, P_π is the distribution on $\mathbb{N}_*$ defined by

$$\begin{array}{ll} P_\pi(j) = \pi(j) & \text{if } 1 \leq j \leq k \\ P_\pi(j) = \pi(0)\,Q_k(j) & \text{if } j \geq k+1 \end{array}$$

where $\pi(0)$ is a notation for $\pi(0) = 1 - \sum_{j=1}^k \pi(j)$.

Proposition 24. *Suppose that*

$$k_n = o\left(\left(n \inf_{0\leq j\leq k} \pi_*(j)\right)^{1/3}\right) \tag{A.9}$$

and Π *contains a neighborhood of* π_* *such that any* $\pi \in \mathbb{R}^{k_n}$ *satisfying*

$$\left|1 - \frac{\pi(j)}{\pi_*(j)}\right| \leq \frac{1}{k_n} \tag{A.10}$$

belongs to Π.

Then there exists a sequence $(M_n)_{n\geq 1}$ *such that Assumptions 1, 2, 3, 4, 5 and 6 hold for* n *large enough.*

Note that π_* can approach the edge of Π as n increases. In situations where π_* is simply the projection of a multinomial distribution P_0 on $\mathbb{N}_*$, this assumption reduces to P_0 being an interior point of the model.

Proposition 24 enables us to retrieve the main Theorem 3.7 of [17]. In particular, Boucheron and Gassiat proved in their Proposition 3.14 that the Dirichlet distributions $D(\beta, \ldots, \beta)$ with $\beta > 0$ fulfill our last assumptions 7 and 8, so that Theorem 11 applies, if

$$\max\left(k_n \ln n, \ln\det I(\pi_*)\right) = o\left(\left(n \inf_{0\leq j\leq k} \pi_*(j)\right)^{1/3}\right).$$

Proof. If (A.9) holds, by setting

$$M_n = \sqrt{k_n}\left(n \inf_{0\leq j\leq k} \pi_*(j)\right)^{1/6} \tag{A.11}$$

we construct a sequence $(M_n)_{n\geq 1}$ such that Assumption 1 holds, and

$$M_n = o\left(\left(n \inf_{0\leq j\leq k} \pi_*(j)\right)^{1/3}\right). \tag{A.12}$$

The Fisher information matrix of Multinomial models is known (see for instance [53]):

$$\ddot{\psi}(\theta) = D - \pi\pi^T$$
$$I(\pi) = D^{-1} + \frac{\mathbf{1}\mathbf{1}^T}{\pi(0)}$$

where D is the diagonal matrix $\mathrm{Diag}(\pi(j), 1 \leq j \leq k)$, and $\mathbf{1} \in \mathbb{R}^k$ is the vector whose all coordinates are 1.

Therefore

$$\|I(\pi_*)\| \leq (k_n + 1) \sup_{0\leq j\leq k_n} \frac{1}{\pi_*(j)}.$$

In the multinomial case, we have also $\|t(x)\| = 1$ for all $x \in \mathcal{X}$. Therefore a consequence of (A.11) and (A.12) is

$$M_n^2 \|I(\pi_*)\| \sup_{x\in\mathcal{X}} \|t(x)\|^2 = o(n).$$

This entails Assumption 4.

We also know explicit expressions for $\det I(\pi)$ and J_π:

$$\det I(\pi) = \prod_{j=0}^{k_n} \frac{1}{\pi(j)} \tag{A.13}$$

$$J_\pi = D^{1/2} - \frac{\pi\pi^T D^{-1/2}}{1+\sqrt{\pi(0)}}$$

$$J_\pi^{-1} = D^{-1/2} + \frac{D^{1/2}\mathbf{1}\mathbf{1}^T}{\pi(0)+\sqrt{\pi(0)}}.$$

Assumption 3 is verified thanks to Lemma 34 below, Assumption 2 thanks to Lemma 35, and Assumption 5 thanks to Lemma 36. Eventually, Assumption 6 comes from Lemma 28. □

Lemma 34. *Assume $k_n = o(M_n)$ and $M_n = o\left(\sqrt{n} \wedge \left(n \inf_{0\le j\le k} \pi_*(j)\right)\right)$. Then Assumption 3 holds.*

We use $a \wedge b$ as a notation for $\min(a, b)$.

Lemma 35. *Assume that (A.9) and (A.10) hold, and choose M_n as in (A.11). Then Assumption 2 holds for n large enough.*

Lemma 36. *If $M_n = o\left(\left(n \inf_{0\le j\le k} \pi_*(j)\right)^{1/3}\right)$, then Assumption 5 is satisfied.*

Proof of Lemma 34. Let $\pi \in \Pi$, and $h = \sqrt{n}\,(\pi - \pi_*)$. $\|J_\pi^T I(\pi_*) J_\pi\|$ is the spectral radius of the symmetric matrix $J_\pi^T I(\pi_*) J_\pi$. It is also the spectral radius of the matrix

$$D_*^{-1} D \left(I_{k_n} + \mathbf{1}(\pi_* - \pi)\right) = D_*^{-1} \ddot{\psi}(\theta) I(\pi_*) D_*.$$

Therefore,

$$\begin{aligned} \|J_\pi^T I(\pi_*) J_\pi\| &\le \|D_*^{-1} D\|\, \|I_{k_n} + \mathbf{1}(\pi_* - \pi)\| \\ &\le \sup_{1\le j\le k_n} \frac{\pi(j)}{\pi_*(j)} \left(1 + \sqrt{k_n}\,\|\pi_* - \pi\|\right). \end{aligned} \tag{A.14}$$

Similarly,

$$\|J_{\pi_*}^T I(\pi) J_{\pi_*}\| \le \sup_{1\le j\le k_n} \frac{\pi_*(j)}{\pi(j)} \left(1 + \sqrt{k_n}\,\|\pi_* - \pi\|\right).$$

On another hand, straightforward computations give

$$\sigma_n^2(h) = n \sum_{j=0}^{k} \frac{(\pi(j) - \pi_*(j))^2}{\pi_*(j)}$$

So

$$\begin{aligned}\|\pi_* - \pi\|^2 &= \sum_{j=1}^{k} \pi_*(j) \, \frac{(\pi(j) - \pi_*(j))^2}{\pi_*(j)} \\ &\leq \frac{\sigma_n^2(h)}{n}. \end{aligned} \tag{A.15}$$

For any $0 \leq j \leq k$, we have

$$\begin{aligned}\left|1 - \frac{\pi(j)}{\pi_*(j)}\right| &= \frac{|\pi(j) - \pi_*(j)|}{\pi_*(j)} \\ &\leq \sup_{0\leq j\leq k} \frac{1}{\sqrt{\pi_*(j)}} \sqrt{\sum_{l=0}^{k} \frac{(\pi(l) - \pi_*(l))^2}{\pi_*(l)}} \\ &\leq \sqrt{\frac{\sigma_n^2(h)}{n \inf_{0\leq j\leq k} \pi_*(j)}}. \end{aligned} \tag{A.16}$$

Therefore, since $M_n = o\left(n \inf\limits_{0\leq j\leq k} \pi_*(j)\right)$,

$$\lim_{n\to\infty} \sup_{\sigma_n^2(h)\leq M_n} \sup_{0\leq j\leq k_n} \frac{\pi_*(j)}{\pi(j)} = 1 \tag{A.17}$$

and

$$\sup_{\sigma_n^2(h)\leq M_n} \|J_{\pi_*}^T I(\pi) J_{\pi_*}\| = 1 + o(1).$$

If now we consider π such as $\sigma_n^2(h) \geq M_n$, then (A.14), (A.15), and (A.16) lead to

$$\begin{aligned}\left\| J_{\pi_*}^{-1} \ddot{\psi}(\theta) (J_{\pi_*}^T)^{-1} \right\| &\leq \left(1 + \frac{\sigma_n(h)}{\sqrt{n \inf_{0\leq j\leq k} \pi_*(j)}}\right) \cdot \left(1 + \frac{\sqrt{k_n}\, \sigma_n(h)}{\sqrt{n}}\right) \\ &\leq \left(1 + \frac{\sigma_n(h)}{\sqrt{M_n}}\, o(1)\right)^2. \end{aligned}$$

In the last line we used the assumption $M_n = o\left(\sqrt{n} \wedge \left(n \inf_{0\leq j\leq k} \pi_*(j)\right)\right)$. □

Proof of Lemma 35. If $\pi \in \mathcal{E}_{n,\pi_*}(M_n)$, $\sigma_n^2(h) \leq M_n$. From (A.16) we get

$$\begin{aligned}\left|1 - \frac{\pi(j)}{\pi_*(j)}\right| &\leq \sqrt{\frac{M_n}{n \inf_{0\leq j\leq k} \pi_*(j)}}. \\ &= o\left(k_n^{-1}\right). \end{aligned}$$

Thus $\pi \in \Pi$ for n large enough. □

Proof of Lemma 36. For $h \in \mathcal{E}(M_n)$ and $u \in (0,1)$, let $\pi = \pi_* + \frac{u}{\sqrt{n}}h$. If $u \neq 0$, we have $h = \frac{\sqrt{n}}{u}(\pi - \pi_*)$. The property $h \in \mathcal{E}(M_n)$ translates into

$$\frac{n}{u^2}\sum_{j=0}^{k}\frac{(\pi(j)-\pi_*(j))^2}{\pi_*(j)} \leq M_n. \tag{A.18}$$

Then

$$h^T I(\pi)(t(X)-\pi) = \frac{\sqrt{n}}{t}\sum_{j=0}^{k}\mathbb{1}_{\{X=j\}}\frac{\pi(j)-\pi_*(j)}{\pi(j)}$$

$$J_\pi^{-1}(t(X)-\pi) = \mathbb{1}_{\{X=0\}}\frac{(3-2\pi(0))}{\sqrt{\pi(0)}}[\sqrt{\pi}] + \sum_{j=1}^{k}\mathbb{1}_{\{X=j\}}\left(\frac{1}{\sqrt{\pi(j)}}e_j + \left(\frac{2(1-\pi(0))}{\sqrt{\pi(0)}} - \frac{1}{(1+\sqrt{\pi(0)})}\right)[\sqrt{\pi_*}]\right)$$

where e_j is the j^{th} canonical vector of $\mathbb{R}^k$, and $[\sqrt{\pi}] = \left(\sqrt{\pi(j)}\right)_{1\leq j\leq k}$.

Therefore

$$\begin{aligned}
&\left\|E_\pi\left[J_\pi^{-1}(t(X)-\pi)\left(h^T I(\pi)(t(X)-\pi)\right)^2\right]\right\| \\
&= \frac{n}{t^2}\left\|\left(\frac{2(1-\pi(0))}{\sqrt{\pi(0)}}\sum_{j=0}^{k}\frac{(\pi(j)-\pi_*(j))^2}{\pi(j)} + \frac{(\pi(0)-\pi_*(0))^2}{\pi(0)^{3/2}}\right.\right. \\
&\quad \left.\left. - \sum_{j=1}^{k}\frac{(\pi(j)-\pi_*(j))^2}{\pi(j)\left(1+\sqrt{\pi(0)}\right)}\right)[\sqrt{\pi}] + \sum_{j=1}^{k}\frac{(\pi(j)-\pi_*(j))^2}{\pi(j)^{3/2}}e_j\right\| \\
&\leq 4M_n \sup_{0\leq j\leq k}\left(\frac{\pi_*(j)}{\pi(j)}\right)^{3/2}\sup_{0\leq j\leq k}\frac{1}{\sqrt{\pi_*(j)}}
\end{aligned}$$

using (A.18) and the fact that $\|[\sqrt{\pi}]\| \leq \|e_j\| = 1$. Using (A.17), we get

$$\sup_{h\in\mathcal{E}(M_n)}\sup_{u\in(0,1)}\left\|E_{\pi_u}\left[J_{\pi_u}^{-1}(t(X)-\pi_u)\left((t(X)-\pi_u)^T I(\pi_u)h\right)^2\right]\right\| = O\left(\frac{M_n}{\inf_{0\leq j\leq k}\sqrt{\pi_*(j)}}\right) = o\left(\sqrt{\frac{n}{M_n}}\right).$$

□

A.A Taylor expansion of log-likelihood ratios

We start with the expression of the densities $f_{\pi_*} = f_{\pi_0}$ and $f_{\pi_*+\frac{h}{\sqrt{n}}} = f_{\pi_1}$.

$$\begin{aligned}
\ln\frac{\mathrm{d}P_{n,h}}{\mathrm{d}P_{n,0}}(X_{1:n}) &= \sum_{i=1}^{n} t(X_i)^T\left(Q(\pi_1)-Q(\pi_0)\right) - n\left(\phi(\pi_1)-\phi(\pi_0)\right) \\
&= \sqrt{n}\,\Delta_n\left(Q(\pi_1)-Q(\pi_0)\right) + n\,\pi_0^T\left(Q(\pi_1)-Q(\pi_0)\right) \\
&\quad - n\left(\phi(\pi_1)-\phi(\pi_0)\right)
\end{aligned}$$

Then, we need to calculate some derivative functions:

$$\frac{\mathrm{d}Q(\pi_u)}{\mathrm{d}u} = \frac{1}{\sqrt{n}} I(\pi_u) h \tag{A.19}$$

$$\frac{\mathrm{d}\phi(\pi_u)}{\mathrm{d}u} = \pi_u^T \frac{\mathrm{d}Q(\pi_u)}{\mathrm{d}u} \tag{A.20}$$

$$\frac{\mathrm{d}^2\phi(\pi_u)}{\mathrm{d}u^2} = \frac{1}{\sqrt{n}} h^T \frac{\mathrm{d}Q(\pi_u)}{\mathrm{d}u} + \pi_u^T \frac{\mathrm{d}^2 Q(\pi_u)}{\mathrm{d}u^2}$$

(A.3) is a straightforward consequence of these equations.

A.B Proof of Proposition 23

Upper bound of $\|A_n(h)\|$.

We start with a simple modification of (A.1):

$$\|A_n(h)\| \leq \sup_{\tilde{u}\in(0,1)} \left\| \frac{\sqrt{n}}{2} J_{\pi_*}^T \left. \frac{\mathrm{d}^2 Q(\pi_u)}{\mathrm{d}u^2} \right|_{u=\tilde{u}} \right\|$$

Now, using (A.19) and the fact that, for any continuously derivable, invertible matrix $A(t)$, $\frac{\mathrm{d}}{\mathrm{d}u}\left(t \mapsto A^{-1}(t)\right) = -A^{-1}(t)\frac{\mathrm{d}A(t)}{\mathrm{d}u}A^{-1}(t)$, we get

$$\frac{\mathrm{d}^2 Q(\pi_u)}{\mathrm{d}u^2} = -\frac{1}{\sqrt{n}} I(\pi_u) \frac{\mathrm{d}\ddot{\psi}(Q(\pi_u))}{\mathrm{d}u} I(\pi_u) h$$

Then

$$\begin{aligned}
\frac{\mathrm{d}\ddot{\psi}(Q(\pi_u))}{\mathrm{d}u} &= \frac{\mathrm{d}}{\mathrm{d}u} \int_{\mathcal{X}} (t(x)-\pi_u)(t(x)-\pi_u)^T e^{t(x)^T Q(\pi_u) - \phi(\pi_u)} f(x)\,\mathrm{d}\mu(x) \\
&= \frac{1}{\sqrt{n}} \int_{\mathcal{X}} \Big[(t(x)-\pi_u)(t(x)-\pi_u)^T I(\pi_u) h\,(t(x)-\pi_u)^T \\
&\qquad - h\,(t(x)-\pi_u)^T - (t(x)-\pi_u) h^T \Big] e^{t(x)^T Q(\pi_u) - \phi(\pi_u)} f(x)\,\mathrm{d}\mu(x)
\end{aligned}$$

The derivation is known to be valid (see for instance [106, p. 38]).

Since $\pi_u = E_{\pi_u}[t(X)]$, we get

$$\frac{\mathrm{d}^2 Q(\pi_u)}{\mathrm{d}u^2} = -\frac{1}{n} E_{\pi_u}\left[I(\pi_u)(t(X)-\pi_u)\left((t(X)-\pi_u)^T I(\pi_u) h\right)^2 \right]. \tag{A.21}$$

Then, we can write

$$\begin{aligned}
\left\| J_{\pi_*}^T \frac{\mathrm{d}^2 Q(\pi_u)}{\mathrm{d}u^2} \right\| \leq \frac{\sqrt{\|J_{\pi_*}^T I(\pi_u) J_{\pi_*}\|}}{n} \\
\left\| E_{\pi_u}\left[J_{\pi_u}^{-1}(t(X)-\pi_u)\left((t(X)-\pi_u)^T I(\pi_u) h\right)^2 \right] \right\|
\end{aligned}$$

With Assumptions 1, 3, and 5, we get

$$\sqrt{k_n} \sup_{h\in\mathcal{E}(M_n)} \|A_n(h)\| = o(1).$$

Upper bound of $|B_n(h)|$.

As before, we use Taylor's formula:

$$B_n(h) = \frac{n}{6}\,\frac{\mathrm{d}^3}{\mathrm{d}u^3}\left(\pi_*^T Q(\pi_u) - \phi(\pi_u)\right)\Big|_{u=\tilde{u}} \qquad \text{for some } \tilde{u} \in (0,1).$$

Let $R(t)$ denote the function

$$R(t) = \pi_*^T Q(\pi_u) - \phi(\pi_u).$$

Then, using again the calculus made above,

$$\begin{aligned} R'(t) &= -\frac{t}{n} h^T I(\pi_u) h = -\frac{t}{n} E_{\pi_u}\left[\left((t(X) - \pi_u)^T I(\pi_u) h\right)^2\right] \\ R''(t) &= -\frac{1}{n} E_{\pi_u}\left[\left((t(X) - \pi_u)^T I(\pi_u) h\right)^2\right] + \frac{t}{n^{3/2}} E_{\pi_u}\left[\left((t(X) - \pi_u)^T I(\pi_u) h\right)^3\right] \end{aligned}$$

To get the second line, we use the following modification of (A.21):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}u}\left(I(\pi_u)h\right) &= \frac{\mathrm{d}}{\mathrm{d}u}\left(E_{\pi_u}\left[I(\pi_u)(t(X) - \pi_u)(t(X) - \pi_u)^T I(\pi_u) h\right]\right) \\ &= -\frac{1}{\sqrt{n}} E_{\pi_u}\left[I(\pi_u)(t(X) - \pi_u)\left((t(X) - \pi_u)^T I(\pi_u) h\right)^2\right] \end{aligned}$$

With similar arguments, we can derive again

$$\begin{aligned} R^{(3)}(t) = {} & \frac{2}{n^{3/2}} E_{\pi_u}\left[\left((t(X) - \pi_u)^T I(\pi_u) h\right)^3\right] + \frac{t}{n^2} E_{\pi_u}\left[\left((t(X) - \pi_u)^T I(\pi_u) h\right)^4\right] \\ & + \frac{3t}{n^2}\left(E_{\pi_u}\left[\left((t(X) - \pi_u)^T I(\pi_u) h\right)^2\right]\right)^2 \\ & - \frac{3t}{n^2} E_{\pi_u}\left[\left((t(X) - \pi_u)^T I(\pi_u) h\right)^2 (t(X) - \pi_u)^T\right] \\ & I(\pi_u) E_{\pi_u}\left[(t(X) - \pi_u)\left((t(X) - \pi_u)^T I(\pi_u) h\right)^2\right] \end{aligned}$$

$$\begin{aligned} |R^{(3)}(t)| \le {} & \frac{2}{n^{3/2}} \sqrt{\|J_{\pi_*}^T I(\pi_u) J_{\pi_*}\|}\ \left\|J_{\pi_*}^{-1} h\right\| \\ & \left\|E_{\pi_u}\left[J_{\pi_u}^{-1}(t(X) - \pi_u)\left((t(X) - \pi_u)^T I(\pi_u) h\right)^2\right]\right\| \\ & + \frac{4}{n^2} E_{\pi_u}\left[\left((t(X) - \pi_u)^T I(\pi_u) h\right)^4\right] \\ & + \frac{1}{n^2}\left\|E_{\pi_u}\left[J_{\pi_u}^{-1}(t(X) - \pi_u)\left((t(X) - \pi_u)^T I(\pi_u) h\right)^2\right]\right\|^2. \end{aligned}$$

Assumptions 3, 5 and 6 allow to conclude:

$$\sup_{h \in \mathcal{E}(M_n)} |B_n(h)| = o(1).$$

A.C MLE concentration

We prove here Lemma 30. Let us consider

$$Z_n = \left\| J_{\pi_*}^{-1} \Delta_n \right\|.$$

In other words, $Z_n^2 = \Delta_n^T I(\pi_*) \Delta_n$. We want to make an empirical process to appear in the expression of Z_n. Let us consider, for $1 \leq i \leq n$ and $a \in \mathbb{R}^{k_n}$,

$$Y_{i,a} = \frac{1}{\sqrt{n}} a^T J_{\pi_*}^{-1} \left(t(X_i) - E_{\pi_*} t(X_i) \right).$$

Then

$$Z_n = \sup_{a : \|a\| = 1} \left| \sum_{i=1}^{n} Y_{i,a} \right|$$

and the following relation, derived from Talagrand's inequality in [75, p. 170], is satisfied for any $x > 0$:

$$P_{\pi_*} \left(Z_n \geq E_{\pi_*} Z_n + 2\sqrt{(2v + 16b\, E_{\pi_*} Z_n)\, x} + 2b\, x \right) \leq e^{-x} \tag{A.22}$$

where

$$v = \sup_{a : \|a\| = 1} \sum_{i=1}^{n} E_\pi\, Y_{i,a}^2 = 1$$

$$b = \sup_{a : \|a\| = 1} \|Y_{1,a}\|_\infty \leq 2\sqrt{\frac{\|I(\pi_*)\|}{n}} \sup_{x \in \mathcal{X}} \|t(x)\|.$$

Let us recall (A.6):

$$E_{\pi_*} Z_n \leq \sqrt{k_n} = o(\sqrt{M_n}).$$

To conclude, we just choose $x = x_n$, with $x_n \to \infty$ and nevertheless $x_n = o\left(\sqrt{M_n}/b\right)$ and $x_n = o(M_n)$: this entails $E_{\pi_*} Z_n + 2\sqrt{(2v + 16b\, E_{\pi_*} Z_n)\, x} + 2b\, x = o(\sqrt{M_n})$.

A.D Distance in variation

We prove here Lemma 31.

$$\begin{aligned}
\left\| P_{\pi_*}^{\otimes n} - P_{n,\mathcal{E}(r_n)} \right\|_{\mathrm{TV}} &= \frac{1}{2} E_{\pi_*} \left| 1 - \int_{\mathcal{E}(r_n)} \frac{\mathrm{d}P_{n,h}}{\mathrm{d}P_{n,0}}(X_{1:n}) \frac{n^{-k_n/2} w\left(\pi_* + \frac{h}{\sqrt{n}}\right) \mathrm{d}h}{W\left(\pi_* + \frac{1}{\sqrt{n}} \mathcal{E}(r_n)\right)} \right| \\
&\leq \sup_{h \in \mathcal{E}(r_n)} \|P_{n,h} - P_{n,0}\|_{\mathrm{TV}} \\
&\leq \sup_{h \in \mathcal{E}(r_n)} \sqrt{D(P_{n,0}, P_{n,h})}
\end{aligned}$$

For n large enough, $r_n \leq M_n$ and for any $h \in \mathcal{E}(r_n)$,

$$\begin{aligned} D(P_{n,0}, P_{n,h}) &= E_{\pi *}\left[-\ln \frac{\mathrm{d}P_{n,h}}{\mathrm{d}P_{n,0}}(X_{1:n})\right] \\ &= \frac{1}{2}\sigma_n^2(h) - E_{\pi *}\left[\Delta_n^T (J_{\pi_*}^T)^{-1} A_n(h) + B_n(h)\right] \\ &\leq r_n + \sqrt{E_{\pi *}\left[\|J_{\pi_*}^{-1}\Delta_n\|^2\right]} \sup_{h \in \mathcal{E}(r_n)} \|A_n(h)\| + \sup_{h \in \mathcal{E}(r_n)} |B_n(h)| \end{aligned}$$

Since

$$E_{\pi *}\left[\|J_{\pi_*}^{-1}\Delta_n\|^2\right] = k_n,$$

a straightforward consequence of Proposition 23 is

$$\sup_{h \in \mathcal{E}(r_n)} D(P_{n,0}, P_{n,h}) \to 0 \quad \text{as} \quad n \to \infty.$$

Bibliographie

[1] J. Äberg, Y. M. Shtarkov, and B. J. M. Smeets. Multialphabet coding with separate alphabet description. In *Proc. of Compression and Complexity of Sequences*, pages 56–65. IEEE Comp. Soc., 1997.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd Internat. Symp. on Information Theory*, pages 267–281, 1973.

[3] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10 :245–279, 2009.

[4] K. Atteson. The asymptotic redundancy of Bayes rules for Markov chains. *IEEE Trans. Inf. Theory*, 45(6) :2104–2109, 1999.

[5] Z. Bai, C. R. Rao, and Y. Wu. Model selection with data-oriented penalty. *J. Statist. Plann. Inference*, 77(1) :102–117, 1999.

[6] A. R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In J. M. Bernardo, J. O. Berger, A. P. David, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 6, pages 27–52. Oxford Univ. Press, 1998.

[7] A. R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory*, 44(6) :2743–2760, 1998.

[8] A. R. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2) :536–561, 1999.

[9] A. R. Barron and Y. Yang. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5) :1564–1599, 1999.

[10] J.-P. Baudry. *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.* PhD thesis, Univ Paris-Sud, dec 2009.

[11] C. Biernacki, G. Celeux, and G. Govaert. Strategies for getting highest likehood in mixture models. Technical Report 4255, INRIA, September 2001.

[12] C. Biernacki, G. Celeux, and G. Govaert. Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *J. Statist. Plann. Inference*, 140 :2991–3002, 2010.

[13] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[14] D. Bontemps. Redondance bayésienne et minimax, sources stationnaires sans mémoire en alphabet infini. Master's thesis, Univ Paris-Sud, sep 2007.

[15] D. Bontemps. Universal coding on infinite alphabets : Exponentially decreasing envelopes. *IEEE Trans. Inf. Theory*, to be published.

[16] S. Boucheron, A. Garivier, and E. Gassiat. Coding on countably infinite alphabets. *IEEE Trans. Inf. Theory*, 55(1) :358–373, 2009.

[17] S. Boucheron and E. Gassiat. A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Statist.*, 3 :114–148, 2009.

[18] L. Breiman. The individual ergodic theorem of information theory. *Ann. Math. Statist.*, 28(3) :809–811, 1957.

[19] I. Castillo. A semi-parametric Bernstein-von Mises theorem. `http://www.proba.jussieu.fr/~castillo/bvm.pdf`.

[20] I. Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Statist.*, 12 :1281–1299, 2008.

[21] O. Catoni. *Statistical Learning Theory and Stochastic Optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, 2001. École d'été de Probabilités de Saint-Flour XXXI.

[22] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge Univ. Press, 2006.

[23] C. Chen, F. Forbes, and O. Francois. fastruct : model-based clustering made faster. *Molecular Ecology Notes*, 6(4) :980–983, 2006.

[24] Y. Choi and W. Szpankowski. Pattern matching in constrained sequences. In *2007 Int. Symp. Information Theory*, pages 2606–2610, Nice, 2007.

[25] B. Clarke and S. Ghosal. Posterior normality and reference priors for exponential families with increasing dimension. `http://www4.stat.ncsu.edu/~sghosal/papers/HighDimRefPr.pdf`.

[26] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory*, 36(3) :453–471, 1990.

[27] B. S. Clarke and A. R. Barron. Jeffrey's prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Inference*, 41 :37–60, 1994.

[28] J. Corander, P. Marttinen, J. Sirén, and J. Tang. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics*, 9 :539, 2008.

[29] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[30] I. Csiszár and J. Körner. *Information Theory : Coding Theorems for Discrete Memoryless Systems*. Academic Press, Inc., New York, USA, 1981.

[31] I. Csiszár and P. C. Shields. Redundancy rates for renewal and other processes. *IEEE Trans. Inf. Theory*, 42(6) :2065–2072, 1996.

[32] L. D. Davisson. Minimax noiseless universal coding for Markov sources. *IEEE Trans. Inf. Theory*, 29(2) :211–214, 1983.

[33] L. D. Davisson and A. Leon-Garcia. A source matching approach to finding minimax codes. *IEEE Trans. Inf. Theory*, 26 :166–174, 1980.

[34] C. de Boor. *A practical guide to splines*. Springer-Verlag, New York, 1978.

[35] R. de Jonge and J. H. van Zanten. Adaptive nonparametric Bayesian inference using location-scale mixture priors. `http://www.win.tue.nl/~jzanten/papers/paper_revision.pdf`.

[36] A. P. Dempster, N. M. Lairdsand, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Series B*, 39 :1–38, 1977.

[37] A. K. Dhulipala and A. Orlitsky. Universal compression of Markov and related sources over arbitrary alphabets. *IEEE Trans. Inf. Theory*, 52(9) :4182–4190, 2006.

[38] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1) :1–21, 1986.

[39] P. Diaconis and D. Freedman. Consistency of Bayes estimates for nonparametric regression : normal theory. *Bernoulli*, 4(4) :411–444, 1998.

[40] M. Drmota and W. Szpankowski. Precise minimax redundancy and regret. *IEEE Trans. Inf. Theory*, 50(11) :2686–2707, 2004.

[41] P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Inf. Theory*, 21(2) :194–203, 1975.

[42] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Inf. Theory*, 38(4) :1258–1270, 1992.

[43] P. Flajolet and W. Szpankowski. Analytic variations on redundancy rates of renewal processes. *IEEE Trans. Inf. Theory*, 48(11) :2911–2921, 2002.

[44] D. P. Foster, R. A. Stine, and A. J. Wyner. Universal codes for finite sequences of integers drawn from a monotone distribution. *IEEE Trans. Inf. Theory*, 48(6) :1713–1720, 2002.

[45] D. Freedman. Wald lecture : On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.*, 27(4) :1119–1141, 1999.

[46] R. G. Gallager. *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[47] A. Garivier. *Modèles contextuels et alphabets infinis en théorie de l'information.* PhD thesis, Univ Paris-Sud, nov 2006.

[48] A. Garivier. A lower bound for the maximin redundancy in pattern coding. *Entropy*, 11(4) :634–642, 2009.

[49] E. Gassiat. Codage universel et sélection de modèles emboités. Notes de cours de master.

[50] G. M. Gemelos and T. Weissman. On the entropy rate of pattern processes. *IEEE Trans. Inf. Theory*, 52(9) :3994–4007, 2006.

[51] C. R. Genoveve and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, 28(4) :1105–1127, 2000.

[52] S. Ghosal. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2) :315–331, 1999.

[53] S. Ghosal. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74 :49–68, 2000.

[54] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *Ann. Statist.*, 27 :143–158, 1999.

[55] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2) :500–531, 2000.

[56] S. Ghosal, J. Lember, and A. W. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electron. J. Statist.*, 2 :49–68, 2008.

[57] S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5) :1233–1263, 2001.

[58] S. Ghosal and A. W. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1) :192–223, 2007.

[59] S. Ghosal and A. W. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2) :697–723, 2007.

[60] L. Györfi, I. Páli, and E. C. van der Meulen. On universal noiseless source coding for infinite source alphabets. *Eur. Trans. Telecom.*, 4 :4–16, 1993.

[61] L. Györfi, I. Páli, and E. C. van der Meulen. There is no universal source code for an infinite source alphabet. *IEEE Trans. Inf. Theory*, 40(1) :267–271, 1994.

[62] D. Haussler. A general minimax result for relative entropy. Technical Report UCSC-CRL-96-26, University of California, UC Santa Cruz, CA 96064, 1996.

[63] D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. *Ann. Statist.*, 25(6) :2451–1492, 1997.

[64] D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proc. IRE*, 40(9) :1098–1101, September 1952.

[65] N. Jevtic, A. Orlitsky, and N. P. Santhanam. A lower bound on compression of unknown alphabets. *Theor. Comput. Sci.*, 332(1-3) :293–311, 2005.

[66] D. ke He and E. hui Yang. The universality of grammar-based codes for sources with countably infinite alphabets. *IEEE Trans. Inf. Theory*, 51(11) :3753–3765, 2005.

[67] J. C. Kieffer. A unified approach to weak universal source coding. *IEEE Trans. Inf. Theory*, 24(6) :674–682, 1978.

[68] Y. Kim. The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.*, 34(4) :1678–1700, 2006.

[69] Y. Kim and J. Lee. A Bernstein–von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.*, 32(4) :1492–1512, 2004.

[70] B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2) :837–877, 2006.

[71] L. G. Kraft. A device for quantizing, grouping and coding amplitude modulated pulses. Master's thesis, Dept. Electrical Engineering, MIT, Cambridge, MA, 1949.

[72] R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Inf. Theory*, 27(2) :199–207, 1981.

[73] E. K. Latch, G. Dharmarajan, J. C. Glaubitz, and O. E. J. Rhodes. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, 7(2) :295, 2006.

[74] É. Lebarbier. *Quelques approches pour la détection de rupture à horizon fini*. PhD thesis, Univ Paris-Sud, F-91405 Orsay, July 2002.

[75] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2007.

[76] C. Maugis and B. Michel. Slope heuristics for variable selection and clustering via Gaussian mixtures. Technical Report 6550, INRIA, 2008.

[77] C. Maugis and B. Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM : P&S*, 2009. accepted for publication.

[78] B. McMillan. The basic theorems of information theory. *Ann. Math. Statist.*, 24(2) :196–219, 1953.

[79] B. McMillan. Two inequalities implied by unique decipherability. *IRE Trans. Inf. Theory*, 2(4) :115–116, 1956.

[80] M. Nadif and G. Govaert. Clustering for binary data and mixture models : choice of the model. *Appl. Stoch. Models Data Anal.*, 13 :269–278, 1998.

[81] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. *IEEE Trans. Inf. Theory*, 52(7) :2954–2964, 2006.

[82] A. Orlitsky, N. P. Santhanam, and J. Zhang. Speaking of infinity [i.i.d. strings]. *IEEE Trans. Inf. Theory*, 50(10) :2215–2230, 2004.

[83] A. Orlitsky, N. P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Trans. Inf. Theory*, 50(7) :1469–1481, 2004.

[84] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2) :945–59, jun 2000.

[85] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inf. Theory*, 30(4) :629–636, 1984.

[86] J. J. Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM J. Res. Dev.*, 20(3) :198–203, 1976.

[87] V. Rivoirard and J. Rousseau. Bernstein von Mises Theorem for linear functionals of the density. `http://arxiv.org/abs/0908.4167`.

[88] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[89] C. Scricciolo. Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.*, 34(6) :2897–2920, 2006.

[90] C. Scricciolo. On rates of convergence for bayesian density estimation. *Scandinavian J. Statist.*, 34(3) :626–642, 2007.

[91] G. I. Shamir. Sequential universal lossless techniques for compression of patterns and their description length. In *Data Compression Conference*, pages 419–428, 2004.

[92] G. I. Shamir. On the MDL principle for i.i.d. sources with large alphabets. *IEEE Trans. Inf. Theory*, 52(5) :1939–1955, 2006.

[93] G. I. Shamir. Patterns of i.i.d. sequences and their entropy - part i : General bounds. *CoRR*, abs/cs/0605046, 2006.

[94] G. I. Shamir. Universal lossless compression with unknown alphabets - the average case. *IEEE Trans. Inf. Theory*, 52(11) :4915–4944, 2006.

[95] G. I. Shamir. Patterns of i.i.d. sequences and their entropy - part ii : Bounds for some distributions. *CoRR*, abs/0711.2102, 2007.

[96] G. I. Shamir. Universal source coding for monotonic and fast decaying monotonic distributions. *CoRR*, abs/0704.0838, 2007.

[97] G. I. Shamir and D. J. J. Costello. On the entropy rate of pattern processes. *IEEE Trans. Inf. Theory*, 50(8) :1620–1635, 2004.

[98] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27 :379–423,623–656, 1948.

[99] X. Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.*, 97(457) :222–235, 2002.

[100] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3) :687–714, 2001.

[101] P. C. Shields. Universal redundancy rates do not exist. *IEEE Trans. Inf. Theory*, 39(2) :520–524, 1993.

[102] Y. M. Shtarkov. Universal sequential coding of single messages. *Probl. Inf. Transm.*, 23(3) :175–186, 1987.

[103] W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Probl. Inf. Transm.*, 34(2) :142–146, 1998.

[104] W. Toussile and E. Gassiat. Variable selection in model-based clustering using multilocus genotype data. *Adv. Data Anal. Classif.*, 3(2) :109–134, September 2009.

[105] A. B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Mathématiques et Applications. Springer-Verlag, Berlin, 2004.

[106] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series Statist. Probab. Math. Cambridge Univ. Press, Cambridge, 1998.

[107] A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric Gaussian process methods. `http://www.win.tue.nl/~jzanten/papers/learning.pdf`.

[108] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3) :1435–1463, 2008.

[109] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B) :2655–2675, 2009.

[110] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, 1996.

[111] N. Verzelen. *Adaptative estimation to regular Gaussian Markov random fields*. PhD thesis, Univ Paris-Sud, Dec. 2009.

[112] F. Villers. *Tests et selection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, Univ Paris-Sud, 2007.

[113] S. G. Walker. New approaches to bayesian consistency. *Ann. Statist.*, 32(5) :2028–2043, 2004.

[114] S. G. Walker and N. L. Hjort. On bayesian consistency. *J. Royal Statist. Soc. Series B*, 63(4) :811–821, 2001.

[115] S. G. Walker, A. Lijoi, and I. Prünster. On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, 35(2) :738–746, 2007.

[116] M. J. Weinberger, N. Merhav, and M. Feder. Optimal sequential probability assignment for individual sequences. *IEEE Trans. Inf. Theory*, 40(2) :384–396, 1994.

[117] E. T. Whittaker and G. N. Watson. *A Course of Modern Analysis*. Cambridge Mathematical Library. Cambridge Univ. Press, Cambridge, fourth edition, 1927. Reprinted 1990.

[118] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method : Basic properties. *IEEE Trans. Inf. Theory*, 41(3) :653–664, 1995.

[119] Q. Xie and A. R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Inf. Theory*, 43(2) :646–657, 1997.

[120] Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inf. Theory*, 46(2) :431–445, 2000.

[121] Y. Xing. Convergence rates of nonparametric posterior distributions. `http://arxiv.org/abs/0804.2733`.

[122] Y. Xing. Convergence rates of posterior distributions for observations without the iid structure. `http://arxiv.org/abs/0811.4677`.

[123] Y. Xing. On adaptive Bayesian inference. *Electron. J. Statist.*, 2 :848–862, 2008.

[124] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, 23(3) :337–343, 1977.

[125] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory*, 24(5) :530–536, 1978.