



Model selection: a decision-theoretic approach

Aurélie Boisbunon

► To cite this version:

Aurélie Boisbunon. Model selection: a decision-theoretic approach. Statistics Theory [stat.TH]. Université de Rouen, 2013. English. NNT: . tel-00793898

HAL Id: tel-00793898

<https://theses.hal.science/tel-00793898>

Submitted on 28 Feb 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

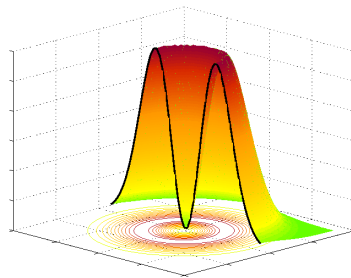
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Laboratoire d'Informatique,
de Traitement de l'Information et des Systèmes
Normandie Université
Université de Rouen

Thèse en vue de l'obtention du titre de
Docteur en Mathématiques de l'Université de Rouen

Sélection de modèle : une approche décisionnelle



Aurélié BOISBUNON

Jury :

Stéphane CANU	INSA Rouen	Directeur
Dominique FOURDRINIER	Université de Rouen	Directeur
Mohamed NADIF	Université Paris-Descartes	Rapporteur
Jean-Michel POGGI	Université Paris-Sud	Examineur
Alain RAKOTOMAMONJY	Université de Rouen	Examineur
Marten WEGKAMP	Cornell University	Rapporteur

Remerciements

Il y a beaucoup de gens que je voudrais remercier ici, j'espère n'en oublier aucun mais le risque est grand ... J'espère aussi que vous ne ferez pas trop attention à l'ordre utilisé, il n'y a pas de message caché à y voir.

Je voudrais commencer par mes directeurs de thèse, Stéphane Canu et Dominique Fourdrinier, qui m'ont tous deux beaucoup apporté pendant ces trois ans, chacun avec son point de vue différent. Ils m'ont beaucoup aidé à acquérir une plus grande rigueur dans ma démarche scientifique, et une plus grande pédagogie pour mes présentations, en mettant en valeur la “big picture” et le “take-home message”. Ils m'ont beaucoup encouragée et soutenue pendant ces trois ans, et je les en remercie énormément.

Je me tourne maintenant vers les membres de mon jury : mes rapporteurs Mohamed Nadif et Marten Wegkamp, qui m'ont fait des commentaires et des remarques très intéressantes (et agréables à lire !), et mes examinateurs Jean-Michel Poggi et Alain Rakotomamonjy, qui n'ont pas manqué d'ajouter une touche d'humour à ma soutenance.

Je voudrais également remercier William E. Strawderman et Martin T. Wells, avec qui j'ai eu le grand plaisir de travailler pendant mon séjour à Cornell, et avec qui j'ai eu également des discussions très intéressantes qui m'ont aidées à y voir plus clair.

Un grand merci aussi à Bruno Portier et Michel Bobbia, sans qui je n'aurais probablement pas eu l'idée de faire une thèse et grâce à qui j'ai eu la motivation d'en faire la démarche, et aux professeurs que j'ai eu pendant mon année de cours de Master Recherche, Guislaine Gayraud et Liliane Bel.

Et maintenant, tous ceux qui n'apparaissent pas dans ma thèse mais dont leur soutien m'a bien aidée à tenir le coup et à aller jusqu'au bout !

Une fois n'est pas coutume, je commence par le meilleur : Rémi, mon amour et mon plus grand soutien, tant par les nombreuses discussions scientifiques que nous avons eues ensemble, avec toutes ses idées géniales pour améliorer mon travail, que par sa présence au quotidien, en me remontant le moral dans les moments difficiles, en me changeant les idées, en me faisant rire et rêver ... Il faut parfois peu de choses (quelques cours de LaTeX et une boîte à outils) pour faire opérer la magie ;)

Ensuite, ma famille : mes parents, frère et sœur, beaux-parents, et famille plus large, qui m'ont tous énormément aidée dans mes projets et sans qui je n'en serais pas là ! Mes amis également, Eugénie et ses séances yoga-massage pour la dernière ligne droite, Chris et nos longues discussions sur nos états d'âmes, Adeline et Joannie avec nos bons quarts d'heure de déconnade, Nanou et Cocotte, mes amies de longue date, Ismaël, Kais et Hicham, avec entre autres un visionnage inoubliable de Total Recall. Et mes amis rencontrés en Espagne: ma très chère Joanna, Piero, Harry, Rob et Jana, avec qui j'ai passé des moments inoubliables.

Mes très chers collègues et anciens collègues de la salle A112 : Guillaume, Julien, Carlo, Zac, et Diana aussi (qui n'a peut-être pas son bureau à côté de nous mais c'est tout comme),

Polo, Damien, Seqvan, André, Ken et Gregory. Combien de parties de QPUC, de TLMVPSP, de Wordslide, de Geoscience et de quizz avons-nous fait, combien de trolls et débats lancés par Carlo en douce avant qu’il ne s’éclipse, commençant souvent par quelque chose du genre “95 % des gens ...” et engendrant parfois des heures de discussion ! Et les doctorants des autres salles aussi : Firas, Xilan, Abou, Benjamin, Abir, Jorge, Amnir, ...

Mes secrétaires préférées : Brigitte, Sandra, Fabienne, Chantal et Laure Paris, qui m’ont accueillie avec un grand sourire et tant de gentillesse à mon arrivée, et qui nous facilitent tellement la vie ! Un grand merci à vous !

Mes chers collègues d’ASI, du LITIS et de l’INSA : les deux Nico, Gillou, Alain, Romain, Géraldine, Pierrick, Elsa, Thierry, qui ont du attendre bien des fois au RU que je finisse mon assiette ; mais aussi Alexandre (toujours avec le sourire), Clément, Florence, Jean-François, Seb, Estelle, Sébastien, Laurent, Jean-Philippe, Nathalie, Fred, Michel, ... Mes collègues de l’université, avec qui j’ai malheureusement moins eu l’occasion de discuter mais dont j’apprécie tout autant la gentillesse : Maxime, Camille, Laurent, Thierry, Pierre, Sébastien, David, Philippe, ...

Les personnes que j’ai eu le grand plaisir de rencontrer au cours de ma thèse : Nicolas Vergne, Margherita Disertori, Nicolas Chenavier, et Mohamed Houda au LMRS, Mohamed Hebiri et Christophe Ambroise au sparsity workshop de Bristol, ainsi que les personnes que j’ai rencontrées à Cornell: mes collocataires Joanna, Becky, Alex et Zineb, et les gens du département de Stats Haim, Raj, Caitlin, Darcy, Irina, Kirsten, Jim, Robert, Bea, Shawna et Todd.

Je terminerai par d’autres personnes qui me tiennent également beaucoup à cœur : mes anciens collègues d’Air Normand, en particulier Céline et David que j’ai eu plusieurs fois l’occasion de voir depuis mon départ ; mes anciens collègues de la Conagua, la famille Ayala-Solano, Griselda, Martin, Emma, Jorge, Fausto, Jose Clemente ; et enfin mes amies de la danse, Céline, Lucille, Marion, Cécile, Diane, Isabelle, Élodie.

Un grand merci du fond du cœur à vous tous !

Résumé étendu

Cette thèse s'articule autour de la problématique de la sélection de modèle, étudiée en particulier dans le contexte de la régression linéaire. L'objectif est de déterminer, à partir de données mesurées, le meilleur modèle de prédiction parmi une collection de modèles. En d'autres termes, on cherche le modèle réalisant le meilleur compromis entre attache aux données et complexité du modèle.

Après l'introduction du Chapitre 1, le Chapitre 2 présente de façon formelle le problème de la sélection de modèle, ainsi que sa résolution avec une procédure générale en trois étapes : la première étape consiste à construire la collection des modèles à comparer ; la deuxième étape est définie à partir d'un critère d'attache aux données (souvent le risque empirique) dont l'optimisation permet de déterminer dans chaque modèle la fonction de prédiction qui semble représenter le mieux les données ; enfin, la troisième et dernière étape repose sur un deuxième critère, basé à la fois sur l'ajustement aux données et sur la complexité du modèle, qui compare les fonctions de prédiction obtenues à l'étape précédente et détermine ainsi le meilleur modèle de la collection.

La suite de ce chapitre passe en revue un certain nombre de méthodes proposées dans la littérature pour chacune des étapes, depuis les critères d'information d'Akaike (AIC) et de Bayes (BIC) jusqu'aux méthodes généralisées telles que la minimisation du risque structurel (SRC) ou l'heuristique de la pente pour l'évaluation des modèles, ainsi que les méthodes dites "stepwise" et les méthodes de régularisation parcimonieuses de type Lasso pour la construction d'une collection de modèles.

Les Chapitres 3 et 4 exposent notre principale contribution en matière d'évaluation des modèles. Nous proposons des critères basés sur des techniques de théorie de la décision, plus précisément l'estimation de coût. Ces critères, appelés estimateurs de coût, reposent sur une hypothèse distributionnelle plus large que l'hypothèse classique gaussienne avec indépendance entre les observations : la famille des lois à symétrie sphérique. Cette famille nous permet à la fois de nous affranchir de l'hypothèse d'indépendance et confère une certaine robustesse. En effet, nos critères ne dépendent pas de la forme spécifique de la distribution mais uniquement de la propriété de sphéricité.

Le Chapitre 3 commence par une brève introduction au principe de l'estimation de coût, puis se concentre sur les estimateurs sans biais du coût, c'est-à-dire tels que leur espérance est égale à l'espérance du vrai coût (ce qui équivaut au risque de la fonction de prédiction considérée). Nous rappelons d'abord les techniques utilisées dans le cas gaussien avec le théorème de Stein, théorème au coeur de l'estimation de coût, et présentons la dérivation de l'estimateur sans biais du coût dans ce contexte avec variance connue. Puis, nous étendons les résultats, d'abord au cas gaussien avec variance inconnue, puis au cas sphérique. Il s'avère que l'estimateur sans biais pour le cas sphérique est égal à celui obtenu dans le cas gaussien avec variance connue. De plus, il est équivalent au C_p de Mallows et à l'AIC avec bruit gaussien, ce qui permet d'expliquer une

certaine robustesse du C_p et de l'AIC face à la famille sphérique. Cependant, les limites connues de ces deux derniers critères, qui sont la sélection de modèles généralement trop complexes, s'étendent bien évidemment à notre estimateur sans biais du coût. Il est donc intéressant d'en chercher des améliorations, ce qui est proposé dans le Chapitre suivant.

Dans le Chapitre 4, nous nous intéressons au problème de la comparaison des critères d'évaluation de modèle et proposons de traiter ce problème au travers du *risque de communication*, mesurant la qualité du critère de façon analogue à l'erreur quadratique (MSE). Nous cherchons ainsi à améliorer l'estimateur sans biais du coût par des estimateurs biaisés mais ayant une variance plus faible, ce qui permet un meilleur contrôle des résultats. Nous proposons deux fonctions de correction additives, toutes les deux basées sur l'estimateur des moindres carrés pour des raisons de commodités mathématiques : la première prend en compte la sélection de variables en cours, tandis que la seconde se base essentiellement sur les variables non sélectionnées pour vérifier la qualité de la sélection. Les deux fonctions de correction proposées dépendent de constantes que l'on cherche à optimiser de façon à obtenir la plus petite différence de risque de communication entre l'estimateur corrigé et l'estimateur sans biais. Dans d'autres travaux sur l'estimation de coût, de telles constantes étaient déterminées exactement pour un estimateur des paramètres du modèle en particulier (estimateur des moindres carrés et estimateur de James-Stein). Dans notre contexte, nous n'avons pu fournir une telle expression des constantes pour tout estimateur. Nous proposons cependant de les estimer à partir des données en minimisant l'estimateur sans biais de la différence des risques de communication. Bien que cette méthode ne garantisse pas l'amélioration théorique de l'estimateur corrigé, l'étude numérique du Chapitre 6 montre que celui-ci peut conduire en pratique à une meilleure sélection de modèle. Le chapitre se termine par une comparaison de la théorie de l'estimation de coût avec d'autres théories générales sur la sélection de modèle, à savoir la théorie de l'apprentissage statistique (en anglais *Statistical Learning Theory*, *STL*) développée par [Vapnik 1998] et l'Heuristique de la pente, développée par [Birgé & Massart 2007].

Le Chapitre 5 s'intéresse aux aspects algorithmiques nécessaires à la comparaison des critères proposés en pratique : nous traitons tout d'abord le problème de la construction d'une collection de modèles, puis celui de la génération aléatoire de vecteurs sphériques pour vérifier la robustesse de nos critères.

Le problème de la construction de modèles est intimement lié à l'exploration des modèles possibles. En effet, si l'on dispose de p variables explicatives pour la prédiction de la variable d'étude, il existe 2^p sous-ensembles de variables à partir desquelles on peut construire le modèle de prédiction. Il est évident qu'on ne peut pas tous les tester dès que p dépasse 10 : tout l'enjeu consiste donc à déterminer une façon efficace d'explorer le moins de modèles possibles, tout en explorant suffisamment pour garantir une bonne solution. Parmi les techniques d'exploration utilisées en littérature, les algorithmes de chemin de régularisation permettent de déterminer les points de transition du problème d'optimisation concerné de façon à ajouter une à une les variables les plus corrélées avec le résidu. En particulier, [Efron *et al.* 2004] ont développé un tel algorithme pour le problème du Lasso, qui connaît un grand succès. Le Lasso est connu pour être un bon sélecteur, mais son biais d'estimation empêche l'obtention d'un bon modèle de prédiction à partir de critères basés sur l'erreur de prédiction. Nous nous sommes donc intéressés à une méthode utilisant le même principe de sélection que le Lasso, mais donnant des estimateurs moins biaisés : il s'agit de la pénalité concave minimax (en anglais *Minimax Concave Penalty*, *MCP*). La difficulté de cette méthode provient de la non convexité et non différentiabilité du problème

d'optimisation associé, rendant impossible l'application des outils usuels (sous-différentielles) utilisés pour déterminer le chemin de régularisation. La généralisation est possible au cas non convexe grâce à la notion de différentielle de Clarke. Bien qu'une telle différentielle soit en général difficile à calculer, les conditions d'optimalité peuvent être facilement dérivées dans le cas où le problème d'optimisation peut être décomposé en un terme non convexe différentiable et un terme convexe non différentiable, ce qui est le cas du MCP. Nous avons ainsi pu proposer un algorithme de chemin de régularisation pour le MCP, et il serait possible d'en faire autant pour d'autres problèmes non convexes (comme le SCAD, par exemple).

La seconde partie de ce chapitre porte sur la génération de vecteurs aléatoires sphériques. Celle-ci peut-être réalisée en écrivant la densité de la loi à symétrie sphérique soit comme un mélange de lois uniformes sur la sphère unité (ce qui est le cas pour toute distribution à symétrie sphérique), soit comme un mélange de lois gaussiennes centrées, quand cela est possible. La principale difficulté est de déterminer la loi de mélange dans les deux cas. Une fois cette difficulté surmontée, il est souvent aisé d'en prendre une transformation simple de loi univariée dont il existe de bons générateurs aléatoires.

Le Chapitre 6 présente une étude numérique comparant sur données simulées les performances de nos critères et des méthodes de la littérature exposées dans le Chapitre 2. L'étude numérique est divisée en deux temps : dans un premier temps, on sélectionne le meilleur modèle (oracle) de la collection à partir du vrai coût, et dans un deuxième temps, on remplace le vrai coût par un critère d'évaluation de modèle (nos estimateurs de coût et critères de la littérature).

L'objectif de la première partie de l'étude numérique est de vérifier la qualité des collections de modèles construites. En effet, la question posée ici est de savoir si le chemin utilisé pour explorer les modèles possibles passe par les modèles les plus intéressants, en particulier le vrai modèle lorsque celui-ci est accessible. Puisque le modèle est sélectionné à partir de la vraie erreur de prédiction, on a donc le meilleur modèle de prédiction possible dans la collection. Mais ce meilleur modèle, cet oracle, correspond-il au vrai modèle ? La réponse est positive pour les collections de modèles présentant les estimateurs les moins biaisés et dès que le nombre d'observation et/ou le rapport signal/bruit est assez grand. On peut donc en conclure qu'il est possible de bien sélectionner et bien prédire en même temps.

Dans la deuxième partie de l'étude, on cherche à déterminer s'il existe de bonnes procédures complètes de sélection de modèles, c'est-à-dire de choix adéquats entre collection de modèles et critère d'évaluation. Nous avons ainsi appliqué une quinzaine de critères d'évaluation de la littérature (dont certains dépendent d'un estimateur de la variance que l'on a également fait varier) à huit collection de modèles différentes. Nous avons ensuite sélectionné, pour chaque collection, les trois critères d'évaluation de modèles donnant les meilleures performances en sélection. Les résultats montrent que, dans le cadre de simulation que nous nous sommes fixés, on est capable de trouver des critères pour chaque collection de façon qu'il ne semble pas y avoir une méthode nettement plus performante qu'une autre en matière de sélection. En revanche, il est évident que le modèle ainsi sélectionné pour le Lasso est moins bon en prédiction que les autres, de par son large biais d'estimation, et il est donc nécessaire d'ajouter une étape post-sélection de modèle à une procédure basée sur le Lasso. Par ailleurs, notons que nos estimateurs de coût obtiennent des performances comparables à ceux de la littérature, et que les estimateurs corrigés obtiennent de meilleures performances en sélection que les estimateurs sans biais du coût. Enfin, il est intéressant de noter que les critères les plus couramment utilisés en pratique (validation croisée) ne sont pas toujours parmi les plus performants.

Le manuscrit se termine sur le Chapitre 7, qui présente les conclusions et perspectives de ces travaux.

Mots-clés : sélection de modèle, sélection de variable, régression linéaire, estimation de coût, distributions à symétrie sphérique, dépendance, Lasso, MCP.

La suite de ce manuscrit est en anglais.

Contents

1	Introduction	1
1.1	Why model selection?	1
1.1.1	The difficulty of predicting or understanding data	1
1.1.2	From predicting/understanding to model selection: the “divide and conquer” dogma	2
1.2	The principle of selecting among several models	2
1.2.1	Making compromises for the best of both worlds	2
1.2.2	A long and hazardous journey to the Truth...	3
1.3	Contributions	4
1.4	Overview of the manuscript	4
1.5	Publications	5
2	State of the art	11
2.1	The general problem of model selection	11
2.1.1	Predicting or explaining the data	12
2.1.2	The art of compromise	17
2.1.3	The general linear model	22
2.2	Estimating the prediction risk	25
2.2.1	The empirical risk	25
2.2.2	Analytical methods	27
2.2.3	Resampling methods	35
2.2.4	Which criterion should we choose?	36
2.3	Construction of the collection of models	38
2.3.1	Stepwise methods	38
2.3.2	Sparse regularization methods	40
2.3.3	Mixed strategies and other approaches	48
2.4	Summary of model selection procedures from literature	50
2.5	Contributions	52
2.5.1	A fairly large distributional framework with a dependence property	52
2.5.2	New criteria with lower risk	52
2.5.3	Numerical study and algorithms	53
3	Unbiased loss estimators for model selection	55
3.1	Origins of loss estimation theory	55
3.1.1	Stein’s Unbiased Risk Estimator (SURE)	56
3.1.2	From risk estimation to loss estimation	59
3.1.3	Loss estimation for model selection	60
3.2	The Gaussian case with known variance	63

3.2.1	Unbiased estimator of the estimation loss	63
3.2.2	Links with C_p , AIC and FPE	64
3.3	The Gaussian case with unknown variance	67
3.3.1	Unbiased estimator of the invariant estimation loss	69
3.3.2	Link with AIC_c	71
3.4	The spherical case	71
3.4.1	The class of multivariate spherically symmetric distributions	71
3.4.2	Unbiased estimator of the estimation loss	81
3.5	Summary	84
4	Corrected loss estimators for model selection	85
4.1	Improving on unbiased estimators of loss	85
4.1.1	A new layer of evaluation	85
4.1.2	Conditions of improvement over the unbiased estimator	87
4.1.3	Choice of the correction function	91
4.2	Corrected loss estimators for the restricted model	92
4.2.1	Condition for improvement with γ_r	92
4.2.2	Application to estimators of the regression coefficient	93
4.3	Corrected loss estimators for the full model	96
4.3.1	Condition for improvement with γ_f	96
4.4	Link with principled methods	99
4.5	Summary	100
5	Algorithmic aspects	103
5.1	Regularization path algorithms	103
5.1.1	Least Angle Regression algorithm for Lasso (LARS)	103
5.1.2	Algorithm for Minimax Concave Penalty	110
5.2	Random variable generation for spherically symmetric distributions	116
5.2.1	Through the stochastic representation	118
5.2.2	Through mixtures of other spherical distributions	122
6	Numerical study	127
6.1	How good is the oracle?	127
6.1.1	Purpose of the study	127
6.1.2	Sparse regularization paths versus stepwise methods	128
6.1.3	Replacing by other estimators	134
6.1.4	Discussion on the first study	136
6.2	Comparison of model evaluation criteria	138
6.2.1	Purpose of the study	138
6.2.2	Unbiased loss estimator vs corrected loss estimator	139
6.2.3	Comparison to existing methods from literature	147
6.2.4	Discussion on the second study	147
7	Conclusion et perspectives	153
7.1	Discussion on contributions and results	153
7.1.1	Summary on model evaluation	153
7.1.2	Summary on algorithmic and numerical aspects	155

7.1.3	Limitations of the present work	156
7.2	Perspectives and future works	157
7.2.1	Extension to elliptical symmetry	157
7.2.2	The Bayesian point of view	157
7.2.3	Other losses for comparing two model evaluation criteria	157
7.2.4	Application to classification and clustering	157
A	Appendix	159
A.1	Woodbury matrix update	160
A.1.1	Woodbury matrix identity	160
A.1.2	Update for adding a column and a line	160
A.1.3	Update for deleting a column and a line	161
A.2	Twice weak differentiability of the correction function γ_f	162
A.3	Subfunctions for LARS-MCP algorithm	164
A.4	Computing the degrees of freedom	165
A.4.1	Analytical form	165
A.4.2	Numerical computation	167
A.5	More results on the simulation study	168
A.5.1	Loss estimators with Student distribution	169
A.5.2	Loss estimators with Kotz distribution	172
A.5.3	Lasso	175
A.5.4	MCP	176
A.5.5	Adaptive lasso	177
A.5.6	Garrote	178
A.5.7	Elastic net	179
A.5.8	Adaptive Elastic net	180
A.5.9	Forward Selection	181
A.5.10	Backward elimination	182
	References	183

Introduction

Contents

1.1 Why model selection?	1
1.1.1 The difficulty of predicting or understanding data	1
1.1.2 From predicting/understanding to model selection: the “divide and conquer” dogma	2
1.2 The principle of selecting among several models	2
1.2.1 Making compromises for the best of both worlds	2
1.2.2 A long and hazardous journey to the Truth...	3
1.3 Contributions	4
1.4 Overview of the manuscript	4
1.5 Publications	5

This chapter introduces the reasons that stimulated research on the problem of model selection as well as the difficulties related to the resolution of this problem.

1.1 Why model selection?

The problem of model selection arises from many real-life situations and in a large range of domains such as Statistics, Machine learning, Bioinformatics, and so on. When little is known about the data under study, a common approach is to propose several representations and choose the best one. This process is known as model selection.

1.1.1 The difficulty of predicting or understanding data

Let us begin with two real-life examples.

The first example is related to Brain-Computer Interfaces (BCI) that aims at controlling a machine, such as a wheeling chair or a computer, with the mind only. One of the objectives underlying this problem is a better understanding of the brain, especially the zones that are activated depending on the mental state or task (see [Dornhege *et al.* 2007]).

The second example is concerned with the prediction of the level of ozone contamination in the atmosphere on the following day. This problem is a public health concern since a high concentration of ozone might result in a (temporary) decrease in physical capacities, especially for asthmatics, children and elderly people. Hence, on days with high ozone contamination, outdoor and/or physical activities should not be held, so that its prediction is important in order to warn the population of the dangers. It is well known that ozone is the product of the

reaction between nitrogene oxides (other pollutants) and sunlight. However, the prediction of ozone cannot be based on these last quantities because they themselves are difficult to predict. An alternative solution is to consider other quantities related to nitrogene oxides and sunlight and much easier to predict. For instance, the temperature is highly dependent on the amount of sunlight and can thus be used as a proxy. Other meteorological quantities favor or reduce the formation of nitrogene oxides and their amount might thus be related to the amount of ozone. In this example, the objective is not to understand the system underlying the production of ozone, which is well known, but rather to modelize the dependencies between ozone and meteorological quantities which are easier to predict. Since these relations are indirect, they are hard to modelize.

The problems in both examples are hard to solve because of the scarce knowledge we have on it. Next paragraph describes a way to cope with the difficulty, namely model selection.

1.1.2 From predicting/understanding to model selection: the “divide and conquer” dogma

The difficulty in the tasks of predicting or understanding the data at hand comes from the ignorance on the relations between the quantities under study. A common way to overcome this problem is to propose several representations of the data modelizing these relations differently. The problem thus becomes one of evaluating the representations, which we call *models* hereafter, comparing them and selecting the “best” one.

Hence, model selection arises as a simplification of other problems since the choice is reduced from an infinite number of possible models to a finite number. However, it is itself a hard problem to solve. Indeed, it requires the definition of a *good* model, and of a measure of its quality. Such definitions should be specified in adequacy with the main objective of the study. This might seem obvious, but in practice some methods used to construct the models or to evaluate them are sometimes inappropriate with the main objective. We recall the main possible objectives [Hocking 1976]:

1. the *description* of the dataset;
2. the *prediction* of future instances of the quantity under study;
3. the *estimation* of statistics linked to that quantity, such as its mean, median, and standard deviation;
4. the *extrapolation* on points in between observations;
5. the *estimation of the paramaters* specifying a model;
6. the *control* of the quantity under study;
7. and *model building*, useful to understand the relationship between quantities.

The rest of the manuscript focuses on the objective of good prediction.

1.2 The principle of selecting among several models

1.2.1 Making compromises for the best of both worlds

With the increasing memory of computers, the datasets are becoming larger and larger, with hundreds, thousands or even millions of observations of the quantity of interest and hundreds,

thousands or millions of quantities that can help in its prediction, which we call explanatory variables hereafter. When practitioners can afford studies with such large datasets, they try to capture as much information as they can so that nothing can be missed.

On the other hand, when using classical statistical tools, the models that best fit the data are the more complex ones, where complex here refers to both the number of explanatory variables used in the model and the form of their link to the quantity under study (for instance, we could use a linear or a polynomial link to model the relationship between the level of ozone and, say, the temperature). Indeed, even though it is always possible to find a model that fits exactly the data observed, such a model will give poor performances of prediction on future instances of the quantity under study.

The challenge in model selection thus consists in finding the model that will make the best tradeoff between model fitting (on the available data) and complexity, so that it can give a good prediction on future observations.

1.2.2 A long and hazardous journey to the Truth...

Several difficulties arise from the process of model selection. We briefly introduce each of them in the following paragraphs.

Measuring the complexity of a model. When the different models have a similar form, for instance they all estimate the link between the explanatory variables and the variable under study to be linear, the complexity can easily be measured by the number of explanatory variables kept in the models and considered as relevant for the prediction problem. However, it is much more difficult to measure the complexity between two models with the same number of explanatory variables, but where for instance one model fits a linear link and the other estimates a polynomial link to the output variable, or when both models are of the same form but one also models possible interactions between explanatory variables. As we can see, a good measure of complexity should take into account the shape of the model, the number of its components, and the possible interactions between its components. But how can we assess such a measure in practice?

Constructing several models that approximate the true underlying system. Another difficulty arises from the construction of the different models representing the data. How many models should we construct? And how should they be constructed? Is it better to construct as many models as the computer can support and be sure that at least one of them is close to the truth, or is it better to construct only a few good ones?

Evaluating and comparing the models. Once the models have been constructed and a measure of the complexity has been defined, the problem still remains to evaluate their qualities by taking into account both their abilities to fit the observed data and their complexity. There exist many criteria in the literature on model selection that manage to do so, but the question then becomes: which one to choose? What are their properties?

All these questions have been tackled in the literature and still continue to be treated, but the solutions are only partly satisfactory and the problem remains open.

1.3 Contributions

In this section, we briefly introduce our contributions to the different subproblems in model selection.

Evaluating the quality of a model and measuring its complexity. In Statistics, the quantities observed in a dataset are often considered to be random variables in order to take into account the uncertainty or the scarcity of knowledge on the underlying system. The existing methods for evaluating the quality of a model often rely on strong assumptions modelizing the randomness of the variables, and therefore they might be sensitive to extreme values. On the opposite direction, some methods make very little assumption on the randomness, but their inherent generality might alter their performances.

We derive criteria for model evaluation that are based on a larger distributional assumption than the former methods, but a tighter one than the latter methods. To be more precise, we model the randomness by *spherically symmetric distributions*. This family of distributions handles some form of dependence between the observations of the quantity under study, an assumption that is seldom made in the literature. Also, our criteria only rely on the spherical property, and they have the same form whatever the distribution is, as long as it is spherically symmetric. This feature gives distributional robustness to our method.

Comparing model evaluation criteria. We provide a new way to evaluate the quality of a model evaluation criterion, which is based on its distance to the theoretical quantity evaluating the performance in prediction of a model. This additional level of evaluation allows the comparison between two model evaluation criteria as well as it results in the derivation in new better criteria.

We also compare our criteria to existing methods in a simulation study.

Constructing models. Finally, we address the problem of constructing models by proposing algorithms and by investigating their appropriateness to the main objective of the study – in our case good prediction – through simulation study. We propose to examine which methods for constructing models and which methods for evaluating them give together the best performances of prediction.

1.4 Overview of the manuscript

The rest of the manuscript is organized as follows.

Chapter 2 develops more deeply the problem of model selection and reviews existing methods on both the evaluation and the construction of models.

Chapter 3 is devoted to our first contributions with unbiased criteria in several distributional settings: the Gaussian assumption with known variance, the Gaussian assumption with unknown variance, and the spherical assumption. The major advantage of our criteria is that they do not rely on the special form of the distribution and thus have the same expression whatever the distribution is, as soon as it is spherically symmetric.

Chapter 4 addresses the problem of comparing two criteria. The comparison is performed on a theoretical level, looking at the (quadratic) risks of the criteria and choosing the one with the

lower value of risk. This additional level of evaluation results in the derivation of better criteria for model selection.

In both Chapter 3 and Chapter 4, we relate our criteria to existing methods.

Chapter 5 is concerned with algorithmic aspects that are useful for the simulation study. The first algorithmic aspect is the proposition of a regularization path algorithm for the Minimax Concave Penalty (MCP), a nonconvex optimization method leading to a sparse and nearly unbiased estimator of the coefficient in linear regression. In the second part of the chapter, we look at the problem of generating spherically symmetric random vectors, as it is our main distributional assumption.

Chapter 6 presents two simulation studies: one on the adequacy of methods constructing models to the objective of prediction, and one where we compare our model evaluation criteria to those of the literature.

Finally, Chapter 7 closes the manuscript with some conclusions and discussions and develops the perspectives for future works.

1.5 Publications

Papers in progress

- [1] A. Boissunon, S. Canu, D. Fourdrinier, W.E. Strawderman, M.T. Wells, "AIC and C_p as loss estimators for spherically symmetric distributions", Work in progress, September 2012.

Conferences and workshops

- [1] A. Boissunon, S. Canu, D. Fourdrinier, "A New Procedure for Model Selection", *Workshop on Co-clustering and Model Selection (ClasSel)*, February 16 2012.
- [2] A. Boissunon, S. Canu, D. Fourdrinier, "Criteria for variable selection with dependence", *Workshop on New Frontiers in Model Order Selection, NIPS 2011*. Video¹.
- [3] A. Boissunon, S. Canu, D. Fourdrinier, "Critères robustes de sélection de variables dans le modèle linéaire via l'estimation de coût", *Proceedings of the 18th Conference of the Franco-phone Clustering Society, 2011*.
- [4] Boissunon, A., Canu, S., Fourdrinier, D., Strawderman, W. E. and Wells, M. T. "Variable selection: a decision theory approach", Invited talk, 39^{ème} congrès annuel de la Société de Statistique du Canada Acadia University, Wolfville, N.-É., Canada, June 12-15 2011.

Technical reports

- [1] A. Boissunon, S. Canu, D. Fourdrinier, "A global procedure for variable selection", Technical report.
- [2] A. Boissunon, S. Canu, D. Fourdrinier, "A trade-off between Gaussian and worst case analysis for model selection", Technical report.

¹ http://videolectures.net/aurelie_boissunon/

- [3] A. Boissunon, S. Canu, D. Fourdrinier, "Sélection de variables dans le modèle linéaire", *Report for the midterm review of project ClasSel, 2010.*

Notations and Acronyms

Sets and spaces

\mathbb{R}	Set of real numbers
\mathcal{Y}	Output space
\mathcal{X}	Input space
\mathcal{F}	Function space
\mathcal{P}	Space of probability distributions
\mathcal{M}	Model

Variables and observations

$y \in \mathbb{R}$	Output random variable
$x \in \mathbb{R}^p$	Input random vector
$x^j \in \mathbb{R}$	j^{th} explanatory variable of x
$Y \in \mathbb{R}^n$	Output random vector
$X \in \mathbb{R}^{n \times p}$	Input random matrix
$Y_i \in \mathbb{R}$	i^{th} component of the vector Y
$X_i \in \mathbb{R}^p$	i^{th} row of the matrix X
$X^j \in \mathbb{R}^n$	j^{th} explanatory variable of X
$X_{i,j} \in \mathbb{R}$	Element (i, j) of the matrix X
$\mathbf{y} \in \mathbb{R}$	Observation of y
$\mathbf{x} \in \mathbb{R}^p$	Observation of x
$\mathbf{Y} \in \mathbb{R}^n$	Observation of Y
$\mathbf{X} \in \mathbb{R}^{n \times p}$	Observation of X
$e \in \mathbb{R}$	Noise
$\varepsilon \in \mathbb{R}^n$	Noise vector

Probabilities and expectations

σ	Noise level
Σ	Covariance matrix
$\mathbb{P}_{x,y}$	Probability distribution of the couple (x, y)
$\mathbb{P}_{y x}$	Conditional probability distribution of y given x
\mathbb{E}_y	Expectation under the distribution of y
\mathbb{E}_β	Expectation under the distribution of y parametrized by β
$\mathbb{E}_{\beta, \sigma^2}$	Expectation under the distribution of y parametrized by (β, σ^2)
cov	Covariance
$p(y)$	Density of y
$p(y \beta)$	Density of y parametrized by β
\mathcal{I}	Fisher Information matrix

Distributions

$\mathcal{N}(\mu, \sigma^2)$	Gaussian univariate distribution with mean μ and variance σ^2
$\mathcal{N}_n(\mu, \Sigma)$	Gaussian multivariate distribution with mean μ and covariance matrix Σ
$\mathcal{T}_n(\mu, \Sigma)$	Student multivariate distribution with mean μ
\mathcal{L}	Laplace univariate distribution
\mathcal{L}_n	Laplace multivariate spherical distribution
\mathcal{K}_n	Kotz distribution

Functions and operators

f	Target function
\hat{f}	Estimator of the target function f
\hat{y}	Prediction for y
\hat{L}	Estimator of loss
\hat{L}_0	Unbiased estimator of loss
γ, ζ	Correction functions
\hat{L}_γ	Corrected estimator of loss
$C(\hat{f})$	Complexity of \hat{f}
$\mathbb{1}$	Indicator function
$\ \cdot\ , \ \cdot\ _2$	Euclidean norm
$\ \cdot\ _1$	ℓ_1 -norm
$\ \cdot\ _A$	ℓ_1 -norm
$ \cdot $	Absolute value
$\text{diag}(\mathbf{M})$	Diagonal elements of the matrix \mathbf{M}
$\text{tr}(\mathbf{M})$	Trace of the matrix \mathbf{M}
sgn	Sign function
div	Divergence operator
∇	Gradient operator
Δ	Laplacian operator
$\mathbf{J}_f(t)$	Jacobian matrix of a function f at point t
rank	Rank
$\#$	Cardinal

Losses and risks

$l(\hat{f}(x), y)$	Loss function in univariate case
$L(\hat{f}(X), Y)$	Loss function in multivariate case
$R_{(x,y)}(\hat{f})$	Univariate prediction risk
$R_{y x}(\hat{f}, \mathbf{x})$	Univariate conditional prediction risk
$R_{(X,Y)}(\hat{f})$	Multivariate prediction risk
$R_{Y X}(\hat{f}, \mathbf{X})$	Multivariate conditional prediction risk
$R_{\text{emp}}(\hat{f})$	Empirical risk
$\mathcal{L}(\hat{\beta}, \beta, \hat{L})$	Communication loss
$\mathcal{R}(\hat{\beta}, \hat{L})$	Communication risk

Miscellaneous

I_n	Identity matrix of size $n \times n$
I, J	Subsets
λ	Hyperparameter
df	Degrees of freedom
\widehat{df}	Generalized degrees of freedom
D_{KL}	Kullback-Leibler divergence

Acronyms

VC-dim	Vapnik-Chervonenkis dimension
MSE	Mean Squared Error
SSE	Sum of Squared Error
PR	Prediction Error
LS	Least-Squares
ML	Maximum Likelihood
JS	James-Stein
GJS	Generalized James-Stein
RR	Ridge Regression
AIC	Aikaike Information Criterion
FPE	Final Prediction Error
AIC_c	Corrected Aikaike Information Criterion
CV	Cross Validation
LOOCV	Leave-One-Out Cross Validation
GCV	Generalized Cross Validation
BIC	Bayes Information Criterion
SBC	Schwarz Bayes Criterion
TIC	Takeuchi Information Criterion
RIC	Risk Inflation Criterion
CAIC	Consistent AIC
HQ	Hannan and Quinn criterion
CAICF	Consistent AIC with Fisher matrix
ICOMP	Information Complexity Criterion
SRM	Structural Risk Minimization
SH	Slope Heuristics
SURE	Stein's Unbiased Risk Estimator
Lasso	Least Absolute Shrinkage and Selection Operator
MCP	Minimax Concave Penalty
SCAD	Smoothly Clipped Absolute Deviation
Adalasso	Adaptive Lasso
Enet	Elastic net
Adanet	Adaptive Elastic net
HT	Hard Threshold
ST	Soft Threshold
FS	Firm Shrinkage
NP	Non Polynomial

State of the art

Contents

2.1 The general problem of model selection	11
2.1.1 Predicting or explaining the data	12
2.1.2 The art of compromise	17
2.1.3 The general linear model	22
2.2 Estimating the prediction risk	25
2.2.1 The empirical risk	25
2.2.2 Analytical methods	27
2.2.3 Resampling methods	35
2.2.4 Which criterion should we choose?	36
2.3 Construction of the collection of models	38
2.3.1 Stepwise methods	38
2.3.2 Sparse regularization methods	40
2.3.3 Mixed strategies and other approaches	48
2.4 Summary of model selection procedures from literature	50
2.5 Contributions	52
2.5.1 A fairly large distributional framework with a dependence property	52
2.5.2 New criteria with lower risk	52
2.5.3 Numerical study and algorithms	53

This chapter reviews a number of existing methods related to the problem of model selection. Section 2.1 begins with an overview of the problem with its formalization through the notion of loss and risk, and ends with the assumptions we will keep in the sequel. Then, some of the methods for evaluating and comparing models are enumerated and explained in Section 2.2. Section 2.3 reviews methods for constructing models and collections of models. Section 2.4 presents a summary of global procedures of model selection (that is, with a solution for both the construction of models and the evaluation of these models). Finally, Section 2.5 introduces our contributions and where they stand in the picture of model selection relatively to existing methods.

2.1 The general problem of model selection

The problem of model selection is closely related to the problem of prediction or understanding of a quantity of interest. Because of that relation, we first begin by exposing the problem of prediction/explication and then we move to the actual problem of model selection.

2.1.1 Predicting or explaining the data

Context and notations

Let y be a quantity of interest. For instance, y can be the concentration of pollutants such as ozone or particles in the atmosphere, and we wish to predict its value for the following day (see [Poggi & Portier 2011]). If this value is high, then the authorities alert the population and advice that outdoor activities should be canceled. In this example, y takes real positive values. Another example is the diagnosis of a patient (see [Friedman 1994]). Here, y can be qualitative with values in {"healthy", "sick"} or with the type of disease. These values generate a distinction between patients and lead to a classification. One objective is thus to predict to which class belongs a new patient. Another objective could be to understand what causes the disease (environmental, clinical or genetical aspects). We formalize both examples by saying that y belongs to a space \mathcal{Y} , that can be (a subset of) \mathbb{R} or $\{0, 1\}$. The variable y will be referred to as the *output variable* or the *study variable*.

Now, in both the examples given, there exist some relations to other quantities. It is indeed well known that the concentration of ozone is higher on sunny and warm days with dense traffic, as ozone is the result of the reaction between nitrogene oxides (produced by cars, among others) and sunlight. However, it might be difficult to predict the concentration of nitrogene oxides as well as the amount of sunlight on the following day. On the other side, the sunlight also causes the temperature to increase, and thus we might take instead measures of temperature as a surrogate of sunlight for predicting the concentration of ozone on the following day. Therefore, we might not look for causal relations, but rather for dependencies [Friedman 1994]. The quantities used in the process of predicting or explaining the output y are called *explanatory variables* or *input variables* and are written as $x = (x^1, \dots, x^p) \in \mathcal{X}$, p being the number of input variables. Each variable x^j can be either quantitative (x^j belongs to a subspace of \mathbb{R}) or qualitative (x^j belongs to $\{0, 1\}$ if it is a binary variable, or to a set of qualitative values if it is a categorical variable, such as {"small", "medium", "large"} for instance [Hastie et al. 2005]), so that \mathcal{X} can be very different from one problem to another.

Figure 2.1 displays the system involving both x and y . It also involves extra information contained in other variables, denoted here by x_0 , which are not observable, unlike x . These unobservable variables can be of different nature, such as measurement error [Hastie et al. 2005], or the symptoms that a patient failed to notice or to notify as relevant in the medical diagnosis example [Friedman 1994]. Also, a patient might not explain its symptoms in the same way depending on its mood. Other examples of unobservable variables can be found in [Cherkassky & Mulier 1998]. In Statistics and Machine Learning, it is thus common to think of y as a random variable to account for the unobservable information. As far as x is concerned, the literature is split between the assumption that x is fixed and the assumption that it is random. The x -fixed case assumption offers a simplification over the x -random case, and is thus often adopted even in cases where it is not grounded. In order to be general and include both cases in this state of the art, we will thus first take the couple (x, y) to be generated by a joint probability distribution $\mathbb{P}_{x,y}$, and then specify our assumptions in Subsection 2.1.3. The distribution $\mathbb{P}_{x,y}$ belongs to a subspace \mathcal{P} of the space of all probability distributions on $\mathcal{X} \times \mathcal{Y}$. We will also give specifications on \mathcal{P} in the sequel, due to some necessary restrictions. Note that, when we observe a sample of n instances $(\mathbf{X}, \mathbf{Y}) = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ of the couple (x, y) , it is generally assumed that each observation $(\mathbf{x}_i, \mathbf{y}_i)$ is identically distributed as $\mathbb{P}_{x,y}$, and it is also often assumed that the observations are independent. However, when the independence assumption does not seem reasonable, one way

to model possible correlations or dependences between the observations is to consider the data (\mathbf{X}, \mathbf{Y}) as one observation of the couple (X, Y) , where $X \in \mathcal{X}^n$ and $Y \in \mathcal{Y}^n$, an assumption that we will often use in the sequel.

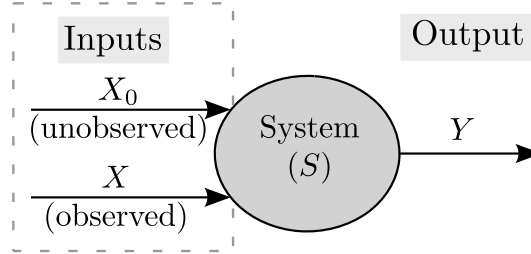


Figure 2.1: True underlying system generating the output Y . Note however that the relation between X , X_0 and Y might not be causal (see [Friedman 1994]).

As mentioned earlier, the objective is either to predict or to explain the variations of y based on x . In order to do so, the Statistician generally assumes that there exists a target function (or a target functional parameter) $f : \mathcal{X} \mapsto \mathcal{Y}$ modelizing the link between x and y and aims at determining an estimator \hat{f} of the target f (more details will be given the following subsections). However, there exists an infinity of functions \hat{f} mapping from \mathcal{X} to \mathcal{Y} . The objective is thus to find the one that seems to be the best in view of the data observed. Indeed, the stated objective hides the true problem that we can sum up as follows (see for instance [Friedman 1994] or [Breiman 1996]): **does the estimator \hat{f} give an accurate approximation to the underlying relationship between x and y ?**

In practice, one often fixes a form or a structure on \hat{f} , that is, one often takes \hat{f} in a class of functions. The restricted class of functions is called *model*, or *hypothesis space* (see [Niyogi & Girosi 1996]). For instance, in regression problems where y takes real values, one might look for linear functions $\hat{f}(x) = x^t \hat{\beta}$, $\hat{\beta}$ being the regression coefficient in \mathbb{R}^p . In such a model, we can also assume that not all the variables in x are relevant, especially when p is quite large, and thus we will look for linear models of lower sizes. The shape or structure of \hat{f} as well as the maximum number of variables it needs as an input are examples of what is called the *complexity* of the function, and more generally the complexity of the model. We can thus define a little more precisely what is a model.

Definition 2.1 (Model). *A model \mathcal{M} is a class of functions $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}$ sharing a similar form or structure and having a fixed maximum complexity.*

The notion of complexity is however only a vague notion and can be measured in several ways, as we will see in Subsection 2.1.2. As we do not know to which model belongs the true underlying model for the system (S) , it could be thought that taking the largest possible model increases the probability to include the true one. However, at the same time, it would also increase the error induced by fitting a complex model on a finite sample. Hence, a common practice is to consider several models $\mathcal{M}_1, \dots, \mathcal{M}_M$ with different complexities, and the objective is to select the best model among the list. This leads to the following definition for model selection, that can be found in a similar formulation in [Guyon et al. 2010] for instance.

Definition 2.2 (Model Selection). *Model selection is the process of evaluating several models $\mathcal{M}_1, \dots, \mathcal{M}_M$ and comparing them to determine which one best predicts or best explains the*

data.

From this definition, the challenge is to formalize the notions of “best predicts” and “best explains” so as to find the model that realizes the best tradeoff between goodness of fit and complexity. Note that, although the problems of prediction and explication (identification of the underlying system) can be solved in the same way, the latter one is a much more complicated task and is hard to formalize [Friedman 1994]. We concentrated our research on the prediction task only.

The notion of accurate prediction has been well formalized through the notion of loss and risk that we develop in the following paragraph. The rest of this section provides a deeper insight into the topic and the notions seen so far, and is essentially based on the following references: [Friedman 1994], [Hastie *et al.* 2008], [Vapnik 1998] and [Cherkassky & Mulier 1998].

Formalization: prediction loss and prediction risks

The formalization of accuracy of a prediction through losses and risks is now generally accepted by researchers from both the Statistics and Machine Learning fields (see [Massart 2007] and [Guyon *et al.* 2010], among others). Note that the point of view we expose in the sequel is a frequentist point of view. An interesting review and discussion on model selection from the Bayesian perspective can be found in [George 2000].

Given $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}$ a function that modelizes the link between x and y , its accuracy can be expressed through the choice of a function

$$\begin{aligned} l & : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_+ \\ (\hat{f}(x), y) & \mapsto l(\hat{f}(x), y) \end{aligned} \quad (2.1)$$

such that $l(\hat{f}(x), y)$ is integrable for any $\hat{f}(x) \in \mathcal{Y}$ and such that $\hat{f}(x) = y$ belongs to the set of functions minimizing l , that is,

$$y \in \arg \min_{\hat{f}(x) \in \mathcal{Y}} l(\hat{f}(x), y).$$

The function l is called a *prediction loss function*, and accounts for the cost incurred by an approximation $\hat{f}(x)$ of y . Often, we choose l so that $l(y, y) = 0$. Hence a *good* prediction is one that has a value of loss $l(\hat{f}(x), y)$ close to 0.

Example 2.1 gives the three losses most commonly used in practice.

Example 2.1. In regression problems, it is common to take l as a function of a distance or a norm, the most common examples being the *squared-error loss* defined as the squared Euclidean distance

$$l(\hat{f}(x), y) = (y - \hat{f}(x))^2, \quad (2.2)$$

and the *absolute error loss* defined as

$$l(\hat{f}(x), y) = |y - \hat{f}(x)|. \quad (2.3)$$

In binary classification ($\mathcal{Y} = \{0, 1\}$ or $\{0, 1\}^n$), it is common to take the *0-1 loss* defined by the discrete metric

$$l(\hat{f}(x), y) = \mathbb{1}_{\{\hat{f}(x) \neq y\}} = \begin{cases} 0 & \text{if } \hat{f}(x) = y, \\ 1 & \text{if } \hat{f}(x) \neq y. \end{cases} \quad (2.4)$$

However, defining the loss $l(\hat{f}(x), y)$ as a metric is not the only choice, and one might want to use an asymmetric function and give a different cost for a prediction lower than the true value y than for a higher prediction (see examples of asymmetric losses in [Berger 1985]). Asymmetric losses can be useful in problems such as prediction in wind energy production, where some governments apply fees that are lower for underpredicting than for overpredicting [Pinson *et al.* 2004].

Now, the output y we wish to predict is assumed to be random, so the actual loss is not considered to be a good criterion of the quality of a prediction $\hat{f}(x)$. Indeed, \hat{f} might give a good prediction for one instance of y , but a poor prediction for the following instance, so that \hat{f} might not be so good overall. Instead, we consider the following risks, namely the *prediction risk* or the *conditional prediction risk*

$$R_{(x,y)}(\hat{f}) = \mathbb{E}_{(x,y)}[l(\hat{f}(x), y)] = \int_{\mathcal{X} \times \mathcal{Y}} l(\hat{f}(x), y) d\mathbb{P}_{x,y}(x, y) \quad (2.5)$$

$$R_{y|x}(\hat{f}, \mathbf{x}) = \mathbb{E}_{y|x}[l(\hat{f}(x), y)|x = \mathbf{x}] = \int_{\mathcal{Y}} l(\hat{f}(x), y) d\mathbb{P}_{y|x}(y). \quad (2.6)$$

The prediction risk $R_{(x,y)}$ is also called the *expected prediction error*, and the conditional prediction risk $R_{y|x}$ is also often referred to as the *generalization error*. Note that these two risks are related in the following way

$$R_{(x,y)}(\hat{f}) = \mathbb{E}_x[R_{y|x}(\hat{f}, x)],$$

where \mathbb{E}_x is the expectation under the marginal probability of x . We will sometimes refer to both of them as the prediction risks, applying for either of them. Both criteria express the *average* accuracy of a prediction. Again, a good approximation should have low (conditional) risk.

Note that Equation (2.5) defining the risk implies a restriction on \mathcal{P} to the space of all probability distributions for which the risk $R_{(x,y)}(\hat{f})$ is finite, and the same occurs with \mathcal{P}_x containing $\mathbb{P}_{y|x}$ restricted to the space with existing conditional risk $R_{y|x}(\hat{f}, \mathbf{x})$. In particular, Equations (2.5) and (2.6) require the existence of the density function as well as the existence of some moments. For instance, with the squared-error loss, we need

$$\int_{\mathcal{X} \times \mathcal{Y}} y^2 d\mathbb{P}_{x,y}(x, y) < \infty \quad \text{and} \quad \int_{\mathcal{X} \times \mathcal{Y}} \hat{f}^2(x) d\mathbb{P}_{x,y}(x, y) < \infty$$

for the corresponding risk to be finite.

The multivariate case. In the case where we consider the random matrix X and the random vector Y , the prediction loss function is written as

$$L(\hat{f}(X), Y) = \sum_{i=1}^n l(\hat{f}(X_i), Y_i), \quad (2.7)$$

where X_i is the i^{th} line of X , Y_i is the i^{th} component of Y , and $\hat{f}(X)$ is an abusive notation for the vector $(\hat{f}(X_1), \dots, \hat{f}(X_n))^t$. Its prediction risk and conditional prediction risk are thus

$$R_{(X,Y)}(\hat{f}) = \mathbb{E}_{(X,Y)}[L(\hat{f}(X), Y)] = \int_{\mathcal{X}^n \times \mathcal{Y}^n} L(\hat{f}(X), Y) d\mathbb{P}_{x,y}(X, Y) \quad (2.8)$$

$$R_{Y|X}(\hat{f}, \mathbf{X}) = \mathbb{E}_{Y|X}[L(\hat{f}(X), Y)|X = \mathbf{X}] = \int_{\mathcal{Y}^n} L(\hat{f}(X), Y) d\mathbb{P}_{Y|X}(Y). \quad (2.9)$$

In the following paragraph, we derive from Equations (2.8) and (2.9) the best possible prediction.

The additive noise model

The prediction risk in Equation (2.8) and the conditional prediction risk in Equation (2.9) both allow to specify the objective. Indeed, the *best predictive model*, or the target model, that is, the best approximation function f , is the function minimizing these risks and is given by

$$f(\cdot) = \arg \min_{c \in \mathcal{Y}} R_{(x,y)}(c)$$

or

$$f(\mathbf{x}) = \arg \min_{c \in \mathcal{Y}} R_{y|x}(c, \mathbf{x}).$$

The objective of good prediction thus reduces to estimating the function $f(x)$. Note that the space of distributions we consider in \mathcal{P} determines the set where $f(x)$ is defined (see [Niyogi & Girosi 1996]).

Example 2.2 specifies the target f for the losses defined in Example 2.1.

Example 2.2. If $l(\hat{f}(x), y)$ is the squared-error loss function defined in Equation (2.2), then $f(x)$ is called the *regression function* and is the *conditional mean* of y , that is,

$$f(\mathbf{x}) = \mathbb{E}_{y|x}[y|x = \mathbf{x}]. \quad (2.10)$$

If $l(\hat{f}(x), y)$ is the absolute error loss function defined in Equation (2.3), then $f(x)$ is the *conditional median* of y , that is,

$$f(\mathbf{x}) = \text{median}(y|x = \mathbf{x}).$$

Finally, if $l(\hat{f}(x), y)$ is the 0-1 loss function defined in Equation (2.4), then $f(x)$ is called the *Bayes classifier* and is defined by

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}[Y = 1|x = \mathbf{x}] > 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

In the sequel, we focus the study only on the **regression** problem optimized under the **squared-error loss**, namely $\mathcal{Y} = \mathbb{R}$ (or a subset of \mathbb{R}) and

$$l(\hat{f}(x), y) = (y - \hat{f}(x))^2. \quad (2.11)$$

We recall that the true underlying system described in Figure 2.1 also relies on unobservable variables x_0 and can thus be modeled as

$$y = s(x, x_0),$$

where s denotes the function of the system. Nevertheless, as we cannot observe x_0 , it is common to approximate the system by an *additive noise model*. Such a model is also consistent with $f(x)$ in Equation (2.10) being the conditional mean of y , and is written as

$$\sigma e = y - f(x) \quad \text{or equivalently} \quad y = f(x) + \sigma e, \quad (2.12)$$

where e is a vector or a scalar and accounts for the total variations in x_0 , and σ represents the noise level. In this model, the variable e is often called the *noise* or the *innovation*. Assumptions on e will be specified in Subsection 2.1.3.

Although Model (2.12) does not represent exactly the truth, it is argued in [Hastie *et al.* 2008] that the additive noise model is a good approximation for the underlying system.

For the multivariate case, the model is written as

$$Y = f(X) + \sigma\varepsilon, \quad (2.13)$$

where $f(X)$ is an abusive notation for the vector $(f(X_1), \dots, f(X_n))^t$, and $\varepsilon = (e_1, \dots, e_n)^t$ is the noise vector.

2.1.2 The art of compromise

This subsection exposes the difficulties to which we are confronted when minimizing the (conditional) prediction risk in a given model \mathcal{M} . Those difficulties are actually related to the choice of the model, and especially to its complexity or capacity. This leads to divide the problem into two levels, one level where we consider a collection of M nested models and find the best prediction for each model, and the second level where these M best predictions are compared.

Prediction risk, estimation loss and other types of error

Before going into the details of the difficulties in model selection, let us expose the types of error that result from the process of minimizing the (conditional) prediction risk in a given model \mathcal{M} .

The conditional prediction risk can indeed be decomposed into several elements. First, we can notice that

$$\begin{aligned} R_{y|x}(\hat{f}, \mathbf{x}) &= \mathbb{E}_{y|x} \left[\left(y - f(x) + f(x) - \hat{f}(x) \right)^2 | x = \mathbf{x} \right] \\ &= \mathbb{E}_{y|x} \left[(y - f(x))^2 + (\hat{f}(x) - f(x))^2 + 2(y - f(x)) (f(x) - \hat{f}(x)) | x = \mathbf{x} \right] \\ &= R_{y|x}(f, \mathbf{x}) + (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2, \end{aligned} \quad (2.14)$$

where $R_{y|x}(f, \mathbf{x})$ is the conditional risk of the target function $f(x)$ and is equal to the variance of y conditionally to $x = \mathbf{x}$. Hence, this term is constant for any estimator \hat{f} and is often referred to as the *irreducible error*. Since the objective is to minimize either the risk or the conditional risk over a given model \mathcal{M} , we thus have that

$$\arg \min_{\hat{f} \in \mathcal{M}} R_{y|x}(\hat{f}, \mathbf{x}) = \arg \min_{\hat{f} \in \mathcal{M}} \left\{ R_{y|x}(f, \mathbf{x}) + (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 \right\} = \arg \min_{\hat{f} \in \mathcal{M}} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 \quad (2.15)$$

or, similarly,

$$\arg \min_{\hat{f} \in \mathcal{M}} R_{(x,y)}(\hat{f}) = \arg \min_{\hat{f} \in \mathcal{M}} \mathbb{E}_x [(\hat{f}(x) - f(x))^2]. \quad (2.16)$$

Noticing that

$$l(\hat{f}(x), f(x)) = (\hat{f}(x) - f(x))^2, \quad (2.17)$$

we can deduce from Equations (2.15) and (2.16) that the problem of prediction of y is therefore equivalent to the problem of estimation of $f(x)$. This allows us to make the link with

Statistical Decision Theory, which roughly aims at estimating parameters (such as the mean or the variance) involved in the distribution of a random variable (see [Wald 1939], [Berger 1985] and [Candès 2006]). Here, the parameter is the conditional mean $f(x)$ of y and fits exactly in the context of Decision Theory. More details will be given in Chapter 3. From Equation (2.14), we can also define the estimation loss as

$$l(\hat{f}(x), f(x)) = R_{y|x}(\hat{f}, x) - R_{y|x}(f, x), \quad (2.18)$$

in which we recognized what is called the *excess loss* [Boucheron et al. 2005]. Under the multivariate assumption, the estimation loss is expressed as

$$L(\hat{f}(X), f(X)) = \|\hat{f}(X) - f(X)\|^2 = R_{Y|X}(\hat{f}, X) - R_{Y|X}(f, X). \quad (2.19)$$

From now on, we will focus more on the estimation loss than on the prediction risk.

In the notation $l(\hat{f}(x), f(x))$, it appears clearly that the criterion of good prediction depends on the target function $f(x)$, which is unknown to us. Hence, the estimation loss $l(\hat{f}(x), f(x))$ and the prediction risks $R_{(x,y)}$ and $R_{y|x}$ are also unknown and need to be estimated. The problem of estimating $l(\hat{f}(x), f(x))$ is deferred to Section 2.2. For the moment, we only assume that we have such an estimator, based on data only, and we call it *crit* (for criterion). Since we do not know the true estimation loss, we use *crit* as a surrogate and select the best predictor (according to *crit*) by

$$\hat{f}_{\mathcal{M}}^{\text{crit}}(x) = \arg \min_{\hat{f} \in \mathcal{M}} \text{crit}(\hat{f}).$$

Second, the estimation loss itself can be decomposed using Equation (2.18) (see [Barron 1994]):

$$\begin{aligned} l(\hat{f}(x), f(x)) &= R_{y|x}(\hat{f}, x) - R_{y|x}(\hat{f}_{\mathcal{M}}^*, x) + R_{y|x}(\hat{f}_{\mathcal{M}}^*, x) - R_{y|x}(f, x) \\ &= R_{y|x}(\hat{f}, x) - R_{y|x}(\hat{f}_{\mathcal{M}}^*, x) + l(\hat{f}_{\mathcal{M}}^*(x), f(x)), \end{aligned}$$

where $\hat{f}_{\mathcal{M}}^*$ is the best prediction in Model \mathcal{M} , that is, $\hat{f}_{\mathcal{M}}^*$ is such that

$$R_{y|x}(\hat{f}_{\mathcal{M}}^*, x) = \inf_{\hat{f} \in \mathcal{M}} R_{y|x}(\hat{f}, x).$$

Note that if f belongs to \mathcal{M} , then $\hat{f}_{\mathcal{M}}^*(x)$ is equal to f . Otherwise, the function $\hat{f}_{\mathcal{M}}^*(x)$ is the best we can hope for by restricting the search to Model \mathcal{M} . This function is not an estimator of $f(x)$ since it depends on the $f(x)$ itself [Massart 2007]. Therefore it is often called the *oracle* (see [Donoho & Johnstone 1994] and [Candès 2006]), and has also been named the *crystal ball model* by [Breiman 1996], both vocables denoting its ideal nature. By definition of the oracle, the estimation loss $l(\hat{f}_{\mathcal{M}}^*(x), f(x))$ represents the distance from the regression function $f(x)$ to the model \mathcal{M} and is often denominated as the *approximation error* (see [Barron 1994], [Bartlett et al. 2002]) or as the *model bias* [Hastie et al. 2008, Chapter 7]. On the other hand, the term $R_{y|x}(\hat{f}, x) - R_{y|x}(\hat{f}_{\mathcal{M}}^*, x)$ is the error we make by minimizing *crit* instead of the risk itself and is referred to as the *estimation error* or the *estimation bias*.

We can also add another type of error that is not taken into account either in the prediction risk or in the estimation loss, the *numerical error*. This type of error occurs with the choice of an algorithm \mathcal{A} for estimating f based on data as well as with the precision used for computing the algorithm.

The link between prediction risk, estimation loss and the different types of error involved in the process of prediction are schematized in Figure 2.2.

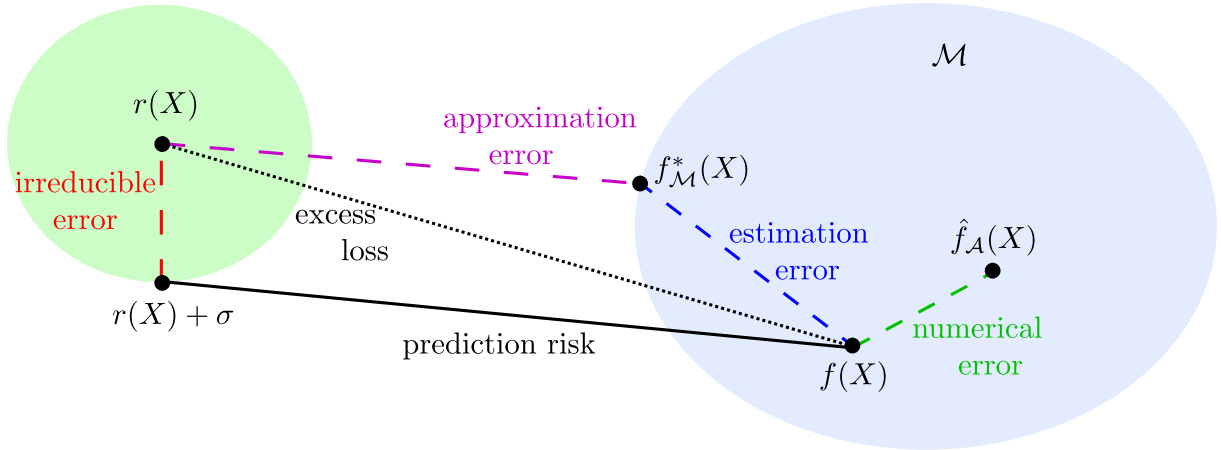


Figure 2.2: Different types of error in modelization and prediction. The functions f , $\hat{f}_{\mathcal{M}}^*$, \hat{f} and $\hat{f}_{\mathcal{A}}$ respectively correspond to the target, the oracle, the estimator, and the computation of the estimator based on algorithm \mathcal{A} . The green disk represents the minimum prediction error possible if we knew the target function f , and the blue disk represents the class of functions defined by model \mathcal{M} .

The problem of complexity

Now that we stated the decomposition of the prediction risks, we can explain how the complexity of Model \mathcal{M} influences the different types of error. Let us first begin with a simple example that will make the problem of complexity clear.

Example 2.3. A common example in wavelet, in the multivariate setting, is to take $n = p$ and $X = \text{I}_n$, so that the objective is to estimate the vector mean μ of the Gaussian vector $Y \sim \mathcal{N}_n(\mu, \sigma^2 \text{I}_n)$. In that case, we can take for instance the maximum likelihood estimator $\hat{\mu}^{ML} = Y$, which has the following quadratic risk

$$R_Y(\hat{\mu}^{ML}) = \mathbb{E}_Y[\|\hat{\mu}^{ML} - \mu\|^2] = \mathbb{E}_Y[\|Y - \mu\|^2] = n\sigma^2.$$

Since $\hat{\mu}^{ML}$ is an unbiased estimator of μ , its risk is equal to its variance. On the other hand, if we take the null estimator $\hat{\mu}_0 = 0$, the risk is equal to its bias since its variance is null:

$$R_Y(\hat{\mu}_0) = \mathbb{E}_Y[\|0 - \mu\|^2] = \|\mu\|^2.$$

We could also take the thresholding estimator $\hat{\mu}^J$ such that, for a given subset J ,

$$\hat{\mu}_i^J = \begin{cases} Y_i & \text{if } i \in J \\ 0 & \text{if } i \notin J \end{cases}.$$

This latter estimator has risk

$$R_Y(\hat{\mu}^J) = \mathbb{E}_Y[\|\hat{\mu}^J - \mu\|^2] = \sum_{i=1}^n \min(\mu_i^2, \sigma^2) = \|\mu_{J^c}\|^2 + n_J \sigma^2,$$

where J^c is the complementary set of J and n_J is the size of J . We can easily notice that the risk $R_Y(\hat{\mu}^J)$ is linear with respect to the dimension n_J of the subset J . Hence, the risk seems

to be lower for subsets of low dimension. However, in such cases, the bias term $\|\mu_{J^c}\|^2$ might be large, so that the challenge is to find the subset J yielding the lower risk.

In a similar way as in Example 2.3, the number k of variables in x used in the model can be taken as a measure of the *complexity* (or *capacity*) of the function \hat{f} and more generally of the model \mathcal{M} . The more variables of x it takes as an input, the more coefficients have to be estimated, and thus the more complex the function \hat{f} is. This example illustrates well the problem of the complexity in risk minimization. Indeed, one might want to take a model \mathcal{M} of high complexity $C(\mathcal{M})$ to include all possible submodels. In such a case, the approximation error will be low but the estimation error might be high since the minimization process will result in the selection of a function \hat{f} with highest complexity $C(\mathcal{M})$, even if the true regression function is not very complex. On the contrary, taking a model \mathcal{M} of low complexity $C(\mathcal{M})$ can reduce the estimation error while substantially increasing the approximation error. See [Niyogi & Girosi 1996], [Arlot & Celisse 2010]. Meanwhile, estimating complex functions or simple functions with many parameters might lead to non negligible numerical errors at the time of computing the function \hat{f} .

For nonlinear functions \hat{f} , the complexity is a vague notion that is not clearly defined in the literature. The best definition can be found in [Bozdogan 2000], which we state here.

Definition 2.3 (Complexity). *The complexity of a system is a measure of the degree of interdependency between the whole system and a simple enumerative composition of its subsystems or parts.*

This definition covers several existing measures of the complexity (see the discussion and references in [Bozdogan 2000]), without specifying it clearly. Some authors argue that it should take the smoothness of \hat{f} into account, either in terms of the number of continuous derivatives, or in terms of the highest moment of the Fourier transform of \hat{f} (see [Barron 1993]). There have been some attempts to give general measures, such as the Vapnik-Chervonenkis dimension (VC-dim) [Vapnik & Chervonenkis 1971] or the effective/generalized degrees of freedom [Hastie & Tibshirani 1990, Ye 1998]. Both measures coincide in the linear case with the dimension k of the model.

Another interesting argument on the problem of complexity can be found in [Hastie et al. 2008]: restricting the model space to fewer dimensions decreases the variance of the estimator $\hat{f}(x)$ and thus ensures a better control of the prediction. This goes in the same direction as [Guyon 2009], where it is argued that the variance of $\hat{f}(x)$ can also be taken as a measure of complexity.

As we can see, the problem of defining a model or a class of models makes the problem of model selection quite complicated, and the challenges are to define a good measure of complexity and to find a model that will have the right complexity so as to keep the approximation error, the estimation error, and also the numerical error to a minimum. The corresponding process is called the *complexity control*.

The following paragraph explains how we can overcome the problem of complexity and gives several examples of measures.

Collection of models based on structure of functions

A solution for a better control of complexity that has been generally applied is to consider a collection of M nested models $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ with increasing complexity. Example 2.4 gives examples of collections of models.

Example 2.4. We have already seen in the previous subsection an example of a collection of models with the linear projections. We can formalize this example as follows:

$$\mathcal{M}_m = \{\hat{f} \in \mathcal{L}^\pi(\mathcal{X}', \mathcal{Y}), \mathcal{X}' \subseteq \mathcal{X} \setminus \dim \mathcal{X}' \leq k\}.$$

In this case, the complexity

$$C(\mathcal{M}_m) = k$$

is the maximum rank of the linear projections, or, to put it in another way, the number of variables relevant to explain the system.

We can extend this example to polynomials of maximum degree k :

$$\mathcal{M}_m = \{\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \dots \in \mathcal{K}(x) \setminus \deg(\hat{f}) \leq k\},$$

where $\mathcal{K}(x)$ is the polynomial ring. Here, the complexity is measured by the maximum degree k .

Another example is based on the smoothness of the functions \hat{f} . This smoothness can be measured by a norm. For instance, we can take

$$\mathcal{M}_m = \{\hat{f} \in \mathcal{F} \setminus \|\hat{f}\|_2^2 \leq \lambda_m\},$$

where \mathcal{F} is the set of all continuous functions from \mathcal{X} to \mathcal{Y} (see [Arlot & Celisse 2010]). In this example, the complexity is equal to the maximum norm of functions in \mathcal{M}_m .

In a general way, we can define any model \mathcal{M} by

$$\mathcal{M}_m = \{\hat{f} \in \mathcal{F} \setminus C(\hat{f}) \leq c_m\},$$

where both the function space \mathcal{F} and the measure of complexity $C(\hat{f})$ have to be specified. Considering the sequence $(c_m)_{m=1}^M$ of constants such that

$$c_1 < c_2 < \dots < c_M$$

yields a collection of nested models.

All this is summarized in the following procedure.

General procedure for model selection

1. Define a collection of models $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ of increasing complexity:

$$C(\mathcal{M}_1) \leq \dots \leq C(\mathcal{M}_M).$$

2. For each model \mathcal{M}_m , that is, fixing the complexity $C(\mathcal{M})$, find the “best” estimator $\hat{f}_m(x)$ of $f(x)$:

$$\hat{f}_m(x) = \arg \min_{\hat{f} \in \mathcal{M}_m} \text{crit}_1(\hat{f}(x)).$$

3. Find the “best” model $\mathcal{M}_{\hat{m}}$ among the collection:

$$\hat{m} = \arg \min_{m \in \{1, \dots, M\}} \text{crit}_2(\hat{f}_m(x)).$$

Note that crit_1 and crit_2 need not be the same since crit_1 is only used for functions of the same complexity. Hence, crit_1 need not take the complexity into account, while it is compulsory for crit_2 in order to realize a good tradeoff between goodness of fit and complexity.

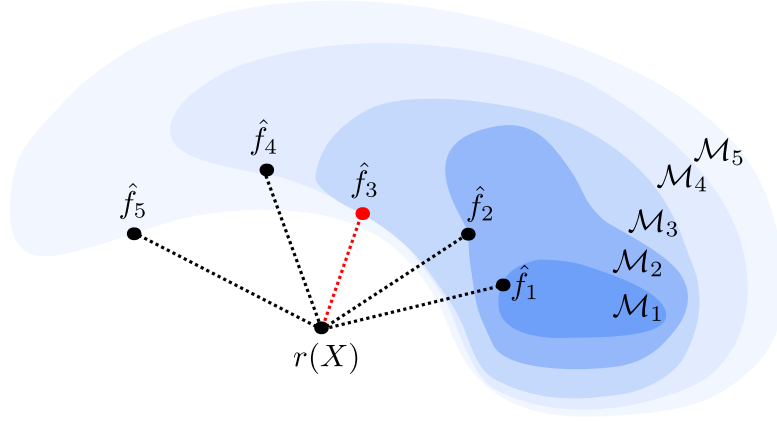


Figure 2.3: Hierarchy of models with increasing complexity. The functions \hat{f}_m , $m = 1, \dots, 5$ minimize the data-based criterion crit_1 for each model \mathcal{M}_m . The red circle corresponds to the function that minimizes the data-based criterion crit_2 for the collection $\{\hat{f}_1, \dots, \hat{f}_M\}$.

Figure 2.3 displays a diagram of the procedure. The collection of models is represented by the blue shapes, the functions \hat{f}_m , $m = 1, \dots, 5$, are those defined in Step 2 of the procedure, and the red circle indicates the estimator selected by the criterion crit_2 in Step 3.

Remark 2.1 (Inference post model selection). Since the main objective is one of good prediction, one might wish to perform a fourth step to the general procedure of model selection where f is estimated again after the model $\mathcal{M}_{\hat{m}}$ has been selected, possibly on a different dataset. This step is referred to as post-selection estimation or inference post model selection. We will not treat this step in the sequel but we refer the interested reader to the enlightening discussions in [Ye 1998] and [Leeb & Pötscher 2005].

The following subsection specifies the assumptions we will keep for the rest of the manuscript.

2.1.3 The general linear model

From now on, we will focus on the linear model only. The main reason of this choice is that model selection is already a hard problem even with this simplification. However, some of the techniques we review and propose can be easily derived in a nonlinear setting.

Linear model and variable selection

We define the univariate linear model with $f(x) = x^t \beta$, leading to

$$y = x^t \beta + \sigma e, \tag{2.20}$$

and its multivariate version

$$Y = X\beta + \sigma\varepsilon, \quad (2.21)$$

where we recall that y and Y takes real values, $x = (x^1, \dots, x^p)^t$, $X = (X^1, \dots, X^p)$ is the design matrix, e is the noise, ε is the noise vector and σ is the noise level. In both cases, the goal is to estimate the unknown regression coefficient β with values in \mathbb{R}^p . The corresponding estimator is denoted $\hat{\beta}$.

As mentioned by [Massart 2007, Chapter 4] the restriction to linear models might seem quite sharp. However, it offers many nice aspects, besides its simplicity, many of which are enumerated in [Hastie *et al.* 2008, Chapter 3].

First of all, it allows an easy interpretation of the results: the components of β are indeed indicators of the linear correlation between y and the variables in x . The components with higher absolute value reflect a strong relation between the corresponding variables in x and y .

On the other hand, the linear model can be useful as a local approximation of more complex functions. In that case, the model can be written as

$$y = \sum_{j=1}^p \beta_j \psi_j(x) + \sigma e,$$

where $(\psi_1(x), \dots, \psi_p(x))$ represents a (predefined) basis expansion of f . For instance, we can take a Fourier basis, wavelets, a trigonometric basis, or any transformation of the original inputs. The set of basis functions is called a *dictionary*.

Another type of transformation is on the output variable, which can sometimes be linear in x [Breiman & Friedman 1985]:

$$\Psi(y) = x^t \beta + \sigma e,$$

where Ψ can be for instance the logarithmic function (see [Rukhin 1986] and references therein for the description of production processes in Economics).

We will assume that the data y and x have already been transformed and that we only have to fit Model (2.21).

Finally, [Hastie *et al.* 2005] argue that the linear model sometimes yields better performances than a nonlinear model, especially when the data is sparse, when few observations are available, or when there is a low signal to noise ratio.

One subproblem of model selection in the linear model defined in Equation (2.21) is the problem of *variable selection*. Indeed, when the number p of variables in x is moderate to large, one might want to select only the variables that are relevant to predict y , and thus reduce the dimensionality of the problem. By doing so, once the best model has been found, the prediction process can be substantially faster (as well as more accurate if p is large). We will thus focus our interest in both a **good prediction** and a **selection of the most relevant variables**.

Before going to the next section on distributional assumptions, we would also like to point out that the rest of the study is done in the case where the input matrix X **is assumed to be fixed**. In such a case, there is equality between the conditional prediction risk $R_{Y|X}$ and the prediction risk $R_{(X,Y)}$. One reason of this choice is again to simplify a problem that is already very hard. Also, [Steinwart 2007] pointed out that the minimization of the risk $R_{(X,Y)}$ can be performed by pointwise minimization of the conditional risk $R_{Y|X}$. Hence, it might be considered as a first step in the minimization of $R_{(X,Y)}$.

Assumptions on the noise

The identifiability of Model (2.12) (and consequently of Model (2.20)) is only possible if we assume that the noise e verifies

$$\mathbb{E}[e] = 0. \quad (2.22)$$

We now specify the other assumptions on the noise. The first assumption is on the dependence between two observations of e . The most common one found in the literature states that two observations of the noise e are independent. Such an hypothesis conveniently offers a simplification of calculations. Also, it can be a useful approximation for instance in cases where the Law of Large Numbers applies. However, the Law of Large Numbers is defined in an asymptotic setting, so that, in practice on finite datasets, the approximation might be far from the truth. Hence, the multivariate setting comes as an alternative for modeling possible dependence between observations of e , which we now consider as components of the vector ε . Condition (2.22) thus becomes

$$\mathbb{E}[\varepsilon] = 0. \quad (2.23)$$

The noise e and the noise vector ε are respectively distributed according to an unknown distribution \mathbb{P}_e and \mathbb{P}_ε ,

$$e \sim \mathbb{P}_e \quad \text{and} \quad \varepsilon \sim \mathbb{P}_\varepsilon.$$

The most common assumption is that \mathbb{P}_ε is the Gaussian distribution $\mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. This assumption offers a substantial simplification since it is completely determined by its first two moments, the mean and the variance. Note that, in this case, the univariate and multivariate settings are equivalent since, for the Gaussian distribution, the noncorrelation assumption, described by the covariance matrix proportional to the identity matrix \mathbf{I}_n , is equivalent to the independence assumption. The distribution \mathbb{P}_e can also be specified to be the Student distribution, or other distributions depending on the context. We will call this assumption as the *fully specified- \mathbb{P}_ε assumption*.

An natural extension of the Gaussian assumption $\mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ is to consider the more general Gaussian distribution $\mathcal{N}_n(0, \Sigma)$, where the covariance matrix Σ is generally unknown and includes the special case of heteroskedasticity with

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}.$$

On the opposite direction, works such as [Vapnik 1998] (and related works) generally consider the univariate independent case where \mathbb{P}_e is completely unknown. The theory based on such assumption is sometimes referred to as *worst-case analysis*. This assumption has often been criticized as it might lead to loose results.

Finally, in-between the latter two assumptions, we can find the following one

$$\mathbb{P}_e \in \mathfrak{E} \quad \text{or} \quad \mathbb{P}_\varepsilon \in \mathfrak{D}$$

where \mathfrak{E} and \mathfrak{D} are families of distributions in the univariate and multivariate setting respectively. Examples of such families are the exponential family of distributions, the family of mixtures of Gaussian distributions, the family of spherically symmetric distributions, and the family of elliptically symmetric distributions. Note that the exponential family of distributions implies

independence between the observations of e , while it is not the case for the other three families (except for the Gaussian distribution $\mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$, which is a member of the four families). We refer to Chapter 3, Section 3.4 for more details on spherically and elliptically symmetric distributions.

2.2 Estimating the prediction risk

This section is devoted to the problem of estimating either the (conditional) prediction risk (2.8) (or (2.9)), or the estimation loss (2.17), which are equivalent up to a constant and are both unknown since they depend on the unknown target function f . We begin with the most natural estimator of the prediction risk, namely the empirical risk, and explain the drawbacks of this estimator along with the problem of overfitting. Then we divide the review into analytical methods, which estimate the prediction risk from the same data used to estimate the regression parameter; and resampling methods, which use a different dataset. This review is not intended to be exhaustive, but rather explains the main principles and origins of each criterion.

We recall that we observe the data $(\mathbf{x}, \mathbf{y}) = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$, which are considered either as n instances of the couple $(x, y) \in \mathcal{X} \times \mathcal{Y}$ (univariate assumption), or as one observation (\mathbf{X}, \mathbf{Y}) of the couple $(X, Y) \in \mathcal{X}^n \times \mathcal{Y}^n$ (multivariate assumption). We denote by N the number of observations ($N = n$ in the first case and $N = 1$ in the latter case). Note that, in both cases, the data observed are the same, that is, $(\mathbf{x}, \mathbf{y}) = (\mathbf{X}, \mathbf{Y})$, and the difference between the univariate and multivariate case is merely conceptual.

In the sequel, the criteria we expose are computed with the n observations. However, since they are also statistics, we write them under the multivariate case.

2.2.1 The empirical risk

The *empirical risk* of an estimator $\hat{\beta}$ of regression parameter β , sometimes also referred to as the *goodness of fit* or the *contrast*, is defined by

$$R_{emp}(\hat{\beta}) = \frac{1}{N} \tilde{L}(X\hat{\beta}, Y) = \frac{1}{N} \sum_{i=1}^n \tilde{l}(X_i\hat{\beta}, Y_i), \quad (2.24)$$

where \tilde{l} can be either the loss function l defined in Equation (2.1) (in our case the squared-error loss), or a surrogate of l .

The choice of the squared-error loss is the most common choice in regression problems, especially for the ease of computation incurred. However, it is often related to a Gaussian assumption on the noise, and can be sensitive to outliers [Steinwart 2007]. In order to cope with such an issue, [Huber 1964] proposed the following surrogate loss function

$$\tilde{L}^\eta(X\hat{\beta}, Y) = \sum_{i=1}^n \tilde{l}^\eta(X_i\hat{\beta}, Y_i), \quad \tilde{l}^\eta(X_i\hat{\beta}, Y_i) = \begin{cases} (X_i\hat{\beta} - Y_i)^2/2 & \text{if } |X_i\hat{\beta} - Y_i| \leq \eta \\ \eta(|X_i\hat{\beta} - Y_i| - \eta/2) & \text{otherwise,} \end{cases}$$

where η depends on the fraction of data affected by gross error. Huber loss $\tilde{L}^\eta(X\hat{\beta}, Y)$ leads to the selection of more robust estimators $\hat{\beta}$.

Another possible surrogate is the estimated log-likelihood (LL)

$$\tilde{L}^{LL}(X\hat{\beta}, Y) = -2 \log \hat{p}(Y|X\hat{\beta}).$$

The log-likelihood is a common loss function for the problem of estimating the density $p(Y)$. In our context, density estimation and estimation of the conditional mean of Y given X are related. For instance, if we take \hat{p} to be the Gaussian density $\mathcal{N}(X\hat{\beta}, \sigma^2)$, then the log-likelihood is defined by

$$\tilde{L}^{LL}(X\hat{\beta}, Y) = n \log(\sigma^2) + n \log(2\pi) + \frac{\|Y - X\hat{\beta}\|^2}{\sigma^2},$$

where we clearly recognized the standardized (or invariant) squared-error loss on the right-most part, and the other components can be thought of as constants if we assume the variance σ^2 to be known. See [Cherkassky & Mulier 1998], Chapter 2.

In the general setting where \hat{f} is not linear, when we consider the empirical risk as a criterion for selecting among models with finite data, we are faced with the issue that

$$\min_{\hat{f} \in \mathcal{M}} R_{emp}(\hat{f}) \quad (2.25)$$

is an ill-posed problem. Indeed, there exist an infinity of solutions such that

$$\hat{f}(\mathbf{X}_i) = \mathbf{Y}_i$$

exactly on the n data points observed. However, very few of these solutions can give a good approximation to new instances \mathbf{Y}_{new} . This phenomenon is known as *overfitting*, and is due to the fact that minimizing the empirical risk yields complex solutions. In the more rigid linear model, a similar phenomenon occurs, where the optimization problem (2.25) gives non null estimates for all the components in β , even when some variables in X are not relevant to the problem.

Therefore, R_{emp} is clearly not a good criterion for model selection when it is computed on the same data as for the estimation of β . However, it can be good in the case where the complexity of the model is fixed. Hence, it is taken as the golden rule for the second step of the general model selection procedure from Subsection 2.1.2:

$$\text{crit}_1(\hat{\beta}) = R_{emp}(\hat{\beta}). \quad (2.26)$$

When the empirical risk is based on the squared-error loss, Problem (2.25) is equivalent to least-squares, while for a loss based on log-likelihood it is equivalent to the maximum likelihood principle. In particular, if we do not make any restrictions on the model space of $\hat{\beta}$, the solution to Problem (2.25) is the least-squares estimator

$$\hat{\beta}^{LS} = (X^t X)^{-1} X^t Y. \quad (2.27)$$

Several solutions have been proposed to overcome the problem of overfitting. In the sequel, we divide them into two categories: the methods from the first category intend to reduce the bias of the empirical risk – as an estimator of the prediction risk – by taking into account the complexity of the model; the methods from the second category estimate the parameter β and the prediction risk $R_{(X,Y)}$ on different datasets. All the methods in the rest of this section correspond to the criterion crit_2 (Step 3) in the model selection procedure. Note that the distinction between these two categories might be fuzzy, as can be seen in [Arlot & Celisse 2010].

2.2.2 Analytical methods

In this section, we review criteria that are estimated on the same data used to estimate the regression coefficient β . They all intend to give a more accurate estimation of the estimation loss $L(X\hat{\beta}, X\beta) = \|X\hat{\beta} - X\beta\|^2$ than the empirical risk. They all can be expressed in one of the following forms:

$$\text{crit}_2(\hat{\beta}) = R_{\text{emp}}(\hat{\beta}) + \lambda \text{pen}(\hat{\beta}) \quad (2.28)$$

or

$$\text{crit}_2(\hat{\beta}) = \lambda \times R_{\text{emp}}(\hat{\beta}) \times \text{pen}(\hat{\beta}), \quad (2.29)$$

where $\text{pen}(\hat{\beta})$ is a measure of the complexity of $\hat{\beta}$ and λ is an hyperparameter trading off the goodness of fit by the complexity, that can depend on the number p of variables, the sample size n , or even the data for what are called the *data-driven penalties*.

These methods are also often referred to as *penalization methods* because of their form. Many of the methods we review in what follows were derived in the case where $\hat{\beta}$ is estimated by restricted least-squares, that is,

$$\hat{\beta}_I^{LS} = (X_I^t X_I)^{-1} X_I^t Y, \quad (2.30)$$

where I is the subset of variables assumed to be relevant, or maximum likelihood. In that case, the most common form of the penalty function pen is

$$\text{pen}(\hat{\beta}_I^{LS}) = \hat{\sigma}^2 k,$$

where $k = \#I$ is the number of variables in the selection and $\hat{\sigma}$ is an estimator of the noise level σ . However, we will see that there exist other possible forms.

Fixed penalties

Mallows' C_p . Mallows' idea in [Mallows 1973] was to propose an unbiased estimator of the scaled expected prediction error $\mathbb{E}_\beta[\|X\hat{\beta}_I - X\beta\|^2/\sigma^2]$, where $\hat{\beta}_I$ is an estimator of β based on the selected variables set $I \subseteq \{1, \dots, p\}$, \mathbb{E}_β denotes the expectation with respect to the distribution of Y in Model (2.21) and $\|\cdot\|$ is the Euclidean norm on \mathbb{R}^n . Assuming Gaussian *i.i.d.* residuals, he came to the following criterion

$$C_p(\hat{\beta}_I) = \frac{\|Y - X\hat{\beta}_I\|^2}{\hat{\sigma}^2} + 2\widehat{df} - n, \quad (2.31)$$

where $\hat{\sigma}^2$ is an estimator of the variance σ^2 , for instance the unbiased estimator based on the full linear model fitted with the least-squares estimator $\hat{\beta}^{LS}$, that is $\hat{\sigma}^2 = \|Y - X\hat{\beta}^{LS}\|^2/(n-p)$, and \widehat{df} , often called the effective or generalized dimension of the model [Hastie & Tibshirani 1990, Meyer & Woodroffe 2000], is an estimator of df , the degrees of freedom. Note that, for the least-squares estimator, $df = k$ the number of components of $\hat{\beta}_I$.

Mallows' C_p relies on the assumption that, if for some subset I of explanatory variables the expected prediction error is low, then we can assume those variables to be relevant for predicting Y . In practice, the rule for selecting the “best” candidate is the minimization of C_p . However, Mallows warns against a systematic application of the minimization of C_p , and advises to look instead at the shape of the C_p -plot and select the models for which $C_p \approx \widehat{df}$, especially when some explanatory variables are highly correlated. In addition, the author argues that the rule is unbiased only in the case where the subset I is independent of Y .

Akaike Information Criterion (AIC). A few years later, Akaike followed Mallows' spirit to propose automatic criteria that would not need a subjective calibration (like for the significance level in hypothesis testing, for instance). His proposal was more general than C_p with application to many problems such as variable selection, factor analysis, analysis of variance, or order selection in auto-regressive models (see [Akaike 1974] and [Akaike 1973]). His motivation was different: he considered the problem of estimating the density $p(\cdot|\beta)$ of a study variable Y , where p is parametrized by $\beta \in \mathbb{R}^p$, by $p(\cdot|\hat{\beta})$. His aim was to generalize the principle of maximum likelihood enabling a selection between maximum likelihood estimators $\hat{\beta}_I^{ML}$ based on several subsets I . The author showed that all the information for discriminating $p(\cdot|\hat{\beta}_I)$ from $p(\cdot|\beta)$ could be summed up by the Kullback-Leibler divergence $D_{KL}(\hat{\beta}_I, \beta) = \mathbb{E}[\log p(Y_{\text{new}}|\beta)] - \mathbb{E}[\log p(Y_{\text{new}}|\hat{\beta}_I)]$ where the expectation is taken over new observations. This divergence can in turn be approximated by its second-order variation when $\hat{\beta}_I$ is sufficiently close to β , which actually corresponds to the distance $\|\hat{\beta}_I - \beta\|_{\mathcal{I}}^2/2$ where $\mathcal{I} = -\mathbb{E}[(\partial^2 \log p / \partial \beta_i \partial \beta_j)_{i,j=1}^p]$ is the Fisher Information matrix and for a vector t , its weighted norm $\|t\|_{\mathcal{I}}$ is defined by $(t^t \mathcal{I} t)^{1/2}$. By means of asymptotical analysis and by considering the expectation of D_{KL} the author came to the following criterion

$$\text{AIC}(\hat{\beta}_I^{ML}) = -2 \sum_{i=1}^n \log p(y_i|\hat{\beta}_I^{ML}) + 2k, \quad (2.32)$$

where k is the number of parameters of $\hat{\beta}_I$. In the special case of a Gaussian distribution, AIC and C_p are equivalent up to a constant for Model (2.21) (see Chapter 3, Section 3.2.2 for more details). Hence Akaike described his criterion as a generalization of C_p for other distribution assumptions. Unlike Mallows, Akaike explicitly recommends the rule of minimization of AIC to identify the best model from data. Note that [Ye 1998] proposed to extend AIC to other estimators than the maximum likelihood estimator by replacing k by the generalized degrees of freedom \widehat{df} .

Final Prediction Error criterion (FPE). This criterion was also proposed by Akaike [Akaike 1970], but in the context of estimation of the parameter in a linear autoregressive model. The term *Final Prediction Error* actually refers to the prediction risk $R_{(X,Y)}$ associated with the squared-error loss. It derives from the fact that the prediction risk of the least-squares estimator for β restricted on a subset I is asymptotically equal to

$$R_{(X,Y)}(\hat{\beta}_I^{LS}) \xrightarrow{n \rightarrow \infty} \left(1 + \frac{k}{n}\right) \sigma^2,$$

under the assumption that the noise components are independent and stationary. Akaike then shows that the statistic

$$\hat{\sigma}^2 = \left(1 - \frac{k}{n}\right)^{-1} \|Y - X\hat{\beta}_I^{LS}\|^2$$

is a good estimator of the noise level σ^2 . Hence, he proposed the following criterion

$$\text{FPE}(\hat{\beta}_I^{LS}) = \frac{n+k}{n-k} \|Y - X\hat{\beta}_I^{LS}\|^2.$$

Rewriting it, we have that

$$\text{FPE}(\hat{\beta}_I^{LS}) = \|Y - X\hat{\beta}_I^{LS}\|^2 + 2k\hat{\sigma}_{restr}^2,$$

where $\hat{\sigma}_{restr} = \|Y - X\hat{\beta}_I^{LS}\|^2/(n - k)$ is an unbiased estimator of the variance when we assume the restricted model

$$Y = X_I\beta + \varepsilon$$

to be true. Hence, FPE is similar to C_p , but with another estimator of the variance.

Note that FPE is equivalent to what [Hocking 1976] calls the *average prediction variance*, defined as

$$\text{APV}(\hat{\beta}_I) = \frac{n + k}{n - k} \times \frac{1}{n} \|Y - X\hat{\beta}_I\|^2,$$

which applies to other estimators than Least-squares.

Corrected AIC (AIC_c). As mentioned in [Sugiura 1978] and [Hurvich & Tsai 1989], AIC is an unbiased estimator of the Kullback-Leibler divergence only in the asymptotic setting, and is thus evidently biased in practical examples since we use finite data. The objective of *Corrected AIC* (AIC_c) is thus to correct AIC's bias in a non asymptotical setting, especially for small sample. It consists in estimating the Kullback-Leibler divergence $D_{KL}(\hat{\beta}, \beta)$ in an autoregressive model, assuming the noise ε to be standard Gaussian $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ and estimating the distribution of Y by $\mathcal{N}_n(X\hat{\beta}, \hat{\sigma}^2 \mathbf{I}_n)$, where $\hat{\sigma}^2 = \|Y - X\hat{\beta}\|^2/n$ is the Maximum likelihood estimator of the noise level σ^2 . In such a case, the Kullback-Leibler divergence is equal to

$$\begin{aligned} D_{KL}(\hat{\beta}, \beta) &= \mathbb{E}_{Y|X} \left[n \log \sigma^2 + \frac{\|Y - X\hat{\beta}\|^2}{\hat{\sigma}^2} + \text{constant terms} \right] \\ &= \mathbb{E}_{Y|X} \left[n \log \sigma^2 + \frac{\|X\beta - X\hat{\beta}\|^2}{\hat{\sigma}^2} + \frac{n\sigma^2}{\hat{\sigma}^2} + \text{constant terms} \right], \end{aligned}$$

From the Gaussian assumption of Y , it follows that both $n\hat{\sigma}^2/\sigma^2$ and $\|X\beta - X\hat{\beta}\|^2/\hat{\sigma}^2$ are distributed according to a $\chi^2(n - p)$ and a $\chi^2(p)$ respectively. This yields the following estimator (ignoring the constant terms)

$$\text{AIC}_c(\hat{\beta}_I^{ML}) = n \log \hat{\sigma}^2 + \frac{n(n + k)}{n - k - 2}.$$

Generalizing it to other problems, [Hurvich *et al.* 1990] obtained the criterion

$$\text{AIC}_c(\hat{\beta}_I^{ML}) = -2 \sum_{i=1}^n \log p(y_i | \hat{\beta}_I^{ML}) + \frac{n(n + k)}{n - k - 2}. \quad (2.33)$$

[Burnham & Anderson 2002] recommend the use of AIC_c instead of AIC as soon as the ratio between the sample size and the maximum number of variables, n/p , is lower than or equal to 40.

Note however that the bias reduction in AIC_c relies on the assumption that the true model belongs to the collection of models (see [Bozdogan 2000]).

Generalized Cross Validation (GCV). The *Generalized Cross Validation criterion* is designed to approximate analytically the resampling method Leave-One-Out Cross Validation (LOOCV) [Golub *et al.* 1979]. We will give more details on LOOCV in the next subsection, but basically it consists in estimating β based on the data leaving out the i^{th} observation, computing the empirical risk R_{emp} on the i^{th} observation, iterating the process for all i and averaging the

n values of the empirical risk. The analytical approximation is derived for the case where β is estimated by the ridge regression estimator

$$\hat{\beta}^{RR} = (X^t X + \lambda I_n)^{-1} X^t Y.$$

See Subsection 2.3.3 for more details on ridge regression. The analytical form of LOOCV can be obtained exactly for $\hat{\beta}^{RR}$ thanks to the Woodbury identity (see Appendix A.1) applied to the matrix $(X^t X + \lambda I_n)^{-1}$. The main idea is then to use the Singular Value Decomposition (SVD)

$$X = UDV^t,$$

where U and V are respectively $n \times n$ and $p \times p$ orthogonal matrices, and D is an $n \times p$ rectangular diagonal matrix with the square roots of the eigenvalues of $X^t X$ as diagonal components. Then, computing the analytical form of LOOCV on a transformation of the linear model results in the criterion GCV with rotational invariance.

The expression of GCV is

$$\text{GCV}(\hat{\beta}) = \frac{n}{(n - \text{tr } H)^2} \|Y - X\hat{\beta}\|^2,$$

where H is such that $X\hat{\beta} = HY$, that is, H is the hat matrix, and $\text{tr } H$ is its trace. An interesting feature of GCV is that, unlike the criterion we have seen so far, it does not require an estimator of the noise level σ^2 .

Bayesian Information Criterion (BIC). This criterion is based on the remark that, for a model with fixed dimension, the maximum likelihood estimator can be obtained as the asymptotic limit of Bayes estimators for arbitrary prior distributions everywhere nonnull. The validity of the Bayes procedure has been established by [Schwarz 1978] for linear models and in the case where the noise components are independent and identically distributed. The procedure relies on the following principle: assuming that Y follows a distribution parametrized by β , which is also random, [Schwarz 1978] argues that the prior distribution on β need not be known exactly, as long as it can be expressed as

$$p(\beta) = \sum_{j=1}^M p(\mathcal{M}_j) p(\beta | \mathcal{M}_j),$$

where $(\mathcal{M}_j)_{j=1}^M$ is the set of models, $p(\mathcal{M}_j)$ is the *a priori* probability that the j^{th} model is the right one, and $p(\beta | \mathcal{M}_j)$ is the *a priori* probability of the parameter β given model \mathcal{M}_j . The Bayes solution then selects the model with highest *a posteriori* probability, that is,

$$\mathcal{M}_{j^*} = \arg \max_{\mathcal{M}_j} \left\{ \Pi(\mathcal{M}_j) = \log \int_{\mathcal{M}_j \cap \mathcal{B}} p(\mathcal{M}_j) p(y | \beta) d p(\beta | \mathcal{M}_j) \right\},$$

where \mathcal{B} is the set of definition of β .

Thanks to an asymptotic expansion of $\Pi(\mathcal{M}_j)$, Schwarz proposed the *Schwarz Bayes Criterion* (SBC) :

$$\text{SBC}(\hat{\beta}_I^{ML}) = \log p(Y | \hat{\beta}_I^{ML}) - k \log n / 2, \quad (2.34)$$

that can also be written as

$$\text{BIC}(\hat{\beta}_I^{ML}) = -2 \log p(Y | \hat{\beta}_I^{ML}) + k \log n, \quad (2.35)$$

the latter one being the *Bayesian Information Criterion* (BIC). Maximizing SBC yields the same selection as minimizing BIC. The latter expression is very close to that of AIC in Equation (2.32). In both cases, the goodness of fit criterion is based on the log-likelihood and the penalty is based on the number k of nonzero coefficients in the maximum likelihood estimator of β . The difference occurs with the hyperparameter trading off the goodness of fit and the penalty, which is set to $\lambda = 2$ for AIC and to $\lambda = \log n$ for BIC. Noticing that $\log n$ is larger than 2 as soon as $n \geq 8$, it is easy to see that BIC penalizes more the complexity than AIC and thus selects simpler models.

Other criteria. The early works of Mallows, Akaike and Schwarz have inspired a huge number of other criteria. We do not intend to review them all, but we just briefly name a few along with little justification.

[Foster & George 1994] derived their *Risk Inflation Criterion* (RIC), of the form

$$\text{RIC}(\hat{\beta}_I^{LS}) = \|Y - X\hat{\beta}_I^{LS}\|^2 + 2k\hat{\sigma}^2 \log p,$$

in order to overcome AIC and C_p 's inability to correctly handle the case where the true model is the null model $\beta = 0$. [Bozdogan 1994] tried several remedies to the tendency of AIC to select complex models by increasing the penalization. Among them are

$$\begin{aligned} \text{AIC}_3(\hat{\beta}_I^{ML}) &= -2 \log p(Y|\hat{\beta}_I^{ML}) + 3k, \\ \text{CAIC}(\hat{\beta}_I^{ML}) &= -2 \log p(Y|\hat{\beta}_I^{ML}) + (\log n + 1)k. \end{aligned}$$

Note that AIC_3 is derived so as to correct AIC's bias on the frontiers of the parameter space (see [Biernacki 1997], Appendix B, for more details). CAIC, standing for *Consistent AIC*, is a combination between AIC and BIC in an attempt to find a middle ground between both criteria. With a similar objective, [Hannan & Quinn 1979] tried to determine a criterion with λ depending on the sample size n and increasing with n at the slowest possible rate. Hence they came to the following criterion

$$\text{HQ}(\hat{\beta}_I^{ML}) = -2 \log p(Y|\hat{\beta}_I^{ML}) + c(\log \log n)k,$$

where $c > 2$.

Other authors, like [Takeuchi 1976] with the *Takeuchi Information Criterion* and [Bozdogan 1987, Bozdogan 1994] with the *Consistent AIC with Fisher matrix* (CAICF) and the *Information Complexity Criterion* (ICOMP), defined more general measures of the complexity involving the trace or the determinant of the Fisher information matrix \mathcal{I} .

Principled methods

In this paragraph, we review two theories, namely the Structural Risk Minimization (SRM) developed by [Vapnik & Chervonenkis 1971], and the Slope Heuristics developed by [Birgé & Massart 2001]. Both theories are more general than the methods we have seen so far in the sense that they consider any empirical risk and any collection of models, and they can be applied in many problems such as regression, classification and density estimation. Although their rationale is similar, that is, they both try to control the deviation of the empirical risk to the actual risk, they differ in that SRM considers the worst case over all the models and hence gives a global control, while slope heuristics intends to control the deviation only for the selected model, which results in a local control.

Structural Risk Minimization (SRM). Most of the criteria introduced up to now rely either on a strong distributional hypothesis, where the prior distribution of Y is assumed to be known, or on asymptotical behaviour, where the statistics of interest are asymptotically Gaussian from the Central Limit Theorem.

In a non asymptotic framework, [Vapnik & Chervonenkis 1971] developed a theory that does not rely at all on the form of the distribution, but rather considers any distribution for Y : this theory is called the *Statistical Learning Theory* (STL). Its principle is based on the general procedure for model selection we discussed on Section 2.1. For a given collection of models and a given empirical risk R_{emp} , STL aims at defining the conditions for which the empirical risk R_{emp} uniformly and almost surely converges to the true prediction risk $R_{(X,Y)}$, that is,

$$\forall \eta, \quad \exists \delta \quad \mathbb{P} \left[\sup_{\hat{\beta} \in \mathcal{M}} |R_{(X,Y)}(\hat{\beta}) - R_{emp}(\hat{\beta})| > \delta \right] \leq \eta, \quad (2.36)$$

for any probability $\mathbb{P}_{x,y}$ on the couple (X, Y) in the space \mathfrak{D} of probabilities. The asymptotic framework led [Vapnik & Chervonenkis 1971] to define a new measure of complexity, the *Vapnik-Chervonenkis dimension* (VC-dim). This measure of complexity is quite complex and often difficult to evaluate, except in a few cases. However, in the case where $\hat{\beta}$ is the (restricted) least-squares estimator, VC-dim matches the number of estimated components in $\hat{\beta}$ (or equivalently the rank of the projection).

The rest of the theory relies on the nonasymptotic framework. It aims at estimating the bound δ in Equation (2.36) as a function of the sample size n , of the level of confidence η , and the Vapnik-Chervonenkis dimension VC-dim,

$$\delta = \delta(n, \eta, \text{VC-dim}).$$

The bound δ is referred to as the *generalization bound*. Finally, having estimated the generalization bound, the uniform convergence equation (2.36) implies that, with probability at least $1 - \eta$,

$$R_{(X,Y)}(\hat{\beta}) \leq R_{emp}(\hat{\beta}) + \delta(n, \eta, \text{VC-dim}).$$

The right-hand side of this inequality is the criterion proposed by [Vapnik & Chervonenkis 1971] for selecting between models and that they called the *Structural Risk Minimization* (SRM):

$$\text{SRM}(\hat{\beta}) = R_{emp}(\hat{\beta}) + \delta(n, \eta, \text{VC-dim}). \quad (2.37)$$

Figure 2.4 shows a visualization of the principle behind Statistical Learning Theory and Structural Risk Minimization.

In the regression framework, the bound on the true prediction risk takes the following form, with probability at least $1 - \eta$,

$$R_{Y,X}(X\hat{\beta}) \leq R_{emp}(Y, X\hat{\beta}) \times \left(1 - c \sqrt{\frac{\text{VC-dim}}{n} \left(\log \left(\frac{a n}{\text{VC-dim}} \right) + 1 \right) - \frac{\log \eta}{n}} \right)_+^{-1}, \quad (2.38)$$

where a and c are constants and $x_+ = \max(x, 0)$. [Cherkassky *et al.* 1999] propose to use the choices $a = c = 1$, and $\eta = n^{-1/2}$. In order to compare SRM to other criteria from the literature, they also approximated VC-dim by the effective degrees of freedom.

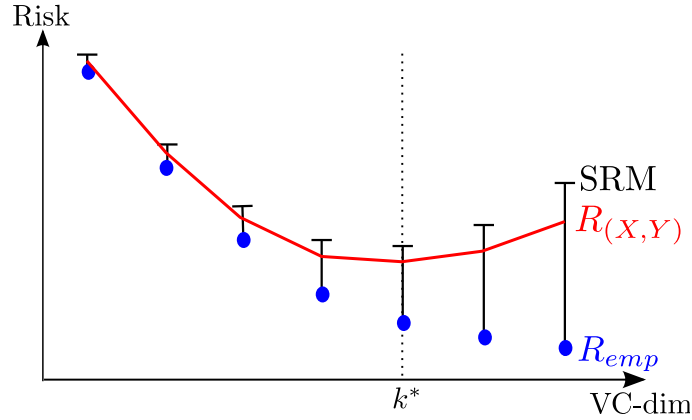


Figure 2.4: Principle of the Structural Risk Minimization (SRM). The empirical risk (blue dots) decreases with the complexity, while SRM (black intervals) gives an upper bound on its difference to the true risk (red line). The estimator selected by SRM is represented by the dotted line and has complexity k^* .

Slope Heuristics (SH). Slope heuristics took its grounds in [Birgé & Massart 2001] and was extended in [Birgé & Massart 2007]. It has been applied in other contexts than regression, like non supervised classification with the special problem of selecting Gaussian mixture models for genomic and genotypic data (see [Bontemps & Toussile 2010] and [Maugis & Michel 2011], among others).

The idea behind slope heuristics comes from the remark that the empirical risk R_{emp} , also called *contrast* by [Massart 2007], is a biased estimator of the true prediction risk $R_{(X,Y)}$ when it is evaluated on the same data as in the estimation of the parameter. We thus wish to correct it by a good penalty that will allow us to estimate the true risk $R_{(X,Y)}$ the more accurately. The best penalty is the penalty, denoted pen^{id} , that gives exactly the true risk, that is,

$$R_{(X,Y)}(\hat{\beta}) = R_{emp}(\hat{\beta}) + \text{pen}^{\text{id}}(\hat{\beta}). \quad (2.39)$$

It is referred to as the *ideal penalty* and is obviously unknown since it depends on the true risk $R_{(X,Y)}$. However, by rewriting Equation (2.39), we can notice that it is actually equal to the difference between the true risk and the empirical risk, that is,

$$\text{pen}^{\text{id}}(\hat{\beta}) = R_{(X,Y)}(\hat{\beta}) - R_{emp}(\hat{\beta}). \quad (2.40)$$

The ideal penalty thus represents the bias of the empirical risk (up to a factor of -1). Using concentration inequalities of the type

$$\mathbb{P}[|Z - \mathbb{E}Z| > \delta] \leq \eta$$

allows a better control of the bound δ and thus induces a lower variability of the criterion

$$\text{SH}(\hat{\beta}) = R_{emp}(\hat{\beta}) + \text{pen}_{\text{slope}}(\hat{\beta}),$$

where $\hat{\beta}$ is the minimum contrast estimator of β , that is, the estimator $\hat{\beta}$ minimizing the empirical risk R_{emp} , and $\text{pen}_{\text{slope}}$ is defined by

$$\text{pen}_{\text{slope}}(\hat{\beta}) = \text{slope} \times C(\hat{\beta}), \quad \text{with} \quad \text{slope} = \lambda \sigma^2,$$

C being a measure of the complexity. In particular, the control is better for the minimizer of SH, namely

$$\hat{\beta}^{(\hat{m})} = \arg \min_{\hat{\beta} \in \mathcal{M}_m} \text{SH}(\hat{\beta}),$$

which guarantees the oracle inequality

$$\mathbb{E}_{(X,Y)}[L(X\hat{\beta}^{(\hat{m})}, f(X))] \leq C_n \times \inf_{m \in \{1, \dots, M\}} \mathbb{E}_{(X,Y)}[L(X\hat{\beta}^{(m)}, f(X))] + R_n, \quad (2.41)$$

where C_n and R_n are constants depending only on the sample size n and the number of variables p .

Slope heuristics is performed in the following way: several models $\mathcal{M}_{\hat{m}(\text{slope})}$ are selected by minimizing the criterion SH for increasing values of slope, starting from slope = 0, which corresponds to the empirical risk minimizer. Increasing only a little the value of slope still yields complex models. At a certain point, the complexity of the selected model results in a significative jump compared to the complexity of the previously selected model. The corresponding value of slope defines the minimal penalty

$$\text{pen}_{\min}(\hat{\beta}) = \text{slope}_{\min} C(\hat{\beta}).$$

The slope heuristics suggests that the optimal value of the slope is twice that of the minimal slope

$$\text{slope}_{\text{opt}} = 2 \text{slope}_{\min}.$$

Figure 2.5a displays a typical example where the jump in dimension is clearly visible, and symbolized by the black cross. Figure 2.5b displays the empirical risk as a function of the complexity. The red line corresponds to the minimal slope reached for complex models.

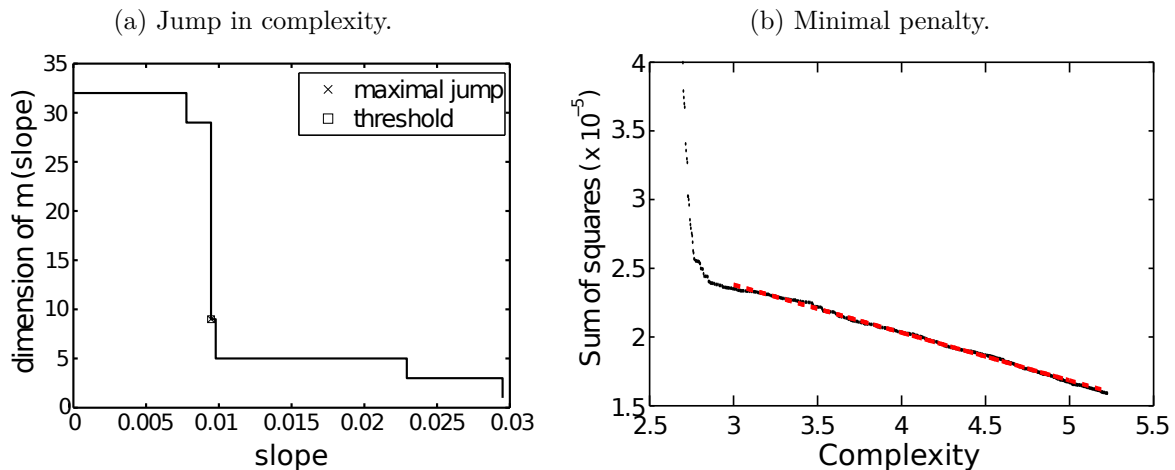


Figure 2.5: Left panel: Evolution of the complexity of models with respect to the slope. The black dot indicates the maximum jump in complexity, corresponding to the value of the minimal slope slope_{\min} . Extracted from [Arlot & Massart 2009]. Right panel: Evolution of the empirical risk (here the sum of squares) as a function of the complexity. The minimal penalty is represented by the red line. Extracted from [Caillierie & Michel 2009].

The main advantage of slope heuristics is that it estimates simultaneously both the hyperparameter λ and the noise level σ^2 based on data through the slope, while most methods plug in an estimator of σ^2 and fix the hyperparameter λ . Hence, slope heuristics is considered to be a *data-driven penalty* method. However, it seems difficult to apply in practice, in particular in situations where there is no clear jump or when the number M of models in the collection is not very large, since it might result in a poor estimation of the minimal penalty.

[Birgé & Massart 2007] propose several forms of complexity measures in the Gaussian regression model taking the least-squares criterion for the contrast. Among them, we have:

$$\begin{aligned} C_1(\hat{\beta}^{(m)}) &= D_m \\ C_2(\hat{\beta}^{(m)}) &= D_m \left(1 + \sqrt{2L_m}\right)^2 \\ C_3(\hat{\beta}^{(m)}) &= D_m \left(1 + 2\sqrt{H(D_m)} + 2H(D_m)\right) \\ C_4(\hat{\beta}^{(m)}) &= D_m \left(\kappa + 2(2 - \theta)\sqrt{L_m} + 2\theta^{-1}L_m\right) \end{aligned}$$

where D_m is the dimension of \mathcal{M}_m , $(L_m)_{m=1}^M$ is a sequence of non negative weights on each model \mathcal{M}_m , such that $\sum_{m \in \{1, \dots, M\}} \exp(-D_m L_m) < \infty$, $H(D_m) = D_m^{-1} \log(\#\{\mathcal{M} \setminus \dim(\mathcal{M}) = D_m\})$, and $\theta \in (0, 1)$ and $\kappa > 2 - \theta$ are constants. More precisely, the authors suggest the choices $\kappa = 2$ and $\theta = 1/2$ for the last measure of complexity C_3 . Note that, here, the collection of models may include several models with the same dimension D , in which case they are not necessarily nested. The measure C_2 is particularly indicated in cases where the number of models of the same dimension is large.

2.2.3 Resampling methods

There exist two main families of resampling methods, namely *cross-validatory* methods and *bootstrap* methods. Here, we just review two of the most commonly used cross-validatory methods, namely the *Leave-One-Out Cross Validation* (LOOCV) and the *V-fold Cross Validation* (CV-V). For the bootstrap family, we refer to [Efron 2004]. An extensive and illuminating review on resampling methods has been done in [Arlot & Celisse 2010].

V-fold Cross Validation (CV-V). The basic procedure of a *V-fold Cross Validation*, introduced by [Geisser 1975], follows the following steps:

1. Split the data $(Y_i, X_i)_{i=1}^n$ into V subsets $\{(Y_i, X_i)_{i \in J_1}, \dots, (Y_i, X_i)_{i \in J_V}\}$.
2. For each v in $\{1, \dots, V\}$, estimate β based on the $V - 1$ subsets excluding the v^{th} one and compute the empirical risk R_{emp} as in Equation (2.24) on the v^{th} subset:

$$R_{emp}(\hat{\beta}^{(-v)}) = \sum_{i \in J_v} \tilde{L}(X \hat{\beta}^{(-v)}, Y), \quad \text{with} \quad \hat{\beta}^{(-v)} = \hat{\beta}((Y_i, X_i), i \notin J_v).$$

3. Finally, compute the average of the empirical risks on the V subsets

$$\text{CV-V}(\hat{\beta}) = \frac{1}{V} \sum_{v=1}^V R_{emp}(\hat{\beta}^{(-v)}). \quad (2.42)$$

Common choices for the number of splits are $V = 5$ and $V = 10$, leading to CV-5 and CV-10.

Leave-One-Out Cross Validation (LOOCV). The *Leave-One-Out Cross Validation* (LOOCV), proposed by [Stone 1974] and [Allen 1974], can be viewed as a particular case of CV- V where we perform n splits, that is, $V = n$. Hence, each time, the estimator of β is computed without the i^{th} observation, and the empirical risk is evaluated on this i^{th} observation. It thus corresponds to the following criterion:

$$\text{LOOCV}(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n R_{\text{emp}}(\hat{\beta}^{(-i)}). \quad (2.43)$$

The Leave-One-Out cross validation is well known for overestimating the true risk $R_{(X,Y)}$ and often results in the selection of complex models. It is also more unstable than CV-5 or CV-10.

2.2.4 Which criterion should we choose?

Sofar, we have seen about a dozen of criteria for model selection while there exists a huge amount of other criteria in the literature on the subject. This fact obviously arises the question of which criterion to choose, and behind that question lies the following problem: **what is a good criterion of model selection?**

This paragraph is largely inspired by [Arlot & Celisse 2010], which give a comprehensive and enlightening discussion on the subject. They divide the answer between two main objectives. The first objective is the one we have been focusing on in the previous section, namely the estimation of the regression function f (which, as we have seen, is closely related to the prediction of new instances of Y). The second objective is the identification of the “true” model.

Efficiency. For the objective of good prediction, the optimality of a criterion is defined in terms of efficiency. [Arlot & Celisse 2010] make another distinction for efficiency depending on whether the optimality is asymptotic or nonasymptotic.

Let $\{\hat{\beta}^{(1)}, \dots, \hat{\beta}^{(M)}\}$ be the collection of estimators associated with the collection of models $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$. The model selected by a criterion crit is the model $\mathcal{M}_{\hat{m}}$ where \hat{m} is such that

$$\hat{m} = \underset{m \in \{1, \dots, M\}}{\text{arg min}} \text{crit}(\hat{\beta}^{(m)}).$$

The best model among the list is the model \mathcal{M}_{m^*} where m^* is such that

$$m^* = \underset{m \in \{1, \dots, M\}}{\text{arg inf}} \left\{ L(\hat{\beta}^{(m)}, f) = \|X\hat{\beta}^{(m)} - f(X)\|^2 \right\}.$$

Definition 2.4 states the notion of efficiency in the asymptotic framework.

Definition 2.4 (Asymptotic efficiency). *A model selection procedure is said to be asymptotically efficient if it verifies the condition*

$$\frac{L(\hat{\beta}^{(\hat{m})}, f)}{L(\hat{\beta}^{(m^*)}, f)} \xrightarrow[n \rightarrow \infty]{a.s.} 1.$$

This definition means that we expect the selected model to have a true estimation loss close to that of the best model in the collection.

In a nonasymptotic framework, the adaptation of this definition corresponds to an oracle inequality. Note that the term *non-asymptotic* includes two cases: the finite sample setting,

where both the sample size n and the number p of variables are assumed to be fixed; and the framework where p can depend on n (see for instance [Massart 2007]), which is often written as $p = p(n)$.

Definition 2.5 (Nonasymptotic efficiency). *A model selection procedure is said to be efficient if it verifies the following oracle inequality either in expectation or with large probability:*

$$L(\hat{\beta}^{(\hat{m})}, f) \leq C_n L(\hat{\beta}^{(m^*)}, f) + R_n, \quad (2.44)$$

where $C_n \geq 1$ and $R_n \geq 1$ are two constants such that

$$\begin{aligned} C_n &\xrightarrow{n \rightarrow \infty} 1, \\ R_n &\ll L(\hat{\beta}^{(m^*)}, f). \end{aligned}$$

Consistency in model selection. The consistency in model selection is the ability of a model selection procedure to recover the “true” model with probability tending to 1 with the sample size n . Note that the common definition of the “true” model is the smallest model \mathcal{M} containing the target function f (see for instance [Yang 2005]). In order to obtain the model consistency, it is generally assumed that the true model belongs to the collection of models. However, when this assumption is false, it is argued in [Lebarbier & Mary-Huard 2006] that model consistent procedures actually tend to recover what [Burnham & Anderson 2002] call the *quasi-true model*, which is defined as the smallest model from the collection yielding the smallest Kullback-Leibler divergence. The quasi-true model can thus be seen as an oracle and leads to the following definition of consistency in model selection.

Definition 2.6 (Consistency in model selection). *A model selection procedure is said to be consistent for model selection if it verifies the condition*

$$\mathbb{P}[\hat{m} = m^*] \xrightarrow{n \rightarrow \infty} 1.$$

It has been shown in [Yang 2005] that BIC is model consistent, and so is any criterion of the same form as in Equation (2.28) for which the parameter λ depends on the sample size n .

Note that, as stated by [Lebarbier & Mary-Huard 2006], the consistency in model selection does not assure a good solution since the quasi-true model could be far from the target function f .

There also exists a different type of oracle inequalities for the consistency in selection, given in [Bunea & Wegkamp 2004]:

$$\frac{1}{n} \|X\hat{\beta}^{(\hat{m})} - X\beta\|^2 + \text{pen}(\hat{m}) \leq \min_m \min_{\hat{\beta} \in \mathcal{M}_m} (1 + 2a) \left\{ \frac{1}{n} \|X\hat{\beta} - X\beta\|^2 + \text{pen}(m) \right\}$$

with probability tending to 1.

The best of both worlds? A legitimate question that has been arisen and answered in [Shao 1997] for deterministic penalties (*i.e.* when λpen does not depend on data) and in [Yang 2005] for data-driven penalties is whether there exist procedures that are both efficient and model consistent. In such a case, we would have some insurance of selecting the oracle, at least asymptotically. Unfortunately, the answer is no in both cases, so that the user has to define a priority in the objectives with regard to the data.

2.3 Construction of the collection of models

In this section, we review a number of methods that construct a collection of models with increasing complexity. These methods are divided into *stepwise methods*, which are probably the earliest procedures found in the literature, and *sparse regularization methods*, which are optimization problem leading to a sparse solution.

2.3.1 Stepwise methods

Stepwise methods are greedy algorithms looking at each step for the next variable to add or remove from the current selection of variables, and then the restricted least-squares estimator $\hat{\beta}_I^{LS}$ is computed on the updated selection of variables. The collections of models can be written as

$$\mathcal{M}_m = \{\hat{f}(X) = X\beta \mid \|\beta\|_0 \leq k\},$$

where $\|t\|_0$ is the ℓ_0 – pseudonorm, that is, the number of nonzero components in t , and $1 \leq k \leq p$. Figure 2.6 shows the path of solutions, starting either from the null model with $\hat{\beta} = 0$, or from the complete model with $\hat{\beta}^{LS}$. Depending on which way the algorithm is computed, it corresponds either to *Forward Selection*, *Backward Elimination* or *Stepwise regression*, which are described into more detail in the following paragraphs.

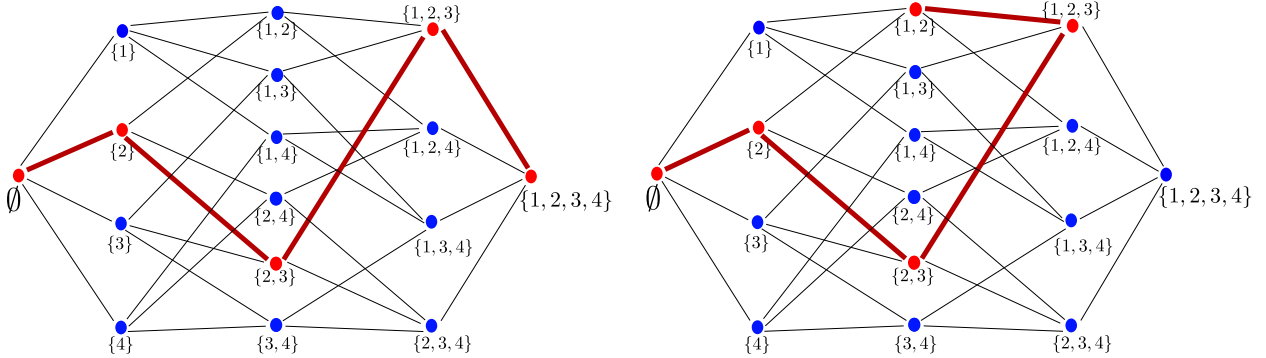


Figure 2.6: A branch of exploration (in red) in the lattice of all the possible subsets of variables, represented by the dots. The exploration with stepwise methods begins either with the null model where $I = \emptyset$ (the left-most red point) or with the full model $I = \{1, \dots, p\}$ (the right-most red point).

Forward Selection

Forward Selection consists in adding to the selection the variables in X one at a time, starting from the null model

$$\hat{\beta}_0 = 0.$$

At each step, the coefficients corresponding to the relevant variables are estimated by restricted least-squares

$$\hat{\beta}_I^{LS} = (X_I^t X_I)^{-1} X_I^t Y,$$

where I is the subset of indices of the relevant variables in X . The next variable to be added is the one maximizing the difference in the Sum of Squared Error (SSE) [Rawlings *et al.* 1998]

$$\begin{aligned} j_{next} &= \arg \max_{j \in \{1, \dots, p\} \setminus I} |\Delta SSE| \\ &= \arg \max_{j \in \{1, \dots, p\} \setminus I} \left| \|Y - X_I \hat{\beta}_I^{LS}\|^2 - \|Y - X_{I \cup \{j\}} \hat{\beta}_{I \cup \{j\}}^{LS}\|^2 \right|. \end{aligned}$$

At the end of each step, the relevance of the resulting subset $I \cup \{j_{next}\}$ is tested and the result of the test determines whether to continue the algorithm or not. The choice of the stopping criterion will be discussed at the end of the subsection.

Backward Elimination

Backward Elimination is very similar to *Forward Selection*, but the procedure is reversed: it starts from the full model estimated by least-squares with $\hat{\beta}^{LS}$, and the irrelevant variables are deleted one by one according to the following rule:

$$j_{del} = \arg \min_{j \in I} \left| \|Y - X_I \hat{\beta}_I^{LS}\|^2 - \|Y - X_{I \setminus \{j\}} \hat{\beta}_{I \setminus \{j\}}^{LS}\|^2 \right|.$$

In the same way as for *Forward Selection*, a stopping criterion is evaluated at the end of each step.

Stepwise Regression

Stepwise Regression has been proposed to overcome a major drawback of *Forward Selection* and *Backward Elimination*: the inability of both procedures to take into account the fact that two variables can be considered as relevant when they both belong to the selection and irrelevant if taken separately, or *vice versa*. Indeed, both procedures do not allow to take a step back and check for the relevance of each variable or each subset of variables.

Stepwise Regression thus consists in performing *Forward Selection* in turns with *Backward Elimination* each time a variable has been added to the selection. In other words, the relevance of each selected variable is tested each time the selection set is increased.

Stopping criterion

Initially, the stopping criterion for the stepwise procedures consisted in a Fisher test with null hypothesis “ $H_0 : \beta_I = 0$ ” versus the alternative hypothesis “ $H_1 : \beta_I \neq 0$ ”. The test statistics used respectively for *Forward Selection* (*forward*) and *Backward Elimination* (*backward*) are

$$\begin{aligned} F^{forward} &= \frac{\left| \|X_I \hat{\beta}_I^{LS}\|^2 - \|X_{I \cup \{j_{next}\}} \hat{\beta}_{I \cup \{j_{next}\}}^{LS}\|^2 \right|}{\|Y - X_{I \cup \{j_{next}\}} \hat{\beta}_{I \cup \{j_{next}\}}^{LS}\|^2 / (n - k - 1)} \\ F^{backward} &= \frac{\left| \|Y - X_I \hat{\beta}_I^{LS}\|^2 - \|Y - X_{I \setminus \{j_{del}\}} \hat{\beta}_{I \setminus \{j_{del}\}}^{LS}\|^2 \right|}{\|Y - X_I \hat{\beta}_I^{LS}\|^2 / (n - k)}, \end{aligned}$$

where $k = \#I$. Under the null hypothesis H_0 , both statistics are distributed as a Fisher $F(1, n - k)$. However, this stopping criterion entails the following problem, raised by [Akaike 1974]:

should the confidence level for accepting H_0 be fixed for all the steps or should it depend on the size of the subset I ? And which value should it be given? Although Akaike seems to think that a fixed level would not report the size of the subset and hereby the possible bias in approximation, the simulation study in [Bendel & Afifi 1977] shows on the contrary that a fixed level gives similar performances (in terms of mean squared error) to other classical criteria such as Mallows C_p , and propose optimal values for the confidence level depending on the difference $n - p$.

Other stopping criterion have been tested, among them the coefficient of determination R^2 , the adjusted coefficient of determination R_{adj}^2 , Mallows C_p , AIC and BIC.

2.3.2 Sparse regularization methods

Relaxing a NP-hard problem. *Regularization methods* or *penalized methods* are optimization problems of the form

$$\min_{\beta \in \mathbb{R}^p} \left\{ J_{\lambda}^{\text{pen}}(\beta) = \text{model fitting} + \lambda \times \text{penalty} \right\}, \quad (2.45)$$

where $\lambda \in \mathbb{R}_+$ is a constant, often referred to as *hyperparameter*, trading off model fitting (or goodness of fit) by penalization. The analytical methods from Section 2.2.2 for estimating the estimation loss belong to this type of methods [Fan & Tang 2012], where the penalty is a function of the number of selected variables or of the degrees of freedom. However, such a penalty function is discrete, and the optimization problem is NP-hard so that the global solution cannot be computed in a reasonable time, which is why we often consider the minimization on a finite collection of models. Another way of overcoming the heavy computational time of such problems is to relax it to another optimization problem that can be computed in polynomial time, such as *sparse regularization methods*. Figure 2.7 displays a visualization of the relaxation.

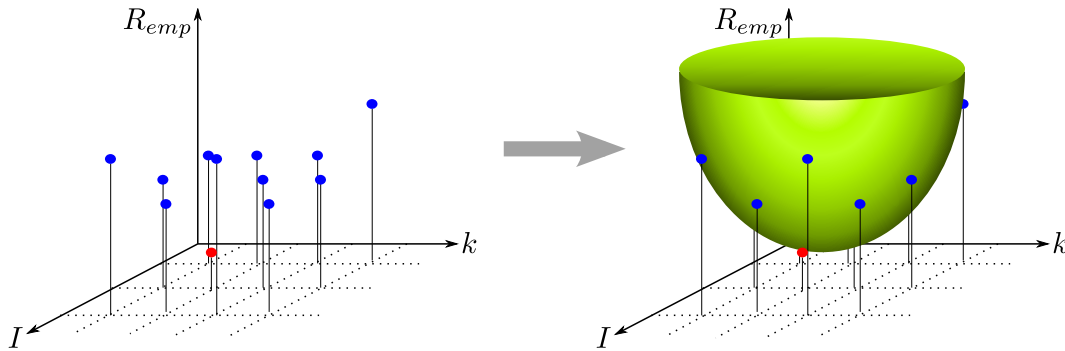


Figure 2.7: Representation of the relaxation of a NP-hard discrete problem (left) into a convex problem (right). The blue points correspond to the empirical risk of each subset I of size k and the red dot is the global solution. On the right panel, the green cup shape represents the continuous penalty function of the new optimization problem. Inspired by <http://www.iet.ntnu.no/~schellew/convexrelaxation/ConvexRelaxation.html>.

Sparse regularization methods are relaxations of the NP-hard problem for which the solution

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} J_{\lambda}^{\text{pen}}(\beta)$$

of the relaxed problem is sparse, *i.e.* it has components set exactly to zero:

$$\hat{\beta}_j = 0 \quad \forall j \notin I.$$

When the penalty function $\text{pen}(x)$ is convex, [Bach *et al.* 2011] state that the conditions for obtaining a sparse solution are

$$\hat{\beta} \text{ is sparse} \Leftrightarrow \begin{cases} (1) \text{ pen}(\cdot) \text{ is non differentiable at } t = 0, \\ (2) 0 \in \partial \text{pen}(0), \end{cases} \quad (2.46)$$

where $\partial \text{pen}(0)$ denotes the subgradient of pen , *i.e.* a generalization of the notion of gradient (see Chapter 5, Section 5.1 for more detail on the subgradient). Note that when pen is not convex, the second condition is changed to one involving the Clarke differential. More details will also be given in Chapter 5. The null components of $\hat{\beta}$ thus correspond to the variables in X that the regularization method considers as most irrelevant. The interest in sparse regularization methods comes from the fact that they propose simultaneously a way of selecting variables and an estimator for the corresponding nonzero coefficients.

Regularization path. The number of zeros and nonzeros components directly depends on the value of the hyperparameter λ : when the latter one is set to $\lambda = 0$, then the functional $J_\lambda^{\text{pen}}(\beta)$ in Equation (2.45) is equal to the least-squares criterion and therefore all the components of $\hat{\beta}$ are nonnull; on the contrary, if λ is sufficiently large so that the penalization takes over the least-squares criterion, then all the components in $\hat{\beta}$ are exactly 0 and the selection is empty. Hence, the choice of the hyperparameter λ plays a key role in the problem of variable selection. For that reason, sparse regularization methods are often considered to be model selection procedures. However, since it is not clear which value the hyperparameter λ should take. Therefore, we believe that they should rather be considered as methods for constructing collections of models by taking several values of λ .

The simplest way to construct collections of models is by taking λ on a grid. The collections of models can be written as

$$\mathcal{M}_m = \{\hat{f}(X) = X\beta \mid \text{pen}(\beta) \leq c_m\},$$

where c_m is linked to λ and $\{c_1, \dots, c_M\}$ is the grid. However, a poor choice of the parameters of the grid (regularity, number of points, interval between two points) might cause to miss the best model from the class. Taking a fine grid with a huge number of points might solve this problem but also results in a large computational cost. There exists an interesting alternative, often referred to as the *regularization path*.

The *regularization path* (for variable selection) consists in starting with a large value of λ for which the solution is the null model with $\hat{\beta} = 0$, and finding at each step the value of λ such that one zero component of $\hat{\beta}$ becomes nonnull. To clarify things, let us take the least-squares criterion $\|Y - X\beta\|^2$ for the model fitting measure in Equation (2.45), which is the choice for all the sparse regularization methods we will expose in the sequel. The criterion that determines the next variable to add to the current selection I is

$$j_{\text{next}} = \arg \max_{j \in \{1, \dots, p\} \setminus I} \left| (Y - X_I \hat{\beta}_I^{(m)})^t X^j \right|, \quad (2.47)$$

where $\hat{\beta}^{(m)}$ is the solution corresponding to model \mathcal{M}_m . Equation (2.47) amounts to finding which variable is most correlated with the current residual $\hat{\epsilon}^{(m)} = Y - X_I \hat{\beta}_I^{(m)}$. Note that this

criterion derives from the non differentiability of the penalty at 0 function and the (Karush-Kuhn-Tucker) optimality conditions associated with the corresponding minimization problem.

The value of λ associated with j_{next} is

$$\lambda_{m+1} = \left| (Y - X_I \hat{\beta}_I^{(m)})^t X^j \right|.$$

This ingenious way of determining the best grid, due to [Efron *et al.* 2004] with the Least Angle Regression (LAR) algorithm, allows to construct a collection of nested models with subsets of selected variables of increasing sizes. The verification of the optimality conditions also allows to delete a variable if the current step is too long. In that sense, it can be viewed as a modification of Stepwise regression with a different criterion for selecting variables. Figure 2.8 shows a simple example of a regularization path with three variables.

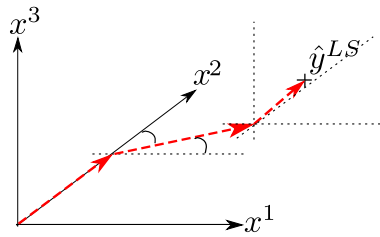


Figure 2.8: Regularization path (in red) for a simple example. The variable x^2 is the most correlated with y , and the variable x^1 is the most correlated with the residual $\hat{e}^{(1)} = y - x^2 \hat{\beta}_2^{(1)}$. The second step is taken in the direction splitting the angle between x^2 and x^1 in 2 equal angles. The path ends with the least-squares estimator for the full model.

Choice of the penalty. The literature is quite extensive on such methods, which penalty is often written in the form

$$\text{pen}(\beta) = \sum_{j=1}^p \rho(|\beta_j|) \quad (2.48)$$

and can express prior knowledge on the structure of the data (for instance ordered or grouped variables). In the following paragraphs, we only review a few sparse regularization methods which seem interesting in our context where no structure is assumed on data. However, we will discuss each time the solution when the design matrix X is orthogonal, that is when $X^t X$ is the identity matrix I_p , a case which can often be encountered when X represents a dictionary, since it allows an easier visualization of the difference in estimation between the methods. Figure 2.9 shows the form of the function $\rho(\cdot)$ in Equation (2.48) while Figure 2.10 displays their respective isocontours in 2D. Also shown are the forms of their shrinkage as a function of the Least-squares when the design matrix X is orthogonal.

Least Absolute Shrinkage and Selection Operator (Lasso)

The Lasso is one of the earliest sparse regularization methods and has been proposed by [Tibshirani 1996] and corresponds to Equation (2.48) with

$$\rho(|\beta_j|) = |\beta_j|.$$

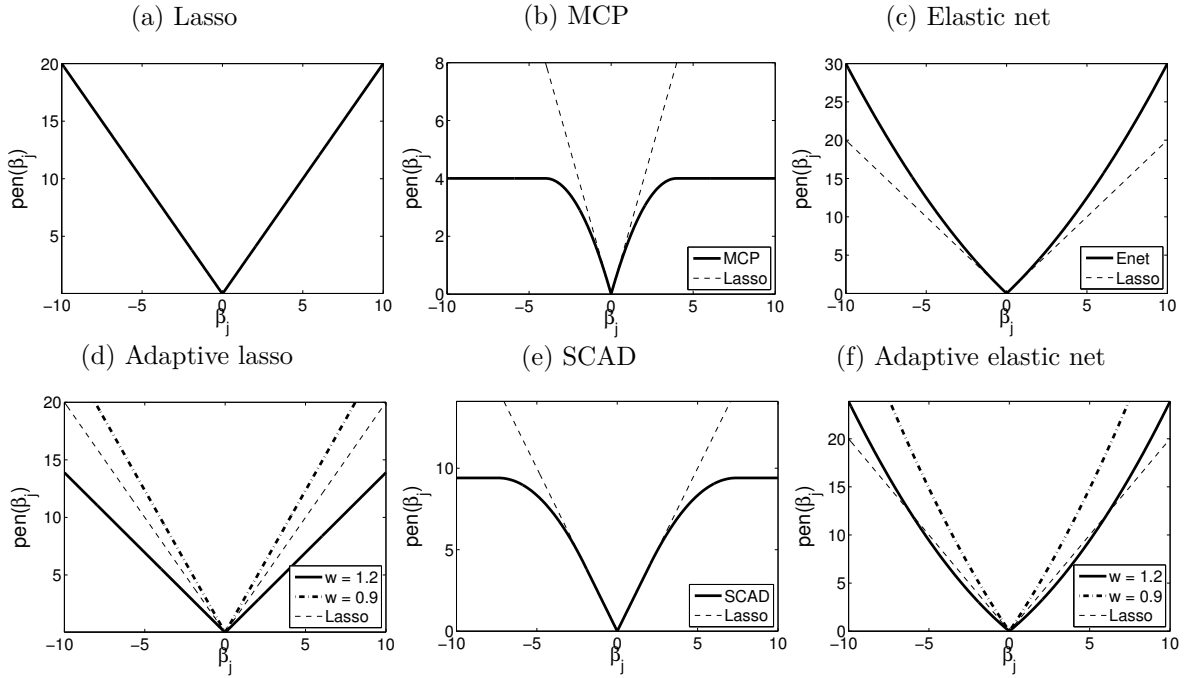


Figure 2.9: Form of the penalty in some sparse regularization methods. The dashed line shows the Lasso penalty, given for comparison purposes.

This actually corresponds to the ℓ_1 -penalty

$$\text{pen}^{\text{lasso}}(\beta) = \|\beta\|_1, \quad (2.49)$$

where $\|t\|_1 = \sum_{j=1}^p |t_j|$ is the ℓ_1 -norm. Figure 2.9a displays the evolution of the penalty function with respect to the value of the component $\hat{\beta}_j$. Lasso generalizes the *soft thresholding*, proposed by [Donoho & Johnstone 1994], for the orthogonal design case, which is a translation by λ of the least-squares estimator, truncated at λ :

$$\hat{\beta}_j = \left(Y^t X^j - \lambda \text{sgn}(Y^t X^j) \right) \mathbf{1}_{\{|Y^t X^j| > \lambda\}}. \quad (2.50)$$

Figure 2.11a displays the evolution of soft shrinkage with respect to the least-squares estimator.

Despite its important success, the Lasso estimator possesses a large bias when the hyperparameter λ is also large. This can be a serious drawback if we are interested in both a good selection and a good prediction, and [Zou 2006] has also shown an example for which Lasso is inconsistent for variable selection, *i.e.* it cannot recover the “true” subset (when its exists) with large probability. Decreasing the value of λ certainly decreases Lasso’s bias, but also its sparsity. According to [Leng *et al.* 2006], if the hyperparameter λ is tuned so that Lasso is consistent for variable selection, then its prediction is not optimal, and *vice versa*. Therefore, [Efron *et al.* 2004] proposed to replace the Lasso estimator by the restricted least-squares once the regularization path has been computed. Other recent works have proposed instead other penalty functions keeping Lasso’s nice sparsity property while leading to a much less biased solution. Among these alternatives, we can cite the *Minimax Concave Penalty* (MCP), the *Smoothly Clipped Absolute Deviation* (SCAD), and the *Adaptive Lasso*, which we develop in the next paragraphs. But before that, we would like to expose another penalty, the *Elastic net*, which does not correct Lasso’s bias but might lead to a different regularization path.

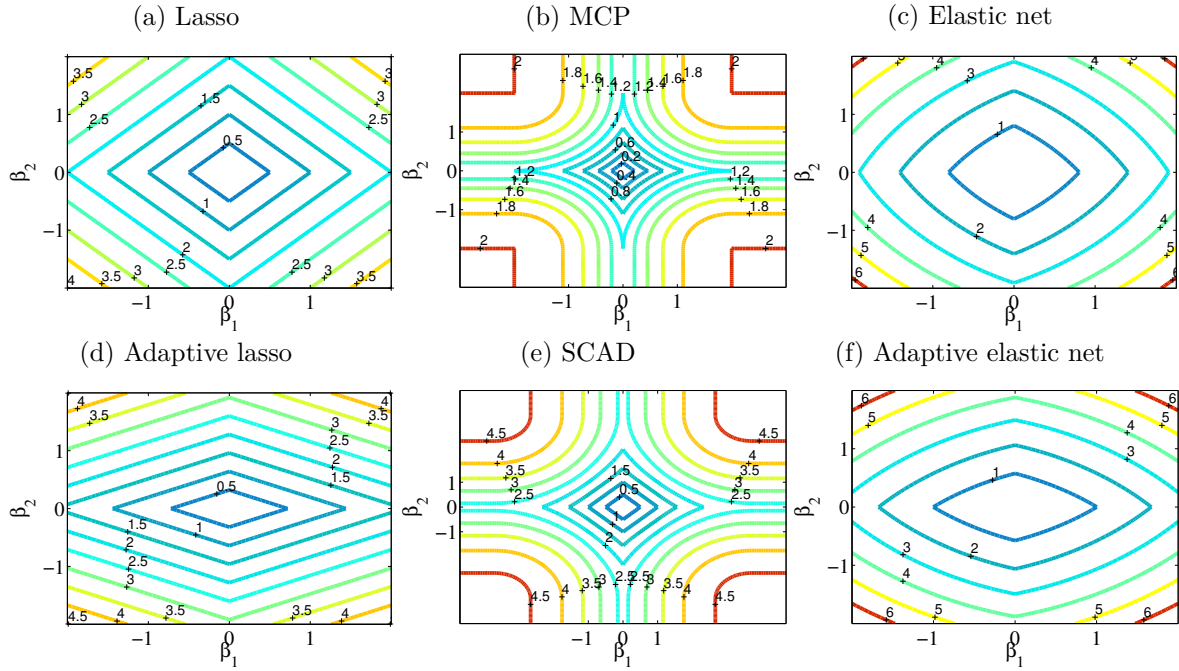


Figure 2.10: Level sets of the penalty as a function of β_1 and β_2 in some sparse regularization methods.

Elastic net

Proposed by [Zou & Hastie 2005], the *Elastic net* adds an ℓ_2 -norm to Lasso, leading to the following optimization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ J_{\lambda}^{enet}(\beta) = \|Y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right\}, \quad (2.51)$$

where λ_1 and λ_2 are two fixed positive hyperparameters. Elastic net is a combination between Lasso and Ridge regression, the latter one corresponding to the ℓ_2 -penalty $\text{pen}^{ridge}(\beta) = \|\beta\|_2^2$. Figure 2.9c shows the form of the Elastic net penalty with $\lambda_1 = 2$ and $\lambda_2 = 0.1$.

It can be easily noticed that problem (2.51) is equivalent to the following one

$$\min_{\beta \in \mathbb{R}^p} \left\{ \left\| \begin{pmatrix} Y \\ 0 \end{pmatrix} - \begin{pmatrix} X \\ \sqrt{\lambda_2} I_p \end{pmatrix} \beta \right\|^2 + \lambda_1 \|\beta\|_1 \right\},$$

which is actually a simple Lasso problem applied with the transformation

$$(X, Y) \mapsto \left(\begin{pmatrix} X \\ \sqrt{\lambda_2} I_p \end{pmatrix}, \begin{pmatrix} Y \\ 0 \end{pmatrix} \right). \quad (2.52)$$

Hence, the Elastic net optimization problem can be solved as easily as Lasso once the transformation has been done.

In the orthogonal design case, the Elastic net estimator takes the form

$$\hat{\beta}_j^{enet} = \frac{1}{1 + \lambda_2} \left(Y^t X^j - \lambda_1 \text{sgn}(Y^t X^j) \right) \mathbf{1}_{\{|Y^t X^j| > \lambda_1\}}. \quad (2.53)$$

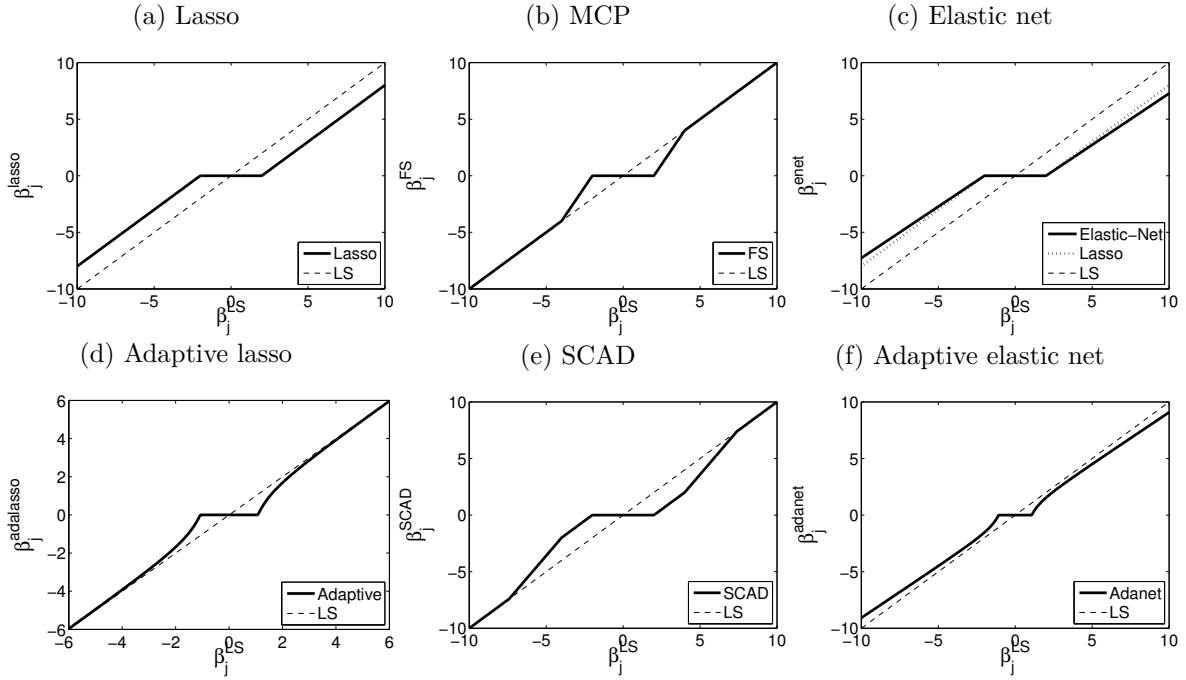


Figure 2.11: Level sets of the penalty as a function of β_1 and β_2 in some sparse regularization methods.

This form clearly exhibits the Ridge regression applied to the Lasso solution through the term $1/(1 + \lambda_2)$. Figure 2.11c displays the evolution of the Elastic net estimator with respect to the least-squares estimator for an orthogonal design matrix X . The comparison with Lasso (dotted line) shows that Elastic net is even more biased. However, in a general design setting, the ℓ_2 -norm is well known for overcoming problems related to a possible ill-conditioned matrix X , which makes the Elastic net appealing.

Minimax Concave Penalty (MCP) and Smoothly Clipped Absolute Deviation (SCAD)

Minimax Concave Penalty (MCP) and *Smoothly Clipped Absolute Deviation* (SCAD) are both quadratic spline penalty functions, respectively proposed by [Zhang 2010] and [Fan & Li 2001] to overcome Lasso's bias. The rationale given in the latter reference is enlightening regarding the interest in such methods. Indeed, when the objectives are both a good prediction and the selection of the relevant variables, [Fan & Li 2001] argue that a good penalty function should result in a **sparse**, **continuous** and **unbiased** solution. Lasso verifies the first two properties, while the restricted Least-squares or Hard Threshold, that can be obtained for instance by the penalty function given by (2.48) with

$$\rho^{HT}(|\beta_j|) = \lambda - \frac{1}{\lambda}(|\beta_j| - \lambda)^2 \mathbf{1}_{\{|\beta_j| < \lambda\}},$$

is both sparse and unbiased but not continuous, which might lead to an instability of the model prediction. Hence, [Fan & Li 2001], and more recently [Zhang 2010], proposed a penalty function that is a linear combination between the Lasso and the Hard Threshold penalty, leading to a

solution verifying all the properties they require. These penalty functions are given by Equation (2.48) with respectively

$$\rho^{MCP}(|\beta_j|) = \left(|\beta_j| - \frac{\beta_j^2}{2\gamma\lambda} \right) \mathbb{1}_{\{|\beta_j| \leq \gamma\lambda\}} + \frac{\gamma\lambda}{2} \mathbb{1}_{\{|\beta_j| > \gamma\lambda\}}, \quad (2.54)$$

$$\rho^{SCAD}(|\beta_j|) = |\beta_j| \mathbb{1}_{\{|\beta_j| \leq \lambda\}} - \frac{(|\beta_j| - \lambda)^2}{2(a-1)\lambda} \mathbb{1}_{\{\lambda < |\beta_j| \leq a\lambda\}} + \frac{(a+1)\lambda}{2} \mathbb{1}_{\{|\beta_j| > a\lambda\}}, \quad (2.55)$$

where λ is the hyperparameter tuning the sparsity as for Lasso, and $\gamma > 1$ and $a > 2$ are both hyperparameters tuning the bias of the resulting estimator. Figures 2.9b and 2.9e display these penalties with $\lambda = 2$, $\gamma = 2$ and $a = 3.7$. Note that both SCAD and MCP are equivalent to the Lasso when γ and a tend to infinity. On the other hand, when γ is close to 1, the solution of MCP tends to the restricted least-squares, which is never the case for SCAD.

The major difficulty arising from both methods comes from nonconvexity of their penalty functions. Hence, the resulting solution is not unique and the problem might be harder to solve than Lasso. However, as we will see in Chapter 5, one branch of the regularization path corresponding to MCP can be obtained within a slightly bigger but still polynomial computational time, a result that could be extended to SCAD. One advantage of MCP over SCAD is that MCP has a less concave penalty than SCAD and than other quadratic spline penalties, which presumably leads to an easier solving of the associated optimization problem.

When X is an orthogonal design matrix, MCP corresponds to the *Firm Shrinkage* estimator, proposed by [Bruce & Gao 1996] as an alternative to soft thresholding. Firm shrinkage (FS) and SCAD are expressed by

$$\hat{\beta}_j^{FS} = \begin{cases} 0 & \text{si } |Y^t X^j| \leq \lambda \\ \frac{\gamma}{\gamma-1}(Y^t X^j - \lambda \operatorname{sgn}(Y^t X^j)) & \text{if } \lambda < |Y^t X^j| \leq \gamma\lambda \\ Y^t X^j & \text{if } |Y^t X^j| \geq \gamma\lambda, \end{cases} \quad (2.56)$$

$$\hat{\beta}_j^{SCAD} = \begin{cases} (Y^t X^j - \lambda \operatorname{sgn}(Y^t X^j)) & \text{if } |Y^t X^j| \leq 2\lambda \\ \frac{a}{a-2}(Y^t X^j - \lambda \operatorname{sgn}(Y^t X^j)) & \text{if } 2\lambda < |Y^t X^j| \leq a\lambda \\ Y^t X^j & \text{if } |Y^t X^j| \geq a\lambda, \end{cases} \quad (2.57)$$

where λ , γ and a are the same as in (2.54) and (2.55). These forms show more clearly the linear combination between soft and hard threshold, the latter one being defined by $\hat{\beta}^{HT} = Y^t X^j \mathbb{1}_{\{|Y^t X^j| > \lambda\}}$. Figure 2.11b and 2.11e display the evolution of Firm Shrinkage and SCAD as a function of the hard threshold estimator (or least-squares estimator). On these figures, it is clear that, even by taking a value of a close to its limit 2, SCAD is more biased than MCP. We can also notice that each component can belong to several subsets: the subset I_0 of null components, the subset I_P of penalized components, and the subset I_N of nonpenalized (unbiased) components:

$$I_0 = \{1 \leq j \leq p \mid \hat{\beta}_j = 0\}, \quad I_P = \{1 \leq j \leq p \mid 0 < \hat{\beta}_j < \hat{\beta}_j^{HT}\}, \quad I_N = \{1 \leq j \leq p \mid \hat{\beta}_j = \hat{\beta}_j^{HT}\}.$$

For SCAD, the subset I_P can even be decomposed into two smaller subsets depending on the amount of penalization or shrinkage.

Adaptive Lasso and Adaptive Elastic net

Another way of reducing Lasso's and Elastic net's bias is to consider their adaptive versions, proposed by [Zou 2006] and [Zou & Zhang 2009] and given by (2.48) with

$$\rho^{adallasso}(|\beta_j|) = w_j |\beta_j| = w_j \rho^{lasso}(|\beta_j|), \quad (2.58)$$

$$\rho^{adanet}(|\beta_j|) = w_j |\beta_j| + \frac{\lambda_2}{\lambda_1} \beta_j^2 \quad (2.59)$$

where in both cases $\mathbf{w} = (w_j)_{j=1}^p$ is a vector of weights chosen beforehand. A typical choice of weights is given by

$$w_j = |\hat{\beta}_j^{init}|^{-q}, \quad (2.60)$$

where $\hat{\beta}^{init}$ is an initial solution for β obtained for instance by Least-squares or Ridge regression (especially if $X^t X$ is not invertible), and q is a positive scalar, common choices being $q = 1$ or $q = 2$. Note that the Adaptive Lasso taken with the choice

$$w_j = 1/|\hat{\beta}_j^{LS}|, \quad q = 1,$$

and with the additional constraint $\hat{\beta}_j \hat{\beta}_j^{LS} \geq 0$ corresponds to the Garrote [Breiman 1995]. Whatever the choice for $\hat{\beta}^{init}$ is, if it estimates one component with a small value, then the corresponding weight w_j will be large and will force $\hat{\beta}_j$ to go faster to 0. On the contrary, a large value for $\hat{\beta}_j^{init}$ means that the corresponding variable is likely to be relevant. This results in a small weight w_j and little penalization of the component $\hat{\beta}_j$, which is thereby nearly unbiased. Figure 2.9d and 2.9f show the shape of the penalties respectively defined through Equation (2.58) with $\lambda = 2$ and Equation (2.59) with $\lambda_1 = 2$ and $\lambda_2 = 0.1$. Different choice of w_j are displayed to show their influence on the penalties.

The main advantage of Adaptive Lasso and Adaptive Elastic net over MCP and SCAD lies in that their respective optimization problem is convex, thereby easier to solve and having a unique solution. They also both greatly benefit from the efficient LAR algorithm with a simple change in variables. For the Adaptive Lasso, the change of variables is

$$X \mapsto \tilde{X} \quad \text{with} \quad \tilde{X}^j = X^j / w_j, \quad j = 1, \dots, p. \quad (2.61)$$

For the Adaptive Elastic net, the change in variables in Formula (2.61) is followed by the one of the Elastic net in (2.52) where the design matrix X is replaced by the new matrix \tilde{X} . On the other side, Adaptive Lasso and Adaptive Elastic net depend on p hyperparameters for the weights (and possibly even $p + 1$ hyperparameters with the choice of Equation (2.60) where we also need to set the power q) and Adaptive Elastic net also depends on the choice of the second hyperparameter λ_2 . Hence, both methods require the optimization of much more parameters than do MCP and SCAD.

Taking w_j as in Equation (2.60) with $\hat{\beta}^{init}$ being the least-squares estimator and $q = 2$, the Adaptive Lasso and the Adaptive Elastic net can easily be derived for the case where X is orthogonal:

$$\hat{\beta}_j^{adallasso} = \left(Y^t X^j - \tilde{\lambda} \frac{\text{sgn}(Y^t X^j)}{|Y^t X^j|^q} \right) \mathbf{1}_{\{|Y^t X^j|^{1+q} > \tilde{\lambda}\}}, \quad (2.62)$$

$$\hat{\beta}_j^{adanet} = \frac{1}{1 + \lambda_2} \left(Y^t X^j - \tilde{\lambda}_1 \frac{\text{sgn}(Y^t X^j)}{|Y^t X^j|^q} \right) \mathbf{1}_{\{|Y^t X^j|^{1+q} > \tilde{\lambda}_1\}}. \quad (2.63)$$

Here, the hyperparameter tuning the sparsity is denoted by $\tilde{\lambda}$ and $\tilde{\lambda}_1$ to emphasize the fact that it cannot be the same as for soft thresholding or elastic net. Indeed, the condition for the components of $\hat{\beta}$ to be all equal to zero is $\tilde{\lambda} \geq \tilde{\lambda}^{(1)} = \max_j |Y^t X^j|^{1+q} = (\lambda^{(1)})^{1+q}$, where $\lambda^{(1)}$ is Lasso's hyperparameter for which the first variable is included in the selection.

Figure 2.11d and 2.11f display the estimators given in (2.62) with $\lambda = 2$ and in with $\lambda_1 = 2$ and $\lambda_2 = 0.1$. Note from Equation (2.62) that the choice $q = 0$ corresponds to Lasso (soft shrinkage) while the choice $q = \infty$ corresponds to hard threshold.

2.3.3 Mixed strategies and other approaches

Mixed strategies

As mentioned earlier, [Efron *et al.* 2004] considered a mixed strategy where the restricted Least-squares estimator is computed on Lasso's regularization path in order to overcome the problem of bias estimation. Following this idea, we can consider mixed strategies with other estimators than the restricted Least-squares. We propose to review here a few alternatives.

James-Stein estimator. [James & Stein 1961] proved that, for a given model with $n \geq 3$, the best linear unbiased estimator of β (in our context, the least-squares estimator) can be improved by biased estimators $\hat{\beta}$ having lower estimation risk

$$R(X\hat{\beta}) = \mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2],$$

which corresponds to the *Mean Squared Error* (MSE) of $X\hat{\beta}$. They propose the following shrinkage estimator¹

$$\hat{\beta}_I^{JS} = \left(1 - \frac{a}{\|X_I \hat{\beta}_I^{LS}\|^2}\right) \hat{\beta}_I^{LS}, \quad (2.64)$$

which yields the lower estimation risk when $a = k - 2$ when Y is Gaussian with covariance matrix I_n .

Generalized James-Stein estimator. For the derivation of the James-Stein estimator, the noise level σ is assumed to be known. [James & Stein 1961] extended their estimator to the case where the noise level is unknown but we have access to another random variable $S \sim \sigma^2 \chi^2(l)$, which is independent of Y . In that case, we get the generalized James-Stein estimator (GJS)

$$\hat{\beta}_I^{GJS} = \left(1 - \frac{aS}{\|X_I \hat{\beta}_I^{LS}\|^2}\right) \hat{\beta}_I^{LS}. \quad (2.65)$$

In particular, we can take S to be

$$S = \|Y - X\hat{\beta}_I^{LS}\|^2,$$

which follows a χ^2 distribution with $l = n - k$ degrees of freedom. The generalized James-Stein estimator can also be optimized so as to achieve the minimum estimation risk, which is obtained for $a = (k - 2)/(l + 2)$.

¹We give here the expression of the James-Stein estimator for our special context, but it was designed for the estimation of the mean vector μ of a multivariate Gaussian random vector $Y \sim \mathcal{N}_n(\mu, I_n)$, without explanatory variables, in which case it has the form

$$\hat{\mu}^{JS} = (1 - a/\|Y\|^2) Y.$$

Ridge Regression. A major drawback of the latter two estimators is that they rely on the least-squares estimator. Hence, in cases where $X^t X$ is ill-conditioned and the least-squares estimator does not give a satisfactory estimation, they can hardly give a good solution. In order to overcome the problems linked to the invertibility of the matrix $X^t X$, [Hoerl & Kennard 1970] developed the *Ridge Regression* (RR), which consists in the following optimization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|^2 \right\}, \quad \lambda > 0. \quad (2.66)$$

Note that this is a regularization method with an ℓ_2 - penalty. The solution of problem (2.66) can be expressed as

$$\hat{\beta}_I^{RR} = (X_I^t X_I + \lambda I_k)^{-1} X_I^t Y. \quad (2.67)$$

A wise choice of λ results in the increase of all the eigenvalues of $X^t X$ so that none are null, and the resulting matrix $X_I^t X_I + \lambda I_k$ is thus invertible. However, the choice of λ adds another unknown to the problem besides the choice of a good subset I of variables, and the issue is to know whether we can compare the solution for different values of λ in Problem (2.66) on the same data used to compare the different subsets I .

Other methods for constructing the collection of models

Besides the choice of a good estimator $\hat{\beta}_I$, we can also question the goodness of the rules for including or deleting a variable from the selection. So far we have seen two options: the difference in the Sum of Squared Error (SSE)

$$\Delta SSE(X_I \hat{\beta}_I, X_{I'} \hat{\beta}_{I'}) = |SSE(X_I \hat{\beta}_I) - SSE(X_{I'} \hat{\beta}_{I'})| = \left| \|Y - X_I \hat{\beta}_I\|^2 - \|Y - X_{I'} \hat{\beta}_{I'}^{LS}\|^2 \right|.$$

used in Stepwise methods, and the correlation with the current residual

$$\text{Corr}(j, I) = \left| (Y - X_I \hat{\beta}_I)^t X^j \right|, \quad \forall j \notin I$$

updating the subset in Sparse regularization methods.

There exist other options worth mentioning. The construction of the collection of models could be performed by adding some randomness to the exploration of the possible subsets to evaluate, with methods such as Monte-Carlo Tree Search (see [Chaslot *et al.* 2008], [Dramiński *et al.* 2010] and [Gaudel & Sebag 2010] for instance).

The last approach we would like to point out is that of [Bennett *et al.* 2006] and [Bennett *et al.* 2008], namely *Bilevel Optimization*. In a nutshell, it consists in writing the whole procedure of model selection (*i.e.* from the construction of the collection of models to the evaluation of the models) as the joint optimization of two objective functions

$$\text{minimize} \quad \text{crit}_2(\hat{\beta}) \quad (2.68)$$

$$\text{subject to} \quad \text{constraints on hyperparameters } \Lambda$$

$$\hat{\beta} \in \arg \min_{\beta \in B} \{ J_\Lambda^{\text{pen}}(\beta) = \text{model fitting}(X, Y, \beta) + \text{penalty}(\beta; \Lambda) \}. \quad (2.69)$$

The authors have only applied it to the special case where the models are evaluated by the V -fold cross validation ($\text{crit}_2 = \text{CV-}V$) with an absolute loss $\tilde{L}(X\hat{\beta}, Y) = \|Y - X\hat{\beta}\|$, and the regression coefficient $\hat{\beta}$ is estimated by Support-Vector Regression (SVR), which corresponds to the ϵ - insensitive loss $\text{model fitting}(X, Y, \beta) = \max(\|Y - X\hat{\beta}\| - \epsilon, 0)$ and the ℓ_2 - penalty (or

by Support-Vector Machine, SVM, in classification). This method allows the optimization of all the hyperparameters in the *inner-level problem* (2.69), for instance both the hyperparameter λ tuning the sparsity in Sparse regularization methods and the extra-hyperparameters tuning the bias in MCP, SCAD, Adaptive Lasso and Adaptive Elastic-net. However, [Guyon 2009] argues that not all the methods can be expressed as a bilevel optimization problem.

2.4 Summary of model selection procedures from literature

This section summarizes the propositions from the literature we have reviewed in the last two sections along with the choices they make for the whole procedure of model selection, including both the construction of the collection of models and the evaluation of the models, when available.

Table 2.1 specifies the different elements of the model selection procedure for each method. We recall that the model \mathcal{M}_m we consider in the context of linear regression is defined by

$$\mathcal{M}_m = \left\{ \hat{f}(X) = X\hat{\beta} \mid C(\hat{\beta}) \leq c_m \right\},$$

where $C(\hat{\beta})$ is a measure of the complexity of the model and c_m is a threshold value for this complexity. The sequence $\{c_1, \dots, c_M\}$ defines the collection of models $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$. The specifications for both quantities are given in the columns labeled “Complexity” and “Choice of c_m ” of the table. The last two columns of the table specify the choice of the criterion crit_1 that selects the best estimator in each model \mathcal{M}_m , and the choice of the criterion crit_2 that selects the best model $\hat{\mathcal{M}}_m$ among the collection $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$.

Table 2.1 is splitted into two categories. The first category corresponds to the methods whose major concern is on model evaluation. The methods from Subsection 2.2.2 (upper part) are mainly based on least-squares or maximum likelihood, thereby the specification on the collection of models is almost straightforward. On the contrary, the methods from Subsection 2.2.2 and 2.2.3 (lower part) are much more general and mainly assume that the collection of models should be specified by the user.

The second category compares methods focusing on the construction of models. The authors of each method always give a suggestion on the criterion to choose for selecting between the models. However, their suggestion might be inappropriate in some cases, the most edifying example being that of [Efron *et al.* 2004]. They indeed recommended the use of C_p to select the hyperparameter λ in their LARS algorithm computing the Lasso’s regularization path. In contrast, in the discussions on their paper, Hemant Ishwaran and Robert Stine criticized this choice and showed that it leads to overfitting. Indeed, because of its bias, the Lasso is a good selector but a poor estimator. Hence, it is more designed as a method for identifying the true underlying model than for predicting. It has indeed been proven in [Zhao & Yu 2007] that the Lasso is consistent in selection, while C_p is efficient (see [Shibata 1983]). It would thus be better to tune the Lasso by a model evaluation criterion having consistency in selection, such as BIC for instance.

We thus believe that **the construction of the collection of models and the evaluation of the models are both inherent parts of model selection and should be chosen with care, especially regarding the adequacy between the objectives of both parts.**

	Name	Complexity	Choice of c_m	crit ₁	crit ₂
FOCUS ON MODEL EVALUATION	C_p , FPE, RIC	$C(\hat{\beta}_I) = \#I$	$c_m \in \{1, \dots, p\}$	$\ Y - X\hat{\beta}_I\ ^2$	C_p , FPE, RIC
	Information criteria or Bayesian methods	$C(\hat{\beta}_I) = \#I$	Not specified	$-2 \sum_{i=1}^n \log \hat{p}(y_i \mathbf{x}_i \hat{\beta}_I)$	AIC, BIC, AIC _c , AIC ₃ , CAIC, HQ, CAICF, TIC
	GCV	$C(\hat{\beta}) = \ \hat{\beta}\ ^2$	Grid	$\ Y - X\hat{\beta}_I\ ^2$	GCV
	SRM Slope heuristics Resampling methods	To be specified by user	To be specified by user	$R_{emp}(\hat{\beta})$	SRM SH LOOCV, CV-V, Bootstrap
FOCUS ON COLLECTION OF MODELS	Exhaustive exploration Stepwise	$C(\hat{\beta}_I) = \#I$	$c_m = m - 1$	$\ Y - X\hat{\beta}_I\ ^2$	Any criterion F-test
	Soft/hard thresholding Firm Shrinkage	$C(\hat{\beta}) = \text{pen}(\beta)$	Grid	$\ Y - X\hat{\beta}\ ^2$	Universal threshold SURE
	Lasso SCAD	$C(\hat{\beta}) = \text{pen}(\beta)$	Grid	$\ Y - X\hat{\beta}\ ^2$	GCV / CV / SURE CV / GCV
	LARS MCP				C_p C_p
	Adaptive Lasso Elastic Net	$C(\hat{\beta}) = \text{pen}(\beta)$	Reg. path	$\ Y - X\hat{\beta}\ ^2$	CV-5 CV-10
	Adaptive Elastic Net				BIC

Table 2.1: Model selection procedures in literature. The column labeled “complexity” is the measure used to define model \mathcal{M}_m , the column labeled “choice of c_m ” is the sequence of thresholds on the complexity used to define the collection of models, the column labeled “crit₁” corresponds to the criterion used for selecting the best estimator in each model \mathcal{M}_m , and finally the column labeled “crit₂” is the criterion used to evaluate and compare the M models. The methods are separated into two categories, depending on the main focus for which it was developed: either the model evaluation or the construction of the collection of models.

2.5 Contributions

This section closes the state of the art on model selection by introducing our contributions and by showing where they stand in this picture relatively to existing methods.

2.5.1 A fairly large distributional framework with a dependence property

Most of the methods for model evaluation we have seen so far usually depends on the strong assumption that the distribution of the noise ε is known at least in form and is generally taken to be Gaussian. This is the case for instance of C_p , FPE, AIC_c and Slope heuristics. The methods based on information theory such as AIC, AIC_3 , CAIC, or the Bayesian methods like BIC and TIC apply to other distributions than the Gaussian law but still rely on the form of the estimated distribution through the log-likelihood criterion.

On the other side, SRM and Cross-validatory methods assume that the distribution of ε is completely unknown. It has been argued that such an assumption might result, in some cases, in loose generalization bounds for SRM and thus limits its performance in selection. Cross-validation could appear as a better choice in such cases, but its large computational cost is often prohibitive. Also, both methods generally assume the noise components ε_i to be independent, which is often a good approximation to the truth in an asymptotical framework but might be a poor representation in a finite-sample setting.

We propose to work instead with the assumption that the noise ε is a spherically symmetric random vector. The family of spherically symmetric distributions is a generalization of the Gaussian law relaxing the independence assumption. Note however that the components are assumed to be uncorrelated. Hence, our work is a first step toward the more general assumption of elliptically symmetric distributions, verifying both dependence and correlation properties. Another feature of our work is that the criteria we propose in Chapters 3 and 4 do not rely on the special form of the distribution, but only on the spherical assumption. Thus, they have the same expression whatever the spherical distribution is. In that sense, they present a robustness property.

2.5.2 New criteria with lower risk

While Chapter 3 is exclusively devoted to the derivation of unbiased criteria under our distributional assumption, Chapter 4 addresses the problem of evaluating a model evaluation criterion through the *loss estimation theory*. This second level of evaluation in the process is defined by a loss measuring the discrepancy of the model evaluation criterion $\hat{L}(\hat{\beta})$ to the actual estimation loss, such as the communication loss

$$\mathcal{L}(X\beta, X\hat{\beta}, \hat{L}) = (\hat{L}(\hat{\beta}) - L(X\beta, X\hat{\beta}))^2,$$

and its corresponding risk

$$\mathcal{R}_\beta(X\hat{\beta}, \hat{L}) = \mathbb{E}_\beta[\mathcal{L}(\beta, \hat{\beta}, \hat{L})].$$

Choosing a model evaluation criteria $\hat{L}(\hat{\beta})$ with lower risk $\mathcal{R}_\beta(X\hat{\beta}, \hat{L})$ leads to a better control of the estimation of the estimation loss $L(X\beta, X\hat{\beta})$. The heuristics behind the loss estimation theory is that a better estimator of the actual estimation loss should have a minimum closer to that of the estimation loss.

2.5.3 Numerical study and algorithms

In the numerical study, we first propose to investigate whether the methods for constructing collections of models, described in Section 2.3, are appropriate with the objective of good prediction. In order to do so, we will look at the selection of the best model in the collection with the actual estimation loss, that is, we will select the oracle and see if it belongs to the true underlying model, when this one belongs to the collection. Indeed, in real-life examples, we have no certainty on the target model; but if it does belong the collection that we built, we want to recover it.

Second, we propose a simulation study to compare the performances of our unbiased criteria to corrected criteria, and to compare both types of criteria to existing methods from the literature.

For these numerical studies, we developed an algorithm for computing the regularization path of the Minimax Concave Penalty (MCP). Such a regularization path is a little more complicated than the one for Lasso since the corresponding optimization problem is nonconvex. However, there exist similar conditions of optimality that follow from *Clarke differentials*, the generalization of the subgradient to nonconvex problems.

Finally, since our criteria are based on the spherical assumption, we investigate algorithms for generating spherically symmetric random vectors from the distributions that will be presented in the following chapter.

Unbiased loss estimators for model selection

Contents

3.1	Origins of loss estimation theory	55
3.1.1	Stein's Unbiased Risk Estimator (SURE)	56
3.1.2	From risk estimation to loss estimation	59
3.1.3	Loss estimation for model selection	60
3.2	The Gaussian case with known variance	63
3.2.1	Unbiased estimator of the estimation loss	63
3.2.2	Links with C_p , AIC and FPE	64
3.3	The Gaussian case with unknown variance	67
3.3.1	Unbiased estimator of the invariant estimation loss	69
3.3.2	Link with AIC_c	71
3.4	The spherical case	71
3.4.1	The class of multivariate spherically symmetric distributions	71
3.4.2	Unbiased estimator of the estimation loss	81
3.5	Summary	84

This chapter presents our contributions to model selection with unbiased loss estimators. We first make a short historical review on the theory of loss estimation, then we divide the study into three settings: in the first one, we consider the case where the error is drawn from the Gaussian distribution $\mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ where we assume the noise level σ to be known or independently estimated; the following setting extends the results to the case where the absence of knowledge on σ is directly taken into account when deriving the estimator of loss; finally, in the last setting, we consider the noise to be spherically symmetric.

3.1 Origins of loss estimation theory

Loss estimation traces back to [Sandved 1968] who, in various settings, introduced a notion of unbiased estimator of loss. It then received more attention after [Stein 1981] developed the theory of unbiased risk estimation. Also, [Johnstone 1988] dealt with (in)admissibility of unbiased estimators of loss, a notion that we will clarify in the sequel but basically consists in the (in)existence of other estimators of loss having lower risk.

We first give an outline of risk estimation before explaining the reasons of the orientation towards loss estimation. Finally, we present how loss estimation can be used for model selection.

3.1.1 Stein's Unbiased Risk Estimator (SURE)

Risk estimation has been initially developed in the following context: let Z be a random vector in \mathbb{R}^d , which we assume to be Gaussian, that is, $Z \sim \mathcal{N}_d(\theta, \sigma^2 \mathbf{I}_d)$, where the variance σ^2 is assumed to be known. The objective is to estimate the mean vector θ of Z , also in \mathbb{R}^d .

Given $\hat{\theta} = \hat{\theta}(Z)$ a chosen estimator of θ , we wish to evaluate its quality. Such an evaluation can be performed through a loss function, or cost function. In the context of estimation of the mean in \mathbb{R}^d , it is common to take the quadratic loss

$$L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2, \quad (3.1)$$

because of its adequacy with the problem as well as its simplicity. Note that this loss reaches its minimum when $\hat{\theta} = \theta$, hence assuring a good solution as soon as the loss is close to 0. We also define the quadratic risk of $\hat{\theta}$ as the expectation of its loss, namely

$$R_\theta(\hat{\theta}) = \mathbb{E}_\theta[L(\hat{\theta}, \theta)] = \mathbb{E}_\theta[\|\hat{\theta} - \theta\|^2], \quad (3.2)$$

where \mathbb{E}_θ denotes the expectation with respect to the density of Z . Note that the risk (3.2) is a generalization of the mean squared error to the multivariate case. The use of the quadratic loss and the quadratic risk is not new since it is the central measure of the Gauss-Markov theorem on best unbiased estimators, as recalled in the introduction of [James & Stein 1961]. In Decision Theory, the estimation risk R_θ is often taken as a golden rule to compare different estimators (also often referred to as *decision rules*) of θ and to define the notion of admissibility, which we give hereafter (see [Berger 1985]).

Definition 3.1 (Domination). *A decision rule $\hat{\theta}_1$ is better than a decision rule $\hat{\theta}_2$, or $\hat{\theta}_1$ dominates $\hat{\theta}_2$, if it verifies*

$$R_\theta(\hat{\theta}_1) \leq R_\theta(\hat{\theta}_2) \quad \forall \theta \in \Theta$$

and if there exists at least one value of θ for which the inequality is strict.

Definition 3.2 (Admissibility and inadmissibility). *A decision rule $\hat{\theta}$ is admissible if there exists no better decision rule. A decision rule $\hat{\theta}$ is inadmissible if there does exist a better decision rule.*

Let us take $\hat{\theta}$ to be the best unbiased estimator of θ . In the current context, this estimator is given by $\hat{\theta}_0 = Z$ since we assume that only one observation of the random vector Z is available. The risk of $\hat{\theta}_0$ is constant and exactly equal to $d\sigma^2$, where σ^2 is the variance of Z . It is obvious that no other unbiased estimator can be better than $\hat{\theta}_0$, since it is the best one (here, the term “better” is taken in the sense of the quadratic risk (3.2)). From this fact came the now well known idea to sacrifice the unbiasedness in order to get estimators with lower variance and lower quadratic risk. If such estimators are available, then we have greater control on the estimation and more certainty of its closeness to the true parameter θ .

[James & Stein 1961] have proposed better estimators of the form

$$\hat{\theta}_a^{JS} = \left(1 - \frac{a}{\|Z\|^2}\right) \hat{\theta}_0 = \left(1 - \frac{a}{\|Z\|^2}\right) Z, \quad (3.3)$$

where a is a constant that can be optimized so as to minimize the quadratic risk. In order to perform such an optimization, Theorem 3.1, known as Stein identity, states a result on scalar products, which was developed in [Stein 1981] many years after James-Stein estimators. This theorem is central for the derivation of loss estimators.

Theorem 3.1 (Stein identity). *Let Z be a Gaussian vector, $Z \sim \mathcal{N}_d(\theta, \sigma^2 \mathbf{I}_d)$, and $g : \mathbb{R}^d \mapsto \mathbb{R}^d$. If g is weakly differentiable, then, provided both expectations exist, we have*

$$\mathbb{E}_\theta \left[(Z - \theta)^t g(Z) \right] = \sigma^2 \mathbb{E}_\theta [\text{div}_Z g(Z)], \quad (3.4)$$

where $\text{div}_Z g(Z) = \sum_{i=1}^d \partial g_i(Z) / \partial Z_i$ is the weak divergence of $g(Z)$.

Stein identity relies on the notion of weak differentiability¹, which we define hereafter.

Definition 3.4 (Weak differentiability). *A function $h : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be weakly differentiable if there exist d functions $\nabla_1 h, \dots, \nabla_d h$ locally integrable on \mathbb{R}^d such that, for any $i = 1, \dots, d$,*

$$\int_{\mathbb{R}^d} h(t) \frac{\partial \phi}{\partial t_i}(t) dt = - \int_{\mathbb{R}^d} \nabla_i h(t) \phi(t) dt$$

for any infinitely differentiable function ϕ on \mathbb{R}^d with compact support. The functions $\nabla_i h$ are the i th partial weak derivatives of h and the vector $\nabla h = (\nabla_1 h, \dots, \nabla_d h)$ is referred to as the weak gradient of h .

In [Fourdrinier et al. 2012], it is argued that the condition of weak differentiability is verified for all the functions in the Sobolev space of order 1

$$W_{loc}^{1,d}(\Omega) = \left\{ h \in L_{loc}^d(\Omega), \quad \nabla_i h \in L_{loc}^d(\Omega), \quad 1 \leq i \leq d \right\},$$

where Ω is an open set in \mathbb{R}^d and $L_{loc}^d(\Omega)$ is the space of locally integrable functions over Ω . We next give a brief outline of the proof of Theorem 3.1.

Proof of Stein identity. The proof of Stein identity follows 4 steps.

First, let us take $d = 1$, $\theta = 0$ and $\sigma^2 = 1$, that is $Z \sim \mathcal{N}_1(0, 1)$. Stein derives Equation (3.4) thanks to a hidden integration by parts combined with the absolute continuity of g implied by the almost differentiability condition.

Second, Stein extends the result to the case $Z \sim \mathcal{N}_1(\theta, \sigma^2)$, where σ^2 is assumed to be known. This extension is performed through the change of variable $Y = (Z - \theta)/\sigma$, so that Y is standard normal.

¹Stein identity originally relied on the notion of almost differentiability of a function, whose following definition is extracted from [Stein 1981].

Definition 3.3 (Almost differentiability). *A function $h : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be almost differentiable if there exists a function $\nabla h : \mathbb{R}^d \mapsto \mathbb{R}$ such that, for all $z \in \mathbb{R}^d$,*

$$h(t + z) - h(t) = \int_0^1 z^t \nabla h(t + \varrho z) d\varrho$$

for almost all $t \in \mathbb{R}^d$. A function $g : \mathbb{R}^d \mapsto \mathbb{R}^d$ is almost differentiable if all its coordinate functions are. Essentially, ∇ is the vector differential operator of first partial derivatives with i^{th} coordinate

$$\nabla_i = \frac{\partial}{\partial t_i}.$$

The almost differentiability notion used in Stein identity has been noticed to be equivalent to the one of weak differentiability by [Johnstone 1988]. A formal proof of that result is given in [Fourdrinier et al. 2012] through the property of absolute continuity.

Third, Stein considers multivariate random vectors $Z \sim \mathcal{N}_d(\theta, I_d)$, where θ is also in \mathbb{R}^d , and functions h mapping from \mathbb{R}^d to \mathbb{R} . Denoting by $Z^{(-j)}$ the vector of Z where the component j has been removed and fixing the components $Z^{(-j)}$ yield

$$\mathbb{E}[(Z_j - \theta_j)h(Z)|Z^{(-j)}] = \mathbb{E}[\nabla_j h(Z)|Z^{(-j)}].$$

Taking the expectation under $Z^{(-j)}$, the independence between Z_j and $Z^{(-j)}$ results in

$$\mathbb{E}_\theta[(Z_j - \theta_j)h(Z)] = \mathbb{E}_\theta[\nabla_j h(Z)]. \quad (3.5)$$

Finally, the last step consists in considering a weakly differentiable function g mapping from \mathbb{R}^d to \mathbb{R}^d and applying (3.5) with $h(Z) = g_j(Z)$, that is, the j^{th} component of $g(Z)$. Remarking that

$$(Z - \theta)^t g(Z) = \sum_{j=1}^d (Z_j - \theta_j) g_j(Z),$$

the desired result is obtained.

Note that another proof can be found in [Fourdrinier *et al.* 2012] through Stokes' theorem (more precisely, its special case the divergence theorem). Indeed, as we will see in Section 3.4, the Gaussian distribution can be seen as a scale mixture of uniforms on the unit sphere. Hence, conditioning on the radius $R = \|Z - \theta\|^2$, the expectation becomes an integral on the surface of the sphere of radius R and Stokes' theorem can be applied. \square

An application of Stein's identity can be illustrated with the derivation of the risk of the James-Stein estimator in (3.3) for which

$$\hat{\theta}_a^{JS}(Z) = Z + a g(Z)$$

with $g(Z) = -Z/\|Z\|^2$. We note that the function g is not differentiable (since it explodes at 0) and that its weak differentiability is satisfied for $d \geq 3$, but not for $d \leq 2$. We have

$$\begin{aligned} \mathbb{E}_\theta[\|(1 - a/\|Z\|^2)Z - \theta\|^2] &= \mathbb{E}_\theta[\|Z - \theta\|^2] + a^2 \mathbb{E}_\theta[\|Z\|^{-2}] - 2 \mathbb{E}_\theta[a Z^t (Z - \theta)/\|Z\|^2] \\ &= d\sigma^2 + a(a - 2(d - 2))\sigma^4 \mathbb{E}_\theta[\|Z\|^{-2}]. \end{aligned} \quad (3.6)$$

since $\text{div}_Z(Z/\|Z\|^2) = (d - 2)/\|Z\|^2$ for $d \geq 3$. The risk we obtain in (3.6) reaches its minimum

$$R_\theta(\hat{\theta}_{d-2}^{JS}) = d\sigma^2 - 2(d - 2)^2 \sigma^4 \mathbb{E}_\theta[\|Z\|^{-2}]$$

when $a = d - 2$. It is easy to see that $R_\theta(\hat{\theta}_{d-2}^{JS})$ is always lower than $d\sigma^2$ when $d \geq 3$, since the second term is negative. Hence the James-Stein estimator $\hat{\theta}_{d-2}^{JS}$ improves on the unbiased one $\hat{\theta}_0$ as soon as $d \geq 3$. For the case where $d < 3$, the unbiased estimator $\hat{\theta}_0$ can be improved, as shown by [Stein 1955].

Note that, when σ^2 is unknown, we can replace the James-Stein estimator by its generalized estimator given in Section 2.3.3 of the previous chapter. Another possible extension is given in [Stein 1981] and takes the form

$$\hat{\theta}^{JS} = \left(I_d - \frac{A}{Z^t B Z} \right) Z, \quad (3.7)$$

where A is a symmetric matrix and $B = \{(\text{tr} A)I_d - 2A\}^{-1}A^2$. This extension is directly related to *smoothing splines*, as shown in [Li 1985].

However, the risk of an estimator of θ is not always easy to compute in practice because of its dependence to the true parameter θ . [Stein 1981] thus proposed to estimate it relying on Theorem 3.1. Indeed, this identity gives an expression of the risk $R_\theta(\hat{\theta})$ that do not depend explicitly on θ , but only indirectly through the law of Z . Thus an unbiased estimator of $R_\theta(\hat{\theta})$ is given by

$$\text{SURE}(\hat{\theta}) = \|\hat{\theta} - Z\|^2 + \sigma^2(2 \operatorname{div}_Z \hat{\theta} - d), \quad (3.8)$$

where σ^2 can be replaced by an unbiased estimator of the variance if it is unknown.

Note also that the quadratic loss in (3.1) is not the only one considered by Stein. He proposes to look at the more general quadratic form

$$L(\theta, a) = (a - \eta(\theta))^t \alpha(\theta) (a - \eta(\theta)),$$

where η is a function mapping from the space Θ of θ to a space \mathcal{A} of actions, a is the chosen action, and α is a function mapping from Θ into the space of symmetric positive definite matrices of size $d \times d$.

The second loss proposed by Stein is useful for the estimation of the covariance matrix Σ of a random vector Z in \mathbb{R}^d when n observations of Z are available, $n \geq p$. This loss is often referred to as Stein's loss and is of the form

$$L(\Sigma, \hat{\Sigma}) = \operatorname{tr}(\Sigma^{-1} \hat{\Sigma}) - \log \det(\Sigma^{-1} \hat{\Sigma}) - d, \quad (3.9)$$

for a given estimator $\hat{\Sigma}$ of the covariance matrix Σ . Note that Stein's loss is actually equal to the Itakura-Saito divergence used for instance for Nonnegative matrix factorization (NMF) in Signal Processing (see the application in denoising and decomposition of sources in a piece of music [Févotte *et al.* 2009]).

To conclude the discussion on Stein's work, we would like to point out the suggestion of the author in [Stein 1981] to derive an unbiased estimator of the statistic

$$\operatorname{Var}(\text{SURE}) = \mathbb{E}_\theta[(\|\theta - \hat{\theta}\|^2 - \text{SURE})^2], \quad (3.10)$$

that is $\operatorname{Var}(\text{SURE})$ is the variance of SURE, the unbiased risk estimator. Stein proposes to use this estimator of the variance of SURE to determine confidence sets for θ of the form

$$I_\alpha = \left\{ \theta \mid \|\theta - \hat{\theta}\|^2 \leq \text{SURE} + c_\alpha \sqrt{\widehat{\operatorname{Var}}(\text{SURE})} \right\},$$

where α is the confidence level, c_α the critical value corresponding to $\alpha/2$, and $\widehat{\operatorname{Var}}(\text{SURE})$ is the unbiased estimator of the variance of SURE.

Confidence sets are not treated in this manuscript, but it is interesting to note that the statistic (3.10) has influenced the comparison between two estimators of loss, as we will see in the sequel.

3.1.2 From risk estimation to loss estimation

We first give the definition of an unbiased estimator of loss in a general setting. This definition of unbiasedness is the one used by [Johnstone 1988]. Special focus will be given on the quadratic loss, since it is the most commonly used and allows simple calculations. In practice, it is a reasonable choice if we are interested in both good selection and good prediction at the same time. Moreover, quadratic loss allows us to link loss estimators to other criteria, such as the popular C_p and AIC.

Definition 3.5 (Unbiasedness). *Let Z be a random vector in \mathbb{R}^d with mean $\theta \in \mathbb{R}^d$, and let $\hat{\theta} \in \mathbb{R}^d$ be any estimator of θ . An estimator $\hat{L}_0(\hat{\theta})$ of the loss $L(\hat{\theta}, \theta)$ is said to be unbiased if for all $\theta \in \mathbb{R}^d$ it satisfies the condition*

$$\mathbb{E}_\theta[\hat{L}_0(\hat{\theta})] = R(\hat{\theta}, \theta),$$

with

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta[L(\hat{\theta}, \theta)],$$

where $R(\hat{\theta}, \theta)$ is the risk of $\hat{\theta}$ at θ , \mathbb{E}_θ denoting the expectation with respect to the distribution of Z .

This definition of unbiasedness of an estimator of the loss is somehow non standard; for Stein it corresponds to unbiasedness of an estimator of the risk. However, as mentioned earlier, the terminology loss estimation and loss estimators are due to [Sandved 1968] and [Li 1985] and has been kept by other authors [Johnstone 1988, Lele 1993, Fourdrinier & Strawderman 2003, Fourdrinier & Wells 2012].

A particularly interesting result of [Li 1985] is the consistency of the estimator of μ estimated by the rule “minimize SURE” (which is equivalent to the rule “minimize the unbiased estimator \hat{L}_0 ”). Although this result has only been proved for the special case of James-Stein type estimators, this is encouraging for choosing such a rule to select the best model from data.

Differences between loss estimation and risk estimation are enlightened by results from [Li 1985]. He proves that SURE estimates the loss consistently over the true mean θ as d goes to infinity. He also constructs a simple example where θ is estimated by a particular form of James-Stein shrinkage estimators for which SURE tends asymptotically to a random variable, and hence is inconsistent for the estimation of the risk, which is not random. Another interesting result of [Li 1985] is the consistency of the estimator of θ selected using the rule “minimize SURE”. Although this result has only been proved for the special case of James-Stein type estimators, this is encouraging for choosing such a rule to select the best model from data.

As mentioned earlier, risk unbiased estimators and unbiased loss estimators are the same. However, loss estimation theory goes beyond the Stein’s Unbiased Risk Estimation principle. Indeed, it aims at finding estimators of the loss that will estimate the true loss more accurately than the unbiased estimator. The heuristics behind loss estimation is that getting better estimators of loss should lead to a selection of an estimator of the parameter θ that has a true loss as close as possible to the best estimator in the class (namely the oracle).

The optimization of such estimators requires a new “layer” of evaluation, this one evaluating the quality of a loss estimator, and can be performed through the minimization of losses and risks just like we have defined for the estimators of β in the previous chapter.

The search for better estimators of loss is the subject of Chapter 4.

3.1.3 Loss estimation for model selection

We recall that, in this manuscript, we are interested in estimating the unknown parameter β in either the full linear model

$$Y = X\beta + \sigma \varepsilon, \tag{3.11}$$

or the linear model restricted to subset $I \in \{1, \dots, p\}$

$$Y = X_I\beta_I + \sigma \varepsilon, \tag{3.12}$$

where ε will be assumed to be successively: Gaussian $\mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ with known variance σ^2 , Gaussian $\mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ with unknown variance σ^2 , and spherical $\mathcal{S}_n(0)$. In this section and in the following ones, we will consider any estimator $\hat{\beta}$ of β among those described in the previous chapter. The only condition for the following results to be valid is that $\hat{\beta}$ should be weakly differentiable with respect to Y (see Definition 3.4). In this context, we consider the estimation loss

$$L(\beta, \hat{\beta}) = \|X\hat{\beta} - X\beta\|^2 = (\hat{\beta} - \beta)^t X^t X (\hat{\beta} - \beta). \quad (3.13)$$

Using estimation of the loss or estimation of the risk for selecting between different models is a common approach in the model selection literature since Mallows and Akaike's works. Especially, [Akaike 1974] described his criterion as a “mathematical formulation of the principle of parsimony in model building”. The heuristic of loss estimation is that, the closer an estimator \hat{L} is from the true loss, the more we expect their respective minima to be close too.

There still are two major issues in this problem: measuring the complexity of the model and estimating the variance. Indeed, as we have seen in the previous chapter, the complexity plays the crucial role of a tradeoff between model fitting and good generalization properties, so that measuring it seems inevitable. On the other hand, many of the model evaluation criteria presented so far rely on an estimator of the variance, and so do the criteria we present in the sequel. However, it is not always clear which estimator of the variance is better to use. We thus discuss both issues in the next paragraphs.

Measuring the complexity

It turns out that the divergence term in SURE (see Equation (3.8)), say $\text{div}_Y X\hat{\beta}$ for the linear model, is related to the estimator of the degrees of freedom \widehat{df} used equation (2.31) for C_p definition, and the number k of parameters proposed in AIC equation (2.32). A convenient way to establish this connection is to follow [Ye 1998] in defining the (generalized) degrees of freedom of an estimator as the trace of the scaled covariance between the prediction $X\hat{\beta}$ and the observation Y

$$df = \frac{1}{\sigma^2} \text{tr} \left(\text{cov}_\beta(X\hat{\beta}, Y) \right). \quad (3.14)$$

This definition has the advantage of encompassing the effective degrees of freedom proposed for generalized linear models and the standard degrees of freedom used when dealing with the least square estimator.

When Stein's identity applies

$$df = \mathbb{E}_\beta[\text{div}_Y X\hat{\beta}].$$

Setting

$$\widehat{df} = \text{div}_Y X\hat{\beta},$$

the statistic \widehat{df} appears as an unbiased estimator of the (generalized) degrees of freedom. In the case of linear estimators, there exists a hat matrix, that is, a matrix H such that $X\hat{\beta} = HY$ and we have

$$\text{div}_Y X\hat{\beta} = \text{div}_Y(HY) = \sum_{i=1}^n \frac{\partial \left(\sum_{j=1}^n H_{i,j} Y_j \right)}{\partial Y_i} = \sum_{i=1}^n H_{i,i} = \text{tr}(H),$$

so that

$$\widehat{df} = \text{tr}(H).$$

This definition of \widehat{df} is the one used by [Mallows 1973] for the extension of C_p to ridge regression. Note that, in this case, \widehat{df} is no longer depending on Y and thus meets its expectation ($df = \widehat{df}$). When H is a projection matrix (*i.e.* when $H^2 = H$), as it is for the least-squares estimator,

$$\text{tr}(H) = k,$$

where k is the rank of the projector which is also the number of linearly independent parameters, and thus $df = k$. In this case the definition of degrees of freedom meets its intuition. It is the number of parameters of the model that are free to vary.

When H is no longer a projector, $\text{rank}(H)$ is no longer a valid measure of complexity since it can be equal to n while $\text{tr}(H)$ is the trace norm of H (also known as the nuclear norm), a measure of the complexity of the associated mapping used as a convex proxy for the rank in some optimization problem [Recht *et al.* 2010]. For non linear estimators, the divergence $\text{div}_Y X\hat{\beta}$ is the trace of the Jacobian matrix (its trace norm) of the mapping that produced a set of fitted values from Y . According to [Ye 1998], it can be interpreted as “the cost of the estimation process” or as “the sum of the sensitivity of each fitted value to perturbations”.

In this work, we will only consider the unbiased estimator of the degrees of freedom provided by Stein as the measure of complexity.

Estimator of the variance: full model versus restricted model

The second issue, that is the estimation of the variance, has been addressed in several ways. The most popular way is to assume first σ^2 to be known, then to derive the model evaluation criteria with this assumption, and finally to plug in an estimator of the variance since it is seldom known in practice. This approach raises however the question of which estimator to use, as clearly pointed out in [Efron 1986] and in [Cherkassky & Ma 2003]. The authors proposed two unbiased estimators of the variance, one for the full model estimated by least-squares

$$\hat{\sigma}_{full}^2 = \frac{\|Y - X\hat{\beta}^{LS}\|^2}{n - p}, \quad (3.15)$$

and the second one for the model restricted to a subset $I \in \{1, \dots, p\}$

$$\hat{\sigma}_{restr}^2 = \frac{\|Y - X\hat{\beta}_I^{LS}\|^2}{n - k}, \quad (3.16)$$

where k is the size of I and $\hat{\beta}_I^{LS}$ is the least-squares estimator for the submodel corresponding to the subset I . If we are concerned with unbiasedness of the loss, so that the estimator of σ^2 should be unbiased and uncorrelated with $\text{div}_Y X\hat{\beta}$. Hence the choice between $\hat{\sigma}_{full}^2$ and $\hat{\sigma}_{restr}^2$ should be made with respect to what we believe is the true model, either the full model in (3.11) or the restricted model in (3.12). According to [Cherkassky & Ma 2003], “there seems to be no consensus on which approach is best for practical model selection”. However, we might find some piece of information on the difference between the two choices in [Efron 1986]. Indeed, in this work, the author argues that the estimator of variance for the restricted model is unbiased only when I is the true subset, otherwise the corresponding C_p overestimates the loss (we will see in Section 3.2 the links between \hat{L}_0 and C_p). If this is also true for the unbiased estimator \hat{L}_0 and if the true subset belongs to the set of subsets being compared, then it might help identifying the true subset more easily. Numerical comparisons between the choices will be given in the

Chapter 6. In Section 3.2, we will derive our estimators of loss under the assumption that σ^2 is known and estimate it thanks to an unbiased estimator (either for the full or for the restricted model). However, if we are not concerned with unbiasedness, then we can also think of other estimators of the variance, such as the estimator of maximum likelihood

$$\hat{\sigma}_{ML}^2 = \frac{\|Y - X\hat{\beta}^{LS}\|^2}{n}, \quad (3.17)$$

or a maximum a posteriori estimator of the variance, if we put a prior on β , both either for the full linear model or its restriction to subset I .

The second approach for treating the variance issue is to consider σ^2 to be unknown and to take this lack of knowledge into account in the expectations. In this case, we consider the invariant loss

$$L^{inv}(X\hat{\beta}, X\beta) = \frac{\|X\hat{\beta} - X\beta\|^2}{\sigma^2}$$

and the statistic S defined by

$$S = \|Y - X\hat{\beta}^{LS}\|^2, \quad (3.18)$$

following a $\sigma^2\chi^2(r)$ and independent of our study variable Y . We will treat this approach in Section 3.3.

Finally, we would like to mention another interesting approach, although not treated here, called the slope heuristics and proposed by [Birgé & Massart 2007]. This latter approach consists in estimating the optimal slope of the regression between the empirical risk $\|Y - X\hat{\beta}\|^2/n$ and the proposed penalty, which may be, for instance, a function of the degrees of freedom. The estimated slope takes into account simultaneously the level of tradeoff between accuracy and complexity (represented by λ in Equation (2.28)) and the variance, that is, the optimal slope is $\hat{\alpha}_{\text{opt}} = \hat{\lambda}_{\text{opt}}\hat{\sigma}^2$.

3.2 The Gaussian case with known variance

In this section, we derive our criterion by first considering the noise level σ^2 to be known. If this assumption is true, we can take it to be equal to 1 without loss of generality, since the model can always be normalized as

$$Y' = \frac{1}{\sigma}Y = \frac{1}{\sigma}(X\beta + \sigma\varepsilon) = X\beta' + \varepsilon.$$

On the contrary, if it is not true, we can replace it a posteriori by an unbiased estimator (such as $\hat{\sigma}_{full}^2$ or $\hat{\sigma}_{restricted}^2$) once the statistics are developed.

3.2.1 Unbiased estimator of the estimation loss

Applying Definition 3.5 to our context, we obtain the following theorem.

Theorem 3.2 (Unbiased estimator of the quadratic loss under Gaussian assumption). *Let $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$. Let $\hat{\beta} = \hat{\beta}(Y)$ be an estimator of β such that $X\hat{\beta}$ is weakly differentiable with respect to Y and let $\hat{\sigma}^2$ be an unbiased estimator of σ^2 independent of $\text{div}_Y(X\hat{\beta})$. Then*

$$\hat{L}_0(\hat{\beta}) = \|Y - X\hat{\beta}\|^2 + (2 \text{div}_Y(X\hat{\beta}) - n)\hat{\sigma}^2 \quad (3.19)$$

is an unbiased estimator of $\|X\hat{\beta} - X\beta\|^2$.

Proof of Theorem 3.2. The risk of $X\hat{\beta}$ at $X\beta$ is

$$\begin{aligned} \mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2] &= \mathbb{E}_\beta[\|X\hat{\beta} - Y\|^2 + \|Y - X\beta\|^2] \\ &\quad + \mathbb{E}_\beta[2(Y - X\beta)^t(X\hat{\beta} - Y)]. \end{aligned} \quad (3.20)$$

Since $Y \sim \mathcal{N}_n(X\beta, \sigma^2 \mathbf{I}_n)$, we have

$$\mathbb{E}_\beta[\|Y - X\beta\|^2] = n\sigma^2$$

leading to

$$\mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2] = \mathbb{E}_\beta[\|Y - X\hat{\beta}\|^2] - n\sigma^2 + 2 \operatorname{tr}(\operatorname{cov}_\beta(X\hat{\beta}, Y - X\beta)). \quad (3.21)$$

Moreover, applying Stein's identity for the right-most part of the expectation in (3.20) with $g(Y) = X\hat{\beta}$ and assuming that $X\hat{\beta}$ is weakly differentiable with respect to Y , we can rewrite (3.20) as

$$\mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2] = \mathbb{E}_\beta[\|Y - X\hat{\beta}\|^2] - n\sigma^2 + 2\sigma^2 \mathbb{E}_\beta[\operatorname{div}_Y X\hat{\beta}].$$

Hence, according to Definition 3.5 and to the development of the risk $R(X\hat{\beta}, X\beta)$, the statistic $\hat{L}_0(\hat{\beta})$ is an unbiased estimator of $\|X\hat{\beta} - X\beta\|^2$ since $\hat{\sigma}^2$ is an unbiased estimator of σ^2 independent from Y . \square

Note that this result is similar to that obtained by [Stein 1981] in the context of estimating a multivariate normal mean.

3.2.2 Links with C_p , AIC and FPE

Same estimator of variance for all submodels

Practical criteria. In order to make the following discussion clearer, we recall here the formula of the three criteria of interest for the Gaussian assumption, namely the unbiased estimator of loss \hat{L}_0 , Mallows' C_p and the extended version of AIC proposed by [Ye 1998]:

$$\begin{aligned} \hat{L}_0(\hat{\beta}) &= \|Y - X\hat{\beta}\|^2 + (2 \operatorname{div}_Y(X\hat{\beta}) - n)\hat{\sigma}^2 \\ C_p(\hat{\beta}) &= \frac{\|Y - X\hat{\beta}\|^2}{\hat{\sigma}^2} + 2 \operatorname{div}_Y(X\hat{\beta}) - n \\ \text{AIC}(\hat{\beta}) &= \frac{\|Y - X\hat{\beta}\|^2}{\hat{\sigma}^2} + 2 \operatorname{div}_Y(X\hat{\beta}). \end{aligned}$$

Using the estimator $\hat{\sigma}_{full}^2$ of the variance in (3.15), we thus obtain the following link between \hat{L}_0 , C_p and AIC:

$$\hat{L}_0(\hat{\beta}) = \hat{\sigma}_{full}^2 \times C_p(\hat{\beta}) = \hat{\sigma}_{full}^2 \times (\text{AIC}(\hat{\beta}) - n). \quad (3.22)$$

For more discussion on equivalence with other model selection criteria see for instance [Li 1985], [Shao 1997] and [Efron 2004].

Theoretical criteria. These links between different criteria function for model selection are due to the fact that, under our working hypothesis (linear model, quadratic loss, normal distribution $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$ for a fixed design matrix X), they can be seen as unbiased estimators of related quantities of interest. We will now recall these quantities. An important quantity for the practitioner is the prediction error $PE(\hat{\beta}_I, \beta)$ measuring the expected discrepancy between the predicted values $X\hat{\beta}_I$ and a new observation Y_{new} for a given estimation $\hat{\beta}_I$:

$$\begin{aligned} PE(\hat{\beta}_I, \beta) &= \mathbb{E}[\|Y_{\text{new}} - X\hat{\beta}_I\|^2] \\ &= \mathbb{E}[\|Y_{\text{new}} - X\beta\|^2] + \|X\beta - X\hat{\beta}_I\|^2 \\ &= n\sigma^2 + \|X\beta - X\hat{\beta}_I\|^2. \end{aligned}$$

The prediction error is minimal when $\hat{\beta}_I = \beta$ and its value is $PE^* = PE(\beta, \beta) = n\sigma^2$. Some prefer focus on the excess of prediction error, namely, the loss $L(\hat{\beta}_I, \beta) = PE(\hat{\beta}_I, \beta) - PE^*$ (sometimes referred to as the total squared error). It turns out that this is also the quadratic loss function since

$$L(\hat{\beta}_I, \beta) = \|X\hat{\beta}_I - X\beta\|^2.$$

and thus prediction and estimation are equivalent goals. This is the reason why the following quantity is also referred to as the *invariant loss*, or the *scaled predictive error*,

$$L^{\text{inv}}(\hat{\beta}_I, \beta) = \frac{L(\hat{\beta}_I, \beta)}{\sigma^2} = \frac{\|X\hat{\beta}_I - X\beta\|^2}{\sigma^2},$$

which is the scale-invariant loss used for instance in minimax analysis.

Under our working hypotheses the Kullback-Leibler divergence $D_{KL}(\hat{\beta}_I, \beta)$ is also related to these quantities since

$$\begin{aligned} D_{KL}(\hat{\beta}_I, \beta) &= \mathbb{E} \left[-\log \left(\frac{f(Y_{\text{new}}|\hat{\beta})}{f(Y_{\text{new}}|\beta)} \right) \right] \\ &= \mathbb{E}[(\|Y_{\text{new}} - X\hat{\beta}_I\|^2 - \|Y_{\text{new}} - X\beta\|^2)/2\sigma^2] \\ &= \frac{PE(\hat{\beta}_I, \beta)}{2\sigma^2} - \frac{n}{2} \\ &= \frac{1}{2}L^{\text{inv}}(\hat{\beta}_I, \beta). \end{aligned} \tag{3.23}$$

On the other hand, the expected log-likelihood is

$$Q(\hat{\beta}_I, \beta) = Q(\beta; \beta) - D_{KL}(\hat{\beta}_I, \beta) = -\frac{PE(\hat{\beta}_I, \beta)}{2\sigma^2}.$$

This quantities are linearly related provided that

$$L(\hat{\beta}_I, \beta) = PE(\hat{\beta}_I, \beta) - n\sigma^2 = \sigma^2 \times L^{\text{inv}}(\hat{\beta}_I, \beta) = \sigma^2 \times (-2Q(\hat{\beta}_I, \beta) - n).$$

Thus the unbiased loss estimation principle of minimizing an unbiased estimator of any of these quantities will provide the same selection.

Mallows' C_p was originally designed as an unbiased estimator of the expected scaled sum of squared errors $\mathbb{E}_\beta[L^{\text{inv}}(\hat{\beta}_I, \beta)] = \mathbb{E}_\beta[(\hat{\beta}_I - \beta)^t X^t X (\hat{\beta}_I - \beta)/\sigma^2]$, which is in fact the scale-invariant risk of $\hat{\beta}_I$, that is, $R^{\text{inv}}(\hat{\beta}) = \mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2/\sigma^2]$. Akaike also originally considered AIC as an

estimator of the expectation of a loss function $\mathbb{E}_\beta[Q(\hat{\beta}_I, \beta)]$. When $\hat{\beta}$ is sufficiently close to β it admits the approximation

$$2 D_{KL}(\hat{\beta}_I, \beta) \approx \|\hat{\beta} - \beta\|_{\mathcal{I}}^2 = (\hat{\beta} - \beta)^t \mathcal{I} (\hat{\beta} - \beta),$$

where \mathcal{I} is the Fisher-information matrix defined by

$$\mathcal{I} = \left(-\mathbb{E} \left[\frac{\partial^2 \log p(Y_{\text{new}} | \hat{\beta})}{\partial \beta_i \partial \beta_j} \right] \right)_{i,j=1}^n$$

and equals $X^t X / \sigma^2$ for linear models.

Model selection. The final objective is to select the “best” model among those at hand. This can be performed by minimizing either of the three proposed criteria, that is the unbiased estimator of loss \hat{L}_0 , C_p and AIC. The idea behind this heuristic is that the best model in terms of prediction is the one minimizing the loss $\|X\hat{\beta} - X\beta\|^2$. All three criteria estimate this loss, and so the hope is that their minimum will coincide with the minimum of the loss, or at least will “mimic” it. Now, from (3.22), it can be easily seen that the three criteria differ from each other only up to a multiplicative and/or additive constant. Hence the models selected by the three criteria will be the same.

We would like to point out that Theorem 3.2 does not use the hypothesis that the model is linear and are valid for nonlinear models

$$Y = f(X) + \sigma \varepsilon.$$

Therefore \hat{L}_0 generalizes C_p to non linear models. Moreover, following its definition (2.32), AIC implementation requires the specification of the underlying distribution. In this sense it is considered as a generalization of C_p for non Gaussian distributions. However, in practice, we might only have a vague intuition of nature of the underlying distribution and we might not be able to give its specific form. We will see in the following section that \hat{L}_0 , which is equivalent to the Gaussian AIC as we have just seen, can be also derived from a more general distribution context, the one of spherically symmetric distributions, with no need to specify the precise form of the distribution.

Different estimator of variance for each submodel with the least-squares estimator

Practical criteria In many articles and books, like for instance [McQuarrie & Tsai 1998], [Burnham & Anderson 2002], or [Claeskens & Hjort 2008], AIC is found in a different form, namely

$$\text{AIC}(\hat{\beta}_I^{LS}) = n \log \left(\frac{\|X_I \hat{\beta}_I^{LS} - Y\|^2}{n} \right) + 2k.$$

This expression actually corresponds to the case where both β and σ^2 are estimated by maximum likelihood based on a subset I . Indeed, the log-likelihood of the Gaussian distribution is

$$\log p(Y|X_I, \beta_I) = -\frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{\|Y - X_I \beta_I\|^2}{2\sigma^2}.$$

Estimating σ^2 by $\hat{\sigma}_{MLE}^2$ in (3.17) thus yields the following estimator of the log-likelihood

$$\begin{aligned}\log p(Y|X_I, \hat{\beta}_I^{LS}) &= -\frac{n}{2} \log \left(\frac{\|Y - X_I \hat{\beta}_I^{LS}\|^2}{n} \right) - \frac{n}{2} \log(2\pi) - \frac{\|Y - X_I \hat{\beta}_I^{LS}\|^2}{2\|Y - X_I \hat{\beta}_I^{LS}\|^2/n} \\ &= -\frac{n}{2} \log \left(\frac{\|Y - X_I \hat{\beta}_I^{LS}\|^2}{n} \right) - \frac{n}{2} \log(2\pi) - \frac{n}{2}.\end{aligned}$$

Noticing that the two last terms are constant with respect to $\hat{\beta}_I^{LS}$, we obtain the desired result. This form is very different from the one given in the previous paragraph, where the estimator of the variance was the same for all submodels, and thus the comparison with C_p and \hat{L}_0 does not stand anymore.

However, in the case where σ^2 is estimated by $\hat{\sigma}_{restr}^2$ in (3.16), and where the estimator of β is taken to be the least-squares estimator on subset I , there is a certain similarity with another criterion developed by Akaike, namely the Final Prediction Error (FPE) criterion [Akaike 1970]. In this particular case, \hat{L}_0 and FPE take the following expressions

$$\hat{L}_0(\hat{\beta}_I^{LS}) = \frac{k}{n-k} \|X_I \hat{\beta}_I^{LS} - Y\|^2, \quad \text{FPE}(\hat{\beta}_I^{LS}) = \frac{n+k}{n-k} \|X_I \hat{\beta}_I^{LS} - Y\|^2,$$

where k is the size of subset I . Note that this expression for \hat{L}_0 is the one derived in [Fourdrinier & Wells 1994].

We can easily see that the link between \hat{L}_0 and FPE is

$$\hat{L}_0(\hat{\beta}_I^{LS}) = \frac{k}{n+k} \text{FPE}(\hat{\beta}_I^{LS}).$$

Theoretical criteria FPE has been derived with another objective: it was built in the context of estimation of the parameter in a linear autoregressive model. By analogy with the linear regression model, we can say that it estimates the prediction error defined in the previous paragraph, namely

$$PE(\hat{\beta}_I, \beta) = \mathbb{E}[\|Y_{\text{new}} - X \hat{\beta}_I\|^2],$$

which is equivalent to $L(\hat{\beta}_I, \beta)$ up to the additive constant $n\sigma^2$. This explains the similarity between the two corresponding practical criteria.

Model selection A disturbing fact is that \hat{L}_0 and FPE are not equivalent since in this paragraph we consider the case where we estimate σ^2 differently for each submodel, even though their corresponding theoretical criteria are equivalent. Therefore there is little chance that \hat{L}_0 and FPE both select the same model.

3.3 The Gaussian case with unknown variance

In this section, we assume σ^2 to be unknown. We consider that the estimator of β can be written as

$$\hat{\beta} = \hat{\beta}^{LS} + g(\hat{\beta}^{LS}, S), \quad (3.24)$$

where, for any s , $g(\cdot, s)$ is a weakly differentiable function and where S is a nonnegative random variable such that $S \sim \sigma^2 \chi^2(d)$. To be more precise, we can take

$$S = \|Y - X \hat{\beta}^{LS}\|^2, \quad (3.25)$$

which is distributed as a χ^2 distribution with $n - p$ degrees of freedom (up to the factor σ^2). The expression in (3.24) takes into account three different cases:

- the case where $\hat{\beta}$ does not depend on S , that is $g(\hat{\beta}^{LS}, S) = g(\hat{\beta}^{LS})$; examples from this case are the restricted least-squares estimator (if we consider the full model), the James-Stein estimator and regularization methods if we do not consider them to depend on the variance;
- the case where $\hat{\beta}$ depends on S through a separable function, that is $g(\hat{\beta}^{LS}, S) = Sg(\hat{\beta}^{LS})$; examples from this case are the generalized James-Stein estimator;
- and the more general case where $\hat{\beta}$ depends on S through a non separable function $g(\hat{\beta}^{LS}, S)$; examples from this case are regularization methods if we consider them to depend on the variance.

Example 3.1 (Lasso). [Tibshirani 1996] showed that the Lasso optimization problem

$$\min_{\beta} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}$$

is equivalent, from a Bayesian point of view, to the maximization of the log-likelihood of the hierarchical model

$$\begin{aligned} Y|X, \beta, \sigma^2 &\sim \mathcal{N}_n(X\beta, \sigma^2 \mathbf{I}_n) \\ \forall 1 \leq j \leq p, \quad \beta_j | \sigma^2 &\sim \mathcal{L}(\sigma^2 / \lambda), \end{aligned}$$

where \mathcal{L} is the Laplace distribution with mean 0 and scale parameter $1/b = \sigma^2 / \lambda$. Hence, we have that $\lambda = b\sigma^2$, so that we can consider that the hyperparameter λ implicitly takes into account an estimator of the variance, that is, $\lambda = b'S$. This decomposition is interesting in particular for the regularization path where λ depends on the data.

Now, it has been shown in [Zou et al. 2007] that, if we knew in advance the subset I of nonzero components in $\hat{\beta}^{lasso}$ for a given hyperparameter λ , then it can be expressed as

$$\hat{\beta}_I^{lasso} = (X_I^t X_I)^{-1} (X_I^t Y - \lambda \operatorname{sgn}(\hat{\beta}_I^{lasso})) = \hat{\beta}_I^{LS} - \lambda (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{lasso}).$$

In view of this expression and of the decomposition of λ into a constant term and an estimator of the variance, we can thus re-express $\hat{\beta}^{lasso}$ as in (3.24). Indeed, under the linear model assumption restricted to the subset I , then we have directly

$$\hat{\beta}_I^{lasso} = \hat{\beta}_I^{LS} + Sg_r^{lasso}(\hat{\beta}_I^{LS}),$$

with $g_r^{lasso}(\hat{\beta}_I^{LS}) = b' (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{lasso})$. On the other hand, under the full model assumption, we have the slightly more complex case

$$\hat{\beta}^{lasso} = \hat{\beta}^{LS} + g_f^{lasso}(\hat{\beta}^{LS}, S),$$

with $g_f^{lasso}(\hat{\beta}^{LS}, S) = \hat{\beta}^{LS} - \hat{\beta}_I^{LS} + Sg_r^{lasso}(\hat{\beta}_I^{LS})$.

In the case of unknown variance, it is common to consider the invariant loss

$$L^{inv}(\beta, \hat{\beta}) = \frac{\|X\hat{\beta} - X\beta\|^2}{\sigma^2} \quad (3.26)$$

instead of the loss $L(\beta, \hat{\beta})$ (see for instance [Brandwein & Strawderman 1991], [Maruyama 2003] and [Fourdrinier & Strawderman 2010]). This way, the influence of the noise level is explicitly modeled.

We thus need an extension of Stein's identity for the case of unknown variance, given in [Fourdrinier & Wells 2012]. In the sequel $\mathbb{E}_{\mu, \sigma^2}$ denotes the expectation of Y where both the mean μ and the variance σ^2 are unknown.

Theorem 3.3 (Stein's identity for the unknown variance case). *Let $Y \sim \mathcal{N}_n(\mu, \sigma^2 \mathbf{I}_n)$ where σ^2 is unknown and is estimated by a function of $S \sim \sigma^2 \chi^2(d)$, and let $h : \mathbb{R}^n \times \mathbb{R}_+ \mapsto \mathbb{R}^n$. If $h(\cdot, s)$ is weakly differentiable, then*

$$\mathbb{E}_{\mu, \sigma^2}[(Y - \mu)^t h(Y, S)/\sigma^2] = \mathbb{E}_{\mu, \sigma^2}[\text{div}_Y h(Y, S)],$$

provided both expectations exist.

Proof of Theorem 3.3. The proof is given in [Fourdrinier & Wells 2012]. □

In order to derive the unbiased estimator of loss and to compare it to corrected estimators, we also need the following result, once again taken from [Fourdrinier & Wells 2012], which consists in applying twice Theorem 3.3.

Corollary 3.1. *Let $Y \sim \mathcal{N}_n(\mu, \sigma^2 \mathbf{I}_n)$ where σ^2 is unknown and is estimated by a function of $S \sim \sigma^2 \chi^2(d)$, and let $\varphi : \mathbb{R}^n \times \mathbb{R}_+ \mapsto \mathbb{R}$. If $\varphi(\cdot, s)$ is twice weakly differentiable, then*

$$\mathbb{E}_{\mu, \sigma^2}[\varphi(Y, S)/\sigma^2] = \mathbb{E}_{\mu, \sigma^2}[2\partial\varphi(Y, S)/\partial S] + \mathbb{E}_{\mu, \sigma^2}[(d - 2)S^{-1}\varphi(Y, S)],$$

provided the expectations exist.

Proof of Corollary 3.1. The proof is given in [Fourdrinier & Wells 2012]. □

3.3.1 Unbiased estimator of the invariant estimation loss

From Theorem 3.3 and Corollary 3.1, we derive the following theorem.

Theorem 3.4 (Unbiased estimator of the invariant quadratic loss under Gaussian assumption with unknown variance). *Let $Y \sim \mathcal{N}_n(X\beta, \sigma^2 \mathbf{I}_n)$ where both β and σ^2 are unknown. Let $\hat{\beta} = \hat{\beta}(Y)$ be an estimator of β such that $X\hat{\beta}$ is weakly differentiable with respect to Y and rewrite it as*

$$\hat{\beta} = \hat{\beta}^{LS} + g(\hat{\beta}^{LS}, S)$$

where $S = \|Y - X\hat{\beta}^{LS}\|^2 \sim \sigma^2 \chi^2(n - p)$. Then

$$\hat{L}_0^{inv}(\hat{\beta}) = (n - p - 2) \frac{\|Y - X\hat{\beta}\|^2}{S} + 2 \text{div}_Y(X\hat{\beta}) - n + 4(X\hat{\beta} - Y)^t X \frac{\partial g(\hat{\beta}^{LS}, S)}{\partial S} \quad (3.27)$$

is an unbiased estimator of the invariant loss $\|X\hat{\beta} - X\beta\|^2/\sigma^2$.

Remark 3.1. Note that, for the estimators we consider in this manuscript, the right-most term in (3.27) actually becomes

$$\frac{\partial g(\hat{\beta}^{LS}, S)}{\partial S} = \frac{\hat{\beta} - \hat{\beta}^{LS}}{S}.$$

This can be easily seen for Example 3.1, even for the full model, since the term $\hat{\beta}^{LS} - \hat{\beta}_I^{LS}$ does not depend on S .

Proof of Theorem 3.4. The invariant quadratic risk of $X\hat{\beta}$ at $X\beta$ is

$$\mathbb{E}_{\beta, \sigma^2} \left[\frac{\|X\hat{\beta} - X\beta\|^2}{\sigma^2} \right] = \mathbb{E}_{\beta, \sigma^2} \left[\frac{\|X\hat{\beta} - Y\|^2}{\sigma^2} + \frac{\|Y - X\beta\|^2}{\sigma^2} + \frac{2(Y - X\beta)^t(X\hat{\beta} - Y)}{\sigma^2} \right],$$

where $\mathbb{E}_{\beta, \sigma^2}$ denotes the expectation under Y parametrized by (β, σ^2) . Since $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, we have that

$$\mathbb{E}_{\beta, \sigma^2} \left[\frac{\|Y - X\beta\|^2}{\sigma^2} \right] = n$$

leading to

$$\mathbb{E}_{\beta, \sigma^2} \left[\frac{\|X\hat{\beta} - X\beta\|^2}{\sigma^2} \right] = \mathbb{E}_{\beta, \sigma^2} \left[\frac{\|Y - X\hat{\beta}\|^2}{\sigma^2} \right] - n + 2 \frac{\text{tr}(\text{cov}_{\beta, \sigma^2}(X\hat{\beta}, Y - X\beta))}{\sigma^2}. \quad (3.28)$$

Moreover, applying Stein's identity for the unknown variance case (Theorem 3.3) to the right-most part of the expectation in (3.28) with $h(Y) = X\hat{\beta}$, we can rewrite (3.20) as

$$\mathbb{E}_{\beta, \sigma^2} \left[\frac{\|X\hat{\beta} - X\beta\|^2}{\sigma^2} \right] = \mathbb{E}_{\beta, \sigma^2} \left[\frac{\|Y - X\hat{\beta}\|^2}{\sigma^2} \right] - n + 2 \mathbb{E}_{\beta, \sigma^2} [\text{div}_Y X\hat{\beta}],$$

where we assumed that $X\hat{\beta}$ is weakly differentiable with respect to Y . Finally, Corollary 3.1 applied with $\varphi(Y) = \|Y - X\hat{\beta}\|^2$ to the left-most part of (3.28) yields

$$\begin{aligned} \mathbb{E}_{\beta, \sigma^2} \left[\frac{\|Y - X\hat{\beta}\|^2}{\sigma^2} \right] &= \mathbb{E}_{\beta, \sigma^2} \left[2 \frac{\partial \|Y - X\hat{\beta}\|^2}{\partial S} + (n - p - 2) \frac{\|Y - X\hat{\beta}\|^2}{S} \right] \\ &= \mathbb{E}_{\beta, \sigma^2} \left[2 \frac{\partial \|Y - X(\hat{\beta}^{LS} + g(\hat{\beta}^{LS}, S))\|^2}{\partial S} + (n - p - 2) \frac{\|Y - X\hat{\beta}\|^2}{S} \right] \\ &= \mathbb{E}_{\beta, \sigma^2} \left[4(Y - X\hat{\beta})^t X \frac{\partial g(\hat{\beta}^{LS}, S)}{\partial S} + (n - p - 2) \frac{\|Y - X\hat{\beta}\|^2}{S} \right], \end{aligned}$$

where $S = \|Y - X\hat{\beta}^{LS}\|^2$.

Hence, according to Definition 3.5 and to the development of the invariant risk $R(X\hat{\beta}, X\beta)$, the statistic $\hat{L}_0^{inv}(\hat{\beta})$ is an unbiased estimator of $\|X\hat{\beta} - X\beta\|^2/\sigma^2$. \square

Note that, when $g(\hat{\beta}^{LS}, S) = g(\hat{\beta}^{LS})$, that is, when $\hat{\beta}$ does not depend on the variance, we have

$$\hat{L}_0^{inv}(\hat{\beta}) = \frac{(n - p - 2)}{\|Y - X\hat{\beta}^{LS}\|^2} \left(\|Y - X\hat{\beta}\|^2 + (2 \text{div}(X\hat{\beta}) - n) \frac{\|Y - X\hat{\beta}^{LS}\|^2}{n - p - 2} \right).$$

Comparing with the unbiased estimator of $\|X\hat{\beta} - X\beta\|^2$ (with independent estimator of the variance)

$$\hat{L}_0(\hat{\beta}) = \|Y - X\hat{\beta}\|^2 + (2 \operatorname{div}(X\hat{\beta}) - n) \frac{\|Y - X\hat{\beta}^{LS}\|^2}{n - p},$$

which we derived in previous section, we can see that the main difference results in the denominator of the estimator of the variance. Indeed, the denominator $n - p$ is transformed into one of $n - p - 2$, which can be interpreted as a correction for not knowing the variance.

3.3.2 Link with AIC_c

AIC_c is a corrected version of AIC proposed by [Sugiura 1978] and extended by Hurvich and Tsai in a series of papers [Hurvich & Tsai 1989, Hurvich & Tsai 1991, Hurvich & Tsai 1993]. It is designed to correct AIC's bias for the finite sample setting, since AIC was derived to be unbiased only asymptotically. Hence, the theoretical criterion that AIC_c intends to estimate is the same as for AIC, namely the expected likelihood (up to the factor -2). We recall that AIC_c takes the form

$$\text{AIC}_c(\hat{\beta}) = -2 \log p(Y|X, \hat{\beta}) + \frac{n(n+p)}{n-p-2}.$$

In the particular case where $\hat{\beta}$ is not a function of $S = \|Y - X\hat{\beta}^{LS}\|^2$ and where we take the full model in S , \hat{L}_0^{inv} and AIC_c take the following expressions

$$\begin{aligned} \hat{L}_0^{inv}(\hat{\beta}) &= (n-p-2) \frac{\|Y - X\hat{\beta}\|^2}{\|Y - X\hat{\beta}^{LS}\|^2} + 2 \operatorname{div}_Y(X\hat{\beta}^{LS}) - n, \\ \text{AIC}_c(\hat{\beta}) &= n \frac{\|Y - X\hat{\beta}\|^2}{\|Y - X\hat{\beta}^{LS}\|^2} + \frac{n(n+p)}{n-p-2}. \end{aligned}$$

We can easily see that the link between \hat{L}_0^{inv} and AIC_c is

$$\hat{L}_0^{inv}(\hat{\beta}) = \frac{n-p-2}{n} \text{AIC}_c(\hat{\beta}) + 2 \operatorname{div}_Y(X\hat{\beta}) - 2n - p.$$

We can clearly notice a resemblance between the two criteria, although the penalty function (the right-most part of each criterion) is different and thus might lead to a different selection of the best model.

3.4 The spherical case

3.4.1 The class of multivariate spherically symmetric distributions

Description and properties of the spherical class

The previous chapter dealt with the Gaussian case with covariance matrix proportional to identity. In that case, the results for univariate and multivariate random vectors only differ up to a factor n , because of the independence between observations.

The wide (and sometimes systematic) use of the Gaussian law arises from its numerous properties, such as easy calculations of probabilities and moments and since it is the limit distribution of many statistical quantities. However, it is not adapted to all kind of data, in particular in presence of extreme values or outliers or in the non asymptotic setting. Moreover,

it is now generally accepted, since Huber's work [Huber 1975], that it is important to propose robust methods. Indeed, distributional robustness preserves good properties of the methods when the true underlying distribution departs from the Gaussian law.

We propose to treat such robustness by enlarging the distributional assumption for the error component to a family of distributions generalizing the Gaussian law. Before going into more details on this generalization, we recall the characterization of the Gaussian distribution. Let $Y = (y_1, \dots, y_n)^t$ be a random vector,

$$Y \text{ is Gaussian} \Leftrightarrow \begin{cases} (1) \forall i \neq j \quad y_i \text{ is independent of } y_j, \\ (2) Y \text{ is spherical.} \end{cases} \quad (3.29)$$

This characterization, reported in [Chmielewski 1981] and in [Kariya & Sinha 1989], has been initially proposed by Maxwell in 1860 [Maxwell 1860]. It allows two natural generalizations, as pointed out by [Fan & Fang 1985]: the first one considers distributions verifying the independence property, such as the exponential family of distributions, while the second one relax the independence assumption to the benefit of spherical symmetry. Both generalizations go in different directions and have led to fruitful works (see [Brown 1986] for the exponential family and [] for the spherical family). Note that their only common member is the Gaussian distribution.

In the sequel, we choose to work with the spherical family. This family preserve some of the interesting properties of the Gaussian distribution, as shown in [Fang *et al.* 1989], and appears to be well suited to our work. These properties are orthogonal invariance, invariance by translation and exchangeability. We develop each property along with their relevance in the following paragraphs.

Orthogonal invariance Let us begin with the definition of orthogonal invariance.

Definition 3.6 (Orthogonal invariance). *Let $\mathcal{O}(n)$ be the set of $n \times n$ orthogonal matrices, that is such that $H^t H = H H^t = I_n$. An n -dimensional random vector Y is said to be orthogonally invariant if, for any orthogonal matrix H in $\mathcal{O}(n)$, the random vector $Z = HY$ is distributed as Y .*

The orthogonal invariance property is actually used to define spherically symmetric distributions.

Definition 3.7 (Spherical symmetry). *A random vector $Y \in \mathbb{R}^n$ (equivalently the distribution of Y) is said to be spherically symmetric around $\mu \in \mathbb{R}^n$ if $Y - \mu$ is orthogonally invariance. We denote this by $Y \sim \mathcal{S}_n(\mu)$.*

Orthogonal invariance incurs as a consequence that several test statistics have unchanged null distribution for all the family, including the Gaussian law. This result is however not always true for the nonnull distribution. [Kariya & Sinha 1989] give conditions for which the null or the nonnull distribution is the same as that under Gaussian assumption. Among those tests, we are particularly interested in Student and Fisher tests for the nullity of one or several regression coefficients. Hence, it is still coherent to use them as stopping criterion for Forward Selection or Backward Elimination mentioned in Chapter 2.

Moreover, this property is also interesting for the extension of our results on loss estimation presented in Section 3.2. Indeed, orthogonal invariance allows the transformation of the usual linear model into its canonical form with residual vector, while keeping the distribution of the model unchanged. The canonical form facilitates the derivation of loss estimators. We will give more details on the canonical form in Section 3.4.1.

Invariance by translation

Definition 3.8 (Invariance by translation). *A random vector Y is said to be invariant by translation if, for all arbitrary vector $b \in \mathbb{R}^n$, the distribution of the vector $Z = Y + b$ is the distribution of Y translated by b .*

The property of invariance by translation is important for the linear model: if we assume ε to be spherically symmetric and if we assume X to be determinist, then the distribution of the vector $Y = X\beta + \varepsilon$ is the distribution of ε translated from the origin to $X\beta$.

Exchangeability

Definition 3.9 (Exchangeability). *A random vector Y is said exchangeable if, for all permutation $\pi = \{i_1, \dots, i_n\}$ of $\{1, \dots, n\}$, the vector $Y_\pi = (Y_{i_1}, \dots, Y_{i_n})$ is distributed as Y .*

Exchangeability is in fact a particular case of orthogonal invariance where the elements of the orthogonal matrix H takes only the values 0 and 1. Hence, we can define the set of permutation matrices by

$$\{P \in \mathcal{O}(n) \mid P_{i,j} \in \{0, 1\} \forall 1 \leq i, j \leq n\}.$$

The case where the components of the vector Y are independent is in turn a particular case of exchangeability. Indeed, in that case, the density of Y , when it exists, can be written as the product of the marginal density of each component. Hence the application of a permutation does not modify the distribution.

Along with these properties of spherical distributions, we can add a forth one characterizing the generalization of the spherical family to the elliptical family: the property of linear invariance.

Linear invariance Linear invariance is a more general property than orthogonal invariance and is defined as follows.

Definition 3.10 (Linear invariance). *Let $\mathcal{LI}(n)$ be the set of $n \times n$ non singular positive definite matrices. A group \mathcal{P} of distributions is said to be linearly invariant if, for any vector Y having distribution in \mathcal{P} , for any non singular matrix M in $\mathcal{LI}(n)$ and for any arbitrary vector $b \in \mathbb{R}^n$, the distribution of the vector $Z = MY + b$ is also a member of \mathcal{P} .*

In the Gaussian case, the linear invariance property extends the spherical Gaussian distribution to the elliptical Gaussian distribution: if $Y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$, then $Z = MY + b \sim \mathcal{N}_n(M\mu + b, \sigma^2 MM^t)$. This property is preserved for spherically symmetric distributions, and also for elliptically symmetric distributions.

Note that, according to [Kariya & Sinha 1989], the relaxation of the characterization in (3.29) to the property of independence assumption destroys the properties of orthogonal and linear invariance of the Gaussian distribution. Hence these latter two properties are not verified for distributions such as the exponential family of distributions.

Up to now, we have characterized the family of spherically symmetric distributions by the properties shared with the Gaussian law. We can also characterize it by its probability density, when it exists.

Definition 3.11 (Probability density of spherically symmetric distributions). *Let Y be a spherically symmetric vector around the location vector μ with scale parameter σ . If its distribution is absolutely continuous with respect to the Lebesgue measure in \mathbb{R}^n , then it has a density of the form*

$$p(y) = \frac{1}{\sigma^n} g\left(\frac{\|y - \mu\|^2}{\sigma^2}\right)$$

for a given function g from \mathbb{R}^n to \mathbb{R}_+ , called the generating function.

Remark 3.2. According to [Kelker 1970], only the distributions with an atom of weight at the origin do not admit a density. In the sequel, as we use results on expectations, we do not consider such distributions.

Among this family of laws, we can mention the Gaussian distribution, the Student distribution, the Kotz distribution, the exponential power distribution (also known as the generalized normal distribution), the spherical logistic distribution, etc. Tables 3.1 and 3.2 display these examples along with their density and its visualization in the bivariate case. Both tables summarize information taken from [Fang *et al.* 1989], [Gupta & Varga 1993], [Kotz *et al.* 2001], [Kotz & Nadarajah 2004]. Figures 3.1 and 3.2 show the differences for a bivariate vector $Y = (Y_1, Y_2)$ between the case where its components are jointly drawn from a bivariate spherically symmetric distribution and the case where its components are independently drawn from a univariate spherically symmetric distribution.

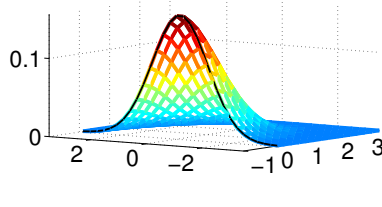
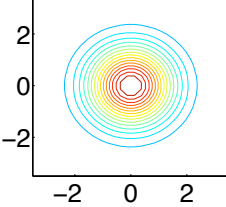
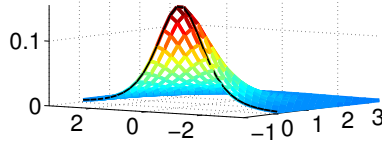
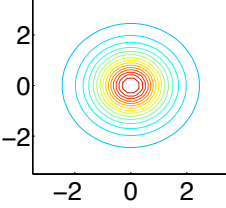
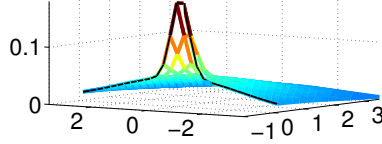
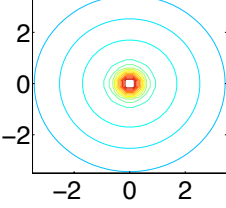
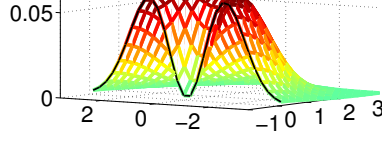
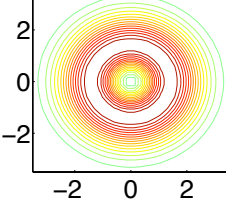
Law	Probability density	Param.	2D-visualization	Contours
Gaussian $\mathcal{N}_n(t; \mu, \sigma^2)$	$p(y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\ y-\mu\ ^2}{2\sigma^2}}$			
Student $\mathcal{T}_n(t; \mu, \sigma^2, \nu)$	$p(y) = \frac{\Gamma(\frac{n+\nu}{2})}{(\pi\sigma^2\nu)^{\frac{n}{2}}\Gamma(\frac{\nu}{2})} \left(1 + \frac{\ y-\mu\ ^2}{\nu\sigma^2}\right)^{-\frac{(\nu+n)}{2}}$	$\nu \geq 1$	$\nu=3$ 	$\nu=3$ 
Gaussian mixtures $\mathcal{GM}_n(t; \mu, \sigma^2, G)$	$p(y) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int_0^\infty \frac{1}{v^{\frac{n}{2}}} e^{-\frac{\ y-\mu\ ^2}{2v\sigma^2}} G(dv)$	$\int G(dv) = 1$	$p(v=0.1)=0.3, p(v=5)=0.7$ 	$p(v=\{0.1;5\})=\{0.3;0.7\}$ 
Kotz $\mathcal{K}_n(t; \mu, \sigma^2, N, r)$	$p(y) = \frac{\Gamma(\frac{n}{2}) r^{\frac{2N-2+n}{2}}}{\pi^{\frac{n}{2}} \sigma^{n+2N-1} \Gamma(\frac{2N-2+n}{2})} \ y-\mu\ ^{2(N-1)} e^{-r \frac{\ y-\mu\ ^2}{\sigma^2}}$	$N > \frac{2-n}{2},$ $r > 0$	$N=2, r=1$ 	$N=2, r=1$ 

Table 3.1: Examples of spherically symmetric distributions and their visualization for $n = 2$ - Part 1.

Law	Probability density	Param.	2D-visualization	Contours
Exponential power $\mathcal{EP}_n(t; \mu, \sigma^2, b)$	$p(y) = \frac{n\Gamma\left(\frac{n}{2}\right)}{(\pi\sigma^2)^{\frac{n}{2}} 2^{1+\frac{1}{2b}} \Gamma\left(1 + \frac{n}{2b}\right)} e^{-\frac{1}{2}\left(\frac{\ y-\mu\ ^2}{\sigma^2}\right)^b}$	$b > 0$		
Logistic $\mathcal{Log}_n(t; \mu, \sigma^2)$	$p(y) = \frac{(\sum_{j=1}^{\infty} (-1)^{j-1} j^{1-n/2})^{-1}}{(2\pi\sigma^2)^{\frac{n}{2}}} \frac{e^{-\frac{\ y-\mu\ ^2}{2\sigma^2}}}{\left(1 + e^{-\frac{\ y-\mu\ ^2}{2\sigma^2}}\right)^2}$			
Laplace $\mathcal{L}_n(t; \mu, \sigma^2)$	$p(y) = \frac{1}{2^{\frac{n}{2}-1} (\pi\sigma^2)^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} K_0\left(\sqrt{2} \frac{\ y-\mu\ }{\sigma}\right)$			
Bessel $\mathcal{B}_n(t; \mu, \sigma^2, q, r)$	$p(y) = \frac{(-1)^{q+1} (2r)^{-q-n} \pi}{\pi^{\frac{n}{2}} \Gamma\left(q + \frac{n}{2}\right) \sin(q\pi)} \ y - \mu\ ^q I_q\left(\frac{\sqrt{\ y - \mu\ ^2}}{r}\right),$ with $I_q(z) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(k+q+1)} \left(\frac{z}{2}\right)^{q+2k}$	$q > -\frac{n}{2},$ $r > 0$		

Table 3.2: Examples of spherically symmetric distributions and their visualization for $n = 2$ - Part 2.

A nice feature of the spherical family is that it brings together a distributions with a large spectrum of tails, from light to heavy, hence enabling the processing of data with a more or less important pourcentage of extreme values. In particular, the works of [Kariya & Sinha 1989] on the robustness of statistical tests shows another approach of distributional robustness than that proposed by Huber [Huber 1981]. Indeed, Huber deals with extreme values by removing part of the data (the components with higher and lower amplitude) in order to obtain robust estimators of the mean, the variance and other statistical quantities.

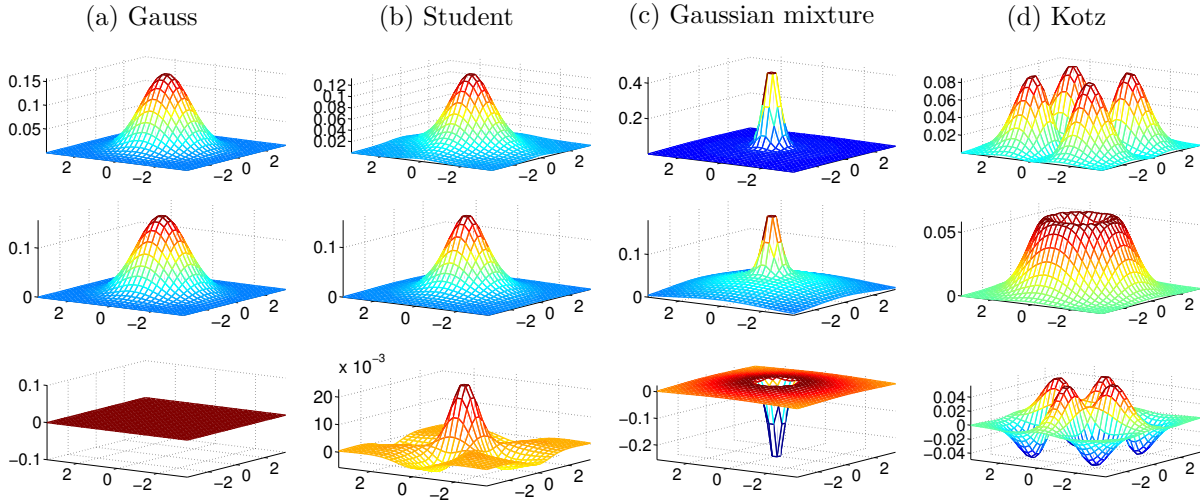


Figure 3.1: Difference between the independent case and the dependent case. Top: The components Y_1 and Y_2 are independently drawn from univariate spherically symmetric distributions. Middle: The components Y_1 and Y_2 are jointly drawn from bivariate spherically symmetric distributions. Bottom: Difference between the dependent and the independent cases. The parameters for each law are the same than in Table 3.1.

We refer the interested reader to [Kelker 1970] for a historical review on spherically symmetric distributions and to [Fang *et al.* 1989] for a more complete presentation (along with other characterizations of spherical distributions).

We can also characterize spherically symmetric distributions by mixtures of uniform distributions on spheres of radius R , where R is a positive random variable independent of the uniform direction U .

Definition 3.12 (Stochastic representation). *If $Y \in \mathbb{R}^n$ is a spherically symmetric random vector around μ , then Y can be decomposed as*

$$Y = \mu + RU$$

where $R = \|Y - \mu\|$ and $U = (Y - \mu)/\|Y - \mu\| \sim \mathcal{U}_{S_1}$, where \mathcal{U}_{S_1} is the uniform distribution on the n -dimensional sphere S_1 of unit radius.

Thereby, from the stochastic representation, we can see that the random vector $(Y - \mu)/\|Y - \mu\|$ follows the same (uniform) distribution whatever the distribution of Y is.

To conclude on this brief overview on spherically symmetric distributions, we recall the following result, taken from [Fang *et al.* 1989].

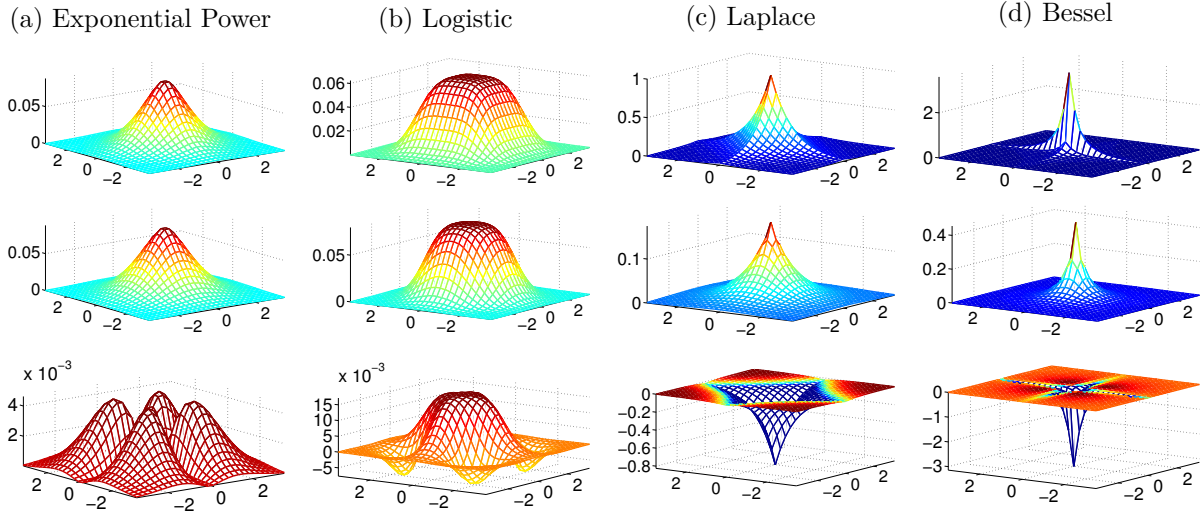


Figure 3.2: Difference between the independent case and the dependent case. Top: The components Y_1 and Y_2 are independently drawn from univariate spherically symmetric distributions. Middle: The components Y_1 and Y_2 are jointly drawn from bivariate spherically symmetric distributions. Bottom: Difference between the dependent and the independent cases. The parameters for each law are the same than in Table 3.2.

Theorem 3.5 (Mean and covariance). *Let Y be a spherically symmetric vector around μ , and let R and U be the radius and direction resulting from the stochastic representation of Y . If $\mathbb{E}[R^2] < \infty$, then the mean and the covariance of Y are given by*

$$\mathbb{E}_Y[Y] = \mu \quad \text{et} \quad \mathbb{E}_Y[YY^t] = \frac{1}{n} \mathbb{E}[R^2] \mathbf{I}_n.$$

This theorem shows that, although the components of a spherical vector are not independent, they are not correlated either. In addition, the variance $\sigma^2 = \mathbb{E}_Y[Y^t Y]/n$ only depends on the variance of R and we can conclude that $\sigma^2 = \mathbb{E}[R^2]/n$. This result can be easily checked for the multivariate Gaussian law $\mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$, whose squared radius is distributed as $\sigma^2 \chi^2(n)$. Similarly, for the Student distribution with ν degrees of freedom, the squared radius is distributed as a Fisher distribution $n \mathcal{Fisher}(n/2, \nu/2)$.

Practical applications

The spherical hypothesis, and more generally the elliptical hypothesis, can be very useful in many real life applications from the following fields: Physics (Wave Mechanics), Communication Theory, Signal Processing, Pattern Recognition, Finance, Economy, or Pharmacokinetics. Several practical examples are given in [Chmielewski 1981], [Du & Ma 2011], and references therein. The use of the elliptical assumption has improved the performances of methods from the state of the art in the description and the estimation of parameters of stochastic processes and random fields (temporal, spatial and spatiotemporal). We can cite for instance their use in random walks and in Ornstein-Uhlenbeck processes (see [Pchelintsev 2011], [Schroeter 1980], and [Lindsey & Jones 2000]). The latter two references respectively the analysis of weapon effectiveness with respect to the target surface and the study of a drug's concentration in blood for

clinical trials. This last quantity is particularly non Gaussian since, for each study, often one or two subjects have an extreme reaction to the drug being tested. Spherically symmetric distributions with heavy tails, such as the Student distribution or the generalized Gaussian distribution (also known as *power exponential distribution*) are thus interesting in such a situation.

Moreover, [Berk 1997] and, more recently, [Hafner & Rombouts 2007] have shown a case where the elliptical family is the largest family being most consistent with the *Capital Asset Pricing Model*, a model that aims at evaluating the return on investment of an asset. Another extension of Gaussian results to elliptical distributions, but this time for spatial and spatiotemporal processes, has been done for estimating variograms, used in kriging [Genton 2000], [Gneiting et al. 2007]. These works are useful for weather forecasting based on measures of temperature, wind speed and direction, and precipitation transmitted by several weather stations.

Most of the references cited in this paragraph have shown better performances than under the Gaussian assumption, especially when the true distribution seems to have heavy tails or when the independence assumption seems too restrictive.

Canonical form of the linear model

In this paragraph, we present the canonical form of the linear model. This form consists in an orthogonal transformation of the data and is closely related to the QR -factorization of the design matrix X . Although such a transformation is not compulsory for the extension of our results on loss estimation to the spherical case, it offers a great simplification. The study of loss estimation (for the estimation of a mean) without using the canonical form has been done in [Fourdrinier & Strawderman 2008] and shows the computational burden that it carries. This is the reason why, in the sequel, we will focus on the canonical form, especially in the proofs of the theorems. However, it has two limitations: the first one is that it relies on the assumption that the target function is linear, that is, $f(X) = X\beta$; the second limitations comes from the fact that it requires the number p of variables to be strictly lower than the number n of observations.

As mentioned earlier, the canonical form can be obtain by applying an orthogonal transformation to the linear model. We recall that, in this section, we have

$$Y = X\beta + \sigma\varepsilon, \quad \text{with} \quad \varepsilon \sim \mathcal{S}_n(0), \quad (3.30)$$

where \mathcal{S}_n denotes any spherically symmetric distribution. From this model, we construct the orthogonal matrix

$$G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}$$

such that the p lines of G_1 form an orthonormal base of the column space of X , namely $\mathcal{C}(X)$, while the $n - p$ lines of G_2 form an orthonormal base of the space orthogonal to $\mathcal{C}(X)$. In other words, we have

$$G_2 X = 0, \quad (3.31)$$

and there exists a $p \times p$ non singular matrix A such that

$$X = G_1^t A. \quad (3.32)$$

Note that, in practice, the matrix G can easily be obtained from the QR -factorization of the design matrix X by taking $G = Q^t$. We recall that, in this factorization, R is composed of two

submatrices R_1 and R_2 where R_1 is upper triangular and R_2 is the null matrix. The factorization can be rewritten as

$$X = Q_1 R_1, \quad (3.33)$$

where Q_1 is the submatrix containing the p first columns of Q . Note that Q_1 is also constructed as an orthonormal base of the column space $\mathcal{C}(X)$, hence the analogy with G_1 follows immediately. However, the main difference arises from the fact that the factorization in (3.33) is unique as soon as X is full rank and the diagonal components of R_1 are constrained to be positive [Golub & Van Loan 1996]. On the other side, the canonical form does not require such a restriction on the matrix $A = G_1 X$ (the analog of R_1), which could also be lower triangular or full. Hence, the decomposition $X = G_1^t A$ is not unique. The QR-factorization only offers one convenient way to efficiently compute G and A .

Applying the orthogonal matrix Q^t to model (3.30) brings

$$W = T + \sigma G\varepsilon, \quad (3.34)$$

where $W = \begin{pmatrix} Z \\ U \end{pmatrix}$ and $T = \begin{pmatrix} \theta \\ 0 \end{pmatrix}$ with $Z = G_1 Y$, $U = G_2 Y$, and $\theta = G_1 X \beta$.

The invariance property of spherically symmetric distributions implies that, if Y is spherically symmetric, then $(Z^t, U^t)^t$ is also spherically symmetric:

$$Y \sim \mathcal{S}_n(\mu) \quad \Rightarrow \quad \begin{pmatrix} Z \\ U \end{pmatrix} \sim \mathcal{S}_n \begin{pmatrix} \theta \\ 0 \end{pmatrix}.$$

From the properties of spherically symmetric distributions, Z and U are independent if and only if \mathcal{S}_n is the Gaussian distribution.

By construction of the canonical form, Z accounts for the information contained in both Y and X at the same time, while U accounts for the information contained in Y only. This is why U is often referred to as the *residual vector*. An interesting aspect of the canonical form is that it presents the model in a purer and simpler form, where the parameter to be estimated is the mean of the quantity of interest itself. This way, we can benefit from some of the results of loss estimation in the context of estimation of the mean.

We recall that in this work we are concerned with good prediction, and that we formalized this objective as

$$L(X\beta, X\hat{\beta}) = \|X\hat{\beta} - X\beta\|^2.$$

We now express this criterion in the canonical form. Replacing X by its decomposition yields

$$\begin{aligned} \|X\hat{\beta} - X\beta\|^2 &= (\hat{\beta} - \beta)^t X^t X (\hat{\beta} - \beta) \\ &= (\hat{\beta} - \beta)^t A^t G_1 G_1^t A (\hat{\beta} - \beta) \\ &= (\hat{\beta} - \beta)^t A^t A (\hat{\beta} - \beta) \\ &= \|\hat{\theta} - \theta\|^2, \end{aligned} \quad (3.35)$$

since $\theta = G_1 X \beta = A \beta$ and $\hat{\theta} = A \hat{\beta}$. From (3.35), we deduce that minimizing $L(X\beta, X\hat{\beta})$ with respect to an estimator $\hat{\beta}$ is equivalent to minimizing $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2$ with respect to its canonical estimator $\hat{\theta}$. In addition, we also have inequality between their risks:

$$R_\beta(\hat{\beta}) = \mathbb{E}_\beta[\|X\hat{\beta} - X\beta\|^2] = \mathbb{E}_\theta[\|\hat{\theta} - \theta\|^2] = R_\theta(\hat{\theta}),$$

where \mathbb{E}_θ denotes the expectation under (Z, U) , and it is thus equivalent to minimize the risk of $\hat{\beta}$ and to minimize the risk of $\hat{\theta}$, if we are more interested in minimizing the risk than the actual loss.

Note that the canonical form of the least-squares estimator $\hat{\beta}^{LS}$ is simply $\hat{\theta}^{LS} = Z$. Indeed, replacing X by its decomposition $G_1^t A$ yields

$$\begin{aligned}\hat{\theta}^{LS} &= G_1 X \hat{\beta}^{LS} \\ &= G_1 X (X^t X)^{-1} X^t Y \\ &= G_1 G_1^t A (A^t G_1 G_1^t A)^{-1} A^t G_1 Y.\end{aligned}$$

Besides, by construction we have that $G_1 G_1^t = I_p$ and $G_1 Y = Z$. Thereby we obtain

$$\hat{\theta}^{LS} = A(A^t A)^{-1} A^t Z.$$

Now, the matrix A being squared and non singular, decomposing the inverse of the product $A^t A$ results in

$$\hat{\theta}^{LS} = A A^{-1} (A^t)^{-1} A^t Z = Z, \quad (3.36)$$

which is what we claimed.

In order to close this paragraph, note that if the estimator of β can be written in the form

$$\hat{\beta} = \hat{\beta}^{LS} + g(\hat{\beta}^{LS}),$$

then, its canonical form can also be expressed as

$$\hat{\theta} = Z + g'(Z),$$

où $g'(Z) = A g(A^{-1} Z)$. Indeed, we have that

$$\hat{\theta} = G_1 X \hat{\beta} = G_1 X \hat{\beta}^{LS} + G_1 X g((G_1 X)^{-1} Z) = Z + G_1 X g((G_1 X)^{-1} Z),$$

from the invertibility of $G_1 X = A$ and since $\hat{\theta}^{LS} = Z$.

We have now all the elements necessary to extend the loss estimators to the spherical case.

3.4.2 Unbiased estimator of the estimation loss

This section is devoted to the generalization of Theorem 3.2 to the class of spherically symmetric distributions, given by Theorem 3.6.

In all that follows, we consider the case where the estimator $\hat{\beta}$ does not depend on the estimator $\hat{\sigma}^2$ of the noise level. Considering $\hat{\beta} = \hat{\beta}(\hat{\sigma}^2)$ would lead to different loss estimators.

Theorem 3.6 (Unbiased estimator of the quadratic loss under spherical assumption). *Let $Y \sim \mathcal{S}_n(X\beta, \sigma^2)$ and $\hat{\sigma}^2 = \|Y - X\hat{\beta}^{LS}\|^2/(n-p)$ be an estimator of the variance $\sigma^2 \mathbb{E}_\beta[\|\varepsilon\|^2/n]$. Given $\hat{\beta} = \hat{\beta}(Y)$ an estimator of β , if $X\hat{\beta}$ is weakly differentiable with respect to Y , then the following estimator of $\|X\hat{\beta} - X\beta\|^2$ is unbiased:*

$$\hat{L}_0(\hat{\beta}) = \|Y - X\hat{\beta}\|^2 + (2 \operatorname{div}_Y(X\hat{\beta}) - n)\hat{\sigma}^2, \quad (3.37)$$

where $\operatorname{div}_Y x = \sum_{i=1}^n \partial x_i / \partial Y_i$ is the weak divergence of x with respect to Y .

Before proving Theorem 3.6, we need an extension of Stein's identity to spherically symmetric distributions. Such an extension has been obtained by [Fourdrinier & Wells 1995a] under the canonical form of the linear model, reported here for the sake of completeness. However, we do not report its proof. We refer the interested reader to [Fourdrinier & Wells 1995a].

Theorem 3.7 (Stein-type identity). *Given $(Z, U) \in \mathbb{R}^n$ a random vector following a spherically symmetric distribution around $(\theta, 0)$, and $g : \mathbb{R}^p \mapsto \mathbb{R}^p$ a weakly differentiable function, we have*

$$\mathbb{E}_\theta[(Z - \theta)^t g(Z)] = \mathbb{E}_\theta[\|U\|^2 \operatorname{div}_Z g(Z)/(n - p)], \quad (3.38)$$

provided both expectations exist.

Note that the divergence in theorem 3.6 is taken with respect to Y while in the Stein type identity 3.38 requires the divergence with respect to Z . Their relationship can be seen in the following lemma.

Lemma 3.1. *Under the hypothesis of Theorems 3.6 and 3.7 and using the canonical transformation of the linear model in (3.34), we have*

$$\operatorname{div}_Y X\hat{\beta} = \operatorname{div}_Z \hat{\theta}. \quad (3.39)$$

Proof. Denoting by $\operatorname{tr}(A)$ the trace of any matrix A and by $J_f(t)$ the Jacobian matrix of any function f at t , we have

$$\operatorname{div}_Y X\hat{\beta} = \operatorname{tr}(J_{X\hat{\beta}}(Y)) = \operatorname{tr}(Q^t J_{X\hat{\beta}}(Y) Q)$$

by definition of the divergence and since Q is an orthogonal matrix. Now, applying the chain rule to the function $\hat{T}(W) = Q^t X\hat{\beta}$ with $Y = QW$, we have

$$J_{\hat{T}}(W) = J_{Q^t X\hat{\beta}}(Y) Q = Q^t J_{X\hat{\beta}}(Y) Q, \quad (3.40)$$

noticing that Q is a linear transformation. Also, as

$$\hat{T} = \begin{pmatrix} \hat{\theta} \\ 0 \end{pmatrix},$$

we have the following decomposition

$$J_{\hat{T}}(W) = \begin{pmatrix} J_{\hat{\theta}}(Z) & 0 \\ J_{\hat{\theta}}(U) & 0 \end{pmatrix},$$

and thus

$$\operatorname{tr}(J_{\hat{T}}(W)) = \operatorname{tr}(J_{\hat{\theta}}(Z)). \quad (3.41)$$

Therefore, according to (3.40) and (3.41), we obtain

$$\operatorname{tr}(J_{\hat{\theta}}(Z)) = \operatorname{tr}(J_{X\hat{\beta}}(Y)),$$

which is (3.39). □

We now have all the elements to prove Theorem 3.6.

Proof of Theorem 3.6. As we did in the Gaussian case, we wish to find an unbiased estimator of the quadratic loss which would not depend explicitly on β . The quadratic loss function of $X\hat{\beta}$ at $X\beta$ can be decomposed

$$\|X\hat{\beta} - X\beta\|^2 = \|Y - X\hat{\beta}\|^2 - \|Y - X\beta\|^2 + 2(Y - X\beta)^t(X\hat{\beta} - X\beta).$$

Using the canonical formulation it becomes

$$\|X\hat{\beta} - X\beta\|^2 = \|Y - X\hat{\beta}\|^2 - \|Y - X\beta\|^2 + 2(Z - \theta)^t(\hat{\theta} - \theta),$$

such that its expectation, the risk is

$$R(X\hat{\beta}, X\beta) = \mathbb{E}_\beta \|Y - X\hat{\beta}\|^2 - \sigma^2 \mathbb{E}_\beta \|\varepsilon\|^2 + 2 \mathbb{E}_\beta (Z - \theta)^t \hat{\theta}.$$

Applying Theorem 3.7 with $g(Z) = \hat{\theta}$ leads to

$$R(X\hat{\beta}, X\beta) = \mathbb{E}_\beta \|Y - X\hat{\beta}\|^2 - \sigma^2 \mathbb{E}_\beta \|\varepsilon\|^2 + 2 \mathbb{E}_\beta [\|U\|^2 \operatorname{div}_Z \hat{\theta} / (n - p)].$$

Applying Lemma 3.1 we get

$$R(X\hat{\beta}, X\beta) = \mathbb{E}_\beta \|Y - X\hat{\beta}\|^2 - \sigma^2 \mathbb{E}_\beta \|\varepsilon\|^2 + 2 \mathbb{E}_\beta \left[\frac{\|U\|^2}{n - p} \operatorname{div}_Y X\hat{\beta} \right].$$

The proof is completed using the fact that $\mathbb{E}_\beta[\hat{\sigma}^2] = \sigma^2 \mathbb{E}_\beta[\|\varepsilon\|^2/n]$ for the middle term and the equality $\|U\|^2 = \|Y - X\hat{\beta}^{LS}\|^2$. \square

The unbiased estimator of the quadratic estimation loss derived under the spherical assumption is actually equal to the one derived under the Gaussian law. Hence, this implies that we do not have to specify the form of the distribution, the only condition being its spherical symmetry.

We have not proved that it was a good estimator, but if unbiasedness is a property we are interested in, then \hat{L}_0 is robust and its performance are similar for the whole family.

Remark 3.3. Note that the extension of Stein's lemma in Theorem 3.7 relaxes the assumption that $\hat{\sigma}^2$ should be uncorrelated with $\hat{df} = \operatorname{div}_Y X\hat{\beta}$, as the equality takes the whole product into account.

The results from this section actually correspond to the implicate assumption that the noise level σ is known, which explains its relation to the unbiased estimator of loss derived under the Gaussian distribution with known variance σ^2 . There have been attempts to generalize the unbiased estimator under spherical symmetry to the case where the noise level σ is unknown [Fourdrinier & Strawderman 2010]. This setting gives rise to the following extension of Stein identity (given in the canonical form)

$$\mathbb{E}_{\theta, \sigma^2}[(Z - \theta)^t g(Z, U)] = c \mathbb{E}_{\theta, \sigma^2}^*[\operatorname{div}_Z g(Z, U)],$$

where $\mathbb{E}_{\theta, \sigma^2}^*$ is the expectation with respect to the distribution

$$\frac{1}{c \sigma^{2n}} P \left(\frac{\|z - \theta\|^2 + \|u\|^2}{\sigma^2} \right) \quad \text{with} \quad P(t) = \frac{1}{2} \int_t^\infty p(u) du, \quad (3.42)$$

and p is the density of (Z, U) . Hence, estimators of loss can be found in a similar form than in Equation (3.27), but they are unbiased for the distribution (3.42) and it is not clear what form they would have for the distribution of (Z, U) .

3.5 Summary

In this chapter, we derived unbiased estimators of the quadratic loss for the linear model with Gaussian noise, where we assumed the variance of the noise to be successively known and unknown. We related them to existing methods from the literature, namely C_p , AIC, FPE and AIC_c , which can thus be viewed through a loss estimation approach. We then derived the unbiased estimator of loss under a wider distributional setting: the family of spherically symmetric distributions. The unbiased estimator of the quadratic estimation loss derived under the spherical assumption is actually equal to the one derived under the Gaussian law with known variance. Hence, this implies that we do not have to specify the form of the distribution, the only condition being its spherical symmetry. From the equivalence between unbiased estimators of loss, C_p and AIC, we conclude that their form for the Gaussian case can be used to handle any spherically symmetric distribution. The spherical family is interesting for many practical cases since it allows a dependence property between the components of random vectors whenever the distribution is not Gaussian. Some members of this family also have heavier tails than the Gaussian law, and thus the unbiased estimator derived here can be robust to outliers.

It is well known that unbiased estimators of loss are not the best estimators and can be improved (see for instance [Johnstone 1988]). This remark is also shared by AIC and C_p . It was not our intention in this chapter to show better results of such estimators, but our result explains why their performances can be similar when departing from the Gaussian assumption. The study of unbiased estimators is however of great importance in loss estimation theory. It indeed allows to compare theoretically the risks of two estimators of loss for estimating the true loss, as we will see in the following chapter. The heuristic of loss estimation is that the closer an estimator is to the true loss, the more we expect their respective minima to be close.

Corrected loss estimators for model selection

Contents

4.1	Improving on unbiased estimators of loss	85
4.1.1	A new layer of evaluation	85
4.1.2	Conditions of improvement over the unbiased estimator	87
4.1.3	Choice of the correction function	91
4.2	Corrected loss estimators for the restricted model	92
4.2.1	Condition for improvement with γ_r	92
4.2.2	Application to estimators of the regression coefficient	93
4.3	Corrected loss estimators for the full model	96
4.3.1	Condition for improvement with γ_f	96
4.4	Link with principled methods	99
4.5	Summary	100

This chapter is devoted to the derivation of estimators of loss that are more accurate than unbiased loss estimators. We first present how we can evaluate and compare estimators of loss. Then, we propose two bias terms based on data, called *correction functions*, one for the restricted linear model and the second for the full linear model. Sections 4.2 and 4.3 specify the conditions of improvement over the unbiased loss estimator for each correction function, and we look at optimal corrections yielding the greatest improvement. Note that the study is done under the spherical assumption, which includes the Gaussian case as we have seen earlier. Finally, Section 4.4 relates the philosophy behind loss estimation to that of Structural Risk Minimization [Vapnik 1998] and Slope Heuristics [Birgé & Massart 2007].

4.1 Improving on unbiased estimators of loss

4.1.1 A new layer of evaluation

As suggested by [Sandved 1968], unbiased loss estimators are not the best estimators for loss estimation, a fact that is well known in risk estimation. Several authors have thus considered the problem of improvement over the unbiased estimator and proposed a way of evaluating estimators of loss. In the context of estimating the mean vector θ of a random vector Z , [Lu &

Berger 1989] proposed to evaluate the quality of a loss estimator \hat{L} through the “communication loss”

$$\mathcal{L}(\theta, \hat{\theta}, \hat{L}) = (\hat{L}(\hat{\theta}) - L(\theta, \hat{\theta}))^2, \quad (4.1)$$

and its risk

$$\mathcal{R}_\theta(\hat{\theta}, \hat{L}) = \mathbb{E}_\theta[\mathcal{L}(\theta, \hat{\theta}, \hat{L})] = \mathbb{E}_\theta[(\hat{L}(\hat{\theta}) - L(\theta, \hat{\theta}))^2]. \quad (4.2)$$

The risk in Equation (4.2) is actually a generalization of Stein’s proposition to evaluate confidence sets, and corresponds to the Mean Squared Error of the estimator \hat{L} of loss $L(\theta, \hat{\theta})$. In other words, we look for biased estimators of $L(\theta, \hat{\theta})$ whose variance is sufficiently low so as to allow a better control of the estimation. This definition of a new loss and its risk leads to the following definition of improvement (or domination) over the unbiased estimator.

Definition 4.1 (Improvement for loss estimation). *Let \hat{L}_0 and \hat{L} be two estimators of the loss $L(\theta, \hat{\theta})$. The estimator \hat{L} is said to be better than, or to dominate, \hat{L}_0 if the condition*

$$\mathcal{R}_\theta(\hat{\theta}, \hat{L}) \leq \mathcal{R}_\theta(\hat{\theta}, \hat{L}_0) \quad (4.3)$$

is verified for all $\theta \in \Theta$, and if the condition

$$\mathcal{R}_\theta(\hat{\theta}, \hat{L}) < \mathcal{R}_\theta(\hat{\theta}, \hat{L}_0)$$

is verified at least for one value of θ .

In addition to proving their existence, [Johnstone 1988] propose new estimators of loss of the form

$$\hat{L}_\gamma(\hat{\theta}) = \hat{L}_0(\hat{\theta}) - \gamma(Z), \quad \text{with } \gamma(Z) = \frac{c}{\|Z\|^2}, \quad (4.4)$$

where c is a constant. The author thus shows that the condition of improvement of $\hat{L}_\gamma(\hat{\theta})$ over $\hat{L}_0(\hat{\theta})$ under the Gaussian assumption, that is the condition for which the inequality (4.1) is valid when $\hat{\theta}$ is the least-squares estimator of θ , is obtained for

$$0 < c < 4(d - 4),$$

where d is the dimension of Z .

This approach is the one we choose for developing better criteria of model selection.

The use of a quadratic risk for $\mathcal{R}_\theta(\hat{\theta}, \hat{L})$ allows easy computations. Note however that there are a few interesting alternatives for \mathcal{R} . The first one would be to consider Stein’s loss in Equation (3.9) applied to the estimation of the loss $L(\theta, \hat{\theta})$ instead of the estimation of the covariance matrix:

$$\mathcal{L}(\theta, \hat{\theta}, \hat{L}) = \frac{L(\theta, \hat{\theta})}{\hat{L}(\hat{\theta})} - \log \left(\frac{L(\theta, \hat{\theta})}{\hat{L}(\hat{\theta})} \right) - d. \quad (4.5)$$

The main advantage of this loss is that it penalizes less the large values of \hat{L} with respect to $L(\theta, \hat{\theta})$ than the quadratic loss, and it is thus a remedy for the leapfrogging effect of the quadratic loss. Another interesting alternative is the *utility function* proposed by [Rukhin 1988a, Rukhin 1988b]

$$\mathcal{L}(\theta, \hat{\theta}, \hat{L}) = \frac{L(\theta, \hat{\theta})}{\sqrt{\hat{L}(\hat{\theta})}} + \sqrt{\hat{L}(\hat{\theta})}.$$

Both solutions generate much more tedious calculation than the quadratic loss. Therefore we will only consider the latter one in the rest of this manuscript.

In the context of estimating the mean of a multivariate Gaussian vector, [Johnstone 1988] proved the inadmissibility of the unbiased estimator of loss of the Least-squares estimator and of the James-Stein estimator, and proposed corrected estimators that improve on it. This work has been extended in [Fourdrinier & Wells 1995a] to the general linear regression model under the spherical assumption. Around the same time, the latter authors also considered the problem of model selection and proposed corrected estimators of the loss when the regression coefficient β is estimated by restricted least-squares (LS) $\hat{\beta}_I^{LS}$ (see [Fourdrinier & Wells 1994]). In this chapter, we extend the latter work on model selection to other estimators of β . Table 4.1 summarizes these historical remarks.

Objective	Reference	Estimator	Distribution
Estimation of a multivariate mean	[Johnstone 1988]	LS, JS	Gaussian
	[Fourdrinier & Wells 1995a]	LS, JS	Spherical
Model Selection	[Fourdrinier & Wells 1994]	Restricted LS	Gaussian
	In this chapter	Restricted LS, restricted JS, Lasso-type ^a .	Spherical

^aWe refer to as Lasso-type methods all the sparse regularization methods described in Chapter 2.

Table 4.1: Works on corrected loss estimators.

The strategy we adopt in the sequel is based on the following steps. First, following what was done in [Johnstone 1988] and [Fourdrinier & Wells 1995b], we propose to work with *corrected estimators of loss* of the form

$$\hat{L}_\gamma(\hat{\beta}) = \hat{L}_0(\hat{\beta}) - \gamma(X\hat{\beta}^{LS}), \quad (4.6)$$

where $\hat{L}_0(\hat{\beta})$ is the unbiased loss estimator derived in the previous chapter and used here as a reference, and the *correction function* $\gamma(t)$ is a general twice weakly differentiable function. Note that the choice of taking a correction based on the least-squares estimator merely comes from the simplicity of calculation, while taking $\gamma(X\hat{\beta})$ for a general estimator $\hat{\beta}$ results in much more tedious algebra. Second, in the following paragraph, we derive conditions on the correction function γ for \hat{L}_γ to improve on \hat{L}_0 . Then, we propose two forms for the correction function γ , depending on whether we consider the full or the restricted model. And finally, we sharpen the conditions of improvement for each correction function in Sections 4.2 and 4.3.

4.1.2 Conditions of improvement over the unbiased estimator

Going back to our context of linear regression model

$$Y = X\beta + \sigma\varepsilon,$$

and its version restricted to a subset $I \subset \{1, \dots, p\}$

$$Y = X_I\beta_I + \sigma\varepsilon,$$

the communication risk becomes

$$\mathcal{R}(\hat{\beta}, \hat{L}) = \mathbb{E}_{\beta}[(\hat{L}(\hat{\beta}) - \|X\hat{\beta} - X\beta\|^2)^2]. \quad (4.7)$$

The definition of improvement given in Definition 4.1 leads to identifying the conditions for which the following inequality holds:

$$\Delta_{\beta}(\hat{L}_{\gamma}, \hat{L}_0) = \mathcal{R}_{\beta}(\hat{\beta}, \hat{L}_{\gamma}) - \mathcal{R}_{\beta}(\hat{\beta}, \hat{L}_0) \leq 0, \quad (4.8)$$

where $\mathcal{R}_{\beta}(\hat{\beta}, \hat{L})$ is the quadratic risk defined in Equation (4.7) for the prediction loss $L(\beta, \hat{\beta}) = \|X\hat{\beta} - X\beta\|^2$. The following lemma gives a general result on the difference $\Delta_{\theta}(\hat{L}_{\gamma}, \hat{L}_0)$ in risks between \hat{L}_{γ} and \hat{L}_0 for any twice weakly differentiable function γ . This result relies on the assumption that the noise level σ is known. Also, in all that follows, we still consider the case where the estimator $\hat{\beta}$ does not depend on the estimator $\hat{\sigma}^2$ of the noise level.

Lemma 4.1. *Let $Y \sim \mathcal{S}_n(X\beta)$. Given $\hat{\beta}$ an estimator of β and $\gamma(\cdot)$ a correction function, if $\hat{\beta}$ is weakly differentiable with respect to Y and γ is twice weakly differentiable, then the condition*

$$\gamma^2(\hat{\beta}^{LS}) + 2 \frac{\|Y - X\hat{\beta}^{LS}\|^2}{n-p} \left\{ \frac{\|Y - X\hat{\beta}^{LS}\|^2}{n-p+2} \Delta_Y \gamma(X\hat{\beta}^{LS}) + 2(X\hat{\beta} - Y)^t \nabla_Y \gamma(X\hat{\beta}^{LS}) \right\} \leq 0 \quad (4.9)$$

is sufficient for $\hat{L}_{\gamma}(\hat{\beta}) = \hat{L}_0(\hat{\beta}) - \gamma(X\hat{\beta}^{LS})$ to dominate \hat{L}_0 .

Before proving Lemma 4.1, we need the following extension of Theorem 3.7 and the corollary of the Stein-type identity for spherically symmetric distributions, derived in a similar fashion as what is done in [Stein 1981].

Theorem 4.1 (Extended Stein-type identity). *Let $(Z, U) \in \mathbb{R}^n$ be a random vector following a spherically symmetric distribution around $(\theta, 0)$ as in Model (3.34), $g : \mathbb{R}^p \mapsto \mathbb{R}^p$, and q be an integer. If g is weakly differentiable, then*

$$\mathbb{E}_{\theta}[\|U\|^q (Z - \theta)^t g(Z)] = \frac{1}{n-p+q} \mathbb{E}_{\theta}[\|U\|^{q+2} \text{div}_Z g(Z)], \quad (4.10)$$

provided both expectations exist.

Proof of Theorem 4.1. The proof can be found in the Appendix A.2 of [Fourdrinier & Wells 2012]. It basically consists in the divergence theorem where the expectations are first conditioned on the radius $R = \|Z - \theta\|^2 + \|U\|^2$ and rely on the property that projections on a lower subspace such as

$$\pi : \begin{pmatrix} Z \\ U \end{pmatrix} \mapsto Z$$

are spherically symmetric. The last step of the proof is obtained by taking the expectations with respect to the distribution of the radius. \square

Corollary 4.1. *Let $(Z, U) \in \mathbb{R}^n$ be a random vector following a spherically symmetric distribution around $(\theta, 0)$ and $h : \mathbb{R}^p \mapsto \mathbb{R}$. If $h(\cdot)$ is twice weakly differentiable, then*

$$\begin{aligned} \mathbb{E}_{\theta}[\|Z - \theta\|^2 h(Z)] &= \frac{p}{n-p} \mathbb{E}_{\theta}[\|U\|^2 h(Z)] \\ &+ \frac{1}{(n-p)(n-p+2)} \mathbb{E}_{\theta}[\|U\|^4 \Delta_Z h(Z)], \end{aligned} \quad (4.11)$$

provided the expectations exist, where $\Delta_Z h(Z)$ denotes the weak Laplacian of $h(Z)$ with respect to Z .

Proof of Corollary 4.1. Corollary 4.1 is obtained by applying Theorem 3.7 with $g(Z) = (Z - \theta)h(Y)$ and its extension Theorem 4.1 with $g(Z) = \nabla_Z h(Z)$. Indeed, we have that

$$\begin{aligned}\mathbb{E}_\theta[\|Z - \theta\|^2 h(Z)] &= \mathbb{E}_\theta[(Z - \theta)^t (Z - \theta) h(Z)] \\ &= \frac{1}{n - p} \mathbb{E}_\theta \left[\|U\|^2 \left(h(Z) \operatorname{div}_Z (Z - \theta) + (Z - \theta)^t \nabla_Z h(Z) \right) \right] \\ &= \frac{1}{n - p} \mathbb{E}_\theta \left[p \|U\|^2 h(Z) + \frac{\|U\|^4}{n - p + 2} \operatorname{div}_Z \{ \nabla_Z h(Y) \} \right],\end{aligned}$$

where the second equality derives from the product rule between a scalar and a vector functions

$$\operatorname{div}_Z \{ (Z - \theta)h(Z) \} = h(Z) \operatorname{div}_Z (Z - \theta) + (Z - \theta)^t \nabla_Z h(Z).$$

The desired result is then obtained by definition of the Laplacian operator:

$$\Delta_Z h(Z) = \operatorname{div}_Z \{ \nabla_Z h(Z) \}.$$

□

We are now able to prove Lemma 4.1.

Proof of Lemma 4.1. The proof follows four steps: first, a development of the difference in risk (4.8), second the transformation of the terms depending on the true parameter β into their canonical form, then the application on these terms of the Stein-type theorem and its corollary for the spherical case, and finally the inverse transformation from the canonical form back to the usual linear form in Y and X .

The difference in risk can be easily developed by replacing \hat{L}_γ by its expression in (4.6), resulting in the following expression

$$\begin{aligned}\Delta_\beta(\hat{L}_\gamma, \hat{L}_0) &= \mathbb{E}_\beta[(\hat{L}_\gamma - \|X\hat{\beta} - X\beta\|^2)^2] - \mathbb{E}_\beta[(\hat{L}_0 - \|X\hat{\beta} - X\beta\|^2)^2] \\ &= \mathbb{E}_\beta[\gamma^2(\hat{\beta}^{LS}) - 2\gamma(X\hat{\beta}^{LS})\{\hat{L}_0 - \|X\hat{\beta} - X\beta\|^2\}].\end{aligned}$$

Now, replacing \hat{L}_0 by its expression given in (3.37) from the previous chapter and developing the loss $\|X\hat{\beta} - X\beta\|^2$ yield

$$\begin{aligned}\Delta_\beta(\hat{L}_\gamma, \hat{L}_0) &= \mathbb{E}_\beta[\gamma^2(\hat{\beta}^{LS}) - 2\gamma(X\hat{\beta}^{LS})\{\|Y - X\hat{\beta}\|^2 + (2 \operatorname{div}(X\hat{\beta}) - n) \hat{\sigma}^2\} \\ &\quad + 2\gamma(X\hat{\beta}^{LS})\{\|X\hat{\beta} - Y\|^2 + \|Y - X\beta\|^2 + 2(Y - X\beta)^t(X\hat{\beta} - Y)\}] \\ &= \mathbb{E}_\beta[\gamma^2(\hat{\beta}^{LS}) - 2\gamma(X\hat{\beta}^{LS})(2 \operatorname{div}(X\hat{\beta}) - n) \hat{\sigma}^2 \\ &\quad + 2\gamma(X\hat{\beta}^{LS})\{\|Y - X\beta\|^2 + 2(Y - X\beta)^t(X\hat{\beta} - Y)\}]\end{aligned}$$

Applying the canonical form on the right-most terms of the equation and letting

$$\zeta(Z) = \gamma(G_1^t Z) = \gamma(X\hat{\beta}^{LS}), \quad (4.12)$$

we obtain that

$$\mathbb{E}_\beta[\gamma(X\hat{\beta}^{LS})\|Y - X\beta\|^2] = \mathbb{E}_\theta[\zeta(Z)\{\|Z - \theta\|^2 + \|U\|^2\}] \quad (4.13)$$

$$\begin{aligned}\mathbb{E}_\beta[\gamma(X\hat{\beta}^{LS})(Y - X\beta)^t(X\hat{\beta} - Y)] &= \mathbb{E}_\theta \left[\zeta(Z) \begin{pmatrix} Z - \theta \\ U \end{pmatrix}^t \begin{pmatrix} \hat{\theta} - Z \\ -U \end{pmatrix} \right] \\ &= \mathbb{E}_\theta [\zeta(Z) \{ (Z - \theta)^t (\hat{\theta} - Z) - \|U\|^2 \}]. \quad (4.14)\end{aligned}$$

Applying Theorem 4.1 with $g(Z) = \zeta(Z)(\hat{\theta} - Z)$ and Corollary 4.1 with $h(Z) = \zeta(Z)$ gives

$$\mathbb{E}_\theta[\zeta(Z)\|Z - \theta\|^2] = \frac{1}{n-p} \mathbb{E}_\theta \left[p\|U\|^2 \zeta(Z) + \frac{1}{n-p+2} \|U\|^4 \Delta_Z \zeta(Z) \right] \quad (4.15)$$

and

$$\begin{aligned} \mathbb{E}_\theta[\zeta(Z)(Z - \theta)^t(\hat{\theta} - Z)] &= \frac{1}{n-p} \mathbb{E}_\theta[\|U\|^2 \operatorname{div}_Z \{\zeta(Z)(\hat{\theta} - Z)\}] \\ &= \frac{1}{n-p} \mathbb{E}_\theta[\|U\|^2 \{\zeta(Z)(\operatorname{div}_Z \hat{\theta} - p) + (\hat{\theta} - Z)^t \nabla_Z \zeta(Z)\}]. \end{aligned} \quad (4.16)$$

The tedious part is now to give the expressions of $\nabla_Z \zeta(Z)$ and $\Delta_Z \zeta(Z)$ in the linear form as functions of Y or $\hat{\beta}^{LS}$. However, both of them can be derived in a similar fashion as we did in the proof of Lemma 3.1 on the equality between the weak divergences. Indeed, noticing that we have in fact

$$\nabla_Z \zeta(Z) = \nabla_W \zeta(\hat{T}^{LS})$$

with the notations in (3.34) and $\hat{T}^{LS} = \begin{pmatrix} Z \\ 0 \end{pmatrix}$ being the canonical form of the least-squares estimator, the chain rule gives

$$\nabla_W \zeta(\hat{T}^{LS}) = \nabla_{GY} \gamma(X\hat{\beta}^{LS}) = G \nabla_Y \gamma(X\hat{\beta}^{LS}). \quad (4.17)$$

In the same way, we can notice that

$$\Delta_Z \zeta(Z) = \Delta_W \zeta(\hat{T}^{LS}).$$

Applying again the chain rule yields

$$\Delta_W \zeta(\hat{T}^{LS}) = \Delta_{GY} \gamma(X\hat{\beta}^{LS}) = \operatorname{tr} \left(G H_{X\hat{\beta}^{LS}}(Y) G^t \right),$$

where $H_{X\hat{\beta}^{LS}}(Y)$ denotes the Hessian matrix of the function γ at Y . Note that the last equality derives from the fact that the Laplacian operator is defined as the trace of the Hessian matrix. Hence, by orthogonality of G , we obtain the equality between both weak Laplacians

$$\Delta_Z \zeta(Z) = \Delta_Y \gamma(\hat{\beta}^{LS}). \quad (4.18)$$

Combining the elements from Equations (4.15) and (4.18) into (4.13) and going back to the linear model yields

$$\begin{aligned} \mathbb{E}_\beta[\gamma(X\hat{\beta}^{LS})\|Y - X\beta\|^2] &= \mathbb{E}_\beta \left[\frac{S}{n-p} \left\{ p\gamma(X\hat{\beta}^{LS}) + \frac{S}{n-p+2} \Delta_Y \gamma(X\hat{\beta}^{LS}) \right\} + S\gamma(X\hat{\beta}^{LS}) \right] \\ &= \mathbb{E}_\beta \left[\frac{S}{n-p} \left\{ n\gamma(X\hat{\beta}^{LS}) + \frac{S}{n-p+2} \Delta_Y \gamma(X\hat{\beta}^{LS}) \right\} \right]. \end{aligned}$$

Similarly, combining Equations (4.16) and (4.17) into (4.14) gives

$$\begin{aligned} \mathbb{E}_\beta[\gamma(X\hat{\beta}^{LS})(Y - X\beta)^t(X\hat{\beta} - Y)] &= \mathbb{E}_\beta \left[S\gamma(X\hat{\beta}^{LS}) \left\{ \frac{1}{n-p} (\operatorname{div}_Y X\hat{\beta} - p) - 1 \right\} \right] \\ &\quad + \mathbb{E}_\beta \left[\frac{S}{n-p} \gamma(X\hat{\beta}^{LS})(X\hat{\beta} - Y)^t G^t G \nabla_Y \gamma(X\hat{\beta}^{LS}) \right] \\ &= \mathbb{E}_\beta \left[\frac{S}{n-p} \gamma(X\hat{\beta}^{LS})(\operatorname{div}_Y X\hat{\beta} - n) \right] \\ &\quad + \mathbb{E}_\beta \left[\frac{S}{n-p} \gamma(X\hat{\beta}^{LS})(X\hat{\beta} - Y)^t \nabla_Y \gamma(X\hat{\beta}^{LS}) \right]. \end{aligned}$$

Substituting these last equalities into the difference of risks, we obtain that

$$\begin{aligned}
\Delta_\beta(\widehat{L}_\gamma, \widehat{L}_0) &= \mathbb{E}_\beta \left[\gamma^2(\widehat{\beta}^{LS}) - \frac{2S}{n-p} \gamma(X\widehat{\beta}^{LS})(2 \operatorname{div}(X\widehat{\beta}) - n) \right] \\
&+ \mathbb{E}_\beta \left[\frac{2}{n-p} \left\{ nS \gamma(X\widehat{\beta}^{LS}) + \frac{1}{n-p+2} S^2 \Delta_Y \gamma(X\widehat{\beta}^{LS}) \right\} \right] \\
&+ \mathbb{E}_\beta \left[\frac{4S}{n-p} \left\{ (\operatorname{div}_Y X\widehat{\beta} - n) \gamma(X\widehat{\beta}^{LS}) + (X\widehat{\beta} - Y)^t \nabla_Y \gamma(X\widehat{\beta}^{LS}) \right\} \right] \\
&= \mathbb{E}_\beta \left[\gamma^2(\widehat{\beta}^{LS}) + \frac{2S}{n-p} \gamma(X\widehat{\beta}^{LS}) \{ -(2 \operatorname{div}(X\widehat{\beta}) - n) + n + 2(\operatorname{div}_Y X\widehat{\beta} - n) \} \right] \\
&+ \mathbb{E}_\beta \left[\frac{2S^2}{(n-p)(n-p+2)} \Delta_Y \gamma(X\widehat{\beta}^{LS}) + \frac{4S}{n-p} (X\widehat{\beta} - Y)^t \nabla_Y \gamma(X\widehat{\beta}^{LS}) \right] \\
&= \mathbb{E}_\beta \left[\gamma^2(\widehat{\beta}^{LS}) + \frac{2S}{n-p} \left\{ \frac{S}{n-p+2} \Delta_Y \gamma(X\widehat{\beta}^{LS}) + 2(X\widehat{\beta} - Y)^t \nabla_Y \gamma(X\widehat{\beta}^{LS}) \right\} \right].
\end{aligned}$$

Finally, a sufficient condition for $\Delta_\beta(\widehat{L}_\gamma, \widehat{L}_0)$ to be negative is that the random variable inside the parenthesis is always negative. This completes the proof. \square

Note that, in the Gaussian case where the variance σ^2 is assumed to be known, the difference in risks results in the following condition

$$\gamma^2(X\widehat{\beta}^{LS}) + 2\sigma^4 \Delta_Y \gamma(X\widehat{\beta}^{LS}) + 4\sigma^2 (X\widehat{\beta} - Y)^t \nabla_Y \gamma(X\widehat{\beta}^{LS}) \leq 0.$$

This result was obtained in [Johnstone 1988] and [Fourdrinier & Wells 2012] under the canonical form and taking $\sigma^2 = 1$. Plugging in the unbiased estimator of σ^2 yields

$$\gamma^2(X\widehat{\beta}^{LS}) + 2 \frac{\|Y - X\widehat{\beta}^{LS}\|^2}{n-p} \left\{ \frac{\|Y - X\widehat{\beta}^{LS}\|^2}{n-p} \Delta_Y \gamma(X\widehat{\beta}^{LS}) + 2(X\widehat{\beta} - Y)^t \nabla_Y \gamma(X\widehat{\beta}^{LS}) \right\} \leq 0.$$

Comparing with the inequality in (4.9), we can see that the only difference between the Gaussian case and the spherical case lies in the denominator in the first term in $\{\}$ brackets. Indeed, in the spherical case, the estimator of the variance is weighted by $n-p+2$, while in the Gaussian case it is weighted by $n-p$, corresponding to the unbiased estimator.

4.1.3 Choice of the correction function

In this paragraph, we consider the two cases of full linear model and restricted linear model separately. We propose one correction function for each case.

[Johnstone 1988] and [Fourdrinier & Wells 1995a] developed improved estimators with correction of the form

$$\gamma(X\widehat{\beta}^{LS}) = \frac{c}{\|X\widehat{\beta}^{LS}\|^2},$$

where c is a constant. Note that, in the full linear model, this correction is constant with respect to the selection subset I and thus its minimum occurs for the same subset as the unbiased estimator's minimum. Hence, for the problem of selecting among several subsets I_1, \dots, I_m , this correction function is only interesting for the restricted model and takes the form

$$\gamma_r(X_I \widehat{\beta}_I^{LS}) = \frac{c_r}{\|X_I \widehat{\beta}_I^{LS}\|^2},$$

where the subscript r stands for “restricted model”.

For the full linear model, we propose the following correction function instead

$$\gamma_f(X\hat{\beta}^{LS}) = c_f \left(k Z_{(k+1)}^2 + \sum_{j=k+1}^p Z_{(j)}^2 \right)^{-1},$$

where $Z_j = (Q^j)^t X \hat{\beta}^{LS}$, Q^j being the j^{th} column of the matrix Q computed by the QR factorization of X , $Z_{(j)}$ is the j^{th} element of the vector Z ordered by decreasing absolute value $|Z_{(1)}| \geq \dots \geq |Z_{(p)}|$, k is the number of selected variables, and c_f is a constant. This form might appear strange in the Gaussian case, but it will make more sense in the spherical case under a transformation of the linear model. The correction γ_f is a function of both the number of selected variables and the information contained in the non-selected variables, through the size k of the submodel considered and through the information of the rejected variables contained in $Z_{(k+1)}, \dots, Z_{(p)}$. This reflects our belief that the non-selected variables can help evaluating how good the selection is.

Remark 4.1. Note that [Fourdrinier & Wells 2012] state that γ_r is twice weakly differentiable for $k > 4$. As far as γ_f is concerned, its twice weakly differentiability is proven in Appendix A.2 under the canonical form. Hence, both correction functions satisfy the condition for Lemma (4.1).

For the corrected estimator \hat{L}_γ to improve on the unbiased estimator \hat{L}_0 , the constants c_r and c_f should be calibrated in a way that the sufficient condition (4.9) holds.

4.2 Corrected loss estimators for the restricted model

4.2.1 Condition for improvement with γ_r

Let us start with the correction function γ_r

$$\gamma_r(X_I \hat{\beta}_I^{LS}) = \frac{c_r}{\|X_I \hat{\beta}_I^{LS}\|^2}, \quad (4.19)$$

for the restricted model. We have the following result.

Theorem 4.2 (Improvement for the restricted model). *Let Y be distributed as $\mathcal{S}_n(X_I \beta_I)$. Let $\hat{\beta}_I$ be an estimator of β , $S_I = \|Y - X_I \hat{\beta}_I^{LS}\|^2$ and $\hat{\sigma}_{restricted}^2 = S_I/(n - k)$ be an estimator of the variance $\sigma^2 \mathbb{E}_{\beta_I}[\|\varepsilon\|^2]$. A sufficient condition for*

$$\hat{L}_\gamma^r(\hat{\beta}_I) = \hat{L}_0(\hat{\beta}_I) - \gamma_r(X \hat{\beta}_I^{LS}),$$

where $\gamma_r(X \hat{\beta}_I^{LS})$ is taken as in (4.19), to improve on

$$\hat{L}_0(\hat{\beta}_I) = \|Y - X_I \hat{\beta}_I\|^2 + (2 \operatorname{div}(X_I \hat{\beta}_I) - n) \hat{\sigma}_{restricted}^2$$

is that

$$\operatorname{sgn}(c_r) \left(c_r - \frac{4 \|Y - X_I \hat{\beta}_I^{LS}\|^2}{n - k} \left\{ \frac{(k - 4) \|Y - X_I \hat{\beta}_I^{LS}\|^2}{n - k + 2} + 2 (X_I \hat{\beta}_I - Y)^t X_I \hat{\beta}_I^{LS} \right\} \right) \leq 0. \quad (4.20)$$

Proof. The weak gradient and weak Laplacian of γ_r are respectively

$$\nabla_Y \gamma_r(X\hat{\beta}_I^{LS}) = -2c_r \frac{X_I \hat{\beta}_I^{LS}}{\|X_I \hat{\beta}_I^{LS}\|^4} \quad \Delta_Y \gamma_r(X\hat{\beta}_I^{LS}) = -2c_r \frac{k-4}{\|X_I \hat{\beta}_I^{LS}\|^4}.$$

Condition (4.9) yields

$$\begin{aligned} \Delta_\beta(\hat{L}_\gamma, \hat{L}_0) &= \mathbb{E}_\beta \left[\frac{c_r^2}{\|X_I \hat{\beta}_I^{LS}\|^4} - \frac{4c_r S_I}{n-k} \left\{ \frac{(k-4)S_I}{(n-k+2)\|X_I \hat{\beta}_I^{LS}\|^4} + \frac{2(X_I \hat{\beta}_I - Y)^t X_I \hat{\beta}_I^{LS}}{\|X_I \hat{\beta}_I^{LS}\|^4} \right\} \right] \\ &= \mathbb{E}_\beta \left[\frac{c_r}{\|X_I \hat{\beta}_I^{LS}\|^4} \left(c_r - \frac{4S_I}{n-k} \left\{ \frac{(k-4)S_I}{n-k+2} + 2(X_I \hat{\beta}_I - Y)^t X_I \hat{\beta}_I^{LS} \right\} \right) \right]. \end{aligned} \quad (4.21)$$

A sufficient condition for (4.21) to be negative is that the statistic

$$c_r \left(c_r - \frac{4S_I}{n-k} \left\{ \frac{(k-4)S_I}{n-k+2} + 2(X_I \hat{\beta}_I - Y)^t X_I \hat{\beta}_I^{LS} \right\} \right)$$

is always negative. As $(X_I \hat{\beta}_I - Y)^t X_I \hat{\beta}_I^{LS}$ can be negative, we do not know the sign of the term in $\{\}$ brackets and thus we do not know the sign of c_r either. Hence, we obtain the desired result. \square

The value of c_r leading to the best improvement is the center of the interval defined by Equation (4.20), that is,

$$c_r^* = \frac{2S_I}{n-k} \left\{ \frac{(k-4)S_I}{n-k+2} + 2(X_I \hat{\beta}_I - Y)^t X_I \hat{\beta}_I^{LS} \right\}. \quad (4.22)$$

As this value still depends on the data and on the estimator, we next investigate possible simplifications for the restricted Least-squares estimator, the restricted James-Stein estimator, and for the Lasso.

4.2.2 Application to estimators of the regression coefficient

Restricted least-squares estimator. When $\hat{\beta}_I$ is taken to be the restricted least-squares estimator $\hat{\beta}_I^{LS}$, the scalar product between $X_I \hat{\beta}_I^{LS} - Y$ and $X_I \hat{\beta}_I^{LS}$ is null since $\hat{\beta}_I$ is the projection of Y on the column space of X_I . Hence c_r is positive and should range over

$$0 \leq c_r \leq \frac{4(k-4)S_I^2}{(n-k)(n-k+2)}. \quad (4.23)$$

In particular, for greater improvement, the optimal value is

$$c_r^* = \frac{2(k-4)S_I^2}{(n-k)(n-k+2)},$$

leading to the following corrected estimator

$$\begin{aligned} \hat{L}_\gamma^r(\hat{\beta}_I^{LS}) &= \|Y - X_I \hat{\beta}_I^{LS}\|^2 + (2k-n) \frac{\|Y - X_I \hat{\beta}_I^{LS}\|^2}{n-k} - \frac{2(k-4)}{(n-k)(n-k+2)} \frac{\|Y - X_I \hat{\beta}_I^{LS}\|^4}{\|X_I \hat{\beta}_I^{LS}\|^4} \\ &= \frac{k}{n-k} \|Y - X_I \hat{\beta}_I^{LS}\|^2 - \frac{2(k-4)}{(n-k)(n-k+2)} \frac{\|Y - X_I \hat{\beta}_I^{LS}\|^4}{\|X_I \hat{\beta}_I^{LS}\|^4}. \end{aligned} \quad (4.24)$$

Note that the improvement over the unbiased estimator is only possible when $k \geq 5$. This result is similar to that obtained for the problem of estimation of the mean of a Gaussian vector in [Johnstone 1988].

James-Stein estimator. We now take $\hat{\beta}_I$ to be the James-Stein estimator $\hat{\beta}_I^{LS}$ on subset I . Taking its expression given in (3.3), the scalar product reduces to

$$\begin{aligned} (X_I \hat{\beta}_I - Y)^t X_I \hat{\beta}_I^{LS} &= \left(X_I \hat{\beta}_I^{LS} - \frac{k-2}{\|X_I \hat{\beta}_I^{LS}\|^2} X_I \hat{\beta}_I^{LS} - Y \right)^t X_I \hat{\beta}_I^{LS} \\ &= -\frac{k-2}{\|X_I \hat{\beta}_I^{LS}\|^2} \hat{\beta}_I^{LS} X_I^t X_I \hat{\beta}_I^{LS} \\ &= -(k-2). \end{aligned}$$

The sufficient condition (4.20) thus results in

$$\text{sgn}(c_r) \left(c_r - 4 \frac{S_I}{n-k} \left\{ (k-4) \frac{S_I}{n-k+2} - 2(k-2) \right\} \right) \leq 0.$$

Note that, in the Gaussian case where the variance σ^2 is assumed to be known and taken to 1, [Johnstone 1988] proved that the condition reduces to

$$\text{sgn}(c_r) (c_r - 4 \{(k-4) - 2(k-2)\}) = c_r (c_r + 4k) \leq 0.$$

which gives a negative constant c_r ranging over

$$-4k \leq c_r \leq 0.$$

In particular, for greater improvement, the optimal value is

$$c_r^* = -2k.$$

Also, we can compute the divergence of the James-Stein estimator from its expression in (3.3):

$$\begin{aligned} \text{div}_Y(X \hat{\beta}_I^{JS}) &= k - (k-2) \text{div}_Y \left(\frac{X_I \hat{\beta}_I^{LS}}{\|X_I \hat{\beta}_I^{LS}\|^2} \right) \\ &= k - (k-2) \left(\frac{k}{\|X_I \hat{\beta}_I^{LS}\|^2} - 2 \frac{(\hat{\beta}_I^{LS})^t X_I X_I^t \hat{\beta}_I^{LS}}{\|X_I \hat{\beta}_I^{LS}\|^4} \right) \\ &= k - \frac{(k-2)^2}{\|X_I \hat{\beta}_I^{LS}\|^2}. \end{aligned}$$

These results lead to the following corrected estimator

$$\hat{L}_\gamma^r(\hat{\beta}_I^{JS}) = \|Y - X_I \hat{\beta}_I^{JS}\|^2 + \left(2k - \frac{2(k-2)^2}{\|X_I \hat{\beta}_I^{LS}\|^2} - n \right) \frac{\|Y - X \hat{\beta}^{LS}\|^2}{n-k} + \frac{2k}{\|X_I \hat{\beta}_I^{LS}\|^2}.$$

Note that a similar result has also been derived in [Johnstone 1988].

Lasso. The Lasso estimator can have an explicit expression for a given λ and assuming we have the knowledge of its null components. This expression is given by [Zou *et al.* 2007]

$$\begin{aligned} X_I \hat{\beta}_I^{lasso} &= X_I (X_I^t X_I)^{-1} (X_I^t Y - \lambda \text{sgn}(\hat{\beta}_I^{lasso})) \\ &= X_I \hat{\beta}_I^{LS} - \lambda X_I (X_I^t X_I)^{-1} \text{sgn}(\hat{\beta}_I^{lasso}). \end{aligned} \tag{4.25}$$

Hence, the scalar product $(X\hat{\beta}_I^{lasso} - Y)^t X_I \hat{\beta}_I^{LS}$ is equal to

$$\begin{aligned} (X_I \hat{\beta}_I^{lasso} - Y)^t X_I \hat{\beta}_I^{LS} &= -\lambda \operatorname{sgn}(\hat{\beta}_I^{lasso})^t (X_I^t X_I)^{-1} X_I^t X_I \hat{\beta}_I^{LS} \\ &= -\lambda \operatorname{sgn}(\hat{\beta}_I^{lasso})^t \hat{\beta}_I^{LS}. \end{aligned}$$

This equality hence results in the following condition on c_r :

$$\operatorname{sgn}(c_r) \left(c_r - 4 \frac{S_I}{n-k} \left\{ (k-4) \frac{S_I}{n-k+2} - 2\lambda \operatorname{sgn}(\hat{\beta}_I^{lasso})^t \hat{\beta}_I^{LS} \right\} \right) \leq 0.$$

Recalling that, when $\lambda = 0$, the Lasso estimator is equal to the Least-Squares estimator, we can easily recover the range given in (4.23). Hence, in that case, c_r is positive. On the opposite limit, when λ is sufficiently large, the most-right term can exceed $-4(k-4)S_I/(n-k+2)$ and results in a negative constant c_r .

From Equation (4.25), it can easily be noticed that

$$\begin{aligned} \|\hat{\beta}_I^{lasso}\|_1 &= \operatorname{sgn}(\hat{\beta}_I^{lasso})^t \hat{\beta}_I^{lasso} \\ &= \operatorname{sgn}(\hat{\beta}_I^{lasso})^t (\hat{\beta}_I^{LS} - \lambda (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{lasso})). \end{aligned}$$

so that

$$\operatorname{sgn}(\hat{\beta}_I^{lasso})^t \hat{\beta}_I^{LS} = \|\hat{\beta}_I^{lasso}\|_1 + \lambda \operatorname{sgn}(\hat{\beta}_I^{lasso})^t (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{lasso}).$$

Also, there is a one-to-one correspondance between λ and t_λ , where t_λ intervenes in the formulation

$$\begin{cases} \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \\ \text{subject to } \|\beta\|_1 \leq t_\lambda \end{cases},$$

and the optimum β^* verifies $\|\beta^*\|_1 = t_\lambda$. Hence, we obtain the following inequality

$$\operatorname{sgn}(c_r) \left(c_r - 4 \frac{S_I}{n-k} \left\{ \frac{(k-4)S_I}{n-k+2} - 2\lambda \left(t_\lambda + \lambda \operatorname{sgn}(\hat{\beta}_I^{lasso})^t (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{lasso}) \right) \right\} \right) \leq 0.$$

In particular, a good value for c_r would be

$$c_r^* = \frac{2(k-4)\|Y - X\hat{\beta}^{LS}\|^2}{(n-k)(n-k+2)} - \frac{2\lambda S_I}{n-k} \left(t_\lambda + \lambda \operatorname{sgn}(\hat{\beta}_I^{lasso})^t (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{lasso}) \right).$$

However, when λ is not fixed but computed so as to be a transition point, namely, a point for which a new variable X^j is added or deleted to the subset, things are more complicated. According to [Zou *et al.* 2007], we can express the Lasso estimator at a transition point as

$$\begin{aligned} X_I \hat{\beta}_I^{lasso} &= X_I \hat{\beta}_I^{LS} - \frac{X_I (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{LS}) (X^j)^t (\mathbf{I}_n - X_I (X_I^t X_I)^{-1} X_I^t) Y}{\operatorname{sgn}(\hat{\beta}_j^{LS}) - (X^j)^t X_I (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{LS})} \\ &= X_I \hat{\beta}_I^{LS} - \frac{X_I (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{LS}) (X^j)^t (Y - X_I \hat{\beta}_I^{LS})}{\operatorname{sgn}(\hat{\beta}_j^{LS}) - (X^j)^t X_I (X_I^t X_I)^{-1} \operatorname{sgn}(\hat{\beta}_I^{LS})}. \end{aligned}$$

Hence the scalar product between $X_I \hat{\beta}_I - Y$ and $X_I \hat{\beta}_I^{LS}$ yields

$$\begin{aligned}
 (X_I \hat{\beta}_I^{lasso} - Y)^t X_I \hat{\beta}_I^{LS} &= - \left(\frac{X_I (X_I^t X_I)^{-1} \text{sgn}(\hat{\beta}_I^{LS}) (X^j)^t (Y - X_I \hat{\beta}_I^{LS})}{\text{sgn}(\hat{\beta}_j^{LS}) - (X^j)^t X_I (X_I^t X_I)^{-1} \text{sgn}(\hat{\beta}_I^{LS})} \right)^t X_I \hat{\beta}_I^{LS} \\
 &= - \frac{(Y - X_I \hat{\beta}_I^{LS})^t X^j \text{sgn}(\hat{\beta}_I^{LS})^t (X_I^t X_I)^{-1} X_I^t X_I \hat{\beta}_I^{LS}}{\text{sgn}(\hat{\beta}_j^{LS}) - (X^j)^t X_I (X_I^t X_I)^{-1} \text{sgn}(\hat{\beta}_I^{LS})} \\
 &= - \frac{(Y - X_I \hat{\beta}_I^{LS})^t X^j \text{sgn}(\hat{\beta}_I^{LS})^t \hat{\beta}_I^{LS}}{\text{sgn}(\hat{\beta}_j^{LS}) - (X^j)^t X_I (X_I^t X_I)^{-1} \text{sgn}(\hat{\beta}_I^{LS})} \\
 &= - \frac{(Y - X_I \hat{\beta}_I^{LS})^t X^j \|\hat{\beta}_I^{LS}\|_1}{\text{sgn}(\hat{\beta}_j^{LS}) - (X^j)^t X_I (X_I^t X_I)^{-1} \text{sgn}(\hat{\beta}_I^{LS})}
 \end{aligned}$$

Here, we cannot use the same technique as earlier since t_λ is not fixed anymore either. Hence, the resulting condition

$$\text{sgn}(c_r) \left(c_r - \frac{4 S_I}{n - k} \left\{ (k - 4) \frac{S_I}{n - k + 2} - \frac{2 (Y - X_I \hat{\beta}_I^{LS})^t X^j \|\hat{\beta}_I^{LS}\|_1}{\text{sgn}(\hat{\beta}_j^{LS}) - (X^j)^t X_I (X_I^t X_I)^{-1} \text{sgn}(\hat{\beta}_I^{LS})} \right\} \right) \leq 0$$

still depends on the data.

Other estimators. Since the other estimators of β we exposed in Chapter 2 are mostly based on the Lasso, there is little hope to obtain a value of c_r that will not depend on the data. In practice, we will use the value c_r^* defined in Equation (4.22), as it yields good performances in the simulation study. However, from a theoretical perspective, this value is not completely satisfying as we would have to verify whether it actually leads to a lower risk $\mathcal{R}(\hat{\beta}, \hat{L}_\gamma)$ than the risk of the unbiased estimator of loss, namely $\mathcal{R}(\hat{\beta}, \hat{L}_0)$. This verification happens to be quite challenging because of the dependence with the data, so we defer it for future works.

4.3 Corrected loss estimators for the full model

4.3.1 Condition for improvement with γ_f

We move now to the correction function γ_f

$$\gamma_f(X \hat{\beta}^{LS}) = c_f \left(k Z_{(k+1)}^2 + \sum_{j=k+1}^p Z_{(j)}^2 \right)^{-1}, \quad (4.26)$$

where $Z_j = (Q^j)^t X \hat{\beta}^{LS}$, Q^j . We obtain an analog result.

Theorem 4.3 (Improvement for the full model). *Let Y be distributed as $\mathcal{S}_n(X\beta)$ where σ^2 is assumed to be known. Let $\hat{\beta} = \hat{\beta}(I)$ be an estimator of β and $S = \|Y - X \hat{\beta}^{LS}\|^2$. A sufficient condition for*

$$\hat{L}_\gamma^f(\hat{\beta}) = \hat{L}_0(\hat{\beta}) - \gamma_f(X \hat{\beta}^{LS}),$$

where $\gamma_f(X \hat{\beta}^{LS})$ is taken as in (4.26), to improve on

$$\hat{L}_0(\hat{\beta}) = \|Y - X \hat{\beta}\|^2 + (2 \text{div}(X \hat{\beta}) - n) \hat{\sigma}_{full}^2$$

is that

$$\text{sgn}(c_f) \left(c_f + \frac{2S}{n-p} \left\{ \frac{S}{n-p+2} \left(-2p + \frac{4k(k+1)Z_{(k+1)}^2}{d(X\hat{\beta}^{LS})} \right) + 4d(X\hat{\beta}^{LS}) \right\} \right) \leq 0, \quad (4.27)$$

with $d(X\hat{\beta}^{LS}) = kZ_{(k+1)}^2 + \sum_{j=k+1}^p Z_{(j)}^2$ is the denominator of $\gamma_f(X\hat{\beta}^{LS})$.

Proof. Note first that $d(Y)$ can be reformulated as

$$d(X\hat{\beta}^{LS}) = (\hat{\beta}^{LS})^t X^t G_1^t M G_1 X \hat{\beta}^{LS} \quad (4.28)$$

with M the diagonal matrix with diagonal components

$$M_{i_j, i_j} = \begin{cases} 0 & \text{for } j = 1, \dots, k \\ k+1 & \text{for } j = k+1 \\ 1 & \text{for } j = k+2, \dots, p \end{cases},$$

the index i_j corresponding to the j th component of the ordered variable $(Z_{(1)}, \dots, Z_{(p)})$. The weak gradient and weak Laplacian of γ_f are respectively

$$\begin{aligned} \nabla_Y \gamma_f(X\hat{\beta}^{LS}) &= -2c_f X(X^t X)^{-1} X^t G_1^t M G_1 X \hat{\beta}^{LS} / d^2(X\hat{\beta}^{LS}) \\ \Delta_Y \gamma_f(X\hat{\beta}^{LS}) &= -\frac{2c_f}{d^2(X\hat{\beta}^{LS})} \text{tr}(H_f) + \frac{4c_f}{d^3(X\hat{\beta}^{LS})} (\hat{\beta}^{LS})^t X^t G_1^t M^2 G_1 X \hat{\beta}^{LS}, \end{aligned}$$

where $H_f = X(X^t X)^{-1} X^t G_1^t M G_1 X (X^t X)^{-1} X^t$. Replacing X by its decomposition $G_1^t A$ and noticing that $G_1 G_1^t = I_p$, the trace is easily computed as follows

$$\begin{aligned} \text{tr}(H_f) &= \text{tr}(G_1^t A (A^t G_1 G_1^t A)^{-1} A^t G_1 G_1^t M G_1 G_1^t A (A^t G_1 G_1^t A)^{-1} A^t G_1) \\ &= \text{tr}(G_1^t A (A^t A)^{-1} A^t M A (A^t A)^{-1} A^t G_1) \\ &= \text{tr}(G_1^t A A^{-1} (A^t)^{-1} A^t M A A^{-1} (A^t)^{-1} A^t G_1) \\ &= \text{tr}(M G_1 G_1^t) \\ &= \text{tr}(M) \\ &= p. \end{aligned}$$

Noticing that $\hat{\beta}_{(j)} = 0$ for $j = k+1, \dots, p$ while $(\nabla_Y \gamma_f(X\hat{\beta}^{LS}))_{(j)} = 0$ for $j = 1, \dots, k$, we obtain the following dot product between $(X\hat{\beta} - Y)$ and the weak gradient of γ_f :

$$(X\hat{\beta} - Y)^t \nabla_Y \gamma_f(X\hat{\beta}^{LS}) = 2c_f / d(X\hat{\beta}^{LS}).$$

Indeed, the dot product is actually equal to

$$\begin{aligned} (X\hat{\beta} - Y)^t \nabla_Y \gamma_f(X\hat{\beta}^{LS}) &= -\frac{2c_f}{d^2(Y)} \left(\hat{\beta}^t X^t G_1^t M G_1 X \hat{\beta}^{LS} - Y^t X (X^t X)^{-1} X^t G_1^t M G_1 X \hat{\beta}^{LS} \right) \\ &= -\frac{2c_f}{d^2(X\hat{\beta}^{LS})} \hat{\beta}^t X^t G_1^t M G_1 X \hat{\beta}^{LS} + \frac{2c_f}{d(X\hat{\beta}^{LS})}, \end{aligned}$$

where the last equation derives from (4.28). Now, $G_1 X$ is a $p \times p$ matrix, which we will call A , and, reorganizing the vectors and matrices so that the indices of I are adjacent, the product $\hat{\beta}^t X^t G_1^t M$

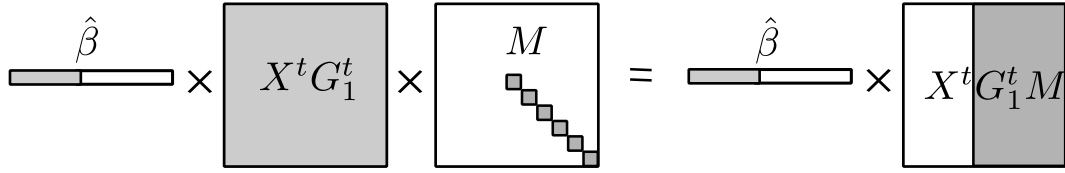


Figure 4.1: Sketch of the product $\hat{\beta}^t X^t G_1^t M$ (reorganized according to the selection I): non null components in grey, null components in white.

can be sketched as in Figure 4.1, where the grey parts represent the non null components and the white one the null components, so that we obtain

$$\hat{\beta}^t X^t G_1^t M = 0.$$

Hence condition (4.9) yields

$$\frac{c_f}{d^2(X\hat{\beta}^{LS})} \left(c_f + \frac{2S}{n-p} \left\{ \frac{S}{n-p+2} \left(-2p + \frac{4k(k+1)Z_{(k+1)}^2}{d(X\hat{\beta}^{LS})} \right) + 4d(X\hat{\beta}^{LS}) \right\} \right) \leq 0,$$

which implies the desired result by positivity of $d^2(X\hat{\beta}^{LS})$.

□

From Equation (4.27), the value of c_f leading to the best improvement is thus

$$c_f^* = \frac{S}{n-p} \left\{ \frac{S}{n-p+2} \left(-2p + \frac{4k(k+1)Z_{(k+1)}^2}{d(X\hat{\beta}^{LS})} \right) + 4d(X\hat{\beta}^{LS}) \right\}. \quad (4.29)$$

Note that, unlike the correction γ_r for the restricted model, the correction γ_f for the full model does not depend on $\hat{\beta}$, but only on its number of non-zero components. However, noticing that

$$(k+1)Z_{(k+1)}^2 \leq d(X\hat{\beta}^{LS}) \quad \Rightarrow \quad -\frac{(k+1)Z_{(k+1)}^2}{d(X\hat{\beta}^{LS})} \geq -1$$

and that

$$d(X\hat{\beta}^{LS}) \leq \|X\hat{\beta}^{LS}\|^2,$$

we can approximate c_f^* by

$$\hat{c}_f = \frac{2(p-2k)S^2}{(n-p)(n-p+2)} - \frac{4S}{(n-p)} \|X\hat{\beta}^{LS}\|^2.$$

Once again, both values of c_f are not completely satisfactory because of their dependence to the data, the verification that \hat{L}_γ^f actually dominates \hat{L}_0 with such values is again challenging. However, we will see in Chapter 6 that the simulation results show good performances in selection with these values.

4.4 Link with principled methods

In this section, we propose to investigate the possible relations between loss estimation theory on one side and the Statistical Learning Theory [Vapnik 1998] and the theory behind Slope heuristics [Birgé & Massart 2007] on the other side, which we will refer to as *data-driven penalties* to include related works such as [Arlot & Massart 2009]. The following discussion concerns the linear regression model only.

Considering the estimator $\hat{\beta} = \hat{\beta}^{(m)}$ relying on Model \mathcal{M}_m , we remind that our corrected estimators and the criteria respectively developed by [Birgé & Massart 2007], namely slope heuristics (SH), and [Vapnik 1998], namely Structural Risk Minimization (SRM), for the regression framework are

$$\begin{aligned}\widehat{L}_\gamma(\hat{\beta}^{(m)}) &= \|Y - X\hat{\beta}^{(m)}\|^2 + (2\widehat{df} - n)\hat{\sigma}^2 - \gamma(X\hat{\beta}^{LS}) \\ \text{SH}(\hat{\beta}^{(m)}) &= \|Y - X\hat{\beta}^{(m)}\|^2 + \text{pen}(m) \\ \text{SRM}(\hat{\beta}^{(m)}) &= \frac{1}{n} \|Y - X\hat{\beta}^{(m)}\|^2 \times \text{pen}(n, \text{VC-dim}(m)).\end{aligned}$$

From these expressions, we notice a similar form between \widehat{L}_γ and SH, while SRM is different in shape because of the multiplicative penalty.

The key equation that links the theories is related to the notion of *ideal penalty* defined by [Birgé & Massart 2007] as

$$\text{pen}_{\text{id}}(m) \triangleq \mathbb{E}_Y[\|Y - X\hat{\beta}^{(m)}\|^2] - \|Y - X\hat{\beta}^{(m)}\|^2. \quad (4.30)$$

In the same spirit, we can define what would be our *ideal correction* by

$$\gamma_{\text{id}}(m) \triangleq \widehat{L}_0(\hat{\beta}^{(m)}) - \|X\hat{\beta}^{(m)} - X\beta\|^2, \quad (4.31)$$

while the ideal penalty for SRM would be

$$\text{pen}_{\text{id}}^{\text{SRM}}(m) \triangleq \frac{\mathbb{E}_y[(y - x^t\hat{\beta})^2] - \frac{1}{n}\|Y - X\hat{\beta}\|^2}{\mathbb{E}_y[(y - x^t\hat{\beta})^2]}. \quad (4.32)$$

From the ideal cases for each criterion, we notice similarities between Equations (4.30) and (4.32). The link with Equation (4.31) is also quite straightforward recording that, for a fixed design matrix X ,

$$\mathbb{E}_Y[\|Y - X\hat{\beta}^{(m)}\|^2] = \|X\hat{\beta}^{(m)} - X\beta\|^2 + n\sigma^2.$$

Indeed, replacing \widehat{L}_0 by its expression $\|X\hat{\beta}^{(m)} - Y\|^2 + (2\widehat{df} - n)\hat{\sigma}^2$ and from the relationship between $\mathbb{E}_Y[\|Y - X\hat{\beta}^{(m)}\|^2]$ and $\|X\hat{\beta}^{(m)} - X\beta\|^2$, we can express the ideal correction as a function of the ideal penalty:

$$\gamma_{\text{id}}(m) = -\text{pen}_{\text{id}}(m) + n\sigma^2 + (2\widehat{df} - n)\hat{\sigma}^2.$$

Birgé and Massart's theory relies in the proposition of good estimators of $\text{pen}_{\text{id}}(m)$ and Vapnik's theory aims at bounding $\text{pen}_{\text{id}}^{\text{SRM}}(m)$, while our theory is to propose good estimators of γ_{id} . Hence, the reasons for improvement over classical criteria such as C_p or the unbiased loss estimator are the same in theory, but the means used to perform such an improvement are different. Indeed, Birgé and Massart use oracle inequalities for assessing the quality of SH, and Vapnik uses uniform deviations in order to get better generalization bounds. On the other hand,

we try to minimize the Mean Squared Error of the model selection criterion $\widehat{L}_\gamma = \widehat{L}_0 - \gamma$ for the estimating the loss $\|X\widehat{\beta}^{(m)} - X\beta\|^2$, that is we try to minimize simultaneously the variance and the bias of the model selection criterion:

$$\forall \beta \quad \min_{\gamma} \left\{ MSE(\widehat{L}_\gamma) = \mathbb{E}_Y \left[\left(\widehat{L}_0 - \gamma - \|X\widehat{\beta}_m - X\beta\|^2 \right)^2 \right] \right\}. \quad (4.33)$$

In the three cases, the aim is to have a better control of the criterion.

Now, it is impossible to perform the minimization (4.33) over any function γ . This is why we proposed two shapes for the correction function. Similarly, [Birgé & Massart 2007] proposed several form for the penalty pen, given in Chapter 2 and recorded in Table 4.2. Note that the main difference between our corrected estimators and slope heuristics is that we consider an additive term to the penalty of the unbiased estimator, namely $\text{pen}(m) = (2\hat{\sigma}^2 - n)\widehat{df}$, while SH modifies C_p 's penalty by a multiplicative term. For both criteria however, the modification is based on data, hence the name *data-dependent penalty* often encountered, while it is not the case for SRM.

The main elements of this discussion are presented in Table 4.2.

4.5 Summary

In this chapter, we discussed the problem of comparing model evaluation criteria. We proposed to perform such a comparison through an additional layer of evaluation, relying on the assessment of the quality of a loss estimator through its risk. This principle is the same as the one used to evaluate the estimators $\widehat{\beta}$ of the regression parameter β , namely the Mean Squared Error principle. We used this mode of evaluation in order to derive corrected estimators of the loss under the spherical assumption, and proposed two different shape of the correction depending on whether we assume the full model or the restricted model to be the true one. We obtained sufficient conditions of improvement (or domination) of the corrected estimators over the unbiased estimators. Then, we tried to optimize the corrections according to these sufficient conditions. However, there is no theoretical guaranty that the corrections we obtained in this way actually yields lower Mean Squared Error. Hence, there is still a lot of work to do to verify the domination. On the other hand, this problem might come from the choice of the corrections and it might be overcome by considering other corrections.

	Loss estimation theory	Data-driven penalties	Statistical Learning Theory
ASSUMPTIONS ON THE NOISE	Any multivariate spherical distributions (including non <i>i.i.d</i> case)	Univariate Gaussian distribution, univariate heteroscedastic distribution	Any univariate distribution, <i>i.i.d</i> case
CRITERION	$crit = \ Y - X\hat{\beta}\ ^2 + (2\text{div}(X\hat{\beta}) - n)\hat{\sigma}^2 - \gamma(X\hat{\beta}^{LS})$	$crit = \ Y - X\hat{\beta}\ ^2 + \text{pen}(m)$	$crit = \frac{1}{n}\ Y - X\hat{\beta}\ ^2 \times \text{pen}(n, \text{VC-dim})$
IDEAL CASE	$\gamma_{\text{id}} = \ X\hat{\beta} - X\beta\ ^2 - \ Y - X\hat{\beta}\ ^2 - (2\text{div}(X\hat{\beta}) - n)\hat{\sigma}^2$	$\text{pen}_{\text{id}} = \mathbb{E}_Y[(Y - X\hat{\beta})^2] - \frac{1}{n}\ Y - X\hat{\beta}\ ^2$	$\text{pen}_{\text{id}} = \left(\mathbb{E}_Y[(Y - X\hat{\beta})^2] - \frac{1}{n}\ Y - X\hat{\beta}\ ^2 \right) \times \left(\mathbb{E}_Y[(Y - X\hat{\beta})^2] \right)^{-1}$
IMPROVEMENT	$\mathbb{E}_\beta[(\hat{L}_0 - \gamma(X\hat{\beta}^{LS}) - \ X\hat{\beta} - X\beta\ ^2)^2]$ $\leq \mathbb{E}_\beta[(\hat{L}_0 - \ X\hat{\beta} - X\beta\ ^2)^2]$	Oracle inequalities, localised uniform deviation	Uniform deviations
PROPOSED SHAPE	$\gamma_r(X_I\hat{\beta}_I^{LS}) = c_r\ X_I\hat{\beta}_I^{LS}\ ^{-2}$ $\gamma_f(X\hat{\beta}^{LS}) = c_f \left(kZ_{(k+1)}^2 + \sum_{i=k+1}^p Z_{(i)}^2 \right)^{-1}$ with $ Z_{(1)} > \dots, Z_{(p)} $ and $Z_j = (Q^j)^t X\hat{\beta}^{LS}$	$\text{pen}_1(m) = c_1 k$ with $k = \dim(\mathcal{M}_m)$ $\text{pen}_2(m) = c_2 k(1 + \sqrt{2L_m})^2$ with $L_m \setminus \sum_{m \geq 1} \exp(-k L_m) < \infty$ $\text{pen}_3(m) = c_3 k(1 + 2\sqrt{H(k)} + 2H(k))$ with $H(k) = \frac{1}{k} \log(\#\{\mathcal{M} \setminus \dim \mathcal{M} = k\})$ $\text{pen}_4(m) = c_4 k \left(\kappa + 2(2 - \varsigma)\sqrt{L_m} + \frac{2L_m}{\varsigma} \right)$ with $\varsigma \in (0, 1), \kappa > 2 - \varsigma$	$\text{pen}(n, v) = \sqrt{n} \times \left(\sqrt{n} - \sqrt{v \left(\log\left(\frac{n}{v}\right) + 1\right) + \frac{\log n}{2}} \right)_+^{-1}$ with $v = \text{VC-dim}$

Table 4.2: Overview of the differences between Data-driven penalties, Statistical Learning Theory and our theory of loss estimation, for the linear regression with fixed design matrix.

Algorithmic aspects

Contents

5.1	Regularization path algorithms	103
5.1.1	Least Angle Regression algorithm for Lasso (LARS)	103
5.1.2	Algorithm for Minimax Concave Penalty	110
5.2	Random variable generation for spherically symmetric distributions	116
5.2.1	Through the stochastic representation	118
5.2.2	Through mixtures of other spherical distributions	122

In this chapter, we will see the algorithmic aspects of regularization paths as well as of simulation. We begin by presenting the Least Angle Regression algorithm with its modification for Lasso (LARS), developed by [Efron *et al.* 2004]. This algorithm finds Lasso's regularization path, that is, it computes the transition points of the hyperparameter tuning the penalty. Then we extend the LARS algorithm to Minimax Concave Penalty (MCP), a method developed by [Zhang 2010] overcoming Lasso's bias. The difficulty in computing MCP's path is that its optimization problem is both nonconvex and non differentiable. However, there exist similar optimality conditions that we derive from Clarke differentials. In the second part, we discuss the generation of spherically symmetric random vectors, that we use in the simulation study (see Chapter 6).

5.1 Regularization path algorithms

In this section, we recall the idea behind the Least Angle Regression algorithm for Lasso (LARS) proposed by [Efron *et al.* 2004] (although in a slightly different form than in the original paper). This algorithm is obviously not new, but helps understand the one we propose for finding MCP' regularization path, that we will expose later in the section.

5.1.1 Least Angle Regression algorithm for Lasso (LARS)

Overview of the problem

We first recall the minimization problem solved by Lasso. Given λ a positive scalar, the Lasso estimator is solution of

$$\min_{\beta \in \mathbb{R}^p} \left\{ J_{\lambda}^{lasso}(\beta) = \frac{1}{2} \| \mathbf{Y} - \mathbf{X}\beta \|^2 + \lambda \| \beta \|_1 \right\}, \quad (5.1)$$

where \mathbf{Y} and \mathbf{X} are respectively an observation of Y and X . Since $J_\lambda^{lasso}(\beta)$ is convex for a fixed λ , the solution

$$\hat{\beta}_\lambda^{lasso} = \arg \min_{\beta \in \mathbb{R}^p} J_\lambda^{lasso}(\beta)$$

is unique. If we take λ equal to 0, Problem (5.1) results in the least squares and the solution is not sparse. On the other hand, if we set λ to a sufficiently high value, the ℓ_1 penalty overrides the squared loss $\|\mathbf{Y} - \mathbf{X}\beta\|^2$ and the solution is $\hat{\beta} = 0$, resulting in the null model with no variable selected. From there came the idea of making λ vary in order to include or delete variables from the current selection. In the sequel, we will consider starting from the null model, where $\hat{\beta}_\lambda^{lasso} = 0$ and $\lambda = +\infty$, and decreasing λ so that the variables are added one at a time until the least-squares solution is reached, where $\lambda = 0$. Hence, we compute the sequence $(\lambda^{(0)}, \dots, \lambda^{(K)})$ of hyperparameters leading to the sequence $(\hat{\beta}_{\lambda^{(0)}}^{lasso}, \dots, \hat{\beta}_{\lambda^{(K)}}^{lasso})$ of solutions, which is called the regularization path. Figure 5.1 shows a simple example of such a path. The problem thus consists in finding the sequence of hyperparameters based on data, which we now discuss.

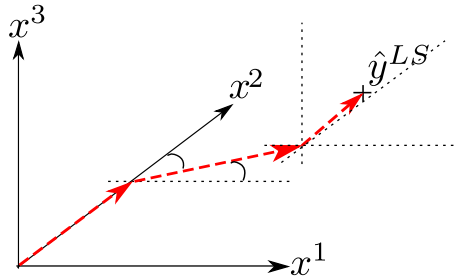


Figure 5.1: Path of solutions (in red) from $\hat{\beta}_\lambda^{lasso} = 0$ to $\hat{\beta}_\lambda^{lasso} = \hat{\beta}^{LS}$.

Optimality conditions for convex non differentiable problems

For a convex and differentiable functional J , the minimization problem

$$\min_{\beta \in \mathbb{R}^p} J(\beta)$$

is easy to solve by finding the root of the gradient of J , $\partial J / \partial \beta$. For instance, if we take the least-squares, we have

$$J(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2,$$

which is convex and differentiable. Cancelling its gradient

$$\nabla_\beta J(\beta) = -2\mathbf{X}^t(\mathbf{Y} - \mathbf{X}\beta),$$

we easily obtain the least-squares solution

$$\hat{\beta}^{LS} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}.$$

However, as mentioned earlier, the functional $J_\lambda^{lasso}(\beta)$ in (5.1) is convex but non differentiable at $\beta = 0$, for λ fixed. In such a case, we can use the extension of the gradient, the subgradient, which we define hereafter.

Definition 5.1 (Subgradients and subdifferential). *Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function. A subgradient of f at a point $\mathbf{x}_0 \in \mathbb{R}^p$ is a vector $\mathbf{g} \in \mathbb{R}^p$ satisfying the inequality*

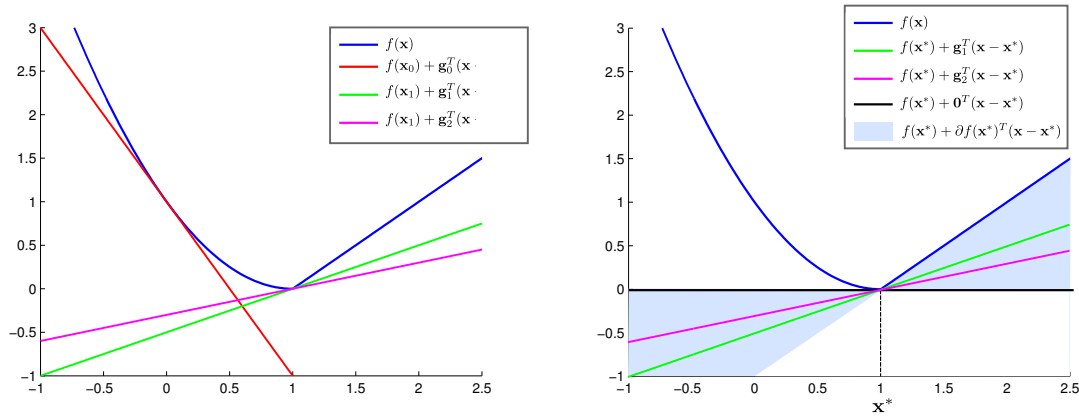
$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{g}^T(\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x} \in \mathbb{R}^p. \quad (5.2)$$

The set of all the subgradients of f at \mathbf{x}_0 is called the subdifferential and is written $\partial f(\mathbf{x}_0)$.

This definition can be found in books on convex optimization, such as [Boyd & Vandenberghe 2004] or [Bertsekas et al. 2003]. Note that the subgradient is unique and equal to the gradient when the function f is differentiable (see [Bertsekas et al. 2003]). Figure (5.2) displays a visualization of subgradients for a non differentiable function. Like for a differentiable convex functional, we look for the roots of $\partial J(\beta)$. If the null vector is a subgradient of $J(\beta)$,

$$0 \in \partial J(\beta), \quad (5.3)$$

then the minimum of J a global minimum (see [Bach et al. 2011]).



(a) Examples of subgradients of $f(\mathbf{x})$ for $\mathbf{x}_0 = 0$ and $\mathbf{x}_1 = 1$

(b) Subdifferential for $\mathbf{x}_1 = 1$

Figure 5.2: Examples of subgradients of a convex and non differentiable function f for $p = 1$.

(a) The function f is differentiable in $\mathbf{x}_0 = 0$ and thus has a unique subgradient (red line). It is non differentiable in $\mathbf{x}_1 = 1$ and has infinitely many subgradients such as \mathbf{g}_1 (green line) and \mathbf{g}_2 (magenta line). (b) The subdifferential of f at $\mathbf{x}_1 = 1$ is depicted by the blue zone. The point $\mathbf{x}^* = 1$ is the minimum of the function $f(\mathbf{x})$ since 0 is a subgradient of $f(\mathbf{x})$. (Extracted from [Flamary 2011])

The subgradient of the ℓ_1 -norm in (5.1) is

$$\forall 1 \leq j \leq p \quad \frac{\partial \|\beta\|_1}{\partial \beta_j} = \begin{cases} 1 & \text{if } \beta_j > 0 \\ -1 & \text{if } \beta_j < 0 \\ \alpha_j & \text{if } \beta_j = 0, \text{ with } -1 < \alpha_j < 1. \end{cases} \quad (5.4)$$

From now on, we assume that we know the selection I obtained for a value of λ , and we reorder and partition the data into the set I of nonzero coefficients, or equivalently the set of variables in the selection, and the set I_0 of zero coefficients. Thus, we obtain the following subgradient of J_λ^{lasso} :

$$\partial J_\lambda^{\text{lasso}}(\beta) = \begin{pmatrix} \mathbf{X}_I^t \mathbf{X}_I & \mathbf{X}_I^t \mathbf{X}_0 \\ \mathbf{X}_0^t \mathbf{X}_I & \mathbf{X}_0^t \mathbf{X}_0 \end{pmatrix} \begin{pmatrix} \beta_I \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{X}_I^t \\ \mathbf{X}_0^t \end{pmatrix} \mathbf{Y} + \lambda \begin{pmatrix} \text{sgn}(\beta_I) \\ \alpha_0 \end{pmatrix}, \quad (5.5)$$

where α_0 is the part of one subgradient (5.4) corresponding to the null components of β , that is $\beta_0 = (\beta_j)_{j \in I_0}$, and \mathbf{X}_0 is the submatrix of \mathbf{X} composed of columns with index in I_0 . This in turn leads to a system of two equations

$$\mathbf{X}_I^t \mathbf{X}_I \beta_I - \mathbf{X}_I^t \mathbf{Y} + \lambda \operatorname{sgn}(\beta_I) = 0 \quad (5.6)$$

$$\mathbf{X}_0^t \mathbf{X}_I \beta_I - \mathbf{X}_0^t \mathbf{Y} + \lambda \alpha_0 = 0. \quad (5.7)$$

This system is true for any $\lambda \in (\lambda_{(m)}, \lambda_{(m+1)})$, since the sets I and I_0 are unchanged. The idea is thus to find the next value of λ such that the sets I and I_0 change, which occurs when one of these equations reaches its limit. Equation (5.7) has limit when one component of α_0 reaches the value ± 1 , meaning that the corresponding β_j is departing from 0 and its index should enter I . On the other side, Equation (5.6), however, has limit when one component of β_I reaches the value 0 and hence its corresponding index goes back to I_0 . The algorithm thus allows to add and delete variables, depending on which of these equations reaches its limit first. Note also that Equation (5.6) recovers the result from Lemma 1 in [Zou *et al.* 2007] about the estimate of β_I for a given λ and assuming we know subset I :

$$\hat{\beta}_\lambda^{lasso} = (\mathbf{X}_I^t \mathbf{X}_I)^{-1} (\mathbf{X}_I^t \mathbf{Y} - \lambda \operatorname{sgn}(\beta_I)), \quad (5.8)$$

while Equation (5.7) computes one subgradient of $\|\beta_0\|_1$:

$$\alpha_0 = \lambda^{-1} \mathbf{X}_0^t (\mathbf{Y} - \mathbf{X}_I \beta_I). \quad (5.9)$$

Finding the path (to wisdom)

The first step is to find the value of λ such that the first variable is to be added to I . At that point, the estimator is $\hat{\beta}_{\lambda^{(0)}}^{lasso} = 0$. Hence, Equation (5.6) does not have a meaning yet and Equation (5.7) yields¹

$$-\mathbf{X}^t \mathbf{Y} + \lambda \alpha_0 = 0, \quad (5.10)$$

so that the first variable j_1 to enter I corresponds to the first component α_{j_1} to reach ± 1 . From (5.10), this can be expressed as

$$\lambda^{(0)} = \max_j |(\mathbf{X}^j)^t \mathbf{Y}| = |(\mathbf{X}^{j_1})^t \mathbf{Y}|. \quad (5.11)$$

For $\lambda \geq \lambda^{(0)}$, the j_1^{th} component of $\hat{\beta}_{\lambda^{(0)}}^{lasso}$ is still equal to 0 and gets non null at $\lambda = \lambda^{(0)} - \eta$, where η is a strictly positive scalar. The notations $\lambda^{(k)}$, $\hat{\beta}_{\lambda^{(k)}}^{lasso}$ and $\alpha_0^{(k)}$ correspond respectively to the values of λ , $\hat{\beta}_\lambda^{lasso}$ and α_0 at step k .

The next step is to find the value $\lambda^{(1)}$ such that the second variable is to be added to I and compute the corresponding estimate $\hat{\beta}_{\lambda^{(1)}}^{lasso}$. In order to do so, recall that $I = \{j_1\}$ and $I_0 = \{1, \dots, j_1 - 1, j_1 + 1, \dots, p\}$ are unchanged for $\lambda \in (\lambda^{(0)}, \lambda^{(1)})$. Hence, Equation (5.7) gives the following equations:

$$\begin{aligned} 0 - \mathbf{X}_0^t \mathbf{Y} + \lambda^{(0)} \alpha_0 &= 0 \\ \mathbf{X}_0^t \mathbf{X}^{j_1} \hat{\beta}_{j_1}' - \mathbf{X}_0^t \mathbf{Y} + \lambda' \alpha_0' &= 0, \end{aligned}$$

¹Note that, if \mathbf{Y} and \mathbf{X} have been standardized to have mean 0 and unit length, Equation (5.10) corresponds to the correlation between Y and each variable X^j .

where $\lambda' \in (\lambda^{(0)}, \lambda^{(1)})$, $\hat{\beta}'$ is its corresponding estimate and α'_0 is a subgradient of $\|\beta'_0\|_1$. Subtracting these equations yields

$$\mathbf{X}_0^t \mathbf{X}^{j_1} \hat{\beta}'_{j_1} + \lambda' \alpha'_0 - \lambda^{(0)} \alpha_0 = 0.$$

Now, we can replace $\hat{\beta}'_{j_1}$ by its expression in (5.8) and reorder the equation so that

$$\lambda' \left(\alpha'_0 - \frac{1}{(\mathbf{X}^{j_1})^t \mathbf{X}^{j_1}} \mathbf{X}_0^t \mathbf{X}^{j_1} \text{sgn}(\hat{\beta}'_{j_1}) \right) = \lambda^{(0)} \alpha_0 - \frac{1}{(\mathbf{X}^{j_1})^t \mathbf{X}^{j_1}} \mathbf{X}_0^t \mathbf{X}^{j_1} (\mathbf{X}^{j_1})^t \mathbf{Y}.$$

The next value of λ is thus the largest one (but smaller than $\lambda^{(0)}$) such that one component α_j of α'_0 reaches ± 1 . Hence, we obtain

$$\lambda^{(1)} = \max_{j \in I_0} \frac{\lambda^{(0)} \alpha_j - \frac{1}{(\mathbf{X}^{j_1})^t \mathbf{X}^{j_1}} (\mathbf{X}^j)^t \mathbf{X}^{j_1} (\mathbf{X}^{j_1})^t \mathbf{Y}}{\pm 1 - \frac{1}{(\mathbf{X}^{j_1})^t \mathbf{X}^{j_1}} (\mathbf{X}^j)^t \mathbf{X}^{j_1} \text{sgn}(\hat{\beta}'_{j_1})}.$$

Note that $\hat{\beta}'_{j_1}$ has the same sign as $(\hat{\beta}_{\lambda^{(0)}}^{lasso})_{j_1}$, except when j_1 goes back to I_0 .

Performing the same step for a given subset I , we have that, from (5.7),

$$\mathbf{X}_0^t \mathbf{X}_I (\hat{\beta}_{\lambda^{(k)}}^{lasso})_I - \mathbf{X}_0^t \mathbf{Y} + \lambda^{(k)} \alpha_0^k = 0 \quad (5.12)$$

and, for $\lambda' \in (\lambda^{(k)}, \lambda^{(k+1)})$,

$$\mathbf{X}_0^t \mathbf{X}_I \hat{\beta}'_I - \mathbf{X}_0^t \mathbf{Y} + \lambda' \alpha'_0 = 0. \quad (5.13)$$

Subtracting (5.12) to (5.13) yields

$$\mathbf{X}_0^t \mathbf{X}_I \left\{ \hat{\beta}'_I - (\hat{\beta}_{\lambda^{(k)}}^{lasso})_I \right\} + \lambda' \alpha'_0 - \lambda^{(k)} \alpha_0^k = 0. \quad (5.14)$$

Again, we replace $\hat{\beta}'_I$ and $(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I$ by the expression in (5.8) and reorder the equation so that

$$\lambda' \left(\alpha'_0 - \mathbf{X}_0^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \text{sgn}(\hat{\beta}'_I) \right) = \lambda^{(k)} \left(\alpha_0^{(k)} - \mathbf{X}_0^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \text{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I \right).$$

Since $\text{sgn}(\hat{\beta}'_I) = \text{sgn}((\hat{\beta}_{\lambda^{(k)}}^{lasso})_I)$ and the next value of λ adding a variable to I is obtained for the first component of α' reaching ± 1 , we obtain

$$\begin{aligned} \lambda_{add}(j) &= \frac{\lambda^{(k)} \left(\alpha_j^{(k)} - (\mathbf{X}^j)^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \text{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I \right)}{\pm 1 - (\mathbf{X}^j)^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \text{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I} \\ &= \lambda^{(k)} + \frac{\lambda^{(k)} \left(\alpha_j^{(k)} - \pm 1 \right)}{\pm 1 - (\mathbf{X}^j)^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \text{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I}. \end{aligned} \quad (5.15)$$

The best value of λ_{add} is the greatest positive one immediately lower than $\lambda^{(k)}$,

$$\lambda_{add}^* = \max_{j \in I_0} \left\{ \lambda_{add}(j) \setminus \lambda_{add}(j) > 0 \text{ and } \lambda_{add}(j) < \lambda^{(k)} \right\} = \lambda_{add}(j^*).$$

However, we might take a step too long by adding a variable, and we have to check at each step if we need to remove first a variable from the current set I . If so, one component of $\hat{\beta}_{\lambda^{(k)}}^{lasso}$ is going to take the value 0. Doing as in (5.14) with Equation (5.6), we have that

$$\mathbf{X}_I^t \mathbf{X}_I (\hat{\beta}_{\lambda^{(k)}}^{lasso})_I - \mathbf{X}_I^t \mathbf{Y} + \lambda^{(k)} \text{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I = 0 \quad (5.16)$$

while, for $\lambda' \in (\lambda^{(k)}, \lambda^{(k+1)})$,

$$\mathbf{X}_I^t \mathbf{X}_I \hat{\beta}'_I - \mathbf{X}_I^t \mathbf{Y} + \lambda' \operatorname{sgn}(\beta'_I) = 0. \quad (5.17)$$

Again, subtracting (5.16) to (5.17) yields

$$(\hat{\beta}'_I - (\hat{\beta}_{\lambda^{(k)}}^{lasso})_I) + (\lambda' - \lambda^{(k)}) (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I = 0, \quad (5.18)$$

since $\operatorname{sgn}(\hat{\beta}'_I) = \operatorname{sgn}((\hat{\beta}_{\lambda^{(k)}}^{lasso})_I)$. Now, as we said earlier, if we need to remove a variable j then $(\hat{\beta}_{\lambda^{(k+1)}}^{lasso})_j$ takes the value 0. Hence, this results in

$$(\hat{\beta}_{\lambda^{(k)}}^{lasso})_j - (\lambda' - \lambda^{(k)}) \left\{ (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I \right\}_j = 0.$$

This last equation implies the following update for λ

$$\lambda_{rem}^* = \max_{j \in I} \left\{ \lambda_{rem}(j) \mid \lambda_{rem}(j) > 0 \text{ and } \lambda_{rem}(j) < \lambda^{(k)} \right\} = \lambda_{rem}(j^*),$$

where

$$\lambda_{rem}(j) = \lambda^{(k)} + \frac{(\hat{\beta}_{\lambda^{(k)}}^{lasso})_j}{\left\{ (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I \right\}_j}. \quad (5.19)$$

The superscript *rem* stands for *remove*.

Finally, we update λ by taking the greatest value between λ_{add}^* and λ_{rem}^* , that is,

$$\lambda^{(k+1)} = \max\{\lambda_{add}^*, \lambda_{rem}^*\}, \quad (5.20)$$

and we either add or remove variable j^* to I depending on whether λ_{add}^* or λ_{rem}^* is the greatest.

Once λ has been updated, we can update $\hat{\beta}_{\lambda}^{lasso}$ and α_0 , too, from (5.18) and (5.14):

$$(\hat{\beta}_{\lambda^{(k+1)}}^{lasso})_I = (\hat{\beta}_{\lambda^{(k)}}^{lasso})_I + (\lambda^{(k)} - \lambda^{(k+1)}) (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I \quad (5.21)$$

$$\begin{aligned} \alpha_0^{(k+1)} &= \frac{\lambda^{(k)}}{\lambda^{(k+1)}} \alpha_0^{(k)} - \frac{1}{\lambda^{(k+1)}} \mathbf{X}_0^t \mathbf{X}_I \left\{ (\hat{\beta}_{\lambda^{(k+1)}}^{lasso})_I - (\hat{\beta}_{\lambda^{(k)}}^{lasso})_I \right\} \\ &= \frac{\lambda^{(k)}}{\lambda^{(k+1)}} \alpha_0^{(k)} + \frac{(\lambda^{(k)} - \lambda^{(k+1)})}{\lambda^{(k+1)}} \mathbf{X}_0^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I)^{-1} \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}}^{lasso})_I. \end{aligned} \quad (5.22)$$

The last step occurs when $\lambda^{(K)} = 0$ and the last variable is added to I . The corresponding update for $\hat{\beta}_{\lambda}^{lasso}$ is

$$\hat{\beta}_{\lambda^{(K)}}^{lasso} = \hat{\beta}_{\lambda^{(K-1)}}^{lasso} + \lambda^{(K-1)} (\mathbf{X}^t \mathbf{X})^{-1} \operatorname{sgn}(\hat{\beta}_{\lambda^{(K-1)}}^{lasso}) \quad (5.23)$$

and equals the least-squares estimator.

The algorithm

All these steps are gathered in Algorithm 5.1. In line 26 of the algorithm, the sign in ± 1 depends on the one in (5.19) giving the best value of λ .

Algorithm 5.1 LARS**Require:** \mathbf{X}, \mathbf{Y} **Ensure:** $(\hat{\beta}^{(k)})_{k=1}^K, (\lambda^{(k)})_{k=1}^K$

```

1:  $\hat{\beta}^{(0)} = \text{zeros}(p, 1)$ 
2:  $I_1^{(0)} = \emptyset$ 
3:  $I_0^{(0)} = \{1, \dots, p\}$ 
4:  $\lambda^{(0)} \leftarrow \max_j \text{abs}((\mathbf{X}^j)^t \mathbf{Y})$ 
5:  $j\_next \leftarrow \arg \max_j \text{abs}((\mathbf{X}^j)^t \mathbf{Y})$ 
6:  $\alpha^{(0)} \leftarrow \mathbf{X}^t \mathbf{Y} / \lambda^{(0)}$ 
7:  $k \leftarrow 0$ 
8: while  $\lambda^{(k)} > 0$  do
9:    $\omega = (\mathbf{X}_{I^{(k)}}^t \mathbf{X}_{I^{(k)}})^{-1} \text{sgn}(\hat{\beta}_{I^{(k)}}^{(k)})$ 
10:   $z = \mathbf{X}_0^t \mathbf{X}_I \omega$ 
11:  Compute  $\lambda^{add}$  and  $\lambda^{rem}$  as in (5.15) and (5.19), and find the corresponding index  $j\_next$ 
    to add or remove.
12:  if  $\lambda^{add} > \lambda^{rem}$  then
13:     $\lambda^{(k+1)} \leftarrow \lambda^{add}$ 
14:     $I^{(k+1)} \leftarrow I^{(k)} \cup \{j\_next\}$ 
15:     $I_0^{(k+1)} \leftarrow I_0^{(k)} \setminus \{j\_next\}$ 
16:     $\hat{\beta}_{I^{(k)}}^{(k+1)} = \hat{\beta}_{I^{(k)}}^{(k)} + (\lambda^{(k)} - \lambda^{(k+1)})\omega$ 
17:     $\hat{\beta}_{j\_next}^{(k+1)} = 0$ 
18:     $\alpha_{I_0^{(k+1)}}^{(k+1)} = \frac{1}{\lambda^{(k+1)}} \left( \lambda^{(k)} \alpha_{I_0^{(k)}}^{(k)} + (\lambda^{(k)} - \lambda^{(k+1)}) z_{I_0^{(k+1)}} \right)$ 
19:  else
20:     $\lambda^{(k+1)} \leftarrow \lambda^{rem}$ 
21:     $I^{(k+1)} \leftarrow I^{(k)} \setminus \{j\_next\}$ 
22:     $I_0^{(k+1)} \leftarrow I_0^{(k)} \cup \{j\_next\}$ 
23:     $\hat{\beta}_{I^{(k+1)}}^{(k+1)} = \hat{\beta}_{I^{(k+1)}}^{(k)} + (\lambda^{(k)} - \lambda^{(k+1)})\omega_{I^{(k+1)}}$ 
24:     $\hat{\beta}_{j\_next}^{(k+1)} = 0$ 
25:     $\alpha_{I_0^{(k)}}^{(k+1)} = \frac{1}{\lambda^{(k+1)}} \left( \lambda^{(k)} \alpha_{I_0^{(k)}}^{(k)} + (\lambda^{(k)} - \lambda^{(k+1)}) z \right)$ 
26:     $\alpha_{j\_next}^{(k+1)} = \pm 1$ 
27:  end if
28:   $k \leftarrow k + 1$ 
29: end while
30: Compute  $\hat{\beta}^{(k+1)}$  as in (5.23).
31:  $\lambda^{(k+1)} \leftarrow 0$ 

```

Convergence and optimization of the algorithm

In the worst possible case, since the algorithm allows to add and remove variables, the regularization path can explore the $2^p + 1$ possible subsets, exactly as an exhaustive exploration. [Mairal & Yu 2012] even showed that the number of linear segments of the path could reach up to $(3^p + 1)/2$, so that it can be even more computationally expensive than an exhaustive exploration. Nevertheless, this is only a worst-case result and in most cases where there are more observations than variables ($p < n$), the path is often of size $p + 1$ (the first solution being the null model). In practice, it is more frequent that a variable is deleted in the setting $p \geq n$, and this often occurs when the size of the selection exceeds the number of observations.

Turning to the issue of optimizing the algorithm, we notice that the most expensive operation is the inversion of matrix $(\mathbf{X}_I^t \mathbf{X}_I)^{-1}$ at each step. This issue is overcome by updating $(\mathbf{X}_{I^{(k+1)}}^t \mathbf{X}_{I^{(k+1)}})^{-1}$ at step $k + 1$ from $(\mathbf{X}_{I^{(k)}}^t \mathbf{X}_{I^{(k)}})^{-1}$, which has already been computed at step k . This can be performed thanks to the Woodbury matrix identity. See the details in Appendix A.1.

5.1.2 Algorithm for Minimax Concave Penalty

Overview of the problem

In this subsection, we are interested in computing the regularization path of the Minimax Concave Penalty (MCP) estimator developed by [Zhang 2010]. [Zhang 2010] proposed an algorithm, called MC+, that appears to be a subgradient descent computing the regularization path but is not really clear to us, while [Breheny & Huang 2011] solve MCP through a coordinate descent algorithm. This latter one does not compute the path but instead requires the user to set the value(s) for the hyperparameter. Hence, we propose here our own LARS-type algorithm that computes MCP's regularization path. We first recall the corresponding optimization problem. Let λ be a fixed positive scalar and let γ be another scalar such that $\gamma > 1$. The MCP estimator is solution of the problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ J_{\lambda}^{mcp}(\beta) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^p \rho(|\beta_j|) \right\}, \quad (5.24)$$

with

$$\rho(u) = \int_0^u \left(1 - \frac{v}{\gamma\lambda} \right)_+ dv = \begin{cases} \lambda u - \frac{u^2}{2\gamma} & \text{if } u < \gamma\lambda \\ \frac{\gamma\lambda^2}{2} & \text{otherwise} \end{cases}, \quad (5.25)$$

where $u_+ = \max(u, 0)$. Note that the penalty term $\rho(\cdot)$ in Equation (5.25) is a linear combination between the Lasso penalty and the hard thresholding penalty as defined by [Fan & Li 2001]:

$$\rho(u) = \lambda - \frac{1}{\lambda}(u - \lambda)^2 \mathbf{1}_{\{u < \lambda\}}.$$

Therefore, only the components of β such that $|\beta_j| < \gamma\lambda$ are subject to the penalization, while the others are unbiased (hence the term *nearly unbiased* used by [Zhang 2010] to qualify the MCP estimator).

Optimality conditions for nonconvex and non differentiable problems

The functional can be decomposed as follows

$$J_\lambda^{mcp}(\beta) = \phi(\beta) + \varphi(\beta) \quad (5.26)$$

with

$$\phi(\beta) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 - h(\beta) \quad (5.27)$$

$$h(\beta) = \lambda \sum_{j=1}^p \left\{ \frac{\beta_j^2}{2\gamma\lambda} \mathbb{1}_{\{|\beta_j| \leq \gamma\lambda\}} + \left(|\beta_j| - \frac{\gamma\lambda}{2} \right) \mathbb{1}_{\{|\beta_j| > \gamma\lambda\}} \right\} \quad (5.28)$$

$$\varphi(\beta) = \lambda \|\beta\|_1. \quad (5.29)$$

Note that ϕ is differentiable but, for some values of λ and γ , not convex while φ is not differentiable but convex. Figure 5.3 displays the shape of the function $h(\beta_j)$. The difficulty in

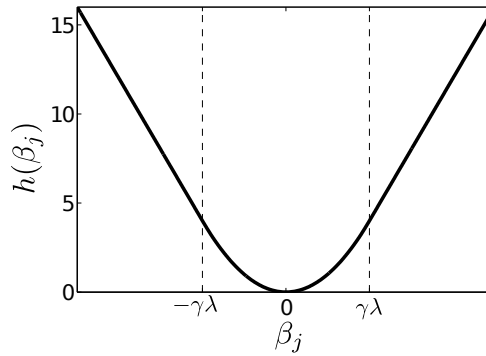


Figure 5.3: Evolution of function $h(\beta_j)$ with respect to the component β_j .

solving Problem (5.24) thus comes from its non differentiability and nonconvexity, and the subdifferential notion defined in 5.1 no longer applies. However, there exists a generalization of the subdifferential to non convex functions, known as the *Clarke differential*, and defined as follows [Clarke 1990].

Definition 5.2 (Clarke differential). *The Clarke differential is defined, for locally Lipschitz functions f , as the convex hull of some generalized gradient and more precisely*

$$\partial_c f(\beta^*) = \{\mathbf{g} \in \mathbb{R}^p \mid \mathbf{g}^t \mathbf{d} \leq D_c f(\beta^*, \mathbf{d}) \text{ for all } \mathbf{d} \in \mathbb{R}^p\}$$

where $D_c f(\beta, \mathbf{d})$ denotes the Clarke directional derivative of the function f at point β in the direction \mathbf{d} , defined by

$$D_c f(\beta, \mathbf{d}) = \limsup_{\epsilon \rightarrow 0+, \delta \rightarrow \beta} \frac{f(\delta + \epsilon \mathbf{d}) - f(\delta)}{\epsilon}.$$

Note that, for J_λ^{mcp} , Clarke directional derivative coincides with the usual directional derivative.

For the minimization of a non smooth and non convex functional, if β^* is a local minima of the proper loss function $J_\lambda^{mcp}(\beta)$ then it verifies the inclusion

$$0 \in \partial_c J_\lambda^{mcp}(\beta^*) \quad (5.30)$$

where $\partial_c J_\lambda^{mcp}(\beta^*)$ denotes the Clarke subdifferential of the functional J_λ^{mcp} at point β^* [Clarke 1990]. Note that Condition (5.30) is the generalization of Condition (5.3) to nonconvex functions.

As noticed in 5.26, our functional J_λ^{mcp} can be split into the strictly differentiable term ϕ and the convex term φ . In such a case, [Clarke 1990, Proposition 2.3.3, Corollary 1] shows that Condition (5.30) becomes

$$\beta^* \text{ is a local minima} \quad \Rightarrow \quad -\nabla_\beta \phi(\beta^*) \in \partial\varphi(\beta^*), \quad (5.31)$$

where $\nabla_\beta \phi(\beta^*)$ denotes the gradient of function ϕ at point β^* and $\partial\varphi(\beta^*)$ is the subdifferential of the convex function φ at point β^* (coinciding with its Clarke subdifferential).

Remark 5.1. Condition (5.31) suggest the use of a proximal algorithm to retrieve a stationary point (see [Hare & Sagastizábal 2009] for details).

Remark 5.2. The question for Condition (5.31) to be a sufficient condition for a feasible point to be a global minimizer remains.

From Condition (5.31), there exists one subgradient v of φ such that we have the equality

$$-\nabla_\beta \phi(\beta^*) = v \in \partial\varphi(\beta^*).$$

Computing the gradient of ϕ and from the subdifferential of the ℓ_1 – norm (5.4), the optimality conditions for MCP are thus

$$\begin{cases} (\mathbf{X}^j)^t(\mathbf{Y} - \mathbf{X}\beta) = \lambda \operatorname{sgn}(\beta_j) \dot{\rho}(|\beta_j|) & \text{if } \beta_j \neq 0 \\ |(\mathbf{X}^j)^t(\mathbf{Y} - \mathbf{X}\beta)| \leq \lambda & \text{if } \beta_j = 0, \end{cases} \quad (5.32)$$

where

$$\dot{\rho}(t) = \left(1 - \frac{t}{\gamma\lambda}\right)_+.$$

These conditions are the same as that given in [Zhang 2010]², and can also be recovered thanks to Difference of Convex (DC) programming [An & Tao 2005]. Indeed, Problem (5.24) can be expressed in a third way as

$$\min_{\beta \in \mathbb{R}^p} \left\{ J_\lambda^{mcp}(\beta) = J_\lambda^{lasso}(\beta) - h(\beta) \right\},$$

where $h(\beta)$ is defined by (5.28), and both $J_\lambda^{lasso}(\beta)$ and $h(\beta)$ are convex. Indeed, [An & Tao 2005] give the following optimality condition for DC programs

$$\beta^* \text{ is a local minima} \quad \Rightarrow \quad \nabla_\beta h(\beta^*) \in \partial J_\lambda^{lasso}(\beta^*), \quad (5.33)$$

which gives exactly (5.32). However, the condition for (5.33) to be valid is that J_λ^{mcp} is a polyhedral convex, which does not seem to be the case here. Nevertheless, Equation (5.33) can be obtained again thanks to Clarke differential, noticing that h is differentiable and J_λ^{lasso} is nondifferentiable but convex. Note that DC algorithms have been developed in [Gasso *et al.* 2009] for several nonconvex optimization problems, but not for MCP.

The conditions in (5.32) are pretty close to the system of equations (5.6) and (5.7), where we had $\dot{\rho}(t) = 1$. Hence, regularization path algorithms can be computed for nonconvex problems.

²Note that [Zhang 2010] gives Condition (5.32) without justifying their origin.

Also, it allows us to use the same trick as for Lasso and partition the system into the selection set I and the rejection set I_0 . However, this time, the system is a little more complicated as the set I itself contains two subsets: the subset $I_P = \{j \in I \mid \lambda \leq |\beta_j| < \gamma\lambda\}$ corresponding to the indices of the penalized nonzero components β_j , and the subset $I_N = \{j \in I \mid |\beta_j| \geq \gamma\lambda\}$ corresponding to the indices of the unpenalized nonzero components β_j . There are thus 4 possible moves of a given index j , as shown on Figure 5.4.

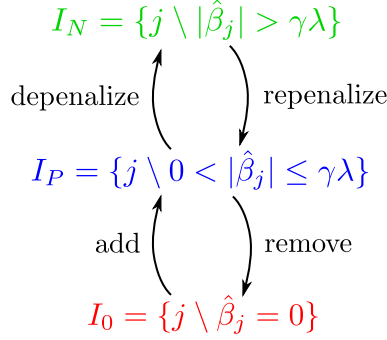


Figure 5.4: Possible moves between the subsets of indices of β .

From conditions (5.32), we derive the following system of equations:

$$\mathbf{X}_I^t \mathbf{X}_I \beta_I - \mathbf{X}_I^t \mathbf{Y} + \begin{pmatrix} 0 \\ \lambda \operatorname{sgn}(\beta_P) - \gamma^{-1} \beta_P \end{pmatrix} = 0 \quad (5.34)$$

$$\mathbf{X}_0^t \mathbf{X}_I \beta_I - \mathbf{X}_0^t \mathbf{Y} + \lambda \alpha_0 = 0, \quad (5.35)$$

where α_0 is defined as in (5.5), that is, a vector of size $p-k$ with components in $[-1; 1]$. Note that Equation (5.35) is exactly the same as (5.7) for Lasso. This follows from the fact that the penalty functions of Lasso and MCP are identical for $|\beta_j| < \lambda$, so that the respective subdifferentials of J_λ^{lasso} and J_λ^{mcp} are the same at $\beta = 0$. Equation (5.34) can be reformulated as

$$(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I) \beta_I - \mathbf{X}_I^t \mathbf{Y} + \lambda s_I = 0, \quad (5.36)$$

where $s_I = \begin{pmatrix} 0 \\ \operatorname{sgn}(\beta_P) \end{pmatrix}$ and $\Upsilon_I = \operatorname{diag}(s_I)$. We thus obtain the following expression of the MCP estimator:

$$\hat{\beta}_{\lambda, \gamma}^{\text{mcp}} = (\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} (\mathbf{X}_I^t \mathbf{Y} - \lambda s_I). \quad (5.37)$$

Here however, special care should be taken on the eigenvalues of the matrix $(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)$. Indeed, as they can be both positive or negative, one of them could also be null or close to null, so that the matrix would be ill-conditioned. In that case, [Zhang 2010] recommends using the following value for γ :

$$\gamma = \frac{2}{1 - \max_{j \neq k} \frac{|(\mathbf{X}^j)^t \mathbf{X}^k|}{n}}.$$

Note also that there is a certain similarity with the Elastic Net estimator proposed by [Zou & Hastie 2005] when $I_N = \emptyset$ and $I_P = I$. But the difference stands in that the identity matrix (up to the factor γ^{-1}) is subtracted to the matrix $\mathbf{X}^t \mathbf{X}$ in MCP, whereas it is added to $\mathbf{X}^t \mathbf{X}$ in Elastic Net. The consequence of that fact is that the Elastic Net estimator is even more biased than the Lasso estimator, while MCP corrects Lasso's bias.

Computing the regularization path of MCP

The algorithm we propose for solving MCP is inspired by the LARS algorithm we presented in the previous subsection. Indeed, we start with the null model $\hat{\beta}_{\lambda^{(0)}, \gamma}^{mcp} = 0$ and add the variables one at a time until we reach the least-squares solution for $\lambda^{(K)} = 0$. The difference here is that we need to check at each step whether there is a variable in I_P that should be “*depenalized*” and should enter I_N , or on the contrary a variable should be “*repenalized*”, meaning that it should move from I_N to I_P . The very first step of the algorithm is exactly the same as for Lasso, that is,

$$\lambda^{(0)} = \max_j |(\mathbf{X}^j)^t \mathbf{Y}| = |(\mathbf{X}^{j_1})^t \mathbf{Y}|. \quad (5.38)$$

The following ones are similar, too, except for the presence of the second hyperparameter γ :

$$\lambda^{(1)} = \max_{j \in I_0} \frac{\lambda^{(0)} \alpha_j - \frac{1}{(\mathbf{X}^{j_1})^t \mathbf{X}^{j_1} - \gamma^{-1}} (\mathbf{X}^j)^t \mathbf{X}^{j_1} (\mathbf{X}^{j_1})^t \mathbf{Y}}{\pm 1 - \frac{1}{(\mathbf{X}^{j_1})^t \mathbf{X}^{j_1} - \gamma^{-1}} (\mathbf{X}^j)^t \mathbf{X}^{j_1} \text{sgn}(\hat{\beta}'_{j_1})}.$$

This goes on until one variable enters the set I_N . Assume that we have run the algorithm up to step k , and we thus have at hand $\lambda^{(k)}$, $I_P^{(k)}$, $I_N^{(k)}$, $I_0^{(k)}$, $\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp}$ and $\alpha_0^{(k)}$. We look for the next value of λ that will make a new variable enter the set I_P . From Equation (5.35), we have that

$$\mathbf{X}_0^t \mathbf{X}_I (\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_I - \mathbf{X}_0^t \mathbf{Y} + \lambda^{(k)} \alpha_0^{(k)} = 0 \quad (5.39)$$

and, for $\lambda' \in (\lambda^{(k)}, \lambda^{(k+1)})$, the subsets stay unchanged so that

$$\mathbf{X}_0^t \mathbf{X}_I \beta'_I - \mathbf{X}_0^t \mathbf{Y} + \lambda' \alpha'_0 = 0. \quad (5.40)$$

Subtracting (5.39) to (5.40) yields

$$\mathbf{X}_0^t \mathbf{X}_I (\beta'_I - (\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_I) + \lambda' \alpha'_0 - \lambda^{(k)} \alpha_0^{(k)} = 0.$$

Replacing β'_I and $(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_I$ by the expression in (5.37) and reordering the equation, we obtain

$$\lambda' \left(\alpha'_0 - \mathbf{X}_0^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I \right) = \lambda^{(k)} \left(\alpha_0^{(k)} - \mathbf{X}_0^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I \right). \quad (5.41)$$

A variable from I_0 enters I_P when its corresponding α_j reaches ± 1 , so that

$$\lambda_{add}(j) = \lambda^{(k)} + \frac{\lambda^{(k)} (\alpha_j^{(k)} - \pm 1)}{\pm 1 - (\mathbf{X}^j)^t \mathbf{X}_I (\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I}. \quad (5.42)$$

A variable thus enters the subset I when

$$\lambda_{add}^* = \max_{j \in I_0} \left\{ \lambda_{add}(j) \mid \lambda_{add}(j) > 0 \text{ and } \lambda_{add}(j) < \lambda^{(k)} \right\} = \lambda_{add}(j^*)$$

does exist.

Doing the same with Equation (5.36), we obtain the following equation

$$(\beta'_I - (\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_I) = (\lambda^{(k)} - \lambda') (\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I.$$

A variable from I_P re-enters I_0 when its coefficient β_j takes the value 0, so that

$$\lambda_{rem}(j) = \lambda^{(k)} + \frac{(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j}{\{(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I\}_j}, \quad (5.43)$$

and the first one to do so is the one for which we have

$$\lambda_{rem}^* = \max_{j \in I_P} \left\{ \lambda_{rem}(j) \mid \lambda_{rem}(j) > 0 \text{ and } \lambda_{rem}(j) < \lambda^{(k)} \right\} = \lambda_{rem}(j^*).$$

The computations of λ_{add}^* and λ_{rem}^* are basically the same than for Lasso, except for the part with γ . Now we need to compute the value λ_{dep}^* such that one variable moves from I_P to I_N (the “*depenalization*” step), and λ_{rep}^* such that one variable moves from I_N to I_P (the “*repenalization*” step). For the first one, we go back to Equation (5.41). The limit is reached when one coefficient β_j takes the value $\gamma \lambda \operatorname{sgn}(\beta_j)$, since it is the limit of the penalized set $\{|\beta_j| \leq \gamma \lambda\}$. This leads to the equation

$$(\gamma \lambda \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j - (\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j) = (\lambda^{(k)} - \lambda') \{(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I\}_j,$$

which in turn yields

$$\begin{aligned} \lambda_{dep}(j) &= \frac{(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j + \lambda^{(k)} \{(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I\}_j}{\gamma \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j + \{(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I\}_j} \\ &= \lambda^{(k)} + \frac{(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j - \gamma \lambda^{(k)} \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j}{\gamma \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j + \{(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I\}_j}. \end{aligned} \quad (5.44)$$

The best value of $\lambda_{dep}(j)$ is thus

$$\lambda_{dep}^* = \max_{j \in I_P} \left\{ \lambda_{dep}(j) \mid \lambda_{dep}(j) > 0 \text{ and } \lambda_{dep}(j) < \lambda^{(k)} \right\} = \lambda_{dep}(j^*).$$

Finally, it is easy to see that the same limit is reached on the other side for a variable to move from I_N to I_P , involving the same equation, so that we obtain

$$\lambda_{rep} = \lambda^{(k)} + \frac{(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j - \gamma \lambda^{(k)} \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j}{\gamma \operatorname{sgn}(\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_j + \{(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I\}_j}, \quad (5.45)$$

and the first variable to do so is the one for which we have

$$\lambda_{rep}^* = \max_{j \in I_N} \left\{ \lambda_{rep}(j) \mid \lambda_{rep}(j) > 0 \text{ and } \lambda_{rep}(j) < \lambda^{(k)} \right\} = \lambda_{rep}(j^*).$$

The only difference occurs on the maximization under the set I_N instead of I_P . Now that we can compute the value of λ for any move, the end of step $k+1$ goes as for the Lasso, that is, we set

$$\lambda^{(k+1)} = \max\{\lambda_{add}^*; \lambda_{rem}^*; \lambda_{dep}^*; \lambda_{rep}^*\}, \quad (5.46)$$

and we perform the corresponding changes for the sets I_0 , I_P and I_N . The step ends with the update of $\hat{\beta}_{\lambda, \gamma}^{mcp}$ and α_0 through

$$(\hat{\beta}_{\lambda^{(k+1)}, \gamma}^{mcp})_I = (\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_I + (\lambda^{(k)} - \lambda^{(k+1)})(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \Upsilon_I)^{-1} s_I \quad (5.47)$$

$$\alpha_0^{(k+1)} = \frac{\lambda^{(k)}}{\lambda^{(k+1)}} \alpha_0^{(k)} - \frac{1}{\lambda^{(k+1)}} \mathbf{X}_0^t \mathbf{X}_I \left\{ (\hat{\beta}_{\lambda^{(k+1)}, \gamma}^{mcp})_I - (\hat{\beta}_{\lambda^{(k)}, \gamma}^{mcp})_I \right\}. \quad (5.48)$$

Just like for Lasso, the final step is computed through (5.47) with $\lambda^{(K)} = 0$:

$$\hat{\beta}_{\lambda^{(K)}, \gamma}^{mcp} = \hat{\beta}_{\lambda^{(K-1)}, \gamma}^{mcp} + \lambda^{(K-1)}(\mathbf{X}^t \mathbf{X} - \gamma^{-1} S)^{-1} s.$$

The algorithm

Algorithm 5.2 merges all these steps and uses a subfunction for the possible moves between I_0 , I_P and I_P , called `update_subset` and given in the Appendix A.3.

Convergence and optimization of the algorithm

The mere fact that the algorithm allows 4 possible moves for a given index (add or removed from selection, and depenalization of repenalization) implies a larger computational time than for Lasso. In practice, we observe that it is around 2 to 3 times longer in most cases where Lasso's regularization paths contains $p + 1$ subsets. Figure 5.5 compares both regularization paths in the case where there are more variables than observations ($n < p$). In this case, MCP's path presents a zone of instability, shown by the grey zone, before converging to one of the Least-squares solutions.

Just like in the LARS algorithm, the most expensive operation is the inverse of matrix $(\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \mathbf{Y}_I)$. We can again use the Woodbury matrix equality in order to update the inverse at step $k + 1$ thanks to the one computed at step k (see Appendix A.1).

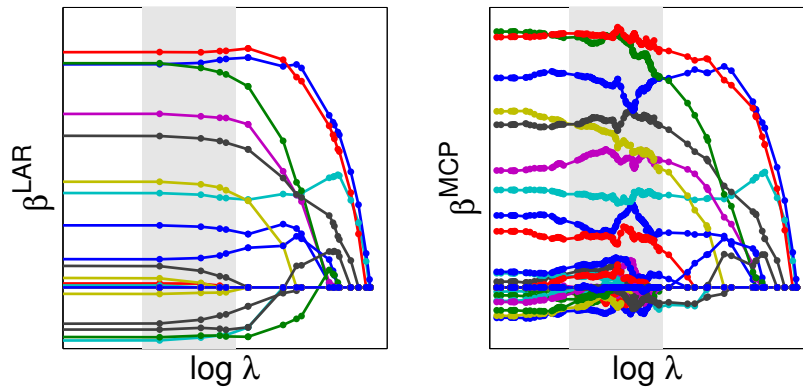


Figure 5.5: Regularization paths for Lasso and MCP, with $p = 50$ variables and $n = 20$ observations. The grey zone shows the instability of MCP before it converges to Least-squares, while Lasso is stable all along the path.

5.2 Random variable generation for spherically symmetric distributions

In this section, we are concerned with the generation of spherically symmetric random vectors. Indeed, we propose to compare on a simulation study the performances of our criteria for several spherical laws, since they are supposed to be independent of the particular form of the distribution. The simulation study is done in Chapter 6, but we study here the generation of random vectors from spherical laws. We expose in the sequel a way to perform such a pseudo-random number generator through their representation as scale mixture of spherical distributions that are easier to generate. We divide the presentation into two parts: representation as a mixture of uniforms on spheres, which exists for any spherical distribution since it corresponds to the

Algorithm 5.2 LARS-MCP: Main program**Require:** $\mathbf{X}, \mathbf{Y}, \gamma$ **Ensure:** $(\hat{\beta}^{(k)})_{k=1}^K, (\lambda^{(k)})_{k=1}^K$

$$\hat{\beta}^{(0)} \leftarrow 0$$

$$I_P^{(0)} \leftarrow \emptyset \quad I_N^{(0)} \leftarrow \emptyset \quad I_0^{(0)} \leftarrow \{1, \dots, p\}$$

$$\lambda^{(0)} \leftarrow \max_j |(\mathbf{X}^j)^t \mathbf{Y}|$$

$$j_next \leftarrow \arg \max_j |(\mathbf{X}^j)^t \mathbf{Y}|$$

$$\alpha^{(0)} \leftarrow \mathbf{X}^t \mathbf{Y} / \lambda^{(0)}$$

$$k \leftarrow 0$$

while $\lambda^{(k)} > 0$ **do**

$$[I_N^{(k+1)}, I_P^{(k+1)}, I_0^{(k+1)}] \leftarrow \text{update_subset}(I_N^{(k)}, I_P^{(k)}, I_0^{(k)}, j_next, \text{move})$$

$$I \leftarrow I_N^{(k)} \cup I_P^{(k)}$$

$$s = \begin{pmatrix} \mathbf{0}_{(n_N, 1)} \\ \text{sign}(\hat{\beta}_P^{(k)}) \end{pmatrix}$$

$$\omega = \{\mathbf{X}_I^t \mathbf{X}_I - \gamma^{-1} \text{diag}(s)\}^{-1} s$$

$$z = \mathbf{X}_0^t \mathbf{X}_I \omega$$

$$j_{add} = \arg \max_{j \in I_0} \left\{ \lambda_{add}(j) = \lambda^{(k)} + \frac{\alpha_j - \pm 1}{\pm 1 - z_j} \mid 0 < \lambda_{add}(j) < \lambda^{(k)} \right\}$$

$$\lambda_{add}^* = \lambda_{add}(j_{add})$$

$$j_{rem} = \arg \max_{j \in I_P} \left\{ \lambda_{rem}(j) = \lambda^{(k)} + \frac{\hat{\beta}_j^{(k)}}{\omega_j} \mid 0 < \lambda_{rem}(j) < \lambda^{(k)} \right\}$$

$$\lambda_{rem}^* = \lambda_{rem}(j_{rem})$$

$$j_{dep} = \arg \max_{j \in I_P} \left\{ \lambda_{dep}(j) = \lambda^{(k)} + \frac{\hat{\beta}_j^{(k)} - \gamma \lambda^{(k)} \text{sign}(\hat{\beta}_j^{(k)})}{\gamma \text{sign}(\beta_j) - \omega_j} \mid 0 < \lambda_{dep}(j) < \lambda^{(k)} \right\}$$

$$\lambda_{dep}^* = \lambda_{dep}(j_{dep})$$

$$j_{rep} = \arg \max_{j \in I_N} \left\{ \lambda_{rep}(j) = \lambda^{(k)} + \frac{\hat{\beta}_j^{(k)} - \gamma \lambda^{(k)} \text{sign}(\hat{\beta}_j^{(k)})}{\gamma \text{sign}(\beta_j) - \omega_j} \mid 0 < \lambda_{rep}(j) < \lambda^{(k)} \right\}$$

$$\lambda_{rep}^* = \lambda_{rep}(j_{rep})$$

$$j_{candidates} = \{j_{add}\} \cup \{j_{rem}\} \cup \{j_{dep}\} \cup \{j_{rep}\}$$

$$\lambda_{candidates} = \{\lambda_{add}^*\} \cup \{\lambda_{rem}^*\} \cup \{\lambda_{dep}^*\} \cup \{\lambda_{rep}^*\}$$

$$\text{tmp} = \arg \max \lambda_{candidates}$$

$$\lambda^{(k+1)} = \lambda_{candidates}(\text{tmp})$$

$$j_next = j_{candidates}(\text{tmp})$$

$$\hat{\beta}_I^{(k+1)} = \hat{\beta}_I^{(k)} + (\lambda^{(k)} - \lambda^{(k+1)}) \omega$$

$$\alpha_0^{(k+1)} = (\lambda^{(k+1)})^{-1} (\lambda^{(k)} \alpha_0^{(k)} + (\lambda^{(k)} - \lambda^{(k+1)}) z)$$

$$k \leftarrow k + 1$$

end while

stochastic representation of spherical vectors, and representation as scale mixtures of centered Gaussians, which are also easy to generate. Other possibilities can be found in [Devroye 1986].

5.2.1 Through the stochastic representation

The first and more general way to generate a spherically symmetric random vector is to consider its stochastic representation. Indeed, we recall that, whatever the density of a spherical vector Y is, this vector can be written as

$$Y = RU \quad R = \|Y\| > 0, \quad U = Y/\|Y\| \sim \mathcal{U}_{S_1}, \quad R, U \text{ are independent} \quad (5.49)$$

where R is the radius of Y , U is its direction, and \mathcal{U}_{S_1} is the uniform distribution on the sphere S_1 of unit radius. Hence, if we can generate U , then any distribution with support $(0, \infty)$ can be used to generate the radius R and thus to generate a spherical random vector Y . Therefore our interest turns now towards the generation of uniform random vectors on spheres.

To do so, we propose to look at the expression of $U \in \mathbb{R}^n$ through its spherical coordinates:

$$\begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \dots \\ U_{n-1} \\ U_n \end{pmatrix} = \begin{pmatrix} \sin \theta_1 \sin \theta_2 \dots \sin \theta_{n-2} \sin \theta_{n-1} \\ \sin \theta_1 \sin \theta_2 \dots \sin \theta_{n-2} \cos \theta_{n-1} \\ \sin \theta_1 \sin \theta_2 \dots \cos \theta_{n-2} \\ \dots \\ \sin \theta_1 \cos \theta_2 \\ \cos \theta_1 \end{pmatrix}, \quad (5.50)$$

where $(\theta_1, \dots, \theta_{n-1}) \in (0, \pi)^{n-2} \times (0, 2\pi)$. According to [Fourdrinier et al. 2012], if U is uniformly distributed on the sphere S_1 , then the density of θ_i is proportional to $\sin^{n-i-1} \theta_i$ on $(0, \pi)$ for $1 \leq i \leq n-2$, and θ_{n-1} is uniformly distributed on the interval $(0, 2\pi)$. We propose to generate the angles θ_i , $1 \leq i \leq n-2$, thanks to an accept-reject method, described in Algorithm 5.3. Our generator for uniforms on the unit sphere can be easily derived from Algorithm 5.3 and is described in Algorithm 5.4.

Algorithm 5.3 randsin

Require: Power $p \in \mathbb{N}_*$

Ensure: Angle $\theta \in (0, \pi)$

```

test = false
while test ≠ true do
  Generate  $u$  and  $v$  as  $\mathcal{U}([0, 1])$ 
  Compute  $t = \sin^p(\pi v)$ .
  if  $u \leq t$  then
     $\theta \leftarrow t$ 
    test ← true
  end if
end while
```

Algorithm 5.4 randSphere**Require:** Length $n \in \mathbb{N}_*$ **Ensure:** Vector $U \in \mathbb{R}^n$ **for** $i = 1, \dots, n - 2$ **do** $\theta_i = \text{randsin}(n - i - 1)$. **end for** Generate θ_{n-1} as $\mathcal{U}([0, 2\pi])$. Compute U as in (5.50).

However, due to the accept-reject method in Algorithm 5.3 and to the large number of operations, the Algorithm 5.4 can be quite slow. If an efficient generator of Gaussian random variables is available, a faster way to generate U can be derived through the Gaussian distribution, as proposed by [Devroye 1986]. Indeed, the stochastic representation (5.49) is valid for any spherically symmetric distribution, in particular for the Gaussian distribution. Hence, if $Z \sim \mathcal{N}_n(0, \sigma^2 I_n)$, then $U = Z/\|Z\|$ is uniform on the sphere S_1 . Thus, the alternative algorithm for randSphere (5.4) can be easily derived as in Algorithm 5.5.

Algorithm 5.5 randSphereGauss**Require:** Length $n \in \mathbb{N}_*$ **Ensure:** Vector $U \in \mathbb{R}^n$ Generate n standard random variables $N_i \sim \mathcal{N}(0, 1)$. Compute $S = \sum_{i=1}^n N_i^2$. Compute $U = (N_1, \dots, N_n)/S$.

Figure 5.6 compares the repartition of vectors generated by both Algorithm 5.4 and Algorithm 5.5 for $n = 2$ and $n = 3$.

We now present a few examples of spherically symmetric random vectors Y that can be generated thanks to their stochastic representation, and we specify each time the distribution of their radius R . Before doing so, we state a result on the link between the density of Y and the density of its radius, given in [Kelker 1970].

Lemma 5.1 (Radial distribution). *Let $Y \sim \mathcal{S}_n(0)$ have a density of the form*

$$p(y) = g(\|y\|^2/\sigma^2).$$

Then, its radius $R = \|Y\|$ has density

$$h(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)} r^{n-1} g(r^2).$$

The function g is called the generating function.

Example 5.1. If Y is a spherical standard Gaussian random vector, $Y \sim \mathcal{N}_n(0, I_n)$, then it is well known that the square of its radius follows a Chi-squared distribution with n degrees of freedom: $R^2 \sim \chi^2(n)$ (see for instance [Fang et al. 1989]). If Y is a spherical Gaussian random vector with variance σ^2 , then its radius is $R = \|Y\|/\sigma$ and its square still follows a Chi-squared distribution.

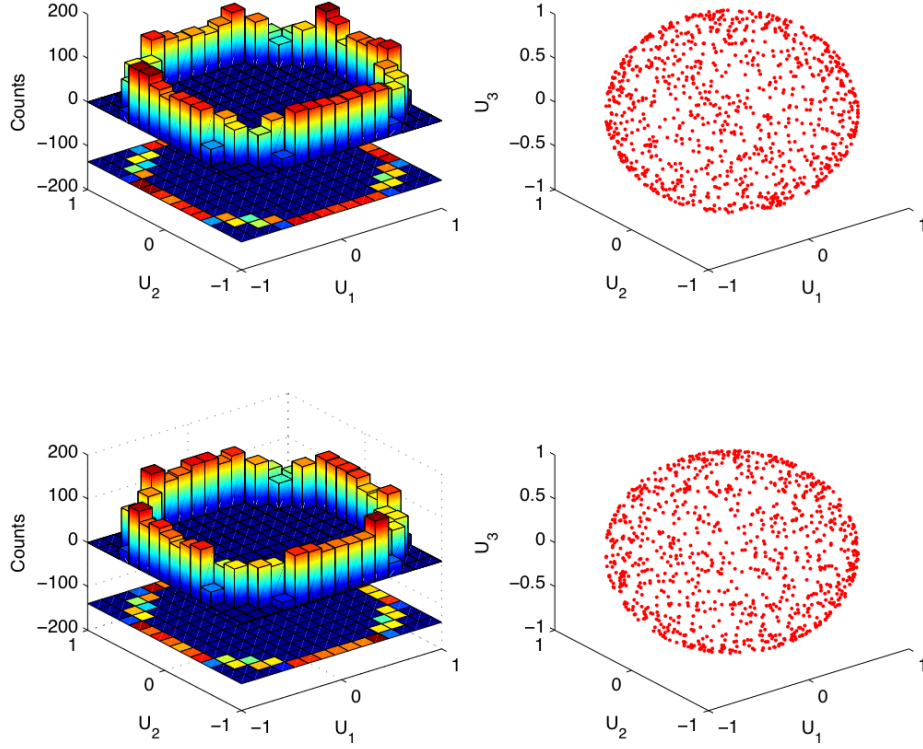


Figure 5.6: Histogram of a uniform vector on the sphere generated by Algorithm 5.4 (top) and Algorithm 5.5 (bottom) for $n = 2$ (left) and repartition for $n = 3$ (right).

Example 5.2. If Y is a Student random vector, $Y \sim \mathcal{T}_n(\nu)$, where ν is the degrees of freedom, then the square of its radius follows a Fisher distribution: $R^2 \sim n\text{Fisher}(n, \nu)$. Indeed, we recall that the density of Y is

$$p(y) = \frac{\Gamma\left(\frac{n+\nu}{2}\right)}{(\pi\sigma^2\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\|y\|^2}{\nu\sigma^2}\right)^{-\frac{(\nu+n)}{2}}.$$

Hence, from Lemma 5.1, we can compute the density of its radius

$$\begin{aligned} h(r) &= \frac{2\pi^{n/2}}{\Gamma(n/2)} r^{n-1} \times \frac{\Gamma\left(\frac{n+\nu}{2}\right)}{(\pi\sigma^2\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{r^2}{\nu\sigma^2}\right)^{-\frac{(\nu+n)}{2}} \\ &= \frac{2r^{n-1}}{(\sigma^2\nu)^{\frac{n}{2}}B\left(\frac{\nu}{2}, \frac{n}{2}\right)} \left(1 + \frac{r^2}{\nu\sigma^2}\right)^{-\frac{(\nu+n)}{2}}. \end{aligned}$$

Now, by the change of variable $t = r^2/n$, this leads to

$$\begin{aligned} h(t) &= \frac{2(nt)^{\frac{n-1}{2}}}{(\sigma^2\nu)^{\frac{n}{2}} B\left(\frac{\nu}{2}, \frac{n}{2}\right)} \left(1 + \frac{nt}{\nu\sigma^2}\right)^{-\frac{(\nu+n)}{2}} \times \frac{1}{2\sqrt{nt}} \\ &= \frac{1}{tB\left(\frac{\nu}{2}, \frac{n}{2}\right)} (nt)^{\frac{n}{2}} (\nu\sigma^2)^{-\frac{n}{2}} (\nu\sigma^2)^{\frac{\nu+n}{2}} (\nu\sigma^2 + nt)^{-\frac{(\nu+n)}{2}} \\ &= \frac{1}{tB\left(\frac{\nu}{2}, \frac{n}{2}\right)} (nt)^{\frac{n}{2}} (\nu\sigma^2)^{\frac{\nu}{2}} (\nu\sigma^2 + nt)^{-\frac{(\nu+n)}{2}}, \end{aligned}$$

which is exactly a Fisher distribution with parameters n and ν .

Example 5.3. If Y is a Kotz random vector, $Y \sim \mathcal{K}_n(N, r, \sigma^2)$, then [Fang *et al.* 1989] showed that the square of its radius follows a Gamma distribution: $R^2 \sim \text{Gamma}(N + n/2 - 1, \sigma^2/r)$. Indeed, the density of Y is

$$p(y) = \frac{\Gamma\left(\frac{n}{2}\right) r^{\frac{2N-2+n}{2}}}{\pi^{\frac{n}{2}} (\sigma^2)^{\frac{n+2N-2}{2}} \Gamma\left(\frac{2N-2+n}{2}\right)} \|y\|^{2(N-1)} e^{-r \frac{\|y\|^2}{\sigma^2}}.$$

Thanks to Lemma 5.1, the density of its radius is thus

$$\begin{aligned} h(R) &= \frac{2\pi^{n/2}}{\Gamma(n/2)} R^{n-1} \times \frac{\Gamma\left(\frac{n}{2}\right) r^{\frac{2N-2+n}{2}}}{\pi^{\frac{n}{2}} (\sigma^2)^{\frac{n+2N-2}{2}} \Gamma\left(\frac{2N-2+n}{2}\right)} R^{2(N-1)} e^{-r \frac{R^2}{\sigma^2}} \\ &= \frac{2r^{\frac{2N-2+n}{2}}}{(\sigma^2)^{\frac{n+2N-2}{2}} \Gamma\left(\frac{2N-2+n}{2}\right)} R^{n+2N-3} e^{-r \frac{R^2}{\sigma^2}}. \end{aligned}$$

With the change of variable $t = R^2$, we obtain

$$\begin{aligned} h(t) &= \frac{2r^{\frac{2N-2+n}{2}}}{(\sigma^2)^{\frac{n+2N-2}{2}} \Gamma\left(\frac{2N-2+n}{2}\right)} t^{\frac{n+2N-3}{2}} e^{-r \frac{t}{\sigma^2}} \frac{1}{2\sqrt{t}} \\ &= \frac{1}{\Gamma\left(\frac{2N-2+n}{2}\right)} \left(\frac{r}{\sigma^2}\right)^{\frac{2N-2+n}{2}} t^{\frac{n+2N-4}{2}} e^{-\frac{r}{\sigma^2} t}, \end{aligned}$$

which is the Gamma distribution with parameters $\frac{n+2N-2}{2}$ and $\frac{\sigma^2}{r}$.

Example 5.4. We now take the example of an exponential power random vector Y , with power b . Its density is

$$p(y) = \frac{n\Gamma\left(\frac{n}{2}\right)}{(\pi\sigma^2)^{\frac{n}{2}} 2^{1+\frac{n}{2b}} \Gamma\left(1 + \frac{n}{2b}\right)} \exp\left\{-\frac{1}{2} \left(\frac{\|y\|^2}{\sigma^2}\right)^b\right\},$$

whose radial density can be computed as follows

$$\begin{aligned} h(r) &= \frac{2\pi^{n/2}}{\Gamma(n/2)} r^{n-1} \times \frac{n\Gamma\left(\frac{n}{2}\right)}{(\pi\sigma^2)^{\frac{n}{2}} 2^{1+\frac{n}{2b}} \Gamma\left(1 + \frac{n}{2b}\right)} \exp\left\{-\frac{1}{2} \left(\frac{r^2}{\sigma^2}\right)^b\right\} \\ &= \frac{n}{(\sigma^2)^{\frac{n}{2}} 2^{\frac{n}{2b}} \Gamma\left(1 + \frac{n}{2b}\right)} r^{n-1} \exp\left\{-\frac{1}{2} \left(\frac{r^2}{\sigma^2}\right)^b\right\}. \end{aligned}$$

Again, by the change of variable $t = r^{2b}$, this leads to

$$\begin{aligned}
 h(t) &= \frac{n}{(\sigma^2)^{\frac{n}{2}} 2^{\frac{n}{2b}} \Gamma(1 + \frac{n}{2b})} t^{\frac{n-1}{2b}} \exp\left\{-\frac{t}{2\sigma^{2b}}\right\} \times \frac{t^{\frac{1}{2b}-1}}{2b} \\
 &= \frac{n}{b(\sigma^2)^{\frac{n}{2}} 2^{1+\frac{n}{2b}} \Gamma(1 + \frac{n}{2b})} t^{\frac{n}{2b}-1} \exp\left\{-\frac{t}{2\sigma^{2b}}\right\} \\
 &= \frac{n}{b(\sigma^2)^{\frac{n}{2}} 2^{1+\frac{n}{2b}} \frac{n}{2b} \Gamma(\frac{n}{2b})} t^{\frac{n}{2b}-1} \exp\left\{-\frac{t}{2\sigma^{2b}}\right\} \\
 &= \frac{1}{(\sigma^2)^{\frac{n}{2}} 2^{\frac{n}{2b}} \Gamma(\frac{n}{2b})} t^{\frac{n}{2b}-1} \exp\left\{-\frac{t}{2\sigma^{2b}}\right\},
 \end{aligned}$$

where the last two equations derive from the property of the gamma function, $\Gamma(z+1) = z\Gamma(z)$. We obtain once again a Gamma distribution, but this time for the squared radius to the power b : $R^{2b} \sim \text{Gamma}\left(\frac{n}{2b}, 2\sigma^{2b}\right)$.

Name	Density of Y	Density of radius $R = \ Y\ $
Gaussian	$Y \sim \mathcal{N}_n(0, \sigma^2 I_n)$	$R^2 \sim \chi^2(n)$
Student	$Y \sim \mathcal{T}_n(\nu)$	$\frac{R^2}{n} \sim \text{Fisher}(n, \nu)$
Kotz	$Y \sim \mathcal{K}_n(N, r, \sigma^2)$	$R^2 \sim \text{Gamma}\left(N + \frac{n}{2} + 1, \frac{r}{\sigma^2}\right)$
Exponential Power	$Y \sim \mathcal{EP}_n(b, \sigma^2)$	$R^{2b} \sim \text{Gamma}\left(\frac{n}{2b}, 2\sigma^{2b}\right)$

Table 5.1: Distribution of the radius for several spherical laws.

Remark 5.3. The examples cited here are only the most known distributions. But we can easily derive other spherically symmetric distributions by taking other densities for the radius. There exist indeed other continuous densities with support $(0, \infty)$ which have already been studied for random generation. For instance, we can think of the Weibull density or the Lévy distribution.

5.2.2 Through mixtures of other spherical distributions

Another way to generate spherically symmetric distributions is through the property that any scale mixture of spherically symmetric distributions is also a spherically symmetric distribution. This property was already used in the previous section since the stochastic representation is a scale mixture of uniforms on the unit sphere, which are spherically symmetric. Here we extend the principle to other scale mixtures, like for instance Gaussian mixtures. Indeed, several distributions can be seen as scale Gaussian mixtures, such as the Student distribution and the Bessel distribution, as we will see next. The general principle is the same as in the previous subsection:

1. Generate a vector $X \in \mathbb{R}^n$ from the chosen spherical density.
2. Generate the scale random variable V according to the density corresponding to the desired mixture.

3. Compute the resulting random vector $Y = \sqrt{V}X$.

We recall that a scale Gaussian mixture is written as

$$p(y) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int_0^\infty \frac{1}{v^{\frac{n}{2}}} e^{-\frac{\|y\|^2}{2v}} g(v) dv \quad (5.51)$$

Example 5.5. The Student distribution with ν degrees of freedom is a mixture of Gaussian with mixing density an inverse Gamma distribution with parameters $\nu/2$ and $\nu/2$. Indeed, the density of an inverse Gamma distribution with parameters α and β is

$$g(v) = \frac{\beta^\alpha}{\Gamma(\alpha)} v^{-\alpha-1} e^{-\frac{\beta}{v}}.$$

Hence, the mixture density (5.51) becomes

$$\begin{aligned} p(y) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \int_0^\infty \frac{1}{v^{\frac{n}{2}}} e^{-\frac{\|y\|^2}{2v}} \times \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \Gamma^{-1}\left(\frac{\nu}{2}\right) v^{-\frac{\nu}{2}-1} e^{-\frac{\nu}{2v}} dv \\ &= \frac{\nu^{\frac{\nu}{2}}}{2^{\frac{n+\nu}{2}} \pi^{\frac{n}{2}} \Gamma\left(\frac{\nu}{2}\right)} \int_0^\infty \frac{1}{v^{\frac{n}{2}+\frac{\nu}{2}+1}} e^{-\left(\frac{\|y\|^2 + \nu}{2v}\right)} dv. \end{aligned}$$

We recognize in the integral an inverse Gamma distribution with parameters $\alpha = (n + \nu)/2$ and $\beta = (\|y\|^2 + \nu)/2$, so that we have

$$\int_0^\infty \frac{1}{v^{\frac{n}{2}+\frac{\nu}{2}+1}} e^{-\left(\frac{\|y\|^2 + \nu}{2v}\right)} dv = \left(\frac{\|y\|^2 + \nu}{2}\right)^{-\frac{n+\nu}{2}} \Gamma\left(\frac{n + \nu}{2}\right).$$

Hence, we obtain

$$\begin{aligned} p(y) &= \frac{\nu^{\frac{\nu}{2}}}{2^{\frac{n+\nu}{2}} \pi^{\frac{n}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\|y\|^2 + \nu}{2}\right)^{-\frac{n+\nu}{2}} \Gamma\left(\frac{n + \nu}{2}\right) \\ &= \frac{\Gamma\left(\frac{n+\nu}{2}\right)}{(\pi\nu)^{\frac{n}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\|y\|^2}{\nu} + 1\right)^{-\frac{n+\nu}{2}}, \end{aligned}$$

which is the Student distribution with parameter ν .

Example 5.6. The last example we are going to treat in this section is the one of the multivariate Bessel distribution. We recall that, if $Y \sim \mathcal{B}_n(q, r, \sigma^2)$, its density is of the form

$$p(y) = \frac{1}{2^{q+n-1} \pi^{\frac{n}{2}} r^{n+q} \Gamma\left(q + \frac{n}{2}\right)} \|y\|^q K_q\left(\frac{\sqrt{\|y\|^2}}{r}\right),$$

where $K_q(z)$ is the modified Bessel function of the third kind defined, for $|\arg z| < \pi$, by

$$K_q(z) = \frac{\pi}{2 \sin(q\pi)} (I_{-q}(z) - I_q(z))$$

with

$$I_q(z) = \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(k + q + 1)} \left(\frac{z}{2}\right)^{q+2k}.$$

This distribution is a scale mixture of Gaussian laws with mixing density a Gamma distribution with parameters $q + n/2$ and $1/(2r^2)$. Indeed, such a mixing distribution has density

$$g(v) = \left((2r^2)^{q+\frac{n}{2}} \Gamma\left(q + \frac{n}{2}\right) \right)^{-1} v^{q+\frac{n}{2}-1} e^{-\frac{v}{2r^2}}.$$

The normal mixture thus becomes

$$\begin{aligned} p(y) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \int_0^\infty \frac{1}{v^{\frac{n}{2}}} e^{-\frac{\|y\|^2}{2v}} \times \left((2r^2)^{q+\frac{n}{2}} \Gamma\left(q + \frac{n}{2}\right) \right)^{-1} v^{q+\frac{n}{2}-1} e^{-\frac{v}{2r^2}} dv \\ &= \frac{1}{2^{q+n} \pi^{\frac{n}{2}} (r^2)^{q+\frac{n}{2}} \Gamma\left(q + \frac{n}{2}\right)} \int_0^\infty v^{q-1} e^{-\frac{v}{2r^2} - \frac{\|y\|^2}{2v}} dv. \end{aligned}$$

We recognized in the integral a generalized inverse Gaussian distribution with parameters $1/r^2$, $\|y\|^2$ and q , which has density

$$g(v) = \frac{1}{2 (r\|y\|)^q K_q\left(\frac{\|y\|}{r}\right)} v^{q-1} e^{-\frac{v}{2r^2} - \frac{\|y\|^2}{2v}}.$$

Hence, the integral can be easily computed since

$$\int_0^\infty g(v) dv = 1,$$

so that

$$\int_0^\infty v^{q-1} e^{-\frac{v}{2r^2} - \frac{\|y\|^2}{2v}} dv = 2 (r\|y\|)^q K_q\left(\frac{\|y\|}{r}\right).$$

Finally, we obtain

$$\begin{aligned} p(y) &= \frac{1}{2^{q+n} \pi^{\frac{n}{2}} (r^2)^{q+\frac{n}{2}} \Gamma\left(q + \frac{n}{2}\right)} \times 2 (r\|y\|)^q K_q\left(\frac{\|y\|}{r}\right) \\ &= \frac{1}{2^{q+n-1} \pi^{\frac{n}{2}} r^{q+n} \Gamma\left(q + \frac{n}{2}\right)} \|y\|^q K_q\left(\frac{\|y\|}{r}\right), \end{aligned}$$

which we recognize as the multivariate Bessel distribution.

Note that the Laplace distribution is a special case of multivariate Bessel distributions with parameters $q = 0$ and $r = \sigma/\sqrt{2}$. Hence the Laplace distribution is a scale mixture of Gaussian laws with mixing density $V \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{\sigma^2}\right)$.

Table 5.2 summarizes these examples, which are based on [Andrews & Mallows 1974], [Fang et al. 1989], [Feller 1966], [West 1987].

Name	Density of Y	Density of scale V
Student	$Y \sim \mathcal{T}_n(\nu)$	$\frac{1}{V} \sim \mathcal{Gamma}\left(\frac{\nu}{2}, \frac{2}{\nu}\right)$
Laplace	$Y \sim \mathcal{L}_n(b)$	$V \sim \mathcal{Gamma}\left(\frac{n}{2}, \frac{1}{\sigma^2}\right)$
Bessel	$Y \sim \mathcal{B}_n(q, r)$	$V \sim \mathcal{Gamma}\left(q + \frac{n}{2}, \frac{1}{2r^2}\right)$
Exponential Power	$Y \sim \mathcal{EP}_n(b, \sigma^2)$	$V \sim \mathcal{Stable\ law}\left(\frac{\alpha}{2}, 1, \gamma, 0\right)$
Logistic	$Y \sim \mathcal{Log}_n(\sigma^2)$	$\frac{\sqrt{V}}{2} \sim \mathcal{Kolmogorov}$

Table 5.2: Distribution of the scale for Gaussian mixtures.

Numerical study

Contents

6.1	How good is the oracle?	127
6.1.1	Purpose of the study	127
6.1.2	Sparse regularization paths versus stepwise methods	128
6.1.3	Replacing by other estimators	134
6.1.4	Discussion on the first study	136
6.2	Comparison of model evaluation criteria	138
6.2.1	Purpose of the study	138
6.2.2	Unbiased loss estimator vs corrected loss estimator	139
6.2.3	Comparison to existing methods from literature	147
6.2.4	Discussion on the second study	147

This chapter presents numerical results on model selection problems. There exist many empirical studies in the literature. But, due to the complexity of the model selection issues, they generally focus on some specific aspects. For instance, algorithmic studies aim at demonstrating the superiority of a given algorithm for a chosen selection procedure (see for instance [El Anbari 2011]). From the model selection point of view, the empirical studies generally aim at demonstrating the superiority of a given criterion for a chosen algorithm (see for instance [Baraud *et al.* 2009]). In this work, we try to discuss both aspects under different noise distributions.

This chapter is divided into two main studies. The first study intends to determine the adequacy of methods used in the construction of collections of model with the objective of good prediction through the estimation loss (or prediction risk), before considering the actual problem of estimating the estimation loss based on data in the second study, where we compare the criteria we developed based on the theory of loss estimation to the existing methods from literature presented in Chapter 2.

6.1 How good is the oracle?

6.1.1 Purpose of the study

This first study compares the different methods presented in Chapter 2, Section 2.3 used for the construction of collections of models and their associated estimator, namely sparse regularization methods and stagewise methods. Before considering the problem of the estimation of the estimation loss or the prediction risk in the following study, we consider the problem of the adequacy of the resulting collections of estimations with an objective of good prediction based

on the true estimation loss (or prediction risk). In order to do so, we will use our knowledge of the true underlying system in simulated data to compute the true estimation loss on the collection of estimations $\{\hat{\beta}_1, \dots, \hat{\beta}_M\}$ and select the estimation having lower estimation loss. The question this simulation study tries to answer is thus: **if we knew the true loss, how good would be the best estimation from the collection $\{\hat{\beta}_1, \dots, \hat{\beta}_M\}$** , meaning how shall we explore the submodels and, for each submodel, how shall we estimate the regression coefficient?

This question will be answered in various settings: first, when the design matrix X is orthogonal, in which case the paths are the same for all the methods and the difference lies only in the estimation; next, when the design matrix X is general so that both the paths and the estimations can be different; and finally, when we ultimately replace their respective estimations by the restricted least-squares solution, hereby comparing the paths.

6.1.2 Sparse regularization paths versus stepwise methods

Protocol

We propose the following example, based on [Fourdrinier & Wells 1994]. The regression coefficient β is set to $(2; 0; 0; 4; 0)^t$. The design matrix X is either orthogonal, or general with correlation matrix Σ . The non diagonal elements of Σ are uniformly drawn between 0 and ρ , where ρ is taken in the set $\{0; 0.2; 0.4; 0.5; 0.6; 0.8\}$. Since the theoretical study of previous chapters was under the fixed design assumption, we thus generate X once and fix it for the rest of the experiment. We then draw $R = 5000$ replicates of the noise vector ε from a Gaussian distribution with covariance matrix $\sigma^2 I_n$, where σ^2 is taken in $\{0.2; 0.4; \dots; 3.8; 4\}$. We also make the number n of observations vary in $\{20; 40; \dots; 80; 100\}$. Finally, we run the experiment 10 times, leading to 10 different examples for X .

For each example and each replicate, we run the algorithms corresponding to the methods described in Chapter 2, Section 2.3, recalled in Table 6.1. No stopping criterion is used, so that the entire regularization path is built, starting with the null model with no variable and ending with the full model (except for Backward Selection which produces the reverse path). Also, we fix the hyperparameters that do not tune the sparsity. Indeed, these extra hyperparameters generally tune the bias. Hence, it could be interesting to set them to a value leading to the lower bias, often leading to an estimator close to the restricted Least-squares estimator. However, it is not always a good choice in view of a selection based on the generalized degrees of freedom \widehat{df} as a measure of complexity, since in some cases low bias also leads to high value of \widehat{df} (see Figure 6.1). Therefore, we considered the values suggested by the authors of each method.

Then, we select the the model minimizing the true loss in the collection $\{\hat{\beta}_1, \dots, \hat{\beta}_M\}$, that is,

$$\hat{\beta}_{m^*} = \arg \min_{\hat{\beta} \in \{\hat{\beta}_1, \dots, \hat{\beta}_M\}} \|X\hat{\beta} - X\beta\|^2.$$

Note that $\hat{\beta}_{m^*}$ corresponds to the oracle.

Finally, we compute the following measures of quality of the selection: the frequency of selection of the true subset (freq), the average number of nonzero coefficients (\bar{k}), and the average F-score ($\overline{\text{F-score}}$), the F-score being defined for each replicate as a combination of the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The terms *positive/negative* relate to the nonzero/zero components in the estimated coefficient $\hat{\beta}$, while *true/false* relate to the correctly/incorrectly classified when compared to the true regression

Name	Method	Extra hyperparameters
lasso	Lasso	—
firm	MCP / Firm Shrinkage	$\gamma = 2$
adalasso	Adaptive Lasso	$w = (\hat{\beta}_j^{LS})^{-2}$
garrote	Garrote	$w = (\hat{\beta}_j^{LS})^{-1}$
enet	Elastic-Net	$\lambda_2 = 0.3$
adanet	Adaptive Elastic-Net	$w = (\hat{\beta}_j^{LS})^{-2}, \lambda_2 = 0.3$
scad	SCAD ^a	$a = 3.7$
forward	Forward Selection	—
backward	Backward Elimination	—

^aonly for the orthogonal design case

Table 6.1: Methods for constructing collections of models.

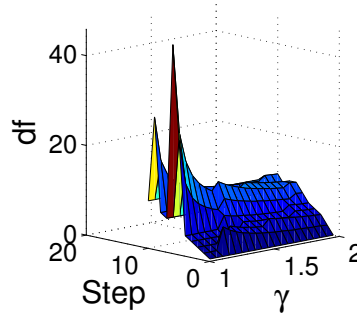


Figure 6.1: The necessity of trading off bias and generalized degrees of freedom. This graph results from the Firm Shrinkage (orthogonal design), where the hyperparameter γ has been taken between 1 and 2.

coefficient β). The corresponding formulas are

$$\text{freq}(\hat{\beta}) = \frac{1}{R} \sum_{r=1}^R \mathbb{1}_{\{\text{sgn}(\hat{\beta}^{(r)}) = \text{sgn}(\beta)\}} \quad (6.1)$$

$$\bar{k}(\hat{\beta}) = \frac{1}{R} \sum_{r=1}^R \sum_{j=1}^p \mathbb{1}_{\{\text{sgn}(\hat{\beta}_j^{(r)}) \neq 0\}} \quad (6.2)$$

$$\begin{aligned} \overline{\text{F-score}}(\hat{\beta}) &= \frac{1}{R} \sum_{r=1}^R \frac{2 \text{TP}(\hat{\beta}^{(r)})}{2 \text{TP}(\hat{\beta}^{(r)}) + \text{FP}(\hat{\beta}^{(r)}) + \text{FN}(\hat{\beta}^{(r)})} \\ &= \frac{1}{R} \sum_{r=1}^R \frac{2 \text{TP}(\hat{\beta}^{(r)})}{p + \text{TP}(\hat{\beta}^{(r)}) - \text{TN}(\hat{\beta}^{(r)})}, \end{aligned} \quad (6.3)$$

where

$$\begin{aligned} \text{TP}(\hat{\beta}) &= \# \{ \hat{\beta}_j \neq 0 \mid \beta_j \neq 0 \}, & \text{TN}(\hat{\beta}) &= \# \{ \hat{\beta}_j = 0 \mid \beta_j = 0 \}, \\ \text{FP}(\hat{\beta}) &= \# \{ \hat{\beta}_j \neq 0 \mid \beta_j = 0 \}, & \text{FN}(\hat{\beta}) &= \# \{ \hat{\beta}_j = 0 \mid \beta_j \neq 0 \}. \end{aligned}$$

Note that the F-score is equal to 0 if all the true nonzero coefficients of β have been incorrectly estimated to be zero ($TP = 0$), while it is equal to 1 when they have been correctly classified as nonzero ($TP = \#\{\beta_j \neq 0\}$) and the true zero coefficients in β have been also correctly classified as null, in which case we have $FP = FN = 0$. Hence, a good selection is represented by a frequency of good selection/recovery and an average F-score both close to 1, while the average number of estimated nonzero coefficients in $\hat{\beta}$ should be close to the true number of nonzero coefficients in β , which we denote k^* (in our small example, $k^* = 2$).

Orthogonal design case

In this paragraph, we compare the methods from Table 6.1 when X is an orthogonal design matrix, *i.e.* when we have $X^t X = I_p$.

Same path, different estimations. A special feature of the orthogonal design case is that the regularization paths provided by all the methods are exactly the same. Indeed, it can be easily verified for the sparse regularization methods since, in that case, they correspond to thresholding methods of the form

$$\hat{\beta}_j = s((X^j)^t Y; \lambda) \mathbf{1}_{\{|(X^j)^t Y| > \lambda\}},$$

where $s(\cdot; \lambda)$ is a shrinkage function depending on the method. For instance, the shrinkage function for Lasso is $s(t; \lambda) = t - \lambda \operatorname{sgn}(t)$. In order to perform the paths so that the variables are added to the selection one at a time, a convenient choice for the sequence of hyperparameters $(\lambda_m)_{m=1}^M$ is

$$\lambda_m = \begin{cases} |(X^{j_m})^t Y|, & \text{if } 1 \leq m \leq M-1, \\ 0, & \text{if } m = M, \end{cases} \quad (6.4)$$

where the index sequence (j_1, \dots, j_{M-1}) is a reordering of $(1, \dots, p)$ such that

$$|(X^{j_1})^t Y| \geq \dots \geq |(X^{j_{M-1}})^t Y|.$$

Since this sequence of hyperparameters does not depend at all on the shrinkage function $s(\cdot; \lambda)$, we can easily notice that the only difference results in the amount of shrinkage on each coefficient $\hat{\beta}_j$ and that the corresponding path of selected variables is (I_1, \dots, I_M) with

$$I_1 = \emptyset \quad \text{and} \quad I_m = \{j_1, \dots, j_{m-1}\}, \quad 2 \leq m \leq M. \quad (6.5)$$

It is also straightforward but a little more tricky to demonstrate that stagewise methods gives the same paths as sparse regularization methods. If we take for instance Forward Selection, we recall that the criterion to maximize for adding the next variable into the selection is, for all $j \in \{1, \dots, p\} \setminus I_{m-1}$,

$$\begin{aligned} \Delta MSE(j) &= \left| \|Y - X_I \hat{\beta}_I^{LS}\|^2 - \|Y - X_{I \cup \{j\}} \hat{\beta}_{I \cup \{j\}}^{LS}\|^2 \right| \\ &= \left| \|Y - X_I X_I^t Y\|^2 - \|Y - X_{I \cup \{j\}} X_{I \cup \{j\}}^t Y\|^2 \right|, \end{aligned}$$

since $\hat{\beta}_I^{LS} = X_I^t Y$ when X is orthogonal. Re-expressing both terms yields

$$\begin{aligned} \Delta MSE(j) &= \left| \|(I_n - X_I X_I^t) Y\|^2 - \|(I_n - X_{I \cup \{j\}} X_{I \cup \{j\}}^t) Y\|^2 \right| \\ &= \left| Y^t \left\{ (I_n - X_I X_I^t)(I_n - X_I X_I^t) - (I_n - X_{I \cup \{j\}} X_{I \cup \{j\}}^t)(I_n - X_{I \cup \{j\}} X_{I \cup \{j\}}^t) \right\} Y \right| \\ &= \left| Y^t \left\{ I_n - X_I X_I^t - I_n + X_{I \cup \{j\}} X_{I \cup \{j\}}^t \right\} Y \right|, \end{aligned}$$

the latter equality resulting from the orthogonality of X , and thus of X_I and $X_{I \cup \{j\}}$. Finally, noticing that

$$X_{I \cup \{j\}} X_{I \cup \{j\}}^t = \begin{pmatrix} X_I & X^j \end{pmatrix} \begin{pmatrix} X_I^t \\ (X^j)^t \end{pmatrix} = X_I X_I^t + X^j (X^j)^t,$$

we obtain

$$\Delta MSE(j) = |Y^t \mathbf{x}^j (X^j)^t Y| = \left((X^j)^t Y \right)^2.$$

This last equality amounts to reorder the variables in X so that

$$((X^{j_1})^t Y)^2 \geq \dots \geq ((X^{j_{M-1}})^t Y)^2.$$

Since the square function is monotonically increasing on $[0; \infty)$, the corresponding path of selected variables is the same as in (6.5). Hence, Forward Selection is equivalent to the Hard Thresholding rule

$$\hat{\beta}_j^{HT} = (X^j)^t Y \mathbf{1}_{\{|(X^j)^t Y| > \lambda\}},$$

with λ taken in (6.4). The same demonstration can be easily applied to Backward Elimination, where the only difference is that the selection is performed on the other side, from the full model to the null model.

Comparison of the estimations. The interest in having the same path for all the methods being tested is that it leads to a better understanding of the adequacy between the estimation itself (without caring for the order of the selection) and the estimation loss, since it is the theoretical criterion we chose here for selecting the best predictive model.

Figures 6.2 displays the empirical probability of selecting the true subset, the average F-score and the average number of non-zero coefficients, computed for all nine methods and for all values of n . The evolution of the measures of quality is shown with respect to the Signal-to-Noise Ratio (SNR) defined by

$$SNR = \frac{\min_{j \in \{1, \dots, p\}} |\beta_j|}{\sigma} = \frac{|\beta_{\min}|}{\sigma}.$$

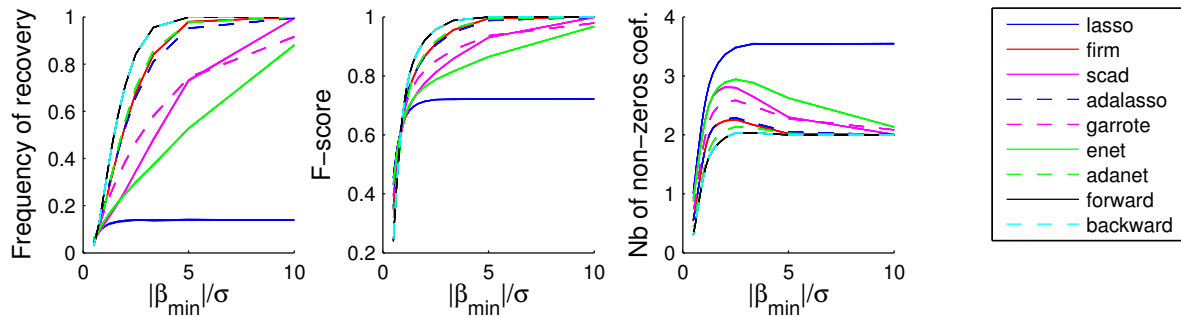


Figure 6.2: Frequency of selection (left panel), average F-score (center panel) and average number of non-zero coefficients (right panel) with respect to signal-to-noise ratio for the orthogonal design case. The curves for different sample sizes n (varying from 20 to 100) are almost perfectly superimposed.

The first striking feature of Figure 6.2 is that the number n of observations does not influence any of the measures of quality since all the curves for the different values of n are almost perfectly

superimposed. This fact was proved for the Lasso by [Leng *et al.* 2006] in the Theorem 4.1 therein, which we recall here:

Theorem 6.1. *When the true coefficient vector is $\beta = (\beta_1, \dots, \beta_{k^*}, 0, \dots, 0)^t$ with $p - k^* > 0$ zero coefficients and $X^t X = I_p$, if the Lasso is tuned according to prediction accuracy, then it selects the right model with a probability less than a constant $C < 1$, where C depends only on σ^2 and k^* , and not on the sample size n .*

According to our results, Theorem 6.1 seems to be extendable to all 9 methods, not just the Lasso.

Now, turning to the comparison between the estimators, the best methods for estimating β are those for which the probability of selecting the true subset goes faster to 1 as σ decreases and the SNR increases, and the same for the average F-score. From Figure 6.2, it can be seen that the best estimators are the Least-Squares Estimator (corresponding to forward and backward on the graphs), the Firm shrinkage, the Adaptive Lasso and the Adaptive Elastic Net, which are nearly unbiased estimators. The Least-Squares Estimator gives even better performances than the other three. On the contrary, using estimators with larger bias, such as Lasso, Elastic-net, Garrote and SCAD, with minimum true estimation loss, results in the selection of too many variables, as displayed in the right panel of Figure 6.2. Indeed, the true number of coefficient is $k^* = 2$ in our example, while the Lasso selects an average of 3.5 coefficients (out of 5!) as soon as the signal-to-noise ratio is greater than 4 (corresponding to a standard deviation lower than 0.5). This average number of nonzero coefficients seems to reach a plateau as the SNR increases. On the other hand, the curves for Garrote, Elastic Net and SCAD eventually tend to that of the Least-squares for high SNR, hence those estimators yields a better selection than Lasso. In view of a first conclusion that a large bias seems to be incompatible with a good selection, it is surprising that Elastic-Net performs better than Lasso since it has an even greater bias, but according to its developers [Zou & Hastie 2005] the ℓ_2 -regularization yields a better selection, a feature our results seem to agree with.

Another important remark about all three graphs is that the best performances are obtained for a signal-to-noise ratio greater than 2, which in our case correspond to a value of 1 for the standard deviation. Even then, the true subset recovery is only possible with probability of around 75%. The probability is (almost) 1 only when $\sigma \leq 0.5$ for the least-squares, and when $\sigma \leq 0.2$ for the Firm Shrinkage, the Adaptive Lasso, and the Adaptive Elastic Net.

General design with low correlation

In this paragraph, we perform the same study for the case where X is a general matrix with maximum correlation $\rho = 0.4$ between each pair of variables (X^i, X^j) . Figures 6.3 and 6.4 display the average frequency of selecting the true subset and the average F-score respectively, with respect to the signal-to-noise ratio and to the sample size n for the 8 methods. Since the general shape of the graph displaying the number of nonzero component is pretty similar to that in Figure 6.2, we leave this measure of quality and concentrate on the other two. Also, we did not run SCAD here because we did not implement it for the nonorthogonal design case.

First, note that unlike the previous case the sample size n plays an important role in the quality of the selection, although maybe not as important as the signal-to-noise ratio. Second, when X is general, the 8 methods might not share the same regularization path, unlike in the orthogonal design case, so that we compare both the paths and the estimators.

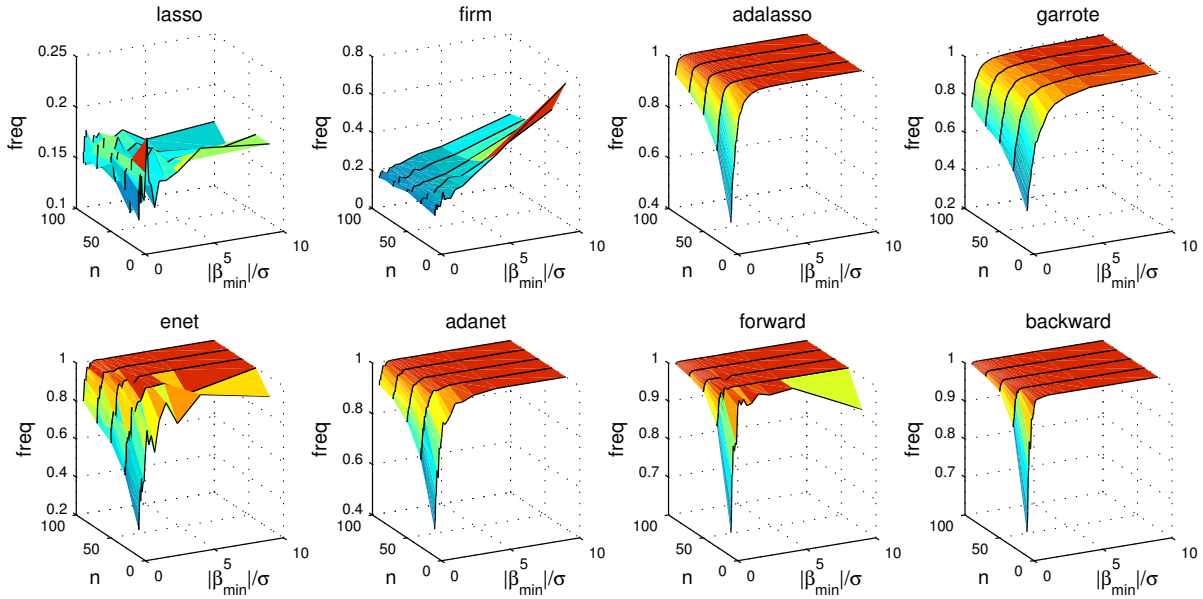


Figure 6.3: Average frequency of selection (freq) with respect to sample size (n) and signal-to-noise ratio ($|\beta_{\min}|/\sigma$) for the general design case with maximum correlation $\rho = 0.4$ between the variables.

The general shape of the curves is actually a bit different from the orthogonal case. Indeed, it can be seen on Figures 6.3 and 6.4 that most methods have their frequency of recovery and F-score going much faster to 1, even the biased Elastic-Net and Garrote. On the contrary, the Lasso still have the poorest performances with a average frequency of recovery ranging from 10% to 20%, and increasing the sample size n does not seem to improve such a bad score.

The most surprising result, compared to the previous ones, is that of MCP (firm). Indeed, this method had a behaviour similar to that of Adaptive Lasso and Adaptive Elastic-Net in the orthogonal design case. For the general case however, its performances have been deteriorated quite a lot. Looking more closely to the results, it appears that, for 2 examples of the matrix X out of the 10 we generated, MCP fails to select one of the two relevant variables at the first step, and is then never able to recover exactly the true subset in the rest of the path. Since the first step for MCP is exactly the same as the first step for Lasso and SCAD, it is clear that the same phenomenon occurs also for the other two methods.

General design with high correlation

In this paragraph, we increased the maximum correlation between two variables in X to $\rho = 0.8$. Figure 6.5 displays the average frequency of selecting the true subset with respect to the signal-to-noise ratio and to the sample size n for all 8 methods. We do not show the average F-score here since the shape of the graphs is pretty similar to that of the average frequency of recovery, but on a different scale.

Here, the discrepancies between the curves for varying sample size n and varying signal-to-noise ratio are much more important than in the previous case. In particular, the worst performances correspond to the case where there are $n = 20$ observations, which is still large compared to the number $p = 5$ of variables in X . Nevertheless, we can easily notice that

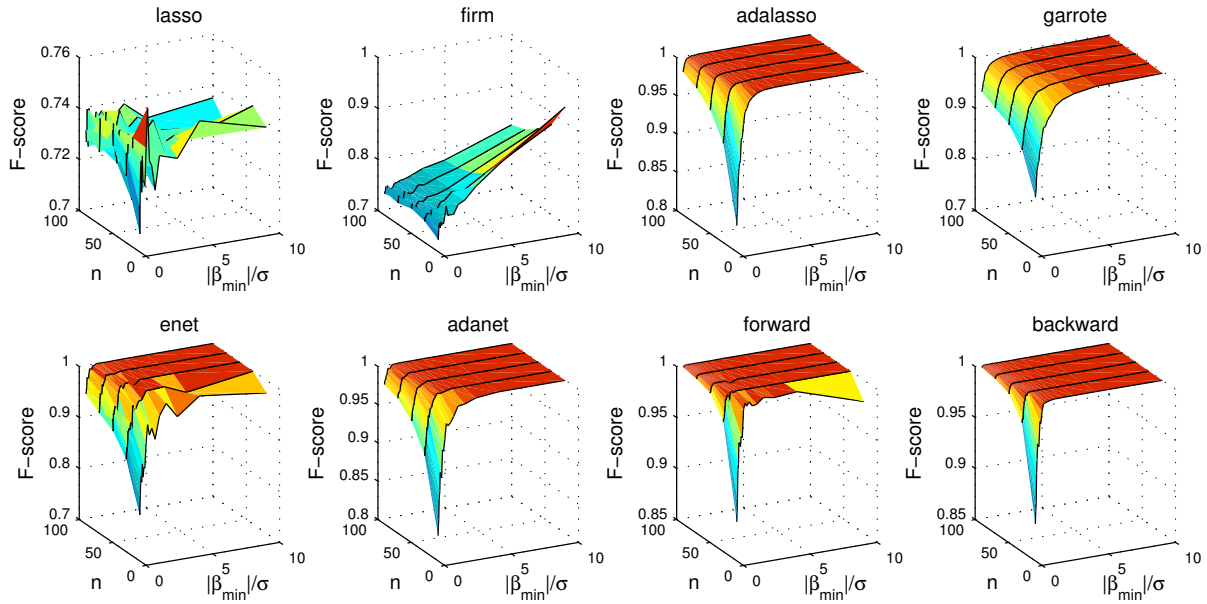


Figure 6.4: Average F-score of selection (F-score) with respect to sample size (n) and signal-to-noise ratio ($|\beta_{\min}|/\sigma$) for the general design case with maximum correlation $\rho = 0.4$ between the variables.

stepwise methods, Adaptive Lasso and Adaptive Elastic-Net still perform quite well, with an average frequency of recovery pretty close to 1 as soon as the signal-to-noise ratio is large enough. A caveat to this remark is that Forward Selection seems to be more unstable here than Backward Elimination and Adaptive Lasso. Finally, the performances of Elastic-Net are also quite unstable and have been deteriorated when compared to the case where the maximum correlation between variables is low. This could only be an artifact of the fact that we fixed the second hyperparameter, while it might be much better if we optimized it somehow.

Conclusions on the first part of the study

From this first part of the study, we can conclude that the less biased the estimation is, the better the selection is obtained based on minimizing the actual estimation loss. It thus makes sense to use this theoretical criterion as a baseline for the selection when using methods such as Forward Selection, Backward Elimination, Adaptive Lasso and Adaptive Elastic-Net. On the contrary, it is not a good theoretical criterion for selecting models when using the Lasso, as the results have shown in all the setups we have tried. This conclusion was also drawn from [Leng *et al.* 2006]. We would like to make it clear however that this does not mean that Lasso is a poor method, but rather that special care should be taken in general to the adequacy between methods for constructing collections of models and estimators with theoretical criteria such as the estimation loss. Since the results are already bad if we knew the actual estimation loss, we cannot expect them to be good when we replace it by an estimator of the estimation loss.

6.1.3 Replacing by other estimators

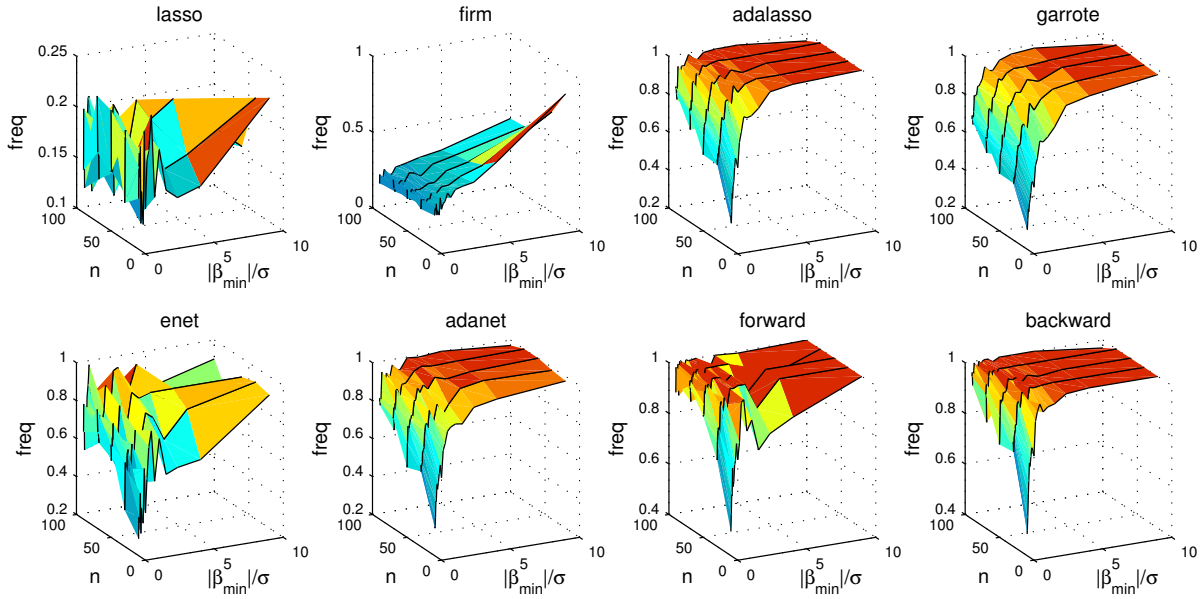


Figure 6.5: Average frequency of selection (freq) with respect to sample size (n) and signal-to-noise ratio ($||\beta_{\min}||/\sigma$) for the general design case with maximum correlation $\rho = 0.8$ between the variables.

The least-squares estimator. In view of the conclusions from the previous part, we decided to run the experiment as in the previous subsection with the following additional step: after each method has been computed with its entire regularization path, the estimation for each step of the path is replaced by the less biased estimator of all: the restricted least-squares estimator. The selection of the best estimation among the collection is still performed by minimizing the true estimation loss. This experiment aims at comparing the regularization paths proposed by each method. Figure 6.6 displays the average frequency of selecting the true subset with respect to the signal-to-noise ratio and to the sample size n for the 8 methods with the least-squares estimator. This figure clearly shows better performances than when we take the estimator corresponding to each method (except Forward Selection and Backward Elimination which already estimates the parameter β by least-squares), with an average recovery of the true subset close to 100% in most cases. This is especially the case for Lasso showing good performances of selection for the first time of the study.

We can also note that the graphs are exactly the same for Lasso and Elastic Net (enet). Since the estimation is the same here, this could mean that both regularization paths share similarities, although they might not be exactly equal. The same remark can be done for Adaptive Lasso, Adaptive Elastic Net and Garrote. There is also a big similarity between the latter ones and Backward Elimination. However, the regularization path of Lasso is clearly not the same than that of Adaptive Lasso and to that of MCP (firm), for instance, since the graphs significantly differ.

Finally, it appears clearly that the best results are obtained for Adaptive Lasso, Adaptive Elastic Net, Garrote, and Backward Elimination.

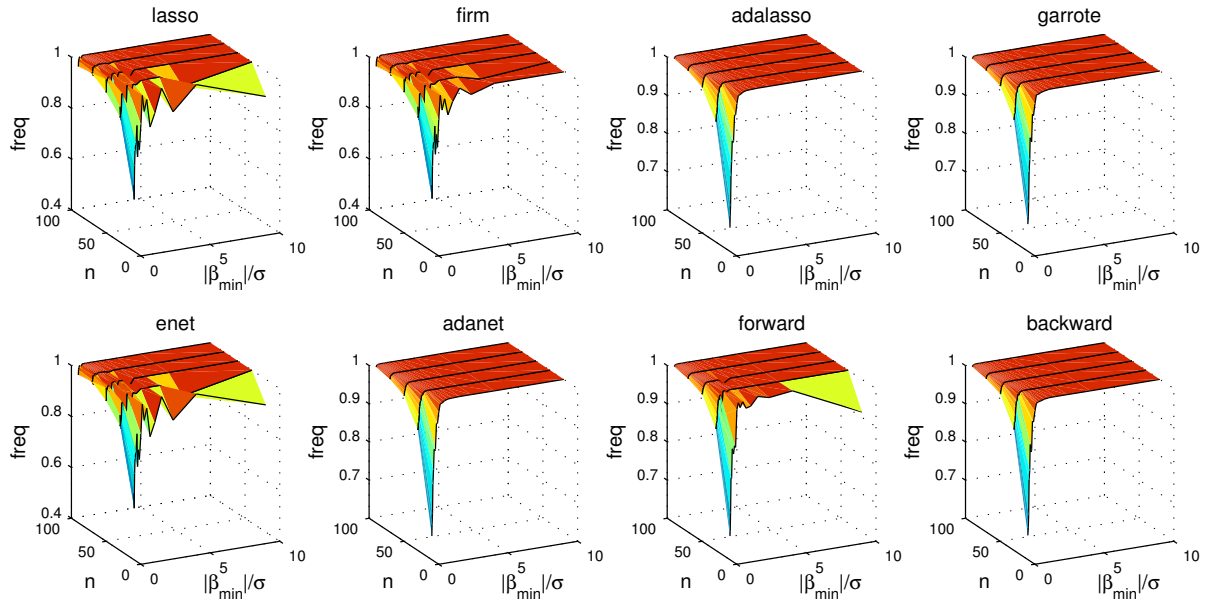


Figure 6.6: Average frequency of selection (freq) with respect to sample size (n) and signal-to-noise ratio ($|\beta_{\min}|/\sigma$) when the estimation is replaced by least-squares after computing the path (general design with maximum correlation $\rho = 0.4$ between the variables).

Other estimators of the regression coefficient. It is well known that the Least-squares estimator can be improved by other estimators such as the James-Stein (type) estimators or the Ridge Regression. In this paragraph, we propose to investigate if such a fact is also true when we are concerned with selecting the most relevant variables. Therefore, we run the same experiment as previously but we replace the estimation for each step of the path by the James-Stein estimator, the generalized James-Stein estimator or the Ridge Regression estimator. Figure 6.7 displays the average frequency of selecting the true subset with respect to the signal-to-noise ratio and to the sample size n for the 8 methods with the James-Stein, generalized James-Stein and Ridge Regression estimators. We do not display here all the 8 methods since we have seen in the previous paragraph that the graphs are exactly the same for some of the methods. In Figure 6.7, it is very striking to see that the results are not so good for the James-Stein (type) estimators than they were with the Least-Squares estimator. For the Ridge Regression estimator, there is little difference with Least-Squares, but the latter one still gives slightly better results.

6.1.4 Discussion on the first study

In view of the results on the study we performed, it appears that the methods for constructing collections of models that are the most appropriate with the actual estimation loss are Backward Elimination, Adaptive Lasso and Adaptive Elastic-Net. The performances of the two latter ones are improved when replacing their estimation by Least-Squares.

In a general way, the actual estimation loss seems to be a good theoretical criterion for variable selection when the corresponding estimator of β has very little bias, the best results being obtained with restricted Least-Squares and Ridge Regression. This result is interesting for Ridge Regression since it can always be computed, whereas Least-Squares has no unique

solution in the case where there is more variables than observations ($p \geq n$). Note that the (generalized) James-Stein estimators could however be used as post-model-selection operators, if needed.

Note also that the results on Lasso agree with its consistency in selection property proved by [Zhao & Yu 2007]. Indeed, replacing its estimator by the restricted least-squares $\hat{\beta}_I^{LS}$ once the path has been computed yields very good performances of selection. This means that the true subset actually belongs to Lasso's path most of the time.

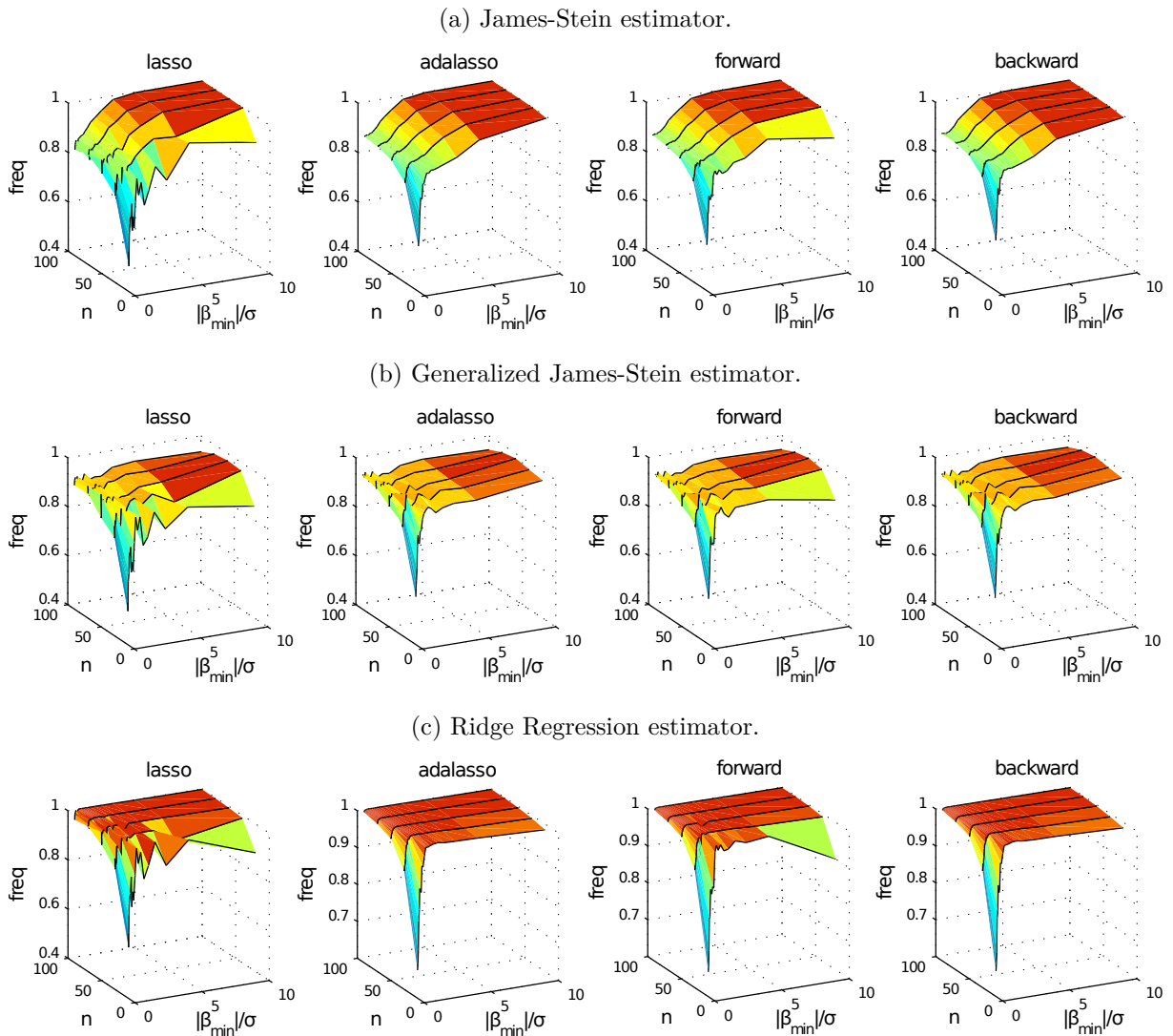


Figure 6.7: Average frequency of selection (freq) with respect to sample size (n) and signal-to-noise ratio ($|\beta_{\min}|/\sigma$) when the estimation is replaced by the James-Stein estimator (top), the generalized James-Stein estimator (center) and the Ridge Regression estimator (bottom) with $\lambda = 0.3$ after computing the path (general design with maximum correlation $\rho = 0.4$ between the variables).

6.2 Comparison of model evaluation criteria

In this section, we propose to turn to the comparison of methods evaluating the collection of model and estimating the prediction risk $R_{(X,Y)}$ or the estimation loss $\|X\hat{\beta} - X\beta\|^2$.

6.2.1 Purpose of the study

In the previous study, we tried to determine whether there is a collection of models yielding better performances when selected by the actual estimation loss. The results when replacing their respective estimator by the least-squares were very good for all of them, meaning that the true subset actually belongs to their paths. Hence, it is hard to tell whether one outperforms the others.

The objective of this new study is to compare, for each collection of models, the performances in selection of the different model selection criteria, our loss estimators on the one side and the methods presented in Chapter 2, Section 2.2 on the other side.

Protocol

We use the same protocol as in previous section, that is, the true regression coefficient is set to $(2, 0, 0, 4, 0)$, the matrix X is generated according to a Gaussian distribution where the \mathbf{x}^j 's have variance 1 and correlation $\rho = 0.4$ (since the performances are more erratic for higher correlation). The sample size n is taken in the set $\{20, 40, 100\}$, and the noise level σ in the set $\{0.5, 1, 2\}$, since this corresponds to the most interesting cases from the previous study.

As exposed in Chapter 3, Section 3.1.3, the main problems in model selection are to measure the complexity and to estimate the variance. For the first problem, we chose to use the generalized degrees of freedom because of the important part they play in Stein's identity. In practice however, the generalized degrees of freedom \widehat{df} are not always easy to compute. In Appendix A.4, we give the expression of \widehat{df} for the collections of model we study, when an analytical form is available. Otherwise, we present how they can be computed through directional derivatives, although this might substantially increase the computational costs.

As far as estimating the variance is concerned, we follow the discussion we had in Chapter 3 and divide the analysis into two cases: the case where the estimator of the variance $\hat{\sigma}_{full}^2$ is the same for all models, corresponding to the full model

$$Y = X\beta + \sigma\varepsilon,$$

and the case where we estimate the variance differently for each subset I , corresponding to the restricted linear model

$$Y = X_I\beta_I + \sigma\varepsilon.$$

We next present the performances in selection of our loss estimators for each case, and add a third case where the estimator of the variance is estimated differently on each model and is defined by

$$\hat{\sigma}_r^2(\hat{\beta}) = \frac{\|Y - X\hat{\beta}\|^2}{n - \widehat{df}}.$$

6.2.2 Unbiased loss estimator vs corrected loss estimator

In this subsection, we compare the corrected loss estimators developed in Chapter 4 to the unbiased estimators developed in Chapter 3.

Same estimator of the variance for all models

In this paragraph, we compare the unbiased estimator,

$$\hat{L}_0(\hat{\beta}) = \|Y - X\hat{\beta}\|^2 + (2 \operatorname{div}(X\hat{\beta}) - n)\hat{\sigma}_{full}^2,$$

the invariant unbiased estimator

$$\hat{L}_0^{inv}(\hat{\beta}) = (n - p - 2) \frac{\|Y - X\hat{\beta}\|^2}{S} + 2 \operatorname{div}_Y(X\hat{\beta}) - n + 4 (X\hat{\beta} - Y)^t X \frac{\partial g(\hat{\beta}^{LS}, S)}{\partial S},$$

where $S = \|Y - X\hat{\beta}^{LS}\|^2$ and the corrected estimator

$$\hat{L}_\gamma^f(\hat{\beta}) = \hat{L}_0(\hat{\beta}) - c_f \left(k Z_{(k+1)}^2 + \sum_{j=k+1}^p Z_{(j)}^2 \right)^{-1},$$

with the choices

$$c_f^* = -\frac{2S}{n-p} \left\{ \frac{S}{n-p+2} \left(-2p + \frac{4k(k+1)Z_{(k+1)}^2}{d(X\hat{\beta}^{LS})} \right) + 4d(X\hat{\beta}^{LS}) \right\} \quad (6.6)$$

$$\widehat{c}_f = \frac{2S^2}{(n-p)(n-p+4)(n-p+6)} \left(p - 2 - 2(\widehat{df} + 1) \frac{\widehat{df}}{p} \right). \quad (6.7)$$

Figure 6.8 displays the evolution of the average frequency of recovery (freq) of the true subset as a function of the noise level σ , for each value of the sample size n and for each methods constructing the models. The black line corresponds to the unbiased estimator \hat{L}_0 , the blue line to the invariant unbiased estimator \hat{L}_0^{inv} , the magenta line to the corrected estimator with c_f^* , the green line to the corrected estimator with \widehat{c}_f , and the red line to the true estimation loss. The dashed lines correspond to the standard deviations from the average frequency.

The first thing we notice at first glance is that, in a general way and for most collections of models, the performances obtained result in the following order of preference between the criteria:

$$\hat{L}_0 < \hat{L}_0^{inv} < \hat{L}_\gamma^f(\widehat{c}_f) < \hat{L}_\gamma^f(c_f^*).$$

This order is not verified for Backward Elimination, however. Also, the standard deviations are quite low and very similar between the two estimators. In view of such results, it seems worthwhile to consider correcting the unbiased estimator.

Second, the four criteria result in very good performances of selection (close to 1) with the Elastic net and the Adaptive Elastic Net, while their performances are average to low for the other collections of models. Indeed, interestingly enough, the results are not so good for the Adaptive Lasso and Forward selection, while the actual estimation loss is able to recover the true subset quite often for these methods. In particular, the performances are low for Lasso and MPC, a result that was expected because of their lack of adequacy with the actual estimation loss.

Different estimators of the variance based on subset size

In this paragraph, we compare the unbiased estimator,

$$\hat{L}_0(\hat{\beta}) = \|Y - X\hat{\beta}\|^2 + (2 \operatorname{div}(X\hat{\beta}) - n)\hat{\sigma}_{restricted}^2,$$

the invariant unbiased estimator

$$\hat{L}_0^{inv}(\hat{\beta}) = (n - k - 2) \frac{\|Y - X\hat{\beta}\|^2}{S_I} + 2 \operatorname{div}_Y(X\hat{\beta}) - n + 4 (X\hat{\beta} - Y)^t X \frac{\partial g(\hat{\beta}_I^{LS}, S_I)}{\partial S_I},$$

where $S_I = \|Y - X_I \hat{\beta}_I^{LS}\|^2$, and the corrected estimators

$$\hat{L}_\gamma^r(\hat{\beta}) = \hat{L}_0(\hat{\beta}) - c_r \left(\|X_I \hat{\beta}_I^{LS}\|^2 \right)^{-1},$$

with the choice

$$c_r^* = \frac{2 S_I}{n - k} \left\{ \frac{(k - 4) S_I}{n - k + 2} + 4 (X\hat{\beta} - Y)^t X_I \hat{\beta}_I^{LS} \right\}.$$

We also compare our results to the loss estimator developed by [Fourdrinier & Wells 1994], which is defined by

$$\hat{L}^*(\hat{\beta}) = \frac{k}{n - k + 2} \|Y - X\hat{\beta}\|^2 - \frac{2(\widehat{df} - 4)}{(n - k + 4)(n - k + 6)} \frac{\|Y - X\hat{\beta}_I^{LS}\|^4}{\|X_I \hat{\beta}_I^{LS}\|^2}.$$

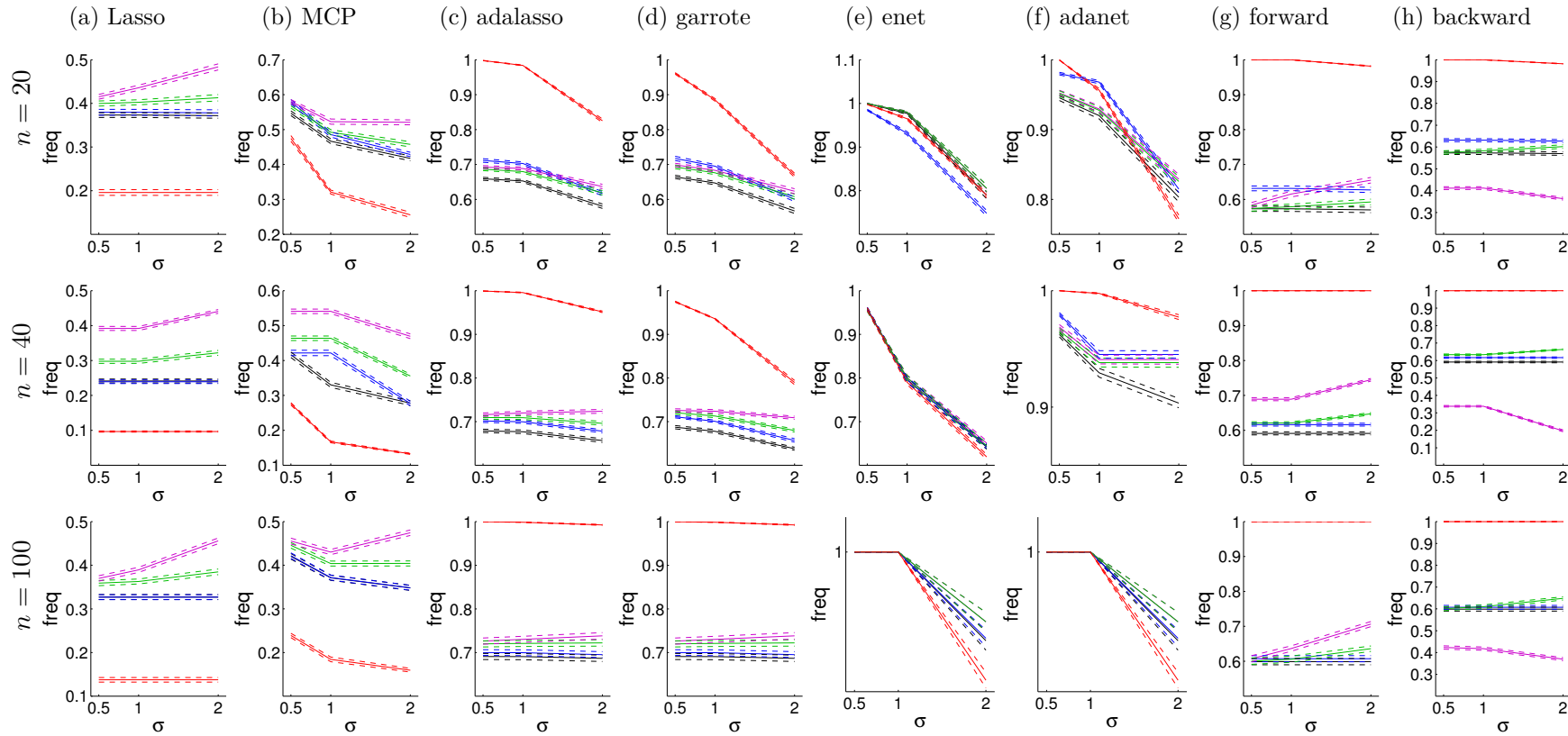


Figure 6.8: Average frequency of recovery of the true subset **under the full model assumption with independent estimator $\hat{\sigma}_{full}^2$ of the variance** for the unbiased loss estimator \hat{L}_0 (black line), for the invariant unbiased loss estimator \hat{L}_0^{inv} (blue line) and for the corrected estimator \hat{L}_γ^f with correction function γ_f with constant c_f^* (magenta line) and \hat{c}_f (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, the middle row to $n=40$ and the bottom row to $n=100$. Each column corresponds to a method for constructing the collection of models.

Figure 6.9 displays the evolution of the average frequency of recovery (freq) of the true subset as a function of the noise level σ , for each value of the sample size n and for each methods constructing the models. The black line corresponds to the unbiased estimator \hat{L}_0 , the blue line to the invariant unbiased estimator \hat{L}_0^{inv} , the magenta line to the corrected estimator \hat{L}_γ^r , the green line to the corrected estimator \hat{L}^* and the red line to the true estimation loss. The dashed lines correspond to the standard deviations from the average frequency.

In this case, the results are more erratic than in the previous one. Indeed, there is no clear ordering of the performances of the criteria, so we analyze them case by case.

For the Lasso, the four criteria obtained poor performances of selection. For the MCP, the performances are slightly better, especially when the variance is low, with a preference for the invariant unbiased estimator \hat{L}_0^{inv} . The performances of the four criteria are much better when the regression parameter β is estimated by Adaptive Lasso or Garrote. In this case, there does not seem to be much difference between the unbiased estimator \hat{L}_0 and the corrected estimator \hat{L}_γ^r , while the invariant unbiased estimator \hat{L}_0^{inv} and the corrected estimator \hat{L}^* clearly outperforms them for the Garrote, but not so clearly for the Adaptive Lasso.

As far as the Elastic net is concerned, the performances are again very close for the unbiased estimator \hat{L}_0 and the corrected estimator \hat{L}_γ^r , but they are strongly affected by the decrease in sample size. Indeed, their results are very good when $n = 100$, average when $n = 40$, and low when $n = 20$. Considering the fact that $p = 5$ is still small compared to $n = 20$, this might indicate their inability to handle the case where the number p of variables is close to the sample size n when using the Elastic net estimator.

Turning now to the Adaptive Elastic net, the performances of the four criteria are all very low. Looking more closely at the results, it turns out that, in most cases, they selected the model with the 4th variable only, while the true model is $\{1, 4\}$.

Finally, the performances for Forward Selection and Backward Elimination are pretty good with the three criteria \hat{L}_0 , \hat{L}_γ^r and \hat{L}^* , with a slight advantage of the corrected estimator \hat{L}^* . These good results are however mitigated for the high variance case ($\sigma = 2$). Note that, here, since β is estimated by Least-squares, the invariant unbiased estimator is actually equal to $k - 2$. It is thus worthless as a selector in this particular case.

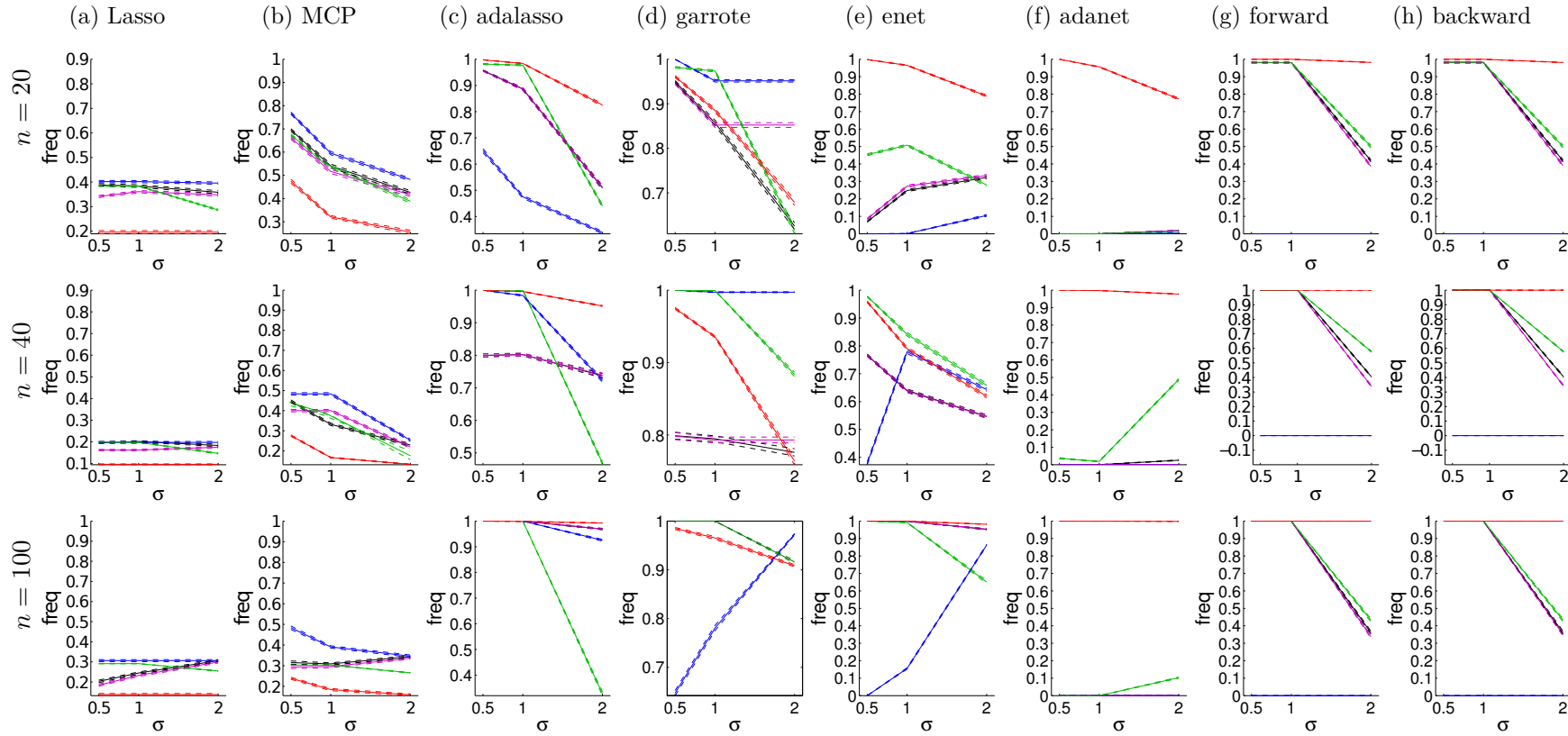


Figure 6.9: Average frequency of recovery of the true subset **under the restricted model assumption with independent estimator $\hat{\sigma}_{restricted}^2$ of the variance** for the unbiased loss estimator \hat{L}_0 (black line), for the invariant unbiased loss estimator \hat{L}_0^{inv} (blue line), for the corrected estimator \hat{L}_γ^r with correction function γ_r (magenta line) and for the corrected estimator \hat{L}^* (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, the middle row to $n = 40$ and the bottom row to $n = 100$. Each column corresponds to a method for constructing the collection of models.

Different estimators of the variance based on model complexity

In this paragraph, we tested a third estimator of the variance based on the collection of models. This estimator is defined by

$$\hat{\sigma}_r^2(\hat{\beta}) = \frac{\|Y - X\hat{\beta}\|^2}{n - \widehat{df}},$$

and leads to the following estimator

$$\widehat{L}_0(\hat{\beta}; \hat{\sigma}_r^2(\hat{\beta})) = \|Y - X\hat{\beta}\|^2 + (2 \operatorname{div}(X\hat{\beta}) - n) \hat{\sigma}_r^2(\hat{\beta}) = \frac{\widehat{df}}{n - \widehat{df}} \|Y - X\hat{\beta}\|^2.$$

Note that this estimator $\widehat{L}_0(\hat{\beta})$ is not unbiased anymore, because of the bias of $\hat{\sigma}_r^2(\hat{\beta})$ as well as its possible correlation to the generalized degrees of freedom \widehat{df} . However, such an estimator of the loss is interesting since its expression is close to that of the Final Prediction Error (FPE) and the Average Prediction Variance (APV) presented in Chapter 2. We also computed the corrected estimator

$$\widehat{L}_\gamma^r(\hat{\beta}; \hat{\sigma}_r^2(\hat{\beta})) = \widehat{L}_0(\hat{\beta}; \hat{\sigma}_r^2(\hat{\beta})) - c_r \left(\|X_I \hat{\beta}_I^{LS}\|^2 \right)^{-1},$$

with the choice

$$c_r^* = \frac{2 S_I}{n - k} \left\{ \frac{(k - 4) S_I}{n - k + 2} + 4 (X\hat{\beta} - Y)^t X_I \hat{\beta}_I^{LS} \right\},$$

and

$$\widehat{L}^*(\hat{\beta}; \hat{\sigma}_r^2(\hat{\beta})) = \frac{\widehat{df}}{n - \widehat{df} + 2} \|Y - X\hat{\beta}\|^2 - \frac{2(\widehat{df} - 4)}{(n - \widehat{df} + 4)(n - \widehat{df} + 6)} \frac{\|Y - X\hat{\beta}\|^4}{\|X\hat{\beta}\|^2}.$$

Figure 6.10 displays the evolution of the average frequency of recovery (freq) of the true subset as a function of the noise level σ , for each value of the sample size n and for each methods constructing the models. The black line corresponds to the unbiased estimator \widehat{L}_0 , the magenta line to the corrected estimator \widehat{L}_γ^r , the green line to the corrected estimator \widehat{L}^* and the red line to the true estimation loss. The dashed lines correspond to the standard deviations from the average frequency. Here again, the invariant unbiased estimator is a linear function of the generalized degrees of freedom and thus we do not display it.

For all the collections of models but the Elastic net and the Adaptive Elastic net, the performances in selection of the three criteria are very good and very close one to another when the variance is set to $\sigma \in \{0.5, 1\}$. The performances tend to decrease quite a lot in the case where $\sigma = 2$ and as n decreases as well.

As far as the Elastic net and the Adaptive Elastic net are concerned, the eperformances are good only for high sample size n and low variance σ .

Note also that, for all collections of models, the estimator \widehat{L}_0 is always outperformed by one of the corrected estimators, especially when $\sigma = 2$ (since in the other cases the difference is not noticeable).

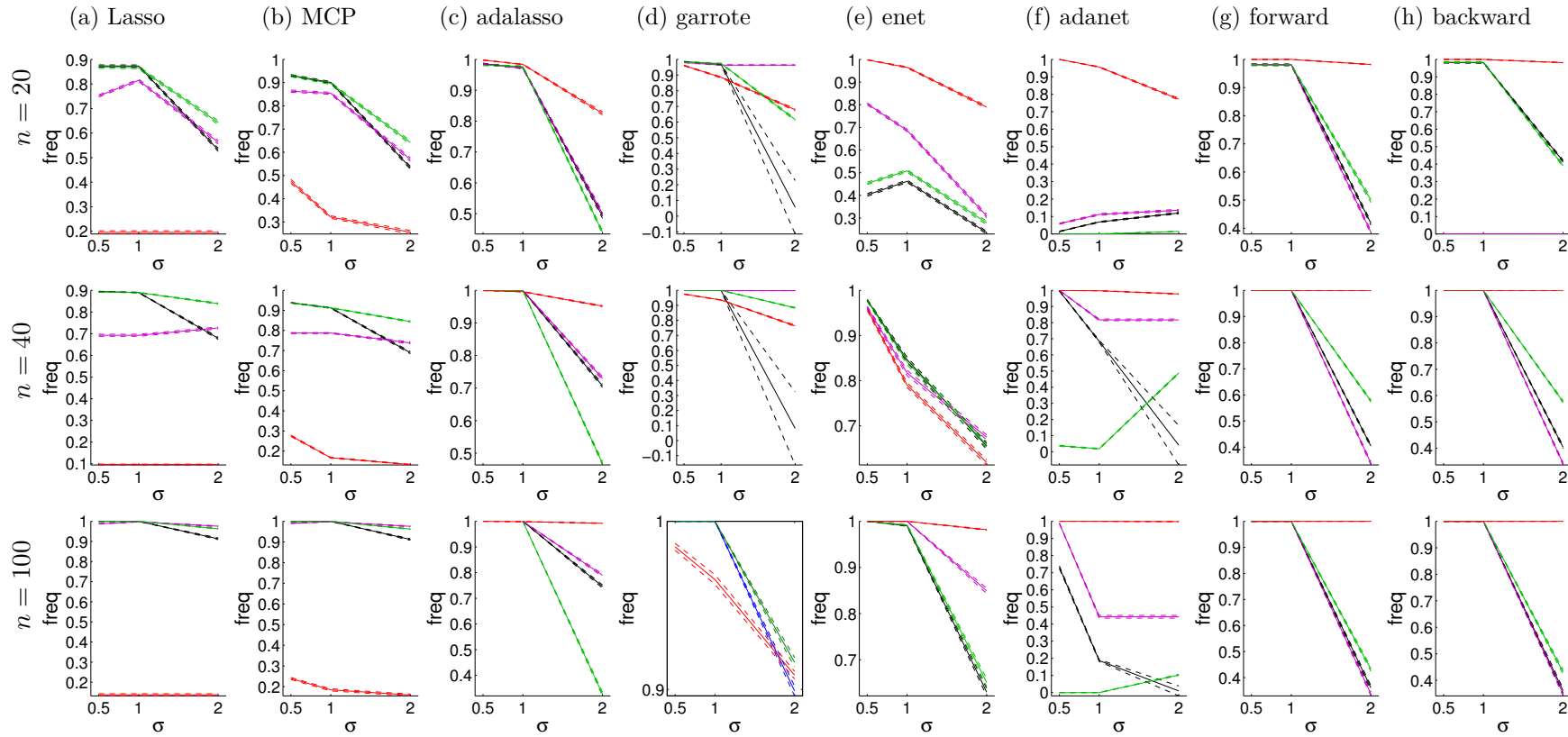


Figure 6.10: Average frequency of recovery of the true subset **under the restricted model assumption with dependent estimator** $\hat{\sigma}_r^2(\hat{\beta})$ **of the variance** for the loss estimator \hat{L}_0 (black line), for the corrected estimator \hat{L}_γ^r with correction function γ_r (magenta line), and for the corrected estimator \hat{L}^* (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, the middle row to $n = 40$ and the bottom row to $n = 100$. Each column corresponds to a method for constructing the collection of models.

Comparison of the three cases

In the three cases we presented, we noticed an improvement in terms of selection of the corrected estimators \hat{L}_γ^r and \hat{L}_γ^f over the reference estimator \hat{L}_0 , especially when considering the full model assumption. This is a very encouraging result which suggests to go on in this direction. Also, the performances for the invariant loss estimator \hat{L}_0^{inv} were often better than that of \hat{L}_0 . Hence, we can expect that correcting the invariant estimator \hat{L}_0^{inv} would result in even better performances than correcting \hat{L}_0 .

Nevertheless, the major change in performances have been obtained with the different estimators of the variance that we tested. Indeed, the performances of selection are much better for the Lasso and the MCP are very good when estimating the variance by $\hat{\sigma}_r^2(\hat{\beta})$, even though the actual estimation loss is barely able to recover the true subset. The Adaptive Lasso and the Garrote are also better selected under the restricted model and estimating the variance by either $\hat{\sigma}_{restricted}^2$ or $\hat{\sigma}_r^2(\hat{\beta})$. On the contrary, for the Elastic net and the Adaptive Elastic net, the performances are much better under the full model assumption, especially when the noise level is high ($\sigma = 2$), since the selected model is generally too small under the restricted model assumption. For Forward selection and Backward elimination, it is not so clear whether there is a better choice for estimating the variance. Indeed, when the noise level is small ($\sigma \leq 1$), the performances are clearly better under the restricted model assumption. However, for a high noise level ($\sigma = 2$), the performances under that same assumption heavily decrease, becoming lower than the performances under the full model assumption. These results clearly indicate that improving on the unbiased estimator is not sufficient, and that the choice of the estimator of the variance is as crucial in practice as it is in theory.

Table 6.2 summarizes these comments.

Method	$\hat{\sigma}_{full}^2$	$\hat{\sigma}_{restricted}^2$	$\hat{\sigma}_r^2(\hat{\beta})$
Lasso	–	–	++
MCP	–	–	++
Adaptive Lasso	–	++	++
Garrote	–	++	++
Elastic net	++	–	–
Adaptive Elastic net	–	++	–
Forward Selection	+	+	+
Backward Elimination	+	+	+

Table 6.2: Estimator of the variance yielding the best performances of selection. The double + sign indicates the best choice of estimator, while the – sign indicates poor performances of selection. The + sign alone indicates that there is no clear consensus.

Other distributions

We run the same example with a multivariate Student noise with $\nu = 5$ degrees of freedom and a multivariate Kotz noise with parameters $r = 0.5$ and $N = 2$. The results are given in Appendix A.5.1 and A.5.2. Basically, the general comments are the same as those given in the Gaussian case. The main differences are in the ability to handle a larger noise level σ . Indeed,

the performances in selection with a Student noise are slightly lower than with a Gaussian noise for the case $\sigma = 1$, while with a Kotz noise they are even better when $\sigma = 2$.

6.2.3 Comparison to existing methods from literature

We now turn to the comparison of the performances of our loss estimators to the model evaluation criteria presented in Chapter 2. For the discussion to be clear, we recall the different criteria we compare in Tables 6.3 and 6.4. Table 6.3 presents the criteria that depend on an estimator of the variance, so that we specify the exact form of the criteria in each case and we use a different notation depending on the estimator of the variance used to compute the criteria. Note that we replaced the subset size k by the generalized degrees of freedom \widehat{df} when necessary for the sake of comparison with our loss estimators. Note also that, for SRM, we used the expression proposed in [Cherkassky & Ma 2003], where the Vapnik-Chervonenkis dimension is approximated by \widehat{df} . On the other hand, we used the size k of the subset for the slope heuristics (SH) since we only have one model of each size in the collections of models we compared.

Figures 6.11 and 6.12 displays the evolution of the average frequency of recovery (freq) of the true subset as a function of the noise level σ , for the three model evaluation criteria yielding the best performances on each collection models. The blue line corresponds to the sample size $n = 20$, the green line to $n = 40$, and the red line to $n = 100$, while the dashed lines represent the standard deviation from the average in each case. The graphs for the other criteria are given in Appendix A.5.

Fist, note that the criteria giving the best performances are the same for the Lasso, the MCP, the Adaptive Lasso and the Garrote, namely our loss estimators $\widehat{L}_0(\hat{\sigma}_r^2(\hat{\beta}))$ and $\widehat{L}^*(\hat{\sigma}_r^2(\hat{\beta}))$, and the corrected AIC criterion (AIC_c), when the variance is estimated by $\hat{\sigma}_r^2(\hat{\beta})$ for the three criteria.

The story is a little different for the Elastic net and the Adaptive Elastic net, which are best selected respectively with the corrected estimator \widehat{L}_γ^f , AIC3 and BIC, and with \widehat{L}_γ^f , GCV and BIC, all of them computed with $\hat{\sigma}_{full}^2$.

Also, note that, in a general way, the results are better for MCP than for Lasso, and they are also better for the corrected estimators \widehat{L}_γ and \widehat{L}^* than for the unbiased estimator \widehat{L}_0 .

6.2.4 Discussion on the second study

This second simulation study clearly confirms the general principle that there is no criterion outperforming the others in all situations. Instead, there seems to be a few good model selection criteria for a given collection of models. Going a little further, more than good couples (collection of models, criterion), we highlighted the fact that there are good triples (collection of models, estimator of the variance, criterion), and we tried to give the triples yielding the best performances. If we had to choose only one of them in view of our results, we would suggest the Adaptive Elastic net selected by Generalized Cross-Validation (GCV)¹.

¹Note that here it is not a triple since GCV does not depend directly on an estimator of the variance.

Notation	Expression
$\widehat{L}_0(\hat{\sigma}_{full}^2)$	$\ Y - X\hat{\beta}\ ^2 + (2\widehat{df} - n) \frac{\ Y - X\hat{\beta}^{LS}\ ^2}{n - p}$
$\widehat{L}_0(\hat{\sigma}_{restricted}^2)$	$\ Y - X\hat{\beta}\ ^2 + (2\widehat{df} - n) \frac{\ Y - X\hat{\beta}_I^{LS}\ ^2}{n - k}$
$\widehat{L}_0(\hat{\sigma}_r^2(\hat{\beta}))$	$\frac{\widehat{df}}{n - \widehat{df}} \ Y - X\hat{\beta}\ ^2$
$\widehat{L}_\gamma^f(\hat{\sigma}_{full}^2)$	$\widehat{L}_0(\hat{\sigma}_{full}^2) - \gamma_f(X\hat{\beta}^{LS})$
$\widehat{L}_\gamma^r(\hat{\sigma}_{restricted}^2)$	$\widehat{L}_0(\hat{\sigma}_{restricted}^2) - \gamma_r(X_I\hat{\beta}_I^{LS})$
$\widehat{L}_\gamma(\hat{\sigma}_r^2(\hat{\beta}))$	$\widehat{L}_0(\hat{\sigma}_r^2(\hat{\beta})) - \gamma_r(X_I\hat{\beta}_I^{LS})$
$AIC(\hat{\sigma}_{full}^2)$	$(n - p) \frac{\ Y - X\hat{\beta}\ ^2}{\ Y - X\hat{\beta}^{LS}\ ^2} + 2\widehat{df}$
$AIC(\hat{\sigma}_r^2(\hat{\beta}))$	$n \log \left(\frac{\ Y - X\hat{\beta}\ ^2}{n} \right) + 2\widehat{df}$
$AIC_c(\hat{\sigma}_{full}^2)$	$(n - p) \frac{\ Y - X\hat{\beta}\ ^2}{\ Y - X\hat{\beta}^{LS}\ ^2} + \frac{2\widehat{df}(\widehat{df} + 1)}{n - \widehat{df} - 1}$
$AIC_c(\hat{\sigma}_r^2(\hat{\beta}))$	$n \log \left(\frac{\ Y - X\hat{\beta}\ ^2}{n} \right) + \frac{2\widehat{df}(\widehat{df} + 1)}{n - \widehat{df} - 1}$
$AIC_3(\hat{\sigma}_{full}^2)$	$(n - p) \frac{\ Y - X\hat{\beta}\ ^2}{\ Y - X\hat{\beta}^{LS}\ ^2} + 3\widehat{df}$
$AIC_3(\hat{\sigma}_r^2(\hat{\beta}))$	$n \log \left(\frac{\ Y - X\hat{\beta}\ ^2}{n} \right) + 3\widehat{df}$
$BIC(\hat{\sigma}_{full}^2)$	$(n - p) \frac{\ Y - X\hat{\beta}\ ^2}{\ Y - X\hat{\beta}^{LS}\ ^2} + \log(n) \widehat{df}$
$BIC(\hat{\sigma}_r^2(\hat{\beta}))$	$n \log \left(\frac{\ Y - X\hat{\beta}\ ^2}{n} \right) + \log(n) \widehat{df}$

Table 6.3: Model evaluation criteria (and their notations) that depend explicitly on an estimator of the variance.

Notation	Expression
GCV	$\frac{n}{(n - \widehat{df})^2} \ Y - X\hat{\beta}\ ^2$
LOOCV	$\frac{1}{n} \sum_{i=1}^n \ Y - X\hat{\beta}^{(-i)}\ ^2$
CV-5	$\frac{1}{5} \sum_{v=1}^5 \ Y - X\hat{\beta}^{(-v)}\ ^2$
CV-10	$\frac{1}{10} \sum_{v=1}^{10} \ Y - X\hat{\beta}^{(-v)}\ ^2$
SRM	$\ Y - X\hat{\beta}\ ^2 \times \left(1 - \sqrt{\frac{\widehat{df}}{n} \left(\log \left(\frac{n}{\widehat{df}} \right) + 1 \right) - \frac{\log \sqrt{(n/2)}}{n}} \right)^{-1}_+$
slope	$\ Y - X\hat{\beta}\ ^2 + \text{slope} \times k$

Table 6.4: Model evaluation criteria (and their notations) that do not depend explicitly on an estimator of the variance.

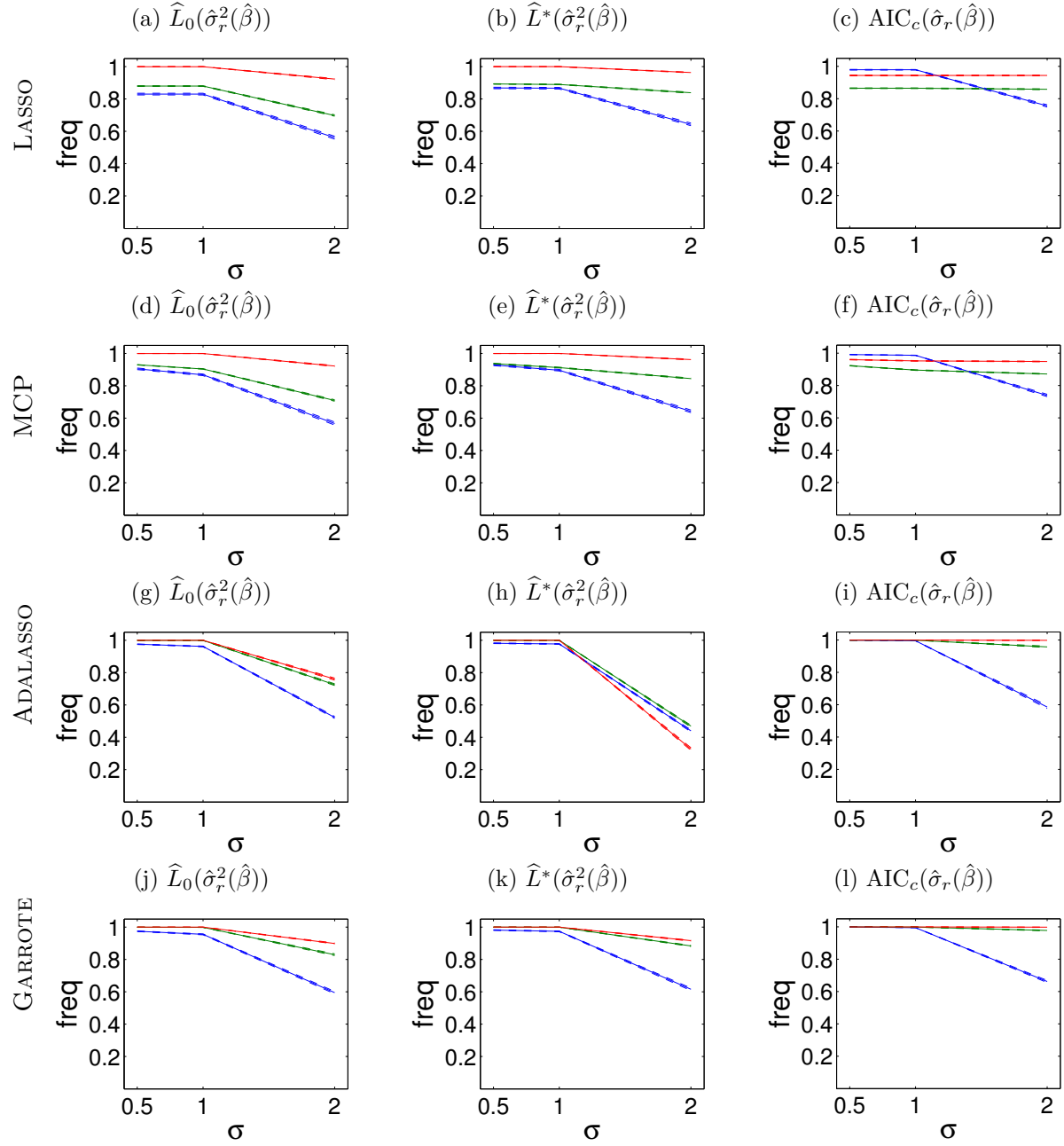


Figure 6.11: Average frequency of recovery of the true subset as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red), with Gaussian noise. The dashed lines display the standard deviation. Only the three model evaluation criteria giving the best results for each collection of models are displayed. The others results are given in Appendix A.5.

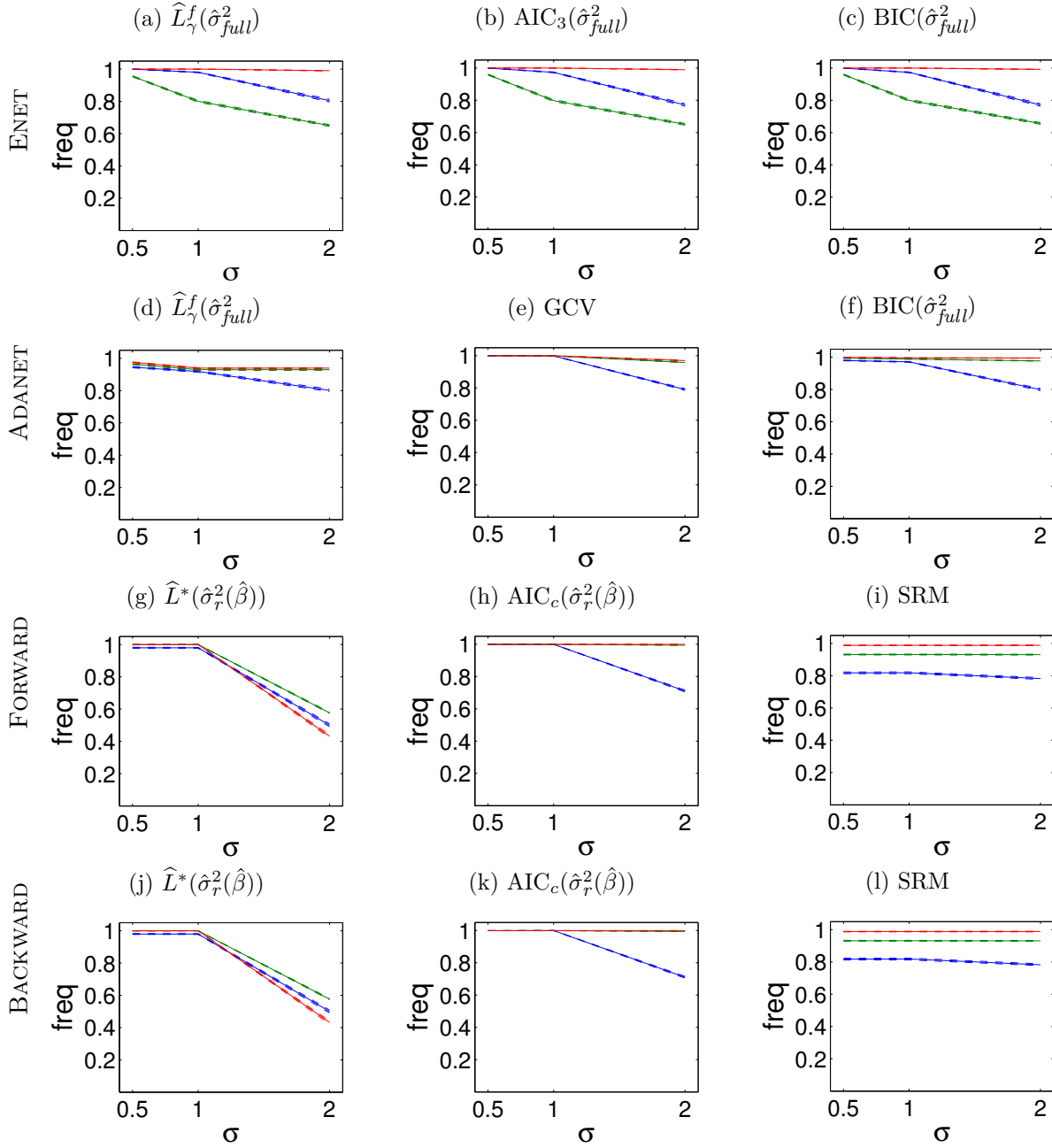


Figure 6.12: Average frequency of recovery of the true subset as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red), with Gaussian noise. The dashed lines display the standard deviation. Only the three model evaluation criteria giving the best results for each collection of models are displayed. The others results are given in Appendix A.5.

Conclusion et perspectives

Contents

7.1 Discussion on contributions and results	153
7.1.1 Summary on model evaluation	153
7.1.2 Summary on algorithmic and numerical aspects	155
7.1.3 Limitations of the present work	156
7.2 Perspectives and future works	157
7.2.1 Extension to elliptical symmetry	157
7.2.2 The Bayesian point of view	157
7.2.3 Other losses for comparing two model evaluation criteria	157
7.2.4 Application to classification and clustering	157

7.1 Discussion on contributions and results

In this manuscript, we studied several aspects of the problem of model selection and proposed to bring our little brick to this mighty bastion. In this section, we summarize our work, both theoretical and practical, before discussing future works and perspectives in the next section.

7.1.1 Summary on model evaluation

A large (but not too large) distributional framework. One of the main innovative feature of our work on model selection is the spherical assumption for the noise. Indeed, we showed that the form of our model evaluation criteria only relies on the spherical symmetry property, and not on the particular form of the distribution of the noise. This way, our criteria can handle a large spectrum of distributions, from light-tailed to heavy-tailed, bringing them distributional robustness to possible outliers. Also, this assumption allows us to consider the non-*i.i.d.* case since non-Gaussian spherical vectors have dependent components. This assumption is seldom made in the literature, even in the worst-case analysis where the distribution is assumed to be completely unknown (for instance in Vapnik's work). Finally, we believe that such a generality is not always an interesting feature and that the restriction to a family of distributions is necessary to construct tools with good performances.

Note that, in some cases, we have shown the equivalence between our unbiased criteria and several existing criteria such as Mallows' C_p and AIC. Hence, the robustness property might be shared with other existing criteria by equivalence, which, to the best of our knowledge, was never formally proved.

Estimating the variance and measuring the complexity. The question of which estimator of the variance to use is a tricky one and is often eluded in the literature, an interesting exception being [Arlot & Bach 2009]. However, as we have seen in the simulation study, it actually plays a crucial role. For instance, if it is estimated differently for each subset, then BIC almost always selects the null model $\beta = 0$. Even though this issue was raised in [Cherkassky & Ma 2003], where the authors considered estimating the variance both for the full model and the restricted model, they computed AIC and BIC with the true value of the noise level σ in their simulation study.

In this work, we have studied two ways to deal with the variance. In the first way, we first assume it to be known for the derivation of the criteria, and then replace it *a posteriori* by an estimator. We considered several estimators of the variance, namely the unbiased estimator based on the full least-squares solution, the unbiased estimators based on the least-squares solution restricted to a subset of variables, and, in the simulation study only, the estimator based on each model of the collection. In contrast, the second way of dealing with the variance we studied relies on the lack of knowledge on the variance from the very beginning and models explicitly its connexion to the problem through the invariant loss.

As far as the complexity is concerned, Stein identity provides an explicit measure through the notion of divergence, which is related to the generalized degrees of freedom \widehat{df} . This measure is reasonable since it corresponds to the rank of the application mapping Y to the prediction \hat{Y} , so that for linear mapping it is actually equal to the dimension of the application. Hence, it takes into account both the number of variables assumed to be relevant and the smoothness of their link to the target Y . In practice, when the analytical form of \widehat{df} is hard to determine, it can always be numerically computed through directional derivatives. Although this might increase significantly the computational time, it seems easier to estimate in regression and for nonlinear estimators than the VC-dimension or than other measures based on the covariance matrix of the estimator, the main alternatives to \widehat{df} .

New criteria for model evaluation with lower risk. We proposed a new way to compare model evaluation criteria, which is different than what is generally used in the literature, namely consistency and efficiency. Our method of comparison basically consists in minimizing the Mean-Squared Error (MSE) and the aim is to find which criterion has both low bias and low variance at the same time, guaranteeing a better control and more stability. Although this method is fairly common for comparing two estimators of the model parameters (see for instance [Hastie *et al.* 2008]), it has surprisingly not yet encountered the same success for comparing two model evaluation criteria.

From this second level of evaluation, we derived new loss estimators under two assumptions: the assumption that the true model is the one with all the p variables being relevant, and the assumption that subset I is the true subset of relevant variables.

Note that this work has been done under the nonasymptotic framework, as it is for oracle inequalities. However, the relation between the two measures of quality of a model evaluation criterion is not straightforward. It would thus be interesting to investigate whether the corrected loss estimators we developed Chapter 4 have better oracle inequalities than the unbiased loss estimators. The simulation study we ran seems to answer positively to that question, but it needs to be verified theoretically.

7.1.2 Summary on algorithmic and numerical aspects

For the comparison of the performances of our loss estimators in practice, we were confronted to several algorithmic problems: the construction of collections of models through regularization path algorithm for nonconvex problems, and the random generation of spherically symmetric vectors.

Regularization path algorithms for nonconvex and nondifferentiable problems. Regularization path algorithms are an efficient way to find the transition values of the hyperparameter λ adding a variable to the current subset. Hence, it is very well suited for constructing collection of models. It has been successfully developed for the Lasso in [Efron *et al.* 2004], and the LARS algorithm also benefits to other estimators such as Adaptive Lasso, Elastic net and Adaptive Elastic net with just a change in variable. However, its application to nonconvex and nondifferentiable problems is not straightforward since the optimality conditions it relies on are only valid for convex functionals.

We used a generalization to nonconvex functionals through the notion of Clarke differential. This way, we were able to prove that similar optimality conditions apply, which enables us to provide a regularization path algorithm for the Minimax Concave Penalty (MCP) [Zhang 2010]. Similar algorithms can be derived for other nonconvex problems, such as the Smoothly Clipped Absolute Deviation (SCAD) [Fan & Li 2001].

Random variable generation for spherically symmetric vectors. We developed all our codes in Matlab, which only provides pseudo-random generators for univariate random variables (*e.g.* Gaussian and Student) or for multivariate copulas. Our distributional assumption of spherically symmetric vectors does not fit in either case, so we studied how to generate non-Gaussian spherically symmetric random vectors. We developed a code for generating the spherical distributions most frequently encountered in the literature, and used it for the simulation study.

Numerical comparison of performances in selection for different collections of models. When we first tested our loss estimators on the Lasso, we were surprised at the poor performances in selection we obtained. It seemed to be contradictory with the works on its oracle inequalities and consistency. The paper of [Leng *et al.* 2006] clearly showed that tuning the Lasso with the estimation loss $\|X\hat{\beta} - X\beta\|^2$ does lead to good prediction performances but with a poor selection. This is due to the fact that, on small models, the Lasso estimator presents a large bias, and this bias decreases as more variables are added to the selection. Inspired by [Leng *et al.* 2006], we decided to verify whether the oracle of each collection of models is actually close to the true underlying model when the latter one belongs to the collection.

It turns out that MCP shares a similar behaviour than the Lasso, although it is a little better. This might come from the fact that the first variables to be selected are the same for both methods. In cases where one fails to select a relevant variable first, so does the other. Their performances in selection can however be much improved by replacing their estimator with the least-squares solution on the selection. On the contrary, the Adaptive Lasso, the Elastic net, the Adaptive Elastic net, and Stepwise methods showed a very good behaviour in our simulation study. This seems to confer them good oracle properties and to guarantee that they can both well predict and well select at the same time.

Numerical comparison of model evaluation criteria. After having verified the oracle properties of the collection of models, we went to the next level by testing whether our loss estimators and other criteria estimating the actual estimation loss $\|X\hat{\beta} - X\beta\|^2$ were also able to recover the true underlying model. For each collection of models, we selected the three model evaluation criteria yielding the best performances in selection. Our loss estimators were always present among the three best criteria. It is quite compelling to see how different these criteria are from what was proposed by the respective authors of the collection of models, which is summarized in Table 2.1.

Also, we clearly pointed out that the estimator of the variance plays a crucial part in the performances of the model evaluation criteria that rely on it (see Table 6.3). We compared the results with several variance estimators and selected the one yielding the best performances in each case (see Table 6.2).

7.1.3 Limitations of the present work

The linear model and the $p < n$ assumption. The derivation of unbiased loss estimators under the spherical assumption relies on the assumption that the true underlying model is linear, essentially because of the use of the canonical form. Although it is possible to derive them without the canonical form, it involves much more tedious calculations and is harder to perform. Under the Gaussian assumption, however, the underlying model need not be linear, nor does the estimator of the mean of Y , so that the extension to the nonlinear case is straightforward.

Also, the same canonical form as well as the estimator of the variance we used constrain the study to the case where the sample size n is larger than the number p of possible explanatory variables. The invariant unbiased loss estimator even requires $p < n - 2$. The use of other estimators of the variance, such as the maximum likelihood estimator, could overcome this problem, but a deeper study is needed to see if it gives similar performances.

The X -fixed design case versus the X -random design case. In this work, we assumed the design matrix X to be fixed since it offers a great simplification. However, this assumption is seldom realistic in practice. Although there exists a relation between the conditional prediction risk $R_{Y|X}$ (corresponding to the X -fixed assumption) and the prediction risk $R_{(X,Y)}$ (corresponding to the X -random assumption), there is no guaranty that their minimum coincides.

Scalability and application to real datasets. We only run our simulation study on one small example with varying noise level σ and varying sample size n . A larger simulation study on other examples is needed to analyze and compare the performances of our loss estimators and methods from the literature when the true subset size is changed and when the true underlying model is not linear.

Also, our methods should be tested on real datasets in order to see whether their performances are still good in practice.

7.2 Perspectives and future works

7.2.1 Extension to elliptical symmetry

The next interesting step would be to consider the more general family of elliptically symmetric distributions. These distributions are a generalization of the spherical family where the covariance matrix is not proportional to the identity matrix. This way, we can actually model both dependence and correlation between the components of the output variable Y . A particularly interesting case of this family is the heteroskedasticity, where the covariance matrix is diagonal but not proportional to the identity matrix.

7.2.2 The Bayesian point of view

Even though we exposed and studied several Bayesian criteria (such as BIC), the discussion on model selection was more from a frequentist point of view. However, loss estimation can also be treated from a Bayesian perspective. Indeed, assuming that the regression coefficient β is also a random variable with prior distribution $\pi(\beta)$, [Fourdrinier & Strawderman 2003] and [Fourdrinier & Wells 2012] considered correction functions taking into account the corresponding marginal density. In both papers, the authors only applied it to the case where $m(Y) = \|Y\|^{-(n-2)}$. Since estimators of β such as the Lasso or the Ridge regression can be seen as Bayes estimators with respectively a Laplace or a Gaussian prior distribution, there is a wide range of new correction functions to investigate. It would thus be interesting to run a similar experiment as we did in the simulation study in order to compare Bayesian methods, and also to compare both Bayesian and frequentist methods. Among the Bayesian literature, we can cite for instance the interesting works on Bayes factors with Zellner's g -prior (see [George & Foster 2000] and [Maruyama & George 2011] for instance).

7.2.3 Other losses for comparing two model evaluation criteria

The comparison of two model evaluation criteria through the quadratic communication risk may not be the most appropriate loss. Indeed, the quadratic estimation loss $L(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|^2$ is comparable to an estimator of a variance term. Hence, we could use another loss such as Stein risk which penalizes less the large values of $L(\theta, \hat{\theta})$, or Rukhin's risk [Rukhin 1988a, Rukhin 1988b] (see Chapter 4).

7.2.4 Application to classification and clustering

Loss estimation is a fairly general theory and the use of a loss function different from the quadratic estimation loss $\|X\hat{\beta} - X\beta\|^2$ could enable its adaptation to other problems than regression. For instance, classification problems deal with the estimation of the 0 – 1 excess loss (see [Boucheron *et al.* 2005]).

Another example is that of clustering with mixture models, where one of the objectives is to estimate the parameters of the mixture density modelling the data (see for instance [Nadif & Govaert 1998] or [Govaert & Nadif 2010]). In such a case, the baseline loss can be defined as the log-likelihood.

APPENDIX **A**

Appendix

A.1 Woodbury matrix update

In this appendix, we recall how we can update an inverse matrix through the Woodbury matrix identity. We first recall the identity in the general case, then we apply it when we want to add or delete a column and a line to a symmetric matrix. This appendix is based on [Hager 1989].

A.1.1 Woodbury matrix identity

The Sherman-Morrison-Woodbury formula helps computing the inverse matrix the matrix

$$(A - UV)^{-1}$$

when we know the inverse matrix A^{-1} . In this formula, A is a square symmetric matrix in $\mathbb{R}^{n \times n}$, and U and V^t are both in $\mathbb{R}^{n \times d}$. We expose it in the following

Lemma A.1 (Woodbury matrix identity.). *Let A be a square matrix in $\mathbb{R}^{n \times n}$, and U and V^t be two matrices in $\mathbb{R}^{n \times d}$. If A and $A - UV$ are invertible, then*

$$(A - UV)^{-1} = A^{-1} + A^{-1}U(I_d - VA^{-1}U)^{-1}VA^{-1}.$$

The special case where U and V are vectors ($d = 1$) corresponds to the Sherman-Morrison identity. An extension of this formula is to replace V by $D^{-1}V$, which results in the following modification the identity

$$(A - UD^{-1}V)^{-1} = A^{-1} + A^{-1}U(D - VA^{-1}U)^{-1}VA^{-1}. \quad (\text{A.1})$$

In this appendix, we are also interested in another form of the identity, where we wish to compute the inverse of the matrix

$$M = \begin{pmatrix} A & U \\ V & D \end{pmatrix},$$

where D is a square matrix in $\mathbb{R}^{d \times d}$. The identity becomes

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}U(D - VA^{-1}U)^{-1}VA^{-1} & -A^{-1}U(D - VA^{-1}U)^{-1} \\ -(D - VA^{-1}U)^{-1}VA^{-1} & (D - VA^{-1}U)^{-1} \end{pmatrix}. \quad (\text{A.2})$$

A.1.2 Update for adding a column and a line

In the algorithms from Chapter 5, Section 5.1, at a given step we compute the inverse of the matrix

$$\Sigma_I = (X_I^t X_I)^{-1},$$

where I is a sequence of indices such that $\#I = d < n$ and X is the design matrix from the linear model

$$Y = X\beta + \varepsilon.$$

On the following step(s), we wish to compute the inverse matrix

$$\Sigma = (X^t X)^{-1},$$

when we add d columns to X (denoted here by X_0):

$$X = \begin{pmatrix} X_I & X_0 \end{pmatrix}.$$

The relation between Σ_I and Σ is thus

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_I^{-1} & X_I^t X_0 \\ X_0^t X_I & X_0^t X_0 \end{pmatrix},$$

which allows us to use directly the formula (A.2) with $A = \Sigma_I^{-1}$, $U = V^t = X_I^t X_0$ and $D = X_0^t X_0$.

A.1.3 Update for deleting a column and a line

This time, we consider the reverse situation where we know the inverse matrix

$$\Sigma = (X^t X)^{-1}$$

and we wish to compute the inverse matrix

$$\Sigma_I = (X_I^t X_I)^{-1},$$

where I is a sequence of indices such that $\#I = d < n$. Without loss of generality, we can consider the case where $I = \{1, \dots, n-d\}$, that is, we wish to update the inverse of Σ when we delete its last d rows and its last d columns. If the d rows and columns we wish to delete are not the last ones of Σ , it suffices to reorder the matrix so that

$$\Sigma^{-1} = \begin{pmatrix} X_I^t X_I & X_I^t X_0 \\ X_0^t X_I & X_0^t X_0 \end{pmatrix}.$$

Now, remarking that

$$\begin{pmatrix} \Sigma_I^{-1} & 0 \\ 0 & I_{n-d} \end{pmatrix} = \Sigma^{-1} - \begin{pmatrix} 0 & X_I^t X_0 \\ X_0^t X_I & X_0^t X_0 - I_{n-d} \end{pmatrix},$$

where the identity matrix I_{n-d} has been added to allow the invertibility of the left-hand side matrix, we can apply the original Woodbury formula (A.1) with $A = \Sigma^{-1}$, and where U , D and V should verify the equality

$$UD^{-1}V = \begin{pmatrix} 0 & X_I^t X_0 \\ X_0^t X_I & X_0^t X_0 - I_{n-d} \end{pmatrix} \quad (\text{A.3})$$

The equality (A.3) is verified for instance when

$$U = \begin{pmatrix} 0 & X_I^t X_0 \\ I_{n-d, n-d} & \frac{1}{2}(X_0^t X_0 - I_{n-d}) \end{pmatrix}, \quad V = U^t, \quad D = \begin{pmatrix} 0 & I_{n-d, n-d} \\ I_{n-d, n-d} & 0 \end{pmatrix}. \quad (\text{A.4})$$

A.2 Twice weak differentiability of the correction function γ_f

In this appendix, we study the weak differentiability of the corrective function defined in Equation (4.26). We recall this function hereafter:

$$\gamma_f(Z) = \frac{a}{\left[(k+1)Z_{(k+1)}^2 + \sum_{i=k+2}^p Z_{(i)}^2\right]}, \quad a \in \mathbb{R} \quad (\text{A.5})$$

Here, in an aim of simplifying the notations, we denote by $Z_{(i)}, i = k+1, \dots, p$, the elements which have absolute value lower than λ . They are organized as if we had sorted them: $|Z_{(1)}| > \dots > |Z_{(p)}|$. Hence $Z_{(k+1)}$ corresponds to the element with highest absolute value among those with absolute value lower than λ .

But we can state this function in terms of λ :

$$\gamma_f(Z) = \frac{a}{kZ_j^2 \mathbf{1}_{\{j=\arg \max_l \{|Z_l| \leq \lambda\}\}} + \sum_{i=1}^p Z_i^2 \mathbf{1}_{\{i \in \{l \mid |Z_l| \leq \lambda\}\}}}$$

Next, we give the definition for the k times weak differentiability.

Definition A.1. A function $u \in L_{loc}^1(\mathbb{R}^p)$ is said to be k times weakly differentiable if, for every multiindex α such that $|\alpha| \leq k$, there is a function $u_\alpha \in L_{loc}^1(\mathbb{R}^p)$ with the following property:

$$\int_{\mathbb{R}^p} u \partial^\alpha \varphi dx = (-1)^{|\alpha|} \int_{\mathbb{R}^p} u_\alpha \varphi dx \quad \forall \varphi \in C_0^\infty(\mathbb{R}^n) \quad (\text{A.6})$$

Here, we need $k = 2$ in order to apply Corollary 4.1 with γ_f , so we have to verify (A.6) for $(\gamma_f)_i$, $(\gamma_f)_{i,j}$ and $(\gamma_f)_{i,i}$. Without loss of generality, we will assume here that $a = 1$.

If these functions exist, they should be equal to:

$$\begin{aligned} (\gamma_f)_i(x) &= \frac{-2x_i}{d^2(x)} \left(k \mathbf{1}_{\{i=\arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbf{1}_{\{i \in \{l \mid |x_l| \leq \lambda\}\}} \right) \\ (\gamma_f)_{i,j}(x) &= \frac{4x_i x_j}{d^3(x)} \left(k \mathbf{1}_{\{i \vee j = \arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbf{1}_{\{i \wedge j \in \{l \mid |x_l| \leq \lambda\}\}} \right) \\ (\gamma_f)_{i,i}(x) &= \frac{2}{d^2(x)} \left((2(k+2)x_i^2 - 1)k \mathbf{1}_{\{i=\arg \max_l \{|x_l| \leq \lambda\}\}} + (2x_i^2 - 1) \mathbf{1}_{\{i \in \{l \mid |x_l| \leq \lambda\}\}} \right) \end{aligned}$$

where $d(x) = kx_i^2 \mathbf{1}_{\{i=\arg \max_l \{|x_l| \leq \lambda\}\}} + \sum_{i=1}^p x_i^2 \mathbf{1}_{\{i \in \{l \mid |x_l| \leq \lambda\}\}}$, $i \wedge j$ means “ i and j ”, and $i \vee j$ means “ i or j ”.

In the first case, we have by Fubini:

$$\int_{\mathbb{R}^p} (\gamma_f)_i \varphi dx = \int_{\mathbb{R}} \dots \int_{\mathbb{R}} (\gamma_f)_i \varphi dx_i dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_p$$

Now, replacing u_i by its value and using integration by part, we have:

$$\begin{aligned} \int_{\mathbb{R}} (\gamma_f)_i \varphi dx_i &= \int_{\mathbb{R}} \frac{-2x_i}{d^2(x)} \left(k \mathbf{1}_{\{i=\arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbf{1}_{\{i \in \{l \mid |x_l| \leq \lambda\}\}} \right) \varphi dx_i \\ &= -2 \left(k \mathbf{1}_{\{i=\arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbf{1}_{\{i \in \{l \mid |x_l| \leq \lambda\}\}} \right) \int_{\mathbb{R}} \frac{x_i}{d^2(x)} \varphi dx_i \\ &= \left[\frac{1}{d(x)} \varphi \right]_{-\infty}^{+\infty} - \int_{\mathbb{R}} \frac{1}{d(x)} \partial^i \varphi dx_i \end{aligned}$$

The first term is zero since φ has a compact support. Hence, we obtain:

$$\begin{aligned} \int_{\mathbb{R}^p} (\gamma_f)_i \varphi \, dx &= - \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \frac{1}{d(x)} \partial^i \varphi \, dx_i dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_p \\ &= - \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \gamma_f \partial^i \varphi \, dx_i dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_p \\ &= - \int_{\mathbb{R}^p} \gamma_f \partial^i \varphi \, dx \end{aligned}$$

Using the same approach, we find that:

$$\begin{aligned} \int_{\mathbb{R}^2} (\gamma_f)_{i,j} \varphi \, dx_j dx_i &= \int_{\mathbb{R}^2} \frac{4x_i x_j}{d^3(x)} \left(k \mathbb{1}_{\{i \vee j = \arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbb{1}_{\{i \wedge j \in \{l \mid |x_l| \leq \lambda\}\}} \right) \varphi \, dx_j dx_i \\ &= - \int_{\mathbb{R}} \left[\left(k \mathbb{1}_{\{i = \arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbb{1}_{\{i \in \{l \mid |x_l| \leq \lambda\}\}} \right) \frac{2x_i}{d^2(x)} \varphi \right]_{-\infty}^{+\infty} dx_i \\ &\quad + \int_{\mathbb{R}^2} \left(k \mathbb{1}_{\{i \vee j = \arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbb{1}_{\{i \wedge j \in \{l \mid |x_l| \leq \lambda\}\}} \right) \frac{2x_i}{d^2(x)} \partial^j \varphi \, dx_j dx_i \\ &= 0 - \int_{\mathbb{R}^2} (\gamma_f)_i \partial^j \varphi \, dx_i dx_j \\ &= \int_{\mathbb{R}^2} \gamma_f \partial^{j,i} \varphi \, dx_i dx_j \end{aligned}$$

using again Fubini's theorem, integration by part, the compact supports of φ and $\partial^j \varphi$, and the previous result with $\partial^j \varphi$ instead of φ .

Hence we have:

$$\begin{aligned} \int_{\mathbb{R}^p} (\gamma_f)_{i,j} \varphi \, dx &= \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \gamma_f \partial^{i,j} \varphi \, dx_i dx_j dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_{j-1} dx_{j+1} \dots dx_p \\ &= \int_{\mathbb{R}^p} \gamma_f \partial^{i,j} \varphi \, dx \end{aligned}$$

Finally, we verify (A.6) for $\alpha = (i, i)$:

$$\begin{aligned} \int_{\mathbb{R}} (\gamma_f)_{i,i} \varphi \, dx_i &= \int_{\mathbb{R}} \frac{2}{d^2(x)} \left((2(k+2)x_i^2 - 1)k \mathbb{1}_{\{i = \arg \max_l \{|x_l| \leq \lambda\}\}} + (2x_i^2 - 1)\mathbb{1}_{\{i \in \{l \mid |x_l| \leq \lambda\}\}} \right) \varphi \, dx_i \\ &= - \left[\left(k \mathbb{1}_{\{i = \arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbb{1}_{\{i \in \{l \mid |x_l| \leq \lambda\}\}} \right) \frac{2x_i}{d^2(x)} \varphi \right]_{-\infty}^{+\infty} \\ &\quad + \int_{\mathbb{R}} \left(k \mathbb{1}_{\{i \vee j = \arg \max_l \{|x_l| \leq \lambda\}\}} + \mathbb{1}_{\{i \wedge j \in \{l \mid |x_l| \leq \lambda\}\}} \right) \frac{2x_i}{d^2(x)} \partial^i \varphi \, dx_i \\ &= 0 - \int_{\mathbb{R}} (\gamma_f)_i \partial^i \varphi \, dx_i \\ &= \int_{\mathbb{R}} \gamma_f \partial^{i,i} \varphi \, dx_i \end{aligned}$$

for the same reasons stated before.

Hence we have:

$$\begin{aligned} \int_{\mathbb{R}^p} (\gamma_f)_{i,i} \varphi \, dx &= \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \gamma_f \partial^{i,i} \varphi \, dx_i dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_p \\ &= \int_{\mathbb{R}^p} \gamma_f \partial^{i,i} \varphi \, dx \end{aligned}$$

which completes the verification that γ_f is twice weakly differentiable.

A.3 Subfunctions for LARS-MCP algorithm

In this appendix, we give the subfunctions useful for the completeness of Algorithm 5.2.

The function called *update_subsets* makes the necessary changes on the subsets depending on which move needs to be performed.

Algorithm A.1 *update_subsets*: Update the subsets I_N , I_P and I_0

Require: $I_N^{\text{current}}, I_P^{\text{current}}, I_0^{\text{current}}, j, \text{move}$

Ensure: $I_N^{\text{next}}, I_P^{\text{next}}, I_0^{\text{next}}$

```

if move == 0 then
     $I_N^{\text{next}} \leftarrow I_N^{\text{current}}, I_P^{\text{next}} \leftarrow I_P^{\text{current}}, I_0^{\text{next}} \leftarrow I_0^{\text{current}}$ 
else
    if move == 1 then
         $I_N^{\text{next}} \leftarrow I_N^{\text{current}}$ 
         $I_P^{\text{next}} \leftarrow [I_P^{\text{current}}, I_0^{\text{current}}(j)]$ 
         $I_0^{\text{next}} \leftarrow I_0^{\text{current}}, I_0^{\text{next}}(j) \leftarrow []$ 
    end if
    if move == 2 then
         $I_N^{\text{next}} \leftarrow [I_N^{\text{current}}, I_P^{\text{current}}(j)]$ 
         $I_P^{\text{next}} \leftarrow I_P^{\text{current}}, I_P^{\text{next}}(j) \leftarrow []$ 
         $I_0^{\text{next}} \leftarrow I_0^{\text{current}}$ 
    end if
    if move == -1 then
         $I_N^{\text{next}} \leftarrow I_N^{\text{current}}$ 
         $I_P^{\text{next}} \leftarrow I_P^{\text{current}}, I_P^{\text{next}}(j) \leftarrow []$ 
         $I_0^{\text{next}} \leftarrow [I_P^{\text{current}}(j), I_0^{\text{current}}]$ 
    end if
    if move == -2 then
         $I_N^{\text{next}} \leftarrow I_N^{\text{current}}, I_N^{\text{next}}(j) \leftarrow []$ 
         $I_P^{\text{next}} \leftarrow [I_N^{\text{current}}(j), I_P^{\text{current}}]$ 
         $I_0^{\text{next}} \leftarrow I_0^{\text{current}}$ 
    end if
end if

```

A.4 Computing the degrees of freedom

In this appendix, we treat the problem of computing in practice the generalized degrees of freedom of a given estimator $X\hat{\beta}$, which is equal to its divergence with respect to Y :

$$\widehat{df}(X\hat{\beta}) = \text{div}_Y(X\hat{\beta}),$$

as it is a central quantity intervening in our model evaluation criteria as a measure of complexity.

The generalized degrees of freedom can be computed analytically for most of the estimators of the parameter β we considered in this manuscript. Section A.4.1 presents the expressions for such estimators. However, for the Adaptive Lasso and the Adaptive Elastic Net, computing the analytical form of \widehat{df} is more delicate because of the dependence to an initial estimate of β (like the least-squares or the ridge estimator). Hence, in Section A.4.2 we also expose how to compute \widehat{df} thanks to the directional derivative.

A.4.1 Analytical form

The easiest case is that of the (restricted or full) least-squares estimator, where, as we have repeatedly indicated, the generalized degrees of freedom are exactly equal to the true degrees of freedom and correspond to the number of nonzero components in $\hat{\beta}$. We give, in a series of lemmas, the form of the estimators whose generalized degrees of freedom are known analytically.

Lemma A.2 (Degrees of freedom of the Least-squares estimator). *The degrees of freedom of the Least-squares estimator and the restricted Least-squares estimator are respectively equal to*

$$\begin{aligned}\widehat{df}(X\hat{\beta}^{LS}) &= p \\ \widehat{df}(X_I\hat{\beta}_I^{LS}) &= k.\end{aligned}$$

Proof. By definition of the least-squares estimator and of the restricted least-squares estimator, we have that

$$\begin{aligned}\widehat{df}(X\hat{\beta}^{LS}) &= \text{div}_Y(X(X^tX)^{-1}X^tY) = \text{tr}(X(X^tX)^{-1}X^t) = \text{tr}((X^tX)^{-1}X^tX) = \text{tr}(I_p) = p, \\ \widehat{df}(X_I\hat{\beta}_I^{LS}) &= \text{tr}(X_I(X_I^tX_I)^{-1}X_I^t) = \text{tr}((X_I^tX_I)^{-1}X_I^tX_I) = \text{tr}(I_k) = k.\end{aligned}$$

□

Lemma A.3 (Generalized degrees of freedom of the James-Stein estimator). *The generalized degrees of freedom of the James-Stein estimator are equal to*

$$\widehat{df}(X_I\hat{\beta}_I^{JS}) = k - \frac{(k-2)^2}{\|X_I\hat{\beta}_I^{LS}\|^2}.$$

Proof. By definition of the James-Stein estimator, we have that

$$\begin{aligned}
\widehat{df}(X_I \hat{\beta}_I^{JS}) &= \text{div}_Y \left(\left(1 - \frac{(k-2)}{\|X_I \hat{\beta}_I^{LS}\|^2} \right) X_I \hat{\beta}_I^{LS} \right) \\
&= k - (k-2) \text{div}_Y \left(\frac{X_I \hat{\beta}_I^{LS}}{\|X_I \hat{\beta}_I^{LS}\|^2} \right) \\
&= k - (k-2) \left(\frac{\text{div}_Y(X_I \hat{\beta}_I^{LS})}{\|X_I \hat{\beta}_I^{LS}\|^2} + \left(\nabla_Y \frac{1}{\|X_I \hat{\beta}_I^{LS}\|^2} \right)^t X_I \hat{\beta}_I^{LS} \right) \\
&= k - (k-2) \left(\frac{k}{\|X_I \hat{\beta}_I^{LS}\|^2} - 2 \left(\frac{X_I \hat{\beta}_I^{LS}}{\|X_I \hat{\beta}_I^{LS}\|^4} \right)^t X_I \hat{\beta}_I^{LS} \right) \\
&= k - \frac{(k-2)^2}{\|X_I \hat{\beta}_I^{LS}\|^2}.
\end{aligned}$$

□

Lemma A.4 (Generalized degrees of freedom of the Ridge regression estimator). *The generalized degrees of freedom of the Ridge regression estimator are equal to*

$$\widehat{df}(X_I \hat{\beta}_I^{RR}) = \sum_{j=1}^k \frac{d_j^2}{d_j^2 + \lambda},$$

where $d_j, j = 1, \dots, k$ are the eigenvalues of the matrix $X_I^t X_I$.

Lemma A.5 (Generalized degrees of freedom of the Lasso). *The generalized degrees of freedom of the Lasso estimator are equal to*

$$\widehat{df}(X \hat{\beta}^{lasso}) = k.$$

Proof. The proof is given in [Zou et al. 2007]. From the expression they give of the nonzero components of Lasso, for a given subset I , we easily obtain

$$\hat{\beta}_I^{lasso} = (X_I^t X_I)^{-1} (X_I^t Y - \lambda \text{sgn}(\hat{\beta}_I^{lasso})) = \hat{\beta}_I^{LS} - \lambda (X_I^t X_I)^{-1} \text{sgn}(\hat{\beta}_I^{lasso}),$$

we can easily see that the right part in parenthesis will have a null derivative, and the divergence of $\hat{\beta}_I^{lasso}$ is equal to that of $\hat{\beta}_I^{LS}$. □

Lemma A.6 (Generalized degrees of freedom of the MCP). *The generalized degrees of freedom of the MCP estimator are equal to*

$$\widehat{df}(X \hat{\beta}^{mcp}) = \text{tr}(X_I (X_I^t X_I + \gamma^{-1} \Upsilon_I)^{-1} X_I^t),$$

where $\Upsilon_I = \text{diag}(\text{sgn}(\hat{\beta}_I^{mcp}))$.

Proof. The proof is given in [Zhang 2010]. □

A.4.2 Numerical computation

For other estimators of β , the generalized degrees of freedom can be computed numerically thanks to the *directional derivative*. We recall the definition of the directional derivative, extracted from [Nocedal & Wright 1999].

Definition A.2 (Directional derivative). *The directional derivative of a function $\hat{f} : \mathbb{R}^n \mapsto \mathbb{R}$ in the direction u is given by*

$$D(f(t); u) = \lim_{\eta \rightarrow 0} \frac{f(t + \eta u) - f(t)}{\eta}. \quad (\text{A.7})$$

The components of the gradient of a function f , namely $(\nabla f)_i = \partial f / \partial t_i$, $1 \leq i \leq n$, are defined as the directional derivative of f in the direction $p = e_i$ where e_i is the i^{th} vector of the canonical basis, that is,

$$(e_i)_j = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases}$$

Now, going back to the definition of the divergence, we have that

$$\text{div}_Y(X\hat{\beta}) = \sum_{i=1}^n \frac{\partial (X\hat{\beta})_i}{\partial Y_i}.$$

Hence, we can compute the generalized degrees of freedom of $X\hat{\beta}$ by

$$\widehat{df}(X\hat{\beta}) = \sum_{i=1}^n D((X\hat{\beta}(Y))_i; e_i). \quad (\text{A.8})$$

A.5 More results on the simulation study

This appendix gives all the results we obtained for the simulation study in Section [6.2](#).

A.5.1 Loss estimators with Student distribution

Noise level σ estimated on the full model by $\hat{\sigma}_{full}^2$

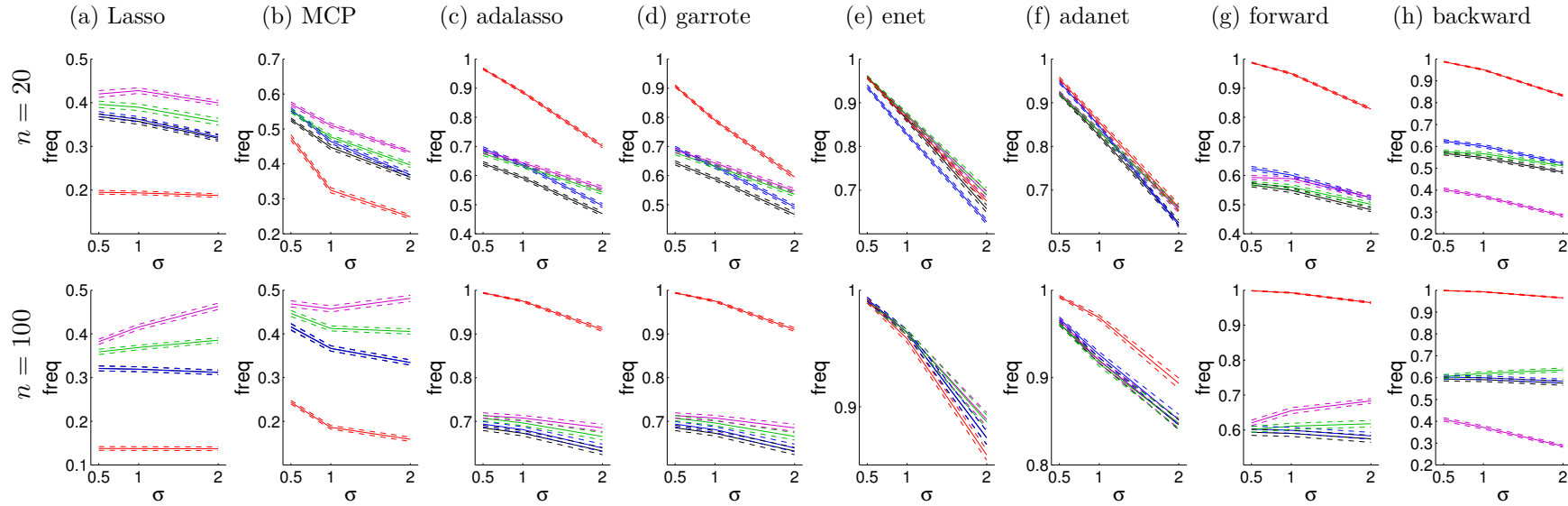


Figure A.1: Average frequency of recovery of the true subset under the full model assumption for the unbiased loss estimator \hat{L}_0 with independent estimator $\hat{\sigma}_{full}^2$ of the variance (black line), for the invariant unbiased loss estimator \hat{L}_0^{inv} (blue line) and for the corrected estimator \hat{L}_{γ}^f with correction function γ_f with constant c_f^* (magenta line) and \hat{c}_f (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, and the bottom row to $n = 100$. Each column corresponds to a method for constructing the collection of models.

Noise level σ estimated on the restricted model by $\hat{\sigma}_{restricted}^2$

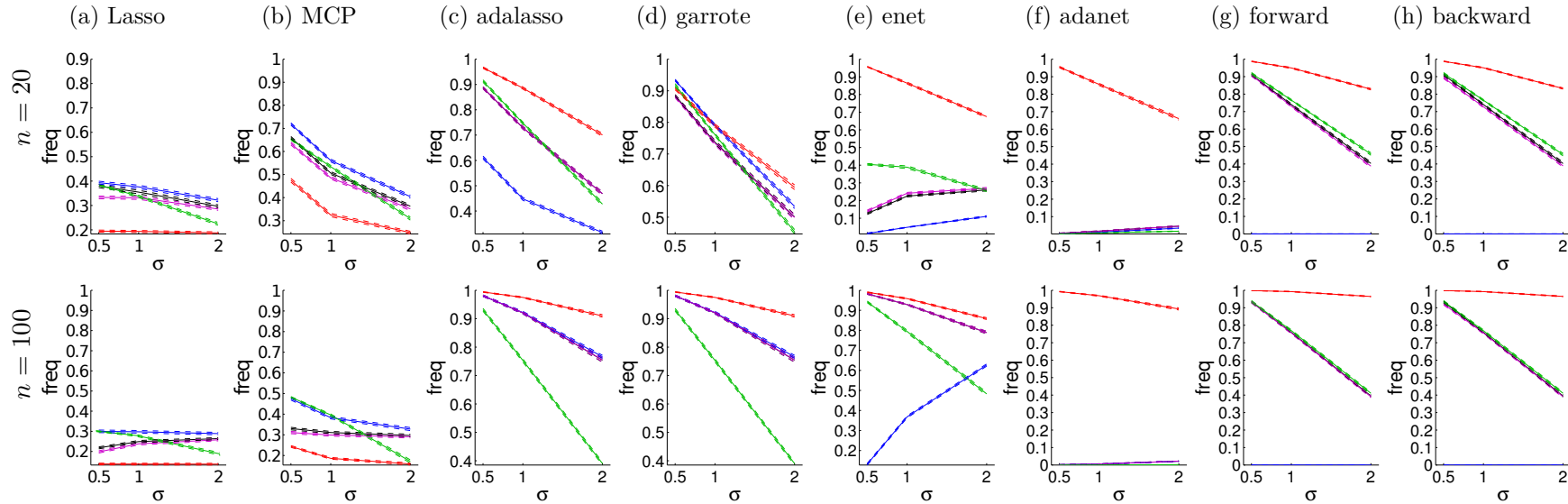


Figure A.2: Average frequency of recovery of the true subset under the restricted model assumption for the unbiased loss estimator \hat{L}_0 with independent estimator $\hat{\sigma}_{restricted}^2$ of the variance (black line), for the invariant unbiased loss estimator \hat{L}_0^{inv} (blue line), for the corrected estimator \hat{L}_γ^r with correction function γ_r (magenta line) and for the corrected estimator \hat{L}^* (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, and the bottom row to $n = 100$. Each column corresponds to a method for constructing the collection of models.

Noise level σ estimated on the restricted model by $\hat{\sigma}_r^2(\hat{\beta})$

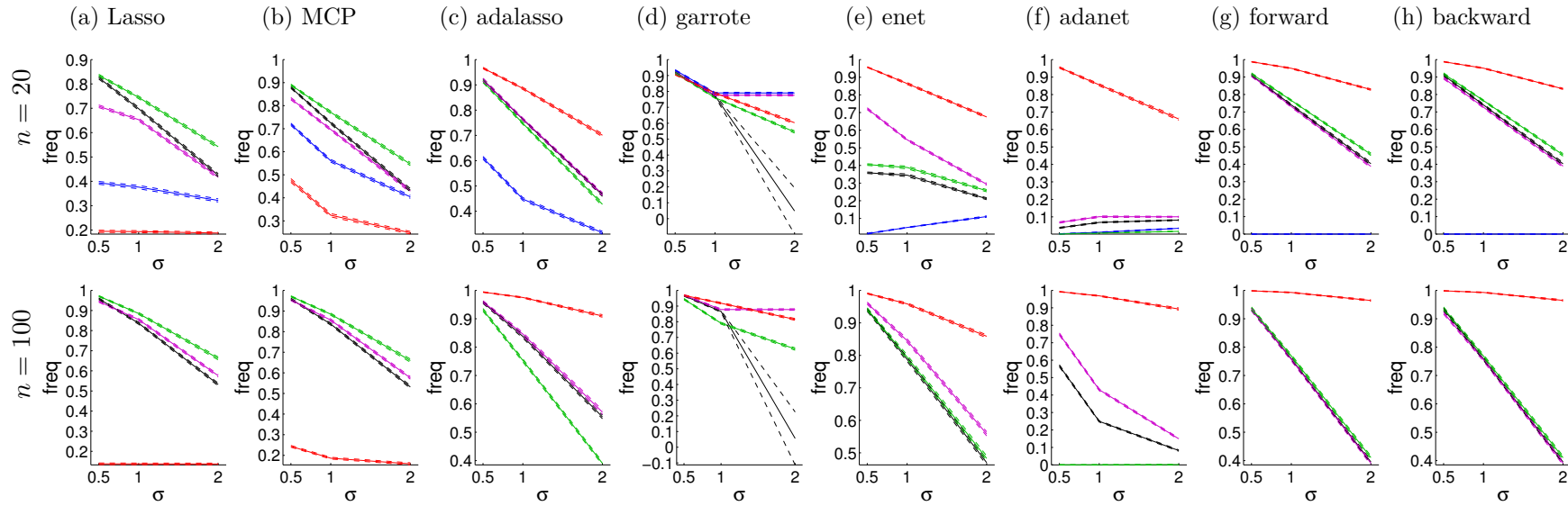


Figure A.3: Average frequency of recovery of the true subset under the restricted model assumption for the loss estimator \hat{L}_0 with dependent estimator $\hat{\sigma}_r^2(\hat{\beta})$ of the variance (black line), for the corrected estimator \hat{L}_γ^r with correction function γ_r (magenta line), and for the corrected estimator \hat{L}^* (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, and the bottom row to $n = 100$. Each column corresponds to a method for constructing the collection of models.

A.5.2 Loss estimators with Kotz distribution

Noise level σ estimated on the full model by $\hat{\sigma}_{full}^2$

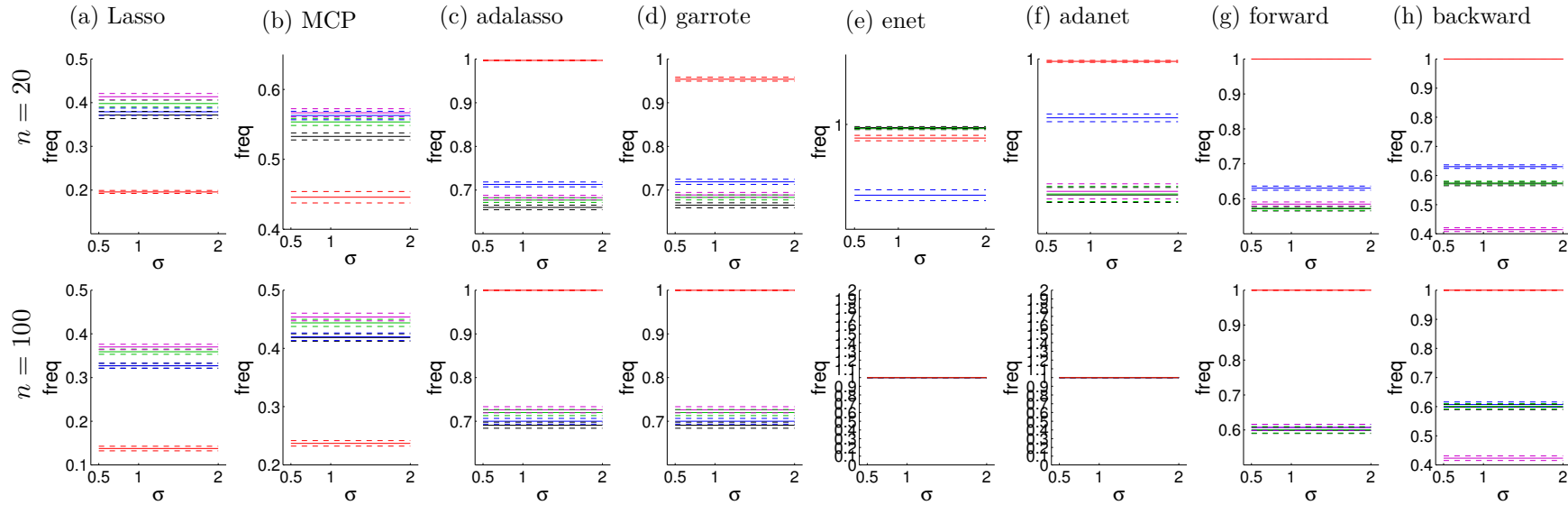


Figure A.4: Average frequency of recovery of the true subset under the full model assumption for the unbiased loss estimator \hat{L}_0 with independent estimator $\hat{\sigma}_{full}^2$ of the variance (black line), for the invariant unbiased loss estimator \hat{L}_0^{inv} (blue line) and for the corrected estimator \hat{L}_{γ}^f with correction function γ_f with constant c_f^* (magenta line) and \hat{c}_f (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, and the bottom row to $n = 100$. Each column corresponds to a method for constructing the collection of models.

Noise level σ estimated on the restricted model by $\hat{\sigma}_{restricted}^2$

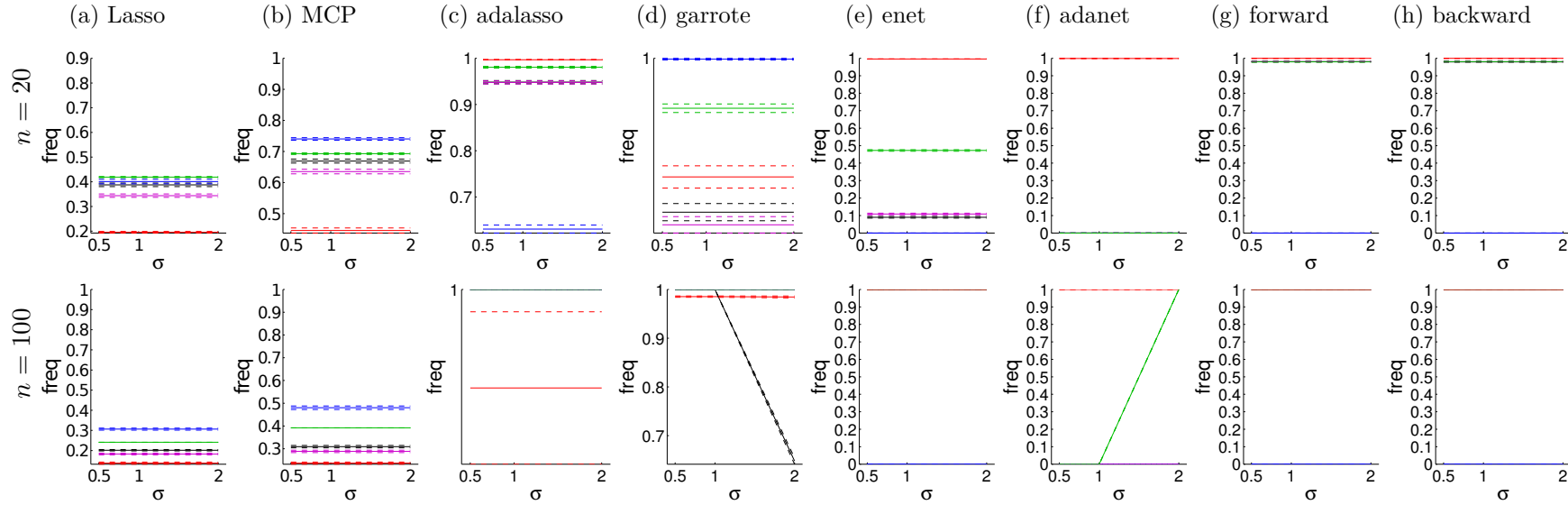


Figure A.5: Average frequency of recovery of the true subset under the restricted model assumption for the unbiased loss estimator \hat{L}_0 with independent estimator $\hat{\sigma}_{restricted}^2$ of the variance (black line), for the invariant unbiased loss estimator \hat{L}_0^{inv} (blue line), for the corrected estimator \hat{L}_γ^r with correction function γ_r (magenta line) and for the corrected estimator \hat{L}^* (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, and the bottom row to $n = 100$. Each column corresponds to a method for constructing the collection of models.

Noise level σ estimated on the restricted model by $\hat{\sigma}_r^2(\hat{\beta})$

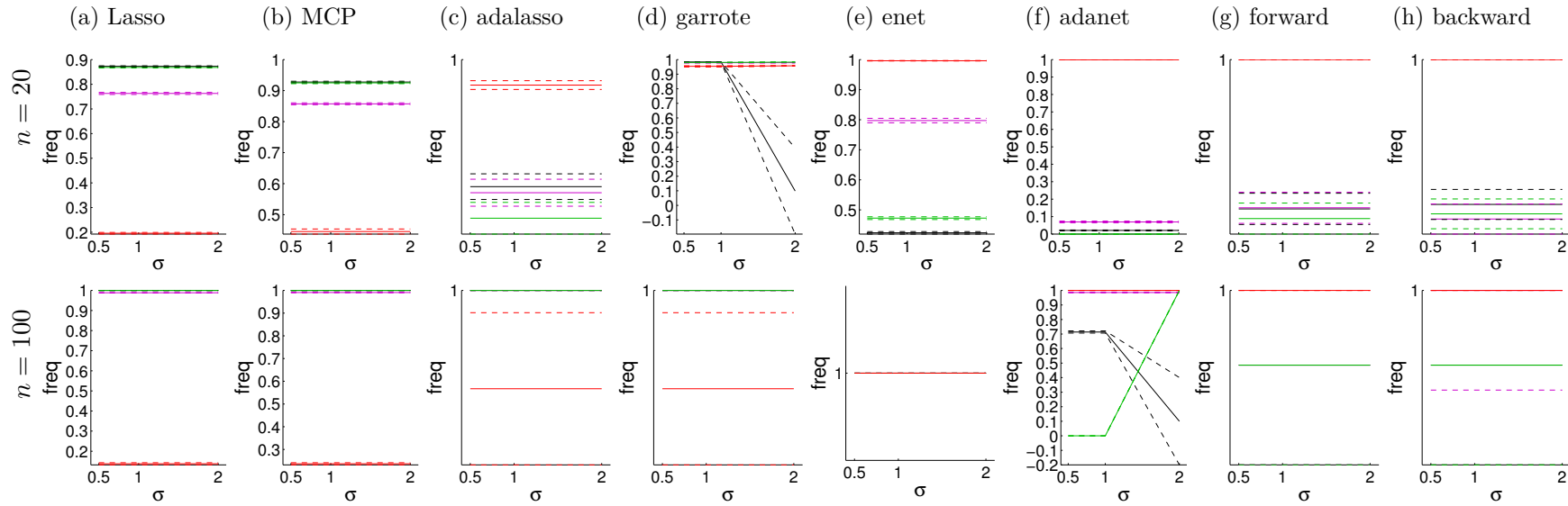


Figure A.6: Average frequency of recovery of the true subset under the restricted model assumption for the loss estimator \hat{L}_0 with dependent estimator $\hat{\sigma}_r^2(\hat{\beta})$ of the variance (black line), for the corrected estimator \hat{L}_γ^r with correction function γ_r (magenta line), and for the corrected estimator \hat{L}^* (green line). The dashed lines display the standard deviation. The true loss has been added in red for reference. The top row corresponds to the case where the sample size n is set to 20, and the bottom row to $n = 100$. Each column corresponds to a method for constructing the collection of models.

A.5.3 Lasso

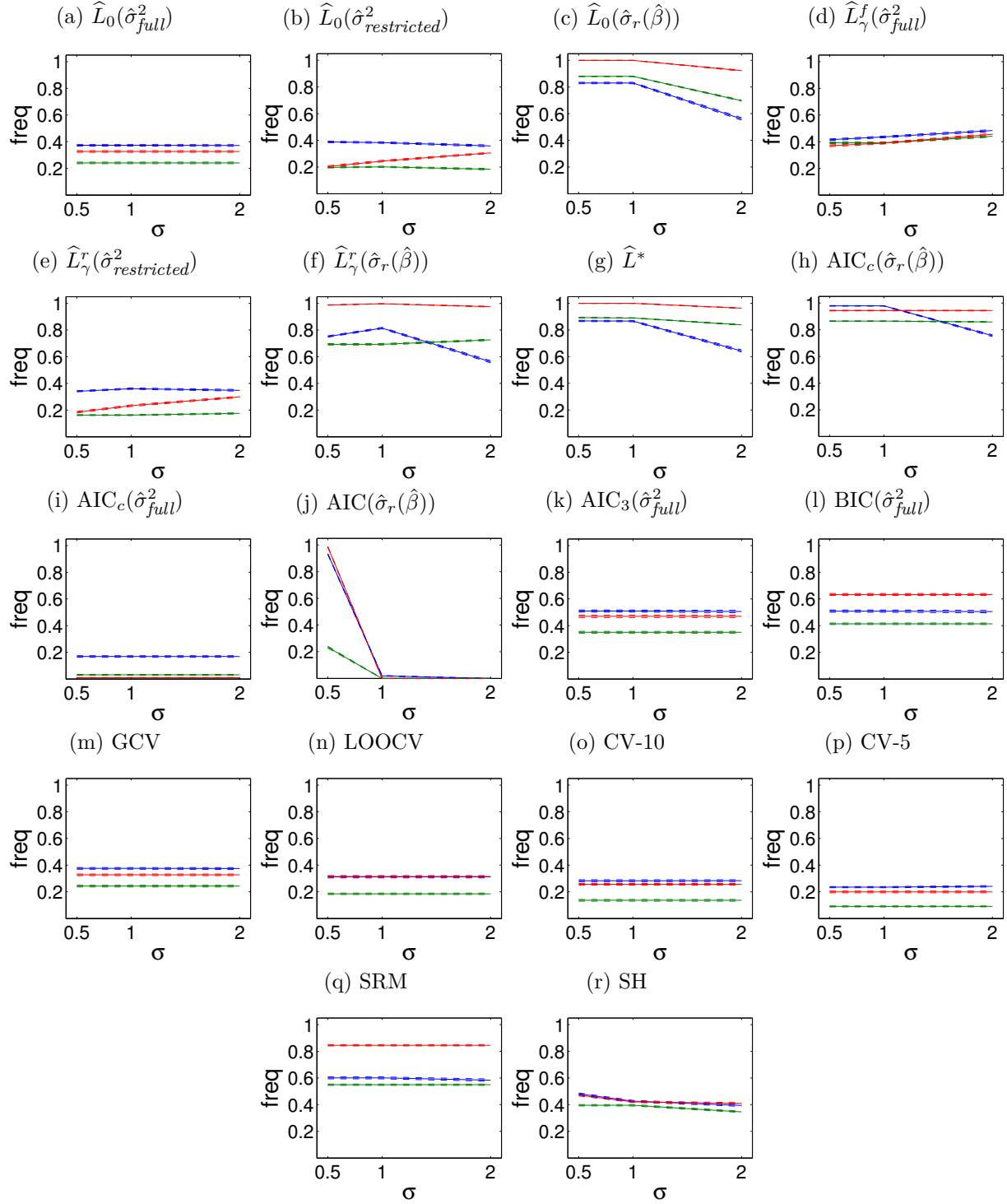


Figure A.7: Average frequency of recovery of the true subset with the Lasso as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red). The dashed lines display the standard deviation.

A.5.4 MCP

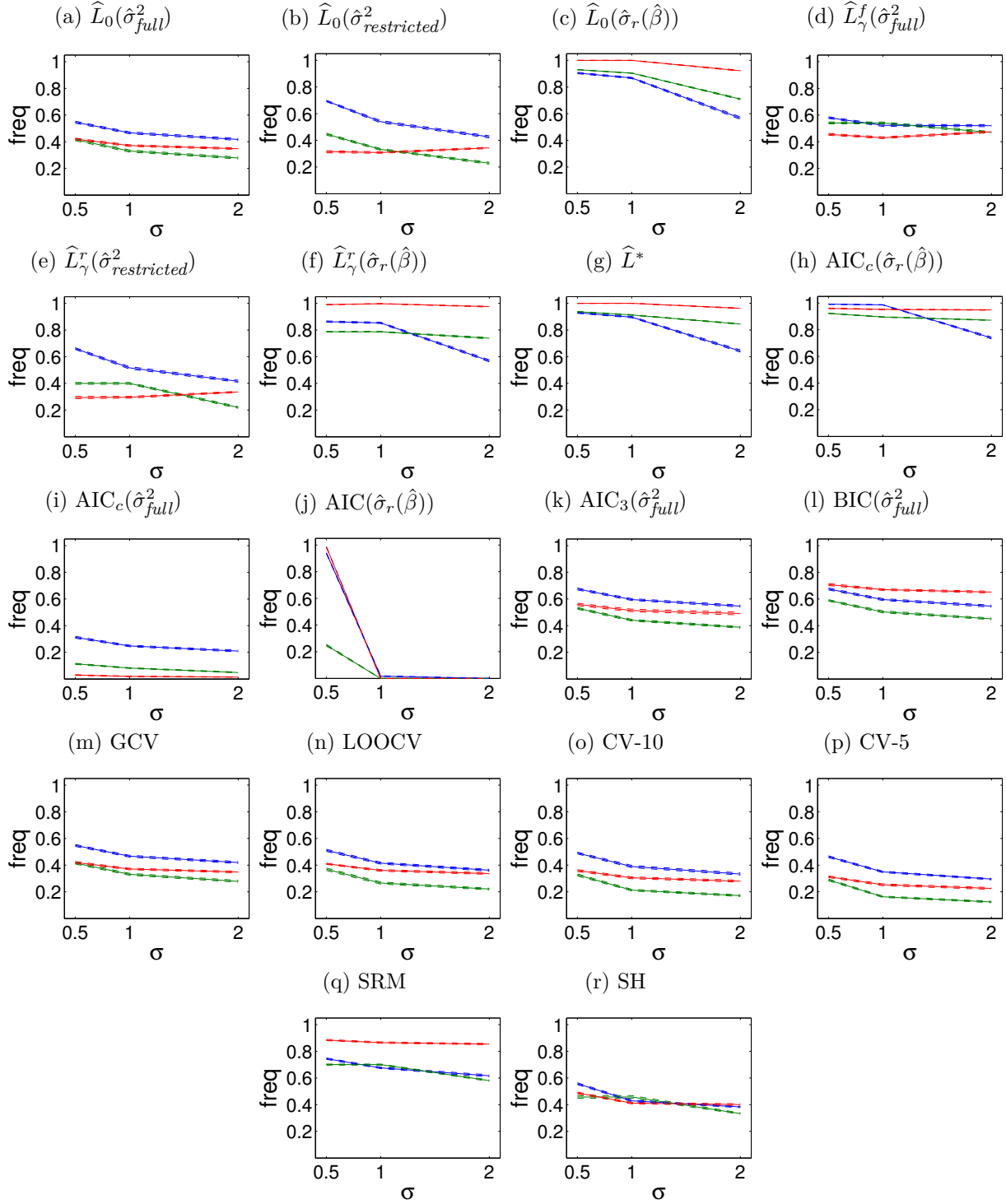


Figure A.8: Average frequency of recovery of the true subset with the MCP as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red). The dashed lines display the standard deviation.

A.5.5 Adaptive lasso

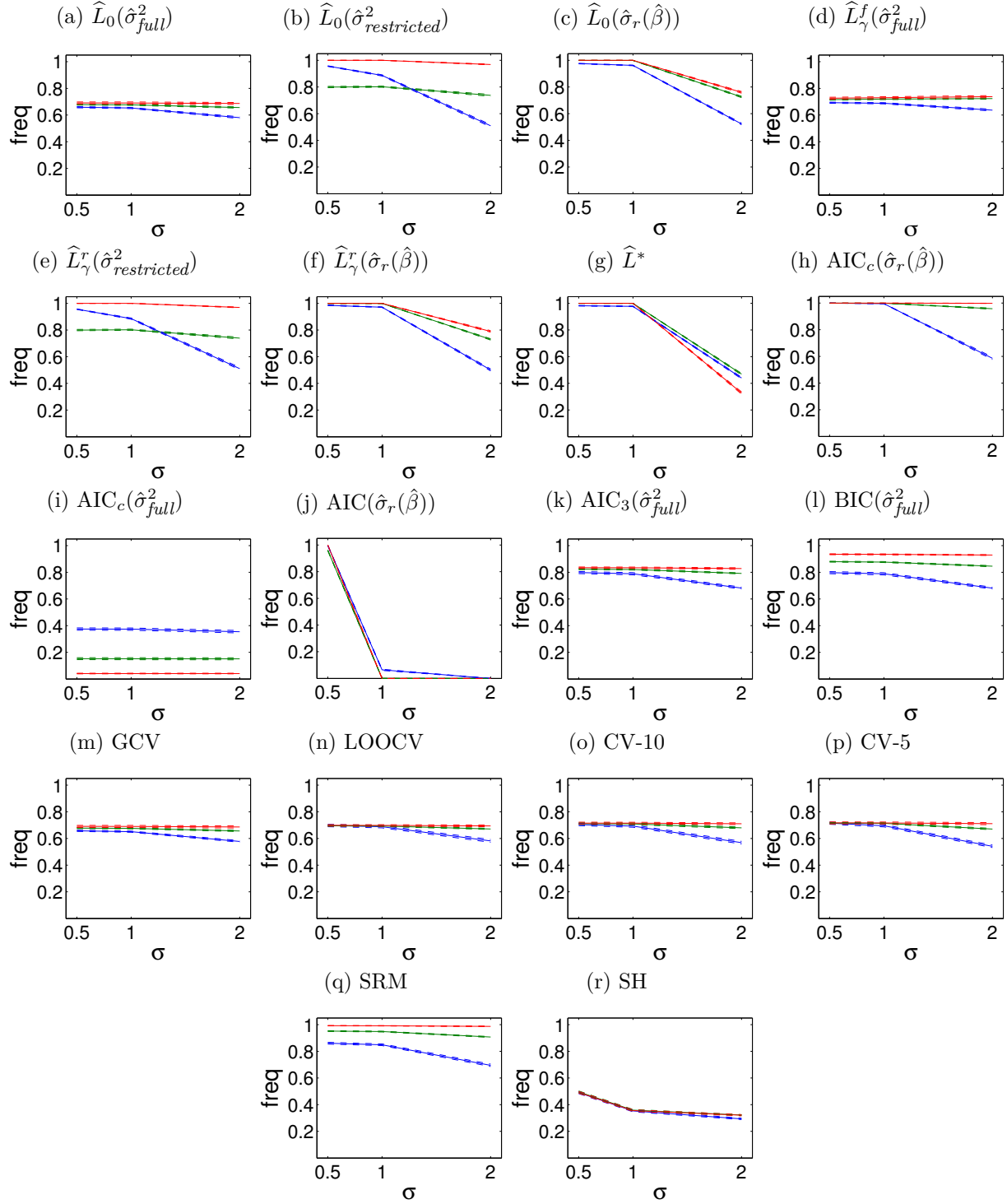


Figure A.9: Average frequency of recovery of the true subset with the Adaptive lasso as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red). The dashed lines display the standard deviation.

A.5.6 Garrote

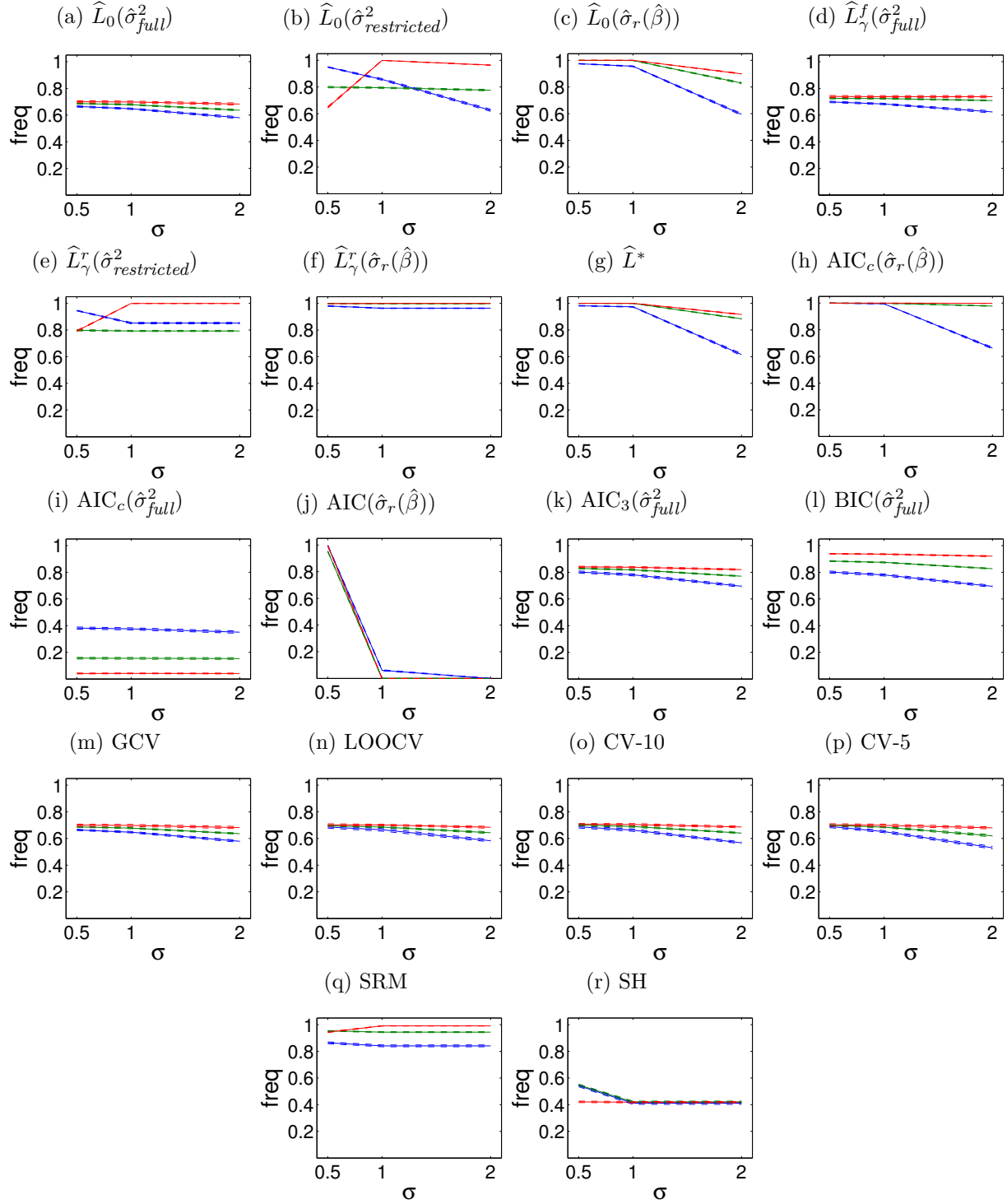


Figure A.10: Average frequency of recovery of the true subset with the Garrote as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red). The dashed lines display the standard deviation.

A.5.7 Elastic net

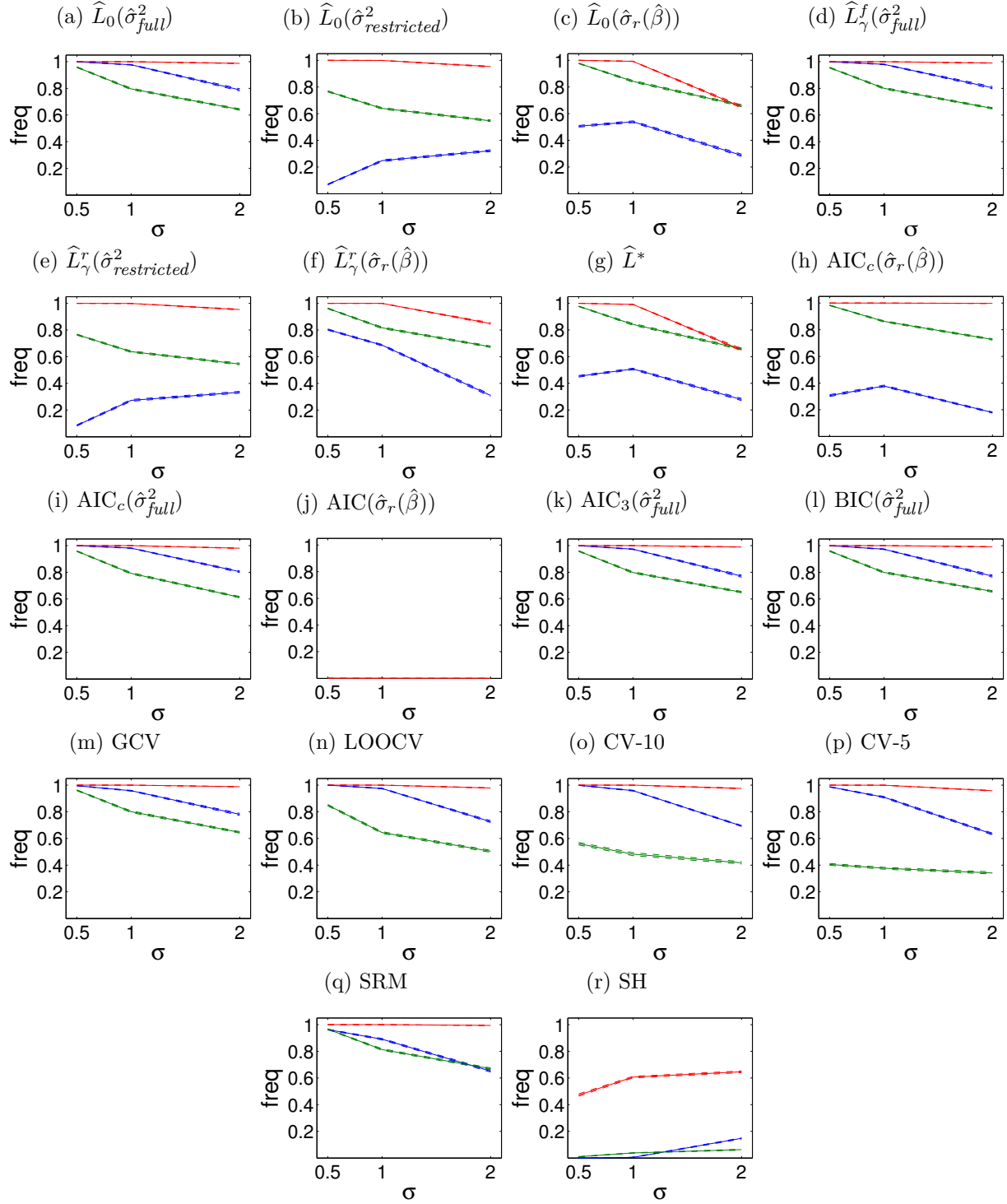


Figure A.11: Average frequency of recovery of the true subset with the Elastic net as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red). The dashed lines display the standard deviation.

A.5.8 Adaptive Elastic net

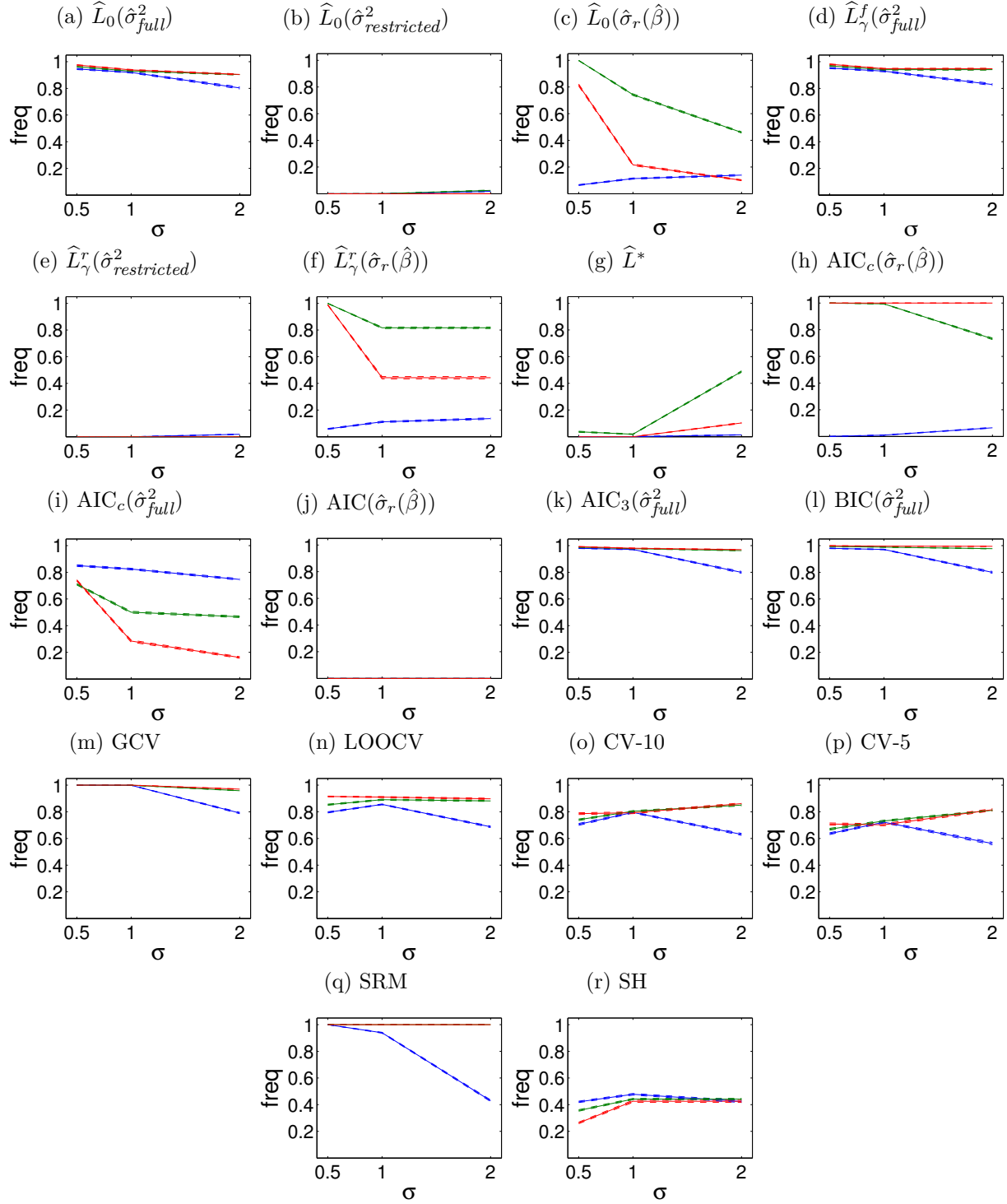


Figure A.12: Average frequency of recovery of the true subset with the Adaptive elastic net as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red). The dashed lines display the standard deviation.

A.5.9 Forward Selection

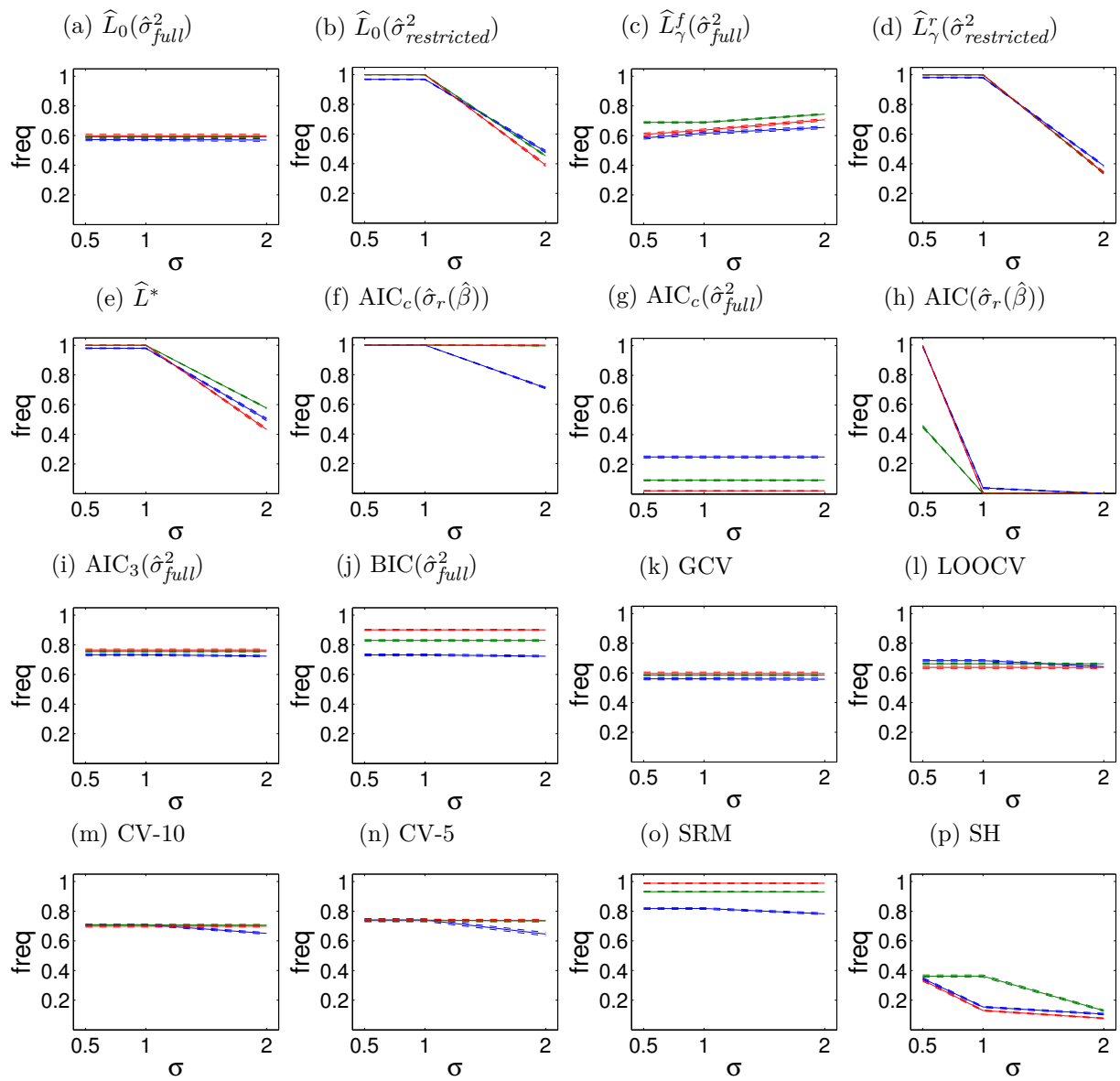


Figure A.13: Average frequency of recovery of the true subset with the Forward selection as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red). The dashed lines display the standard deviation.

A.5.10 Backward elimination

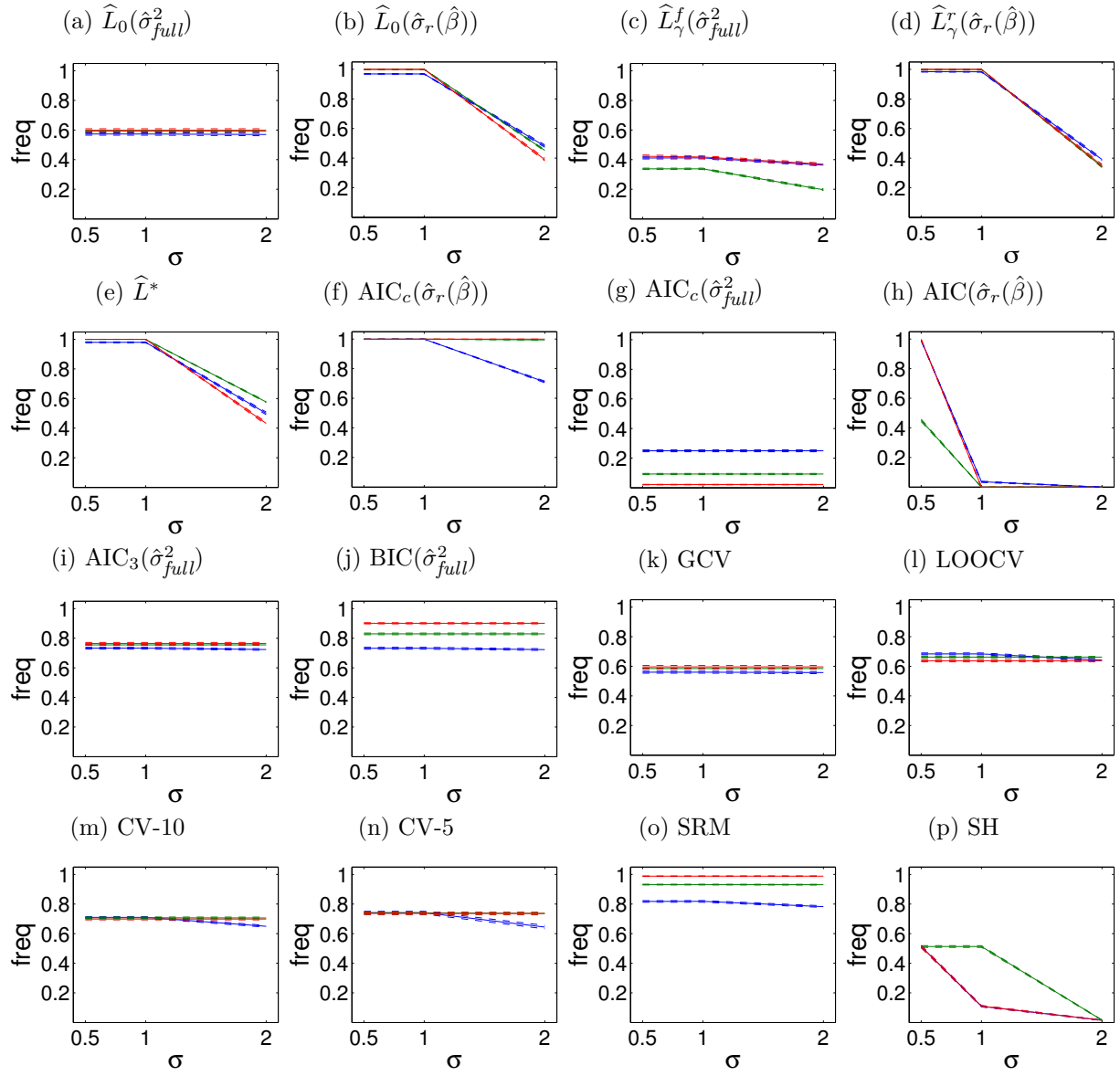


Figure A.14: Average frequency of recovery of the true subset with the Backward elimination as a function of the noise level σ for the sample sizes $n = 20$ (blue), $n = 40$ (green) and $n = 100$ (red). The dashed lines display the standard deviation.

References

- [Akaike 1970] H. Akaike. *Statistical predictor identification*. Annals of the Institute of Statistical Mathematics, vol. 22, no. 1, pages 203–217, 1970. (Cited in pages 28 et 67.)
- [Akaike 1973] H. Akaike. *Information theory and an extension of the maximum likelihood principle*. In Second International Symposium on Information Theory, volume 1, pages 267–281. Akademiai Kiado, 1973. (Cited in page 28.)
- [Akaike 1974] H. Akaike. *A new look at the statistical model identification*. IEEE Transactions on Automatic Control, vol. 19, no. 6, pages 716–723, 1974. (Cited in pages 28, 39 et 61.)
- [Allen 1974] D.M. Allen. *The relationship between variable selection and data agumentation and a method for prediction*. Technometrics, vol. 16, no. 1, pages 125–127, 1974. (Cited in page 36.)
- [An & Tao 2005] L.T.H. An and P.D. Tao. *The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems*. Annals of Operations Research, vol. 133, no. 1, pages 23–46, 2005. (Cited in page 112.)
- [Andrews & Mallows 1974] D.F. Andrews and C.L. Mallows. *Scale mixtures of normal distributions*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 36, pages 99–102, 1974. (Cited in page 124.)
- [Arlot & Bach 2009] S. Arlot and F. Bach. *Data-driven calibration of linear estimators with minimal penalties*. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, editors, Advances in Neural Information Processing Systems 22, pages 46–54. NIPS, 2009. (Cited in page 154.)
- [Arlot & Celisse 2010] S. Arlot and A. Celisse. *A survey of cross-validation procedures for model selection*. Statistics Surveys, vol. 4, pages 40–79, 2010. (Cited in pages 20, 21, 26, 35 et 36.)
- [Arlot & Massart 2009] S. Arlot and P. Massart. *Data-driven calibration of penalties for least-squares regression*. Journal of Machine Learning Research, vol. 10, pages 245–279, 2009. (Cited in pages 34 et 99.)
- [Bach et al. 2011] F. Bach, R. Jenatton, J. Mairal and G. Obozinski. Optimization for Machine Learning, Chapter Convex optimization with sparsity-inducing norms, pages 19–54. MIT Press, 2011. (Cited in pages 41 et 105.)
- [Baraud et al. 2009] Y. Baraud, C. Giraud and S. Huet. *Gaussian model selection with an unknown variance*. The Annals of Statistics, vol. 37, no. 2, pages 630–672, 2009. (Cited in page 127.)

- [Barron 1993] A. Barron. *Universal approximation bounds for superpositions of a sigmoidal function*. IEEE Transactions on Information Theory, vol. 39, pages 291–319, 1993. (Cited in page 20.)
- [Barron 1994] A.R. Barron. *Approximation and estimation bounds for artificial neural networks*. Machine Learning, vol. 14, no. 1, pages 115–133, 1994. (Cited in page 18.)
- [Bartlett *et al.* 2002] P.L. Bartlett, S. Boucheron and G. Lugosi. *Model selection and error estimation*. Machine Learning, vol. 48, no. 1, pages 85–113, 2002. (Cited in page 18.)
- [Bendel & Afifi 1977] R.B. Bendel and A.A. Afifi. *Comparison of stopping rules in forward "stepwise" regression*. Journal of the American Statistical Association, vol. 72, pages 46–53, 1977. (Cited in page 40.)
- [Bennett *et al.* 2006] K.P. Bennett, J. Hu, X. Ji, G. Kunapuli and J.S. Pang. *Model selection via bilevel optimization*. In Neural Networks, 2006. IJCNN'06. International Joint Conference on, pages 1922–1929. IEEE, 2006. (Cited in page 49.)
- [Bennett *et al.* 2008] Kristin Bennett, Gautam Kunapuli, Jing Hu and Jong-Shi Pang. *Bilevel Optimization and Machine Learning*. In Jacek Zurada, Gary Yen and Jun Wang, editors, Computational Intelligence: Research Frontiers, volume 5050 of *Lecture Notes in Computer Science*, pages 25–47. Springer Berlin / Heidelberg, 2008. (Cited in page 49.)
- [Berger 1985] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985. (Cited in pages 15, 18 et 56.)
- [Berk 1997] J.B. Berk. *Necessary conditions for the CAPM*. Journal of Economic Theory, vol. 73, no. 1, pages 245–257, 1997. (Cited in page 79.)
- [Bertsekas *et al.* 2003] D.P. Bertsekas, A. Nedić and A.E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific optimization and computation series. Athena Scientific, 2003. (Cited in page 105.)
- [Biernacki 1997] C. Biernacki. *Choix de modèles en classification*. PhD thesis, Université de Technologie de Compiègne, 1997. (Cited in page 31.)
- [Birgé & Massart 2001] L. Birgé and P. Massart. *Gaussian model selection*. Journal of the European Mathematical Society, vol. 3, no. 3, pages 203–268, 2001. (Cited in pages 31 et 33.)
- [Birgé & Massart 2007] L. Birgé and P. Massart. *Minimal penalties for Gaussian model selection*. Probability Theory and Related Fields, vol. 138, no. 1, pages 33–73, 2007. (Cited in pages iv, 33, 35, 63, 85, 99 et 100.)
- [Bontemps & Toussile 2010] Dominique Bontemps and Wilson Toussile. *Clustering et sélection de variables sur des données génétiques*. In 42èmes Journées de Statistique, Marseille, France, France, 2010. (Cited in page 33.)
- [Boucheron *et al.* 2005] S. Boucheron, O. Bousquet and G. Lugosi. *Theory of classification: A survey of some recent advances*. ESAIM Probability and Statistics, vol. 9, pages 323–375, 2005. (Cited in pages 18 et 157.)

- [Boyd & Vandenberghe 2004] S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ Pr, 2004. (Cited in page 105.)
- [Bozdogan 1987] H. Bozdogan. *Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions*. Psychometrika, vol. 52, no. 3, pages 345–370, 1987. (Cited in page 31.)
- [Bozdogan 1994] H. Bozdogan. *Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity*. In H. Bozdogan, editor, *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling*, volume 2, *Multivariate Statistical Modeling*, pages 69–113. Kluwer Academic Publishers, 1994. (Cited in page 31.)
- [Bozdogan 2000] H. Bozdogan. *Akaike's information criterion and recent developments in information complexity*. Journal of mathematical psychology, vol. 44, no. 1, pages 62–91, 2000. (Cited in pages 20 et 29.)
- [Brandwein & Strawderman 1991] A.C. Brandwein and W.E. Strawderman. *Generalizations of James-Stein estimators under spherical symmetry*. Annals of Statistics, vol. 19, no. 3, pages 1639–1650, 1991. (Cited in page 69.)
- [Breheny & Huang 2011] P. Breheny and J. Huang. *Coordinate descent algorithms for non-convex penalized regression, with applications to biological feature selection*. Annals of Applied Statistics, vol. 5, no. 1, page 232, 2011. (Cited in page 110.)
- [Breiman & Friedman 1985] L. Breiman and J.H. Friedman. *Estimating optimal transformations for multiple regression and correlation*. Journal of the American Statistical Association, vol. 80, pages 580–598, 1985. (Cited in page 23.)
- [Breiman 1995] L. Breiman. *Better Subset Regression Using the Nonnegative Garrote*. Technometrics, vol. 37, no. 4, pages 373–384, 1995. (Cited in page 47.)
- [Breiman 1996] L. Breiman. *Heuristics of instability and stabilization in model selection*. Annals of Statistics, vol. 24, no. 6, pages 2350–2383, 1996. (Cited in pages 13 et 18.)
- [Brown 1986] L.D. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Lecture notes-monograph series. Institute of Mathematical Statistics, 1986. (Cited in page 72.)
- [Bruce & Gao 1996] A.G. Bruce and H.Y. Gao. *Understanding WaveShrink: Variance and bias estimation*. Biometrika, vol. 83, no. 4, pages 727–745, 1996. (Cited in page 46.)
- [Bunea & Wegkamp 2004] F. Bunea and M.H. Wegkamp. *Two-stage model selection procedures in partially linear regression*. Canadian Journal of Statistics, vol. 32, no. 2, pages 105–118, 2004. (Cited in page 37.)
- [Burnham & Anderson 2002] K.P. Burnham and D.R. Anderson. *Model Selection and Multi-model Inference: a Practical Information-Theoretic Approach*. Springer Verlag, 2002. (Cited in pages 29, 37 et 66.)

- [Caillerie & Michel 2009] Claire Caillerie and Bertrand Michel. *Model selection for simplicial approximation*. Rapport de recherche RR-6981, INRIA, 2009. (Cited in page 34.)
- [Candès 2006] E.J. Candès. *Modern statistical estimation via oracle inequalities*. Acta Numerica, vol. 15, pages 257–326, 2006. (Cited in page 18.)
- [Chaslot *et al.* 2008] G. Chaslot, S. Bakkes, I. Szita and P. Spronck. *Monte-Carlo Tree Search: A new framework for game AI*. In Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference, pages 216–217, 2008. (Cited in page 49.)
- [Cherkassky & Ma 2003] V. Cherkassky and Y. Ma. *Comparison of model selection for regression*. Neural Computation, vol. 15, no. 7, pages 1691–1714, 2003. (Cited in pages 62, 147 et 154.)
- [Cherkassky & Mulier 1998] V.S. Cherkassky and F. Mulier. Learning from data: Concepts, Theory, and Methods. John Wiley & Sons, Inc., 1998. (Cited in pages 12, 14 et 26.)
- [Cherkassky *et al.* 1999] V. Cherkassky, X. Shao, F.M. Mulier and V.N. Vapnik. *Model complexity control for regression using VC generalization bounds*. IEEE Transactions on Neural Networks, vol. 10, no. 5, pages 1075–1089, 1999. (Cited in page 32.)
- [Chmielewski 1981] M.A. Chmielewski. *Elliptically symmetric distributions: A review and bibliography*. International Statistical Review/Revue Internationale de Statistique, vol. 49, pages 67–74, 1981. (Cited in pages 72 et 78.)
- [Claeskens & Hjort 2008] G. Claeskens and N.L. Hjort. Model Selection and Model Averaging. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press, 2008. (Cited in page 66.)
- [Clarke 1990] F.H. Clarke. Optimization and nonsmooth analysis, volume 5 of *Classics In Applied Mathematics*. Society for Industrial and Applied Mathematics, 1990. (Cited in pages 111 et 112.)
- [Devroye 1986] Luc Devroye. Non-Uniform Random Variate Generation. Springer-Verlag, 1986. (Cited in pages 118 et 119.)
- [Donoho & Johnstone 1994] D.L. Donoho and I.M. Johnstone. *Ideal spatial adaptation by wavelet shrinkage*. Biometrika, vol. 81, no. 3, page 425, 1994. (Cited in pages 18 et 43.)
- [Dornhege *et al.* 2007] G. Dornhege, J.R. Millán, T. Hinterberger, D. McFarland and K.R. Müller. Toward brain-computer interfacing, volume 74. MIT press Cambridge, MA, 2007. (Cited in page 1.)
- [Dramiński *et al.* 2010] M. Dramiński, M. Kierczak, J. Koronacki and J. Komorowski. *Monte Carlo feature selection and interdependency discovery in supervised classification*. In Jacek Koronacki, Zbigniew Ras, Slawomir Wierzchon and Janusz Kacprzyk, editors, Advances in Machine Learning II, volume 263 of *Studies in Computational Intelligence*, pages 371–385. Springer Berlin / Heidelberg, 2010. (Cited in page 49.)
- [Du & Ma 2011] J. Du and C. Ma. *Spherically invariant vector random fields in space and time*. IEEE Transactions on Signal Processing, vol. 59, no. 12, pages 5921–5929, 2011. (Cited in page 78.)

- [Efron *et al.* 2004] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. *Least angle regression (with discussions and authors reply)*. Annals of Statistics, vol. 32, no. 2, pages 407–451, 2004. (Cited in pages [iv](#), [42](#), [43](#), [48](#), [50](#), [103](#) et [155](#).)
- [Efron 1986] B. Efron. *How biased is the apparent error rate of a prediction rule?* Journal of the American Statistical Association, vol. 81, no. 394, pages 461–470, 1986. (Cited in page [62](#).)
- [Efron 2004] B. Efron. *The Estimation of Prediction Error*. Journal of the American Statistical Association, vol. 99, no. 467, pages 619–632, 2004. (Cited in pages [35](#) et [64](#).)
- [El Anbari 2011] Mohammed El Anbari. *Regularization and variable selection using penalized likelihood*. PhD thesis, Université Paris–Sud 11 et Université Cadi Ayyad, 2011. (Cited in page [127](#).)
- [Fan & Fang 1985] JQ Fan and KT Fang. *Inadmissibility of sample mean and regression coefficients for elliptically contoured distributions*. Northeastern Mathematical Journal, vol. 1, pages 68–81, 1985. (Cited in page [72](#).)
- [Fan & Li 2001] J. Fan and R. Li. *Variable selection via nonconcave penalized likelihood and its oracle properties*. Journal of the American Statistical Association, vol. 96, no. 456, pages 1348–1360, 2001. (Cited in pages [45](#), [110](#) et [155](#).)
- [Fan & Tang 2012] Y. Fan and C.Y. Tang. *Tuning Parameter Selection in High-Dimensional Penalized Likelihood*. Journal of the Royal Statistical Society Series B, 2012. To appear. (Cited in page [40](#).)
- [Fang *et al.* 1989] K.T. Fang, S. Kotz and K.W. Ng. Symmetric Multivariate and Related Distributions, volume 36 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, 1989. (Cited in pages [72](#), [74](#), [77](#), [119](#), [121](#) et [124](#).)
- [Feller 1966] W. Feller. *An Introduction to Probability Theory and its Applications*. Series in Probability and Mathematical Statistics. John Wiley & Sons, 1966. (Cited in page [124](#).)
- [Févotte *et al.* 2009] C. Févotte, N. Bertin and J.L. Durrieu. *Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis*. Neural Computation, vol. 21, no. 3, pages 793–830, 2009. (Cited in page [59](#).)
- [Flamary 2011] R. Flamary. *Apprentissage statistique pour le signal: applications aux interfaces cerveau-machine*. PhD thesis, Université de Rouen, 2011. (Cited in page [105](#).)
- [Foster & George 1994] D.P. Foster and E.I. George. *The risk inflation criterion for multiple regression*. Annals of Statistics, vol. 22, no. 4, pages 1947–1975, 1994. (Cited in page [31](#).)
- [Fourdrinier & Strawderman 2003] D. Fourdrinier and W.E. Strawderman. *On Bayes and unbiased estimators of loss*. Annals of the Institute of Statistical Mathematics, vol. 55, no. 4, pages 803–816, 2003. (Cited in pages [60](#) et [157](#).)
- [Fourdrinier & Strawderman 2008] D. Fourdrinier and W.E. Strawderman. *Generalized Bayes minimax estimators of location vectors for spherically symmetric distributions*. Journal of Multivariate Analysis, vol. 99, no. 4, pages 735–750, 2008. (Cited in page [79](#).)

- [Fourdrinier & Strawderman 2010] D. Fourdrinier and W.E. Strawderman. *Robust generalized Bayes minimax estimators of location vectors for spherically symmetric distribution with unknown scale*. Borrowing Strength: Theory Powering Applications-A Festschrift for Lawrence D. Brown, vol. 6, pages 249–262, 2010. (Cited in pages 69 et 83.)
- [Fourdrinier & Wells 1994] D. Fourdrinier and MT Wells. *Comparaisons de procédures de sélection d'un modèle de régression: une approche décisionnelle*. Comptes rendus de l'Académie des sciences. Série 1, Mathématique, vol. 319, no. 8, pages 865–870, 1994. (Cited in pages 67, 87, 128 et 140.)
- [Fourdrinier & Wells 1995a] D. Fourdrinier and M.T. Wells. *Estimation of a loss function for spherically symmetric distributions in the general linear model*. Annals of Statistics, vol. 23, no. 2, pages 571–592, 1995. (Cited in pages 82, 87 et 91.)
- [Fourdrinier & Wells 1995b] D. Fourdrinier and M.T. Wells. *Loss Estimation for Spherically Symmetrical Distributions*. Journal of multivariate analysis, vol. 53, no. 2, pages 311–331, 1995. (Cited in page 87.)
- [Fourdrinier & Wells 2012] D. Fourdrinier and M.T. Wells. *On Improved Loss Estimation for Shrinkage Estimators*. Statistical Science, vol. 27, no. 1, pages 61–81, 2012. (Cited in pages 60, 69, 88, 91, 92 et 157.)
- [Fourdrinier et al. 2012] D. Fourdrinier, W.E. Strawderman and M.T Wells. *Shrinkage estimation*, 2012. to appear in 2013. (Cited in pages 57, 58 et 118.)
- [Friedman 1994] J.H. Friedman. From Statistics to Neural Networks. Theory and Pattern Recognition Applications, Chapter An overview of predictive learning and function approximation, pages 1–61. Springer Verlag, 1994. (Cited in pages 12, 13 et 14.)
- [Gasso et al. 2009] G. Gasso, A. Rakotomamonjy and S. Canu. *Recovering sparse signals with a certain family of nonconvex penalties and DC programming*. Signal Processing, IEEE Transactions on, vol. 57, no. 12, pages 4686–4698, 2009. (Cited in page 112.)
- [Gaudel & Sebag 2010] R. Gaudel and M. Sebag. *Feature selection as a one-player game*. In International Conference on Machine Learning, pages 359–366, 2010. (Cited in page 49.)
- [Geisser 1975] S. Geisser. *The predictive sample reuse method with applications*. Journal of the American Statistical Association, vol. 70, no. 350, pages 320–328, 1975. (Cited in page 35.)
- [Genton 2000] M.G. Genton. *The correlation structure of Matheron's classical variogram estimator under elliptically contoured distributions*. Mathematical Geology, vol. 32, no. 1, pages 127–137, 2000. (Cited in page 79.)
- [George & Foster 2000] E.I. George and D.P. Foster. *Calibration and empirical Bayes variable selection*. Biometrika, vol. 87, no. 4, pages 731–747, 2000. (Cited in page 157.)
- [George 2000] E.I. George. *The Variable Selection Problem*. Journal of the American Statistical Association, vol. 95, no. 452, pages 1304–1308, 2000. (Cited in page 14.)

- [Gneiting *et al.* 2007] T. Gneiting, M.G. Genton and P. Guttorp. Statistical Methods for Spatio-Temporal Systems, volume 107 of *Monographs on Statistics and Applied Probability*, Chapter Geostatistical space-time models, stationarity, separability, and full symmetry, pages 151–175. Chapman & Hall, 2007. (Cited in page 79.)
- [Golub & Van Loan 1996] G.H. Golub and C.F. Van Loan. Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 1996. (Cited in page 80.)
- [Golub *et al.* 1979] G.H. Golub, M. Heath and G. Wahba. *Generalized cross-validation as a method for choosing a good ridge parameter*. Technometrics, vol. 21, no. 2, pages 215–223, 1979. (Cited in page 29.)
- [Govaert & Nadif 2010] G. Govaert and M. Nadif. *Latent block model for contingency table*. Communications in Statistics – Theory and Methods, vol. 39, no. 3, pages 416–425, 2010. (Cited in page 157.)
- [Gupta & Varga 1993] A.K. Gupta and T. Varga. Elliptically Contoured Models in Statistics. Kluwer Academic Publishers, 1993. (Cited in page 74.)
- [Guyon *et al.* 2010] I. Guyon, A. Saffari, G. Dror and G. Cawley. *Model selection: Beyond the bayesian/frequentist divide*. Journal of Machine Learning Research, vol. 11, pages 61–87, 2010. (Cited in pages 13 et 14.)
- [Guyon 2009] I. Guyon. Machine learning summer school, Chapter A practical guide to model selection. Springer, 2009. (Cited in pages 20 et 50.)
- [Hafner & Rombouts 2007] C.M. Hafner and J.V.K. Rombouts. *Semiparametric multivariate volatility models*. Econometric Theory, vol. 23, no. 02, pages 251–280, 2007. (Cited in page 79.)
- [Hager 1989] W.W. Hager. *Updating the inverse of a matrix*. SIAM review, vol. 31, no. 2, pages 221–239, 1989. (Cited in page 160.)
- [Hannan & Quinn 1979] E.J. Hannan and B.G. Quinn. *The determination of the order of an autoregression*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 41, no. 2, pages 190–195, 1979. (Cited in page 31.)
- [Hare & Sagastizábal 2009] W. Hare and C. Sagastizábal. *Computing proximal points of non-convex functions*. Mathematical Programming, vol. 116, no. 1, pages 221–258, 2009. (Cited in page 112.)
- [Hastie & Tibshirani 1990] T.J. Hastie and R.J. Tibshirani. Generalized Additive Models. Chapman & Hall/CRC, 1990. (Cited in pages 20 et 27.)
- [Hastie *et al.* 2005] T.J. Hastie, R.J. Tibshirani, J. Friedman and J. Franklin. The Elements of Statistical Learning: Data mining, Inference and Prediction, volume 27. Springer, 2005. (Cited in pages 12 et 23.)
- [Hastie *et al.* 2008] T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd Edition), volume 1. Springer Series in Statistics, 2008. (Cited in pages 14, 17, 18, 20, 23 et 154.)

- [Hocking 1976] R.R. Hocking. *A Biometrics invited paper. The analysis and selection of variables in linear regression*. Biometrics, vol. 32, no. 1, pages 1–49, 1976. (Cited in pages 2 et 29.)
- [Hoerl & Kennard 1970] A.E. Hoerl and R.W. Kennard. *Ridge regression: applications to nonorthogonal problems*. Technometrics, vol. 12, no. 1, pages 69–82, 1970. (Cited in page 49.)
- [Huber 1964] P.J. Huber. *Robust estimation of a location parameter*. The Annals of Mathematical Statistics, vol. 35, no. 1, pages 73–101, 1964. (Cited in page 25.)
- [Huber 1975] P.J. Huber. A Survey of Statistical Design and Linear Models, Chapter Robustness and designs, pages 287–303. North Holland, Amsterdam, 1975. (Cited in page 72.)
- [Huber 1981] P.J. Huber. Robust Statistics, volume 67 of *Wiley series in probability and mathematical statistics*. John Wiley & Sons, Inc., 1981. (Cited in page 77.)
- [Hurvich & Tsai 1989] C.M. Hurvich and C.L. Tsai. *Regression and time series model selection in small samples*. Biometrika, vol. 76, no. 2, pages 297–307, 1989. (Cited in pages 29 et 71.)
- [Hurvich & Tsai 1991] C.M. Hurvich and C.L. Tsai. *Bias of the corrected AIC criterion for underfitted regression and time series models*. Biometrika, vol. 78, no. 3, pages 499–509, 1991. (Cited in page 71.)
- [Hurvich & Tsai 1993] C.M. Hurvich and C.L. Tsai. *A corrected Akaike information criterion for vector autoregressive model selection*. Journal of time series analysis, vol. 14, no. 3, pages 271–279, 1993. (Cited in page 71.)
- [Hurvich et al. 1990] C.M. Hurvich, R. Shumway and C.L. Tsai. *Improved estimators of Kullback–Leibler information for autoregressive model selection in small samples*. Biometrika, vol. 77, no. 4, pages 709–719, 1990. (Cited in page 29.)
- [James & Stein 1961] W. James and C. Stein. *Estimation with quadratic loss*. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, June 20–July 30, 1960, page 361. University of California Press, 1961. (Cited in pages 48 et 56.)
- [Johnstone 1988] I.M. Johnstone. *On inadmissibility of some unbiased estimates of loss*. Statistical Decision Theory and Related Topics, vol. 4, no. 1, pages 361–379, 1988. (Cited in pages 55, 57, 59, 60, 84, 86, 87, 91, 93 et 94.)
- [Kariya & Sinha 1989] T. Kariya and B.K. Sinha. Robustness of Statistical Tests, volume 1. Academic Press, 1989. (Cited in pages 72, 73 et 77.)
- [Kelker 1970] D. Kelker. *Distribution theory of spherical distributions and a location-scale parameter generalization*. Sankhyā: The Indian Journal of Statistics, Series A, vol. 32, no. 4, pages 419–430, 1970. (Cited in pages 74, 77 et 119.)
- [Kotz & Nadarajah 2004] S. Kotz and S. Nadarajah. Multivariate t Distributions and their Applications. Cambridge University Press, 2004. (Cited in page 74.)

- [Kotz *et al.* 2001] S. Kotz, T.J. Kozubowski and K. Podgorski. The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance. No. 183. Birkhauser, 2001. (Cited in page 74.)
- [Lebarbier & Mary-Huard 2006] E. Lebarbier and T. Mary-Huard. *Une Introduction au Critère BIC: Fondements Théoriques et Interprétation*. Journal de la Société française de statistique, vol. 147, no. 1, pages 39–57, 2006. (Cited in page 37.)
- [Leeb & Pötscher 2005] H. Leeb and B.M. Pötscher. *Model selection and inference: Facts and fiction*. Econometric Theory, vol. 21, no. 1, pages 21–59, 2005. (Cited in page 22.)
- [Lele 1993] C. Lele. *Admissibility results in loss estimation*. Annals of Statistics, vol. 21, no. 1, pages 378–390, 1993. (Cited in page 60.)
- [Leng *et al.* 2006] C. Leng, Y. Lin and G. Wahba. *A note on the lasso and related procedures in model selection*. Statistica Sinica, vol. 16, no. 4, page 1273, 2006. (Cited in pages 43, 132, 134 et 155.)
- [Li 1985] K.C. Li. *From Stein's unbiased risk estimates to the method of generalized cross validation*. Annals of Statistics, vol. 13, no. 4, pages 1352–1377, 1985. (Cited in pages 58, 60 et 64.)
- [Lindsey & Jones 2000] J.K. Lindsey and B. Jones. *Modeling pharmacokinetic data using heavy-tailed multivariate distributions*. Journal of Biopharmaceutical Statistics, vol. 10, no. 3, pages 369–381, 2000. (Cited in page 78.)
- [Lu & Berger 1989] KL Lu and J.O. Berger. *Estimation of Normal Means: Frequentist Estimation of Loss*. Annals of Statistics, vol. 17, no. 2, pages 890–906, 1989. (Cited in page 85.)
- [Mairal & Yu 2012] J. Mairal and B. Yu. *Complexity analysis of the lasso regularization path*. In Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012. (Cited in page 110.)
- [Mallows 1973] CL Mallows. *Some comments on C_p* . Technometrics, vol. 15, no. 4, pages 661–675, 1973. (Cited in pages 27 et 62.)
- [Maruyama & George 2011] Y. Maruyama and E.I. George. *Fully Bayes Factors with a Generalized g -prior*. The Annals of Statistics, vol. 39, no. 5, pages 2740–2765, 2011. (Cited in page 157.)
- [Maruyama 2003] Y. Maruyama. *A robust generalized Bayes estimator improving on the James-Stein estimator for spherically symmetric distributions*. Statistics & Decisions/International Mathematical Journal for Stochastic Methods and Models, vol. 21, no. 1/2003, pages 69–78, 2003. (Cited in page 69.)
- [Massart 2007] P. Massart. Concentration Inequalities and Model Selection: École d'Été de Probabilités de Saint-Flour XXXIII-2003. No. 1896. Springer-Verlag, 2007. (Cited in pages 14, 18, 23, 33 et 37.)

- [Maugis & Michel 2011] C. Maugis and B. Michel. *Data-driven penalty calibration: a case study for Gaussian mixture model selection*. ESAIM: Probability and Statistics, vol. 15, no. 1, pages 320–339, 2011. (Cited in page 33.)
- [Maxwell 1860] J.C. Maxwell. *V. Illustrations of the dynamical theory of gases. – Part I. On the motions and collisions of perfectly elastic spheres*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 19, no. 124, pages 19–32, 1860. (Cited in page 72.)
- [McQuarrie & Tsai 1998] A.D.R. McQuarrie and C.L. Tsai. *Regression and Time Series Model Selection*. World Scientific, 1998. (Cited in page 66.)
- [Meyer & Woodroffe 2000] M. Meyer and M. Woodroffe. *On the degrees of freedom in shape-restricted regression*. Annals of Statistics, vol. 28, no. 4, pages 1083–1104, 2000. (Cited in page 27.)
- [Nadif & Govaert 1998] M. Nadif and G. Govaert. *Clustering for binary data and mixture models – choice of the models*. Applied stochastic models and data analysis, vol. 13, no. 3-4, pages 269–278, 1998. (Cited in page 157.)
- [Niyogi & Girosi 1996] P. Niyogi and F. Girosi. *On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions*. Neural Computation, vol. 8, no. 4, pages 819–842, 1996. (Cited in pages 13, 16 et 20.)
- [Nocedal & Wright 1999] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer verlag, 1999. (Cited in page 167.)
- [Pchelintsev 2011] E. Pchelintsev. *Improved estimation in a non-Gaussian parametric regression*. Technical report, Department of Mathematics and Mechanics, Tomsk State University, 2011. (Cited in page 78.)
- [Pinson *et al.* 2004] P. Pinson, C. Chevallier and G. Kariniotakis. *Optimizing benefits from wind power participation in electricity market using advanced tools for wind power forecasting and uncertainty assessment*. In Proceedings of the 2004 European Wind Energy Conference, London, November 2004. (Cited in page 15.)
- [Poggi & Portier 2011] J.M. Poggi and B. Portier. *PM 10 forecasting using clusterwise regression*. Atmospheric Environment, vol. 45, no. 38, pages 7005–7014, 2011. (Cited in page 12.)
- [Rawlings *et al.* 1998] J.O. Rawlings, S.G. Pantula and D.A. Dickey. *Applied Regression Analysis: A Research Tool*. Springer Verlag, 1998. (Cited in page 39.)
- [Recht *et al.* 2010] B. Recht, M. Fazel and P.A. Parrilo. *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*. SIAM Review, vol. 52, no. 3, pages 471–501, 2010. (Cited in page 62.)
- [Rukhin 1986] A.L. Rukhin. *Improved estimation in lognormal models*. Journal of the American Statistical Association, vol. 81, no. 396, pages 1046–1049, 1986. (Cited in page 23.)

- [Rukhin 1988a] A.L. Rukhin. *Estimated Loss and Admissible Loss Estimators*. Statistical Decision Theory and Related Topics IV, vol. 1, page 409, 1988. (Cited in pages 86 et 157.)
- [Rukhin 1988b] A.L. Rukhin. *Loss functions for loss estimation*. Annals of Statistics, vol. 16, no. 3, pages 1262–1269, 1988. (Cited in pages 86 et 157.)
- [Sandved 1968] E. Sandved. *Ancillary Statistics and Estimation of the Loss in Estimation Problems*. Annals of Mathematical Statistics, vol. 39, no. 5, pages 1756–1758, 1968. (Cited in pages 55, 60 et 85.)
- [Schroeter 1980] G. Schroeter. *Application of the Hankel transformation to spherically symmetric coverage problems*. Journal of Applied Probability, vol. 17, no. 4, pages 1121–1126, 1980. (Cited in page 78.)
- [Schwarz 1978] G. Schwarz. *Estimating the dimension of a model*. Annals of Statistics, vol. 6, no. 2, pages 461–464, 1978. (Cited in page 30.)
- [Shao 1997] J. Shao. *An asymptotic theory for linear model selection*. Statistica Sinica, vol. 7, pages 221–242, 1997. (Cited in pages 37 et 64.)
- [Shibata 1983] R. Shibata. *Asymptotic mean efficiency of a selection of regression variables*. Annals of the Institute of Statistical Mathematics, vol. 35, no. 1, pages 415–423, 1983. (Cited in page 50.)
- [Stein 1955] C.M. Stein. *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 197–206, 1955. (Cited in page 58.)
- [Stein 1981] C.M. Stein. *Estimation of the mean of a multivariate normal distribution*. Annals of Statistics, vol. 9, no. 6, pages 1135–1151, 1981. (Cited in pages 55, 56, 57, 58, 59, 64 et 88.)
- [Steinwart 2007] I. Steinwart. *How to compare different loss functions and their risks*. Constructive Approximation, vol. 26, no. 2, pages 225–287, 2007. (Cited in pages 23 et 25.)
- [Stone 1974] M. Stone. *Cross-validatory choice and assessment of statistical predictions*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 36, no. 2, pages 111–147, 1974. (Cited in page 36.)
- [Sugiura 1978] N. Sugiura. *Further analysts of the data by Akaike's information criterion and the finite corrections*. Communications in Statistics-Theory and Methods, vol. 7, no. 1, pages 13–26, 1978. (Cited in pages 29 et 71.)
- [Takeuchi 1976] K. Takeuchi. *Distribution of informational statistics and a criterion of model fitting*. Suri-Kagaku (Mathematical Sciences), vol. 153, pages 12–18, 1976. (Cited in page 31.)
- [Tibshirani 1996] R.J. Tibshirani. *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pages 267–288, 1996. (Cited in pages 42 et 68.)

- [Vapnik & Chervonenkis 1971] V.N. Vapnik and A.Y. Chervonenkis. *On uniform convergence of the frequencies of events to their probabilities*. Teoriya veroyatnostei i ee primeneniya, vol. 16, no. 2, pages 264–279, 1971. (Cited in pages 20, 31 et 32.)
- [Vapnik 1998] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998. (Cited in pages iv, 14, 24, 85 et 99.)
- [Wald 1939] A. Wald. *Contributions to the theory of statistical estimation and testing hypotheses*. Annals of Mathematical Statistics, vol. 10, no. 4, pages 299–326, 1939. (Cited in page 18.)
- [West 1987] M. West. *On scale mixtures of normal distributions*. Biometrika, vol. 74, no. 3, pages 646–648, 1987. (Cited in page 124.)
- [Yang 2005] Y. Yang. *Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation*. Biometrika, vol. 92, no. 4, pages 937–950, 2005. (Cited in page 37.)
- [Ye 1998] J. Ye. *On measuring and correcting the effects of data mining and model selection*. Journal of the American Statistical Association, vol. 93, no. 441, pages 120–131, 1998. (Cited in pages 20, 22, 28, 61, 62 et 64.)
- [Zhang 2010] C.H. Zhang. *Nearly unbiased variable selection under minimax concave penalty*. Annals of Statistics, vol. 38, no. 2, pages 894–942, 2010. (Cited in pages 45, 103, 110, 112, 113, 155 et 166.)
- [Zhao & Yu 2007] P. Zhao and B. Yu. *On model selection consistency of Lasso*. Journal of Machine Learning Research, vol. 7, no. 2, page 2541, 2007. (Cited in pages 50 et 137.)
- [Zou & Hastie 2005] H. Zou and T.J. Hastie. *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pages 301–320, 2005. (Cited in pages 44, 113 et 132.)
- [Zou & Zhang 2009] H. Zou and H.H. Zhang. *On the adaptive elastic-net with a diverging number of parameters*. Annals of statistics, vol. 37, no. 4, page 1733, 2009. (Cited in page 47.)
- [Zou et al. 2007] H. Zou, T.J. Hastie and R.J. Tibshirani. *On the "degrees of freedom" of the lasso*. Annals of Statistics, vol. 35, no. 5, pages 2173–2192, 2007. (Cited in pages 68, 94, 95, 106 et 166.)
- [Zou 2006] H. Zou. *The adaptive lasso and its oracle properties*. Journal of the American Statistical Association, vol. 101, no. 476, pages 1418–1429, 2006. (Cited in pages 43 et 47.)

Sélection de modèle : une approche décisionnelle

Résumé : Cette thèse s'articule autour de la problématique de la sélection de modèle, étudiée dans le contexte de la régression linéaire. L'objectif est de déterminer le meilleur modèle de prédiction à partir de données mesurées, c'est-à-dire le modèle réalisant le meilleur compromis entre attache aux données et complexité du modèle.

La contribution principale consiste en la dérivation de critères d'évaluation de modèles basés sur des techniques de théorie de la décision, plus précisément l'estimation de coût. Ces critères reposent sur une hypothèse distributionnelle plus large que l'hypothèse classique gaussienne avec indépendance entre les observations : la famille des lois à symétrie sphérique. Cette famille nous permet à la fois de nous affranchir de l'hypothèse d'indépendance et d'ajouter une plus grande robustesse puisque nos critères ne dépendent pas de la forme spécifique de la distribution. Nous proposons également une méthode de comparaison des critères dérivés au travers d'une mesure de type Erreur quadratique (MSE), qui permet de déterminer si un critère d'évaluation de modèle est meilleur qu'un autre.

La seconde contribution attaque le problème de la construction des différents modèles comparés. Les collections de modèles considérées sont celles issues des méthodes de régularisation parcimonieuses, de type Lasso. En particulier, nous nous sommes intéressés à la Pénalité Concave Minimax (MCP), qui garde la sélection du Lasso tout en corrigeant son biais d'estimation. Cette pénalité correspond cependant à un problème non différentiable et non convexe. La généralisation des outils habituels de sous-différentielles grâce aux différentielles de Clarke a permis de déterminer les conditions d'optimalité et de développer un algorithme de chemin de régularisation pour le MCP.

Enfin, nous comparons nos propositions avec celles de la littérature au travers d'une étude numérique, dans laquelle nous vérifions la qualité de la sélection. Les résultats montrent notamment que nos critères obtiennent des performances comparables à ceux de la littérature, et que les critères les plus couramment utilisés en pratique (validation croisée) ne sont pas toujours parmi les plus performants.

Mots clés : *sélection de modèle, sélection de variable, régression linéaire, estimation de coût, distributions à symétrie sphérique, dépendance, Lasso, MCP.*

Model Selection: a decision-theoretic approach

Abstract: This manuscript addresses the problem of model selection, studied in the linear regression framework. The objective is to determine the best predictive model based on observed data, that is, the model realizing the best tradeoff between goodness of fit and complexity.

Our main contribution consists in deriving model evaluation criteria based on tools from Decision Theory, in particular loss estimation. Such criteria rely on a distributional assumption larger than the classical Gaussian hypothesis with independent observations: the family of spherically symmetric distributions. This family of laws allows us to relax the independence assumption and thus brings robustness, since our criteria do not depend on the specific form of the distribution. We also propose a method for comparing model evaluation criteria through a Mean-Squared Error type measure.

Our second contribution tackles the problem of constructing the models we compare. The conditions of models considered are obtained from sparse regularization methods, namely the Lasso and related methods. In particular, we studied the Minimax Concave Penalty (MCP), which keeps Lasso's selection while correcting its estimation bias. However, this penalty corresponds to a non differentiable and non-convex optimization problem. The generalization of subdifferentials with Clarke differentials allowed us to derive the optimality conditions d'optimalité and to propose a regularization path algorithm for MCP.

Finally, we compare our propositions to the literature through a numerical study, in which we verify the quality of the selection. The results especially show that our criteria yield performances similar to the literature, and that frequently used criteria such as Cross Validation do not always result in good performances.

Keywords: *model selection, variable selection, linear regression, loss estimation, spherically symmetric distributions, dependence, Lasso, MCP.*