Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) et Laboratoire d'Étude et de Recherches en Statistique et Développement (LERSTAD)

Thèse de doctorat

présentée en première version en vu d'obtenir le grade de Docteur, spécialité « Mathématique appliquée »

par

Marius Kwémou Djoukoué

Réduction de dimension en régression logistique, application aux données Actu-Palu

Thèse soutenue le 29 septembre 2014 devant le jury composé de :

M.	Jean-Marc Bardet	Université de Paris 1- Panthéon-Sorbonne	(Examinateur)
Mme	Elisabeth Gassiat	Université Paris-Sud	(Examinateur)
M.	Abdou Kâ Diongue	Université Gaston Berger de Saint-Louis	(Co-directeur)
Mme	Béatrice Laurent-Bonneau	INSA de Toulouse	(Rapporteur)
Mme	Adeline Leclercq Samson	Université Joseph Fourier de Grenoble	(Rapporteur)
M.	Jean-Yves Le Hesran	IRD - Université Paris Descartes	(Examinateur)
Mme	Marie-Luce Taupin	Université d'Evry Val d'Essonne	(Directeur)
Mme	Anne-Sophie Tocquet	Université d'Evry Val d'Essonne	(Examinateur)

À mes parents Hélène et Emmanuel Djoukoué

Remerciements

C'est le moment d'être reconnaissant et de dire merci à tous ceux qui ont cru en moi et qui m'ont permis d'arriver au bout de cette thèse.

Je voudrais tout d'abord exprimer mes plus profonds remerciements à mes directeurs de thèse, Marie-Luce Taupin et Abdou Kâ Diongue. Marie-Luce, tu as été pour moi une directrice de thèse extraordinaire et ceci à plusieurs points. Déjà tu as accepté de diriger ma thèse alors que tu ne m'avais jamais rencontré (j'espère qu'un jour tu me diras pourquoi !). Tu as toujours été disponible et as toujours su m'encourager même dans les moments de doute. Enfin, pour tes qualités humaines incomparables, tu as toujours anticipé sur ce qui pouvait m'être utile pour mon intégration en France (J'ai encore le recueil de recettes de cuisine que tu m'avais donné !). Abdou, merci d'avoir accepté de m'encadrer, merci pour tes conseils et ton soutien. Je te remercie pour la confiance que tu m'as témoignée.

J'exprime toute ma gratitude à Jean-Yves Le Hesran, qui a été présent depuis le début de cette aventure. Je te remercie pour ta disponibilité, tes nombreux conseils et pour la collaboration fructueuse durant mon doctorat (et mon Master aussi !). Ma reconnaissance va également à l'endroit de Stéphanie Dos Santos, qui a fourni les données Actu-Palu, données qui ont motivé cette thèse. À travers vous, j'aimerais remercier l'Institut de Recherche pour le Développement (IRD), en particulier les membres de l'équipe Actu-Palu.

Merci à Béatrice Laurent-Bonneau et Adeline Leclercq Samson pour m'avoir fait l'honneur de rapporter ma thèse.

Je remercie Jean-Marc Bardet, Elisabeth Gassiat et Anne-Sophie Tocquet d'avoir accepté de faire partie de mon jury de thèse.

Je tiens à remercier tous les membres du laboratoire de Mathématiques et Modélisation d'Évry (LaMME) qui ont contribué, de près ou de loin, à faire de ce doctorat une magnifique expérience. Je pense tout particulièrement à mes collègues de bureau, Sarah et Van Hanh, avec qui j'ai partagé les doutes et les joies d'un doctorant. Sarah, merci pour ta gentillesse et pour les longues discussions autour du Lasso ou non. Mention toute particulière aux secrétaires, Michèle et Valerie, pour la disponibilité et la bonne humeur. Je souhaite tout le meilleur à l'équipe des doctorants, Alia, Jean-Michel , Sarah, Morganne, Quentin ...

Merci également à tous les membres du Laboratoire d'Étude et de Recherches en Statistique et Développement (LERSTAD) du Sénégal. Je profite pour dire ma reconnaissance à Aliou Diop, responsable du Master STAFAV de Saint-Louis, et à tous mes enseignants. De façon plus générale merci au peuple Sénégalais, en séjournant au Sénégal, on est inévitablement touché par la téranga (hospitalité) simple et chaleureuse qui est accordée aux non sénégalais.

À mes ami-e-s, du Cameroun, du Sénégal, de France, ou d'ailleurs, je voudrais ici vous dire merci pour les différents moments passés ensemble. J'ai eu la chance de faire des rencontres magnifiques, au Sénégal et en France. Je veux dire ma reconnaissance et mon amitié à Elodie (merci pour ton oreille attentive !), Innes, Gaelle,

Sonia, Myriam et Webo, pour tous ces moments agréables passés en votre compagnie. Mes acolytes de Saint-Louis, Donald et Billy, pour nos équipées euphoriques. Mes compagnons de France, Louis-Joe, Sylviane, Merveille, Anna, Erick, merci pour votre amitié. Mes compagnons d'ailleurs, Davain, Félix, Délphine, Yacine, Sévérine et Bertin, pour votre sympathie, et surtout pour nos longs moments passés au téléphone (et sur internet !) à papoter. À la famille Parisot, merci pour l'accueil et pour tous les autres services.

À ma famille, pour la confiance et le soutien. Papa et maman, trouvez en ce travail le fruit des efforts que vous avez consentis à mon éducation et ma formation. Aux autres membres de la famille, ce travail a été rendu possible grâce à vous, Merci. Rosalie, toi qui a partagé cette aventure doctoresque en temps réel, merci pour tellement...

Évry, le 21 septembre 2014.

Table des matières

Liste des figures

Introduction

1

Sommaire

Ce chapitre présente la problématique et les objectifs de la thèse, ainsi que les outils et les différentes contributions.

Cette thèse a été réalisée en cotutelle entre l'Université d'Évry Val d'Essonne en France et l'Université Gaston Berger de Saint-Louis au Sénégal. Elle a été financée par l'Institut de Recherche pour le Développement (IRD).

L'Institut de Recherche pour le Développement (IRD) est un établissement français public à caractère scientifique et technique (EPST) placé sous la tutelle des ministères chargés de la Recherche et de la Coopération. Il a pour mission de développer des projets scientifiques centrés sur les relations entre l'homme et son environnement dans la zone intertropicale.

1.1 Contexte

Cette thèse a pour point de départ l'étude des données récoltées dans le cadre du projet interdisciplinaire Actu-Palu (ANR 07 – *SEST* – 001), *paludisme et diversité de l'environnement urbain Africain* : Un enjeu majeur pour la mise en place des thérapies à base d'artémisinine. L'objet principal de ce projet interdisciplinaire est d'aider à l'amélioration de l'efficacité des nouvelles stratégies thérapeutiques de lutte contre le paludisme.

Le paludisme est une maladie potentiellement mortelle. Il est dû à des parasites du genre *Plasmodium* transmis d'une personne à l'autre par des piqûres de moustiques *Anophèles* infectés, appelés "vecteurs du paludisme" (Figure 1.1).

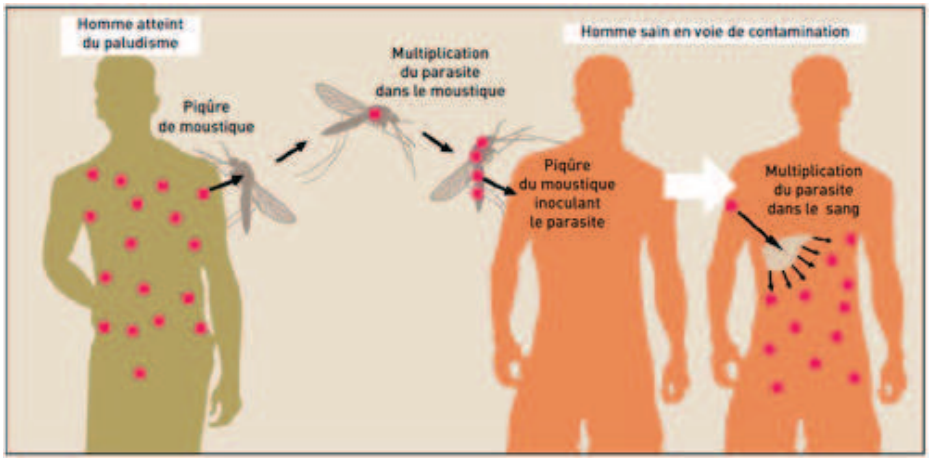


Figure 1.1 – *Transmission du paludisme (Source : www.docvadis.fr)*

Selon l'OMS, le paludisme a tué plus de 600 000 personnes en 2010, principalement des enfants africains de moins de 5 ans. Jusqu'à la fin des années 90, la lutte contre le paludisme était basée sur la chloroquine. L'extension de la résistance à la chloroquine a amené l'OMS et les pays africains à préconiser une bithérapie à base d'artémisinine (ACT). Les ACT sont les traitements les plus efficaces pour soigner le paludisme non compliqué et leurs coûts subventionnés en font, en principe, des médicaments de plus en plus faciles d'accès.

La fièvre a longtemps été le symptôme utilisé comme diagnostic présomptif du paludisme dans les centres de santé. Depuis 2006, la mise en place de Test de Diagnostic Rapide (TDR) permet un diagnostic biologique plus facile. Cependant, à domicile, les familles continuent de faire de la fièvre le symptôme principal du paludisme et ces familles ont souvent recours à une automédication. Plusieurs études ont montré qu'en cas de fièvre, plus de 50% des familles ont recours à une automédication, parfois à base d'antipaludique. Toutefois, l'usage irrationnel des

antipaludiques par une automédication est une raison souvent avancée pour justifier l'apparition et la propagation de la chimiorésistance de *plasmodium falciparum* à la chloroquine. Les familles ayant un risque élevé d'avoir un épisode de fièvre sont donc, de ce fait, plus exposées à une utilisation anarchique des antipaludiques. Par conséquent, ces familles sont les plus à risque de développer les résistances au traitement. Une conséquence possible (comme avec la chloroquine) est un échec des nouvelles stratégies de lutte contre le paludisme.

Il est donc maintenant urgent de tirer les leçons de l'échec de la chloroquine et de s'assurer d'une bonne mise en place des nouveaux traitements pour garantir durablement leur efficacité maximale. De simples messages au niveau des dispensaires risquent de ne pas suffire à une bonne utilisation de ces nouveaux médicaments. Pour améliorer l'efficacité des nouvelles stratégies de lutte contre le paludisme, il est important d'identifier les foyers à risque d'avoir un épisode fébrile. Cette identification passe par une analyse des facteurs économiques, sociologiques et médicaux qui expliquent qu'un foyer soit à risque.

1.1.1 Problématique et objectifs de l'analyse des données Actu-Palu

L'un des objectifs de l'analyse des données Actu-Palu est la détermination des variables importantes pour la prédiction des foyers à risque. Un foyer est dit à risque si il contient au moins un enfant de 2 à 10 ans qui a eu une fièvre. La variable d'intérêt est une variable binaire qui code les foyers (*foyers à risque* **vs** *foyers non à risque*). Il s'agit donc d'un problème de classification supervisée, car la variable d'intérêt est binaire et observée d'avance. Un modèle simple et pertinent pour prédire une variable binaire est le modèle de régression logistique, permettant d'établir une relation paramétrique entre une variable binaire et des variables explicatives.

L'une des particularités des données Actu-Palu est le nombre important de variables explicatives. En effet, les questionnaires qui ont été passés dans les foyers explorent de nombreux aspects de la vie quotidienne, mode de vie, économie, organisation du ménage, lieu de vie, caractéristiques individuelles, mode d'accès aux soins, connaissance de la maladie *etc*. Le nombre de variables est important (plusieurs centaines) et le contexte est donc, par nature, dit de grande dimension.

Sans réduction préalable du nombre de variables explicatives, la régression logistique n'est pas très performante. En effet, si le nombre d'individus n'est pas nettement supérieur au nombre de variables, alors la variance des estimateurs sera importante, aboutissant à des prédictions imprécises (Bull et al. (2007), Greenland et al. (2000)). Ainsi l'utilisation de toutes les variables dans le modèle introduit du bruit via les variables qui n'ont pas de lien avec la variable d'intérêt, ce qui peut fausser ou détériorer l'analyse. Ces variables sont nuisibles pour le modèle. D'autres variables, sans être nuisibles, peuvent être redondantes par rapport à des variables pertinentes, ces variables sont inutiles pour le modèle.

L'idée est de réduire la dimension de l'espace des variables explicatives, c'est-à-dire passer d'un nombre important de variables à un nombre relativement faible sans perte significative des performances de prédiction des méthodes utilisées. L'objectif de la réduction de dimension est d'éviter tout risque de surapprentissage.

1.1.2 Objectifs méthodologiques

L'analyse des données Actu-Palu soulève naturellement plusieurs questions méthodologiques dans un contexte de classification supervisée, plus précisément dans un modèle de régression logistique.

Après l'étude des données Actu-Palu, nous nous sommes intéressés à des méthodes toutes reliées par l'idée de sélection de variables ou de modèles, en régression logistique. Plus précisément, nous avons établi des inégalités oracles non asymptotiques pour des estimateurs obtenus par maximisation de vraisemblance pénalisée pour deux types de pénalités : les pénalités ℓ_1, de type Lasso et les pénalités ℓ_0.

Avant de présenter les résultats, nous introduisons d'abord les outils associés, présentés suivant la structure de la thèse. Nous commençons par définir le modèle de régression logistique classique. Puis nous décrivons les méthodes de réduction de dimension (Lasso, Group Lasso, forêts aléatoires). Enfin, nous présentons le principe de sélection de modèles développé par Birgé et Massart (2001; 2007).

1.2 Modèle de régression logistique

Le modèle de régression logistique (McCullagh et Nelder (1983), Draper et Smith (1966), Dobson (1990)) permet d'établir une relation paramétrique entre une variable binaire $Y \in \{0,1\}$ et le vecteur de covariables (ou variables explicatives) $z = (z_1, \ldots, z_d)^T$. Supposons que l'on observe n couples $(z_1, Y_1), \ldots, (z_n, Y_n) \in \mathbb{R}^d \times \{0,1\}$, le modèle de régression logistique est défini par la relation suivante :

$$\mathbb{P}(Y_i = 1 | z_i = t_i) = \frac{\exp(t_i^T \beta_0)}{1 + \exp(t_i^T \beta_0)} \tag{1.1}$$

où β_0 est un paramètre inconnu à estimer. Ce modèle bénéficie d'une grande notoriété dans les domaines tels que l'épidémiologie, la génomique, la sociologie, *etc*. Il peut être vu comme un cas particulier de la famille des modèles linéaires généralisés (McCullagh et Nelder (1983)) utilisant la fonction de lien *logit*. La fonction de lien *logit* a l'avantage de rendre facile l'estimation de Odds-Ratio (OR), utilisé comme approximation du risque relatif et permettant de mesurer l'effet d'un facteur. Nous renvoyons aux livres de Hilbe (2009), Menard (2002) et Hosmer Jr et al. (2013) pour des détails et exemples d'applications du modèle de régression logistique.

Il existe une littérature abondante sur l'estimation et l'inférence statistique en régression logistique. La procédure d'estimation, implémentée dans la plupart des logiciels standards de statistique (R, SAS, STATA, *etc*.), est généralement basée sur la minimisation de la log vraisemblance négative, conditionnellement aux $z_1, \ldots, z_n$. Plus précisément, β_0 est estimé par

$$\hat{\beta}_{MLE} = \arg\min_{\beta} L_n(\beta), \tag{1.2}$$

où $L_n(.)$ est l'opposé de la log vraisemblance, défini par

$$L_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ \log(1 + \exp(z_i^T \beta)) - Y_i z_i^T \beta \Big\}. \tag{1.3}$$

Notons que $\hat{\beta}_{MLE}$ est l'estimateur du maximum de vraisemblance conditionnel. Le problème (1.2) a rarement une solution explicite. En pratique, $\hat{\beta}_{MLE}$ peut être approché par différentes méthodes d'optimisation comme l'algorithme de Newton Raphson (Ypma (1995)), voir aussi Givens et Hoeting (2012), Mak (1993). Les résultats asymptotiques, tels que la consistance et la normalité de l'estimateur $\hat{\beta}_{MLE}$, sont maintenant bien connus et peuvent être trouvés dans Gourieroux et Monfort (1981), Fahrmeir et Kaufmann (1985), par exemple.

Cette procédure d'estimation est bien adaptée pour les petites valeurs de d (nombre de variables). Cependant, dès que d dépasse quelques dizaines de variables, comme c'est le cas pour les données Actu-Palu, le maximum de vraisemblance n'est plus approprié. En effet, lorsque le nombre de variables explicatives est important, l'estimation par le maximum de vraisemblance pose les difficultés classiques suivantes :

— Surapprentissage : le modèle obtenu a en général une petite erreur de prédiction sur l'échantillon qui a servi à l'estimer (échantillon d'apprentissage) mais perd ses pouvoirs de prédiction sur de nouveaux échantillons.

— Instabilité : un petit changement dans les données peut conduire à des estimations très différentes. En effet, le maximum de vraisemblance croît avec la complexité (nombre de paramètres) du modèle. Le critère (1.2) aura donc tendance à sélectionner le modèle le plus complexe, qui en grande dimension a une variance importante. Ce phénomène bien connu est illustré par la figure 1.2 : la variance croît avec la complexité du modèle pendant que le biais décroît.

— Non unicité des solutions : lorsque $d \gg n$, l'estimateur $\hat{\beta}_{MLE}$ n'est pas défini de manière unique, les résultats obtenus ne sont donc pas interprétables.

Pour pallier à ces défauts de l'estimation par maximum de vraisemblance en grande dimension, nous avons procédé en deux étapes : étape 1, réduction de la dimension des variables explicatives, étape 2, estimation de β_0 comme en (1.2) en utilisant le sous-ensemble de variables sélectionné à l'étape 1.

1.3 Méthodes de réduction de dimension ou de sélection de variables

Le paramètre β_0 dans le modèle (1.1) traduit le poids des variables explicatives sur la variable réponse Y. En d'autres termes, si $\beta_{0j} = 0$, alors la variable explicative associée à cette composante n'a pas d'influence sur Y. Lorsque le nombre de variables explicatives est important, un objectif peut être de sélectionner parmi ces variables celles qui ont une influence sur la variable réponse, c'est-à-dire identifier les composantes $\beta_{0j} \neq 0$. Ce type d'approche est connu sous le nom de sélection de variables ou réduction de dimension, et fournit des modèles qui ont l'avantage d'être facilement interprétables.

On parle généralement de grande dimension quand le nombre total de variables explicatives est du même ordre de grandeur ou est supérieur au nombre d'individus. En d'autres termes, il y a trop de variables pour pouvoir directement appliquer un modèle de régression logistique.

Nous partons du postulat qu'il existe un modèle incluant un petit nombre de variables explicatives permettant de bien prédire la variable réponse. Cette hypothèse semble raisonnable, car elle traduit le fait que le nombre important de variables à notre disposition (dans le cas des données socio-épidémiologiques) est dû au fait

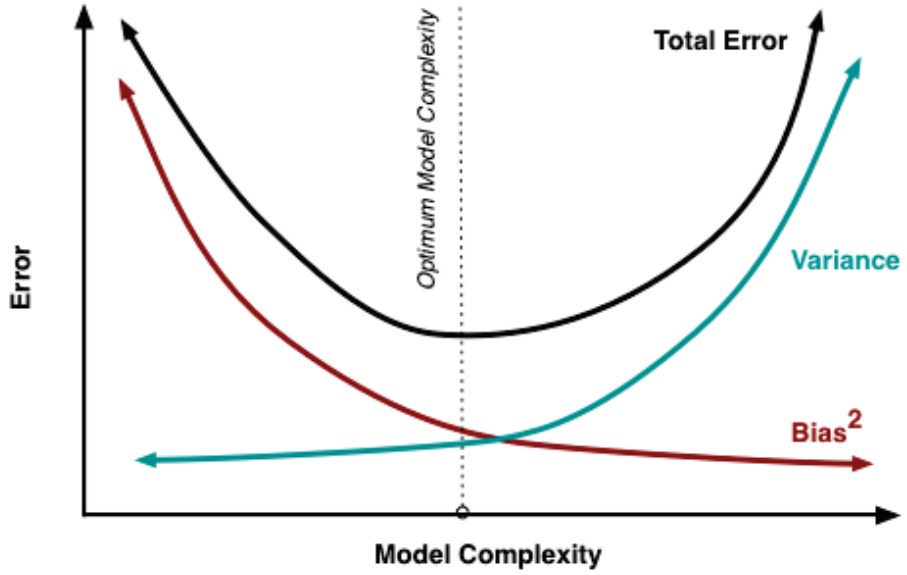


FIGURE 1.2 – *Compromis biais variance*

que les questionnaires sont issus des experts de différents domaines (médecins, sociologues, économistes *etc*) qui peuvent poser des questions très proches. Même si cette hypothèse n'est pas vérifiable, le fait de décrire les données par un nombre restreint de variables permet d'avoir des modèles facilement interprétables. Plusieurs stratégies de réduction de dimension existent, nous présentons ici celles basées sur la minimisation d'un contraste pénalisé et celles basées sur la construction d'une hiérarchie des variables explicatives dans les forêts aléatoires.

1.3.1 Réduction de dimension par pénalisation

En grande dimension, une alternative à l'estimation directe par le maximum de vraisemblance est souvent de considérer des estimateurs minimisant un critère pénalisé. Un estimateur minimisant un critère pénalisé est défini par,

$$\hat{\beta}\text{pen} = \arg\min_{\beta} \Big\{ \gamma_n(\beta) + \text{pen}(\beta) \Big\}, \tag{1.4}$$

où le terme $\gamma_n(\beta)$ est un critère empirique (moindre carré ou opposée de la log vraisemblance), qui quantifie la qualité d'ajustement du modèle. Sauf mention contraire, dans cette thèse $\gamma_n(\beta)$ désignera la log vraisemblance négative, $L_n(\beta)$ définie par (1.3). Le terme $\text{pen}(\beta)$ est une fonction positive appelée pénalité, qui contrôle la complexité du modèle en pénalisant les modèles complexes. La minimisation de (1.4) revient donc à chercher le meilleur compromis entre la qualité d'ajustement du modèle et sa complexité. La pénalité peut être choisie de façon à contraindre les estimateurs à avoir beaucoup de composantes nulles, ce qui permet de mieux appréhender le rôle de chaque variable explicative. Plusieurs choix de pénalités assurent que les estimateurs seront parcimonieux, *i.e.* auront peu de composantes non nulles. Nous nous concentrons ici sur les pénalités de type ℓ_0, de type Lasso (ou ℓ_1) et de type Group Lasso (ou ℓ_1/ℓ_2), que nous allons maintenant décrire.

Pénalisation ℓ_0

Les premières méthodes de sélection de variables ou de sélection de modèles utilisant les critères pénalisés ont été introduites par Mallows (1973) (voir aussi Akaike (1973; 1974) pour le AIC, Schwarz (1978a) pour le BIC). Ces critères utilisent des pénalités proportionnelles à la complexité du modèle, définie comme la pseudo-norme ℓ_0 du paramètre. Plus précisément, pour tout $\beta \in \mathbb{R}^d$ la pseudo-norme ℓ_0 de β est définie par

$$\|\beta\|_0 = \text{Card}\Big\{j \in \{1,\dots,d\},\ \beta_j \neq 0\Big\}.$$

L'estimateur obtenu en minimisant la log vraisemblance négative pénalisée par la pseudo-norme ℓ_0 est

$$\hat{\beta}_{\ell_0} \in \arg\min_{\beta}\{\gamma_n(\beta) + \lambda\|\beta\|_0\}. \tag{1.5}$$

En pénalisant par le nombre de composantes non nulles de β, ce critère contraint l'estimateur $\hat{\beta}_{\ell_0}$ à être parcimonieux. En d'autres termes, $\hat{\beta}_{\ell_0}$ est un vecteur avec plusieurs coordonnées nulles ($\hat{\beta}_{\ell_0 j} = 0$). Cela conduit à une sélection de variables, car seules les variables dont la coordonnée associée est non nulle ($\hat{\beta}_{\ell_0 j} \neq 0$) sont considérées comme pertinentes. La constante de pénalisation $\lambda > 0$ permet de gérer le compromis entre la qualité d'ajustement du modèle et la parcimonie. Le problème d'optimisation (1.5) est non convexe (donc algorithmiquement incalculable en un temps polynomial). Ce problème de non convexité peut être résolu en faisant par exemple une recherche exhaustive sur la famille de modèles $\mathcal{M} = \Big\{m, m \subset \{1,\dots,d\}\Big\}$. Plus précisément, soit $m \in \mathcal{M}$, notons S_m le sous-espace vectoriel engendré par la famille de vecteurs $\{z^j, j \in m\}$, où $z^j = (z_{1j},\dots,z_{nj})^T$ est la variable explicative associée à β_{0j}. Pour chaque $m \in \mathcal{M}$ on calcule $\hat{\beta}_m$ défini par

$$\hat{\beta}_m = \arg\min_{\beta \in S_m} \gamma_n(\beta).$$

On choisit alors $\hat{m}$ tel que

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \Big\{\gamma_n(\hat{\beta}_m) + \lambda\|\hat{\beta}_m\|_0\Big\}.$$

La recherche exhaustive se fait sur 2^d modèles ce qui, au vu des performances informatiques actuelles, est hors de porté lorsque d dépasse quelques dizaines. De plus, comme mentionné plus haut, l'estimateur $\hat{\beta}_m$ est inapproprié quand la dimension de S_m est grande.

En pratique, on considère souvent une famille restreinte $\widetilde{\mathcal{M}} \subset \mathcal{M}$ de modèles (par exemple les modèles emboîtés) et on cherche le modèle $\tilde{m}$ défini par

$$\tilde{m} = \arg\min_{m \in \widetilde{\mathcal{M}}} \Big\{\gamma_n(\hat{\beta}_m) + \lambda\|\hat{\beta}_m\|_0\Big\}. \tag{1.6}$$

Pour $\lambda = 1/n$ et $\lambda = \log(n)/2n$, le critère (1.6) correspond respectivement aux critères AIC et BIC. En régression linéaire par exemple, ces critères présentent de bonnes performances théoriques, sous certaines conditions sur la famille de modèles $\widetilde{\mathcal{M}}$. Nous reviendrons sur la pénalisation ℓ_0 (sélection de modèles) à la Section 1.5. Cependant, notons que le modèle $\tilde{m}$ n'est pas forcément une solution globale du problème (1.5).

Une autre façon de surmonter ce problème de non convexité consiste à utiliser d'autres pénalités, similaires à la pénalité ℓ_0 *i.e.* assurant des solutions parcimonieuses, et convexes, pour être résolvable en un temps raisonnable. On parle alors de convexification du problème à résoudre. C'est l'approche qui prévaut dans la pénalisation ℓ_1 et ℓ_1/ℓ_2.

Pénalisation ℓ_1

L'estimateur basé sur la minimisation de la log vraisemblance négative pénalisée par la norme ℓ_1 du paramètre, connu sous le nom de Lasso (Least Absolute Shrinkage and Selection Operator, (Tibshirani (1996))), ou *basis pursuit* (Chen et al. (1998)), est défini par

$$\hat{\beta}_L(\lambda) \in \arg\min_{\beta} \Big\{ \gamma_n(\beta) + \lambda \sum_{j=1}^{d} |\beta_j| \Big\}, \tag{1.7}$$

où λ est appelé paramètre de régularisation. Il doit être choisi de façon à assurer l'équilibre entre la qualité d'ajustement du modèle et la parcimonie (plus de détails sur le choix de λ à la Section 1.3.1).

La méthode Lasso présente trois principaux avantages qui justifient l'intérêt qui lui est accordé ces dernières années. Elle permet de sélectionner automatiquement les variables, car pour les grandes valeurs de λ, certaines composantes de β_0 sont estimées égales à zéro. Elle est applicable en grande dimension, y compris quand $d \gg n$. De plus, le problème d'optimisation (1.7) est convexe, donc relativement facile à résoudre. Même si en général il n'existe pas de forme analytique de la solution, des algorithmes existent pour résoudre ce problème, par exemple, l'algorithme *coordinate descent* introduit par Friedman et al. (2010) ou l'algorithme *predictor-corrector* introduit par Park et Hastie (2007).

Bien que l'estimateur Lasso permette d'avoir des modèles parcimonieux, ces performances reposent sur l'hypothèse de faibles corrélations entre les variables. En cas de forte corrélation entre plusieurs variables explicatives, l'estimateur Lasso a tendance à choisir une seule d'entre elles. Si les variables explicatives ont une structure connue à priori, par exemple la corrélation entre certaines variables, il peut être avantageux d'envisager une sélection par groupes de variables. Cette sélection par groupes peut se faire en utilisant des pénalités de type norme ℓ_1/ℓ_2 par exemple.

Pénalisation ℓ_1/ℓ_2

Nous considérons ici le cas où les variables explicatives ont une structure en groupe qui est connue à priori, structure que l'on souhaite prendre en compte dans la procédure d'estimation. La structure en groupe des variables est présente par exemple en biologie, où un groupe peut être constitué des variables qui partagent une même propriété biologique ou chimique. C'est aussi le cas des variables catégorielles (nombreuses dans les données Actu-Palu), où chacune d'entre elles est représentée dans la matrice du *design* par un groupe d'indicatrices de modalités. Plus précisément, une variable V ayant trois modalités a, b, c, est représentée par deux indicatrices V_a, V_b, (où $V_{i,a} = 1$ si l'individu i a la modalité a et 0 si non). Dans le cas des variables catégorielles, l'estimateur Lasso sélectionne les indicatrices des modalités et non le groupe d'indicatrices, *i.e.* la variable dans sa totalité. Dans de telles situations, il est plus judicieux d'envisager de sélectionner (ou rejeter) les

variables par groupes. Cette structure en groupes des variables peut être prise en compte en utilisant l'estimateur Group Lasso introduit par Yuan et Lin (2006) pour le modèle linéaire, et par Meier et al. (2008) pour le modèle logistique. Cette méthode considère les groupes de variables au lieu des variables individuelles de la façon suivante : notons $(G_\ell)_{\ell=1,\dots,g}$ une partition de $\{1,\dots,d\}$ en g groupes. Pour tout $\beta \in \mathbb{R}^d$, on note $\beta = (\beta_1,\dots,\beta_d) = (\beta^1,\dots,\beta^g)$, où $\beta^\ell = (\beta_j)_{j\in G_\ell}$. L'estimateur Group Lasso $\hat{\beta}_{GL}$ est défini par

$$\hat{\beta}_{GL} \in \arg\min_{\beta} \Big\{ \gamma_n(\beta) + r \sum_{l=1}^{g} \|\beta^l\|_2 \Big\}, \tag{1.8}$$

où $r > 0$ est le paramètre de régularisation. Si chaque groupe contient exactement une variable, on retrouve l'estimateur Lasso. Il s'agit donc d'une extension du Lasso. Le Group Lasso permet de faire la sélection de variables par groupes, *i.e.* tous les coefficients d'un groupe sont généralement tous nuls ou tous non nuls.

L'estimateur Group Lasso défini en (1.8) repose sur l'hypothèse que les groupes forment une partition de $\{1,\dots,d\}$. Les cas où les groupes ne forment pas une partition sont traités par Jacob et al. (2009), Huang et al. (2011), Jenatton et al. (2011) entre autres.

Il existe d'autres variantes de l'estimateur Lasso comme : *elastic net*, (Zou et Hastie (2005)), *fused Lasso*, (Tibshirani et al. (2005)), *latent Group Lasso* (Jacob et al. (2009)). Ces estimateurs diffèrent de l'estimateur Lasso par le choix de la fonction de pénalité, à adopter selon l'objectif de l'analyse.

Choix du paramètre de régularisation λ

Les estimateurs Lasso et Group Lasso dépendent du choix du paramètre de régularisation λ. Si $\lambda = 0$, l'estimateur Lasso et Group Lasso coïncident avec l'estimateur du maximum de vraisemblance qui est inadéquat en grande dimension. Si $\lambda \to \infty$, la procédure Lasso ne sélectionne aucune variable, car toutes les coordonnées de β_0 sont estimées à zéro. Donc $\lambda = 0$ et $\lambda \to \infty$ sont inadéquats. L'estimateur Lasso sélectionne d'autant plus de variables explicatives que λ est petit, et plus λ est grand, plus les coordonnées de $\hat{\beta}_L$ sont contraintes à être nulles. La Figure 1.3 illustre l'évolution du nombre de variables sélectionnées par le Lasso en fonction de λ. L'objectif est de déterminer une valeur de λ qui permet de sélectionner les variables pertinentes et ainsi d'améliorer les performances en prédiction du modèle. Il existe deux méthodes classiques pour choisir λ :

— **Validation croisée**, qui consiste à se donner une grille de valeurs de λ : $\lambda_1,\dots,\lambda_t$. Pour chaque $i \in \{1,\dots,t\}$ on répète les étapes suivantes.
  1- Partitionner l'ensemble des individus en k groupes, $V_1,\dots,V_k$.
  2- Pour chaque $j \in \{1,\dots,k\}$, l'estimation des paramètres se fait sur V_j^c et l'erreur de prédiction $E_j(\lambda_i)$ est calculée sur V_j.
  3- Ensuite on calcule une estimation de l'erreur de prédiction définie par :

$$er_i = \frac{1}{k} \sum_{j=1}^{k} E_j(\lambda_i).$$

  Le paramètre optimal est $\lambda_{i_{opt}}$, où $i_{opt} = \arg\min_{1\le i\le t} er_i$. La validation croisée est recommandée lorsque l'objectif de l'analyse est la prédiction (Leng et al. (2006), Hesterberg et al. (2008)). Mais elle est en général couteuse en temps de calcul.

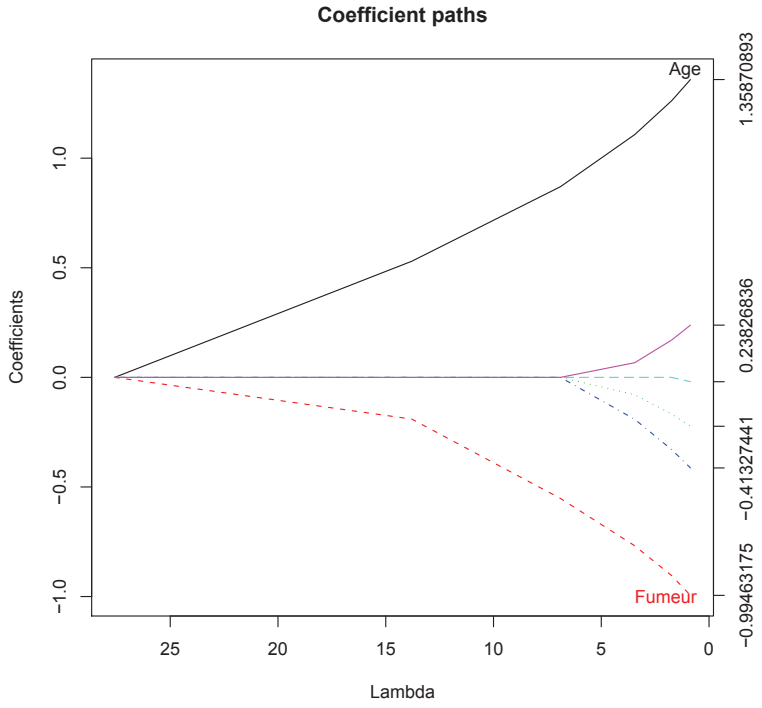


FIGURE 1.3 – *Exemple de chemin de régularisation*

— **Critère d'information**, le choix de λ peut être fait en utilisant les critères d'information de type AIC et BIC. Les critères de type AIC, BIC peuvent être définis pour les estimateurs Lasso et Group :

$$\begin{aligned} AIC(\lambda) &= \gamma_n(\hat{\beta}(\lambda)) + \tfrac{1}{n} df(\lambda), \\ BIC(\lambda) &= \gamma_n(\hat{\beta}(\lambda)) + \tfrac{\log n}{2n} df(\lambda), \end{aligned}$$

où $df(\lambda)$ est le degré de liberté de l'estimateur Lasso ou Group Lasso pour λ donné. Dans ce contexte, on choisit la valeur de λ qui minimise

$$\lambda_{opt} = \arg\min_{\lambda} AIC(\lambda) \text{ ou } \lambda_{opt} = \arg\min_{\lambda} BIC(\lambda),$$

le minimum étant choisi sur une grille de valeurs de λ donnée. Cette méthode est plus rapide en temps de calcul que la validation croisée. Pour plus de détails sur le choix de λ par critère d'information nous renvoyons le lecteur aux articles suivants : Zou et al. (2007), Fadili et al. (2012), Tibshirani et al. (2012), Vaiter et al. (2012).

1.3.2 Réduction de dimension via les forêts aléatoires

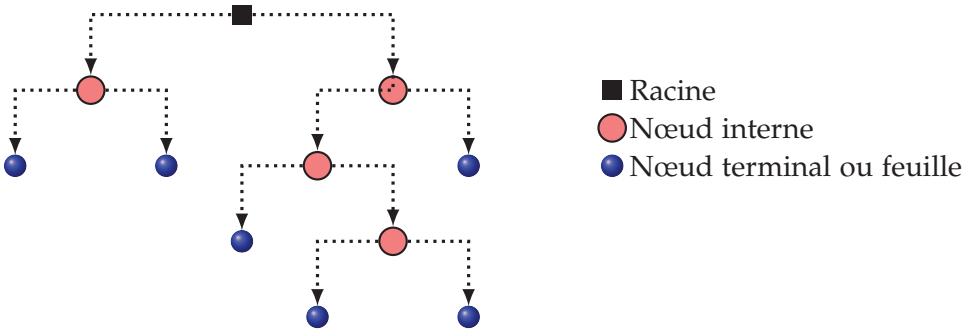
Les méthodes de réduction de dimension présentées ci-dessus reposent toutes sur l'hypothèse que les données sont générées suivant un modèle paramétrique, le modèle de régression logistique. Cette hypothèse peut parfois être restrictive. Utiliser directement les données pour "apprendre" le lien entre les variables explicatives et la variable réponse peut être bénéfique. C'est l'approche utilisée dans la méthode des forêts aléatoires. Cette dernière permet d'extraire des informations sur la loi qui a généré les données des données elles mêmes.

Pour réduire la dimension des variables explicatives via les forêts aléatoires, nous avons utilisé l'indice d'importance des variables explicatives fourni par les forêts aléatoires. Ces indices permettent de construire une hiérarchie des variables explicatives. Cette hiérarchie permet de sélectionner les variables en utilisant deux techniques : la première basée sur le choix d'un seuil et la deuxième basée sur l'utilisation des modèles emboîtés. Avant de décrire ces deux techniques, décrivons d'abord la méthode de construction des forêts aléatoires.

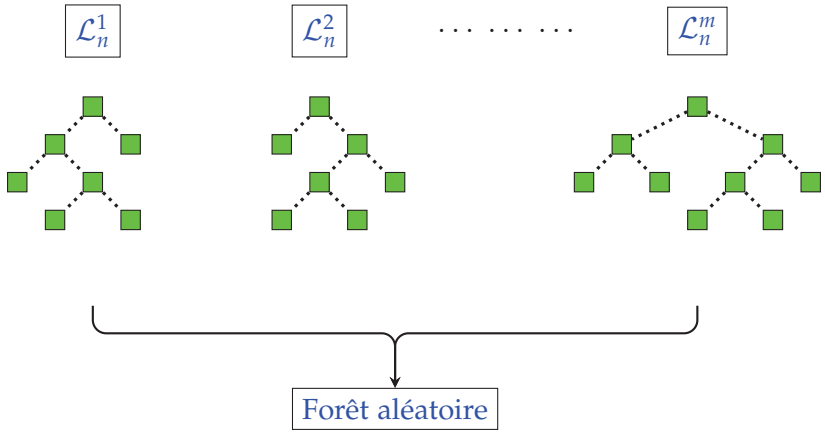
Forêts aléatoires

La méthode des forêts aléatoires est une technique d'apprentissage statistique introduite par Breiman (2001) (voir aussi Biau et al. (2008), Biau (2012)) basée sur l'agrégation d'arbres de classification CART (*Classification And Regression Tree*). Une forêt étant construite à partir des arbres CART, nous rappelons d'abord comment on construit un arbre CART.

CART est une méthode non paramétrique d'apprentissage qui construit un arbre de décision aussi bien en régression qu'en classification (Breiman et al. (1984)). Dans cette méthode, l'arbre est construit de la façon suivante : partant de la racine (les données complètes), on choisit la variable qui produit la meilleure coupure en deux des données. La coupure des données portant sur une variable z^j se fait en partitionnant les observations en deux groupes ($\{z^j \leq a\}$et$\{z^j > a\}$) qui prédisent le mieux la variable réponse Y. Les nœuds de l'arbre sont associés aux éléments de la partition. La même procédure est appliquée à chaque nœud "fils". On arrête la procédure lorsqu'il n'y a plus assez d'observations dans un nœud pour être partitionné en deux. L'arbre final est ensuite élagué pour éviter le surapprentissage. Les nœuds terminaux, encore appelés feuilles, sont associés aux partitions les plus fines de l'arbre. Ils sont utilisés comme prédictions. Dans le cas des données Actu-Palu par exemple, pour prédire le statut d'un nouveau ménage, on lui associera la réponse (*foyer à risque* **vs** *foyer non à risque*) majoritairement présentée dans le nœud terminal.



Une forêt aléatoire est construite en agrégeant les informations fournies par m arbres de classification. Notons $\mathcal{L}_n = \{(z_1, Y_1), \ldots, (z_n, Y_n)\}$. Chaque arbre noté $r_k(., \theta_k, \mathcal{L}_n)$, $k = 1, \ldots, m$ est construit en introduisant de l'aléatoire représenté par θ_k, d'où le nom forêt aléatoire. L'aléatoire est dû au fait que chaque arbre est construit sur un échantillon bootstrap $\mathcal{L}_n^l$, $l = 1, \ldots, m$, et à chaque nœud on tire $mtry < d$ variables de façon aléatoire et c'est dans cet ensemble de variables que l'on cherche celle qui réalise la coupure optimale. Le choix d'un petit nombre de variables à chaque nœud permet de réduire la complexité de l'algorithme.



Pour une nouvelle variable explicative z, la prédiction par une forêt aléatoire se fait en prenant la majorité des votes de chacun des arbres,

$$RF(z) = \begin{cases} 1 & \text{si } \frac{1}{m}\sum_{k=1}^{m} r_k(z, \theta_k, \mathcal{L}_n) \geq \frac{1}{2} \\ 0 & \text{si non} \end{cases} \tag{1.9}$$

Les arbres de la forêt ne sont pas élagués, ils ont donc une grande variance et un petit biais. L'agrégation des arbres permet d'avoir une forêt aléatoire avec une petite variance (Breiman (2001)). Les forêts aléatoires sont utilisables en grande dimension et permettent de prendre en compte la corrélation et les interactions entre les variables explicatives (voir Chen et Ishwaran (2012)).

La construction d'une forêt aléatoire fait intervenir deux paramètres importants :

— Le nombre m d'arbres de la forêt. Il doit être choisi de façon à assurer la stabilité de la forêt.

— Le nombre $mtry$ des variables choisies à chaque nœud de l'arbre. Il est compris entre 1 et d, c'est le paramètre le plus important. Une petite valeur de $mtry$ réduit la probabilité de choisir les variables importantes à chaque nœud, ce qui peut dégrader les performances de la forêt aléatoire. Une grande valeur de $mtry$ augmente la complexité de l'algorithme. Breiman a suggéré de prendre $mtry = \sqrt{d}$ pour des problèmes de classification. Ce choix a ensuite été confirmé par plusieurs travaux, voir par exemple Liaw et Wiener (2002), Díaz-Uriarte et De Andres (2006).

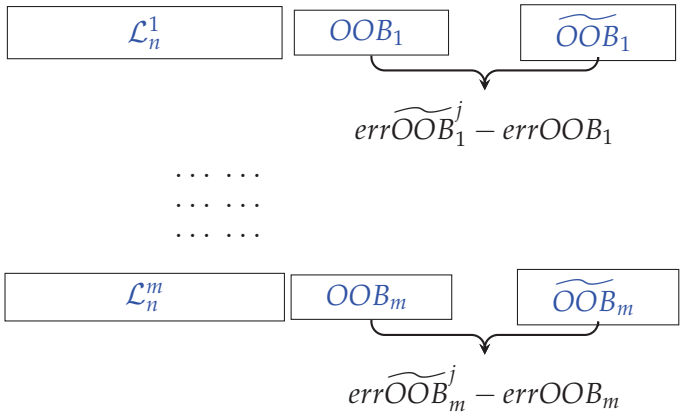
Comme nous l'avons dit plus haut, la sélection de variables via les forêts aléatoires se fait ensuite en utilisant les indices d'importance de variable.

Indice d'importance d'une variable dans les forêts aléatoires

L'indice d'importance d'une variable est construit à partir de l'échantillon "Out-Of-Bag". Un échantillon "Out-Of-Bag" est un échantillon en dehors de l'échantillon bootstrap. Pour une observation (z_i, Y_i) donnée, la prédiction $\hat{Y}_i$ se fait en agrégeant uniquement les valeurs prédites par les arbres qui ont été construits sans utiliser cette observation. On réitère cette procédure sur toutes les observations. Ensuite on calcule l'erreur de prédiction associée. Cette erreur est appelée erreur "Out-Of-Bag" (erreur OOB). Le principe de calcul de l'erreur de prédiction est similaire à celui de la validation croisée, car les données à prédire n'ont pas été utilisées pour construire les prédictions. Cette erreur est une estimation de l'erreur de généralisation de la forêt. Elle n'utilise pas les prédictions de la forêt, mais les prédictions d'arbres

agrégés de cette forêt. Notons que pour chaque observation ce n'est pas le même ensemble d'arbres qui est agrégé.

L'indice d'importance d'une variable est une mesure quantitative qui renseigne sur l'importance de la variable dans la prédiction. Pour une variable donnée, il est défini comme la différence en moyenne de la performance de l'arbre avant et après avoir perturbé les valeurs de cette variable. L'heuristique de cet indice est le suivant : si une variable est importante, la perturbation de ses valeurs va conduire à une augmentation de l'erreur de prédiction. Inversement si elle n'est pas importante sa perturbation n'aura presque aucun effet sur les prédictions, donc sur l'erreur. L'indice d'importance d'une variable z^j se calcule de la façon suivante. Soit $\mathcal{L}_n^l$ un échantillon bootstrap et OOB_l l'échantillon Out-Of-Bag associé, *i.e.* l'ensemble des observations qui ne sont pas dans $\mathcal{L}_n^l$. On perturbe les valeurs de z^j dans OOB_l, cela conduit à un échantillon perturbé $\widetilde{OOB}_l^j$. Ensuite on calcule les erreurs $errOOB_l$ et $err\widetilde{OOB}_l^j$ sur les échantillons OOB_l et $\widetilde{OOB}_l^j$ respectivement. On fait de même sur tous les échantillons bootstrap.



L'indice d'importance de la variable z^j s'exprime donc de la façon suivante

$$imp(z^j) = \frac{1}{m}\sum_{l=1}^{m}\left(er\widetilde{OOB}_l^j - erOOB_l\right).$$

Les indices d'importance des variables explicatives fournissent une structure hiérarchique des variables, structure qui sera utilisée pour sélectionner les variables. Comme nous l'avons dit plus haut nous avons utilisé deux techniques de sélection de variables, toutes les deux basées sur la hiérarchie des variables explicatives.

Méthode à seuil

La méthode à seuil est basée sur la définition d'un seuil à partir duquel une variable sera considérée comme pertinente. On sélectionne les variables dont l'indice d'importance est supérieur au seuil. Nous avons utilisé le seuil proposé par Strobl et al. (2009) : prendre la valeur absolue du plus petit indice d'importance (car il y a des indices d'importance négatifs). L'idée sous-jacente est que les variables qui ont un indice inférieur à ce seuil peuvent être considérées comme ayant une importance qui fluctue autour de la valeur zéro. Cette méthode, dans la suite, sera appelée *RFtreshold*.

Méthode basée sur les modèles emboîtés

On construit d modèles (forêts aléatoires) emboîtés : le premier modèle avec la variable la plus importante, le deuxième avec les deux variables les plus importantes, ainsi de suite jusqu'au modèle avec toutes les variables. Pour chacun de ces modèles on calcule l'erreur OOB. On prend comme modèle optimal celui qui a la plus petite erreur OOB. Cette méthode de sélection dans la suite sera appelée *RFnested*.

1.4 Inégalités oracles et pondération pour les estimateurs Lasso et Group Lasso

Cette section est consacrée à la présentation des inégalités oracles pour les estimateurs Lasso et Group Lasso pondérés. Dans un soucis de clarté, nous faisons cette présentation dans le cas plus simple du modèle de régression additif. Supposons que l'on observe des couples $(z_1, Y_1), \ldots, (z_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ tels que pour tout $i = 1, \ldots, n$,

$$Y_i = f_0(z_i) + W_i, \tag{1.10}$$

où $f_0 : \mathbb{R}^d \to \mathbb{R}$ est une fonction inconnue, à estimer. Dans le cas particulier du modèle de régression linéaire, *i.e* où $f_0(z) = z^T\beta_0$, estimer f_0 revient à estimer le paramètre $\beta_0 \in \mathbb{R}^d$. Les variables aléatoires $W_1, \ldots, W_n$ sont indépendantes et identiquement distribuées (i.i.d) et $z_1, \ldots, z_n$ sont déterministes.

Notations et définitions

Soit un dictionnaire $D = \{\phi_1, \ldots, \phi_p\}$ de fonctions $\phi_j : \mathbb{R}^d \to \mathbb{R}$, pour tout $j = 1, \ldots, p$ (des exemples de choix de dictionnaire sont donnés à la Section 1.6.2). Pour tout $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$, on note $f_\beta(z) = \sum_{j=1}^p \beta_j \phi_j(z)$. Pour tout $J \subset \{1, \ldots, p\}$, on note $|J|$ le cardinal de J. Pour tout $f : \mathbb{R}^d \to \mathbb{R}$, notons

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n f(z_i)^2.$$

1.4.1 Estimateur Lasso pondéré

En régression additive, on utilise généralement le critère des moindres carrés défini pour tout $t : \mathbb{R}^d \to \mathbb{R}$ par

$$MC_n(t) = \frac{1}{n} \sum_{i=1}^n (Y_i - t(z_i))^2. \tag{1.11}$$

Dans ce contexte, l'estimateur Lasso pondéré est défini par

$$\hat{\beta}_L = \arg\min_{\beta \in \mathbb{R}^d} \Big\{ MC_n(f_\beta) + \lambda \sum_{j=1}^p \omega_j |\beta_j| \Big\}. \tag{1.12}$$

Comme à la Section 1.3.1, le paramètre de régularisation λ est à choisir de façon à assurer le meilleur compromis entre la qualité d'ajustement mesurée par $MC_n(f_\beta)$ et la parcimonie mesurée par $\sum_{j=1}^p \omega_j |\beta_j|$. Pour de grandes valeurs des poids $w_j > 0$, $j \in \{1, \ldots, p\}$, les coefficients associés $\hat{\beta}_{L,j}$ sont estimés égaux à 0

et les autres sont rétrécis vers 0. Si l'on connaissait à l'avance les coefficients non significatifs, il suffirait de leur affecter un poids important. L'idée d'utiliser une pénalité pondérée a été proposée par exemple par Zou (2006). Dans cet article l'auteur utilise comme poids $\omega_j = 1/|\hat{\beta}_{MLE,j}|$ où $\hat{\beta}_{MLE}$ est l'estimateur du maximum de vraisemblance. Notons que dans le cas gaussien, l'estimateur du maximum de vraisemblance coïncide avec l'estimateur des moindres carrés. Cette procédure est connue sous le nom de *adaptive* Lasso. Zou (2006) a par ailleurs montré par des études de simulation que l'*adaptive* Lasso a de bonnes propriétés en sélection de variables comparé à l'estimateur Lasso non pondéré. Le principal inconvénient de l'*adaptive* Lasso est qu'il n'est applicable qu'en petite dimension. En effet, les poids sont estimés à l'aide de l'estimateur du maximum de vraisemblance, qui est un mauvais estimateur en grande dimension (quand il est défini). Un autre choix de poids applicable en grande dimension est donné par $\omega_j = 2\|\phi_j\|_n$ (voir Bickel et al. (2009)). Ces poids sont utilisés pour faciliter l'obtention des inégalités oracles. Le cas particulier où $\omega_j = 1$ pour tout $j = \{1, \dots, p\}$ correspond à l'estimateur Lasso défini par Tibshirani (1996).

Les propriétés théoriques permettant d'évaluer les performances de l'estimateur Lasso dans le modèle de régression additif ou linéaire sont maintenant bien connues. Le type de résultat recherché diffère selon l'objectif et le type de modèle. Dans le modèle de régression linéaire ($f_0(z) = z^T\beta_0$), ces propriétés sont de trois types. Il s'agit à la fois des résultats de convergence et des résultats non asymptotiques.

- **Prédiction :** l'objectif est de prédire la variable Y i.e. produire une meilleure approximation de $X\beta_0$, où X est la matrice du *design*, $X = (z_{i,j})_{1\leq i\leq n, 1\leq j\leq d}$. On s'intéresse donc aux propriétés de convergence vers 0 de l'erreur de prédiction $\|X\hat{\beta}_L - X\beta_0\|_2$.
- **Estimation :** l'objectif est de produire une estimation de β_0. On s'intéresse donc aux propriétés de convergence vers 0 de l'erreur d'estimation $\|\hat{\beta}_L - \beta_0\|_2$.
- **Sélection :** l'objectif est d'identifier les indices j qui appartiennent au support de β_0 ($supp(\beta) = \{j,\ \beta_j \neq 0\}$). Il s'agit de montrer que $\mathbb{P}(supp(\hat{\beta}_L) = supp(\beta_0))$ est proche de 1.

Les résultats en *sélection* et *estimation* ont été établis entre autres par Zhao et Yu (2006), Wainwright (2009), Bunea (2008a), Knight et Fu (2000), Meinshausen et Bühlmann (2006), Osborne et al. (2000), Zhang et Huang (2008), Meinshausen et Yu (2009) ; pour des résultats en *prédiction* voir par exemple Bickel et al. (2009), Bunea et al. (2006; 2007b;a), Massart et Meynet (2011). Dans cette thèse, nous nous intéressons aux résultats non asymptotiques.

Dans le modèle de régression additif, *i.e.* où f_0 n'est pas forcément linéaire, les performances de l'estimateur Lasso sont décrites en prédiction et généralement à travers des inégalités oracles non asymptotiques.

1.4.2 Inégalité oracle non asymptotique

En régression additive, une des exigences lorsque l'on construit un estimateur $f_{\hat{\beta}}$ de f_0 est que le risque soit le plus proche possible de 0, *i.e* que $\|f_{\hat{\beta}} - f_0\|_n^2$ soit proche de 0 avec grande probabilité ou que $\mathbb{E}\left[\|f_{\hat{\beta}} - f_0\|_n^2\right]$ soit proche de 0. Si on ne suppose pas que f_0 s'écrit comme combinaison linéaire des éléments du

dictionnaire de fonctions, les propriétés de $f_{\hat{\beta}}$ sont décrites par une inégalité qui a la forme suivante :

$$\|f_{\hat{\beta}} - f_0\|_n^2 \leq C \inf_{\beta \in \mathbb{R}^p} \left\{ \|f_\beta - f_0\|_n^2 + \Delta_n(\beta) \right\}. \tag{1.13}$$

L'inégalité (1.13) est vérifiée en espérance ou avec grande probabilité. Le premier terme $\|f_\beta - f_0\|_n^2$ correspond à l'erreur d'approximation ou terme de biais. L'idée implicite étant que le dictionnaire présente une bonne qualité d'approximation de f_0. Le terme $\Delta_n(\beta)$ est le terme de variance, qui décroît vers 0 avec n. Cette inégalité signifie que le risque de l'estimateur $f_{\hat{\beta}}$ est, à une constance multiplicative C près, du même ordre que le risque de la fonction qui réalise le meilleur compromis entre le terme de biais et celui de variance. La fonction qui réalise ce compromis est généralement appelée fonction oracle, ainsi l'inégalité (1.13) est une inégalité oracle. Elle est dite non asymptotique car est valable pour tout n. Le terme de variance est généralement utilisé pour décrire la vitesse de l'inégalité oracle. Quand il est de l'ordre de $\sqrt{\log(p)/n}$, on parle de vitesse lente, et quand il est de l'ordre de $\log(p)/n$, on parle de vitesse rapide. La quantité $C \geq 1$ est déterministe, pour $C = 1$ on parle d'inégalité oracle exacte. Dans le cas du modèle de régression additif, l'estimateur Lasso satisfait une inégalité oracle de la forme suivante (voir Bickel et al. (2009)) :

$$\|f_{\hat{\beta}_{Lasso}} - f_0\|_n^2 \leq C \inf_{\beta \in \mathbb{R}^p} \left\{ \|f_\beta - f_0\|_n^2 + A \frac{\log p}{n} \|\beta\|_0 \right\}. \tag{1.14}$$

L'inégalité (1.14) signifie que le risque de l'estimateur $f_{\hat{\beta}_{Lasso}}$ est, à une constante multiplicative près, du même ordre que le risque de la fonction qui réalise le meilleur compromis entre le biais et la parcimonie [1], mesurée par la norme ℓ_0 du paramètre. Cette fonction est appelée oracle ℓ_0. Lorsque l'oracle à des propriétés statistiques intéressantes, l'inégalité oracle permet de garantir les mêmes propriétés pour l'estimateur.

Hypothèse RE pour l'estimateur Lasso

Les inégalités oracles à vitesse rapide pour l'estimateur Lasso défini en (1.12) sont généralement obtenues en faisant une hypothèse sur la matrice de Gram Φ_n définie par

$$\Phi_n = X^T X / n, \quad \text{où} \quad X = \left(\phi_j(z_i)\right)_{1 \leq i \leq n,\ 1 \leq j \leq p}.$$

Soit $\Delta \in \mathbb{R}^p$ et $K \subset \{1, \ldots, p\}$ on note Δ_K un vecteur de $\mathbb{R}^p$ qui a les mêmes coordonnées que Δ pour les indices $j \in K$ et les coordonnées nulles ailleurs. Nous définissons l'hypothèse dite de valeur propre restreinte (*restricted eigenvalue condition*) :

Soit s un entier tel que $1 \leq s \leq p$ et a_0 une constante positive nous supposons que **(RE$_1$)**

$$\mu(s, a_0) := \min_{K \subseteq \{1, \ldots p\} : |K| \leq s} \ \min_{\Delta \neq 0 : \|\Delta_{K^c}\|_1 \leq a_0 \|\Delta_K\|_1} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_K\|_2} > 0,$$

1. Le terme de variance ici étant proportionnel à la norme ℓ_0 de β

Pour bien comprendre cette hypothèse, rappelons que l'estimateur des moindres carrées existe si la matrice de Gram est définie positive *i.e.*

$$\min_{\Delta\in\mathbb{R}^p,\Delta\neq 0} \frac{(\Delta^T\Phi_n\Delta)^{1/2}}{\|\Delta\|_2} = \min_{\Delta\in\mathbb{R}^p,\Delta\neq 0} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta\|_2} > 0. \tag{1.15}$$

En d'autres termes, la plus petite valeur propre de la matrice de Gram est strictement positive. Cependant, en grande dimension ($p \gg n$) la matrice de Gram est dégénérée, donc l'hypothèse (1.15) n'est jamais vérifiée. C'est pourquoi on prend le minimum de (1.15) dans un ensemble restreint à $\{K \subseteq \{1,\dots,p\} : |K| \leq s,\ \Delta \neq 0 : \|\Delta_{K^c}\|_1 \leq a_0\|\Delta_K\|_1\}$, d'où le nom de valeur propre restreinte (*restricted eigenvalue condition*). Cette hypothèse a été introduite dans Bickel et al. (2009) pour établir les inégalités oracles pour l'estimateur Lasso dans le modèle régression additif (avec les poids $\omega_j = 2\|\phi_j\|_n$). Dans cet article les auteurs décrivent des conditions simples pour que cette hypothèse soit vérifiée. Elle est connue comme l'une des hypothèses les moins restrictives utilisée pour obtenir les inégalités oracles à vitesse rapide. Par exemple les hypothèses utilisées dans Bunea et al. (2006; 2007b;a) pour établir les inégalités oracles pour l'estimateur Lasso dans le modèle de régression additif sont plus restrictives que l'hypothèse des valeurs propres restreintes. Pour une comparaison exhaustive des hypothèses sur la matrice de Gram utilisées pour établir les inégalités oracles en régression additive, nous renvoyons le lecteur à l'article de van de Geer et Bühlmann (2009).

1.4.3 Group Lasso pondéré

Considérons maintenant le cas où les variables ont une structure de groupes connue. Soit $(G_\ell)_{\ell=1,\dots,g}$ une partition de $\{1,\dots,p\}$. Pour tout $\beta = (\beta_1,\dots,\beta_p)$, on note $\beta^\ell = (\beta_j)_{j\in G_\ell}$. L'estimateur Group Lasso pour le modèle (1.10) est défini par :

$$f_{\hat\beta_{GL}} := \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \left\{ MC_n(f_\beta) + \lambda \sum_{\ell=1}^{g} \omega_\ell \|\beta^\ell\|_2 \right\}, \tag{1.16}$$

où ω_ℓ , $\ell \in \{1,\dots,g\}$ sont des poids, et $\lambda > 0$ est le paramètre de régularisation qui cherche le bon compromis entre la qualité d'ajustement du modèle, mesurée par $MC_n(f_\beta)$, et la parcimonie, mesurée par $\sum_{\ell=1}^{g} \omega_\ell\|\beta^\ell\|_2$. Yuan et Lin (2006) ont proposé de prendre des poids qui dépendent de la taille des groupes, plus précisément $\omega_\ell = \sqrt{|G_\ell|}$.

Les propriétés théoriques de l'estimateur Group Lasso ont beaucoup été étudiées pour le modèle de régression linéaire ou de régression additif. Citons par exemple les propriétés en *sélection* ou en *estimation* établis par Obozinski et al. (2010), Kolar et al. (2011), Huang et al. (2010), Lounici et al. (2009; 2011), Chesneau et Hebiri (2008), Nardi et Rinaldo (2008) pour le modèle linéaire ; et par Ravikumar et al. (2009), Meier et al. (2009), Huang et Zhang (2010) pour le modèle de régression additif.

Hypothèse RE pour l'estimateur Group Lasso

Comme avec l'estimateur Lasso, pour établir des inégalités oracles à vitesse rapide, il faut faire une hypothèse sur la matrice de Gram. Pour tout $\Delta \in \mathbb{R}^p$, notons

$$\|\Delta\|_{2,1}^2 = \sum_{\ell=1}^{g} \|\Delta^\ell\|_2.$$

Nous considérons une hypothèse analogue à l'hypothèse RE pour l'estimateur Lasso :

Soit s un entier tel que $1 \leq s \leq g$ et a_0 une constante positive nous supposons que **($\mathbf{RE_2}$)**

$$\mu_1(s, a_0) := \min_{K \subseteq \{1,\dots p\}: |K| \leq s} \min_{\Delta \neq 0: \|\Delta_{K^c}\|_{2,1} \leq a_0 \|\Delta_K\|_{2,1}} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_K\|_2} > 0.$$

Cette hypothèse a été utilisée par Lounici et al. (2011) pour établir des inégalités oracles pour l'estimateur Group Lasso dans le modèle de régression linéaire gaussien. Ils ont aussi montré que le Group Lasso améliore le Lasso lorsque les groupes sont bien choisis. En effet, ils ont montré un gain théorique du Group Lasso par rapport au Lasso en terme de vitesse. Ce gain est démontré dans le modèle de régression linéaire gaussien. Plus précisément, ils ont obtenu une vitesse de l'ordre de $\log(g)/n$, alors que la vitesse obtenue avec l'estimateur Lasso est de l'ordre de $\log(p)/n$. Si il y a donc peu de groupes ($g < p$), la vitesse de l'estimateur Group Lasso est meilleure que celle de l'estimateur Lasso.

1.5 Sélection de modèles

Nous rappelons dans cette section le principe de la sélection de modèles dans le cas particulier du modèle de régression additif défini en (1.10). La qualité d'un prédicteur t est mesurée par la perte relative

$$l(f_0, t) = \mathbb{E}[MC_n(t)] - \mathbb{E}[MC_n(f_0)] = \frac{1}{n}\sum_{i=1}^{n}(f_0(z_i) - t(z_i))^2 := \|f_0 - t\|_n^2. \quad (1.17)$$

où MC_n est le critère des moindres carrés défini en (1.11). La fonction f_0 dans le modèle (1.10) étant inconnue, on veut construire un estimateur $\hat{f}$ à partir des données qui soit le plus proche possible de f_0 au sens où son risque $\mathbb{E}[\|f_0 - \hat{f}\|_n^2]$ est le plus petit possible. Une méthode raisonnable pour estimer f_0 consiste à minimiser le critère MC_n sur un modèle S *i.e*

$$\hat{f}_S = \arg\min_{t \in S}\{MC_n(t)\}.$$

Le risque de l'estimateur $\hat{f}_S$ s'écrit :

$$\mathbb{E}(\|f_0 - \hat{f}_S\|_n^2) = \inf_{t \in S}\|f_0 - t\|_n^2 + \frac{\sigma^2}{n}\dim(S), \quad \text{où } \sigma^2 = \mathbb{E}(W_i^2). \quad (1.18)$$

Le premier terme, appelé terme de biais, représente l'erreur d'approximation du modèle S. Le deuxième terme, appelé terme de variance, représente l'erreur d'estimation dans le modèle S. Le terme de biais et celui de variance varient en sens inverse. C'est-à-dire que le terme de biais diminue quand la dimension de S augmente, tandis que le terme de variance augmente avec la dimension de S. Pour obtenir un bon estimateur de f_0, il faut déterminer un modèle S qui réalise un bon compromis entre le biais et la variance. Ce dernier point est l'objectif de la sélection de modèles. Nous présentons ici l'approche non asymptotique de sélection de modèles par pénalisation développée par Birgé et Massart (Birgé et Massart (2001; 2007)).

On se donne une collection de modèles $(S_m)_{m\in\mathcal{M}}$. Soit $(\hat{f}_m)_{m\in\mathcal{M}}$ la collection des estimateurs des moindres carrés associée à cette collection de modèles. Le modèle idéal m^* est celui dont l'estimateur associé $\hat{f}_{m^*}$ minimise le risque :

$$m^* = \arg\min_{m\in\mathcal{M}} \mathbb{E}(\|f_0 - \hat{f}_m\|_n^2).$$

Comme le risque de m^* dépend de la vraie fonction inconnue f_0, ce risque n'est pas accessible. Par conséquent $\hat{f}_{m^*}$ ne peut pas être considéré comme un estimateur de f_0. Le but de la sélection de modèles est de sélectionner un modèle $\hat{m}$ à partir des données tel que le risque de l'estimateur associé $\hat{f}_m$ soit le plus proche possible du risque de l'estimateur idéal : $\mathbb{E}(\|f_0 - \hat{f}_{m^*}\|_n^2)$. L'estimateur idéal est également appelé oracle. D'après l'expression (1.18) du risque, $\hat{m}$ doit pour cela faire un bon compromis entre le biais et la variance. L'idée consiste donc à sélectionner un modèle qui minimise un critère des moindres carrés pénalisés, *i.e.* à considérer

$$\hat{m} = \arg\min_{m\in\mathcal{M}} \Big\{ MC_n(\hat{f}_m) + \text{pen}(m) \Big\}, \tag{1.19}$$

où $\text{pen}(m)$ est un terme qui pénalise les gros modèles (au sens de l'inclusion). L'estimateur $\hat{f}_{\hat{m}}$ associé au modèle $\hat{m}$ ainsi choisi est appelé l'estimateur des moindres carrés pénalisés. La construction de $\hat{f}_{\hat{m}}$ passe par la détermination de la pénalité $\text{pen}(m)$ qui assure que le risque de $\hat{f}_{\hat{m}}$ soit proche de celui de l'oracle. Dans l'approche non asymptotique, on va chercher à montrer que pour tout n, l'estimateur des moindres carrés pénalisés vérifie :

$$\begin{aligned} \mathbb{E}(\|f_0 - \hat{f}_{\hat{m}}\|_n^2) &\leq C \inf_{m\in\mathcal{M}} \mathbb{E}(\|f_0 - \hat{f}_m\|_n^2) + \Delta_n \\ &= C \inf_{m\in\mathcal{M}} \Big\{ \inf_{t\in S_m} \|f_0 - t\|_n^2 + \frac{\sigma^2}{n} \dim(S_m) \Big\} + \Delta_n, \end{aligned} \tag{1.20}$$

où $C \geq 1$ est une constante idéalement proche de 1 et Δ_n est un terme résiduel qui tend vers 0 quand n tend vers l'infini. Une telle inégalité est appelée inégalité oracle non asymptotique. Elle montre que l'estimateur $\hat{f}_{\hat{m}}$ a un risque aussi petit, à une constante multiplicative près, et à terme de reste près, que le meilleur des risques possible dans une collection d'estimateurs.

Le premier critère pénalisé de type (1.19) est dû à Mallows (1973). Il est fondé sur l'heuristique qui suit. Soit $f_m = \arg\min_{t\in S_m} \|f_0 - t\|_n^2$ et $D_m = \dim(S_m)$. D'après (1.18) et le théorème de Pythagore

$$m^* = \arg\min_{m\mathcal{M}} \Big\{ -\|f_m\|_n^2 + \frac{\sigma^2}{n} D_m \Big\}. \tag{1.21}$$

L'heuristique de Mallows consiste à remplaçer $\|f_m\|_n^2$ dans (1.21) par un estimateur sans biais. Comme $\|\hat{f}_m\|_n^2 - \frac{\sigma^2}{n} D_m$ est un estimateur sans biais de $\|f_m\|_n^2$, car $\mathbb{E}(\|\hat{f}_m\|_n^2) = \|f_m\|_n^2 + \frac{\sigma^2}{n} D_m$, en remplaçant donc $\|f_m\|_n^2$ dans (1.21) par cet estimateur sans biais on obtient un nouveau critère

$$-\|\hat{f}_m\|_n^2 + 2\frac{\sigma^2}{n} D_m = -\frac{1}{n}\sum_{i=1}^n Y_i^2 + MC_n(\hat{f}_m) + 2\frac{\sigma^2}{n} D_m.$$

Comme $(\sum_{i=1}^n Y_i^2)/n$ ne dépend pas de m, on en déduit le critère C_p de Mallows :

$$C_p(m) = MC_n(\hat{f}_m) + 2\frac{\sigma^2}{n} D_m.$$

Ce critère est un critère pénalisé de type (1.19) avec $\text{pen}(m) = 2\frac{\sigma^2}{n}D_m$. Lorsque la variance σ^2 est inconnue, on peut la remplacer par un estimateur.

Le critère C_p de Mallows ne donne de bons résultats que si le nombre de modèles de dimension donnée n'est pas trop grand (Birgé et Massart (2007)). Lorsque le nombre de modèles de dimension donnée est grand, il faut étudier les déviations de $\|\hat{f}_m\|_n^2 - \|f_m\|_n^2$ autour de son espérance $(\frac{\sigma^2}{n}D_m)$, et choisir une pénalité qui les compense. L'étude de ces déviations se fait en utilisant les inégalités de concentration. C'est par exemple cette approche que Birgé et Massart (2007) ont utilisé pour établir une inégalité oracle dans le cas où les bruits W_i sont gaussiens. Ils ont obtenu une pénalité de la forme

$$\text{pen}(m) = \mu\frac{\sigma^2}{n}\Big(D_m + a\sqrt{D_m L_m} + bL_m\Big),$$

où $\mu > 1$, $a > 2$ et b>2 sont trois constantes, et où $(L_m)_{m\in\mathcal{M}}$ est une famille de poids vérifiant :

$$\sum_{m\in\mathcal{M}} e^{-L_m} \leq \infty. \tag{1.22}$$

Ils ont établi une inégalité oracle non asymptotique et validé le critère de C_p de Mallows lorsque le nombre de modèles à dimension fixée n'est pas trop grand *i.e.* lorsque $Card\{m \in \mathcal{M}; D_m = D\} \leq \xi D^r$, avec $\xi > 0$ et $r \in \mathbb{N}$. Plus précisément, en choisissant les poids $L_m = LD_m$, l'inégalité (1.22) est vérifiée pour tout $L > 0$, et donc la pénalité $\text{pen}(m) = \mu'\frac{\sigma^2}{n}D_m$ avec $\mu' > 1$ convient. En particulier pour $\mu' = 2$ le C_p de Mallows est validé. C'est-à-dire que l'estimateur des moindres carrés pénalisés (1.19), avec $\text{pen}(m) = 2\frac{\sigma^2}{n}D_m$ (qui correspond au C_p de Mallows), vérifie une inégalité oracle non asymptotique. Un résultat similaire à été établi dans un cadre non gaussien par Baraud (2000) en supposant que les bruits W_i ont des moments d'ordre k>2r+6. Notons que ces résultats ne nécessitent pas d'hypothèse sur la matrice de Gram comme avec le Lasso ou Group Lasso. En général les pénalités sont connues à une constante multiplicative près. Comme avec le Lasso la constante est très importante. Elle peut être calibrée en utilisant l'heuristique de pente introduite par Birgé et Massart (2007) (voir Section 4.5.2 pour une présentation de cette heuristique).

La sélection de modèles a été étudiée dans plusieurs contextes. Citons par exemple Baraud (2000), Birgé et Massart (2001), Yang (1999) pour le modèle linéaire ; Birgé (2014a), Castellan (2003b) pour l'estimation de densité ; et Lebarbier (2005), Durot et al. (2009), Braun et al. (2000) pour la segmentation. À notre connaissance il n'existe pas de résultats sur la sélection de modèles en régression logistique.

Notons que dans le modèle de régression additif, le critère des moindres carrés utilisé facilite l'obtention des inégalités oracles. En effet, il existe une relation simple entre les moindres carrés et la norme $\|.\|_n$. Cette relation rend facile le contrôle de la déviance et l'obtention des inégalités oracles. Dans le modèle de régression logistique, le critère utilisé est le maximum de vraisemblance. Ce critère induit une fonction de perte qui se connecte à la divergence de Kullback plutôt qu'à la norme $\|.\|_n$.

1.6 Présentation générale de nos résultats

L'analyse des données Actu Palu, données qui ont motivé nos travaux, a donné lieu au chapitre 2. Ce chapitre présente quelques stratégies de sélection de variables

dans des grandes enquêtes socio-épidémiologiques, avec une application à l'étude des épisodes fébriles chez les enfants de deux à dix ans dans les données Actu-Palu. Ce chapitre fait l'objet d'un article, Kwemou et al. (2014), soumis dans une revue internationale à comité de lecture.

Motivés par l'étude des données Actu-Palu, au Chapitre 3, nous avons étudié les propriétés théoriques des estimateurs Lasso et Group Lasso en régression logistique. L'étude de leurs propriétés se fait par la construction d'inégalités oracles non asymptotiques. Ce chapitre fait l'objet d'un article, Kwemou (2012), soumis dans une revue internationale à comité de lecture.

Dans le Chapitre 4, nous avons transposé les techniques de sélection de modèles introduites par Birgé et Massart (2001) au cas de la régression logistique. Nous avons établi des inégalités oracles non asymptotiques pour les estimateurs qui en découlent. Ce chapitre fait l'objet d'un article rédigé en collaboration avec Marie-Luce Taupin et Anne Sophie-Tocquet.

1.6.1 Chapitre 2 Stratégies de sélection de variables pour la prédiction des foyers à risque d'avoir un enfant atteint de fièvre à Dakar

L'un des objectifs de l'analyse des données Actu Palu est de sélectionner les variables pertinentes pour prédire les foyers à risque d'avoir un épisode fébrile dans Dakar. Comme mentionné plus haut les données Actu-Palu ont un nombre important de variables explicatives, ce qui rend le modèle de régression logistique inefficace. Ce grand nombre de variables a motivé le choix d'une procédure en deux étapes :

- Réduire le nombre de variables explicatives à l'aide des méthodes Lasso, Group Lasso et forêts aléatoires (RF).
- Ensuite, utiliser la régression logistique qui prend en compte les variables sélectionnées à l'étape précédente.

Présentation des données Actu-Palu

Les données Actu-Palu sont issues d'une enquête par questionnaires auprès de 379 ménages de Pikine, dans la banlieue de Dakar (ANR $07 - SEST - 001$).
La variable d'intérêt est binaire et code les foyers à risque : *foyers à risque* **vs** *foyers non à risque*.
Les variables explicatives sont issues des questionnaires qui ont été passés dans les foyers, qui explorent de nombreux aspects de la vie quotidienne tels que le mode de vie, l'économie, l'organisation du ménage, le lieu de vie, les caractéristiques du chef de ménage (le parent qui s'occupe des questions de santé dans le ménage), le mode d'accès aux soins, la connaissance de la maladie *etc*. Après un pré-traitement de la base de données, les analyses ont été effectuées sur 71 variables explicatives en majorité catégorielles.

Approches considérées pour l'analyse des données

Pour réduire le nombre de variables explicatives dans ce type de grandes bases de données, deux approches sont classiquement utilisées. La première consiste à faire des tests de corrélation entre chacune des variables explicatives et la variable réponse. On sélectionne alors les variables qui sont statistiquement liées à la variable réponse (voir Dudoit et al. (2002)). Mais cette technique de sélection ne per-

met pas de prendre en compte la possible interaction entre les variables explicatives. Une autre approche consiste à utiliser les méthodes de compression ou de transformation des variables explicatives. Elle se fait par exemple en utilisant les méthodes factorielles pour construire et sélectionner des axes informatifs ou "*super variable*" (voir Nguyen et Rocke (2002)). Elles permettent de réduire la dimension des variables explicatives. Cependant, les axes sélectionnés sont des combinaisons linéaires des variables explicatives. Ils font donc intervenir toutes les variables même les moins importantes. De plus il est en général difficile de donner un sens biologique ou socio-épidémiologique aux axes sélectionnés.

Nous avons proposé de réduire le nombre de variables explicatives en utilisant les méthodes Lasso, Group Lasso et forêts aléatoires (voir Section 1.3). Ces méthodes ont l'avantage de sélectionner les variables explicatives en gardant leurs structures initiales, ce qui rend facile l'interprétation des résultats. Elles permettent aussi de prendre en compte les interactions entre les variables explicatives. Elles ont été utilisées avec succès dans de nombreuses études génomiques, (voir Wu et al. (2009a), Li et al. (2011), Legarra et al. (2011), Garcia-Magariños et al. (2010), pour le Lasso et Group Lasso ; et Goldstein et al. (2010; 2011), Meng et al. (2009), Bureau et al. (2005), Dìaz-Uriarte et De Andres (2006) pour les forêts aléatoires) qui ont la particularité d'avoir beaucoup de variables explicatives, comme dans les données Actu-Palu. Les données Actu-Palu, à la différence des données génomiques, font intervenir des variables de différente nature : un mélange de variables quantitatives et catégorielles, certaines présentant un grand nombre de modalités éventuellement peu représentées dans la population.

Pour chaque sous-ensemble de variables sélectionné, nous avons mis en oeuvre un modèle de régression logistique. Les résultats sont alors comparés à partir des erreurs de prédiction. Le sous-ensemble de variables sélectionné par la méthode optimale a par la suite servi à prédire les foyers à risque dans un modèle de régression logistique.

Résultats

Les données comportent 23,6% de foyers à risque contre 76,84% de foyers non à risque. La premier constat (sans surprise) est que la réduction du nombre de variables explicatives a amélioré les performances en terme de prédiction du modèle de régression logistique. En effet, le modèle de régression logistique utilisant toutes les variables a une erreur de prédiction supérieure à celles de tous les modèles de régression logistique qui utilisent les sous-ensembles de variables sélectionnés (voir Table 1.1). Le Group Lasso est la méthode optimale car le modèle logistique utilisant le sous-ensemble de variables qu'il a sélectionné a la plus petite erreur de prédiction. Le modèle de régression logistique sur ce sous-ensemble de variables sélectionné permet de faire les constats suivants : certaines variables augmentent la probabilité qu'un foyer soit à risque (*nombre d'enfants de 2 à 10 ans*, *ménage utilisant les réseaux d'approvisionnement en médicaments moins chers*) et les autres la diminuent (*ménages qui dépensent plus pour les soins de santé*, *Age du chef de ménage*, *ménage achetant les médicaments sur le marché*). Toutes choses égales par ailleurs, la probabilité qu'un foyer soit à risque est plus élevée chez les ménages qui ont beaucoup d'enfants. La probabilité qu'un foyer soit à risque est plus élevée chez les ménages qui utilisent les réseaux d'approvisionnement en médicaments moins chers. La probabilité qu'un foyer soit à risque est moins élevée chez les ménages qui dépensent le plus pour les soins de santé. La probabilité qu'un ménage soit à risque est moins

élevée chez les ménages dont le chef est âgé. La probabilité qu'un foyer soit à risque est moins élevée chez les ménages qui achètent les médicaments sur le marché.

Méthodes de réduction	.	Lasso	Group Lasso	RFtreshold	RFnested
Erreur de prédiction (%)	36.11	22.22	**19.44**	25.39	25.39
Nbre de variables	71	3	15	9	9

TABLE 1.1 – *Erreur de prédiction : erreur de prédiction du modèle logistique qui prend en compte les variables sélectionnées par chaque méthode de réduction de dimension. Nbre de variables : nombre de variables sélectionnées par les méthodes de réduction de dimension.*

1.6.2 Chapitre 3 Inégalités oracles non asymptotiques pour les estimateurs Group Lasso et Lasso en régression logistique

Modèle

Considérons maintenant une extension du modèle (1.1), définie par

$$\mathbb{P}(Y_i = 1|z_i) = \frac{\exp(f_0(z_i))}{1+\exp(f_0(z_i))}, \tag{1.23}$$

où $f_0 : \mathbb{R}^d \to R$ est la fonction inconnue à estimer (Hastie (1983)). Le cas particulier où f_0 est linéaire ($f_0(z) = z^T\beta_0$ pour tout $z \in \mathbb{R}^d$) correspond au modèle (1.1).

Nous nous proposons de construire une stratégie d'estimation de f_0. Pour cela, on se munit d'un dictionnaire

$$\mathbb{D} = \{\phi_1, \dots, \phi_p\} \tag{1.24}$$

de fonctions $\phi_1, \dots, \phi_p : \mathbb{R}^d \to \mathbb{R}$. Notre objectif est d'estimer f_0 par une combinaison linéaire *parcimonieuse* des fonctions du dictionnaire. Plusieurs méthodes de choix des fonctions du dictionnaire existent. Elles dépendent de l'objectif de l'étude. Si l'objectif est de sélectionner les variables explicatives, les fonctions du dictionnaires peuvent être des identités ($\phi_j(z_i) = z_{ij}$ pour tout i et j). Elles peuvent aussi constituer une base de fonctions pouvant bien approximer f_0 (base d'histogrammes, de splines, *etc*). Un autre choix consiste à prendre des estimateurs de f_0 construits sur des échantillons indépendants. Ces estimateurs peuvent être issus de méthodes d'estimation structurellement différentes ou de même nature avec des paramètres de lissage ou de régularisation différents. Ce dernier cas est connu sous le nom d'agrégation d'estimateurs, et est très utilisé en apprentissage statistique. Nous supposons ici que nous sommes dans un contexte de grande dimension, *i.e.* où p est grand devant n, ou est du même ordre que n. Ce paradigme s'est imposé en statistique ces dernières années avec l'émergence du *high dimensional data* ou *big data*, fréquent par exemple en génétique, en text mining, en imagerie, *etc.* Rappelons que $z_1, \dots, z_n$ sont supposées déterministes.

Dans le contexte du modèle (1.23), nous considérons γ_n défini par

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\log(1+\exp(t(z_i)) - Y_i t(z_i)\Big\}. \tag{1.25}$$

Ce contraste empirique est une généralisation de l'opposé de la log vraisemblance défini en (1.3).

Estimateur Lasso pondéré

Nous considérons une version pondérée de l'estimateur Lasso définie par :

$$f_{\hat{\beta}_L} := \underset{f_\beta \in \Gamma}{\operatorname{argmin}} \left\{ \gamma_n(f_\beta) + r \sum_{j=1}^{p} \omega_j |\beta_j| \right\}, \tag{1.26}$$

où

$$\Gamma \subseteq \left\{ f_\beta(.) = \sum_{j=1}^{p} \beta_j \phi_j(.),\ \beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p \right\}$$

et $r > 0$ désigne le paramètre de régularisation. Comme à la Section 1.3.1, le paramètre de régularisation r permet de gérer le compromis entre la qualité d'ajustement mesurée par $\gamma_n(f_\beta)$ et la parcimonie mesurée par $\sum_{j=1}^{p} \omega_j |\beta_j|$. Les quantités ω_j, $j = 1, \ldots, p$ sont des poids (des exemples de poids sont donnés à la Section 1.4). Nous proposons une pondération, qui découle d'une l'inégalité de concentration de type Bernstein.

Résultats

La performance des estimateurs que nous proposons est établie via des inégalités oracles non asymptotiques. Notons $R(f_\beta) = \mathbb{E}(\gamma_n(f_\beta))$ la fonction de risque, et $R(f_\beta) - R(f_0)$ l'excès de risque. L'excès de risque est analogue à la perte relative dans les moindres carrés (voir Section 1.4).

Nous avons dans un premier temps établi deux inégalités oracles non asymptotiques et exactes.

Théorème 1.1 *Soit $f_{\hat{\beta}_L}$ l'estimateur Lasso défini en (1.26). Supposons que pour tout $i = 1, \ldots, n$ et $j = 1, \ldots, p$, $|\phi_j(z_i)| \leq c_2$.*

A-) Soit $x > 0$ et $r \geq 1$. Pour tout $j = \{1, \ldots, p\}$, posons

$$\omega_j = \frac{2}{n} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \phi_j^2(z_i)(x + \log p)} + \frac{2c_2(x + \log p)}{3n}. \tag{1.27}$$

Avec une probabilité supérieure à $1 - 2\exp(-x)$,

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq \inf_{\beta \in \mathbb{R}^p} \left\{ R(f_\beta) - R(f_0) + 2\|\beta\|_1 r \max_{1 \leq j \leq p} \omega_j \right\}.$$

B-) Soit $A > 2\sqrt{c_2}$. Supposons $\omega_j = 1$ pour tout $j = \{1, \ldots, p\}$, et

$$r = A \sqrt{\frac{\log p}{n}}.$$

Alors avec une probabilité supérieure à $1 - 2p^{1 - A^2/4c_2}$,

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq \inf_{\beta \in \mathbb{R}^p} \left\{ R(f_\beta) - R(f_0) + 2A\|\beta\|_1 \sqrt{\frac{\log p}{n}} \right\}.$$

Les résultats du Théorème 1.1 portent sur la version pondérée et non pondérée de l'estimateur Lasso. Le terme de vitesse est de l'ordre de $\|\beta\|_1 \sqrt{\log p / n}$ pour tout β, ce qui correspond à une vitesse lente. Ces inégalités oracles lentes sont établies

sans aucune hypothèse sur la matrice de Gram. Elles sont à notre connaissance les premières inégalités oracles non asymptotiques et exactes pour l'estimateur Lasso en régression logistique, obtenues sans aucune hypothèse sur la matrice de Gram. Ces résultats montrent que l'excès de risque de l'estimateur Lasso peut être majoré par le meilleur compromis entre le terme de biais (erreur d'approximation) et le terme de variance. Cela signifie que l'estimateur Lasso se comporte aussi bien que le meilleur compromis entre le terme de biais et de variance.

Peu de résultats existent sur l'estimateur Lasso dans le modèle de régression logistique, la plupart étant asymptotique ou faisant l'hypothèse que la fonction f_0 est linéaire. Citons par exemple Zou (2006), Huang et al. (2008), Bunea (2008b), pour les résultats en *sélection* ; Bach (2010) pour des résultats en *estimation* et en *prédiction*. À notre connaissance le seul résultat qui ne fait pas l'hypothèse que f_0 est linéaire est dû à van de Geer (2008). Elle a établi sous certaines hypothèses (notamment sur la matrice de Gram) une inégalité oracle non asymptotique pour l'estimateur Lasso dans le modèle de régression logistique.

Les résultats du Théorème 1.1 sont obtenus sans aucune hypothèse sur la matrice de Gram, ils se démarquent ainsi de l'inégalité oracle établie par van de Geer (2008). Pour obtenir les inégalités oracles à vitesse rapide, il est nécessaire de faire une hypothèse sur la matrice de Gram.
Comme dans le modèle de régression additif, nous faisons l'hypothèse de valeur propre restreinte :

Soit s un entier tel que $1 \leq s \leq p$ et a_0 une constante positive nous supposons que **(RE$_3$)**

$$\mu(s, a_0) := \min_{K \subseteq \{1,\dots p\}: |K| \leq s} \ \min_{\Delta \neq 0: \|\Delta_{K^c}\|_1 \leq a_0 \|\Delta_K\|_1} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_K\|_2} > 0,$$

où Δ_K est un vecteur de $\mathbb{R}^p$ qui a les mêmes coordonnées que Δ pour les indices $j \in K \subset \{1, \dots, p\}$ et les coordonnées nulles ailleurs. Des commentaires sur cette hypothèse sont faites à la Section 1.4.2. Présentons maintenant notre deuxième résultat, qui porte sur les inégalités oracles non asymptotiques avec vitesse rapide obtenues sous l'hypothèse des valeurs propres restreintes (**RE$_3$**).

Théorème 1.2 *Soit $\eta > 0$ et $1 \leq s \leq p$. Supposons que (**RE$_3$**) soit satisfaite avec $a_0 = 3 + 4/\eta$. Sous des hypothèses techniques nous avons les résultats suivants :*

A-) Soit $x > 0$ et $r \geq 1$. Avec probabilité supérieure $1 - 2\exp(-x)$,

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta) \inf_{f_\beta \in \Gamma} \left\{ R(f_\beta) - R(f_0) + \frac{c(\eta)\|\beta\|_0 r^2 \left(\max_{1 \leq j \leq p} \omega_j\right)^2}{c_0 \epsilon_0 \mu^2(s, 3+4/\eta)} \right\}, \tag{1.28}$$

et

$$\|f_{\hat{\beta}_L} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta) \inf_{f_\beta \in \Gamma} \left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)\|\beta\|_0 r^2 \left(\max_{1 \leq j \leq p} \omega_j\right)^2}{c_0' c_0 \epsilon_0^2 \mu^2(s, 3+4/\eta)} \right\}. \tag{1.29}$$

B-) Soit $A > 2\sqrt{c_2}$. *Posons* $\omega_j = 1$ *pour tout* $j = \{1, \ldots, p\}$, *et*

$$r = A\sqrt{\frac{\log p}{n}}.$$

Alors avec probabilité au moins $1 - 2p^{1-A^2/4c_2}$,

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta) \inf_{f_\beta \in \Gamma} \left\{ R(f_\beta) - R(f_0) + \frac{A^2 c(\eta)}{c_0 \epsilon_0 \mu^2(s, 3+4/\eta)} \frac{\|\beta\|_0 \log p}{n} \right\}, \tag{1.30}$$

et

$$\|f_{\hat{\beta}_L} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta) \inf_{f_\beta \in \Gamma} \left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)A^2}{c_0' c_0 \epsilon_0^2 \mu^2(s, 3+4/\eta)} \frac{\|\beta\|_0 \log p}{n} \right\}. \tag{1.31}$$

Dans les deux cas, $c(\eta)$ *est une constante dépendant uniquement de* η *;* c_0, c_0', *et* ϵ_0 *sont des constantes positives.*

Les inégalités oracles dans le Théorème 1.2 sont dites à vitesse rapide. En effet, le terme de variance est de l'ordre de $\|\beta\|_0 \log p/n$. Cette vitesse est similaire à celle des inégalités oracle pour l'estimateur Lasso dans le modèle de régression additif, établies dans Bickel et al. (2009), sous l'hypothèse des valeurs propres restreintes. À notre connaissance, les inégalités oracles portant sur la norme L_2 empirique dans le Théorème 1.2 sont nouvelles pour l'estimateur Lasso dans le modèle logistique. Ces inégalités oracles portant sur la norme L_2 empirique ont été établies grâce à un résultat démontré au Chapitre 3, qui connecte la norme L_2 empirique à l'excès de risque (voir Lemme 4.4). Des inégalités oracles non asymptotiques portant sur l'excès de risque ont été établies par van de Geer (2008) pour le Lasso dans le modèle de régression logistique et sous des hypothèses différentes des nôtres. En effet, nos résultats sont établis sous l'hypothèse RE qui peut être vue comme une version empirique de l'hypothèse C dans van de Geer (2008). La confiance (probabilité que nos résultats soient vrais) ne dépend ni de la taille n de l'échantillon, ni du nombre p de fonctions dans le dictionnaire, contrairement à celle de van de Geer (2008). De plus nous établissons des inégalités pour la version pondérée et non pondérée de l'estimateur Lasso. Les poids que nous proposons sont différents de ceux de van de Geer (2008) et présentent de bonnes performances, comme le montrent les études de simulation.

Estimateur Group Lasso pondéré

Considérons maintenant le cas où les variables ont une structure de groupes connue. Nous proposons ici une version pondérée de l'estimateur Group Lasso définie par :

$$f_{\hat{\beta}_{GL}} := \underset{f_\beta \in \Gamma_1}{\operatorname{argmin}} \left\{ \gamma_n(f_\beta) + r \sum_{\ell=1}^{g} \omega_\ell \|\beta^\ell\|_2 \right\}, \tag{1.32}$$

où ω_ℓ , $\ell \in \{1, \ldots, g\}$ sont des poids, et $r > 0$ est le paramètre de régularisation qui cherche le bon compromis entre la qualité d'ajustement des données mesurée par $\gamma_n(f_\beta)$, et la parcimonie mesurée par $\sum_{\ell=1}^{g} \omega_\ell \|\beta^\ell\|_2$.

Résultats

Nous avons dans un premier temps établi une inégalité oracle non asymptotique et exacte pour l'estimateur Group Lasso défini en (1.32). Ce résultat est l'analogue au Théorème 1.1 pour l'estimateur Group Lasso.

Théorème 1.3 *Soit $f_{\hat{\beta}_{GL}}$ l'estimateur Group Lasso défini en (1.32) avec $r \geq 1$ et*

$$\omega_l = \frac{2|G_l|}{n}\sqrt{\frac{1}{2}\max_{j\in G_l}\sum_{i=1}^{n}\phi_j^2(z_i)\,(x+\log p)} + \frac{2c_2|G_l|}{3n}\,(x+\log p)\,, \tag{1.33}$$

avec $x > 0$. Supposons que pour tout i=1,...,n, j=1,...,p, la fonction $\phi_j(z_i)$ est bornée. Alors avec une probabilité supérieure à $1-2\exp(-x)$,

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq \inf_{\beta\in\mathbb{R}^p}\left\{R(f_\beta) - R(f_0) + 2r\|\beta\|_{2,1}\max_{1\leq l\leq g}\omega_l\right\}. \tag{1.34}$$

Comme dans le Théorème 1.1, l'inégalité oracle du Théorème 1.3 est à vitesse lente. Elle est obtenue sans aucune hypothèse sur la matrice de Gram et ne fait pas l'hypothèse que la fonction f_0 est linéaire. À notre connaissance, cette inégalité est la première inégalité oracle non asymptotique et exacte pour le Group Lasso dans le modèle de régression logistique, établie sans aucune hypothèse sur la matrice de Gram.

Comme avec l'estimateur Lasso, pour obtenir les inégalités oracles à vitesse rapide, il est nécessaire de faire une hypothèse sur la matrice de Gram. Nous considérons une hypothèse analogue à l'hypothèse **RE$_3$** pour le Lasso :

> Soit s un entier tel que $1 \leq s \leq g$ et a_0 une constante positive nous supposons que **(RE$_4$)**
>
> $$\mu_1(s,a_0) := \min_{K\subseteq\{1,\dots p\}:|K|\leq s}\ \min_{\Delta\neq 0:\|\Delta_{K^c}\|_{2,1}\leq a_0\|\Delta_K\|_{2,1}} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_K\|_2} > 0.$$

Théorème 1.4 *Soit $f_{\hat{\beta}_{GL}}$ l'estimateur Group Lasso défini en (1.32) avec les poids ω_l dédinis en (1.33). Soit $\eta > 0$ et $1 \leq s \leq g$, supposons que l'hypothèse (**RE$_4$**) est satisfaite avec $a_0 = 3 + 4/\eta$. Sous certaines hypothèses techniques, avec une probabilité supérieure à $1-2\exp(-x)$,*

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta)\inf_{f_\beta\in\Gamma_1}\left\{R(f_\beta) - R(f_0) + \frac{c(\eta)|J(\beta)|r^2\left(\max_{1\leq l\leq g}\omega_l\right)^2}{c_0\epsilon_0\mu_1(s,a_0)^2}\right\}, \tag{1.35}$$

et

$$\|f_{\hat{\beta}_{GL}} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta)\inf_{f_\beta\in\Gamma_1}\left\{\|f_\beta - f_0\|_n^2 + \frac{4c(\eta)|J(\beta)|r^2\left(\max_{1\leq l\leq g}\omega_l\right)^2}{c_0'c_0\epsilon_0^2\mu_1(s,a)^2}\right\}. \tag{1.36}$$

Où

$$J(\beta) = \left\{l\in\{1,\dots,g\},\ \|\beta^l\|_2 \neq 0\right\},$$

et $c(\eta)$, c_0, c_0', C_0, c_1 ; ϵ_0 , $r \geq 1$ sont des constantes positives.

Dans le Théorème 1.4, la vitesse des inégalités est de l'ordre de $\log(p)/n$. A notre connaissance ces inégalités sont les premières inégalités oracles non asymptotiques pour le Group Lasso dans le modèle de régression logistique (1.23).

Trois résultats existent à notre connaissance pour l'estimateur Group Lasso en régression logistique. Sous des hypothèses sur la matrice de Gram et quelques hypothèses techniques, Meier et al. (2008) ont établi un résultat asymptotique en *prédiction*, Negahban et al. (2012) ont établi des résultats non asymptotiques en *estimation*. Ces deux articles font l'hypothèse selon laquelle la fonction f_0 à estimer est linéaire. Tout récemment, un troisième résultat a été établi par Blazère et al. (2014) pour le Group Lasso dans le modèle logistique. Plus précisément, ils ont établi des résultats en *prédiction* et en *estimation* pour le Group Lasso appliqué à la famille des modèles linéaires généralisés (GLM). Mentionnons plusieurs différences entre ces résultats et les nôtres. Ils considèrent une famille de modèles plus générale, GLM, à laquelle appartient le modèle logistique. Ces résultats ont été établis en faisant une hypothèse sur la matrice de Gram et quelques hypothèses techniques dont une hypothèse de bornitude sur les paramètres de la fonction f_0 à estimer. Contrairement à nous, ils supposent que la fonction f_0 à estimer est linéaire, c'est-à-dire qu'elle peut s'écrire comme une combinaison linéaire des éléments du dictionnaire. Dans ce contexte du modèle de régression logistique, avec f_0 linéaire, leurs résultats sont similaires à ceux du Corollaire 3.1, qui est un cas particulier du Théorème 1.4 de notre travail. De plus, dans ce cas particulier où la fonction f_0 est linéaire, nous avons établi des résultats en *prédiction* et en *estimation* sans faire d'hypothèse de bornitude sur la vraie fonction f_0 ou sur ses paramètres (voir Théorème 3.3), comme c'est fait dans Blazère et al. (2014). Enfin, nous proposons une version pondérée du Group Lasso et établissons, entre autres, une inégalité oracle non asymptotique et exacte sans aucune hypothèse sur la matrice de Gram et sur la fonction f_0 à estimer (Théorème 1.3).

Notons que la vitesse des inégalités dans le Théorème 1.4 est légèrement différente de celle trouvée par Lounici et al. (2011) pour l'estimateur Group Lasso dans le modèle linéaire gaussien. En effet, ils ont obtenu une vitesse de l'ordre de $\log(g)/n$. La vitesse est obtenue en contrôlant un processus à l'aide des inégalités de concentration. Bien que nous contrôlons un processus de la même forme que le processus (3.7) dans Lounici et al. (2011), nous n'avons pas un terme en $\log(g)$ tout le temps. Cette amélioration repose clairement sur l'hypothèse de résidus gaussiens dans le modèle linéaire.
Dans certains cas, nous retrouvons les vitesses avec un terme en $\log(g)$:
Supposons (sans perte de généralité) que $|G_1| = \cdots = |G_g| = m$, donc $p = m \times g$. Soit q une constante positive telle que $x = qlog(g) - log(m) > 0$, nous avons donc les poids

$$\omega_l = \frac{2|G_l|}{n}\sqrt{\frac{1}{2}\max_{j\in G_l}\sum_{i=1}^{n}\phi_j^2(z_i)\Big((1+q)\log(g)\Big)} + \frac{2c_2|G_l|}{3n}\Big((1+q)\log(g)\Big).$$

Ainsi

$$\omega_l^2 \sim \sqrt{\frac{\log(g)}{n}},$$

et les résultats du Théorème 1.4 sont vrais avec une probabilité supérieure à

$$1 - 2\frac{m}{g^q}.$$

Si $g > 2m$ alors les résultats sont vrais pour tout $q > 1$.

L'un des points forts de nos résultats est que la probabilité qu'ils soient vrais ne dépend ni de p, g ou n, contrairement à Lounici et al. (2011).

Simulations

Pour illustrer les performances des estimateurs théoriques que nous proposons, nous avons réalisé une étude par simulations. Nous avons comparé nos estimateurs à leur version canonique définie dans Tibshirani (1996) et Meier et al. (2008) pour le Lasso et le Group Lasso respectivement dans le modèle de régression logistique. Les résultats montrent que nos estimateurs ont de bonnes propriétés en sélection de variables. Par exemple l'estimateur Group Lasso pondéré a un taux bonnes sélections proche de 99% pour certaines valeurs du paramètre de régularisation. Cela signifie que sous réserve que le paramètre de régularisation soit bien choisi, le Group Lasso pondéré sélectionne les bons groupes dans 99% des cas. Les résultats montrent aussi que les versions pondérées ont des propriétés en sélection de variables meilleures que leurs versions canoniques.

1.6.3 Chapitre 4 Sélection de modèles en régression logistique

Principe de sélection de modèles

Nous considérons comme précédemment l'extension du modèle de régression logistique définie en (1.23). Nous utilisons le principe de sélection de modèles développé par Birgé et Massart (2001; 2007) (et brièvement décrit en Section 1.5 pour le modèle linéaire). Nous décrivons ce principe dans le cas de la régression logistique.

Soit une collection donnée de modèles $(S_m)_{m\in\mathcal{M}}$, où $\mathcal{M}$ dépend éventuellement de n, et les estimateurs associés $(\hat{f}_m)_{m\in\mathcal{M}}$ définis pour tout m par

$$\hat{f}_m = \arg\min_{t\in S_m} \gamma_n(t),$$

où $\gamma_n(.)$ est le contraste défini en (1.25). Idéalement, on aimerait choisir dans cette collection le modèle qui est le plus "proche" de f_0 au sens du risque, *i.e.*

$$m^* = \arg\min_{m\in\mathcal{M}} \Big[R(\hat{f}_m) - R(f_0)\Big],$$

où pour toute fonction t, $R(t) = \mathbb{E}[\gamma_n(t)]$. Cependant, m^* est inaccessible car dépend de la loi inconnue des variables $Y_1, \ldots, Y_n$. La fonction f_{m^*} (ou le modèle S_{m^*}) est un oracle pour le problème de sélection. La sélection de modèles a pour but de bâtir un critère qui permet de choisir le modèle qui imite le comportement et les performances de l'oracle en terme de risque. Une procédure pour sélectionner un tel modèle consiste à utiliser un critère pénalisé. La sélection de modèles via un critère pénalisé consiste à choisir $\hat{m}$ qui minimise le critère pénalisé suivant

$$\hat{m} = \arg\min \Big\{\gamma_n(\hat{f}_m) + \text{pen}(m)\Big\}, \tag{1.37}$$

où pen : $\mathcal{M}_n \longrightarrow \mathbb{R}^+$ est la pénalité. Le point crucial est de proposer une pénalité qui conduise à sélectionner un modèle dont le risque est proche de celui de l'oracle. Plus précisément, d'un point de vue non asymptotique, cela revient à construire une pénalité qui permet de sélectionner un modèle $S_{\hat{m}}$ qui vérifie l'inégalité oracle :

$$R(\hat{f}_{\hat{m}}) - R(f_0) \leq C \inf_{m\in\mathcal{M}_n} \Big[R(\hat{f}_m) - R(f_0)\Big] + \Delta_n,$$

avec grande probabilité, ou en espérance.

Résultats

Nous avons considéré deux situations, le cas d'une collection quelconque de modèles, et le cas particulier des fonctions constantes par morceaux. Dans chacune de ces situations nous avons proposé une forme de pénalité, et montré que l'estimateur qui en découle vérifie une inégalité de type oracle.

Dans un premier temps, nous considérons la collection de modèles définis pour tout $m \in \mathcal{M}$ par

$$\mathcal{S}_m := \Big\{ f_\beta = \sum_{j \in m} \beta_j \phi_j \Big\}, \tag{1.38}$$

où $\{\phi_1, \ldots, \phi_M\}$ sont les fonctions d'un dictionnaire (vois Section 1.6.2 pour des exemples de dictionnaire), $\mathcal{M}$ un sous- ensemble de l'ensemble des parties de $\{1, \ldots, M\}$.

Nous proposons une pénalité pour cette collection de modèles et établissons des inégalités oracles non asymptotiques à la fois pour la divergence de Kullback-Leibler et pour la norme L_2 empirique. Notons

$$L_\infty(C_0) = \Big\{ f : \mathbb{R}^d \to \mathbb{R}, \ \max_{1 \leqslant i \leqslant n} |f(x_i)| \leqslant C_0 \Big\}.$$

Théorème 1.5 *Soit $\{\mathcal{S}_m\}_{m \in \mathcal{M}}$ une collection de modèles définie en (1.38). Soit $\hat{f}_m = \arg\min_{t \in \mathcal{S}_m \cap L_\infty(C_0)} \gamma_n(t)$ et $f_m = \arg\min_{t \in \mathcal{S}_m \cap L_\infty(C_0)} \mathbb{E}[\gamma_n(t)]$. On note D_m la dimension de $\mathcal{S}_m$, $m \in \mathcal{M}$. Soit $\{L_m\}_{m \in \mathcal{M}}$ une suite de nombres positifs vérifiant*

$$\Sigma = \sum_{m \in \mathcal{M}} \exp(-L_m D_m) < \infty.$$

Considérons une pénalité pen : $\mathcal{M} \to \mathbb{R}_+$, telle que,

$$pen(m) \geqslant \lambda \frac{D_m}{n} \left(\frac{1}{2} + \sqrt{5 L_m} \right)^2,$$

où λ est une constante positive. Supposons que $\max_{1 \leq i \leq n} |f_0(z_i)| \leq c_1$ alors

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}})] \leqslant C \inf_{m \in \mathcal{M}} \big\{ \mathcal{K}(\mathbb{P}_{f_0}, \mathbb{P}_{f_m}) + pen(m) \big\} + C_1 \frac{\Sigma}{n}$$

et

$$\mathbb{E}_{f_0} \parallel \hat{f}_{\hat{m}} - f_0 \parallel_n^2 \leqslant C' \inf_{m \in \mathcal{M}} \big\{ \parallel f_0 - f_m \parallel_n^2 + pen(m) \big\} + C_1' \frac{\Sigma}{n}.$$

où C, C', C_1, C_1' sont des constantes.

Le Théorème 1.5 fournit une forme de pénalité garantissant que le modèle sélectionné soit "proche" de l'oracle en terme de risque. En effet, sous réserve que la pénalité soit bien choisie, le risque du modèle sélectionné est, à un constante près, proche de celui de l'oracle. Ces résultats sont obtenus en faisant l'hypothèse de bornitude sur la vraie fonction f_0 et les fonctions dans les modèles. Cette hypothèse joue un rôle central dans la preuve de ce théorème car elle permet de connecter la norme L_2 empirique et divergence de Kullback-Leibler (excès de risque) comme dans Kwemou (2012). Cependant, la pénalité dépend d'une constante inconnue λ. Cette constante dépend de la borne imposée à la vraie fonction f_0. En pratique,

cette constante peut être calibrée en utilisant le principe de l'*heuristique de pente* introduit dans Birgé et Massart (2007) (voir Section 4.5.2). Il est possible dans un cas particulier de modèles d'obtenir une pénalité qui ne dépend pas des hypothèses faites sur la vraie fonction, c'est l'objet du prochain résultat.

Nous considérons maintenant une collection de modèles constituée de fonctions constantes par morceaux. Avant de présenter le résultat faisons un petit rappel : toute fonction $g : \mathbb{R}^d \to \mathbb{R}$ peut être représentée par $(g(x_i))_{i\in\{1,\dots,n\}}$. Nous allons donc pour simplifier les notations supposer que la fonction g est définie par g : $\{1,\dots,n\} \to \mathbb{R}$, $i \mapsto g(x_i)$.

Soit $\mathcal{M}$ une collection de partitions de $\{1,\dots,n\}$ et S_m le sous-espace vectoriel engendré par $\{\mathbb{1}_J, J \in m\}$. La dimension de S_m est simplement le cardinal de m ($|m|$ ou D_m). Considérons l'hypothèse suivante :

$$\text{Il existe une constante } \rho > 0 \text{ telle que } \min_{i=1,\cdots,n} \pi_{f_0}(x_i) \geq \rho \text{ et} \quad \min_{i=1,\cdots,n} [1 - \pi_{f_0}(x_i)] \geq \rho. \tag{$\mathbf{A_1}$}$$

Théorème 1.6 *Soit $\{S_m, m \in \mathcal{M}_n\}$ une collection de modèle constituée de fonctions constantes par morceaux, où $\mathcal{M}_n$ est un ensemble de partitions construites à partir de la partition m_f, i.e. que m_f est un raffinement de chaque $m \in \mathcal{M}$. Supposons que pour tout $J \in m_f$, $|J| \geq \Gamma \log^2(n)$ où $|J|$ est le cardinale de J et Γ est une constante positive. Soit $(L_m)_{m\in\mathcal{M}_n}$ une famille de poids vérifiant*

$$\Sigma = \sum_{m\in\mathcal{M}_n} \exp(-L_m|m|) < +\infty. \tag{1.39}$$

Soit pen : $\mathcal{M}_n \to \mathbb{R}_+$ vérifiant pour tout $m \in \mathcal{M}_n$, et $\mu > 1$,

$$pen(m) \geqslant \mu \frac{D_m}{n} \left(1 + 6L_m + 8\sqrt{L_m}\right).$$

Soit $\tilde{f} = \hat{f}_{\hat{m}}$ où

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \gamma_n(\hat{f}_m) + pen(m),$$

sous l'hypothèse ($\mathbf{A_1}$),

$$\mathbb{E}_{f_0}[h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\tilde{f}})] \leqslant C_\mu \inf_{m\in\mathcal{M}_n} \left\{\mathcal{K}(\mathbb{P}_{f_0}, \mathbb{P}_{f_m}) + pen(m)\right\} + \frac{C(\rho,\mu,\Gamma,\Sigma)}{n} \tag{1.40}$$

où $C_\mu = \frac{2\mu^{1/3}}{\mu^{1/3}-1}$.

Le résultat du Théorème 1.6, contrairement à celui du Théorème 1.5, propose une pénalité qui s'affranchit de toute hypothèse sur la vraie fonction inconnue.

Simulations

Nous avons fait des études de simulations pour étudier les performances des pénalités que nous proposons. Les simulations portent sur le cas particulier où les modèles sont des fonctions constantes par morceaux. Nos critères ont été comparés aux AIC et BIC. Nous avons considéré deux situations : le cas où la vraie fonction est constante par morceaux, et le cas où elle ne l'est pas. Nous avons considéré plusieurs tailles d'échantillon, $n = 100, 200, \dots, 1000$. Les résultats des simulations ont permis de voir que le choix de la pénalité dépend de la taille de l'échantillon.

Les modèles sélectionnés par nos critères ont des performances en prédiction supérieures à celui sélectionné par le critère AIC pour tout n. Pour de petites tailles d'échantillon (inférieure à 200), le critère BIC a de meilleures performances. Pour des tailles comprises entre 200 et 400 nos critères ont des performances similaires à celles du BIC. Pour des tailles supérieures à 400, nos critères ont de meilleures performances que le BIC.

Calibration de la pénalité via l'heuristique de pente

Lorsque une pénalité est connue à une constante multiplicative près, *i.e.* $\text{pen}(m) = \mu \times \text{pen}_{shape}(m)$ où μ est un constante et $\text{pen}_{shape}(m)$ la forme générique de la pénalité alors la constante μ peut être estimée via l'heuristique de la pente introduite par Birgé et Massart (2007).

Rappelons que le modèle $\hat{m}$ est défini par

$$\hat{m} = \arg\min \Big\{ \gamma_n(\hat{f}_m) + \text{pen}(m) \Big\},$$

et l'on attend que le risque soit de l'ordre de

$$\min_{m \in \mathcal{M}} \Big[R(\hat{f}_m) - R(f_0) \Big].$$

La pénalité idéale, sélectionnant l'oracle m^*, s'écrit donc

$$\text{pen}_{id}(m) = R(\hat{f}_m) - R(f_0) - \gamma_n(\hat{f}_m),\ m \in \mathcal{M}_n,$$

soit ainsi

$$\text{pen}_{id}(m) = R(\hat{f}_m) - \gamma_n(\hat{f}_m) = \mathbb{E}[\gamma_n(\hat{f}_m)] - \gamma_n(\hat{f}_m).$$

Cette pénalité idéale dépend de la vraie fonction inconnue f_0. Une idée naturelle est de choisir une pénalité proche de la pénalité idéale. C'est l'objectif de la l'heuristique de pente. La pénalité idéale peut être décomposée comme suit

$$\begin{aligned} \text{pen}_{id}(m) &= R(\hat{f}_m) - \gamma_n(\hat{f}_m) \\ &= R(\hat{f}_m) - R(f_m) + \gamma_n(f_m) - \gamma_n(\hat{f}_m) + R(f_m) - \gamma_n(f_m) \\ &= v_m + \hat{v}_m + e_m, \end{aligned}$$

où $v_m = R(\hat{f}_m) - R(f_m)$, $\hat{v}_m = \gamma_n(f_m) - \gamma_n(\hat{f}_m)$, et $e_m = R(f_m) - \gamma_n(f_m)$. L'heuristique de pente repose sur deux points :

1- L'existence d'une pénalité minimale $\text{pen}_{\text{min}}(m) = \hat{v}_m$ telle que les pénalités inférieures à pen_{min} permettent de sélectionner les modèles les plus complexes. Tandis que les pénalités supérieures à pen_{min} permettent de sélectionner les modèles de complexité raisonnable.

2- D'après la loi des grands nombres, on peut supposer que $\gamma_n(f_m)$ est proche de son espérance $R(f_m)$ et donc $e_m \approx 0$. Par ailleurs, $\hat{v}_m$ est la version empirique de v_m, il est raisonnable de supposer qu'ils sont du même ordre, *i.e.* $v_m \approx \hat{v}_m$. La pénalité idéale est donc approchée par

$$2\hat{v}_m = 2\text{pen}_{min}(m) \approx \text{pen}_{id}(m).$$

Comme la pénalité est connue à une constant multiplicative près, la pénalité idéale est donc $\text{pen}_{id}(.) = \mu_{id}\text{pen}_{shape}(.)$. Ainsi,

$$\frac{\mu_{id}}{2}\text{pen}_{shape}(.)$$

est une valeur approchée de la pénalité minimale. En résumé pour estimer la pénalité idéale il faut estimer d'abord la pénalité minimale en cherchant μ_{min} tel que

$$\text{pen}_{min}(m) = \mu_{min} \times \text{pen}_{shape}(m).$$

En pratique, on se donne une grille de valeurs de μ : $\mu_1, \ldots, \mu_V$ où chaque μ_j conduit à la sélection du modèle $\hat{m}_{\mu_j}$ de dimension $D_{\hat{m}_{\mu_j}}$. La constante μ_{min} est estimée en utilisant le premier point de l'heuristique de pente. En effet, si on représente $D_{\hat{m}_{\mu_j}}$ en fonction μ_j alors μ_{min} est tel que $D_{\hat{m}_{\mu_j}}$ est grand pour les valeurs de $\mu_j < \mu_{min}$, et est raisonnablement petit pour les valeurs de $\mu_j > \mu_{min}$. Ainsi la constante μ_{min} correspond à la position du plus grand saut. Pour plus de détails sur cette méthode nous renvoyons le lecteur à l'article de Baudry et al. (2012) et Arlot et Massart (2009).

Cette heuristique a été validée théoriquement dans plusieurs contextes : dans le modèle de régression linéaire gaussien homoscédastique (Birgé et Massart (2007)) ; dans le modèle de régression linéaire hétéroscédastique (Arlot et Massart (2009)) ; en estimation de densité par les moindres carrés (Lerasle (2012)). Sans être validée théoriquement, elle a été utilisée avec succès dans plusieurs autres études : sélection de variables en apprentissage non supervisée (Bontemps et Toussile (2013)) ; en détection de ruptures (Lebarbier (2005)) parmi d'autres. D'autres exemples d'applications de l'heuristique de pente sont donnés dans Baudry et al. (2012). Ces nombreux exemples d'application de l'heuristique de pente laissent penser qu'elle peut être adaptée en régression logistique, y compris théoriquement.

2

Variables selection for identification of households at risk of having febrile episode in Dakar, Senegal

Sommaire

Most of large socio-epidemiological surveys involve a large number of explanatory variables. In such case, applying classical models such as logistic regression provides unstable and inefficient estimators and predictions. We consider here the problem of dimension reduction as a preliminary step before applying logistic regression model. Hence, our strategy consists in two steps. First, we propose to apply dimension reduction methods such as Lasso, Group Lasso and Random Forests to reduce the number of variables. Secondly, we perform logistic regression model by taking into account set of variables selected by previous dimension reduction methods. The prediction performances are thus evaluated and compared by using leave-one-out cross validation.

We apply our strategy to data collected in Actu-Palu study. This study was carried out among 379 households in Dakar. One of the aims of the study was to investigate association between socio-epidemiological characteristics related to household and the risk of having febrile episode in the household. Our strategy has reduced the number of variables from 71 to less than 15, which leads to a substantial gain of interpretability. Furthermore we compare the performance of each logistic models based on selected variables by computing the prediction errors. It appears that logistic regression models using selected variables outperformed the

one using all variables (full logistic regression model). Indeed, the prediction error for logistic regression using all variables (36.11%) was greater than the prediction error for logistic regression using selected variables, 19,44%, 22.22% and 25.39% for the variables selected by respectively Group Lasso, Lasso and Random Forests. The subset of variables selected by the Group Lasso was optimal (minimal prediction error). According to logistic regression model on the optimal subset of variables, households with more children from 2 to 10 years old were significantly more likely of having febrile episode. Households where household's head bought less expensive medications were significantly more likely of having children with febrile episode. Households that spent the most for the care of hospitalization were significantly less likely of having febrile episode. Households where Household's head were older had less risk of having febrile episode. Finally households that bought drugs on market were less likely of having febrile episode. This work underlies the importance of dimension reduction and proposes a strategy in two steps to deal with large socio-epidemiological surveys involving a large number of explanatory variables.

2.1 Introduction

Large public health surveys incorporate lot of various informations from very different nature and in large amount. These informations, besides medical and biological data, and the incidence of a phenomenon being studied may be social, economic or environmental in order to study the phenomenon in its contextual framework. Exploration of these informations requires suitable survey, with especially, long and multidirectional questions. The goal is clearly to enrich and refine the findings of these investigations. However, this enrichment is often followed by the difficulties to extract relevant information according to the aim of modelisation, interpretation or prevision. To explain a dependent variable as a function of the explanatory variables, the common strategy is to build a robust and interpretable model, using limited number of variables. With an high number of variables, which would contain many close informations, or be less informative, standard statistical methods and then, data analysis become inefficient. In statistics, the dimension is defined by the number of explanatory variables by individual (biological measures, socio-demographic informations *etc*). The term "*high dimension*" describes the situation where the number of explanatory variables per individual is large, *i.e.* of the same order as the number of individuals or higher. In such contexts most of usual methods fail to extract relevant information and to propose a robust and adapted model with good properties, for example prediction. In the presence of a large number of variables, dimension reduction problem arises therefore with acuity. Therefore, dimension reduction refers here to the reduction of the number of explanatory variables without significant loss of performance (interpretability, quality of prediction) of the models used. In other words the dimension reduction or the reduction of the number of variables will aim to select the most informative explanatory variables. In high dimensional situation, it is then important to develop dimension reduction strategies in order to select variables that will be introduced in the model. Dimension reduction will lead to increasing learning accuracy, and improving result comprehensibility.

This problem of dimension reduction is well known especially in Genome wide association studies (GWAS), where the number of variables (SNPs (Single Nucleotide Polymorphism...), genes expression level *etc* on the order of several thousands) is considerably higher than the number of individuals (on the order of 1,000 in favorable situations). This situation is also frequent in the large socio-epidemiological and contextual surveys where the accumulation of informations coming from different sources (epidemiological, sociologic, demographic, *etc*...) can lead to many variables (see McCarthy (2000), Marmot and Wilkinson (2005), Ompad *et al.* (2007).

To reduce the number of explanatory variables, the first classical approaches rely on univariate statistical tools, that is variable by variable. In this way, one can perform correlation tests between each explanatory variable and the dependent variable, and choose to only keep in the model, the variables related to the dependent variable. For instance, in GWAS, Dudoit *et al* (2002) propose to apply a preliminary reduction of the number of variables (genes) before performing a systematic comparison of several methods of discrimination for the classification of tumors based on microarray experiments. Although those unidimensional statistical approaches often allow to reduce significantly the number of explanatory variables, their main drawback is that the possible interaction between explanatory variables is not taken into account. This partly comes from the selection process, performed variable by variable.

Other approaches, based on transformation or compression of explanatory variables have been proposed to reduce the number of explanatory variables. These approaches include factorial methods such as Principal Component Analysis PCA (see Jolliffe (2002) and Massy (1965)) or Partial Least Squares (PLS) (de Jong (1993)). These methods allow to build "*informative components*" or "*crucial variables* ", which are the linear combinations of initial explanatory variables. The most informative components are then selected. The selected components, generally in few number, will then be used as a new explanatory variables in the model. For example Nguyen and Rocke (2002), in order to classify various human tumor based on the explanatory variables (genes), in very large numbers, proceeded in two steps : firstly a dimension reduction step using PCA and PLS regression to construct and select the most informative components ; secondly a classification step using Logistic Discrimination and Quadratic discriminant Analysis (QDA). So they used PCA and PLS to reduce the high p-dimensional gene space to a few r-dimensional gene component space which explains the total gene expression variation as much as possible. These r gene components were then used as new explanatory variables in a LD and QDA to classify human tumor. These dimension reduction methods using transformations of initial explanatory variables still have two major flaws :

- Since the informative components are the linear combinations of the initial variables, they involve all variables, even the less informative.
- In general, it is difficult to give a biological, sociological or epidemiological meaning to the selected components. These components are therefore hardly interpretable.

More recently, in the context of large databases, other dimension reduction approaches have been developed, such as the Lasso Tibshirani (1996) and Random Forests Breiman (2001). These methods could be an interesting solution, because of their numerous advantages. The Lasso type methods and Random Forests, unlike ACP, PLS, allow to reduce the number of explanatory variables while keeping their original structure, that is to say without alter the original representation of the initial explanatory variables. So they preserve the original meaning of the variables, hence offering the advantage of interpretability. In addition, unlike the methods based on univariate statistical techniques, these methods also allow to take into account the structure of explanatory variables and possible interactions between them. These methods were widely used in genomic analysis where databases reach hundreds of variables. For instance, Ghosh and Chinnaiyan (2005) have used the Lasso to select the bio-markers in the study of prostate cancer ; Dìaz-Uriarte and De Andres 2006 have used Random Forests on 9 sets of data to select the most discriminating genes. Other encouraging applications of these methods in the analysis of genomic data, where databases reach hundreds of variables, can be found for instance in Wu *et al.* (2009a), Li *et al.* (2011), Legarra *et al.* (2011) and Garcia-Magariños (2010) for the Lasso ; and in Goldstein *et al.* (2010, 2011), Meng *et al.* (2009), Bureau *et al.* (2005) for Random Forests. These methods still work when the number of variables is greater than the number of individuals.

Unlike genomic data, the variables of socio-epidemiological surveys, are from very different nature, a mixture of qualitative variables, including those with a large number of modalities, and quantitative variables. The variables of socio-epidemiological surveys are often derived from declarative questions. The common point between genomic and large socio-epidemiological studies is to offer a lot of variables that have to be reduced in order to analyze efficiently.

Our aim was to compare three dimension reduction methods, Lasso, Group

Lasso and Random Forests in the data obtained during a socio-epidemiological and contextual investigation concerning fever amongst children. The purpose of the study was to identify relevant variables to explain febrile episode amongst children from 2 to 10 years old in a household (home at risk). The explanatory variables concerned the characteristics of the members in household and domestic environment, material and monetary resources, socio-demographic and cultural characteristics. They also concerned practices of access to health care, particularly the habits of the household head in the event of fever amongst children from 2 to 10 years old. The dependent variable was a binary variable which encodes the home at risk (*home at risk* **vs** *home not at risk*), and was observed for each queried household. We are thus in a context of supervised classification.

The chapter is organized as follows. In Section 2.2, we start with database presentation, and then, briefly describe the statistical tools. In Section 2.3 we present our results , variables selection and models validations, through the study of prediction errors. The results are followed by discussion and conclusion.

2.2 Methods

2.2.1 Population

The data set relies on the follow-up of a cohort of 379 households located in eight districts of Pikine (see figure 2.1), on the outskirts of Dakar and carried out by ISED of UCAD (Senegal), IRD, (UMR 216) and the CERDI. They monitored households and noted the occurrence of health problems. In each household, they collected sociodemographic, cultural and environemental informations. The purpose was to identify risk factors of health problems. We considered a follow up of 4 months. In this analysis, we considered fever amongst children. The study population was a subgroup of a transversal study on malaria and the use of care, in case of fever amongst children in the urban area of Dakar : ACTU-PALU project (see Diallo *et al.* (2012, 2012)), which concern a representative sample of the city of Dakar (3000 households, from 50 districts in the urban area of Dakar).

2.2.2 Data collection

Explanatory variables

Socio-demographic and contextual data were collected using two questionnaires : a household questionnaire, which has collected the socio-demographic and economic characteristics, lifestyle informations of the family (income, environment *etc*). Head of household questionnaire, which has collected information concerning the head of household : his knowledge about the risks of health, his level of study, modes of managing fever amongst children *etc*.

Dependent variable : indicator of home at risk

Investigators have visited household each two months. During each of these visits, the febrile episodes amongst children from 2 to 10 years old in the household were count. A home was said at risk when there was at least one febrile episode declared. The dependent variable was a binary variable which encodes the homes at risk : *home at risk* **vs** *home not at risk*.

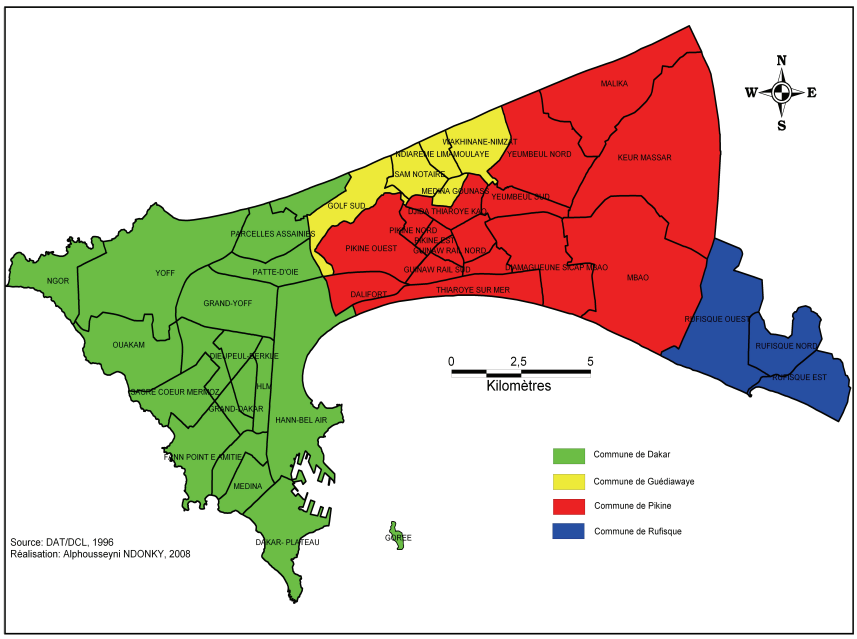


FIGURE 2.1 – *Urban area of Dakar*

Preprocessing of data

We have initially made a preliminary processing of dataset to make it efficiently usable. For instance, this treatment was about variables from nested questions. For example, concerning the activity of the head of household, some questions were asked and the answer to a question conditioned the answer to the following : *Do you have an activity ? What is this activity ? What is the sector of this activity ?* It is clear that these three questions contain close informations. Then these three variables should not be used together in the analysis. Just only one variable can be used. The second preprocessing of data concerns aggregation of information from several variables. This is the case for example of the variables which dealt with knowledge about disease and treatment. A score of 1 was assigned to a correct answer and 0 to a false answer. We then created a new variable, which is the average of scores obtained on each of the questions about the knowledge. Finally, for categorical variables with poorly represented modalities, we have grouped these modalities. This regroupment of modalities was done by caring about biological or socio-epidemiological sense. After this preliminary treatment the final database contained 71 explanatory variables for 379 households.

2.2.3 Statistical methods

In the dataset, the variable of interest is binary : indicator of *home at risk* **vs** *homes not at risk*. In this context of supervised classification, we choose a simple and relevant model to predict the homes at risk : logistic regression model (see Hilbe (2009) and Menard (2002)). Let us start by its description.

Logistic regression model

One observes $(x_1, Y_1), \ldots, (x_n, Y_n)$, where Y_i is a binary variable (which can take the values 1 = home at risk, 0 = home not at risk) to be explained and $x_i = (x_{i1}, \ldots, x_{ip})$ is the whole set of explanatory variables. For $i = 1, \ldots, n$, logistic regression model proposes to modelize the link between Y_i and x_i by the following relation

$$\mathbb{P}(Y_i = 1 \mid x_i) = \frac{\exp(\beta_0 x_i)}{1 + \exp(\beta_0 x_i)}.$$

The parameter β_0 is unknown and has to be estimated using the observations $(x_1, Y_1), \ldots, (x_n, Y_n)$ on n individuals.

As we mentioned above, logistic regression with all explanatory variables (that is without priorly reducing the number of explanatory variables), will not be very successful since it would introduce noise through the variables that have a low prediction power. Consequently, this will degrade the performance of logistic model, especially in term of prediction accuracy. This partly comes from the fact that the number of individuals is not considerably higher than the number of variables. Moreover, the explanatory variables are mostly correlated. Hence, the complete logistic model will furnish estimators with high variance and this will lead to small accurate prediction (see Bull (2007) and Greenland *et al.* (2000). That is why, we choose a two step methodology. The first step consists in reducing the number of explanatory variables. The second step relies on performing logistic regression models based on the resulting sets of selected variables after step 1, and compare their prediction errors. We have used two main approaches to reduce the number of variables : the first, based on penalized negative log likelihood : Lasso, Group Lasso ; and the second method based on the construction of an hierarchy of explanatory variables using Random Forests.

Dimension reduction methods based on Lasso and Group Lasso

• The Lasso

The Lasso (Least Absolute Shrinkage and Selection Operator) Tibshirani (1996) is a popular method for variables selection or dimension reduction. Lasso regression is widely used in domains with massive datasets, such as genomics, where the number p of variables may be of the same order or largely higher than the number of individuals n. The lasso estimator $\hat{\beta}_{Lasso}$ is the solution of the following optimization problem.

$$\hat{\beta}_{Lasso} = \arg\min_{\beta \in \mathbb{R}^p} \Big\{ \gamma_n(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \Big\}, \tag{2.1}$$

where $\gamma_n(\beta)$ is the negative log likelihood function and $\lambda > 0$ the regularization parameter. The second term is called "ℓ_1 penalty" since relies on the usual ℓ_1-norm of β which offers selection properties as well as sparsity. Sparsity means here that we enforce the model to include few explanatory variables, at least few with respect to p, which corresponds to the full model. The lasso presents some advantages

- Lasso automatically select variables : some coefficients are estimated to be exactly zero for sufficiently high values of λ. These will represent variables that have no discriminatory power. The model thus gains in interpretability.

- The optimization problem to solve is convex, namely relatively easy to resolve, even if there is no analytical form of the solution in the general case. Algorithms such as *coordinate descent* by Friedman *et al.* (2010) and *predictor-corrector* by Park and Hastie (2007) resolve the issue (2.1).
- Lasso is important for its stability and its parsimony.

• The Group Lasso

Situations where prior information of grouped explanatory variables may appear. Group Lasso Yuan and Lin (2006) or Meier *et al.* (2008) is a group version of the Lasso that uses a penalty which allows to select variables in groups. Variable groups must therefore be known in advance. An important context where the explanatory variables fall within a structure of group, appears in the case of categorical explanatory variables. In this case for example, the Lasso select just a part of the modalities of a categorical variable. While the Group Lasso selects all modalities of a categorical variable or reject all. The Group Lasso is thus a good alternative in this case. The Group Lasso $\hat{\beta}_{GL}$ estimator is defined as

$$\hat{\beta}_{GL} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \gamma_n(\beta) + \lambda \sum_{g=1}^{G} \|\beta_g\|_2 \right\} \tag{2.2}$$

The regularization parameter $\lambda > 0$ is used to adjust the trade-off between minimizing the loss and finding a solution which is sparse at the group level. Group Lasso shares the same advantage as the Lasso.

• Optimal choice of λ

For both, Lasso and Group Lasso methods, when $\lambda = 0$, we find the classical maximum likelihood (not penalized) for the full model. And for very high values of λ, all the parameters are estimated exactly equal to zero. So λ has to be well chosen in order to make a good balance between goodness-of-fit and parsimony. All questions related to correct estimation and model selection are actually conditional to the correct choice of λ, since this value roughly speaking determines the size of the model selected by the Lasso. The parameter λ is tuned by cross-validation (see Tibshirani (1996)) or using AIC and BIC criterion minimization. Those criteria provide guides on how to tune the amount of regularization in the light of prediction problem but does not provide a desirable guide in the light of estimation problem or selection problem. Here we focus on prediction properties. Practically, given a grid of tuning parameters $\lambda_1, \ldots, \lambda_k$, for each λ_j, we perform Lasso (or Group Lasso) algorithm. Each λ_j is related to a subset S_{λ_j} of selected variables. For each λ_j we perform a logistic regression model using the subset S_{λ_j} of selected variables, and compute prediction errors using cross-validation (leave-one-out). Then, the optimal subset of selected variables is chosen as the subset that minimizes the prediction error. These methods have been implemented through *glmnet* and *grplasso* R software packages.

Dimension reduction based on Random Forests and variable importance

By definition, the Lasso and Group Lasso are strongly related to the assumption that data are distributed according to a logistic model (parametric). This assumption may not be satisfied in practice. So we choose to consider an alternative non-parametric reduction method, called Random Forests. One important point is that Random Forests method is more robust because its use does not rely on known model structure.

Random Forests, is an ensemble learning method firstly introduced by Breiman (2001). It is based on an aggregation of decision trees, more precisely, Random forests are based on aggregation CART trees, these trees being randomly built. Each classification tree is constructed using a bootstrap sample, and at each node a random subset of the variables is selected, and searched over to find the optimal split. The trees here are not pruned, so they have a large variance. While the CART algorithm uses pruned tree for increased stability, RF leaves the tree unpruned, as bagging is used to decrease the variance created by the lack of pruning. Prediction of a new data is obtained by aggregating the predictions of trees, using majority votes. Random Forest has several characteristics that make it suitable for Actu-Palu data :

- It can be used when there are many variables (even more than n the number of observations).
- It can be used with qualitative and quantitative explanatory variables, and is able of capturing interactions between them.
- It presents good predictive performance even with many irrelevant variables.

Variable importance which is computed for each explanatory variable, refers to a quantitative measure of importance of explanatory variables. For each variable, it is defined as the difference in average of the tree performance before, and after randomly disrupts values observed by this variable. Intuitively, for important variables, lot of random permutations of their values will induce a huge prediction error. Conversely, if the permutations have almost no impact on the error, the variable is considered as less important.

We have used two methods of variables selection by random forests, both based on the hierarchy of variables given by value of importance.

• Selection using threshold

Dimension reduction using threshold has been proceeded by ranking variables according to their value of importance. Only variables with value of importance higher than a fixed threshold are selected. The choice of the threshold is crucial in the selection process. Following Strobl *et al.* (2009), we have chosen the absolute value of the smallest value of importance. The idea around this thresholding is that irrelevant variables have a value of importance which fluctuates around the value of importance zero. In the rest of the chapter this method will be called *RF.threshold*

• Selection using nested models

We built p nested models (random forests). The first with the most important variable, the second with the two most important variables and so on until the model with all variables. For each of these models we calculate the OOB error. We choose the model that has the smallest OOB error. In the following, this method of selection will be called *RFnested*.

Random Forests have been implemented using the R software package called *randomForest*. In this algorithm, two parameters have to be tuned. Those paramaters are *ntree*, the number of trees in the forest and *mtry*, the number of input variables randomly chosen at each split. Here we have used *ntree* = 2000 and the default value *mtry* = $\sqrt{p}$, where p is the maximal number of explanatory variables.

Comparison of the dimension reduction methods

We have compared the different dimension reduction methods by assessing the ability of subsets selected by each methods to predict the homes at risk in a logistic

model. For this, we have calculated the rate prediction errors using cross validation (leave-one-out) : for each household $i = 1, \ldots, n$, we have performed the following steps :

1 We have estimated the parameters of logistic regression model using all households except the *i th* household ;
2 We have predicted the response $\hat{y}_i$ (home at risk or not at risk) of the *i th* household using the parameters estimated in step 1 ;
3 We have calculated the prediction error of the household i, by comparing the value $\hat{y}_i$ predicted in step 2 with the true (observed value) value of y_i.

The rate of bad predictions is then the average of errors over all households.

The performance of dimension reduction methods has also been evaluated by performing Hosmer-Lemershow (H-L) test . The H-L test is a statistical test of goodness-of-fit for logistic regression models. It is based on the following hypothesis test (H_0) : The fitted model is correct versus (H_1) : The fitted model is not correct. We accept the hypothesis (H_0) if the $p-$value is greater than 5% and reject otherwise. Thus, the larger the p-value, the better logistic regression model.

2.3 Results

We recall that an household is said at risk if there were at least one febrile episode in at least one of the two visits of investigators. This leads to a binary dependent variable with 23.16% of homes at risk against 76.84% of homes not at risk.

Dimension reduction based on Random Forests

As we mention above, we have first applied a dimension reduction by Random Forests using a threshold proposed by Strobl (2009) (*RF.threshold*, see Figure 4.2). The variables that have value of importance higher than the threshold have been identified as being relevant variables for the prediction of the homes at risk. According to Table 2.2 and Table 2.3, the following nine variables were selected : *" Number of children from 2 to 10 years old in the household ", " Marital status ", " Level of study", "Time spends from home to the activity ", "Sector of activity", "The type of toilet of the head of household ", "Having friends or confidants in the district", "Knowledge on the treatment", and "Do you ever buy drugs to the sellers in the market ?"*. Note that a large proportion of variables (87%) present a value of importance lower than the threshold value (Figure 4.2). According to the construction of the threshold (see Strobl (2009)), the values of importance of these variables fluctuate around the value of importance zero. These variables seem to provide no additional information in the prediction of homes at risk.

Dimension reduction using nested models (*RF.nested*) has selected exactly the same variables as *RF.threshold* (see Table 2.2 and Table 2.3).

Dimension reduction based on Lasso and Group Lasso

Dimension reduction using Lasso or Group Lasso was done by estimating coefficients of some variables to be exactly zero. These variables correspond to ones that have no discriminatory power between homes at risk and homes not at risk. While those with non zero coefficients represent variables that can successfully discriminate homes at risk and homes not at risk. According to Table 2.2 and Table 2.3, with

the Lasso method, the following six variables have been selected : *"Do you ever buy drugs to the sellers in the market ?", "Marital status", and " The type of toilet of the head of household"*. For its part the Group Lasso allowed to select 15 variables : *"Number of children from 2 to 10 years old in the household ", " Knowledge on the treatment", " Having friends or confidants in the district ", "Do you know that the infusion/palujec/quinine injection treats malaria ?", "Cost for hospital care during the last 12 months ", " Which fuel do you often use for the preparation of the meal ?", "What is your main lighting mode ?", "Do you know that there are some medications far less expensive than others in drugstores ?", "Do you ever buy drugs to the sellers in the market ?", "When you did not fully understand the explanations to give medication to your child, how do you do ?", " Do you discuss about health issues ?", "Time spends from home to the activity", "Level of study", "Marital status", and "Age"*.

Comparison of dimension reduction methods

Table2.1 shows the rate of prediction errors and the p-value of H - L test for logistic regression models that take into account each subset of variables selected by each dimension reduction method. This table also shows the number of variables selected by each dimension reduction methods. The p-values of H-L test for logistic regression models were higher than 0.05, since p-values are respectively $p = 0.85$, $p = 0.57$, and $p = 0.78$ for the variables selected by the Lasso, Group Lasso, and Random Forests respectively. We conclude that logistic regression models taking into account variables selected by each reduction dimension method fit well the data. Group Lasso has selected 15 variables , Lasso 3 variables and Random Forests 9 variables. Then these dimension reduction methods helped significantly to reduce the number of variables, from 71 variables to less than 15. Logistic regression model using all the 71 variables (full logistic regression model) had the highest prediction error (36.11%). Prediction error was 19.44% for the Group Lasso, 22.22% for the Lasso, and 25.39% for the Random Forests. We infer that dimension reduction methods greatly improved the qualities of logistic regression models, compared with the full logistic regression model. The method providing the lowest prediction error, *i.e.* the best prediction performance amongst the three dimension reduction methods was the Group Lasso with prediction error of 19.44%.

Logistic regression on the most predictive subset of selected variables

Let us briefly study the resulting logistic regression model obtained using the variables selected by the Group Lasso. Group Lasso has reduced the number of variables from 71 to 15. Since 15 variables seemed quite large, in order to improve estimations, we have used stepwise selection by AIC to select the most relevant variables amongst them. This second selection allowed to select 8 variables. The estimated values of coefficients for logistic regression model using these 8 variables are presented in the Table 2.5. According to this table, mainly five variables were linked with the risk of having child fever in household : households with more children from 2 to 10 years old were significantly more likely of having febrile episode ($p = 0.0036$). Households where household's head bought less expensive medications were significantly more likely of having children with febrile episode ($p = 0.0381$). Households that spent the most for the care of hospitalization were significantly less likely of having febrile episode ($p = 0.0437$). Households where Household's head were older had less risk of having febrile episode ($p = 0.0294$).

Finally households that bought drugs on market were less likely of having febrile episode ($p = 0.0244$).

2.4 Discussion

The aim of this work was to identify a strategy for reducing the number of the explanatory variables in a socio-epidemiological and contextual survey, in order to use logistic regression to predict the homes at risk. In other words, our aim was to select from the large number of explanatory variables collected, a small number of informative variables that allows to predict home at risk. We have used first a parametric approaches based on the use of the penalized criteria : Lasso and Group Lasso. Secondly, we have used a non-parametric approach based on the value of importance of variables provided by Random Forests method.

The originality of these approaches of dimension reduction lies on their ability to take into account all the variables and the possible interaction between them. These methods are not based on screening step using correlation tests, which have drawbacks as presented in the introduction. These dimension reduction methods are also original because, unlike PCA and PLS, there is no transformation of the initial variables in the reduction process. Hence they provide interpretable models, involving initial variables rather than linear combinations of variables, preserving the original semantics of the variables.

After reducing the number of variables, we have used, for each reduced subset of variables, a logistic regression model to predict home at risk. We have then compared these methods by evaluating their prediction quality.

Group Lasso as the most predictive method

In this study, Group Lasso was retained as the optimal dimension reduction method according to his predictive performance. Indeed, the logistic regression model using selected variables by Group Lasso (see Table 2.1) has the smallest prediction error. In addition, 5 variables on 8 were significantly related to homes at risk in logistic regression model (Table 2.5). The advantage of Group Lasso method compared to the Random Forests method is probably due to the use of logistic model to evaluate the performance of dimension reduction methods. Indeed, the Group Lasso method is based on assumption of logistic model (unlike the Random Forests), which is coherent with the use of logistic regression model again. The advantage of Group Lasso method compared to Lasso method is probably due to the nature of the explanatory variables. Indeed most of explanatory variables are categorial, which is typically a case for which Group Lasso (unlike the Lasso) outperforms, since its definition takes into account the group structure of categorial variables in the selection procedure Meier *et al.* (2008).

Group Lasso as a screening procedure

Dimension reduction using Group Lasso has retained 15 variables. It is known that Group Lasso often selects higher set of variables. At this stage we choose to perform an additional variables selection procedure among these 15 variables. For such a number of explanatory variables, it is also known that variables selection using AIC criterion perform better than Lasso. For those reasons we have proceeded to a last variables selection procedure using stepwise model selection based on AIC

criterion. In that way we have more accurate selection and better logistic regression analysis. Of course we could not directly use stepwise selection on the 71 variables because it is adapted to modest number of variables.

High dimension setting

It can be surprising that genomic studies use thousands of variables in dimension reduction methods (Ghosh and Chinnaiyan (2005) Wu *et al.* (2009a) Li *et al.* (2011)), while socio-epidemiological data are limited to tens. Indeed large socio-epidemiological and contextual data are of order of few hundreds. As explained above, with these data, a preprocessing can be easily done by considering nested, filter or redundant questions. Redundant questions are sometimes due to the fact that questionnaires are often developed by specialists in diverse fields (doctors, sociologists, economists *etc*). They can ask quite close questions that we have to identify before using statistical analysis. However after such a preprocessing for genomic data, there still remains many variables.

In this work, after a preprocessing of data, we have applied dimension reduction methods to 71 explanatory variables. Although this number is considerably lower than the number of variables (genes) in the genomic data, it remains quite large, in the sense that they should not be all used in a logistic regression model. Indeed, for the sake of numerical precision or for interpretability, logistic regression with 71 variables is not efficient. We stress that in this study *high dimension* means that the number of variables is too large to use logistic regression.

Improvement due to dimension reduction procedure

As showed in Table 2.1, logistic regression with all 71 explanatory variables is not very efficient, since it has high prediction error, 36.11%. The dimension reduction, from 71 variables to less than 15, allowed to reduce the prediction error. It therefore reflected a gain of information and an improvement in the discrimination of homes at risk due to the dimension reduction. However the best prediction error (19.44%) is not too small, and thus highlights the difficulties of discriminate homes at risk using these explanatory variables. These difficulties can be due to two reasons : firstly, explanatory variables come from declarative questions, which can present some incorrect responses. Indeed, persons interviewed, driven by shame, fear, a desire to show a correct attitude, may give incorrect or simply imprecise answers. Secondly, concerning the response variable (*home at risk* **vs** *home not at risk*), some families had no thermometer and consequently the declaration of fever can be variable. This approximation could explain a quite high prediction error. This difficulty of analysing socio-epidemiological data is frequent and can be found for instance in Rondet *et al.* (2013) and Vallée and Chauvin (2012). However, this analysis bring a lots of interesting results and all approximation presented above are constitutive of these studies and are taking in account in the interpretation of the results.

The rate of missing values for the dimension reduction methods

There were missing data in the dataset (no answer to a question) and a missing data excluded all the data of the household. These problem of missing data is recurrent in large socio-epidemiological surveys where families hesitate to answer all questions. We have analyzed these missing data. They were randomly dis-

tributed between the households. Households with missing data did not have a specific profile. Thus we have chosen to delete households with missing data. To our knowledge, exclude these households does not, made bias in dimension reduction methods. We have worked with 66.3% (n=252) of the included households for dimension reduction methods.

However, this exclusion of households with missing data concerned only dimension reduction methods. Once number of variables has been reduced, we have applied logistic model to households that have no missing data in selected variables. By doing so we have considerably reduced the number of missing data, and then, the number of excluded households (see Table 2.5). Indeed, the more variables the are, the more data are missing, and then more household to exclude. The variables selected with the Lasso for example did not have missing data. Thus for these variables no household has been excluded before using logistic regression. For Group Lasso, 94.2% (n=357) of included households in logistic regression model.

The selection of the variable *"number of children from 2 to 10 years old in the household"*

The variable *"number of children from 2 to 10 years old in the household"* seems to play a capital role in the prediction of homes at risk. Indeed, on the one hand, it was selected by 2 methods of dimension reduction. On the other hand, in RF method, it emerges as the third most important, (see Figure 4.2). Finally, according to Table 2.5, it is the most significant (p = 0.0036) in logistic regression model using the subset of variables selected by the optimal method according to predictive performance. Indeed, the probability that there is at least one febrile episode in the household increases with the number of children from 2 to 10 years old in the household. Moreover, contamination between children increases the proportional risk for children to get fever.

These variables which are linked with economic environment of the household or sociodemographic information are in adequacy with results in others studies. This result shows the capacity of reduction method to identify coherent risk factors for fever in household.

2.5 Conclusion

Large socio-epidemiological surveys involve a large number of explanatory variables. In such case, applying classical models such as logistic regression becomes inefficient. We need to reduce the number of variables proposed in the analysis model. The current method used to reduce the number of variables like univariate statistical techniques , factorial methods, PCA and PLS regression are imperfect .

Reduction methods were widely used in genomic analysis (Lasso, Group Lasso or Random Forests method). These methods allow to reduce the number of explanatory variables while keeping their original structure and allow to take into account possible interactions between them. Unlike genomic data, variables of socio-epidemiological surveys are of very different nature, a mixture of qualitative variables, including some with a large number of modalities. So we have tested these methods with socio-epidemiological dataset. Our results show that the reduction method permit to identify an acceptable dataset to use with logistic regression. Significative variables linked with fever in household are in accordance with findings

published in other study, confirming the interest of group Lasso method to reduce number of variables within socio-epidemiological data set.

Although these dimension reduction methods have been used on the ACTUPALU dataset, this work constitutes an important afterthought, that may interest many socio-epidemiological and contextual studies. These studies are frequent, specially in Africa, and involve many contextual data and are often difficult to analyze without specific statistical tools.

However, the main objective of this article was not to get a precise analysis of risk factors but we just wanted to test reduction methods in contextual studies with qualitative and quantitative variables. We need to push forward the pre-selection of variables and compare the result of many sets of variables. Moreover, identification of variables like "*buy medicine in informal market* " or "*age of mother*" suggests cross informations with scholar level or social network and need to be interpreted more precisely.

methods	.	Lasso	Group Lasso	*RF.threshold*	*RF.nested*
Error (%)	36.11	22.22	**19.44**	25.39	25.39
H-L test	1	**0.85**	0.57	0.78	0.78
Number of selected variables	71	3	15	9	9

Table 2.1 – *Error : rate of bad predictions in a logistic regression model that takes into account the variables selected by each dimension reduction method. H - L test : p-value of Hosmer and Lemeshow test.*

Variables	Lasso	Group Lasso	*RF.threshold*	*RF.nested*
F100		√		
F101	√	√	√	√
F114		√	√	√
F204			√	√
F212		√	√	√
F800		√		
F808		√		
F812	√	√	√	√
F830		√		
M307	√		√	√
M313		√		
M315		√		
M606		√		
amisConfidents		√	√	√
ScoreConTrait		√	√	√
varINJ		√		
Nbre_total_d_enfants_de_2__10 ans		√	√	√

Table 2.2 – *Variables selected by each dimension reduction method*

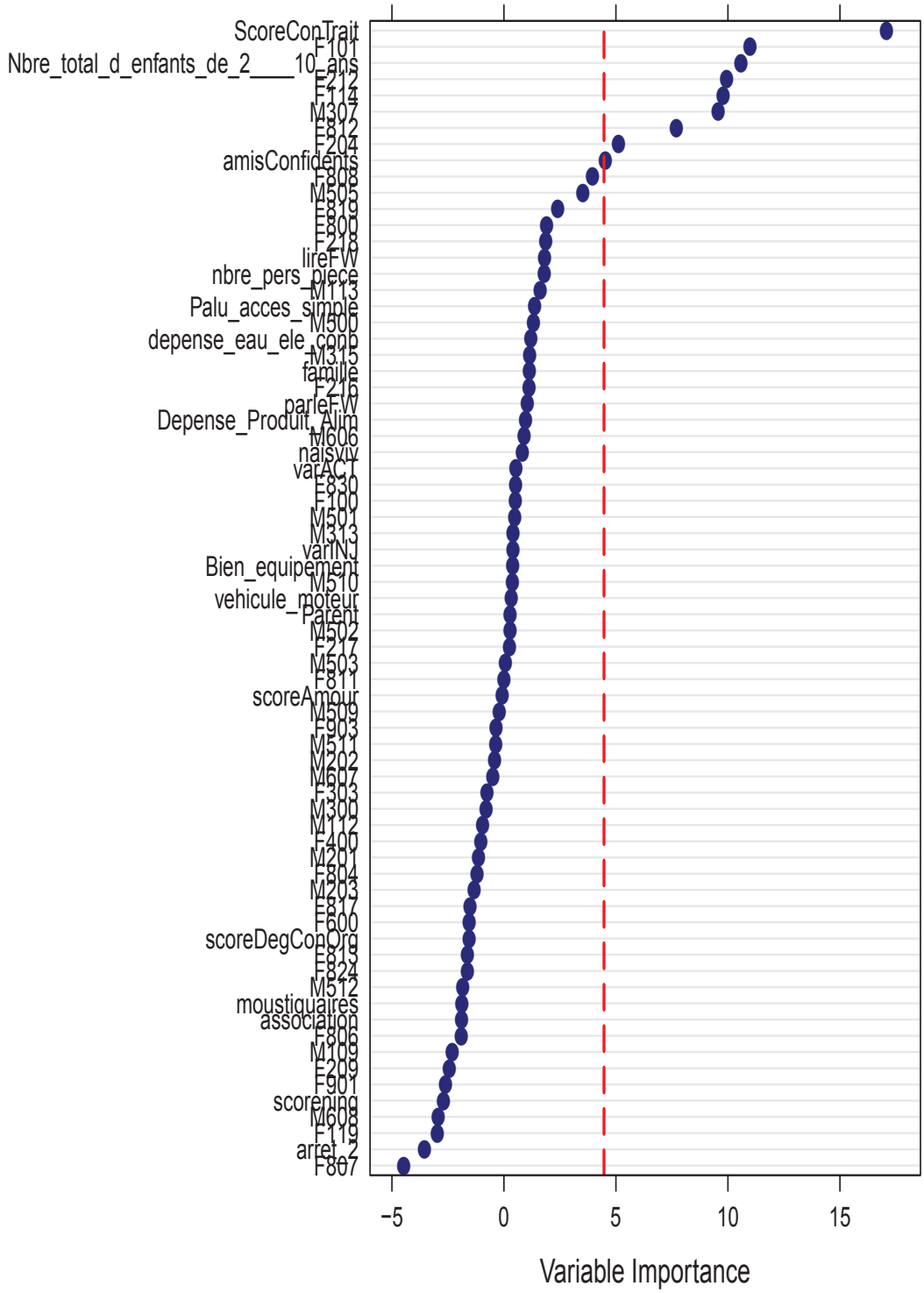


FIGURE 2.2 – *Value of importance for each variable*

Variables	labels
F100	Age
F101	Marital status
F114	Level of study
F204	Sector of activity
F212	Time spends from home to the activity
F800	Do you discuss about health issues ?
F808	When you did not fully understand the explanations to give medication to your child, how do you do ?
F812	Do you ever buy drugs to the sellers in the market ?
F830	Do you know that there are some medications far less expensive than others in drugstores ?
M307	The type of toilet of the head of household
M313	What is your main lighting mode ?
M315	Which fuel do you often use for the preparation of the meal ?
M606	Cost for hospital care during the last 12 months
amisConfidents	Having friends or confidants in the district
ScoreConTrait	Knowledge on the treatment
varINJ	Do you know that the infusion/palujec/quinine injection treats malaria ?
Nbre_total_d_enfants_de_2__10 ans	Number of children from 2 to 10 years old in the household

Table 2.3 – *Label of variables selected by at least one dimension reduction method.*

methods	Lasso	Group Lasso	*RF.threshold*	*RF.nested*
rate of households	0.00 %	5.8%	4.2%	4.2%

Table 2.4 – *Rate of households with missing data in the subsets of variables selected by each of the dimension reduction methods.*

		estimate	OR	IC	pvalue
(Intercept)		-1.1357	0.3212	[0.0398 ; 2.2049]	0.2622
F812	No	-	-	-	0.0244 *
	yes	-0.8530	0.4261	[0.1927 ; 0.8636]	
F830	No	-	-	-	0.0381*
	Yes	0.6096	1.8396	[1.0456 ; 3.3207]	
M315	Other	-	-	-	0.1092
	Gas	1.0655	2.9024	[0.8959 ; 13.1820]	
F100		-0.0310	0.9694	[0.9419 ; 0.9961]	0.0294 *
F212		0.0065	1.0065	[0.9974 ; 1.0153]	0.1447
ScoreConTrait		-1.1422	0.3191	[0.0778 ; 1.3050]	0.1111
M606		-0.0000	1.0000	[1.0000 ; 1.0000]	0.0437*
Nbre_total_d_enfants_de_2__10 ans		0.2496	1.2836	[1.0854 ;1.5208]	0.0036 **

Table 2.5 – *Logistic regression model using variables selected by the optimal method, Group Lasso*

3 Non-asymptotic Oracle Inequalities for the Lasso and Group Lasso in high dimensional logistic model

Sommaire

We consider the problem of estimating a function f_0 in logistic regression model. We propose to estimate this function f_0 by a sparse approximation build as a linear combination of elements of a given dictionary of p functions. This sparse approximation is selected by the Lasso or Group Lasso procedure. In this context, we state non asymptotic oracle inequalities for Lasso and Group Lasso under restricted eigenvalue assumption as introduced in Bickel et al. (2009). Those theoretical results are illustrated through a simulation study.

3.1 Introduction

During the last few years, logistic regression problems with more and more high-dimensional data occur in a wide variety of scientific fields, especially in studies that attempt to find risk factors for disease and clinical outcomes. For example in gene expression data analysis or in genome wide association analysis the number p of predictors may be of the same order or largely higher than the sample size n (thousands p of predictors for only a few dozens of individuals n, see for instance Garcia-Magariños et al. (2010) or Wu et al. (2009b)). In this context the considered model is often what we call here *usual* logistic regression. It is given by

$$\mathbb{P}(Y_i = 1) = \pi(z_i^T \beta_0) = \frac{\exp(z_i^T \beta_0)}{1 + \exp(z_i^T \beta_0)}, \tag{3.1}$$

where one observes n couples (z_1, Y_1),...,$(z_n, Y_n) \in \mathbb{R}^d \times \{0, 1\}$, and β_0 is the unknown parameter to estimate. Throughout the chapter, we consider a fixed design setting (i.e $z_1, \ldots, z_n$ are considered deterministic).

In this chapter, we consider a more general logistic model described by

$$\mathbb{P}(Y_i = 1) = \frac{\exp(f_0(z_i))}{1 + \exp(f_0(z_i))}, \tag{3.2}$$

where the outputs $Y_i \in \{0, 1\}$, $i = 1, \ldots, n$ are independent and f_0 (not necessarily linear) is an unknown function (Hastie (1983)). We aim at estimating f_0 by constructing a suitable approximation. More precisely we estimate f_0 by a sparse approximation of linear combination of elements of a given dictionary of functions $\mathbb{D} = \{\phi_1, \ldots, \phi_p\} : \hat{f}(.) := \sum_{j=1}^{p} \hat{\beta}_j \phi_j(.)$. Our purpose expresses the belief that, in many instances, even if p is large, only a subset of $\mathbb{D}$ may be needed to approximate f_0 well. This construction can be done by minimizing the empirical risk. However, it is well-known that with a large number of parameters in high dimensional data situations, direct minimization of empirical risk can lead to *Overfitting* : the classifier can only behave well in training set, and can be bad in test set. The procedure would also be unstable : since empirical risk is data dependent, hence random, small change in the data can lead to very different estimators. Penalization is used to overcome those drawbacks. One could use ℓ_0 penalization, *i.e.* penalized by the number of non zero coefficients (see for instance AIC, BIC Akaike (1974), Schwarz (1978a)). Such a penalization would produce interpretable models, but leads to non convex optimization and there is not efficient algorithm to solve this problem in high dimensional framework. Tibshirani (1996) proposes to use ℓ_1 penalization, which is a regularization technique for simultaneous estimation and selection. This penalization leads to convex optimization and is important from computational point of view (as well as from theoretical point of view). As a consequence of the optimality conditions, regularization by the ℓ_1 penalty tends to produce some coefficients that are exactly zero and shrink others, thus the name of Lasso (Least Absolute Shrinkage and Selection Operator). There exist some algorithms to solve this convex problem, *glmnet* (see Friedman et al. (2010)), *predictor-corector* (see Park et Hastie (2007)) among the others.

A related Lasso-type procedure is the Group Lasso, where the covariates are assumed to be clustered in groups, and instead of ℓ_1-penalty (summing the absolute values of each individual loading) the sum of Euclidean norms of the loadings in each group is used. It shares the same kind of properties as the Lasso, but encourages

predictors to be selected in groups. This is useful when the set of predictors is partitioned into prescribed groups, only few being relevant in the estimation process. Group Lasso has numerous applications : when categorical predictors (factors) are present, the Lasso solution is not adequate since it only selects individual dummy variables instead of whole factors. In this case, categorical variables are usually represented as groups of dummy variables. In speech and signal processing for example, the groups may represent different frequency bands (see McAuley et al. (2005)).

Previously known results. Recently, a great deal of attention has been focused on ℓ_1-penalized based estimators. Most of this attention concerns regression models and ℓ_1-penalized least squares estimator of parameters in high dimensional linear and non linear additive regression. Among them one can cite Bunea et al. (2006; 2007b;a), Massart et Meynet (2011), who have studied the Lasso for linear model in nonparametric setting and proved sparsity oracle inequalities. Similar sparsity oracle inequalities are proved in Bickel et al. (2009), and those results hold under the so-called *restricted eigenvalue assumption* on the Gram matrix. Those kind of results have been recently stated for the variants of the Lasso. For instance Lounici et al. (2011) under a group version of *restricted eigenvalue assumption* stated oracle inequalities in linear gaussian noise model under Group sparsity. Those results lead to the refinements of their previous results for multi-task learning (see Lounici et al. (2009)). The behavior of the Lasso and Group Lasso regarding their selection and estimation properties have been studied in : Knight et Fu (2000), Meinshausen et Bühlmann (2006), Zhao et Yu (2006), Osborne et al. (2000), Zhang et Huang (2008), Meinshausen et Yu (2009) for Lasso in linear regression ; Chesneau et Hebiri (2008), Nardi et Rinaldo (2008) for Group Lasso in linear regression ; Ravikumar et al. (2009), Meier et al. (2009), Huang et al. (2010) for additive models. Few results on the Lasso and Group Lasso concern logistic regression model. Most of them are asymptotic results and concern the *usual* logistic regression model defined by (3.1). Zou (2006) shows consistency in variable selection for adaptive Lasso in generalized linear models when the number of covariables p is fixed. Huang et al. (2008) prove sign consistency and estimation consistency for high-dimensional logistic regression. Meir et al. (2008) shown consistency for the Group Lasso in *usual* logistic model (3.1). To our knowledge there are only two non asymptotic results for the Lasso in logistic model : the first one is from Bach (2010), who provided bounds for excess risk (generalization performance) and estimation error in the case of *usual* logistic regression model under *restricted eigenvalue assumption* on the weighted Gram matrix. The second one is from van de Geer (2008), who established non asymptotic oracle inequality for Lasso in high dimensional generalized linear models with Lipschitz loss functions. Non asymptotic results concerning Group Lasso for logistic regression model have been established by Negahban et al. (2012), and more recently by Blazère et al. (2014), both with the assumption that f_0 is linear.

In this chapter, we state general non asymptotic oracle inequalities for the Lasso and Group Lasso in logistic model within the framework of high-dimensional statistics. We do not assume that f_0 is linear. We first state "slow" oracle inequalities (see Theorem 3.1 and Theorem 3.4) with no assumption on the Gram matrix, on the regressors nor on the margin. Secondly we provide "fast" oracle inequalities (see Theorem 3.2 and Theorem 4.1) under *restricted eigenvalue assumption* and some technical assumptions on the regressors. In each case, we give, as a consequence, the bounds for excess risk, $L_2(\frac{1}{n}\sum_{i=1}^{n}\delta_{z_i})$-norm and estimation errors for Lasso and Group Lasso in the *usual* logistic regression (*i.e.* when f_0 is linear). Our non asymp-

totic results lead to an adaptive data-driven weighting of the ℓ_1-norm (for the Lasso) and group norm (for the Group Lasso). Simulation study is given to illustrate the numerical performance of Group Lasso and Lasso with such weights.

This chapter is organized as follows. In Section 3.2, we describe our weighted Group Lasso estimation procedure and state non asymptotic oracle inequalities for the Group Lasso estimator. In Section 3.3 we describe our weighted Lasso estimation procedure and state non asymptotic oracle inequalities for the Lasso estimator. In Section 3.2.3 and Section 3.3.3 we give as a consequence the bounds for excess risk, $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ and estimation errors for Lasso and Group Lasso in the *usual* logistic regression (3.1) . Section 3.4 is devoted to simulation study. The proofs are gathered in Section 3.6 and Appendix.

Definitions and notations

Consider the matrix $X = \left(\phi_j(z_i)\right)_{1\le i\le n,\ 1\le j\le p}$ and $\{G_l, l = 1,\dots,g\}$ the partition of $\{1,\dots,p\}$. For any $\beta = (\beta_1,\dots,\beta_p)^T = (\beta^1,\dots,\beta^g)^T \in \mathbb{R}^p$, where $\beta^l = (\beta_j)_{j\in G_l}$ for $l = 1,\dots,g$. Let $f_\beta(.) = \sum_{j=1}^p \beta_j\phi_j(.) = \sum_{l=1}^g \sum_{j\in G_l} \beta_j\phi_j(.)$. With our notations

$$(f_\beta(z_1),\dots,f_\beta(z_n))^T = X\beta.$$

We define the group norm of β as

$$\|\beta\|_{2,q} = \left(\sum_{l=1}^g \left(\sum_{j\in G_l} \beta_j^2\right)^{\frac{q}{2}}\right)^{\frac{1}{q}} = \left(\sum_{l=1}^g \|\beta^l\|_2^q\right)^{\frac{1}{q}},$$

for every $1 \le q < \infty$. For $\beta \in \mathbb{R}^p$ $K(\beta) = \{j \in \{1,\dots,p\} : \beta_j \neq 0\}$ and $J(\beta) = \{l \in \{1,\dots,g\} : \beta^l \neq 0\}$, respectively the set of relevant coefficients (which characterizes the sparsity of the vector β) and the set of relevant groups. For all $\delta \in \mathbb{R}^p$ and a subset $I \subset \{1,\dots,p\}$, we denote by δ_I the vector in $\mathbb{R}^p$ that has the same coordinates as δ on I and zero coordinates on the complement I^c of I. Moreover $|I|$ denotes the cardinality of I. For all $h, f, g : \mathbb{R}^d \to R$, we define the scalar products

$$\langle f, h\rangle_n = \frac{1}{n}\sum_{i=1}^n h(z_i)f(z_i),$$

and

$$\langle f, h\rangle_g = \frac{1}{n}\sum_{i=1}^n h(z_i)f(z_i)\pi(g(z_i))(1-\pi(g(z_i))), \text{ where } \pi(t) = \frac{\exp(t)}{1+\exp(t)}.$$

We use the notation

$$q_f(h) = \frac{1}{n}\sum_{i=1}^n h(z_i)(Y_i - \pi(f(z_i))),$$

$\|h\|_\infty = \max_i |h(z_i)|$ and $\|h\|_n = \sqrt{\langle h, h\rangle_n} = \sqrt{\frac{1}{n}\sum_{i=1}^n h^2(z_i)}$ which denote the $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ norm (empirical norm). We consider empirical risk (logistic loss) for logistic model

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^n \log(1+\exp(f(z_i))) - Y_i f(z_i). \tag{3.3}$$

We denote by R the expectation of $\hat{R}$ with respect to the distribution of $Y_1, \ldots, Y_n$, *i.e*

$$R(f) = \boldsymbol{E}(\hat{R}(f)) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(f(z_i))) - \boldsymbol{E}(Y_i) f(z_i).$$

It is clear that $R(.)$ is a convex function and f_0 is a minimum of $R(.)$ when the model is well-specified (*i.e.* when (4.1) is satisfied). Note that with our notations

$$R(f) = \boldsymbol{E}(\hat{R}(f)) = \hat{R}(f) + q_{f_0}(f). \tag{3.4}$$

We shall use both the excess risk of $f_{\hat{\beta}}$, $R(f_{\hat{\beta}}) - R(f_0)$ and the prediction loss $\|f_{\hat{\beta}} - f_0\|_n^2$ to evaluate the quality of the estimator. Note that $R(f_{\hat{\beta}})$ corresponds to the average Kullback-Leibler divergence to the best model when the model is well-specified, and is common for the study of logistic regression.

3.2 Group Lasso for logistic regression model

3.2.1 Estimation procedure

The goal is not to estimate the parameters of the "true" model (since there is no true parameter) but rather to construct an estimator that mimics the performance of the best model in a given class, whether this model is true or not. Our aim is then to estimate f_0 in Model (4.1) by a linear combination of the functions of a dictionary

$$\mathbb{D} = \{\phi_1, \ldots, \phi_p\},$$

where $\phi_j : \mathbb{R}^d \to \mathbb{R}$ and p possibly $\gg n$. The functions ϕ_j can be viewed as estimators of f_0 constructed from independent training sample, or estimators computed using p different values of the tuning parameter of the same method. They can also be a collection of basis functions, that can approximate f_0, like wavelets, splines, kernels, etc... We implicitly assume that f_0 can be well approximated by a linear combination

$$f_\beta(.) = \sum_{j=1}^{p} \beta_j \phi_j(.),$$

where β has to be estimated.

In this section we assume that the set of relevant predictors have known group structure, for example in gene expression data these groups may be gene pathways, or factor level indicators in categorical data. And we wish to achieve sparsity at the level of groups. This group sparsity assumption suggests us to use the Group Lasso method. We consider the Group Lasso for logistic regression (see Meier et al. (2008), Yuan et Lin (2006)), where predictors are included or excluded in groups. The logistic Group Lasso is the minimizer of the following optimization problem

$$f_{\hat{\beta}_{GL}} := \underset{f_\beta \in \Gamma_1}{\operatorname{argmin}} \left\{ \hat{R}(f_\beta) + r \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 \right\}, \tag{3.5}$$

where

$$\Gamma_1 \subseteq \left\{ f_\beta(.) = \sum_{l=1}^{g} \sum_{j \in G_l} \beta_j \phi_j(.),\ \beta \in \mathbb{R}^p \right\}.$$

The tuning parameter $r > 0$ is used to adjust the trade-off between minimizing the loss and finding a solution which is sparse at the group level, i.e., to a vector

β such that $\beta^l = 0$ for some of the groups $l \in \{1, \ldots, g\}$. Sparsity is the consequence of the effect of non-differentiable penalty. This penalty can be viewed as an intermediate between ℓ_1 and ℓ_2 type penalty, which has the attractive property that it does variables selection at the group level. The weights $\omega_l > 0$, which we will define later, are used to control the amount of penalization per group.

3.2.2 Oracle inequalities

In this section we state non asymptotic oracle inequalities for excess risk and $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ loss of Group Lasso estimator. Consider the following assumptions :

There exists a constant $0 < c_1 < \infty$ such that $\max_{1\leq i\leq n} |f_0(z_i)| \leq c_1$. **($\mathbf{A_2}$)**

There exists a constant $0 < c_2 < \infty$ such that $\max_{1\leq i\leq n} \max_{1\leq j\leq p} |\phi_j(z_i)| \leq c_2$. **($\mathbf{A_3}$)**

For all $f_\beta \in \Gamma_1$, there is some universal constant C_0 such that $\max_{1\leq i\leq n} |f_\beta(z_i)| \leq C_0$. **($\mathbf{A_4}$)**

Assumptions ($\mathbf{A_5}$) and ($\mathbf{A_4}$) are technical assumptions useful to connect the excess risk and the $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ loss (see Lemma 4.4). An assumption similar to ($\mathbf{A_5}$) has been used in Bunea et al. (2007b) to prove oracle inequality in gaussian regression model. The same kind of assumption as ($\mathbf{A_4}$) has been made in Tarigan et van de Geer (2006) to prove oracle inequality for support vector machine type with ℓ_1 complexity regularization.

Theorem 3.1 *Let $f_{\hat{\beta}_{GL}}$ be the Group Lasso solution defined in (3.5) with $r \geq 1$ and*

$$\omega_l = \frac{2|G_l|}{n}\sqrt{\frac{1}{2}\max_{j\in G_l}\sum_{i=1}^n \phi_j^2(z_i)\,(x+\log p)} + \frac{2c_2|G_l|}{3n}\,(x+\log p)\,, \tag{3.6}$$

where $x > 0$. Under Assumption ($\mathbf{A_3}$), with probability at least $1 - 2\exp(-x)$ we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq \inf_{\beta\in\mathbb{R}^p}\left\{R(f_\beta) - R(f_0) + 2r\|\beta\|_{2,1}\max_{1\leq l\leq g}\omega_l\right\}. \tag{3.7}$$

The first part of the right hand of Inequality (3.7) corresponds to the approximation error (bias). The selection of the dictionary can be very important to minimize this approximation error. It is recommended to choose a dictionary $\mathbb{D}$ such that f_0 could well be approximated by a linear combination of the functions of $\mathbb{D}$. The second part of the right hand of Inequality (3.7) is the variance term and is usually referred as the rate of the oracle inequality. In Theorem 3.1, we speak about "slow" oracle inequality, with the rate at the order $\|\beta\|_{2,1}\sqrt{\log p/n}$ for any β. Moreover this is a sharp oracle inequality in the sense that there is a constant 1 in front of term $\inf_{\beta\in\mathbb{R}^p}\{R(f_\beta) - R(f_0)\}$. This result is obtained without any assumption on the Gram matrix ($\Phi_n = X^TX/n$). In order to obtain oracle inequality with a "fast rate" of order $\log p/n$ we need additional assumption on the restricted eigenvalue of the Gram matrix, namely the *restricted eigenvalue assumption*.

For some integer s such that $1 \leq s \leq g$ and a positive number a_0, **($\mathbf{RE_5}$)**
the following condition holds

$$\mu_1(s, a_0) := \min_{K\subseteq\{1,\ldots p\}:|K|\leq s}\ \min_{\Delta\neq 0:\|\Delta_{K^c}\|_{2,1}\leq a_0\|\Delta_K\|_{2,1}} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_K\|_2} > 0.$$

This is a natural extension to the Group Lasso of *restricted eigenvalue assumption* introduced in Bickel et al. (2009) (or Assumption ($\mathbf{RE_7}$) used below) for the usual Lasso. The only difference lies on the set where the minimum is taken : for the Lasso the minimum is taken over $\{\Delta \neq 0 : \|\Delta_{K^c}\|_1 \leq a_0\|\Delta_K\|_1\}$ whereas for the Group Lasso the minimum is over $\{\Delta \neq 0 : \|\Delta_{K^c}\|_{2,1} \leq a_0\|\Delta_K\|_{2,1}\}$. This assumption has already been used in Lounici et al. (2009; 2011) to prove oracle inequality for linear gaussian noise model under Group sparsity and for multi-task learning. To emphasize the dependency of Assumption ($\mathbf{RE_5}$) on s and a_0 we will sometimes refer to it as $RE(s, a_0)$.

Theorem 3.2 *Let $f_{\hat{\beta}_{GL}}$ be the Group Lasso solution defined in (3.5) with ω_l defined as in (3.6). Fix $\eta > 0$ and $1 \leq s \leq g$, assume that ($\mathbf{A_5}$), ($\mathbf{A_3}$), ($\mathbf{A_4}$) and ($\mathbf{RE_5}$) are satisfied, with $a_0 = 3 + 4/\eta$. Thus with probability at least $1 - 2\exp(-x)$ we have*

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta) \inf_{f_\beta \in \Gamma_1} \left\{ R(f_\beta) - R(f_0) + \frac{c(\eta)|J(\beta)|r^2 \left(\max_{1\leq l\leq g} \omega_l\right)^2}{c_0 \epsilon_0 \mu_1(s, a_0)^2} \right\}, \tag{3.8}$$

and

$$\|f_{\hat{\beta}_{GL}} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta) \inf_{f_\beta \in \Gamma_1} \left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)|J(\beta)|r^2 \left(\max_{1\leq l\leq g} \omega_l\right)^2}{c_0' c_0 \epsilon_0^2 \mu_1(s, a)^2} \right\}. \tag{3.9}$$

Where $c(\eta)$ is a constant depending only on η ; $c_0 = c_0(C_0, c_1)$ and $c_0' = c_0'(C_0, c_1)$ are constants depending on C_0 and c_1 ; $\epsilon_0 = \epsilon_0(c_1)$ is a constant depending on c_1 ; and $r \geq 1$.

In Theorem 3.2, the variance terms are of order $\log p/n$. Hence we say that the corresponding non asymptotic oracle inequalities have "fast rates". For the best of our knowledge, Inequalities (3.7), (3.8) and (3.9) are the first non asymptotic oracle inequalities for the Group Lasso in logistic regression model. These inequalities allow us to bound the prediction errors of Group Lasso by the best sparse approximation and a variance term. The major difference with existing results concerning Group Lasso for logistic regression model (see Negahban et al. (2012), Meier et al. (2008), Blazère et al. (2014)) is that f_0 is not necessarily linear. And we also demonstrated a sharp non asymptotic oracle inequality without any assumption on the Gram matrix (see Theorem 3.1).

Remark 3.1 *Our results remain true if we assume that we are in the "neighborhood" of the target function. If we suppose that there exists ζ such that $\max_{1\leq i\leq n} |f_\beta(z_i) - f_0(z_i)| \leq \zeta$, then Lemma 4.4 is still true.*

Remark 3.2 *The choice of the weights ω_ℓ comes from Bernstein's inequality. We could also use the following weights*

$$\omega_l' = \frac{2|G_l|}{n}\sqrt{2\max_{j\in G_l}\sum_{i=1}^n \mathbb{E}[\phi_j^2(z_i)\epsilon_i^2]\,(x+\log p)} + \frac{2|G_l|\max_{1\leq i\leq n}\max_{j\in G_l}|\phi_j(z_i)|}{3n}(x+\log p),$$

with $\epsilon_i = Y_i - \mathbb{E}[Y_i]$. Theorems 3.1 and 3.2 still hold true with such weights ω_l'. But these weights depend on the unknown function f_0 to be estimated through $\mathbb{E}(\epsilon_i^2) =$

$\pi(f_0(z_i))(1-\pi(f_0(z_i)))$. This is the reason for using weights ω_l slightly greater than ω'_l. We will show in simulation study (Section 3.4) how to use the weights ω'_l to improve the Group Lasso defined in Meier et al. (2008) which used $\sqrt{|G_l|}$ as weight for the group l.

3.2.3 Special case : f_0 linear

In this section we assume that f_0 is a linear function *i.e.* $f_0(z_i) = f_{\beta_0}(z_i) = \sum_{l=1}^{g}\sum_{j\in G_l}\beta_j z_{ij}$. Denote by $X = (z_{ij})_{1\le i\le n, 1\le j\le p}$, the design matrix. Let $z_i = (z_{i1},\dots,z_{ip})^T$ be the ith row of the matrix X and $z^{(j)} = (z_{1j},\dots,z_{nj})^T$ is jth column. For $i = 1,\dots,n$

$$\mathbb{P}(Y_i = 1) = \frac{\exp(z_i^T\beta_0)}{1+\exp(z_i^T\beta_0)}. \tag{3.10}$$

This corresponds to the *usual* logistic regression (3.1) *i.e.* logistic model that allows linear dependency between z_i and the distribution of Y_i. In this context, the Group Lasso estimator of β_0 is defined by

$$\hat{\beta}_{GL} := \underset{\beta:\, f_\beta\in\Gamma_1}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\Big\{\log(1+\exp(z_i^T\beta)) - Y_i z_i^T\beta\Big\} + r\sum_{l=1}^{g}\omega_l\|\beta^l\|_2. \tag{3.11}$$

Corollary 3.1 *Let assumption $\mathbf{RE_5}(s,3)$ be satisfied and $|J(\beta_0)|\le s$, where $1\le s\le g$. Consider the Group Lasso estimator $f_{\hat{\beta}_{GL}}$ defined by (3.11) with*

$$\omega_l = \frac{2|G_l|}{n}\sqrt{\frac{1}{2}\max_{j\in G_l}\sum_{i=1}^{n} z_{ij}^2\,(x+\log p)} + \frac{2c_2|G_l|}{3n}\,(x+\log p) \tag{3.12}$$

where $x > 0$. Under the assumptions of Theorem 3.2, with probability at least $1-2\exp(-x)$ we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \le \frac{9sr^2\left(\max\limits_{1\le l\le g}\omega_l\right)^2}{\mu^2(s,3)c_0\epsilon_0} \tag{3.13}$$

$$\|f_{\hat{\beta}_{GL}} - f_{\beta_0}\|_n^2 \le \frac{9sr^2\left(\max\limits_{1\le l\le g}\omega_l\right)^2}{\mu^2(s,3)c_0^2\epsilon_0^2} \tag{3.14}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,1} \le \frac{12rs\left(\max\limits_{1\le l\le g}\omega_l\right)^2}{\mu^2(s,3)c_0\epsilon_0(\min\limits_{1\le l\le g}\omega_l)} \tag{3.15}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,q}^q \le \left(\frac{12rs\left(\max\limits_{1\le l\le g}\omega_l\right)^2}{\mu^2(s,3)c_0\epsilon_0(\min\limits_{1\le l\le g}\omega_l)}\right)^q \quad \text{for all } 1 < q \le 2. \tag{3.16}$$

Remark 3.3 *In logistic regression model (3.27), if vector β_0 is sparse, i.e. $|J(\beta_0)|\le s$, then Assumption ($\mathbf{RE_5}$) implies that β_0 is uniquely defined. Indeed, if there exists β^* such that for $i = 1,\dots,n$, $\pi(z_i^T\beta_0) = \pi(z_i^T\beta^*)$, it follows that $X\beta_0 = X\beta^*$ and $|J(\beta^*)|\le s$. Then according to assumption $RE(s,a_0)$ with $a_0 \geqslant 1$, we necessarily have $\beta_0 = \beta^*$. Indeed if $RE(s,a_0)$ is satisfied with $a_0\geqslant 1$, then $\min\{\|X\beta\|_2 : |J(\beta)|\le 2s, \beta\neq 0\} > 0$.*

Remark 3.4 *(**Theoretical advantage of Group Lasso over the Lasso**) Concerning results on oracle inequality for the Group Lasso few results exist. The first oracle inequality for the Group Lasso in the additive regression model is due to Nardi et Rinaldo (2008). Since then, some of these inequalities have been improved in Lounici et al. (2011) Lounici et al. (2011), concerning in particular the gain on order rate. More precisely, Lounici et al. (2011) Lounici et al. (2011) have found a rate of order* $\log g/n$ *for Group Lasso in gaussian linear model, which is better than is corresponding rate for the Lasso,* $\log p/n$ *(since* $g \leq p$*). This improvement seems mainly based on the assumption that the noise is gaussian. In our case (see proof of Theorem 3.1, formula (3.35)) the empirical process involves non gaussian variables and thus their method should not apply in our context. However the probability that their results are true depends on* g *whereas the probability that our results hold does not depend on* g*.*

We can find the rate of order $\log g/n$ *by choosing this constant* x *in the weights in a certain manner. Indeed, let us assume (without loss of generality) that the groups are all of equal size* $|G_1| = \cdots = |G_g| = m$*, so that* $p = m.g$*. Since the weights in (3.6) are defined for all* $x > 0$*, if we take* $x = q \log g - \log m > 0$ *where* q *is a positive constant such that* $g^q > m$*. Then the weights in (3.6) become*

$$\omega_l = \frac{2|G_l|}{n}\sqrt{\frac{1}{2}\max_{j\in G_l}\sum_{i=1}^{n}\phi_j^2(z_i)\left[(1+q)\log g\right]} + \frac{2c_2|G_l|}{3n}\left[(1+q)\log g\right],$$

thus

$$\omega_l^2 \sim \frac{\log g}{n},$$

and the results in Theorem 3.1 and Theorem 3.2 hold with probability at least

$$1 - 2\frac{m}{g^q}.$$

In the special case where the $g > 2m$ *these results are true for all* $q > 0$*.*

3.2.4 Non bounded functions

The results of Corollary 3.1 are obtained (as the consequence of Theorem 3.2) with the assumptions that f_{β_0} and all $f_\beta \in \Gamma_1$ are bounded. In some situations these assumptions could not be verified. In this section we will establish the same results without assuming $(\mathbf{A_5})$ or $(\mathbf{A_4})$ *i.e.* neither f_{β_0} nor f_β is bounded. We consider the Group Lasso estimator defined in (3.11) and the following assumption :

For some integer s such that $1 \leq s \leq g$ and a positive number a_0, $(\mathbf{RE_6})$
the following condition holds

$$\mu_2(s, a_0) := \min_{K\subseteq\{1,\dots p\}:|K|\leq s}\ \min_{\Delta\neq 0:\|\Delta_{K^c}\|_{2,1}\leq a_0\|\Delta_K\|_{2,1}} \frac{\Delta^T X^T D X \Delta}{n\|\Delta_K\|_2^2} > 0,$$

where $D = \mathrm{Diag}\left(\mathrm{var}(Y_i)\right)$.

This is an extension of the Assumption $\mathbf{RE_5}$ to the weighted Gram matrix $X^T DX/n$.

Theorem 3.3 *Consider the Group Lasso estimator* $f_{\hat\beta_{GL}}$ *defined by (3.11) with* w_l *defined as in (3.12) where* $x > 0$*. Set* $v = \max_{1\leq i\leq n}\max_{1\leq l\leq g}\|z_i^l\|_2$*. Let Assumptions (*$\mathbf{A_3}$*) and (*$\mathbf{RE_6}$*) be satisfied with*

$$a_0 = \frac{3\max_{1\leq l\leq g}\omega_l}{\min_{1\leq l\leq g}\omega_l}.$$

If $r(1+a_0)^2 \max_{1\le l\le g} \omega_l \le \frac{\mu_2^2}{3v|J|}$, *with probability at least* $1-2\exp(-x)$ *we have*

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \le \frac{9(1+a_0)^2 J(\beta_0)|r^2 \left(\max_{1\le l\le g} \omega_l\right)^2}{\mu_2^2(s,3)} \tag{3.17}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,1} \le \frac{6(1+a_0)^2|J(\beta_0)|r \left(\max_{1\le l\le g} \omega_l\right)}{\mu_2^2(s,3)} \tag{3.18}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,q}^q \le \left(\frac{6(1+a_0)^2|J(\beta_0)|r \left(\max_{1\le l\le g} \omega_l\right)}{\mu_2^2(s,3)}\right)^q \quad \textit{for all } \ 1 < q \le 2. \tag{3.19}$$

Moreover if we assume that there exists $0 < \epsilon_0 \le 1/2$ *such that*

$$\epsilon_0 \le \pi(f_{\beta_0}(z_i))[1-\pi(f_{\beta_0}(z_i))] \quad \textit{for all } \ i = 1,\dots,n$$

then,

$$\|X\hat{\beta}_{GL} - X\beta_0\|_n^2 \le \frac{36(1+a_0)^2|J(\beta_0)|r^2 \left(\max_{1\le l\le g} \omega_l\right)^2}{\mu^2(s,3)\epsilon_0}. \tag{3.20}$$

Inequalities (3.18) and (3.19) are the extensions of the results in Bach (2010) for the Lasso to Group Lasso in logistic regression model.

In this section we studied some properties of the Group Lasso. However the Group Lasso is based on prior knowledge that the set of relevant predictors have known group structure. If this group sparsity condition is not satisfied, the sparsity can be achieve by simply using the Lasso. We will show in the next section how to adapt the results of this section to the Lasso.

3.3 Lasso for logistic regression

3.3.1 Estimation procedure

The Lasso estimator $f_{\hat{\beta}_L}$ is defined as a minimizer of the following ℓ_1-penalized empirical risk

$$f_{\hat{\beta}_L} := \underset{f_\beta \in \Gamma}{\operatorname{argmin}} \left\{ \hat{R}(f_\beta) + r\sum_{j=1}^{p} \omega_j|\beta_j| \right\}, \tag{3.21}$$

where the minimum is taken over the set

$$\Gamma \subseteq \left\{ f_\beta(.) = \sum_{j=1}^{p} \beta_j \phi_j(.),\ \beta = (\beta_1,\dots,\beta_p) \in \mathbb{R}^p \right\}$$

and ω_j are positive weights to be specified later. The "classical" Lasso penalization corresponds to $\omega_j = 1$, where r is the tuning parameter which makes balance between goodness-of-fit and sparsity. The Lasso estimator has the property that it does predictors selection and estimation at the same time. Indeed for large values of ω_j, the related components $\hat{\beta}_j$ are set exactly to 0 and the other are shrunken toward zero.

3.3.2 Oracle inequalities

In this section we provide non asymptotic oracle inequalities for the Lasso in logistic regression model.

Theorem 3.4 *Let $f_{\hat{\beta}_L}$ be the ℓ_1-penalized minimum defined in (3.21). Let Assumption ($\mathbf{A_3}$) be satisfied.*

A-) Let $x > 0$ be fixed and $r \geq 1$. For $j = \{1, \dots, p\}$, let

$$\omega_j = \frac{2}{n}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\phi_j^2(z_i)(x + \log p)} + \frac{2c_2(x + \log p)}{3n}. \tag{3.22}$$

Thus with probability at least $1 - 2\exp(-x)$ we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq \inf_{\beta \in \mathbb{R}^p}\left\{R(f_\beta) - R(f_0) + 2\|\beta\|_1 r \max_{1\leq j\leq p}\omega_j\right\}.$$

B-) Let $A > 2\sqrt{c_2}$. For $j = \{1, \dots, p\}$, let $\omega_j = 1$, and

$$r = A\sqrt{\frac{\log p}{n}}.$$

Thus with probability at least $1 - 2p^{1-A^2/4c_2}$ we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq \inf_{\beta \in \mathbb{R}^p}\left\{R(f_\beta) - R(f_0) + 2A\|\beta\|_1\sqrt{\frac{\log p}{n}}\right\}.$$

As previously, the variance terms are of order $\|\beta\|_1\sqrt{\log p/n}$ for any β. Hence these are sharp oracle inequalities with "slow" rates. These results are obtained without any assumption on the Gram matrix. To obtain oracle inequalities with a "fast rate", of order $\log p/n$, we need the restricted eigenvalue condition.

For some integer s such that $1 \leq s \leq p$ and a positive number a_0, the following condition holds ($\mathbf{RE_7}$)

$$\mu(s, a_0) := \min_{K\subseteq\{1,\dots p\}:|K|\leq s}\ \min_{\Delta\neq 0:\|\Delta_{K^c}\|_1\leq a_0\|\Delta_K\|_1}\frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_K\|_2} > 0.$$

This assumption has been introduced in Bickel et al. (2009), where several sufficient conditions for this assumption are described. This condition is known to be one of the weakest to derive "fast rates" for the Lasso. For instance conditions on the Gram matrix used to prove oracle inequality in Bunea et al. (2006; 2007b;a) are more restrictive than *restricted eigenvalue assumption*. In those papers either Φ_n is positive definite, or mutual coherence condition is imposed. We refer to van de Geer et Bühlmann (2009) for a complete comparison of the assumptions used to prove oracle inequality for the Lasso. Especially it is proved that *restricted eigenvalue assumption* is weaker than the neighborhood stability or irrepresentable condition.

Theorem 3.5 *Let $f_{\hat{\beta}_L}$ be the ℓ_1-penalized minimum defined in (3.21). Fix $\eta > 0$ and $1 \leq s \leq p$. Assume that ($\mathbf{A_5}$), ($\mathbf{A_3}$), ($\mathbf{A_4}$) and ($\mathbf{RE_7}$) are satisfied, with $a_0 = 3 + 4/\eta$.*

A-) Let $x > 0$ be fixed and $r \geq 1$. For $j = \{1, \dots, p\}$, ω_j defined as in (3.22). Thus with probability at least $1 - 2\exp(-x)$ we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta) \inf_{f_\beta \in \Gamma} \left\{ R(f_\beta) - R(f_0) + \frac{c(\eta)|K(\beta)|r^2 \left(\max_{1 \leq j \leq p} \omega_j \right)^2}{c_0 \epsilon_0 \mu^2(s, 3 + 4/\eta)} \right\}, \tag{3.23}$$

and

$$\|f_{\hat{\beta}_L} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta) \inf_{f_\beta \in \Gamma} \left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)|K(\beta)|r^2 \left(\max_{1 \leq j \leq p} \omega_j \right)^2}{c_0' c_0 \epsilon_0^2 \mu^2(s, 3 + 4/\eta)} \right\}. \tag{3.24}$$

B-) Let $A > 2\sqrt{c_2}$. For $j = \{1, \dots, p\}$, let $\omega_j = 1$, and

$$r = A\sqrt{\frac{\log p}{n}}.$$

Thus with probability at least $1 - 2p^{1 - A^2/4c_2}$ we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta) \inf_{f_\beta \in \Gamma} \left\{ R(f_\beta) - R(f_0) + \frac{A^2 c(\eta)}{c_0 \epsilon_0 \mu^2(s, 3 + 4/\eta)} \frac{|K(\beta)| r^2 \log p}{n} \right\}, \tag{3.25}$$

and

$$\|f_{\hat{\beta}_L} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta) \inf_{f_\beta \in \Gamma} \left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)A^2}{c_0' c_0 \epsilon_0^2 \mu^2(s, 3 + 4/\eta)} \frac{|K(\beta)| r^2 \log p}{n} \right\}. \tag{3.26}$$

In both cases $c(\eta)$ is a constant depending only on η; $c_0 = c_0(C_0, c_1)$ and $c_0' = c_0'(C_0, c_1)$ are constants depending on C_0 and c_1; and $\epsilon_0 = \epsilon_0(c_1)$ is a constant depending on c_1.

In this theorem the variance terms are of order $|K(\beta)| \log p / n$. Such order in sparse oracle inequalities usually refer to "fast rate". This rate is of same kind of the one obtain in Bickel et al. (2009) for linear regression model. For the best of our knowledge, (3.24) and (3.26) are the first non asymptotic oracle inequalities for the $L_2(\frac{1}{n}\sum_i^n \delta_{z_i})$ norm in logistic model. Some non asymptotic oracle inequalities for excess risk like (3.23) or (3.25) have been established in van de Geer (2008) under different assumptions. Indeed, she stated oracle inequality for high dimensional generalized linear model with Lipschitz loss function, where logistic regression is a particular case. Her result assumes to be hold in the "neighborhood" of the target function, while our result is true for all bounded functions. Note also that our results hold under *RE* condition, which can be seen as empirical version of Assumption C in van de Geer (2008). The confidence (probability that result holds true) of Inequality (3.23) does not depend on n or p while the confidence of her results depends on n and p. Moreover, the weights we proposed from Bernstein's inequality are different and exhibit better performance, at least in the specific cases studied in the simulation part (see Section 3.4).

3.3.3 Special case : f_0 linear

In this section we assume that f_0 is a linear function that is $f_0(z_i) = f_{\beta_0}(z_i) = \sum_{j=1}^{p} \beta_{0j} z_{ij} = z_i^T \beta_0$, where $z_i = (z_{i1}, \dots, z_{ip})^T$. Denote $X = (z_{ij})_{1 \le i \le n, 1 \le j \le p}$ the design matrix. Thus for $i = 1, \dots, n$

$$\mathbb{P}(Y_i = 1) = \pi(z_i^T \beta_0) = \frac{\exp(z_i^T \beta_0)}{1 + \exp(z_i^T \beta_0)}. \tag{3.27}$$

The Lasso estimator of β_0 is thus defined as

$$\hat{\beta}_L := \underset{\beta:\ f_\beta \in \Gamma}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1 + \exp(z_i^T \beta)) - Y_i z_i^T \beta \right\} + r \sum_{j=1}^{p} \omega_j |\beta_j| \right\}. \tag{3.28}$$

When the design matrix X has full rank, the solution of optimization Problem (3.28) is usually unique. When $p \gg n$ this infimum might not be unique.

Corollary 3.2 *Let assumption RE(s,3) be satisfied and $|K(\beta_0)| \le s$, where $1 \le s \le p$. Consider the Lasso estimator $f_{\hat{\beta}_L}$ defined by (3.28) with*

$$\omega_j = \frac{2}{n} \sqrt{\frac{1}{2} \sum_{i=1}^{n} z_{ij}^2 (x + \log p)} + \frac{2c_2(x + \log p)}{3n}$$

Under the assumptions of Theorem 4.1 with probability at least $1 - \exp(-x)$ we have

$$R(f_{\hat{\beta}_L}) - R(f_{\beta_0}) \le \frac{9sr^2 \left(\max_{1 \le j \le p} \omega_j \right)^2}{\mu^2(s,3) c_0 \epsilon_0} \tag{3.29}$$

$$\|f_{\hat{\beta}_L} - f_{\beta_0}\|_n^2 \le \frac{9s^2 r^2 \left(\max_{1 \le j \le p} \omega_j \right)^2}{\mu^2(s,3) c_0^2 \epsilon_0^2} \tag{3.30}$$

$$\|\hat{\beta}_L - \beta_0\|_1 \le \frac{12sr \left(\max_{1 \le j \le p} \omega_j \right)^2}{\mu^2(s,3) c_0 \epsilon_0 \left(\min_{1 \le j \le p} \omega_j \right)} \tag{3.31}$$

$$\|\hat{\beta}_L - \beta_0\|_q^q \le \left(\frac{12sr \left(\max_{1 \le j \le p} \omega_j \right)^2}{\mu^2(s,3) c_0 \epsilon_0 \left(\min_{1 \le j \le p} \omega_j \right)} \right)^q \quad \text{for all } 1 < q \le 2. \tag{3.32}$$

If $r = A\sqrt{\log p / n}$ and $\omega_j = 1$ for all $j \in \{1, \dots, p\}$ we have the same results with probability at least $1 - 2p^{1 - A^2/4c_2}$.

Line (3.29) and Line (3.31) of the corollary are similar to those of Theorem 5 in Bach (2010). Note that, up to differences in constant factors, the rates obtained in this corollary are the same as those obtained in Theorem 7.2 in Bickel et al. (2009) for linear model with an s-sparse vector. Remark 3.3 remains true in this section.

3.4 Simulation study

To illustrate the theoretical part of this chapter we provide in this section some experimental results on simulated data. Our aim is to compare the Group Lasso using the weights we proposed to the Group Lasso proposed in Meier et al. (2008). Recall that Group Lasso for logistic regression proposed in Meier et al. (2008) used $\sqrt{|G_l|}$ as weight for the group l. We consider the Group Lasso defined in (3.11), with the weights defined in (3.12), which we denote by *weight.GL*. We also consider the Group Lasso estimator defined in (3.11) with weights

$$\omega_l' = \frac{2|G_l|}{n}\sqrt{2\max_{j\in G_l}\sum_{i=1}^{n}\mathbb{E}[z_{ij}^2\epsilon_i^2]\,(2+\log p)} + \frac{2|G_l|\max\limits_{1\le i\le n}\max\limits_{j\in G_l}|z_{ij}|}{3n}\,(2+\log p)\,,$$

which we denote by *weight.theoretical.GL*. Note that these are the exact weights, and as mentioned in Remark 3.2, all our results remain true with these weights. But the only drawback is that, these weights depend on the unknown β_0 (the parameter to be estimated) through $\mathbb{E}[\epsilon_i^2] = \pi(z_i^T\beta_0)[1-\pi(z_i^T\beta_0)]$. Later, we will show how to estimate $\mathbb{E}[\epsilon_i^2]$ in order to estimate ω_l.

3.4.1 Data generation

► *Data generation for the Group Lasso.* We simulated our covariate matrix X with different numbers of covariates, observations and groups. The columns of X were independent and identically distributed (i.i.d.) gaussian, and the response y was constructed from logistic model (3.10) with $\beta_0^1 = (1,\dots,1)$, $\beta_0^2 = (-1.5,\dots,-1.5)$, $\beta_0^3 = (2,\dots,2)$ and $\beta_0^l = (0,\dots,0)$ for $l \notin \{1,2,3\}$. This corresponds to the model with $|J(\beta_0)|$=3. We consider different values of $|J^c(\beta_0)|$ to change the amount of sparsity. Denote by $nk = |G_l|, l \in \{1,2,3\}$ and $nkc = |G_l|, l \notin \{1,2,3\}$. For each data set we calculate the prediction error, $\|f_{\hat\beta_{GL}} - f_{\beta_0}\|_n^2$; estimation error, $\|\hat\beta_{GL} - \beta_0\|_2$. We also calculate the rate of *true selection* ; and the rate of *false relevant and irrelevant coefficients*. *True selection* corresponds to the situation where the procedure selects exactly the true relevant coefficients. The rate of *false relevant and irrelevant coefficients* is the rate of bad selection in an estimation (the procedure declares that a coefficient is relevant yet it is irrelevant or declares irrelevant yet it is relevant).
► *Data generation for the Lasso.* We simulated 500 datasets consisting of n observations from logistic model (3.27), with $\beta_0 = (1.5,-1,2,0,0,\dots,0) \in \mathbb{R}^p$ where $|K(\beta_0)| = 3$ and $p = 3+k \in \{200,500,1000\}$. The columns of X were i.i.d. gaussian. We first consider the Lasso with the weights ω_j' which we denote *weight.theoretical*. As we can not compute these weights in practice, we propose to estimate them as follows. Since the only unknown term in ω_j' is $\boldsymbol{E}(\epsilon_i^2)$, we propose two estimators of $\boldsymbol{E}(\epsilon_i^2) = \pi(z_i^T\beta_0)(1-\pi(z_i^T\beta_0)$:

1. estimate by $\hat\sigma_i^2 = \pi(z_i^T\hat\beta_L)(1-\pi(z_i^T\hat\beta_L)$ where $\hat\beta_L$ is the "classical" Lasso estimator of β_0 (without weight) ;
2. the second estimator is $\hat\sigma_i^2 = \pi(z_i^T\hat\beta_{Logit})(1-\pi(z_i^T\hat\beta_{Logit})$, where $\hat\beta_{Logit}$ is an estimator of β_0 obtained after successively using the Lasso to screen coefficients and a logistic model which take into account coefficients different to zero in the Lasso.

The results for the four methods are presented in the Figure 3.4, Figure 3.5 and Figure 3.6. Lasso represents the "classical" Lasso (without weight) ; *weight.Logit*is

the Lasso with weights estimated using procedure (2); *weight.Lasso* is the Lasso with weights estimated by the procedure (1); *weight.theoretical* is the Lasso with theoretical weights. For all the methods, r will be estimated by cross-validation.

3.4.2 Comments

Referring to Figure 3.1, Figure 3.2 and Figure 3.3, we see that the performance of all methods increases until some optimal performance, and then decreases. This means that when we reach the optimal model, which corresponds to the model with r_{opt}-value, nothing is gained by adding other variables. Moreover it is important to note that the optimal value (r_{opt}) in prediction or estimation is different to the optimal value in selection. In prediction, estimation or selection, the Group Lasso using our weights outperforms the Group Lasso defined in Meier et al. (2008). According to Figure 3.1 for instance, *weight.theoretical.GL* reaches 99% of true selection rate, while *weight.GL* peaks at 97% and the Group Lasso comes in last with 66% of true selection.

According to Figure 3.4, Figure 3.5 and Figure 3.6 we can see that for estimation or prediction error the performance of all the methods are almost the same. When the number of sample n increases, the performance of all the methods also increases. The strength of the methods decreases with the number k of null coefficients. The real difference is in rate of *true selection* and the rate of *false relevant and irrelevant coefficients*, where the *weight.theoretica*l, *weight.Logit* and *weight.Lasso* outperform the Lasso. And *weight.Logit* seems to be better than *weight.Lasso*.

3.5 Conclusion

In this chapter we stated non asymptotic oracle inequalities for the Lasso and Group Lasso. Our results are non asymptotic : the number n of observations is fixed while the number p of covariates can grow with respect to n and can be much larger than n. The major difference with existing results concerning Group Lasso or Lasso for logistic regression model is that we do not assume that f_0 is linear. First we provided sharp oracle inequalities for excess risk, with "slow" rates, with no assumption on the Gram matrix, on the regressors nor on the margin. Secondly, under RE condition we provided "fast" oracle inequalities for excess risk and $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ loss. We also provided as a consequence of oracle inequalities the bounds for excess risk, $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ error and estimation error in the case where the true function f_0 is linear (*usual* logistic regression (3.1)). We shown in simulation study that the weighted versions of Lasso and Group Lasso we proposed exhibit better properties than the canonical Lasso and Group Lasso.

3.6 Proofs of main results

Proof of Theorem 3.1

Since $\hat{\beta}_{GL}$ is the minimizer of $\hat{R}(f_\beta) + r\sum_{l=1}^g \omega_l \|\beta^l\|_2$, we get

$$R(f_{\hat{\beta}_{GL}}) - \frac{1}{n}\varepsilon^T X\hat{\beta}_{GL} + r\sum_{l=1}^g \omega_l \|\hat{\beta}^l_{GL}\|_2 \leq R(f_\beta) - \frac{1}{n}\varepsilon^T X\beta + r\sum_{l=1}^g \omega_l \|\beta^l\|_2.$$

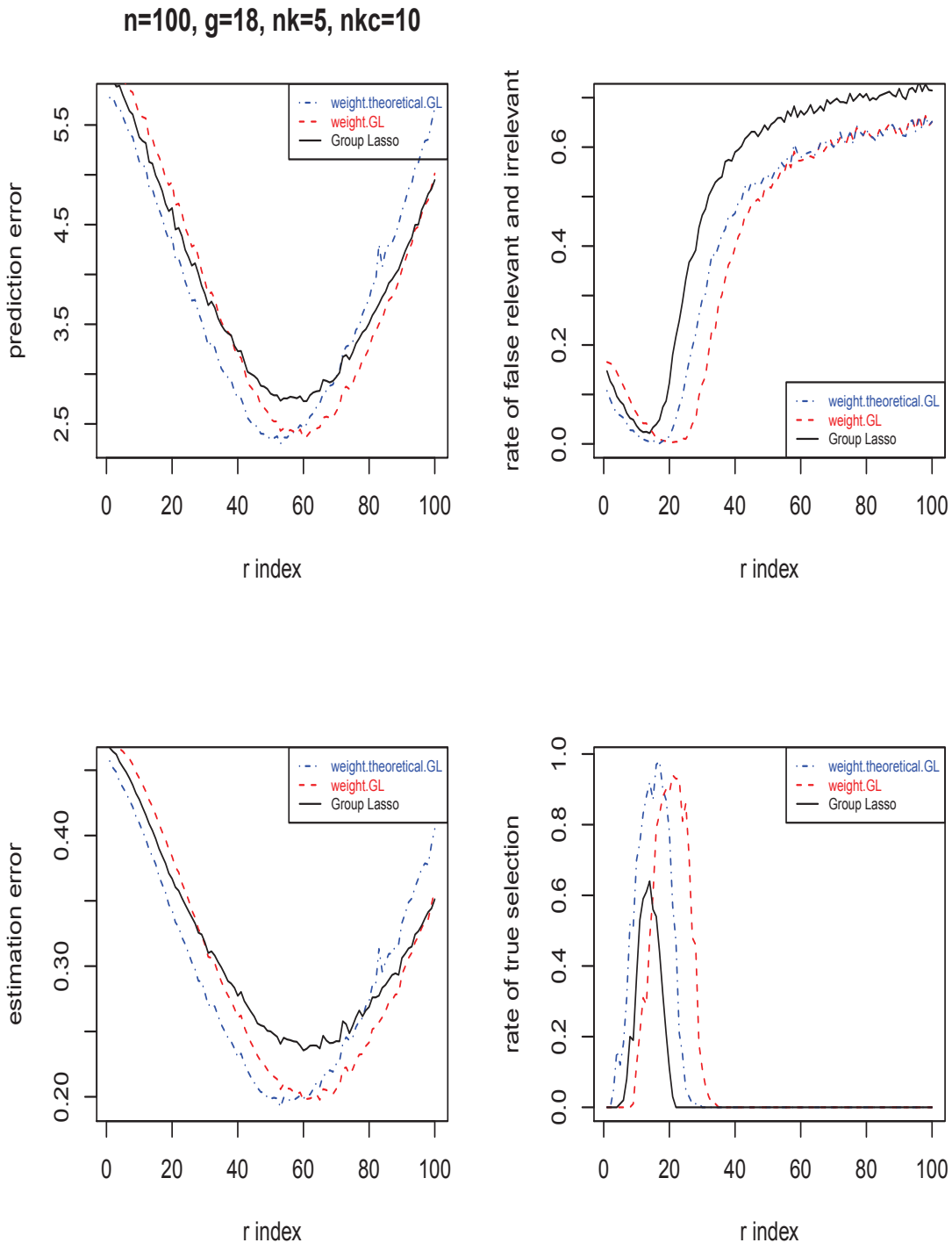


FIGURE 3.1 – *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients (see Section 3.4). All methods were fit from a path of 100 tuning parameters r from r_{max} to r_{min}. Each point corresponds to the average after 100 simulations from the setup described in Section 3.4.*

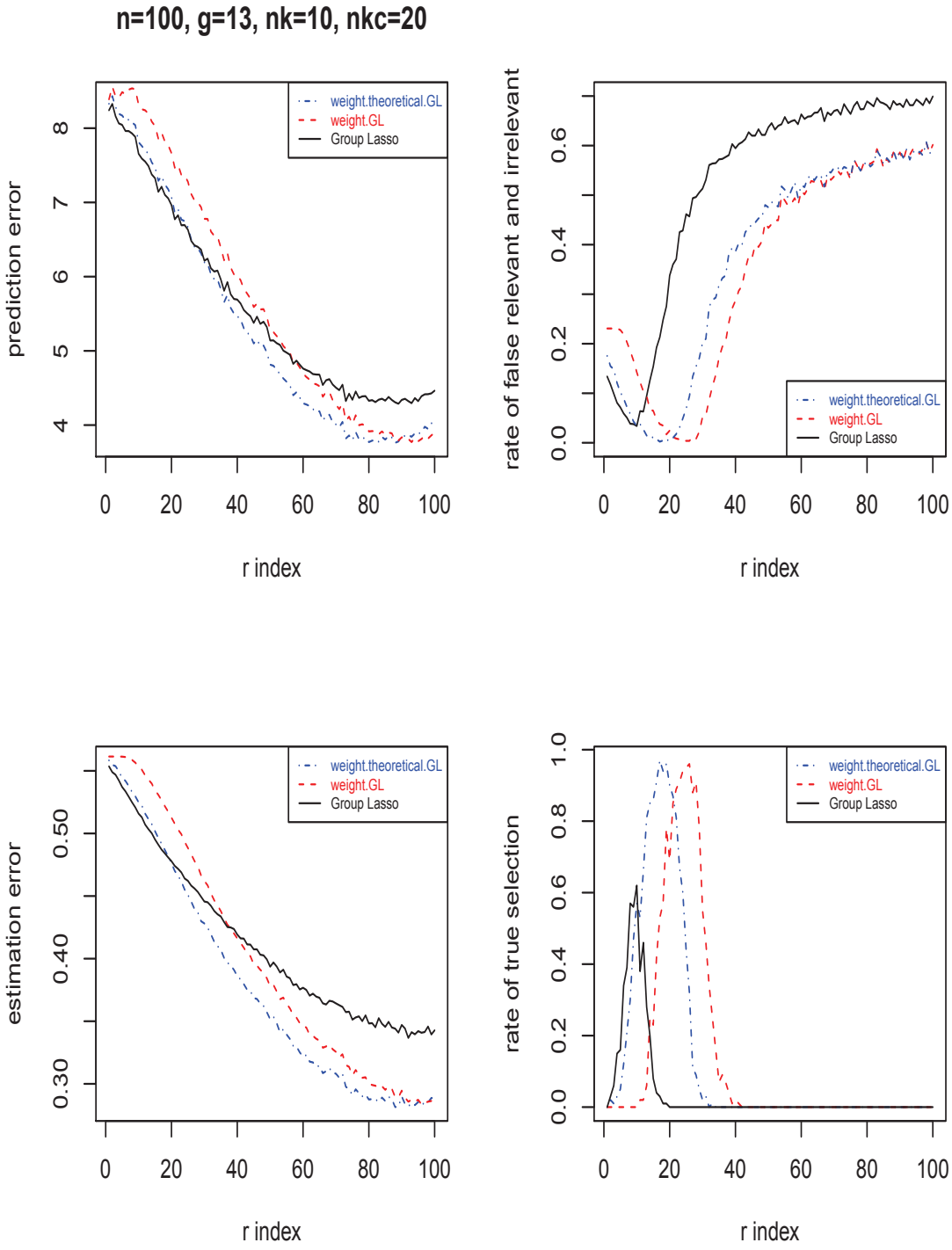


FIGURE 3.2 – *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients (see Section 3.4). All methods were fit from a path of 100 tuning parameters r from r_{max} to r_{min}. Each point corresponds to the average after 100 simulations from the setup described in Section 3.4.*

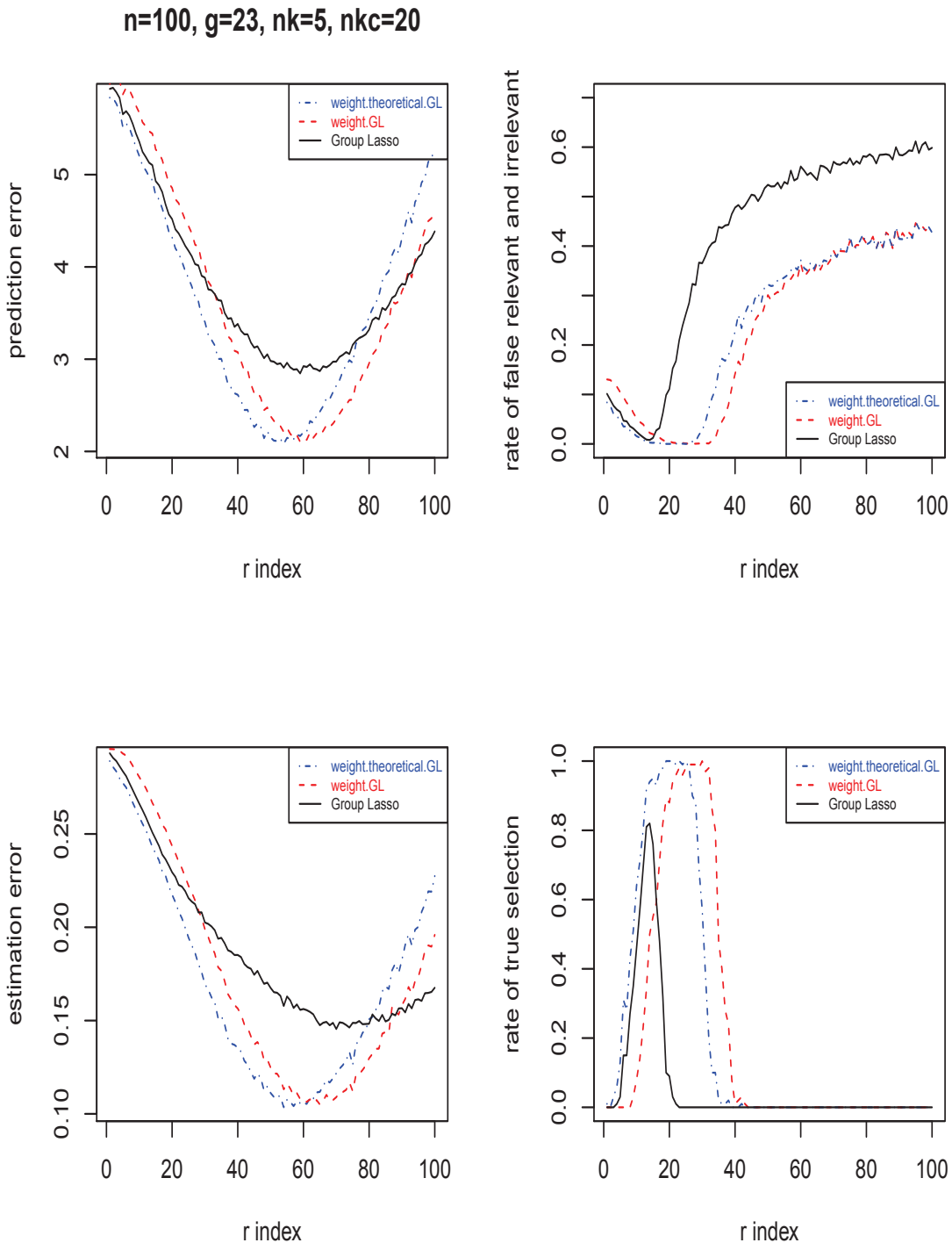


FIGURE 3.3 – *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients (see Section 3.4). All methods were fit from a path of 100 tuning parameters r from r_{max} to r_{min}. Each point corresponds to the average after 100 simulations from the setup described in Section 3.4.*

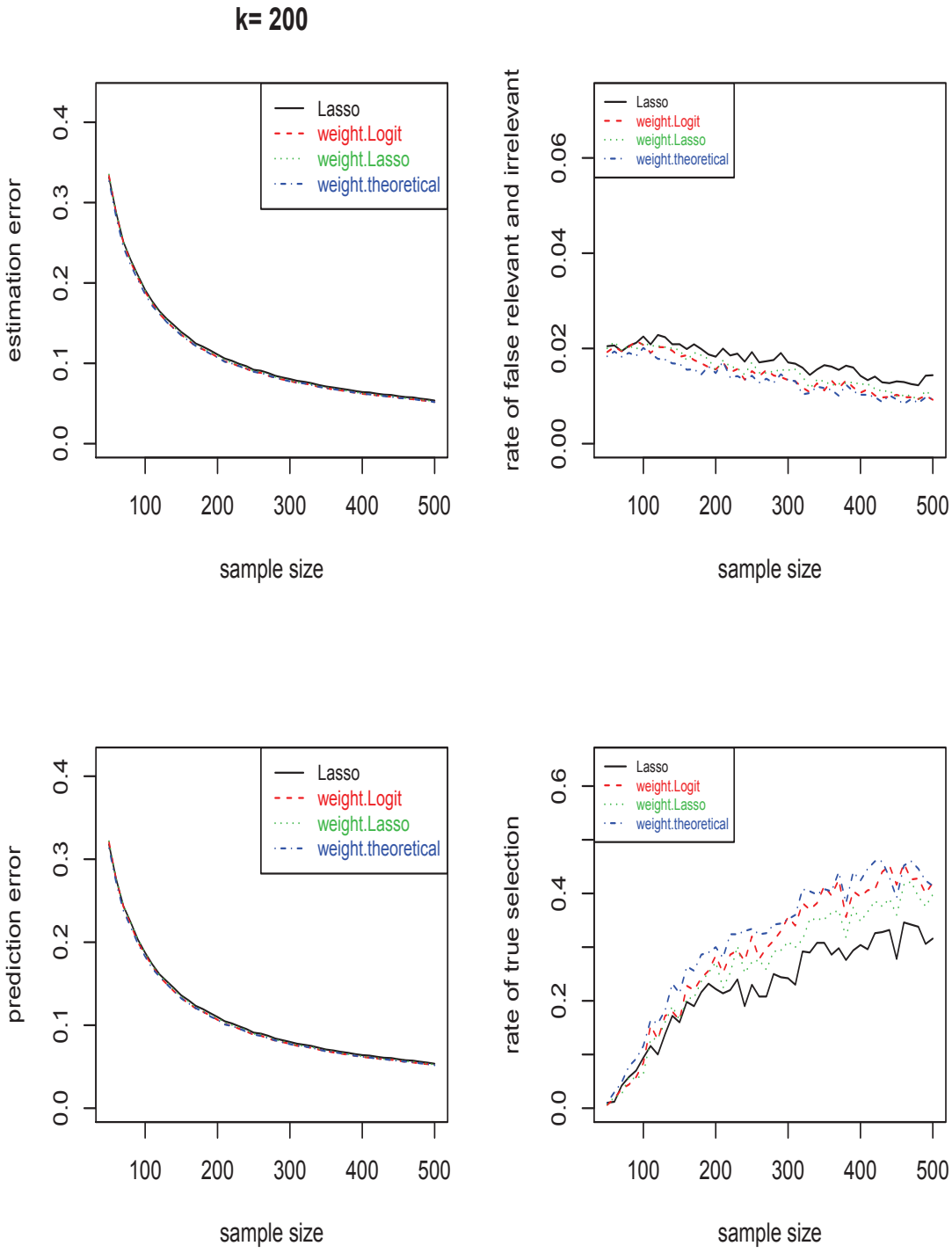


FIGURE 3.4 – *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients (see Section 3.4). k=200 from the setup described in Section 3.4*

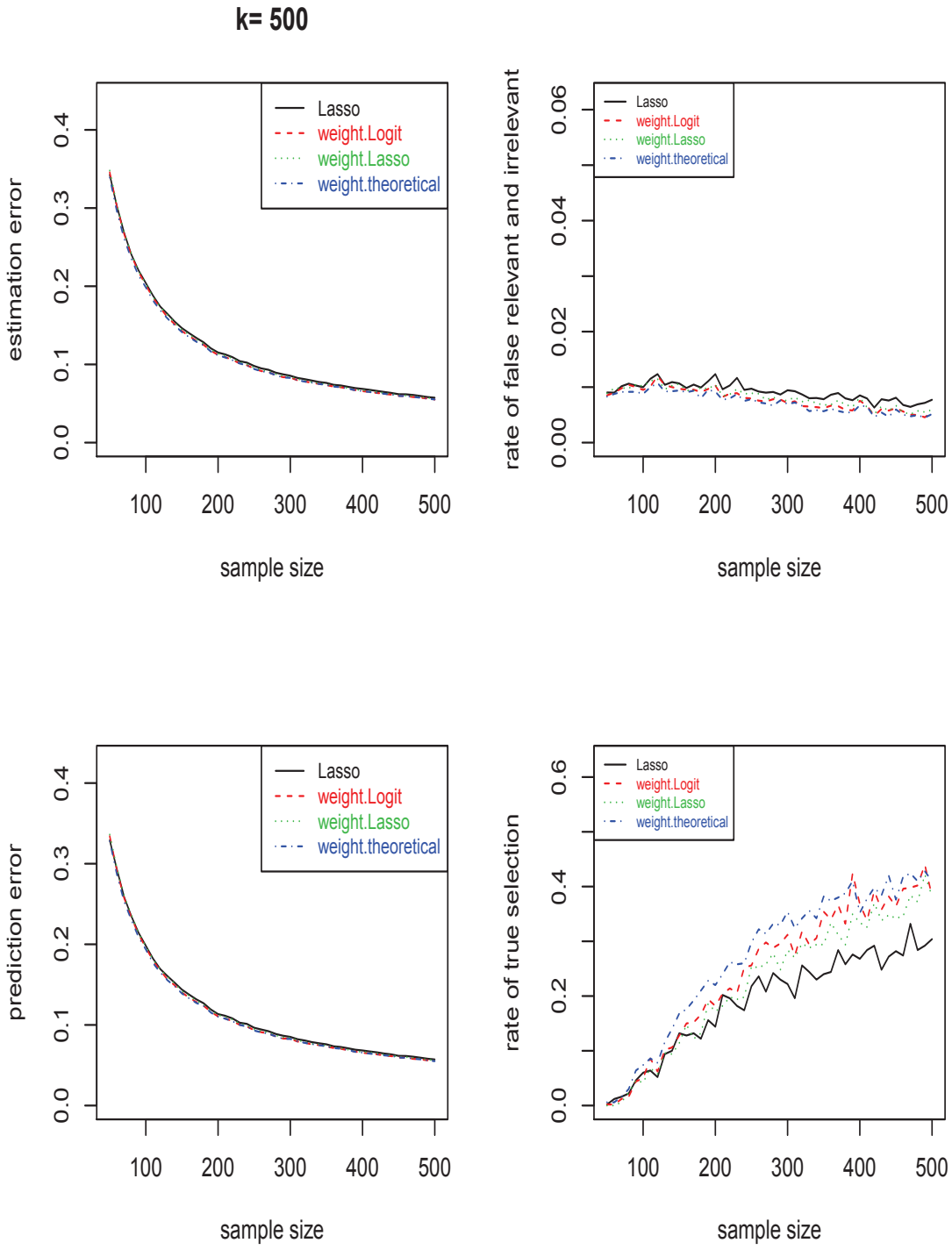


FIGURE 3.5 – *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients (see Section 3.4). k=500 from the setup described in Section 3.4*

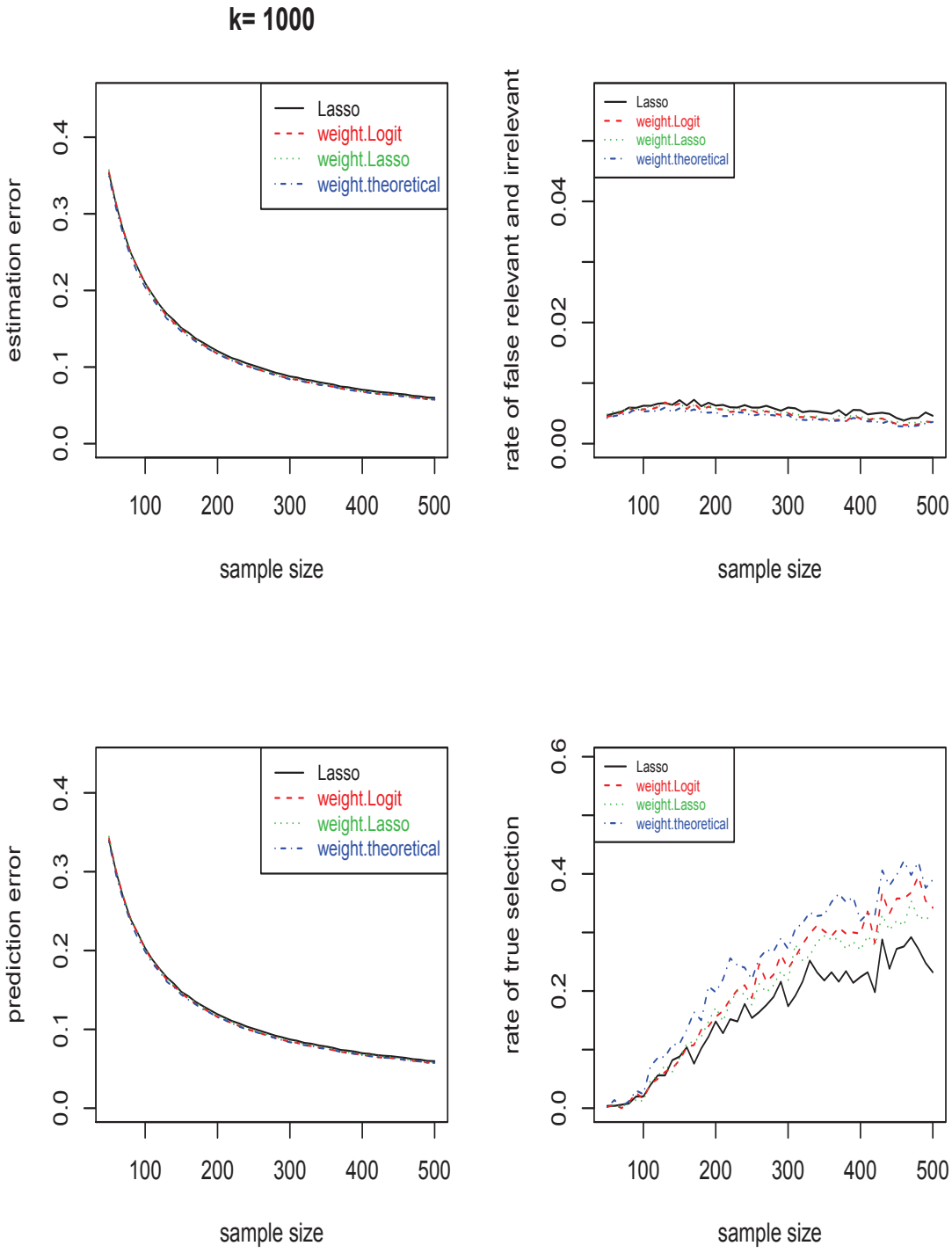


FIGURE 3.6 – *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficient (see Section 3.4). k=1000 from the setup described in Section 3.4*

By applying Cauchy-Schwarz inequality, we obtain

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq R(f_\beta) - R(f_0) + \sum_{l=1}^{g} \frac{1}{n} \sqrt{\sum_{j \in G_l} \left(\sum_{i=1}^{n} \phi_j(z_i)\epsilon_i \right)^2} \|(\hat{\beta}_{GL} - \beta)^l\|_2 + r \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 - r \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l_{GL}\|_2. \quad (3.33)$$

Set $Z_l = n^{-1} \sqrt{\sum_{j \in G_l} \left(\sum_{i=1}^{n} \phi_j(z_i)\epsilon_i \right)^2}$, for $l \in \{1, \ldots, g\}$ and the event

$$\mathcal{A} = \bigcap_{l=1}^{g} \{ Z_l \leq r\omega_l/2 \}. \quad (3.34)$$

We state the result on event $\mathcal{A}$ and find an upper bound of $\mathbb{P}(\mathcal{A}^c)$.
On the event $\mathcal{A}$:

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq R(f_\beta) - R(f_0) + r \sum_{l=1}^{g} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 + r \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 - r \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l_{GL}\|_2.$$

This implies that

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq R(f_\beta) - R(f_0) + 2r \sum_{l=1}^{g} \omega_l \|\beta^l\|_2.$$

We conclude that on the event $\mathcal{A}$ we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq \inf_{\beta \in \mathbb{R}^p} \left\{ R(f_\beta) - R(f_0) + 2r\|\beta\|_{2,1} \max_{1 \leq l \leq g} \omega_l \right\}.$$

We now come to the bound of $\mathbb{P}(A^c)$ and write

$$\mathbb{P}(\mathcal{A}^c) = \mathbb{P}\left(\bigcup_{l=1}^{g} \{ \sqrt{\sum_{j \in G_l} \left(\sum_{i=1}^{n} \phi_j(z_i)\epsilon_i \right)^2} > nr\omega_l/2 \} \right) \quad (3.35)$$

$$\leq \sum_{l=1}^{g} \mathbb{P}\left(\sqrt{\sum_{j \in G_l} \left(\sum_{i=1}^{n} \phi_j(z_i)\epsilon_i \right)^2} > nr\omega_l/2 \right). \quad (3.36)$$

For $j \in G_l$ set $T^l_j = \sum_{i=1}^{n} \phi_j(z_i)\epsilon_i$, we have

$$\mathbb{P}(\mathcal{A}^c) \leq \sum_{l=1}^{g} \mathbb{P}\left(\sqrt{\sum_{j \in G_l} (T^l_j)^2} > nr\omega_l/2 \right) \leq \sum_{l=1}^{g} \mathbb{P}\left(\sum_{j \in G_l} |T^l_j| > nr\omega_l/2 \right).$$

Using the fact that, for all $l \in \{1, \ldots, g\}$

$$\left\{ \sum_{j \in G_l} |T^l_j| > nr\omega_l/2 \right\} \subset \bigcup_{j \in G_l} \left\{ |T^l_j| > \frac{nr\omega_l}{2|G_l|} \right\}, \quad (3.37)$$

it follows that

$$\mathbb{P}(\mathcal{A}^c) \leq \sum_{l=1}^{g} \sum_{j \in G_l} \mathbb{P}\left(|T_j^l| > \frac{nr\omega_l}{2|G_l|} \right).$$

For $j \in G_l$, set $v_j^l = \sum_{i=1}^n \boldsymbol{E}(\phi_j^2 \epsilon_i^2)$. Since $\sum_{i=1}^n \phi_j^2(z_i) \geqslant 4v_j^l$, we have

$$\mathbb{P}(|T_j^l| > \frac{nr\omega_l}{2|G_l|}) \leq \mathbb{P}\left(|T_j^l| > \sqrt{2v_j^l\,(x + \log p)} + \frac{c_2}{3}\,(x + \log p) \right), \quad r \geq 1.$$

By applying Bernstein's inequality (see Lemma 3.5) to the right hand side of the previous inequality we get

$$\mathbb{P}(|T_j^l| > \frac{n\omega_l}{2|G_l|}) \leq 2\exp\left(-x - \log p\right).$$

It follows that

$$\mathbb{P}(\mathcal{A}^c) \leq \sum_{l=1}^{g} \sum_{j \in G_l} \mathbb{P}\left(|T_j^l| > \frac{n\omega_l}{2|G_l|} \right) \leq 2\exp(-x). \tag{3.38}$$

This ends the proof of the Theorem 3.1. ■

Proof of Theorem 3.2

Fix an arbitrary $\beta \in \mathbb{R}^p$ such that $f_\beta \in \Gamma_1$. Set $\delta = W(\hat{\beta}_{GL} - \beta)$ where $W = \mathrm{Diag}(W_1, \ldots, W_p)$ is a block diagonal matrix, with $W_l = \mathrm{Diag}(\omega_l, \ldots, \omega_l)$. Since $\hat{\beta}_{GL}$ is the minimizer of $\hat{R}(f_\beta) + r\sum_{l=1}^g \omega_l \|\beta^l\|_2$, we get

$$R(f_{\hat{\beta}_{GL}}) - \frac{1}{n}\varepsilon^T X\hat{\beta}_{GL} + r\sum_{l=1}^{g} \omega_l \|\hat{\beta}_{GL}^l\|_2 \leq R(f_\beta) - \frac{1}{n}\varepsilon^T X\beta + r\sum_{l=1}^{g} \omega_l \|\beta^l\|_2.$$

On the event $\mathcal{A}$ defined in (3.34), adding the term $\frac{r}{2}\sum_{l=1}^g \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2$ to both sides of Inequality (3.33) yields to

$$R(f_{\hat{\beta}_{GL}}) + \frac{r}{2}\sum_{l=1}^{g} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq R(f_\beta) + r\sum_{l=1}^{g} \omega_l (\|(\hat{\beta}_{GL} - \beta)^l\|_2 - \|\hat{\beta}_{GL}^l\|_2 + \|\beta^l\|_2).$$

Since $\|(\hat{\beta}_{GL} - \beta)^l\|_2 - \|\hat{\beta}_{GL}^l\|_2 + \|\beta^l\|_2 = 0$ for $l \notin J(\beta) = J$, we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) + \frac{r}{2}\sum_{l=1}^{g} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq R(f_\beta) - R(f_0) + 2r\sum_{l \in J} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2. \tag{3.39}$$

we get from Equation (3.39) that

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq R(f_\beta) - R(f_0) + 2r\sum_{l \in J} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 \tag{3.40}$$

Consider separately the two events :

$$\mathcal{A}_1 = \{2r\sum_{l \in J} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq \eta(R(f_\beta) - R(f_0))\},$$

and

$$\mathcal{A}_1^c = \{\eta(R(f_\beta) - R(f_0)) < 2r\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2\}. \tag{3.41}$$

On the event $\mathcal{A} \cap \mathcal{A}_1$, we get from (3.40)

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta)(R(f_\beta) - R(f_0)), \tag{3.42}$$

and the result follows. On the event $\mathcal{A} \cap \mathcal{A}_1^c$, all the following inequalities are valid. On one hand, by applying Cauchy Schwarz inequality, we get from (3.40) that

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) &\leq R(f_\beta) - R(f_0) + 2r\sqrt{|J(\beta)|}\sqrt{\sum_{l\in J}\omega_l^2\|(\hat{\beta}_{GL} - \beta)^l\|_2^2} \\ &\leq R(f_\beta) - R(f_0) + 2r\sqrt{|J(\beta)|}\|\delta_J\|_2. \end{aligned} \tag{3.43}$$

On the other hand we get from Equation (3.39) that

$$\frac{r}{2}\sum_{l=1}^{g}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq R(f_\beta) - R(f_0) + 2r\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2,$$

and using (3.41) we obtain

$$\frac{1}{2}\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 + \frac{1}{2}\sum_{l\in J^c}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq \frac{2}{\eta}\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 + 2\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2,$$

which implies

$$\|\delta_{J^c}\|_{2,1} \leq (3 + 4/\eta)\|\delta_J\|_{2,1}.$$

We can therefore apply Assumption (**RE$_5$**) with $a_0 = 3 + 4/\eta$, and conclude that

$$\mu_1^2\|\delta_J\|_2^2 \leq \frac{\|X\delta\|_2^2}{n} = \frac{1}{n}(\hat{\beta}_{GL} - \beta)^T W X^T X W(\hat{\beta}_{GL} - \beta) \leq (\max_{1\leq l\leq g}\omega_l)^2\|f_{\hat{\beta}_{GL}} - f_\beta\|_n^2. \tag{3.44}$$

Gathering Equations (3.43) and (3.44) we get

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) &\leq R(f_\beta) - R(f_0) + 2r(\max_{1\leq l\leq g}\omega_l)\sqrt{|J(\beta)|}\mu_1^{-1}\|f_{\hat{\beta}_{GL}} - f_\beta\|_n \\ &\leq R(f_\beta) - R(f_0) + 2r(\max_{1\leq l\leq g}\omega_l)\sqrt{|J(\beta)|}\mu_1^{-1}(\|f_{\hat{\beta}_{GL}} - f_0\|_n + \|f_\beta - f_0\|_n). \end{aligned}$$

We now use Lemma 4.4 which compares excess risk to empirical norm.

Lemma 3.1 *Under assumptions (**A$_5$**) and (**A$_4$**) we have*

$$c_0\epsilon_0\|f_\beta - f_0\|_n^2 \leq R(f_\beta) - R(f_0) \leq \frac{1}{4}c_0'\|f_\beta - f_0\|_n^2.$$

where c_0 and c_0' are constants depending on C_0; and ϵ_0 is a constant depending on c_1 and c_2.

(See the Appendix for the proof of Lemma 4.4).
Consequently

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) \quad \leq \quad & R(f_\beta) - R(f_0) + \frac{2r(\max_{1\leq l\leq g}\omega_l)\sqrt{|J(\beta)|}\mu_1^{-1}}{\sqrt{c_0\epsilon_0}}\sqrt{R(f_{\hat{\beta}_{GL}}) - R(f_0)} \\ & + \frac{2r(\max_{1\leq l\leq g}\omega_l)\sqrt{|J(\beta)|}\mu_1^{-1}}{\sqrt{c_0\epsilon_0}}\sqrt{R(f_\beta) - R(f_0)}. \end{aligned}$$

Using inequality $2uv < u^2/b + bv^2$ for all $b > 1$, with $u = r(\max_{1\le l\le g}\omega_l)\frac{\sqrt{|J(\beta)|}\mu_1^{-1}}{\sqrt{c_0\epsilon_0}}$ and v being either
$\sqrt{R(f_{\hat\beta_{GL}}) - R(f_0)}$ or $\sqrt{R(f_\beta) - R(f_0)}$ we have

$$\begin{aligned} R(f_{\hat\beta_{GL}}) - R(f_0) \quad &\le \quad R(f_\beta) - R(f_0) + 2b\left(\frac{r(\max_{1\le l\le g}\omega_l)\sqrt{|J(\beta)|}\mu_1^{-1}}{\sqrt{c_0\epsilon_0}}\right)^2 \\ &\quad + \frac{R(f_{\hat\beta_{GL}}) - R(f_0)}{b} + \frac{R(f_\beta) - R(f_0)}{b}. \end{aligned}$$

This implies that

$$R(f_{\hat\beta_{GL}}) - R(f_0) \le \frac{b+1}{b-1}\left\{R(f_\beta) - R(f_0) + \frac{2b^2r^2(\max_{1\le l\le g}\omega_l)^2|J(\beta)|}{(b+1)\mu_1^2c_0\epsilon_0}\right\}. \tag{3.45}$$

Now taking $b = 1 + 2/\eta$ leads to

$$R(f_{\hat\beta_{GL}}) - R(f_0) \le (1+\eta)\left\{R(f_\beta) - R(f_0) + \frac{c(\eta)r^2(\max_{1\le l\le g}\omega_l)^2|J(\beta)|}{\mu_1^2c_0\epsilon_0}\right\}. \tag{3.46}$$

According to Inequalities (3.42) and (3.46) we conclude that on event $\mathcal{A}$,

$$R(f_{\hat\beta_{GL}}) - R(f_0) \le (1+\eta)\left\{R(f_\beta) - R(f_0) + \frac{c(\eta)r^2(\max_{1\le l\le g}\omega_l)^2|J(\beta)|}{\mu_1^2c_0\epsilon_0}\right\}, \tag{3.47}$$

where $c(\eta) = 2(1+2/\eta)^2/(2+2/\eta)$. Inequality (3.8) of the Theorem 3.2 follows. Inequality (3.9) follows from Lemma 4.4. This ends the proof of the Theorem 3.2 by considering (3.38). ■

Proof of Corollary 3.1

Set $\delta = W(\hat\beta_{GL} - \beta_0)$, Line (3.13) of Corollary 3.1 follows directly from Equation (3.47) with $\beta = \beta_0$ and $\eta = 1$. Note that on the event $\mathcal{A}$ defined in (3.34), we have

$$\|\delta_{J(\beta_0)^c}\|_{2,1} \le 3\|\delta_{J(\beta_0)}\|_{2,1}. \tag{3.48}$$

Indeed, since $\hat\beta_{GL}$ is the minimizer of $\hat R(f_\beta) + r\sum_{l=1}^g \omega_l\|\beta^l\|_2$,

$$R(f_{\hat\beta_{GL}}) - R(f_{\beta_0}) + r\sum_{l=1}^g \omega_l\|\hat\beta^l_{GL}\|_2 \le \frac{1}{n}\varepsilon^T X(\hat\beta_{GL} - \beta_0) + r\sum_{l=1}^g \omega_l\|\beta^l_0\|_2$$

which implies

$$r\|W\hat\beta_{GL}\|_{2,1} \le \sum_{l=1}^g \frac{1}{n}\sqrt{\sum_{j\in G_l}\left(\sum_{i=1}^n (z_{ij})\epsilon_i\right)^2}\,\|(\hat\beta_{GL} - \beta_0)^l\|_2 + r\|W\beta_0\|_{2,1}$$

On the event A we have

$$\begin{aligned}\|W(\hat{\beta}_{GL})_{J(\beta_0)}\|_{2,1} + \|W(\hat{\beta}_{GL})_{J^c(\beta_0)}\|_{2,1} \quad &\leq \quad \frac{1}{2}(\|W(\hat{\beta}_{GL} - \beta_0)_{J(\beta_0)}\|_{2,1} + \|W(\hat{\beta}_{GL})_{J^c(\beta_0)}\|_{2,1}) \\ &\quad + \|W(\beta_0)_{J(\beta_0)}\|_{2,1}.\end{aligned}$$

This yields to (3.48). Line (3.14) follows from Line (3.13) by applying Lemma 4.4. Line (3.15) follows from Line (3.14) by using Equation (3.44) and $\|\delta\|_{2,1}^2 \leq 16s\|\delta_{J(\beta_0)}\|_2^2$. Line (3.16) is the consequence of the Lemma 3.4 with $a_l = \|(\hat{\beta}_{GL} - \beta_0)^l\|_2$ and

$$b_1 = \frac{12rs\left(\max\limits_{1\leq l\leq g}\omega_l\right)^2}{\mu^2(s,3)c_0\epsilon_0(\min\limits_{1\leq l\leq g}\omega_l)}. \quad \blacksquare$$

Proof of Theorem 3.3

On the event $\mathcal{A}$ defined in (3.34), using Inequality (3.33) with $\beta = \beta_0$ yields

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \leq \sum_{l=1}^{g} \frac{3r\omega_l}{2}\|(\hat{\beta}_{GL} - \beta_0)^l\|_2. \tag{3.49}$$

By Lemma 3.2 we have,

$$\frac{\langle h, h\rangle_{f_{\beta_0}}}{\|h\|_\infty^2}(\exp(-\|h\|_\infty) + \|h\|_\infty - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \tag{3.50}$$

where

$$h(z_i) = (f_{\hat{\beta}_{GL}} - f_{\beta_0})(z_i) = \sum_{l=1}^{g}\sum_{j\in G_l}(\hat{\beta}_{GL,j} - \beta_{0j})z_{ij}.$$

One can easily verify that$\|h\|_\infty \leq v\|\delta'\|_{2,1}$ with $\delta' = \hat{\beta}_{GL} - \beta_0$. Equation (3.50) and the decreasing of $t \mapsto \frac{\exp(-t)+t-1}{t^2}$ lead to

$$\frac{\delta'^T X^T D X \delta'}{n(v\|\delta'\|_{2,1})^2}(\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}).$$

Now, Inequality (3.48) implies

$$\|\delta'_{J(\beta_0)^c}\|_{2,1} \leq 3\frac{\left(\max\limits_{1\leq l\leq g}\omega_l\right)}{\min\limits_{1\leq l\leq g}\omega_l}\|\delta'_{J(\beta_0)}\|_{2,1}.$$

We can therefore apply Assumption (**RE$_6$**) with $a_0 = 3(\max\limits_{1\leq l\leq g}\omega_l)/\min\limits_{1\leq l\leq g}\omega_l$ and get that

$$\frac{\mu_2^2\|\delta'_{J(\beta_0)}\|_2^2}{v^2\|\delta'\|_{2,1}^2}(\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}).$$

We can use that $\|\delta'\|_{2,1}^2 \leq (1+a_0)^2|J|\|\delta'_J\|_2^2$, with $J = J(\beta_0)$ to write

$$\frac{\mu_2^2}{(1+a_0)^2|J|v^2}(\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}).$$

According to Equation (3.49) we have

$$\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1 \leq \frac{3r(1+a_0)^2\left(\max_{1\leq l\leq g}\omega_l\right)v^2|J|}{2\mu_2^2}\|\delta'\|_{2,1}. \tag{3.51}$$

Now, a short calculation shows that for all $a \in (0,1]$,

$$e^{\frac{-2a}{1-a}} + (1-a)\frac{2a}{1-a} - 1 \geqslant 0 \tag{3.52}$$

Set $a = v\|\delta'\|_{2,1}/(v\|\delta'\|_{2,1}+2)$. Thus $v\|\delta'\|_{2,1} = 2a/(1-a)$ and we have

$$e^{-v\|\delta'\|_{2,1}} + v\|\delta'\|_{2,1} - 1 \geqslant \frac{v^2\|\delta'\|_{2,1}^2}{v\|\delta'\|_{2,1}+2}. \tag{3.53}$$

This implies using Equation (3.51) that

$$v\|\delta'\|_{2,1} \leq \frac{3r(1+a_0)^2\left(\max_{1\leq l\leq g}\omega_l\right)|J|v/\mu_2^2}{1-3r(1+a_0)^2\left(\max_{1\leq l\leq g}\omega_l\right)|J|v/2\mu_2^2}.$$

Now if $r(1+a_0)^2\max_{1\leq l\leq g}\omega_l \leq \frac{\mu_2^2}{3v|J|}$, we have $v\|\delta'\|_{2,1} \leq 2$ and consequently

$$\frac{\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1}{v^2\|\delta'\|_{2,1}^2} \geqslant 1/4.$$

Now, Inequality (3.51) implies

$$\|\delta'\|_{2,1} \leq \frac{6(1+a_0)^2|J|r\left(\max_{1\leq l\leq g}\omega_l\right)}{\mu_2^2}.$$

This proves the Line (3.18). Line (3.17) follows from (3.18) by using Inequality (3.49). Line (3.19) is the consequence of Lemma 3.4 taking $a_l = \|(\hat{\beta}_{GL} - \beta_0)^l\|_2$ and $b_1 = 6(1+a_0)^2|J|r(\min_{1\leq l\leq g}\omega_l)/\mu_2^2(s,3)$. Line (3.20) follows from Line (3.17) and Inequality (3.50). ■

Proof of Theorem 3.4

Note that Lasso can be derived by Group Lasso by taking one predictor per group i.e $p = g$ and $G_j = \{j\}$ for $j \in \{1,\ldots,p\}$. This implies, using (3.33) that

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq R(f_\beta) - R(f_0) + \sum_{j=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_j(z_i)\varepsilon_i\right||\hat{\beta}_{L,j} - \beta_j| + r\sum_{j=1}^{p}\omega_j|\beta_j| - r\sum_{j=1}^{p}\omega_j|\hat{\beta}_{L,j}|.$$

For $1 \leq j \leq p$, set $S_j = \sum_{i=1}^{n}\phi_j(z_i)\varepsilon_i$ and let us denote by E, the event

$$E = \bigcap_{j=1}^{p}\left\{|S_j| \leq nr\omega_j/2\right\}. \tag{3.54}$$

We state the results on the event E and then find an upper bound of $\mathbb{P}(E^c)$.
On the event E :

$$\begin{aligned} R(f_{\hat{\beta}_L}) - R(f_0) &\leq R(f_\beta) - R(f_0) + r\sum_{j=1}^{p}\omega_j|\hat{\beta}_{L,j} - \beta_j| + r\sum_{j=1}^{p}\omega_j|\beta_j| - r\sum_{j=1}^{p}\omega_j|\hat{\beta}_{L,j}| \\ &\leq R(f_\beta) - R(f_0) + 2r\sum_{j=1}^{p}\omega_j|\beta_j|. \end{aligned}$$

We conclude that on the event E we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq \inf_{\beta\in\mathbb{R}^p}\left\{R(f_\beta) - R(f_0) + 2r\|\beta\|_1 \max_{1\leq j\leq p}\omega_j\right\}.$$

Now we are going to find an upper bound of $\mathbb{P}(E^c)$:

$$\begin{aligned} \mathbb{P}(E^c) &\leq \mathbb{P}\left(\bigcup_{j=1}^{p}\{|\sum_{i=1}^{n}\phi_j(z_i)(Y_i - \boldsymbol{E}(Y_i))| > r\omega_j n/2\}\right) \\ &\leq \sum_{j=1}^{p}\mathbb{P}(|\sum_{i=1}^{n}\phi_j(z_i)(Y_i - \boldsymbol{E}(Y_i))| > r\omega_j n/2). \end{aligned}$$

For $j \in \{1,\ldots,p\}$, set $v_j = \sum_{i=1}^{n}\boldsymbol{E}(\phi_j^2\epsilon_i^2)$. Since $\sum_{i=1}^{n}\phi_j^2(z_i) \geqslant 4v_j$, we have

$$\mathbb{P}(|S_j| > nr\omega_j/2) \leq \mathbb{P}\left(|S_j| > \sqrt{2v_j(x+\log p)} + \frac{c_2}{3}(x+\log p)\right), \quad r \geq 1.$$

By applying Bernstein's inequality (see Boucheron et al. (2004), Massart (2007)) to the right hand side of the previous inequality we get

$$\mathbb{P}(|S_j| > nr\omega_j/2) \leq 2\exp(-x - \log p).$$

It follows that

$$\mathbb{P}(E^c) \leq \sum_{j=1}^{p}\mathbb{P}(|S_j| > r\omega_j n/2) \leq 2\exp(-x). \tag{3.55}$$

When $\omega_j = 1$, for all $j \in \{1,\ldots,p\}$ and $r = A\sqrt{\frac{\log p}{n}}$, we apply Hoeffding's inequality (see Boucheron et al. (2004), Massart (2007)). This leads to

$$\begin{aligned} \mathbb{P}(E^c) &= \mathbb{P}\left(\bigcup_{j=1}^{p}\{|\sum_{i=1}^{n}\phi_j(z_i)(Y_i - \boldsymbol{E}(Y_i))| > rn/2\}\right) \\ &\leq \sum_{j=1}^{p}\mathbb{P}(|\sum_{i=1}^{n}\phi_j(z_i)(Y_i - \boldsymbol{E}(Y_i))| > rn/2) \\ &\leq 2p\exp\left(-\frac{2(rn/2)^2}{\sum_{i=1}^{n}2c_2}\right) = 2p\exp\left(-\frac{r^2n}{4c_2}\right) = 2p^{1-\frac{A^2}{4c_2}}. \end{aligned} \tag{3.56}$$

This ends the proof of Theorem 3.4. ■

Proof of Theorem 4.1

Fix an arbitrary $\beta \in \mathbb{R}^p$ such that $f_\beta \in \Gamma$, and set $\delta = W(\hat{\beta}_L - \beta)$, where $W = \mathrm{Diag}(w_1, \ldots, w_p)$. It follows from Inequality (3.47) that

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta)\left\{ R(f_\beta) - R(f_0) + \frac{c(\eta) r^2 \left(\max_{1\leq j\leq p} \omega_j\right)^2 |K(\beta)|}{\mu^2 c_0 \epsilon_0} \right\}, \quad (3.57)$$

where $c(\eta) = 2(1+2/\eta)^2/(2+2/\eta)$. This ends the proof of Inequality (3.23) of the Theorem 4.1. Inequality (3.24) follows from Lemma 4.4. To prove Inequalities (3.25) and (3.26) we just replace ω_j by $A\sqrt{\frac{\log p}{n}}$.
This ends the proof of the Theorem 4.1 by using (3.55) and (3.56). ■

Proof of Corollary 3.2

Set $\delta = W(\hat{\beta}_L - \beta_0)$. The result (3.29) directly comes by taking $\beta = \beta_0$ and $\eta = 2$ in (3.57). Note that, on the event E defined in (3.54), we have

$$\|\delta_{K(\beta_0)^c}\|_1 \leq 3\|\delta_{K(\beta_0)}\|_1. \quad (3.58)$$

Indeed, since $\hat{\beta}_L$ is the minimizer of $\hat{R}(f_\beta) + r\sum_{j=1}^p \omega_j|\beta_j|$, then

$$R(f_{\hat{\beta}_L}) - R(f_{\beta_0}) + r\sum_{j=1}^p \omega_j|\hat{\beta}_{L,j}| \leq \frac{1}{n}\varepsilon^T X(\hat{\beta}_L - \beta_0) + r\sum_{j=1}^p \omega_j|\beta_{0j}|,$$

which implies that

$$r\|W\hat{\beta}_L\|_1 \leq \sum_{j=1}^p \left|\frac{1}{n}\sum_{i=1}^n \phi_j(z_i)\varepsilon_i\right| |\hat{\beta}_{L,j} - \beta_j| + r\|W\beta_0\|_1.$$

On the event E we have

$$\begin{aligned} \|W(\hat{\beta}_L)_{K(\beta_0)}\|_1 + \|W(\hat{\beta}_L)_{K^c(\beta_0)}\|_1 \quad &\leq \quad \frac{1}{2}(\|W(\hat{\beta}_L - \beta_0)_{K(\beta_0)}\|_1 + \|W(\hat{\beta}_L)_{K^c(\beta_0)}\|_1) \\ &\quad + \|W(\beta_0)_{K(\beta_0)}\|_1. \end{aligned}$$

Thus (3.58) follows. Line (3.30) follows from Line (3.29) by applying Lemma 4.4. Line (3.31) follows from Line(3.30) by using Inequality (3.44) and $\|\delta\|_1^2 \leq 16s\|\delta_{K(\beta_0)}\|_2^2$. The last line follows from Lemma 3.4 in Appendix with $a_j = |\hat{\beta}_{L,j} - \beta_{0j}|$ and

$$b_1 = \frac{12sr\left(\max_{1\leq j\leq p} \omega_j\right)^2}{\mu^2(s,3)c_0\epsilon_0\left(\min_{1\leq j\leq p} \omega_j\right)}. \quad ■$$

Appendix

The proof of Lemma 4.4 are based on property of self concordant function (see for instance Nesterov et Nemirovskii (1994)), *i.e.*, the functions whose third derivatives are controlled by their second derivatives. A one-dimensional, convex function g is called self concordant if

$$|g^{'''}(x)| \leq Cg^{''}(x)^{3/2}.$$

The function we use ($g(t) = \hat{R}(g + th)$) is not really self concordant but we can bound his third derivative by the second derivative times a constant. Our results on self-concordant functions are based on the ones of Bach (2010). He has used and extended tools from convex optimization and self-concordance to provide simple extensions of theoretical results for the square loss to logistic loss. We use the same kind of arguments and state some relations between excess risk and prediction loss in the context of nonparametric logistic model, where f_0 is not necessarily linear as assumed in Bach (2010). Precisely we extend Proposition 1 in Bach (2010) to the functions which are not necessarily linear (see Lemma 3.2). This allows us to establish Lemma 4.4.

Lemma 3.2 *For all $h, f : \mathbb{R}^d \to \mathbb{R}$, we have*

$$\frac{\langle h,h\rangle_f}{\|h\|_\infty^2}(\exp(-\|h\|_\infty) + \|h\|_\infty - 1) \leq R(f+h) - R(f) + (q_f - q_{f_0})(h), \tag{3.59}$$

$$R(f+h) - R(f) + (q_f - q_{f_0})(h) \leq \frac{\langle h,h\rangle_f}{\|h\|_\infty^2}(\exp(\|h\|_\infty) - \|h\|_\infty - 1), \tag{3.60}$$

and

$$\langle h,h\rangle_f e^{-\|h\|_\infty} \leq \langle h,h\rangle_{f+h} \leq \langle h,h\rangle_f e^{\|h\|_\infty}. \tag{3.61}$$

Proof of Lemma 3.2

We use the following lemma (see Bach (2010) Lemma 1) that we recall here :

Lemma 3.3 *Let g be a convex three times differentiable function $g : \mathbb{R} \to \mathbb{R}$ such that for all $t \in \mathbb{R}$ $|g^{'''}(t)| \leq Sg^{''}(t)$, for some $S \geq 0$. Then , for all $t \geq 0$:*

$$\frac{g^{''}(0)}{S^2}(\exp(-St) + St - 1) \leq g(t) - g(0) - g'(0)t \leq \frac{g^{''}(0)}{S^2}(\exp(St) - St - 1). \tag{3.62}$$

We refer to Appendix A of Bach (2010) for the proof of this lemma.
Set

$$g(t) = \hat{R}(f + th) = \frac{1}{n}\sum_{i=1}^{n} l((f+th)(z_i)) - Y_i(f+th)(z_i), \quad f, h \in H,$$

where $l(u) = \log(1 + \exp(u))$. A short calculation leads to $l'(u) = \pi(u)$, $l^{''}(u) = \pi(u)(1-\pi(u))$, $l^{'''}(u) = \pi(u)[1-\pi(u)][1-2\pi(u)]$. It follows that

$$g^{''}(t) = \frac{1}{n}\sum_{i=1}^{n} h^2(z_i) l^{''}((f+th)(z_i)) = \langle h,h\rangle_{f+th},$$

and

$$g^{'''}(t) = \frac{1}{n}\sum_{i=1}^{n} h^3(z_i) l^{'''}((f+th)(z_i)).$$

Since $l'''(u) \leq l''(u)$ we have,

$$|g'''(t)| = \left| \frac{1}{n} \sum_{i=1}^{n} h^3(z_i) l'''((f+th)(z_i)) \right|$$
$$\leq \frac{1}{n} \sum_{i=1}^{n} h^2(z_i) l''((f+th)(z_i)) \|h\|_\infty = \|h\|_\infty g''(t).$$

We now apply Lemma 3.3 to $g(t)$ with $S = \|h\|_\infty$, taking $t = 1$. Using Equation (3.4) we get the first and second inequality of Lemma 3.2. Now by considering $g(t) = \langle h, h \rangle_{f+th}$, a short calculation leads to $|g'(t)| \leq \|h\|_\infty g(t)$ which implies $g(0)e^{-\|h\|_\infty t} \leq g(t) \leq g(0)e^{\|h\|_\infty t}$. By applying the last inequality to $g(t)$, and taking $t = 1$ we get the third inequality of Lemma 3.2.

Proof of Lemma 4.4

Set $h_0 = f_\beta - f_0$ from Lemma 3.2 below,

$$\frac{\langle h_0, h_0 \rangle_{f_0}}{\|h_0\|_\infty^2} (\exp(-\|h_0\|_\infty) + \|h_0\|_\infty - 1) \leq R(f_\beta) - R(f_0).$$

Using Assumptions $(\mathbf{A_4})$, $(\mathbf{A_5})$ and the decreasing of $t \mapsto \frac{\exp(-t)+t-1}{t^2}$, we claim that there exists $c_0 = c_0(C_0, c_1) > 0$ such that

$$c_0 \leq \frac{\exp(-\|h_0\|_\infty) + \|h_0\|_\infty - 1)}{\|h_0\|_\infty^2}.$$

According to Assumption $(\mathbf{A_5})$, there exists $0 \leq \epsilon_0 \leq 1/2$ such that for $1 \leq i \leq n$

$$\epsilon_0 \leq \pi(f_0(z_i))(1 - \pi(f_0(z_i))) \leq 1 - \epsilon_0.$$

The proof of the left hand side of Lemma 4.4 follows from the fact that $\epsilon_0 \|h_0\|_n^2 \leq \langle h_0, h_0 \rangle_{f_0}$. From the second line of Lemma 3.2 we have

$$R(f_\beta) - R(f_0) \leq \frac{\langle h_0, h_0 \rangle_{f_0}}{\|h_0\|_\infty^2} (\exp(\|h_0\|_\infty) - \|h_0\|_\infty - 1).$$

Using assumption $(\mathbf{A_4})$ and increasing of $t \mapsto \frac{\exp(t)-t-1}{t^2}$ thus there exists $c_0' = c_0'(C_0, c_1) > 0$ such that

$$R(f_\beta) - R(f_0) \leq c_0' \langle h_0, h_0 \rangle_{f_0}$$
$$\leq c_0' \frac{1}{4} \|h_0\|_n^2.$$

This end the proof of the right hand side of the Lemma 4.4.

Lemma 3.4 *If we assume that $\sum_{i=1}^p a_j \leq b_1$ with $a_j > 0$, this implies that $\sum_{i=1}^p a_j^q \leq b_1^q$, with $1 \leq q \leq 2$.*

Proof of Lemma 3.4

We start by writing

$$\sum_{i=1}^{p} a_j^q = \sum_{i=1}^{p} a_j^{2-q} a_j^{2q-2}$$
$$\leq \left(\sum_{i=1}^{p} a_j \right)^{2-q} \left(\sum_{i=1}^{p} a_j^2 \right)^{q-1}.$$

Since $\sum_{i=1}^{p} a_j^2 \leq \left(\sum_{i=1}^{p} a_j\right)^2 \leq b_1^2$, thus

$$\sum_{i=1}^{p} a_j^q \leq b_1^{2-q} b_1^{2q-2} = b_1^q. \tag{3.63}$$

This ends the proof.

Lemma 3.5 (Bernstein's inequality) *Let $X_1, \ldots, X_n$ be independent real valued random variables such that for all $i \leq n$, $X_i \leq b$ almost surely, then for all $x > 0$, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}(X_i)\right| \geqslant \sqrt{2vx} + bx/3\right] \leq 2\exp(-x),$$

where $v = \sum_{i=1}^{n} \mathbb{E}(X_i^2)$.

This lemma is obtain by gathering Proposition 2.9 and inequality (2.23) from Massart (2007).

Lemma 3.6 (Hoeffding's inequality) *Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ almost surely for all $i \leq n$. Then for any positive x, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}(X_i)\right| \geqslant x\right] \leq 2\exp\left(-\frac{2x^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

This lemma is a consequence of Proposition 2.7 in Massart (2007).

Model selection for logistic regression

4

Sommaire

This chapter is devoted to model selection in logistic regression. We extend the model selection principle introduced by Birgé and Massart (2001) to logistic regression model. This selection is done by using penalized maximum likelihood criteria. We propose in this context a completely data-driven criteria based on the slope heuristics. We prove non asymptotic oracle inequalities for selected estimators. Theoretical results are illustrated through simulation studies.

4.1 Introduction

Consider the following generalization of the logistic regression model : let $(Y_1, x_1), \cdots, (Y_n, x_n)$, be a sample of size n such that $(Y_i, x_i) \in \{0,1\} \times \mathcal{X}$ and

$$\mathbb{E}_{f_0}(Y_i) = \pi_{f_0}(x_i) = \frac{\exp f_0(x_i)}{1 + \exp f_0(x_i)},$$

where f_0 is an unknown function to be estimated and the design points $x_1, ..., x_n$ are deterministic. This model can be viewed as a nonparametric version of the "classical" logistic model which relies on the assumption that $x_i \in \mathbb{R}^d$, and that there exists $\beta_0 \in \mathbb{R}^d$ such that $f_0(x_i) = \beta_0^\top x_i$.

Logistic regression is a widely used model for predicting the outcome of binary dependent variable. For example logistic model can be used in medical study to predict the probability that a patient has a given disease (e.g. cancer), using observed characteristics (explanatory variables) of the patient such as weight, age, patient's gender *etc*. However in the presence of numerous explanatory variables with potential influence, one would like to use only a few number of variables, for the sake of interpretability or to avoid overfitting. But it is not always obvious to choose the adequate variables. This is the well-known problem of variables selection or model selection. In this chapter, the unknown function f_0 is not specified and not necessarily linear. Our aim is to estimate f_0 by a linear combination of given functions, often called dictionary. The dictionary can be a basis of functions, for instance spline or polynomial basis.

A nonparametric version of the classical logistic model has already been considered by Hastie (1983), where a nonparametric estimator of f_0 is proposed using local maximum likelihood. The problem of nonparametric estimation in additive regression model is well known and deeply studied. But in logistic regression model it is less studied. One can cite for instance Lu (2006), Vexler (2006), Fan *et al.* (1998), Farmen (1996), Raghavan (1993), and Cox (1990).

Recently few papers deal with model selection or nonparametric estimation in logistic regression using ℓ_1 penalized contrast Bunea (2008b), Bach (2010), van de Geer (2008), Kwemou (2012). Among them, some establish non asymptotic oracle inequalities that hold even in high dimensional setting. When the dimension of $\mathcal{X}$ is high, that is greater than dozen, such ℓ_1 penalized contrast estimators are known to provide reasonably good results. When the dimension of $\mathcal{X}$ is small, it is often better to choose different penalty functions. One classical penalty function is what we call ℓ_0 penalization. Such penalty functions, built as increasing function of the dimension of $\mathcal{X}$, usually refers to model selection. The last decades have witnessed a growing interest in the model selection problem since the seminal works of Akaike (1973), Schwarz (1978b). In additive regression one can cite among the others Baraud (2000), Birgé and Massart (2001), Yang (1999), in density estimation Birgé (2014b), Castellan (2003b) and in segmentation problem Lebarbier (2005), Durot *et al.* (2009), and Braun *et al.* (2000). All the previously cited papers use ℓ_0 penalized contrast to perform model selection. But model selection procedures based on penalized maximum likelihood estimators in logistic regression are less studied in the literature.

In this chapter we focus on model selection using ℓ_0 penalized contrast for logistic regression model and in this context we state non asymptotic oracle inequalities. More precisely, given some collection functions, we consider estimators of f_0 built as linear combination of the functions. The point that the true function is not sup-

posed to be linear combination of those functions, but we expect that the spaces of linear combination of those functions would provide suitable approximation spaces. Thus, to this collection of functions, we associate a collection of estimators of f_0. Our aim is to propose a data driven procedure, based on penalized criterion, which will be able to choose the "best" estimator among the collection of estimators, using ℓ_0 penalty functions.

The collection of estimators is built using minimization of the opposite of logarithm likelihood. The properties of estimators are described in term of Kullback-Leibler divergence and the empirical L_2 norm. Our results can be splitted into two parts.

First, in a general model selection framework, with general collection of functions we provide a completely data driven procedure that automatically selects the best model among the collection. We state non asymptotic oracle inequalities for Kullback-Leibler divergence and the empirical L_2 norm between the selected estimator and the true function f_0. The estimation procedure relies on the building of a suitable penalty function, suitable in the sense that it performs best risks and suitable in the sense that it does not depend on the unknown smoothness parameters of the true function f_0. But, the penalty function depends on a bound related to target function f_0. This can be seen as the price to pay for the generality. It comes from needed links between Kullback-Leibler divergence and empirical L_2 norm.

Second, we consider the specific case of collection of piecewise functions which provide estimator of type regressogram. In this case, we exhibit a completely data driven penalty, free from f_0. The model selection procedure based on this penalty provides an adaptive estimator and state a non asymptotic oracle inequality for Hellinger distance and the empirical L_2 norm between the selected estimator and the true function f_0. In the case of piecewise constant functions basis, the connection between Kullback-Leibler divergence and the empirical L_2 norm are obtained without bound on the true function f_0. This last result is of great interest for example in segmentation study, where the target function is piecewise constant or can be well approximated by piecewise constant functions.

Those theoretical results are illustrated through simulation studies. In particular we show that our model selection procedure (with the suitable penalty) have good non asymptotic properties as compared to usual known criteria such as AIC and BIC. A great attention has been made on the practical calibration of the penalty function. This practical calibration is mainly based on the ideas of what is usually referred as slope heuristic as proposed in Birgé and Massart (2007) and developed in Arlot and Massart (2009).

The chapter is organized as follow. In Section 4.2 we set our framework and describe our estimation procedure. In Section 4.3 we define the model selection procedure and state the oracle inequalities in the general framework. Section 4.4 is devoted to regressogram selection, in this section, we establish a bound of the Hellinger risk between the selected model and the target function. The simulation study is reported in Section 4.5. The proofs of the results are postponed to Section 4.6 and 4.6.

4.2 Model and framework

Let $(Y_1, x_1), \cdots, (Y_n, x_n)$, be a sample of size n such that $(Y_i, x_i) \in \{0,1\} \times \mathcal{X}$. Throughout the chapter, we consider a fixed design setting *i.e.* $x_1, \ldots, x_n$ are considered as deterministic. In this setting, consider the extension of the "classical"

logistic regression model (4.1) where we aim at estimating the unknown function f_0 in

$$\mathbb{E}_{f_0}(Y_i) = \pi_{f_0}(x_i) = \frac{\exp f_0(x_i)}{1+\exp f_0(x_i)}. \tag{4.1}$$

We propose to estimate the unknown function f_0 by model selection. This model selection is performed using penalized maximum likelihood estimators. In the following we denote by $\mathbb{P}_{f_0}(x_1)$ the distribution of Y_1 and by $\mathbb{P}^{(n)}_{f_0}(x_1,\cdots,x_n)$ the distribution of $(Y_1,\ldots,Y_n)$ under Model (4.1). Since the variables Y_i's are independent random variables,

$$\mathbb{P}^{(n)}_{f_0}(x_1,\cdots,x_n) = \Pi_{i=1}^n \mathbb{P}_{f_0}(x_i) = \prod_{i=1}^n \pi_{f_0}(x_i)^{Y_i}(1-\pi_{f_0}(x_i))^{1-Y_i}.$$

It follows that for a function f mapping $\mathcal{X}$ into $\mathbb{R}$, the likelihood is defined as :

$$L_n(f) = \mathbb{P}^{(n)}_{f}(x_1,\cdots,x_n) = \prod_{i=1}^n \pi_f(x_i)^{Y_i}(1-\pi_f(x_i))^{1-Y_i},$$

where

$$\pi_f(x_i) = \frac{\exp\left(f(x_i)\right)}{1+\exp(f(x_i))}. \tag{4.2}$$

We choose the opposite of the log-likelihood as the estimation criterion that is

$$\gamma_n(f) = -\frac{1}{n}\log(L_n(f)) = \frac{1}{n}\sum_{i=1}^n \Big\{ \log(1+e^{f(x_i)}) - Y_i f(x_i)\Big\}. \tag{4.3}$$

Associated to this estimation criterion we consider the Kullback-Leibler information divergence $\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f})$ defined as

$$\mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f}) = \frac{1}{n}\int \log\left(\frac{\mathbb{P}^{(n)}_{f_0}}{\mathbb{P}^{(n)}_{f}}\right) d\mathbb{P}^{(n)}_{f_0}.$$

The loss function is the excess risk, defined as

$$\mathcal{E}(f) := \gamma(f) - \gamma(f_0) \text{ where, for any } f, \quad \gamma(f) = \mathbb{E}_{f_0}[\gamma_n(f)]. \tag{4.4}$$

Easy calculations show that the excess risk is linked to the Kullback-Leibler information divergence through the relation

$$\mathcal{E}(f) = \gamma(f) - \gamma(f_0) = \mathcal{K}(\mathbb{P}^{(n)}_{f_0},\mathbb{P}^{(n)}_{f}).$$

It follows that, f_0 minimizes the excess risk, that is

$$f_0 = \arg\min_f \gamma(f).$$

As usual, one can not estimate f_0 by the minimizer of $\gamma_n(f)$ over any functions space, since it is infinite. The usual way is to minimize $\gamma_n(f)$ over a finite dimensional collections of models, associated to a finite dictionary of functions $\phi_j : \mathcal{X} \to \mathbb{R}$

$$\mathcal{D} = \{\phi_1,\ldots,\phi_M\}.$$

For the sake of simplicity we will suppose that $\mathcal{D}$ is a orthonormal basis of functions. Indeed, if $\mathcal{D}$ is not an orthonormal basis of functions, we can always find an orthonormal basis of functions $\mathcal{D}' = \{\psi_1, \dots, \psi_{M'}\}$ such that

$$\langle \phi_1, \dots, \phi_M \rangle = \langle \psi_1, \dots, \psi_{M'} \rangle.$$

Let $\mathcal{M}$ the set of all subsets $m \subset \{1, \dots, M\}$. For every $m \in \mathcal{M}$, we call $\mathcal{S}_m$ the model

$$\mathcal{S}_m := \left\{ f_\beta = \sum_{j \in m} \beta_j \phi_j \right\} \tag{4.5}$$

and D_m the dimension of the span of $\{\phi_j, j \in m\}$. Given the countable collection of models $\{\mathcal{S}_m\}_{m \in \mathcal{M}}$, we define $\{\hat{f}_m\}_{m \in \mathcal{M}}$ the corresponding estimators, *i.e.* the estimators obtaining by minimizing γ_n over each model $\mathcal{S}_m$. For each $m \in \mathcal{M}$, $\hat{f}_m$ is defined by

$$\hat{f}_m = \arg\min_{t \in \mathcal{S}_m} \gamma_n(t). \tag{4.6}$$

Our aim is choose the "best" estimator among this collection of estimators, in the sense that it minimizes the risk. In many cases, it is not easy to choose the "best" model. Indeed, a model with small dimension tends to be efficient from estimation point of view whereas it could be far from the "true" model. On the other side, a more complex model easily fits data but the estimates have poor predictive performance (overfitting). We thus expect that this best estimator mimics what is usually called the oracle defined as

$$m^* = \arg\min_{m \in \mathcal{M}} \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)}). \tag{4.7}$$

Unfortunately, both, minimizing the risk and minimazing the kulback-leibler divergence, require the knowledge of the true (unknown) function f_0 to be estimated.

Our goal is to develop a data driven strategy based on data, that automatically selects the best estimator among the collection, this best estimator having a risk as close as possible to the oracle risk, that is the risk of $\hat{f}_{m^*}$. In this context, our strategy follows the lines of model selection as developed by Birgé and Massart (2001). We also refer to the book Massart (2007) for further details on model selection.

We use penalized maximum likelihood estimator for choosing some data-dependent $\hat{m}$ nearly as good as the ideal choice m^*. More precisely, the idea is to select $\hat{m}$ as a minimizer of the penalized criterion

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \left\{ \gamma_n(\hat{f}_m) + \text{pen}(m) \right\}, \tag{4.8}$$

where pen : $\mathcal{M} \longrightarrow \mathbb{R}^+$ is a data driven penalty function. The estimation properties of $\hat{f}_m$ are evaluated by non asymptotic bounds of a risk associated to a suitable chosen loss function. The great challenge is choosing the penalty function such that the selected model $\hat{m}$ is nearly as good as the oracle m^*. This penalty term is classically based on the idea that

$$m^* = \arg\min_{m \in \mathcal{M}} \mathbb{E}_{f_0} \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)}) = \arg\min_{m \in \mathcal{M}} \left[\mathbb{E}_{f_0} \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathbb{E}_{f_0} \mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)}) \right]$$

where f_m is defined as

$$f_m = \arg\min_{t \in S_m} \gamma(t).$$

Our goal is to build a penalty function such that the selected model $\hat{m}$ fulfills an oracle inequality :

$$\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) \leq C_n \inf_{m \in \mathcal{M}} \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)}) + R_n.$$

This inequality is expected to hold either in expectation or with high probability, where C_n is as close to 1 as possible and R_n is a remainder term negligible compared to $\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{m*}}^{(n)})$.

In the following we consider two separated case. First we consider general collection of models under boundedness assumption. Second we consider the specific case of regressogram collection.

4.3 Oracle inequality for general models collection under boundedness assumption

Consider model (4.1) and $(\mathcal{S}_m)_{m \in \mathcal{M}}$ a collection of models defined by (4.5). Let $C_0 > 0$ and $\mathbb{L}_\infty(C_0) = \left\{ f : \mathcal{X} \to \mathbb{R}, \ \max_{1 \leqslant i \leqslant n} |f(x_i)| \leqslant C_0 \right\}$. For $m \in \mathcal{M}$, γ_n given in (4.3), and γ is given by (4.4), we define

$$\hat{f}_m = \arg \min_{t \in \mathcal{S}_m \cap \mathbb{L}_\infty(C_0)} \gamma_n(t) \text{ and } f_m = \arg \min_{t \in S_m \cap \mathbb{L}_\infty(C_0)} \gamma(t). \tag{4.9}$$

The first step consists in studying the estimation properties of $\hat{f}_m$ for each m, as it is stated in the following proposition.

Proposition 4.1 *Let $C_0 > 0$ and $\mathcal{U}_0 = e^{C_0}/(1+e^{C_0})^2$. For $m \in \mathcal{M}$, let $\hat{f}_m$ and f_m as in (4.9). We have*

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)})] \leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \frac{D_m}{2n\mathcal{U}_0^2}$$

This proposition says that the "best" estimator amoung the collection $\{\hat{f}_m\}_{m \in \mathcal{M}}$, in the sense of the Kullback-Leibler risk, is the one which makes a balance between the bias and the complexity of the model. In the ideal situation where f_0 belongs to $\mathcal{S}_m$, we have that

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)})] \leqslant \frac{1}{\mathcal{U}_0^2} \frac{D_m}{2n}.$$

To derive the model selection procedure we need the following assumption :

There exists a constant $0 < c_1 < \infty$ such that $\max_{1 \leqslant i \leqslant n} |f_0(x_i)| \leqslant c_1$. **($\mathbf{A_5}$)**

In the following theorem we propose a choice for the penalty function and we state non asymptotic risk bounds.

Theorem 4.1 *Given $C_0 > 0$, for $m \in \mathcal{M}$, let $\hat{f}_m$ and f_m be defined as (4.9). Let us denote $\| f \|_n^2 = n^{(-1)} \sum_{i=1}^n f^2(x_i)$. Let $\{L_m\}_{m \in \mathcal{M}}$ some positive numbers satisfying*

$$\Sigma = \sum_{m \in \mathcal{M}} \exp(-L_m D_m) < \infty.$$

We define pen : $\mathcal{M} \to \mathbb{R}_+$ *, such that, for* $m \in \mathcal{M}$*,*

$$pen(m) \geqslant \lambda \frac{D_m}{n} \left(\frac{1}{2} + \sqrt{5L_m} \right)^2,$$

where λ *is a positive constant depending on* c_1*. Under Assumption (*$\mathbf{A_5}$*) we have*

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})] \leqslant C \inf_{m \in \mathcal{M}} \left\{ \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + pen(m) \right\} + C_1 \frac{\Sigma}{n}$$

and

$$\mathbb{E}_{f_0} \parallel \hat{f}_{\hat{m}} - f_0 \parallel_n^2 \leqslant C' \inf_{m \in \mathcal{M}} \{ \parallel f_0 - f_m \parallel_n^2 + pen(m) \} + C_1' \frac{\Sigma}{n}.$$

where C, C', C_1, C_1' *are constants depending on* c_1 *and* C_0*.*

This theorem provides oracle inequalities for L_2-norm and for K-L divergence between the selected model and the true function. Provided that penalty has been properly chosen, one can bound the L_2-norm and the K-L divergence between the selected model and the true function. The inequalities in Theorem 4.1 are non-asymptotic inequalities in the sense that the result is obtain for a fixed n. This theorem is very general and does not make specific assumption on the dictionary. However, the penalty function depends on some unknown constant λ which depends on the bound of the true function f_0 through Condition (4.5). In practice this constant can be calibrated using "slope heuristics" proposed in Birgé and Massart (2007). In the following we will show how to obtain similar result with a penalty function not connected to the bound of the true unknown function f_0 in the regressogram case.

4.4 Regressogram functions

4.4.1 Collection of models

In this section we suppose (without loss of generality) that $f_0 : [0,1] \to \mathbb{R}$. For the sake of simplicity, we use the notation $f_0(x_i) = f_0(i)$ for every $i = 1, \ldots, n$. Hence f_0 is defined from $\{1, \ldots, n\}$ to $\mathbb{R}$. Let $\mathcal{M}$ be a collection of partitions of intervals of $\mathcal{X} = \{1, \ldots, n\}$. For any $m \in \mathcal{M}$ and $J \in m$, let $\mathbb{1}_J$ denote the indicator function of J and S_m be the linear span of $\{\mathbb{1}_J, J \in m\}$. When all intervals have the same length, the partition is said regular, and is irregular otherwise.

4.4.2 Collection of estimators : regressogram

For a fixed m, the minimizer $\hat{f}_m$ of the empirical contrast function γ_n, over S_m, is called the *regressogram*. That is, f_0 is estimated by $\hat{f}_m$ given by

$$\hat{f}_m = \arg\min_{f \in S_m} \gamma_n(f). \tag{4.10}$$

where γ_n is given by (4.3). Associated to S_m we have

$$f_m = \arg\min_{f \in S_m} \gamma(f) - \gamma(f_0) = \arg\min_{f \in S_m} \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_f^{(n)}). \tag{4.11}$$

In the specific case where S_m is the set of piecewise constant functions on some partition m, $\hat{f}_m$ and f_m are given by the following lemma.

Lemma 4.1 *For $m \in \mathcal{M}$, let f_m and $\hat{f}_m$ be defined by (4.11) and (4.10) respectively . Then, $f_m = \sum_{J\in m} \overline{f}_m^{(J)} \mathbb{1}_J$ and $\hat{f}_m = \sum_{J\in m} \hat{f}_m^{(J)} \mathbb{1}_J$ with*

$$\overline{f}_m^{(J)} = \log\left(\frac{\sum_{i\in J} \pi_{f_0}(x_i)}{|J|(1-\sum_{i\in J} \pi_{f_0}(x_i)/|J|)}\right) \text{ and } \hat{f}_m^{(J)} = \log\left(\frac{\sum_{i\in J} Y_i}{|J|(1-\sum_{i\in J} Y_i/|J|)}\right).$$

Moreover, $\pi_{f_m} = \sum_{J\in m} \pi_{f_m}^{(J)} \mathbb{1}_J$ and $\pi_{\hat{f}_m} = \sum_{J\in m} \pi_{\hat{f}_m}^{(J)} \mathbb{1}_J$ with

$$\pi_{f_m}^{(J)} = \frac{1}{|J|}\sum_{i\in J} \pi_{f_0}(x_i), \text{ and } \pi_{\hat{f}_m}^{(J)} = \frac{1}{|J|}\sum_{i\in J} Y_i.$$

Consequently, $\pi_{f_m} = \arg\min_{\pi\in S_m} \| \pi - \pi_{f_0} \|_n^2$ is the usual projection of π_{f_0} on to S_m.

4.4.3 First bounds on $\hat{f}_m$

Consider the following assumptions :

$$\text{There exists a constant } \rho > 0 \text{ such that } \min_{i=1,\cdots,n} \pi_{f_0}(x_i) \geq \rho \text{ and } \min_{i=1,\cdots,n} [1-\pi_{f_0}(x_i)] \geq \rho. \tag{$\mathbf{A_6}$}$$

Proposition 4.2 *Consider Model (4.1) and let $\hat{f}_m$ be defined by (4.10) with m such that for all $J \in m$, $|J| \geqslant \Gamma[\log(n)]^2$ for a positive constant Γ. Under Assumption ($\mathbf{A_6}$), for all $\delta > 0$ and $a > 1$, we have*

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)})] \leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)})) + \frac{(1+\delta)D_m}{(1-\delta)^2 n} + \frac{\kappa(\Gamma,\rho,\delta)}{n^a}.$$

4.4.4 Adaptive estimation and oracle inequality

The following result provides an adaptive estimation of f_0 and a risk bound of the selected model.

Definition 4.1 *Let $\mathcal{M}$ be a collection of partitions of $\mathcal{X} = \{1,\ldots,n\}$ constructed on the partition m_f i.e. m_f is a refinement of every $m \in \mathcal{M}$.*

In other words, a partition m belongs to $\mathcal{M}$ if any element of m is the union of some elements of m_f. Thus S_{m_f} contains every model of the collection $\{S_m\}_{m\in\mathcal{M}}$.

Theorem 4.1 *Consider Model (4.1) under Assumption ($\mathbf{A_6}$). Let $\{S_m, m \in \mathcal{M}\}$ be a collection of models defined in Section 4.4.1 where $\mathcal{M}$ is a set of partitions constructed on the partition m_f such that*

$$\text{for all } J \in m_f, |J| \geq \Gamma \log^2(n), \tag{4.1}$$

where Γ is a positive constant. Let $(L_m)_{m\in\mathcal{M}}$ be some family of positive weights satisfying

$$\Sigma = \sum_{m\in\mathcal{M}} \exp(-L_m D_m) < +\infty. \tag{4.2}$$

Let pen : $\mathcal{M} \to \mathbb{R}_+$ satisfying for $m \in \mathcal{M}$, and for $\mu > 1$,

$$pen(m) \geqslant \mu \frac{D_m}{n}\left(1 + 6L_m + 8\sqrt{L_m}\right).$$

Let $\tilde{f} = \hat{f}_{\hat{m}}$ *where*

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \left\{ \gamma_n(\hat{f}_m) + pen(m) \right\},$$

then, for $C_\mu = 2\mu^{1/3}/(\mu^{1/3} - 1)$, *we have*

$$\mathbb{E}_{f_0}[h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\tilde{f}}^{(n)})] \leqslant C_\mu \inf_{m \in \mathcal{M}} \left\{ \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + pen(m) \right\} + \frac{C(\rho, \mu, \Gamma, \Sigma)}{n}. \quad (4.3)$$

This theorem provides a non asymptotic bound for the Hellinger risk between the selected model and the true one. On the opposite of Theorem 4.1, the penalty function does not depend on the bound of the true function. The selection procedure based only on the data offers the advantage to free the estimator from any prior knowledge about the smoothness of the function to estimate. The estimator is therefore adaptive. As we bound Hellinger risk in (4.3) by Kulback-Leibler risk, one should prefer to have the Hellinger risk on the right hand side instead of the Kulback-Leibler risk. Such a bound is possible if we assume that $\log(\|\pi_{f_0}/\rho\|_\infty)$ is bounded. Indeed if we assume that there exists T such that $\log(\|\pi_{f_0}/\rho\|_\infty) \leq T$, this implies that $\log(\|\pi_{f_0}/\pi_{f_m}\|_\infty) \leq T$ uniformly for all partitions $m \in \mathcal{M}$. Now using Inequality (7.6) p. 362 in Birgé and Massart (1998) we have that $\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) \leq (4 + 2\log(M))h^2(\mathbb{P}_{f_0}, \mathbb{P}_{f_m})$ which implies,

$$\mathbb{E}_{f_0}[h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\tilde{f}}^{(n)})] \leqslant C_\mu.C(T) \inf_{m \in \mathcal{M}} \left\{ h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m) \right\} + \frac{C(\rho, \mu, \Gamma, \Sigma)}{n}.$$

Choice of the weights $\{L_m, m \in \mathcal{M}\}$

According to Theorem 4.1, the penalty function depends on the collection $\mathcal{M}$ through the choice of the weights L_m satisfying (4.2), *i.e.*

$$\Sigma = \sum_{m \in -M} \exp(-L_m D_m) = \sum_{D \geq 1} e^{-L_D D} Card\{m \in \mathcal{M}, |m| = D\} < \infty. \quad (4.4)$$

Hence the number of models having the same dimension D plays an important role in the risk bound.

If there is only one model of dimension D, a simple way of choosing L_D is to take them constant, *i.e.* $L_D = L$ for all $m \in \mathcal{M}$, and thus we have from (4.4)

$$\Sigma = \sum_{D \geq 1} e^{-LD} < \infty.$$

This is the case when $\mathcal{M}$ is a family of regular partitions. Consequently, the choice *i.e.* $L_D = L$ for all $m \in \mathcal{M}$ leads to a penalty proportional to the dimension D_m, and for every $D_m \geq 1$,

$$\mathrm{pen}(m) = \mu\left(1 + 6L + 8\sqrt{L}\right)\frac{D_m}{n} = c \times \frac{D_m}{n}. \quad (4.5)$$

In the more general context, that is in the case of irregular partitions, the numbers of models having the same dimension D is exponential and satisfies

$$Card\Big\{m \in \mathcal{M}, |m| = D\Big\} = \binom{n-1}{D-1} \leq \binom{n}{D}.$$

In that case we choose L_m depending on the dimension D_m. With L depending on D, Σ in (4.2) satisfies

$$\begin{aligned}\Sigma &= \sum_{D\geq 1} e^{-L_D D} Card\{m \in \mathcal{M}, |m| = D\} \\ &\leq \sum_{D\geq 1} e^{-L_D D} \binom{n}{D} \\ &\leq \sum_{D\geq 1} e^{-L_D D} \left(\frac{en}{D}\right)^D \\ &\leq \sum_{D\geq 1} e^{-D\left(L_D - 1 - \log\left(\frac{n}{D}\right)\right)}\end{aligned}$$

So taking $L_D = 2 + \log\left(\frac{n}{D}\right)$ leads to $\Sigma < \infty$ and the penalty becomes

$$\text{pen}(m) = \mu \times \text{pen}_{\text{shape}}(m), \tag{4.6}$$

where

$$\text{pen}_{\text{shape}}(m) = \frac{D_m}{n}\left[13 + 6\log\left(\frac{n}{D_m}\right) + 8\sqrt{2 + \log\left(\frac{n}{D_m}\right)}\right]. \tag{4.7}$$

The constant μ can be calibrated using the slope heuristics Birgé and Massart (2007) (see Section 4.5.2).

Remark 4.1 *In Theorem 4.1, we do not assume that the target function f_0 is piecewise constant. However in many contexts, for instance in segmentation, we might want to consider that f_0 is piecewise constant or can be well approximated by piecewise constant functions. That means there exists of partition of $\mathcal{X}$ within which the observations follow the same distribution and between which observations have different distributions.*

4.5 Simulations

In this section we present numerical simulation to study the non-asymptotic properties of the model selection procedure introduced in Section 4.4.4. More precisely, the numerical properties of the estimators built by model selection with our criteria are compared with those of the estimators resulting from model selection using the well known criteria AIC and BIC.

4.5.1 Simulations frameworks

We consider the model defined in (4.1) with $f_0 : [0,1] \to \mathbb{R}$. The aim is to estimate f_0. We consider the collection of models $(S_m)_{m\in\mathcal{M}}$, where

$$S_m = \text{Vect}\{\mathbb{1}_{\left[\frac{k-1}{D_m}, \frac{k}{D_m}\right[} \text{ such that } 1 \leq k \leq D_m\},$$

and $\mathcal{M}$ is the collection of regular partitions

$$m = \left\{\left[\frac{k-1}{D_m}, \frac{k}{D_m}\right[, \text{ such that } 1 \leq k \leq D_m, \right\},$$

where

$$D_m \leq \frac{n}{\log n}.$$

The collection of estimators is defined in Lemma 4.1. Let us thus consider four penalties.

— the AIC criretion defined by

$$\text{pen}_{\text{AIC}} = \frac{D_m}{n};$$

— the BIC criterion defined by

$$\text{pen}_{\text{BIC}} = \frac{\log n}{2n} D_m;$$

— the penalty proportional to the dimension as in (4.5) defined by

$$\text{pen}_{\text{lin}} = c \times \frac{D_m}{n};$$

— and the penalty defined in (4.6) by

$$\text{pen} = \mu \times \text{pen}_{\text{shape}}(m).$$

pen_{lin} and pen are penalties depending on some unknown multiplicative constant (c and μ respectively) to be calibrated. As previously said we will use the "slope heuristics" introduced in Birgéa nd Massart (2007) to calibrate the multiplicative constant. We have distinguished two cases :

— The case where there exists $m_o \in \mathcal{M}$ such that the true function belong to S_{m_o} *i.e.* where f_0 is piecewise constant,

$$\begin{aligned} \text{Mod1}: \ f_0 &= 0.5\mathbb{1}_{[0,1/3)} + \mathbb{1}_{[1/3,0.5)} + 2\mathbb{1}_{[0.5,2/3)} + 0.25\mathbb{1}_{[2/3,1]} \\ \text{Mod2}: \ f_0 &= 0.75\mathbb{1}_{[0,1/4]} + 0.5\mathbb{1}_{[1/4,0.5)} + 0.2\mathbb{1}_{[0.5,3/4)} + 0.3\mathbb{1}_{[3/4,1]}. \end{aligned}$$

— The second case, f_0 does not belong to any S_m, $m \in \mathcal{M}$ and is chosen in the following way :

$$\begin{aligned} \text{Mod3}: \ f_0(x) &= \sin(\pi x) \\ \text{Mod4}: \ f_0(x) &= \sqrt{x}. \end{aligned}$$

In each case, the x_i's are simulated according to uniform distribution on $[0, 1]$.

The Kullback-Leibler divergence is definitely not suitable to evaluate the quality of an estimator. Indeed, given a model S_m, there is a positive probability that on one of the interval $I \in m$ we have $\pi^{(I)}_{\hat{f}_m} = 0$ or $\pi^{(I)}_{\hat{f}_m} = 1$, which implies that $\mathcal{K}(\pi^{(n)}_{f_0}, \pi^{(n)}_{\hat{f}_m}) = +\infty$. So we will use the Hellinger distance to evaluate the quality of an estimator.

Even if an oracle inequality seems of no practical use, it can serve as a benchmark to evaluate the performance of any data driven selection procedure. Thus model selection performance of each procedure is evaluated by the following benchmark

$$C^* := \frac{\mathbb{E}\left[h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_{\hat{m}}})\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}} h^2(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{\hat{f}_m})\right]}. \tag{4.8}$$

C^* evaluate how far is the selected estimator to the oracle. The values of C^* evaluated for each procedure with different sample size $n \in \{100, 200, \ldots, 1000\}$ are reported in Figure 4.2 , Figure 4.4, Figure 4.3 and Figure 4.5. For each sample size $n \in \{100, 200, \ldots, 1000\}$, the expectation was estimated using mean over 1000 simulated datasets.

4.5.2 Slope heuristics

The aim of this section is to show how the penalty in Theorem 4.1 can be calibrated in practice using the main ideas of data-driven penalized model selection criterion proposed by Birgé and Massart (2007). We calibrate penalty using "slope heuristics" first introduced and theoretically validated by Birgé and Massart (2007) in a gaussian homoscedastic setting. Recently it has also been theoretically validated in the heteroscedastic random-design case by Arlot (2009) and for least squares density estimation by Lerasle (2012). Several encouraging applications of this method are developed in many other frameworks (see for instance in clustering and variable selection for categorical multivariate data Bontemps and Toussile (2013), for variable selection and clustering via Gaussian mixtures Maugis and Michel (2011), in multiple change points detection Lebarbier (2005)). Some overview and implementation of the slope heuristics can be find in Baudry *et al.* (2012).

We now describe the main idea of those heuristics, starting from that main goal of the model selection, that is to choose the best estimator of f_0 among a collection of estimators $\{\hat{f}_m\}_{m\in\mathcal{M}}$. Moreover, we expect that this best estimator mimics the so-called oracle defined as (4.7). To this aim, the great challenge is to build a penalty function such that the selected model $\hat{m}$ is nearly as good as the oracle. In the following we call the ideal penalty the penalty that leads to the choice of $m*$. Using that

$$\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)}) = \gamma(\hat{f}_m) - \gamma(f_0),$$

then, by definition, $m*$ defined in (4.7) satisfies

$$m* = \arg\min_{m\in\mathcal{M}} [\gamma(\hat{f}_m) - \gamma(f_0)] = \arg\min_{m\in\mathcal{M}} \gamma(\hat{f}_m).$$

The ideal penalty, leading to the choice of the oracle $m*$, is thus $[\gamma(\hat{f}_m) - \gamma_n(\hat{f}_m)]$, for $m \in \mathcal{M}$. As the matter of fact, by replacing $\text{pen}_{id}(\hat{f}_m)$ by its value, we obtain

$$\begin{aligned}
\arg\min_{m\in\mathcal{M}} [\gamma_n(\hat{f}_m) + \text{pen}_{id}(\hat{f}_m)] &= \arg\min_{m\in\mathcal{M}} [\gamma_n(\hat{f}_m) + \gamma(\hat{f}_m) - \gamma_n(\hat{f}_m)] \\
&= \arg\min_{m\in\mathcal{M}} [\gamma(\hat{f}_m)] \\
&= m*.
\end{aligned}$$

Of course this ideal penalty always selects the oracle model but depends on the unknown function f_0 throught the sample distribution, since $\gamma(t) = \mathbb{E}_{f_0}[\gamma_n(t)]$. A natural idea is to choose $\text{pen}(m)$ as close as possible to $\text{pen}_{id}(m)$ for every $m \in \mathcal{M}$. Now, we use that this ideal penalty can be decomposed into

$$\text{pen}_{id}(m) = \gamma(\hat{f}_m) - \gamma_n(\hat{f}_m) = v_m + \hat{v}_m + e_m,$$

where

$$v_m = \gamma(\hat{f}_m) - \gamma(f_m), \quad \hat{v}_m = \gamma_n(f_m) - \gamma_n(\hat{f}_m), \text{ and } \quad e_m = \gamma(f_m) - \gamma_n(f_m).$$

The slope heuristics relies on two points :

— The existence of a minimal penalty $\text{pen}_{\min}(m) = \hat{v}_m$ such that when the penalty is smaller than $\text{pen}_{\min}$ the selected model is one of the most complex models. Whereas, penalties larger than $\text{pen}_{\min}$ lead to a selection of models with "reasonable" complexity.

— Using concentration arguments, it is reasonable to consider that uniformly over $\mathcal{M}$, $\gamma_n(f_m)$ is close to its expectation which implies that $e_m \approx 0$. In the same way, since $\hat{v}_m$ is a empirical version of v_m, it is also reasonable to consider that $v_m \approx \hat{v}_m$. Ideal penalty is thus approximately given by $2\hat{v}_m$, and thus

$$\text{pen}_{id}(m) \approx 2\text{pen}_{min}(m).$$

In practice, $\hat{v}_m$ can be estimated from the data provided that ideal penalty $\text{pen}_{id}(.) = \kappa_{id}\text{pen}_{shape}(.)$ is known up to a multiplicative factor. A major point of the slope heuristics is that

$$\frac{\kappa_{id}}{2}\text{pen}_{shape}(.)$$

is a good estimator of $\hat{v}_m$ and this provides the minimal penalty.

Provided that $\text{pen} = \kappa \times \text{pen}_{shape}$ is known up to a multiplicative constant κ that is to be calibrated, we combine the previously heuristic to the method usually known as dimension jump method. In practice, we consider a grid $\kappa_1, \ldots, \kappa_M$, where each κ_j leads to a selected model $\hat{m}_{\kappa_i}$ with dimension $D_{\hat{m}_{\kappa_i}}$. The constant κ_{min} which corresponds to the value such that $\text{pen}_{min} = \kappa_{min} \times \text{pen}_{shape}$, is estimated using the first point of the "slope heuristics". If $D_{\hat{m}_{\kappa_j}}$ is plotted as a function of κ_j, κ_{min} is such that $D_{\hat{m}_{\kappa_j}}$ is "huge" for $\kappa < \kappa_{min}$ and "reasonably small" for $\kappa > \kappa_{min}$. So κ_{min} is the value at the position of the biggest jump. For more details about this method we refer the reader to Baudry *et al.* (2012) and Arlot and Massart (2009).

Figures 4.2 and 4.3 are the cases where the true function is piecewise constant. Figure 4.4 and Figure 4.5 are situations where the true function does not belong to any model in the given collection. The performance of criteria depends on the sample size n. In these two situations we observe that our two model selection procedures are comparable, and their performance increases with n. While the performance of model selected by BIC decreases with n. Our criteria outperformed the AIC for all n. The BIC criterion is better than our criteria for $n \leq 200$. For $200 < n \leq 400$, the performance of the model selected by BIC is quite the same as the performance of models selected by our criteria. Finally for $n > 400$ our criteria outperformed the BIC.

Theoretical results and simulations raise the following question : why our criteria are better than BIC for quite large values of n yet theoretical results are non asymptotic ? To answer this question we can say that, in simulations, to calibrate our penalties we have used "slope heuristics", and those heuristic are based on asymptotic arguments (see Section 4.5.2).

4.6 Proofs

Notations and technical tools

Subsequently we will use the following notations. Denote by $\| f \|_n$ and $\langle f, g \rangle_n$ the empirical euclidian norm and the inner product

$$\| f \|_n^2 = \frac{1}{n}\sum_{i=1}^{n} f^2(x_i), \text{ and } \langle f, g \rangle_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(x_i).$$

Note that $\| . \|_n$ is a semi norm on the space $\mathcal{F}$ of functions $g : \mathcal{X} \longrightarrow \mathbb{R}$, but is a norm in the quotient space $\mathcal{F}/\mathcal{R}$ associated to the equivalence relation $\mathcal{R}$:

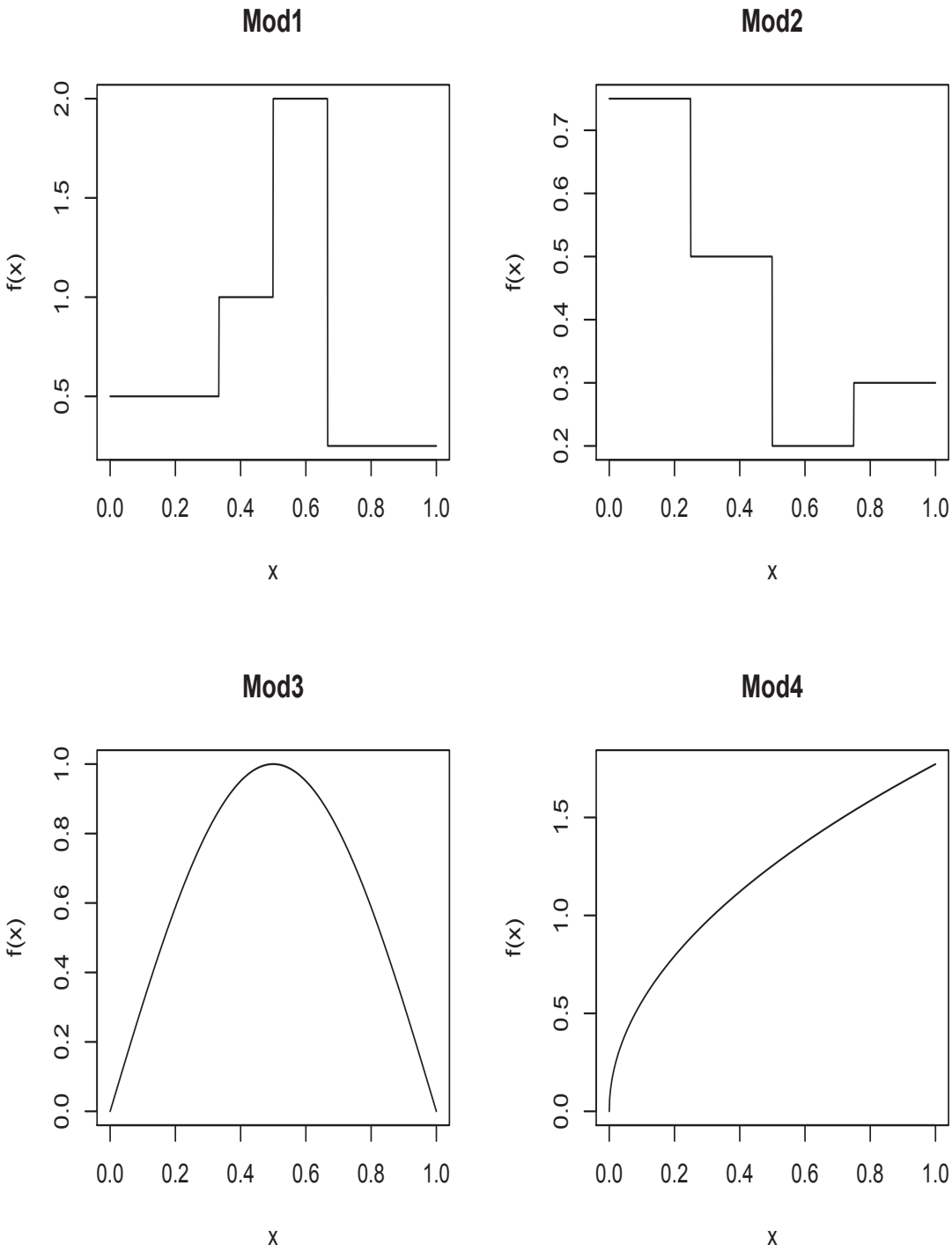


FIGURE 4.1 – *Different functions f_0 to be estimated*

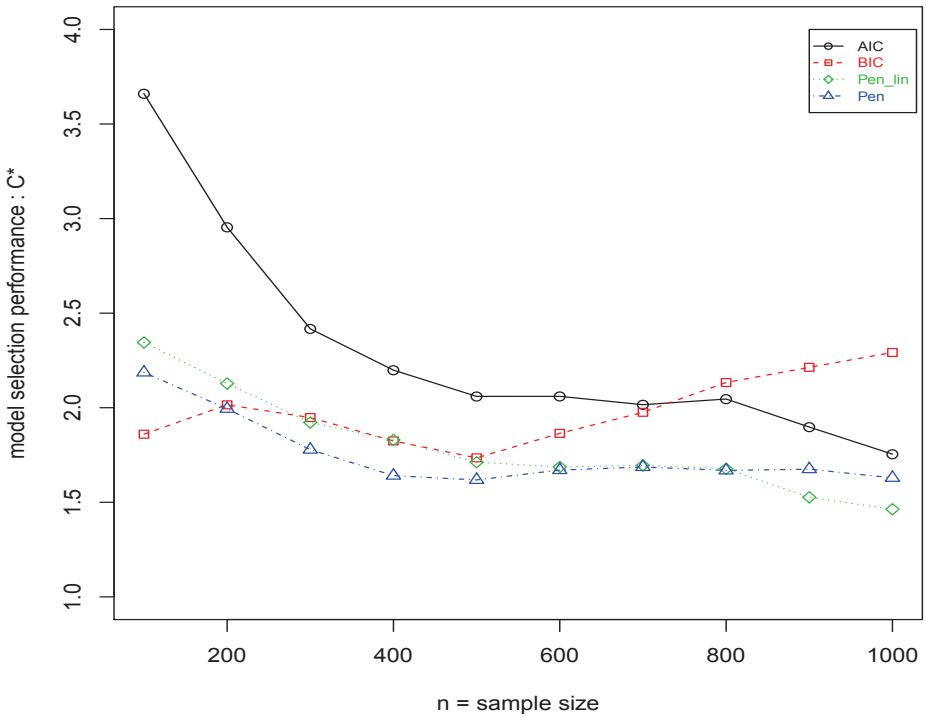


FIGURE 4.2 – *Model selection performance (C^*) as a function of sample size n, with each penalty, Mod1.*

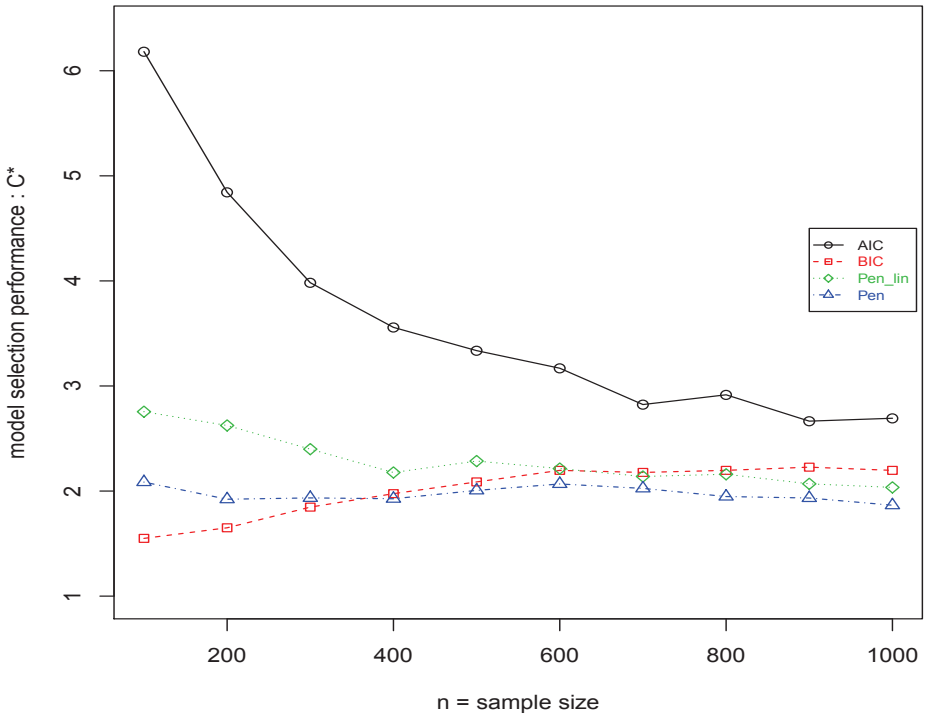


FIGURE 4.3 – *Model selection performance (C^*) as a function of sample size n, with each penalty, Mod2.*

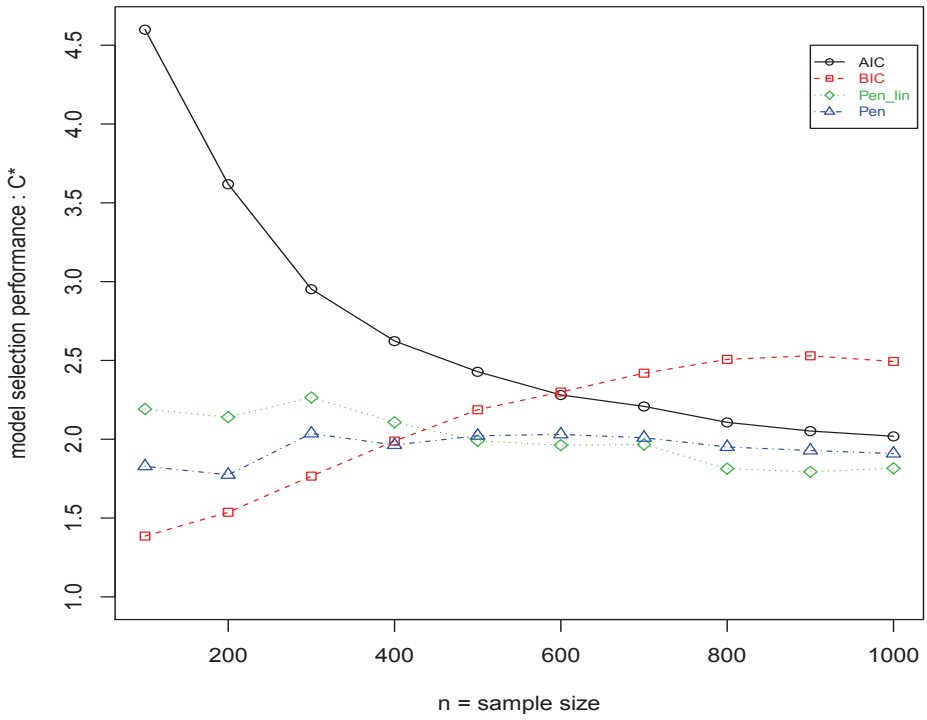


FIGURE 4.4 – *Model selection performance (C^*) as a function of sample size n, with each penalty, Mod3.*

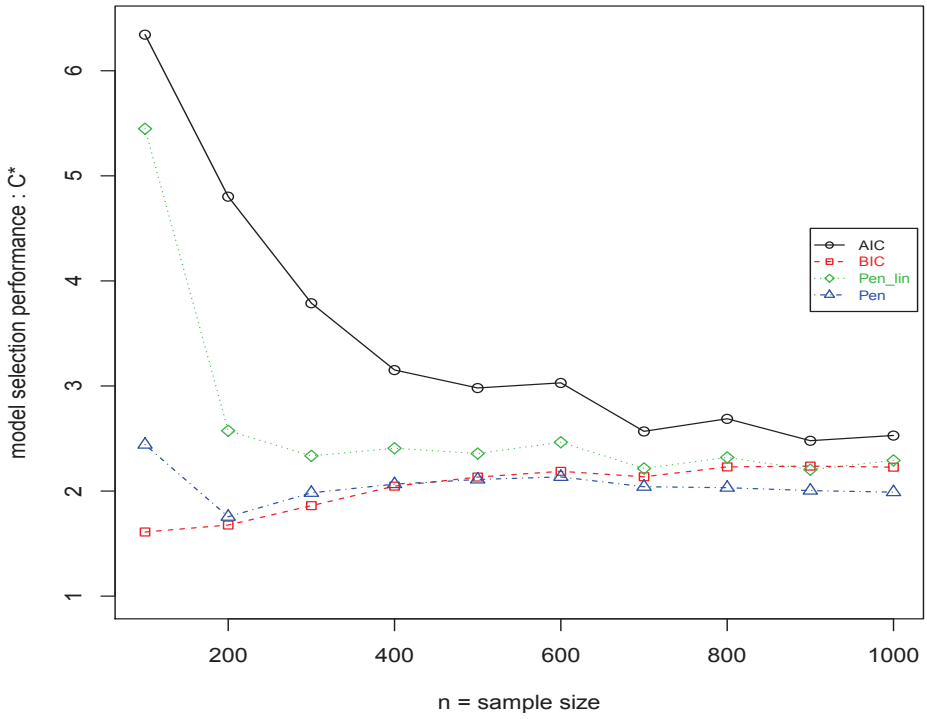


FIGURE 4.5 – *Model selection performance (C^*) as a function of sample size n, with each penalty, Mod4.*

$g \mathcal{R} h$ if and only if $g(x_i) = h(x_i)$ for all $i \in \{1, \dots, n\}$. It follows from (4.3) that γ defined in (4.4) can be expressed as the sum of a centered empirical process and of the estimation criterion γ_n. More precisely, denoting by $\vec{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$, with $\varepsilon_i = Y_i - \mathbb{E}_{f_0}(Y_i)$, for all f, we have

$$\gamma(f) = \gamma_n(f) + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i) := \gamma_n(f) + \langle \vec{\varepsilon}, f \rangle_n. \tag{4.1}$$

Easy calculations show that for γ defined in (4.4) we have,

$$\begin{aligned} \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_f^{(n)}) &= \frac{1}{n}\int \log\left(\frac{\mathbb{P}_{f_0}^{(n)}}{\mathbb{P}_f^{(n)}}\right) d\mathbb{P}_{f_0}^{(n)} = \gamma(f) - \gamma(f_0) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\pi_{f_0}(x_i)\log\left(\frac{\pi_{f_0}(x_i)}{\pi_f(x_i)}\right) + (1-\pi_{f_0}(x_i))\log\left(\frac{1-\pi_{f_0}(x_i)}{1-\pi_f(x_i)}\right)\right]. \end{aligned}$$

Let us recall the usual bounds (see Castellan (2003a)) for kullback-Leibler information :

Lemma 4.1 *For positive densities p and q with respect to μ, if $f = \log(q/p)$, then*

$$\frac{1}{2}\int f^2(1 \wedge e^f) p \, d\mu \leqslant \mathcal{K}(p,q) \leqslant \frac{1}{2}\int f^2(1 \vee e^f) p \, d\mu.$$

Proof of Proposition 4.1 :

By definition of $\hat{f}_m$, for all $f \in S_m \cap \mathbb{L}_\infty(C_0)$, $\gamma_n(\hat{f}_m) - \gamma_n(f) \leqslant 0$. We apply (4.1), with $f = f_m$ and $f = \hat{f}_m$,

$$\gamma(\hat{f}_m) - \gamma(f_0) \leqslant \gamma(f_m) - \gamma(f_0) + \langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n.$$

As usual, the main part of the proof relies on the study of the empirical process $\langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n$. Since $\hat{f}_m - f_m$ belongs to S_m, $\hat{f}_m - f_m = \sum_{j=1}^{D_m} \alpha_j \psi_j$, where $\{\psi_1, \dots, \psi_{D_m}\}$, is an orthonormal basis of S_m and consequently

$$\langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n = \sum_{j=1}^{D_m} \alpha_j \langle \vec{\varepsilon}, \psi_j \rangle_n.$$

Applying Cauchy-Schwarz inequality we get

$$\begin{aligned} \langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n &\leqslant \sqrt{\sum_{j=1}^{D_m} \alpha_j^2} \sqrt{\sum_{j=1}^{D_m} \left(\langle \vec{\varepsilon}, \psi_j \rangle_n\right)^2} \\ &= \|\hat{f}_m - f_m\|_n \sqrt{\sum_{j=1}^{D_m}\left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \psi_j(x_i)\right)^2}. \end{aligned}$$

We now apply Lemma 4.2 (See Section 4.6 for the proof of Lemma 4.2)

Lemma 4.2 *Let $\mathcal{S}_m$ the model defined in (4.5) and $\{\psi_1, \dots, \psi_{D_m}\}$ an orthonormal basis of the linear span $\{\phi_k,\ k \in m\}$. We also denote by Λ_m the set of $\beta = (\beta_1, ..., \beta_D)$ such that $f_\beta(.) =$*

$\sum_{j=1}^{D} \beta_j \psi_j(.)$ *satisfies* $f_\beta \in \mathcal{S}_m \cap \mathbb{L}_\infty(C_0)$. *Let* β^* *be any minimizer of the function* $\beta \to \gamma(f_\beta)$ *over* Λ_m, *we have*

$$\frac{\mathcal{U}_0^2}{2} \| f_\beta - f_{\beta^*} \|_n^2 \leq \gamma(f_\beta) - \gamma(f_{\beta^*}), \tag{4.2}$$

where $\mathcal{U}_0 = e^{C_0}/(1+e^{C_0})^2$.

Then we have

$$\langle \vec{\varepsilon}, \hat{f}_m - f_m \rangle_n \quad \leqslant \quad \sqrt{\sum_{j=1}^{D_m} \left(\langle \vec{\varepsilon}, \psi_j \rangle_n \right)^2} \frac{\sqrt{2}}{\mathcal{U}_0} \sqrt{\gamma(\hat{f}_m) - \gamma(f_m)}$$

Now we use that for every positive numbers, a, b, x, $ab \leqslant (x/2)a^2 + [1/(2x)]b^2$, and infer that

$$\gamma(\hat{f}_m) - \gamma(f_0) \leq \gamma(f_m) - \gamma(f_0) + \frac{x}{\mathcal{U}_0^2} \sum_{j=1}^{D_m} \left(\langle \vec{\varepsilon}, \psi_j \rangle_n \right)^2 + (1/2x)(\gamma(\hat{f}_m) - \gamma(f_m)).$$

For $x > 1/2$, it follows that

$$\mathbb{E}_{f_0}[\gamma(\hat{f}_m) - \gamma(f_0)] \leqslant \gamma(f_m) - \gamma(f_0) + \frac{2x^2}{(2x-1)\mathcal{U}_0^2} \mathbb{E}_{f_0} \left[\sum_{j=1}^{D_m} \left(\langle \vec{\varepsilon}, \psi_j \rangle_n \right)^2 \right].$$

We conclude the proof by using that

$$\mathbb{E}_{f_0} \left[\sum_{j=1}^{D_m} \left(\langle \vec{\varepsilon}, \psi_j \rangle_n \right)^2 \right] \leqslant \frac{D_m}{4n}.$$

□

Proof of Theorem 4.1

By definition, for all $m \in \mathcal{M}$,

$$\gamma_n(\hat{f}_{\hat{m}}) + \text{pen}(\hat{m}) \leqslant \gamma_n(\hat{f}_m) + \text{pen}(m) \leqslant \gamma_n(f_m) + \text{pen}(m).$$

Applying (4.1) we have

$$\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) \leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n + \text{pen}(m) - \text{pen}(\hat{m}). \tag{4.3}$$

It remains to study $\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n$, using the following lemma, which is a modification of Lemma 1 in Durot *et al.* (2009).

Lemma 4.3 *For every* D, D' *and* $x \geqslant 0$ *we have*

$$\mathbb{P} \left(\sup_{u \in \left(S_D \cap \mathbb{L}_\infty(C_0) + S_{D'} \cap \mathbb{L}_\infty(C_0) \right)} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n} \rangle \sqrt{\frac{D+D'}{4n}} + \sqrt{\frac{5x}{n}} \right) \leqslant \exp(-x).$$

Fix $\xi > 0$ and let $\Omega_\xi(m)$ denote the event

$$\Omega_\xi(m) = \bigcap_{m' \in \mathcal{M}} \left\{ \sup_{u \in \left(S_m \cap \mathbb{L}_\infty(C_0) + S_{m'} \cap \mathbb{L}_\infty(C_0) \right)} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n} \leq \sqrt{\frac{D_m + D_{m'}}{4n}} + \sqrt{5(L_{m'} D_{m'} + \xi)/n} \right\}.$$

Then we have

$$\mathbb{P}\left(\Omega_\xi(m)\right) \geqslant 1 - \Sigma \exp(-\xi). \tag{4.4}$$

See the Appendix for the proof of this lemma. Fix $\xi > 0$, applying Lemma 4.3, we infer that on the event $\Omega_\xi(m)$,

$$\begin{aligned}
\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n &\leqslant \left(\sqrt{\frac{D_m + D_{\hat{m}}}{4n}} + \sqrt{5\frac{L_{\hat{m}} D_{\hat{m}} + \xi}{n}} \right) \| \hat{f}_{\hat{m}} - f_m \|_n \\
&\leqslant \left(\sqrt{\frac{D_m + D_{\hat{m}}}{4n}} + \sqrt{5\frac{L_{\hat{m}} D_{\hat{m}} + \xi}{n}} \right) \left(\| \hat{f}_{\hat{m}} - f_0 \|_n + \| f_0 - f_m \|_n \right) \\
&\leqslant \left(\sqrt{D_{\hat{m}}} \left(\frac{1}{\sqrt{4n}} + \sqrt{\frac{5L_{\hat{m}}}{n}} \right) + \sqrt{\frac{D_m}{4n}} + \sqrt{5\frac{\xi}{n}} \right) \left(\| \hat{f}_{\hat{m}} - f_0 \|_n + \| f_0 - f_m \|_n \right).
\end{aligned}$$

Applying that $2xy \leqslant \theta x^2 + \theta^{-1} y^2$, for all $x > 0$, $y > 0$, $\theta > 0$, we get that on $\Omega_\xi(m)$ and for every $\eta \in]0,1[$

$$\begin{aligned}
\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n &\leqslant \left(\frac{1-\eta}{2}\right) \left[(1+\eta) \| \hat{f}_{\hat{m}} - f_0 \|_n^2 + (1+\eta^{-1}) \| f_0 - f_m \|_n^2 \right] \\
&+ \frac{1}{2(1-\eta)} \left[(1+\eta) D_{\hat{m}} \left(\frac{1}{\sqrt{4n}} + \sqrt{\frac{5L_{\hat{m}}}{n}} \right)^2 + (1+\eta^{-1}) \left(\sqrt{\frac{D_m}{4n}} + \sqrt{\frac{5\xi}{n}} \right)^2 \right] \\
&\leqslant \frac{1-\eta^2}{2} \| \hat{f}_{\hat{m}} - f_0 \|_n^2 + \frac{\eta^{-1} - \eta}{2} \| f_0 - f_m \|_n^2 + \frac{1+\eta}{2(1-\eta)} D_{\hat{m}} \left(\frac{1}{\sqrt{4n}} + \sqrt{\frac{5L_{\hat{m}}}{n}} \right)^2 \\
&\quad + \frac{1+\eta^{-1}}{1-\eta} \left(\frac{D_m}{4n} + \frac{5\xi}{n} \right).
\end{aligned}$$

If $\mathrm{pen}(m) \geqslant \left(\lambda D_m \left(\frac{1}{2} + \sqrt{5L_m} \right)^2 \right) / n$, with $\lambda > 0$, we have

$$\begin{aligned}
\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n &\leqslant \frac{1-\eta^2}{2} \| \hat{f}_{\hat{m}} - f_0 \|_n^2 + \frac{\eta^{-1} - \eta}{2} \| f_0 - f_m \|_n^2 + \frac{1+\eta}{2(1-\eta)\lambda} \mathrm{pen}(\hat{m}) \\
&\quad + \frac{1+\eta^{-1}}{(1-\eta)\lambda} \mathrm{pen(m)} + \frac{1+\eta^{-1}}{1-\eta} \frac{5\xi}{n}.
\end{aligned}$$

It follows from (4.3) that

$$\begin{aligned}
\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) &\leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \frac{1-\eta^2}{2} \| \hat{f}_{\hat{m}} - f_0 \|_n^2 + \frac{\eta^{-1} - \eta}{2} \| f_0 - f_m \|_n^2 \\
&\quad + \frac{1+\eta}{2(1-\eta)\lambda} \mathrm{pen}(\hat{m}) + \frac{1+\eta^{-1}}{(1-\eta)\lambda} \mathrm{pen(m)} + \frac{1+\eta^{-1}}{1-\eta} \frac{5\xi}{n} + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}).
\end{aligned}$$

Taking $\lambda = (\eta+1)/(2(1-\eta))$, we have

$$\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) \leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \frac{4\lambda}{(2\lambda+1)^2} \parallel \hat{f}_{\hat{m}} - f_0 \parallel_n^2 + \frac{4\lambda}{4\lambda^2-1} \parallel f_0 - f_m \parallel_n^2 + \frac{6\lambda+1}{2\lambda-1}\mathrm{pen}(m) + \frac{10\lambda(2\lambda+1)}{2\lambda-1}\frac{\xi}{n}.$$

Now we use the following lemma (see Lemma 6.1 in Kwemou (2012)) that allows to connect empirical norm and Kullback-Leibler divergence.

Lemma 4.4 *Under Assumptions (**A**$_5$), for all $m \in \mathcal{M}$ and all $t \in S_m \cap \mathbb{L}_\infty(C_0)$, we have*

$$c_{min}\|t - f_0\|_n^2 \leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_t^{(n)}) \leqslant c_{max}\|t - f_0\|_n^2.$$

where c_{min} and c_{max} are constants depending on C_0 and c_1.

Consequently

$$\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) \leqslant C(c_{min})\left\{\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m)\right\} + C_1(c_{min})\frac{\xi}{n},$$

where

$$C(c_{min}) = \max\left\{\frac{1+\frac{4\lambda}{(4\lambda^2-1)c_{min}}}{1-\frac{4\lambda}{c_{min}(2\lambda+1)^2}}; \frac{\frac{6\lambda+1}{2\lambda-1}}{1-\frac{4\lambda}{c_{min}(2\lambda+1)^2}}\right\} \text{ and } C_1(c_{min}) = \frac{\frac{10\lambda(2\lambda+1)}{2\lambda-1}}{1-\frac{4\lambda}{c_{min}(2\lambda+1)^2}}.$$

Thus we take λ such that

$$1 - \frac{4\lambda}{c_{min}(2\lambda+1)^2} > 0, \tag{4.5}$$

where c_{min} depends on the bound of the true function f_0. By definition of $\Omega_\xi(m)$ and (4.4), there exists a random variable $V \geqslant 0$ with $\mathbb{P}(V > \xi) \leqslant \Sigma \exp(-\xi)$ and $\mathbb{E}_{f_0}(V) \leqslant \Sigma$, such that

$$\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) \leqslant C(c_{min})\left\{\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m)\right\} + C_1(c_{min})\frac{V}{n},$$

which implies that for all $m \in \mathcal{M}$,

$$\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})] \leqslant C(c_{min})\left\{\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m)\right\} + C_1(c_{min})\frac{\Sigma}{n}.$$

This concludes the proof. □

Proof of Proposition 4.2 :

Let f_m, $\hat{f}_m$, π_{f_m} and $\pi_{\hat{f}_m}$ given in Lemma 4.1, proved in appendix. In the following, $D_m = |m|$. For $\delta > 0$, let $\Omega_m(\delta)$ be the event

$$\Omega_m(\delta) = \bigcap_{J \in m}\left\{\left|\frac{\pi_{\hat{f}_m}^{(J)}}{\pi_{f_m}^{(J)}} - 1\right| \leqslant \delta\right\} \bigcap \left\{\left|\frac{1-\pi_{\hat{f}_m}^{(J)}}{1-\pi_{f_m}^{(J)}} - 1\right| \leqslant \delta\right\}. \tag{4.6}$$

According to pythagore's type identity and Lemma 4.1 we write

$$\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) = \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{f_m}^{(n)})\mathbb{1}_{\Omega_m(\delta)} + \mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{f_m}^{(n)})\mathbb{1}_{\Omega_m^c(\delta)},$$

where

$$\mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)}) = \frac{1}{n}\sum_{i=1}^{n}\left[\pi_{f_m}(x_i)\log\left(\frac{\pi_{f_m}(x_i)}{\pi_{\hat{f}_m}(x_i)}\right) + (1-\pi_{f_m}(x_i))\log\left(\frac{1-\pi_{f_m}(x_i)}{1-\pi_{\hat{f}_m}(x_i)}\right)\right] \quad (4.7)$$
$$= \frac{1}{n}\sum_{J\in m}|J|\left[\pi_{f_m}^{(J)}\log\left(\frac{\pi_{f_m}^{(J)}}{\pi_{\hat{f}_m}^{(J)}}\right) + (1-\pi_{f_m}^{(J)})\log\left(\frac{1-\pi_{f_m}^{(J)}}{1-\pi_{\hat{f}_m}^{(J)}}\right)\right].$$

The first step consists in showing that

$$\frac{1-\delta}{2(1+\delta)^2}\mathcal{X}_m^2\mathbb{1}_{\Omega_m(\delta)} \leqslant \mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)})\mathbb{1}_{\Omega_m(\delta)} \leqslant \frac{1+\delta}{2(1-\delta)^2}\mathcal{X}_m^2\mathbb{1}_{\Omega_m(\delta)}, \quad (4.8)$$

where

$$\mathcal{X}_m^2 = \frac{1}{n}\sum_{J\in m}\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|\pi_{f_m}^{(J)}[1-\pi_{f_m}^{(J)}]}, \text{ with } \quad \frac{4\rho^2 D_m}{n} \leqslant \mathbb{E}_{f_0}[\mathcal{X}_m^2] \leqslant \frac{2D_m}{n}. \quad (4.9)$$

The second step relies on the proof of

$$\left|\mathbb{E}_{f_0}\left(\mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)})\mathbb{1}_{\Omega_m^c(\delta)}\right)\right| \leqslant 2\log\left(\frac{1}{\rho}\right)\mathbb{P}[\Omega_m^c(\delta)]. \quad (4.10)$$

The last step consists in showing that for $\epsilon > 0$, since for all $J \in m$, $|J| \geq \Gamma[\log(n)]^2$, where $\Gamma > 0$ is an absolute constant, then we have

$$\mathbb{P}[\Omega_m^c(\delta)] \leqslant 4|m|\exp\left(-\frac{\delta^2}{2(1+\delta/3)}\rho^2\Gamma[\log(n)]^2\right) \leq \frac{\kappa(\rho,\delta,\Gamma,\epsilon)}{n^{(1+\epsilon)}}. \quad (4.11)$$

Gathering (4.8)-(4.11), we conclude that

$$\begin{aligned}\mathbb{E}_{f_0}[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)})] &\leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \frac{(1+\delta)|m|}{(1-\delta)^2 n} + 2\log\left(\frac{1}{\rho}\right)\mathbb{P}[\Omega_m^c(\delta)] \\ &\leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \frac{(1+\delta)|m|}{(1-\delta)^2 n} + \frac{\kappa(\rho,\delta,\Gamma,\epsilon)}{n^{(1+\epsilon)}}.\end{aligned}$$

We finish by proving (4.8), (4.9), (4.10) and (4.11).

• **Proof of (4.8) and (4.9) :** Arguing as in Castellan (2003a) and using Lemma 4.1 we have

$$\mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)}) \geqslant \frac{1}{2n}\sum_{J\in m}|J|\left[\pi_{f_m}^{(J)}\left(1\wedge\frac{\pi_{\hat{f}_m}^{(J)}}{\pi_{f_m}^{(J)}}\right)\log^2\left(\frac{\pi_{f_m}^{(J)}}{\pi_{\hat{f}_m}^{(J)}}\right) + (1-\pi_{f_m}^{(J)})\left(1\wedge\frac{1-\pi_{\hat{f}_m}^{(J)}}{1-\pi_{f_m}^{(J)}}\right)\log^2\left(\frac{1-\pi_{f_m}^{(J)}}{1-\pi_{\hat{f}_m}^{(J)}}\right)\right]$$

and

$$\mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)}) \leqslant \frac{1}{2n}\sum_{J\in m}|J|\left[\pi_{f_m}^{(J)}\left(1\vee\frac{\pi_{\hat{f}_m}^{(J)}}{\pi_{f_m}^{(J)}}\right)\log^2\left(\frac{\pi_{f_m}^{(J)}}{\pi_{\hat{f}_m}^{(J)}}\right) + (1-\pi_{f_m}^{(J)})\left(1\vee\frac{1-\pi_{\hat{f}_m}^{(J)}}{1-\pi_{f_m}^{(J)}}\right)\log^2\left(\frac{1-\pi_{f_m}^{(J)}}{1-\pi_{\hat{f}_m}^{(J)}}\right)\right].$$

It follows that

$$\frac{1-\delta}{2}V^2(\pi_{f_m}, \pi_{\hat{f}_m})\mathbb{1}_{\Omega_m(\delta)} \leqslant \mathcal{K}(\mathbb{P}_{f_m}^{(n)}, \mathbb{P}_{\hat{f}_m}^{(n)})\mathbb{1}_{\Omega_m(\delta)} \leqslant \frac{1+\delta}{2}V^2(\pi_{f_m}, \pi_{\hat{f}_m})\mathbb{1}_{\Omega_m(\delta)}, \quad (4.12)$$

where $V^2(\pi_{f_m}, \pi_{\hat{f}_m})$ is defined by

$$V^2(\pi_{f_m}, \pi_{\hat{f}_m}) = \frac{1}{n}\sum_{J\in m}|J|\frac{[\pi^{(J)}_{\hat{f}_m} - \pi^{(J)}_{f_m}]^2}{\pi^{(J)}_{f_m}}\left[\frac{\log[\pi^{(J)}_{\hat{f}_m}/\pi^{(J)}_{f_m}]}{\pi^{(J)}_{\hat{f}_m}/\pi^{(J)}_{f_m} - 1}\right]^2 + \frac{1}{n}\sum_{J\in m}|J|\frac{[\pi^{(J)}_{\hat{f}_m} - \pi^{(J)}_{f_m}]^2}{1-\pi^{(J)}_{f_m}}\left[\frac{\log[(1-\pi^{(J)}_{\hat{f}_m})/(1-\pi^{(J)}_{f_m})]}{(1-\pi^{(J)}_{\hat{f}_m})/(1-\pi^{(J)}_{f_m}) - 1}\right]^2. \quad (4.13)$$

Now we use that, for all $x > 0$,

$$\frac{1}{1\vee x} \leqslant \frac{\log(x)}{x-1} \leqslant \frac{1}{1\wedge x}. \quad (4.14)$$

Hence we infer that

$$\frac{1}{(1+\delta)^2}\mathcal{X}^2_m \mathbb{1}_{\Omega_m(\delta)} \leqslant V^2(\pi_{f_m}, \pi_{\hat{f}_m})\mathbb{1}_{\Omega_m(\delta)} \leqslant \frac{1}{(1-\delta)^2}\mathcal{X}^2_m \mathbb{1}_{\Omega_m(\delta)},$$

with $\mathcal{X}^2_m$ defined in (4.9). This entails that (4.8) is proved. It remains now to check that

$$\frac{4\rho^2|m|}{n} \leqslant \mathbb{E}_{f_0}[\mathcal{X}^2_m] \leqslant \frac{2|m|}{n}.$$

According to Lemma 4.1 , for all partition $J \in m$ and for any $x_i \in J$,

$$\pi_{\hat{f}_m}(x_i) = \pi^{(J)}_{\hat{f}_m}, \qquad \text{with} \qquad \pi^{(J)}_{\hat{f}_m} = \frac{1}{|J|}\sum_{i\in J} Y_i,$$

$$\text{and } \pi_{f_m}(x_i) = \pi^{(J)}_{f_m}, \qquad \text{with} \qquad \pi^{(J)}_{f_m} = \frac{1}{|J|}\sum_{i\in J}\pi_{f_0}(x_i).$$

Consequently,

$$\mathcal{X}^2_m = \frac{1}{n}\sum_{J\in m}|J|\frac{(\sum_{k\in J}\varepsilon_k)^2}{\sum_{k\in J}\pi_{f_0}(x_k)[|J| - \sum_{k\in J}\pi_{f_0}(x_k)]} = \frac{1}{n}\sum_{J\in m}\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|\pi^{(J)}_{f_m}[1-\pi^{(J)}_{f_m}]},$$

and finally

$$\mathbb{E}_{f_0}(\mathcal{X}^2_m) = \frac{1}{n}\sum_{J\in m}\mathbb{E}\left(\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|\pi^{(J)}_{f_m}[1-\pi^{(J)}_{f_m}]}\right) = \frac{1}{n}\sum_{J\in m}\left(\frac{1}{|J|\pi^{(J)}_{f_m}[1-\pi^{(J)}_{f_m}]}\right)\sum_{k\in J}\mathrm{Var}\,(Y_k).$$

Consequently

$$\mathbb{E}_{f_0}(\mathcal{X}^2_m) = \frac{1}{n}\sum_{J\in m}\frac{\sum_{i\in J}\pi_{f_0}(x_i)(1-\pi_{f_0}(x_i))}{|J|\pi^{(J)}_{f_m}[1-\pi^{(J)}_{f_m}]}.$$

Now, according to Assumption ($\mathbf{A_6}$), and Lemma 4.1, for all partition m, all $J \in m$, and all $x_i \in J$

$$0 < \rho^2 \leqslant \pi_{f_0}(x_i)(1-\pi_{f_0}(x_i)) \leqslant 1/4, \text{ and } 0 < \rho \leqslant \pi^{(J)}_{f_m} \text{ and } 0 < \rho \leqslant (1-\pi^{(J)}_{f_m}).$$

It follows that

$$4\rho^2 \leqslant \frac{\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))}{|J|\pi^{(J)}_{f_m}[1-\pi^{(J)}_{f_m}]} = \frac{\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))}{|J|\pi^{(J)}_{f_m}} + \frac{\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))}{|J|[1-\pi^{(J)}_{f_m}]} \leqslant 2,$$

and thus

$$\frac{4\rho^2|m|}{n} \leqslant \frac{1}{n}\sum_{J\in m}\frac{\sum_{i\in J}\pi_{f_0}(x_i)(1-\pi_{f_0}(x_i))}{|J|\pi^{(J)}_{f_m}[1-\pi^{(J)}_{f_m}]} \leqslant \frac{2|m|}{n}.$$

In other words,

$$\frac{4\rho^2|m|}{n} \leqslant \mathbb{E}_{f_0}(\mathcal{X}^2_m) \leqslant \frac{2|m|}{n}.$$

The ends up the proof of (4.8) and (4.9).

• **Proof of (4.10) :** We start from (4.7), apply Assumption ($\mathbf{A_6}$) and Lemma 4.1, to obtain that and (4.10) is checked since

$$\begin{aligned}
|\mathbb{E}\left(\mathcal{K}(\mathbb{P}^{(n)}_{f_m},\mathbb{P}^{(n)}_{\hat{f}_m})\mathbb{1}_{\Omega^c_m(\delta)}\right)| &\leqslant \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left|\left[\log\left(\frac{\pi_{f_m}(x_i)}{\pi_{\hat{f}_m}(x_i)}\right)\mathbb{1}_{\Omega^c_m(\delta)}\right]\right| \\
&+ \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left|\left[\log\left(\frac{(1-\pi_{f_m}(x_i))}{(1-\pi_{\hat{f}_m}(x_i))}\right)\mathbb{1}_{\Omega^c_m(\delta)}\right]\right| \\
&\leqslant 2\log\left(\frac{1}{\rho}\right)\mathbb{P}[\Omega^c_m(\delta)].
\end{aligned}$$

• **Proof of (4.11) :** We come to the control of $\mathbb{P}_{f_0}[\Omega^c_m(\delta)]$. Since

$$\mathbb{P}[\Omega^c_m(\delta)] \leqslant \sum_{J\in m}\mathbb{P}\left\{\left|\frac{\pi^{(J)}_{\hat{f}_m}}{\pi^{(J)}_{f_m}}-1\right|\geqslant\delta\right\}+\sum_{J\in m}\mathbb{P}\left\{\left|\frac{1-\pi^{(J)}_{\hat{f}_m}}{1-\pi^{(J)}_{f_m}}-1\right|\geqslant\delta\right\},$$

by applying Lemma 4.1, we infer that

$$\mathbb{P}\left\{\left|\frac{\pi^{(J)}_{\hat{f}_m}}{\pi^{(J)}_{f_m}}-1\right|\geqslant\delta\right\} = \mathbb{P}\left\{\left|\frac{\sum_{k\in J}\varepsilon_k}{\sum_{k\in J}\pi_{f_0}(x_k)}\right|\geqslant\delta\right\} = \mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)\right\},$$

and

$$\mathbb{P}\left\{\left|\frac{1-\pi^{(J)}_{\hat{f}_m}}{1-\pi^{(J)}_{f_m}}-1\right|\geqslant\delta\right\} = \mathbb{P}\left\{\left|\frac{\sum_{k\in J}\varepsilon_k}{\sum_{k\in J}(1-\pi_{f_0}(x_k))}\right|\geqslant\delta\right\} = \mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}(1-\pi_{f_0}(x_k))\right\}.$$

We write

$$\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)\right\} \leqslant \mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\}$$

and

$$\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}(1-\pi_{f_0}(x_k))\right\} \leqslant \mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\}.$$

Then we have

$$\mathbb{P}[\Omega^c_m(\delta)] \leqslant 2\sum_{J\in m}\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right|\geqslant\delta\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\}.$$

Now, we apply Bernstein Concentration Inequality (see Massart (2007) for example) to the right hand side of previous inequality, starting by recalling this Bernstein inequality.

Theorem 4.1 *Let $Z_1, \cdots, Z_n$ be independent real valued random variables. Assume that there exist some positive numbers v and c such that for all $k \geqslant 2$,*

$$\sum_{i=1}^{n} \mathbb{E}\left[|Z_i|^k\right] \leqslant \frac{k!}{2} v c^{k-2}.$$

Then for any positive z,

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}(Z_i) \geqslant \sqrt{2vz} + cz\right) \leqslant \exp(-z), \text{ and } \mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}(Z_i) \geqslant z\right) \leqslant \exp\left(-\frac{z^2}{2(v+cz)}\right).$$

Especially, if $|Z_i| \leqslant b$ for all i, then

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}(Z_i) \geqslant z\right) \leqslant \exp\left(-\frac{z^2}{2(\sum_{i=1}^{n}\mathbb{E}(Z_i^2) + bz/3)}\right). \tag{4.15}$$

Applying (4.15) with $z = \delta \sum_{k\in J} \pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))$, $b = 1$ and $v = \sum_{k\in J} \pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))$, we get that

$$\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right| \geqslant \delta \sum_{k\in J} \pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\}$$

is less than

$$2\exp\left(-\frac{\delta^2[\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))]^2}{2\left(\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k)) + (\delta/3)\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right)}\right),$$

and consequently

$$\begin{aligned}\mathbb{P}\left\{\left|\sum_{k\in J}\varepsilon_k\right| \geqslant \delta \sum_{k\in J} \pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right\} &\leqslant 2\exp\left[-\frac{\delta^2}{2(1+\delta/3)}\left(\sum_{k\in J}\pi_{f_0}(x_k)(1-\pi_{f_0}(x_k))\right)\right] \\ &\leqslant 2\exp\left[-\frac{\delta^2}{2(1+\delta/3)}|J|\rho^2\right].\end{aligned}$$

Consequently,

$$\mathbb{P}[\Omega_m^c(\delta)] \leqslant 4|m|\exp(-\Delta\rho^2\Gamma[\log(n)]^2), \qquad \text{with} \qquad \Delta = \frac{\delta^2}{2(1+\delta/3)},$$

where Γ is given by (4.1). For $\epsilon > 0$ and δ such that

$$\frac{\delta^2}{2(1+\delta/3)}\rho^2\Gamma\log(n) \geqslant 2+\epsilon, \tag{4.16}$$

using that $|m| \leqslant n$ implies that

$$4|m|\exp\left(-\frac{\delta^2}{2(1+\delta/3)}\rho^2\Gamma[\log(n)]^2\right) \leqslant \frac{\kappa}{n^{(1+\epsilon)}}.$$

And Result (4.11) follows.

Proof of Theorem 4.1

By definition, for all $m \in \mathcal{M}$,

$$\gamma_n(\hat{f}_{\hat{m}}) + \text{pen}(\hat{m}) \leqslant \gamma_n(\hat{f_m}) + \text{pen}(m) \leqslant \gamma_n(f_m) + \text{pen}(m).$$

Applying Formula (4.1), we have

$$\gamma(\hat{f}_{\hat{m}}) - \gamma(f_0) \leqslant \gamma(f_m) - \gamma(f_0) + \langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n + \text{pen}(m) - \text{pen}(\hat{m}). \tag{4.17}$$

Following Baraud (2000) or Castellan (2003a), instead of bounding the supremum of the empirical process $\langle \vec{\varepsilon}, \hat{f}_{\hat{m}} - f_m \rangle_n$, we split it in three terms. Let

$$\overline{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}_{f_0}(\gamma_n(t)) = - < \vec{\varepsilon}, f >_n$$

with $< \vec{\varepsilon}, f >_n$ defined in (4.1), and write

$$\begin{aligned} \gamma(\hat{f}_{\hat{m}}) - \gamma(f_0) \quad \leqslant \quad & \gamma(f_m) - \gamma(f_0) + \text{pen}(m) - \text{pen}(\hat{m}) \\ & + \overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0) + \overline{\gamma}_n(f_0) - \overline{\gamma}_n(f_{\hat{m}}) + \overline{\gamma}_n(f_{\hat{m}}) - \overline{\gamma}_n(\hat{f}_{\hat{m}}). \end{aligned}$$

In other words,

$$\begin{aligned} \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) \quad \leqslant \quad & \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \text{pen}(m) - \text{pen}(\hat{m}) \\ & + \overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0) + \overline{\gamma}_n(f_0) - \overline{\gamma}_n(f_{\hat{m}}) + \overline{\gamma}_n(f_{\hat{m}}) - \overline{\gamma}_n(\hat{f}_{\hat{m}}). \end{aligned} \tag{4.18}$$

The proof of Theorem 4.1 can be decomposed in three steps :

1. We prove that for $\epsilon > 0$,

$$\mathbb{E}_{f_0}\left[(\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)}\right] \leqslant \frac{\kappa'(\rho, \delta, \Gamma, \epsilon)}{n^{(1+\epsilon)}}.$$

2. Let $\Omega_1(\xi)$ be the event

$$\Omega_1(\xi) = \bigcap_{m' \in \mathcal{M}} \left\{ \chi^2_{m'} \mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant \frac{2}{n}|m'| + \frac{16}{n}\left(1 + \frac{\delta}{3}\right)\sqrt{(L_{m'}|m'| + \xi)|m'|} + \frac{8}{n}\left(1 + \frac{\delta}{3}\right)(L_{m'}|m'| + \xi) \right\},$$

where $(L_{m'})_{m' \in \mathcal{M}}$ satisfies Condition (4.2) and m_f is given by Definition 4.1. For all m' in $\mathcal{M}$ we prove that on $\Omega_1(\xi)$

$$\begin{aligned} \left(\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'})\right)\mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant \quad & \frac{1}{2n}\left(\frac{1+\delta}{1-\delta}\right)|m'|\left[2 + \left(1 + \frac{\delta}{3}\right)\left(2\delta + 8L_{m'} + 16\sqrt{L_{m'}}\right)\right] \\ & + \frac{4\xi}{n}\left(\frac{1+\delta}{1-\delta}\right)\left(1 + \frac{\delta}{3}\right)\left(1 + \frac{4}{\delta}\right) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}_{f_{m'}}^{(n)}, \mathbb{P}_{\hat{f}_{m'}}^{(n)})\mathbb{1}_{\Omega_{m_f}(\delta)} \end{aligned} \tag{4.19}$$

and

$$\mathbb{P}(\Omega_1(\xi)^c) \leqslant 2\Sigma e^{-\xi}. \tag{4.20}$$

3. Let $\Omega_2(\xi)$ be the event

$$\Omega_2(\xi) = \bigcap_{m' \in \mathcal{M}} \left[(\overline{\gamma}_n(f_0) - \overline{\gamma}_n(f_{m'})) \leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{m'}}^{(n)}) - 2h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{m'}}^{(n)}) + \frac{2}{n}(L'_m|m'| + \xi) \right].$$

We prove that, $\mathbb{P}(\Omega_2(\xi)^c) \leqslant \Sigma e^{-\xi}$.

Now, we will prove the result of Theorem 4.1 using (R-1), (R-2) and (R-3).
According to (4.18), we can write

$$\begin{aligned} \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\mathbb{1}_{\Omega_{m_f}(\delta)} \quad \leqslant \quad & \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \text{pen}(m) - \text{pen}(\hat{m}) \\ & + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} + (\overline{\gamma}_n(f_0) - \overline{\gamma}_n(f_{\hat{m}}))\mathbb{1}_{\Omega_{m_f}(\delta)} \\ & + (\overline{\gamma}_n(f_{\hat{m}}) - \overline{\gamma}_n(\hat{f}_{\hat{m}})\mathbb{1}_{\Omega_{m_f}(\delta)}. \end{aligned}$$

Combining (R-2) and (R-3) with $m' = \hat{m}$, we infer that on $\Omega_1(\xi) \bigcap \Omega_2(\xi)$

$$\begin{aligned}\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\mathbb{1}_{\Omega_{m_f}(\delta)} &\leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\ &\quad + \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)|\hat{m}|\Big[2 + \Big(1+\frac{\delta}{3}\Big)\Big(2\delta + 8L_{\hat{m}} + 16\sqrt{L_{\hat{m}}}\Big)\Big] + 2L_{\hat{m}}\frac{|\hat{m}|}{n} \\ &\quad + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big] \\ &\quad + \Big[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)}) - 2h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)}) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}_{f_{\hat{m}}}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\Big]\mathbb{1}_{\Omega_{m_f}(\delta)}.\end{aligned}$$

This implies that

$$\begin{aligned}\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\mathbb{1}_{\Omega_{m_f}(\delta)} &\leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\ &\quad + \frac{|\hat{m}|}{n}\Big[\Big(\frac{1+\delta}{1-\delta}\Big) + \Big(\frac{\delta(1+\delta)^2}{1-\delta}\Big) + \Big(\frac{(1+\delta)^2}{1-\delta}\Big)\Big(6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big)\Big] \\ &\quad + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big] \\ &\quad + \Big[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)}) - 2h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)})) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}_{f_{\hat{m}}}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\Big]\mathbb{1}_{\Omega_{m_f}(\delta)}.\end{aligned}$$

Since

$$\Big\{\Big(\frac{1+\delta}{1-\delta}\Big)(1+\delta(1+\delta)) \vee \Big(\frac{(1+\delta)^2}{1-\delta}\Big)\Big\} \leqslant C(\delta) \text{ with } C(\delta) := \Big(\frac{1+\delta}{1-\delta}\Big)^3,$$

we infer

$$\begin{aligned}\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\mathbb{1}_{\Omega_{m_f}(\delta)} &\leqslant \mathcal{K}(\mathbb{P}_{f}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\ &\quad + \frac{|\hat{m}|}{n}C(\delta)\Big[1 + 6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big] + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big] \\ &\quad + \Big[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)}) - 2h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)}) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}_{f_{\hat{m}}}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\Big]\mathbb{1}_{\Omega_{m_f}(\delta)}.\end{aligned}$$

Using Pythagore's type identity $\mathcal{K}(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}}) = \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)}) + \mathcal{K}(\mathbb{P}_{f_{\hat{m}}}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})$ (see Equation (7.42) in Massart (2007)) we have

$$\begin{aligned}\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\mathbb{1}_{\Omega_{m_f}(\delta)} &\leqslant \mathcal{K}(\mathbb{P}_{f}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\ &\quad + \frac{|\hat{m}|}{n}C(\delta)\Big[1 + 6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big] + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big] \\ &\quad + \Big[\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) - 2h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)}) - \frac{\delta}{1+\delta}\mathcal{K}(\mathbb{P}_{f_{\hat{m}}}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\Big]\mathbb{1}_{\Omega_{m_f}(\delta)}.\end{aligned}$$

Now, we successively use

(i) the relation between Kullback-Leibler information and the Hellinger distance $\mathcal{K}(\mathbb{P}_{f_{\hat{m}}}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) \geq 2h^2(\mathbb{P}_{f_{\hat{m}}}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})$ (see Lemma 7.23 in Massart (2007)),

(ii) and inequality $h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)}) \leqslant 2[h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_{\hat{m}}}^{(n)}) + h^2(\mathbb{P}_{f_{\hat{m}}}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})]$.

Consequently, on $\Omega_1(\xi) \bigcap \Omega_2(\xi)$

$$\begin{aligned}\frac{\delta}{1+\delta}h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{\hat{f}_{\hat{m}}}^{(n)})\mathbb{1}_{\Omega_{m_f}(\delta)} &\leqslant \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)} \\ &\quad + \frac{|\hat{m}|}{n}C(\delta)\Big[1 + 6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big] + \frac{4\xi}{n}\Big[\frac{1}{2} + \Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big)\Big].\end{aligned}$$

Since $\mathrm{pen}(\hat{m}) \geq \mu|\hat{m}|\Big[1 + 6L_{\hat{m}} + 8\sqrt{L_{\hat{m}}}\Big]/n$, by taking $\mu = C(\delta)$ yields that on $\Omega_1(\xi) \bigcap \Omega_2(\xi)$

$$h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}})\mathbb{1}_{\Omega_{m_f}(\delta)} \leqslant \frac{2\mu^{1/3}}{\mu^{1/3}-1}\Big(\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_{f_m}^{(n)}) + \mathrm{pen}(m) + (\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)}\Big) + \frac{\xi}{n}C_1(\mu).$$

Then, using that

$$\mathbb{P}(\Omega_1(\xi)^c \cup \Omega_2(\xi)^c) \leqslant 3\Sigma e^{-\xi},$$

we deduce that $\mathbb{P}(\Omega_1(\xi) \cap \Omega_2(\xi)) \geq 1 - 3\Sigma e^{-\xi}$. We now integrating with respect to ξ, and use (R-1) to write that

$$\mathbb{E}_{f_0}\Big[h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}})\mathbb{1}_{\Omega_{m_f}(\delta)}\Big] \quad \leqslant \quad \frac{2\mu^{1/3}}{\mu^{1/3}-1}\Big(\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m)\Big) + \frac{\kappa_1(\rho,\mu,\Gamma,\epsilon)}{n^{(1+\epsilon)}} + \frac{C_2(\mu,\Sigma)}{n}.$$

Furthermore, since $h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}}) \leqslant 1$, by applying Inequality (4.11) we have,

$$\mathbb{E}_{f_0}\Big[h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}})\mathbb{1}_{\Omega^c_{m_f}(\delta)}\Big] \leq \frac{\kappa_2(\rho,\mu,\Gamma,\epsilon)}{n^{(1+\epsilon)}}.$$

Hence we conclude that

$$\mathbb{E}_{f_0}\Big[h^2(\mathbb{P}_{f_0}, \mathbb{P}_{\hat{f}_{\hat{m}}})\Big] \quad \leqslant \quad \frac{2\mu^{1/3}}{\mu^{1/3}-1}\Big(\mathcal{K}(\mathbb{P}^{(n)}_{f_0}, \mathbb{P}^{(n)}_{f_m}) + \mathrm{pen}(m)\Big) + \frac{\kappa_3(\rho,\mu,\Gamma,\epsilon)}{n^{(1+\epsilon)}} + \frac{C_2(\mu,\Sigma)}{n},$$

and minimizing over $\mathcal{M}$ leads to the result of Theorem 4.1.
We now come to the proofs of (R-1), (R-2) and (R-3).
• Proof of (R-1)
We know that

$$\begin{aligned}
\Big|\mathbb{E}_{f_0}\Big[(\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)}\Big]\Big| &= \Big|\mathbb{E}_{f_0}\Big[(\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega^c_{m_f}(\delta)}\Big]\Big| \\
&\leq \mathbb{E}_{f_0}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big\{\Big|\epsilon_i \log\{\frac{\pi_{f_m}(x_i)}{\pi_{f_0}(x_i)}\}\Big| + \Big|\epsilon_i \log\{\frac{1-\pi_{f_m}(x_i)}{1-\pi_{f_0}(x_i)}\}\Big|\Big\}\mathbb{1}_{\Omega^c_{m_f}(\delta)}\Big] \\
&\leq 2\log\Big\{\frac{1}{\rho}\Big\}\,\mathbb{P}(\Omega^c_{m_f}(\delta)).
\end{aligned}$$

We conclude the proof of (R-1) by using Inequality (4.11), which implies that

$$\Big|\mathbb{E}_{f_0}\Big[(\overline{\gamma}_n(f_m) - \overline{\gamma}_n(f_0))\mathbb{1}_{\Omega_{m_f}(\delta)}\Big]\Big| \leq 2\log\Big\{\frac{1}{\rho}\Big\}\frac{\kappa(\rho,\delta,\Gamma,\epsilon)}{n^{(1+\epsilon)}} = \frac{\kappa'(\rho,\delta,\Gamma,\epsilon)}{n^{(1+\epsilon)}}.$$

• Proof of (R-2)
We start by the proof of (4.19)

$$\begin{aligned}
\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'}) &= -\frac{1}{n}\sum_{i=1}^{n}\Big\{\epsilon_i \log\Big(\frac{\pi_{f_{m'}}(x_i)}{\pi_{\hat{f}_{m'}}(x_i)}\Big) - \epsilon_i \log\Big(\frac{1-\pi_{f_{m'}}(x_i)}{1-\pi_{\hat{f}_{m'}}(x_i)}\Big)\Big\} \\
&= -\frac{1}{n}\sum_{J\in m'}\Big(\sum_{i\in J}\epsilon_i\Big)\Big[\frac{\sqrt{|J|\pi^{(J)}_{f_{m'}}}}{\sqrt{|J|\pi^{(J)}_{f_{m'}}}}\log\Big(\frac{\pi^{(J)}_{f_{m'}}}{\pi^{(J)}_{\hat{f}_{m'}}}\Big) - \frac{\sqrt{|J|1-\pi^{(J)}_{f_{m'}}}}{\sqrt{|J|(1-\pi^{(J)}_{f_{m'}})}}\log\Big(\frac{1-\pi^{(J)}_{f_{m'}}}{1-\pi^{(J)}_{\hat{f}_{m'}}}\Big)\Big].
\end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'}) \leq &\sqrt{\frac{1}{n}\sum_{J\in m'}|J|\Big[\pi^{(J)}_{f_{m'}}\log^2\Big(\frac{\pi^{(J)}_{f_{m'}}}{\pi^{(J)}_{\hat{f}_{m'}}}\Big) + (1-\pi^{(J)}_{f_{m'}})\log^2\Big(\frac{1-\pi^{(J)}_{f_{m'}}}{1-\pi^{(J)}_{\hat{f}_{m'}}}\Big)\Big]} \\
&\times\sqrt{\frac{1}{n}\sum_{J\in m'}\Big[\frac{\Big(\sum_{i\in J}\epsilon_i\Big)^2}{|J|\pi^{(J)}_{f_{m'}}} + \frac{\Big(\sum_{i\in J}\epsilon_i\Big)^2}{|J|(1-\pi^{(J)}_{f_{m'}})}\Big]},
\end{aligned}$$

and in other words

$$\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'}) \leq \sqrt{\mathcal{X}^2_{m'}} \times \sqrt{V^2(\pi_{f_{m'}}, \pi_{\hat{f}_{m'}})},$$

where $\mathcal{X}^2_{m'}$ and $V^2(\pi_{f_{m'}}, \pi_{\hat{f}_{m'}})$ are defined respectively in (4.9) and (4.13) . Using both that inequality $2xy \leqslant \theta x^2 + \theta^{-1}y^2$, for all $x > 0$, $y > 0$ with $\theta = (1+\delta)/(1-\delta)$, and Inequality (4.12), we obtain on $\Omega_{m_f}(\delta)$ that,

$$\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'})) \;\leq\; \frac{1}{2}\Big(\frac{1+\delta}{1-\delta}\Big)\chi^2_{m'} + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}}, \mathbb{P}^{(n)}_{\hat{f}_{m'}}).$$

Consequently, on $\Omega_1(\xi)$

$$\begin{aligned}(\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'}))\mathbb{1}_{\Omega_{m_f}(\delta)} \;&\leq\; \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)\Big[2|m'| + 16\Big(1+\frac{\delta}{3}\Big)\sqrt{(L_{m'}|m'|+\xi)|m'|} + 8\Big(1+\frac{\delta}{3}\Big)(L_{m'}|m'|+\xi)\Big] \\ &+\; \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}}, \mathbb{P}^{(n)}_{\hat{f}_{m'}})\mathbb{1}_{\Omega_{m_f}(\delta)}.\end{aligned}$$

Using inequalities $|x+y|^{1/2} \leqslant |x|^{1/2} + |y|^{1/2}$ and $2xy \leqslant \theta x^2 + \theta^{-1}y^2$ with $\theta = \delta/4$, we infer that (4.19) follows since

$$\begin{aligned}\overline{\gamma}_n(f_{m'}) - \overline{\gamma}_n(\hat{f}_{m'}))\mathbb{1}_{\Omega_{m_f}(\delta)} \;&\leq\; \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)\Big[2|m'| + \Big(1+\frac{\delta}{3}\Big)\Big(16\sqrt{L_{m'}}|m'| + 8L_{m'}|m'| + 2\delta|m'|\Big) \\ &\quad +8\xi\Big(1+\frac{\delta}{3}\Big)(1+\frac{4}{\delta})\Big] + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}}, \mathbb{P}^{(n)}_{\hat{f}_{m'}})\mathbb{1}_{\Omega_{m_f}(\delta)} \\ &\leq\; \frac{1}{2n}\Big(\frac{1+\delta}{1-\delta}\Big)|m'|\Big[2 + \Big(1+\frac{\delta}{3}\Big)\Big(2\delta + 8L_{m'} + 16\sqrt{L_{m'}}\Big)\Big] \\ &\quad +\frac{4\xi}{n}\Big(\frac{1+\delta}{1-\delta}\Big)\Big(1+\frac{\delta}{3}\Big)\Big(1+\frac{4}{\delta}\Big) + \frac{1}{1+\delta}\mathcal{K}(\mathbb{P}^{(n)}_{f_{m'}}, \mathbb{P}^{(n)}_{\hat{f}_{m'}})\mathbb{1}_{\Omega_{m_f}(\delta)}.\end{aligned}$$

• Proof of (4.20) :
Write $\mathcal{X}^2_{m'} = \sum_{J\in m'}\{Z_{1,J} + Z_{2,J}\}$, where

$$Z_{1,J} = \frac{1}{n}\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|\pi^{(J)}_{f_{m'}}} \;\text{ and }\; Z_{2,J} = \frac{1}{n}\frac{(\sum_{k\in J}\varepsilon_k)^2}{|J|(1-\pi^{(J)}_{f_{m'}})}.$$

We will control $\sum_{J\in m'} Z_{1,J}$ and $\sum_{J\in m'} Z_{2,J}$ separately. In order to use Bernstein inequality (see Theorem 4.1), we need an upper bound of $\sum_{J\in m'}\mathbb{E}[Z^p_{1,J}\mathbb{1}_{\Omega_{m_f}(\delta)}]$, for every $p \geq 2$. By definition

$$\mathbb{E}[Z^p_{1,J}\mathbb{1}_{\Omega_{m_f}(\delta)}] \;=\; \frac{1}{\Big(n|J|\pi^{(J)}_{f_{m'}}\Big)^p}\int_0^\infty 2px^{2p-1}\mathbb{P}\Big(\Big\{|\sum_{k\in J}\varepsilon_k| \geq x\Big\} \cap \Omega_{m_f}(\delta)\Big)dx.$$

For every m' constructed on the grid m_f, for all $J \in m'$, on $\Omega_{m_f}(\delta) \cap \Big\{x \leqslant |\sum_{k\in J}\varepsilon_k|\Big\}$, we have

$$x \leqslant |\sum_{k\in J}\varepsilon_k| \leqslant \delta\sum_{i\in J}\pi_{f_0}(x_i).$$

Combining the previous inequality, the Bernstein inequality (4.15) with the fact that $\varepsilon_k \leqslant 1$,

we infer that

$$\begin{aligned}
\mathbb{E}[Z_{1,J}^p \mathbb{1}_{\Omega_{m_f}(\delta)}] &\leq \frac{1}{\Big(n\sum_{k\in J}\pi_{f_0}(x_k)\Big)^p}\int_0^{\delta\sum_{k\in J}\pi_{f_0}(x_k)} 2px^{2p-1}\mathbb{P}\Big(|\sum_{k\in J}\varepsilon_k|\geq x\Big)dx \\
&\leq \frac{1}{\Big(n\sum_{k\in J}\pi_{f_0}(x_k)\Big)^p}\int_0^{\delta\sum_{i\in J}\pi_{f_0}(x_i)} 4px^{2p-1}\exp\Big(-\frac{x^2}{2\Big(\frac{x}{3}+\sum_{k\in J}\pi_{f_0}(x_k)\Big)}\Big)dx \\
&\leq \frac{1}{\Big(n\sum_{k\in J}\pi_{f_0}(x_k)\Big)^p}\int_0^{\delta\sum_{i\in J}\pi_{f_0}(x_i)} 4px^{2p-1}\exp\Big(-\frac{x^2}{2\Big(1+\frac{\delta}{3}\Big)\sum_{k\in J}\pi_{f_0}(x_k)}\Big)dx \\
&\leq \frac{1}{n^p}2^{p+1}(1+\frac{\delta}{3})^p p\int_0^\infty t^{p-1}\exp(-t)dt \\
&\leq \frac{1}{n^p}2^{p+1}p(1+\frac{\delta}{3})^p(p!).
\end{aligned}$$

Consequently

$$\sum_{J\in m'}\mathbb{E}[Z_{1,J}^p\mathbb{1}_{\Omega_{m_f}(\delta)}]\leqslant\frac{1}{n^p}2^{p+1}p(1+\frac{\delta}{3})^p(p!)\times|m'|.$$

Now, since $p\leqslant 2^{p-1}$, we have

$$\sum_{J\in m'}\mathbb{E}[Z_{1,J}^p\mathbb{1}_{\Omega_{m_f}(\delta)}]\leqslant\frac{p!}{2}\times\Big[\frac{32}{n^2}(1+\frac{\delta}{3})^2|m'|\Big]\times\Big[\frac{4}{n}(1+\frac{\delta}{3})\Big]^{p-2}.$$

Using Bernstein inequality and that $\mathbb{E}\Big[\sum_{J\in m'}Z_{1,J})\Big]\leqslant|m'|/n$, we have that for every positive x

$$\mathbb{P}\Big(\sum_{J\in m'}Z_{1,J}\mathbb{1}_{\Omega_{m_f}(\delta)}\geq\frac{|m'|}{n}+\frac{8}{n}(1+\frac{\delta}{3})\sqrt{x|m'|}+\frac{4}{n}(1+\frac{\delta}{3})x\Big)\leqslant\exp(-x).$$

In the same way we prove that

$$\mathbb{P}\Big(\sum_{J\in m'}Z_{2,J}\mathbb{1}_{\Omega_{m_f}(\delta)}\geq\frac{|m'|}{n}+\frac{8}{n}(1+\frac{\delta}{3})\sqrt{x|m'|}+\frac{4}{n}(1+\frac{\delta}{3})x\Big)\leqslant\exp(-x).$$

Hence

$$\mathbb{P}\Big(\mathcal{X}^2_{m'}\mathbb{1}_{\Omega_{m_f}(\delta)}\geq\frac{2|m'|}{n}+\frac{16}{n}(1+\frac{\delta}{3})\sqrt{x|m'|}+\frac{8}{n}(1+\frac{\delta}{3})x\Big)\leqslant 2\exp(-x),$$

and we conclude that $\mathbb{P}(\Omega_1^c(\xi))\leqslant 2\sum_{m'}\exp(-L'_m|m'|-\xi)=2\Sigma e^{-\xi}$. This ends the proof of (R-2).

• Proof of (R-3)

Recall that $\overline{\gamma}_n(f)=\gamma_n(f)-\mathbb{E}(\gamma_n(f))$ for every f. According to Markov inequality, for $b>0$,

$$\begin{aligned}
\mathbb{P}((\overline{\gamma}_n(f_0)-\overline{\gamma}_n(g))\geq b) &= \mathbb{P}\Big(\exp\Big(\frac{n}{2}(\overline{\gamma}_n(f_0)-\overline{\gamma}_n(g))\Big)\geq\exp\Big(\frac{nb}{2}\Big)\Big) \\
&\leq \exp\Big(\frac{-nb}{2}\Big)\mathbb{E}\Big[\exp\Big(\frac{n}{2}(\overline{\gamma}_n(f_0)-\overline{\gamma}_n(g))\Big)\Big] \\
&= \exp\Big[\frac{-nb}{2}+\log\mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0)-\gamma_n(g)\Big)+\frac{n}{2}\mathbb{E}\Big[\gamma_n(g)-\gamma_n(f_0)\Big]\Big)\Big]\Big] \\
&\leq \exp\Big[\frac{-nb}{2}+\frac{n}{2}\mathcal{K}(\mathbb{P}_{f_0}^{(n)},\mathbb{P}_g^{(n)})+\log\mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0)-\gamma_n(g)\Big)\Big)\Big]\Big].
\end{aligned}$$

Now,

$$\begin{aligned}
\log \mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0)-\gamma_n(g)\Big)\Big)\Big] &= \log \mathbb{E}\Big[\exp\Big(\frac{1}{2}\sum_{i=1}^{n} Y_i \log(\frac{\pi_g(x_i)}{\pi_{f_0}(x_i)}) + (1-Y_i)\log(\frac{1-\pi_g(x_i)}{1-\pi_{f_0}(x_i)})\Big)\Big] \\
&= \log \mathbb{E}\Big[\Pi_{i=1}^{n}\Big\{\Big(\frac{\pi_g(x_i)}{\pi_{f_0}(x_i)}\Big)^{Y_i/2} \times \Big(\frac{1-\pi_g(x_i)}{1-\pi_{f_0}(x_i)}\Big)^{(1-Y_i)/2}\Big\}\Big] \\
&= \log \Pi_{i=1}^{n}\Big\{\sqrt{\frac{\pi_g(x_i)}{\pi_{f_0}(x_i)}}\pi_{f_0}(x_i) + \sqrt{\frac{1-\pi_g(x_i)}{1-\pi_{f_0}(x_i)}}(1-\pi_{f_0}(x_i))\Big\} \\
&= \sum_{i=1}^{n} \log\Big\{\sqrt{\pi_g(x_i)\pi_{f_0}(x_i)} + \sqrt{(1-\pi_g(x_i))(1-\pi_{f_0}(x_i))}\Big\}.
\end{aligned}$$

In other words we have

$$\begin{aligned}
&\log \mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0)-\gamma_n(g)\Big)\Big) = \\
&\qquad \sum_{i=1}^{n} \log\Big\{1 - \frac{1}{2}\Big[\Big(\sqrt{\pi_{f_0}(x_i)} - \sqrt{\pi_g(x_i)}\Big)^2 + \Big(\sqrt{1-\pi_{f_0}(x_i)} - \sqrt{1-\pi_g(x_i)}\Big)^2\Big]\Big\}.
\end{aligned}$$

This implies that

$$\begin{aligned}
\log \mathbb{E}\Big[\exp\Big(\frac{n}{2}\Big(\gamma_n(f_0)-\gamma_n(g)\Big)\Big)\Big] &\leq \sum_{i=1}^{n} -\frac{1}{2}\Big[\Big(\sqrt{\pi_{f_0}(x_i)} - \sqrt{\pi_g(x_i)}\Big)^2 + \Big(\sqrt{1-\pi_{f_0}(x_i)} - \sqrt{1-\pi_g(x_i)}\Big)^2\Big] \\
&= -nh^2(\mathbb{P}_{f_0}, \mathbb{P}_g).
\end{aligned}$$

Consequently

$$\mathbb{P}(\overline{\gamma}_n(f_0) - \overline{\gamma}_n(g) \geq b) \leqslant \exp\Big[\frac{-nb}{2} + \frac{n}{2}\mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_g^{(n)}) - nh^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_g^{(n)})\Big],$$

and, if we choose for positive x,

$$b = \frac{2x}{n} + \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_g^{(n)}) - 2h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_g^{(n)}) > 0,$$

we have,

$$\mathbb{P}\Big(\overline{\gamma}_n(f_0) - \overline{\gamma}_n(g) \geq \frac{2x}{n} + \mathcal{K}(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_g^{(n)}) - 2h^2(\mathbb{P}_{f_0}^{(n)}, \mathbb{P}_g^{(n)})\Big) \leqslant \exp(-x).$$

We conclude that $\mathbb{P}(\Omega_2^c(\xi)) \leqslant \sum_{m'} \exp(-L'_m|m'| - \xi) \leq \Sigma e^{-\xi}$, which ends the proof of (R-3). □

Appendix

Proof of Lemma 4.1.

By definition

$$f_m = \arg\min_{f \in S_m}\left[\sum_{i=1}^{n} \log(1+\exp(f(x_i))) - \pi_{f_0}(x_i)f(x_i)\right].$$

For all $f \in S_m$, for all $J \in m$ and for all $x \in J$, we have $f(x) = f^{(J)}$. Hence $f_m(x) = \overline{f}_m^{(J)}$ for all x in J, and for all J in m, we aim at finding $\overline{f}_m^{(J)}$ such that

$$\overline{f}_m^{(J)} = \arg\min_{f^{(J)}}\left[|J|\log(1+\exp(f^{(J)})) - \sum_{i \in J}\pi_{f_0}(x_i)f^{(J)}\right]$$

where $|J| = card\{i \in \{1,...,n\}; x_i \in J\}$. Easy calculations show that he coefficient $\overline{f}_m^{(J)}$ satisfies

$$|J|\frac{\exp(\overline{f}_m^{(J)})}{1+\exp(\overline{f}_m^{(J)})} - \sum_{i\in J}\pi_{f_0}(x_i) = 0,$$

that is

$$\overline{f}_m^{(J)} = \log\left(\frac{\sum_{i\in J}\pi_{f_0}(x_i)}{|J|(1-\sum_{i\in J}\pi_{f_0}(x_i)/|J|)}\right). \tag{4.1}$$

Consequently, π_{f_m} defined as in (4.2) satisfies that $\pi_{f_m}(x) = \pi_{f_m}^{(J)}$ for all $x \in J$, where

$$\pi_{f_m}^{(J)} = \frac{1}{|J|}\sum_{i\in J}\pi_{f_0}(x_i),$$

and hence $\pi_{f_m} = \arg\min_{t\in S_m} \| t - \pi_{f_0} \|_n$ is the usual projection of π_{f_0} on to $S_m =< \Phi_j, j \in m >$. In the same way, $\hat{f}_m$ defined by (4.10) satisfies $\hat{f}_m(t) = \hat{f}_m^{(J)}$ for all $t \in J$, where

$$\hat{f}_m^{(J)} = \log\left(\frac{\sum_{i\in J} Y_i}{|J|(1-\sum_{i\in J} Y_i/|J|)}\right).$$

In other words, $\pi_{\hat{f}_m}$, defined as π_f with f replaced by $\pi_{\hat{f}_m}$, satisfies $\pi_{\hat{f}_m}(x) = \pi_{\hat{f}_m}^{(J)}$, for all $x \in J$, with

$$\pi_{\hat{f}_m}^{(J)} = \frac{1}{|J|}\sum_{i\in J} Y_i.$$

Proof of Lemma 4.2.

In the following, for the sake of notation simplicity, we will use $\gamma(\beta)$ for $\gamma(f_\beta)$. A second-order Taylor expansion of the function $\gamma()$ around β^* gives for any $\beta \in \Lambda_m$

$$\begin{aligned}&\gamma(\beta) = \gamma(\beta^*) + \nabla_\beta\gamma(\beta^*)(\beta-\beta^*)\\ &+ \int_0^1 (1-t)\sum_{i_1+\cdots+i_D=2}\frac{2!}{i_1!\dots i_D!}(\beta_1-\beta_1^*)^{i_1}\dots(\beta_D-\beta_D^*)^{i_D}\frac{\partial\gamma^2}{\partial\beta_1\dots\partial\beta_D}(\beta^*+t(\beta-\beta*))dt.\end{aligned}$$

Easy calculation shows that

$$\begin{aligned}&\sum_{i_1+\cdots+i_D=2}\frac{2!}{i_1!\dots i_D!}(\beta_1-\beta_1^*)^{i_1}\dots(\beta_D-\beta_D^*)^{i_D}\frac{\partial\gamma^2}{\partial\beta_1\dots\partial\beta_D}(\beta^*+t(\beta-\beta*))\\ =\;& \sum_{j=1}^D\frac{1}{n}\sum_{i=1}^n\psi_j^2(x_i)(\beta_j-\beta_j^*)^2\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\left[1-\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\right]\\ +\;& 2\sum_{l\neq k}\frac{1}{n}\sum_{i=1}^n\psi_l(x_i)\psi_k(x_i)(\beta_l-\beta_l^*)(\beta_k-\beta_k^*)\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\left[1-\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\right]\\ =\;& \frac{1}{n}\sum_{i=1}^n\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\left[1-\pi\left(f_{\beta^*+t(\beta-\beta*)}(x_i)\right)\right](f_\beta(x_i)-f_{\beta^*}(x_i))^2.\end{aligned}$$

This implies that

$$\gamma(\beta) \geq \gamma(\beta^*) + \nabla_\beta\gamma(\beta^*)(\beta-\beta^*) + \frac{\mathcal{U}_0^2}{2}\|f_\beta - f_{\beta^*}\|_n^2.$$

Since β^* is the minimizer of $\gamma(.)$ over the set Λ_m, we have $\nabla_\beta\gamma(\beta^*)(\beta-\beta^*) \geq 0$ for all $\beta \in \Lambda_m$. Thus the result follows.

Proof of Lemma 4.3

Let S_D and $S_{D'}$ two vector spaces of dimension D and D' respectively. Set $S = S_D \cap \mathbb{L}_\infty(C_0) + S_{D'} \cap \mathbb{L}_\infty(C_0)$ and $\vec{\varepsilon}'$ be an independent copie of $\vec{\varepsilon}$. Set

$$Z = \sup_{u \in S} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n}, \text{ and for all } i = 1, \dots, n, \quad Z^{(i)} = \sup_{u \in S} \frac{1}{\| u \|_n} \left(\frac{1}{n} \sum_{k \neq i} \varepsilon_k u(x_k) + \varepsilon_i' u(x_i) \right). \tag{4.2}$$

By Cauchy-Schwarz Inequality the supremum in (4.2) is achieved at $\Pi_S(\vec{\varepsilon})$. Consequently,

$$Z - Z^{(i)} \leqslant \frac{(\varepsilon_i - \varepsilon_i')(\Pi_S(\vec{\varepsilon})(x_i)}{n \| \Pi_S(\vec{\varepsilon}) \|_n}, \qquad \text{and} \qquad \mathbb{E}_{f_0}[(Z - Z^{(i)})^2 | \vec{\varepsilon}] \leq \mathbb{E}_{f_0} \left[\frac{(\varepsilon_i - \varepsilon_i')^2 [\Pi_S(\vec{\varepsilon})(x_i)]^2}{n^2 \| \Pi_S(\vec{\varepsilon}) \|_n^2} | \vec{\varepsilon} \right]$$

with

$$\begin{aligned} \mathbb{E}_{f_0} \left[\frac{(\varepsilon_i - \varepsilon_i')^2 [\Pi_S(\vec{\varepsilon})(x_i)]^2}{n^2 \| \Pi_S(\vec{\varepsilon}) \|_n^2} | \vec{\varepsilon} \right] &= \frac{[\Pi_S(\vec{\varepsilon})(x_i)]^2}{n^2 \| \Pi_S(\vec{\varepsilon}) \|_n^2} \mathbb{E}_{f_0} \left[(\varepsilon_i - \varepsilon_i')^2 | \vec{\varepsilon} \right] \\ &= \frac{[\Pi_S(\vec{\varepsilon})(x_i)]^2}{n^2 \| \Pi_S(\vec{\varepsilon}) \|_n^2} \left(\varepsilon_i^2 + \mathbb{E}_{f_0}(\varepsilon_i^2) \right) \leq \frac{5 [\Pi_S(\vec{\varepsilon})(x_i)]^2}{4n^2 \| \Pi_S(\vec{\varepsilon}) \|_n^2}. \end{aligned}$$

This implies that

$$\sum_{i=1}^n \mathbb{E}_{f_0}[(Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}} | \vec{\varepsilon}] \leq \frac{5}{4n}.$$

We now apply Lemma 4.1 from Boucheron *et al.* (2004)), that is recalled here.

Lemma 4.1 *Let $X_1, \dots, X_n$ independent random variables taking values in a measurable space $\mathcal{X}$. Denote by X_1^n the vector of these n random variables. Set $Z = f(X_1, \dots, X_n)$ and $Z^{(i)} = f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$, where $X_1', \dots, X_n'$ denote independent copies of $X_1, \dots, X_n$ and $f : \mathcal{X}^n \to \mathbb{R}$ some measurable function. Assume that there exists a positive constant c such that, $\mathbb{E}_{f_0} \left[\sum_{i=1}^n (Z - Z^{(i)})^2 \mathbb{1}_{Z > Z^{(i)}} | X_1^n \right] \leqslant c$. Then for all $t > 0$,*

$$\mathbb{P}_{f_0}(Z > \mathbb{E}_{f_0}(Z) + t) \leqslant e^{-t^2/4c}.$$

Applying Lemma 4.1 to Z defined in (4.2), we obtain that for all $x > 0$,

$$\mathbb{P} \left(\sup_{u \in S} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n} > \mathbb{E}_{f_0} \left[\sup_{u \in S} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n} \right] + \sqrt{\frac{5x}{n}} \right) \leqslant \exp(-x).$$

Let $\{\psi_1, \dots, \psi_{D+D'}\}$ be an orthonormal basis of $S_D + S_{D'}$. Using Jensen's Inequality, we write

$$\begin{aligned} \mathbb{E}_{f_0} \left[\sup_{u \in S} \frac{\langle \vec{\varepsilon}, u \rangle_n}{\| u \|_n} \right] = \mathbb{E}_{f_0}(\| \Pi_S(\vec{\varepsilon}) \|_n) = \mathbb{E}_{f_0} \left[\left(\sum_{k=1}^{D+D'} (\langle \vec{\varepsilon}, \psi_k \rangle_n)^2 \right)^{1/2} \right] \\ \leq \left(\sum_{k=1}^{D+D'} \mathbb{E}_{f_0}(\langle \vec{\varepsilon}, \psi_k \rangle_n)^2 \right)^{1/2} \\ \leqslant \sqrt{\frac{D+D'}{4n}}. \end{aligned}$$

This concludes the proof of Lemma 4.3.

Conclusion et perspectives

Les travaux présentés dans cette thèse peuvent être divisés en deux parties : une partie appliquée et une partie méthodologique. La partie appliquée, présentée dans le Chapitre 2 et à l' Appendix A.1, porte sur l'analyse des données Actu-Palu. La partie méthodologique présentée dans les Chapitres 3 et 4 est consacrée à l'étude des propriétés des estimateurs du maximum de vraisemblance pénalisé dans le modèle de régression logistique. L'étude de ces propriétés se fait par la démonstration des inégalités oracles non asymptotiques.

Au Chapitre 2, nous avons proposé et mis en oeuvre des stratégies de sélection de variables dans des grandes enquêtes socio-épidémiologiques. Ces stratégies s'effectuent en deux étapes. La première étape est une étape de réduction du nombre de variables par l'une des méthodes suivantes : Lasso, Group Lasso ou forêts aléatoires. La deuxième étape consiste à prédire par le modèle de régression logistique en prenant en compte les variables sélectionnées par les méthodes de la première étape. Ces stratégies ont été mises en oeuvre sur les données Actu-Palu pour sélectionner les variables pertinentes pour la prédiction des foyers à risque d'avoir un épisode fébrile à Dakar. Ce travail met en évidence plusieurs aspects : d'une part l'importance, dans les grandes enquêtes socio-épidémiologiques, de réduire le nombre de variables explicatives, à l'aide d'outils appropriés, avant l'utilisation des méthodes d'analyses statistiques standards telle que la régression logistique. D'autre part l'intérêt des méthodes Lasso, Group Lasso et forêts aléatoires, peu utilisées dans ce type d'enquêtes, pour cette réduction. En effet ces méthodes nous ont permis d'exhiber un modèle plus interprétable et qui a de meilleurs qualités de prédiction, en particulier que le modèle complet. Enfin, la méthode optimale au sens de la prédiction est le Group Lasso, qui a la particularité de prendre en compte de manière groupée les modalités des variables qualitatives. Cet aspect est important puisque ces grandes enquêtes présentent très souvent un nombre important de variables qualitatives.

Les données Actu-Palu sont issues d'une enquête par questionnaire plus générale auprès de 50 quartiers de la conurbation de Dakar-Pikine-Guédiawaye-Rufisque. Lors de ma première année de thèse je me suis intéressé à une autre variable à expliquer, il s'agit du recours aux soins en cas de fièvre chez un enfant de 2 à 10 ans. Cette partie est présentée à l' Appendix A.1. L'objectif était de mettre en évidence les variables explicatives permettant de prédire le type de recours aux soins. Lors de cette étude, plusieurs difficultés sont apparues. La difficulté principale a été qu'aucune méthode de réduction de dimension n'a permis d'obtenir une erreur de prédiction acceptable. En effet toutes les erreurs de prédiction étaient autour de 40%. En réalité, ce résultat bien qu'apparemment négatif soulève de nombreuses questions qui ont dépassé le cadre de cette thèse avec la direction donnée finalement. La première chose à remarquer c'est qu'une étude rapide des données montre une proportion affichée des foyers annonçant le recours aux soins externes largement supérieure à ce qui est usuellement constaté sur le terrain. Cette partie soulève donc de nombreuses questions : Le modèle logistique a été utilisé pour prédire le type de recours aux soins. Cependant il ne permet pas de prendre en compte l'effet quartier. Il serait donc intéressant de prendre en compte cette effet quartier en utilisant par exemple un modèle de régression logistique mixte (voir Groll et Tutz (2012), Schelldorfer et al. (2011)). Ainsi la réduction de dimension peut se faire en utilisant le Lasso pour le modèle de régression logistique mixte (Groll et Tutz (2012)). A notre connaissance, le Group Lasso pour le modèle logistique mixte n'a pas encore été étudié. Deux pistes de recherche peuvent donc être envisagées. D'une part la définition et l'étude des propriétés théoriques du Group Lasso pour le modèle de régression logistique mixte. D'autre part la proposition d'une procédure d'implémentation.

L'une des méthodes de réduction que nous avons utilisé est basée sur les forêts aléatoires et l'indice d'importance des variables. Bien que la méthode des forêts aléatoires soit largement plébiscitée dans des études expérimentales, très peu de résultats permettant d'évaluer ses propriétés théoriques existent (voir Biau (2012) pour un exemple de résultat théorique). L'étude de ses propriétés peut donc constituer une piste de recherche intéressante. De même il n'existe pas à notre connaissance une étude des propriétés théoriques de l'indice d'importance des variables. Une telle étude serait d'une grande importance pour confirmer leurs bonnes propriétés observées dans des études pratiques (Strobl et al. (2009)).

Dans le Chapitre 3, nous avons proposé des versions pondérées des estimateurs Lasso et Group Lasso pour le modèle de régression logistique. Dans un contexte de grande dimension, nous avons établi des inégalités oracles non asymptotiques pour ces estimateurs. Ces inégalités oracles montrent que les estimateurs ont un risque aussi petit, à une constante multiplicative près, que le meilleur compromis entre le bais et la variance. Nos résultats ne font pas l'hypothèse que la vraie fonction à estimer est linéaire, ce qui les démarquent des résultats dans la littérature sur le modèle de régression logistique. Nous avons montré par des études de simulations que nos estimateurs ont de bonnes propriétés en sélection, et qu'ils sont meilleurs que le Lasso et Group Lasso canoniques, au moins dans les cas considérés dans les simulations.

Dans la continuité de ce chapitre, il serait intéressant d'établir des inégalités oracles non asymptotique pour d'autres variantes du Lasso dans le modèle de régression logistique (par exemple *elastic net* (Zou et Hastie (2005)), *fused Lasso* (Tibshirani et al. (2005)), *latent Group Lasso* (Jacob et al. (2009))).

Dans le Chapitre 4, nous avons étendu la notion de sélection de modèle développée par Birgé et Massart (2001) à la régression logistique. Nous avons établi des inégalités oracles non asymptotiques pour les estimateurs du maximum de vraisemblance pénalisé. Ces inégalités oracles montrent que ces estimateurs ont un risque aussi petit, à une constante multiplicative près, et à terme de reste près, que le risque du meilleur estimateur de la collection d'estimateurs, *i.e.* celui qui a le risque le plus petit. Les études de simulations ont montré que les critères que nous proposons, basés sur l'heuristique de pente, ont de bonnes performances.

Il pourrait être intéressant d'étudier les propriétés d'optimalité au sens du risque minimax des estimateurs proposés dans le Chapitre 4. Un estimateur a une vitesse optimale au sens du risque minimax sur une classe de fonctions $\mathcal{S}$ si sa vitesse est la meilleure possible pour estimer les fonctions f appartenant à la classe $\mathcal{S}$. Une autre piste de recherche serait d'établir des inégalités oracles non asymptotiques, dans le même esprit que celui du Théorème 4.1, en considérant d'autres bases que les fonctions constantes par morceaux, par exemple les bases d'ondelettes, les bases trigonométriques *etc.* Par ailleurs, les études de simulations, basées sur l'heuristique de pente, laissent penser que cette heuristique peut être validée théoriquement en régression logistique. La validation théorique de cette heuristique en régression logistique reste une question ouverte qui mérite d'être étudiée.

Plus généralement, il serait intéressant d'étendre les résultats des Chapitres 3 et 4 à la famille des modèles linéaires généralisés (McCullagh et Nelder 1983) à laquelle appartient le modèle logistique et d'autres modèles comme le modèle linéaire gaussien et le modèle de Poisson *etc*. Les résultats de ces chapitres sont établis dans le cas d'un *design* fixe ($z_1, \ldots, z_n$, déterministes), il serait intéressant d'étudier le cas du *design* aléatoire.

Annexes

A

Sommaire

A.1 Sélection des variables pour la prédiction du type de recours aux soins

Cette partie porte sur la sélection de variables pertinentes pour la prédiction du recours aux soins en cas de fièvre. En d'autres termes, on cherche ce qui explique qu'en cas de fièvre chez un enfant de 2 à 10 ans, sa mère choisit de le soigner par automédication ou de recourir aux services de santé externes. L'automédication est l'utilisation des médicaments hors prescription médicale . Comme nous l'avons dit à l'introduction, l'automédication est connue pour être l'une des causes de l'apparition et la propagation de la chimiorésistance de *plasmodium falciparum* aux antipaludiques. L'apparition de ces chimiorésistances a comme possible conséquence l'échec des stratégies de lutte contre le paludisme. Il est donc important d'étudier les déterminants du recours à l'automédication pour améliorer durablement l'efficacité des nouvelles stratégies de lutte contre le paludisme.

A.1.1 Données Actu-Palu utilisées

Les données Actu-Palu utilisées ici sont issues d'une enquête par questionnaire auprès de la population [1] de la conurbation de Dakar-Pikine-Guédiawaye-Rufisque. Cinquante quartiers ont été enquêtés, dans lesquels 60 ménages ont été visités, soit un échantillon de 3000 ménages. Deux catégories d'informations ont été recueillies : l'une portant sur les caractéristiques du ménage, et l'autre sur le mode de vie dans le ménage notamment sur l'accès aux soins. Ces informations apparaissent dans 2 questionnaires : le questionnaire ménage et le questionnaire femme.

• Questionnaire ménage

Il porte sur les caractéristiques des membres du ménage : caractéristiques de l'habitat et de l'environnement domestique, ressources matérielles et monétaires, *etc*.

• Questionnaire femme

Dans ce questionnaire une femme du ménage (le plus souvent la mère) a été interrogée sur les caractéristiques socio-épidémiologiques et culturelles, sur les pratiques d'accès aux soins en général et tout particulièrement sur la démarche suivie si un enfant de 2 à 10 ans a eu une fièvre dans le mois précédent la visite de l'enquêteur.

Variable à expliquer

Nous nous intéressons ici au recours au soins en cas de fièvre chez un enfant de 2 à 10 ans. Il s'agit ici du premier recours aux soins (car il peut y en avoir plusieurs). Nous avons retrouvé dans les données 5 types de recours aux soins : l'automédication moderne, l'automédication traditionnelle, le recours aux services de santé ou médecin, le recours aux guérisseurs, et les non recours. 43,7% des femmes de l'échantillon ont eu recours à l'automédication moderne, contre 48,1% qui ont eu recours aux services de santé (figure A.1). Dans cette partie, la question porte essentiellement sur le problème de l'automédication qu'elle soit moderne ou traditionnelle. Nous avons fait le choix de fusionner les modalités automédication moderne et automédication traditionnelle en une modalité *automédication*. Les modalités service de santé ou médecin privé et guérisseur seront regroupées en une modalité correspondant au *recours externe*. La modalité non recours (*n'a rien fait*) sera retirée de l'analyse. En effet, elle ne peut pas être fusionnée avec l'une des modalités précédentes et ne peut constituer une modalité à elle seule à cause de son effectif trop faible (2.7%). La variable d'intérêt aura donc 2 modalités : automédication, recours externe.

Prétraitement

On dénombre 2952 femmes qui ont participé à l'enquête. Parmi elles, 1273 femmes vérifiaient le critère d'inclusion : avoir un enfant de 2 à 10 ans qui a eu une fièvre au cours des 30 derniers jours précédant la visite de l'enquêteur dont il est guéri depuis plus

1. Au Chapitre2 nous avons utilisé les données sur Pikine.

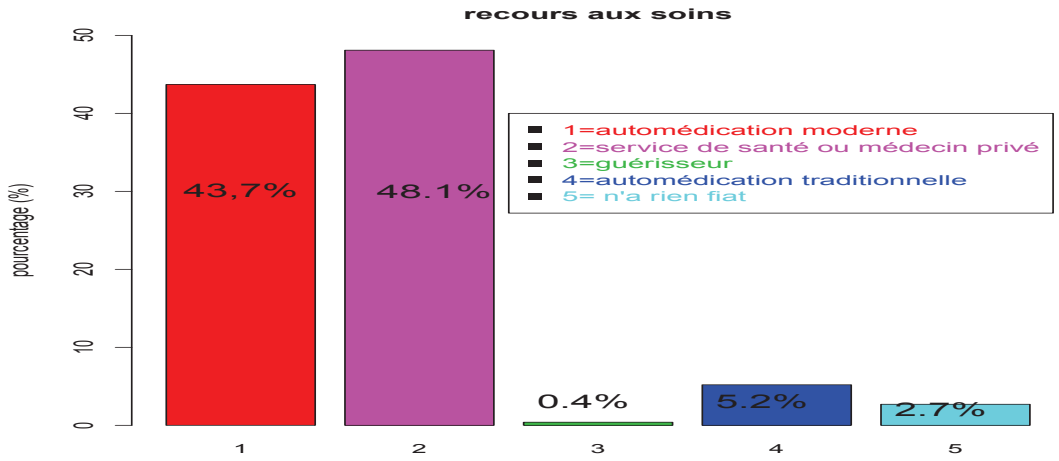


FIGURE A.1 – *Repartition des modalités de la variable d'intérêt*

de 3 jours. Nous avons donc au total 1273 femmes dans l'étude pour 73 variables explicatives (en majorité catégorielles). Les données comportaient des "données manquantes" : non réponse à une question. Ces données manquantes sont inhérentes aux enquêtes socio-épidémiologiques, où les personnes ont du mal à répondre à certaines questions. Nous avons analysé les données manquantes, elles sont réparties de façon aléatoire dans la base de données. Nous avons choisi d'enlever de la base de données les femmes ayant au moins une valeur manquante. La suppression des femmes ayant des données manquantes n'a, à notre avis, pas causé de biais car les valeurs manquantes étaient reparties de façons aléatoires dans la base de données. La base finale avec laquelle nous travaillons est donc constituée de 745 femmes et 73 variables explicatives.

A.1.2 Approches considérées

La variable d'intérêt est une variable binaire (*automédication* **vs** *recours externe*). Un modèle simple et pertinent pour prédire cette variable est le modèle de régression logistique. Comme dans l'étude des foyers à risque (Chapitre 2), le nombre de variables explicatives ici est important, nous procédons en deux étapes : une étape de réduction de dimension et une étape de prédiction par le modèle de régression logistique.

A.1.3 Méthodes de réduction de dimension

Dans un premier temps nous avons utilisé les méthodes de réduction de dimension présentées à la Section *Methods* du Chapitre 2 (voir aussi Section 1.3). Il s'agit du Lasso, Group Lasso (G-L), et des méthodes de réduction via les forêts aléatoires (*RFnested*, *RFtreshold*). En plus de ces méthodes, nous avons utilisé d'autres méthodes que nous décrivons rapidement.

Estimateur ridge

L'estimateur ridge est défini par :

$$\widehat{\beta}_{ridge}(\lambda) = \arg\min_{\beta \in R^p} L_n(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{A.1}$$

où γ_n est l'opposé de la log de vraisemblance défini en 1.2. L'estimateur ridge permet de contourner les problèmes de multicolinéarité même en présence d'un nombre important de variables explicatives $(p > n)$. Le principal défaut de cet estimateur est lié aux difficultés d'interprétation car, aucune sélection de variables n'étant faite, toutes les variables sont

concernées dans le modèle. Nous nous intéressons ici aux approches par pénalisation permettant également une sélection de variables, c'est le cas du Lasso (voir Tibshirani (1996) et aussi Section 1.3.1) et de ses variantes.

Il existe plusieurs variantes de l'estimateur Lasso, chacune d'elles étant proposée pour apporter une amélioration au Lasso dans un contfexte bien particulier. En plus du Group Lasso défini à la Section 1.3.1, nous utilisons ici d'autres variantes du Lasso telles que l'*elastic net*, *adoptive Lasso* et *bolasso*.

Elastic net

Les résultats théoriques qui garantissent la consistance de l'estimateur Lasso portent en général sur une hypothèse de faible corrélation entre les variables. Le Lasso a donc de mauvaises performances en cas de forte multicolinéarité entre les variables explicatives. En effet, lorsque plusieurs variables explicatives sont fortement corrélées, le lasso risque de n'en conserver qu'une. Ce qui masque une partie du phénomène à étudier. Pour pallier à cette faiblesse du Lasso, Zou et Hastie (2005) ont proposé l'*elastic net*, qui est une variante de l' estimateur Lasso utilisant une pénalité proportionnelle à la combinaison linéaire convexe des pénalités ℓ_1 et ℓ_2. L'estimateur *elasctic net* noté $\widehat{\beta}_{elnet}(\lambda)$ est defini par :

$$\widehat{\beta}_{elnet}(\lambda) = \arg\min_{\beta \in R^p} \Big\{ L_n(\beta) + \lambda P_\alpha(\beta) \Big\} \tag{A.2}$$

où

$$P_\alpha(\beta) = \alpha \sum_{j=1}^{p} |\beta_j| + \frac{1}{2}(1-\alpha) \sum_{j=1}^{p} \beta_j^2$$

P_α est une pénalité intermédiaire entre la pénalité ridge ($\alpha = 0$) et la pénalité lasso ($\alpha = 1$). Cette pénalité a l'avantage de sélectionner les variables tout en prenant en compte les corrélations entre celles-ci. En effet le premier terme de pénalité (ℓ_1) assure la sélection de variables c'est à dire la *sparsité* de la solution $\widehat{\beta}_{elnet}(\lambda)$ et le second terme (ℓ_2) permet de prendre en compte la corrélation entre les variables (en encourageant les variables corrélées à être sélectionnées ensemble).

Adaptive Lasso

Dans la méthode Lasso, il est bien connu que, plus le paramètre de régularisation est grand plus le coefficient a de forte chance d'être estimé égal à zéro et inversement, plus le paramètre de régularisation est petit, plus le coefficient a de forte chance d'être estimé différent de zéro. Il est donc judicieux de pénaliser différemment les coefficients du vecteur β : affecter aux coefficients non significatifs une pénalité considérable (un poids important) et aux coefficients significatifs une petite pénalité (un petit poids). Pour $\lambda > 0$ fixé, l'estimateur Adaptive Lasso (A-L) est défini comme suit :

$$\widehat{\beta}_{adap}(\lambda) = \arg\min_{\beta \in R^p} L_n(\beta) + \lambda \sum_{j=1}^{p} \omega_j |\beta_j|.$$

Le problème est que l'on ne connait pas à l'avance les paramètres significatifs. En pratique on utilise généralement les poids $\omega_j = 1/|\hat{\beta}_j|$, où $\hat{\beta}_j$ est soit l'estimateur du maximum de vraisemblance (voir Zou (2006)) soit l'estimateur ridge (voir Section A.1).

Bolasso

Le Bolasso (*BoL*) (voir Bach (2008)) est une méthode qui combine le Bootstrap et le Lasso. Il consiste à appliquer la méthode Lasso sur des échantillons bootstrap et de faire

l'intersection des sous ensembles sélectionnés par chaque méthode Lasso. Il est défini par l'algorithme suivant :

Algorithme 1 : Bolasso

Data : $(X, Y) \in R^{n\times(p+1)}$
Nombre de bootstrap : B;
for k =1 *to* B **do**
  Générer un échantillon bootstrap $(X^{(k)}, Y^{(k)})$;
  Calculer l'estimateur Lasso $\hat{\beta}^{(k)}$ en utilisant $(X^{(k)}, Y^{(k)})$;
  Générer le support $J_k = \{j, \hat{\beta}_j^{(k)} \neq 0\}$;
$J = \bigcap_{k=1}^{B} J_k$;
Estimer $\hat{\beta}_J$ sur (X_J, Y);
Retourner J et $\hat{\beta}_J$

L'algorithme est basé sur le fait que la méthode Lasso sélectionne en général tous les coefficients significatifs, plus quelques coefficients non significatifs. L'intersection de plusieurs sous ensembles sélectionnés par la méthode Lasso permet de réduire le nombre de coefficients non significatifs, car ce ne sont pas les mêmes coefficients non significatifs qui sont sélectionnés par chaque méthode Lasso. Le sous ensemble qui résulte de l'intersection est donc proche du "*vrai sous ensemble*" de coefficients significatifs.

Séparateur à Vaste Marge

Le Séparateur à Vaste Marge (SVM) est une méthode de classification binaire par apprentissage supervisé introduite par Vapnik (2000). Supposons que les données sont des couples $(z_i, Y_i)_{1\leq i\leq n} \in \mathcal{X} \times \{-1, +1\}$ où $\mathcal{X}$ désigne l'espace des variables explicatives souvent pris dans $\mathbb{R}^p$. L'appartenance d'une observation z_i à une classe ou à une autre est matérialisée par la valeur -1 ou +1 de son étiquette Y_i. L'objectif est de trouver une fonction qui permet de classer au mieux les données, c'est à dire une fonction qui, pour une nouvelle observation dont nous avons seulement mesuré z nous pussions prédire sa classe Y. Le séparateur à vaste marge repose sur l'existence d'une fonction de prédiction ($signe(f(z_i))$) dans un espace approprié. Elle s'appuie sur l'utilisation de fonctions appelées noyau qui permettent une séparation optimale des données. En classification linéaire la fonction f est linéaire en z et prend la forme générale suivante :

$$f(z_i) = \langle w, z_i \rangle + b,$$

où $(w, b) \in \mathbb{R}^p \times \mathbb{R}$ sont les paramètres de la fonction de décision f à estimer, et l'opérateur $\langle , \rangle$ est le produit scalaire usuel dans $\mathbb{R}^p$. La règle de décision est donc donnée par $signe(f(z_i))$. Géométriquement ce classifieur divise l'espace des variables explicatives en deux demi espaces correspondant chacun à une classe. Cette séparation est réalisée par l'hyperplan $H_{(w,b)}$ défini par l'équation $\langle w, z_i \rangle + b = 0$. La distance d'un point au plan est donnée par $d(x) = |\langle w, z_i \rangle + b|/||w||$. L'hyperplan optimal est celui pour lequel la distance aux points les plus proches (marge) est maximale. Un développement basé sur un jeu d'échelles montre que l'hyperplan à marge maximale est la solution du problème d'optimisation suivant :

$$\text{Minimiser}_{w,b} ||w||^2, \quad \text{sous la contrainte} \quad Y_i(\langle w, z_i \rangle + b) \geqslant 1, \ i = 1, \ldots, n. \tag{A.3}$$

Notons que ce procédé fait comme hypothèse que les deux classes sont linéairement séparables *i.e.* qu'il existe un hyperplan qui permet de séparer parfaitement les deux classes. Dans le cas non linéairement séparable, l'utilisation des fonctions à noyau permet de plonger les données dans un espace de dimension plus élevée où un séparateur linéaire peut être trouvé. Dans les méthodes à noyaux, on considère la transformation de l'espace des variables explicatives $\mathcal{X}$ en un espace de caractéristiques (feature space, en anglais) $\mathcal{H}$ par une application non linéaire :

$$\mathcal{X} \to \mathcal{H}, z \mapsto \phi(z).$$

La dimension de $\mathcal{H}$ est généralement supérieure à celle de $\mathcal{X}$ et $\mathcal{H}$ est en plus muni d'un produit sclalaire $\langle .,. \rangle_{\mathcal{H}}$. La règle de décision permettant de séparer au mieux les exemples positifs des exemples négatifs est donnée par :

$$r(z) = \text{sign}(\sum_i \alpha_i \langle \phi(z), \phi(z_i) \rangle_{\mathcal{H}} + b)$$

où α_i *et* b sont les paramètres à optimiser. La transformation ϕ est souvent définie par le biais du noyau comme suit :

$$\phi : \mathcal{X} \to \mathcal{H},\ z \mapsto \mathcal{K}(z, .),$$

avec

$$\langle \phi(z), \phi(z') \rangle_{\mathcal{H}} = \mathcal{K}(z, z').$$

La fonction $\mathcal{K}(.,.)$ est appelée noyau. Ainsi, lorsqu'on applique un noyau à deux observations issues de l'espace des variables $\mathcal{X}$, on calcule en fait leur produit scalaire dans l'espace des caractéristiques $\mathcal{H}$. La sélection de variables par la méthode SVM se fait en utilisant la hiérarchie de variables donnée par cette méthode.

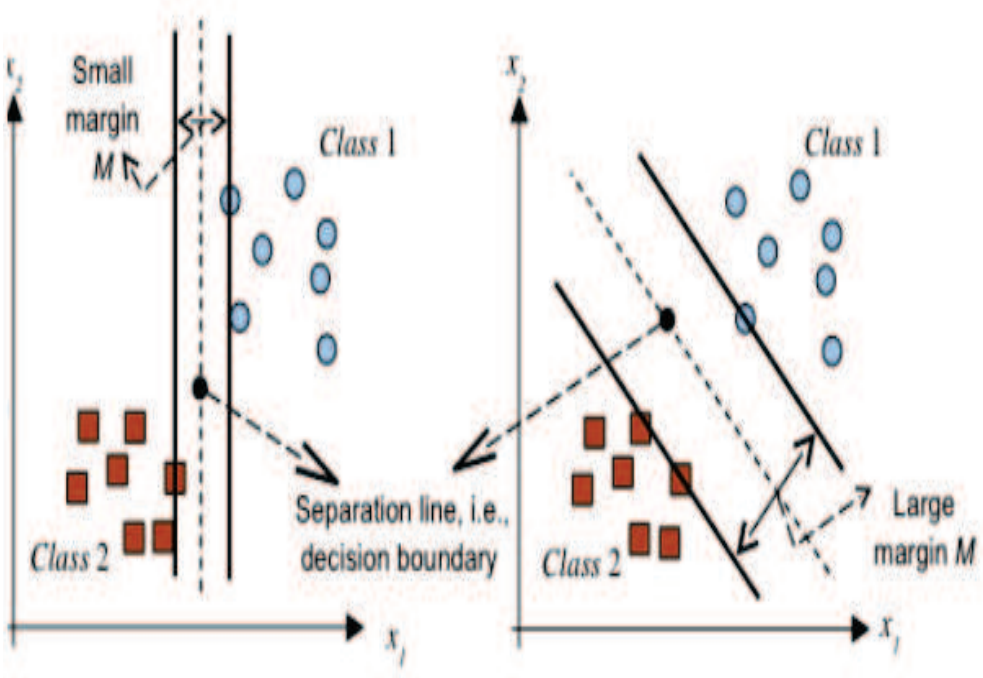


FIGURE A.2 – *Séparateur à vaste marge*

A.1.4 Résultats

L'automédication a été pratiquée par 48,9% des femmes intérogées. La Figure A.4 représente les erreurs OOB des modèles (forêts aléatoires) emboités (*RFnested*). À partir du modèle avec la variable la plus importante, on constate en moyenne une décroissance des erreurs jusqu'au modèle optimal. Ensuite, partant du modèle optimal, l'ajout d'une variable dans le modèle augmente l'erreur de prédiction.

La Table A.1 présente les erreurs de prédiction du modèle de régression logistique qui prend en compte chacun des sous ensembles de variables sélectionnés par chacune des méthodes de réduction de dimension. Sans surprise, le modèle de régression logistique sur chacun des sous ensembles de variables sélectionnés a une erreur de prédiction inférieure à l'erreur du modèle logistique qui prend en compte toutes les variables. Le modèle logistique utilisant le sous ensemble de variables sélectionné par le Group Lasso a l'erreur de prédiction la plus petite (37.1%). Le sous ensemble sélectionné par le Group Lasso est donc optimal.

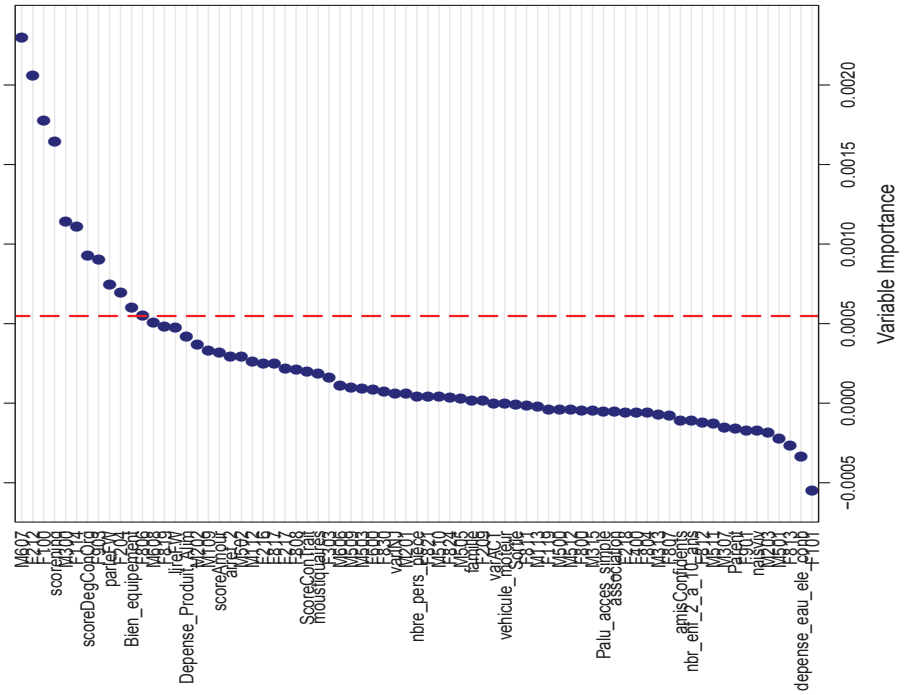


FIGURE A.3 – *Importance des variables*

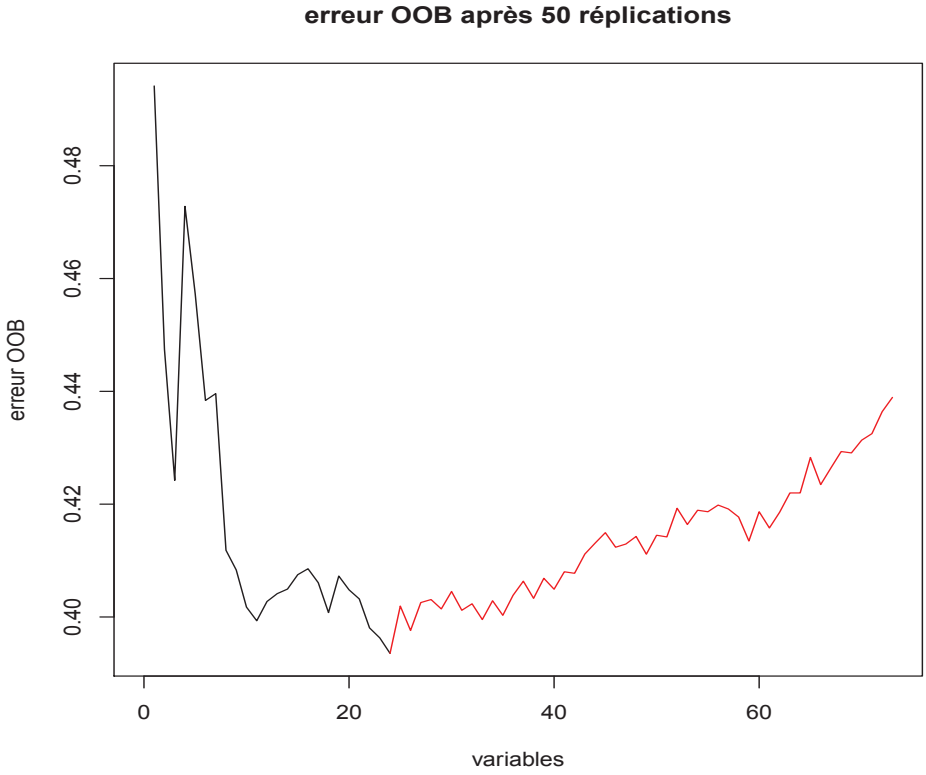


FIGURE A.4 – *Erreur Out Of Bag (OOB) des modèles (forêts aléatoires) emboités où les variables sont introduites par ordre d'importance*

Méthodes	.	Lasso	G-L	A-L	BoL	E-n	*RFtreshold*	*RFnested*	SVM
Erreur (%)	43,5	39.1	**37.1**	38.6	38.2	39	40.2	40	39.3
Nbre variables	73	9	13	6	5	12	12	23	15

TABLE A.1 – *Erreur : erreur de prédiction du modèle logistique qui prend en compte les variables sélectionnées par chaque méthode de réduction de dimension.*

A.1.5 Discussion

Les erreurs de prédiction des différents modèles logistiques prenant en compte les variables sélectionnées par les méthodes de réduction sont proches de 40%. Ces erreurs sont très grandes. Les différents modèles logistiques ne permettent pas "*d'apprendre*" des données.

L'erreur obtenue avec la méthode des forêts aléatoires est proche de 40% (voir Figure A.4). Ce fort taux d'erreur montre que les informations dont nous disposons permettent difficilement de prédire le recours aux soins avec ces données. En effet, les forêts aléatoires sont reconnues pour être robustes et bien adaptées à l'analyse des données complexes (voir Chen et Ishwaran (2012)). Ce constat soulève la question de la pertinence du côté déclaratif de la variable d'intérêt. En d'autres termes, les femmes enquêtées ont elles déclaré le vrai type de recours aux soins en cas de fièvre ? Plusieurs études montrent qu'en cas de fièvre, beaucoup plus de 50% des femmes ont recours a l'automédication, ce qui n'apparait pas dans les données Actu-Palu. La variable d'intérêt semble donc être mal déclarée. En effet, les personnes interrogées, guidées par la volonté de paraître ou par une démarche calculée, ont souvent tendance à donner des réponses erronées. Il est donc important de réfléchir à une autre approche pour questionner le recours aux soins.

Bibliographie

H. Akaike. Information theory and an extension of the maximum likelihood principle. Dans *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973. (Cité pages 8 et 89.)

Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6) :716–723, 1974. (Cité pages 8 et 55.)

Sylvain Arlot et Pascal Massart. Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research*, 10 :245–279, 2009. (Cité pages 34, 90, 99 et 100.)

F.R. Bach. Bolasso : model consistent lasso estimation through the bootstrap. Dans *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008. (Cité page 127.)

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4 :384–414, 2010. ISSN 1935-7524. URL http://dx.doi.org/10.1214/09-EJS521. (Cité pages 26, 56, 63, 66, 83 et 89.)

Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493, 2000. ISSN 0178-8051. URL http://dx.doi.org/10.1007/PL00008731. (Cité pages 21, 89 et 112.)

Jean-Patrick Baudry, Cathy Maugis, et Bertrand Michel. Slope heuristics : overview and implementation. *Stat. Comput.*, 22(2) :455–470, 2012. ISSN 0960-3174. URL http://dx.doi.org/10.1007/s11222-011-9236-1. (Cité pages 34, 99 et 100.)

Gérard Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13 :1063–1095, 2012. ISSN 1532-4435. (Cité pages 12 et 122.)

Gérard Biau, Luc Devroye, et Gábor Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9 :2015–2033, 2008. ISSN 1532-4435. (Cité page 12.)

Peter J. Bickel, Ya'acov Ritov, et Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics*, 37(4) :1705–1732, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/08-AOS620. (Cité pages 16, 17, 18, 27, 53, 56, 60, 64, 65 et 66.)

Lucien Birgé. Model selection for density estimation with $\mathbb{L}_2$-loss. *Probab. Theory Related Fields*, 158(3-4) :533–574, 2014a. ISSN 0178-8051. URL http://dx.doi.org/10.1007/s00440-013-0488-x. (Cité page 21.)

Lucien Birgé. Model selection for density estimation with $\mathbb{L}_2$-loss. *Probab. Theory Related Fields*, 158(3-4) :533–574, 2014b. ISSN 0178-8051. URL http://dx.doi.org/10.1007/s00440-013-0488-x. (Cité page 89.)

Lucien Birgé et Pascal Massart. Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3) :329–375, 1998. ISSN 1350-7265. URL http://dx.doi.org/10.2307/3318720. (Cité page 96.)

Lucien Birgé et Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268, 2001. ISSN 1435-9855. URL http://dx.doi.org/10.1007/s100970100031. (Cité pages 5, 19, 21, 22, 30, 87, 89, 92, 122 et 146.)

Lucien Birgé et Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007. ISSN 0178-8051. URL http://dx.doi.org/10.1007/s00440-006-0011-8. (Cité pages 5, 19, 21, 30, 32, 33, 34, 90, 94, 97, 98 et 99.)

Mélanie Blazère, Jean-Michel Loubes, et Fabrice Gamboa. Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. *IEEE Trans. Inform. Theory*, 60(4) :2303–2318, 2014. ISSN 0018-9448. URL http://dx.doi.org/10.1109/TIT.2014.2303121. (Cité pages 29, 56 et 60.)

Dominique Bontemps et Wilson Toussile. Clustering and variable selection for categorical multivariate data. *Electron. J. Stat.*, 7 :2344–2371, 2013. ISSN 1935-7524. URL http://dx.doi.org/10.1214/13-EJS844. (Cité pages 34 et 99.)

S. Boucheron, G. Lugosi, et O. Bousquet. Concentration inequalities. *Advanced Lectures on Machine Learning*, pages 208–240, 2004. (Cité pages 81 et 119.)

J. V. Braun, R. K. Braun, et H.-G. Müller. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2) :301–314, 2000. ISSN 0006-3444. URL http://dx.doi.org/10.1093/biomet/87.2.301. (Cité pages 21 et 89.)

Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001. (Cité pages 12, 13, 38 et 43.)

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, et Charles J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984. ISBN 0-534-98053-8 ; 0-534-98054-6. (Cité page 12.)

B Bull, Shelley, Lewinger Juan, Pablo, et Lee Sophia, SF. Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine*, 26 :903–918, 2007. (Cité pages 4 et 41.)

Florentina Bunea. Consistent selection via the Lasso for high dimensional approximating regression models. Dans *Pushing the limits of contemporary statistics : contributions in honor of Jayanta K. Ghosh*, volume 3 de *Inst. Math. Stat. Collect.*, pages 122–137. Inst. Math. Statist., Beachwood, OH, 2008a. URL http://dx.doi.org/10.1214/074921708000000101. (Cité page 16.)

Florentina Bunea. Honest variable selection in linear and logistic regression models via l_1 and $ł_1 + ł_2$ penalization. *Electron. J. Stat.*, 2 :1153–1194, 2008b. ISSN 1935-7524. URL http://dx.doi.org/10.1214/08-EJS287. (Cité pages 26 et 89.)

Florentina Bunea, Alexandre Tsybakov, et Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1 :169–194, 2007a. ISSN 1935-7524. URL http://dx.doi.org/10.1214/07-EJS008. (Cité pages 16, 18, 56 et 64.)

Florentina Bunea, Alexandre B. Tsybakov, et Marten H. Wegkamp. Aggregation and sparsity via l_1 penalized least squares. Dans *Learning theory*, volume 4005 de *Lecture Notes in Comput. Sci.*, pages 379–391. Springer, Berlin, 2006. URL http://dx.doi.org/10.1007/11776420_29. (Cité pages 16, 18, 56 et 64.)

Florentina Bunea, Alexandre B. Tsybakov, et Marten H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35(4) :1674–1697, 2007b. ISSN 0090-5364. URL http://dx.doi.org/10.1214/009053606000001587. (Cité pages 16, 18, 56, 59 et 64.)

Alexandre Bureau, Josée Dupuis, Kathleen Falls, Kathryn L Lunetta, Brooke Hayward, Tim P Keith, et Paul Van Eerdewegh. Identifying snps predictive of phenotype using random forests. *Genetic epidemiology*, 28(2) :171–182, 2005. (Cité pages 23 et 38.)

G. Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8) :2052–2060, 2003a. ISSN 0018-9448. URL http://dx.doi.org/10.1109/TIT.2003.814485. (Cité pages 104, 108 et 112.)

Gwénaëlle Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8) :2052–2060, 2003b. ISSN 0018-9448. URL http://dx.doi.org/10.1109/TIT.2003.814485. (Cité pages 21 et 89.)

Scott Shaobing Chen, David L. Donoho, et Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1) :33–61, 1998. ISSN 1064-8275. URL http://dx.doi.org/10.1137/S1064827596304010. (Cité page 9.)

Xi Chen et Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6) : 323–329, 2012. (Cité pages 13 et 131.)

Ch. Chesneau et M. Hebiri. Some theoretical results on the grouped variables Lasso. *Math. Methods Statist.*, 17(4) :317–326, 2008. ISSN 1066-5307. URL http://dx.doi.org/10.3103/S1066530708040030. (Cité pages 18 et 56.)

Dennis D. Cox et Finbarr O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.*, 18(4) :1676–1695, 1990. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1176347872. (Cité page 89.)

Sijmen de Jong. Simpls : an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18 :251–263, 1993. (Cité page 38.)

A. Diallo, S. Dos Santos, R. Lalou, et J.-Y. Le Hesran. Perceived malaria in the population of an urban setting : a skipped reality in dakar, senegal. *Malaria Journal*, 11(1) :340, 2012. (Cité page 39.)

R Dìaz-Uriarte et A De Andres, S. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7 :3, 2006. (Cité pages 23 et 38.)

Ramón Díaz-Uriarte et Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1) :3, 2006. (Cité page 13.)

A. Dillo, N.-T. Ndam, A. Moussiliou, S. Dos Santos, A. Ndonky, M. Borderon, S. Oliveau, R. Lalou, et J.-Y. Le Hesran. Asymptomatic carriage of plasmodium in urban dakar : The risk of malaria should not be underestimated. *PLoS ONE*, 7(2), 2012. (Cité page 39.)

Annette J. Dobson. *An introduction to generalized linear models*. Chapman and Hall Ltd., London, 1990. ISBN 0-412-31100-3. Second edition of 1t Introduction to statistical modelling. (Cité page 5.)

N. R. Draper et H. Smith. *Applied regression analysis*. John Wiley & Sons Inc., New York, 1966. (Cité page 5.)

Sandrine Dudoit, Jane Fridlyand, et Terence P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97 :77–87, 2002. (Cité pages 22 et 37.)

C. Durot, E. Lebarbier, et A.-S. Tocquet. Estimating the joint distribution of independent categorical variables via model selection. *Bernoulli*, 15(2) :475–507, 2009. ISSN 1350-7265. URL http://dx.doi.org/10.3150/08-BEJ155. (Cité pages 21, 89 et 105.)

Jalal Fadili, Gabriel Peyré, Charles-Alban Deledalle, et Samuel Vaiter. The degrees of freedom of the group lasso. *preprint arXiv :1205.1481*, 2012. (Cité page 11.)

Ludwig Fahrmeir et Heinz Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368, 1985. (Cité page 6.)

Jianqing Fan, Mark Farmen, et Irène Gijbels. Local maximum likelihood estimation and inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(3) :591–608, 1998. ISSN 1369-7412. URL http://dx.doi.org/10.1111/1467-9868.00142. (Cité page 89.)

Mark William Farmen. *The smoothed bootstrap for variable bandwidth selection and some results in nonparametric logistic regression*. ProQuest LLC, Ann Arbor, MI, 1996. URL http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:9631903. Thesis (Ph.D.)–The University of North Carolina at Chapel Hill. (Cité page 89.)

J. Friedman, T. Hastie, et R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1) :1, 2010. (Cité pages 9, 42 et 55.)

Manuel Garcia-Magariños, Anestis Antoniadis, Ricardo Cao, et Wenceslao González-Manteiga. Lasso logistic regression, GSoft and the cyclic coordinate descent algorithm : application to gene expression data. *Stat. Appl. Genet. Mol. Biol.*, 9 :Art. 30, 30, 2010. ISSN 1544-6115. URL http://dx.doi.org/10.2202/1544-6115.1536. (Cité pages 23, 38 et 55.)

Debashis Ghosh et M Chinnaiyan, Arul. Classification and selection of biomarkers in genomic data using lasso. *BioMed Research International*, 2 :147–154, 2005. (Cité pages 38 et 47.)

Geof H Givens et Jennifer A Hoeting. *Computational statistics*, volume 708. John Wiley & Sons, 2012. (Cité page 6.)

Benjamin A Goldstein, Alan E Hubbard, Adele Cutler, et Lisa F Barcellos. An application of random forests to a genome-wide association dataset : Methodological considerations & new findings. *BMC genetics*, 11(1) :49, 2010. (Cité pages 23 et 38.)

Benjamin A Goldstein, Eric C Polley, et Farren Briggs. Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011. (Cité pages 23 et 38.)

Christian Gourieroux et Alain Monfort. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics*, 17(1) :83–97, 1981. (Cité page 6.)

Sander Greenland, A Schwartzbaum, J, et D Finkle, W. Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology*, 151 :531–539, 2000. (Cité pages 4 et 41.)

Andreas Groll et Gerhard Tutz. Variable selection for generalized linear mixed models by l 1-penalized estimation. *Statistics and Computing*, pages 1–18, 2012. (Cité page 121.)

T Hastie. Non-parametric logistic regression. *SLAC PUB-3160, June*, 1983. (Cité pages 24, 55 et 89.)

Tim Hesterberg, Nam Hee Choi, Lukas Meier, et Chris Fraley. Least angle and l_1 penalized regression : a review. *Stat. Surv.*, 2 :61–93, 2008. ISSN 1935-7516. URL http://dx.doi.org/10.1214/08-SS035. (Cité page 10.)

Joseph M. Hilbe. *Logistic regression models*. Chapman & Hall/CRC Texts in Statistical Science Series. CRC Press, Boca Raton, FL, 2009. ISBN 978-1-4200-7575-5. (Cité pages 5 et 40.)

David W Hosmer Jr, Stanley Lemeshow, et Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013. (Cité page 5.)

J. Huang, S. Ma, et CH Zhang. The iterated lasso for high–dimensional logistic regression. *Technical Report 392*, 2008. (Cité pages 26 et 56.)

Jian Huang, Joel L. Horowitz, et Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4) :2282–2313, 2010. ISSN 0090-5364. URL http://dx.doi.org/10.1214/09-AOS781. (Cité pages 18 et 56.)

Junzhou Huang et Tong Zhang. The benefit of group sparsity. *Ann. Statist.*, 38(4) :1978–2004, 2010. ISSN 0090-5364. URL http://dx.doi.org/10.1214/09-AOS778. (Cité page 18.)

Junzhou Huang, Tong Zhang, et Dimitris Metaxas. Learning with structured sparsity. *J. Mach. Learn. Res.*, 12 :3371–3412, 2011. ISSN 1532-4435. URL http://dx.doi.org/10.1145/1553374.1553429. (Cité page 10.)

Laurent Jacob, Guillaume Obozinski, et Jean-Philippe Vert. Group lasso with overlap and graph lasso. Dans *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009. (Cité pages 10 et 122.)

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et Francis Bach. Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.*, 12 :2297–2334, 2011. ISSN 1532-4435. (Cité page 10.)

I. T. Jolliffe. *Principal component analysis*. Springer Series in Statistics. Springer-Verlag, New York, second édition, 2002. ISBN 0-387-95442-2. (Cité page 38.)

Keith Knight et Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5) :1356–1378, 2000. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1015957397. (Cité pages 16 et 56.)

Mladen Kolar, John Lafferty, et Larry Wasserman. Union support recovery in multi-task learning. *J. Mach. Learn. Res.*, 12 :2415–2435, 2011. ISSN 1532-4435. (Cité page 18.)

M. Kwemou. Non-asymptotic oracle inequalities for the lasso and group lasso in high dimensional logistic model. *preprint arXiv :1206.0710*, 2012. (Cité pages 31, 89 et 107.)

Émilie Lebarbier. Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing*, 85(4) :717–736, 2005. (Cité pages 21, 34, 89 et 99.)

A Legarra, C Robert-GraniÃ©, P Croiseau, F Guillaume, et Fritz. Improved lasso for genomic selection. *Genetics research*, 20 :77, 2011. (Cité pages 23 et 38.)

Chenlei Leng, Yi Lin, et Grace Wahba. A note on the lasso and related procedures in model selection. *Statist. Sinica*, 16(4) :1273–1284, 2006. ISSN 1017-0405. (Cité page 10.)

Matthieu Lerasle. Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(3) :884–908, 2012. ISSN 0246-0203. URL http://dx.doi.org/10.1214/11-AIHP425. (Cité pages 34 et 99.)

Jiahan Li, Kiranmoy Das, Guifang Fu, Runze Li1, et Rongling Wu. The bayesian lasso for genome-wide association studies. *Bioinformatics*, 27 :516–523, 2011. (Cité pages 23, 38 et 47.)

Andy Liaw et Matthew Wiener. Classification and regression by randomforest. *R news*, 2 (3) :18–22, 2002. (Cité page 13.)

K. Lounici, M. Pontil, A.B. Tsybakov, et S. Van De Geer. Taking advantage of sparsity in multi-task learning. *In COLT'09*, 2009. (Cité pages 18, 56 et 60.)

Karim Lounici, Massimiliano Pontil, Sara van de Geer, et Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39(4) :2164–2204, 2011. ISSN 0090-5364. URL http://dx.doi.org/10.1214/11-AOS896. (Cité pages 18, 19, 29, 30, 56, 60 et 62.)

Fan Lu. *Regularized nonparametric logistic regression and kernel regularization*. ProQuest LLC, Ann Arbor, MI, 2006. ISBN 978-0542-88702-4. URL http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:3234696. Thesis (Ph.D.)–The University of Wisconsin - Madison. (Cité page 89.)

Tak K Mak. Solving non-linear estimation equations. *Journal of the Royal Statistical Society. Series B. Methodological*, 55(4) :945–955, 1993. (Cité page 6.)

Colin L Mallows. Some comments on c p. *Technometrics*, 15(4) :661–675, 1973. (Cité pages 8 et 20.)

Michael Marmot et Richard Wilkinson. *Social determinants of health*. Oxford University Press, 2005. (Cité page 37.)

Pascal Massart. *Concentration inequalities and model selection*, volume 1896 de *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4 ; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. (Cité pages 81, 85, 92, 110 et 113.)

Pascal Massart et Caroline Meynet. The Lasso as an ℓ_1-ball model selection procedure. *Electronic Journal of Statistics*, 5 :669–687, 2011. ISSN 1935-7524. URL http://dx.doi.org/10.1214/11-EJS623. (Cité pages 16 et 56.)

William F Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60 :234–256, 1965. (Cité page 38.)

Cathy Maugis et Bertrand Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.*, 15 :41–68, 2011. ISSN 1292-8100. URL http://dx.doi.org/10.1051/ps/2009004. (Cité page 99.)

J. McAuley, J. Ming, D. Stewart, et P. Hanna. Subband correlation and robust speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 13(5) :956–964, 2005. (Cité page 56.)

Mark McCarthy. Social determinants and inequalities in urban health. *Reviews on environmental health*, 15(1-2) :97–108, 2000. (Cité page 37.)

P. McCullagh et J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1983. ISBN 0-412-23850-0. (Cité pages 5 et 122.)

Lukas Meier, Sara van de Geer, et Peter Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, 70(1) :53–71, 2008. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x. (Cité pages 10, 29, 30, 42, 46, 56, 58, 60, 61, 67 et 68.)

Lukas Meier, Sara van de Geer, et Peter Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B) :3779–3821, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/09-AOS692. (Cité pages 18 et 56.)

Nicolai Meinshausen et Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3) :1436–1462, 2006. ISSN 0090-5364. URL http://dx.doi.org/10.1214/009053606000000281. (Cité pages 16 et 56.)

Nicolai Meinshausen et Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1) :246–270, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/07-AOS582. (Cité pages 16 et 56.)

Scott Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002. (Cité pages 5 et 40.)

Yan A Meng, Yi Yu, L Adrienne Cupples, Lindsay A Farrer, et Kathryn L Lunetta. Performance of random forest when snps are in linkage disequilibrium. *BMC bioinformatics*, 10 (1) :78, 2009. (Cité pages 23 et 38.)

Yuval Nardi et Alessandro Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.*, 2 :605–633, 2008. ISSN 1935-7524. URL http://dx.doi.org/10.1214/08-EJS200. (Cité pages 18, 56 et 62.)

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, et Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4) :538–557, 2012. ISSN 0883-4237. URL http://dx.doi.org/10.1214/12-STS400. (Cité pages 29, 56 et 60.)

Yurii Nesterov et Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13 de *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. ISBN 0-89871-319-6. URL http://dx.doi.org/10.1137/1.9781611970791. (Cité page 83.)

Danh V Nguyen et David M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18 :39–50, 2002. (Cité pages 23 et 38.)

Guillaume Obozinski, Ben Taskar, et Michael I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Stat. Comput.*, 20(2) :231–252, 2010. ISSN 0960-3174. URL http://dx.doi.org/10.1007/s11222-008-9111-x. (Cité page 18.)

Danielle C Ompad, Sandro Galea, Waleska T Caiaffa, et David Vlahov. Social determinants of the health of urban populations : methodologic considerations. *Journal of Urban Health*, 84(1) :42–53, 2007. (Cité page 37.)

M. R. Osborne, Brett Presnell, et B. A. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3) :389–403, 2000. ISSN 0272-4979. URL http://dx.doi.org/10.1093/imanum/20.3.389. (Cité pages 16 et 56.)

Mee Young Park et Trevor Hastie. L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B*, 69(4) :659–677, 2007. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2007.00607.x. (Cité pages 9, 42 et 55.)

Nandini Raghavan. *Bayesian inference in nonparametric logistic regression*. ProQuest LLC, Ann Arbor, MI, 1993. URL http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:9411757. Thesis (Ph.D.)–University of Illinois at Urbana-Champaign. (Cité page 89.)

Pradeep Ravikumar, John Lafferty, Han Liu, et Larry Wasserman. Sparse additive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5) :1009–1030, 2009. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2009.00718.x. (Cité pages 18 et 56.)

Claire Rondet, Marion Soler, Virginie Ringa, Isabelle Parizot, et Pierre Chauvin. The role of a lack of social integration in never having undergone breast cancer screening : Results from a population-based, representative survey in the paris metropolitan area in 2010. *Preventive medicine*, 57(4) :386–391, 2013. (Cité page 47.)

Jürg Schelldorfer, Peter Bühlmann, GEER DE, et SARA VAN. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38(2) :197–214, 2011. (Cité page 121.)

Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2) :461–464, 1978a. ISSN 0090-5364. (Cité pages 8 et 55.)

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978b. (Cité page 89.)

Caroline Strobl, J. Malley, et G. Tutz. An introduction to recursive partitioning : Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, 14(4) :323–348, 2009. (Cité pages 14, 43, 44 et 122.)

Bernadetta Tarigan et Sara A. van de Geer. Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli*, 12(6) :1045–1076, 2006. ISSN 1350-7265. URL http://dx.doi.org/10.3150/bj/1165269150. (Cité page 59.)

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1) :267–288, 1996. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1996)58:1<267:RSASVT>2.0.CO;2-G&origin=MSN. (Cité pages 9, 16, 30, 38, 41, 42, 55 et 127.)

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, et Keith Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1) :91–108, 2005. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00490.x. (Cité pages 10 et 122.)

Ryan J Tibshirani, Jonathan Taylor, et al. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2) :1198–1232, 2012. (Cité page 11.)

Samuel Vaiter, Charles Deledalle, Gabriel Peyré, Jalal Fadili, et Charles Dossal. The degrees of freedom of the group lasso for a general design. *arXiv preprint arXiv :1212.6478*, 2012. (Cité page 11.)

Julie Vallée, Pierre Chauvin, et al. Investigating the effects of medical density on health-seeking behaviours using a multiscale approach to residential and activity spaces : Results from a prospective cohort study in the paris metropolitan area, france. *International journal of health geographics*, 11(1) :54, 2012. (Cité page 47.)

Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2) :614–645, 2008. ISSN 0090-5364. URL http://dx.doi.org/10.1214/009053607000000929. (Cité pages 26, 27, 56, 65 et 89.)

Sara A. van de Geer et Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3 :1360–1392, 2009. ISSN 1935-7524. URL http://dx.doi.org/10.1214/09-EJS506. (Cité pages 18 et 64.)

V.N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000. (Cité page 128.)

A. Vexler et G. Gurevich. Guaranteed local maximum likelihood detection of a change point in nonparametric logistic regression. *Comm. Statist. Theory Methods*, 35(4-6) :711–726, 2006. ISSN 0361-0926. URL http://dx.doi.org/10.1080/03610920500498923. (Cité page 89.)

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5) : 2183–2202, 2009. ISSN 0018-9448. URL http://dx.doi.org/10.1109/TIT.2009.2016018. (Cité page 16.)

Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, et Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25 :714–721, 2009a. (Cité pages 23, 38 et 47.)

T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel, et K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6) :714–721, 2009b. (Cité page 55.)

Yuhong Yang. Model selection for nonparametric regression. *Statist. Sinica*, 9(2) :475–499, 1999. ISSN 1017-0405. (Cité pages 21 et 89.)

Tjalling J Ypma. Historical development of the newton-raphson method. *SIAM review*, 37 (4) :531–551, 1995. (Cité page 6.)

Ming Yuan et Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1) :49–67, 2006. ISSN 1369-7412. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x. (Cité pages 10, 18, 42 et 58.)

Cun-Hui Zhang et Jian Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Annals of Statistics*, 36(4) :1567–1594, 2008. ISSN 0090-5364. URL http://dx.doi.org/10.1214/07-AOS520. (Cité pages 16 et 56.)

Peng Zhao et Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7 : 2541–2563, 2006. ISSN 1532-4435. (Cité pages 16 et 56.)

Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101 (476) :1418–1429, 2006. ISSN 0162-1459. URL http://dx.doi.org/10.1198/016214506000000735. (Cité pages 16, 26, 56 et 127.)

Hui Zou et Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320, 2005. (Cité pages 10, 122 et 127.)

Hui Zou, Trevor Hastie, et Robert Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5) :2173–2192, 2007. ISSN 0090-5364. URL http://dx.doi.org/10.1214/009053607000000127. (Cité page 11.)

Notations

$\mathbb{N}, \mathbb{N}^*$	ensemble des entiers naturels, des entiers strictement positifs
$\mathbb{R}, \mathbb{R}_+$	ensembles des réels et des réels positifs
$\mathbb{R}^d$	ensemble des vecteurs réels à d dimensions
$\mathbb{P}, \mathbb{E}$	probabilité et espérance
$\|E\|$	cardinal de l'ensemble E
Z^T	transposé du vecteur Z

Ce document a été préparé à l'aide de l'éditeur de texte GNU Emacs et du logiciel de composition typographique LaTeX 2_ε.

Titre Réduction de dimension en régression logistique, application aux données Actu-Palu

Résumé Cette thèse est consacrée à la sélection de variables ou de modèle en régression logistique. Elle peut être divisée en deux parties, une partie appliquée et une partie méthodologique. La partie appliquée porte sur l'analyse des données d'une grande enquête socio-épidémiologique dénommée Actu-Palu. Ces grandes enquêtes socio-épidémiologiques impliquent généralement un nombre considérable de variables explicatives. Le contexte est par nature dit de grande dimension. En raison du *fléau de la dimension*, le modèle de régression logistique n'est pas directement applicable. Nous procédons en deux étapes, une première étape de réduction du nombre de variables par les méthodes Lasso, Group Lasso et forêts aléatoires. La deuxième étape consiste à appliquer le modèle logistique au sous-ensemble de variables sélectionné à la première étape. Ces méthodes ont permis de sélectionner les variables pertinentes pour l'identification des foyers à risque d'avoir un épisode fébrile chez un enfant de 2 à 10 ans dans Dakar.
La partie méthodologique, composée de deux sous-parties, porte sur l'établissement de propriétés théoriques d'estimateurs dans le modèle de régression logistique non paramétrique. Ces estimateurs sont obtenus par maximum de vraisemblance pénalisé, dans un cas avec une pénalité de type Lasso ou Group Lasso et dans l'autre cas avec une pénalité de type ℓ_0. Dans un premier temps, nous proposons des versions pondérées des estimateurs Lasso et Group Lasso pour le modèle logistique non paramétrique. Nous établissons des inégalités oracles non asymptotiques pour ces estimateurs. Un deuxième ensemble de résultats vise à étendre le principe de sélection de modèle introduit par Birgé et Massart (2001) à la régression logistique. Cette sélection se fait via des critères du maximum de vraisemblance pénalisé. Nous proposons dans ce contexte des critères de sélection de modèle, et nous établissons des inégalités oracles non asymptotiques pour les estimateurs sélectionnés. La pénalité utilisée, dépendant uniquement des données, est calibrée suivant l'idée de l'heuristique de pente. Tous les résultats de la partie méthodologique sont illustrés par des études de simulations numériques.

Mots-clés Régression logistique, Lasso, Group Lasso, forêts aléatoires, sélection de modèle, inégalités oracles, heuristique de pente

Title Dimension reduction in logistic regression model, application to Actu-Palu data

Abstract This thesis is devoted to variables selection or model selection in logistic regression. It can be divided into two parts, an applied part and a methodological part. The applied part focuses on the analysis of data from a large socioepidemiological survey, called Actu-Palu. These large socioepidemiological surveys typically involve a considerable number of explanatory variables. This is well known as high-dimensional setting. Due to the *curse of dimensionality*, logistic regression model is no longer reliable. We proceed in two steps, a first step of reducing the number of variables by the Lasso, Group Lasso and random forests methods. The second step is to apply the logistic model to the sub-set of variables selected in the first step. These methods have helped to select relevant variables for the identification of households at risk of having febrile episode amongst children from 2 to 10 years old in Dakar. In the methodological part, as a first step, we propose weighted versions of Lasso and Group Lasso estimators for nonparametric logistic model. We prove non asymptotic oracle inequalities for these estimators. Secondly we extend the model selection principle introduced by Birgé and Massart (2001) to logistic regression model. This selection is done using penalized maximum likelihood criteria. We propose in this context a completely data-driven criteria based on the slope heuristics. We prove non asymptotic oracle inequalities for selected estimators. The results of the methodological part are illustrated through simulation studies.

Keywords Logistic regression, Lasso, Group Lasso, random forests, model selection, oracle inequality, slope heuristics