Thresholding estimators for high-dimensional data: model selection, testing and existence

Giacobino, Caroline Laura

Université de Genève
Section de mathématiques

Faculté des Sciences
Professeur Sylvain Sardy

Thresholding Estimators for High-Dimensional Data: Model Selection, Testing and Existence

Thèse

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention statistique

par

Caroline Giacobino

de

Corsier (GE)

Thèse n° 5054

Genève
2017

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès Sciences
Mention statistique

Thèse de *Madame Caroline GIACOBINO*

intitulée :

"Thresholding Estimators for High-Dimensional Data: Model Selection, Testing and Existence"

La Faculté des sciences, sur le préavis de Monsieur S. SARDY, professeur associé et directeur de thèse (Section de mathématiques), Madame E. CANTONI, professeure associée (Geneva School of Economics and Management), Madame D. PICARD, professeure (Laboratoire de probabilités et modèles aléatoires, Université Paris-Diderot Paris 7, Paris, France) et Monsieur N. HENGARTNER, docteur (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, New Mexico, USA), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 27 février 2017

Thèse - 5054 -

Le Décanat

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Acknowledgements

This thesis would not have been possible without the help, support and guidance of many people.

First of all, I would like to express my gratitude to my advisor, Sylvain Sardy, for his advice, patience, and giving me freedom to conduct my own research.

I wish to thank Professors Eva Cantoni, Nicolas Hengartner and Dominique Picard for kindly agreeing to serve on my thesis committee.

I also wish to thank Nicolas Hengartner for his hospitality during my visits to the Los Alamos National Laboratory. I truly appreciated his enthusiasm as well as our conversations.

I would like to thank Jairo Diaz Rodriguez for producing figures and tables in Section 3.4 and Appendix B. Thanks to Pascaline Descloux and Máté Lehel Juhász for carefully reading my thesis and providing valuable feedback.

I acknowledge the support provided by the Swiss National Science Foundation and by the Swiss Doctoral Program in Statistics and Applied Probability for traveling to several conferences in Switzerland and abroad.

I am grateful to the secretaries and librarians from the Section de Mathématiques for assisting me in many different ways.

More personally, I would like to thank all my friends from the Section de Mathématiques for making these years such an enjoyable experience and helping me keep things in perspective.

A special thanks goes to my friends Helen, Luc and Philippe for coming to my thesis defense. I will always remember our first years at university. I also wish to thank Alvaro for his support and help from a very long distance.

Thank you to my family and parents who have always shown love and support in what I do. They have never pushed me into any particular career path and I have always appreciated that.

Finally, thank you Frédéric for your endless support, encouragement and for all that you have done. You believed in me, no matter what.

Contents

Résumé

Dans des domaines tels que la génomique, la finance et la classification d'images, pour n'en citer que quelques-uns, les exemples sont fréquents où le nombre de paramètres d'un modèle est considérablement plus important que la taille de l'échantillon. Cela suscite un intérêt grandissant pour l'analyse de données de haute dimension. L'estimateur du maximum de vraisemblance n'est plus approprié dans ces cas et des contraintes additionnelles de régularisation, reflétant des propriétés ou des suppositions spécifiques au domaine, doivent être introduites.

Dans cette thèse, nous considérons une classe de techniques de régularisation, dites de *seuillage*, qui suppose qu'une certaine transformation $\boldsymbol{\xi}^*$ du vecteur des paramètres est sparse, ce qui signifie que seules quelques entrées sont différentes de zero. L'hypothèse de parcimonie est naturelle dans de nombreuses applications telles que les données sur l'expression génique produites par des puces à ADN où l'on suppose que seul un petit nombre de gènes sont responsables d'un certain sous-type de tumeur.

Ces techniques sont indexées par un paramètre de régularisation positif λ qui gouverne le niveau de sparsité de l'estimé $\hat{\boldsymbol{\xi}}_\lambda$. Cela soulève la question du choix du paramètre de seuillage afin de ne retenir que les variables « correctes ».

Nous montrons d'abord que le vecteur des paramètres estimés est identiquement nul pour de nombreux estimateurs de seuillage lorsque λ est plus grand ou égal à une certaine statistique, la *fonction zero-thresholding*, et dérivons une forme explicite. Cela nous conduit à l'introduction du *quantile universal threshold*, une méthodologie de sélection du paramètre de seuillage qui suit la même logique dans divers domaines. Des inégalités oracles de sparsité sont dérivées et une étude de simulation effectuée afin de montrer l'efficacité de notre approche en ce qui concerne la sélection de modèles et la prédiction pour les modèles linéaires généralisés.

Nous introduisons ensuite une nouvelle classe de procédures de test pour les modèles linéaires, les *tests de seuillage*, qui sont basés sur les estimateurs de seuillage. Une étude comparative de la puissance est conduite pour des alternatives sparses

et denses et une procédure de test adaptative est suggérée. De plus, nous montrons qu'une propriété de screening peut être satisfaite.

Finalement, nous démontrons l'existence de nombreux estimateurs définis comme un minimiseur de la somme d'une fonction de perte et de pénalité, toutes deux convexes. Nous dérivons des conditions nécessaires et suffisantes pour l'existence d'estimateurs régularisés dans les modèles linéaires généralisés lorsque certains paramètres ne sont pas pénalisés du fait qu'ils sont supposés non nuls. De plus, une procédure numérique permettant de vérifier l'existence de tels estimateurs en régression logistique et régression de Poisson est décrite.

Summary

Many real world examples in which the number of features of the model can be dramatically larger than the sample size have been identified in various domains such as genomics, finance and image classification, to name a few. This has led to increased interest in high-dimensional data analysis. The maximum likelihood principle fails in those instances and additional regularization constraints, reflecting properties or assumptions specific to the domain, must be introduced.

In this thesis, we consider a class of regularization techniques, called *thresholding*, which assumes a certain transform $\boldsymbol{\xi}^*$ of the parameter vector is sparse, meaning it has only few nonzero coordinates. The parsimony assumption is natural in many applications such as in gene expression microarrays where few genes are believed to be responsible for a particular tumor subtype.

These techniques are indexed by a nonnegative tuning parameter λ which governs the sparsity level of the estimate $\hat{\boldsymbol{\xi}}_\lambda$. This raises the question of how to choose the thresholding parameter in order to select the "right" variables.

We first show that many thresholding estimators set the estimated coefficients to the null vector for all threshold parameters greater than a certain statistic, the *zero-thresholding function*, and derive an explicit formulation. This leads us to the introduction of the *quantile universal threshold*, a tuning parameter selection methodology which follows the same paradigm in various domains. Sparsity oracle inequalities are derived and a simulation study conducted to show the effectiveness of our approach in terms of model selection and prediction in generalized linear models.

We then introduce a new class of testing procedures in linear models, *thresholding tests*, which are based on thresholding estimators. A comparative power study is performed under sparse and dense alternatives and an adaptive testing procedure is suggested. Moreover, we show that a screening property can be achieved.

Finally, we prove existence of a large class of thresholding estimators defined as a minimizer of the sum of a loss and a penalty function, both convex. We derive

necessary and sufficient conditions for the existence of regularized estimators in generalized linear models when some parameters are left unpenalized, since they are assumed a priori to be nonzero. In addition, a numerical procedure to check for the existence of such estimators in logistic and Poisson regression is described.

Chapter 1

Introduction

Production, storage and processing of massive data sets is now commonplace due to scientific breakthroughs in various domains such as genomics, finance and image classification. As a consequence, high-dimensional data analysis is attracting a lot of attention. This is the study of models where the number of parameters P is larger than the sample size N. In those instances, the maximum likelihood estimator is no longer uniquely defined. Even when $P \leqslant N$ with P close to N, it tends to perform poorly due to its high variance. Motivated by the seminal papers of James and Stein [1961] and Tikhonov [1963], a considerable amount of literature has concentrated over the last half-century on parameter estimation using regularization techniques. In both parametric and nonparametric models, a reasonable prior or constraints reflecting properties or assumptions specific to a domain are set on the parameters in order to reduce the variance of the estimator and the complexity of the fitted model, at the price of a bias increase. A common assumption is that the solution is *sparse*. In other words, the response variable depends upon very few nonzero parameters. This assumption arises naturally for example in DNA microarrays. Expression levels of thousands of genes are measured simultaneously in a given cell and it is believed that few genes are responsible for a particular tumor subtype. Another application is in signal denoising. The function to be estimated is expanded on a wavelet basis and it is known that many natural signals have sparse basis coefficients. This sparsity pattern is referred to as *coordinate sparsity*. Only a few entries of the parameter vector $\boldsymbol{\beta}^*$ are different from zero. Various other sparsity patterns such as the following are commonly assumed.

- *Group sparsity.* Given a partition $\{1, \ldots, P\} = \bigcup_{k=1}^{M} G_k$, only a few vectors $\boldsymbol{\beta}^*_{G_k} = (\beta^*_j)_{j \in G_k}$ are nonzero. There is a group structure of predictors in several applications. In linear regression for example, it is customary to represent a categorical predictor taking on several levels by a group of dummy variables. In

fMRI analysis, a set of voxels from the same brain areas are naturally related.

- *Variation sparsity.* The vector $\Delta\boldsymbol{\beta}^* = (\beta_2^* - \beta_1^*, \ldots, \beta_P^* - \beta_{P-1}^*)$ is coordinate sparse. This assumption arises when predictors are closely related to their neighbours, for example when denoising a univariate signal believed to be piecewise constant.

In this thesis, we consider a class of regularization techniques, called *thresholding*, which assumes a certain transform $\boldsymbol{\xi}^* = g(\boldsymbol{\beta}^*)$ is sparse. These techniques are indexed by a tuning parameter $\lambda \geqslant 0$ and yield a sparse $\hat{\boldsymbol{\xi}}_\lambda = g(\hat{\boldsymbol{\beta}}_\lambda)$.

Thresholding techniques, which include the lasso [Tibshirani, 1996], are employed in various settings such as linear regression [Donoho and Johnstone, 1994; Tibshirani, 1996], generalized linear models [Park and Hastie, 2007], low-rank matrix estimation [Shabalin and Nobel, 2013; Donoho and Gavish, 2014], density estimation [Donoho et al., 1996; Sardy and Tseng, 2010], linear inverse problems [Donoho, 1995], compressed sensing [Donoho, 2006; Candès and Romberg, 2007] and time series [Neto et al., 2012].

In Chapter 2, we review non exhaustively thresholding estimators and show that for many estimators, the estimated vector of coefficients is set equal to the null vector for all threshold parameters greater than a certain statistic, the *zero-thresholding function*.

There are several goals when using these methods. A first one is to find models with good predictive accuracy. This could allow for example to classify tumor subtypes in an early phase of a disease. A second one is to estimate the true sparsity pattern, that is, the active set of relevant variables, $\mathcal{S}^* = \{q \mid \xi_q^* \neq 0\}$. It amounts to selecting basis coefficients in wavelet denoising, or determining the gene expression signature of a tumor subtype. These goals are quite different and the threshold aimed at prediction optimality often differs from the optimal threshold for model identification [Yang, 2005; Leng et al., 2006; Meinshausen and Bühlmann, 2006]. A third one is variable screening which requires that the estimated support $\hat{\mathcal{S}}_\lambda = \{q \mid \hat{\xi}_{\lambda,q} \neq 0\}$ contains $\mathcal{S}^*$ with high probability. Under some assumptions and for a suitably chosen λ, certain thresholding estimators attain these goals.

To perform effective model selection, the choice of the threshold is crucial since the latter governs the complexity of the estimated model. A too large λ results in a simplistic model missing important features whereas a too small λ leads to a model including unnecessary features. Classical methodologies consist in minimizing a criterion. Examples include cross-validation, AIC [Akaike, 1998], BIC [Schwarz, 1978] and Stein unbiased risk estimation (SURE) [Stein, 1981]. Because traditional information criteria do not adapt well to the high-dimensional setting, generalizations

such as GIC [Fan and Tang, 2013] and EBIC [Chen and Chen, 2008] have been suggested.

In Chapter 3, we propose a new threshold selection method, the *quantile universal threshold*, based on the idea that if the null model $\boldsymbol{\xi}^* = \mathbf{0}$ is true, it should be recovered with high probability. As it turns out, this delivers good theoretical properties, even when $\boldsymbol{\xi}^* \neq \mathbf{0}$.

Significance testing of the parameter $\boldsymbol{\beta}^*$, or more generally of a linear transform of $\boldsymbol{\beta}^*$, is of fundamental interest in statistics. In linear regression models, one of the most popular procedure is Fisher's F-test which was introduced for the study of data from agricultural experiments [Fisher, 1973]. It is routinely used in many instances including classical stepwise model selection procedures which sequentially add or delete variables.

In Chapter 4, we propose testing procedures in linear models, *thresholding tests*, based on thresholding estimators. The null hypothesis $\boldsymbol{\xi}^* = \mathbf{0}$ is rejected if $\hat{\boldsymbol{\xi}}_\lambda \neq \mathbf{0}$. The power of several of these procedures is studied and it is shown they inherit a screening property.

A large class of thresholding estimators are defined as a minimizer of the sum of a loss and a penalty function, both convex. Although efficient algorithms exist to compute the estimated models over a regularization path (see e.g. Friedman et al. [2010]), no indication of nonexistence of the estimate is given. As a result, practitioners tend to overlook the problem of nonexistence. Moreover, the statistical inference based on the result of the algorithm might be incorrect.

In Chapter 5, we prove existence of several classical thresholding estimators and derive necessary and sufficient conditions for the existence of regularized estimators in generalized linear models. Our approach is related to the behaviour of the objective function along certain directions.

This thesis is organized as follows. We start in Section 2.1 by reviewing non-exhaustively thresholding estimators in linear regression, generalized linear models and low-rank matrix estimation and give a formal definition of such estimators. We then introduce the key concept of a *zero-thresholding function* in Section 2.2 and derive an explicit formulation in many instances.

Chapter 3 adresses the problem of selecting the thresholding parameter. In Section 3.1, we review existing techniques which set the parameter such that the null model $\boldsymbol{\xi}^* = \mathbf{0}$ is recovered with high probability and we introduce the *null-thresholding statistic*. In Section 3.2, we propose the *quantile universal threshold* as a tuning parameter methodology that follows the same paradigm in various domains. Theoretical properties are then derived. An explicit formulation of the quantile

universal threshold is given for several examples in Section 3.3 and numerical results of our methodology applied to the lasso for generalized linear models are contained in Section 3.4. In particular, we illustrate the effectiveness of our methodology in terms of model selection and prediction error in Section 3.4.1 and in terms of false discovery rate and true positive rate in Section 3.4.2. We observe in Section 3.4.3 a phase transition property and make concluding remarks in Section 3.4.4.

In Chapter 4, we consider hypothesis testing in linear models. Testing procedures based on thresholding estimators, *thresholding tests*, are proposed in Section 4.1. The *affine lasso* and several variants are introduced in Section 4.2. They allow to test hypotheses of the form $H_0 : A\boldsymbol{\beta}^* = \mathbf{c}$ for a full row rank matrix A and a given vector $\mathbf{c}$. Confidence regions are then derived. A comparative power study is performed under sparse and dense alternatives in Section 4.3 and an adaptive testing procedure suggested. We show in Section 4.4 that these testing procedures inherit a screening property. Moreover, we derive a necessary and sufficient condition for a compatibility condition to hold in balanced analysis of variance (ANOVA) designs. This condition ensures variable screening is achieved under an additional assumption on the magnitude of the nonzero entries of $\boldsymbol{\beta}^*$.

Chapter 5 adresses the problem of existence of thresholding estimators which minimize the sum of a loss and a penalty function, both convex. In Section 5.1, we give sufficient conditions for the existence of such estimators and prove existence for a large class. In Section 5.2, we consider the special case of estimators in generalized linear models with convex penalties, allowing some parameters to be unpenalized. We give necessary and sufficient conditions for their existence. In addition, a numerical procedure to check for the existence of such estimators in logistic and Poisson regression is described in Section 5.3.

There are two appendices: Appendix A contains a review of objects in convex analysis and Appendix B contains a supplementary simulation study to Section 3.4 and considers the problem of variance estimation in linear models.

Finally, our notation in the thesis is as follows. The transpose of a given $N \times P$ matrix X is written X^t. We denote by $\mathbf{x}^i$ and $\mathbf{x}_j$ its ith row and jth column, respectively. The set $\overline{\mathcal{I}}$ is the complement of the set $\mathcal{I}$. For sets of indices $\mathcal{A} = \{j_1, \ldots, j_M\}$ and $\mathcal{B} = \{k_1, \ldots, k_Q\}$, we let $X_{\mathcal{A}} = [\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_M}]$ and

$$X^{\mathcal{B}} = \begin{bmatrix} \mathbf{x}^{k_1} \\ \vdots \\ \mathbf{x}^{k_Q} \end{bmatrix}$$

denote the $N \times M$ and $Q \times P$ submatrices. Similarly, for a vector $\mathbf{v} = (v_1, \ldots, v_P)^t \in \mathbb{R}^P$ we let $\mathbf{v}_{\mathcal{A}} = (\mathbf{v}_{j_1}, \ldots, \mathbf{v}_{j_M})^t$. We write $C(X)$ for the column space of X and

P_X for the orthogonal projection matrix onto $C(X)$. A column vector of ones is denoted by $\mathbf{1}$, a column vector of zeros by $\mathbf{0}$ and an $N \times P$ matrix of zeros by $0_{N \times P}$.

Chapter 2

The Zero-Thresholding Function

We start in Section 2.1 by reviewing non-exhaustively thresholding estimators in linear regression, generalized linear models and low-rank matrix estimation. We then introduce in Section 2.2 the key concept of a zero-thresholding function and give an explicit formulation in several examples.

2.1 Review of Thresholding Estimators

Thresholding estimators are routinely used in many settings including the following.

Linear regression. Consider the linear model

$$\mathbf{Y} = X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I_N), \tag{2.1}$$

where X is an $N \times P$ matrix of covariates or discretized basis functions and $\boldsymbol{\beta}^*$ is the true unknown coefficient vector.

In wavelet denoising, Donoho and Johnstone [1994] introduced WaveShrink. For an observed $\mathbf{y}$ from (2.1), where $X = W$ is a wavelet matrix and $\boldsymbol{\beta}^*$ is the vector of wavelet coefficients of the signal to be recovered, they suggest applying a thresholding estimator to $\boldsymbol{\theta} = W^t\mathbf{y}$, the vector of wavelet coefficients of $\mathbf{y}$. To obtain an estimate of the signal, the inverse wavelet transform is then applied . The hard and soft thresholding functions are defined by

$$\begin{aligned}\eta_H(\theta_i, \lambda) &= \theta_i \cdot 1\{|\theta_i| > \lambda\},\\ \eta_S(\theta_i, \lambda) &= \mathrm{sign}(\theta_i)(|\theta_i| - \lambda) \cdot 1\{|\theta_i| > \lambda\},\end{aligned}$$

where $1\{\cdot\}$ is the indicator function. Clearly, sparsity is induced since θ_i is set to zero if $|\theta_i| \leqslant \lambda$.

For an observed $\mathbf{y}$, a large class of estimators, including the hard and soft thresholding functions, is of the form

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}}\ \ell_{\mathbf{y}}(X\boldsymbol{\beta}) + p_\lambda(g(\boldsymbol{\beta})), \tag{2.2}$$

for a certain loss function $\ell_{\mathbf{y}}$ and penalty p_λ inducing sparsity in $\hat{\boldsymbol{\xi}}_\lambda = g(\hat{\boldsymbol{\beta}}_\lambda)$. The element notation "$\in$" indicates the minimizer might not be unique. The lasso [Tibshirani, 1996]

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}}\ \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

is amongst the most popular techniques. Other examples include the least absolute deviation (LAD) lasso [Wang et al., 2007]

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}}\ \|\mathbf{y} - X\boldsymbol{\beta}\|_1 + \lambda\|\boldsymbol{\beta}\|_1$$

and the $\sqrt{\text{lasso}}$ [Belloni et al., 2011]

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}}\ \|\mathbf{y} - X\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1.$$

As an alternative, Candès and Tao [2007] proposed the Dantzig selector

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\,:\,\|X^t(\mathbf{y}-X\boldsymbol{\beta})\|_\infty \leqslant \lambda}{\operatorname{argmin}}\ \|\boldsymbol{\beta}\|_1.$$

The smoothly clipped absolute deviation (SCAD) estimator [Fan and Peng, 2004] and the minimax concave penalty (MCP) estimator [Zhang, 2010] suggest using nonconvex penalty functions.

A more traditional thresholding estimator, which is computationally feasible for only small values of P, is best subset selection. It minimizes $\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2/2 + \lambda\|\boldsymbol{\beta}\|_0$ where $\|\boldsymbol{\beta}\|_0 = \sum_{p=1}^P 1\{\beta_p \neq 0\}$.

Given a partition $\{1,\dots,P\} = \bigcup_{k=1}^M G_k$, group sparsity inducing thresholding estimators were introduced by Yuan and Lin [2006] which define the group lasso as

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}}\ \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\sum_{k=1}^M \|\boldsymbol{\beta}_{G_k}\|_2,$$

as well as [Bunea et al., 2014] who proposed the group $\sqrt{\text{lasso}}$ by substituting the ℓ_2 norm for the squared ℓ_2 norm divided by two.

Variation sparsity is induced by one-dimensional total variation [Rudin et al., 1992]

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}}\ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\beta}\|_2^2 + \lambda\|D\boldsymbol{\beta}\|_1,$$

with

$$D = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \tag{2.3}$$

the $(P-1) \times P$ first order difference matrix.

The generalized lasso introduced by Tibshirani and Taylor [2011],

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|A\boldsymbol{\beta}\|_1$$

allows more generally for sparsity in $A\hat{\boldsymbol{\beta}}_\lambda$, for a certain matrix A.

Composite penalties with a two-dimensional tuning parameter have also been proposed such as the elastic net [Zou and Hastie, 2005] where $p_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2$, and the fused lasso [Tibshirani et al., 2005] with penalty function $\lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{p=2}^P |\beta_p - \beta_{p-1}|$ inducing not only coordinate sparsity but also variation sparsity.

Note that although ridge regression [Hoerl and Kennard, 1970], bridge [Fu, 1998] and smoothing splines [Wahba, 1990] are of the form (2.2), they do not yield a sparse $\hat{\boldsymbol{\xi}}_\lambda$.

Generalized linear models (GLMs). GLMs encompass Gaussian linear models, logistic regression for binary responses and Poisson regression for count data. The canonical model assumes the negative log-likelihood is of the form

$$\ell_{\mathbf{y}}(\boldsymbol{\beta}) = \sum_{n=1}^N \left[-y_n\theta_n + b(\theta_n)\right] \quad \text{with} \quad \theta_n = \mathbf{x}^n\boldsymbol{\beta}, \tag{2.4}$$

$b(\cdot)$ a known function and $\mathbf{x}^n$ denoting the nth row of the design matrix X [Nelder and Wedderburn, 1972]. As an extension of lasso, Park and Hastie [2007] define

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta} \in \mathcal{F}}{\operatorname{argmin}} \ell_{\mathbf{y}}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.5}$$

where $\mathcal{F} := \left\{\boldsymbol{\beta} \in \mathbb{R}^P \,|\, X\boldsymbol{\beta} \in \Theta^N\right\}$ and $\Theta := \{\theta \in \mathbb{R} \,|\, b(\theta) < \infty\}$. Several other penalties such as the group lasso penalty [Meier et al., 2008] have been proposed to regularize GLMs.

Low-rank matrix estimation. In applications such as image denoising and signal processing, it is customary to consider a model of the form

$$Y = X^* + Z,$$

where X^* is an unknown $N \times P$ matrix assumed to have low-rank and Z has i.i.d. normally distributed entries with zero mean and variance σ^2. Few singular values

of X^* are nonzero since X^* is assumed to have low-rank. Let $Y = \sum_{i=1}^{N} d_i \mathbf{u}_i \mathbf{v}_i^t$ be the singular value decomposition of Y, where $\mathbf{u}_i \in \mathbb{R}^N$ and $\mathbf{v}_i \in \mathbb{R}^P$ for $i = 1, \ldots, N$ are the left and right singular vectors with singular value d_i. The classical approach consists in hard thresholding the singular values so that

$$\hat{X}_\lambda = \sum_{i=1}^{N} \eta_H(d_i, \lambda) \mathbf{u}_i \mathbf{v}_i^t.$$

Inspired by lasso, a more recent estimator solves

$$\underset{X \in \mathbb{R}^{N \times P}}{\operatorname{argmin}} \ \frac{1}{2} \|Y - X\|_F^2 + \lambda \|X\|_*,$$

where $\|\cdot\|_F$ and $\|\cdot\|_*$ respectively denote the Frobenius and trace norm. This corresponds to soft thresholding the singular values [Mazumder et al., 2010; Cai et al., 2010].

Motivated by the preceding examples in GLMs and low-rank matrix estimation, a definition of a thresholding estimator is the following.

Definition 2.1 *Assume* $\mathbf{Y}$ *has density* $f_{(\boldsymbol{\eta}^*, \boldsymbol{\beta}^*)}$ *with* $\boldsymbol{\xi}^* = g(\boldsymbol{\beta}^*)$ *sparse for a certain function* g*. Let* $\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y})$ *be an estimator indexed by* $\lambda \geqslant 0$*. We call* $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y}) = g \circ \hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y})$ *a thresholding estimator if*

$$\mathbb{P}(\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y}) = \mathbf{0}) > 0 \quad \textit{for some finite } \lambda.$$

The parameter $\boldsymbol{\eta}^*$ is a nuisance parameter which does not possess any sparsity structure, such as the variance and the intercept in linear regression.

We will make use of this definition when introducing the zero-thresholding function in the next section and our threshold selection methodology in Section 3.2. In Chapter 4, tests based on thresholding estimators will be derived.

2.2 Definition and Derivations

A key property shared by a large class of classical thresholding estimators is to set the estimated parameters to zero for any threshold parameter greater than a finite statistic $\lambda(\mathbf{Y})$. This leads to the following definition.

Definition 2.2 *A thresholding estimator* $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y})$ *admits a zero-thresholding function* $\lambda(\mathbf{Y})$ *if*

$$\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y}) = \mathbf{0} \quad \iff \quad \lambda \geqslant \lambda(\mathbf{Y}) \quad \textit{almost everywhere.}$$

The zero-thresholding function is hence determined uniquely up to sets of measure zero. Note that the equivalence implies equiprobability between setting all

coefficients to zero and selecting the threshold large enough. As we now show, a closed form expression can be derived for many of the estimators reviewed in Section 2.1.

- *WaveShrink and the Dantzig selector.* It is immediate from the definition of these estimators that $\lambda(\mathbf{y})$ is respectively given by $\|W^t\mathbf{y}\|_\infty$ with W the wavelet matrix and $\|X^t\mathbf{y}\|_\infty$.

- *SCAD and MCP with orthonormal columns.* In this setting, an explicit solution is known (see e.g. Breheny and Huang [2015]) and it follows that $\lambda(\mathbf{y}) = \|X^t\mathbf{y}\|_\infty$.

- *Lasso, $\sqrt{lasso}$ and LAD lasso.* We consider the Karush-Kuhn-Tucker (KKT) optimality conditions [Rockafellar, 1970]. For the lasso, they are given by

$$X^t(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = \lambda\boldsymbol{\gamma}, \tag{2.6}$$

$$\gamma_i \in \begin{cases} \{\text{sign}(\hat{\beta}_i)\} & \text{if } \hat{\beta}_i \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_i = 0 \end{cases}, \quad \text{for } i = 1, \ldots, P. \tag{2.7}$$

  The vector $\boldsymbol{\gamma} \in \mathbb{R}^P$ is a subgradient of $f(\mathbf{x}) = \|\mathbf{x}\|_1$ evaluated at $\mathbf{x} = \hat{\boldsymbol{\beta}}$ (see Appendix A.2). Hence, a necessary and sufficient condition for $\hat{\boldsymbol{\beta}}$ to be a solution to the lasso is that there exists some $\boldsymbol{\gamma}$ satisfying (2.6) and (2.7). This leads to $\lambda(\mathbf{y}) = \|X^t\mathbf{y}\|_\infty$. For $\sqrt{\text{lasso}}$, $\hat{\boldsymbol{\beta}}$ is a solution if and only if there exists some $\boldsymbol{\gamma}$ satisfying (2.7) and some $\mathbf{u}$ such that

$$X^t\mathbf{u} = \lambda\boldsymbol{\gamma},$$

$$\mathbf{u} \in \begin{cases} \{(\mathbf{y} - X\hat{\boldsymbol{\beta}})/\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|_2\} & \text{if } \mathbf{y} - X\hat{\boldsymbol{\beta}} \neq 0 \\ \{\mathbf{v} \mid \|\mathbf{v}\|_2 \leqslant 1\} & \text{if } \mathbf{y} - X\hat{\boldsymbol{\beta}} = 0 \end{cases}.$$

  It follows that $\lambda(\mathbf{y}) = \|X^t\mathbf{y}\|_\infty/\|\mathbf{y}\|_2$. For LAD lasso, a necessary and sufficient condition for $\hat{\boldsymbol{\beta}}$ to be a solution is that there exists some $\boldsymbol{\gamma}$ satisfying (2.7) and some $\mathbf{w}$ such that

$$X^t\mathbf{w} = \lambda\boldsymbol{\gamma},$$

$$w_i \in \begin{cases} \{\text{sign}(y_i - \mathbf{x}_i^t\hat{\boldsymbol{\beta}})\} & \text{if } y_i - \mathbf{x}_i^t\hat{\boldsymbol{\beta}} \neq 0 \\ [-1, 1] & \text{if } y_i - \mathbf{x}_i^t\hat{\boldsymbol{\beta}} = 0 \end{cases}.$$

  Hence, $\lambda(\mathbf{y}) = \|X^t\,\text{sgn}(\mathbf{y})\|_\infty$ with $\text{sgn}(\cdot)$ denoting the sign function applied componentwise.

- *Group lasso and group $\sqrt{lasso}$.* Let the partition $\{1, \ldots, P\} = \bigcup_{k=1}^M G_k$. From the KKT optimality conditions, $\lambda(\mathbf{y}) = \max_{k=1,\ldots,M} \|X_{G_k}^t\mathbf{y}\|_2$ for the group lasso and $\lambda(\mathbf{y}) = \max_{k=1,\ldots,M} \|X_{G_k}^t\mathbf{y}\|_2/\|\mathbf{y}\|_2$ for the group $\sqrt{\text{lasso}}$.

- *One-dimensional total variation.* Here, $g(\boldsymbol{\beta}^*) = (\beta^*_{j+1} - \beta^*_j)_{j=1,\dots,P-1}$ so that $g(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ is equivalent to $\hat{\boldsymbol{\beta}} = c\mathbf{1}$ for a certain constant c. From the KKT optimality conditions, such a $\hat{\boldsymbol{\beta}}$ is a solution if and only if there exists a vector $\mathbf{v}$ such that

$$\mathbf{y} - \hat{\boldsymbol{\beta}} = \lambda D^t \mathbf{v},$$

$$v_i \in [-1, 1], \quad i = 1, \dots, P-1,$$

with D defined in (2.3). It can then be shown that $\lambda(\mathbf{y}) = \|(DD^t)^{-1}D\mathbf{y}\|_\infty$.

- *Generalized lasso.* In this setting, $g(\boldsymbol{\beta}^*) = A\boldsymbol{\beta}^*$. Assume A is $R \times P$ of full row rank and let $\mathcal{I}$ denote a set of column indices such that $A_\mathcal{I}$ is invertible. Letting

$$\Phi = \begin{bmatrix} A_\mathcal{I} & A_{\overline{\mathcal{I}}} \\ 0 & I \end{bmatrix},$$

we change variables to $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t)^t = \Phi\boldsymbol{\beta}$, so that the generalized lasso problem becomes

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{2}\|\mathbf{y} - X\Phi^{-1}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}_1\|_1.$$

Let us write $X\Phi^{-1}\boldsymbol{\theta} = B_1\boldsymbol{\theta}_1 + B_2\boldsymbol{\theta}_2$, with $B_1 = X_\mathcal{I}A_\mathcal{I}^{-1}$ and $B_2 = X_{\overline{\mathcal{I}}} - X_\mathcal{I}A_\mathcal{I}^{-1}A_{\overline{\mathcal{I}}}$. Solutions to the initial problem are then given by $\hat{\boldsymbol{\beta}} = \Phi^{-1}\hat{\boldsymbol{\theta}}$, with

$$\hat{\boldsymbol{\theta}}_1 \in \operatorname*{argmin}_{\boldsymbol{\theta}_1} \frac{1}{2}\|(I - P_{B_2})(\mathbf{y} - B_1\boldsymbol{\theta}_1)\|_2^2 + \lambda\|\boldsymbol{\theta}_1\|_1,$$

and $\hat{\boldsymbol{\theta}}_2$ such that $B_2\hat{\boldsymbol{\theta}}_2 = P_{B_2}(\mathbf{y} - B_1\hat{\boldsymbol{\theta}}_1)$. Finally, $\lambda(\mathbf{y}) = \|B_1^t(I - P_{B_2})\mathbf{y}\|_\infty$.

- *Best Subset.* We first note that for almost all $\mathbf{y}$, $\hat{\boldsymbol{\beta}} = \mathbf{0}$ is a solution if and only if

$$\frac{1}{2}\|\mathbf{y} - P_{C(X_\mathcal{I})}\mathbf{y}\|_2^2 + \lambda p \geqslant \frac{1}{2}\|\mathbf{y}\|_2^2$$

for all index sets $\mathcal{I} \subset \{1, \dots, P\}$ of cardinality p such that $\operatorname{rank}(X_\mathcal{I}) = p$, $1 \leqslant p \leqslant \operatorname{rank}(X)$. Let $\Delta_p(\mathbf{y})$ denote the maximum of $\|P_{C(X_\mathcal{I})}\mathbf{y}\|_2^2/2$ over all such index sets. Since $\|\mathbf{y} - P_{C(X_\mathcal{I})}\mathbf{y}\|_2^2 = \|\mathbf{y}\|_2^2 - \|P_{C(X_\mathcal{I})}\mathbf{y}\|_2^2$, we obtain

$$\lambda(\mathbf{y}) = \max_{1 \leqslant p \leqslant \operatorname{rank}(X)} \frac{\Delta_p(\mathbf{y})}{p}.$$

If X has orthogonal columns, then for any index set of cardinality p, $\|P_{C(X_\mathcal{I})}\mathbf{y}\|_2^2 \leqslant p\Delta_1(\mathbf{y})$ from Pythagoras' theorem and hence $\lambda(\mathbf{y}) = \Delta_1(\mathbf{y})$.

- *Fused Lasso.* Assuming X is orthonormal, from Lemma A.1 in Friedman et al. [2007]

$$\hat{\boldsymbol{\beta}}_{(\lambda_1,\lambda_2)}(\mathbf{y}) = \eta_S(\hat{\boldsymbol{\beta}}_{(0,\lambda_2)}(\mathbf{y}), \lambda_1)$$

where, with a slight abuse of notation, $\eta_S(\cdot, \lambda_1)$ denotes the soft thresholding operator applied componentwise. It follows that $\lambda_1(\mathbf{y}; \lambda_2) = \|\hat{\boldsymbol{\beta}}_{(0,\lambda_2)}(\mathbf{y})\|_\infty$.

- *Elastic Net.* As in Lemma 1 of Zou and Hastie [2005], we convert the elastic net into a lasso problem by artificially augmenting the dataset. More precisely, letting

$$\tilde{X} = (1+\lambda_2)^{-1/2} \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix}, \quad \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix},$$

and $\boldsymbol{\theta} = \sqrt{1+\lambda_2}\boldsymbol{\beta}$, the elastic net becomes

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{2}\|\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\theta}\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\|\boldsymbol{\theta}\|_1.$$

Solutions to the initial problem are then obtained as $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\theta}}/\sqrt{1+\lambda_2}$. Finally, $\lambda_1(\mathbf{y}; \lambda_2) = \|X^t\mathbf{y}\|_\infty$ regardless of λ_2.

Remark 2.1 *It is often the case that some parameters are assumed a priori to be nonzero. It is then customary to let these parameters unpenalized. The lasso for example becomes*

$$\operatorname*{argmin}_{(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathbb{R}^{P_0+P}} \frac{1}{2}\|\mathbf{y} - X_0\boldsymbol{\beta}_0 - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

with X_0 of size $N \times P_0$. The parameter $\boldsymbol{\beta}_0^*$ is a nuisance parameter and the zero-thresholding function associated to $g(\boldsymbol{\beta}^*)$ is obtained by substituting $\mathbf{y}$ with $(I - P_{X_0})\mathbf{y}$ and X with $(I - P_{X_0})X$ in the above formulas. For example, the zero-thresholding function of lasso is given by $\|X^t(I - P_{X_0})\mathbf{y}\|_\infty$ and that of group lasso by $\max_{k=1,\dots,M} \|X_{G_k}^t(I - P_{X_0})\mathbf{y}\|_2$. Moreover, the zero-thresholding function of $\sqrt{lasso}$ is given by $\|X^t(I - P_{X_0})\mathbf{y}\|_\infty / \|(I - P_{X_0})\mathbf{y}\|_2$ and that of group $\sqrt{lasso}$ by $\max_{k=1,\dots,M} \|X_{G_k}^t(I - P_{X_0})\mathbf{y}\|_2 / \|(I - P_{X_0})\mathbf{y}\|_2$.

Derivation of the zero-thresholding function of the LAD lasso remains an open problem when some parameters are unpenalized.

- *Lasso and group lasso GLM.* We let the parameter $\boldsymbol{\beta}_0$ unpenalized so that (2.5) becomes

$$(\hat{\boldsymbol{\beta}}_{0\lambda}, \hat{\boldsymbol{\beta}}_\lambda) \in \operatorname*{argmin}_{(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathcal{F}} \ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.8}$$

where $\mathcal{F} := \left\{(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathbb{R}^{P_0+P} \mid X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta} \in \Theta^N\right\}$ and $\ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ is defined as in (2.4) with $\theta_n = \mathbf{x}_0^n\boldsymbol{\beta}_0 + \mathbf{x}^n\boldsymbol{\beta}$. We set $\hat{\boldsymbol{\beta}}_\lambda = \mathbf{0}$ if $\lambda = +\infty$. Although the lasso GLM solution might not be unique, its fit is unique.

Lemma 2.1 *Assume b is strictly convex on Θ. For any fixed X_0, X, $\mathbf{y}$ and $\lambda \geqslant 0$, $X_0\hat{\boldsymbol{\beta}}_{0\lambda} + X\hat{\boldsymbol{\beta}}_\lambda$ is unique.*

Proof It follows from the strict convexity of b on Θ and the convexity of $f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ on $\mathcal{F}$ that the objective function in (2.8) is convex on $\mathcal{F}$. The solution set is thus convex.

Assume there exists two solutions $(\hat{\boldsymbol{\beta}}_0^{(1)}, \hat{\boldsymbol{\beta}}^{(1)})$ and $(\hat{\boldsymbol{\beta}}_0^{(2)}, \hat{\boldsymbol{\beta}}^{(2)})$ such that $X_0\hat{\boldsymbol{\beta}}_0^{(1)} + X\hat{\boldsymbol{\beta}}^{(1)} \neq X_0\hat{\boldsymbol{\beta}}_0^{(2)} + X\hat{\boldsymbol{\beta}}^{(2)}$. Because the solution set is convex, $(\hat{\boldsymbol{\beta}}_0^{(3)}, \hat{\boldsymbol{\beta}}^{(3)}) := \delta(\hat{\boldsymbol{\beta}}_0^{(1)}, \hat{\boldsymbol{\beta}}^{(1)}) + (1-\delta)(\hat{\boldsymbol{\beta}}_0^{(2)}, \hat{\boldsymbol{\beta}}^{(2)})$ is a solution for any $0 < \delta < 1$. However,

$$\ell_{\mathbf{y}}\left(\hat{\boldsymbol{\beta}}_0^{(3)}, \hat{\boldsymbol{\beta}}^{(3)}\right) + \lambda\|\hat{\boldsymbol{\beta}}^{(3)}\|_1 < m,$$

where m denotes the minimum value of the objective function and the strict inequality follows from the strict convexity of b and the convexity of $f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. In other words, $(\hat{\boldsymbol{\beta}}_0^{(3)}, \hat{\boldsymbol{\beta}}^{(3)})$ is not in the solution set, a contradiction. ■

From the KKT optimality conditions, $(\hat{\boldsymbol{\beta}}_0, \mathbf{0})$ is a solution for a fixed finite λ if and only if there exists some $\boldsymbol{\gamma} \in \mathbb{R}^P$ such that

$$\begin{cases} X_0\hat{\boldsymbol{\beta}}_0 \in \Theta^N \\ {X_0}^t\mathbf{y} = {X_0}^t\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_0) \\ X^t[\mathbf{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_0)] = \lambda\boldsymbol{\gamma} \end{cases},$$

$$\gamma_i \in [-1, 1], \quad i = 1, \ldots, P,$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}_0) = (b'({\mathbf{x}_0}^1\boldsymbol{\beta}_0), \ldots, b'({\mathbf{x}_0}^N\boldsymbol{\beta}_0))^t$. Hence, for a strictly convex b,

$$\lambda(\mathbf{y}) = \begin{cases} \|X^t[\mathbf{y} - \boldsymbol{\mu}(\mathbf{v})]\|_\infty & \text{if } \mathbf{y} \in \mathcal{D} \\ +\infty & \text{otherwise} \end{cases}, \tag{2.9}$$

with $\mathbf{v}$ any vector such that

$$\begin{cases} X_0\mathbf{v} \in \Theta^N \\ {X_0}^t\mathbf{y} = {X_0}^t\boldsymbol{\mu}(\mathbf{v}) \end{cases} \tag{2.10}$$

and $\mathcal{D} = \{\mathbf{y} \mid \exists \mathbf{v} \in \mathbb{R}^{P_0} \text{ solution to (2.10)}\}$.

For the group lasso GLM, one obtains similarly

$$\lambda(\mathbf{y}) = \begin{cases} \max_{k=1,\ldots,M} \|X_{G_k}^t[\mathbf{y} - \boldsymbol{\mu}(\mathbf{v})]\|_2 & \text{if } \mathbf{y} \in \mathcal{D} \\ +\infty & \text{otherwise} \end{cases}. \tag{2.11}$$

Lemma 2.1 implies $\lambda(\mathbf{y})$ does not depend on which solution $\mathbf{v}$ to (2.10) is chosen. The set $\mathcal{D}$ is the set of values based on which the maximum likelihood estimate (MLE) of $(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ with constraint $\hat{\boldsymbol{\beta}} = 0$ exists. If the response variable is Gaussian, note that $\mathcal{D} = \mathbb{R}^N$. An explicit formulation when the intercept is unpenalized ($X_0 = \mathbf{1}$) is given in Table 2.1. For an arbitrary matrix X_0, a characterization appears in Section 5.2 (see Remark 5.1).

- *Low-rank matrix estimation.* Since $\hat{X} = 0_{N\times P}$ is equivalent to $\hat{\boldsymbol{d}}_\lambda = \mathbf{0}$, $\lambda(\mathbf{y}) = \|\mathbf{d}\|_\infty$, whether soft or hard thresholding is applied to the singular value vector $\mathbf{d}$ of Y.

Table 2.1 – Values of $b'(\beta_0^*)$, $\mathcal{D}$ and $\mathbb{P}(\mathbf{Y} \in \mathcal{D})$ when $X_0 = \mathbf{1}$.

Response distribution	$\mu = b'(\beta_0^*)$	$\mathcal{D}$	$\mathbb{P}(\mathbf{Y} \in \mathcal{D})$
Gaussian	β_0^*	$\mathbb{R}^N$	1
Poisson	$\exp(\beta_0^*)$	$\mathbb{N}^N \setminus \{\mathbf{0}\}$	$1 - \exp(-N\mu)$
Bernoulli	$\dfrac{\exp(\beta_0^*)}{1+\exp(\beta_0^*)}$	$\{0,1\}^N \setminus \{\mathbf{0},\mathbf{1}\}$	$1 - \mu^N - (1-\mu)^N$
Binomial $(m,p)/m$	$\dfrac{\exp(\beta_0^*)}{1+\exp(\beta_0^*)}$	$\{0,1/m,\ldots,1\}^N \setminus \{\mathbf{0},\mathbf{1}\}$	$1 - (\mu)^{mN} - (1-\mu)^{mN}$

Chapter 3

Model Selection with the Quantile Universal Threshold

By setting estimated coefficients to zero, thresholding estimators make it possible to perform variable selection in high dimensions. Selection of the threshold is crucial since it governs the complexity of the chosen model. Classical methodologies consist in minimizing a criterion. In low-rank matrix estimation, Owen and Perry [2009] and Josse and Husson [2012] employ cross-validation whereas Candès et al. [2013] and Josse and Sardy [2016] apply SURE. The latter methodology is also used in regression [Donoho and Johnstone, 1994; Zou et al., 2007; Tibshirani and Taylor, 2012], and reduced rank regression [Mukherjee et al., 2015].

In this chapter, a new threshold selection method is proposed. We motivate our approach in Section 3.1 before introducing the quantile universal threshold and deriving theoretical properties in Section 3.2. Some examples for which an explicit formulation of the proposed threshold exists are considered in Section 3.3. Finally, we conduct a simulation study in Section 3.4 to illustrate the effectiveness of our methodology on real and synthetic data in GLMs.

3.1 Thresholding Under the Null

We consider the idea of choosing a threshold based on the null model $\boldsymbol{\xi}^* = \mathbf{0}$. The methodology selects λ large enough such that $\boldsymbol{\xi}^*$ is recovered with high probability under the null model. Such a choice results in good empirical and theoretical properties even in the case $\boldsymbol{\xi}^* \neq \mathbf{0}$. As we will see, this approach is intimately linked to the zero-thresholding function. We describe three settings in which this idea has appeared.

- *Wavelet denoising.* Donoho and Johnstone [1994] and Donoho et al. [1995] set the threshold of WaveShrink via soft thresholding as $\lambda_P^{\text{universal}} = \sigma\sqrt{2\log P}$. Under the null model with wavelet coefficients $\boldsymbol{\beta}^* = \mathbf{0}$, $\mathbb{P}(\hat{\boldsymbol{\beta}}_{\lambda_P^{\text{universal}}} = \mathbf{0}) \to 1$ as P tends to infinity. It turns out that an oracle inequality and minimax properties hold with $\lambda = \lambda_P^{\text{universal}}$ over a wide class of functions with $\boldsymbol{\beta}^* \neq \mathbf{0}$.

- *Linear regression.* Desirable properties of estimators such as the lasso, group lasso, $\sqrt{\text{lasso}}$, group $\sqrt{\text{lasso}}$ or the Dantzig selector are satisfied if the tuning parameter is set to $\lambda = c\lambda_0$ for a certain $c \geqslant 1$, such that the event $\{\hat{\boldsymbol{\beta}}_{\lambda_0} = \mathbf{0} \mid \boldsymbol{\beta}^* = \mathbf{0}\}$ holds with high probability. More precisely, upper bounds on the estimation and prediction error, as well as the screening property hold with high probability assuming certain conditions on the regression matrix, the support $\mathcal{S}^*$ of the coefficients and their magnitude; see Bühlmann and van de Geer [2011]; Belloni et al. [2011]; Bunea et al. [2014] and references therein.

- *Low-rank matrix estimation.* Gavish and Donoho [2014] and Gavish and Donoho [2016] derive optimal singular value thresholding operators with threshold parameter λ_P for a variety of loss functions which are characterized by the fact that under the null model $X^* = 0_{N\times P}$ with a noise level of $1/\sqrt{N}$, $\mathbb{P}(\hat{X}_{\lambda_P} = 0_{N\times P}) \to 1$ as P tends to infinity.

Going back to Definition 2.2, the events $\{\hat{\boldsymbol{\xi}}_\lambda = \mathbf{0}\}$ and $\{\lambda(\mathbf{Y}) \leqslant \lambda\}$ are equiprobable. Consequently, selecting a threshold λ such that $\{\hat{\boldsymbol{\xi}}_\lambda = \mathbf{0} \mid \boldsymbol{\xi}^* = \mathbf{0}\}$ holds with high probability conducts us to the zero-thresholding function under the null model.

Definition 3.1 *Assume $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y})$ admits a zero-thresholding function $\lambda(\mathbf{Y})$. The null-thresholding statistic is*

$$\Lambda := \lambda(\mathbf{Y}_0) \tag{3.1}$$

with $\mathbf{Y}_0 =_d \mathbf{Y}$ under $H_0 : \boldsymbol{\xi}^* = \mathbf{0}$, provided Λ is not almost everywhere zero.

3.2 Definition and Theoretical Properties

The following selection rule allows to consider in a unified way the settings reviewed in the previous section.

Definition 3.2 *The quantile universal threshold λ^{QUT} is the upper α-quantile of Λ defined in* (3.1).

Remark 3.1 *If the distribution of Λ is unknown but the distribution of $\mathbf{Y}_0$ known, one can easily compute λ^{QUT} with a Monte Carlo simulation.*

Remark 3.2 *For estimators with a two-dimensional parameter* $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ *as elastic net and fused lasso, we define the quantile universal threshold* $\lambda_1^{\mathrm{QUT}}(\lambda_2)$ *as the upper* α*-quantile of* $\Lambda_{\lambda_2} := \lambda_1(\mathbf{Y}_0; \lambda_2)$ *for any fixed value of* λ_2*. Selection of* λ_2 *can then be performed by cross-validation.*

Remark 3.3 *In the random design setting, we define the quantile universal threshold as the upper* α*-quantile of* $\Lambda = \lambda(\mathbf{Y}_0, [X_0, X])$*, with* $[X_0, X]$ *consisting of independent identically distributed rows. As before, if the distribution of* Λ *is unknown,* λ^{QUT} *is easy to compute with a Monte Carlo simulation which requires bootstrapping the rows of* $[X_0, X]$*.*

Before considering the choice of α, some theoretical properties are derived. The first property concerns the control of the familywise error rate and false discovery rate. When performing multiple hypothesis tests, the familywise error rate is defined as the probability of incorrectly rejecting at least one null hypothesis. In the context of variable selection, it is the probability of erroneously selecting at least one variable. The false discovery rate is the expected proportion of incorrectly selected features among all selected features [Benjamini and Hochberg, 1995]. Under the overall null hypothesis, it can be shown that both quantities are equal. Hence, Definition 3.2 implies the following property.

Property 3.1 *Any thresholding estimator tuned with* λ^{QUT} *controls the familywise error rate as well as the false discovery rate at level* α *in the weak sense, that is, under the overal null hypothesis* $H_0 : \boldsymbol{\xi}^* = \mathbf{0}$*.*

Sparsity oracle inequalities are now derived. These hold under a certain compatibility condition which we first review. If there are more covariates than observations in the linear model

$$Y = X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I),$$

the least squares estimate is no longer defined since it requires $X^t X$ to be invertible, which can be written as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^P : \boldsymbol{\beta} \neq \mathbf{0}} \frac{\|X\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_1} > 0.$$

For regularized estimators such as the lasso, Dantzig selector and $\sqrt{\text{lasso}}$, a weaker condition, the $(L, \mathcal{S}^*)$-compatibility condition [Van de Geer, 2007]

$$\phi_{\mathrm{comp}}(L, \mathcal{S}^*) := \min_{\substack{\boldsymbol{\beta} \neq \mathbf{0}, \\ \|\boldsymbol{\beta}_{\overline{\mathcal{S}^*}}\|_1 \leqslant L \|\boldsymbol{\beta}_{\mathcal{S}^*}\|_1}} \frac{\sqrt{s^*}\|X\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}_{\mathcal{S}^*}\|_1} > 0, \tag{3.2}$$

with $\mathcal{S}^* = \{i \mid \beta_i^* \neq 0\}$ of cardinality s^* and a certain $L > 0$, ensures nice statistical properties.

Property 3.2 *The lasso tuned with $\lambda = \lambda^{\mathrm{QUT}}$ satisfies with probability at least $1-\alpha-\mathbb{P}(\lambda_0 \leqslant \Lambda \leqslant \lambda)$, provided the $(L, \mathcal{S}^*)$-compatibility condition is satisfied with $L = (\lambda+\lambda_0)/(\lambda-\lambda_0)$, for a certain λ_0, $0 < \lambda_0 < \lambda$,*

1. $\|X(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)\|_2^2/2 \leqslant 8(\lambda+\lambda_0)^2 s^*/\phi^2_{\mathrm{comp}}(L, \mathcal{S}^*)$,
2. $\|(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)_{\mathcal{S}^*}\|_1 \leqslant A, \quad A = 4(\lambda+\lambda_0)s^*/\phi^2_{\mathrm{comp}}(L, \mathcal{S}^*)$,
3. $\|(\hat{\boldsymbol{\beta}}_\lambda)_{\overline{\mathcal{S}^*}}\|_1 \leqslant 4(\lambda+\lambda_0)^2 s^*/\{(\lambda-\lambda_0)\phi^2_{\mathrm{comp}}(L, \mathcal{S}^*)\}$.

If, in addition,

$$\min_{p \in \mathcal{S}^*} |\beta_p^*| > A,$$

then with the same probability

$$\hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^*.$$

Remark that for F_Λ continuous, $\mathbb{P}(\lambda_0 \leqslant \Lambda \leqslant \lambda)$ can be made arbitrarily small for a well-chosen λ_0 as long as the $(L, \mathcal{S}^*)$-compatibility condition is met.

Property 3.3 *Under the $(L, \mathcal{S}^*)$-compatibility condition with $L = 1$, the conclusions of Property 3.2 hold when $\lambda_0 = 0$ with probability at least $1-\alpha$ for the Dantzig selector tuned with $\lambda = \lambda^{\mathrm{QUT}}$.*

The proof is omitted as it is essentially the same as for Theorem 7.1 in Bickel et al. [2009] using the fact that $\Lambda =_d \|X^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty$.

Remark 3.4 *Similar results can be shown for the $\sqrt{lasso}$ as well as the group lasso and group $\sqrt{lasso}$ tuned with our methodology, the latter two requiring a group compatibility condition.*

Proof [Property 3.2] The proof follows closely that of Theorem 6.1 in Bühlmann and van de Geer [2011]. We first show that for $\lambda^{\mathrm{QUT}} = \lambda > \lambda_0 > 0$, on the event $\{\|X^t\boldsymbol{\epsilon}\|_\infty \leqslant \lambda_0\}$, which has probability $\mathbb{P}(\Lambda \leqslant \lambda) - \mathbb{P}(\lambda_0 \leqslant \Lambda \leqslant \lambda)$,

$$\|(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)_{\overline{\mathcal{S}^*}}\|_1 \leqslant \frac{\lambda+\lambda_0}{\lambda-\lambda_0}\|(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)_{\mathcal{S}^*}\|_1. \tag{3.3}$$

In order to simplify notations in the proof, we omit the subscript of $\hat{\boldsymbol{\beta}}_\lambda$ and write $\mathcal{S}$ for $\mathcal{S}^*$. From the definition of the lasso estimator,

$$\begin{aligned}
\|\mathbf{Y} - X\hat{\boldsymbol{\beta}}\|_2^2/2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 &= \|X\boldsymbol{\beta}^* + \boldsymbol{\epsilon} - X\hat{\boldsymbol{\beta}}\|_2^2/2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
&= \|X\boldsymbol{\beta}^* - X\hat{\boldsymbol{\beta}}\|_2^2/2 + \|\boldsymbol{\epsilon}\|_2^2/2 \\
&\quad - 2\boldsymbol{\epsilon}^t X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)/2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
&\leqslant \|\mathbf{Y} - X\boldsymbol{\beta}^*\|_2^2/2 + \lambda\|\boldsymbol{\beta}^*\|_1 \\
&= \|\boldsymbol{\epsilon}\|_2^2/2 + \lambda\|\boldsymbol{\beta}^*\|_1,
\end{aligned}$$

which implies

$$\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leqslant \boldsymbol{\epsilon}^t X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \lambda\|\boldsymbol{\beta}^*\|_1. \tag{3.4}$$

From Hölder's inequality,

$$|\boldsymbol{\epsilon}^t X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \leqslant \|X^t\boldsymbol{\epsilon}\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1,$$

and hence

$$|\boldsymbol{\epsilon}^t X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \leqslant \lambda_0\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1.$$

Since

$$\|\hat{\boldsymbol{\beta}}\|_1 = \|\hat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1 + \|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1 \geqslant \|\boldsymbol{\beta}^*_{\mathcal{S}}\|_1 - \|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1 + \|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1,$$

it follows that

$$\begin{aligned}
\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/2 &\leqslant \lambda_0\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \lambda\|\boldsymbol{\beta}^*\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1 \quad \text{by (3.4)},\\
&\leqslant \lambda_0\{\|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1 + \|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1\} + \lambda\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_1\\
&\quad - \lambda\|\boldsymbol{\beta}^*_{\mathcal{S}}\|_1 + \lambda\|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1 - \lambda\|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1\\
&= (\lambda + \lambda_0)\|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1 + (\lambda_0 - \lambda)\|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1
\end{aligned} \tag{3.5}$$

and

$$(\lambda - \lambda_0)\|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1 \leqslant \|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + (\lambda - \lambda_0)\|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1 \leqslant (\lambda + \lambda_0)\|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1.$$

This achieves the proof of (3.3). Then, from (3.5),

$$\begin{aligned}
\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/2 &+ (\lambda - \lambda_0)\|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1\\
&+(\lambda + \lambda_0)\|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1 \leqslant 2(\lambda + \lambda_0)\|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1\\
&\leqslant 2(\lambda + \lambda_0)\sqrt{s^*}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2/\phi_{\text{comp}}\\
&\leqslant \|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/4 + 4(\lambda + \lambda_0)^2 s^*/\phi^2_{\text{comp}}
\end{aligned}$$

with the second and third inequality resulting respectively from the compatibility condition and the identity $4uv \leqslant u^2 + 4v^2$ for all real u, v. As a consequence,

$$\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2/4 + (\lambda - \lambda_0)\|\hat{\boldsymbol{\beta}}_{\bar{\mathcal{S}}}\|_1 + (\lambda + \lambda_0)\|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\|_1 \leqslant 4(\lambda + \lambda_0)^2 s^*/\phi^2_{\text{comp}}.$$

The inequalities of the theorem are then immediate. The last assertion follows from the inequality $\|\mathbf{v}\|_\infty \leqslant \|\mathbf{v}\|_1$, $\forall \mathbf{v} \in \mathbb{R}^P$ and from the fact that $\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{S}^*}\|_\infty \leqslant C$ for a certain $C > 0$ implies $\{i \mid \hat{\beta}_i \neq 0\} \supseteq \{i \mid |\beta^*_i| > C\}$. ■

The probability of the previous properties is determined by α; we recommend $\alpha = 0.05$ as Belloni et al. [2011]. An alternative is to set α_P tending to zero as the number P of covariates goes to infinity. Donoho and Johnstone [1994] implicitly select a rate of convergence of $\alpha_P = O(1/\sqrt{\log P})$ (Josse and Sardy [2016] also select this rate).

3.3 Derivation of λ^{QUT}

In some settings, we derive an explicit formulation of the quantile universal threshold such that α_P is $O(1/\sqrt{\log P})$.

In orthonormal regression with best subset selection, $\lambda(\mathbf{y})$ reduces to $\|X^t\mathbf{y}\|_\infty^2/2$ and $\Lambda =_d \|\mathbf{Z}\|_\infty^2/2$, where $Z \sim \mathrm{N}(0, \sigma^2 I)$. Setting $\lambda_P^{\mathrm{QUT}} = \sigma^2 \log P$, $\mathbb{P}(\Lambda > \lambda_P^{\mathrm{QUT}})$ is asymptotically equivalent to $1/\sqrt{\pi \log P}$ as we now show. The following result is classical [Williams, 1991].

Lemma 3.1 *For $t \geqslant 1$,*

$$\left(1 - \frac{1}{t^2}\right) \frac{\phi(t)}{t} \leqslant \bar{\Phi}(t) \leqslant \frac{\phi(t)}{t},$$

where ϕ and $\bar{\Phi}$ denote respectively the probability density and survival function of a standard normal variable.

Hence, $1 - 2\bar{\Phi}(t) = 1 - 2\phi(t)/t + O(\phi(t)/t^3)$ when $t \to \infty$. Setting $\lambda = \lambda_P^{\mathrm{QUT}}$,

$$\begin{aligned}
\mathbb{P}(2\Lambda/\sigma \leqslant \lambda) &= [1 - 2\Phi(-\lambda)]^P \\
&= [1 - 2\phi(\lambda)/\lambda + O(\phi(\lambda)/\lambda^3)]^P \\
&= 1 - 1/\sqrt{\pi \log P} + O(1/\log P),
\end{aligned}$$

with the last equality following from $(1+x)^P = \exp(P\log(1+x))$ and the fact that $\log(1+x) = x + O(x^2)$ and $\exp(x) = 1 + x + O(x^2)$ when $x \to 0$. It is then straightforward to show that $\mathbb{P}(\Lambda > \lambda_P^{\mathrm{QUT}}) \sim 1/\sqrt{\pi \log P}$. Note that $\lambda_P^{\mathrm{QUT}} = 2\lambda^{\mathrm{BIC}} = \sigma^2 \log P$. Generalizations such as GIC and EBIC also select a larger tuning parameter than BIC which is known to perform poorly in the high-dimensional setting.

In one-dimensional total variation, $\lambda(\mathbf{y}) = \|(DD^t)^{-1}D\mathbf{y}\|_\infty$ with D defined in (2.3), and the distribution of $(DD^t)^{-1}D\mathbf{Y}_0/\sigma\sqrt{P}$, $\mathbf{Y}_0 \sim \mathrm{N}(c\mathbf{1}, \sigma^2 I)$ for a certain constant c, is that of a discretized Brownian bridge W on an equispaced grid $t_i = i/P$, $i = 1, \ldots, P-1$, of $[0, 1]$ [Sardy and Tseng, 2004]. Then,

$$\begin{aligned}
\mathbb{P}(\Lambda \geqslant \sigma\sqrt{P} b_P) &\leqslant \mathbb{P}(\|W\| \geqslant b_P) \\
&= 2\sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 b_P^2),
\end{aligned}$$

with $\|\cdot\|$ denoting the sup norm and the equality being given in Shorack and Wellner [1986]. Setting $b_P = \sqrt{\log\log P}/2$, the quantity of interest converges to 0 at the rate $O(1/\sqrt{\log P})$ and $\lambda_P^{\mathrm{QUT}} = \sigma\sqrt{P \log\log P}/2$.

3.4 Numerical Results of Lasso GLM

Our methodology for lasso GLM is implemented in the **qut** package which is available from the Comprehensive R Archive Network (CRAN).

Assuming a Gaussian response, the lasso null-thresholding statistic (see Remark 2.1) is ancillary for $\boldsymbol{\beta}_0^*$. In contrast, $\sqrt{\text{lasso}}$'s null-thresholding statistic is pivotal with respect to both $\boldsymbol{\beta}_0^*$ and σ, and LAD lasso's is pivotal with respect to σ when $P_0 = 0$.

In regularized Poisson and logistic regression, the null-thresholding statistic depends on $\boldsymbol{\beta}_0^*$ which we estimate with the following procedure. First, calculate the MLE of $\boldsymbol{\beta}_0$ based on the observed value $\mathbf{y}$ with the constraint $\hat{\boldsymbol{\beta}} = \mathbf{0}$ (it is the solution to (2.10)). Then, solve lasso GLM with the corresponding quantile universal threshold. Finally, the estimate is $\hat{\boldsymbol{\beta}}_0^{\text{MLE}}$ where $(\hat{\boldsymbol{\beta}}_0^{\text{MLE}}, \hat{\boldsymbol{\beta}}^{\text{MLE}})$ denotes the MLE based on $\mathbf{y}$ with covariates selected by the previous procedure. In Appendix B.1, an empirical investigation of the sensitivity of λ^{QUT} to the estimation of $\boldsymbol{\beta}_0^*$ is conducted.

In the following, $\text{QUT}_{\text{lasso}}$ and $\text{QUT}_{\sqrt{\text{lasso}}}$ stand for the quantile universal threshold applied respectively to lasso and $\sqrt{\text{lasso}}$. CVmin refers to cross-validation, CV1se to a conservative variant of CVmin which takes into account the variability of the cross-validation error [Breiman et al., 1984], SS to stability selection [Meinshausen and Bühlmann, 2010] and GIC to the generalized information criterion [Fan and Tang, 2013]. We apply the latter selection rules to the lasso. For GIC and $\text{QUT}_{\text{lasso}}$, the variance is estimated with (B.2) and (B.3) respectively (see Appendix B.2). The parameter α is set to 0.05.

3.4.1 Real Data

We first briefly describe the four data sets considered to illustrate our approach in Gaussian and logistic regression.

- `riboflavin` [Bühlmann et al., 2014]: Riboflavin production rate measurements from a population of Bacillus subtilis with sample size $N = 71$ and expressions from $P = 4088$ genes.
- `chemometrics` [Sardy, 2008]: Fuel octane level measurements with sample size $N = 434$ and $P = 351$ spectrometer measurements.
- `leukemia` [Golub et al., 1999]: Cancer classification of human acute leukemia cancer types based on $N = 72$ samples of $P = 3571$ gene expression microarrays.

- `internetAd` [Kushmerick, 1999]: Classification of $N = 2359$ possible advertisements on internet pages based on $P = 1430$ features.

We randomly split one hundred times each data set into a training and a test set of equal size. Several selection rules are compared including QUT for random scenario. Except for CV1se, the final model is fitted by MLE with the previously selected covariates in order to improve prediction. In Figure 3.1, we report the number of nonzero coefficients selected on the training set, as well as the test set mean-squared prediction error and correct classification rate.

By selecting a large number of variables CV1se results in efficient prediction, whereas SS shows poor predictive performance due to the low complexity of the model. Good predictive performance is achieved by $\mathrm{QUT}_{\mathrm{lasso}}$ and $\mathrm{QUT}_{\sqrt{\mathrm{lasso}}}$ as well as GIC with a median model complexity between SS and CV1se. Moreover, GIC and $\mathrm{QUT}_{\sqrt{\mathrm{lasso}}}$ exhibit a larger variability than $\mathrm{QUT}_{\mathrm{lasso}}$ in terms of number of nonzero coefficients.

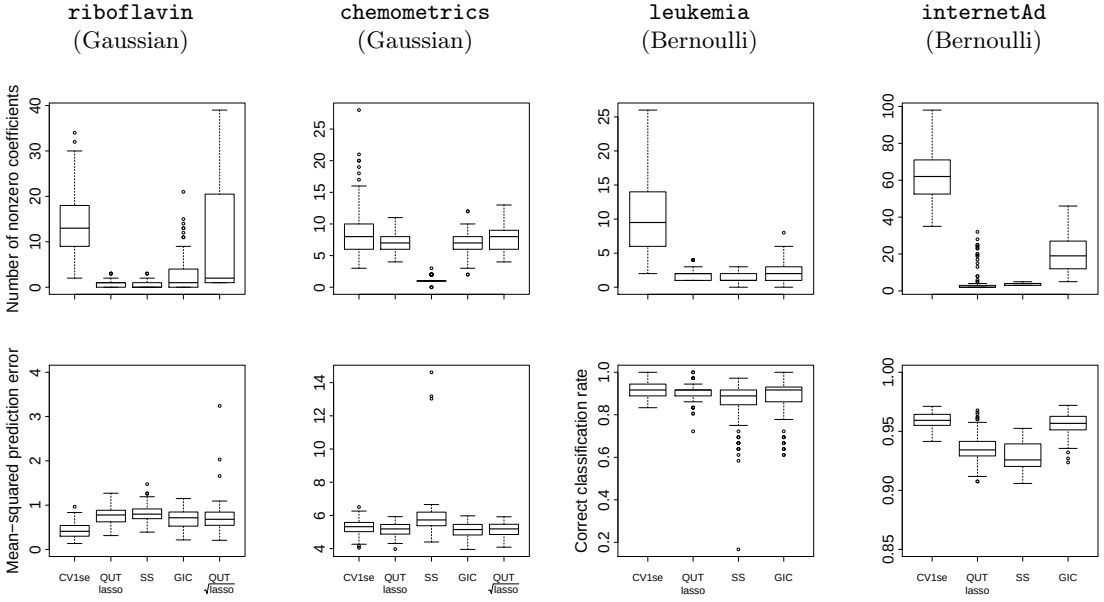

Figure 3.1 – Monte Carlo simulation based on four data sets: `riboflavin` (Gaussian), `chemometrics` (Gaussian), `leukemia` (Bernoulli) and `internetAd` (Bernoulli). We report the boxplots of the following statistics: the number of nonzero coefficients obtained from the training sets (top); the test set mean-squared prediction error for Gaussian responses and the correct classification rate for Bernoulli responses (bottom).

3.4.2 Synthetic Data

Two prominent quality measures of model selection are the true positive rate $\mathrm{TPR} := \mathbb{E}[\mathrm{TPr}]$ and the false discovery rate $\mathrm{FDR} := \mathbb{E}[\mathrm{FDr}]$, where $\mathrm{TPr} := |\hat{\mathcal{S}}_\lambda \cap \mathcal{S}^*|/|\mathcal{S}^*|$, the proportion of selected nonzero features among all nonzero features, and

FDr := $|\hat{\mathcal{S}}_\lambda \cap \overline{\mathcal{S}^*}|/|\hat{\mathcal{S}}_\lambda|$, the proportion of falsely selected features among all selected features.

We perform a simulation based on Reid et al. [2014]. Responses are generated from the linear, logistic and Poisson regression model with a sample size of $N = 100$ and $P = 1000$ covariates. The intercept is set to one and unit noise variance assumed in linear regression. The true parameter $\boldsymbol{\beta}^*$ and predictor matrix X are obtained as follows:

- Elements of X are generated randomly as $x_{ij} \sim \mathrm{N}(0, 1)$ with correlation between columns set to ω.
- The support of $\boldsymbol{\beta}^*$ is of cardinality $s^* = \lceil N^\theta \rceil$ and selected uniformly at random. Entries are generated from a Laplace(1) distribution and scaled according to a certain signal to noise ratio, snr $= \boldsymbol{\beta}^{*t}\Sigma_\omega\boldsymbol{\beta}^*/\sigma^2$, Σ_ω being the covariance matrix of a single row of X.

Table 3.1 – Estimated true positive rate/false discovery rate based on the simulation of Section 3.4.2.

Method	Response variable distribution with *(θ,ω,snr)*					
	Gaussian		Bernoulli		Poisson	
	(0.5,0,1)	*(0.1,0,1)*	*(0.5,0,10)*	*(0.1,0,10)*	*(0.5,0,1)*	*(0.3,0,1)*
lasso						
CV1se	0.21/0.26	0.76/0.17	0.25/0.43	0.74/0.37	0.50/0.60	0.82/0.60
QUT	0.08/0.00	0.69/0.00	0.09/0.02	0.62/0.00	0.54/0.69	0.82/0.65
SS	0.11/0.02	0.73/0.01	0.10/0.04	0.64/0.03	0.13/0.02	0.49/0.02
GIC	0.09/0.05	0.74/0.07	0.12/0.14	0.69/0.14	0.56/0.70	0.82/0.61
$\sqrt{\textbf{lasso}}$						
QUT	0.24/0.26	0.74/0.02				
	(0.5,0.4,1)	*(0.5,0,10)*	*(0.5,0.4,10)*	*(0.5,0,20)*	*(0.5,0.4,1)*	*(0.5,0,20)*
lasso						
CV1se	0.16/0.80	0.73/0.58	0.19/0.80	0.32/0.50	0.39/0.81	0.05/0.56
QUT	0.13/0.72	0.21/0.00	0.14/0.68	0.11/0.02	0.39/0.81	0.63/0.89
SS	0.03/0.03	0.27/0.00	0.03/0.04	0.13/0.03	0.04/0.08	0.02/0.50
GIC	0.06/0.26	0.52/0.20	0.07/0.37	0.17/0.14	0.39/0.81	0.77/0.93
$\sqrt{\textbf{lasso}}$						
QUT	0.16/0.78	0.32/0.02				

Table 3.1 contains estimated TPR and FDR based on one hundred replications. The high complexity of CV1se and the low complexity of SS are again observed. Moreover, QUT applied to lasso, GIC and QUT applied to $\sqrt{\text{lasso}}$ are comparable.

3.4.3 Phase Transition Property

In this section, we investigate the variable screening property and observe a phase transition. Given a thresholding estimator, if several tuning parameter values yield $\hat{\mathcal{S}}_\lambda$ containing the true support $\mathcal{S}^*$, the smallest estimated model can be of interest since it minimizes the FDr. We call it the optimal inclusive model. This leads to the definition of the oracle inclusive rate which measures its cardinality relative to the estimated support.

Definition 3.3 *Assume $\mathcal{S}^* \neq \varnothing$ and let $s_{\min} = \min_\lambda \{|\hat{\mathcal{S}}_\lambda| \mid \hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^*\}$ if it exists. Let $\hat{s}_\lambda = |\hat{\mathcal{S}}_\lambda|$ be the cardinality of $\hat{\mathcal{S}}_\lambda$. The oracle inclusive rate (OIR) is defined as $\mathbb{E}[\mathrm{OIr}]$, where*

$$\mathrm{OIr} := \begin{cases} \frac{s_{\min}}{\hat{s}_\lambda} & \text{if } \hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^* \\ 0 & \text{otherwise} \end{cases}.$$

Models with OIr $\neq 0$ have TPr $= 1$, whereas those with OIr $= 1$ have minimum FDr amongst all models with TPr $= 1$. Moreover, OIR $\leqslant \mathbb{P}(\hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^*)$. A small OIr results from a complex model containing $\mathcal{S}^*$, whereas a null OIr results from $\hat{\mathcal{S}}_\lambda \not\supseteq \mathcal{S}^*$. The latter could be due to a simplistic model or to the variable screening property being unachievable, in which case $s_{\min}$ does not exist.

To assess the performance of the quantile universal threshold in terms of OIR, we extend the simulation of Donoho and Tanner [2010] in compressed sensing to model (2.1) with unit noise variance assumed to be known. The entries of the $N \times P$ X matrix are assumed to be i.i.d. standard Gaussian. We set $P = 1600$ and vary the number of rows $N \in \{160, 320, 480, 640, 800, 960, 1120, 1280, 1440\}$ as well as the cardinality of the support of $\boldsymbol{\beta}^*$, $s^* \in \{1, \ldots, N\}$. Nonzero entries are set to ten. One hundred predictor matrices X and responses $\mathbf{y}$ are generated for each pair (N, s^*).

On the left and middle panel of Figure 3.2, we respectively report the estimated OIR of the oracle lasso selection rule which retains the optimal inclusive model if it exists and the estimated OIR of QUT. Values are plotted as a function of $\delta = N/P$, the undersampling factor, and of $\rho = s^*/N$, the sparsity factor. The right panel contains the estimated OIR of various selection rules for a fixed $\delta = 0.2$. The following interesting behaviours are observed.

- Phase transition of Oracle and QUT. Two regions can be clearly distinguished: a high OIR region due to a selected model containing few covariates outside the optimal model and a zero OIR region in which $s_{\min}$ does not exist. The change between these regions is abrupt, as observed in compressed sensing.

- Near oracle performance of QUT. Comparing the left and middle panels, the performance of QUT is nearly as good as that of the oracle selection rule, with the phase transition occurring at similar values of ρ.
- Low complexity of QUT. Comparing several rules on the right panel, QUT has a high OIR. Moreover, CVmin has lower OIR than CV1se and is comparable to SURE. The low OIR of the three latter selection rules is due to the complexity of their selected model. This goes along the fact they are prediction-based methodologies whereas QUT aims at a good identification of the parameters.

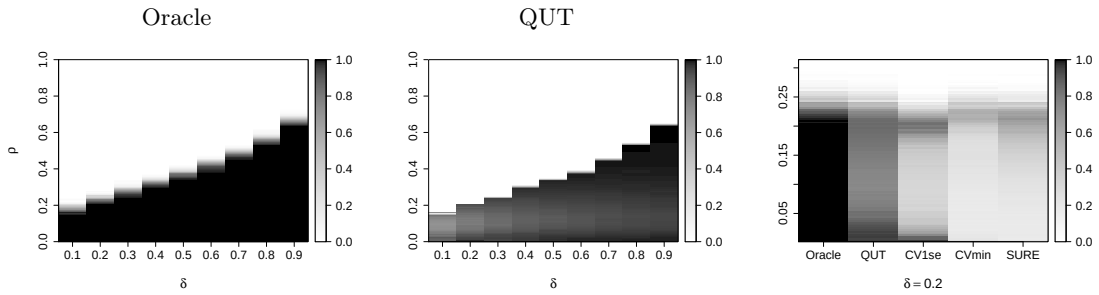

Figure 3.2 – Estimated OIR of the oracle lasso selection rule (left) and QUT (middle) as a function of $(\delta, \rho) = (N/P, s^*/N)$. The right panel contains the estimated OIR of several selection rules for a fixed $\delta = 0.2$.

3.4.4 Conclusion

According to Ockham's razor, if two selected models yield comparable predictive performances, the sparsest should be preferred. QUT tends to be in accordance with this principle. Moreover, a good compromise between high TPR and low FDR is obtained. Finally, near oracle performance is achieved in terms of OIR.

Chapter 4

Thresholding Tests

In this chapter, we consider the linear regression model

$$\mathbf{Y} = X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I), \tag{4.1}$$

and the problem of testing for $H_0 : A\boldsymbol{\beta}^* = \mathbf{c}$ for a given $R \times P$ full row rank matrix A and vector $\mathbf{c}$. A special case is the classical problem of testing whether a subset of coordinates have a linear relationship with the response.

We introduce a new class of testing procedures which are based on thresholding estimators and give examples in Section 4.1. Thresholding estimators are proposed in Section 4.2 which allow to test hypotheses of the form $H_0 : A\boldsymbol{\beta}^* = \mathbf{c}$. In addition, confidence regions for $A\boldsymbol{\beta}^*$ are derived. A comparative power study is performed in Section 4.3 under sparse and dense alternatives and an adaptive testing procedure is suggested. We show in Section 4.4 that the proposed testing procedures inherit a screening property. Finally, a necessary and sufficient condition is derived for a compatibility condition to hold in balanced one-way and two-way ANOVA designs. This condition ensures variable screening is achieved under an additional assumption on the magnitude of the nonzero entries of $\boldsymbol{\beta}^*$.

4.1 Methodology and Examples

We start by introducing the concept of a thresholding test.

Definition 4.1 *Assume $\hat{\boldsymbol{\xi}}_\lambda = A\hat{\boldsymbol{\beta}}_\lambda - \mathbf{c}$ is a thresholding estimator and consider testing*

$$H_0 : A\boldsymbol{\beta}^* = \mathbf{c}$$

for the linear model (4.1), where A is $R \times P$ of full row rank. The test whose test

function is given by

$$\phi(\mathbf{y}) = \begin{cases} 1 & \text{if } A\hat{\boldsymbol{\beta}}_\lambda \neq \mathbf{c} \\ 0 & \text{else} \end{cases}$$

is called a thresholding test with test-threshold λ.

The following proposition shows how to set the test-threshold to ensure the testing procedure has level α.

Proposition 4.1 *Assume* $A\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y}) - \mathbf{c}$ *admits a pivotal null-thresholding statistic*

$$\Lambda = \lambda(\mathbf{Y}_0),$$

where $\mathbf{Y}_0 =_d \mathbf{Y}$ *under* $H_0 : A\boldsymbol{\beta}^* = \mathbf{c}$. *Letting* $\lambda_\alpha = F_\Lambda^{-1}(1-\alpha)$, *the test*

$$\phi(\mathbf{y}) = \begin{cases} 1 & \text{if } \lambda(\mathbf{y}) > \lambda_\alpha \\ 0 & \text{otherwise} \end{cases}$$

is a thresholding test with test-threshold λ_α *of level* α.

The proof is immediate from Definition 3.1 of a null-thresholding statistic. Note that in the terminology of Section 3.2, λ_α is the quantile universal threshold at level α.

Remark 4.1 *If* F_Λ *is not invertible, we let* $F_\Lambda^{-1}(1-\alpha) := \inf\{\mathbf{x} \in \mathbb{R} \mid F_\Lambda(\mathbf{x}) \geqslant 1-\alpha\}$. *In that case, it might be that the test is conservative, that is, the maximum type I error probability is less than* α.

Consider testing for $\boldsymbol{\beta}^* = \mathbf{0}$ or for $\boldsymbol{\beta}^*_{\mathcal{A}} = \mathbf{0}$ with $\mathcal{A}$ an index subset of $\{1, \ldots, P\}$. Examples of thresholding estimators which admit a pivotal null-thresholding statistic Λ include the $\sqrt{\text{lasso}}$ and group $\sqrt{\text{lasso}}$, as well as the LAD lasso (see Section 2.2 and Remark 2.1). However, the lasso, group lasso and Dantzig selector do not admit a pivotal Λ with respect to σ. Considering a weighted form of these estimators, that is, substituting λ for $\lambda w(\mathbf{y})$, one can pivotize the test statistic. In effect, selecting the weight such that $w(\mathbf{Y}) > 0$ almost everywhere and $w(\mathbf{Y}_0)/\sigma$ is pivotal with respect to σ leads to the pivotal null-thresholding statistic $\Lambda/w(\mathbf{Y}_0)$, where Λ is the null-thresholding statistic of the non-weighted estimator. For example, one can set $w(\mathbf{Y}) = \|(I - P_X)\mathbf{Y}\|_2/\sqrt{\text{rank}(I - P_X)}$ as long as $C(X) \subset \mathbb{R}^N$ strictly.

4.2 The Affine Lasso

More generally, to test $H_0 : A\boldsymbol{\beta}^* = \mathbf{c}$ for an arbitrary full row rank matrix A, we introduce a new thresholding estimator, the *affine lasso*, defined by

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\text{argmin}} \ \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|A\boldsymbol{\beta} - \mathbf{c}\|_1. \tag{4.2}$$

If $\mathbf{c} = \mathbf{0}$, we retrieve the generalized lasso and if additionally $A = I$, the lasso. Similarly, the *group affine lasso* is defined by

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \sum_{k=1}^{M} \|A^{G_k}\boldsymbol{\beta} - \mathbf{c}_{G_k}\|_2.$$

We also introduce the *affine* $\sqrt{lasso}$ as well as the *group affine* $\sqrt{lasso}$ by substituting the ℓ_2 norm for the squared ℓ_2 norm divided by two. The resulting estimators of $g(\boldsymbol{\beta}^*) = A\boldsymbol{\beta}^* - \mathbf{c}$ admit a zero-thresholding function which we derive in the following proposition.

Proposition 4.2 *Let $\hat{\boldsymbol{\beta}}_\lambda$ denote the affine lasso, group affine lasso, affine $\sqrt{lasso}$ or group affine $\sqrt{lasso}$ estimator. Then, for any full row rank matrix A, $A\hat{\boldsymbol{\beta}}_\lambda - \mathbf{c}$ admits a zero-thresholding function $\lambda(\mathbf{y})$.*

Remark 4.2 *An explicit formulation of $\lambda(\mathbf{y})$ can be found in Table 4.1.*

Proof Consider the constrained least squares problem

$$\text{minimize: } \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 \qquad \text{subject to: } A\boldsymbol{\beta} = \mathbf{c}, \tag{4.3}$$

and let us denote any solution by $\hat{\boldsymbol{\beta}}^{H_0}$. From the Kuhn-Tucker theorem (Corollary 28.3.1 in Rockafellar [1970]), a given vector $\boldsymbol{\beta}$ is an optimal solution if and only if there exists a vector $\mathbf{z}$ such that

$$\begin{cases} A\boldsymbol{\beta} = \mathbf{c} \\ X^t(X\boldsymbol{\beta} - \mathbf{y}) + A^t\mathbf{z} = \mathbf{0} \end{cases}. \tag{4.4}$$

Since A has full row rank and the constrained least squares fit $X\hat{\boldsymbol{\beta}}^{H_0}$ is unique, the vector $\mathbf{z}$ is uniquely determined and we let $\mathbf{z}^{H_0} = (AA^t)^{-1}AX^t(\mathbf{y} - X\hat{\boldsymbol{\beta}}^{H_0})$. Let $\boldsymbol{\beta}_{\mathbf{c}}^{\perp}$ denote the unique element of $\ker A^{\perp}$ such that $A\boldsymbol{\beta}_{\mathbf{c}}^{\perp} = \mathbf{c}$. Then, the constrained least squares problem (4.3) is equivalent to that of minimizing $\|\mathbf{y} - X\boldsymbol{\beta}_{\mathbf{c}}^{\perp} - X\boldsymbol{\beta}\|_2^2/2$ subject to $A\boldsymbol{\beta} = \mathbf{0}$. Letting U denote a matrix whose columns span $\ker A$, the minimum is given by $\|\mathbf{y} - X\boldsymbol{\beta}_{\mathbf{c}}^{\perp} - P_{XU}(\mathbf{y} - X\boldsymbol{\beta}_{\mathbf{c}}^{\perp})\|_2^2/2$. It follows that $X\hat{\boldsymbol{\beta}}^{H_0} = X\boldsymbol{\beta}_{\mathbf{c}}^{\perp} + P_{XU}(\mathbf{y} - X\boldsymbol{\beta}_{\mathbf{c}}^{\perp})$. Hence, $\mathbf{z}^{H_0} = (AA^t)^{-1}AX^t(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}_{\mathbf{c}}^{\perp})$.

From the affine lasso KKT optimality conditions, a given vector $\boldsymbol{\beta}$ is a solution to the affine lasso (4.2) such that $A\boldsymbol{\beta} = \mathbf{c}$ if and only if there exists a vector $\mathbf{v}$ such that

$$\begin{cases} \|\mathbf{v}\|_\infty \leqslant 1 \\ X^t(X\boldsymbol{\beta} - \mathbf{y}) + \lambda A^t\mathbf{v} = \mathbf{0} \end{cases}. \tag{4.5}$$

Now assume the affine lasso admits a solution $\hat{\boldsymbol{\beta}}_\lambda$ such that $A\hat{\boldsymbol{\beta}}_\lambda - \mathbf{c} = \mathbf{0}$. It is easy to show that $\hat{\boldsymbol{\beta}}_\lambda$ is also a solution to the constrained least squares problem (4.3)

and hence both (4.4) and (4.5) hold with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_\lambda$. Since A is of full row rank, it follows that $\lambda \geqslant \|\mathbf{z}^{H_0}\|_\infty$. Conversely, if $\lambda \geqslant \|(AA^t)^{-1}AX^t(\mathbf{y} - X\hat{\boldsymbol{\beta}}^{H_0})\|_\infty$, then there exists $\mathbf{v}$ such that $\|\mathbf{v}\|_\infty \leqslant 1$ and $\lambda\mathbf{v} = \mathbf{z}^{H_0}$. Hence, (4.5) holds with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{H_0}$ and the affine lasso admits a solution $\hat{\boldsymbol{\beta}}_\lambda$ such that $A\hat{\boldsymbol{\beta}}_\lambda - \mathbf{c} = \mathbf{0}$.

The proofs for the group affine lasso, affine $\sqrt{\text{lasso}}$ and group affine $\sqrt{\text{lasso}}$ cases are similar and omitted. ■

Table 4.1 – Zero-thresholding function associated to different thresholding estimators.

	Zero-thresholding function $\lambda(\mathbf{y})$
Affine lasso	$\|(AA^t)^{-1}AX^t(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}_\mathbf{c}^\perp)\|_\infty$
Group affine lasso	$\max_{1\leqslant k\leqslant M} \|[(AA^t)^{-1}AX^t(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}_\mathbf{c}^\perp)]_{G_k}\|_2$
Affine $\sqrt{\text{lasso}}$	$\dfrac{\|(AA^t)^{-1}AX^t(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}_\mathbf{c}^\perp)\|_\infty}{\|(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}_\mathbf{c}^\perp)\|_2}$
Group affine $\sqrt{\text{lasso}}$	$\max_{1\leqslant k\leqslant M} \dfrac{\|[(AA^t)^{-1}AX^t(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}_\mathbf{c}^\perp)]_{G_k}\|_2}{\|(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}_\mathbf{c}^\perp)\|_2}$

As the following lemma shows, the zero-thresholding function of affine lasso and its variants is pivotal with respect to $\boldsymbol{\beta}^*$ under H_0. Proposition 4.1 then implies thresholding test procedures can be derived based on these estimators.

Lemma 4.3 *Let U denote a matrix whose columns span* $\ker A$ *and let $\boldsymbol{\beta}_\mathbf{c}^\perp$ denote the unique element of* $\ker A^\perp$ *such that $A\boldsymbol{\beta}_\mathbf{c}^\perp = \mathbf{c}$. Then, $(I - P_{XU})(\mathbf{Y} - X\boldsymbol{\beta}_\mathbf{c}^\perp)$ is pivotal with respect to $\boldsymbol{\beta}^*$ under the assumption that $A\boldsymbol{\beta}^* = \mathbf{c}$.*

Proof Since $A\boldsymbol{\beta}^* = \mathbf{c}$, $\boldsymbol{\beta}^* = \boldsymbol{\beta}^0 + \boldsymbol{\beta}_\mathbf{c}^\perp$ for a certain $\boldsymbol{\beta}^0 \in \ker A$. It follows that $(I - P_{XU})(\mathbf{Y} - X\boldsymbol{\beta}_\mathbf{c}^\perp) = (I - P_{XU})(X\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}) = (I - P_{XU})\boldsymbol{\epsilon}$ which concludes the proof. ■

As before, one can pivotize the null-thresholding statistic with respect to σ, if necessary, by substituting λ with $\lambda w(\mathbf{y})$ where $w(\mathbf{Y}) > 0$ almost everywhere and $w(\mathbf{Y_0})/\sigma$ pivotal with respect to σ.

In turns out the classical Fisher F-test is a thresholding test under certain assumptions. We let $\text{MSE} = \|(I - P_X)\mathbf{y}\|_2^2/\operatorname{rank}(I - P_X)$.

Proposition 4.4 *The weighted one group affine lasso thresholding test with*

$w(\mathbf{y}) = \sqrt{R \cdot \text{MSE}}$ is equivalent to Fisher's F-test if X is of full column rank and

$$A(X^tX)^{-1}A^t = dI$$

for a certain constant d. If $A = BX$ for a matrix B, it suffices that

$$A(X^tX)^{-}A^t = dI$$

where $(X^tX)^{-}$ denotes any generalized inverse of X^tX.

Proof Let $\hat{\boldsymbol{\beta}}$ denote any maximum likelihood estimator and $\hat{\boldsymbol{\beta}}^{H_0}$ any solution to the constrained least squares problem (4.3). If X has full column rank, it can be shown [Seber and Lee, 2003] that

$$X^tX\hat{\boldsymbol{\beta}}^{H_0} = X^t\mathbf{y} - A^t[A(X^tX)^{-1}A^t]^{-1}(A\hat{\boldsymbol{\beta}} - \mathbf{c}),$$

and if $A = BX$,

$$X^tX\hat{\boldsymbol{\beta}}^{H_0} = X^t\mathbf{y} - A^t[A(X^tX)^{-}A^t]^{-1}(A\hat{\boldsymbol{\beta}} - \mathbf{c}).$$

The proof then follows from the fact that the weighted one group affine lasso testing procedure rejects for large values of

$$\|(AA^t)^{-1}AX^t(\mathbf{y} - X\hat{\boldsymbol{\beta}}^{H_0})\|_2/\sqrt{R \cdot \text{MSE}},$$

whereas Fisher's testing procedure rejects for large values of

$$(A\hat{\boldsymbol{\beta}} - \mathbf{c})^t[A(X^tX)^{-1}A^t]^{-1}(A\hat{\boldsymbol{\beta}} - \mathbf{c})/(R \cdot \text{MSE})$$

if X has full column rank, and for large values of

$$(A\hat{\boldsymbol{\beta}} - \mathbf{c})^t[A(X^tX)^{-}A^t]^{-1}(A\hat{\boldsymbol{\beta}} - \mathbf{c})/(R \cdot \text{MSE})$$

if $A = BX$. ■

In particular, the two testing procedures are equivalent when testing for the significance of the main effect $\boldsymbol{\beta}^*$ in a balanced one-way ANOVA design

$$Y_{ij} = \mu^* + \beta_i^* + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2), \quad i = 1, \ldots, T,\ j = 1, \ldots, R, \tag{4.6}$$

and when testing for the significance of an interaction effect $\boldsymbol{\gamma}^*$, or for the significance of a main effect $\boldsymbol{\alpha}^*$ or $\boldsymbol{\beta}^*$, in a balanced two-way ANOVA design

$$Y_{ijk} = \mu^* + \alpha_i^* + \beta_j^* + \gamma_{ij}^* + \epsilon_{ijk}, \quad \epsilon_{ijk} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2), \tag{4.7}$$

$i = 1, \ldots, T_1$, $j = 1, \ldots, T_2$ and $k = 1, \ldots, R$.

We emphasize the fact that contrarily to Fisher's F-test, the $\sqrt{\text{lasso}}$ and group $\sqrt{\text{lasso}}$ tests do not require $C(X) \subset \mathbb{R}^N$ strictly. This implies testing can be performed even in the high-dimensional case $P > N$. However, all procedures require that the null hypothesis implies $\mathbb{E}(\mathbf{Y})$ lies in some subspace of $C(X)$. This is equivalent to $C(XU) \subset C(X)$ strictly, where the columns of U span $\ker A$. As a result, in the high-dimensional setting with $C(X) = \mathbb{R}^N$, one can test for the significance of a subset of coordinates of $\boldsymbol{\beta}^*$, say $\boldsymbol{\beta}^*_{\mathcal{A}}$, provided $C(X_{\bar{\mathcal{A}}}) \subset \mathbb{R}^N$.

Remark 4.3 (Confidence Regions) *Consider the problem of deriving a confidence region for $A\boldsymbol{\beta}^*$. The set $\text{CR}_{(1-\alpha)}(\mathbf{y})$ is said to be a $(1-\alpha)$ confidence region estimator of $A\boldsymbol{\beta}^*$ if*

$$\mathbb{P}(A\boldsymbol{\beta}^* \in \text{CR}_{(1-\alpha)}(\mathbf{Y})) = 1 - \alpha, \quad \forall \boldsymbol{\beta}^* \in \mathbb{R}^P.$$

It consists of all the vectors $\mathbf{c}$ that would not be rejected by an α level test of $A\boldsymbol{\beta}^* = \mathbf{c}$. Hence, confidence regions can be derived based on thresholding tests. For example, the affine $\sqrt{lasso}$ leads to the rejection region consisting of all the vectors $\mathbf{c}$ that satisfy the inequality

$$\|(AA^t)^{-1}AX^t(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}^{\perp}_{\mathbf{c}})\|_\infty \leqslant \lambda_\alpha \|(I - P_{XU})(\mathbf{y} - X\boldsymbol{\beta}^{\perp}_{\mathbf{c}})\|_2,$$

where $\lambda_\alpha = F^{-1}_\Lambda(1-\alpha)$ and $\Lambda =_d \|(AA^t)^{-1}AX^tZ\|_\infty / \|Z\|_2$ with $Z \sim \text{N}(\mathbf{0}, I - P_{XU})$.

4.3 Power Study

In this section, the performance of different testing procedures in terms of power is investigated when testing for

$$H_0 : \boldsymbol{\beta}^* = \mathbf{0}$$

against alternatives of the type

$$H_1^{s,\theta} : \text{Exactly } s \text{ entries of } \boldsymbol{\beta}^* \text{ are equal to } \theta$$

in two different models. A small value of s corresponds to a sparse alternative, whereas a large value of s corresponds to a dense alternative.

We first consider a balanced one-way ANOVA model (4.6). For an overall mean $\mu = 0$ and variance $\sigma^2 = 1$, an explicit formulation of the test-threshold and power of the one group lasso, lasso and LAD lasso tests is given in Table 4.2. Rewriting (4.6) as $\mathbf{Y} = B\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2 I_{TR})$, we make use of the fact that $B^t\mathbf{Y} \sim \text{N}(R\boldsymbol{\beta}^*, RI_T)$. For the LAD lasso, we make use of the independence of $(B_i^t \operatorname{sgn}(\mathbf{Y}) + R)/2$, $i = 1, \ldots, T$, which have binomial distribution with parameters R and $1 - \Phi(-\beta_i^*)$.

Table 4.2 – Test-threshold and power of different level α testing procedures under $H_1^{s,\theta}$-type alternatives. Notation: $\Delta_\theta\Phi(\lambda;R) = \Phi((\lambda - R\theta)/\sqrt{R}) - \Phi((-\lambda - R\theta)/\sqrt{R})$, $\Delta_\theta B(\lambda;R) = \mathbb{P}(\frac{-\lambda+R}{2} \leqslant W \leqslant \frac{\lambda+R}{2})$ with W a binomial random variable with parameters R and $p(\theta) = 1 - \Phi(-\theta)$ and $E = \{R - 2k \mid 0 \leqslant k \leqslant \lfloor R/2 \rfloor\}$.

	Test-threshold	Power(θ)
One group lasso	$\lambda_\alpha^{\mathrm{O}} = \sqrt{R}\sqrt{F^{-1}_{\chi^2_T}(1-\alpha)}$	$1 - F_{\chi^2_{T,Rs\theta^2}}(\lambda_\alpha^{\mathrm{O}2}/R)$
Lasso	$\lambda_\alpha^{+} = -\sqrt{R}\Phi^{-1}\left(\frac{1-(1-\alpha)^{1/T}}{2}\right)$	$1 - [\Delta_\theta\Phi(\lambda_\alpha^{+};R)]^s[\Delta_0\Phi(\lambda_\alpha^{+};R)]^{T-s}$
LAD lasso	$\lambda_\alpha^{\blacktriangle} = \min\{\lambda \in E \mid [\Delta_0 B(\lambda;R)]^T \geqslant 1-\alpha\}$	$1 - [\Delta_\theta B(\lambda_\alpha^{\blacktriangle};R)]^s[\Delta_0 B(\lambda_\alpha^{\blacktriangle};R)]^{T-s}$

In Figure 4.1, the power is reported as a function of θ for different sparsity levels when $T = 10$ and $R = 20$. For a highly sparse signal, the lasso test is more powerful than the group lasso test. As the signal becomes less sparse, the group lasso test should be preferred. In all settings, the LAD lasso, which is conservative, is the less powerful procedure.

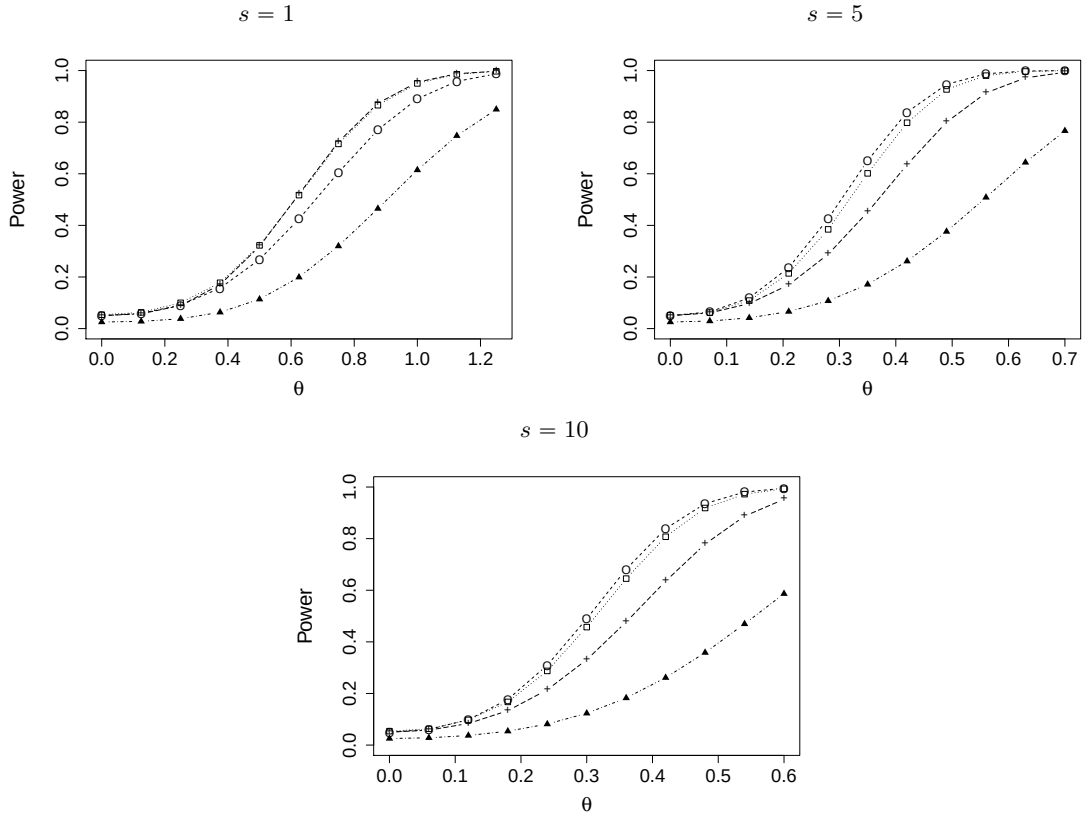

Figure 4.1 – Power of the one group lasso (circles), lasso (plus signs), LAD lasso (triangles) and empirical power of maximum (squares) testing procedures in a balanced one-way ANOVA model with $T = 10$ treatments and $R = 20$ replicates under $H_1^{s,\theta}$-type alternatives.

The second model considered is $\mathbf{Y} = \mathbf{1}\beta_0^* + X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, I)$ with X a 100×200 matrix with i.i.d. standard normal entries mean centered and rescaled to have unit ℓ_2 norm columns. The noise variance and β_0^* are assumed to be unknown. In Figure 4.2, we observe that for a highly sparse signal, the $\sqrt{\text{lasso}}$ testing procedure should be preferred. As the signal becomes less sparse, the group $\sqrt{\text{lasso}}$ is more powerful.

If one has no prior information to prefer one statistic over the other, we suggest to consider the maximum testing procedure. More precisely, for a known variance, the procedure rejects for large values of $\max(\|B^t\mathbf{y}\|_2/\lambda_\alpha^{\mathrm{O}}, \|B^t\mathbf{y}\|_\infty/\lambda_\alpha^+)$. The standardization ensures both individual test statistics possess the same rejection region at level α. In Figure 4.1, we observe this test has power close to the most powerful procedure across all the sparsity levels. Analogously, for an unknown variance, we consider the testing procedure which rejects for large values of the maximum of the $\sqrt{\text{lasso}}$ and group $\sqrt{\text{lasso}}$ test statistics, each of them standardized such that their respective rejection region at level α coincide. Again, it exhibits high power in all the settings of Figure 4.2. In summary, the maximum test is adaptive in the sense that it achieves high power across all the sparsity levels.

4.4 Screening Property

Let the linear model

$$\mathbf{Y} = X_0\boldsymbol{\beta}_0^* + X\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, I), \tag{4.8}$$

and consider testing $\boldsymbol{\beta}^* = \mathbf{0}$. Because the proposed testing procedures are based on thresholding estimators, the nonzero estimated entries give insight on the support of $\boldsymbol{\beta}^*$, $\mathcal{S}_{\boldsymbol{\beta}^*} := \{p \mid \beta_p^* \neq 0\}$. More formally, having rejected the null hypothesis $H_0 : \boldsymbol{\beta}^* = \mathbf{0}$, the screening property

$$\hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}_{\boldsymbol{\beta}^*},$$

where $\hat{\mathcal{S}}_\lambda = \{p \mid \hat{\beta}_p \neq 0\}$, holds with high probability under certain assumptions.

As a generalization of the compatibility condition (3.2), we first introduce the $(L, \mathcal{S}_{\boldsymbol{\beta}^*})$-intercept compatibility condition

$$\phi_{\text{intcomp}}(L, \mathcal{S}_{\boldsymbol{\beta}^*}) := \min_{\substack{\boldsymbol{\beta}\neq\mathbf{0}, \\ \|\boldsymbol{\beta}_{\overline{\mathcal{S}_{\boldsymbol{\beta}^*}}}\|_1 \leq L\|\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}^*}}\|_1}} \frac{\sqrt{s^*}\|(I - P_{X_0})X\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}^*}}\|_1} > 0,$$

with $\mathcal{S}_{\boldsymbol{\beta}^*}$ of cardinality s^* and a certain $L > 0$. We let $\text{MSE} = \|(I - P_{[X_0\, X]})\mathbf{y}\|_2^2/\operatorname{rank}(I - P_{[X_0\, X]})$.

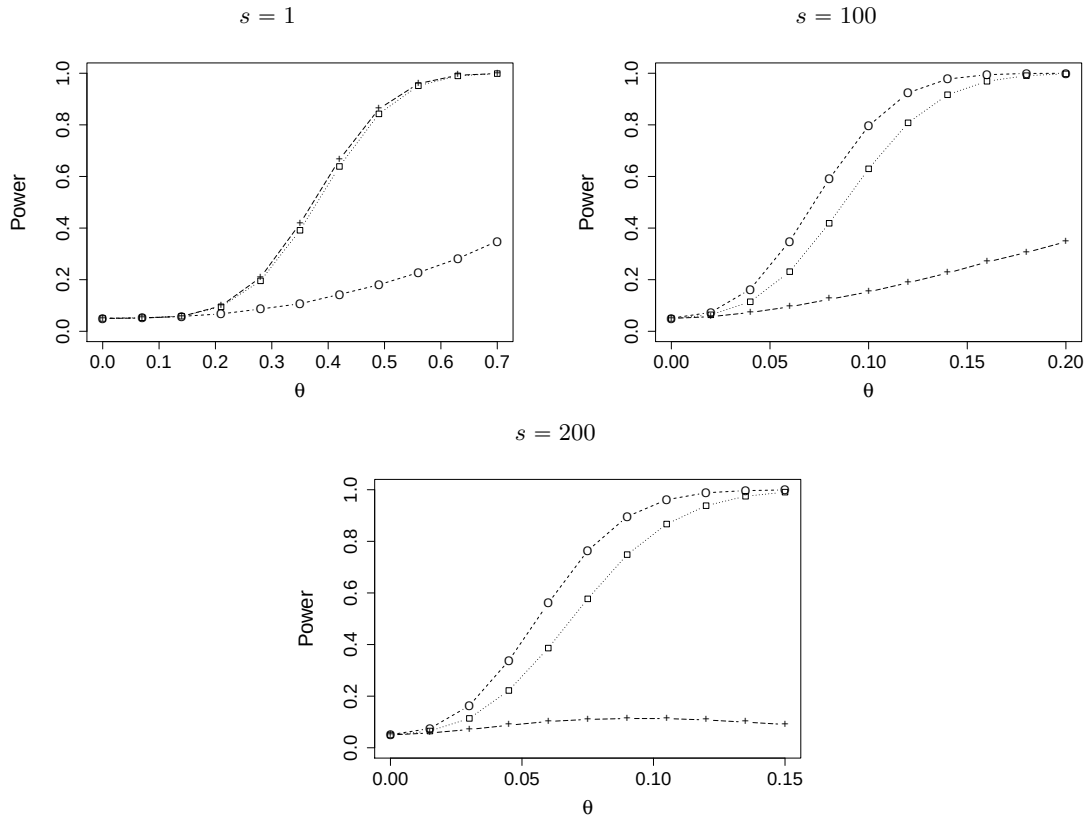

Figure 4.2 – Empirical power of the one group $\sqrt{\text{lasso}}$ (circles), $\sqrt{\text{lasso}}$ (plus signs) and maximum (squares) testing procedures in a linear model under $H_1^{s,\theta}$-type alternatives. The design matrix is of size 100×201 with a first column of ones and all other entries i.i.d. standard normal. The noise variance as well as the intercept value are assumed to be unknown.

Proposition 4.5 *Let λ_α be the test-threshold at level α associated to the weighted lasso testing procedure with weight $w(\mathbf{y}) = \sqrt{\text{MSE}}$ for testing $H_0 : \boldsymbol{\beta}^* = \mathbf{0}$ in the linear model (4.8). Assume the $(L, \mathcal{S}_{\boldsymbol{\beta}^*})$-intercept compatibility condition is satisfied with $L = (\lambda_\alpha + \lambda_0)/(\lambda_\alpha - \lambda_0)$, for a certain λ_0, $0 < \lambda_0 < \lambda_\alpha$. Assume $w(\mathbf{Y}) \leqslant \omega$ with probability at least $1-\delta$. Then, with probability at least $1-\alpha-\delta-\Pr(\lambda_0 \leqslant \Lambda \leqslant \lambda_\alpha)$,*

$$\hat{\mathcal{S}}_{\lambda_\alpha} \supseteq \mathcal{S}_{\boldsymbol{\beta}^*}$$

under the beta-min condition

$$\min_{p \in \mathcal{S}_{\boldsymbol{\beta}^*}} |\beta_p^*| > 4(\lambda_\alpha + \lambda_0) s^* \omega / \phi^2_{\text{intcomp}}(L, \mathcal{S}_{\boldsymbol{\beta}^*}).$$

Proof The proof follows closely that of Property 3.2. It consists in showing that on the event $\{\|X^t(I - P_{X_0})\boldsymbol{\epsilon}\|_\infty / w(\mathbf{Y}) \leqslant \lambda_0\}$,

$$\|(\hat{\boldsymbol{\beta}}_{\lambda_\alpha} - \boldsymbol{\beta}^*)_{\overline{\mathcal{S}_{\boldsymbol{\beta}^*}}}\|_1 \leqslant \frac{\lambda_\alpha + \lambda_0}{\lambda_\alpha - \lambda_0} \|(\hat{\boldsymbol{\beta}}_{\lambda_\alpha} - \boldsymbol{\beta}^*)_{\mathcal{S}_{\boldsymbol{\beta}^*}}\|_1.$$

The intercept compatibility condition then implies

$$\|(\hat{\boldsymbol{\beta}}_{\lambda_\alpha} - \boldsymbol{\beta}^*)_{\mathcal{S}_{\boldsymbol{\beta}^*}}\|_1 \leqslant 4(\lambda_\alpha + \lambda_0)s^* w(\mathbf{Y})/\phi^2_{\text{intcomp}}(L, \mathcal{S}_{\boldsymbol{\beta}^*}).$$

Since $w(\mathbf{Y}) = w(\boldsymbol{\epsilon})$, $\|X^t(I - P_{X_0})\boldsymbol{\epsilon}\|_\infty / w(\mathbf{Y}) \leqslant \lambda_0$ has probability $1 - \alpha - \Pr(\lambda_0 \leqslant \Lambda \leqslant \lambda_\alpha)$. Moreover, $w(\mathbf{Y}) \leqslant \omega$ has probability at least $1-\delta$. Hence, with probability at least $1 - \alpha - \delta - \Pr(\lambda_0 \leqslant \Lambda \leqslant \lambda_\alpha)$,

$$\|(\hat{\boldsymbol{\beta}}_{\lambda_\alpha} - \boldsymbol{\beta}^*)_{\mathcal{S}_{\boldsymbol{\beta}^*}}\|_1 \leqslant 4(\lambda_\alpha + \lambda_0)s^* \omega/\phi^2_{\text{intcomp}}(L, \mathcal{S}_{\boldsymbol{\beta}^*}).$$

We then use the fact that $\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{S}_{\boldsymbol{\beta}^*}}\|_\infty \leqslant C$ for a certain $C > 0$ implies $\{p \mid \hat{\beta}_p \neq 0\} \supseteq \{p \mid |\beta^*_p| > C\}$. ■

Remark 4.4 *A similar result can be shown for the weighted group lasso. For the* $\sqrt{\text{lasso}}$ *and group* $\sqrt{\text{lasso}}$*, we only consider the case where* $w(\mathbf{Y}) = 1$ *and* $\omega = 1$ *and hence* $\delta = 0$ *since the null-thresholding statistic does not need to be pivotized.*

The following lemma gives an explicit formulation of ω appearing in Proposition 4.5.

Lemma 4.6 *For a given* $0 < \delta < 1$*, let* $t = \sqrt{\frac{4\log(1/\delta)}{\text{rank}(I-P)}} + \frac{2\log(1/\delta)\|I-P\|}{\text{rank}(I-P)}$*, where* P *denotes the orthogonal projection matrix onto* $C([X_0\, X])$ *and* $\|\cdot\|$ *denotes the spectral norm of a matrix. Then,*

$$\mathbb{P}(\sqrt{\text{MSE}} \leqslant \sigma\sqrt{1+t}) \geqslant 1 - \delta.$$

The proof of this result is a direct application of Proposition 1.1 in Hsu et al. [2012].

Several simple conditions for the compatibility condition to hold are given in [Bickel et al., 2009]. The following proposition establishes that when testing for significance of a main effect, in balanced one-way and two-way ANOVA models, the intercept compatibility reduces to a constraint on the cardinality s^*.

Proposition 4.7 *In a balanced one-way ANOVA model* (4.6)*, the* $(L, \mathcal{S}_{\boldsymbol{\beta}^*})$*-intercept compatibility condition is equivalent to* $|\overline{\mathcal{S}_{\boldsymbol{\beta}^*}}| > L|\mathcal{S}_{\boldsymbol{\beta}^*}|$*, that is,* $s^* < T/(L+1)$*.*

In a balanced two-way ANOVA model without interaction, the $(L, \mathcal{S}_{\boldsymbol{\alpha}^*})$*- and* $(L, \mathcal{S}_{\boldsymbol{\beta}^*})$*-intercept compatibility condition hold respectively if and only if* $|\overline{\mathcal{S}_{\boldsymbol{\alpha}^*}}| > L|\mathcal{S}_{\boldsymbol{\alpha}^*}|$ *and* $|\overline{\mathcal{S}_{\boldsymbol{\beta}^*}}| > L|\mathcal{S}_{\boldsymbol{\beta}^*}|$*.*

Proof We reformulate the balanced one-way ANOVA model as $\mathbf{Y} = \mathbf{1}\mu^* + B\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$. By definition, the $(L, \mathcal{S}_{\boldsymbol{\beta}^*})$-intercept compatibility condition holds if and only if

$$\ker[(I - P_{\mathbf{1}})B] \cap \{\boldsymbol{\beta} \in \mathbb{R}^T \backslash \{\mathbf{0}\} \mid \|\boldsymbol{\beta}_{\overline{\mathcal{S}_{\boldsymbol{\beta}^*}}}\|_1 \leqslant L\|\boldsymbol{\beta}_{\mathcal{S}_{\boldsymbol{\beta}^*}}\|_1\} = \varnothing.$$

Some computations lead to $\ker[(I - P_{\mathbf{1}})B] = \{\boldsymbol{\beta} \in \mathbb{R}^T \mid \beta_1 = \beta_2 = \ldots = \beta_T\}$, and it is then easy to show that the $(L, \mathcal{S}_{\boldsymbol{\beta}^*})$-intercept compatibility condition is equivalent to $|\overline{\mathcal{S}_{\boldsymbol{\beta}^*}}| > L|\mathcal{S}_{\boldsymbol{\beta}^*}|$.

The proof for the balanced two-way ANOVA model is analogous. We reformulate the model as $\mathbf{Y} = \mathbf{1}\mu^* + A\boldsymbol{\alpha}^* + B\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, and the result follows from $\ker[(I - P_{[\mathbf{1}\,B]})A] = \{\boldsymbol{\alpha} \in \mathbb{R}^{T_1} \mid \alpha_1 = \alpha_2 = \ldots = \alpha_{T_1}\}$ and $\ker[(I - P_{[\mathbf{1}\,A]})B] = \{\boldsymbol{\beta} \in \mathbb{R}^{T_2} \mid \beta_1 = \beta_2 = \ldots = \beta_{T_2}\}$. ■

Remark 4.5 *Consider the balanced two-way ANOVA model with interaction* (4.7) *which we reformulate as* $\mathbf{Y} = \mathbf{1}\mu^* + A\boldsymbol{\alpha}^* + B\boldsymbol{\beta}^* + C\boldsymbol{\gamma}^* + \boldsymbol{\epsilon}$. *After some derivations, we obtain*

$$\ker[(I - P_{[\mathbf{1}\,A\,B]})C] = \{\boldsymbol{\gamma} \in \mathbb{R}^{T_1 T_2} \mid \gamma_{ij} = \bar{\gamma}_{\cdot j} + \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot\cdot}\},$$

with $\bar{\gamma}_{\cdot j} = \sum_{i=1}^{T_1} \gamma_{ij}/T_1$, $\bar{\gamma}_{i\cdot} = \sum_{j=1}^{T_2} \gamma_{ij}/T_2$ *and* $\bar{\gamma}_{\cdot\cdot} = \sum_{i=1}^{T_1}\sum_{j=1}^{T_2} \gamma_{ij}/(T_1T_2)$. *The* $(L, \mathcal{S}_{\boldsymbol{\gamma}^*})$*-intercept compatibility is then equivalent to*

$$\max_{\substack{\boldsymbol{\gamma}\neq\mathbf{0},\\ \boldsymbol{\gamma}\in\Gamma}} \frac{\|\boldsymbol{\gamma}_{\mathcal{S}_{\boldsymbol{\gamma}^*}}\|_1}{\|\boldsymbol{\gamma}\|_1} < \frac{1}{L+1},$$

with $\Gamma := \ker[(I - P_{[\mathbf{1}\,A\,B]})C]$. *If* $\pi_{\mathcal{S}_{\boldsymbol{\gamma}^*}}$ *denotes the linear application defined on* Γ *such that* $\pi_{\mathcal{S}_{\boldsymbol{\gamma}^*}}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}_{\mathcal{S}_{\boldsymbol{\gamma}^*}}$, *the* $(L, \mathcal{S}_{\boldsymbol{\gamma}^*})$*-intercept compatibility reduces to the compact form* $\|\pi_{\mathcal{S}_{\boldsymbol{\gamma}^*}}\|_1 < 1/(L+1)$.

Chapter 5

Existence of Estimates of GLMs with Convex Penalties

Many of the thresholding estimators we have encountered so far are defined as

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\operatorname{argmin}} \ \ell_{\mathbf{y}}(\boldsymbol{\beta}) + p_\lambda(\boldsymbol{\beta}),$$

for a certain loss function $\ell_{\mathbf{y}}$ and penalty function p_λ, both convex. This chapter is concerned with the problem of existence of such estimators.

We start in Section 5.1 by giving sufficient conditions on the loss and penalty functions for the existence of a solution and show that it is guaranteed for a large class of estimators. Section 5.2 adresses the case where some parameters are left unpenalized since they are assumed a priori to be nonzero. It contains necessary and sufficient conditions for the minimum set to be nonempty when the loss function is the negative log-likelihood of a canonical GLM. Finally, a numerical procedure to check for existence of regularized logistic and Poisson regression estimators is given in Section 5.3. The approach taken is linked to several objects in convex analysis which are reviewed in Appendix A.

5.1 The Fully Penalized Case

We first consider the case where all parameters are penalized.

Theorem 5.1 *Assume $\ell_{\mathbf{y}}$ is a proper closed convex function with nonnegative recession function $\ell_{\mathbf{y}}0^+$ and p_λ is a finite closed convex function with positive recession function $p_\lambda 0^+$ for all $\boldsymbol{\beta} \neq \mathbf{0}$. Then, the minimum set of*

$$\ell_{\mathbf{y}}(\boldsymbol{\beta}) + p_\lambda(\boldsymbol{\beta}) \tag{5.1}$$

is nonempty.

Proof The objective function is proper and Proposition A.2 implies it is closed and convex. Moreover, its recession function is positive for any $\boldsymbol{\beta} \neq \mathbf{0}$ since it is given by $\ell_{\mathbf{y}}0^+ + p_\lambda 0^+$ by Proposition A.4. Hence, the objective function has no directions of recession and by Theorem A.6 the infimum is attained. ■

As a consequence, we obtain existence of many classical penalized estimators. The following lemma establishes that assumptions of Theorem 5.1 reduce to $\ell_{\mathbf{y}}0^+ \geqslant 0$ if $\ell_{\mathbf{y}}$ is the negative log-likelihood of a canonical GLM, that is,

$$\ell_{\mathbf{y}}(\boldsymbol{\beta}) = -\langle \mathbf{y}, X\boldsymbol{\beta} \rangle + b(X\boldsymbol{\beta}),$$

for $\boldsymbol{\beta}$ such that $X\boldsymbol{\beta} \in \Theta := \{\boldsymbol{\theta} \in \mathbb{R}^N \mid b(\boldsymbol{\theta}) < \infty\}$. We extend $\ell_{\mathbf{y}}$ to $\mathbb{R}^P$ by setting $b(X\boldsymbol{\beta}) = +\infty$ if $X\boldsymbol{\beta} \notin \Theta$, so that we consider minimizing (5.1) over $\mathbb{R}^P$.

Lemma 5.2 *The negative log-likelihood of a canonical GLM takes its values in* $]-\infty, \infty]$*, is convex and lower semicontinuous.*

Proof From Proposition A.1 and A.2, it suffices to show that

$$b(\boldsymbol{\theta}) = \log \int e^{\langle \boldsymbol{\theta}, \mathbf{y} \rangle} d\mu(\mathbf{y})$$

takes its values in $]-\infty, \infty]$, is convex and lower semicontinuous.

Since $e^{\langle \boldsymbol{\theta}, \mathbf{y} \rangle} > 0$, $b(\boldsymbol{\theta})$ is never equal to $-\infty$.

Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ such that $b(\boldsymbol{\theta}_1) < \infty$ and $b(\boldsymbol{\theta}_2) < \infty$. Then, by Hölder's inequality,

$$\int e^{\langle \lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2, \mathbf{y} \rangle} d\mu(\mathbf{y}) \leqslant \left[\int e^{\langle \boldsymbol{\theta}_1, \mathbf{y} \rangle} d\mu(\mathbf{y})\right]^{\lambda} \left[\int e^{\langle \boldsymbol{\theta}_2, \mathbf{y} \rangle} d\mu(\mathbf{y})\right]^{1-\lambda},$$

for $0 < \lambda < 1$. Hence, the set Θ is convex and the restriction of b to Θ is convex. Since b is equal to $+\infty$ on $\Theta \backslash C$, it is convex on all of $\mathbb{R}^N$.

By Fatou's lemma, if $\boldsymbol{\theta}_n$ converges to $\boldsymbol{\theta}$ as n tends to ∞,

$$\int e^{\langle \boldsymbol{\theta}, \mathbf{y} \rangle} d\mu(\mathbf{y}) \leqslant \liminf_{n \to \infty} \int e^{\langle \boldsymbol{\theta}_n, \mathbf{y} \rangle} d\mu(\mathbf{y}).$$

The lower semicontinuity of $b(\boldsymbol{\theta})$ then follows from the fact it is the composition of the logarithmic function with a positive lower semicontinuous function. ■

We now compute explicitly the recession function of several loss functions using formula (A.2).

- *Squared ℓ_2 loss.* If $\ell_{\mathbf{y}}(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$, then $\ell_{\mathbf{y}}0^+(\boldsymbol{\beta}) = 0$ if $X\boldsymbol{\beta} = 0$ and $+\infty$ otherwise.
- *ℓ_1 and ℓ_2 loss.* If $\ell_{\mathbf{y}}(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|_i$ for $i = 1$ or 2, then $\ell_{\mathbf{y}}0^+(\boldsymbol{\beta}) = \|X\boldsymbol{\beta}\|_i$.

- *Logistic regression loss.* If $\ell_{\mathbf{y}}(\boldsymbol{\beta}) = \sum_{i=1}^{N} -y_i(\mathbf{x}^i\boldsymbol{\beta}) + \log(1 + e^{\mathbf{x}^i\boldsymbol{\beta}})$, then

$$\ell_{\mathbf{y}}0^+(\boldsymbol{\beta}) = \textstyle\sum_{\{i|\mathbf{x}^i\boldsymbol{\beta}>0\}}(1-y_i)\mathbf{x}^i\boldsymbol{\beta} - \sum_{\{i|\mathbf{x}^i\boldsymbol{\beta}<0\}} y_i(\mathbf{x}^i\boldsymbol{\beta}).$$

- *Poisson regression loss.* If $\ell_{\mathbf{y}}(\boldsymbol{\beta}) = \sum_{i=1}^{N} -y_i(\mathbf{x}^i\boldsymbol{\beta}) + e^{\mathbf{x}^i\boldsymbol{\beta}}$, then

$$\ell_{\mathbf{y}}0^+(\boldsymbol{\beta}) = \begin{cases} +\infty & \text{if } \mathbf{x}^i\boldsymbol{\beta} > 0 \text{ for at least one } i \\ \sum_{i=1}^{N} -y_i(\mathbf{x}^i\boldsymbol{\beta}) & \text{else} \end{cases}.$$

- *Multinomial logistic regression loss.* In this setting, each entry of the response variable $\mathbf{G}$ can take on one of K possible values. Letting Y be the $N \times K$ indicator response matrix with elements $y_{ik} = \mathbf{1}_{\{g_i=k\}}$, and setting

$$\mathbb{P}(G_i = k) = \frac{e^{\mathbf{x}^i\boldsymbol{\beta}^{(k)}}}{\sum_{j=1}^{K} e^{\mathbf{x}^i\boldsymbol{\beta}^{(j)}}},$$

  the negative log-likelihood takes the form

$$\ell_Y(\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(K)}) = \textstyle\sum_{i=1}^{N} \Big(\sum_{k=1}^{K} -y_{ik}(\mathbf{x}^i\boldsymbol{\beta}^{(k)}) + \log\big(\sum_{k=1}^{K} e^{\mathbf{x}^i\boldsymbol{\beta}^{(k)}}\big)\Big).$$

  One can then show that

$$\ell_Y 0^+(\boldsymbol{\beta}) = \textstyle\sum_{i=1}^{N} \Big(\sum_{k=1}^{K} -y_{ik}(\mathbf{x}^i\boldsymbol{\beta}^{(k)}) + \max_{1\leqslant k\leqslant K} \mathbf{x}^i\boldsymbol{\beta}^{(k)}\Big).$$

It is easy to show conditions of Theorem 5.1 are met for the squared ℓ_2, ℓ_1 and ℓ_2 losses. The hypotheses are also satisfied for the other loss functions considered which are the negative log-likelihood of a canonical GLM, since their recession function is positive.

Penalty functions satisfying the required assumptions include the lasso $p_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$, the group lasso $p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}_{G_k}\|_2$, the generalized lasso, the affine lasso and composite penalties such as the elastic net and fused lasso. This follows from the fact they all satisfy $p_\lambda 0^+(\boldsymbol{\beta}) = p_\lambda(\boldsymbol{\beta}) > 0$ for $\boldsymbol{\beta} \neq 0$.

5.2 The Partially Penalized Case

If some parameters are assumed a priori to be nonzero, it is customary to let these parameters unpenalized. The objective function (5.1) then becomes

$$\ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + p_\lambda(\mathbf{0}, \boldsymbol{\beta}). \tag{5.2}$$

Assume p_λ as in Theorem 5.1 such that $p_\lambda 0^+(\mathbf{0}, \mathbf{0}) = \mathbf{0}$. Then, for the ℓ_1, ℓ_2 and squared ℓ_2 losses, existence is guaranteed. This follows from Theorem A.6, since the directions of recession, which are given by $\{(\boldsymbol{\beta}_0, \mathbf{0}) \neq \mathbf{0} \mid X_0\boldsymbol{\beta}_0 = \mathbf{0}\}$, are all directions of constancy. If $\ell_{\mathbf{y}}$ is taken to be the negative log-likelihood of a

canonical GLM, there exists directions of recession of (5.2) which are not directions of constancy for certain response distributions.

We now give necessary and sufficient conditions for the existence of a regularized GLM estimator. As we will see, it is equivalent to the existence of a solution to a constrained least squares problem. Our treatment is based on that of Geyer [2009] for the existence of MLEs in GLMs. We denote by $\pi_1 : \mathbb{R}^{P_0+P} \to \mathbb{R}^{P_0}$ the projection onto the first P_0 coordinates. The normal cone and tangent cone of a convex set D in $\mathbb{R}^N$ at a point $\mathbf{x} \in D$ are respectively denoted by $N_D(\mathbf{x})$ and $T_D(\mathbf{x})$.

Theorem 5.3 *Assume the response vector $\mathbf{Y} = (Y_1, \ldots, Y_N)^t$ has density of the form*

$$\exp\{\langle \mathbf{y}, X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta}\rangle - b(X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta}) + c(\mathbf{y})\},$$

with $(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ such that $X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta} \in \Theta := \{\boldsymbol{\theta} \in \mathbb{R}^N \mid b(\boldsymbol{\theta}) < \infty\}$ and let

$$\ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = -\langle \mathbf{y}, X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta}\rangle + b(X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta}) \quad \text{and} \quad \ell^0_{\mathbf{y}}(\boldsymbol{\beta}_0) = \ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \mathbf{0}).$$

Let p_λ be a finite closed convex function such that $p_\lambda 0^+(\mathbf{0}, \boldsymbol{\beta}) \geqslant 0$ with equality if and only if $\boldsymbol{\beta} = \mathbf{0}$. Moreover, let C denote the smallest closed convex set containing $(X_0^t\mathbf{Y}, X^t\mathbf{Y})$ with probability one.

Then, for an observed value $X_0^t\mathbf{y} \in \pi_1(C)$ such that $\ell_{\mathbf{y}}0^+ \geqslant 0$, the following are equivalent:

(a) The minimum set of $\ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + p_\lambda(\mathbf{0}, \boldsymbol{\beta})$ is non-empty.

(b) The minimum set of $\ell^0_{\mathbf{y}}(\boldsymbol{\beta}_0)$ is non-empty.

(c) Every direction of recession of $\ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + p_\lambda(\mathbf{0}, \boldsymbol{\beta})$ is a direction of constancy of $\ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + p_\lambda(\mathbf{0}, \boldsymbol{\beta})$.

(d) Every direction of recession of $\ell^0_{\mathbf{y}}$ is a direction of constancy of $\ell^0_{\mathbf{y}}$.

(e) $N_{\pi_1(C)}(X_0^t\mathbf{y})$ is a vector subspace.

(f) $T_{\pi_1(C)}(X_0^t\mathbf{y})$ is a vector subspace.

Remark 5.1 *We obtain a characterization of the set $\mathcal{D}$ appearing in lasso and group lasso GLM zero-thresholding functions in* (2.9) *and* (2.11) *since (b) of Theorem 5.3 is equivalent to $\mathbf{y}$ belonging to $\mathcal{D}$.*

Throughout the proof, we will make use of the following lemma.

Lemma 5.4 *Under the assumptions of Theorem 5.3, the directions of recession of $\ell^0_{\mathbf{y}}$ are given by $\{\boldsymbol{\phi}_0 \mid (\boldsymbol{\phi}_0, \mathbf{0})$ is a direction of recession of $\ell_{\mathbf{y}}\}$ and its directions of constancy by $\{\boldsymbol{\phi}_0 \mid (\boldsymbol{\phi}_0, \mathbf{0})$ is a direction of constancy of $\ell_{\mathbf{y}}\}$.*

Moreover, the directions of recession of $g_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + p_\lambda(\mathbf{0}, \boldsymbol{\beta})$ are given by $\{(\boldsymbol{\phi}_1, \mathbf{0}) \mid \boldsymbol{\phi}_1$ is a direction of recession of $\ell^0_{\mathbf{y}}\}$ and its directions of constancy by $\{(\boldsymbol{\phi}_1, \mathbf{0}) \mid \boldsymbol{\phi}_1$ is a direction of constancy of $\ell^0_{\mathbf{y}}\}$.

Proof By Theorem A.5, a nonzero vector $\boldsymbol{\phi}_0$ is a direction of constancy of $\ell_{\mathbf{y}}^0$ if and only if $\ell_{\mathbf{y}}^0(\mathbf{u} + s\boldsymbol{\phi}_0) = \ell_{\mathbf{y}}(\mathbf{u} + s\boldsymbol{\phi}_0, \mathbf{0})$ is constant as a function of s, for all $\mathbf{u}$. Since $\ell_{\mathbf{y}}^0$ and $\ell_{\mathbf{y}}$ are both closed, Theorem A.5 again implies this is in turn equivalent to $\ell_{\mathbf{y}}(\mathbf{u} + s\boldsymbol{\phi}_0, \mathbf{v})$ being constant for all $\mathbf{u}$ and $\mathbf{v}$ and to $(\boldsymbol{\phi}_0, \mathbf{0})$ being a direction of constancy of $\ell_{\mathbf{y}}$.

By Proposition A.4, $g_{\mathbf{y}}0^+$ is given by $\ell_{\mathbf{y}}0^+ + p_\lambda 0^+$. It then follows that for a nonzero vector $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$, $g_{\mathbf{y}}0^+(\boldsymbol{\phi}) \leqslant 0$ if and only if $\boldsymbol{\phi}_2 = 0$ and $\ell_{\mathbf{y}}0^+(\boldsymbol{\phi}_1, 0) = 0$, the latter being equivalent to $\ell_{\mathbf{y}}^0 0^+(\boldsymbol{\phi}_1) = 0$ by the first assertion of the lemma.

The proof concerning directions of recession is analogous. ∎

Proof [Theorem 5.3] We let $g_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + p_\lambda(\mathbf{0}, \boldsymbol{\beta})$. Since $g_{\mathbf{y}}$ and $\ell_{\mathbf{y}}^0$ are proper closed convex functions by Lemma 5.2 and Proposition A.1, (c) and (d) imply (a) and (b) respectively from Theorem A.6.

By the polarity relationship of tangent and normal cones (see Appendix A.4), (e) and (f) are equivalent.

Assertion (d) is equivalent to (c) by the second assertion of Lemma 5.4.

Assume there is a direction of recession of $\ell_{\mathbf{y}}^0$ which is not a direction of constancy. Then, from Lemma 5.4, $\ell_{\mathbf{y}}$ admits a direction of recession of the form $(\boldsymbol{\phi}_0, \mathbf{0})$ which is not a direction of constancy. Hence, for all $(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ such that $X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta} \in \Theta$, $\ell_{\mathbf{y}}((\boldsymbol{\beta}_0, \boldsymbol{\beta}) + s(\boldsymbol{\phi}_0, \mathbf{0}))$ is nonincreasing as a function of s by Theorem A.5, and strictly convex on the interval where it is finite from Theorem 1 in Geyer [2009]. A nonincreasing strictly convex function is strictly decreasing where it is finite. Hence, $\ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) > \ell_{\mathbf{y}}(\boldsymbol{\beta}_0 + s\boldsymbol{\phi}_0, \boldsymbol{\beta})$ and $\ell_{\mathbf{y}}(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + p_\lambda(\mathbf{0}, \boldsymbol{\beta}) > \ell_{\mathbf{y}}(\boldsymbol{\beta}_0 + s\boldsymbol{\phi}_0, \boldsymbol{\beta}) + p_\lambda(\mathbf{0}, \boldsymbol{\beta})$ for $s > 0$, for all $(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ such that $X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta} \in \Theta$ and the minimum set of both $\ell_{\mathbf{y}}$ and $g_{\mathbf{y}}$ is empty. Moreover, $\ell_{\mathbf{y}}^0(\boldsymbol{\beta}_0) > \ell_{\mathbf{y}}^0(\boldsymbol{\beta}_0 + s\boldsymbol{\phi}_0)$, for all $\boldsymbol{\beta}_0$ such that $X_0\boldsymbol{\beta}_0 \in \Theta$ which implies the minimum set of $\ell_{\mathbf{y}}^0$ is empty.

Assertion (d) is equivalent by Lemma 5.4 to the assertion: every direction of recession of $\ell_{\mathbf{y}}$ of the form $(\boldsymbol{\phi}_0, \mathbf{0})$ is a direction of constancy. Moreover, it is easy to show that $(\boldsymbol{\phi}_0, \mathbf{0}) \in N_C((X_0^t\mathbf{y}, X^t\mathbf{y}))$ if and only if $\boldsymbol{\phi}_0 \in N_{\pi_1(C)}(X_0^t\mathbf{y})$. The equivalence between (d) and (e) then follows from (g) of Theorem 1 and (e) of Theorem 3 in Geyer [2009]. ∎

5.3 Checking Existence for Poisson and Bernoulli

In this section, the problem of determining whether $T_{\pi_1(C)}(X_0^t\mathbf{y})$ is a vector subspace when the response has a normal or Poisson distribution is considered. Recall from Theorem 5.3 that a certain regularized GLM estimator exists if and only if

$T_{\pi_1(C)}(X_0^t\mathbf{y})$ is a vector subspace with C denoting the smallest closed convex set containing $(X_0^t\mathbf{Y}, X^t\mathbf{Y})$ with probability one. We start by computing $\pi_1(C)$.

Lemma 5.5 *Let D be the smallest closed convex set containing $\mathbf{Y}$ with probability one and let $A \in \mathbb{R}^{M\times N}$. If $\ker A = \{\mathbf{0}\}$ or if $\mathbf{Y}$ is discrete with D polyhedral, AD is the smallest closed convex set containing $A\mathbf{Y}$ with probability one.*

Proof Assume $\ker A = \{\mathbf{0}\}$. The fact that AD is convex follows from the fact that the image by a linear application of a convex set is convex. It is closed as the image of a closed set by an injective linear application and $P(A\mathbf{Y} \in AD) = 1$ since D contains $\mathbf{Y}$ with probability one. Let us assume that there exists a set $E \subset AD$ closed, convex and containing $A\mathbf{Y}$ with probability one. Then, its preimage by A would be closed as the preimage of a closed set by a linear application, convex as the preimage of a convex set and containing $\mathbf{Y}$ with probability one. Moreover, it would be strictly included in D. A contradiction.

Assume $\mathbf{Y}$ is a discrete random variable with D polyhedral. Let F denote the smallest set containing $\mathbf{Y}$ with probability one. It is countable as $\mathbf{Y}$ is a discrete random variable. Moreover, the smallest set containing $A\mathbf{Y}$ with probability one is AF. Hence, $D = \overline{\text{co}}\, F$, the closed convex hull of F, and the smallest closed convex set containing $A\mathbf{Y}$ with probability one is given by $\overline{\text{co}}\, AF$. It remains to show that $\overline{\text{co}}\, AF = A\overline{\text{co}}\, F$. It follows from the fact that $\overline{\text{co}}\, AF = \overline{A\overline{\text{co}}\, F}$ which is closed as the image by a linear application of a polyhedral set. ■

In logistic and Poisson regression, the smallest closed convex set containing $\mathbf{Y}$ with probability one, denoted D, is given by $\overline{\text{co}}\,\{0,1\}^N = [0,1]^N$ and $\overline{\text{co}}\,\mathbb{N}^N = [0,\infty[^N$ respectively. Since both sets are polyhedral, Lemma 5.5 implies $\pi_1(C)$ is given by $X_0^t D$ which is again polyhedral. The following theorem shows that one can determine whether $T_{\pi_1(C)}(X_0^t\mathbf{y})$ is a vector subspace by solving a linear programming problem. We first note that for a polyhedral set E, $T_E(\mathbf{u})$ is polyhedral for $\mathbf{u} \in E$ (Theorem 6.46 in Rockafellar and Wets [1998]). By Theorem A.3, this implies $T_E(\mathbf{u})$ is finitely generated by a certain set V, that is, $T_E(\mathbf{u}) = \text{cone}\, V$.

Theorem 5.6 (Theorem 7 and 10 in [Geyer, 2009]) *Assume E is polyhedral so that $T_E(\mathbf{u}) = \text{cone}\, V$ for a certain set V. Then, $T_E(\mathbf{u})$ is a vector subspace if and only if the optimal value of the following linear programming problem is nonpositive for all $\mathbf{w} \in V$,*

$$\text{maximize: } \langle \mathbf{w}, \boldsymbol{\delta} \rangle \qquad \text{subject to: } \langle \mathbf{v}, \boldsymbol{\delta} \rangle \geqslant 0, \quad \forall \mathbf{v} \in V \backslash \{\mathbf{w}\}.$$

The R function `linearity` available from the `rcdd` package [Geyer and Meeden, 2008] solves the linear programming problem, so it remains to determine V such that $T_{\pi_1(C)}(X_0^t\mathbf{y}) = \text{cone}\, V$.

Let D denote the smallest closed convex set containing $\mathbf{Y}$ with probability one. First (see example 6.10 in Rockafellar and Wets [1998]), $T_D(\mathbf{y}) = \Pi_{i=1}^N F_i$, where

$$F_i = \begin{cases} [0, \infty) & \text{if } y_i = 0 \\ (-\infty, 0] & \text{if } y_i = 1 \end{cases}, \quad \text{for logistic regression,}$$

and

$$F_i = \begin{cases} [0, \infty) & \text{if } y_i = 0 \\ (-\infty, \infty) & \text{if } y_i \in \mathbb{N} \backslash \{0\} \end{cases}, \quad \text{for Poisson regression.}$$

It then follows that $T_D(\mathbf{y}) = \text{cone} \bigcup_{i=1}^N E_i$, where

$$E_i = \begin{cases} \{\mathbf{e}_i\} & \text{if } y_i = 0 \\ \{-\mathbf{e}_i\} & \text{if } y_i = 1 \end{cases}, \quad \text{for logistic regression,}$$

and

$$E_i = \begin{cases} \{\mathbf{e}_i\} & \text{if } y_i = 0, \\ \{\mathbf{e}_i, -\mathbf{e}_i\} & \text{if } y_i \in \mathbb{N} \backslash \{0\} \end{cases}, \quad \text{for Poisson regression,}$$

with $\mathbf{e}_i$ denoting the ith canonical basis vector, $i = 1, \ldots, N$.

Moreover, it can be shown that for a polyhedral set E, if $T_E(\mathbf{u})$ is finitely generated by a set F, then $T_{AE}(A\mathbf{u})$ is finitely generated by AF, where A denotes a linear transform (Geyer [2009] and Theorems 6.43 and 6.46 in Rockafellar and Wets [1998]). Finally, $T_{\pi_1(C)}(X_0^t \mathbf{y}) = T_{X_0^t D}(X_0^t \mathbf{y}) = \text{cone} \bigcup_{i=1}^N X_0^t E_i$. Equivalently, $T_{\pi_1(C)}(X_0^t \mathbf{y}) = \text{cone} \bigcup_{i=1}^N G_i$, where

$$G_i = \begin{cases} \{\mathbf{x}_0{}^i\} & \text{if } y_i = 0 \\ \{-\mathbf{x}_0{}^i\} & \text{if } y_i = 1 \end{cases}, \quad \text{for logistic regression,}$$

and

$$G_i = \begin{cases} \{\mathbf{x}_0{}^i\} & \text{if } y_i = 0 \\ \{\mathbf{x}_0{}^i, -\mathbf{x}_0{}^i\} & \text{if } y_i \in \mathbb{N} \backslash \{0\} \end{cases}, \quad \text{for Poisson regression,}$$

with $\mathbf{x}_0{}^i$ denoting the ith row vector of X_0, $i = 1, \ldots, N$. Hence, we have provided a numerical procedure to check for existence of penalized logistic and Poisson regression estimators.

Appendix A

Convex Analysis

A.1 Convex Sets and Convex Functions

A subset C of $\mathbb{R}^N$ is said to be *convex* if $(1-\lambda)\mathbf{x}+\lambda\mathbf{y} \in C$ whenever $\mathbf{x}$, $\mathbf{y} \in C$ and $0 < \lambda < 1$.

The *convex hull* of a nonempty set $S \subset \mathbb{R}^N$, denoted $\operatorname{co} S$, is the smallest convex set that includes S. It can equivalently be defined as the intersection of all convex sets which contain S. It consists of all the convex combinations of elements of S,

$$\operatorname{co} S = \{\, \textstyle\sum_{i=0}^{P} \lambda_i \mathbf{x}_i \mid \mathbf{x}_i \in S,\ \lambda_i \geqslant 0,\ \sum_{i=0}^{P} \lambda_i = 1,\ P \geqslant 0 \,\}.$$

The *closed convex hull* of S, denoted $\overline{\operatorname{co}}\, S$, is the smallest closed convex set that includes S. It is the closure of $\operatorname{co} S$, that is, $\overline{\operatorname{co}}\, S = \overline{\operatorname{co} S}$.

A function f from a set S to $\mathbb{R} \cup \{-\infty, +\infty\}$ is defined to be *convex* if its *epigraph*

$$\operatorname{epi} f = \{\, (\mathbf{x}, \mu) \mid \mathbf{x} \in S,\ \mu \in \mathbb{R},\ \mu \geqslant f(\mathbf{x}) \,\}$$

is a convex subset of $S \times \mathbb{R}$. A convex function f on S can always be extended to a convex function on all of $\mathbb{R}^N$ by setting $f(\mathbf{x}) = +\infty$ for $\mathbf{x} \notin S$.

For a function f from a convex set C to $(-\infty, \infty]$, convexity is equivalent to

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leqslant \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \quad 0 < \lambda < 1, \tag{A.1}$$

for any $\mathbf{x}$, $\mathbf{y} \in C$. A real-valued function on a convex set C is said to be *strictly convex* if (A.1) is satisfied with strict inequality whenever $\mathbf{x} \neq \mathbf{y}$.

A convex function is said to be *proper* if it is not identically $+\infty$ and $f(\mathbf{x}) > -\infty$ for every $\mathbf{x}$.

For any convex function f on S, the set $\operatorname{dom} f = \{\, \mathbf{x} \mid f(\mathbf{x}) < +\infty \,\}$ is called the *effective domain* of f.

A.2 Subgradients

A vector $\mathbf{v} \in \mathbb{R}^N$ is said to be a *subgradient* of a convex function f at a point $\mathbf{x} \in \mathbb{R}^N$ if

$$f(\mathbf{z}) \geqslant f(\mathbf{x}) + \mathbf{v}^t(\mathbf{z} - \mathbf{x}), \quad \forall \mathbf{z} \in \mathbb{R}^N.$$

The set of all subgradients of f at $\mathbf{x}$ is called the *subdifferential* of f at $\mathbf{x}$ and is denoted by $\partial f(\mathbf{x})$. If $\partial f(\mathbf{x})$ is nonempty, f is said to be *subdifferentiable* at $\mathbf{x}$.

Let f be a convex function and $\mathbf{x}$ a point where f is finite. If f is differentiable at $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$. Conversely, if f has a unique subgradient at $\mathbf{x}$, then f is differentiable at $\mathbf{x}$.

For $\lambda \geqslant 0$, $\partial(\lambda f)(\mathbf{x}) = \lambda \partial f(\mathbf{x})$. If A is a linear transformation and $h(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$ with f convex finite, then $\partial h(\mathbf{x}) = A^t \partial f(A\mathbf{x} + \mathbf{b})$. Moreover, if f_i is convex finite for $i = 1, \ldots, M$ and and $f = \sum_{i=1}^M f_i$, then $\partial f(\mathbf{x}) = \sum_{i=1}^M \partial f_i(\mathbf{x})$.

A necessary and sufficient condition for a given point $\mathbf{x}$ to belong to the minimum set of a convex function f is that $\mathbf{0} \in \partial f(\mathbf{x})$, i.e. that $\mathbf{v} = \mathbf{0}$ be a subgradient of f at $\mathbf{x}$.

A.3 Lower Semicontinuity and Closedness

An extended real-valued function f given on a set $S \subset \mathbb{R}^N$ is said to be *lower semicontinuous* at a point $\mathbf{x}$ of S if

$$f(\mathbf{x}) = \liminf_{\mathbf{y} \to \mathbf{x}} f(\mathbf{y}).$$

The *closure* of a convex function which nowhere has the value $-\infty$, denoted $\operatorname{cl} f$, is the convex function such that $\operatorname{epi} \operatorname{cl} f = \overline{\operatorname{epi} f}$.

A convex function is said to be *closed* if $\operatorname{cl} f = f$. For a proper convex function, closedness is equivalent to lower semicontinuity.

The following propositions show that convexity and closedness are preserved under composition with a linear application and under positive linear combination.

Proposition A.1 *Let $f : \mathbb{R}^M \mapsto (-\infty, \infty]$, let $A \in \mathbb{R}^{M \times N}$, and let $h : \mathbb{R}^N \mapsto (-\infty, \infty]$ be defined by $h(\mathbf{x}) = f(A\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^N$. If f is convex, then h is also convex, while if f is closed, then h is also closed.*

Proposition A.2 *Let $f_i : \mathbb{R}^N \mapsto (-\infty, \infty]$, $i = 1, \ldots, M$, let $\gamma_1, \ldots, \gamma_M$ be positive scalars, and let $f : \mathbb{R}^N \mapsto (-\infty, \infty]$ be defined by $f(\mathbf{x}) = \sum_{i=1}^M \gamma_i f_i(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^N$. If $f_1, \ldots, f_M$ are convex, then f is also convex, while if $f_1, \ldots, f_M$ are closed, then f is also closed.*

A.4 Cones and Polyhedral Sets

A set $C \subset \mathbb{R}^N$ is said to be *polyhedral* if there exists $A \in \mathbb{R}^{M\times N}$ and $\mathbf{b} \in \mathbb{R}^M$ such that $C = \{\, \mathbf{x} \mid A\mathbf{x} \leqslant \mathbf{b} \,\}$.

We call $K \subset \mathbb{R}^N$ a *cone* if $\mathbf{0} \in K$ and $\lambda\mathbf{x} \in K$ for all $\mathbf{x} \in K$ and $\lambda > 0$.

The smallest convex cone containing $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ is the set

$$\operatorname{cone}\{\mathbf{x}_1, \ldots, \mathbf{x}_M\} = \{\, \textstyle\sum_{i=1}^M \lambda_i \mathbf{x}_i \mid \lambda_i \geqslant 0,\ i = 1, \ldots, M \,\}.$$

Such a cone is called *finitely generated*. The following theorem characterizes these cones.

Theorem A.3 (Farkas-Minkowski-Weyl theorem) *A convex cone is polyhedral if and only if it is finitely generated.*

The *tangent cone* of a convex set C at a point $\mathbf{x} \in C$ is

$$T_C(\mathbf{x}) = \overline{\{\, s(\mathbf{w} - \mathbf{x}) \mid \mathbf{w} \in C \text{ and } s \geqslant 0 \,\}}.$$

The *normal cone* of a convex set C in $\mathbb{R}^N$ at a point $\mathbf{x} \in C$ is

$$N_C(\mathbf{x}) = \{\, \boldsymbol{\delta} \in \mathbb{R}^N \mid \langle \mathbf{w} - \mathbf{x}, \boldsymbol{\delta} \rangle \leqslant 0 \text{ for all } \mathbf{w} \in C \,\}.$$

Tangent and normal cones are polars of each other, that is,

$$N_C(\mathbf{x}) = \{\, \mathbf{w} \in \mathbb{R}^N \mid \langle \mathbf{w},\, \mathbf{v} \rangle \leqslant 0 \text{ for all } \mathbf{v} \in T_C(\mathbf{x}) \,\},$$

and

$$T_C(\mathbf{x}) = \{\, \mathbf{v} \in \mathbb{R}^N \mid \langle \mathbf{w},\, \mathbf{v} \rangle \leqslant 0 \text{ for all } \mathbf{w} \in N_C(\mathbf{x}) \,\}.$$

A.5 Recession Cones and Recession Functions

The *recession cone* of a nonempty convex set C in $\mathbb{R}^N$, denoted by 0^+C is the set of vectors $\mathbf{y}$ such that $C + \mathbf{y} \subseteq C$.

Let f be a convex function on $\mathbb{R}^N$ which is not identically $+\infty$. The function whose epigraph is given by $0^+(\operatorname{epi} f)$ is called the *recession function* of f and we denote it by $f0^+$. By definition, we have $\operatorname{epi}(f0^+) = 0^+(\operatorname{epi} f)$. For a closed proper convex function f, the recession function takes the form

$$f0^+(\mathbf{x}) = \lim_{\lambda \downarrow 0} \lambda f(\mathbf{x}/\lambda) \tag{A.2}$$

for every $\mathbf{x} \in \operatorname{dom} f$. Moreover, if $\mathbf{0} \in \operatorname{dom} f$, this formula actually holds for every $\mathbf{x} \in \mathbb{R}^N$.

The following proposition allows to compute the recession function of the sum of functions.

Proposition A.4 *Let $f_i : \mathbb{R}^N \to (-\infty, \infty]$, $i = 1, \ldots, M$, be closed proper convex such that $f = f_1 + \ldots + f_M$ is proper. Then*

$$f0^+(\mathbf{x}) = f_1 0^+(\mathbf{x}) + \ldots + f_M 0^+(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^N.$$

A.6 Directions of Recession and Constancy

The *directions of recession* of a convex function f are the nonzero vectors $\mathbf{y}$ such that $f0^+(\mathbf{y}) \leqslant 0$, and the *directions of constancy* are the nonzero vectors $\mathbf{y}$ such that $f0^+(\mathbf{y}) \leqslant 0$ and $f0^+(-\mathbf{y}) \leqslant 0$. These directions have the following two important properties.

Theorem A.5 *For a proper convex function f, a nonzero vector $\mathbf{y}$ is a direction of recession (respectively constancy) of f if and only if the function*

$$s \mapsto f(\mathbf{x} + s\mathbf{y})$$

is nonincreasing (respectively constant) for every $\mathbf{x}$. When f is closed, this holds for every $\mathbf{x}$ if it holds for even one $\mathbf{x} \in \operatorname{dom} f$.

Theorem A.6 *Let f be a closed proper convex function. If every direction of recession of f is a direction of constancy, the infimum of f is attained.*

Appendix B

Supplementary Material to Section 3.4

B.1 Sensitivity Study

As noted in Section 3.4, the null-thresholding statistic and therefore the quantile universal threshold are functions of the unknown intercept $\boldsymbol{\beta}_0^*$ in logistic and Poisson lasso. In Figure B.1, we empirically investigate the sensitivity of our method to the estimation of $\boldsymbol{\beta}_0^* = 1$ on the Poisson distributed data of Section 3.4.2. On the top left panel, estimation of $\boldsymbol{\beta}_0^*$ (dark grey) described in Section 3.4 has low bias. Moreover we observe the relative median insensitivity of λ^{QUT}, TPr and FDr to the estimate.

B.2 Variance Estimation in Linear Models

When $P > N$ in the linear model (2.1), constructing a reliable estimator for σ^2 is a challenging task and several estimators have been proposed. Reid et al. [2014] consider an estimator of the form

$$\hat{\sigma}^2_{\mathrm{CV}} = \frac{1}{N - \hat{s}_\lambda} \|\mathbf{Y} - X\hat{\boldsymbol{\beta}}_\lambda\|_2^2, \tag{B.1}$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the lasso estimator tuned with cross-validation and $\hat{s}_\lambda$ denotes the number of estimated nonzero entries. Fan et al. [2012] propose refitted cross-validation (RCV). The data set is split into two equal parts, $(X^{(1)}, \mathbf{Y}^{(1)})$ and $(X^{(2)}, \mathbf{Y}^{(2)})$. On each part, a model selection procedure, the lasso tuned with cross-validation, for example, is applied resulting in two different sets of nonzero indices $\hat{\mathcal{S}}_1$, $\hat{\mathcal{S}}_2$ with respective cardinality $\hat{s}^{(1)}$ and $\hat{s}^{(2)}$. This allows to compute

$$\hat{\sigma}_1^2 = \frac{1}{N/2 - \hat{s}^{(1)}} \|(I - P_{X^{(2)}_{\hat{\mathcal{S}}_1}})\mathbf{Y}^{(2)}\|_2^2 \quad \text{and} \quad \hat{\sigma}_2^2 = \frac{1}{N/2 - \hat{s}^{(2)}} \|(I - P_{X^{(1)}_{\hat{\mathcal{S}}_2}})\mathbf{Y}^{(1)}\|_2^2,$$

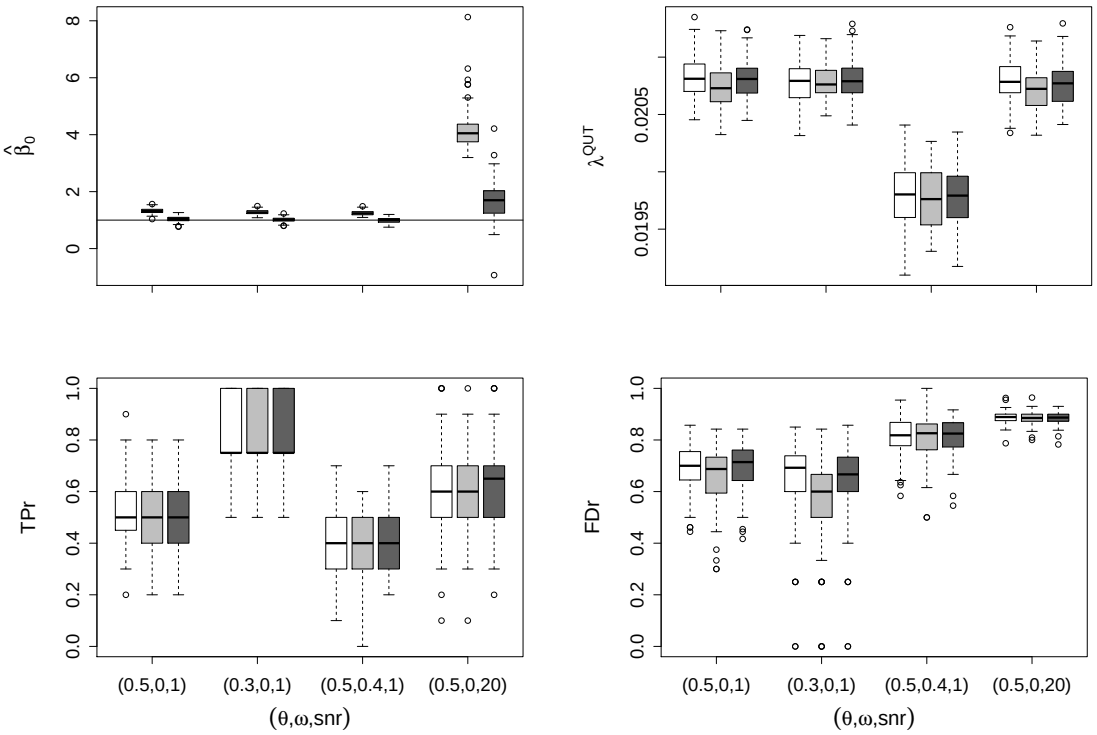

Figure B.1 – Estimation of $\beta_0^* = 1$ (top left panel) and its effect on λ^{QUT} (top right panel), TPr (lower left panel) and FDr (lower right panel). White, light grey and dark grey boxplots correspond respectively to the oracle estimator $\hat{\beta}_0 = 1$, initial step and final step of our estimation procedure.

where $P_{X^{(i)}_{\hat{\mathcal{S}}_j}}$ is the orthogonal projection matrix onto the range of the submatrix of $X^{(i)}$ with columns indexed by $\hat{\mathcal{S}}_j$. Finally, the RCV estimator is defined as

$$\hat{\sigma}^2_{\mathrm{RCV}} = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}. \tag{B.2}$$

We propose a new estimator of σ^2, refitted QUT, defined as

$$\hat{\sigma}^2_{\mathrm{RQUT}} = \underset{\sigma^2>0}{\operatorname{argmin}} \left|\sigma^2 - \hat{\sigma}^2_{\mathrm{RCV}}(\sigma^2)\right|, \tag{B.3}$$

where $\hat{\sigma}^2_{\mathrm{RCV}}(\sigma^2)$ is the RCV estimate with the lasso tuned with $\lambda^{\mathrm{QUT}}(\sigma^2)$. Figure B.2 shows boxplots of the three estimators of variance applied to the Gaussian data of Section 3.4.1. Refitted QUT has smallest variability and a median comparable to RCV.

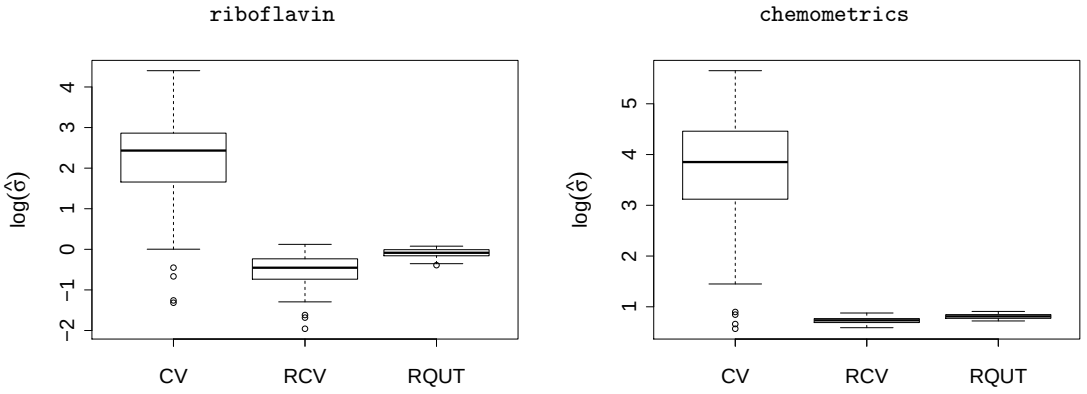

Figure B.2 – Results of a Monte Carlo simulation based on `riboflavin` and `chemometrics` data of Section 3.4.1 for the estimation of σ with cross-validation (CV) defined in (B.1), refitted cross-validation (RCV) defined in (B.2) and refitted QUT (RQUT) defined in (B.3).

Bibliography

Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

Belloni, A., Chernozhukov, V., and Wang, L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Breheny, P. and Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187, 2015.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and Regression Trees*. Wadsworth and Brooks/Cole Advanced Books & Software, Monterey, CA, 1984.

Bühlmann, P. and van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg, 2011.

Bühlmann, P., Kalisch, M., and Meier, L. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

Bunea, F., Lederer, J., and She, Y. The group square-root lasso: theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2): 1313–1325, 2014.

Cai, J.-F., Candès, E. J., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Candès, E. and Romberg, J. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

Candès, E. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Candès, E. J., Sing-Long, C. A., and Trzasko, J. D. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19):4643–4657, 2013.

Chen, J. and Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Donoho, D. L. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis*, 2(2):101–126, 1995.

Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory*, 52 (4):1289–1306, 2006.

Donoho, D. L. and Gavish, M. Minimax risk of matrix denoising by singular value thresholding. *Annals of Statistics*, 42(6):2413–2440, 2014.

Donoho, D. L. and Johnstone, I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

Donoho, D. L. and Tanner, J. Precise undersampling theorems. *Proceedings of the IEEE*, 98(6):913–924, 2010.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B*, 57(2):301–369, 1995.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996.

Fan, J. and Peng, H. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

Fan, J., Guo, S., and Hao, N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B*, 74(1):37–65, 2012.

Fan, Y. and Tang, C. Y. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B*, 75(3):531–552, 2013.

Fisher, R. A. *Statistical Methods for Research Workers*. Hafner Publishing Co., New York, 14th edition, 1973.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

Fu, W. J. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

Gavish, M. and Donoho, D. L. The optimal hard threshold for singular values is 4/sqrt(3). *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

Gavish, M. and Donoho, D. L. Optimal shrinkage of singular values. arXiv:1405.7511v3, 2016.

Geyer, C. J. Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, 3:259–289, 2009.

Geyer, C. J. and Meeden, G. D. R package rcdd, version 1.1. 2008.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

Hoerl, A. E. and Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Hsu, D., Kakade, S. M., and Zhang, T. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52): 1–6, 2012.

James, W. and Stein, C. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, California, 1961. University of California Press.

Josse, J. and Husson, F. Selecting the number of components in PCA using cross-validation approximations. *Computational Statististics and Data Analysis*, 56(6): 1869–1879, 2012.

Josse, J. and Sardy, S. Adaptive shrinkage of singular values. *Statistics and Computing*, 26(3):715–724, 2016.

Kushmerick, N. Learning to remove internet advertisements. In *Proceedings of the third international conference on Autonomous Agents*, pages 175–181. ACM, 1999.

Leng, C., Lin, Y., and Wahba, G. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

Meier, L., van de Geer, S., and Bühlmann, P. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70(1):53–71, 2008.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

Meinshausen, N. and Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010.

Mukherjee, A., Chen, K., Wang, N., and Zhu, J. On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 102(2):457–477, 2015.

Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384, 1972.

Neto, D., Sardy, S., and Tseng, P. ℓ_1-penalized likelihood smoothing and segmentation of volatility processes allowing for abrupt changes. *Journal of Computational and Graphical Statistics*, 21(1):217–233, 2012.

Owen, A. B. and Perry, P. O. Bi-cross-validation of the svd and the nonnegative matrix factorization. *Annals of Applied Statistics*, 3(2):564–594, 2009.

Park, M. Y. and Hastie, T. L_1-regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B*, 69(4):659–677, 2007.

Reid, S., Tibshirani, R., and Friedman, J. A study of error variance estimation in lasso regression. arXiv:1311.5274v2, 2014.

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, 1970.

Rockafellar, R. T. and Wets, R. J.-B. *Variational Analysis*. Springer-Verlag, Berlin, 1998.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

Sardy, S. On the practice of rescaling covariates. *International Statistical Review*, 76(2):285–297, 2008.

Sardy, S. and Tseng, P. On the statistical analysis of smoothing by maximizing dirty markov random field posterior distributions. *Journal of the American Statistical Association*, 99(465):191–204, 2004.

Sardy, S. and Tseng, P. Density estimation by total variation penalized likelihood driven by the sparsity ℓ_1 information criterion. *Scandinavian Journal of Statistics*, 37(2):321–337, 2010.

Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

Seber, G. A. F. and Lee, A. J. *Linear Regression Analysis*. Wiley, New York, 2nd edition, 2003.

Shabalin, A. A. and Nobel, A. B. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013.

Shorack, G. and Wellner, J. *Empirical Processes With Applications to Statistics*. Wiley, New York, 1986.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, 67(1):91–108, 2005.

Tibshirani, R. J. and Taylor, J. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

Tibshirani, R. J. and Taylor, J. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

Tikhonov, A. N. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4(4):1035–1038, 1963.

Van de Geer, S. The deterministic lasso. In *JSM proceedings*. American Statistical Association, 2007.

Wahba, G. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

Wang, H., Li, G., and Jiang, G. Robust regression shrinkage and consistent variable selection through the LAD-lasso. *Journal of Business & Economic Statistics*, 25 (3):347–355, 2007.

Williams, D. *Probability with Martingales*. Cambridge University Press, Cambridge, 1991.

Yang, Y. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

Zou, H., Hastie, T., and Tibshirani, R. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.