Quantile universal threshold for model selection

Caroline Giacobino* Sylvain Sardy†

Jairo Diaz Rodriguez ‡ Nick Hengartner §

In various settings such as regression, wavelet denoising and low rank matrix estimation, statisticians seek to find a low dimensional structure based on high dimensional data. In those instances, the maximum likelihood is unstable or might not be unique. By introducing some constraint on the parameters, a class of regularized estimators such as subset selection and the lasso allow to perform variable selection and parameter estimation by thresholding some parameters to zero. Yet one crucial step challenges all these estimators: the selection of the threshold parameter. If too large, important features are missing; if too small, false features are included.

After defining a zero-thresholding function, we present a unified framework to select features by identifying a threshold λ at the detection edge under the null model. We apply our methodology to existing estimators, with a particular emphasis on ℓ_1-regularized generalized linear models with the lasso as a special case. Numerical results show the effectiveness of our approach in terms of model selection and prediction.

Keywords: Convex optimization, high dimension, variable screening, zero-thresholding function.

*Department of Mathematics, University of Geneva; caroline.giacobino@unige.ch

†Department of Mathematics, University of Geneva; sylvain.sardy@unige.ch

‡Department of Mathematics, University of Geneva; jairo.diaz@unige.ch

§Theoretical Biology and Biophysics group, Los Alamos National Laboratory; nickh@lanl.gov

1 Introduction

R. A. Fisher was a strong advocate of maximum likelihood estimation (MLE) as a general principle for parameter estimation of models based on data. Modeling of genomic data, finance data, image classification, and more generally of data mining, have identified many real world examples in which the number of parameters P of the models can be dramatically larger than the sample size N. In those instances, MLE principle fails. Beyond existence and uniqueness issues, the MLE tends to perform poorly when P is large relative to N because of 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 sixty years on parameter estimation using regularization techniques. In both parametric and nonparametric models, regularization puts some reasonable prior or constraints on the parameters to better control the variance of the estimate and the complexity of the fitted model, at the price of a bias increase. Variance and model complexity are governed by a so-called regularization parameter.

We consider a class of regularization techniques, called *thresholding*, which:

1. assumes a *sparse* model; that is, a large majority but an unknown number of parameters are zero;

2. estimates some, if not all, parameters as zero.

Subset selection and more recent thresholding techniques are employed in various settings including density estimation [Donoho et al., 1996, Sardy and Tseng, 2010], linear inverse problems [Donoho, 1995], compressed sensing [Donoho, 2006, Candès and Romberg, 2006], time series [Neto et al., 2012] and generalized linear models (GLMs) [Park and Hastie, 2007], which we investigate in Section 3. Examples of thresholded parameters include coefficients of a linear model, singular values of a low-rank matrix and jumps in a function.

Ideally, thresholding estimators should identify the nonzero parameters, or at least the most salient ones. The *threshold* λ determines which parameters are set to zero.

1.1 Examples

We illustrate sparse models and thresholding estimators in linear regression and low-rank matrix estimation.

Linear regression. Many thresholding estimators have emerged in this setting. Consider the linear model

$$\mathbf{Y} = X\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}, \tag{1}$$

where X is an $N \times P$ matrix of covariates or discretized basis functions, $\boldsymbol{\beta}^0$ the vector of unknown regression coefficients and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I_N)$. Letting B be a $Q \times P$ matrix and $\boldsymbol{\xi}^0 = B\boldsymbol{\beta}^0$, thresholding assumes the underlying model is sparse in the sense that

$$\mathcal{S}^0 = \{q \in \{1, \ldots, Q\} : \ \xi_q^0 \neq 0\} \tag{2}$$

has a small cardinality $s^0 := |\mathcal{S}^0|$. The challenge lies in identifying the active subset $\mathcal{S}^0$. In gene expression microarrays, this could amount to selecting genes responsible for a cancer type. We now list estimators which can determine $\mathcal{S}^0$ with some success. Consider in particular a class of estimators of $\boldsymbol{\beta}^0$ that are defined by

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}\in\mathbb{R}^P}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|B\boldsymbol{\beta}\|_\nu^\eta, \quad \nu \geq 0,\ \eta \geq 0,\ \lambda \geq 0, \tag{3}$$

where $\|\boldsymbol{\xi}\|_\nu = (\sum_{q=1,\ldots,Q} |\xi_q|^\nu)^{1/\nu}$ with the convention that $0^0 = 0$. Ridge regression [Hoerl and Kennard, 1970] and smoothing splines [Wahba, 1990] are of the form (3) with $\nu = \eta = 2$, but do not set any entry of $B\hat{\boldsymbol{\beta}}_\lambda$ to zero for a finite λ. On the other hand, many estimators enjoy the thresholding property of setting some or all entries of $B\hat{\boldsymbol{\beta}}_\lambda$ to zero:

- Best subset selection ($\nu = \eta = 0$, $B = I$), which is a discrete optimization problem.
- Lasso-type estimators: Lasso [Tibshirani, 1996] ($\nu = \eta = 1$, $B = I$), the closest convexification to best subset selection, which includes soft-Waveshrink [Donoho and Johnstone, 1994]; Adaptive lasso [Zou, 2006] ($\nu = \eta = 1$, B a diagonal matrix of weights); Subbotin lasso [Sardy, 2009] ($\nu = \eta \leq 1$); Group lasso [Yuan and Lin, 2006] ($\nu = 2$ and $\eta = 1$); Generalized lasso [Tibshirani and Taylor, 2011] ($\nu = \eta = 1$) from which total variation denoising [Rudin et al., 1992] is a special case employing the matrix of finite differences for B; LAD-lasso which substitutes the ℓ_2-loss with the ℓ_1-loss [Wang et al., 2007].
- SCAD [Fan and Peng, 2004] and the Dantzig selector [Candes and Tao, 2007], which cannot be expressed as (3).

Low-rank matrix estimation [Mazumder et al., 2010, Cai et al., 2010]. Consider the model $Y = X^0 + \sigma Z$, where Y, X^0 and Z are $N \times P$ matrices with $P \geq N$ and $Z_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$. One assumes X^0 has low rank and estimates it by setting $\hat{X} = U\mathrm{diag}(\hat{\boldsymbol{\sigma}}_\lambda)V^{\mathrm{T}}$ with $Y = U\mathrm{diag}(\boldsymbol{\sigma})V^{\mathrm{T}}$ the singular value decomposition of Y. The standard approach retains only the λ largest singular values. Inspired by lasso, a more recent estimate of X^0 is given by

$$\underset{X\in\mathbb{R}^{N\times P}}{\operatorname{argmin}} \frac{1}{2}\|Y - X\|_F^2 + \lambda\|X\|_*, \tag{4}$$

where $\|\cdot\|_F$ and $\|\cdot\|_*$ respectively denote the Frobenius and the trace norm. For a given $\lambda > 0$, this approach results in $\hat{\sigma}_{i,\lambda} = \max(\sigma_i - \lambda, 0)$.

1.2 The challenge: selection of the threshold λ

Thresholding estimators result in small or null cardinality of $\hat{\mathcal{S}}_\lambda = \{q \in \{1,\ldots,Q\} : \hat{\xi}_{\lambda,q} \neq 0\}$ for large enough λ. Examples of parameters ξ_q include regression coefficients

$B\boldsymbol{\beta}$ as in (2), singular values $\boldsymbol{\sigma}$ in low-rank matrix estimation and wavelet coefficients in nonparametric function estimation. Selection of the threshold λ is crucial to perform effective model selection, that is, effective recovery of the support $\mathcal{S}^0$ of the parameters. A too large λ results in a simplistic model missing important features whereas a too small λ leads to a model including many spurious features.

Classical methodologies to select λ consist in minimizing criteria such as cross-validation, AIC [Akaike, 1973], BIC [Schwarz, 1978] or Stein unbiased risk estimation (SURE) [Stein, 1981]. In low-rank matrix estimation, Owen and Perry [2009] and Josse and Husson [2012] employ cross validation whereas Candes et al. [2013] employ SURE. This methodology is also used in regression [Zou et al., 2007, Tibshirani and Taylor, 2012], and reduced rank regression [Mukherjee et al., 2015]. 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 proposed.

A desirable property is variable screening: it requires that with high probability

$$\hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^0 = \{q \in \{1, \dots, Q\} : \xi_q^0 \neq 0\}. \tag{5}$$

In lasso, this property holds for suitably chosen λ under assumptions on the design matrix and nonzero regression coefficients (see Section 2.1). The smaller the set $\hat{\mathcal{S}}_\lambda$ satisfying (5), the fewer false detections there will be.

We now pause to make the important remark that the optimal threshold for recovering the active set $\mathcal{S}^0$ often differs from the threshold chosen to minimize the prediction error. The conflict between model identification and prediction optimality has been observed by Yang [2005] in subset selection, and by Leng et al. [2006] and Meinshausen and Bühlmann [2006] with the lasso. It turns out that models selected for prediction are typically more complex than those selected to perform variable screening. Variable screening and rank estimation are not strongly dependent on the bias-variance trade-off, and can therefore be achieved by a larger threshold.

1.3 Our contribution

We propose a new threshold selection method that aims at a good identification of $\mathcal{S}^0$ with few false detections, and that follows the same paradigm whether in the setting of regression, low-rank matrix estimation, inverse problems or density estimation. We introduce the unified framework in Section 2, formally defining the zero-thresholding function in Section 2.2 and the quantile universal threshold in Section 2.3. We then present a case study in Section 3: the new threshold selection is applied to lasso in generalized linear models, with a particular attention in Section 3.3 to linear, Poisson and logistic regression. We illustrate the effectiveness of our methodology in Section 4 on simulated data and on four real data sets in regression and classification. The appendices contain the proofs and technical details.

2 A unified framework for threshold selection

2.1 Thresholding under the null

The idea of choosing a null model-based threshold has appeared in some instances. It has shown good empirical and theoretical properties. We illustrate with three settings.

Wavelet denoising. Donoho and Johnstone [1994] consider an orthonormal $P \times P$ wavelet matrix and select the threshold of Waveshrink as $\lambda_P^{\text{universal}} = \sigma\sqrt{2\log P}$. Let us define $\Lambda := \|X^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty$, and note that $\Lambda \stackrel{d}{=} \|X^{\mathrm{T}}\mathbf{Y}\|_\infty$ under the null model $\boldsymbol{\beta}^0 = \mathbf{0}$. Alongside oracle and minimax properties [Donoho et al., 1995], this choice leads to

$$\mathbb{P}(\hat{\boldsymbol{\beta}}_{\lambda_P^{\text{universal}}} = \mathbf{0}; \boldsymbol{\beta}^0 = \mathbf{0}) = \mathbb{P}(\Lambda \leq \lambda_P^{\text{universal}}) \stackrel{P\to\infty}{\longrightarrow} 1 \tag{6}$$

at the rate $O(1/\sqrt{\log P})$. This equality follows from the equivalence for all $\mathbf{y} \in \mathbb{R}^N$

$$\hat{\boldsymbol{\beta}}_\lambda = \mathbf{0} \quad \Leftrightarrow \quad \lambda \geq \|X^{\mathrm{T}}\mathbf{y}\|_\infty. \tag{7}$$

Linear regression with lasso. Assuming the design matrix satisfies a compatibility condition and the nonzero regression coefficients are sufficiently large [Bühlmann and van de Geer, 2011], the lasso with tuning parameter $\lambda_{N,P}$ satisfies the screening property (5) with probability at least as large as that of

$$\mathcal{T}_{\lambda_{N,P}} := \{\Lambda \leq \lambda_{N,P}\} \quad \text{with} \quad \Lambda = \|X^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty. \tag{8}$$

This event is equivalent to $\{\hat{\boldsymbol{\beta}}_{\lambda_{N,P}} = \mathbf{0}; \boldsymbol{\beta}^0 = \mathbf{0}\}$ since (7) holds for any X matrix. It is a direct consequence of the Karush-Kuhn-Tucker (KKT) conditions which have been derived for the lasso for instance by Osborne et al. [2000].

Low-rank matrix estimation. Under the null model $X^0 = 0_{N\times P}$, the empirical distribution of the singular values of the data with noise level $1/\sqrt{N}$ converges to a compactly supported distribution. Gavish and Donoho [2014] derive optimal singular value thresholding operators and set any singular value less than the upper bound of the support to zero.

The common approach of these methodologies is to select a threshold below which a statistic (which we call the zero-thresholding function in Section 2.2) lies with high probability if the null model is true.

2.2 The zero-thresholding function

The examples in the previous subsection point towards a general method for selecting the threshold. The unified framework which emerges leads us to the following definitions.

Definition 1. *Assume* $Y \sim f_{\boldsymbol{\xi}^0}$. *An estimator* $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y})$ *indexed by* $\lambda \geq 0$ *is called a* thresholding estimator *if*

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

Note that ridge regression is not a thresholding estimator since $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{y}) \neq \mathbf{0}$ for all finite λ and $\mathbf{y} \in \mathbb{R}^n$.

Definition 2. *A thresholding estimator* $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y})$ *admits a* zero-thresholding function $\lambda(\mathbf{Y})$ *if*

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

This equivalence implies equiprobability between setting all coefficients to zero and selecting the threshold large enough, as in (6). Quite surprisingly to us, best subset admits a zero-thresholding function, although it is defined as a solution to a non-convex optimization problem. A zero-thresholding function exists in the following cases:

- Low-rank matrix estimation (4): $\lambda(Y) = \|\boldsymbol{\sigma}\|_\infty$, the largest singular value of Y.
- Best subset selection selects the model complexity parameter as the smallest p minimizing

  $$C(p) = \frac{1}{2}\|\mathbf{y} - X\hat{\boldsymbol{\beta}}^p\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}^p\|_0^0,$$

  where $\hat{\boldsymbol{\beta}}^p$ is the best least squares vector with p nonzero entries. It follows that $\|\hat{\boldsymbol{\beta}}^p\|_0^0 = p$ and that

  $$C(p) = C(0) - \Delta_p(\mathbf{y}) + \lambda p \quad \text{with} \quad \Delta_p(\mathbf{y}) = \frac{1}{2} \max_{\{\mathcal{I} \subset \{1,\ldots,P\} : |\mathcal{I}| = p\}} \|P_{X_\mathcal{I}}\mathbf{y}\|_2^2,$$

  where P_X is the orthogonal projection matrix onto the range of X and $X_\mathcal{I}$ denotes the submatrix of X with columns indexed by $\mathcal{I}$. Hence, the zero-thresholding function for best subset selection is

  $$\lambda(\mathbf{y}) = \max_{p=1,\ldots,\operatorname{rank}(X)} \frac{\Delta_p(\mathbf{y})}{p}. \tag{9}$$

  Calculating $\lambda(\mathbf{y})$ is prohibitive when P is large. The expression simplifies to $\lambda(\mathbf{y}) = \Delta_1(\mathbf{y})$ when X has orthogonal columns since $\Delta_p(\mathbf{y}) \leq p\Delta_1(\mathbf{y})$.
- Lasso and the Dantzig selector: $\lambda(\mathbf{y}) = \|X^{\mathrm{T}}\mathbf{y}\|_\infty$.
- Adaptive lasso: $\lambda(\mathbf{y}) = \|B^{-1}X^{\mathrm{T}}\mathbf{y}\|_\infty$.
- Subbotin lasso: we conjecture that

  $$\lambda(\mathbf{y}) = \frac{2(1-\nu)}{(2-\nu)^2} \max_{p=1,\ldots,\operatorname{rank}(X)} \max_{\{\mathcal{I} \subset \{1,\ldots,P\} : |\mathcal{I}| = p\}} \frac{\|P_{X_\mathcal{I}}\mathbf{y}\|_2^2}{\|\hat{\boldsymbol{\beta}}_\mathcal{I}^{(p,\nu)}\|_\nu^\nu},$$

where $\hat{\boldsymbol{\beta}}_{\mathcal{I}}^{(p,\nu)} = \frac{2(1-\nu)}{2-\nu}(X_{\mathcal{I}}^{\mathrm{T}}X_{\mathcal{I}})^{-1}X_{\mathcal{I}}^{\mathrm{T}}\mathbf{y}$ is the Subbotin-lasso estimate based on $X_{\mathcal{I}}$ for any $\nu \in [0,1)$. This expression simplifies to (9) if $\nu = 0$, and to $\lambda(\mathbf{y}) = \{\|X^{\mathrm{T}}\mathbf{y}\|_\infty/(2-\nu)\}^{2-\nu}/\{2(1-\nu)\}^{\nu-1}$ if X is orthonormal.

- Group lasso: $\lambda(\mathbf{y}) = \|X^{\mathrm{T}}\mathbf{y}\|_2$ for one group. When the parameters are separated into G groups so that the penalty is $\lambda\sum_{g=1,\dots,G}\|\boldsymbol{\beta}_g\|_2$, the zero-thresholding function is $\lambda(\mathbf{y}) = \max_{g=1,\dots,G}\|X_g^{\mathrm{T}}\mathbf{y}\|_2$.

- Generalized lasso: assuming B has full row rank Q and denoting by $\mathcal{I}$ a set of column indices such that $B_{\mathcal{I}}$ is invertible and by $\mathcal{I}^c$ its complement,

$$\lambda(\mathbf{y}) = \|A_1^{\mathrm{T}}(I - P_{A_2})\mathbf{y}\|_\infty,$$

  where $A_1 = X_{\mathcal{I}}B_{\mathcal{I}}^{-1}$ and $A_2 = X_{\mathcal{I}^c} - X_{\mathcal{I}}B_{\mathcal{I}}^{-1}B_{\mathcal{I}^c}$. This follows from the fact that generalized lasso can be transformed into a lasso problem by a change of variable when B has full row rank. In total variation denoising, one gets $\lambda(\mathbf{y}) = \|(BB^{\mathrm{T}})^{-1}B\mathbf{y}\|_\infty$.

- LAD-lasso: $\lambda(\mathbf{y}) = \min_{\mathbf{v}\in\partial\|\mathbf{y}\|_1}\|X^{\mathrm{T}}\mathbf{v}\|_\infty$, where $\mathbf{v}$ is a subgradient of the ℓ_1-norm evaluated at $\mathbf{y}$.

Some of these zero-thresholding functions can be inferred from the KKT conditions [Rockafellar, 1970]. Interestingly, although the elastic net [Zou and Hastie, 2005] penalty $\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\beta}\|_2^2$ depends on two regularization parameters, the zero-thresholding function is $\lambda_1(\mathbf{y}) = \|X^{\mathrm{T}}\mathbf{y}\|_\infty$ regardless of the value of λ_2.

2.3 The quantile universal threshold

Given an estimator admits a zero-thresholding function, we propose the following unified framework to select a threshold, called *quantile universal threshold*.

Definition 3. *Assume $Y \sim f_{\boldsymbol{\xi}^0}$ and let $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y})$ be an estimator with associated zero-thresholding function $\lambda(\cdot)$. Let $\Lambda := \lambda(\mathbf{Y}_0)$, where $\mathbf{Y}_0 \sim f_{\boldsymbol{\xi}_0}$ with null model parameter $\boldsymbol{\xi}_0 = \mathbf{0}$. The quantile universal threshold λ^{QUT} is the upper $(1-\alpha)$ quantile of the thresholding statistic Λ, namely*

$$\lambda^{\mathrm{QUT}} := F_\Lambda^{-1}(1-\alpha).$$

We recommend $\alpha = O(1/\sqrt{\log P})$ by analogy to the rate of convergence of $F_\Lambda(\lambda_P^{\mathrm{universal}})$ to one in the context of orthonormal regression with the lasso (see Section 2.1) and allow $P = P_N$ to grow with N. In Section 4, we set $\alpha = 1/\sqrt{\pi\log P}$ by extension of $\bar{F}_\Lambda(\lambda_P^{\mathrm{universal}}) \sim 1/\sqrt{\pi\log P}$. In the following examples, a closed form expression of λ^{QUT} is derived.

- In best subset selection with orthonormal design matrix, $\lambda_P = \sigma^2 \log P$ satisfies $\bar{F}_\Lambda(\lambda_P) \sim 1/\sqrt{\pi \log P}$. This result can be inferred from the thresholding statistic $\Lambda = \lambda(\mathbf{Y_0}) =_d \|\mathbf{Z}\|_\infty^2/2$, where $Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$. This is larger than the traditional BIC penalty $\lambda^{\text{BIC}} = \sigma^2/2 \log P$ which is known to perform poorly in the high dimensional setting. Generalizations which have been proposed such as EBIC also select a larger tuning parameter.

- In total variation denoising, the thresholding statistic tends in distribution to the infinite norm of a Brownian bridge, leading to $\lambda_P = \sigma\sqrt{P \log\log P}/2$ for $\alpha_P = O(1/\sqrt{\log P})$ [Sardy and Tseng, 2004].

- In group lasso regression with orthonormal groups, each of size Q, extreme value theory leads to $\lambda_P = \sigma\sqrt{2\log P + (Q-1)\log\log P - 2\log\Gamma(Q/2)}$ for $\alpha_P = O(1/\sqrt{\log P})$ [Sardy, 2012].

If no closed form expression is known, a Monte-Carlo simulation allows to evaluate λ^{QUT} (e.g., see Algorithm 1). For low-rank matrix estimation, we refer the reader to Josse and Sardy [2015] who point to good estimation of the rank with QUT.

We now state two important properties satisfied by our methodology.

Property 1. *If the design matrix and regression coefficients satisfy the sufficient conditions for variable screening (see Bühlmann and van de Geer [2011]), lasso tuned with λ^{QUT} achieves variable screening with probability at least $1 - \alpha$.*

This is a consequence from the lasso property $\mathbb{P}(\mathcal{T}_{\lambda^{\text{QUT}}}) = 1 - \alpha$ with $\mathcal{T}_\lambda$ defined in (8).

Recall that when performing multiple hypotheses tests, the familywise error rate is defined as the probability of rejecting at least one true null hypothesis. In the context of variable selection, this is the probability of falsely selecting at least one variable. It can be shown that if the null model is true, the familywise error rate is equal to the false discovery rate defined in Section 4.1.

Property 2. *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.*

This follows immediately from α being equal to the familywise error rate under the assumption that the null model is true.

3 QUT for lasso-GLM: a case study

In this section, we derive an explicit formulation of the zero-thresholding function and hence of the quantile universal threshold for ℓ_1-regularized generalized linear models.

3.1 Background

Generalized linear models (GLMs) provide a framework to model the association between a dependent variable and a set of explanatory variables [Nelder and Wedderburn, 1972]. GLMs encompass Gaussian linear models, logistic regression for binary responses, Poisson regression for count data and log-linear models for contingency tables. It is assumed that each component $\mathbf{Y}_i$ of the response vector $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ is independent and has distribution in the exponential family with probability density (with respect to a given dominating measure ν) of the form

$$f_{\mathbf{Y}_i}(\mathbf{y}_i; \boldsymbol{\theta}_i) = \exp\left\{\mathbf{y}_i\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i) + c(\mathbf{y}_i)\right\}, \quad i = 1, \ldots, N,$$

where b and c are known functions, and the parameter set Θ^N is open with $\Theta := \{\theta \in \mathbb{R} \,|\, b(\theta) < \infty\}$. Remark that Θ is convex and that b is strictly convex on Θ unless the response vector is constant almost everywhere. We consider the canonical model which relates linearly the parameter to the explanatory variables by assuming $\boldsymbol{\theta}_i = \mathbf{a}_i^{\mathrm{T}}\boldsymbol{\alpha}^0 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0$. Here $\boldsymbol{\alpha}^0$ corresponds to the P_0 parameters we expect to be nonzero, as is the case for the intercept; these parameters will not be thresholded. In contrast, we assume the P entries of $\boldsymbol{\beta}^0$ might correspond to spurious covariates and allow them to be thresholded to zero. We denote by A and X the $N \times P_0$ and $N \times P$ matrices of covariates and by $f_{(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0)}$ the distribution of a response vector from a GLM model.

Park and Hastie [2007] define an extension of lasso to GLM by

$$(\hat{\boldsymbol{\alpha}}_\lambda(\mathbf{y}), \hat{\boldsymbol{\beta}}_\lambda(\mathbf{y})) \in \underset{(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{F}}{\operatorname{argmin}} -\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}) + \lambda\|\boldsymbol{\beta}\|_{\mathbf{1}}, \tag{10}$$

where

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{n=1}^{N} \left[y_n\left(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}\right) - b\left(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}\right)\right]$$

and $\mathcal{F} := \left\{(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{P_0+P} \,|\, A\boldsymbol{\alpha} + X\boldsymbol{\beta} \in \Theta^N\right\}$. Note that the element notation "$\in$" in (10) indicates the minimizer might not be unique. We set $\hat{\boldsymbol{\beta}}_\lambda = \mathbf{0}$ if $\lambda = +\infty$.

3.2 Zero-thresholding function and QUT for lasso-GLM

In the following Lemma some important properties of lasso-GLM solutions are derived. It is a slight modification of Lemma 1 appearing in Tibshirani [2013] taking into account $\boldsymbol{\alpha}$ is not penalized.

Lemma 1. *Consider the minimization problem* (10) *where b is assumed to be strictly convex on Θ. For any A, X, $\mathbf{y}$ and $\lambda \geq 0$: (a) The solution set is convex; (b) For every solution $(\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\beta}}_\lambda)$, $A\hat{\boldsymbol{\alpha}}_\lambda + X\hat{\boldsymbol{\beta}}_\lambda$ is unique; (c) If in addition $\lambda > 0$, $\|\hat{\boldsymbol{\beta}}_\lambda\|_1$ is unique for every solution $(\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\beta}}_\lambda)$.*

The proof of Lemma 1 can be found in Appendix A. It follows that if for a given λ there exists a solution of the form $(\hat{\boldsymbol{\alpha}}_\lambda, \mathbf{0})$, then all other solutions satisfy $\hat{\boldsymbol{\beta}}_\lambda = \mathbf{0}$. Moreover if $\ker(A) = \{\mathbf{0}\}$ and $(\hat{\boldsymbol{\alpha}}_\lambda, \mathbf{0})$ is a solution, then it is the unique solution.

The derivation of the zero-thresholding function for lasso-GLM is based on the following Theorem proved in Appendix A.

Theorem 1. *Assume b is convex on Θ, and define the vector $\boldsymbol{\mu}(\boldsymbol{\alpha}) = (b'(\mathbf{a}_1^{\mathrm{T}}\boldsymbol{\alpha}), \ldots, b'(\mathbf{a}_N^{\mathrm{T}}\boldsymbol{\alpha}))^{\mathrm{T}}$. For any A, X, $\mathbf{y}$ and $0 \leq \lambda < \infty$,*

$$(\hat{\boldsymbol{\alpha}}_\lambda, \mathbf{0}) \in \underset{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{F}}{\operatorname{argmin}} -\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}) + \lambda\|\boldsymbol{\beta}\|_1 \iff \begin{cases} A\hat{\boldsymbol{\alpha}} \in \Theta^N & \text{(11a)} \\ A^{\mathrm{T}}\mathbf{y} = A^{\mathrm{T}}\boldsymbol{\mu}(\hat{\boldsymbol{\alpha}}) & \text{(11b)} \\ \hat{\boldsymbol{\alpha}}_\lambda = \hat{\boldsymbol{\alpha}} & \text{(11c)} \\ \|X^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\alpha}}_\lambda))\|_\infty \leq \lambda. & \text{(11d)} \end{cases}$$

We observe that if $\mathbf{y} \notin \mathcal{D}$, where

$$\mathcal{D} = \{\mathbf{y} : \exists\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{P_0} \text{ satisfying (11a)} - \text{(11b)}\},$$

there is no finite λ such that $\hat{\boldsymbol{\beta}}_\lambda = \mathbf{0}$. This assumption is equivalent to the assumption that there is no MLE of $\boldsymbol{\alpha}$ with regression matrix A based on $\mathbf{y}$.

The zero-thresholding function for lasso-GLM is now derived as an immediate consequence from Theorem 1.

Corollary 1. *The zero-thresholding function associated to $\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y})$ is*

$$\lambda^{\mathrm{GLM}}(\mathbf{y}) = \begin{cases} \|X^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\mu}(\hat{\boldsymbol{\alpha}}_\lambda))\|_\infty & \text{if } \mathbf{y} \in \mathcal{D}, \\ +\infty & \text{otherwise.} \end{cases}$$

In the Gaussian setting, $\lambda^{\mathrm{GLM}}(\mathbf{y}) < \infty$, as opposed to Poisson and Binomial [Barndorff-Nielsen, 1978, Geyer, 2009]. An explicit characterization of $\mathcal{D}$ when $A = \mathbf{1}$ is given in Table 1. We are now equipped to define the quantile universal threshold in lasso-GLM.

Definition 4. *QUT for lasso-GLM in the fixed scenario. Assume $Y \sim f_{(\boldsymbol{\alpha}^0,\boldsymbol{\beta}^0)}$. The quantile universal threshold $\lambda^{\mathrm{QUT}}_{\boldsymbol{\alpha}^0}$ associated to the lasso-GLM estimator $\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y})$ is the upper $(1-\alpha)$ quantile of $\Lambda := \lambda^{\mathrm{GLM}}(\mathbf{Y_0})$, where $\mathbf{Y_0} \sim f_{(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}$ with null model parameters $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = (\boldsymbol{\alpha}^0, \mathbf{0})$.*

Until now we have assumed $[A, X]$ is fixed and only the responses are random. This setting should be distinguished from the random design setting.

Definition 5. *QUT for lasso-GLM in the random scenario. Assume $[A, X]$ is a random matrix with N independent and identically distributed rows. The quantile universal threshold is the upper $(1-\alpha)$ quantile of $\Lambda = \lambda^{\mathrm{GLM}}(\mathbf{Y}_0, A, X)$, where $\mathbf{Y_0} \mid [A, X] \sim f_{(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)}$ with null model parameters $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = (\boldsymbol{\alpha}^0, \mathbf{0})$.*

In practice the distribution of the rows is unknown. By bootstrapping the rows of the observed matrix one can mimic the distribution as described in Algorithm 1. Both fixed and random alternatives are available in our R library.

3.3 Examples

3.3.1 Linear regression

Assuming a Gaussian distribution, b' reduces to the identity function, $\mathcal{D} = \mathbb{R}^N$ and the thresholding statistic is simply given by

$$\Lambda = \|X^{\mathrm{T}}(I - P_A)\mathbf{Y}_0\|_\infty,$$

where P_A is the orthogonal projection onto the range of A and the null model is $\mathbf{Y}_0 \sim \mathrm{N}(A\boldsymbol{\alpha}^0, \sigma^2 I)$. Note that Λ is an ancillary statistic for $\boldsymbol{\alpha}^0$. The quantile universal threshold can equivalently be defined as $\lambda^{\mathrm{QUT}} = \sigma\lambda_Z$ with λ_Z the $1-\alpha$ quantile of $\Lambda_Z = \|X^{\mathrm{T}}(I - P_A)\mathbf{Y}_0\|_\infty$ where $\mathbf{Y}_0 \sim \mathrm{N}(\mathbf{0}, I_N)$. A simple Monte Carlo simulation as in Algorithm 1 allows to evaluate it.

Algorithm 1. *(Calculation of QUT in the Gaussian case)*

- *For* $m = 1, \ldots, M$*:*
  1. *Generate* $\mathbf{z}^{(m)} \sim \mathrm{N}(\mathbf{0}, I_N)$.
  2. *Case* $[A, X]$*-fixed scenario: Calculate* $\lambda^{(m)} = \|X^{\mathrm{T}}(I - P_A)\mathbf{z}^{(m)}\|_\infty$.
  2. *Case* $[A, X]$*-random scenario:*
     - *Generate a bootstrap sample* $[A_{\mathrm{boot,m}}, X_{\mathrm{boot,m}}]$ *of* $[A, X]$ *row-wise.*
     - *Calculate* $\lambda^{(m)} = \|X^{\mathrm{T}}_{\mathrm{boot,m}}(I - P_{A_{\mathrm{boot,m}}})\mathbf{z}^{(m)}\|_\infty$.
- *Calculate* λ_Z *as the upper* $(1-\alpha)$ *empirical quantile of* $\{\lambda^{(m)}\}_{m=1,\ldots,M}$.
- $\lambda^{\mathrm{QUT}} = \sigma\lambda_Z$.

If the variance is unknown, one needs to estimate it as is the case for other methodologies such as SURE, AIC, BIC and SIC. The reader is referred to Appendix B for a discussion on its estimation.

3.3.2 Poisson and logistic regression

When $A = \mathbf{1}$, that is we do not penalize the intercept, explicit formulations of $\mathcal{D}$ and $\gamma = \mathbb{P}(\mathbf{Y} \in \mathcal{D})$ are given in Table 1. In logistic regression with $A \neq \mathbf{1}$, Albert and Anderson [1984] characterize $\mathcal{D}$. Moreover, existence of a solution to (11a)–(11b) can be checked by the R package **impressionableness** [Konis, 2007].

The distribution of the thresholding statistic depends on the value of the nuisance parameter $\boldsymbol{\alpha}_0$ which we estimate with the following scheme: start with $\alpha^{(0)} = \hat{\boldsymbol{\alpha}}$ solving (11a)–(11b), and for $i \geq 0$ iteratively compute $\lambda^{\mathrm{QUT}}_{\boldsymbol{\alpha}^{(i)}}$ by Monte Carlo similarly as in Algorithm 1 and its corresponding lasso estimate $\boldsymbol{\alpha}^{(i+1)}$.

Table 1: Values of $(\mu, \mathcal{D}, \gamma)$ for some distributions when $A = \mathbf{1}$.

Distribution	$\mu = \mathbb{E}(Z)$	$\mathcal{D}$	γ
Gaussian	α^0	$\mathbb{R}^N$	1
Poisson	$\exp(\alpha^0)$	$\mathbb{N}^N \setminus \{\mathbf{0}\}$	$1 - \exp(-n\mu)$
Bernoulli	$\exp(\alpha^0) / (1 + \exp(\alpha^0))$	$\{0,1\}^N \setminus \{\mathbf{0}, \mathbf{1}\}$	$1 - \mu^N + (1-\mu)^N$
Binomial $(k, \mu/k)$	$k \exp(\alpha^0) / (1 + \exp(\alpha^0))$	$\{0, \ldots, k\}^N \setminus \{\mathbf{0}, k\mathbf{1}\}$	$1 - \mu^{kn} + (1-\mu)^{kn}$

4 Numerical results

In this section we introduce two numerical simulations to evaluate the performance of our method in terms of variable selection. We then apply our approach to four real data sets.

4.1 Phase transition

Consider a selection rule for λ and let us recall two prominent quality measures of model selection: true positive rate TPR := $\mathbb{E}$[TPr] and false discovery rate FDR := $\mathbb{E}$[FDr], where TPr := $|\hat{\mathcal{S}}_\lambda \cap \mathcal{S}^0|/|\mathcal{S}^0|$, the proportion of falsely selected features among all selected features, and FDr := $|\hat{\mathcal{S}}_\lambda \cap \bar{\mathcal{S}}^0|/|\hat{\mathcal{S}}_\lambda|$, the proportion of selected nonzero features among all nonzero features. We now define a related quality measure called *oracle inclusive rate*.

Definition 6. *Let* $\hat{s}_\lambda := |\hat{\mathcal{S}}_\lambda|$ *be the cardinality of* $\hat{\mathcal{S}}_\lambda$ *and assume* $s^* := \min_\lambda\{|\hat{\mathcal{S}}_\lambda| : \hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^0\} > 0$. *The oracle inclusive rate (OIR) is defined as* $\mathbb{E}$[OIr]*, where*

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

Models with OIr = 1 have TPr = 1 and minimum FDr amongst all models with TPr = 1. Moreover, OIR $\leq \mathbb{P}(\hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^0)$.

We extend to the noisy setting an experiment designed by Donoho and Tanner [2010] in compressed sensing by considering model (1) with X an $N \times P$ matrix with i.i.d. standard Gaussian entries and $P = 1600$. Two parameters vary: the number of rows $N \in \{160, 320, 480, 640, 800, 960, 1120, 1280, 1440\}$ and the number $s^0 \in \{1, \ldots, N\}$ of nonzero entries of $\boldsymbol{\beta}^0$ whose values we set to ten. For every pair (N, s^0), we generate 100 matrices X and responses $\mathbf{y}$. On the left plot of Figure 1 we report the estimated OIR for QUT as a function of (δ, ρ), where $\delta = N/P$ is the undersampling factor and $\rho = s^0/N$ is the sparsity factor. OIR is near one in the lower right region, implying our selected model contains the correct model with high probability without many unnecessary covariates. In the upper left region, OIR drops

abruptly to zero, a phenomenon known as *phase transition* analyzed by Donoho and Tanner [2010] in compressed sensing.

The right plot of Figure 1 compares, for a fixed value of $\delta = N/P = 0.2$, QUT's performance with three selection rules: SURE, cross-validation (CVmin) which chooses the model with minimum cross validation error and one standard error rule (CV1se) which selects a simpler model than CVmin taking into account the variability of the cross-validation error [Breiman et al., 1984]. QUT is nearly optimal in that regime, whereas CVmin and SURE both have a low OIR, and CV1se is over conservative.

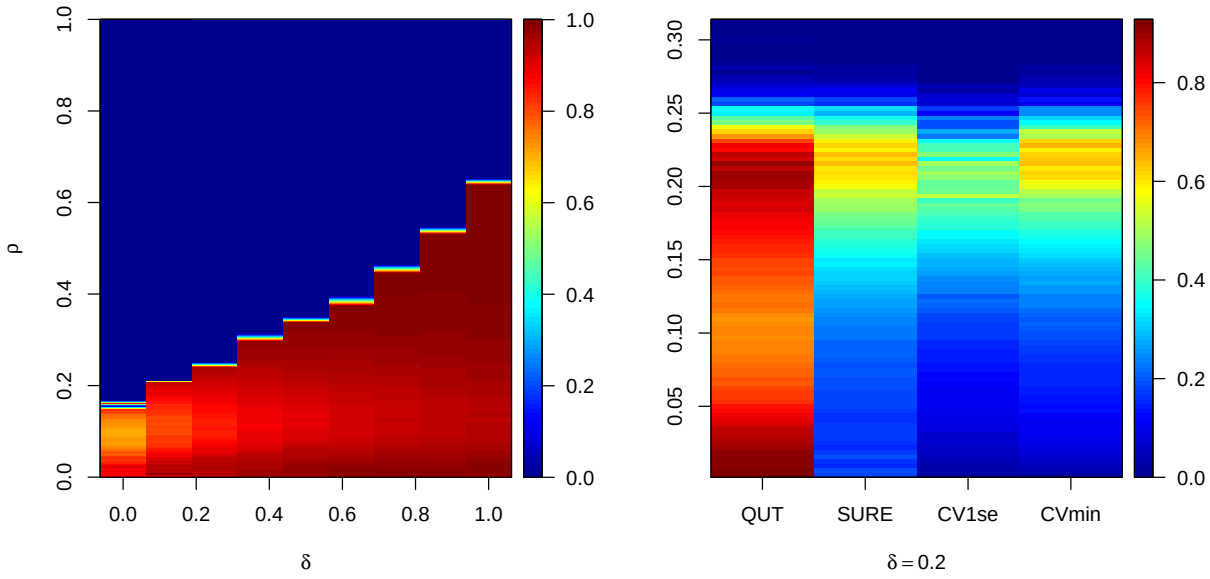

Figure 1: Oracle inclusive rate (OIR): Left, as a function of $(\delta, \rho) = (s^0/N, N/P)$ with QUT; Right, as a function of ρ for a fixed $\delta = 0.2$ with four selection rules .

4.2 True positive rate versus false discovery rate

In order to further compare existing selection rules with our approach, we perform a simulation based on Reid et al. [2014]. Responses are generated according to linear, logistic and Poisson regression for a fixed sample size $N = 100$ and $P = 1000$ covariates. Unit noise variance is assumed in linear regression. Elements of the predictor matrix X are generated randomly following a standard normal distribution and all the correlations between columns of X are set to a positive ω. The resulting correlation matrix of each row Σ_ω guarantees variable screening [Bühlmann and van de Geer, 2011]. The nonzero entries are selected uniformly at random and the corresponding values are generated from a Laplace(1) distribution. Their number is set to $s^0 = \lceil N^\theta \rceil$,

Table 2: TPR/FDR results of Section 4.2 for linear, logistic and Poisson regression simulations.

Method	Response variable distribution					
	Gaussian		Binomial		Poisson	
(θ,ω,snr)	*(0.5,0,1)*	*(0.1,0,1)*	*(0.5,0,10)*	*(0.1,0,10)*	*(0.5,0,1)*	*(0.3,0,1)*
CV1se	0.19/0.26	0.70/0.17	0.17/0.29	0.73/0.36	0.67/0.73	0.86/0.62
QUT	0.13/0.06	0.67/0.05	0.12/0.07	0.63/0.01	0.66/0.72	0.88/0.67
SS	0.12/0.05	0.68/0.02	0.11/0.03	0.65/0.03	0.17/0.04	0.58/0.01
GIC	0.08/0.03	0.69/0.07	0.11/0.09	0.68/0.16	0.71/0.74	0.88/0.65
(θ,ω,snr)	*(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)*
CV1se	0.18/0.75	0.71/0.57	0.13/0.81	0.26/0.40	0.55/0.80	0.01/0.06
QUT	0.17/0.78	0.29/0.01	0.12/0.81	0.14/0.05	0.53/0.79	0.62/0.89
SS	0.03/0.02	0.28/0.00	0.02/0.04	0.12/0.01	0.06/0.11	0.02/0.54
GIC	0.05/0.22	0.55/0.21	0.06/0.31	0.15/0.08	0.54/0.79	0.43/0.87

where $0 \leq \theta \leq 1$ controls the sparsity level of $\boldsymbol{\beta}^0$. The elements of $\boldsymbol{\beta}^0$ are then scaled such that the signal to noise ratio $\text{snr} = \boldsymbol{\beta}^{0\mathrm{T}} \Sigma_\omega \boldsymbol{\beta}^0 / \sigma^2$ takes on specific values.

Table 2 contains estimated TPR and FDR based on 100 replications with four selection rules: CV1se, QUT for random scenario, stability selection (SS) which identifies a set of stable variables that are selected with high probability when repeatedly perturbing the data [Meinshausen and Bühlmann, 2010] and generalized information criterion (GIC). We estimate σ^2 in GIC and QUT with (13) and (14), respectively. The overall conclusions of these simulations that cover a wide range of scenarios are that SS results in a low TPR and CV1se in a high FDR, and QUT is a compromise between the two, similarly to GIC. In other words, our proposal presents a good trade-off between low FDR and high TPR.

4.3 Data analysis

We consider four data sets to illustrate our approach in Gaussian and logistic regression. Below is a brief description of the data sets, specifying the number of observations N, the number of covariates P, and the undersampling factor $\delta^{\text{train}} = (N/2)/P$ of the training sets which are of size $N/2$.

- **Riboflavin** [Buehlmann et al., 2014]: Riboflavin production rate measurements from a population of Bacillus subtilis with sample size $N = 71$ and expressions from $P = 4088$ genes, resulting in $\delta^{\text{train}} = 0.0086$.

- **Chemometrics** [Sardy, 2008]: Fuel octane level measurements with sample size $N = 434$ and $P = 351$ spectrometer measurements, resulting in $\delta^{\text{train}} = 0.62$.

- **Leukemia** [Golub et al., 1999]: Cancer classification of human acute Leukemia cancer types based on $N = 72$ samples of $P = 3571$ gene expression microarrays,

resulting in $\delta^{\text{train}} = 0.01$.

- **InternetAd** [Kushmerick, 1999]: Classification of $N = 2359$ possible advertisements on internet pages based on $P = 1430$ features, resulting in $\delta^{\text{train}} = 0.82$.

We randomly split one thousand times each data set in two, a training and a test set of equal sizes. As in Section 4.2, four different selection rules are compared: CV1se, QUT for random scenario, SS and GIC. To improve the predictive risk, the final model is fitted by MLE with covariates selected by the respective methods (except CV1se which already possesses a good bias-variance trade-off).

In Figure 2, we report the number of nonzero coefficients selected on the training set with lasso, as well as the test set mean-squared prediction error and correct classification rate. We estimate σ^2 in GIC and QUT with (13) and (14), respectively. Following Ockham's razor, if two choices of λ yield comparable predictive performances, the sparser model is preferred. 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 models it selects. The observation that QUT offers a compromise between SS and CV1se carries to real data: good predictive performance is achieved with QUT through a model with complexity between SS and CV1se.

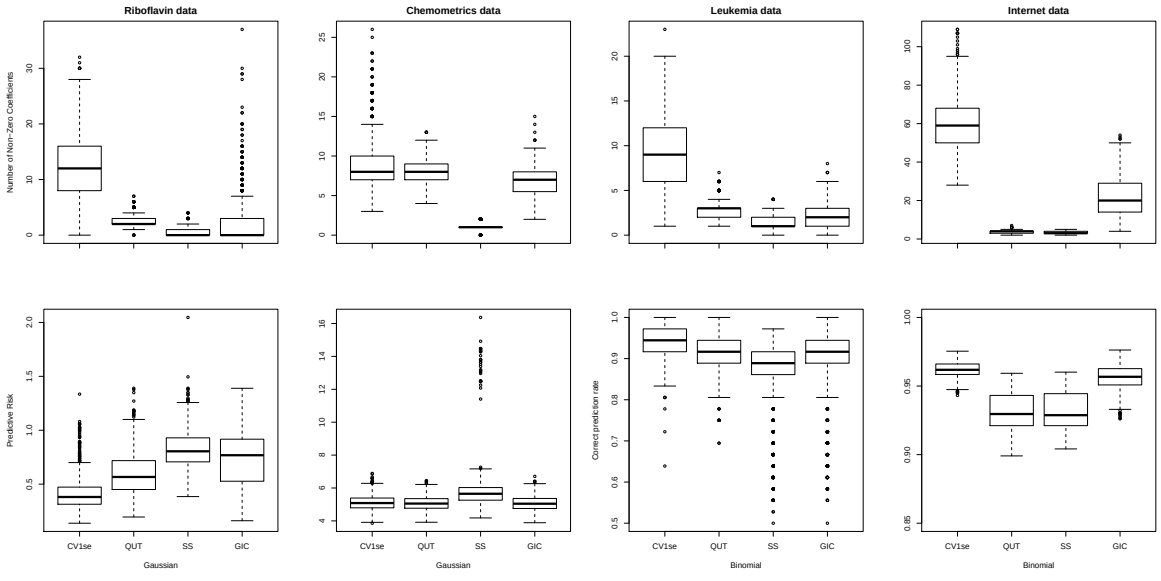

Figure 2: Monte Carlo simulation based on four data sets: Riboflavin (Gaussian), Chemometrics (Gaussian), Leukemia (Binomial) and Internet (Binomial). The reported statistics are the number of nonzero coefficients obtained from the training sets (top) and the test set mean-squared prediction error for Gaussian responses or correct classification rate for binomial responses (bottom).

5 Conclusion

We proposed a unified rule to select a tuning parameter. Our approach relies on the concept of a zero-thresholding function whose explicit formulation we derived for several estimators. The idea is based on the fact that some estimators set to zero all parameters of the models for a finite threshold. Our methodology is recommended for identifying sparse models with few spurious covariates. It is moreover computationally efficient. New applications shall reveal useful to select the regularization parameter of current and future estimators.

The **kt** package which is available from the Comprehensive R Archive Network (CRAN) allows to apply our methodology to ℓ_1-regularized generalized linear models.

6 Acknowledgements

We thank Julie Josse for interesting discussions. The authors of the University of Geneva are supported by the Swiss National Science Foundation.

A Proofs

Proof of Lemma 1. (a) It follows from the strict convexity of b on Θ and the convexity of $f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ on $\mathcal{F}$ that the objective function in (10) is convex on $\mathcal{F}$. The solution set is thus convex.

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

$$-\ell\left(\hat{\boldsymbol{\alpha}}_\lambda^{(3)}, \hat{\boldsymbol{\beta}}_\lambda^{(3)}; \mathbf{y}\right) + \lambda\|\hat{\boldsymbol{\beta}}_\lambda^{(3)}\|_1 < m,$$

where m denotes the minimum value of (10) and the strict inequality follows from the strict convexity of b and the convexity of $f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. In other words, $(\hat{\boldsymbol{\alpha}}_\lambda^{(3)}, \hat{\boldsymbol{\beta}}_\lambda^{(3)})$ is not in the solution set, a contradiction.

(c) For every solution $(\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\beta}}_\lambda)$, $-\ell\left(\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\beta}}_\lambda; \mathbf{y}\right) + \lambda\|\hat{\boldsymbol{\beta}}_\lambda\|_1 = m$. Moreover, $-\ell\left(\hat{\boldsymbol{\alpha}}_\lambda, \hat{\boldsymbol{\beta}}_\lambda; \mathbf{y}\right)$ is unique by (b) because $-\ell$ depends on $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ only through $A\boldsymbol{\alpha} + X\boldsymbol{\beta}$. It follows that if $\lambda > 0$, $\|\hat{\boldsymbol{\beta}}_\lambda\|_1$ is unique.

□

Proof of Theorem 1. It is clear that minimizing (10) over $\mathcal{F}$ is equivalent to minimizing

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} -\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}) + \lambda\|\boldsymbol{\beta}\|_1 & \text{if } (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{F}, \\ +\infty & \text{if } (\boldsymbol{\alpha}, \boldsymbol{\beta}) \notin \mathcal{F}, \end{cases}$$

over all of $\mathbb{R}^{P_0+P}$. Assuming f is convex, a given point $\left(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}\right)$ belongs to the minimum set of f if and only if $\mathbf{0}$ is a subgradient of f at $\left(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}\right)$. This is equivalent to

$$\begin{cases} A\tilde{\boldsymbol{\alpha}} + X\tilde{\boldsymbol{\beta}} \in \Theta^N, \\ A^{\mathrm{T}}(\mathbf{y} - b'\left(A\tilde{\boldsymbol{\alpha}} + X\tilde{\boldsymbol{\beta}}\right)) = 0, \\ X^{\mathrm{T}}(\mathbf{y} - b'\left(A\tilde{\boldsymbol{\alpha}} + X\tilde{\boldsymbol{\beta}}\right)) = \lambda\boldsymbol{\gamma}, \end{cases}$$

for some $\boldsymbol{\gamma} \in \mathbb{R}^P$ such that

$$\gamma_p \in \begin{cases} \left\{\operatorname{sign}(\tilde{\beta}_p)\right\} & \text{if } \tilde{\beta}_p \neq 0, \\ [-1, 1] & \text{if } \tilde{\beta}_p = 0, \end{cases}, \quad p = 1, \ldots, P.$$

Setting $\left(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}\right) = (\boldsymbol{\alpha}_0, \mathbf{0})$ and assuming b is convex, the assertion in Theorem 1 follows. □

B Variance estimation in linear models

Let us consider the linear model (1). When $P > N$, constructing a reliable estimator for σ^2 is a challenging task and several estimators have been proposed. Reid et al. [2014] review some of these estimators and suggest an estimator of the form

$$\hat{\sigma}^2 = \frac{1}{N - \hat{s}_\lambda} \|\mathbf{Y} - X\hat{\boldsymbol{\beta}}_\lambda\|_2^2, \tag{12}$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the lasso estimate with regularization parameter selected via cross validation and $\hat{s}_\lambda$ denotes the number of estimated nonzero entries. Fan et al. [2012] propose a refitted cross validation (RCV) estimator. They split the data set into two equal parts, $(X^{(1)}, \mathbf{Y}^{(1)})$ and $(X^{(2)}, \mathbf{Y}^{(2)})$. On each part, they apply a model selection procedure resulting in two different sets of nonzero indices $\hat{M}_1$, $\hat{M}_2$ with respective cardinality $\hat{m}_1$ and $\hat{m}_2$. This allows to compute

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

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

$$\hat{\sigma}^2_{\mathrm{RCV}} = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}. \tag{13}$$

Consistency and asymptotic normality of this estimator are shown to hold under some regularity assumptions. In practice, they apply the lasso with a cross validation tuned parameter as the model selection procedure in the first stage.

We propose a new estimator of σ^2, refitted QUT, which is defined by

$$\hat{\sigma}^2_{\mathrm{QUT}} = \underset{\sigma^2>0}{\operatorname{argmin}} \left|\sigma^2 - \hat{\sigma}^2_{\mathrm{RCV}}(\sigma^2)\right|, \tag{14}$$

where $\hat{\sigma}^2_{\mathrm{RCV}}(\sigma^2)$ denotes the RCV estimate with the lasso as the model selection procedure and $\lambda^{\mathrm{QUT}}(\sigma^2)$ as the tuning parameter. If consistency of $\hat{\sigma}^2_{\mathrm{RCV}}$ holds, the quantity we seek to minimize converges in probability to zero for σ^2 equal to the noise variance.

Figure 3 shows boxplots of the three estimators of variance applied to the Gaussian data sets of Section 4.3. Refitted QUT has the smallest variance and a median comparable to that of RCV.

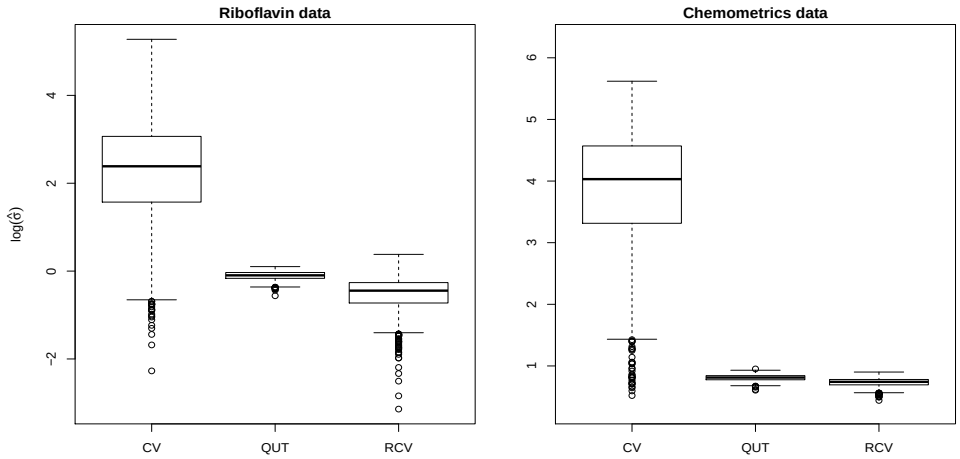

Figure 3: Results of Monte Carlo simulation based on Riboflavin and Chemometrics data of Section 4.3 for the estimation of σ with three estimators: cross-validation (CV) defined in (12), refitted QUT defined in (14) and refitted cross-validation (RCV) defined in (13).

References

H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory*, pages 267–281. Eds. B.N. Petrov and F. Csaki, Budapest: Akademiai Kiado, 1973.

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984.

O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, New York; Chichester, 1978.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

P. Buehlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and its Applications*, 1: 255–278, 2014.

P. Bühlmann and S. van de Geer. In *Statistics for High-Dimensional Data*. Springer-Verlag, 2011.

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

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

E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23:969–985, 2006.

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

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

D. L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied Computational Harmonic Analysis*, 2:101–26, 1995.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52: 1289–306, 2006.

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

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

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

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

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

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

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

M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *arXiv:1405.7511*, 2014.

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

T. R. Golub, P. Slonim, D. K. Tamayo, C. Huard, J. P. Gaasenbeek, M. andMesirov, H. Coller, M. L. Loh, J. R. Downing, C.D. Caligiuri, M. A. AndeBloomfield, and E.S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531–537, 1999.

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

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

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

J. Josse and S. Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, page to appear, 2015.

K. Konis. *Linear programming algorithms for detecting separated data in binary logistic regression models*. PhD thesis, University of Oxford, 2007.

N. Kushmerick. Learning to remove internet advertisements. In *Proceedings of the Third Annual Conference on Autonomous Agents*, pages 175–181, New York, NY, USA, 1999.

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

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 99:2287–2322, 2010.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

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

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

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

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

M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

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

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

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

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

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

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

S. Sardy. Adaptive posterior mode estimation of a sparse sequence for model selection. *Scandinavian Journal of Statistics*, 36:577–601, 2009.

S. Sardy. Smooth blockwise iterative thresholding: a smooth fixed point estimator based on the likelihood's block gradient. *Journal of the American Statistical Association*, 107:800–813, 2012.

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

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

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

C. M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9:1135–1151, 1981.

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

R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

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

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

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

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

Y. Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4), 2005.

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

H. Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

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

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