

# Quantile universal threshold\*

Caroline Giacobino, Sylvain Sardy, Jairo Diaz-Rodriguez

*Department of Mathematics*

*University of Geneva*

*e-mail:* [caroline.giacobino@unige.ch](mailto:caroline.giacobino@unige.ch); [sylvain.sardy@unige.ch](mailto:sylvain.sardy@unige.ch); [jairo.diaz@unige.ch](mailto:jairo.diaz@unige.ch)

and

Nick Hengartner

*Theoretical Biology and Biophysics group*

*Los Alamos National Laboratory*

*e-mail:* [nickh@lanl.gov](mailto:nickh@lanl.gov)

**Abstract:** Efficient recovery of a low-dimensional structure from high-dimensional data has been pursued in various settings including wavelet denoising, generalized linear models and low-rank matrix estimation. By thresholding some parameters to zero, estimators such as lasso, elastic net and subset selection perform variable selection. One crucial step challenges all these estimators: the amount of thresholding governed by a threshold parameter  $\lambda$ . If too large, important features are missing; if too small, incorrect features are included. Within a unified framework, we propose a selection of  $\lambda$  at the detection edge. To that aim, we introduce the concept of a zero-thresholding function and a null-thresholding statistic, that we explicitly derive for a large class of estimators. The new approach has the great advantage of transforming the selection of  $\lambda$  from an unknown scale to a probabilistic scale. Numerical results show the effectiveness of our approach in terms of model selection and prediction.

**Keywords and phrases:** Convex optimization, high-dimensionality, sparsity, regularization, thresholding.

Received March 2017.

## Contents

1	Introduction . . . . .	4702
2	Review of thresholding estimators . . . . .	4703
3	The zero-thresholding function . . . . .	4705
4	The quantile universal threshold . . . . .	4708
4.1	Thresholding under the null . . . . .	4708
4.2	Instances of QUT . . . . .	4708
4.3	Pivotal null-thresholding statistic . . . . .	4709
4.4	Properties of QUT . . . . .	4710

---

\*We would like to thank the Editor, Associate Editor and a reviewer for several useful comments

5	Numerical results of lasso GLM . . . . .	4711
5.1	Based on data sets . . . . .	4712
5.2	Based on simulations . . . . .	4713
5.3	Phase transition property . . . . .	4713
5.4	Conclusion . . . . .	4716
A	Proofs . . . . .	4716
A.1	Proof of Lemma 1 . . . . .	4716
A.2	Proof of Theorem 1 . . . . .	4716
B	Variance estimation in linear models . . . . .	4717
C	Sensitivity study . . . . .	4717
	Acknowledgements . . . . .	4718
	References . . . . .	4718

## 1. Introduction

Many real world examples in which the number of features  $P$  of the model can be dramatically larger than the sample size  $N$  have been identified in various domains such as genomics, finance and image classification, to name a few. In those instances, the maximum likelihood estimation principle fails. Beyond existence and uniqueness issues, it tends to perform poorly when  $P$  is large relative to  $N$  due to its high variance. Motivated by the seminal papers of James and Stein [27] and Tikhonov [58], a considerable amount of literature has concentrated on parameter estimation using regularization techniques. In both parametric and nonparametric models, a reasonable prior or constraints 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.

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

- (i) assumes a certain transform  $\boldsymbol{\xi}^* = g(\boldsymbol{\beta}^*) \in \mathbb{R}^Q$  of the true model parameter  $\boldsymbol{\beta}^* \in \mathbb{R}^P$  is *sparse*, meaning

$$\mathcal{S}^* := \{q \in \{1, \dots, Q\} : \xi_q^* \neq 0\} \quad (1.1)$$

has small cardinality. For example, coordinate-sparsity is induced by  $g(\boldsymbol{\beta}^*) = \boldsymbol{\beta}^*$ , whereas variation-sparsity is induced by  $g(\boldsymbol{\beta}^*) = B\boldsymbol{\beta}^*$  with  $B$  the first order difference matrix;

- (ii) results in an estimated support

$$\hat{\mathcal{S}}_\lambda := \{q \in \{1, \dots, Q\} : \hat{\xi}_{\lambda,q} \neq 0\} \quad (1.2)$$

whose cardinality is governed by the choice of a threshold parameter  $\lambda \geq 0$ .

Thresholding techniques are employed in various settings such as linear regression [15, 54], generalized linear models [40], low-rank matrix estimation [32, 8], density estimation [18, 49], linear inverse problems [13], compressed sensing [14, 9] and time series [38].

Selection of the threshold is crucial to perform effective model selection. It amounts to selecting basis coefficients in wavelet denoising, or genes responsible for a cancer type in microarray data analysis. In change-point detection, it is equivalent to detecting locations of jumps. A too large  $\lambda$  results in a simplistic model missing important features whereas a too small  $\lambda$  leads to a model including many features outside the true model. A typical goal is *variable screening*, that is,

$$\hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^* \quad (1.3)$$

holds with high probability, along with few false detections  $\{q : \hat{\xi}_{\lambda,q} \neq 0, \xi_q^* = 0\}$ . For a suitably chosen  $\lambda$ , certain estimators allow variable screening. The optimal threshold for model identification often differs from the threshold aimed at prediction optimality [61, 31, 34, 64], and it turns out that models aimed at good prediction are typically more complex.

Classical methodologies to select  $\lambda$  consist in minimizing a criterion. Examples include cross-validation, AIC [1], BIC [51] and Stein unbiased risk estimation (SURE) [53]. In low-rank matrix estimation, Owen and Perry [39] and Josse and Husson [28] employ cross-validation whereas Candès et al. [11] and Josse and Sardy [29] apply SURE. The latter methodology is also used in regression [15, 66, 57], and reduced rank regression [36]. Because traditional information criteria do not adapt well to the high-dimensional setting, generalizations such as GIC [21] and EBIC [12] have been suggested.

In this paper, we propose a new threshold selection method that aims at a good identification of the support  $\mathcal{S}^*$ . Our approach has the advantage of transforming the selection of  $\lambda$  from an unknown scale to a probabilistic scale with the simple selection of a probability level. In Section 2, we first review thresholding estimators in linear regression, generalized linear models, low-rank matrix estimation and density estimation. We then introduce the key concept of a zero-thresholding function in Section 3 and derive explicit formulations. In Section 4, we define the null-thresholding statistic, which leads to our proposal: the quantile universal threshold. Some properties are derived. We illustrate the effectiveness of our methodology in Section 5 with four real data sets and simulations. The appendices contain proofs and technical details.

## 2. Review of thresholding estimators

Thresholding estimators are extensively used in the following domains.

**Linear regression.** Consider the linear model

$$\mathbf{Y} = X_0\boldsymbol{\beta}_0^* + X\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I_N), \quad (2.1)$$

where  $X_0$  and  $X$  are matrices of covariates or discretized basis functions of sizes  $N \times P_0$  and  $N \times P$  respectively, and  $\boldsymbol{\beta}_0^*$ ,  $\boldsymbol{\beta}^*$  are unknown coefficients. The vector  $\boldsymbol{\beta}_0^*$  corresponds to  $P_0$  parameters assumed a priori to be nonzero, as is the case for the intercept.

For an observed  $(\mathbf{y}, X_0, X)$ , a large class of estimators is of the form

$$(\hat{\boldsymbol{\beta}}_{0\lambda}, \hat{\boldsymbol{\beta}}_\lambda) \in \underset{(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathbb{R}^{P_0+P}}{\operatorname{argmin}} \mathcal{L}(X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta}, \mathbf{y}) + p_\lambda(g(\boldsymbol{\beta})), \quad (2.2)$$

for a given loss  $\mathcal{L}$  and function  $g$ . A penalty  $p_\lambda$  induces sparsity in  $\hat{\boldsymbol{\xi}}_\lambda = g(\hat{\boldsymbol{\beta}}_\lambda)$ . The element notation “ $\in$ ” indicates the minimizer might not be unique. The lasso [54] defined by

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^P}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \quad (2.3)$$

is among the most popular techniques (we assume here that  $P_0 = 0$  for simplicity). Other examples include:

- (i) Total variation [44], WaveShrink [15], adaptive lasso [64], group lasso [62], generalized lasso [56], sparse group lasso [52], least absolute deviation (LAD) lasso [60], which minimizes  $\|\mathbf{y} - X\boldsymbol{\beta}\|_1 + \lambda \|\boldsymbol{\beta}\|_1$ , square root lasso [3], which minimizes  $\|\mathbf{y} - X\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_1$  and group square root lasso [7].
- (ii) Subbotin lasso [46] where  $p_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_\nu^\nu$ ,  $\nu \leq 1$ , best subset selection, which is equivalent to Subbotin lasso with  $\nu = 0$ , smoothly clipped absolute deviation (SCAD) [19], minimax concave penalty (MCP) [63] and smooth lasso [47].

Convex methodologies (i) also include the Dantzig selector [10]. Although ridge regression [26], bridge [22] and smoothing splines [59] are of the form (2.2), they do not threshold.

**Generalized linear models (GLMs).** The canonical model assumes the log-likelihood is of the form

$$\ell(\boldsymbol{\beta}_0, \boldsymbol{\beta}; \mathbf{y}) = \sum_{n=1}^N [y_n \theta_n - b(\theta_n)] \quad \text{with} \quad \theta_n = \mathbf{x}_{0n}^\top \boldsymbol{\beta}_0 + \mathbf{x}_n^\top \boldsymbol{\beta}, \quad (2.4)$$

$b$  a known function,  $\mathbf{x}_{0n}$  and  $\mathbf{x}_n$  denoting the  $n$ th row of  $X_0$  and  $X$  respectively [37]. As an extension of lasso, Sardy et al. [50] and Park and Hastie [40] define

$$(\hat{\boldsymbol{\beta}}_{0\lambda}, \hat{\boldsymbol{\beta}}_\lambda) \in \underset{(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathcal{F}}{\operatorname{argmin}} -\ell(\boldsymbol{\beta}_0, \boldsymbol{\beta}; \mathbf{y}) + \lambda \|\boldsymbol{\beta}\|_1, \quad (2.5)$$

where  $\mathcal{F} := \{(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \in \mathbb{R}^{P_0+P} \mid X_0\boldsymbol{\beta}_0 + X\boldsymbol{\beta} \in \Theta^N\}$  and  $\Theta := \{\theta \in \mathbb{R} \mid b(\theta) < \infty\}$ . Other penalties such as group lasso [33] have been proposed.

**Low-rank matrix estimation.** Consider the model  $Y = X^* + \sigma Z$ , where  $X^*$  is a low-rank matrix and  $Z_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ . Inspired by lasso, an estimate of  $X^*$  [32, 8] is given by

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

where  $\|\cdot\|_F$  and  $\|\cdot\|_*$  respectively denote the Frobenius and trace norm. For a fixed  $\lambda$ , the solution is  $\hat{X} = U \text{diag}(\hat{\mathbf{d}}_\lambda) V^T$  with  $Y = U \text{diag}(\mathbf{d}) V^T$  the singular value decomposition of  $Y$ , and  $\hat{d}_{\lambda,i} = \max(d_i - \lambda, 0)$ .

**Density estimation.** Let  $Y_1, \dots, Y_N \stackrel{\text{i.i.d.}}{\sim} \varphi$ . A regularized estimate  $\hat{\varphi}_\lambda$  of the discretized density  $\varphi^* = [\varphi(\mathbf{y}_{(1)}), \dots, \varphi(\mathbf{y}_{(N)})]$  is

$$\hat{\varphi}_\lambda = \underset{\varphi \in \mathbb{R}^N}{\text{argmin}} - \sum_{n=1}^N \log \varphi_n + \lambda \|B\varphi\|_1 \quad \text{s.t.} \quad \mathbf{a}^T \varphi = 1,$$

where  $a_1 = (\mathbf{y}_{(2)} - \mathbf{y}_{(1)})/2$ ,  $a_n = (\mathbf{y}_{(n+1)} - \mathbf{y}_{(n-1)})/2$ ,  $n = 2, \dots, N - 1$ ,  $a_N = (\mathbf{y}_{(N)} - \mathbf{y}_{(N-1)})/2$ ,  $\mathbf{y}_{(k)}$  denotes the  $k$ th order statistic and  $B$  is the first order difference matrix [49].

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

**Definition 1.** Assume  $\mathbf{Y} \sim f(\boldsymbol{\eta}^*, \boldsymbol{\beta}^*)$ , with  $\boldsymbol{\xi}^* = g(\boldsymbol{\beta}^*)$  sparse for a certain function  $g$  and  $\boldsymbol{\eta}^*$  a vector of nuisance parameters. Let  $\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y})$  be an estimator indexed by  $\lambda \geq 0$ . We call  $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y}) = g \circ \hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y})$  a thresholding estimator if

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

### 3. The zero-thresholding function

A key property shared by a class of estimators is to set the estimated parameters to zero for a sufficiently large but finite threshold  $\lambda$ . This leads to the following definition.

**Definition 2.** A thresholding estimator  $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y})$  admits a zero-thresholding function  $\lambda_0(\mathbf{Y})$  if

$$\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y}) = \mathbf{0} \quad \Leftrightarrow \quad \lambda \geq \lambda_0(\mathbf{Y}) \quad \text{almost everywhere.}$$

The zero-thresholding function is hence determined uniquely up to sets of measure zero. The equivalence implies equiprobability between setting all coefficients to zero and selecting the threshold large enough. It turns out that such a function has a closed form expression in many instances. Below we derive a catalogue for the estimators reviewed in Section 2.

**Linear regression.** Explicit formulations are the following:

- Lasso, WaveShrink and the Dantzig selector:  $\lambda_0(\mathbf{y}) = \|X^T \mathbf{y}\|_\infty$ ; SCAD and MCP share the same zero-thresholding function when  $X$  is orthonormal. For adaptive lasso,  $\lambda_0(\mathbf{y}) = \|W X^T \mathbf{y}\|_\infty$ , where  $W$  is a diagonal matrix of weights, for LAD-lasso,  $\lambda_0(\mathbf{y}) = \|X^T \text{sgn}(\mathbf{y})\|_\infty$ , where  $\text{sgn}(\cdot)$  is the sign function applied componentwise, and for square root lasso,  $\lambda_0(\mathbf{y}) = \|X^T \mathbf{y}\|_\infty / \|\mathbf{y}\|_2$ .

- Group lasso and square root lasso: if the parameters are partitioned into  $G$  prescribed groups so that  $p_\lambda(\beta) = \lambda \sum_{g=1, \dots, G} \|\beta_g\|_2$ , the zero-thresholding function is  $\lambda_0(\mathbf{y}) = \max_{g=1, \dots, G} \|X_g^T \mathbf{y}\|_2$  for group lasso and  $\lambda_0(\mathbf{y}) = \max_{g=1, \dots, G} \|X_g^T \mathbf{y}\|_2 / \|\mathbf{y}\|_2$  for square root lasso.
- Generalized lasso: Assuming  $B$  has full row rank, let  $\mathcal{I}$  denote a set of column indices such that  $B_{\mathcal{I}}$ , the submatrix of  $B$  with columns indexed by  $\mathcal{I}$ , is invertible. Then,  $\lambda_0(\mathbf{y}) = \|A_1^T (I - P_{A_2}) \mathbf{y}\|_\infty$ , where  $P_X$  is the orthogonal projection matrix onto the range of  $X$ ,  $A_1 = X_{\mathcal{I}} B_{\mathcal{I}}^{-1}$ ,  $A_2 = X_{\bar{\mathcal{I}}} - X_{\mathcal{I}} B_{\mathcal{I}}^{-1} B_{\bar{\mathcal{I}}}$  and  $\bar{\mathcal{I}}$  is the complement of  $\mathcal{I}$ . In one-dimensional total variation,  $\lambda_0(\mathbf{y}) = \|(BB^T)^{-1} B \mathbf{y}\|_\infty$ .
- Best subset:

$$\lambda_0(\mathbf{y}) = \max_{p=1, \dots, \text{rank}(X)} \frac{\Delta_p(\mathbf{y})}{p}, \tag{3.1}$$

where  $\Delta_p(\mathbf{y}) = \frac{1}{2} \max_{\mathcal{I} \subset \{1, \dots, P\}: |\mathcal{I}|=p} \|P_{X_{\mathcal{I}}} \mathbf{y}\|_2^2$ . When  $X$  is orthogonal,  $\lambda_0(\mathbf{y}) = \Delta_1(\mathbf{y})$ .

- Subbotin lasso:  $\lambda_0(\mathbf{y}) = \{\|X^T \mathbf{y}\|_\infty / (2 - \nu)\}^{2-\nu} / \{2(1 - \nu)\}^{\nu-1}$  if  $X$  is orthonormal.

For a convex objective function, derivation of the zero-thresholding function can be inferred from the Karush-Kuhn-Tucker conditions [43]. As an example, we consider LAD-lasso. A given  $\beta \in \mathbb{R}^P$  is a minimum of the objective function  $f(\beta) = \|\mathbf{y} - X\beta\|_1 + \lambda \|\beta\|_1$  if and only if  $\mathbf{0} \in \partial f(\beta)$ , the subdifferential of  $f$  evaluated at  $\beta$ . The zero-thresholding function  $\lambda_0(\mathbf{y}) = \|X^T \text{sgn}(\mathbf{y})\|_\infty$  then follows from the result for all  $\mathbf{y} \in (\mathbb{R}^*)^N$ ,  $\partial f(\mathbf{0}) = -X^T \text{sgn}(\mathbf{y}) + \lambda [-1, 1]^P$ .

Such a derivation can also be performed for estimators with a composite penalty involving a two-dimensional parameter  $\lambda = (\lambda^{(1)}, \lambda^{(2)})$ , for example:

- Elastic net [65] where  $p_\lambda(\beta) = \lambda^{(1)} \|\beta\|_1 + \lambda^{(2)} \|\beta\|_2^2$ : regardless of  $\lambda^{(2)}$ ,  $\lambda_0^{(1)}(\mathbf{y}; \lambda^{(2)}) = \|X^T \mathbf{y}\|_\infty$ .
- Fused lasso [55] where  $p_\lambda(\beta) = \lambda^{(1)} \|\beta\|_1 + \lambda^{(2)} \sum_{p=2}^P |\beta_p - \beta_{p-1}|$ : assuming  $X$  is orthonormal,  $\lambda_0^{(1)}(\mathbf{y}; \lambda^{(2)}) = \|\hat{\beta}_{(0, \lambda^{(2)})}(\mathbf{y})\|_\infty$ .

**Generalized linear models.** The following Lemma shows that although the lasso GLM solution (2.5) might not be unique, its fit is unique (see proof in Appendix A.1).

**Lemma 1.** *Assume  $b$  is strictly convex on  $\Theta$ . For any fixed  $X_0, X, \mathbf{y}$  and  $0 \leq \lambda < \infty$ ,  $X_0 \hat{\beta}_{0\lambda} + X \hat{\beta}_\lambda$  is unique.*

The zero-thresholding function of  $\hat{\beta}_\lambda$  is given in (3.2) below. Its derivation is based on Theorem 1 whose proof can be found in Appendix A.2.

**Theorem 1.** *Assume  $b$  in (2.4) is convex on  $\Theta$  open, and let  $\mu(\beta_0) = (b'(\mathbf{x}_0^T \beta_0), \dots, b'(\mathbf{x}_0^T_N \beta_0))^T$ . For any fixed  $X_0, X, \mathbf{y}$  and  $0 \leq \lambda < \infty$ ,*

$$(\hat{\beta}_{0\lambda}, \mathbf{0}) \in \underset{(\beta_0, \beta) \in \mathcal{F}}{\text{argmin}} -\ell(\beta_0, \beta; \mathbf{y}) + \lambda \|\beta\|_1 \iff \begin{cases} X_0 \hat{\beta}_{0\lambda} \in \Theta^N \\ X_0^T \mathbf{y} = X_0^T \mu(\hat{\beta}_{0\lambda}) \\ \|X^T [\mathbf{y} - \mu(\hat{\beta}_{0\lambda})]\|_\infty \leq \lambda. \end{cases}$$

TABLE 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	$\beta_0^*$	$\mathbb{R}^N$	1
Poisson	$\exp(\beta_0^*)$	$\mathbb{N}^N \setminus \{\mathbf{0}\}$	$1 - \exp(-N\mu)$
Bernoulli	$\frac{\exp(\beta_0^*)}{(1 + \exp(\beta_0^*))}$	$\{0, 1\}^N \setminus \{\mathbf{0}, \mathbf{1}\}$	$1 - \mu^N - (1 - \mu)^N$
Binomial $(m, p) / m$	$\frac{\exp(\beta_0^*)}{(1 + \exp(\beta_0^*))}$	$\{0, 1/m, \dots, 1\}^N \setminus \{\mathbf{0}, \mathbf{1}\}$	$1 - (\mu)^{mN} - (1 - \mu)^{mN}$

Hence, for a strictly convex  $b$  and setting  $\hat{\beta}_\lambda = \mathbf{0}$  if  $\lambda = +\infty$ , the zero-thresholding function of lasso GLM is

$$\lambda_0(\mathbf{y}) = \begin{cases} \|X^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{v})]\|_\infty & \text{if } \mathbf{y} \in \mathcal{D}, \\ +\infty & \text{otherwise,} \end{cases} \tag{3.2}$$

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

$$\begin{cases} X_0 \mathbf{v} \in \Theta^N \\ X_0^T \mathbf{y} = X_0^T \boldsymbol{\mu}(\mathbf{v}) \end{cases} \tag{3.3}$$

and  $\mathcal{D} = \{\mathbf{y} \mid \exists \mathbf{v} \in \mathbb{R}^{P_0} \text{ solution to (3.3)}\}$ . For the group lasso GLM, one obtains

$$\lambda_0(\mathbf{y}) = \begin{cases} \max_{g=1, \dots, G} \|X_g^T[\mathbf{y} - \boldsymbol{\mu}(\mathbf{v})]\|_2 & \text{if } \mathbf{y} \in \mathcal{D}, \\ +\infty & \text{otherwise.} \end{cases}$$

Lemma 1 implies  $\lambda_0(\mathbf{y})$  does not depend on which solution  $\mathbf{v}$  to (3.3) is chosen. The set  $\mathcal{D}$  is the set of values based on which the maximum likelihood estimate (MLE) of  $(\beta_0, \boldsymbol{\beta})$  with constraint  $\hat{\boldsymbol{\beta}} = \mathbf{0}$  exists. If the response variable is Gaussian, then  $\mathcal{D} = \mathbb{R}^N$ . An explicit formulation of  $\mathcal{D}$  when the intercept is unpenalized ( $X_0 = \mathbf{1}$ ) is given in Table 1. For an arbitrary matrix  $X_0$  and under certain assumptions, Giacobino [24] shows that  $\mathcal{D}$  coincides with the set of values  $\mathbf{y}$  such that lasso GLM admits a solution. In particular, the following property holds.

**Property 1.** *Consider a Poisson or (multinomial) logistic regression model. Then, for any fixed  $X_0$ ,  $X$  and  $0 < \lambda < \infty$ , and any observed value  $\mathbf{y}$ , lasso GLM defined in (2.5) admits a solution if and only if a MLE of  $(\beta_0, \boldsymbol{\beta})$  with constraint  $\hat{\boldsymbol{\beta}} = \mathbf{0}$  exists.*

**Low-rank matrix estimation.** The zero-thresholding function is  $\lambda_0(Y) = \|\mathbf{d}\|_\infty$ , the largest singular value of the noisy matrix  $Y$ .

**Density estimation.** The zero-thresholding function is  $\lambda_0(\mathbf{y}) = \|\mathbf{w}\|_\infty$  with  $w_k = N \sum_{i=1}^k a_i - k \sum_{i=1}^N a_i$ ,  $k = 1, \dots, N - 1$ .

## 4. The quantile universal threshold

### 4.1. Thresholding under the null

Inspired by Donoho and Johnstone [15], we now consider the idea of choosing a threshold based on the null model  $\boldsymbol{\xi}^* = \mathbf{0}$ , that is, selecting a threshold  $\lambda$  such that  $\{\hat{\boldsymbol{\xi}}_\lambda = \mathbf{0} \mid \boldsymbol{\xi}^* = \mathbf{0}\}$  holds with high probability. From Definition 2, the events  $\{\hat{\boldsymbol{\xi}}_\lambda = \mathbf{0}\}$  and  $\{\lambda \geq \lambda_0(\mathbf{Y})\}$  are equiprobable. This conducts us to the zero-thresholding function under the null model.

**Definition 3.** Assume  $\hat{\boldsymbol{\xi}}_\lambda(\mathbf{Y})$  admits a zero-thresholding function  $\lambda_0(\mathbf{Y})$ . The null-thresholding statistic is

$$\Lambda_0 := \lambda_0(\mathbf{Y}_0) \quad (4.1)$$

with  $\mathbf{Y}_0 =_d \mathbf{Y}$  under  $H_0 : \boldsymbol{\xi}^* = \mathbf{0}$ .

Given a thresholding estimator and its null-thresholding statistic, selecting  $\lambda$  large enough such that  $\boldsymbol{\xi}^*$  is recovered with probability  $1 - \alpha$  under the null model  $\boldsymbol{\xi}^* = \mathbf{0}$  leads to the following new selection rule.

**Definition 4.** The quantile universal threshold  $\lambda^{\text{QUT}}$  is the upper  $\alpha$ -quantile of  $\Lambda_0$  defined in (4.1).

We discuss the selection of  $\alpha$  in Section 4.4. As we will see, it turns out such a choice results in good empirical and theoretical properties also in the case  $\boldsymbol{\xi}^* \neq \mathbf{0}$ .

If the distribution of  $\Lambda_0$  is unknown,  $\lambda^{\text{QUT}}$  can be computed numerically by Monte Carlo simulation. For instance, one can easily simulate realizations of square root lasso's null-thresholding statistic  $\Lambda_0 = \|X^T \mathbf{Y}_0\|_\infty / \|\mathbf{Y}_0\|_2$ , and compute  $\lambda^{\text{QUT}}$  by taking the appropriate upper quantile. Section 4.4 considers situations where a closed form expression of  $\lambda^{\text{QUT}}$  can be derived.

With the quantile universal threshold, selection of the regularization parameter is now redefined on a probabilistic scale through the probability level  $\alpha$ . QUT is a selection rule designed for model selection as it aims at good identification of the support of the estimand  $\boldsymbol{\xi}^*$ . If one is instead interested in good prediction, then the sparse model identified by QUT can be refitted by maximum likelihood. Such a two step approach has been considered (see Bühlmann and van de Geer [5], Belloni and Chernozhukov [2]) to mimic the behavior of adaptive lasso [64] and results in a smaller bias for large coefficients.

### 4.2. Instances of QUT

A QUT-like selection rule supported by theoretical results has appeared in the following three settings, where a null model for the selection of  $\lambda$  plays a key role.

**Wavelet denoising.** Donoho and Johnstone [15] and Donoho et al. [17] consider an orthonormal  $P \times P$  wavelet matrix and select the threshold of soft-

WaveShrink as  $\lambda_P^{\text{universal}} = \sigma\sqrt{2\log P}$ . Under the null model with wavelet coefficients  $\beta^* = \mathbf{0}$ ,  $\mathbb{P}(\hat{\beta}_{\lambda_P^{\text{universal}}} = \mathbf{0}) \xrightarrow{P \rightarrow \infty} 1$ . An oracle inequality and minimax properties hold with  $\lambda = \lambda_P^{\text{universal}}$  over a wide class of functions, that is, when  $\beta^* \neq \mathbf{0}$ .

**Linear regression.** Desirable properties of estimators such as the lasso, group lasso, square root lasso, group square root lasso or the Dantzig selector are satisfied if the tuning parameter is set to  $\lambda = c\lambda^{(0)}$  for a certain  $c \geq 1$ , such that the event  $\{\hat{\beta}_{\lambda^{(0)}} = \mathbf{0} \mid \beta^* = \mathbf{0}\}$  holds with high probability, for instance with  $\lambda^{(0)} = \lambda^{\text{QUT}}$  for a small  $\alpha$ . More precisely, upper bounds on the estimation and prediction error, as well as the screening property (1.3) 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 [5], Belloni et al. [3], Bunea et al. [7] and references therein.

**Low-rank matrix estimation.** Under the null model  $X^* = 0_{N \times P}$ , it can be shown that with a noise level of  $1/\sqrt{N}$ , the empirical distribution of the singular values of the response matrix converges to a compactly supported distribution. By setting any singular value smaller than the upper bound of the support to zero, Gavish and Donoho [23] derive optimal singular value thresholding operators.

#### 4.3. Pivotal null-thresholding statistic

Assuming a Gaussian distribution, the zero-thresholding function of lasso (3.2) yields the null-thresholding statistic

$$\Lambda_0 = \|X^T(I - P_{X_0})\mathbf{Y}_0\|_\infty,$$

where  $P_{X_0}$  is the orthogonal projection onto the range of  $X_0$  and the null model is  $\mathbf{Y}_0 \sim N(X_0\beta_0^*, \sigma^2 I)$ . Since  $\Lambda_0$  is an ancillary statistic for  $\beta_0^*$ , the quantile universal threshold can equivalently be defined as  $\lambda^{\text{QUT}} = \sigma\lambda_Z$ ,  $\lambda_Z$  being the upper  $\alpha$ -quantile of  $\Lambda_Z = \|X^T(I - P_{X_0})\mathbf{Z}\|_\infty$ , where  $\mathbf{Z} \sim N(\mathbf{0}, I_N)$ . Alike other criteria such as SURE, AIC, BIC and GIC, an estimate of  $\sigma$  is required; see Appendix B for a possible approach.

In contrast, square root lasso's null-thresholding statistic  $\Lambda_0 = \|X^T(I - P_{X_0})\mathbf{Y}_0\|_\infty / \|(I - P_{X_0})\mathbf{Y}_0\|_2$  is pivotal with respect to both  $\beta_0^*$  and  $\sigma$ , and LAD-lasso's is pivotal with respect to  $\sigma$  when  $P_0 = 0$ . This is a clear advantage over lasso, especially in a situation where the nuisance parameter  $\sigma$  is difficult to estimate.

In Poisson and logistic regression, the null-thresholding statistic depends on  $\beta_0^*$  which we estimate with the following procedure. First, calculate the MLE of  $\beta_0$  based on the observed value  $\mathbf{y}$  with the constraint  $\hat{\beta} = \mathbf{0}$  (it is the solution to (3.3)). Then, solve (2.5) with the corresponding quantile universal threshold. Finally, the estimate is  $\hat{\beta}_0^{\text{MLE}}$  where  $(\hat{\beta}_0^{\text{MLE}}, \hat{\beta}^{\text{MLE}})$  denotes the MLE based on  $\mathbf{y}$  with covariates selected by the previous procedure. In Appendix C, we conduct

an empirical investigation of the sensitivity of our approach to the estimation of  $\beta_0^*$ .

The random design setting is the situation where not only the response vector but also the matrix of covariates is random, like all four data sets in Section 5.1. To account for the variability due to random design, we define the quantile universal threshold as the upper  $\alpha$ -quantile of  $\Lambda_0 = \lambda_0(\mathbf{Y}_0, [X_0, X])$ , with  $[X_0, X]$  consisting of independent identically distributed rows. To account for the variability in the design matrix, the rows of  $[X_0, X]$  can be bootstrapped within the Monte Carlo simulation to evaluate  $\lambda^{\text{QUT}}$ . Both fixed and random alternatives are implemented in our R package **qut**.

#### 4.4. Properties of QUT

Before considering the choice of  $\alpha$  and deriving an explicit formulation of the quantile universal threshold in some settings, more theoretical properties are derived. Upper bounds on the estimation and prediction error of the lasso tuned with  $\lambda^{\text{QUT}}$  as well as a sufficient condition for the screening property (1.3) follow from the next property.

**Property 2.** *Assume the  $(L, \mathcal{S}^*)$ -compatibility condition is satisfied for  $\mathcal{S}^*$  of cardinality  $s^*$  with  $L = (\lambda + \lambda^{(0)})/(\lambda - \lambda^{(0)})$ , for a certain  $\lambda^{(0)}$ ,  $0 < \lambda^{(0)} < \lambda$ , that is,*

$$\phi_{\text{comp}}(L, \mathcal{S}^*) := \min \left\{ \sqrt{s^*} \|X\beta\|_2 / \|\beta_{\mathcal{S}^*}\|_1 \mid \|\beta_{\mathcal{S}^c}\|_1 \leq L \|\beta_{\mathcal{S}^*}\|_1, \beta \neq \mathbf{0} \right\} > 0.$$

Then lasso (2.3) with  $\lambda = \lambda^{\text{QUT}}$  satisfies with probability at least  $1 - \alpha - \mathbb{P}(\lambda^{(0)} \leq \Lambda_0 \leq \lambda)$

- (i)  $\|X(\hat{\beta}_\lambda - \beta^*)\|_2^2 / 2 \leq 8(\lambda + \lambda^{(0)})^2 s^* / \phi_{\text{comp}}^2(L, \mathcal{S}^*)$ ,
- (ii)  $\|(\hat{\beta}_\lambda - \beta^*)_{\mathcal{S}^*}\|_1 \leq A$ ,  $A = 4(\lambda + \lambda^{(0)}) s^* / \phi_{\text{comp}}^2(L, \mathcal{S}^*)$ ,
- (iii)  $\|(\hat{\beta}_\lambda)_{\mathcal{S}^c}\|_1 \leq 4s^*(\lambda + \lambda^{(0)})^2 / \{(\lambda - \lambda^{(0)})\phi_{\text{comp}}^2(L, \mathcal{S}^*)\}$ .

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  $\mathbb{P}(\lambda^{(0)} \leq \Lambda_0 \leq \lambda)$  can be made arbitrarily small for a well-chosen  $\lambda^{(0)}$  as long as the  $(L, \mathcal{S}^*)$ -compatibility condition is met. The proof of the property is omitted as it is essentially the same as for Theorem 6.1 in Bühlmann and van de Geer [5] using the fact that the key statistic they bound with high probability is the null-thresholding statistic  $\Lambda_0 = \|X^T \epsilon\|_\infty = \lambda_0(\mathbf{Y}_0)$  defined in (4.1). The screening property is a direct consequence of (ii). Similar results can be shown for the group lasso, square root lasso, group square root lasso and the Dantzig selector.

Another important property of our methodology concerns the familywise error rate. Recall that when performing multiple hypothesis tests, it 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. 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 5.2. Hence, Definition 4 implies the following property.

**Property 3.** *Any thresholding estimator tuned with  $\lambda^{\text{QUT}}$  controls the familywise error rate as well as the false discovery rate at level  $\alpha$  in the weak sense.*

The probability of the previous properties is determined by  $\alpha$ ; we recommend  $\alpha = 0.05$  as Belloni et al. [3]. An alternative is to set  $\alpha_P$  tending to zero as the number  $P$  of covariates goes to infinity. Donoho and Johnstone [15] implicitly select a rate of convergence of  $\alpha_P = O(1/\sqrt{\log P})$  (Josse and Sardy [29] also select this rate).

Finally, an explicit formulation of the quantile universal threshold can be derived in the following settings:

- (i) In orthonormal regression with best subset selection and threshold equal  $\sigma\sqrt{2\log P}$  discussed in Section 4.2, the equivalent penalty is  $\lambda^{\text{QUT}} = 2\lambda^{\text{BIC}} = \sigma^2 \log P$  satisfying  $\bar{F}_{\Lambda_0}(\lambda^{\text{QUT}}) \sim 1/\sqrt{\pi \log P}$ . This result can be inferred from the null-thresholding statistic  $\Lambda_0 =_d \|\mathbf{Z}\|_\infty^2/2$  using (3.1), where  $Z_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ . Generalizations such as GIC and EBIC also select a larger tuning parameter than BIC which performs poorly in the high-dimensional setting.
- (ii) In total variation, the null-thresholding statistic converges in distribution to the infinite norm of a Brownian bridge, then  $\lambda^{\text{QUT}} = \sigma\sqrt{P \log \log P}/2$  for  $\alpha_P = O(1/\sqrt{\log P})$  [48]. For block total variation, the null-thresholding statistic tends to the maximum of a Bessel bridge, which distribution is known [41].
- (iii) In group lasso with orthonormal groups, each of size  $Q$ , extreme value theory leads to  $\lambda^{\text{QUT}} = \sigma\sqrt{2 \log P + (Q-1) \log \log P - 2 \log \Gamma(Q/2)}$  [47].

## 5. Numerical results of lasso GLM

The QUT methodology for lasso and square root lasso is implemented in the **qut** package which is available from the Comprehensive R Archive Network (CRAN). In the following,  $\text{QUT}_{\text{lasso}}$  and  $\text{QUT}_{\sqrt{\text{lasso}}}$  stand for QUT applied to lasso and square root lasso. CVmin refers to cross-validation, CV1se to a conservative variant of CVmin which takes into account the variability of the cross-validation error [4], SS to stability selection [35] and GIC to the generalized information criterion [21]. When applying GIC and  $\text{QUT}_{\text{lasso}}$ , the variance is estimated with (B.2) and (B.3) respectively. The level  $\alpha$  is set to 0.05.

### 5.1. Based on data sets

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

- **riboflavin** [6]: Riboflavin production rate measurements from a population of *Bacillus subtilis* with sample size  $N = 71$  and expressions from  $P = 4088$  genes.
- **chemometrics** [45]: Fuel octane level measurements with sample size  $N = 434$  and  $P = 351$  spectrometer measurements.
- **leukemia** [25]: Cancer classification of human acute leukemia cancer types based on  $N = 72$  samples of  $P = 3571$  gene expression microarrays.
- **internetAd** [30]: Classification of  $N = 2359$  possible advertisements on internet pages based on  $P = 1430$  features.

We randomly split one hundred times into a training and a test set of equal size. Five lasso selection rules are compared including QUT. Except for CV1se, the final model is fitted by MLE with the previously selected covariates in order to improve prediction. In Figure 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.

Good predictive performance is achieved by  $\text{QUT}_{\text{lasso}}$  as well as GIC with a median model complexity between SS and CV1se.  $\text{QUT}_{\text{lasso}}$  works remarkably

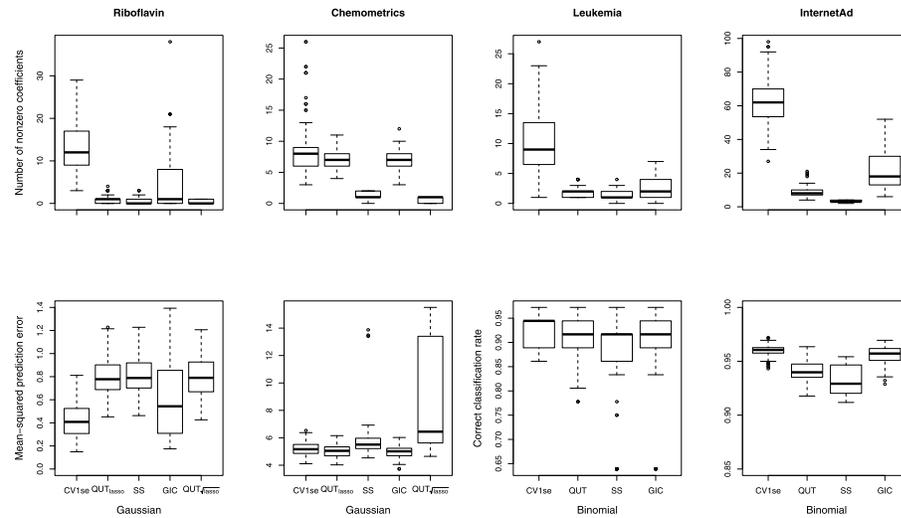


FIG 1. Monte Carlo simulation based on four data sets: **riboflavin** (Gaussian), **chemometrics** (Gaussian), **leukemia** (Binomial) and **internetAd** (Binomial). We report the boxplots of the following statistics: the number of nonzero coefficients obtained from the training sets (top); the test set mean-squared prediction error for Gaussian responses and the correct classification rate for binomial responses (bottom).

well for **chemometrics** and **leukemia**. By selecting a large number of variables CV1se results in efficient prediction, whereas SS and  $\sqrt{\text{lasso}}$  show poor predictive performance due to the low complexity of their estimated model. Moreover, GIC exhibits a larger variability than  $\text{QUT}_{\text{lasso}}$  and  $\text{QUT}_{\sqrt{\text{lasso}}}$  in terms of number of nonzero coefficients.

### 5.2. Based on simulations

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

We perform a simulation based on Reid et al. [42]. Responses are generated from the linear, logistic and Poisson regression model with  $N = 100$  and  $P = 1000$ . The intercept is set to one and unit noise variance is assumed in linear regression. The true parameter  $\beta^*$  and predictor matrix  $X$  are obtained as follows:

- Entries of  $X$  are standard Gaussian with correlation between columns set to  $\omega$ .
- The support of  $\beta^*$  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,  $\text{snr} = \beta^{*\text{T}} \Sigma_\omega \beta^*$ ,  $\Sigma_\omega$  being the covariance matrix of a single row of  $X$  and for a noise variance  $\sigma^2 = 1$  in the Gaussian case.

Table 2 contains estimated TPR and FDR based on one hundred replications. We also report the predictive root mean squared error defined by  $\text{RMSE}^2 = \mathbb{E}\{(\mathbf{x}_{\text{new}}^{\text{T}} \beta^* - \mathbf{x}_{\text{new}}^{\text{T}} \hat{\beta})^2\} / \text{snr}$ ; here the expectation is taken over training sets and new predictive locations  $\mathbf{x}_{\text{new}}$ . Looking at TPR and FDR, the high complexity of CV1se and the low complexity of SS and  $\sqrt{\text{lasso}}$  are again observed. Looking at RMSE,  $\text{QUT}_{\text{lasso}}$  often performs best thanks to a good sparse model before fitting by MLE. Finally,  $\text{QUT}_{\text{lasso}}$  and GIC are comparable in terms of RMSE, but  $\text{QUT}_{\text{lasso}}$  often has a better compromise between TPR and FDR.

### 5.3. Phase transition property

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 5.** Assume  $\mathcal{S}^* \neq \emptyset$  and let  $s_{\min} := \min_\lambda \{\hat{s}_\lambda = |\hat{\mathcal{S}}_\lambda| : \hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}^*\}$  if it exists. The oracle inclusive rate (OIR) for a selection rule  $\lambda$  is defined as

TABLE 2  
*Estimated TPR/FDR/RMSE based on the simulation of Section 5.2.*

Method	Response variable distribution			
	Gaussian	Binomial	Poisson	
	$(\theta, \omega, snr)$	$(0.5, 0, 1)$	$(0.5, 0, 10)$	$(0.5, 0, 0.5)$
<b>lasso</b>				
CV1se		0.23/0.27/0.87	0.26/0.42/0.10	0.28/0.34/3.35
QUT <sub>lasso</sub>		0.09/0.02/0.85	0.10/0.02/0.10	0.37/0.57/2.94
SS		0.12/0.03/0.81	0.11/0.03/0.10	0.13/0.02/3.27
GIC		0.10/0.04/0.85	0.13/0.12/0.10	0.35/0.50/2.98
$\sqrt{\text{lasso}}$				
QUT $\sqrt{\text{lasso}}$		0.05/0.01/0.92		
	$(\theta, \omega, snr)$	$(0.1, 0, 1)$	$(0.1, 0, 10)$	$(0.3, 0, 0.5)$
<b>lasso</b>				
CV1se		0.70/0.25/0.57	0.83/0.50/0.06	0.62/0.38/2.51
QUT <sub>lasso</sub>		0.61/0.00/0.35	0.67/0.00/0.04	0.64/0.44/1.96
SS		0.66/0.03/0.31	0.74/0.01/0.04	0.40/0.02/2.12
GIC		0.68/0.13/0.36	0.78/0.13/0.04	0.64/0.47/2.03
$\sqrt{\text{lasso}}$				
QUT $\sqrt{\text{lasso}}$		0.24/0.00/0.80		
	$(\theta, \omega, snr)$	$(0.5, 0.4, 1)$	$(0.5, 0.4, 10)$	$(0.5, 0.4, 0.5)$
<b>lasso</b>				
CV1se		0.18/0.79/0.67	0.15/0.80/0.08	0.24/0.82/2.56
QUT <sub>lasso</sub>		0.13/0.71/0.63	0.12/0.78/0.09	0.26/0.82/2.41
SS		0.03/0.03/0.92	0.02/0.08/0.11	0.03/0.03/3.64
GIC		0.06/0.37/0.83	0.06/0.48/0.10	0.24/0.81/2.37
$\sqrt{\text{lasso}}$				
QUT $\sqrt{\text{lasso}}$		0.02/0.25/0.92		
	$(\theta, \omega, snr)$	$(0.5, 0, 10)$	$(0.5, 0, 20)$	$(0.5, 0, 2)$
<b>lasso</b>				
CV1se		0.76/0.61/0.39	0.33/0.50/0.07	0.58/0.73/10.78
QUT <sub>lasso</sub>		0.20/0.00/0.66	0.12/0.02/0.07	0.64/0.77/ 9.04
SS		0.26/0.00/0.59	0.14/0.02/0.07	0.11/0.14/12.14
GIC		0.55/0.22/0.42	0.18/0.15/0.07	0.65/0.78 / 9.10
$\sqrt{\text{lasso}}$				
QUT $\sqrt{\text{lasso}}$		0.06/0.00/0.88		

$\mathbb{E}[\text{OIr}]$ , where

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

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

We extend the simulation of Donoho and Tanner [16] in compressed sensing to model (2.1) with unit noise variance assumed to be known. The entries of the

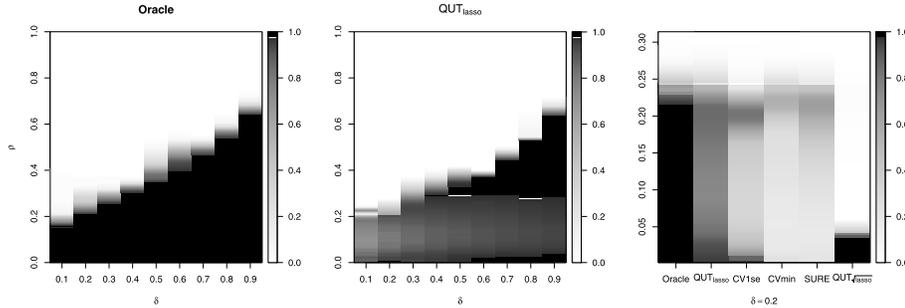


FIG 2. Estimated OIR of the oracle lasso selection rule (left) and QUT (middle) as a function of  $(\delta, \rho) = (N/P, s^*/N)$ . The right panel contains the estimated OIR of several selection rules for a fixed  $\delta = 0.2$ .

$N \times P$   $X$  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  $\beta^*$ ,  $s^* \in \{1, \dots, N\}$ . 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 panel of Figure 2, we report OIR for the oracle lasso selection rule which retains the optimal inclusive model if it exists. Values are plotted as a function of  $\delta = N/P$ , the undersampling factor, and of  $\rho = s^*/N$ , the sparsity factor. On the middle and right panel, we report OIR for  $\text{QUT}_{\text{lasso}}$  along other methodologies as well as  $\text{QUT}_{\sqrt{\text{lasso}}}$ . The following interesting behaviors 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  $\rho$ .
- Low complexity of  $\text{QUT}_{\text{lasso}}$ . 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 high 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.
- Low OIR of  $\text{QUT}_{\sqrt{\text{lasso}}}$ . This could be due to the fact that  $\sqrt{\text{lasso}}$  requires stronger conditions than lasso with a known variance to achieve variable screening [3]. Considering its zero-thresholding function, not only its numerator increases as the model deviates from the null model (as lasso), but also its denominator, making screening harder to reach with  $\alpha = 0.05$ .

#### 5.4. Conclusion

According to Ockham's razor, if two selected models yield comparable predictive performances, the sparsest should be preferred. Lasso with QUT tends to be in accordance with this principle by selecting low complexity models that achieve good predictive performance. A good compromise between high TPR and low FDR is obtained.

### Appendix A: Proofs

#### A.1. Proof of Lemma 1

It follows from the strict convexity of  $b$  on  $\Theta$  and the convexity of  $f(\beta_0, \beta) = \|\beta\|_1$  on  $\mathcal{F}$  that the objective function in (2.5) is convex on  $\mathcal{F}$ . The solution set is thus convex.

Assume there exists two solutions  $(\hat{\beta}_{0\lambda}^{(1)}, \hat{\beta}_\lambda^{(1)})$  and  $(\hat{\beta}_{0\lambda}^{(2)}, \hat{\beta}_\lambda^{(2)})$  such that  $X_0\hat{\beta}_{0\lambda}^{(1)} + X\hat{\beta}_\lambda^{(1)} \neq X_0\hat{\beta}_{0\lambda}^{(2)} + X\hat{\beta}_\lambda^{(2)}$ . Because the solution set is convex,

$$(\hat{\beta}_{0\lambda}^{(3)}, \hat{\beta}_\lambda^{(3)}) := \delta(\hat{\beta}_{0\lambda}^{(1)}, \hat{\beta}_\lambda^{(1)}) + (1 - \delta)(\hat{\beta}_{0\lambda}^{(2)}, \hat{\beta}_\lambda^{(2)})$$

is a solution for any  $0 < \delta < 1$ . However,

$$-\ell(\hat{\beta}_{0\lambda}^{(3)}, \hat{\beta}_\lambda^{(3)}; \mathbf{y}) + \lambda\|\hat{\beta}_\lambda^{(3)}\|_1 < m,$$

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  $f(\beta_0, \beta) = \|\beta\|_1$ . In other words,  $(\hat{\beta}_{0\lambda}^{(3)}, \hat{\beta}_\lambda^{(3)})$  is not in the solution set, a contradiction.

#### A.2. Proof of Theorem 1

Minimizing (2.5) over  $\mathcal{F}$  is equivalent to minimizing

$$f(\beta_0, \beta) = \begin{cases} -\ell(\beta_0, \beta; \mathbf{y}) + \lambda\|\beta\|_1 & \text{if } (\beta_0, \beta) \in \mathcal{F}, \\ +\infty & \text{if } (\beta_0, \beta) \notin \mathcal{F}, \end{cases}$$

over all of  $\mathbb{R}^{P_0+P}$ . Assuming  $f$  is convex, a given point  $(\beta_0, \beta)$  belongs to the minimum set of  $f$  if and only if  $\mathbf{0}$  is a subgradient of  $f$  at  $(\beta_0, \beta)$ . This is equivalent to

$$\begin{cases} X_0\beta_0 + X\beta \in \Theta^N, \\ X_0^T(\mathbf{y} - b'(X_0\beta_0 + X\beta)) = 0, \\ X^T(\mathbf{y} - b'(X_0\beta_0 + X\beta)) = \lambda\gamma, \end{cases}$$

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

$$\gamma_p \in \begin{cases} \{\text{sign}(\beta_p)\} & \text{if } \beta_p \neq 0, \\ [-1, 1] & \text{if } \beta_p = 0, \end{cases}, \quad p = 1, \dots, P.$$

Setting  $(\beta_0, \beta) = (\hat{\beta}_{0,\lambda}, \mathbf{0})$  and assuming  $b$  is convex, the assertion in Theorem 1 follows.

## Appendix B: Variance estimation in linear models

When  $P > N$  in (2.1) ( $P_0 = 0$  is assumed for simplicity), constructing a reliable estimator for  $\sigma^2$  is a challenging task and several estimators have been proposed. Reid et al. [42] consider an estimator of the form

$$\hat{\sigma}^2 = \frac{1}{N - \hat{s}_\lambda} \|\mathbf{Y} - X\hat{\beta}_\lambda\|_2^2, \quad (\text{B.1})$$

where  $\hat{\beta}_\lambda$  is the lasso estimator tuned with cross-validation and  $\hat{s}_\lambda$  denotes the number of estimated nonzero entries. Fan et al. [20] 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 is applied 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

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

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

$$\hat{\sigma}_{\text{RCV}}^2 := \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}. \quad (\text{B.2})$$

Consistency and asymptotic normality hold under some regularity assumptions. In practice, the lasso tuned with cross-validation is applied in the first stage.

We propose a new estimator of  $\sigma^2$ , refitted QUT, which is defined as

$$\hat{\sigma}_{\text{QUT}}^2 := \underset{\sigma^2 > 0}{\operatorname{argmin}} |\sigma^2 - \hat{\sigma}_{\text{RCV}}^2(\sigma^2)|, \quad (\text{B.3})$$

where  $\hat{\sigma}_{\text{RCV}}^2(\sigma^2)$  is the RCV estimate with the lasso tuned with  $\lambda^{\text{QUT}}(\sigma^2)$ . Figure 3 shows boxplots of the three estimators of variance applied to the Gaussian data of Section 5.1. Refitted QUT has smallest variability and seems slightly more conservative than CV and RCV.

## Appendix C: Sensitivity study

As noted in Section 4.3, the null-thresholding statistic and therefore the quantile universal threshold are functions of the unknown intercept  $\beta_0^*$ . In Figure 4, we empirically investigate the sensitivity of our method to the estimation of  $\beta_0^* = 1$  on the Poisson distributed data of Section 5.2. On the left panel, estimation of  $\beta_0^*$  (dark grey) described at the end of Section 4.3 has low bias. Moreover we observe the relative median insensitivity of TPr and FDr to the estimate.

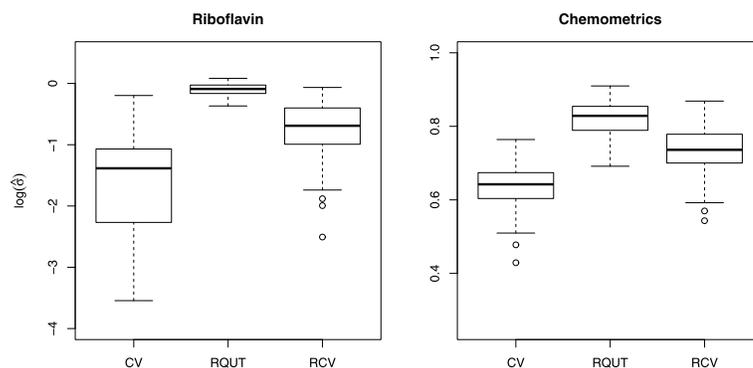


FIG 3. Results of Monte Carlo simulation based on riboflavin and chemometrics data of Section 5.1 for the estimation of  $\sigma$  with cross-validation (CV) defined in (B.1), refitted QUT defined in (B.3) and refitted cross-validation (RCV) defined in (B.2).

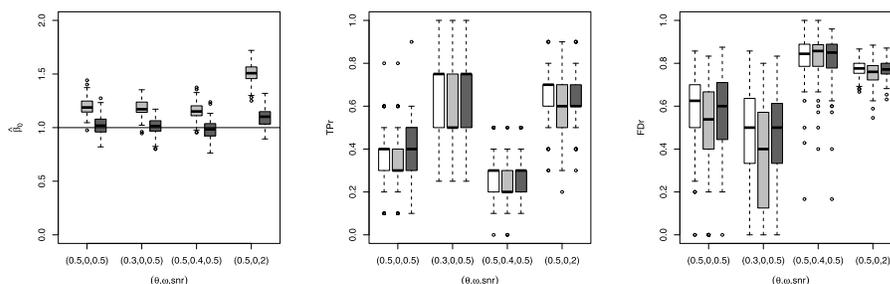


FIG 4. Estimation of  $\beta_0^* = 1$  (left) and its effect on  $TPr$  (middle) and  $FDr$  (right). White, light grey and dark grey boxplots correspond respectively to the oracle estimator  $\beta_0 = 1$ , initial step and final step of our estimation procedure.

## Acknowledgements

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

## References

- [1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998. [MR1486823](#)
- [2] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. [MR3037163](#)
- [3] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. [MR2860324](#)

- [4] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole Advanced Books & Software, Monterey, CA, 1984. [MR0726392](#)
- [5] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg, 2011. [MR2807761](#)
- [6] P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.
- [7] F. Bunea, J. Lederer, and Y. She. The group square-root lasso: theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2):1313–1325, 2014. [MR3164977](#)
- [8] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. [MR2600248](#)
- [9] E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007. [MR2329927](#)
- [10] E. Candès and T. Tao. The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics*, 35(6):2313–2351, 2007. [MR2382644](#)
- [11] E. J. Candès, C. A. Sing-Long, and J. D. Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19):4643–4657, 2013. [MR3105401](#)
- [12] J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008. [MR2443189](#)
- [13] D. L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis*, 2(2):101–126, 1995. [MR1325535](#)
- [14] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. [MR2241189](#)
- [15] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. [MR1311089](#)
- [16] D. L. Donoho and J. Tanner. Precise undersampling theorems. *Proceedings of the IEEE*, 98(6):913–924, 2010.
- [17] D. L. Donoho, I. M. Johnstone, G. Kerkycharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B*, 57(2):301–369, 1995. [MR1323344](#)
- [18] D. L. Donoho, I. M. Johnstone, G. Kerkycharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996. [MR1394974](#)
- [19] J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004. [MR2065194](#)
- [20] 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. [MR2885839](#)

- [21] Y. Fan and C. Y. Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B*, 75(3):531–552, 2013. [MR3065478](#)
- [22] W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998. [MR1646710](#)
- [23] M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. [arXiv:1405.7511v2](#), 2014. [MR3626861](#)
- [24] C. Giacobino. *Thresholding estimators for high-dimensional data: model selection, testing and existence*. PhD thesis, University of Geneva, 2017.
- [25] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.
- [26] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. [MR0611894](#)
- [27] W. James and C. Stein. 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. [MR0133191](#)
- [28] J. Josse and F. Husson. Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*, 56(6):1869–1879, 2012. [MR2892383](#)
- [29] J. Josse and S. Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, 26(3):715–724, 2016. [MR3489867](#)
- [30] N. Kushmerick. Learning to remove internet advertisements. In *Proceedings of the third international conference on Autonomous Agents*, pages 175–181. ACM, 1999.
- [31] 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. [MR2327490](#)
- [32] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010. [MR2719857](#)
- [33] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70(1):53–71, 2008. [MR2412631](#)
- [34] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006. [MR2278363](#)
- [35] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010. [MR2758523](#)
- [36] A. Mukherjee, K. Chen, N. Wang, and J. Zhu. On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 102(2):457–477, 2015. [MR3371016](#)
- [37] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384, 1972.

- [38] D. Neto, S. Sardy, and P. Tseng.  $\ell_1$ -penalized likelihood smoothing and segmentation of volatility processes allowing for abrupt changes. *Journal of Computational and Graphical Statistics*, 21(1):217–233, 2012. [MR2913364](#)
- [39] A. B. Owen and P. O. Perry. Bi-cross-validation of the svd and the non-negative matrix factorization. *Annals of Applied Statistics*, 3(2):564–594, 2009. [MR2750673](#)
- [40] M. Y. Park and T. Hastie.  $L_1$ -regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B*, 69(4):659–677, 2007. [MR2370074](#)
- [41] J. Pitman and M. Yor. The law of the maximum of a besel bridge. *Electronic Journal of Probability*, 4:1–35, 1999.
- [42] S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. [arXiv:1311.5274v2](#), 2014.
- [43] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.
- [44] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.
- [45] S. Sardy. On the practice of rescaling covariates. *International Statistical Review*, 76(2):285–297, 2008.
- [46] S. Sardy. Adaptive posterior mode estimation of a sparse sequence for model selection. *Scandinavian Journal of Statistics*, 36(4):577–601, 2009.
- [47] 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(498):800–813, 2012.
- [48] 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(465):191–204, 2004.
- [49] S. Sardy and P. Tseng. Density estimation by total variation penalized likelihood driven by the sparsity  $\ell_1$  information criterion. *Scandinavian Journal of Statistics*, 37(2):321–337, 2010.
- [50] S. Sardy, A. Antoniadis, and P. Tseng. Automatic smoothing with wavelets for a wide class of distributions. *Journal of Computational and Graphical Statistics*, 13(2):399–421, 2004.
- [51] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [52] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.
- [53] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.
- [54] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [55] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, 67(1):91–108, 2005.
- [56] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

- [57] R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.
- [58] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4(4):1035–1038, 1963.
- [59] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.
- [60] 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.
- [61] Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950, 2005. [MR2234196](#)
- [62] 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. [MR2212574](#)
- [63] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. [MR2604701](#)
- [64] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. [MR2279469](#)
- [65] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005. [MR2137327](#)
- [66] H. Zou, T. Hastie, and R. Tibshirani. On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007. [MR2363967](#)