Smooth blockwise iterative thresholding: a smooth fixed point estimator based on the likelihood's block gradient

SYLVAIN SARDY*
Department of Mathematics, University of Geneva

The proposed smooth blockwise iterative thresholding estimator (SBITE) is a model selection technique defined as a fixed point reached by iterating a likelihood gradient-based thresholding function. The smooth James-Stein thresholding function has two regularization parameters λ and ν, and a smoothness parameter s. It enjoys smoothness like ridge regression and selects variables like lasso. Focusing on Gaussian regression, we show that SBITE is uniquely defined, and that its Stein unbiased risk estimate is a smooth function of λ and ν, for better selection of the two regularization parameters. We perform a Monte-Carlo simulation to investigate the predictive and oracle properties of this smooth version of adaptive lasso.

The motivation is a gravitational wave burst detection problem from several concomitant time series. A nonparametric wavelet-based estimator is developed to combine information from all captors by block-thresholding multiresolution coefficients. We study how the smoothness parameter s tempers the erraticity of the risk estimate, and derive a universal threshold, an information criterion and an oracle inequality in this canonical setting.

Keywords: adaptive lasso, information criterion, iterative block thresholding, James-Stein estimator, multivariate time series, sparse model selection, universal threshold, wavelet smoothing

*2-4 rue du Lièvre, CP 64, 1211 Genève 4, Switzerland; sylvain.sardy@unige.ch

1 Introduction

Assuming a simple Gaussian model $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\alpha}, I)$ with $\boldsymbol{\alpha} \in \mathbb{R}^{\mathrm{P}}$, the James and Stein (1961) estimator $\hat{\boldsymbol{\alpha}}^{\mathrm{JS}} = c\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}$ with $c = 1 - (P-2)/\|\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}\|_2^2$ proved that the maximum likelihood estimate $\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}$ is not admissible when $P > 2$, since James Stein's mean squared error is smaller for all coefficients. This gave birth to a class of shrinkage or thresholding estimators of the form

$$\hat{\boldsymbol{\alpha}} = \eta_\lambda(\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}), \tag{1}$$

where λ controls the regularization. *Shrinkage* means that $\|\hat{\boldsymbol{\alpha}}\| \leq \|\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}\|$, and *thresholding* means that entries of $\hat{\boldsymbol{\alpha}}$ are set to zero to achieve variable selection. When applied coordinatewise to $\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}$, thresholding sets some entries of $\hat{\boldsymbol{\alpha}}$ to zero, and when applied blockwise, then all entries are set to zero at once. The original James-Stein estimator neither shrink nor threshold, but its truncated version $\hat{\boldsymbol{\alpha}}^{\mathrm{JS}+} = c_+\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}}$ does both blockwise by taking the positive part (i.e., $c_+ = \max(c, 0)$) of the multiplicative factor. Waveshrink (Donoho and Johnstone, 1994) is a famous example of coordinatewise thresholding for wavelet smoothing.

Consider now generalized linear models (Nelder and Wedderburn, 1972) with observed response y_n and P corresponding covariates $\tilde{\boldsymbol{x}}_n = (\tilde{x}_{n,1}, \ldots, \tilde{x}_{n,P})$ organized in a matrix $\tilde{\mathbf{X}}$ for $n = 1, \ldots, N$, negative log-likelihood $-l$ (including a possible link function) and coefficients $\boldsymbol{\alpha}$. In the following, covariates have been mean-centered and Σ-rescaled into a matrix $\mathbf{X}$ with corresponding coefficients $\boldsymbol{\beta}$, so that $\hat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ is homoscedastic in the rescaled basis (Sardy, 2008). In that general regression setting, another class of regularization defines the estimate as a minimizer to a penalized likelihood function,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} -l(\mathbf{X}\boldsymbol{\beta}; \mathbf{y}) + \lambda\|\boldsymbol{\beta}\|, \tag{2}$$

where $\|\cdot\|$ is a norm or a semi-norm, and λ is the regularization parameter. Famous examples of such methods are ridge regression (Hoerl and Kennard, 1970), nonparametric smoothing splines (Wahba, 1990), waveshrink, nonnegative garrote (Breiman, 1995) or lasso (Tibshirani, 1996). The last three are variable selection estimators. There exist exact links between (1) and (2), for example waveshrink.

One goal of this paper is to achieve variable selection with a new class of estimators called *smooth blockwise iterative thresholding estimators (SBITE)*, that iteratively apply a new thresholding function called *smooth James-Stein*, which is governed by a thresholding parameter λ, a shrinkage parameter ν and a smoothness parameter s. We will see this class of estimators can minimize penalized likelihood functions (2) by iteratively applying smooth James-Stein thresholding for (λ, ν, s) set to specific values. In that sense iterative thresholding encompasses existing variable selection methods of the type (1) and (2) as particular cases. Iterative thresholding goes beyond these methods by adding flexibility, stability,

smoothness and uniqueness properties. Iterative thresholding can be done coordinatewise or blockwise. As far as flexibility is concerned, recent estimators are governed by several regularization parameters, for example, bridge (Fu, 1998), SCAD (Fan and Li, 2001a), the penalized least squares estimator of Antoniadis and Fan (2001), EBayesThresh (Johnstone and Silverman, 2004, 2005), fused lasso (Tibshirani et al., 2005), elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006) and ℓ_ν-regularization (Sardy, 2009). It is often believed that, although these estimators are more flexible, their risk may be worse in some situations because selection of more than one regularization parameter is unstable. SBITE owes its stability (Breiman, 1996) to its smoothness like ridge regression. Moreover a smooth estimation of its risk allows a more stable selection of its regularization parameters. As far as uniqueness and smoothness are concerned, we will see that smooth James-Stein iterative thresholding smoothly and uniquely extends lasso, which otherwise is not smooth and not necessarily uniquely defined.

This paper is organized as follows. Section 2 presents our original motivation, the detection of gravitational wave bursts using information from several simultaneously recorded time series. It motivates the need for a wavelet-based smoother that thresholds blocks of multiresolution coefficients across captors. Section 3 presents the new estimator in the generalized linear regression setting, discusses its links to existing estimators, presents uniqueness and smoothness properties, and derives its Stein unbiased risk estimate. A Monte-Carlo experiment investigates its finite sample properties. Section 4 focuses on block canonical regression, where we study tempering of the erratic behavior of the Stein unbiased risk estimate by means of the smoothness coefficient of SBITE, and derive a universal threshold, an information criterion and an oracle inequality. Finally the estimator is applied to gravitational wave bursts and simulated data. Section 5 discusses two extensions.

2 Motivation

Gravitational wave bursts are produced by energetic cosmic phenomena such as the collapse of a supernova. They are rare, highly oscillating and of small intensity compared to the instrumental noise, so only the concomitant recording by Q captors (typically $Q = 3$) of the cosmic phenomena may help prove the existence of such wave bursts. The captors are located far apart from each other to avoid recording local earth phenomena (such as an earthquake) on all Q captors. Another difficulty is the non-white nature of the noise, possibly non-Gaussian. The measurements are recorded at a high frequency of 5 KHz: one minute of recording has $3Q \cdot 10^5$ noisy measurements. A good model for these data is

$$\tilde{S}_t^{(q)} = \mu^{(q)}(t) + \tilde{\epsilon}_t^{(q)}, \quad t = 1, \dots, T, \ q = 1 \dots, Q \tag{3}$$

where the noises $\tilde{\epsilon}^{(q)}$ and $\tilde{\epsilon}^{(q')}$ are independent between captors $q \neq q'$, but where the noise is temporally correlated for a given captor. Importantly, most of the

time the underlying signal $\mu^{(q)}(t) = 0$ for all q, and, if $\mu^{(q)}(t) \neq 0$ for a given time t and captor q, then the same is true for all the other captors. Finally because the incoming wave burst may not hit the captors with the same angle, we may not have $\mu^{(q)}(t) = \mu^{(q')}(t)$, but only a proportionality constant relates them.

Like Klimenko and Mitselmakher (2004), we assume each underlying signal $\mu^{(q)}$ expands on $T = N$ orthonormal approximation ϕ and fine scale ψ wavelets:

$$\mu^{(q)}(t) = \sum_{k=0}^{2^{j_0}-1} \beta_{0k}^{(q)} \phi_{j_0,k}(t) + \sum_{j=j_0}^{J} \sum_{n=0}^{N_j-1} \beta_{j,n}^{(q)} \psi_{j,n}(t), \tag{4}$$

where the wavelets are obtained by dilation j and translation n, $J = \log_2(N)$ and $N_j = 2^j$; see Donoho and Johnstone (1994). One can extract an orthonormal regression matrix $\mathbf{W} = [\mathbf{\Phi}_0 \mathbf{\Psi}_{j_0} \dots \mathbf{\Psi}_J]$ from this representation such that (3) becomes $\tilde{\boldsymbol{S}}^{(q)} = \mathbf{W}\boldsymbol{\beta}^{(q)} + \tilde{\boldsymbol{\varepsilon}}^{(q)}$. Applying the orthonormal wavelet decomposition $\mathbf{W}^{\mathrm{T}}$, the model can also be written as $\boldsymbol{S}^{(q)} = \boldsymbol{\beta}^{(q)} + \boldsymbol{\varepsilon}^{(q)}$, where $\boldsymbol{S}^{(q)} = \mathbf{W}^{\mathrm{T}}\tilde{\boldsymbol{S}}^{(q)}$ and $\boldsymbol{\varepsilon}^{(q)} = \mathbf{W}^{\mathrm{T}}\tilde{\boldsymbol{\varepsilon}}^{(q)}$. This latter model is interesting in three aspects. First Johnstone

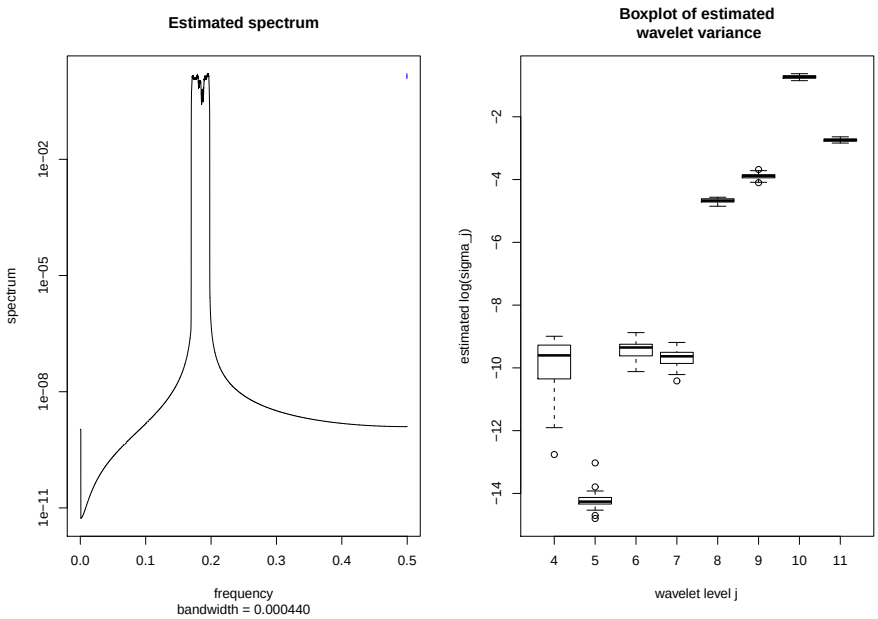

Figure 1: EDA from 26 seconds recording: (left) estimated spectrum and (right) boxplots of estimated wavelet variances on a log-scale for level 4 to 11.

and Silverman (1997) show that stationary correlated noise $\tilde{\boldsymbol{\varepsilon}}_q$ is well decorrelated by a wavelet transform within and between levels. And for a given level j, the nearly white noise process has its own wavelet variance $\sigma_j^{2,(q)}$ (Percival, 1995; Serroukh et al., 2000) that can be estimated from the data (Donoho and Johnstone, 1995). The right plot of Figure 1 represents boxplots of such wavelet variances

estimated from 31 disjoint signals of length $N = 4096$ for $j = 4, \ldots, 11$. The left plot of Figure 1 shows an estimated spectrum from 26 seconds of recording: the noise of the captor is a band-pass filtered colored noise process. Second, the data are Gaussianized by the linear wavelet transformation. Third, owing to sparse wavelet representation, the vectors $\boldsymbol{\beta}^{(q)}$ are sparse, and the amount of sparsity varies between levels, so the selection of hyperparameters should be level dependent. Hence organizing the coefficients of captor q by levels, model (3) and (4) can be well approximated at a given level j by

$$\boldsymbol{Y}_j^{(q)} = \boldsymbol{\alpha}_j^{(q)} + \boldsymbol{z}_j^{(q)} \quad \text{with} \quad \boldsymbol{Y}_j^{(q)} = \boldsymbol{S}_j^{(q)}/\sigma_j^{(q)} \quad \text{and} \quad \boldsymbol{\alpha}_j^{(q)} = \boldsymbol{\beta}_j^{(q)}/\sigma_j^{(q)}, \tag{5}$$

where $\boldsymbol{\alpha}_j^{(q)} = (\alpha_{j,1}^{(q)}, \ldots, \alpha_{j,N_j}^{(q)})$ is a sparse vector and $\boldsymbol{z}_j^{(q)} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$. Importantly, given a dilation j and a translation n, then $\boldsymbol{\alpha}_{j,n} = (\alpha_{j,n}^{(1)}, \ldots, \alpha_{j,n}^{(Q)})$ is either null or, when $\boldsymbol{\alpha}_{j,n} \neq \mathbf{0}$, then its entries are different. Our goal is to derive an estimator that adapts levelwise to block sparsity.

3 Smooth blockwise iterative thresholding

3.1 Review of block coordinate relaxation

Recent estimators consider the situation where the coefficients are blocked into J groups $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ of respective sizes $p_1, \ldots, p_J$ with $\sum_{j=1}^J p_j = P$. Correspondingly, let $\mathbf{X} = [\mathbf{X}_1 \ldots \mathbf{X}_J]$. For instance, for gravitational wave burst detection, wavelet coefficients are grouped into blocks of size Q, the number of captors, and $\mathbf{X}_j$ is the $Q \times Q$ identity matrix for all $j = 1, \ldots, J$.

We recall an optimization technique upon which we elaborate a new estimator in the following section. Suppose for now we want to calculate $\hat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ solution to (2) for $\lambda = 0$. Block coordinate relaxation (BCR) works as follows: start with any initial guess $\boldsymbol{\beta}$, choose a block $j \in \{1, \ldots, J\}$ and update only the jth block $\boldsymbol{\beta}_j$ conditional on the values of the other blocks $\boldsymbol{\beta}_i$ for all $i \neq j$, that is

$$\boldsymbol{\beta}_j^{|\mathrm{MLE}} = \arg\min_{\boldsymbol{\beta}_j \in \mathcal{B}_j} -l(\sum_{i \neq j} \mathbf{X}_i \boldsymbol{\beta}_i + \mathbf{X}_j \boldsymbol{\beta}_j; \mathbf{y}), \tag{6}$$

and leave the other unchanged to obtain the next iterate

$$\boldsymbol{\beta}^{(j)} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{j-1}, \quad \boldsymbol{\beta}_j^{|\mathrm{MLE}} \quad , \boldsymbol{\beta}_{j+1}, \ldots, \boldsymbol{\beta}_J). \tag{7}$$

Note that j is the index of the updated block, but not of the iteration.

Property 1: Assuming that $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathcal{B} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_J$ is a product of closed convex sets, that (6) has a unique solution and that the log-likelihood is continuously differentiable, then the algorithm converges to a stationary point (Bertsekas, 1999, Proposition 2.7.1). If the negative log-likelihood is also strictly convex, then the algorithm finds the MLE.

Property 2: After updating $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^{(j)}$ according to (7), the gradient of the likelihood with respect to the jth block is null, that is $\nabla_{\boldsymbol{\beta}_j} l(\mathbf{X}\boldsymbol{\beta}^{(j)}; \mathbf{y}) = \mathbf{0}$.

3.2 Smooth blockwise iterative thresholding estimator

The MLE does not achieve variable selection however. To do so, one can test the significance of the jth block based on the likelihood and its gradient in the following way. Suppose we are at the MLE where the entire gradient vector is null. The covariates have been Σ-rescaled for all MLE coefficients to have unit variance as discussed in Section 1. So if after thresholding the jth block of the MLE to zero the gradient with respect to the jth block remains small (compared to a threshold λ), then we declare this block not significant. This suggests calculating the likelihood's block gradient at each BCR iteration:

- at the current iterate $\boldsymbol{\beta}^{(j)}$ defined by (7). According to Property 2, the gradient with respect to the jth block is null;
- at the current iterate with its jth block values thresholded to $\mathbf{0}$, namely at

$$\boldsymbol{\beta}^{(j)\to 0} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{j-1}, \quad \mathbf{0} \quad , \boldsymbol{\beta}_{j+1}, \ldots, \boldsymbol{\beta}_J). \tag{8}$$

  The gradient with respect to the jth block is $\nabla_{\boldsymbol{\beta}_j} l(\mathbf{X}\boldsymbol{\beta}^{(j)\to 0}; \mathbf{y})$.

A difference larger than a threshold λ between the two likelihood's block gradient norms, that is, $\|\nabla_{\boldsymbol{\beta}_j} l(\mathbf{X}\boldsymbol{\beta}^{(j)\to 0}; \mathbf{y})\| \geq \lambda$, shows that the jth block is significant given the value of the other blocks. This leads to the following estimator.

Smooth block iterative thresholding estimator (SBITE) (algorithmic definition). Choose a threshold $\lambda \geq 0$, a shrinkage parameter $\nu \geq 1$ and a smoothness parameter $s \geq 1$. Let $\tilde{\boldsymbol{\beta}}^*$ be a root-N-consistent estimate of $\boldsymbol{\beta}$.

1. Start with any initial value;
2. Choose a block j, and calculate $\boldsymbol{\beta}_j^{|\mathrm{MLE}}$ according to (6) and the gradient $\nabla_{\boldsymbol{\beta}_j} l(\mathbf{X}\boldsymbol{\beta}^{(j)\to 0}; \mathbf{y})$;
3. Update the jth block according to

$$\boldsymbol{\beta}_j^{\mathrm{update}} = (1 - \frac{\lambda^\nu}{\|\tilde{\boldsymbol{\beta}}_j^*\|^{\nu-1} \|\nabla_{\boldsymbol{\beta}_j} l(\mathbf{X}\boldsymbol{\beta}^{(j)\to 0}; \mathbf{y})\|})_+^s \boldsymbol{\beta}_j^{|\mathrm{MLE}}; \tag{9}$$

4. Go back to step 2 until convergence.

The thresholding function (9) is called *smooth James-Stein*: when $\|\tilde{\boldsymbol{\beta}}_j^*\|^{\nu-1}$ and $\|\nabla_{\boldsymbol{\beta}_j} l(\mathbf{X}\boldsymbol{\beta}^{(j)\to 0}; \mathbf{y})\|$ are small, thresholding sets the jth block to zero. We study the advantage of the smoothness parameter s later. For $s = 1$ and certain values of ν, SBITE is linked to existing estimators, as established in the following property.

Property 3 (Gaussian case with a smoothness parameter $s = 1$): The SBITE iterations converge at the limit to the estimate of:

1. lasso (Tibshirani, 1996) for $\nu = 1$ and blocks of size one. Lasso is not an oracle procedure (Fan and Li, 2001b).

2. group lasso (Bakin, 1999; Yuan and Lin, 2006) for $\nu = 1$, with blocks.

3. adaptive lasso, which is oracle (Zou, 2006), for $\nu > 1$ and blocks of size one. Adaptive lasso is the motivation for including the norm $\|\tilde{\boldsymbol{\beta}}_j^*\|$ in (9). Hence not only a larger gradient but also a larger root-N-consistent estimate of the jth block leads to milder shrinkage.

4. waveshrink for an orthonormal wavelet matrix $\mathbf{X}$: soft-waveshrink for $\nu = 1$ and hard-waveshrink when $\nu \to \infty$.

5. truncated James-Stein for $\nu = 2$, $\lambda = \sqrt{P-2}$ and a block of size $P > 2$.

6. block thresholding for wavelet smoothing for $\nu = 2$, groups of size $L > 1$ and an orthonormal wavelet matrix $\mathbf{X}$ (Cai, 1999).

For a proof, observe that the SBITE algorithm corresponds to the shooting algorithm (Fu, 1998) for lasso, the BCR algorithm (Sardy et al., 2000) for basis pursuit (Chen et al., 1999) and to the iterative algorithm of Yuan and Lin (2006) to minimize penalized least squares problems of the form

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{J} \frac{1}{\|\tilde{\boldsymbol{\beta}}_j^*\|^{\nu-1}} \|\boldsymbol{\beta}_j\|_2. \tag{10}$$

Note that the last three estimators above (4, 5, 6) converge after one iteration, and (10) can also be solved by another class of iterative algorithms developed for inverse problems (Daubechies et al., 2004).

Hence SBITE provides a new interpretation of lasso as a sequence of tests based on the block gradient of the likelihood evaluated at successive null hypothesis $H_0 : \boldsymbol{\beta}_j = \mathbf{0}$ given the values of the other coefficients, until an equilibrium is reached. A legitimate question addressed in the following section is whether such an equilibrium can be reached at a unique point. If so, SBITE is defined uniquely.

3.3 Uniqueness

Lasso ($s = 1$) does not necessarily define a unique estimate if the kernel of the regression matrix $\mathbf{X}$ is not the $\mathbf{0}$ singleton (Sardy, 2009). On the contrary SBITE is uniquely defined under a milder condition when $s > 1$, as stated in Theorem 1 below. We first give its fixed point definition.

Smooth block iterative thresholding estimator (SBITE) (fixed point definition). Choose a threshold $\lambda \geq 0$, a shrinkage parameter $\nu \geq 1$ and a smoothness parameter $s \geq 1$. Let $\tilde{\boldsymbol{\beta}}^*$ be a root-N-consistent estimate of $\boldsymbol{\beta}$. SBITE is a fixed point

to the SBITE algorithm, which is the solution to the set of P nonlinear equations

$$\boldsymbol{\beta}_j = (1 - \frac{\lambda^\nu}{\|\tilde{\boldsymbol{\beta}}_j^*\|^{\nu-1}\|\nabla_{\boldsymbol{\beta}_j} l(\mathbf{X}\boldsymbol{\beta}^{(j)\to 0}; \mathbf{y})\|})_+^s \boldsymbol{\beta}_j^{|\mathrm{MLE}}, \quad j = 1, \ldots, J, \tag{11}$$

where $\boldsymbol{\beta}_j^{|\mathrm{MLE}}$ is defined by (6), and each $\boldsymbol{\beta}_j$ is a vector of length p_j with $P = \sum_{j=1}^J p_j$.

Despite being highly non-linear and employing a non-convex thresholding function, these equations define SBITE uniquely in the Gaussian case when $s > 1$.

Theorem 1: Let $\mathbf{X} = [\mathbf{X}_1 \ldots \mathbf{X}_J]$ be a matrix such that $\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j = I_{p_j}$ for all $j = 1, \ldots, J$. Then the solution to (11) with $s > 1$ is uniquely defined for the Gaussian likelihood. It is moreover continuously differentiable with respect to the data.

Note that the condition is milder than $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ being positive definite in terms of colinearity, but requires blockwise orthonormalization like group lasso. Convergence of the SBITE algorithm is proved when $s = 1$ for the Gaussian likelihood (Fu, 1998; Sardy et al., 2000) and for more general likelihoods (Sardy and Tseng, 2004). Convergence to the unique fixed point has always been observed when $s > 1$, but remains to be proved.

3.4 Equivalent degrees of freedom

SBITE is governed by two regularization parameters λ and ν, and a smoothness parameter s. For $s = 1$ and for the Gaussian linear model with mean $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\alpha}$ and P variables grouped into blocks of unit size, Zou (2006) selects the two regularization parameters λ and ν of adaptive lasso by cross-validation, a rule known for its high computational cost and instability. This section derives instead the Stein unbiased risk estimate for any combination of the three parameters (λ, ν, s).

Stein (1981) showed that for an estimator of the form $\hat{\boldsymbol{\mu}} = g(\mathbf{Y}) + \mathbf{Y}$ and unit variance, then the quadratic risk can be estimated unbiasedly if g is almost differentiable. For SBITE with blocks of unit size, we have $g(\mathbf{Y}; \lambda, \nu, s) = \bar{\mathbf{Y}}\mathbf{1} + \mathbf{X}\hat{\boldsymbol{\beta}}_{\lambda,\nu;s}(\mathbf{Y}) - \mathbf{Y}$, where $\hat{\boldsymbol{\beta}}_{\lambda,\nu;s}(\mathbf{Y})$ is the solution to (11) for Gaussian likelihood. SBITE is almost differentiable for $s = 1$ and differentiable for $s > 1$, so the Stein unbiased risk estimate (SURE) for SBITE is

$$\mathrm{SURE}(\lambda, \nu; s) = \mathrm{RSS}(\hat{\boldsymbol{\mu}}_{\lambda,\nu,s}) + N + 2\sum_{n=1}^N \partial g_n(\mathbf{Y}; \lambda, \nu, s)/\partial Y_n, \tag{12}$$

where the last term is the so-called equivalent degrees-of-freedom. SURE involves the partial derivatives $\partial g_n(\mathbf{Y}; \lambda, \nu, s)/\partial Y_n = 1/N + \mathbf{x}_n^{\mathrm{row}} \cdot \nabla_n \hat{\boldsymbol{\beta}}_{\lambda,\nu,s}(\mathbf{Y}) - 1$, for $n = 1, \ldots, N$, where $\nabla_n \hat{\boldsymbol{\beta}}_{\lambda,\nu,s}(\mathbf{Y})$ are the derivatives of $\hat{\boldsymbol{\beta}}_{\lambda,\nu,s}(\mathbf{Y})$ with respect to Y_n, and $\mathbf{x}_n^{\mathrm{row}}$ is the nth row of $\mathbf{X}$. The following theorem states that $\nabla_n \hat{\boldsymbol{\beta}}_{\lambda,\nu,s}(\mathbf{Y})$ is

explicitly defined as solution to a full rank system of linear equations when $s > 1$. Hence the risk of SBITE can be estimated unbiasedly for all regression matrix.

Theorem 2: The gradient of the SBITE estimate $\hat{\boldsymbol{\beta}}_{\lambda,\nu,s}(\mathbf{Y})$ with respect to Y_n is the solution to a system of linear equations (26) that is full rank when $s > 1$, regardless of the existence of a kernel for $\mathbf{X}$, for $n = 1, \ldots, N$.

Interestingly also from Theorem 2, letting the smoothness parameter s tend to one leads to the equivalent degrees of freedom of adaptive lasso; if moreover $\nu = 1$, then the solution to (26) is $\mathbf{h}_n^{\bar{\mathcal{I}}_0} = ((\mathbf{X}^{\bar{\mathcal{I}}_0})^{\mathrm{T}}\mathbf{X}^{\bar{\mathcal{I}}_0})^{-1}(\mathbf{x}_n^{\mathrm{row}})^{\mathrm{T}}$. Hence, we see that

$$\begin{aligned}\sum_{n=1}^{N} \partial g_n(\mathbf{Y};\lambda,\nu,s)/\partial Y_n &= 1 + \sum_{n=1}^{N} \mathbf{x}_n^{\mathrm{row}} \cdot ((\mathbf{X}^{\bar{\mathcal{I}}_0})^{\mathrm{T}}\mathbf{X}^{\bar{\mathcal{I}}_0})^{-1}(\mathbf{x}_n^{\mathrm{row}})^{\mathrm{T}} - N \\ &= 1 + \mathrm{trace}(((\mathbf{X}^{\bar{\mathcal{I}}_0})^{\mathrm{T}}\mathbf{X}^{\bar{\mathcal{I}}_0})^{-1}((\mathbf{X}^{\bar{\mathcal{I}}_0})^{\mathrm{T}}\mathbf{X}^{\bar{\mathcal{I}}_0})) - N \\ &= 1 + |\bar{\mathcal{I}}_0| - N,\end{aligned}$$

where $|\bar{\mathcal{I}}_0|$ is lasso's degrees of freedom previously found by Zou et al. (2007).

The smoothness parameter should not be considered as a third regularization parameter like λ and ν, but more like a device to bring smoothness to the estimator and combat the increasing erraticity of the two-dimensional SURE function as ν grows. Section 4.1 quantifies SURE's erraticity tempered with the smoothness parameter s. The smoothness parameter should not be too large however, since it contradicts the goal of a large ν to approach hard thresholding, and since the constant of the oracle inequality (21) of Theorem 4 increases with s. A good trade-off is for instance $s(\nu) = 2\log\nu + 1$ (see Theorem 3 below). Figure 2 illustrates the gain in smoothness by calculating SURE for the prostate cancer data with $P = 8$ covariates (Tibshirani, 1996), and by comparing the smoothness of the estimated risk either with adaptive lasso (left) or with its smooth extension (right). Both risk estimates are unbiased, but the second is less erratic thanks to $s > 1$.

We have assumed unit variance. In practice, one can either estimate the variance and rescale responses to have approximate unit variance, or, in the spirit of generalized cross validation (Golub et al., 1979), one can define generalized SURE

$$\mathrm{GSURE}(\lambda,\nu;s) = \frac{\mathrm{RSS}(\hat{\boldsymbol{\mu}}_{\lambda,\nu;s})/N}{(1 - \frac{1}{N}\sum_{n=1}^{N} \partial g_n(\mathbf{Y};\lambda,\nu,s)/\partial Y_n)^2}. \tag{13}$$

To minimize SURE or GSURE over (λ,ν) for $s = 2\log\nu + 1$, our strategy consists in minimizing it first for $s = 1$ (i.e., adaptive lasso) which can be done efficiently on a fine grid thanks to lars (Efron et al., 2004). This provides a neighborhood for λ, namely $[\hat{\lambda}^{(s=1)}/10, 10\hat{\lambda}^{(s=1)}]$, over which we then minimize SURE on a more local grid for smooth adaptive lasso ($s > 1$) calculated with the SBITE algorithm. This strategy provides a good and efficient selection of the pair of hyperparameters (λ,ν), as demonstrated by Monte-Carlo below.

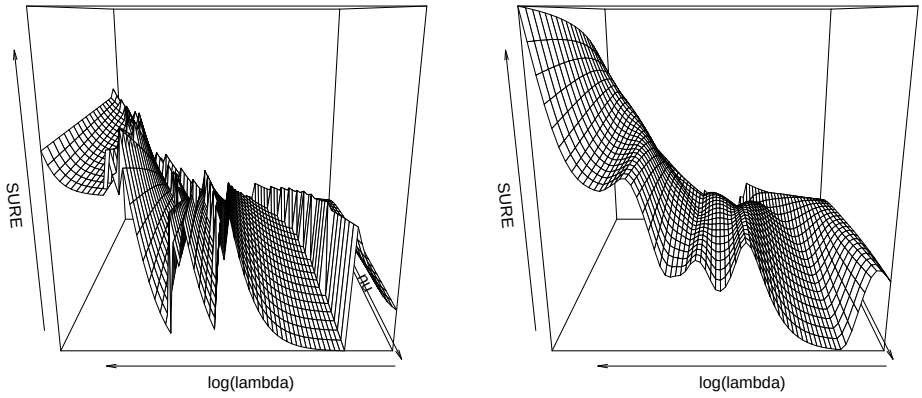

Figure 2: Prostate cancer data: Stein unbiased risk estimate as a function of λ and ν for adaptive lasso (left, $s = 1$) and smooth adaptive lasso (right, $s = 2\log\nu + 1$).

3.5 Monte-Carlo simulation

We replicate the Monte-Carlo simulation of Zou (2006) with $P = 8$ covariates with corresponding coefficients either sparse $\boldsymbol{\alpha} = (3, 1.5, 0, 0, 2, 0, 0, 0)$ for Model 1 with $N \in \{20, 60\}$, or not sparse $\boldsymbol{\alpha} = (.85, .85, .85, .85, .85, .85, .85, .85)$ for Model 2 with $N \in \{40, 80\}$. The covariates $\tilde{\mathbf{x}}_n$ are i.i.d. Gaussian vectors with pairwise correlation between $\tilde{x}_{n,i}$ and $\tilde{x}_{n,j}$ given by $\mathrm{cor}(i, j) = (.5)^{|i-j|}$. The noise is Gaussian with standard error $\sigma \in \{1, 3, 6\}$. Like Zou (2006), we consider the relative prediction error $\mathrm{RPE} = \mathrm{E}[\tilde{\mathbf{x}}_{\mathrm{test}}^{\mathrm{T}}\{\hat{\boldsymbol{\alpha}}([\tilde{\mathbf{X}}, \mathbf{Y}]_{\mathrm{training}}) - \boldsymbol{\alpha}\}]/\sigma^2$, where the expectation is taken over training and test sets, and response. Note that we exactly calculate RPE given the training set by using knowledge of the distribution of the covariates (Zou relies on 10,000 test observations instead). This predictive measure is reported in Table 1 to compare lasso, adaptive lasso and smooth adaptive lasso. To compare estimators fairly, we consider the same selection rule for all, here two-fold cross-validation. Since the estimators are used on the same 100 training sets, then the numbers we see in the tables reveal significant differences, even though marginal standard errors are large. We observe that SBITE improves significantly over lasso and adaptive lasso when the underlying model is sparse and the noise is small; the estimation is slightly worse for the non-sparse model.

We also consider SURE as a selection rule for lasso and its smooth adaptive version SBITE, that we compare based on their relative prediction error conditional on the covariates of the training set, namely $\mathrm{RPE} \mid \tilde{\mathbf{X}} = \sum_{n=1}^{N} \mathrm{E}[\tilde{\mathbf{x}}_{\mathrm{training},n}^{\mathrm{T}}\{\hat{\boldsymbol{\alpha}}([\tilde{\mathbf{X}}, \mathbf{Y}]_{\mathrm{training}}) - \boldsymbol{\alpha}\}]/\sigma^2$, where the expectation is taken over the response only. Although less useful in practice unless prediction is sought at the same locations as the training set, this measure helps quantify the improvement of using smooth James-Stein thresh-

olding. Table 2 reports the results, which shows a systematic gain of SBITE over adaptive lasso. Lasso is often better for the non-sparse model.

Finally Table 3 reports the number C of selected nonzero components and the number I of zero components incorrectly selected. We observe that the selection is correct when the noise is small ($\sigma = 1$) with SBITE and adaptive lasso using SURE, but that false detection grows with noise.

4 Block canonical regression

We consider block canonical regression, that is when the regression matrix is the identity and the coefficients are organized in blocks of size Q, namely,

$$\mathbf{Y}_n = \boldsymbol{\alpha}_n + \boldsymbol{\epsilon}_n \quad \text{with} \quad \mathbf{Y}_n = \begin{pmatrix} Y_n^{(1)} \\ \vdots \\ Y_n^{(Q)} \end{pmatrix}, \ \boldsymbol{\alpha}_n = \begin{pmatrix} \alpha_n^{(1)} \\ \vdots \\ \alpha_n^{(Q)} \end{pmatrix} \text{ and } \boldsymbol{\epsilon}_n = \begin{pmatrix} \epsilon_n^{(1)} \\ \vdots \\ \epsilon_n^{(Q)} \end{pmatrix} \quad (14)$$

for $n = 1, \ldots, N$, where the noise is i.i.d. Gaussian (independent between and within sequences). If the standard deviation is not known, then Donoho and Johnstone (1994) proposed an efficient estimate based on the median absolute deviation; another possibility is to use generalized SURE (13). This setting applies to the gravitational wave bursts detection problem (5) of Section 2 with Q captors. Since $\mathbf{X}$ is the identity, rescaling has no impact, that is $\tilde{\mathbf{X}} = \mathbf{X}$ and $\boldsymbol{\beta} = \boldsymbol{\alpha}$.

Block sparsity assumes most blocks $\boldsymbol{\alpha}_n$ are the $\mathbf{0}$-vector. SBITE (11) achieves block sparsity and has the closed form expression

$$(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s} \quad = \quad (1 - \frac{\lambda^\nu}{\|\mathbf{Y}_n\|_2^\nu})_+^s \mathbf{Y}_n, \quad n = 1, \ldots, N. \quad (15)$$

in this canonical setting.

Table 1: Zou's Monte-Carlo simulation with 100 training sets: median RPE using two-fold cross-validation to select the hyperparameter(s).

	$\sigma = 1$		$\sigma = 3$		$\sigma = 6$	
Model 1	$N = 20$	$N = 60$	$N = 20$	$N = 60$	$N = 20$	$N = 60$
lasso	$0.367_{(0.048)}$	$0.089_{(0.009)}$	$0.419_{(0.069)}$	$0.089_{(0.008)}$	$0.369_{(0.021)}$	$0.096_{(0.010)}$
adaptive lasso	$0.360_{(0.051)}$	$0.052_{(0.009)}$	$0.435_{(0.057)}$	$0.085_{(0.009)}$	$0.308_{(0.021)}$	$0.097_{(0.011)}$
SBITE	$0.328_{(0.046)}$	$0.054_{(0.009)}$	$0.424_{(0.056)}$	$0.085_{(0.009)}$	$0.330_{(0.020)}$	$0.098_{(0.010)}$
Model 2	$N = 40$	$N = 80$	$N = 40$	$N = 80$	$N = 40$	$N = 80$
lasso	$0.238_{(0.014)}$	$0.104_{(0.005)}$	$0.231_{(0.020)}$	$0.108_{(0.005)}$	$0.163_{(0.010)}$	$0.087_{(0.005)}$
adaptive lasso	$0.238_{(0.015)}$	$0.104_{(0.005)}$	$0.233_{(0.021)}$	$0.108_{(0.005)}$	$0.181_{(0.010)}$	$0.091_{(0.005)}$
SBITE	$0.238_{(0.015)}$	$0.104_{(0.005)}$	$0.240_{(0.021)}$	$0.109_{(0.005)}$	$0.172_{(0.010)}$	$0.090_{(0.005)}$

Table 2: Zou's Monte-Carlo simulation with 100 training sets: median RPE | $\mathbf{X}$ at the training covariates $\mathbf{X}$ using SURE to select the hyperparameter(s).

	$\sigma = 1$		$\sigma = 3$		$\sigma = 6$	
Model 1	$N = 20$	$N = 60$	$N = 20$	$N = 60$	$N = 20$	$N = 60$
lasso	$0.264_{(0.021)}$	$0.082_{(0.008)}$	$0.258_{(0.020)}$	$0.082_{(0.008)}$	$0.219_{(0.017)}$	$0.082_{(0.007)}$
adaptive lasso	$0.231_{(0.023)}$	$0.075_{(0.008)}$	$0.280_{(0.023)}$	$0.090_{(0.008)}$	$0.289_{(0.019)}$	$0.096_{(0.007)}$
SBITE	$0.228_{(0.023)}$	$0.065_{(0.008)}$	$0.279_{(0.025)}$	$0.084_{(0.008)}$	$0.255_{(0.019)}$	$0.097_{(0.007)}$
Model 2	$N = 40$	$N = 80$	$N = 40$	$N = 80$	$N = 40$	$N = 80$
lasso	$0.187_{(0.010)}$	$0.092_{(0.004)}$	$0.187_{(0.010)}$	$0.090_{(0.004)}$	$0.137_{(0.008)}$	$0.075_{(0.003)}$
adaptive lasso	$0.191_{(0.011)}$	$0.092_{(0.004)}$	$0.237_{(0.013)}$	$0.127_{(0.006)}$	$0.172_{(0.009)}$	$0.105_{(0.004)}$
SBITE	$0.191_{(0.011)}$	$0.092_{(0.004)}$	$0.219_{(0.013)}$	$0.113_{(0.006)}$	$0.164_{(0.008)}$	$0.099_{(0.004)}$

Table 3: Median number of Selected Variables for Model 1 with $n = 60$

	$\sigma = 1$		$\sigma = 3$	
	C	I	C	I
Truth	3	0	3	0
Lasso†	3	2	3	2
Adaptive lasso†	3	1	3	1
SCAD†	3	0	3	1
Garotte†	3	1	3	1.5
Adaptive lasso SURE	3	0	4	1
SBITE SURE	3	0	5	2

† Results taken from the Monte-Carlo simulation of Zou (2006).

4.1 Total variation of SURE

The Stein unbiased risk estimate also has the closed form expression $\text{SURE}(\lambda, \nu, s) = \sum_{n=1}^{N} \hat{\rho}_n((\lambda, \nu, s), \boldsymbol{\alpha}_n)$ with

$$\hat{\rho}_n((\lambda, \nu, s), \boldsymbol{\alpha}_n) = \{1 - (1 - \frac{\lambda^\nu}{\|\mathbf{Y}_n\|_2^\nu})_+^s\}^2 \|\mathbf{Y}_n\|_2^2 - Q + 2 \sum_{q=1}^{Q} \frac{\partial(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s}}{\partial Y_n^{(q)}},$$

where

$$\frac{\partial(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s}}{\partial Y_n^{(q)}} = \begin{cases} 0 & \text{if } \|\mathbf{Y}_n\|_2 < \lambda \\ (1 - \frac{\lambda^\nu}{\|\mathbf{Y}_n\|_2^\nu})^{s-1} (\nu s \lambda^\nu \frac{(Y_n^{(q)})^2}{\|\mathbf{Y}_n\|_2^{\nu+2}} + 1 - \frac{\lambda^\nu}{\|\mathbf{Y}_n\|_2^\nu}) & \text{if } \|\mathbf{Y}_n\|_2 \geq \lambda \end{cases}. \quad (16)$$

For thresholding functions employing no smoothness, that is $s = 1$ here, SURE has N discontinuity points as a function of λ for a fixed ν. Indeed (16) is discontinuous at $\lambda = \|\mathbf{Y}_n\|_2$ for all $n = 1, \ldots, N$ when $s = 1$, and the size of each jump is equal to 2ν. We had already observed on the left graph of Figure 2 that the larger ν the more erratic the SURE surface for the prostate cancer data when $s = 1$. There are two negative consequences for the selection of λ and ν. First the SURE two-dimensional surface will have minima that will be difficult to localize from an optimization point-of-view. Second, the location of the global minima will be sensitive, in particular with large ν, to changes in the data $\mathbf{Y}_n$.

Donoho and Johnstone (1995) studied the asymptotic properties of selecting λ by minimizing SURE for $\nu = s = 1$. To show their *SureShrink* estimator is optimally smoothness adaptive, a key ingredient is the deviation of SURE around its mean when $\nu = 1$. Theorem 3 below shows that the total variation of SURE

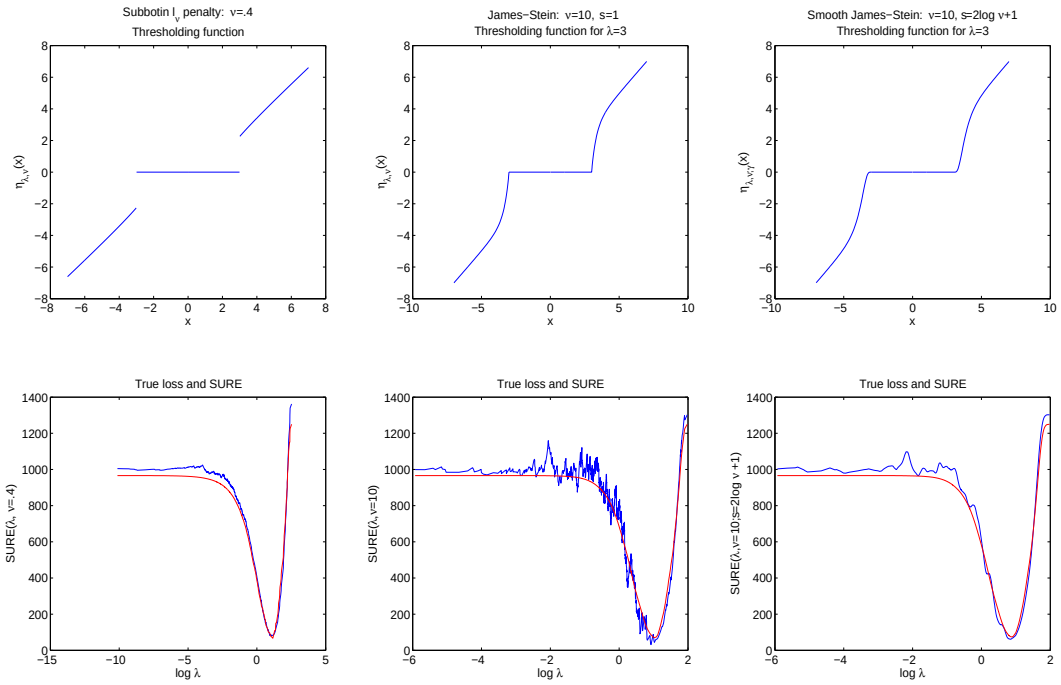

Figure 3: Thresholding functions and corresponding SURE for $Q = 1$ and $\nu = 10$ on simulated data of length $N = 1000$. Left: Subbotin ℓ_ν penalized least squares; Middle: James-Stein ($s = 1$); Right: smooth James-Stein ($s = 2\log\nu + 1$). Top: thresholding functions with parameters chosen to approximate hard thresholding. The left one is discontinuous at the threshold, but with a small slope at the threshold; the middle one has a discontinuous derivative at the threshold; and the right one has a smooth change of derivative at the threshold. Bottom: corresponding Stein unbiased risk estimate (least smooth curve) and true loss (smoothest curve).

grows when ν increases, and that employing smooth James-Stein thresholding with a smoothness parameter larger than one tempers this erratic effect by removing the jumps and decreasing the erraticity of SURE. Figure 3 illustrates the advantage of increasing s when ν gets large on simulated data. We observe that while the two thresholding functions for $s = 1$ and $s > 1$ only differ slightly, the latter is smoother near the threshold value and the corresponding SURE curve is less wiggly around the true loss. Section 4.5 reports results of a Monte-Carlo simulation that quantifies the improvement in mean squared error obtained by adding smoothness.

A measure of erraticity of SURE that is defined not only for $s > 1$ but also for $s = 1$ is its total variation as a function of λ. The total variation of a function f in

the space of functions of bounded variation (that is, not necessarily continuous) is $\mathrm{TV}(f) = \sup \sum_j |f(\lambda_{j+1}) - f(\lambda_j)|$, where the supremum is taken over all possible partitions $[\lambda_j, \lambda_{j+1}]$, $j = 1, \ldots, M$, of the domain of f. (If f is moreover absolutely continuous, then TV reduces to the more conventional smoothness measure $\mathrm{TV}(f) = \int_\Lambda |f'(\lambda)| d\lambda$.) The following theorem quantifies the erraticity of SURE and shows the tempering effect of the smoothness parameter s.

Theorem 3: Consider $\mathrm{SURE}(\lambda; \nu, s) = \sum_{n=1}^N \hat{\rho}_n((\lambda; \nu, s), \boldsymbol{\alpha}_n)$ as a function of λ for $Q = 1$. Its total variation for a given $\nu \geq 1$ satisfies

$$\mathrm{TV}^{(s=1)}(\mathrm{SURE}) = \sum_{n=1}^N \mathrm{TV}^{(s=1)}(\hat{\rho}_n) > \sum_{n=1}^N \mathrm{TV}^{(s>1)}(\hat{\rho}_n) \geq \mathrm{TV}^{(s>1)}(\mathrm{SURE}).$$

Moreover erraticity increases less with $s > 1$ when ν or $|Y_n|$ grows since $\frac{\partial}{\partial \nu} \mathrm{TV}^{(s>1)}(\hat{\rho}_n) \leq \frac{\partial}{\partial \nu} \mathrm{TV}^{(s=1)}(\hat{\rho}_n)$ and $\frac{\partial}{\partial Y_n} \mathrm{TV}^{(s>1)}(\hat{\rho}_n) \leq \frac{\partial}{\partial Y_n} \mathrm{TV}^{(s=1)}(\hat{\rho}_n)$ for $Y_n \geq 0$. In particular when $Y_n \to 0$ and for ν large, then $\frac{\partial}{\partial \nu} \mathrm{TV}^{(s>1)}(\hat{\rho}_n) \to 4(1 - 1/s)^{s-1} \geq 4 \exp(-1)$ for s fixed. Letting s grow slowly with ν, for instance $s(\nu) = 2 \log \nu + 1$, then the lower bound $4 \exp(-1)$ is reached to lower erraticity most.

4.2 Universal threshold and information criterion

To derive a universal threshold (Donoho and Johnstone, 1994) and an information criterion for SBITE, we approximate below the distribution of the smallest threshold $\lambda_{\mathcal{Y}}$ that, for a sample $\mathcal{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ of size N, sets to zero all N blocks of length Q when the true underlying model is made of zero vectors. Controlling the maximum of $\lambda_{\mathcal{Y}}$ leads to a finite sample $\tilde{\lambda}_{N,Q}$ and asymptotic $\lambda_{N,Q}$ universal thresholds, and a prior distribution π_λ for λ.

Assuming $\mathbf{Y}_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}_Q(\mathbf{0}, \mathbf{I}_Q)$ for $n = 1, \ldots, N$, we seek the smallest threshold $\lambda_{N,Q}$ such that SBITE estimates the right model with a probability tending to one:

$$\mathrm{P}((\hat{\boldsymbol{\alpha}}_1)_{\lambda_{N,Q},\nu;s} = \mathbf{0}, \ldots, (\hat{\boldsymbol{\alpha}}_N)_{\lambda_{N,Q},\nu;s} = \mathbf{0}) = \mathrm{P}(\max_{n=1,\ldots,N} \|\mathbf{Y}_n\|_2^2 \leq \lambda_{N,Q}^2) \overset{N \to \infty}{\longrightarrow} 1. \tag{17}$$

The distribution of $M_N = \max_{n=1}^N \|\mathbf{Y}_n\|_2^2$, where $\|\mathbf{Y}_n\|_2^2 \overset{\text{i.i.d.}}{\sim} \chi_Q^2 = \Gamma(Q/2, 1/2)$, is degenerate. Extreme value theory provides proper rescaling of M_N for $c_N^{-1}(M_N - d_N(Q)) \longrightarrow_d G_0(x)$, where $G_0(x) = \exp(-\exp(-x))$ is the Gumbel distribution, $c_N = 2$ and $d_N(Q)$ is the root in ξ to

$$\log N - \log \Gamma(Q/2) = (1 - Q/2) \log(\xi/2) + \xi/2. \tag{18}$$

The normalizing constant $d_N(Q) = 2(\log N + (Q/2 - 1) \log \log N - \log \Gamma(Q/2))$ given by Embrechts et al. (1997, p.156) for the Gamma distribution is the asymptotic root of (18), which provides a good Gumbel approximation when N is large compared

to $\Gamma(Q/2)$. In that case we define the asymptotic universal threshold $\lambda_{N,Q} = \sqrt{2(\log N + (Q/2)\log\log N - \log\Gamma(Q/2))}$, for which (17) is satisfied since

$$\mathrm{P}(\max_{n=1,\dots,N} \|\mathbf{Y}_n\|_2^2 \leq \lambda_{N,Q}^2) \doteq G_0(\log\log N) \approx 1 - 1/\log N \overset{N\to\infty}{\longrightarrow} 1. \tag{19}$$

Note that we get back the standard universal threshold $\sqrt{2\log N}$ up to a small term for $Q = 1$, and the universal threshold of Sardy (2000) for denoising complex-valued signals for $Q = 2$. When Q gets large, the proposed normalizing constant $d_N(Q)$ is too far from the exact root to provide a useful approximation, so we find the root $d_N(Q)$ of (18) numerically. The finite sample universal threshold is then defined as

$$\tilde{\lambda}_{N,Q} = \sqrt{d_N(Q) + c_N \log\log N} \quad \text{with} \quad d_N(Q) \text{ root of (18)} \tag{20}$$

to have the same rate of convergence for all Q with $\tilde{\lambda}_{N,Q}$ in place $\lambda_{N,Q}$ in (19).

More than a bound the asymptotic Gumbel pivot for M_N leads to a prior distribution $F_\lambda(\lambda) = G_0((\lambda^2 - d_N(Q))/2)$ of the threshold λ to reconstruct true zero vectors from noisy measurements. When $s = 1$, Bayes theorem provides the joint posterior distribution of the coefficients and the hyperparameters. Taking its negative logarithm leads to the following information criterion in the spirit of the sparsity ℓ_ν information criterion $\mathrm{SL}_\nu\mathrm{IC}$ (Sardy, 2009).

Definition. Suppose model (14) or model (3.1) of Cai (1999) holds. The sparsity weighted ℓ_2 information criterion for the estimation of $(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_N)$ and the selection of (λ, ν) with SBITE (15) for $s = 1$ is defined as

$$\begin{aligned} \mathrm{SL}_2^w\mathrm{IC}(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_N, \lambda, \nu) \;=\; & \frac{1}{2}\sum_{n=1}^N \|\mathbf{Y}_n - \boldsymbol{\alpha}_n\|_2^2 + \lambda^\nu \sum_{n=1}^N \frac{1}{\|\mathbf{Y}_n\|_2^{\nu-1}} \|\boldsymbol{\alpha}_n\|_2 \\ & - N\log\Big(\frac{\Gamma(Q/2)}{2\pi^{Q/2}\Gamma(Q)}\Big) + Q(\nu-1)\sum_{n=1}^N \log\|\mathbf{Y}_n\|_2 \\ & - QN\nu\log\lambda - \log\pi_\lambda(\lambda;\tau_{N,Q}) - \log\pi_\nu(\nu), \end{aligned}$$

where π_ν is a prior for ν that we choose Uniform on $[1,\infty)$, $\pi(\lambda;\tau) = F'(\lambda;\tau)$ with $F_\lambda(\lambda;\tau) = G_0((\lambda^2/\tau^2 - d_N(Q))/2)$ and τ is calibrated to $\tau_{N,Q}^2 = \tilde{\lambda}_{N,Q}^2/(QN\nu + 1)$ to match the asymptotic model consistency when $\boldsymbol{\alpha}_n = \mathbf{0}$ for $n = 1, \dots, N$.

In practice, one minimizes $\mathrm{SL}_2^w\mathrm{IC}$ like AIC or BIC to both select the hyperparameters (λ, ν) and estimate the sequences $\boldsymbol{\alpha}_n$ for $n = 1, \dots, N$. The information criterion could also be derived for $s > 1$ if we knew the definition of SBITE as a penalized least squares, which is an open problem.

4.3 Oracle inequality

Candès (2005) provides an interesting review on oracle inequalities. Here we derive an oracle inequality for SBITE employing smooth James-Stein thresholding when the block size $Q \geq 2$ is fixed. Cai (1999) derived an oracle inequality for block sizes increasing with the sample size. The ℓ_2 risk for model (14) is $\mathrm{R}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) = \sum_{n=1}^{N} \rho_n(\boldsymbol{\alpha}_n) = \sum_{n=1}^{N} \mathrm{E}\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_n\|_2^2$. Following Donoho and Johnstone (1994), the oracle predictive performance of the block diagonal projection estimator $\hat{\boldsymbol{\alpha}}_n = \delta_n \mathbf{Y}_n$, where $\delta_n \in \{0, 1\}$ is

$$\rho_n(\delta_n, \boldsymbol{\alpha}_n) = \begin{cases} \|\boldsymbol{\alpha}_n\|_2^2, & \text{if } \delta_n = 0, \\ Q, & \text{if } \delta_n = 1. \end{cases}$$

Hence the oracle hyperparameters are $\delta_n^* = 1_{\{\|\boldsymbol{\alpha}_n\|_2^2 > Q\}}$ for $n = 1, \ldots, N$, and the corresponding oracle overall risk is $\mathrm{R}^*(\boldsymbol{\delta}, \boldsymbol{\alpha}) = \sum_{n=1}^{N} \min(\|\boldsymbol{\alpha}_n\|^2, Q)$. The following theorem extends the oracle inequality obtained by Donoho and Johnstone (1994) and Zou (2006) for $Q = s = 1$ to block thresholding with $Q \geq 2$ and $s > 1$.

Theorem 4: For any fixed $Q \geq 2$, there exists a sample size N_0 such that, for all $N \geq N_0$ and with the universal threshold $\tilde{\lambda}_{N,Q}$ defined in (20), then SBITE defined by (15) for $\nu \geq 1$ and $s \geq 1$ achieves the oracle inequality

$$\mathrm{R}(\hat{\boldsymbol{\alpha}}^{\mathrm{SBITE}}_{\tilde{\lambda}_{N,Q},\nu;s}, \boldsymbol{\alpha}) \leq (Q + 1 + 2\nu s + c_{\nu,s,Q}\lambda_{N,Q}^2)(Q + \mathrm{R}^*(\boldsymbol{\delta}, \boldsymbol{\alpha})), \tag{21}$$

where $c_{\nu,s,Q} = \max(1 + \frac{\nu s}{Q}, s^2)$ and $\lambda_{N,Q}^2 = 2\log N + Q\log\log N - 2\log\Gamma(Q/2)$.

This result shows we can mimic the overall oracle risk achieved with N oracle hyperparameters within a factor of essentially $\lambda_{N,Q}^2$ with the single hyperparameter $\tilde{\lambda}_{N,Q}$. The smallest sample size N_0 for which the inequality holds is quite small in practice; more work is needed to get a tight expression. Note that for $s = 1$, the inequality differs from Zou (2006) which had the ν-term in the denominator (this seems to be due to an error in $d\hat{\mu}_i^*(\lambda)/dy_i$ right above (A.13) p. 1427). This result shows that increasing ν or s increases the oracle inequality constant. But this does not prevent the estimator with $\nu > 1$ to be oracle (Zou, 2006), which is not true for lasso with $\nu = 1$. Likewise using a larger s improves predictive performance in practice although the oracle inequality constant increases.

4.4 Application to wave burst detection and estimation

We employ SBITE blockwise and levelwise to $Q = 3$ concomitant time series of length $T = 2^{14}$ (about 3.27 seconds of recording) to detect gravitational wave bursts, as described in Section 2. Taking $J = 4$ in the wavelet expansion (4), SBITE has a total of 22 hyperparameters to select (11 levels with two hyperparameters λ and ν each). Data are pure electronic noise, so we add three proportional so-called "injections" at time $t = 1500$, to mimic a wave burst.

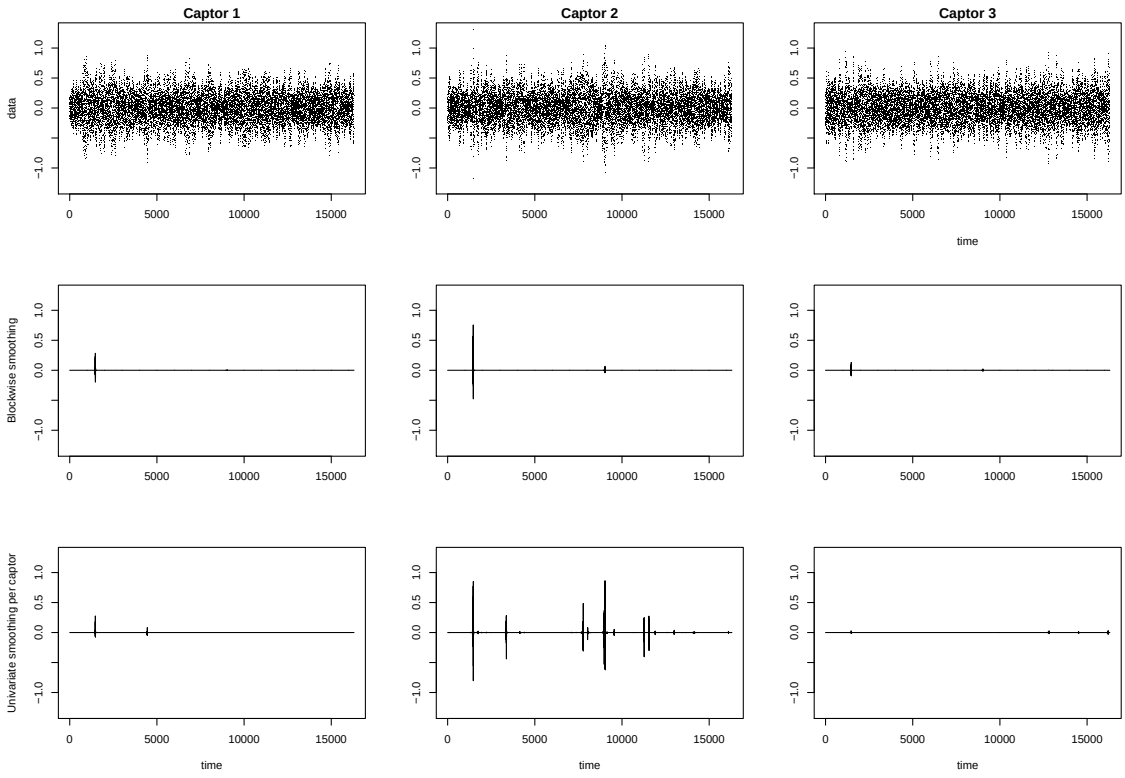

Figure 4: A wave burst injection at time $t = 1500$ added on three real captors noise (columnwise). Top: Time series of length $T = 2^{14}$. Middle: SBITE employing SURE levelwise and blockwise. Bottom: univariate smoothing per captor levelwise.

Figure 4 shows the data (first line) for each captor (columnwise), the SBITE estimate (second line) and estimates employing a univariate smoothing per captor (third line). We observe that, as opposed to univariate smoothing, blockwise smoothing detects the injection and has no false detection except right before time $t = 10'000$. Figure 5 zooms around the injections that are three times larger on the second captor, and five times smaller on the third captor. We see that blockwise estimation of the injections is better than coordinatewise.

4.5 Monte-Carlo simulation

We reproduce the Monte-Carlo simulation of Johnstone and Silverman (2004) for $Q = 1$ captor to estimate a sparse sequence of length $N = 1000$ and of varying degrees of sparsity, as measured by the number of nonzero terms taken in $\{5, 50, 500\}$ and by the value of the nonzero terms μ taken in $\{3, 4, 5, 7\}$. Table 4 reports estimated risks of four estimators: SBITE with ($s > 1$) and without ($s = 1$) smooth-

ness, Subbotin ℓ_ν penalized likelihood (Sardy, 2009) and EBayesThresh (Johnstone and Silverman, 2004). The results clearly show the superiority of SBITE with SURE thanks to more smoothness. In case of extreme sparsity, only SBITE using the $\mathrm{SL}_2^w\mathrm{IC}$ information criterion and EBayesThresh perform better; this drawback of SURE has been explained by Donoho and Johnstone (1995, Section 2.4).

We also perform a Monte-Carlo simulation with $Q = 3$ concomitantly observed sequences: all three underlying sequences are identical in the location of the non zero entries, but not in their amplitude. Since three sequences carry more information than a single one, we may hope to distinguish noise from a signal with a smaller signal-to-noise ratio, so we consider value of the nonzero terms fixed to $\mu_1 = 1$, $\mu_2 = 2$ and $\mu_3 = \mu$ taken in $\{3, 4, 5, 7\}$. The estimated risks (divided by Q to allow some comparison with $Q = 1$) are reported in Table 4. SBITE with smoothness ($s > 1$) again performs best overall, while the $\mathrm{SL}_2^w\mathrm{IC}$ selection rule performs better than for $Q = 1$, even more so with very sparse sequences.

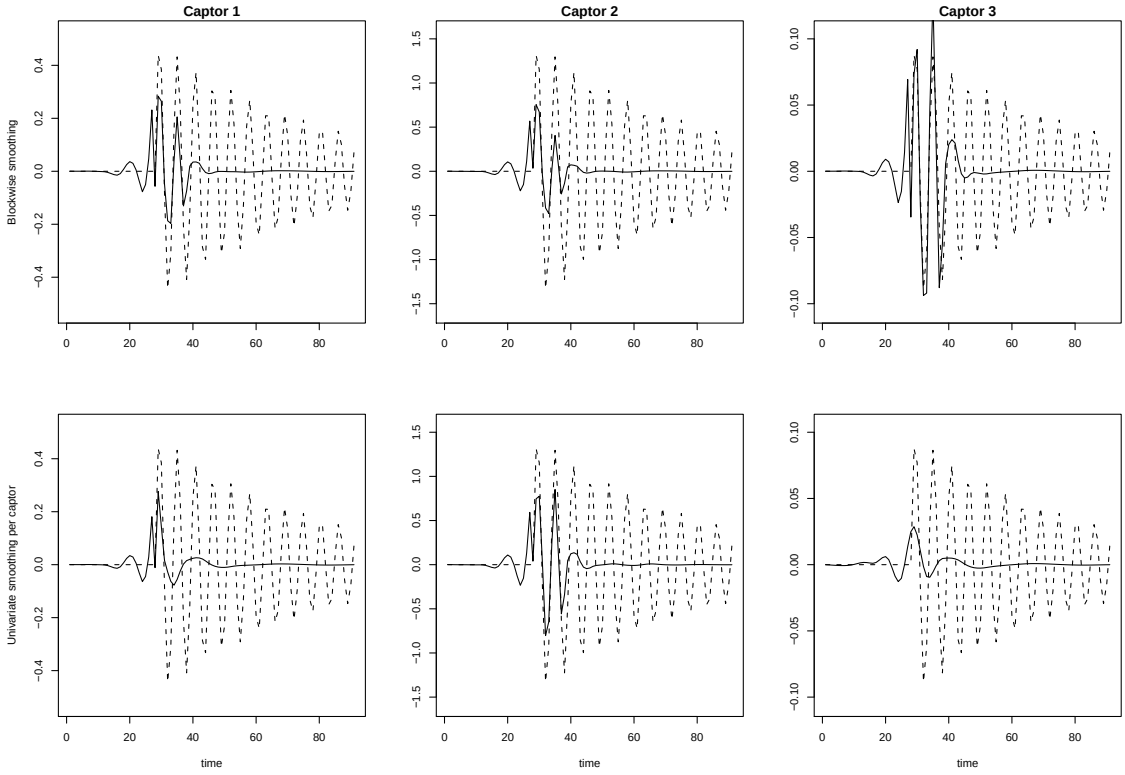

Figure 5: Zooming on the "injection" represented by the dotted line (the y-scales differs between captors). The estimates are plotted with a continuous line: blockwise smoothing (first line), and a univariate smoothing per captor (second line).

Table 4: Monte-Carlo simulation for a sequence of length $N = 1000$. Average total squared loss of: SBITE using smooth SURE; SBITE with smoothness parameter fixed to $s = 1$ using SURE or $\mathrm{SL}_2^w\mathrm{IC}$; the Subbotin$(\lambda, \nu)$ posterior mode estimator using SURE or $\mathrm{SL}_\nu\mathrm{IC}$; and the EBayesThresh estimator with Cauchy-like prior.

Number nonzero	5				50				500			
Value nonzero $\mu =$	3	4	5	7	3	4	5	7	3	4	5	7
$Q = 1$ captor												
SBITE $s > 1$												
SURE	37	37	26	15	202	165	109	65	826	752	614	521
SBITE $s = 1$												
SURE	45	42	31	24	213	174	119	77	848	760	624	551
$\mathrm{SL}_2^w\mathrm{IC}$	39	41	23	6	380	389	213	54	3350	2688	1532	532
Subbotin												
SURE	39	35	27	25	232	167	107	97	1239	794	607	533
$\mathrm{SL}_\nu\mathrm{IC}$	38	36	19	9	356	296	132	59	849	831	839	859
EBayesThresh	37	36	19	8	268	177	104	77	924	899	831	743
$Q = 3$ captors												
SBITE $s > 1$	18	17	14	8.7	106	94	75	56	615	615	558	510
SBITE $s = 1$												
SURE	22	21	18	16	110	98	80	63	612	613	563	516
$\mathrm{SL}_2^w\mathrm{IC}$	19	18	11	5.7	181	168	109	52	1600	1350	807	548

5 Further extensions

Smooth James-Stein thresholding (15) relies on the ℓ_2 norm, like most block thresholding we are aware of. This measure may not be appropriate in certain applications. Indeed if one wants to measure departure from the zero vector in a sense that all entries must be different from zero, then $\|\mathbf{Y}_n\|_2$ in (15) should be replaced by $\min_{q=1,\dots,Q} |Y_n^{(q)}|$ leading to robust SBITE:

$$(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s} = (1 - \frac{\lambda^\nu}{\min_{q=1,\dots,Q} |Y_n^{(q)}|^\nu})_+^s \mathbf{Y}_n;$$

a simple calculation leads to $\lambda_{N,Q} = \sqrt{2/Q \log N}$ for its corresponding universal threshold. Other quantiles, or a norm like the ℓ_1 norm, could also be considered. Importantly also, the use of a common threshold λ for each block implicitly assumes blocks of equal size; if not, then the threshold $\lambda(p_j)$ must grow with block size p_j.

For the group lasso, Yuan and Lin (2006) provided an approximate degrees of freedom. One can instead follow the derivation of Theorem 2 to derive the exact one, as we did in Section 3.4 for adaptive lasso. Consider the smooth generalization of adaptive group lasso (10), defined as solution to

$$\hat{\boldsymbol{\beta}}_j(\mathbf{Y}) \;=\; (c_j)_+^s \mathbf{s}_j \quad \text{with} \quad \left\{ \begin{array}{l} c_j = 1 - \lambda^\nu / (\|\tilde{\boldsymbol{\beta}}_j^*\|_2^{\nu-1} \|\mathbf{s}_j\|_2) \\ \mathbf{s}_j = \mathbf{X}_j^{\mathrm{T}} \mathbf{Y} - \sum_{k \neq j} \mathbf{X}_j^T \mathbf{X}_k \hat{\boldsymbol{\beta}}_k \end{array} \right. \; j = 1, \dots, J, \quad (22)$$

where the coefficients vector is segmented as $\boldsymbol{\beta} = (\boldsymbol{\beta}, \dots, \boldsymbol{\beta}_J) \in \mathbb{R}^{\mathrm{P}}$ in J blocks of respective size p_j such that $\sum_{j=1}^J p_j = P$. Note that (22) is the smooth extension of Yuan and Lin (2006, (2.4)) taking $\mathbf{K}_j = \mathbf{I}_{p_j}$ using their notation. Deriving SURE

in that setting requires calculating the nonzero elements of the block gradient $\nabla_n \hat{\boldsymbol{\beta}}(\mathbf{Y})$ of the estimated vector defined by (22) with respect to the data Y_n for $n = 1, \ldots, N$. Following similar derivation as in Appendix B, one finds they are defined by a set of linear equations of the form (26), that has a unique solution when $s > 1$. Hence the exact equivalent degrees of freedom of smooth adaptive group lasso can be calculated, and in particular for adaptive lasso by letting the smoothness parameter tend to one.

6 Conclusions

We developed SBITE variable selection defined as the fixed point of an iterative sequence employing the smooth James-Stein thresholding function. SBITE can be employed blockwise or coordinatewise, and can control sparsity, shrinkage and smoothness by means of three parameters. For any combination of these three parameters, we have derived the Stein unbiased risk estimate that is smoother the larger s for a better selection of the regularization parameters. Letting the smoothness parameter tend to one, we obtained the equivalent degrees of freedom of lasso, adaptive lasso and group lasso. For block canonical regression, we derived a universal rule, an information criterion and an oracle inequality. The estimator is promising for gravitational wave burst detection and estimation: we are currently conducting an analysis with physicists on several months of recordings to quantify type I and type II errors, as well as false discovery rate. Also, in the spirit of Park and Hastie (2007), generalized linear models could be regularized via smooth James-Stein thresholding. More generally, SBITE can be employed in other settings than regression to provide both sparsity and smoothness.

7 Acknowledgements

I would like to thank C. Giacobino, L. Lang and Y. Velenik for helpful discussions, and S. Foffa, R. Terenzi and the ROG group for providing a sample of the astrophysics data. The associate editor and two anonymous referees helped improve the quality of the paper. Partially supported through Swiss National Science Foundation.

A Proof of theorem 1

Let $s > 1$ and $\nu \geq 1$. To fix notation, let $b_j = \|\tilde{\boldsymbol{\beta}}_j^*\|^{\nu-1} > 0$, $\mathbf{C}_{j,k} = \mathbf{X}_j^T \mathbf{X}_k$ and $\mathbf{r}_j = \mathbf{X}_j^{\mathrm{T}} \mathbf{Y} - \sum_{k \neq j} \mathbf{C}_{j,k} \boldsymbol{\beta}_k$ for all blocks $j = 1, \ldots, J$.

For the Gaussian likelihood, (11) is given by $\boldsymbol{\beta}_j = \{1 - \frac{\lambda^\nu}{b_j \|\mathbf{r}_j\|}\}_+^s \mathbf{C}_{j,j}^{-1} \mathbf{r}_j$, $j = 1, \ldots, J$. Let $F : \mathbb{R}^{\mathrm{P}} \to \mathbb{R}^{\mathrm{P}}$ defined by $F = (f_1, \ldots, f_J)$ with $f_j(\boldsymbol{\beta}) = \boldsymbol{\beta}_j - \{1 -$

$\frac{\lambda^\nu}{b_j\|\mathbf{r}_j\|}\}_+^s \mathbf{C}_{j,j}^{-1}\mathbf{r}_j$, where $f_j : \mathbb{R}^{\mathrm{P}} \to \mathbb{R}^{\mathrm{p_j}}$, $j = 1, \ldots, J$. F being differentiable, we can apply the fundamental global univalence theorem of Gale and Nikaido (1965) which states that F is globally univalent on $\mathbb{R}^{\mathrm{P}}$ provided its Jacobian matrix $\mathbf{J}(\boldsymbol{\beta})$ is a P-matrix for every $\boldsymbol{\beta} \in \mathbb{R}^{\mathrm{P}}$. Global univalence implies that the $\mathbf{0}$-vector has at most one preimage, i.e., the estimate $\hat{\boldsymbol{\beta}}^{\mathrm{SBITE}} = F^{-1}(\mathbf{0})$ is unique.

Recall that a real square matrix is a P-matrix if all of its principal minors are positive. To prove the Jacobian matrix $\mathbf{J}(\boldsymbol{\beta})$ is a P-matrix, let us determine its entries. Clearly $\mathbf{J}(\boldsymbol{\beta})$ has ones on its diagonal since $\mathbf{r}_j$ does not depend on $\boldsymbol{\beta}_j$. To compute the other entries, consider any point $\boldsymbol{\beta} \in \mathbb{R}^{\mathrm{P}}$, and let $\mathcal{I}_0$ be the set of indices j for which the inequality $b_j\|\mathbf{r}_j\| \leq \lambda^\nu$ is true; let $p_0 = \sum_{j\in\mathcal{I}_0} p_j$ and $j_0 = |\mathcal{I}_0|$. Permuting variables if necessary, the satisfied inequalities are for $j = 1, \ldots, j_0$. Hence, the first p_0 lines of the Jacobian matrix at $\boldsymbol{\beta}$ are the $p_0 \times P$ matrix $[\mathbf{I}_{p_0}\ \mathbf{0}]$, where $\mathbf{0}$ is a matrix of zeros. For every remaining block $j \in \{j_0+1, \ldots, J\}$, the Jacobian matrix is a block matrix with blocks

$$\mathbf{J}_{j,k} = \begin{cases} \mathbf{C}_{j,j}^{-1} Q_j \mathbf{C}_{j,k}, & j \neq k \\ \mathbf{I}_{p_j} = \mathbf{C}_{j,j}^{-1}\mathbf{C}_{j,j}, & j = k \end{cases} \text{ for } k = j_0 + 1, \ldots, J,$$

where $Q_j = sc_j^{s-1}(1 - c_j)R_j + c_j^s I_{p_j}$, $R_j = \mathbf{r}_j\mathbf{r}_j^{\mathrm{T}}/\|\mathbf{r}_j\|^2$ is a projection matrix and $c_j = 1 - \frac{\lambda^\nu}{b_j\|\mathbf{r}_j\|} \in (0, 1)$. It is straightforward to show Q_j is symmetric definite positive with eigen values in $(0, 1)$. Note that this is not true when $s = 1$.
Hence the Jacobian matrix is

$$\mathbf{J}(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{I}_{p_0} & \mathbf{0} \\ \mathbf{B} & \mathbf{D}^{\bar{\mathcal{I}}_0}\mathbf{C}^{\bar{\mathcal{I}}_0} \end{pmatrix}, \tag{23}$$

where $\mathbf{B}$ is some $(P - p_0) \times p_0$ matrix, $\mathbf{D}^{\bar{\mathcal{I}}_0} = \mathrm{diag}(\mathbf{C}_{j,j}^{-1}Q_j, j = j_0 + 1, \ldots, J)$, and $\mathbf{C}^{\bar{\mathcal{I}}_0} = (\mathbf{X}^{\bar{\mathcal{I}}_0})^{\mathrm{T}}\mathbf{X}^{\bar{\mathcal{I}}_0} + \mathrm{diag}((Q_j^{-1} - I)X_j^{\mathrm{T}}X_j, j = j_0 + 1, \ldots, J)$. The assumption $X_j^{\mathrm{T}}X_j = I_{p_j}$ now implies that the latter matrix is symmetric positive definite since the block diagonal matrix has symmetric positive definite blocks.
Consequently, $|\mathbf{J}(\boldsymbol{\beta})| = |\mathbf{D}^{\bar{\mathcal{I}}_0}||\mathbf{C}^{\bar{\mathcal{I}}_0}| > 0$. Positivity is also verified for all principal minors of $\mathbf{J}(\boldsymbol{\beta})$ since they have the same structure as (23).

The Jacobian matrix being invertible, the implicit function theorem guarantees the preimage is also continuously differentiable with respect to the data.

B Proof of theorem 2

For Gaussian likelihood and for a *linear* estimate $\tilde{\boldsymbol{\beta}}^* = \mathbf{AY}$ (e.g., least squares or ridge regression) where the entries of $\mathbf{A}$ are noted a_{pn}, the solution to (11) is the system of nonlinear equations:

$$\begin{aligned} \hat{\beta}_p \quad &= \quad \{1 - \frac{\lambda^\nu}{|\tilde{\beta}_p^*|^{\nu-1}|r_p|}\}_+^s r_p/\|\mathbf{x}_p\|_2^2 \\ &\text{with} \quad r_p = \mathbf{x}_p^{\mathrm{T}}\mathbf{Y} - \sum_{q\neq p} \mathbf{x}_p^T\mathbf{x}_q\hat{\beta}_q, \quad p = 1, \ldots, P. \end{aligned} \tag{24}$$

Hence one finds

$$\frac{\partial \hat{\beta}_p(\mathbf{Y})}{\partial Y_n} = \begin{cases} 0 & \text{if } \hat{\beta}_p(\mathbf{Y}) = 0, \\ \frac{1}{\|\mathbf{x}_p\|_2^2}(v_p(x_{np} - \mathbf{x}_p^{\mathrm{T}}\mathbf{X}_{-p}\nabla_n\hat{\boldsymbol{\beta}}(\mathbf{Y})) + u_p) & \text{else,} \end{cases} \tag{25}$$

where

$$u_p = \frac{s(\nu-1)a_{pn}(w_p-1)r_p}{\tilde{\beta}_p^* w_p^s},\ v_p = \frac{1-s+sw_p}{w_p^s} \quad \text{and} \quad 1/w_p = 1 - \frac{\lambda^\nu}{|\tilde{\beta}_p^*|^{\nu-1}|r_p|}.$$

Let $\mathcal{I}_0 = \{p \in \{1,\dots,P\} : \hat{\beta}_p(\mathbf{Y}) = 0\}$, and let $\mathbf{X}^{\bar{\mathcal{I}}_0}$ be the columns of $\mathbf{X}$ with an index in $\bar{\mathcal{I}}_0$. Rewriting (25), the entries of $\nabla_n\hat{\boldsymbol{\beta}}(\mathbf{Y}) =: \mathbf{h}_n$ are

$$(\nabla_n\hat{\boldsymbol{\beta}}(\mathbf{Y}))_p = \begin{cases} 0 & p \in \mathcal{I}_0 \\ h_{n,p} & p \in \bar{\mathcal{I}}_0 \end{cases},$$

where $\mathbf{h}_n^{\bar{\mathcal{I}}_0}$ are solution to the following system of $|\bar{\mathcal{I}}_0|$ linear equations:

$$\mathbf{x}_p^{\mathrm{T}}\mathbf{X}^{\bar{\mathcal{I}}_0}\mathbf{D}_p^{\bar{\mathcal{I}}_0}\mathbf{h}_n^{\bar{\mathcal{I}}_0} = x_{n,p} + \frac{u_p}{v_p} \quad \text{for all } p \in \bar{\mathcal{I}}_0. \tag{26}$$

Here $\mathbf{D}_p^{\bar{\mathcal{I}}_0}$ is the identity matrix except that its pth diagonal element is $D_{p,p}^{\bar{\mathcal{I}}_0} = v_p^{-1}$.

The matrix of the linear system (26) is $(\mathbf{X}^{\bar{\mathcal{I}}_0})^{\mathrm{T}}\mathbf{X}^{\bar{\mathcal{I}}_0}$ which diagonal elements are multiplied by all the $D_{p,p}^{\bar{\mathcal{I}}_0} = v_p^{-1}$, $p \in \bar{\mathcal{I}}_0$. Moreover $w_p > 1$, so all the $D_{p,p}^{\bar{\mathcal{I}}_0} > 1$ since $f(w) = w^s/(1-s+sw)$ satisfies $f(1) = 1$ and $f'(w) > 0$ for $w > 1$. This guarantees existence of a solution $\mathbf{h}_n^{\bar{\mathcal{I}}_0}$ when $s = 1$ if the column of $\mathbf{X}$ are linearly independent, and when $s > 1$ regardless of any collinearity (i.e., s plays the role of a ridge parameter).

C Proof of theorem 3

For $s = 1$, each $\hat{\rho}_n$ is strictly increasing from $\lambda = 0$ to $\lambda = |Y_n|$, and is then constant after a jump of size 2ν. So $\mathrm{TV}^{(s=1)}(\hat{\rho}_n) = Y_n^2 + 4\nu - 2$ and $\mathrm{TV}^{(s=1)}(\mathrm{SURE}) = \sum_{n=1}^N \mathrm{TV}^{(s=1)}(\hat{\rho}_n)$. For $s > 1$, the triangular inequality gives $\mathrm{TV}^{(s>1)}(\mathrm{SURE}) \le \sum_{n=1}^N \mathrm{TV}^{(s>1)}(\hat{\rho}_n)$, and simple calculations lead to $\mathrm{TV}^{(s>1)}(\hat{\rho}_n) = 2\hat{\rho}_n((\tilde{\lambda}_n,\nu,s),\alpha_n) - Y_n^2$, where $\tilde{\lambda}_n$ is solution to $\tilde{x}_n = (1 - \tilde{\lambda}_n^\nu/|Y_n|^\nu)$ with $\tilde{x}_n \in (0,1)$ the unique root to

$$\frac{\partial}{\partial x}\hat{\rho}_n((x;\nu,s),\alpha_n) \propto -sY_n^2(1-x^s)x^{s-1} + (s-1)x^{s-2}(\nu s(1-x)+x) + x^{s-1}(-\nu s+1) \equiv 0$$

i.e., $Y_n^2 x(1-x^s) = \nu(s-1) + x(1-\nu s)$. Hence $\mathrm{TV}^{(s>1)}(\hat{\rho}_n) = Y_n^2 + 4\nu\tilde{x}^{s-1} - 2 - 2Y_n^2\tilde{x}^{2s} \le \mathrm{TV}^{(s=1)}(\hat{\rho}_n)$. Moreover $\frac{\partial}{\partial Y_n}\mathrm{TV}^{(s>1)}(\hat{\rho}_n) = 4Y_n(1-\tilde{x}_n^s)^2 - 2Y_n \le \frac{\partial}{\partial Y_n}\mathrm{TV}^{(s=1)}(\hat{\rho}_n) = 2Y_n$ for $Y_n \ge 0$, and $\frac{\partial}{\partial \nu}\mathrm{TV}^{(s>1)}(\hat{\rho}_n) = 4s\tilde{x}_n^{s-1}(1-\tilde{x}_n) \le 4(1 - 1/s)^{s-1} \le \frac{\partial}{\partial \nu}\mathrm{TV}^{(s=1)}(\hat{\rho}_n) = 4$. At the limit when Y_n tends to zero and for large ν, we have $\frac{\partial}{\partial \nu}\mathrm{TV}^{(s>1)}(\hat{\rho}_n) \doteq 4(1-\frac{1}{s})^{s-1} \ge 4\exp(-1)$ and the lower bound is reached as ν grows if for instance $s = 2\log\nu + 1$.

D Proof of theorem 4

The SBITE estimator $(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s}$ defined in (15) with Q fixed has risk

$$\begin{aligned}\rho_n((\lambda,\nu,s),\boldsymbol{\alpha}_n) &= \mathrm{E}\|(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s}-\boldsymbol{\alpha}_n\|_2^2 \\ &= Q+\mathrm{E}\|(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s}-\mathbf{Y}_n\|_2^2-2Q+2\mathrm{E}(\mathbf{Y}_n-\boldsymbol{\alpha}_n)^{\mathrm{T}}(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s} \\ &= -Q+\mathrm{E}\|(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s}-\mathbf{Y}_n\|_2^2+2\sum_{q=1}^{Q}\mathrm{E}\frac{\partial(\hat{\boldsymbol{\alpha}}_n)_{\lambda,\nu;s}}{\partial Y_n^{(q)}}, \qquad (27)\end{aligned}$$

for all $n=1,\ldots,N$, where we used Stein's lemma for the last term, and where

$$\{(\hat{\boldsymbol{\alpha}}_n^{(q)})_{\lambda,\nu;s}-Y_n^{(q)}\}^2=\begin{cases}(Y_n^{(q)})^2 & \text{if } \|\mathbf{Y}_n\|_2<\lambda \\ (Y_n^{(q)})^2\{1-(1-\frac{\lambda^\nu}{\|\mathbf{Y}_n\|_2^\nu})^s\}^2 & \text{if } \|\mathbf{Y}_n\|_2\geq\lambda\end{cases}$$

and $\partial(\hat{\boldsymbol{\alpha}}_{\lambda,\nu;s})_n/\partial Y_n^{(q)}$ is given by (16). From (27) and using the inequality $(1-(1-\epsilon)^s))^2\leq s^2\epsilon^2$ for $0\leq\epsilon\leq 1$ and $s\geq 1$, one gets two inequalities. First we have

$$\begin{aligned}\rho_n((\lambda,\nu,s),\boldsymbol{\alpha}_n) &\leq -Q+\lambda^2\mathrm{P}(\|\mathbf{Y}_n\|_2<\lambda)+s^2\lambda^2\mathrm{P}(\|\mathbf{Y}_n\|_2>\lambda)+2(\nu s+Q)\mathrm{P}(\|\mathbf{Y}_n\|_2>\lambda) \\ &\leq Q+2\nu s+s^2\lambda^2 \\ &\leq \begin{cases}(Q+2\nu s+s^2\lambda^2)(Q/N+Q) & \text{if } Q\geq 1 \\ (Q+2\nu s+s^2\lambda^2)(Q/N+\|\boldsymbol{\alpha}_n\|_2^2) & \text{if } \|\boldsymbol{\alpha}_n\|_2^2\geq 1\end{cases}. \qquad (28)\end{aligned}$$

Second, we show below that

$$\rho_n((\lambda,\nu,s),\boldsymbol{\alpha}_n)\leq(1+Q+\lambda_{N,Q}^2)(\frac{Q+\nu s}{Q})(Q/N+\|\boldsymbol{\alpha}_n\|_2^2) \quad \text{if } \|\boldsymbol{\alpha}_n\|_2^2\leq 1 \qquad (29)$$

for N large enough. So putting (28) and (29) together, and summing over all $n=1,\ldots,N$ leads to the oracle inequality (21).

To show (29) and complete the proof, note that

$$\begin{aligned}\rho_n((\lambda,\nu,s),\boldsymbol{\alpha}_n) &= \mathrm{E}\|\mathbf{Y}_n\|_2^2-Q+\sum_{q=1}^{Q}\mathrm{E}\{(Y_n^{(q)})^2[\{1-(1-\frac{\lambda^\nu}{\|\mathbf{Y}_n\|_2^\nu})^s\}^2-1]1(\|\mathbf{Y}_n\|_2>\lambda)\} \\ &\quad +2\sum_{q=1}^{Q}\mathrm{E}\{(1-\frac{\lambda^\nu}{\|\mathbf{Y}_n\|_2^\nu})^{s-1}(\nu s\lambda^\nu\frac{(Y_n^{(q)})^2}{\|\mathbf{Y}_n\|_2^{\nu+2}}+1-\frac{\lambda^\nu}{\|\mathbf{Y}_n\|_2^\nu})1(\|\mathbf{Y}_n\|_2>\lambda)\} \\ &\leq \|\boldsymbol{\alpha}_n\|_2^2+2(\nu s+Q)\mathrm{P}(\|\mathbf{Y}_n\|_2>\lambda)=:\|\boldsymbol{\alpha}_n\|_2^2+\frac{\nu s+Q}{Q}g(\mu;\lambda) \qquad (30)\end{aligned}$$

with $g(\mu;\lambda)=2Q\{1-\exp(-\mu^2/2)\sum_{j=0}^{\infty}\frac{(\mu^2/2)^j}{j!}\frac{s(j+Q/2,\lambda^2/2)}{\Gamma(j+Q/2)}\}$ since $\|\mathbf{Y}_n\|_2^2$ is non-central chi-square with Q degrees of freedom and noncentrality parameter $\mu^2=\|\boldsymbol{\alpha}_n\|_2^2<1$. Considering even Q's for simplicity, Taylor's expansion gives $g(\mu;\lambda)\leq$

$g(0;\lambda) + \mu g'(0;\lambda) + \mu^2/2 \sup_{x\in[0,1)} |g''(x;\lambda)|$. First

$$\begin{aligned} g(0;\tilde{\lambda}_{N,Q}) &= 2Q(1 - \frac{s(Q/2, \tilde{\lambda}^2_{N,Q}/2)}{\Gamma(Q/2)}) \\ &= 2Q\exp(-\tilde{\lambda}^2_{N,Q}/2) \sum_{j=0}^{Q/2-1} \frac{(\tilde{\lambda}^2_{N,Q}/2)^j}{\Gamma(j+1)} \\ &\leq 2Q\exp(-\lambda^2_{N,Q}/2) \sum_{j=0}^{Q/2-1} \frac{(\lambda^2_{N,Q}/2)^j}{\Gamma(j+1)} \\ &= \frac{2Q}{N}\frac{\Gamma(Q/2)}{(\log N)^{Q/2}}(1 + \lambda^2_{N,Q}/2 + \sum_{j=2}^{Q/2-1} \frac{(\lambda^2_{N,Q}/2)^j}{\Gamma(j+1)}), \end{aligned}$$

where the inequality stems from the fact that $\tilde{\lambda}^2_{N,Q} \geq \lambda^2_{N,Q}$ for all $Q \geq 2$ and all $N \geq N_0 = \exp(\Gamma(Q/2)^{1/(Q/2-1)})$, and $\tilde{\lambda}^2_{N,Q} \dot{\sim} \lambda^2_{N,Q}$ as $N \to \infty$. Then

$$\begin{aligned} g(0;\tilde{\lambda}_{N,Q}) &\leq \frac{Q}{N}(1)[2 + \lambda^2_{N,Q} + 2\sum_{j=2}^{Q/2-1} e(j,Q,N)] \\ &\leq \frac{Q}{N}[2 + \lambda^2_{N,Q} + 2(Q/2-2)(1)] \leq \frac{Q}{N}[Q + \lambda^2_{N,Q}], \end{aligned}$$

since $\frac{\Gamma(Q/2)}{(\log N)^{Q/2}} \leq 1$ and $e(j,Q,N) = (1 + \frac{Q/2\log\log N - \log\Gamma(Q/2)}{\log N})^j \frac{\Gamma(Q/2)}{(\log N)^{Q/2-j}\Gamma(j+1)} \leq 1$ for N large enough. Second, note that

$$g'(x;\lambda) = 2Qx\exp(-x^2/2)\exp(-\lambda^2/2)\sum_{j=0}^{\infty} \frac{(x^2/2)^j}{\Gamma(j+1)}\frac{(\lambda^2/2)^{j+Q/2}}{\Gamma(j+Q/2+1)}]$$

so $g'(0;\lambda) = 0$. Finally

$$\begin{aligned} g''(x;\lambda) &= 2Q\exp(-\lambda^2/2)\exp(-x^2/2)(1-x^2)\sum_{j=0}^{\infty} \frac{(x^2/2)^j}{\Gamma(j+1)}\frac{(\lambda^2/2)^{j+Q/2}}{\Gamma(j+Q/2+1)} \\ &\quad +2Q\exp(-\lambda^2/2)\exp(-x^2/2)x^2\sum_{j=0}^{\infty} \frac{(x^2/2)^j}{\Gamma(j+1)}\frac{(\lambda^2/2)^{j+Q/2+1}}{\Gamma(j+Q/2+2)} \\ &\leq 2Q + 2Qx^2S(\frac{\lambda^2/2}{Q/2} - 1) \leq 2Q + 2\lambda^2 \end{aligned}$$

with $S = \exp(-\lambda^2/2)\exp(-x^2/2)\sum_{j=0}^{\infty} \frac{(x^2/2)^j}{\Gamma(j+1)}\frac{(\lambda^2/2)^{j+Q/2}}{\Gamma((j+Q/2+1)} \leq 1$. The same inequality holds for $-g''(x;\lambda)$. Consequently for N larger than N_0, we have $g(\mu,\lambda) \leq Q/N(Q + \lambda^2_{N,Q}) + \mu^2/2(2Q + 2\lambda^2_{N,Q})$ for $\mu = \|\boldsymbol{\alpha}_n\|_2 < 1$.

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations (with discussion). *Journal of the American Statistical Association* **96**, 939–967.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*. Ph. D. thesis, Australian National University, Canberra.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Belmont, MA: Athena Scientific.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24**, 2350–2383.

Cai, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *The Annals of Statistics 27*(3), 898–924.

Candès, E. (2005). Modern statistical estimation via oracle inequalities. *Acta Numerica* **15**, 257–325.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20(1)**, 33–61.

Daubechies, I., Defrise, M., and Mol, C. D. (2004). A iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics* **57**, 1413–1457.

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

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

Embrechts, P., Kluppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events: For Insurance and Finance*. Springer-Verlag Inc.

Fan, J. and Li, R. (2001a). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and Li, R. (2001b). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Propoerties. *Journal of the American Statistical Association* **96**, 1348–1360.

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

Gale, D. and Nikaido, Y. (1965). The Jacobian matrix and global univalence of mappings. *Mathematischen Annalen* **159**, 81–93.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

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

James, W. and Stein, C. (1961). Estimation with quadratic loss. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 361–379. University of California Press.

Johnstone, I. M. and Silverman, B. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* **32**, 1594–1649.

Johnstone, I. M. and Silverman, B. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics* **33**, 1700–1752.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B: Methodological* **59**, 319–351.

Klimenko, S. and Mitselmakher, G. (2004). A wavelet method for detection of gravitational wave bursts. *Classical and Quantum Gravity* **21**, 1819–1830.

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

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 69*(4), 659–677.

Percival, D. P. (1995). On estimation of the wavelet variance. *Biometrika* **82**, 619–631.

Sardy, S. (2000). Minimax threshold for denoising complex signals with Waveshrink. *IEEE Transactions on Signal Processing 48*(4), 1023–1028.

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

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

Sardy, S., Bruce, A. G., and Tseng, P. (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of Computational and Graphical Statistics* **9**, 361–379.

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

Serroukh, A., Walden, A. T., and Percival, D. B. (2000). Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. *Journal of the American Statistical Association 95*(449), 184–196.

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

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

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

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

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

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

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

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