Generalized and smooth James-Stein model selection

SARDY, Sylvain

Abstract

The generalized and smooth James-Stein thresholding functions link and extend the thresholding functions employed by the James-Stein estimator, the block- and adaptive-lasso in variable selection, and the soft-, hard- and block-thresholding in wavelet smoothing. The estimator is indexed by two hyperparameters for more flexibility and a smoothness parameter for better estimation of its $l_2$-risk with the Stein unbiased risk estimate (SURE). For blocks of a fixed size, a situation that arises when observing concomitant signals (e.g., gravitational wave bursts), we derive a universal threshold, an information criterion and an oracle inequality for block thresholding. Smooth James-Stein thresholding can also be employed in parametric regression for variable selection. In that case a unique smooth estimate is defined, its smooth SURE is derived, which provides the equivalent degrees of freedom of adaptive lasso as a side result. The new estimator enjoys smoothness like ridge regression and performs variable selection like lasso.

Reference

SARDY, Sylvain. Generalized and smooth James-Stein model selection.

Available at:
http://archive-ouverte.unige.ch/unige:12040

Disclaimer: layout of this document may differ from the published version.
Generalized and smooth James-Stein model selection

SYLVAIN SARDY*

Department of Mathematics, University of Geneva

The generalized and smooth James-Stein thresholding functions link and extend the thresholding functions employed by the James-Stein estimator, the block- and adaptive-lasso in variable selection, and the soft-, hard- and block-thresholding in wavelet smoothing. The estimator is indexed by two hyperparameters for more flexibility and a smoothness parameter for better estimation of its $\ell_2$-risk with the Stein unbiased risk estimate (SURE). For blocks of a fixed size, a situation that arises when observing concomitant signals (e.g., gravitational wave bursts), we derive a universal threshold, an information criterion and an oracle inequality for block thresholding.

Smooth James-Stein thresholding can also be employed in parametric regression for variable selection. In that case a unique smooth estimate is defined, its smooth SURE is derived, which provides the equivalent degrees of freedom of adaptive lasso as a side result. The new estimator enjoys smoothness like ridge regression and performs variable selection like lasso.

Keywords: information criterion, James-Stein estimator, sparse model selection, Stein unbiased risk estimate, universal threshold, wavelet smoothing

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

James and Stein (1961) showed the surprising result that for estimating a vector $\alpha \in \mathbb{R}^P$ with $P > 2$ from a Gaussian measurement $Y \sim N(\alpha, I)$, the maximum likelihood estimate (MLE) is not admissible with respect to the $\ell_2$-risk, $R(\hat{\alpha}, \alpha) = E\|\hat{\alpha} - \alpha\|_2^2$. Indeed they showed that
\[
\hat{\alpha}^{JS}(Y) = (1 - \frac{P - 2}{\|Y\|_2^2})Y
\]
verifies
\[
R(\hat{\alpha}^{JS}, \alpha) < R(\hat{\alpha}^{MLE}, \alpha) \quad \forall \alpha \in \mathbb{R}^P, \ P > 2.
\]
The James-Stein estimator above does not shrink (because the factor $1 - \frac{P - 2}{\|Y\|_2^2}$ can be less than $-1$) and does not threshold (i.e., $P(\hat{\alpha}^{JS}(Y) = 0) = 0$). The drawback that the factor can be negative has been alleviated by taking its positive part
\[
\hat{\alpha}^{JS+}(Y) = (1 - \frac{P - 2}{\|Y\|_2^2})_+ Y,
\]
which makes it not only a shrinkage estimator but also a thresholding estimator. Note that this latter estimator, although better than its previous version in terms of $\ell_2$-risk, remains inadmissible. Interestingly the James-Stein estimator has the following oracle inequality
\[
R(\hat{\alpha}^{JS+}, \alpha) \leq 2 + \inf_{c \in \mathbb{R}} R(\hat{\alpha}^c, \alpha),
\]
with respect to the linear estimator $\hat{\alpha}^c = c\hat{\alpha}$, for which the oracle factor is $c^* = \|\alpha\|_2^2/(P + \|\alpha\|_2^2)$ and the optimal $\ell_2$-risk is $R^*(c, \alpha) = \inf_{c} R(\hat{\alpha}^c, \alpha) = P\|\alpha\|_2^2/(P + \|\alpha\|_2^2)$. Candès (2005) provides an interesting review on oracle inequalities.

This result lead to the theme of regularization in the form of shrinkage, penalized likelihood, thresholding, or posterior distributions. Hoerl and Kennard (1970) with ridge regression, Good and Gaskins (1971) with penalties, Wahba’s smoothing splines (see Wahba (1990) for a recent account) are examples of early key contributions. More recent contributions are Donoho and Johnstone (1994) with WaveShrink and Tibshirani (1996) with lasso. Both estimators assume the model $Y \sim N(X\alpha, I)$, where $X$ is an $N \times P$ matrix of wavelets and of covariates, respectively. They aim at estimating $\alpha$ while enforcing sparsity: some entries of the estimated coefficients $\hat{\alpha}$ are set to zero for sparse wavelet representation and model selection purposes, respectively.

On the one hand WaveShrink (which uses an orthonormal matrix $X$) regularizes the wavelet coefficients $\hat{\alpha}^{MLE} = X^T Y$ by thresholding them componentwise towards zero with the nonlinear “hard” or “soft” thresholding functions
\[
\eta^\text{hard}_\varphi (x) = x \cdot 1_{\{|x| \geq \varphi\}},
\]
\[
\eta^\text{soft}_\varphi (x) = \text{sign}(x)(|x| - \varphi)_+,
\]
where $\varphi$ is the threshold: the estimate $\hat{\alpha}_{n,\varphi} = \eta_\varphi(\hat{\alpha}^{MLE}_n)$ is zero if $|\hat{\alpha}^{MLE}_n| \leq \varphi$ for $n = 1, \ldots, N$. Interestingly, when choosing the “universal” threshold $\varphi_N = \sqrt{2 \log N}$,
the estimators have the following oracle inequality
\[ R(\hat{\alpha}_{\text{hard/soft}}^{\text{MLE}}, \alpha) \leq (2 \log N + 1)(1 + \inf_{\delta \in \{0,1\}^N} R(\hat{\delta}, \alpha)), \] (5)
with respect to the diagonal projector \( \hat{\alpha}^{\delta} = \text{diag}(\hat{\delta})\hat{\alpha}^{\text{MLE}} \), for which the oracle binary factors are \( \delta^*_n = 1_{\{|\hat{\alpha}^{\text{MLE}}_n| \geq 1\}} \) and the optimal \( \ell_2 \)-risk is \( R^*(\delta, \alpha) = \inf_{\delta \in \{0,1\}^N} R(\hat{\delta}, \alpha) = \sum_{n=1}^N |\hat{\alpha}^{\text{MLE}}_n|_2^{-1} |\alpha_n| \). Comparing (5) with (2), one sees that the diagonal projector has a bound \( \log N \) increasing with \( N \) as opposed to the constant bound 2 of the linear estimator, but the oracle risk of (5) uses \( N \) oracle parameters as opposed to one for (2) and has therefore more degrees of freedom to be smaller. The universal threshold has asymptotic minimax properties (Donoho et al., 1995).

On the other hand lasso regularizes the MLE by adding an \( \ell_1 \) penalty to the likelihood part and by estimating \( \alpha \) solving
\[ \min_{\alpha} \frac{1}{2} \|Y - X\alpha\|_2^2 + \lambda \sum_{p=1}^P |\hat{\alpha}^{\text{MLE}}_p|^{\nu-1} |\hat{\alpha}_p|, \] (6)
for a given penalty parameter \( \lambda \geq 0 \). Interestingly, when \( X = I \) and \( \lambda = \varphi \) then the soft shrinkage function applied componentwise to \( Y \) is the closed form solution to (6) (Donoho et al., 1992); for other thresholding penalties see Antoniadis and Fan (2001). Lasso is not an oracle procedure (Fan and Li, 2001), so, for a given \( \nu > 1 \), Zou (2006) proposed the adaptive lasso by solving
\[ \min_{\alpha} \frac{1}{2} \|Y - X\alpha\|_2^2 + \lambda \sum_{p=1}^P \frac{1}{|\hat{\alpha}^{\text{MLE}}_p|^{\nu-1}} |\hat{\alpha}_p|, \] (7)
which is oracle provided \( \lambda \) is properly chosen and \( \hat{\alpha}_p \) are root-\( N \)-consistent estimates for \( p = 1, \ldots, P \), e.g., the least squares estimate.

WaveShrink and lasso have been extended to achieve thresholding blockwise, as opposed to componentwise. For instance Cai (1999) proposed the James-Stein-like estimator
\[ \hat{\alpha}^{\text{Block}}_{jb}(\alpha^{\text{MLE}}_{jb}) = (1 - \frac{\lambda L}{\|\alpha^{\text{MLE}}_{jb}\|_2^2}) + \hat{\alpha}^{\text{MLE}}_{jb} \] (8)
to threshold block \( b \) of length \( L = \log N \) raw wavelet coefficients \( \alpha^{\text{MLE}}_{jb} \) at each resolution level \( j \). He showed excellent asymptotic properties and an oracle inequality similar to (2) but with a constant factor of 4.50524 (instead of one) in front of the oracle risk. Likewise Bakin (1999) and Yuan and Lin (2006) proposed group lasso to threshold blocks of coefficients by arbitrarily splitting the vector \( \alpha \) into \( B \) blocks \((\alpha_1, \ldots, \alpha_B)\) and solving
\[ \min_{\alpha} \frac{1}{2} \|Y - X\alpha\|_2^2 + \lambda \sum_{b=1}^B \|\alpha_b\|_2, \] (9)
where \( \|\alpha_b\|_2 = \sqrt{\sum_{l=1}^{L_b} \alpha_b^2} \) and \( L_b \) is the length of the vector \( \alpha_b, b = 1, \ldots, B \). Hence when \( B = P \), then all blocks have size one and the estimator becomes the standard lasso (6). Importantly the group lasso sets more blocks to zero as \( \lambda \) increases.
The paper proposes to unify and extend these thresholding estimators with two goals in mind: offer a continuum of estimators for more flexibility, and select the indexes of the continuum in an efficient way for better estimation, first in the canonical regression setting, then in standard parametric regression. The paper is organized as follows. In Section 2 we consider the canonical regression problem for concomitant measurements and define the generalized and smooth James-Stein estimators that operate blockwise and that are governed by two hyperparameters. Section 3 is concerned with the selection of the two hyperparameters for which we propose two rules: (1) the Stein unbiased risk estimate rule of Section 3.1 shows that the more smooth the estimator with respect to the data, the less erratic, yet still unbiased, the estimation of its risk; (2) the information criterion rule of Section 3.2 is based on the asymptotic property of the universal threshold that a zero sequence should be correctly estimated with a probability tending to one for a threshold increasing at a slow rate. For blocks of a constant size, Section 4 derives an oracle inequality for block thresholding with the universal threshold. Section 5 investigates the finite sample performance of the proposed estimator on a Monte-Carlo experiment. Section 6 illustrates the methodology on simulated gravitational wave data the denoising of three concomitant signals with wavelets. Finally, motivated by the excellent performance of smooth James-Stein thresholding, Section 7 extends its use to the classical regression setting for the selection of covariates in linear models. The estimator is defined implicitly as a fixpoint and its equivalent degrees of freedom is derived as a function of the two hyperparameters, which gives the equivalent degrees of freedom of the adaptive lasso as a particular case. We illustrate the methodology on the prostate cancer data.

2 The generalized and smooth James-Stein estimators

A motivation for block thresholding arises in the following situation. Suppose a sparse sequence of vectors of length $Q$ is observed (e.g., a phenomenon is observed with $Q$ different captors concomitantly), so that at each index $n$ we observe a vector $Y_n = (Y_{n1}, \ldots, Y_{nQ})$ that measures $\alpha_n = (\alpha_{n1}, \ldots, \alpha_{nQ})$ according to

$$Y_{qn} = \alpha_{qn} + \epsilon_{qn}, \quad n = 1, \ldots, N, \quad q = 1, \ldots, Q$$

where the Gaussian noise is i.i.d. $N(0, 1)$ (independent between and within sequences). Note that in wavelet smoothing $Y_{qn}$ plays the role of MLE wavelet coefficients. This situation arises for instance for gravitational wave detection (Klimenko and Mitselmakher, 2004) where a small number of $Q$ captors measures gravitational waves with low signal to noise ratio. Theory predicts rare arrivals of such waves so that most vectors $\alpha_n = 0$. Moreover captors are facing waves with different angles so when $\|\alpha_n\| \neq 0$ then the entries of $\alpha_n$ are typically all different. This situation differs from Cai’s block thresholding in that the number $Q$ is fixed and cannot increase with $N$. 

Adaptive sparse model selection

4
To threshold vectors, soft-WaveShrink can be generalized in the spirit of group lasso to the following block-soft-WaveShrink estimator that solves

$$
\min_{\alpha_n \in \mathbb{R}^Q, n=1,\ldots,N} \frac{1}{2} \sum_{n=1}^{N} \|Y_n - \alpha_n\|_2^2 + \lambda \sum_{n=1}^{N} \|\alpha_n\|_2.
$$

This penalized least squares optimization can be separated in $N$ subproblems of dimension $Q$. In turn one can show that the solution to each $Q$-variate subproblem is a vector collinear with $Y_n$, but thresholded towards zero. Indeed the solution to (11) are the vectors $\hat{\alpha}_{n,\lambda} = Y_n/\|Y_n\|_2(\|Y_n\|_2 - \lambda)_+$, $n = 1,\ldots,N$. Interestingly the solution can also be written as

$$
\hat{\alpha}_{n,\lambda} = (1 - \frac{\lambda}{\|Y_n\|_2^2})_+ Y_n, \quad n = 1,\ldots,N,
$$

which is reminiscent of the James-Stein estimator (1) and Cai’s estimator (8). The difference stands in that the 2-norm in (12) is not squared.

To bridge between the two estimators and extend them, we propose first the generalized James-Stein (GJS) thresholding estimator

$$
\hat{\alpha}_{n,\lambda,\nu} = \eta_{\lambda,\nu}^{GJS}(Y_n)
$$

$$
:= (1 - \frac{\lambda}{\|Y_n\|_2^\nu})_+ Y_n \quad \lambda \geq 0, \; \nu > 0, \; Y_n \in \mathbb{R}^Q
$$

that encompasses (12) for $\nu = 1$ and (1) for $\nu = 2$. Expressed as penalized least squares, the GJS estimator is solution to

$$
\min_{\alpha_n} \frac{1}{2} \|Y_n - \alpha_n\|_2^2 + \lambda \nu \frac{1}{\|Y_n\|_\nu^{-1}} \|\alpha_n\|_2,
$$

which can also be seen as a group version of the adaptive lasso. This parametrization is convenient in the sense that the threshold is always $\lambda$ regardless of $\nu > 0$. Note that $\nu = 1$ corresponds to soft-thresholding (4), $\nu \to \infty$ corresponds to hard (3), and $0 < \nu < 1$ is a thresholding function that biases the estimate even more than soft thresholding. Since the GJS thresholding function is almost differentiable, we consider the Stein unbiased risk estimate (Stein, 1981) to select both hyperparameters in the next section.

The univariate thresholding function of the generalized James-Stein estimator is plotted on the top left corner of Figure 3.1 for $Q = 1$, $\nu = 10$ and $\lambda = 3$. The choice $\nu = 10$ makes the thresholding function close to hard thresholding. A property that will reveal important in Section 3.1 is that the thresholding function converges to hard when $\nu \to \infty$. This motivates in the next section the smooth James-Stein estimator

$$
\hat{\alpha}_{n,\lambda,\nu,\gamma} = \eta_{\lambda,\nu,\gamma}^{SJS}(Y_n)
$$

$$
:= (1 - \frac{\lambda}{\|Y_n\|_2^\nu})_+ Y_n \quad \lambda \geq 0, \; \nu > 0, \; \gamma \geq 1, \; Y_n \in \mathbb{R}^Q
$$

Adaptive sparse model selection
that uses a thresholding function with smooth higher derivatives for reasons we now explain.

3 Hyperparameters selection

The added flexibility in thresholding controlled by $\nu$ translates into better estimation of $\alpha$ provided both $\lambda$ and $\nu$ can be well selected from the data. To that aim we propose two selection rules. The first one minimizes an unbiased estimate of the risk and the second one minimizes an information criterion. The second one is preferable in a situation of strong sparsity and the first one when the underlying vector is more dense.

3.1 Stein unbiased risk estimate

For the multivariate canonical model (10), we first consider the generalized James-Stein estimator (13). Its corresponding thresholding function is almost differentiable in Stein sense (Stein, 1981), and one finds that the Stein unbiased risk estimate is

$$\text{SURE}(\lambda, \nu) = \sum_{n=1}^{N} \|Y_n\|_2^2 1(\|Y_n\|_2 \leq \lambda) + \sum_{n=1}^{N} \frac{\lambda^{2\nu} \|Y_n\|_2^2}{\|Y_n\|_2^{2\nu}} 1(\|Y_n\|_2 > \lambda) + 2 \sum_{n=1}^{N} \frac{(Q - \lambda \nu(Q - \nu))}{\|Y_n\|_2} 1(\|Y_n\|_2 > \lambda) - NQ. \quad (16)$$

This bivariate function can be minimized over $\lambda$ and $\nu$ to select both hyperparameters. Section 5 investigates the finite sample performance of this selection with a Monte-Carlo simulation. It reveals not as good as the extension of SURE for the non-almost differentiable thresholding obtained by $\ell_\nu$ penalized likelihood (Sardy, 2009). This drawback of SURE applied to the generalized James-Stein estimator can be alleviated in the following way.

The SURE function (16) is discontinuous in $\lambda$ at all $\|Y_n\|_2$, $n = 1, \ldots, N$. Consequently the unbiased risk estimate function is erratic in $\lambda$ (i.e., high variance) and hard to minimize. This is particularly true when $\nu$ is large compared to $Q$ because the jumps at the discontinuity points are large when $Q - \nu$ is large, as we can see on the second line of (16). This is due to the fact that the generalized James-Stein thresholding function is discontinuous in its first derivative and that the larger $\nu$ the steeper the slope at the threshold value to approximate the hard thresholding function. To overcome this drawback, we propose the smooth James-Stein estimator (15), which not only thresholds but also has smoother higher derivatives for $\gamma > 1$. Note that $\gamma = 1$ corresponds to (13). The corresponding Stein unbiased risk estimate

$$\text{SURE}(\lambda, \nu; \gamma) = \sum_{n=1}^{N} \|Y_n\|_2^2 1(\|Y_n\|_2 \leq \lambda) + 2Q \sum_{n=1}^{N} \left(1 - \left[\frac{\lambda}{\|Y_n\|_2}\right]^{\nu}\right)^\gamma \quad (17)$$
is consequently less erratic. Figure 3.1 illustrates the advantage of using a large $\gamma$ when estimating the risk with SURE for a large $\nu$. In practice we set $\gamma = 2 \log \nu + 1$ so that a smoother thresholding function is used when $\nu$ is large (i.e., near hard thresholding). Note that for $\nu = 1$ then $\gamma = 1$ so that we fall back to the standard SURE formula for soft thresholding (Donoho and Johnstone, 1995). Section 5 shows that this choice leads to better risk estimation with (17) for free $\gamma$ than with (16) for $\gamma = 1$. This property implies better selection of the two hyperparameters, and hence provides the smooth James-Stein estimator with better $\ell_2$ risk, as the Monte-Carlo simulation shows in Section 5.

### 3.2 Universal threshold and information criterion

We derive a universal threshold and an information criterion for the generalized James-Stein estimator. The idea is to obtain an approximation for the distribution of the smallest threshold $\lambda_Y$ that for a sample $Y = (Y_1, \ldots, Y_N)$ of size $N$ sets all of the $N$ blocks of length $Q$ to zero when the true underlying model is that all vectors are null. Controlling the maximum of $\lambda_Y$, then leads to a finite sample $\tilde{\lambda}_{N,Q}$ and asymptotic $\lambda_{N,Q}$ universal threshold, and a prior $\pi_\lambda$, which are used to define the information criterion as for the sparsity $\ell_\nu$ information criterion $SL_\nu IC$ (Sardy, 2009), as we now show.

With the assumption that $Y_n \sim N_Q(0, I_Q)$ for $n = 1, \ldots, N$, we seek the smallest threshold $\lambda_{N,Q}$ such that the generalized James-Stein estimator (13) estimates the right model with a probability tending to one:

$$P(\hat{\alpha}_1, \ldots, \hat{\alpha}_N = 0) = P(\|Y_1\|_2^2 \leq \lambda_{N,Q}^2, \ldots, \|Y_N\|_2^2 \leq \lambda_{N,Q}^2) \xrightarrow{N \to \infty} 1. \quad (18)$$

The distribution of $M_N = \max_{n=1}^N \|Y_n\|_2^2$, where $\|Y_n\|_2^2 \overset{i.i.d.}{\sim} \lambda_Q^2 = \Gamma(Q/2, 1/2)$, is degenerate. Extreme value theory provides proper rescaling of $M_N$ for

$$c_N^{-1}(M_N - d_N(Q)) \xrightarrow{d} G_0(x),$$

where $G_0(x) = \exp(-\exp(-x))$ is the Gumbel distribution, $c_N = 2$ and $d_N(Q)$ is the root in $\xi$ to

$$\log N - \log \Gamma(Q/2) = (1 - Q/2) \log(\xi/2) + \xi/2. \quad (19)$$

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 an asymptotic approximation to the root of (19), and leads to a good Gumbel approximation when $N$ is
Figure 1: Thresholding functions and corresponding Stein unbiased risk estimation for $Q = 1$ and $\nu = 10$ on simulated data of length $N = 1000$. Left: Subbotin $\ell_\nu$ penalized least squares; Middle: generalized James-Stein ($\gamma = 1$); Right: smooth James-Stein for a smoothness parameter $\gamma = 2 \log \nu + 1$. Top: thresholding functions with parameters chosen to resemble hard thresholding. The left one is discontinuous at the threshold, but with a small slope at the threshold, the middle one has a sharp change of the left and right derivatives at the threshold, and the right one has a smooth change of derivative at the threshold. Bottom: corresponding Stein unbiased risk estimate (blue curve) and true loss (red curve). We observe on the same data that the right estimation is less erratic than the middle one.

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 the asymptotic property (18) is satisfied since

$$
P(\|Y_1\|_2^2 \leq \lambda_{N,Q}^2, \ldots, \|Y_N\|_2^2 \leq \lambda_{N,Q}^2) = G_0(\log \log N) \approx 1 - 1/\log N \xrightarrow{N \to \infty} 1.
$$

Note that for $Q = 1$ we get back the standard universal threshold $\sqrt{2 \log N}$ up to a small term, and for $Q = 2$ we get back the universal threshold of Sardy (2000) for denoising complex-values wavelet coefficients. When $Q$ gets large however, the proposed normalizing constant $d_N(Q)$ in Embrechts et al. (1997) is too far from the exact root to provide a useful approximation, so we find the root $d_N(Q)$ of (19) numerically, which
exists provided \( N \) is large enough (e.g., \( N > \log(\Gamma(Q/2)) + (Q/2 - 1)(1 - \log(Q/2 - 1)) \)). The finite sample universal threshold is then defined as

\[
\hat{\lambda}_{N,Q} = \sqrt{d_N(Q) + c_N \log \log N} \quad \text{with} \quad d_N(Q) \text{ root of (19)} \quad (22)
\]
to have the same rate of convergence for all \( Q \) with \( \hat{\lambda}_{N,Q} \) in place of \( \lambda_{N,Q} \) in (21).

**Property 1.** The inequality \( \hat{\lambda}_{N,Q}^2 \geq \lambda_{N,Q}^2 \) holds for all \( Q \geq 2 \) and all \( N \geq N_0(Q) = \exp(\Gamma(Q/2)^{1/(Q/2-1)}) \), and \( \hat{\lambda}_{N,Q}^2 \sim \lambda_{N,Q}^2 \) as \( N \to \infty \).

**Proof:** Studying (19) as a function of \( \xi \), one sees that the root for a finite sample is larger than the asymptotic root given by Embrechts et al. (1997, p.156) provided \( (Q/2 - 1) \log \log N - \log \Gamma(Q/2) \geq 0 \).

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 \( \lambda \) to reconstruct true zero vectors from noisy measurements. 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 \( \ell_2 \) information criterion (Sardy, 2009).

**Definition.** Suppose model (10) or model (3.1) of Cai (1999) for block thresholding in wavelet smoothing holds. The sparsity weighted \( \ell_2 \) information criterion for estimation of \( (\alpha_1, \ldots, \alpha_N) \) and selection of \( (\lambda, \nu) \) of the generalized James-Stein estimator (14) is

\[
SL^2_{\nu,IC}(\alpha_1, \ldots, \alpha_N, \lambda, \nu) = \frac{1}{2} \sum_{n=1}^{N} ||Y_n - \alpha_n||^2_2 + \nu \sum_{n=1}^{N} \frac{1}{||Y_n||^2_2} ||\alpha_n||^2_2 \\
- N \log(\frac{\Gamma(Q/2)}{2\pi^{Q/2}\Gamma(Q)}) \\
- QN\nu \log \lambda + Q(\nu - 1) \sum_{n=1}^{N} \log ||Y_n||_2 \\
- \log \pi_{\lambda}(\lambda; \tau_{N,Q}) - \log \pi_{\nu}(\nu), \quad (23)
\]

where \( \pi_{\nu} \) is a prior for \( \nu \) (that we choose Uniform on \( \Omega_{\nu} \subset (0, \infty) \) (possibly improper) where thresholding occurs), \( \pi(\lambda; \tau) = F(\lambda; \tau) \) with \( F(\lambda; \tau) = G_0((\lambda^2 / \tau^2 - d_N(Q))/2) \) and \( \tau \) is calibrated to \( \tau_{N,Q}^2 = \lambda_{N,Q}^2/(QN\nu + 1) \) to match the asymptotic model consistency when the true sequences are null, i.e., when \( \alpha_n = 0, \ n = 1, \ldots, N \).

In practice, one minimizes \( SL^2_{\nu,IC} \) like AIC or BIC to select the hyperparameters \( (\lambda, \nu) \) and estimate the sequences \( \alpha_n, \ n = 1, \ldots, N \). Deriving an information criterion for the smooth James-Stein estimator (i.e., \( \gamma > 1 \)) requires knowing the equivalent optimization problem it solves, which is an open problem.
4 Oracle inequality

The $\ell_2$ risk for model (10) is

$$R(\hat{\alpha}, \alpha) = \sum_{n=1}^{N} \rho_n(\alpha_n) = \sum_{n=1}^{N} E\|\hat{\alpha}_n - \alpha_n\|^2.$$ 

Following Donoho and Johnstone (1994), the oracle predictive performance of the block diagonal projection estimator $\hat{\alpha}_n = \delta Y$, where $\delta_n \in \{0, 1\}$ is

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

Hence the oracle hyperparameters are $\delta_n^* = 1(|\alpha_n|^2 > Q)$ for $n = 1, \ldots, N$, and the corresponding oracle overall risk for this block-thresholding estimator is

$$R^*(\delta, \alpha) = \sum_{n=1}^{N} \min(|\alpha_n|^2, Q).$$

The following theorem extends the oracle inequality obtained by Donoho and Johnstone (1994) for $Q = 1$ to block thresholding for $Q \geq 2$.

**Theorem:** Consider any fixed $Q \geq 2$. Then there exists a sample size $N_0(Q)$ such that, for all $N \geq N_0(Q)$ and with the universal threshold $\hat{\lambda}_{N,Q}$ defined in (22), then the smooth James-Stein estimator (15) for $\nu = 1$ and $\gamma = 1$ achieves the oracle inequality

$$R(\hat{\alpha}_{\text{JS}}^{\hat{\lambda}_{N,Q,1;1}}, \alpha) \leq (Q + \lambda_{N,Q}^2)(Q + R^*(\delta, \alpha)),$$

where $\lambda_{N,Q}^2 = 2 \log N + Q \log \log N - 2 \log \Gamma(Q/2)$.

**Proof:** See Appendix A.

This result shows we can mimic the overall oracle risk achieved with $N$ oracle hyperparameters within a factor of essentially $2 \log N + Q \log \log N - 2 \log \Gamma(Q/2)$ with a block thresholding estimator driven by a single hyperparameter. The smallest sample size $N_0(Q)$ for which the inequality holds is quite small in practice; more work is needed to get a tight expression.

5 Monte-Carlo simulation

5.1 $Q = 1$ sequence

We reproduce the Monte-Carlo simulation of Johnstone and Silverman (2004) to estimate sparse sequences 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 $\mu$ taken in $\{3, 4, 5, 7\}$. Table 1 reports $\ell_1$ and $\ell_2$ estimated risks of three estimators: smooth James-Stein, generalized James-Stein, Subbotin $\ell_\nu$ penalized likelihood (Sardy, 2009) and EBayesThresh (Johnstone and Silverman, 2004).

The results clearly show the superiority of the smooth James-Stein estimator with SURE in terms of $\ell_2$ risk, except in case of extreme sparsity where generalized James-Stein based on the information criterion and EBayesThresh perform better (note that this drawback of SURE has been explained by Donoho and Johnstone (1995, Section 2.4)). In particular it is striking to see that, thanks to the induced smoothness, all $\ell_2$ risks are better with the smooth version of the James-Stein estimator when the two hyperparameters are selected based on SURE.

Table 1: Monte-Carlo simulation for $Q = 1$ sequence of length $N = 1000$. Average total squared ($\ell_2$ loss) and absolute ($\ell_1$ loss) errors of: the smooth James-Stein (SJS) estimator with hyperparameters selected with SURE, the generalized James-Stein (GJS) estimator with hyperparameters selected with SURE and SL $w_2^\nu$IC, the Subbotin($\lambda, \nu$) posterior mode estimator with hyperparameters selected with SURE and SL $\nu$IC, and the EBayesThresh estimator with Cauchy-like prior.

<table>
<thead>
<tr>
<th>Number nonzero $\mu$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
<th>50</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value nonzero $\mu$</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SJS</td>
<td>37</td>
<td>37</td>
<td>26</td>
<td>15</td>
<td>202</td>
<td>165</td>
</tr>
<tr>
<td>GJS</td>
<td>28</td>
<td>26</td>
<td>18</td>
<td>10</td>
<td>137</td>
<td>94</td>
</tr>
<tr>
<td>SURE $\gamma = \hat{\omega}$</td>
<td>45</td>
<td>42</td>
<td>31</td>
<td>24</td>
<td>213</td>
<td>174</td>
</tr>
<tr>
<td>GJS $\gamma = 1$</td>
<td>26</td>
<td>22</td>
<td>16</td>
<td>13</td>
<td>124</td>
<td>93</td>
</tr>
<tr>
<td>SL $w_2^\nu$IC</td>
<td>39</td>
<td>41</td>
<td>23</td>
<td>6</td>
<td>380</td>
<td>389</td>
</tr>
<tr>
<td>Subbotin</td>
<td>13</td>
<td>12</td>
<td>7</td>
<td>4</td>
<td>132</td>
<td>115</td>
</tr>
</tbody>
</table>

5.2 $Q = 3$ concomitant sequences

We consider a similar Monte-Carlo simulation where now $Q = 3$ sequences are concomitantly observed, so that all three underlying sequences are identical in the location of the nonzero entries, but not in their amplitude. Since three sequences carry more information than a single one, we may hope to distinguish noise from signal with a smaller signal to noise ratio. Hence we consider that the value of the nonzero terms are $\mu_1 = 1, \mu_2 = 2$ and $\mu_3$ taken in $\{3, 4, 5, 7\}$. The estimated risks of the block smooth James-Stein estimator for the selection rules of the hyperparameters considered are reported in Table 2. To allow some comparison with Table 1, the estimated overall risk of $Q \times N$ sequences is divided by $Q$. 
The smooth James-Stein estimator again performs best overall, while the SL\textsubscript{w}\textsuperscript{2}IC selection rule of the two hyperparameters performs better than for a single observed sequence. Comparison with Table 1 shows that the improvement with \( Q = 3 \) compared to the \( Q = 1 \) case is larger the more sparse the sequence.

Table 2: Monte-Carlo simulation for \( Q = 3 \) sequences of length \( N = 1000 \). Average total squared (\( \ell^2 \) loss) and absolute (\( \ell^1 \) loss) errors divided by \( Q \) of the smooth James-Stein (SJS) estimator with hyperparameters selected with SURE and the generalized James-Stein (GJS) estimator with hyperparameters selected with SURE and SL\textsubscript{w}IC.

<table>
<thead>
<tr>
<th>Number nonzero ( \mu )</th>
<th>5</th>
<th>50</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ell^2 ) loss SJS</td>
<td>18</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>SURE, ( \gamma = \hat{\omega} )</td>
<td>18</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>SURE, ( \gamma = 1 )</td>
<td>22</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>SL\textsubscript{w}IC</td>
<td>19</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>SURE, ( \gamma = \hat{\omega} )</td>
<td>18</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>SURE, ( \gamma = 1 )</td>
<td>22</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>SL\textsubscript{w}IC</td>
<td>19</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>( \ell^1 ) loss SJS</td>
<td>8.7</td>
<td>106</td>
<td>75</td>
</tr>
<tr>
<td>SURE, ( \gamma = \hat{\omega} )</td>
<td>6.8</td>
<td>64</td>
<td>53</td>
</tr>
<tr>
<td>SURE, ( \gamma = 1 )</td>
<td>7.4</td>
<td>68</td>
<td>54</td>
</tr>
<tr>
<td>SL\textsubscript{w}IC</td>
<td>7.7</td>
<td>71</td>
<td>55</td>
</tr>
</tbody>
</table>

6 Wavelet smoothing of concomitant Doppler waves

Gravitational wave bursts are expected to be produced by energetic cosmic phenomena such as the collapse of a supernova. They are believed to be rare, and only a few detectors have been constructed (i.e., \( Q \) is small), which can record concomitant signals at a high sampling frequency but with low signal-to-noise ratio (see Klimenko and Mitselmakher (2004) for more details). We illustrate how the smooth James-Stein thresholding with hyperparameters selected by SURE or set to the universal threshold can be employed blockwise and levelwise for wavelet smoothing of \( Q = 3 \) concomitant signals similar to gravitational waves. To that aim each of the three signals is the Doppler function sampled at 512 points, preceded with a zero-sequence of length 512, and followed with a zero-sequence of length 1024. Each Doppler signal has a different and low signal-to-noise ratio (snr) of 2, 1 and 1/2. Figure 6 shows the denoising, either taking each signal separately using SURE (left three plots), treating the three signal blockwise using SURE (middle three plots), or treating the three signal blockwise using the block universal threshold (right three plots). The smooth James-Stein estimator applied blockwise provides the best reconstruction of the Doppler signals along with the lowest false detections, especially using the block universal threshold levelwise.

7 Extension to the non-canonical model

We are now considering a regression setting that is more common in parametric regression, the case

\[
Y = \mu + \epsilon \quad \text{with} \quad \mu = X\alpha \quad \text{and} \quad \epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1),
\] (24)
Adaptive sparse model selection

Figure 2: Simulated noisy signals detected by three captors at three signal-to-noise ratio: $\text{snr}=2, 1, 1/2$. The denoising is performed taking each of the three signals at a time on the left, and together blockwise on the right using the smooth James-Stein thresholding with hyperparameters selected by SURE (least asymmetric wavelet of order 8 with 4 coarse levels are used)

where $X$ is any $N \times P$ regression matrix with a single response vector (i.e., $Q = 1$). We believe only a subset of the $P$ covariates forms the true linear association. Lasso (Tibshirani, 1996) for $\nu = 1$ and adaptive lasso (Zou, 2006) for $\nu > 1$, defined as the solution to

$$
\min_{\alpha} \frac{1}{2} \|Y - X\alpha\|_2^2 + \lambda^\nu \sum_{p=1}^{P} \frac{|\alpha_p|}{|\hat{\alpha}_p|^{\nu-1}},
$$

have been developed for that setting. Note that the columns $x_1, \ldots, x_P$ of $X$ must have been properly rescaled (Sardy, 2008). Adaptive lasso enjoys an oracle property in the sense that it can guess the true underlying model with a probability tending to one and behaves essentially like the least squares estimate on the correct model. Adaptive lasso also has the advantage of being efficiently calculated with the lars algorithm (Efron et al., 2004). In practice the selection of $\nu$ remains an open problem, to which we
provide a solution below.

We saw in the previous section (i.e., for $X = I$) both that the parameter $\nu$ can be selected by minimizing the Stein unbiased risk estimate and that its smooth James-Stein extension enjoys better predictive performance. Our goal here is to generalize both ideas to the regression setting (24). The same ideas could be employed in other settings as well.

Relaxation, (called backfitting in Statistics) which is another possible algorithm to solve (25) (see Fu (1998); Sardy et al. (2000)), allows to implicitly define the estimate as a fixpoint to the algorithm in the following way. Starting with any initial guess for $\alpha$, the relaxation algorithm keeps all entries constant except the $p$th entry updated to

$$
\alpha_p^{\text{new}} = \frac{1}{|\hat{\alpha}_p^{LS}|^{\nu-1}||x_p||^2_{2}} \eta_{\lambda}^{\text{soft}} (|\hat{\alpha}_p^{LS}|^{\nu-1}r_p^T x_p),
$$

which is the solution to a convex univariate optimization subproblem, where $r_{-p} = Y - X_{-p}\alpha_{old}$ are the residuals without the $p$th term, and $X_{-p}$ is the original $X$ matrix where the $p$th column is replaced by a column of zeros. At the limit, the fixpoint defines the solution which solves a system of nonlinear equations

$$
\hat{\alpha}_p(Y) = \frac{1}{|\hat{\alpha}_p^{LS}|^{\nu-1}||x_p||^2_{2}} \eta_{\lambda}^{\text{soft}} (|\hat{\alpha}_p^{LS}|^{\nu-1}r_p^T x_p)
$$

with $r_{-p} = Y - X_{-p}\hat{\alpha}(Y), \quad p = 1, \ldots, P.$

Instead of employing the almost differentiable soft thresholding function, we can think about iteratively employing the differentiable smooth James-Stein thresholding function. Based on Banach’s fixed point theorem, the following theorem proves that a unique smooth James-Stein estimate is defined.

**Theorem:** Suppose $X$ has $P$ linearly independent columns. For a given $(\lambda > 0, \nu \geq 1; \gamma \geq 1)$, start with any initial value $\alpha$ and iteratively cycle through $p = 1, \ldots, P$ to define the sequence

$$
\alpha_p^{\text{new}} = \frac{r_{-p}^T x_p}{||x_p||^2_{2}} \left\{ \frac{\eta_{\lambda}^{\text{soft}} (|\hat{\alpha}_p^{LS}|^{\nu-1}|r_{-p}^T x_p|)}{|\hat{\alpha}_p^{LS}|^{\nu-1}|r_{-p}^T x_p|} \right\} \gamma \quad \text{with} \quad r_{-p} = Y - X_{-p}\alpha_{old},
$$

where $X_{-p}$ is the original $X$ matrix with its $p$th column replaced by a column of zeros. Then a unique fixed point

$$
\hat{\alpha}_p(Y) = \frac{r_{-p}^T x_p}{||x_p||^2_{2}} \left\{ \frac{\eta_{\lambda}^{\text{soft}} (|\hat{\alpha}_p^{LS}|^{\nu-1}|r_{-p}^T x_p|)}{|\hat{\alpha}_p^{LS}|^{\nu-1}|r_{-p}^T x_p|} \right\} \gamma \quad \text{with} \quad r_{-p} = Y - X_{-p}\hat{\alpha}(Y), \quad p = 1, \ldots, P,
$$

(26)

is reached, which implicitly defines the smooth James-Stein estimate.

**Proof:** See Appendix B.
One checks that when $X = I$, then $\hat{\alpha}_p^{LS} = y_p$ for $p = 1, \ldots, N$, so that (26) is the smooth James-Stein thresholding previously defined (15). The assumption that the columns of $X$ are linearly independent is sufficient to guarantee convergence to a unique (adaptive) lasso solution, but can be relaxed under certain conditions (Sardy, 2009, Theorem 3). To solve the lasso, a systematic cyclic rule to select the next $p$ upon which to iterate has been found efficient when $P$ is relatively small (Sardy et al., 2000), and an optimal descent rule has been found to considerably increase speed of convergence when $P$ is large. These descent rules can also be employed by the smooth James-Stein estimator.

Provided the estimator $\hat{\mu} = g(Y) + Y$ with $g(Y) = X\hat{\alpha}(Y) - Y$ and $\hat{\alpha}(Y)$ defined in (26) is almost differentiable with respect to the data, we can employ the Stein unbiased risk estimate to select $(\lambda, \nu)$. This requires the partial derivatives

$$\nabla g_n(Y)/\partial Y_n = \hat{x}_n^T \nabla_n \hat{\alpha}(Y) - 1 \quad n = 1, \ldots, N,$$

where $\nabla_n \hat{\alpha}(Y)$ are the derivative of $\hat{\alpha}(Y)$ with respect to $Y_n$ and $\hat{x}_1, \ldots, \hat{x}_N$ are the rows of $X$. The following theorem shows how this can be done.

**Theorem:** Let $\hat{\alpha}^{LS} = AY$ where $A = (X^TX)^{-}X^T = (a_{pn})$ (using the generalized inverse if necessary). For a given $(\lambda > 0, \nu \geq 1)$, the smooth James-Stein (26) estimate $\hat{\alpha}(Y)$ is continuously differentiable for $\gamma > 1$. Moreover let $\mathcal{I}_0 = \{p \in \{1, \ldots, P\} : \hat{\alpha}_p(Y) = 0\} = \{p_i\}_{i=1, \ldots, |\mathcal{I}_0|}$ of cardinal $|\mathcal{I}_0|$, let $\bar{\mathcal{I}}_0$ be its complement, and let $X_{\mathcal{I}^0}$ be the columns of $X$ with an index in $\bar{\mathcal{I}}_0$. For a given $n$, the non-zero elements $h_{\mathcal{I}^0}$ of $\nabla_n \hat{\alpha}(Y)$ are solution to the following system of $|\bar{\mathcal{I}}_0|$ linear equations with as many unknowns:

$$x_p^T X_\mathcal{I^0} D_p^\mathcal{I^0} h_{\mathcal{I}^0} = z_{ni} \quad p = p_i, \ i = 1, \ldots, |\bar{\mathcal{I}}_0|,$$

where

$$z_{ni} = x_{np} + \frac{u_p}{v_p},$$

and $D_p^\mathcal{I^0}$ is the identity matrix except that its $i$th diagonal element is

$$D_p^{\mathcal{I}^0}_{p,ii} = \frac{w_p^\gamma}{1 - \gamma + \gamma w_p} =: v_p^{-1}$$

with

$$w_p = \frac{|\hat{\alpha}_p^{LS}|^{\nu - 1} |r_p X_p|}{\eta_{\nu}^{soft}(|\hat{\alpha}_p^{LS}|^{\nu - 1} |r_p^T X_p|)} \quad \text{and} \quad u_p = \frac{\gamma (\nu - 1) a_{pn}(w_p - 1) r_p^T X_p}{\hat{\alpha}_p^{LS} w_p^\gamma}.$$

Solving (28) leads to the gradient $\nabla_n \hat{\alpha}(Y)$ needed in (27) to calculate $\text{SURE}(\lambda, \nu; \gamma)$ for all $\lambda, \nu$ and $\gamma$.

**Proof:** See Appendix C.

Note that we used the least squares estimate as raw estimate of the coefficients, but the formula would still hold for other linear estimators. At the limit when $\gamma$ tends
to one, the adaptive lasso has $D_{p,n}^2 = 1$ for all $i$ in (30). A particular case of interest is the lasso (i.e., $\nu = 1$) for which we moreover have $z_{ni} = x_{np}$ in (29), so that (28) leads to $\hat{h}_{n,0}^2 = ((X\bar{I}_0)^TX\bar{I}_0)^{-1}(\bar{x}_{n,0}^I)^T$. Hence, from (27), we see that

$$\sum_{n=1}^{N} \partial g_n(Y)/\partial Y_n = -N + \sum_{n=1}^{N} \bar{x}_n^T \nabla_n \hat{\alpha}(Y)$$

$$=-N + \sum_{n=1}^{N} \bar{x}_n^T ((X^T\bar{I}_0)^TX^T\bar{I}_0)^{-1}((X^T\bar{I}_0)^TX^T\bar{I}_0)^T \bar{x}_n^T$$

$$=-N + \text{trace}(((X^T\bar{I}_0)^TX^T\bar{I}_0)^{-1}((X^T\bar{I}_0)^TX^T\bar{I}_0))$$

$$=-N + |\bar{I}_0|,$$

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

We illustrate the estimation of the $\ell_2$ risk and the smoothness gained with the smooth James-Stein estimator on the prostate cancer data with $P = 8$ covariates used by Tibshirani (1996). Figure 7 shows the estimated risk as a function of the two hyperparameters $\lambda$ and $\nu$, either for adaptive lasso (left) or for the smooth James-Stein estimator (right). Both estimations of the risk are unbiased, but the latter is less erratic. The advantage of smoothness would be even more pronounced with data containing a larger collection of covariates. The selected hyperparameters are $(\hat{\lambda}, \hat{\nu}) = (1.12, 10)$, for which the selected subset has five covariates out of the eight original ones. Finally we report that the estimated risk with $\nu = 1$ (i.e., lasso) is roughly 15, while it drops to roughly 13 with $\nu = 10$, showing the potential advantage of adaptive lasso and the smooth James-Stein estimator on this particular data set.

8 Conclusion

Two of Stein’s important contributions have been extended to develop a new variable selection technique driven by two hyperparameters and a smoothness parameter for a better fit of the data and a better estimation of its prediction performance. The technique iteratively applies the smooth James-Stein threshold to a covariate at a time projected on the partial residuals to essentially test the relative significance of its coefficient. The new estimator enjoys smoothness like ridge regression and performs variable selection like lasso.

Block thresholding using the smooth James-Stein thresholding function can be extended in many ways. For instance Cai’s block thresholding of wavelet coefficients can be employed for concomitant measurements (i.e., $Q > 1$), which would entail blocking in two directions: across and between time; the relative block sizes, oracle inequalities, the corresponding Stein unbiased risk estimate formula are open issues related to such an extension of the methodology.

Finally the smooth James-Stein thresholding function originally developed for canonical regression can be further applied by existing estimators seeking smooth sparsity in
Figure 3: Prostate cancer data: Stein unbiased risk estimate as a function of $\lambda$ and $\nu$ for adaptive lasso (left, with $\gamma = 1$) and for the smooth James-Stein estimator (right, with $\gamma = 1 + 2 \log \nu$).

other settings, in the same way it was extended here to the general regression setting, for instance for Markov random field smoothing, classification or blockwise additive models.

9 Acknowledgements

I would like to thank Yvan Velenik for helpful discussions about the convergence issue of Theorem 2. Partially supported through Swiss National Science Foundation.

A Oracle inequality

Consider the smooth James-Stein block thresholding estimator for $\nu = 1$ and $\gamma = 1$ for the canonical regression problem with $N$ measurements of $Q$ (fixed) concomitant observations of a phenomenon. The individual risks have the form

$$\rho_n(\lambda, \alpha_n) = \mathbb{E}[\left(\frac{|Y_n| - \lambda}{|Y_n|}\right)^+ Y_n - \alpha_n]^2 = Q + \mathbb{E}\min(\lambda^2, |Y_n|^2) - 2Q + 2\mathbb{E}(Y_n - \alpha_n)^T \frac{|Y_n| - \lambda}{|Y_n|} Y_n.$$

The last scalar product is the sum of $Q$ terms of the form
\[
E(Y_{ni} - \alpha_{ni})(|Y_n| - \lambda)y_i Y_{ni} = E(Y_{ni} - \alpha_{ni})\frac{(|Y_n| - \lambda)}{|Y_n|} Y_{ni} - E(Y_{ni} - \alpha_{ni})\frac{(|Y_n| - \lambda)}{|Y_n|} Y_{ni1}\{|Y_n| < \lambda\} = I_{i1} - I_{i2}, \quad i = 1, \ldots, Q.
\]
Consider $i = 1$ and partition $Y_n = (Y_1, U)$, where $U = (Y_{n2}, \ldots, Y_{nQ})$. Then
\[
I_{i1} = C \int \int \frac{(|Y| - \lambda)}{|Y|} y_1(y_1 - \mu_1) \exp(-\frac{1}{2}(y_1 - \mu_1)^2)dy_1 \exp(-\frac{1}{2}u^T u)du.
\]
Integrating by part the inner integral, one finds
\[
I_{i1} = C \int \int \exp(-\frac{1}{2}(y_1 - \mu_1)^2)(1 - \lambda \frac{|u|^2}{|y|^2})dy_1 \exp(-\frac{1}{2}u^T u)du
\]
\[
= 1 - CL \int \int \exp(-\frac{1}{2}(y_1 - \mu_1)^2)\frac{|u|^2}{|y|^3}dy_1 \exp(-\frac{1}{2}u^T u)du,
\]
where the last multiple integral is positive. Likewise
\[
I_{i2} = C \int \int \frac{\sqrt{\lambda^2 - |u|^2}}{\sqrt{\lambda^2 - |u|^2}} \exp(-\frac{1}{2}(y_1 - \mu_1)^2)(1 - \lambda \frac{|u|^2}{|y|^2})dy_1 \exp(-\frac{1}{2}u^T u)du
\]
\[
= P(|Y| \leq \lambda)
\]
\[
-CL \int \int \frac{\sqrt{\lambda^2 - |u|^2}}{\sqrt{\lambda^2 - |u|^2}} \exp(-\frac{1}{2}(y_1 - \mu_1)^2)\frac{|u|^2}{|y|^3}dy_1 \exp(-\frac{1}{2}u^T u)du,
\]
where the last multiple integral is positive but smaller than the previous multiple integral, so that the difference $\Delta_i = I_{i1} - I_{i2} \leq 0$ between them is negative. Putting all terms together, one finds that
\[
\rho_n(\lambda, \alpha_n) = Q + E \min(\lambda^2, |Y_n|^2) - 2QP(|Y_n| < \lambda) + 2 \sum_{i=1}^Q \Delta_i 
\]
\[
\leq Q + E \min(\lambda^2, |Y_n|^2) - 2QP(|Y_n| < \lambda).
\]
On the one hand from (32) we have
\[
\rho_n(\lambda, \alpha_n) \leq Q + \lambda^2
\]
\[
\leq \begin{cases} (Q + \lambda^2)(Q/N + \lambda) & Q \geq 1 \\ (Q + \lambda^2)(Q/N + |\alpha_n|^2) & |\alpha_n|^2 \geq 1 \end{cases}
\]
If we can show the second inequality also holds for $|\alpha_n|^2 < 1$, the proof is complete. To that aim, note that on the other hand from (32) we have
\[
\rho_n(\lambda, \alpha_n) \leq Q + E|Y_n|^2 - 2QP(|Y_n| < \lambda)
\]
\[
= |\alpha_n|^2 + 2Q(1 - P(|Y_n|^2 < \lambda^2))
\]
\[
= |\alpha_n|^2 + g(\mu; \lambda)
\]
with $g(\mu; \lambda) = 2Q \{1 - \exp(-\mu^2/2) \sum_{j=0}^{\infty} \frac{(\mu^2/2)^j}{j! \Gamma(j+Q/2)}\}$ since $|Y_n|^2$ has a noncentral chi-square distribution with $Q$ degrees of freedom and noncentrality parameter $\mu^2 = |\alpha_n|^2 < 1$. Considering even $Q$’s for simplicity, Taylor’s expansion gives

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

where firstly

$$g(0; \tilde{\lambda}_{N,Q}) = 2Q \left(1 - \frac{\gamma(Q/2, \tilde{\lambda}_{N,Q}^2/2)}{\Gamma(Q/2)}\right)$$

$$= 2Q \exp(-\tilde{\lambda}_{N,Q}^2/2) \sum_{j=0}^{Q/2-1} \frac{\tilde{\lambda}_{N,Q}^2/2^j}{j! \Gamma(j+1)}$$

$$\leq 2Q \exp(-\lambda_{N,Q}^2/2) \sum_{j=0}^{Q/2-1} \frac{\lambda_{N,Q}^2/2^j}{j! \Gamma(j+1)}$$

$$= \frac{2Q}{N} \frac{\Gamma(Q/2)}{(\log N)^{Q/2}} \left(1 + \lambda_{N,Q}^2/2 + \sum_{j=2}^{Q/2-1} \frac{(\lambda_{N,Q}^2/2)^j}{j! \Gamma(j+1)}\right),$$

where the inequality stems from Property 1. So with $N$ large enough for $\Gamma(Q/2)/(\log N)^{Q/2} \leq 1$, then

$$g(0; \tilde{\lambda}_{N,Q}) \leq \frac{Q}{N}[2 + \lambda_{N,Q}^2$$

$$+ 2 \sum_{j=2}^{Q/2-1} (1 + \frac{Q/2 \log \log N - \log \Gamma(Q/2)}{\log N})^j \frac{\Gamma(Q/2)}{(\log N)^{Q/2-j} \Gamma(j+1)}].$$

Each term in the sum is strictly less than one for $N$ large enough. Hence for $N$ large enough then

$$g(0; \tilde{\lambda}_{N,Q}) \leq \frac{Q}{N}[2 + \lambda_{N,Q}^2 + 2(Q/2 - 2)] \leq \frac{Q}{N}[Q + \lambda_{N,Q}^2].$$

Note also that

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

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

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

$$+ 2Q \exp(-\lambda^2/2) \exp(-x^2/2) x^2 \sum_{j=0}^{\infty} \frac{(x^2/2)^j}{j! \Gamma(j+1)} \frac{\lambda^2/2^j+Q/2+1}{\Gamma(j+Q/2+2)}$$

$$\leq 2Q + 2Qx^2 S(\frac{\lambda^2/2}{Q/2} - 1) \leq 2Q + 2\lambda^2.$$
Adaptive sparse model selection

with \( S = \exp(-\lambda^2/2) \exp(-x^2/2) \sum_{j=0}^{\infty} (\lambda^2/2)^j (\lambda^2/2)^{j+1} \) ≤ 1. It is also easy to show that \(-g''(x; \lambda) \leq 2Q \) for \( x \in [0, 1] \). Consequently for \( N \) larger than say \( N_0(Q) \), we have

\[
g(\mu) \leq Q/N(Q + \lambda^2_{N,Q}) + \mu^2/2(2Q + 2\lambda^2_{N,Q})
\]

for \( \mu = |\alpha_n| < 1 \). Putting (34) and (35) together, we get that \( \rho_n(\lambda; \alpha_n) \leq (Q + \lambda^2_{N,Q})(Q/N + \min(|\alpha_n|^2, Q)) \) for \( N \) large enough. Summing over all \( n = 1, \ldots, N \) leads to the oracle inequality.

B Convergence of the iterated SJS thresholding sequence

The sequence is generated by iteratively applying \( P \) composed applications \( T^{(\gamma)} = T_1^{(\gamma)} \circ \cdots \circ T_P^{(\gamma)} \), where each \( T_p^{(\gamma)} : \mathbb{R}^p \rightarrow \mathbb{R}^p \) for all \( p = 1, \ldots, P \) is defined as

\[
(T_p^{(\gamma)}(\alpha))_i = \begin{cases} 
\alpha_i, & i \neq p, \\
\frac{r_{T_p} x_p}{\|x_p\|_2} \left( \frac{\rho_n^{	ext{soft}}(|\alpha|^2_{p-1}|x_p)}{\|x_p\|\|\hat{r}_p\|} \right) \gamma & \text{with } r_{-p} = Y - X_{-p}\alpha, \ i = p.
\end{cases}
\]

To show that \( T^{(\gamma)} \) is a contraction for all \( \gamma \geq 1 \), first note that when \( \gamma = 1 \) the sequence obtained from iteratively applying \( T = T^{(1)} \) converges to the unique fixed point of \( T \) which is solution to a strictly convex optimization problem (since the columns of \( X \) are linearly dependent); for a proof see Fu (1998) and Sardy et al. (2000). Hence \( T^{(1)} = T, T^2 = T \circ T, \) and in general all \( T^n \) for all \( n \geq 1 \) have at most one fixed point. Hence a reverse of Banach’s fixed point theorem guarantees there exists a metric \( \rho \) on \( \mathbb{R}^p \) such that \( T \) is a contraction of constant \( K \in (0, 1) \) (Bessaga (1959); see also Deimling (1985, Theorem 17.5 p.191)). Moreover for \( \gamma > 1 \) note that \( T_p^{(\gamma)}(\alpha) = \omega^\delta_p T_p(\alpha) < T_p(\alpha) \) since \( \delta > 0 \) and \( \omega_p = \rho_n^{	ext{soft}}(|\alpha|^2_{p-1}|x_p)/(|\hat{r}_p|^2_{p-1}|x_p|) \in [0, 1] \) for all \( \alpha_p \) and all \( p \). So \( T^{(\gamma)} \) is itself a contraction of constant \( K \) for all \( \gamma > 0 \) with the \( \rho \) metric. Hence Banach’s fixed point theorem guarantees convergence to a unique fixed point \( \hat{\alpha} = T^{(\gamma)}(\hat{\alpha}) \), which defines the smooth James-Stein estimate.

C SURE for the SJS estimate

Using the implicit definition of the smooth James-Stein estimate (26), one finds

\[
\frac{\partial \hat{\alpha}_p(Y)}{\partial Y_n} = \begin{cases} 
0 & \text{if } \hat{\alpha}_p(Y) = 0, \\
\frac{1}{\|x_p\|_2^2} (v_p(x_{np} - x_p^T X_{-p} \nabla_x \alpha(Y)) + u_p) & \text{else},
\end{cases}
\]

where \( u_p \) and \( v_p \) are given in (30) and (31). Let now \( I_0 = \{ p \in \{1, \ldots, P\} : \hat{\alpha}_p(Y) = 0 \} = \{ p_i \}_{i=1,\ldots,|I_0|} \) of cardinal \( |I_0| \) and \( \bar{I}_0 \) be its complement. Let also \( h_n^{I_0} \) be the
elements of $\nabla_n \hat{\alpha}(Y)$ that are unknown (the other ones being null) and $X_{\bar{I}_0}$ be the columns of $X$ with an index in $\bar{I}_0$. From (36), we see that for a given $n$, then $h_{\bar{I}_0}^{I_0}$ solves

$$x_p^T X_{\bar{I}_0} D_{\bar{I}_0} h_{\bar{I}_0}^{I_0} = z_{ni} \quad p = p_i, \quad i = 1, \ldots, |I_0|$$

where $D_{\bar{I}_0}$ is given in (30) and $z_{ni}$ is given in (29). These equations define a system of $|\bar{I}_0|$ linear equations with as many unknowns. Solving it leads to the gradient $\nabla_n \hat{\alpha}(Y)$ needed in (27). Note that the matrix of the linear system (28) is essentially the symmetric positive definite matrix $(X_{\bar{I}_0})^T X_{\bar{I}_0}$ which diagonal elements are multiplied by $D_{\bar{I}_0,ii}$. Moreover $w_p > 1$ in (31) since the denominator is the soft-thresholded version of the numerator, hence all $D_{\bar{I}_0,ii} \geq 1$ since $f(w) = w^\gamma / (1 - \gamma + \gamma w)$ satisfies $f(1) = 1$ and $f'(w) \geq 0$ for $w \geq 1$. This guarantees existence of a solution to the linear system with $\gamma = 1$ if the column of $X$ are linearly independent, and a fortiori with $\gamma > 1$ even if the columns of $X$ are linearly dependent (in this second case, $\gamma$ plays the role of a ridge parameter). It also means the implicit function theorem can be employed to guarantee that the smooth James-Stein estimate is continuously differentiable with respect to the data when $\gamma > 1$.

References


Adaptive sparse model selection


