Adaptive Posterior Mode Estimation of a Sparse Sequence for Model Selection

SYLVAIN SARDY
Department of Mathematics, University of Geneva

For the problem of estimating a sparse sequence of coefficients of a parametric or nonparametric generalized linear model, posterior mode estimation with a Subbotin(λ, ν) prior achieves thresholding and therefore model selection when $\nu \in [0, 1]$ for a class of likelihood functions. The proposed estimator also offers a continuum between the (forward/backward) best subset estimator ($\nu = 0$), its approximate convexification called lasso ($\nu = 1$) and ridge regression ($\nu = 2$).

Rather than fixing ν, selecting the two hyperparameters λ and ν adds flexibility for a better fit, provided both are well selected from the data. Considering first the canonical Gaussian model, we generalize the Stein unbiased risk estimate SURE(λ, ν) to the situation where the thresholding function is not almost differentiable (i.e., $\nu < 1$). We then propose a more general selection of λ and ν by deriving an information criterion that can be employed for instance for the lasso or wavelet smoothing.

We investigate some asymptotic properties in parametric and nonparametric settings. Simulations and applications to real data show excellent performance.

Keywords: extreme value theory, generalized linear model, Gumbel and Fréchet prior, information criterion, lasso, ℓ_ν-penalized likelihood, model selection, sparsity, Stein unbiased risk estimate, threshold, wavelet smoothing.

1 Introduction

1.1 Background

The central problem of this paper is that of estimating a sequence of P coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_P)$ from N noisy data $\mathbf{Y} = (Y_1, \ldots, Y_N)$ with the knowledge that the sequence should be sparse: some unknown entries of $\boldsymbol{\alpha}$ are null, while the relative magnitude of the others with respect to the noise is also unknown. In a first stage, we consider the canonical estimation problem of Johnstone and Silverman (2004) who suppose that for $P = N$:

$$Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\alpha_n, 1) \qquad n = 1, \ldots, N \tag{1}$$

The seminal thresholding approach of Donoho and Johnstone (1994) provides a sparse estimate by applying the hard- or soft-thresholding to Y_n:

$$\eta_\varphi^{(\text{hard})}(Y_n) = Y_n \cdot 1_{\{|Y_n| \geq \varphi\}}(Y_n), \tag{2}$$

$$\eta_\varphi^{(\text{soft})}(Y_n) = \operatorname{sign}(Y_n)(|Y_n| - \varphi)_+, \tag{3}$$

where φ is the threshold: for fixed φ, the estimate $\hat{\alpha}_n = \eta_\varphi(Y_n)$ is zero if $|Y_n| \leq \varphi$. They proposed minimax and universal rules ($\varphi_N = \sqrt{2 \log N}$) for the selection of φ, and Donoho and Johnstone (1995) proposed minimizing the Stein unbiased risk estimate over φ when using soft-thresholding. Recently Johnstone and Silverman (2004) derived EBayesThresh, a posterior median estimate which offers a continuum between the two thresholding functions, with the following methodology:

- an independent mixture distribution is assumed on each α_n:

$$\pi(\alpha_n) = (1 - w)\delta_0(\alpha_n) + w a \gamma(a\alpha_n), \tag{4}$$

  where δ_0 is the Dirac mass at zero, and the nonzero part γ is heavy-tailed (e.g., Laplace or quasi-Cauchy) with scale parameter a;

- the posterior median estimate of each α_n is calculated via a closed form expression that thresholds;

- the empirical Bayes selection of the hyperparameters w and a provides an adaptive fit to the sparse sequence by maximizing over w and a the marginal likelihood, which has a tractable form for Gaussian noise.

EBayesThresh provides excellent estimation of sparse sequences, both from empirical and theoretical point of views, because it can adapt thresholding to the data, rather than fixing it to hard or soft.

In a second stage, we consider in the spirit of generalized linear model (Nelder and Wedderburn 1972) the estimation problem with sparsity constraints on $\boldsymbol{\alpha}$ to the more general model:

$$Y_n \overset{\text{i.i.d.}}{\sim} F(y; \mu_0 + \mathbf{x}_n^{\mathrm{T}} \boldsymbol{\alpha}, \psi) \qquad n = 1, \ldots, N \tag{5}$$

where F is the distribution of continuous or discrete random variables Y_n parametrized by a location parameter $\mu_0 + \mu_n$ with $\mu_n = \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\alpha}$ and nuisance parameter ψ. For instance, such models are used in regression, classification and inverse problems with $N \times P$ matrix $X = [\mathbf{x}_1 \mid \ldots \mid \mathbf{x}_N]^{\mathrm{T}}$, where F could be Gaussian, Poisson or Bernoulli, and $\mathbf{x}_n = (x_{n,1}, \ldots, x_{n,P})^{\mathrm{T}}$ are covariates (parametric) or discretized basis functions (nonparametric). For this general setting (5), the oldest way of seeking sparsity is best subset variable selection driven by information criteria like AIC (Akaike 1973), C_p (Mallows 1973) or BIC (Schwarz 1978). Variable selection is the generalization of hard-thresholding to (5) and corresponds to ℓ_0-penalized likelihood. In the general setting, EBayesThresh faces the problem of defining and calculating a *multivariate* posterior median. More recently, lasso-type posterior mode estimators (Donoho and Johnstone 1994; Tibshirani 1996; Park and Hastie 2007) alternatively provide sparsity by assuming a Laplace ℓ_1-prior for $\boldsymbol{\alpha}$; the hyperparameter is selected based on an AIC-like criterion (Zou, Hastie, and Tibshirani 2007). Lasso-type selection is the generalization of soft-thresholding to (5). So both best subset- and lasso-variable selection are posterior mode estimators for ℓ_0 and ℓ_1 priors.

1.2 Proposal

To achieve model selection for the canonical model (1) and then its generalization (5), we attain sparse sequence estimation by conjunction of using the Subbotin(λ, ν) prior and of employing posterior mode estimation. The posterior mean or median could be estimated for instance by means of Markov chain Monte Carlo methods, but would not lead to a sparse estimation with the Subbotin prior. Likewise the Subbotin prior does not have the prerogative of sparse posterior mode estimation: for instance, Antoniadis and Fan (2001) give sufficient conditions on the prior for sparse wavelet smoothing, and Griffin and Brown (2007) propose scale mixtures of Gaussian distributions for the priors of a generalization of lasso. The reasons for our choice of Subbotin posterior mode are the following. First, it links three essential posterior mode estimators: subset variable selection (ℓ_0), lasso (ℓ_1) and ridge regression (ℓ_2) (Hoerl and Kennard 1970). Second, model selection is achieved by posterior mode for ℓ_ν with $\nu \in [0, 1]$, not only for Gaussian likelihood, but for a class of distributions. Third, since the Subbotin(λ, ν) prior can be seen as a continuous approximation to EBayesThresh(w, a)'s priors (4), its empirical performances are expected to be competitive with EBayesThresh in the canonical setting (1). Fourth, Subbotin posterior mode entails solving a continuous optimization problem which gives the possibility to extend the range of applicability of the estimator from model (1) to model (5). The corresponding multivariate optimization, although non-convex when $\nu < 1$, is at least continuous.

The proposed Subbotin posterior mode aims at extending existing estimators to the continuous choice of ν, to a class of likelihood functions, and to any linear association. In Section 2, we first consider the canonical setting (1) where we can calculate the exact Subbotin posterior mode in the nonconvex case $\nu < 1$. We define

the posterior mode estimator in Section 2.1, derive two methods for the adaptive selection of the hyperparameters λ and ν in Section 2.2: by a generalization of Stein unbiased risk estimate (Stein 1981), and by means of an information criterion. We perform a Monte Carlo simulation in Section 2.3 to compare the finite sample performance of the new estimator to that of EBayesThresh for the canonical model and for Gaussian wavelet smoothing. We consider in Section 3 the more general setting (5), we investigate existence, uniqueness and thresholding properties for a class of distributions F and matrices X, and we generalize the information criterion. In Section 4 we apply the method to various settings, where we can calculate the exact Subbotin posterior mode: lasso regression in Section 4.1 (X is not the identity and $\nu = 1$), the Poisson canonical model in Section 4.2 (X is the identity, the likelihood is non-Gaussian and ν is selected in $(0, 1]$), and Poisson wavelet smoothing in Section 4.3 (X is not the identity, the likelihood is non-Gaussian and $\nu = 1$). Section 5 explores some asymptotic properties of Subbotin posterior mode estimation, and in particular its asymptotic minimaxity. Section 5 also shows that the adaptive lasso (Zou 2006) can be seen as an approximation to Subbotin posterior mode estimation. We make some final remarks in Section 6 and postpone technical derivations to the appendices.

2 Subbotin(λ, ν) posterior mode estimate

2.1 Posterior mode with the Subbotin prior

The Subbotin (also called power exponential) distribution has density

$$\pi(\alpha_n \mid \lambda, \nu) = \frac{\lambda^{1/\nu}}{2\Gamma(1 + \frac{1}{\nu})} \exp(-\lambda|\alpha_n|^\nu)$$

parametrized by two hyperparameters λ and ν. The Subbotin can be seen as a continuous approximation to EBayesThresh prior (4) since it tends to the point mass at zero for instance with $\nu = 1/\lambda \to 0$ and to the Laplace when $\nu \to 1$, as illustrated in the top row of Figure 1. Using Subbotin as prior for the canonical model (1), the univariate posterior mode estimate $\hat{\alpha}_{n,\lambda,\nu}$ of each α_n solves

$$\min_{\alpha_n} \frac{1}{2}(Y_n - \alpha_n)^2 + \lambda|\alpha_n|^\nu - \log(\frac{\lambda^{1/\nu}}{2\Gamma(1 + \frac{1}{\nu})}), \quad n = 1, \ldots, N. \tag{6}$$

Hence the multivariate version involves the penalty $+\lambda\|\boldsymbol{\alpha}\|_\nu^\nu = +\lambda\sum_{n=1}^N |\alpha_n|^\nu$ by assuming independent coefficients $\boldsymbol{\alpha}$. The simplicity of the ℓ_ν penalty contrasts with the sum of logarithm of sums one obtains with EBayesThresh's type priors (4).

The ℓ_ν penalized least squares function (6) is non-convex for $\nu < 1$ and has at most two local minima (always one at zero) among which the global one. Theorem 1 of Antoniadis and Fan (2001) states that the posterior mode thresholds the

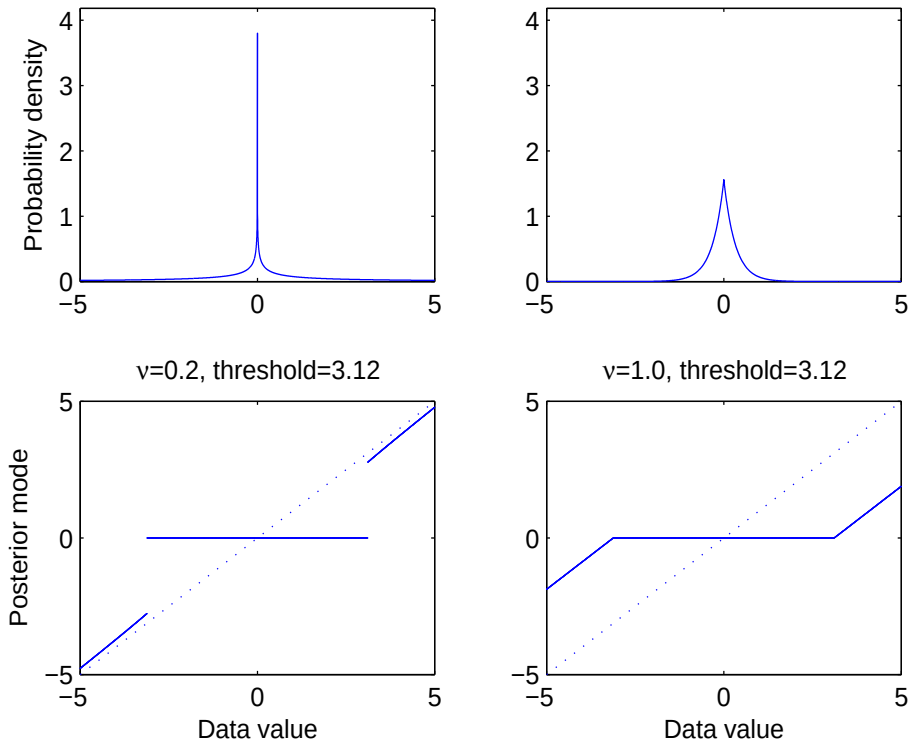

Figure 1: Two examples of Subbotin(λ, ν) prior density (top row) and corresponding posterior mode thresholding (bottom row) functions for $\nu = 0.2$ (left) and $\nu = 1$ (right), with λ such that $\varphi(\lambda; \nu) = \sqrt{2 \log N}$ and $N = 128$ using (8).

maximum likelihood estimate $\hat{\alpha}_{\mathrm{MLE}} = Y_n$ for a class of penalties. We recall that an estimator thresholds when there exists a threshold value $\varphi(\lambda; \nu)$ such that

$$\hat{\alpha}_{\lambda,\nu}(Y_n) = 0 \quad \text{if and only if} \quad |Y_n| \leq \varphi(\lambda; \nu). \tag{7}$$

The result of Antoniadis and Fan (2001) applies in particular to the ℓ_ν penalty for any $\nu \leq 1$. The bottom row of Figure 1 illustrates the thresholding property of the posterior mode for $\nu = 0.2$ (left) and $\nu = 1$ (right), and the flexibility gained by not fixing ν with our approach. In Section 3, Theorem 5 extends the thresholding property of ℓ_ν-based posterior mode estimation to a class of likelihood functions.

The non-zero part of the thresholding function must be found numerically, except when $\nu = 0^+$ (hard (2)) and $\nu = 1$ (soft (3)). Moreover, the thresholding function has a discontinuity at $\pm\varphi(\lambda; \nu)$ with jump $\kappa(\lambda; \nu)$ given by

$$\varphi(\lambda; \nu) = (2 - \nu)[\lambda\{2(1 - \nu)\}^{\nu - 1}]^{1/(2-\nu)}, \tag{8}$$

$$\kappa(\lambda; \nu) = \varphi(\lambda; \nu)\frac{2(1 - \nu)}{(2 - \nu)}, \tag{9}$$

both of which are found by solving system (17) for Gaussian likelihood (for details see Appendix C). The expression for the threshold (8) can also be found in Knight and Fu (2000) after reparametrization; Antoniadis and Fan (2001, p_0 formula p. 944) found a lower bound for it. For $\nu = 0$, the threshold formula $\varphi(\lambda;\nu) = \sqrt{2\lambda}$ reveals that BIC$= \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2/2 + \lambda_{\text{BIC}} \cdot (\#n : \{\hat{\alpha}_n \neq 0\})$ with $\lambda_{\text{BIC}} = (\log N)/2$ is less conservative than the universal rule $\varphi_N = \sqrt{2\log N}$ since the universal penalty is $\lambda_N(\nu) = \log N = 2\lambda_{\text{BIC}}$.

Computationally, posterior mode requires no calculation when $|Y_n| < \varphi(\lambda;\nu)$, since the solution is zero in this case. An iterative method finds the unique mode in the interval $[\text{sign}(Y_n) \cdot \kappa(\lambda;\nu), Y_n]$ otherwise.

2.2 Hyperparameter selection

Flexibility of the prior leads to good estimation of sparse sequences, provided the two hyperparameters can be efficiently selected to adapt to the degree of sparsity. Gao and Bruce (1997) with minimax rules, Antoniadis and Fan (2001) with universal rules, and Fan and Li (2001) and Fan and Peng (2004) with approximate generalized cross validation considered priors with two hyperparameters and ways to select them.

2.2.1 Extension of SURE

Stein (1981) considered model (1) and estimates of the form $\hat{\boldsymbol{\alpha}} = \mathbf{Y} + g(\mathbf{Y})$, where $g : \mathbb{R}^{\text{N}} \to \mathbb{R}^{\text{N}}$. For almost differentiable functions g, he derived an unbiased estimate of the ℓ_2 risk between $\hat{\boldsymbol{\alpha}}$ and the true $\boldsymbol{\alpha}$. Donoho and Johnstone (1995) observed that the soft shrinkage ($\nu = 1$) is almost differentiable and used the Stein unbiased risk estimate (SURE) to select the threshold φ. The two hyperparameters of SCAD thresholding (Fan and Li 2001) could also be selected using SURE. The Subbotin thresholding function $\tilde{g}$ is not almost differentiable when $\nu < 1$ because of the jump κ at $\pm\varphi$. In fact $\tilde{g}$ can be written as the sum of an almost differentiable function g and two Heaviside functions at the discontinuity points $\pm\varphi(\lambda;\nu)$ with jump $\kappa(\lambda;\nu)$ given by (9). Since Stein's derivation is essentially based on integration by part, we generalize his formula by means of the Heaviside function H and its derivative, the Dirac measure δ.

Theorem 1 *Consider model (1) and the estimate $\hat{\boldsymbol{\alpha}} = \mathbf{Y} + \tilde{g}(\mathbf{Y})$, where $\tilde{g} = g - \kappa \cdot (1 - H_{-\varphi}) + \kappa \cdot H_\varphi$ with g satisfying the conditions of Stein (1981, Theorem 1), namely $g : \mathbb{R}^{\text{N}} \to \mathbb{R}^{\text{N}}$ is almost differentiable in Stein's sense and $\text{E}_{\boldsymbol{\alpha}} \sum_{n=1}^{N} |\nabla_{Y_n} g_n(\mathbf{Y})| < \infty$, H_φ is the Heaviside function at φ applied componentwise to $\mathbf{Y}$ and $\kappa > 0$ is the height of the jump of $\tilde{g}$ at its discontinuity points $\pm\varphi$. Then*

$$\begin{aligned}
\text{E}_{\boldsymbol{\alpha}}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2 &= N + \text{E}_{\boldsymbol{\alpha}}\left\{\|\tilde{g}(\mathbf{Y})\|_2^2 + 2\sum_{n=1}^{N} \nabla_{Y_n} g_n(\mathbf{Y}) + 2\kappa \sum_{n=1}^{N} (\delta_{-\varphi}(Y_n) + \delta_{\varphi}(Y_n))\right\} \\
&= N + \text{E}_{\boldsymbol{\alpha}}\left\{\|\tilde{g}(\mathbf{Y})\|_2^2 + 2\sum_{n=1}^{N} \nabla_{Y_n} g_n(\mathbf{Y})\right\}
\end{aligned}$$

$$+2\kappa \sum_{n=1}^{N} (\phi(\varphi - \alpha_n) + \phi(-\varphi - \alpha_n)). \tag{10}$$

Proof: for each $n \in \{1, \ldots, N\}$, from Stein (1981)

$$\mathrm{E}_{\boldsymbol{\alpha}}(Y_n + \tilde{g}_n(\mathbf{Y}) - \alpha_n)^2 \;=\; 1 + \mathrm{E}_{\boldsymbol{\alpha}}\{\tilde{g}_n^2(\mathbf{Y}) + 2\nabla_{Y_n}\tilde{g}_n(\mathbf{Y})\}.$$

And the last expression is $\nabla_{Y_n}\tilde{g}_n(\mathbf{Y}) = \nabla_{Y_n} g_n(\mathbf{Y}) + \kappa(\delta_{-\varphi}(Y_n) + \delta_{\varphi}(Y_n))$ since, integrating by parts,

$$\mathrm{E}_{\boldsymbol{\alpha}}\{(Y_n - \alpha_n)H_{\varphi}(Y_n)\} \;=\; -\phi(y - \alpha_n)H_{\varphi}(y)|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \phi(y - \alpha_n)\delta_{\varphi}(y)dy,$$

where the Gaussian density ϕ belongs to Schwartz space of rapidly decreasing functions, so that the last integral term is well defined and equal to $\phi(\varphi - \alpha_n)$. □

The last term with the Dirac measure in (10) cannot be estimated empirically with no bias however. Indeed for any Gaussian datum y_n, we have $\delta_{\varphi}(y_n) = 0$ with probability one, while $\mathrm{E}_{\alpha_n}\{\delta_{\varphi}(Y_n)\} = \phi(\varphi - \alpha_n) > 0$. We propose to estimate this term by replacing α_n in $\phi(\varphi - \alpha_n)$ with estimates $\tilde{\boldsymbol{\alpha}}$ that achieve minimax or asymptotically minimax risk (Donoho and Johnstone 1994), for instance

$$\tilde{\boldsymbol{\alpha}} = \eta_{\varphi_N}^{(\mathrm{hard})}(\mathbf{Y}) \quad \text{with} \quad \varphi_N = \sqrt{2\log N}. \tag{11}$$

Arguing heuristically, one expects that for large N, the formula will provide a nearly unbiased estimate of the risk, as shown on Figure 2 for a sparse sequence of length $N = 128$ with 33 non-zero coefficients. On this figure, the true loss of $\hat{\boldsymbol{\alpha}}_{\lambda,\nu}$ defined by (10) is estimated with the extension of SURE for three values of $\nu \in \{0, 0.6, 1\}$.

In practice, we select the two hyperparameters of the posterior mode estimate $\hat{\boldsymbol{\alpha}}_{\lambda,\nu}$ defined in (6) by solving

$$\min_{\lambda>0,\ \nu>0} \mathrm{SURE}(\lambda, \nu),$$

where, from Theorem 1,

$$\begin{aligned} \mathrm{SURE}(\lambda,\nu) \;=\; & N + \sum_{n=1}^{N} (\hat{\alpha}_{n,\lambda,\nu} - y_n)^2 + 2\sum_{n=1}^{N} \nabla_{Y_n} g_n(y_n) \\ & + 2\kappa(\lambda;\nu) \sum_{n=1}^{N} (\phi(\varphi(\lambda;\nu) - \tilde{\alpha}_n) + \phi(-\varphi(\lambda;\nu) - \tilde{\alpha}_n)), \end{aligned} \tag{12}$$

with threshold $\varphi(\lambda;\nu)$ given by (8) and discontinuity jump $\kappa(\lambda;\nu)$ given by (9) for $\nu \leq 1$ and equal to zero for $\nu > 1$, $\tilde{\boldsymbol{\alpha}}$ given by (11) and gradient $\nabla_{Y_n} g_n(Y_n) = (1 + \lambda\nu(\nu-1)|\hat{\alpha}_{\lambda,\nu}(Y_n)|^{\nu-2})^{-1} \cdot 1_{[\varphi(\lambda;\nu),\infty)}(|Y_n|) - 1$. The Monte Carlo simulation of Section 2.3 reveals good performance of SURE to select both λ and ν, except

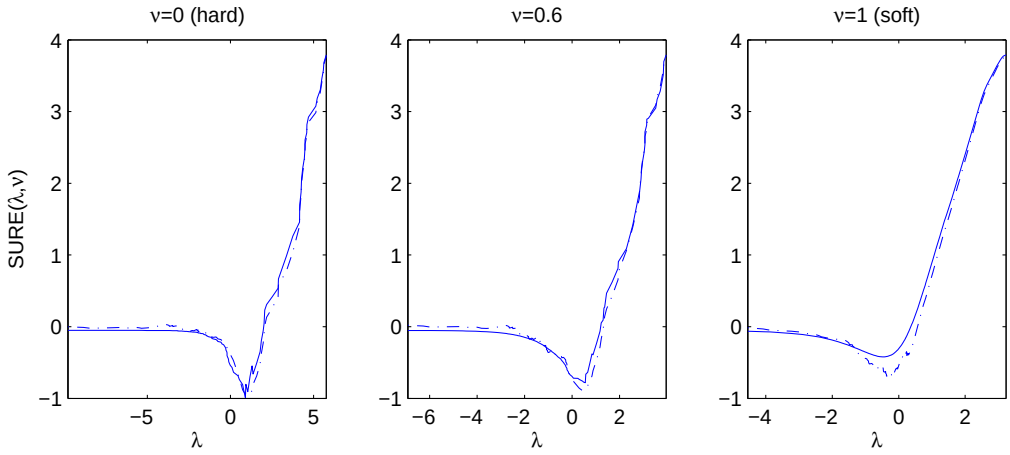

Figure 2: Illustration of the estimation of the true ℓ_2 loss (line) with the extension of SURE (dotted line) as a function of λ for three values of ν: 0 (hard), 0.6 and 1 (soft). Both axes are on a log-scale. For a given ν, the values of λ plotted on the horizontal axis are obtained from (8) using the order statistics $|y_{(n)}|$ for φ.

in situations of extreme sparsity as found by Donoho and Johnstone (1995, Section 2.4) in the case $\nu = 1$.

The case $\nu = 0$ (best subset variable selection) is interesting since, with $\varphi(\lambda; 0) = \sqrt{2\lambda} = \kappa(\lambda; 0)$ and $\nabla_{Y_n} g_n(Y_n) = 1_{[\varphi(\lambda;0),\infty)}(|Y_n|) - 1$,

$$\begin{aligned}
\mathrm{SURE}(\lambda, 0) &= N + \|\hat{\boldsymbol{\alpha}}_{\lambda,0} - \mathbf{Y}\|_2^2 + 2(\#n : \{\hat{\alpha}_{n,\lambda,0} \neq 0\} - N) \\
&\quad + 2\sqrt{2\lambda} \sum_{n=1}^{N} (\phi(-\sqrt{2\lambda} - \tilde{\alpha}_n) + \phi(\sqrt{2\lambda} - \tilde{\alpha}_n)) \\
&= \mathrm{AIC}(\lambda) + 2\sqrt{2\lambda} \sum_{n=1}^{N} (\phi(-\sqrt{2\lambda} - \alpha_n) + \phi(\sqrt{2\lambda} - \alpha_n)) - N
\end{aligned}$$

reveals the bias of AIC for estimating the risk.

2.2.2 Information criterion

Alternatively, we propose an information criterion that not only applies to the canonical model (1), but also applies to more general models (5). It is based on the two following properties.

Property 1 *Suppose* $Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(c\alpha_n, 1)$, *where* c *is known (e.g., equal to one). Consider the Subbotin posterior mode estimate* $\hat{\boldsymbol{\alpha}}_{\lambda,\nu}$ *solution to*

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{Y} - c\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_\nu^\nu.$$

For any $\nu \in (0,1]$ and when the true model is $\alpha_n = 0$ for $n = 1, \ldots, N$, then the estimated model is consistent,

$$\mathrm{P}(\hat{\boldsymbol{\alpha}}_{\lambda_N(\nu),\nu} = \mathbf{0}) \to 1,$$

provided $\lambda_N(\nu)$ satisfies $\varphi(\lambda_N(\nu)/c^\nu; \nu) = \sqrt{2 \log N}$, where the threshold $\varphi(\lambda; \nu)$ is given in (8). Its explicit expression is

$$\lambda_N(\nu) = c^\nu \left(\frac{\sqrt{2 \log N}}{2 - \nu} \right)^{2-\nu} (2(1-\nu))^{1-\nu}. \tag{13}$$

This result is a direct consequence of:

- the thresholding property (7) that $\hat{\boldsymbol{\alpha}}_{\lambda,\nu} = \mathbf{0}$ if and only if $\varphi(\lambda; \nu) \geq \varphi_{\mathbf{Y}} := \max_{n=1,\ldots,N}(|Y_n|)$, where $\varphi(\lambda; \nu)$ is the threshold (8);
- the universal rule (Donoho and Johnstone 1994) that $\varphi_N = \sqrt{2 \log N}$ controls the extremal behavior of N i.i.d. standard Gaussian,

so that $\mathrm{P}(\hat{\boldsymbol{\alpha}}_{\lambda_N(\nu),\nu} = \mathbf{0}) = \mathrm{P}(\varphi_N \geq \varphi_{\mathbf{Y}}) \to 1$ when $Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$.

More than a bound φ_N, the following property derives the distribution of the threshold $\varphi_{\mathbf{Y}}$ that sets all estimated coefficients to zero for a given sample $\mathbf{Y}$, based on an asymptotic pivot for $\varphi_{\mathbf{Y}}$ under the assumptions of Property 1.

Property 2 *Suppose $Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(c\alpha_n, 1)$, where c is known. Let $G_0(x) = \exp(-\exp(-x))$ be the Gumbel distribution, $d_N = \varphi_N - (\log\log N + \log 4\pi - 2\log 2)/(2\varphi_N)$ and $\varphi_N = \sqrt{2 \log N}$ be normalizing constants. Then an asymptotic pivot for $\varphi_{\mathbf{Y}} := \max_{n=1,\ldots,N}(|Y_n|)$ is*

$$\varphi_N \left(\varphi_{\mathbf{Y}} - d_N \right) \to_d G_0$$

when the true model is $\alpha_n = 0$ for $n = 1, \ldots, N$.

The proof is given in Appendix D. This pivot gives the asymptotic distribution, $\pi_\varphi(\varphi) = G'(\varphi)$ with $G(\varphi) = G_0(\varphi_N(\varphi - d_N))$, of the threshold φ to reconstruct the true zero sequence with a probability tending to one. Equivalently, since $\varphi(\lambda; \nu)$ in (8) is strictly increasing in λ for a given ν, the pivot gives the asymptotic distribution of the penalty $\lambda \mid \nu$: from Property 1, $F_{\lambda|\nu}(\lambda) = G_0(\varphi_N(\varphi(\lambda/c^\nu; \nu) - d_N))$.

The universal threshold φ_N has nice statistical properties. For wavelet smoothing for instance, Donoho, Johnstone, Kerkyacharian, and Picard (1995) showed that φ_N provides nearly minimax results for a class of loss functions and smoothness classes. Based on this property, we use the Gumbel-based prior for $\lambda \mid \nu$ to estimate non-zero sequences belonging to the considered smoothness classes. Because it is based on the universal rule, we call it the universal prior. Other priors could be employed, but the Gumbel-based prior has the property to match the distribution of the sample-based threshold $\varphi_{\mathbf{Y}}$ under the null model that all coefficients are zero.

Hence, assuming a prior π_ν for ν, Bayes theorem leads to the posterior distribution of $(\boldsymbol{\alpha}, \lambda, \nu)$ given the data $\mathbf{Y}$. We can then take its expectation or mode to estimate and select jointly $\boldsymbol{\alpha}$ and (λ, ν). We consider in this paper the former alternative that has the advantage to threshold. The posterior mode estimate is defined by taking the negative posterior likelihood and minimizing it as a joint optimization problem, which defines the following information criterion derived in Appendix A.

Definition 1 *Suppose* $Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(c\alpha_n, 1)$ *for* $n = 1, \ldots, N$, *and* c *is known. The sparsity* ℓ_ν *information criterion for estimation of* $\boldsymbol{\alpha}$ *and selection of* (λ, ν) *is*

$$\begin{aligned} \mathrm{SL}_\nu\mathrm{IC}(\boldsymbol{\alpha}, \lambda, \nu) &= \frac{1}{2}\|\mathbf{Y} - c\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_\nu^\nu - \frac{N}{\nu}\log\lambda + N\log\Gamma(1+\frac{1}{\nu}) \\ &\quad - \log\pi_{\lambda|\nu}(\lambda \mid \nu; \tau_N(\nu)) - \log\pi_\nu(\nu), \end{aligned} \tag{14}$$

where $\pi_{\lambda|\nu}(\lambda \mid \nu; \tau) = F'_{\lambda|\nu}(\lambda; \tau)$ *with* $F_{\lambda|\nu}(\lambda; \tau) = G_0(\varphi_N(\varphi(\frac{\lambda}{c^\nu}; \nu)/\tau - d_N))$, *and* τ *is calibrated to* $\tau_N(\nu) = \nu\varphi_N^2/(N(2-\nu))$ *to match the asymptotic model consistency of Property 1 when* $\boldsymbol{\alpha} = \mathbf{0}$ *is the true sequence.*

In practice, in the spirit of AIC and BIC, one minimizes $\mathrm{SL}_\nu\mathrm{IC}$ to select the hyperparameters (λ, ν) and estimate the sequence $\boldsymbol{\alpha}$. We know from Section 2.1 that the solution in $\boldsymbol{\alpha}$ is unique with probability one for a given set of hyperparameters. The following property states the uniqueness in λ selected with $\mathrm{SL}_\nu\mathrm{IC}$ for a given $\boldsymbol{\alpha}$ and ν. We will see in Section 5 that the choice of the universal prior for $\boldsymbol{\alpha}$ will also guarantee asymptotic minimaxity of the estimator with $\mathrm{SL}_\nu\mathrm{IC}$.

Property 3 *For a given sequence* $\boldsymbol{\alpha}$ *and a given hyperparameter* $\nu \in (0, 1]$, *the minimum of* $\mathrm{SL}_\nu\mathrm{IC}(\boldsymbol{\alpha}, \lambda, \nu)$ *in* λ *exists and is unique.*

Proof: penalty λ and threshold φ are one to one for a given $\nu \in (0, 1]$ from (8). The information criterion (14) writes up to a constant as a function of φ as

$$\mathrm{SL}_\nu\mathrm{IC}(\varphi) = \frac{\{2(1-\nu)\}^{1-\nu}}{(2-\nu)^{2-\nu}}\varphi^{2-\nu}\|\boldsymbol{\alpha}\|_\nu^\nu - \frac{N}{\nu}(2-\nu)\log\varphi + \exp(-\varphi_N(\frac{\varphi}{\tau_N} - d_N)) + \varphi_N(\frac{\varphi}{\tau_N} - d_N).$$

The function is strictly convex and $\lim_{\varphi\to 0}\mathrm{SL}_\nu\mathrm{IC}(\varphi) = \lim_{\varphi\to\infty}\mathrm{SL}_\nu\mathrm{IC}(\varphi) = \infty$ on the open interval $(0, \infty)$, therefore admits a unique minimum. □

2.2.3 Other approaches

Empirical Bayes approaches could also be used. The one employed by EBayesThresh, which maximizes the marginal likelihood over the hyperparameters, would require here an expression for, or at least fast computation of, the convolution of the Subbotin with the Gaussian. Instead of integrating with respect to the prior, a method of moment-based empirical Bayes approach would use the tractable first two even moments $\mathrm{E}\alpha_n^r = \lambda^{-r/\nu}\Gamma((r+1)/\nu)/\Gamma(1/\nu)$ of the symmetric Subbotin distribution to select the hyperparameters. We tried this approach with little success, owing to the fact that higher moments are not taken into account.

2.3 Gaussian Monte Carlo simulation

2.3.1 Direct sequence estimation

We consider the 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 taken in $\{3, 4, 5, 7\}$. Table 1 reports ℓ_1 and ℓ_2 losses, as well as type I and type II errors. We observe empirically that the posterior mode Subbotin estimator is competitive with the posterior median EBayesThresh estimator, which possesses optimal asymptotic rates for wide classes of sparse sequences. And both outperform estimators such as False Discovery Rate, soft- and hard-shrinkage using SURE and universal rule, based on the results reported in Johnstone and Silverman (2004, Table 1). Looking at the average number of type I and type II errors, we observe that SL_νIC tends to select a higher threshold than SURE and EBayesThresh. We also observe empirically that the selection of the two hyperparameters of the Subbotin(λ, ν) posterior mode estimator is good with SL_νIC when the sequence is sparse and becomes better with SURE when the sequence becomes more dense, as observed by Donoho and Johnstone (1995, Section 2.4) in the case $\nu = 1$. This suggests a hybrid method for the selection of the hyperparameters to take advantage of both. Finally the ℓ_1 performance measure is favorable to the posterior mode estimator.

2.3.2 Wavelet smoothing

Gaussian orthonormal wavelet smoothing falls back into the canonical model (1). We consider the nonparametric regression problem of estimating a function μ sampled with noise at N equispaced locations t_n according to

$$Y_n = \mu(t_n) + \epsilon_n,$$

where the ϵ_n are independent Gaussian $\mathrm{N}(0, \sigma^2)$. Note that the equispaced assumption can be relaxed by employing isometric wavelets (Sardy, Percival, Bruce, Gao, and Stuetzle 1999), or equivalently warped wavelets (Kerkyacharian and Picard 2004). The standard deviation of the noise can be well estimated by the median absolute deviation of the least squares fine scale wavelet coefficients at the highest level (Donoho and Johnstone 1995), so we assume in the following that $\sigma = 1$. Wavelet-based smoothers assume that μ can be well represented by a linear combination of approximation ϕ and fine scale ψ wavelets. Standard wavelets are a set of orthonormal multi-resolution functions that are locally supported and indexed by a location parameter k and a scale parameter j. A father wavelet ϕ such that $\int_0^1 \phi(t)dt = 1$ generates $P_0 = 2^{j_0}$ approximation wavelets by means of the dilation and translation relation $\phi_{j_0,k}(t) = 2^{j_0/2}\phi(2^{j_0}t - k)$, $k = 0, 1, \ldots, 2^{j_0} - 1$; they capture the coarse features of the signal. Similarly, a mother wavelet ψ such that $\int_0^1 \psi(t)dt = 0$ generates $N - P_0$ fine scale wavelets $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$, $j = j_0, \ldots, J$; $k = 0, 1, \ldots, 2^j - 1$,

Table 1: Gaussian Monte Carlo simulation. Average total squared (ℓ_2 loss) and absolute (ℓ_1 loss) errors, average number of type I and II errors of EBayesThresh (Laplace and Cauchy-like γ) and the Subbotin(λ, ν) posterior mode estimator with hyperparameters selected either with SL$_\nu$IC or SURE on a mixed signal of length 1000. In **bold**, the best between all methods for each loss.

Number nonzero	5				50				500			
Value nonzero	3	4	5	7	3	4	5	7	3	4	5	7
EBayesThresh												
Laplace (w, a)												
ℓ_2 loss	**35**	**33**	**19**	**9**	**211**	**154**	**102**	72	856	873	782	661
ℓ_1 loss	**13**	11	8	**5**	95	74	59	49	709	721	620	502
Type I	2	1	1	0.5	16	12	8	4	500	500	310	98
Type II	3	1	0.3	0	14	3	0.6	0	0	0	0	0
Cauchy w												
ℓ_2 loss	37	37	20	**9**	266	174	105	77	923	898	828	745
ℓ_1 loss	**13**	11	**7**	**5**	102	73	58	52	703	683	656	628
Type I	0.2	0.3	0.3	0.4	3	6	7	7	500	500	500	500
Type II	4	2	0.4	0	26	6	0.6	0	0	0	0	0
Posterior mode (λ, ν)												
SL$_\nu$IC												
ℓ_2 loss	38	37	**19**	**9**	354	293	130	**60**	**848**	830	835	861
ℓ_1 loss	**13**	**10**	**7**	**5**	125	88	**53**	**43**	634	653	670	697
Type I	0.2	0.3	0.2	0.2	0.3	0.4	0.3	0.5	268	329	361	400
Type II	4	2	0.5	0	37	17	3	0	4	0	0	0
SURE												
ℓ_2 loss	38	37	28	26	231	165	110	97	1243	**798**	**604**	**535**
ℓ_1 loss	16	14	13	12	**95**	**71**	59	56	**590**	**468**	**430**	**410**
Type I	5	4	3	3	9	7	6	5	18	14	10	4
Type II	2	1	0.2	0	18	5	0.8	0	94	19	2	0

where $J = \log_2(N) - 1$. Because they are locally supported and orthogonal to polynomials, only a few fine scale wavelets are necessary to approximate μ well: this is known as the *sparse wavelet representation*. Moreover, while the number $N_j = 2^j$ of wavelets within level j increases exponentially with j, the number of fine scale wavelets needed to reproduce local features typically decreases with j: the sparsity models should therefore be *level dependent*. In other words, assuming μ expands on N orthonormal wavelets, $\mu(t) = \sum_{k=0}^{2^{j_0}-1} \alpha_{0k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J}\sum_{\kappa=0}^{N_j-1} \alpha_{j,k}\psi_{j,k}(t)$, the proportion of (near) non-zero wavelet coefficients $\alpha_{j,k}$ is small, and the larger j the smaller the proportion.

To estimate the wavelet coefficients $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$ from the data, an orthonormal matrix $[X_0\ X]$ can be extracted from the continuous expansion to write the sampled $\boldsymbol{\mu} = (\mu(t_1), \ldots, \mu(t_N))$ as $\boldsymbol{\mu} = X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha}$, where X_0 is the $N \times P_0$ matrix of approximation wavelets, X is the $N \times (N - P_0)$ matrix of fine scale wavelets, and $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$ are the corresponding coefficients. Owing to the fact that the ℓ_2 loss is isometric to an orthonormal transform and that the wavelet matrix $[X_0\ X]$ is orthonormal, wavelet smoothing falls back into to the canonical model (1), with the difference that the sparse estimation is employed levelwise, so that hyperparameters (λ_j, ν_j) are selected independently at each level $j = j_0, \ldots, J$. Using the Monte Carlo simulation of Donoho and Johnstone (1994) with signal-to-noise ratio equal to seven,

we adaptively select the smoothness and the sparsity levelwise with:

- six methods for Waveshrink: BIC and SURE for $\nu = 0$ (hard), SL_1IC and SURE for $\nu = 1$ (soft), and SL_νIC and SURE for free $\nu \in (0, 1]$;
- two methods for EBayesThresh: Laplace (with the default value $a_j = 0.5$) and Cauchy-like prior.

When using SLIC levelwise, the prior π_{λ_j} of Definition 1 is used with $N = N_j = 2^j$ the number of coefficients at level j. We draw the following conclusions from the results of the simulation reported in Table 2:

- Whether using the information criterion SLIC or the extension of SURE (12), selecting $\nu \in [0, 1]$ is often better than, or at least as good as, fixing it to $\nu = 0$ or $\nu = 1$. The gain of additional flexibility leads to a gain of goodness-of-fit showing good behavior of the proposed selections of the two hyperparameters;
- SLIC and SURE lead to comparable performance for wavelet smoothing;
- ℓ_ν penalized least squares is competitive with EBayesThresh (Johnstone and Silverman 2005) that works particularly well in this setting.

3 Generalized Subbotin(λ, ν) posterior mode estimate

We now extend the methodology to the general model (5) for a class of continuous or discrete distributions F and matrices $[X_0\ X]$ with $Q_0 + Q$ columns; Q is the number of covariates or the number of basis functions. Let the likelihood be $l_N(\boldsymbol{\mu}, \psi; \mathbf{Y}) = \sum_n l(\mu_n, \psi; Y_n)$ with $l(\mu_n, \psi; Y_n) = \log F(Y_n; \mu_n, \psi)$; we write Ω_μ for the domain of l, so the domain of l_N is the product space Ω_μ^N. Let the location parameters of F be $\boldsymbol{\mu} = X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha}$; we write $\Gamma_{\boldsymbol{\alpha}_0,\boldsymbol{\alpha}}$ for the corresponding domain of the coefficients $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$. Hence the Subbotin$(\lambda, \nu)$ posterior mode estimate solves

$$\min_{(\boldsymbol{\alpha}_0,\boldsymbol{\alpha})\in\Gamma_{\boldsymbol{\alpha}_0,\boldsymbol{\alpha}}} -l_N(X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha}, \psi; \mathbf{Y}) + \lambda\|\boldsymbol{\alpha}\|_\nu^\nu - \frac{Q}{\nu}\log\lambda + Q\log\Gamma(1+\frac{1}{\nu}), \quad (15)$$

where we do not penalize parameters $\boldsymbol{\alpha}_0$, as it is done for instance with the intercept or father wavelets in regression, and where X must have rescaled columns as described in Section 3.2 below.

Table 2: Results of Monte Carlo simulation to compare wavelet smoothers and hyperparameter(s) selection (average MISE $\times 100$). From left to right: BIC and SURE levelwise for $\nu = 0$, SL_1IC and SURE for $\nu = 1$, SL_νIC and SURE for $\nu \in (0, 1]$, EBayesThresh with Laplace and Cauchy priors. (least asymmetric wavelet of order 8 with $j_0 = 4$ are used.)

	Waveshrink levelwise						EBayesThresh levelwise	
	ℓ_0 (hard)		ℓ_1 (soft)		ℓ_ν			
N	BIC	SURE	SL_1IC	SURE	SL_νIC	SURE	Laplace	Cauchy
blocks								
256	73	77	68	66	64	68	63	65
1024	39	37	43	36	34	35	32	32
4096	18	15	21	16	15	15	14	13
bumps								
256	88	124	75	83	90	92	87	91
1024	40	47	52	44	44	40	40	39
4096	17	15	18	16	15	14	14	13
heavisine								
256	35	22	23	22	22	22	22	22
1024	19	11	8.9	9.5	9.5	9.9	8.7	8.6
4096	10	3.6	3.6	3.7	3.3	3.8	3.1	3.1
Doppler								
256	38	49	69	52	45	47	49	50
1024	21	19	25	19	19	18	18	18
4096	10	4.9	8.1	6.4	5.3	5.0	4.5	4.3
zero								
256	29	15	6	11	9	12	8	8
1024	16	4.1	1.7	3.8	2.8	4.3	2.4	2.3
4096	8.5	1.1	0.4	1.0	0.8	1.2	0.6	0.6

3.1 Existence, uniqueness and thresholding

We establish conditions for existence, uniqueness and thresholding of the Subbotin posterior mode estimate (15). We also prove the important monotonicity of the threshold $\varphi(\lambda; \nu)$ for a given ν. Proofs are postponed to Appendix B. In the following $\mathrm{K}(A)$ stands for kernel of the matrix A and $\mathrm{Rg}(A)$ for its range.

Theorem 2 *(Existence) Suppose that in (15) each univariate likelihood $-l(\cdot, \psi; Y_n)$ contributing to l_N is a continuous and finite-valued function on the interval $\Omega_\mu \subseteq \mathbb{R}$, that $-l$ is coercive (i.e., $\lim_{\mu \to \inf \Omega_\mu} -l(\mu, \psi; Y_n) = \lim_{\mu \to \sup \Omega_\mu} -l(\mu, \psi; Y_n) = +\infty$), and that $\mathrm{Rg}([X_0\ X]) \bigcap \Omega_\mu^N \neq \emptyset$. Then a solution $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_{\lambda,\nu}$ to (15) exists.*

A sufficient condition for $\mathrm{Rg}([X_0\ X]) \bigcap \Omega_\mu^N \neq \emptyset$ is that X_0 contains a column of one (i.e., an intercept). Indeed in that case, the inverse image $\Gamma_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}} = \{(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}) :$

$X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha} \in \Omega_\mu^N\}$ of Ω_μ^N, is not empty (for instance set all entries of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}$ to zero except that corresponding to the intercept column in X_0 which is set to any value in Ω_μ). For instance the Gaussian and Poisson distributions satisfy the conditions of Theorem 2. For distributions that are not negative log-convex, a link function can make it convex.

Uniqueness of the estimate cannot be guaranteed for any X matrix when $\nu < 1$ because the Subbotin penalty is not convex. The case $\nu = 1$, studied in detail by Osborne, Presnell, and Turlach (2000) for the Gaussian distribution, still does not guarantee uniqueness because the level set of the ℓ_1 norm is a subspace that can potentially intersect the null space of $[X_0\ X]$ in an infinite number of points, as stated in the following theorem.

Theorem 3 *(Uniqueness, case $\nu = 1$) Additionally to the assumptions of Theorem 2, suppose that $-l(\cdot, \psi; Y_n)$ is strictly convex on Ω_μ. If $\mathrm{K}([X_0\ X]) = \{\mathbf{0}\}$, then $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_\lambda$ is the unique strict minimizer of (15). Otherwise, let $\mathrm{K}_{(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_\lambda}([X_0\ X])$ be the affine space parallel to $\mathrm{K}([X_0\ X])$ going through $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_\lambda$ and let $L_{\hat{\boldsymbol{\alpha}}_\lambda}$ be the subspace of points $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}) \in \mathbb{R}^{\mathrm{P}_0+\mathrm{P}}$ such that $\boldsymbol{\alpha}$ belongs to the boundary of the ℓ_1 ball of $\mathbb{R}^{\mathrm{P}}$ of radius $\|\hat{\boldsymbol{\alpha}}_\lambda\|_1$. Then $\mathrm{K}_{(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_\lambda}([X_0\ X]) \bigcap \Gamma_{(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})} \bigcap L_{\hat{\boldsymbol{\alpha}}_\lambda}$ is the set of strict minimizers of (15).*

The following theorem states a general sparsity result for any X matrix, when $\nu = 1$ and when $\boldsymbol{\alpha}_0$ is known in (15). This sparsity property forms the basis for deriving the universal threshold and universal prior.

Theorem 4 *Suppose that $\nu = 1$ and, in addition to the existence assumptions of Theorem 2, $-l(\cdot, \psi; Y_n)$ is convex and $\boldsymbol{\alpha}_0$ is known and such that $(\boldsymbol{\alpha}_0, \mathbf{0}) \in \mathrm{int}(\Gamma_{(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})})$, the interior of $\Gamma_{(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})}$. Let $\mathbf{g}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}; \mathbf{Y})$ be the gradient of $-l_N(X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha}, \psi; \mathbf{Y})$ with respect to $\boldsymbol{\alpha}$ at $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$. Then, for any $(\boldsymbol{\alpha}_0, \mathbf{0}) \in \Gamma_{(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})}$, if λ is at least as large as $\lambda_{\mathbf{Y}} = \|\mathbf{g}(\boldsymbol{\alpha}_0, \mathbf{0}; \mathbf{Y})\|_\infty < \infty$, then the sparse estimate $(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\alpha}})_{\lambda_{\mathbf{Y}}} = (\boldsymbol{\alpha}_0, \mathbf{0})$ is a global minimum of (15).*

The assumption that $(\boldsymbol{\alpha}_0, \mathbf{0}) \in \mathrm{int}(\Gamma_{(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})})$ is a mild restriction since the sequence $\boldsymbol{\alpha}$ represents a deviation around the null model $\boldsymbol{\mu}_0 = X_0\boldsymbol{\alpha}_0$ which belongs to the interior of the domain of the objective function. Thresholding still holds when $\boldsymbol{\alpha}_0$ is estimated independently, as long as $(\hat{\boldsymbol{\alpha}}_0, \mathbf{0}) \in \mathrm{int}(\Gamma_{(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})})$.

The penalty parameter $\lambda = \lambda_{\mathbf{Y}}$ of Theorem 4 guarantees complete sparsity when solving the multivariate problem (15) for a general matrix X in the Gaussian case for $\nu = 1$, but to analyze how the hyperparameters λ and ν can progressively control the proportion of null entries in the posterior mode estimate we must study the case $X = I$ and $X_0\boldsymbol{\alpha}_0 = \alpha_0\mathbf{1}$, where α_0 is a fixed scalar. In this case, the multivariate posterior mode problem (15) separates into N univariate problems:

$$\min_{\alpha_n} -l(\alpha_0 + \alpha_n, \psi; Y_n) + \lambda|\alpha_n|^\nu - \log\left(\frac{\lambda^{1/\nu}}{2\Gamma(1+\frac{1}{\nu})}\right) \quad n = 1, \ldots, N. \tag{16}$$

We now set the conditions on the likelihood l to guarantee thresholding. We first give a definition of thresholding that also applies to asymmetrical distributions.

Definition 2 *(Shrinking and thresholding) Let $\hat{\alpha}_{\mathrm{MLE}}(Y)$ be a scalar maximum likelihood estimate. An estimator $\hat{\alpha}_{\lambda,\nu}(Y)$ indexed by the hyper-parameters (λ,ν) is a shrinking estimator if $|\hat{\alpha}_{\lambda,\nu}(Y)| \leq |\hat{\alpha}_{\mathrm{MLE}}(Y)|$ for all Y, λ and ν, and is a thresholding estimator if for all Y there exists a lower $\varphi^{\mathrm{l}}(\lambda;\nu)$ and an upper $\varphi^{\mathrm{u}}(\lambda;\nu)$ thresholds such that*

$$\hat{\alpha}_{\lambda,\nu}(Y) = 0 \quad \textit{if and only if} \quad -\varphi^{\mathrm{l}}(\lambda;\nu) \leq \hat{\alpha}_{\mathrm{MLE}}(Y) \leq \varphi^{\mathrm{u}}(\lambda;\nu).$$

Thresholding induces a sparser estimate the larger the thresholds $\varphi^{\mathrm{l}}(\lambda;\nu)$ and $\varphi^{\mathrm{u}}(\lambda;\nu)$. Theorem 5 below establishes conditions for the posterior mode (16) to shrink and threshold, and proves the useful monotonicity of the lower and upper thresholds as a function of λ. To deal with both continuous and discrete distributions at once, we require the discrete distribution (originally defined for ordered data in the countable set Ω_Y) to be prolonged on the interval $\Omega_Y^I = [\inf \Omega_Y, \sup \Omega_Y]$ (e.g., for Poisson $\Omega_Y^I = [0,\infty)$ and $l(\alpha_0+\alpha,\psi;Y)$ exists as a function of Y in Ω_Y^I). The following theorem remains true when α_0 is unknown and estimated independently of α.

Theorem 5 *Suppose that α_0 is known and that, in addition to the existence assumptions of Theorem 2, $-l(\alpha_0+\cdot,\psi;Y)$ in (16) is differentiable and strictly convex with $-\ddot{l}_{\alpha\alpha}(\alpha_0+\cdot,\psi;Y) > 0$ on Γ_α. Then the posterior mode estimate (16) shrinks the MLE. If moreover the penalized likelihood (16) has at most one local maximum and one local minimum on either side of zero, and if $\dot{l}_\alpha(\alpha_0+\alpha,\psi;\cdot)$ is differentiable and monotone on Ω_Y^I for all $\alpha \in \Gamma_\alpha$, then it also thresholds the MLE. And the thresholds $\varphi^{\mathrm{l}}(\lambda;\nu)$ and $\varphi^{\mathrm{u}}(\lambda;\nu)$ are non-decreasing with λ for a given ν.*

Note that a unique solution is defined by taking the solution at zero when the two local minima have the same objective function (which happens with probability zero for continuous distributions). For Gaussian likelihood, we saw that the conditions of Theorem 5 are satisfied, that the entire thresholding function for $|Y_n| \geq \varphi^{\mathrm{l}}(\lambda;\nu) = \varphi^{\mathrm{u}}(\lambda;\nu)$ must be found numerically when $0 < \nu < 1$, and that the function has a jump $\kappa(\lambda;\nu)$ at the threshold. The threshold and jump values are given by (8) and (9) for the Gaussian likelihood with $\alpha_0 = 0$ (see Appendix C). More generally for distributions satisfying the assumptions of Theorem 5, the threshold and jump values $(\varphi^{\mathrm{l}}(\lambda;\nu),\kappa^{\mathrm{l}}(\lambda;\nu))$ and $(\varphi^{\mathrm{u}}(\lambda;\nu),\kappa^{\mathrm{u}}(\lambda;\nu))$ solve in $(\alpha,y) = (\kappa,\varphi)$ the system of equations

$$\left\{ \begin{array}{rcl} f(0;y) & = & f(\alpha,y) \\ \nabla_\alpha f(\alpha;y) & = & 0 \end{array} \right. , \tag{17}$$

where $f(\alpha,y) = -l(\alpha_0+\alpha,\psi;y) + \lambda|\alpha|^\nu$ is the penalized likelihood.

3.2 Information criterion and rescaling

With the prior π_φ derived by controlling the extremal behavior of $\max_{n=1,\ldots,N} |\hat{\alpha}_{\mathrm{MLE}}(Y_n)|$ using results from extreme value theory, and assuming the threshold $\varphi(\lambda;\nu)$ is monotone for a given ν based on Theorem 5, then the sparsity ℓ_ν information criterion

$$\begin{aligned}\mathrm{SL}_\nu\mathrm{IC}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha},\lambda,\nu) &= -l_N(X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha},\psi;\mathbf{Y}) + \lambda\|\boldsymbol{\alpha}\|_\nu^\nu - \frac{Q}{\nu}\log\lambda + Q\log\Gamma(1+\frac{1}{\nu}) \\ &\quad - \log\pi_{\lambda|\nu}(\lambda \mid \nu;\tau_Q(\nu)) - \log\pi_\nu(\nu)\end{aligned}$$

can be employed, where $\tau_Q(\nu)$ is calibrated to achieve asymptotic model consistency for a zero-sequence.

Importantly the columns of X_0 and X must be rescaled for the following two reasons. First since $\boldsymbol{\alpha}_0$ is not penalized in (15), the Q columns of the regression matrix X must be adjusted to remove collinearity with the columns of X_0; this is called mean-centering when $X_0 = \mathbf{1}$ is the intercept. Second, since the ℓ_ν penalty in (15) is isotropic, the columns of X must also be rescaled. Indeed the isotropic penalty intrinsically assumes equal variance in the estimation of the coefficients, in particular for the maximum likelihood estimate when $\lambda = 0$. The homoscedasticity of the MLE of the Gaussian canonical model (1) is in general no longer true when the matrix X is not orthonormal or when the distribution is not Gaussian. We adopt Σ-rescaling (Sardy 2008) based on the diagonal elements of the covariance matrix Σ of the MLE. For Gaussian data for instance, the rescaled matrix XD_Σ with $D_\Sigma^2 = \mathrm{diag}(\Sigma)$ has the required homoscedasticity property since the diagonal of the covariance matrix, $\mathrm{diag}((D_\Sigma X^{\mathrm{T}} X D_\Sigma)^{-1}) = I$, is constant. Another commonly used rescaling divides each column of X by its standard error, but it does not seem appropriate for isotropic penalties. With Σ-rescaling, it makes sense to use the isotropic Subbotin penalty $+\lambda\|\boldsymbol{\alpha}\|_\nu^\nu$ and to generalize the sparsity ℓ_ν information criterion to the non-canonical model.

4 Subbotin posterior mode, information criterion and rescaling in practice

The three situations considered below illustrate sparse estimation beyond the canonical model to the extend that we can solve exactly the corresponding posterior mode estimation problems.

4.1 Gaussian parametric regression

We consider the case when the regression matrix X in (15) is not the identity, nor an orthonormal wavelet matrix, but a matrix of measured covariates. Assuming Gaussian noise with variance one, the Subbotin posterior mode estimator solves

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_\nu^\nu - \frac{P}{\nu}\log\lambda + P\log\Gamma(1+\frac{1}{\nu}),$$

where response and covariates have been mean-centered to avoid dealing with an intercept, and Σ-rescaled as discussed in Section 3.2. For the selection of the hyper-parameters (λ, ν), consider the following proposition.

Proposition 1 *Suppose* $\mathbf{Y} = cX\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, *where* X *is the* $N \times P$ *matrix of* Σ*-rescaled covariates,* c *is known and* $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, I_N)$. *The sparsity* ℓ_ν *information criterion for estimation of* $\boldsymbol{\alpha}$ *and selection of* (λ, ν) *is defined as*

$$\begin{aligned} \mathrm{SL}_\nu\mathrm{IC}(\boldsymbol{\alpha}, \lambda, \nu) &= \frac{1}{2}\|\mathbf{Y} - cX\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_\nu^\nu - \frac{P}{\nu}\log\lambda + P\log\Gamma(1+\frac{1}{\nu}) \\ &\quad - \log\pi_{\lambda|\nu}(\lambda \mid \nu; \tau_P(\nu)) - \log\pi_\nu(\nu), \end{aligned}$$

where $\pi_{\lambda|\nu}(\lambda \mid \nu; \tau) = F'_{\lambda|\nu}(\lambda; \tau)$ *with* $F_{\lambda|\nu}(\lambda; \tau) = G_0(\varphi_P(\varphi(\lambda/c^\nu; \nu)/\tau - d_P))$, $d_P = \varphi_P - (\log\log P + \log 4\pi - 2\log 2)/(2\varphi_P)$, $\varphi_P = \sqrt{2\log P}$, *and* τ *is calibrated to* $\tau_P(\nu) = \nu\varphi_P^2/(P(2-\nu))$.

We show in Section 5 that this selection of the penalty provides $\sqrt{N}$-consistency for lasso ($\nu = 1$). Another possibility is to calibrate the prior with a diverging threshold $\varphi_N = \sqrt{2\log N}$, in the spirit of BIC.

Our application consists of $P = 8$ clinical measures on $N = 97$ men who were about to receive a radical prostatectomy (Tibshirani 1996). The goal is to employ a linear model and to select significant variables among the eight to predict a response value Y, the level of prostate specific antigen. Minimizing $\mathrm{SL}_\nu\mathrm{IC}$ over $\boldsymbol{\alpha}$ for given (λ, ν) entails solving a multivariate non-convex optimization that goes beyond the scope of this paper. Moreover the number of covariates $P = 8$ is too small to hope for a gain in prediction by selecting the second hyperparameter ν. So we illustrate this proposition in the lasso case $\nu = 1$. We consider six estimators: least squares (MLE), stepwise, three versions of lasso, and boosting (see Buehlmann and Hothorn (2007) for a review). The first version of lasso employs the information criterion $\mathrm{SL}_1\mathrm{IC}$, Σ-rescaling and a relaxation algorithm (Sardy, Bruce, and Tseng 2000) to solve lasso for a few λ's. The second and third versions of lasso (using `glmpath` function available in `R`) employ standard rescaling and selects λ based on BIC and AIC (Zou, Hastie, and Tibshirani 2007). The noise variance is estimated with the unbiased estimate using the residuals of the least squares fit. Table 3 reports the estimated coefficients for the six methods. In the last column we report the mean of prediction errors estimated by 500 repeated two-fold random partitioning of the data into training and test sets; the 500 values are visualized in the box plots of Figure 3. Note that the difference of prediction errors between paired-samples reveals even more significant (not reported here). Stepwise-BIC ($\nu = 0$) estimates the sparsest sequence and provides the best prediction. Boosting and Lasso-$\mathrm{SL}_1\mathrm{IC}$ come next with very similar coefficient estimates and predictive performance. Lasso-BIC and AIC do not perform as well.

Table 3: Prostate cancer data. Coefficient values estimated with six methods, and corresponding estimated prediction squared error $\widehat{\mathrm{PE}}$ (standard error 0.004).

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	$\widehat{\mathrm{PE}}$
MLE	0.587	0.454	-0.020	0.107	0.766	-0.105	0.045	0.005	0.630
Stepwise ℓ_0	0.552	0.509	0	0	0.666	0	0	0	**0.557**
Lasso-SL_1IC ℓ_1	0.532	0.397	-0.009	0.078	0.604	0	0.005	0.003	0.606
Boosting	0.532	0.405	-0.010	0.082	0.610	0	0.008	0.003	0.600
Lasso-BIC ℓ_1	0.516	0.346	0	0.051	0.567	0	0	0.002	0.619
Lasso-AIC ℓ_1	0.529	0.392	-0.008	0.075	0.601	0	0	0.002	0.618

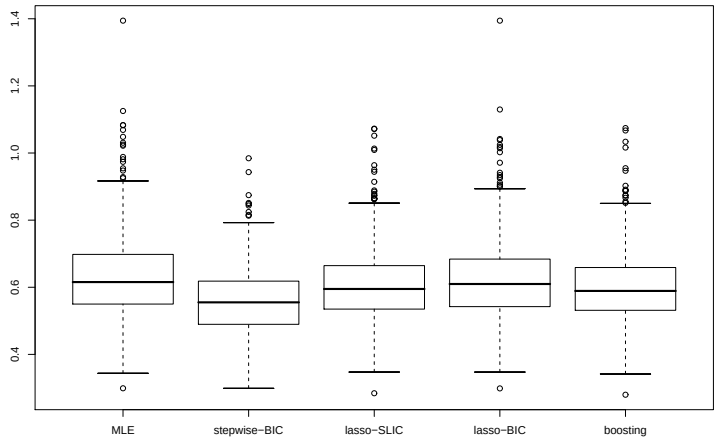

Figure 3: Prostate cancer data. Box plots of estimated prediction errors.

4.2 Poisson likelihood

The model of Johnstone and Silverman (2004) extends to Poisson data as

$$Y_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\alpha_0 + \alpha_n),$$

where α_0 is a known positive background intensity. For this distribution, Theorem 2 guarantees existence of the posterior mode estimate for all $\nu \leq 1$, and Theorem 3 guarantees uniqueness for $\nu = 1$ since $-l(\alpha, \psi; Y) = \alpha_0 + \alpha - Y \log(\alpha_0 + \alpha)$ is strictly convex on $\Gamma_\alpha = (-\alpha_0, \infty)$ and $\lim_{\alpha \to -\alpha_0} \alpha - Y \log(\alpha_0 + \alpha) = \lim_{\alpha \to +\infty} \alpha - Y \log(\alpha_0 + \alpha) = +\infty$ (perturbing null observations to a small positive ϵ). Finally since the Poisson posterior distribution has at most two inflection points on either side of zero and $\dot{l}_\alpha$ is monotone with respect to Y ($\ddot{l}_{\alpha Y} = 1/(\alpha_0 + \alpha) > 0$ for all $\alpha \in \Gamma_\alpha$), Theorem 5 guarantees thresholding.

To derive the information criterion for Poisson likelihood, consider first the case

$\nu = 1$: the posterior mode has a closed form expression

$$\hat{\alpha}_\lambda = \begin{cases} \frac{(\hat{\alpha}_{\mathrm{MLE}}+\varphi^{\mathrm{l}})_-}{1-\lambda} & \text{when} \quad Y \le \alpha_0 \\ \frac{(\hat{\alpha}_{\mathrm{MLE}}-\varphi^{\mathrm{u}})_+}{1+\lambda} & \text{when} \quad Y \ge \alpha_0 \end{cases},$$

where $\hat{\alpha}_{\mathrm{MLE}} = Y - \alpha_0$, and the lower and upper thresholds are $\varphi^{\mathrm{l}}(\lambda; 1) = \min(\alpha_0, \lambda\alpha_0)$ and $\varphi^{\mathrm{u}}(\lambda; 1) = \lambda\alpha_0$. The Poisson distribution has a positive skewness and the domain of Y has a bounded lower endpoint, so we consider the upper threshold to derive the universal prior and control the asymptotic behavior of a sample $\mathbf{Y} = (Y_1, \ldots, Y_N) \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\alpha_0)$. A well known result in extreme value theory states that the Poisson distribution, like other discrete distributions, does not belong to the domain of attraction of an extreme value distribution. To cope with discreteness, a possibility suggested by a referee is to use the fact that $\mathrm{P}(\|\mathbf{Y} - \alpha_0\|_\infty \le \varphi) \doteq \mathrm{P}(\|\mathbf{U}\|_\infty \le 1)$, where $U_n \overset{\text{i.i.d.}}{\sim} I\Gamma(\varphi + \alpha_0, 1/\alpha_0)$ (i.e., inverse Gamma), $n = 1, \ldots, N$. One can show that $I\Gamma(\alpha, \beta)$ is in the maximum domain of attraction of the Fréchet Φ_α distribution, but, for large α, here large $\varphi + \alpha_0$, our findings show that convergence is too slow for practical use as a prior for φ. Another possibility is to use Anscombe transform $\tilde{Y}_n = 2\sqrt{Y_n + 3/8} \dot{\sim} \mathrm{N}(2\sqrt{\alpha_n}, 1)$ to Gaussianize the data. We use instead the approximation $(Y_n - \alpha_0)/\sqrt{\alpha_0} \dot{\sim} \mathrm{N}(0, 1)$ which is good when α_0 is large to establish that

$$\mathrm{P}(\hat{\boldsymbol{\alpha}}_\lambda = \mathbf{0}) = \mathrm{P}(\|\frac{\mathbf{Y} - \alpha_0}{\sqrt{\alpha_0}}\|_\infty \le \frac{\varphi(\lambda; 1)}{\sqrt{\alpha_0}}) \approx \mathrm{P}(\|\mathbf{Z}\|_\infty \le \frac{\varphi(\lambda; 1)}{\sqrt{\alpha_0}}),$$

where $\mathbf{Z} = (Z_1, \ldots, Z_N)$ is a standard Gaussian sample. So the universal threshold is $\sqrt{\alpha_0}\varphi_N$ with $\varphi_N = \sqrt{2 \log N}$. More generally for all $\nu \le 1$, Theorem 5 guarantees that the upper threshold $\varphi^{\mathrm{u}}(\lambda; \nu)$ is strictly increasing in λ, so the universal penalty $\lambda_N(\nu)$ is uniquely defined as the solution in λ to $\varphi^{\mathrm{u}}(\lambda; \nu) = \sqrt{\alpha_0}\varphi_N$, which leads to the following information criterion for Poisson data.

Definition 3 *Suppose* $Y_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\alpha_0 + \alpha_n)$ *for* $n = 1, \ldots, N$. *The sparsity* ℓ_ν *information criterion for estimation of* $\boldsymbol{\alpha}$ *and selection of* (λ, ν) *is defined as*

$$\begin{aligned} \mathrm{SL}_\nu\mathrm{IC}(\boldsymbol{\alpha}, \lambda, \nu) \;=\; & \sum_{n=1}^{N} \alpha_0 + \alpha_n - y_n \log(\alpha_0 + \alpha_n) + \lambda\|\boldsymbol{\alpha}\|_\nu^\nu - \frac{N}{\nu}\log\lambda + N\log\Gamma(1 + \frac{1}{\nu}) \\ & - \log \pi_{\lambda|\nu}(\lambda \mid \nu; \tau_N(\nu)) - \log \pi_\nu(\nu), \end{aligned}$$

where $\pi_{\lambda|\nu}(\lambda \mid \nu; \tau) = F'_{\lambda|\nu}(\lambda; \tau)$ *with* $F_{\lambda|\nu}(\lambda; \tau) = G_0(\varphi_N(\varphi(\lambda; \nu)/(\tau\sqrt{\alpha_0}) - d_N))$, *and* τ *is calibrated to* $\tau_N(\nu) = \nu\varphi_N\lambda_N(\nu)(\varphi'(\lambda_N(\nu); \nu))^2/\sqrt{\alpha_0}/(N\varphi'(\lambda_N(\nu); \nu) + \nu\lambda_N(\nu)\varphi''(\lambda_N(\nu); \nu))$.

To test the Subbotin posterior mode estimate and the information criterion on Poisson data, we consider the following Monte Carlo simulation indexed by the underlying known background Poisson intensity $\alpha_0 \in \{1, 3, 10\}$, the number of nonzero

parameters $K \in \{10, 50\}$ and the 'standard deviation' $\sigma\sqrt{\alpha_0}$ with $\sigma \in \{5, 7, 10\}$ added or subtracted to α_0 to define the nonzero part of the sequence. For very low background $\alpha_0 \in \{1, 3\}$, we consider adding non-zero signal $\alpha_n = \alpha_0 + \sigma\sqrt{\alpha_0},\ n = 1, \ldots, K$, and for higher background $\alpha_0 = 10$, half of the K nonzero added signal is $\alpha_n = \alpha_0 + \sigma\sqrt{\alpha_0},\ n = 1, \ldots, K/2$ and the remaining is set to $\alpha_n = 0.01,\ n = K/2 + 1, \ldots, K$. EBayesThresh is applied to the Anscombe transformed Poisson data. The results of the simulations are reported in Table 4: as expected, the Subbotin posterior mode estimate is better than EBayesThresh on the Gaussianized data when the background intensity is low.

Table 4: Poisson Monte Carlo simulation. Average negative log-likelihood criterion for Anscombe-EBayesThresh and Subbotin(λ, ν) posterior mode estimate on a mixed signal of length 1000. In **bold**, the best between both.

Number nonzero	10			50		
Value nonzero: σ	5	7	10	5	7	10
Background: $x_0 = 1$						
Anscombe-EBayesThresh	969	916	833	765	527	**150**
Posterior mode (λ, ν)	**931**	**890**	**817**	**714**	**506**	153
Background: $x_0 = 3$						
Anscombe-EBayesThresh	-452	-537	-703	**-1142**	**-1588**	**-2342**
Posterior mode (λ, ν)	**-471**	**-547**	**-709**	-1135	-1583	**-2341**
Background: $x_0 = 10$						
Anscombe-EBayesThresh	-13204	**-13289**	**-13533**	**-13854**	-14392	-15183
Posterior mode (λ, ν)	**-13208**	**-13291**	-13499	**-13852**	**-14398**	**-15195**

4.3 Poisson wavelet smoothing

The burst and transient source experiment (BATSE) instruments on board of NASA's Compton Gamma Ray Observatory measure arrival times of high energy gamma rays. Using a partition of time, the data consists of counts of gamma rays in each bin; for details see Meegan, Fishman, Wilson, Paciesas, Pendleton, Horack, Brock, and Kouveliotou (1992). We focus on particular on the trigger 551 data used by Besbeas, De Feis, and Sapatinas (2004): the signal recorded during 0.94 seconds has length $N = 1024$, but as pulses occurred in the first half (until 0.47 seconds), we show only this half on Figure 4 to zoom on the relevant area.

To preserve the sharp features of the underlying signal, we employ a wavelet-based ℓ_1-penalized Poisson likelihood estimator (Sardy, Antoniadis, and Tseng 2004) defined as the solution to

$$\min_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}} -l(X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha}; \mathbf{Y}) + \sum_{j=j_0}^{J} \lambda_j \sum_{k=0}^{2^j-1} |\alpha_{j,k}|,$$

where l is the Poisson likelihood, $\mathbf{Y}$ are the Poisson counts and $[X_0\ X]$ is the Σ-rescaled Haar orthonormal wavelet matrix with $j_0 = 4$ levels and with associated coefficients $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha})$ (see Section 2.3.2 for details on wavelet matrices). To select the hyperparameters λ_j levelwise, we employ $\mathrm{SL}_1\mathrm{IC}$ with the prior π_{λ_j} of the previous section. Figure 4 shows the raw data (top left) and the estimated signal (bottom left): the two dips at times 0.22 and 0.24 are preserved, as well as a few bursts. The wavelet coefficients (top right) along with their corresponding hyperparameters λ_j selected by the information criterion (bottom right) are plotted on the right side of Figure 4. We observe that as the level j increases, the λ_j sequence increases, because the higher the level j, the sparser the corresponding wavelet sequence.

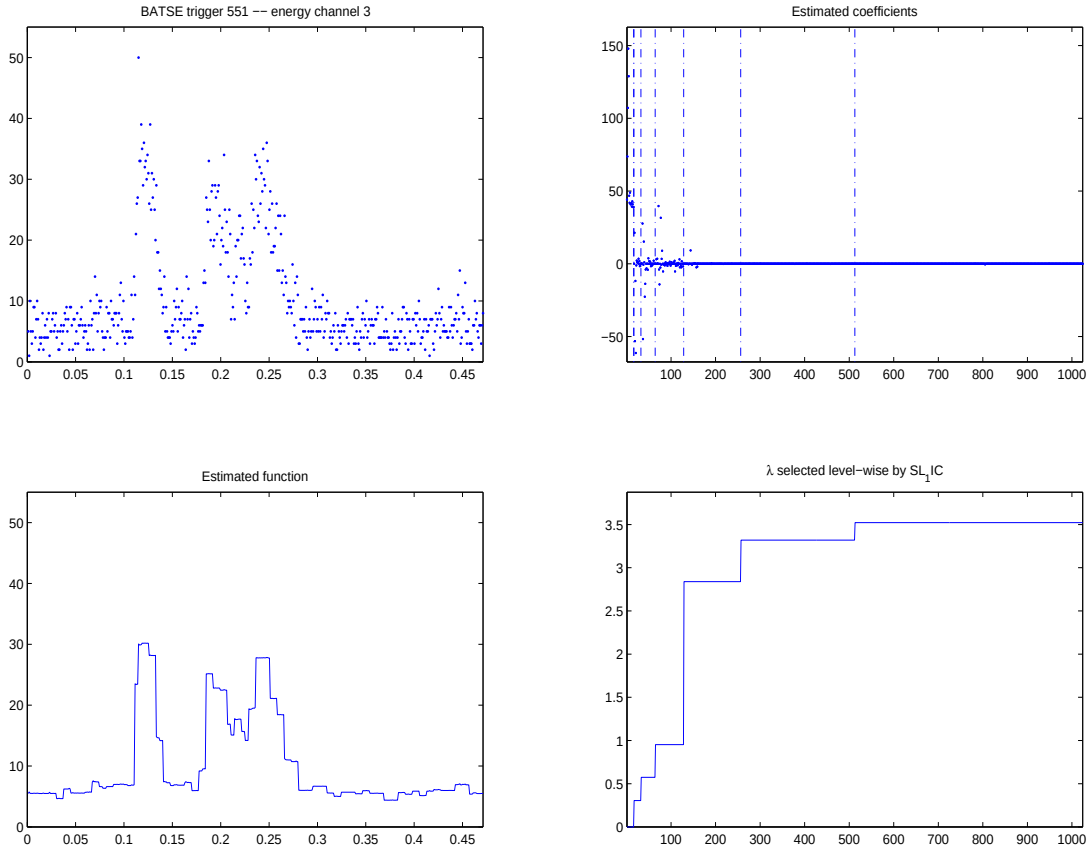

Figure 4: Poisson wavelet smoothing. Top left: BATSE trigger 551 data until 0.47 seconds. Bottom left: estimated intensities. Top right: estimated wavelet coefficient by level. Bottom right: hyperparameters λ_j selected levelwise with $\mathrm{SL}_1\mathrm{IC}$.

5 Asymptotic properties

We investigate some asymptotic properties of the Subbotin posterior mode estimator for Gaussian data when the hyperparameters are selected using the sparsity ℓ_ν information criterion. We consider first the parametric case and recall the definition of Fan and Li (2001) that a parametric sparse estimator is called asymptotically oracle if it identifies the correct model and if the non-zero coefficients estimates behave like standard maximum likelihood estimate (i.e., asymptotically unbiased and Gaussian with root-N convergence). The case $\nu = 1$ has been well studied (Knight and Fu 2000; Fan and Peng 2004; Meinshausen and Bühlmann 2006; Zou 2006). In particular lasso cannot simultaneously achieve the two properties at once. Either $\lambda(N)/\sqrt{N} \to \lambda_0 \geq 0$, then the convergence rate is root-N, but the model estimate cannot be consistent. Or $\lambda(N)/N \to \lambda_0 \geq 0$ and $\lambda(N)/\sqrt{N} \to \infty$, then the model estimate is consistent, but the convergence is slower than root-N. The selection of λ with the sparsity ℓ_1 information criterion of Proposition 1 falls in the first case.

Property 4 *Let $\tilde{X}$ be the original matrix of P mean-centered explanatory variables (P is fixed but large enough for $\exp(-P) \ll 1$) and let $X = \tilde{X} D_\Sigma$ be the rescaled matrix, where $D_\Sigma^2 = \mathrm{diag}(\Sigma)$ and $\Sigma = (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}$. We assume moreover that $X^{\mathrm{T}}X \to C$ as N tends to infinity, where C is a positive definite matrix. Then the SL_1IC-lasso estimate $\hat{\boldsymbol{\alpha}}_{\mathrm{SL_1IC}}$ is $\sqrt{N}$-consistent, as solution to*

$$\min_{\boldsymbol{\alpha},\lambda} \frac{1}{2}\|\mathbf{Y} - \sqrt{N}X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 - P\log\lambda - \log\pi_\lambda(\lambda;\tau_P),$$

where $\pi_\lambda(\lambda;\tau) = F'_{\lambda|\nu}(\lambda;\tau)$ with $F_\lambda(\lambda;\tau) = G_0(\varphi_P(\lambda/(\tau c) - d_P))$, $d_P = \varphi_P - (\log\log P + \log 4\pi - 2\log 2)/(2\varphi_P)$, $\varphi_P = \sqrt{2\log P}$, $c = \sqrt{N}$, τ is calibrated to $\tau_P = \varphi_P^2/P$.

Proof: The Σ-rescaled matrix has the property that $\mathrm{diag}((X^{\mathrm{T}}X)^{-1}) = I$ by definition of D_Σ: the least squares coefficients are homoscedastic with unit variance. Based on Property 1 with $c = \sqrt{N}$, the universal penalty is $\lambda_{P,N} = \sqrt{N}\varphi_P$. From Knight and Fu (2000), $\sqrt{N}$-consistency is therefore guaranteed if we can show that $\lambda_{\mathrm{SL_1IC}} = O(\sqrt{N})$. The $\mathrm{SL_1IC}$ first order optimality condition in λ is

$$\|\hat{\boldsymbol{\alpha}}_{\mathrm{SL_1IC}}\|_1 - \frac{P}{\lambda} - \frac{\pi'}{\pi}(\lambda;\tau_P) = 0 \tag{18}$$

with $\frac{\pi'}{\pi}(\lambda;\tau) = \varphi_P/(c\tau)\cdot(\exp(-\varphi_P(\lambda/(\tau c) - d_P)) - 1)$. Following similar derivations as Appendix A, the choice $\tau_P = \varphi_P^2/P$ guarantees that the universal penalty is selected when the true coefficients are null (i.e., $\|\hat{\boldsymbol{\alpha}}_{\mathrm{SL_1IC}}\|_1 = 0$) and the number of covariates is large enough for $\exp(-P) \ll 1$. Moreover (18) writes explicitly as

$$\|\hat{\boldsymbol{\alpha}}_{\mathrm{SL_1IC}}\|_1 + \frac{P}{\varphi_P\sqrt{N}} = \frac{P}{\lambda} + \frac{P\exp(\varphi_P d_P)}{\varphi_P\sqrt{N}}\exp(-\frac{P}{\varphi_P\sqrt{N}}\lambda), \tag{19}$$

which right hand side is strictly decreasing in $\lambda \in (0, \infty)$ from $+\infty$ to 0. Since τ has been calibrated for $\lambda_{P,N} = \sqrt{N}\varphi_P$ to be the root when $\|\hat{\boldsymbol{\alpha}}_{\mathrm{SL_1IC}}\|_1 = 0$, then the root to (19) is bounded by $\lambda_{P,N} = O(\sqrt{N})$. □

Since lasso cannot be asymptotically oracle, Zou (2006, equation (4)) proposed the adaptive lasso which uses a weighted ℓ_1 penalty, $+\lambda w_p|\alpha_p|$, with weights w_p inversely proportional to the magnitude of the corresponding coefficient, so as to bias less large coefficients; in practice the weights must be estimated, for instance by least squares. He then showed that the adaptive lasso is asymptotically oracle. The Subbotin posterior mode aims at a similar behavior by adapting the penalty with a single parameter ν rather than weighting each ℓ_1 term with a random variable: the smaller ν, the less bias is introduced when shrinking non-zero coefficients. In fact a potential algorithm to find the non-zero Subbotin solutions is to use the adapted lasso iteratively with weights $w_p = \nu/|\alpha_p^{(i)}|^{1-\nu}$, where $\alpha_p^{(i)}$ is the current iterate. In that sense the adaptive lasso can be seen as an approximation to the Subbotin posterior mode.

The advantage of the adaptive selection of ν becomes apparent for large P, in particular in the nonparametric situation where P grows with N. For the canonical model and in Gaussian wavelet smoothing, Antoniadis and Fan (2001) showed that penalized least squares is comparable with the oracle estimator within a logarithmic factor for the universal threshold and for a class of penalties. Since the ℓ_ν penalty satisfies the conditions of their theorem, the same result applies to the Subbotin posterior mode estimator if the penalty is set to the conservative universal penalty $\lambda_N(\nu)$ in (13) such that $\varphi(\lambda_N(\nu); \nu) = \sqrt{2 \log N}$ for any ν. The sparsity ℓ_ν information criterion seeks a less conservative thresholding by selecting and adaptive λ and ν for a better fit. Since $\mathrm{SL}_\nu\mathrm{IC}$ is calibrated with respect to the universal threshold (22), then the selected penalty $\lambda_{\mathrm{SLIC}}(\nu)$ is always bounded by the universal penalty $\lambda_N(\nu)$ for any given ν. Moreover the universal penalty is asymptotically minimax (Johnstone and Silverman 2004; Donoho, Johnstone, Kerkyacharian, and Picard 1995), and the following property proved in Appendix E states that the selected penalty $\lambda_{\mathrm{SLIC}}(\nu)$ converges in probability to the universal penalty.

Property 5 *Suppose that ν is fixed and that the true sequence is ℓ_ν-bounded (i.e., $\|\boldsymbol{\alpha}\|_{N,\nu}^\nu \leq C$, where C is a constant). Then the penalty $\lambda_{\mathrm{SLIC}}(\nu)$ selected by sparsity ℓ_ν information criterion tends in probability to the universal penalty in that $\lambda_N(\nu)/\lambda_{\mathrm{SLIC}}(\nu) \to_p 1$ as $N \to \infty$.*

Since ν is also selected adaptively we expect improvement compared to fixing it to a pre-set value in $(0, 1]$. Section 2.3 confirms the intuition in direct sequence estimation and wavelet smoothing. Theoretical grounds supporting these empirical results need to be studied in more details.

6 Conclusion

Subbotin posterior mode estimation is a way to achieve model selection. The asymptotic minimaxity result proved in the canonical setting should be generalizable to more general settings and could be made more precise with respect to a class of loss functions and Besov spaces. In particular it is interesting to investigate in what the adaptive selection of the parameter ν through $\mathrm{SL}_\nu\mathrm{IC}$ or SURE improves on existing theoretical results. The generalization of Stein unbiased risk estimate beyond the canonical model also remains a challenging problem, to which Zou, Hastie, and Tibshirani (2007) contributed for $\nu = 1$. Finally the connection to the adaptive lasso could be pursued as a potential algorithm to solve, with multiple starts, the non-convex ℓ_ν penalized likelihood with $\nu < 1$.

7 Acknowledgment

I would like to thank an Associate Editor and two referees for valuable comments that helped improved the quality of the paper, Paul Tseng for help about optimization issues, and my colleagues Fred Dumas, Lionel Pournin, Olivier Renaud and Yvan Velenik for interesting discussions. This work was partially funded by FNS grant 100012-109532 and the Office du placement du Canton de Vaud.

References

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

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

Besbeas, P., I. De Feis, and T. Sapatinas (2004). A Comparative Simulation Study of Wavelet Shrinkage Estimators for Poisson Counts. *International Statistical Review 72*, 209–237.

Buehlmann, P. and T. Hothorn (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science 22*, 477–505.

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

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

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

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

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

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

Gao, H.-Y. and A. Bruce (1997). Waveshrink with firm shrinkage. *Statistica Sinica 7*, 855–874.

Griffin, J. E. and P. J. Brown (2007). Bayesian adaptive lassos with non-convex penalization. *http://www.kent.ac.uk/ims/personal/jeg28/BALasso.pdf*.

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

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

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

Kerkyacharian, G. and D. Picard (2004). Regression in random design and warped wavelets. *Bernoulli 10*, 1053–1105.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*(5), 1356–1378.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics 15*, 661–675.

Meegan, C. A., G. J. Fishman, R. B. Wilson, W. S. Paciesas, G. N. Pendleton, J. M. Horack, M. N. Brock, and C. Kouveliotou (1992). The Spatial Distribution of Gamma Ray Bursts Observed by BATSE. *Nature 355*, 143–145.

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

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

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

Park, M.-Y. and T. Hastie (2007). An l1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B 69*, 659–677.

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

Sardy, S., A. Antoniadis, and P. Tseng (2004). Automatic smoothing with wavelets for a wide class of distributions. *Journal of Computational and Graphical Statistics*, 399–421.

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

Sardy, S., D. B. Percival, A. G. Bruce, H.-Y. Gao, and W. Stuetzle (1999). Wavelet de-noising for unequally spaced data. *Statistics and Computing 9*, 65–75.

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

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.

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

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

Sylvain Sardy, Department of Mathematics, University of Geneva, 2-4 rue du Lièvre, Case postale 64, 1211 Genève 4, Switzerland.
E-mail: Sylvain.Sardy@unige.ch

A Derivation of the information criterion

Given priors $\pi_{\lambda|\nu}(\lambda \mid \nu;\tau)$ and $\pi_\nu(\nu)$, the posterior distribution is $f(\boldsymbol{\alpha},\lambda,\nu \mid \mathbf{Y}) = f(\mathbf{Y} \mid \boldsymbol{\alpha})\pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha} \mid \lambda,\nu)\pi_{\lambda|\nu}(\lambda \mid \nu;\tau)\pi_\nu(\nu)$. With $Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(c\alpha_n, 1)$, the negative log-posterior distribution is

$$\frac{1}{2}\|\mathbf{Y}-c\boldsymbol{\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_\nu^\nu-\frac{N}{\nu}\log\lambda+N\log\Gamma(1+\frac{1}{\nu})-\log\pi_{\lambda|\nu}(\lambda \mid \nu;\tau)-\log\pi_\nu(\nu). \quad (20)$$

Minimizing it over $(\boldsymbol{\alpha},\lambda,\nu)$ defines the posterior mode estimate $\hat{\boldsymbol{\alpha}}_{\hat\lambda,\hat\nu}$ for selected posterior modes $\hat\lambda$ and $\hat\nu$. To guarantee the universal property that $\mathrm{P}(\hat{\boldsymbol{\alpha}}_{\hat\lambda,\hat\nu}=\mathbf{0}) \overset{N\to\infty}{\longrightarrow} 1$ when the true sequence is $\boldsymbol{\alpha}=\mathbf{0}$, we calibrate τ such that the universal penalty $\lambda_N(\nu)$ is the root to the first order optimality condition of (20) with respect to λ:

$$\|\boldsymbol{\alpha}\|_\nu^\nu-\frac{N}{\nu\lambda}-\frac{\pi'_{\lambda|\nu}(\lambda \mid \nu;\tau)}{\pi_{\lambda|\nu}(\lambda \mid \nu;\tau)}=0 \quad \text{with} \quad \begin{cases} \|\boldsymbol{\alpha}\|_\nu^\nu=0 \\ \lambda=\lambda_N(\nu) \end{cases}. \quad (21)$$

Since $\pi_{\lambda|\nu}(\lambda \mid \nu;\tau) = F'_{\lambda|\nu}(\lambda;\tau)$ with $F_{\lambda|\nu}(\lambda;\tau) = G_0(\varphi_N(\varphi(\lambda/c^\nu;\nu)/\tau - d_N))$ and $G_0(\varphi) = \exp(-\exp(-\varphi))$, then (21) is equivalent to

$$-\frac{N}{\nu\lambda} - \frac{\varphi_N}{\tau}\{\exp(-\varphi_N(\frac{\varphi_N}{\tau} - d_N)) - 1\}\frac{\varphi'(\lambda/c^\nu;\nu)}{c^\nu} - \frac{\varphi''(\lambda/c^\nu;\nu)}{c^\nu\varphi'(\lambda/c^\nu;\nu)} = 0, \qquad (22)$$

with $\lambda = \lambda_N(\nu)$ and $\varphi_N = \sqrt{2\log N}$. With $\varphi''(\lambda;\nu)/\varphi'(\lambda;\nu) = (\nu-1)/(2-\nu)/\lambda$ and $\varphi_N d_N = o(N)$, the asymptotic solution to (22) is $\tau_N(\nu) = \nu\varphi_N^2/(N(2-\nu))$ since $\exp(-\varphi_N(\varphi_N/\tau_N(\nu) - d_N)) = o(1)$.

B Proof of Theorems

Proof of Theorem 2: (Existence) We consider a more general assumption where the penalty function $\varpi(\boldsymbol{\alpha})$ is coercive and continuous, for instance the Subbotin negative log-likelihood. The assumption on the univariate negative log-likelihood implies that $-l_N(\mu,\psi;\mathbf{Y}) \to +\infty$ whenever μ approaches a boundary point of Ω_μ^N or $\|\mu\| \to \infty$. Moreover (15) has a feasible solution since $\mathrm{Rg}([X_0\ X]) \bigcap \Omega_\mu^N \neq \emptyset$. Since the penalty $\varpi(\boldsymbol{\alpha})$ is coercive and continuous, then variant of Weierstrass' theorem ensures that a solution $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})$ to (15) exists.

Proof of Theorem 3: (Uniqueness, case $\nu = 1$) When $\mathrm{K}([X_0\ X]) = \{\mathbf{0}\}$, then $-l \circ [X_0\ X]$ is strictly convex. So is the objective function (15), since the sum of the strictly convex $-l \circ [X_0\ X]$ and the convex ℓ_1-norm is strictly convex. Moreover $\Gamma_{(\boldsymbol{\alpha}_0,\boldsymbol{\alpha})} = \{(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}) :\ X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha} \in \Omega_\mu^N\}$, the inverse image of Ω_μ^N, is a convex set in $\mathbb{R}^{\mathrm{P}_0+\mathrm{P}}$. Therefore $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_\lambda$ is the unique strict global minimizer of (15).
When $\mathrm{K}([X_0\ X]) \neq \{\mathbf{0}\}$, then another solution must write as $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_\lambda + (\mathbf{k}_0, \mathbf{k})$ for $(\mathbf{k}_0, \mathbf{k}) \in \mathrm{K}([X_0\ X])$. Indeed the objective function is convex, so any convex combination of two solutions $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_\lambda$ and $(\hat{\mathbf{z}}_0, \hat{\mathbf{z}})_\lambda$ is in $\Gamma_{\boldsymbol{\alpha}_0,\boldsymbol{\alpha}}$ and is another solution; moreover $-l$ is strictly convex on $\Gamma_{(\boldsymbol{\alpha}_0,\boldsymbol{\alpha})}$, so it is necessary that $X_0\hat{\boldsymbol{\alpha}}_{0\lambda} + X\hat{\boldsymbol{\alpha}}_\lambda = X_0\hat{\mathbf{z}}_{0\lambda} + X\hat{\mathbf{z}}_\lambda$, i.e., $(\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}})_\lambda = (\hat{\mathbf{z}}_0, \hat{\mathbf{z}})_\lambda + (\mathbf{k}_0, \mathbf{k})$ with $(\mathbf{k}_0, \mathbf{k}) \in \mathrm{K}([X_0\ X])$, otherwise the convex combination would attain a smaller objective, which contradicts the hypothesis. Hence, the objective does not change at the minimum if and only if $\|\hat{\boldsymbol{\alpha}}_\lambda\|_1 = \|\hat{\boldsymbol{\alpha}}_\lambda + \mathbf{k}\|_1$. So if $\mathbf{k}$ is moreover a vector generator of the boundary of the ℓ_1 ball containing $\hat{\boldsymbol{\alpha}}_\lambda$, then the overall objective function keeps the same minimal value. □

Proof of Theorem 4: Let $\mathbf{g}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}; \mathbf{Y})$ be the gradient of $-l_N(X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha}, \psi; \mathbf{Y})$ with respect to $\boldsymbol{\alpha}$. The ℓ_1 norm is not differentiable at $\mathbf{0}$, but let the modulus of the generalized gradient with respect to $\boldsymbol{\alpha}$ of the objective function (15) be $\mathbf{r}(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}; \mathbf{Y}) = (r_1(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}; \mathbf{Y}), \ldots, r_P(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}; \mathbf{Y}))$ with

$$r_p(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}; \mathbf{Y}) = \begin{cases} |g_p(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}; \mathbf{Y}) + \lambda\frac{\alpha_p}{|\alpha_p|}| & \text{if} \quad |\alpha_p| \neq 0; \\ \min_{0\leq|\eta|\leq\lambda} |g_p(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}; \mathbf{Y}) + \eta| & \text{if} \quad |\alpha_p| = 0. \end{cases}$$

If $(\boldsymbol{\alpha}_0, \mathbf{0}) \in \mathrm{int}(\Gamma_{(\boldsymbol{\alpha}_0,\boldsymbol{\alpha})})$, then a necessary and sufficient condition for $(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\alpha}}) = (\boldsymbol{\alpha}_0, \mathbf{0})$ to be a global minimizer is that $\mathbf{r}(\boldsymbol{\alpha}_0, \mathbf{0}; \mathbf{Y}) = \mathbf{0}$. Hence, if

$$\lambda \geq \max_{p=1,\ldots,P} |g_p(\boldsymbol{\alpha}_0, \mathbf{0}; \mathbf{Y})| =: \lambda_{\mathbf{Y}},$$

then $(\boldsymbol{\alpha}_0, \mathbf{0})$ is a global minimizer. Finally to prove the universal rule $\lambda_{\mathbf{Y}}$ is finite, note that the log-likelihood is convex and $(\boldsymbol{\alpha}_0, \mathbf{0})$ is in the interior of the domain. □

Proof of Theorem 5: We treat the case $\hat{\alpha}_{\mathrm{MLE}}(Y) > 0$ and $\dot{l}_\alpha(\alpha_0+\alpha, \psi; \cdot)$ monotone increasing w.r.t. Y (i.e., $\ddot{l}_{\alpha Y} > 0$). For fixed $\nu \in (0,1)$ and $\lambda > 0$, let

$$f(\alpha; Y) = -l(\alpha_0 + \alpha, \psi; Y) + \lambda|\alpha|^\nu. \tag{23}$$

be the function to minimize w.r.t. α in (16).
Shrinkage: Note that the second term in (23) is an even function increasing in α on $(0, \infty)$ and independent of Y. Since $\dot{f}_\alpha(0_+; Y) = -\dot{f}_\alpha(0_-; Y) = +\infty$, then the point $\xi_0 = 0$ is always a local solution. If $\hat{\alpha}_{\lambda,\nu} = \xi_0$ is the global solution then the estimator shrinks since $\hat{\alpha}_{\mathrm{MLE}}(Y) > 0$. Otherwise, since $f(\cdot; Y)$ is decreasing on the left side of zero and increasing on the right side of $\hat{\alpha}_{\mathrm{MLE}}(Y)$, the global minimum can only be $\hat{\alpha}_{\lambda,\nu} \in (0, \hat{\alpha}_{\mathrm{MLE}}(Y))$.
Thresholding: The estimate $\hat{\alpha}_{\mathrm{MLE}}(Y)$ is strictly increasing in Y. To see that, observe that $\dot{l}_\alpha(\alpha_0+\hat{\alpha}_{\mathrm{MLE}}(Y), \psi; Y) = 0$ by definition of the MLE. This implies in particular that $\nabla_Y \dot{l}_\alpha(\alpha_0+\hat{\alpha}_{\mathrm{MLE}}(Y), \psi; Y)$ is null, so $\ddot{l}_{\alpha\alpha}(\alpha_0+\hat{\alpha}_{\mathrm{MLE}}(Y), \psi; Y)\dot{\hat{\alpha}}_{\mathrm{MLE}}(Y)+\ddot{l}_{\alpha Y}(\alpha_0+\hat{\alpha}_{\mathrm{MLE}}(Y), \psi; Y) = 0$ for all $Y \in \Omega_Y$, where $-\ddot{l}_{\alpha\alpha}(\alpha_0+\hat{\alpha}_{\mathrm{MLE}}(Y), \psi; Y) > 0$ and $\ddot{l}_{\alpha Y} > 0$ by assumption. So $\dot{\hat{\alpha}}_{\mathrm{MLE}}(Y) > 0$ for all $Y \in \Omega_Y$.

By assumption, the penalized likelihood (23) has at most one local maximum $\xi_1(Y)$ and one local minimum $\xi_2(Y)$. The latter $\xi_2(Y) \in (\xi_1(Y), \hat{\alpha}_{\mathrm{MLE}}(Y))$ is the potential global minimum. With $\ddot{l}_{\alpha Y} > 0$, $-\dot{l}_\alpha(\alpha_0+\alpha, \psi; Y)$ is decreasing in Y for any given $\alpha > 0$, so let $Y_1 = \inf\{Y \in \Omega_Y : \min_{\alpha \in (0, \sup \Gamma_\alpha]} -\dot{l}_\alpha(\alpha_0+\alpha, \psi; Y)+\lambda\nu\alpha^{\nu-1} = 0\}$. Hence for $Y = Y_1$, the penalized likelihood (23) has a saddle point. If Y_1 does not exists in Ω_Y, then the estimator always thresholds since $f(\cdot; Y)$ is always increasing in α in that case. Note that Y_1 plays the role of p_0 in Antoniadis and Fan (2001).
Define on $[Y_1, \sup \Omega_Y] \times (0, \infty)$ the difference $\delta(Y, \lambda) = f(\xi_2(Y); Y) - f(0; Y)$ between the values of the objective function (23) at the two local minima $\xi_2(Y)$ and ξ_0 for a given λ. Clearly $\delta(Y_1, \lambda) > 0$ since $\dot{f}_\alpha(0_+; Y) = +\infty$. When $\delta(Y, \lambda) < 0$, then $\xi_2(Y)$ becomes the global minimum: the estimator shrinks without thresholding. Suppose we can show that $\delta(Y, \lambda)$ is strictly decreasing on $[Y_1, \sup \Omega_Y]$ for a given λ, then the estimator thresholds. Indeed, define

$$Y_\lambda = \sup\{Y \in [Y_1, \sup \Omega_Y] : \delta(Y, \lambda) \geq 0\}, \tag{24}$$

then

$$\begin{aligned} \hat{\alpha}_\lambda(Y) = 0 &\iff Y \leq Y_\lambda \\ &\iff \hat{\alpha}_{\mathrm{MLE}}(Y) \leq \hat{\alpha}_{\mathrm{MLE}}(Y_\lambda) =: \varphi^{\mathrm{u}}, \end{aligned}$$

since we showed that $\hat{\alpha}_{\mathrm{MLE}}(Y)$ is strictly increasing.

It remains to prove that $\delta(Y,\lambda)$ is decreasing on $[Y_1, \sup\Omega_Y]$ for a given λ. Since the local minimum $\xi_2(Y)$ satisfies by definition that $\dot{f}_\alpha(\xi_2(Y),Y)=0$, then

$$\begin{aligned}\dot{f}_Y(\xi_2(Y),Y) &= (\dot{f}_\alpha(\alpha,Y),\dot{f}_Y(\alpha,Y))\Big|_{(\alpha,Y)=(\xi_2(Y),Y)}(\xi_2'(Y),1)^{\mathrm{T}} \\ &= -\dot{l}_Y(\alpha_0+\alpha,\psi;Y)\Big|_{\alpha=\xi_2(Y)}.\end{aligned}$$

So $\dot{\delta}_Y(Y,\lambda) = -\dot{l}_Y(\alpha_0+\xi_2(Y),\psi;Y) - (-\dot{l}_Y(\alpha_0+0,\psi;Y))$, which is negative since $\xi_2(Y)>0$ and $-\ddot{l}_{\alpha Y}(\alpha_0+\alpha,\psi;Y)<0$. In the discrete case, the fact that $\delta(Y,\lambda)$ is decreasing on Ω_Y^I implies that it is decreasing on Ω_Y for a given λ. This completes the proof that Subbotin posterior mode thresholds.

Monotonicity of the threshold $\varphi^{\mathrm{u}}(\lambda;\nu)$ in λ for a given ν: $\delta(Y,\lambda)$ increases in λ for a given Y, and decreases in Y for a given λ. So Y_λ in (24) is non-decreasing in λ. Consequently, since the MLE is strictly increasing in Y, the upper threshold $\varphi^{\mathrm{u}} = \hat{\alpha}_{\mathrm{MLE}}(Y_\lambda)$ is non-decreasing in λ. □

C Threshold and jump formulae (8) and (9)

We consider the case of a strictly positive jump $\alpha>0$, and consequently of a strictly positive threshold $y>0$; the strictly negative case can be derived likewise for the left threshold and jump. System (17) for $\alpha_0=0$ writes:

$$\begin{aligned}\left\{\begin{array}{rcl} f(0;y) &=& f(\alpha,y) \\ \nabla_\alpha f(\alpha;y) &=& 0\end{array}\right. &\Leftrightarrow \left\{\begin{array}{rcl} \frac{1}{2}y^2 &=& \frac{1}{2}(\alpha-y)^2+\lambda\alpha^\nu \\ (\alpha-y)+\lambda\nu\alpha^{\nu-1} &=& 0\end{array}\right. \\ &\Leftrightarrow \left\{\begin{array}{rcl} 0 &=& \frac{\nu}{2}(\alpha-2y)+\lambda\nu\alpha^{\nu-1} \\ 0 &=& (\alpha-y)+\lambda\nu\alpha^{\nu-1}\end{array}\right. \\ &\Leftrightarrow \left\{\begin{array}{rcl} \alpha &=& y\frac{2(1-\nu)}{2-\nu} \\ y &=& y\frac{2(1-\nu)}{2-\nu}+\lambda\nu(y\frac{2(1-\nu)}{2-\nu})^{\nu-1}\end{array}\right. \\ &\Leftrightarrow \left\{\begin{array}{rcl} \alpha &=& y\frac{2(1-\nu)}{2-\nu} \\ y^{2-\nu} &=& \lambda(2(1-\nu))^{\nu-1}(2-\nu)^{2-\nu}\end{array}\right..\end{aligned}$$

□

D Asymptotic pivot

With the assumption that $\alpha_n=0$ for $n=1,\ldots,N$, the distribution of $X_n=|Y_n|$ with $Y_n\sim\mathrm{N}(0,1)$ is $F(x)=1-2\Phi(-x)$ on $[0,\infty)$, where Φ is the standard Gaussian cumulative distribution function. The distribution of $\varphi_{\mathbf{Y}} = \max_{n=1,\ldots,N}(|Y_n|)$ is degenerate, but results from extreme value theory guarantee a non-degenerate limit

law under a proper affine transformation $c_N^{-1}(\varphi_{\mathbf{Y}} - d_N)$ that we now derive following Embrechts, Kluppelberg, and Mikosch (1997, Example 3.3.29).

The distribution F has right endpoint $x_F = \infty$, is twice differentiable on its domain and $F''(x) = -2x\varphi(x) < 0$, where φ is the standard Gaussian density. Moreover

$$\begin{aligned} \bar{F}(x)F''(x)/f^2(x) &= 2\Phi(-x)(-2\varphi'(-x))/(4\varphi^2(-x)) \\ &\sim (-1/x)\varphi'(-x)/\varphi(-x) = -1 \end{aligned}$$

using $\Phi(-x) \sim \varphi(x)/x$ as $x \to \infty$. So F is a von Mises function with auxiliary function $a(x) = \bar{F}(x)/f(x) = 2\Phi(-x)/(2\varphi(x)) \sim 1/x$. Consequently F belongs to the maximum domain of attraction of the Gumbel distribution. A possible choice of norming constants is $d_N = F^{-1}(1-1/N)$ and $c_N = a(d_N)$. Since $\bar{F}(x) \sim 2\varphi(x)/x =: \bar{G}(x)$ as $x \to \infty$, then we look for a solution of $-\log \bar{G}(d_N) = \log N$, i.e.,

$$\frac{1}{2}d_N^2 + \log d_N + \frac{1}{2}\log(2\pi) - \log 2 = \log N.$$

Then a Taylor expansion around $\varphi_N = \sqrt{2\log N}$ yields

$$d_N = \varphi_N - (\log\log N + \log 4\pi - 2\log 2)/(2\varphi_N)$$

and $c_N = a(d_N) \sim 1/\varphi_N$. □

E Asymptotic minimaxity

Consider ν fixed in $(0, 1]$. Property 3 states the uniqueness of λ; its first order optimality condition (21) obtained by differentiating with respect to λ the information criterion for the calibrated $\tau = \tau_N(\nu)$ leads to the equation

$$\lambda = \frac{\varphi_N}{\nu\|\hat{\boldsymbol{\alpha}}_{\lambda,\nu}\|_{N,\nu}^{\nu}\varphi_N/N + (2-\nu)\varphi'(\lambda;\nu)},$$

to which $\lambda_{\mathrm{SLIC}}(\nu)$ is the root. Note first that $\varphi'(\lambda_N(\nu);\nu) = \varphi_N^{\nu-1}\left((2-2\nu)/(2-\nu)\right)^{\nu-1}$ and that $\varphi_N/((2-\nu)\varphi'(\lambda_N(\nu);\nu)) = \lambda_N(\nu)$ is the universal penalty from Property 1 with $c = 1$. Since $\lambda_N(\nu)/\lambda_{N,\mathrm{SLIC}}(\nu) = 1+\nu(2(1-\nu))^{1-\nu}(2-\nu)^{\nu-2}\varphi_N^{2-\nu}\|\hat{\boldsymbol{\alpha}}_{\lambda_N(\nu),\nu}\|_{N,\nu}^{\nu}/N$, then $\lambda_N(\nu)$ is the asymptotic root chosen by SLIC provided $\varphi_N^{2-\nu}\|\hat{\boldsymbol{\alpha}}_{\lambda_N(\nu),\nu}\|_{N,\nu}^{\nu}/N = o_p(1)$.

With $Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\alpha_n, 1)$, the ℓ_ν-Subbotin coefficients estimate are $\hat{\alpha}_{n,\lambda,\nu} = \eta_{\lambda,\nu}(Y_n)$ for $n = 1, \ldots, N$, where $\eta_{\lambda,\nu}$ is the thresholding function. By definition of the universal penalty $\lambda = \lambda_N(\nu)$, the threshold is $\varphi_N(\nu) = \sqrt{2\log N}$ for all ν. Moreover outside $\pm\varphi_N$ the datum Y_n is shrunk toward zero for all $\nu > 0$ and unshrunk if $\nu = 0$

(the hard thresholding function (2)). Consequently $|\hat{\alpha}_{n,\lambda_N(\nu),\nu}|^\nu \leq |\hat{\alpha}_{n,\lambda_N(0),0}|^\nu = |Y_n \cdot 1_{\{|Y_n|\geq\varphi_N\}}(Y_n)|^\nu$, and

$$\|\hat{\boldsymbol{\alpha}}_{\lambda_N(\nu),\nu}\|_{N,\nu}^\nu \leq \sum_{n=1}^N X_n^\nu, \tag{25}$$

where $X_n = |Y_n| \cdot 1_{\{|Y_n|\geq\varphi_N\}}(Y_n)$.

Let $X = |Y| \cdot 1_{\{|Y|\geq\varphi\}}(Y)$, where $Y \sim \mathrm{N}(\alpha, 1)$. Its density is $f_X(x) = (\phi_{-\alpha}(x) + \phi_\alpha(x))1_{\{x\geq\varphi\}}(x) + \delta_0(x)A_\varphi$, where $A_\varphi = \int_{-\varphi}^{\varphi} \phi_\alpha(u)du$ and ϕ_α is the Gaussian density with mean α and variance one. Its moment generating function $M_X(t) = \exp(t^2/2 - t\alpha)(1 - \Phi(-t + \alpha + \varphi)) + \exp(t^2/2 + t\alpha)(1 - \Phi(\varphi - \alpha - t)) + A_\varphi$. Consequently $\mathrm{E}(X) = \alpha(\Phi(\alpha + \varphi) - \Phi(\varphi - \alpha)) + \phi(\varphi + \alpha) + \phi(\varphi - \alpha)$, which is an even function of α, so that we consider the case $\alpha \geq 0$ to majorate the expected value using Mill's ratio:

$$\mathrm{E}(X) < \frac{\varphi}{\alpha + \varphi}\phi(\varphi + \alpha) + \frac{\varphi}{\varphi - \alpha}\phi(\varphi - \alpha) + \frac{\alpha}{(\alpha + \varphi)^3}\phi(\varphi + \alpha) \tag{26}$$

for all $\alpha \geq 0$ and φ. Hence the distribution of each X_n is parametrized by α_n and $\varphi = \varphi_N = \sqrt{2\log N}$. Since $\lim_{N\to\infty} \varphi_N = \infty$ and α_n is bounded by C by assumption, then for N large enough, α_n is negligible compared to φ_N, so the expected value $\mathrm{E}X_n$ is bounded from (26) as $N \to \infty$ by $2\phi(\varphi_N) = 2/(N\sqrt{2\pi})$ for all $n = 1, \ldots, N$.

For any $\epsilon > 0$, using (25) and Markov's inequality,

$$\mathrm{P}(\|\hat{\boldsymbol{\alpha}}_{\lambda_N(\nu),\nu}\|_{N,\nu}^\nu \frac{\varphi_N^{2-\nu}}{N} > \epsilon) \leq \frac{\varphi_N^{2-\nu}}{N\epsilon} \sum_{n=1}^N \mathrm{E}X_n^\nu.$$

Since $\nu \leq 1$, Jensen's inequality on a concave function guarantees that

$$\mathrm{P}(\|\hat{\boldsymbol{\alpha}}_{\lambda_N(\nu),\nu}\|_{N,\nu}^\nu \frac{\varphi_N^{2-\nu}}{N} > \epsilon) \leq \frac{\varphi_N^{2-\nu}}{N\epsilon} \sum_{n=1}^N (\mathrm{E}X_n)^\nu = (\frac{2}{\sqrt{2\pi}})^\nu \frac{(2\log N)^{1-\nu/2}}{\epsilon N^\nu}$$

which tends to zero as $N \to \infty$. □