Density estimation by total variation penalized likelihood driven by the sparsity ℓ_1 information criterion

SYLVAIN SARDY

Department of Mathematics, University of Geneva

and

PAUL TSENG

University of Washington

We propose a nonlinear density estimator, which is locally adaptive, like wavelet estimators, and positive everywhere, without a log- or root-transform. This estimator is based on maximizing a nonparametric log-likelihood function regularized by a total variation penalty. The smoothness is driven by a single penalty parameter, and to avoid cross validation, we derive an information criterion based on the idea of universal penalty.

The penalized log-likelihood maximization is reformulated as an ℓ_1-penalized strictly convex program whose unique solution is the density estimate. A Newton-type method cannot be applied to calculate the estimate because the ℓ_1-penalty is non-differentiable. Instead, we use a dual block-coordinate relaxation method that exploits the problem structure.

By comparing with kernel, spline and *taut string* estimators on a Monte Carlo simulation, and by investigating the sensitivity to ties on two real data sets, we observe that the new estimator achieves good L_1 and L_2 risk for densities with sharp features, and behaves well with ties.

Key Words: convex program, dual block coordinate relaxation, extreme value theory, ℓ_1-penalization, smoothing, total variation, universal penalty parameter.

1 Introduction

An old problem in statistics (Silverman 1986; Scott 1992; Simonoff 1996) is the estimation of a univariate density function f sampled at N points in some real domain Ω. For simplicity, we assume for now that the points are distinct and we let $\mathbf{x} = (x_1, \ldots, x_N)$ be the distinct order statistics. Section 6 addresses the ties issue. Maximizing the nonparametric log-likelihood function

$$l(f; \mathbf{x}) = \sum_{i=1}^{N} \log f(x_i), \tag{1}$$

over all nonnegative-valued integrable measures f that integrate to unity leads to the degenerate estimate $\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(x)$, where $\delta_{x_i}(\cdot)$ is the Dirac measure at x_i. It is degenerate in that its total variation (TV) is unbounded: $\hat{f}$ does not belong to the space of measures of bounded variation.

Many nonparametric estimators either add to (1) a penalty on nonsmoothness, as in Tikhonov regularization (Tikhonov 1963), or impose some smoothness constraint to obtain a smoother and more accurate estimate. For instance, the histogram is the unique maximum likelihood estimate constrained to be piecewise-constant on bins. In a pioneering paper, Good and Gaskins (1971) proposed adding to (1) a functional $\Phi(f)$ that penalizes nonsmoothness, multiplied by a penalty parameter $\lambda > 0$. They defined the functional estimate $\hat{f}_\lambda$ as the solution to

$$\max_{f \in \mathcal{F}} \; l(f; \mathbf{x}) - \lambda \Phi(f) \quad \text{s.t.} \quad \int f(x) dx = 1, \tag{2}$$

("s.t." is short for "subject to"), where $\mathcal{F}$ is a prescribed class of nonnegative-valued functions on Ω. The estimate $\hat{f}_\lambda$ tends to a high variance estimate when $\lambda \to 0^+$, and to the density in $\mathcal{F}$ of minimum penalty Φ when $\lambda \to \infty$. Various

penalty functionals, including $\Phi(f) = 4\int\{\nabla\sqrt{f}\}^2$ (Good and Gaskins 1971) and $\Phi(f) = \int\{\nabla^3 \log f\}^2$ (Silverman 1982), have been proposed, but each has drawbacks. In particular, these penalty functionals intrinsically assume f is smooth on Ω (i.e., $\mathcal{F} = \mathcal{C}^1(\Omega)$) or the points of nondifferentiability have no practical significance and can be ignored. Accordingly, they penalize high values of $|f'|$ more heavily and can result in oversmoothing when the underlying density is not differentiable everywhere (e.g., its graph has jumps or corners or cusps). Also, with root- or log-transforms, nonsmoothness is penalized more heavily at low density values than at high density values, which may lead to uneven smoothing. It is often argued that basing the penalty on the derivatives of $\sqrt{f}$ or $\log f$ provides a barrier from nonpositivity of f. However, the terms $\log f(x_i)$ in (1) already provide barriers against nonpositivity of f at the sample points $x_1, \ldots, x_N$, and f would be positive everywhere on $[x_1, x_N]$ if the functions in $\mathcal{F}$ are restricted to be monotone on each subinterval $[x_i, x_{i+1}]$, $i = 1, \ldots, N-1$.

Penalized log-likelihood served as a framework for smoothing splines (Wahba 1990). O'Sullivan (1988) developed an approximation to the log-density estimator of Silverman (1982). Kooperberg and Stone (1991) developed *logspline* by selection of knots among potential knots near order statistics, their number being automatically selected by an AIC-like criterion. Silverman (1982) and Stone (1990) derived asymptotic optimal convergence properties for their penalized log-likelihood estimators under smoothness assumptions. Eggermont and LaRiccia (1999) derived optimal convergence rate for Good (1971)'s estimator.

For the estimation of nonsmooth functions, nonlinear wavelet-based estimators pioneered by Waveshrink in regression (Donoho and Johnstone 1994) have been developed for density estimation as well; see Vidakovic (1999) for a review. To guarantee positivity of the wavelet estimate, Penev and Dechevsky (1997) and Pinheiro and Vi-

dakovic (1997) estimated $\sqrt{f}$ at the cost of losing local adaptivity, showing again that the use of a transform is not innocuous. Wavelet estimators are also sensitive to the choice of the dyadic grid (Renaud 2002) as the histogram is sensitive to the choice of bins. Recently, Willett and Nowak (2003) proposed to adaptively prune a multiscale partition, Davies and Kovac (2004) proposed *taut string*, a simple and yet efficient locally adaptive estimator which measures complexity by the number of modes, and Koenker and Mizera (2006) proposed a log-density estimator regularized by a TV penalty on the first derivative.

The density estimator which we propose in Section 2 is based on the TV penalty, a functional that does not even assume first-order differentiability. The resulting convex ℓ_1-norm has been much used for finding sparse overcomplete representations of noisy signals, as it imparts remarkable sparsity or local adaptivity properties to estimators (e.g., *Lasso*, *soft-waveshrink*, ℓ_1-*Markov random field*). However, it has been little used for density estimation, possibly due to difficulties in selecting the penalty parameter and applying a Newton-type method when the penalized log-likelihood function is not differentiable. The goal of this paper is to show that a nonsmooth penalty can be beneficial for density estimation, by overcoming the aforementioned difficulties. We derive in Section 3 a practical rule to select the penalty parameter. We describe in Section 4 an iterative method, based on dual block coordinate relaxation, that exploits the special structure of ℓ_1-penalized log-likelihood and efficiently computes a solution. Section 5 studies the finite sample properties of the new estimator in comparison with existing estimators on a Monte Carlo simulation. We analyze sensitivity to ties on real and simulated data sets in Section 6, and draw some conclusions in Section 7.

2 Total-variation-penalized log-likelihood

To estimate possibly nonsmooth densities f in the space of functions of bounded variation, we use in (2) the TV penalty functional

$$\Phi_{\text{TV}}(f) = \sup \sum_j |f(u_{j+1}) - f(u_j)|, \tag{3}$$

where the supremum is taken over all possible partitions $[u_j, u_{j+1}]$, $j = 1, \ldots, M$, of Ω, the domain of the univariate density. If we assume that f is absolutely continuous, then TV would reduce to the more conventional smoothness measure $\Phi_{\text{TV}}(f) = \int_\Omega |f'(x)|dx$. In general, the total variation (3) cannot be evaluated because enumerating all possible partitions (which includes the histogram bins) is computationally intractable, even if Ω is discretized by a fine grid. Moreover, minimizing the penalized-TV functional yields piecewise-constant function with jumps at the knot points (see Appendix A). The 2-sided discontinuities of the piecewise constant function at the knots is undesirable for the purpose of smoothing. We could enforce more smoothness by restricting f to be left-continuous or right-continuous, but each introduces asymmetry arbitrarily.

In contrast, the classes $\mathcal{F}_u^0$ and $\mathcal{F}_x^1$ of piecewise-constant and piecewise-linear functions f with breakpoints at the u- and x-knots proposed below exhibit more symmetry as well as smoothness. This regularization of the smoothness class is analogous to O'Sullivan (1988)'s proposal to approximate the functional estimate of Silverman (1982). Let $u_0 = x_1$, $u_N = x_N$, and $u_1, \ldots, u_{N-1}$ be the midpoints between the order statistics of the sample points, i.e., $u_i = (x_i + x_{i+1})/2$. Consider the space of zeroth-order splines on the $\mathbf{u}$-partition,

$$\mathcal{F}_u^0 = \{f : [x_1, x_N] \to \Re \mid f \text{ is constant on } (u_{i-1}, u_i), i = 1, \ldots, N\}.$$

or the space of first-order splines on the **x**-partition,

$$\mathcal{F}_x^1 = \left\{ f \in \mathcal{C}^0[x_1, x_N] \mid f \text{ is linear on } [x_{i-1}, x_i], i = 2, \ldots, N \right\}.$$

For a given $\lambda > 0$, we define the density estimates TV^0 (respectively, TV^1) to be the solution to (2) using TV penalty (3) restricted to the function class $\mathcal{F}_u^0$ (respectively, $\mathcal{F}_x^1$). Thus TV^0 is piecewise-constant on the **u**-partition and TV^1 is continuous and piecewise-linear on the **x**-partition. Notice that any density $f \in \mathcal{F}_u^0$ that is not monotone on $[u_{i-1}, u_i]$, $i = 1, \ldots, N$, can be made so by suitably adjusting the value of f at $u_0, u_1, \ldots, u_N$ without decreasing $l(f; \mathbf{x}) - \lambda \Phi_{\mathrm{TV}}(f)$. Thus there exists a TV^0 that is piecewise-monotone on the **u**-partition. Also, any density in $\mathcal{F}_x^1$ is monotone on $[x_{i-1}, x_i]$, $i = 2, \ldots, N$, so any TV^1 is piecewise-monotone on the **x**-partition. Since the TV of a monotone univariate function f on a closed interval $[\alpha, \beta]$ is $|f(\beta) - f(\alpha)|$, it follows from direct integration that such TV^0 and TV^1 are solutions to the following finite-dimensional optimization problem, with $f_i = f(x_i)$ at the order statistics,

$$\min_{\mathbf{f} \in \Re^N} -\sum_{i=1}^{N} \log f_i + \lambda \|B\mathbf{f}\|_1, \quad \text{s.t.} \quad \mathbf{a}^T \mathbf{f} = 1, \tag{4}$$

where B is the $(N-1) \times N$ matrix such that $\|B\mathbf{f}\|_1 = \sum_{i=1}^{N-1} |f_{i+1} - f_i|$ (i.e., $B_{i,i} = -B_{i,i+1} = 1$ for $i = 1, \ldots, N-1$ and zero otherwise), $a_1 = (x_2 - x_1)/2$, $a_N = (x_N - x_{N-1})/2$ and $a_i = (x_{i+1} - x_{i-1})/2$ for $i = 2, \ldots, N-1$. Notice that the integration coefficients **a** depend only on the sample **x**, not on the density **f**. This is a convex program with strictly convex (though nondifferentiable) objective function and linear constraint, so it has a unique solution.

We now make some important observations about the approximation (4). First, the estimates use adaptive partition, reminiscent of logspline which places knots near order statistics, in the sense that the modeling of the underlying density depends on

the sample. We will see that, for TV-penalization, the "knot selection" has the advantage of being driven by a single penalty parameter. Second, the piecewise-monotone interpolation (either constant or linear) avoids introducing spurious modes between observations where no information is available. Third, the TV of the density is evaluated on the range of the observation $[x_1, x_N]$, the domain of the density being unknown. Evaluating the smoothness measure on $[x_1, x_N]$ is also required for the derivation of the universal penalty parameter (see Section 3.1). Fourth, the log-likelihood ensures positivity of f_i, $i = 1, \ldots, N$, which, together with the piecewise-constant or piecewise-linear interpolation, results in an estimate that is positive everywhere on $[x_1, x_N]$. Fifth, the proposed TV estimator is linked to *taut string* that has been shown to be solution to

$$\min_{\mathbf{f}} \sum_{i=1}^{N-1} (x_{i+1} - x_i)\{(x_{i+1} - x_i)f_i - \frac{1}{N-1}\}^2 + \lambda \sum_{i=1}^{N-1} |f_{i+1} - f_i|.$$

Finally, using results from optimization theory (Bertsekas 1999; Rockafellar 1970), TV^0 and TV^1 have the following properties which will be useful for selecting the penalty parameter λ; see Section 3.

Property 1. Suppose N is a multiple of a blocksize $K \in \{1, \ldots, N\}$ (i.e., $N = n_N K$) and let B_K be the matrix operating finite differences skipping every other K:

$$\|B_K \mathbf{f}\|_1 = \sum_{i=1}^{n_N} \sum_{j=1}^{K-1} |f_{K(i-1)+j+1} - f_{K(i-1)+j}|, \tag{5}$$

with the convention that $\sum_{j=1}^{0} = \sum_{j=1}^{1}$ so that $B_1 = B$; note that $B_K \mathbf{f} = \mathbf{0}$ if and only if $\mathbf{f}$ is piecewise constant density taking identical values at K successive points:

$$f_{K(i-1)+1} = \ldots = f_{K(i-1)+K}, \; i = 1, \ldots, n_N. \tag{6}$$

With $B = B_K$ in (4), there exists a *finite* threshold $\lambda_{\mathbf{x}}$ such that the solution is a piecewise constant according to (6) if and only if $\lambda \geq \lambda_{\mathbf{x}}$. Specifically, we rewrite (4) as

$$\min_{\mathbf{f}\in\Re^N, \mathbf{u}\in\Re^{N-1}} -\sum_{i=1}^{N} \log f_i + \lambda\|\mathbf{u}\|_1, \quad \text{s.t.} \quad B_K\mathbf{f} - \mathbf{u} = 0,\ \mathbf{a}^T\mathbf{f} = 1, \tag{7}$$

whose corresponding Lagrangian is

$$L(\mathbf{f}, \mathbf{u}, \mathbf{w}, z) = -\sum_{i=1}^{N} \log f_i + \lambda\|\mathbf{u}\|_1 + \mathbf{w}^T(B_K\mathbf{f} - \mathbf{u}) + z(\mathbf{a}^T\mathbf{f} - 1). \tag{8}$$

Due to the linear constraints, the Karush–Kuhn–Tucker condition $\mathbf{0} \in \partial L(\mathbf{f}, \mathbf{u}, \mathbf{w}, z)$ is both necessary and sufficient for optimality (Rockafellar 1970, Theorem 28.2). Here we use the subdifferential ∂ since $\|\cdot\|_1$ is not differentiable. Since the subdifferential of $\|\cdot\|_1$ at $\mathbf{0}$ is the unit-ball with respect to the ℓ_∞-norm, it is readily seen that $\mathbf{0} \in \partial L(\mathbf{f}, \mathbf{u}, \mathbf{w}, z)$ at $\mathbf{u} = B_K\mathbf{f} := \mathbf{0}$ if and only if

$$\begin{aligned} -1/f_i + (B_K^T\mathbf{w})_i + za_i &= 0, \quad i = 1, \ldots, N, && (9)\\ \|\mathbf{w}\|_\infty &\leq \lambda && (10)\\ \mathbf{a}^T\mathbf{f} - 1 &= 0 && (11) \end{aligned}$$

The solution $\mathbf{w}$ to (9)-(11) has entries

$$w_{K(i-1)+k} = -N\sum_{j=1}^{k} a_{K(i-1)+j} + \frac{Nk}{K}\sum_{j=1}^{K} a_{K(i-1)+j}, \quad i = 1, \ldots, n_N,\ k = 1, \ldots, K \tag{12}$$

for λ large enough. The smallest possible λ is therefore

$$\lambda_{\mathbf{x}} = \|\mathbf{w}\|_\infty \quad \text{with} \quad \mathbf{w} \text{ in (12)}, \tag{13}$$

which is finite.

At the other end, as $\lambda \to 0^+$, a continuity argument shows that the solution to (4) converges to the empirical estimate $\hat{f}_{\lambda,i} = \frac{1}{a_i N}$, $i = 1, \ldots, N$, which is the unique solution to (4) with $\lambda = 0$.

Property 2. By strict convexity of the objective function and by linearity of the constraint, any local minimizer of (4) is its unique global minimizer. This nice property is similar to the convexity property of the cost function used in O'Sullivan (1988) and contrasts with the multimodality of the criterion encountered with *logspline* (Kooperberg and Stone 1991; Kooperberg and Stone 2002).

Property 3. For each $1 \leq i \leq N-1$ with $\frac{1}{a_i} \geq \frac{1}{a_{i+1}}$ (respectively, $\frac{1}{a_i} \leq \frac{1}{a_{i+1}}$), we have $\hat{f}_{\lambda,i} \geq \hat{f}_{\lambda,i+1}$ (respectively, $\hat{f}_{\lambda,i} \leq \hat{f}_{\lambda,i+1}$) for all $\lambda \geq 0$.
For the proof, see Appendix D. This shows that the estimate $\hat{f}_{\lambda,i}$ is ordered (relative to the neighboring values) consistently with the ratio $\frac{1}{a_i}$ for all λ.

Property 4. Under affine transformation of the data $Y = \alpha + \beta X$ with $\beta > 0$, the estimate (4) satisfies $\hat{f}^X_\lambda(x) = \beta \hat{f}^Y_{\lambda\beta}(\alpha + \beta x)$.

3 Selection of the penalty parameter

The penalty parameter $\lambda \geq 0$ indexes a continuous class of models. Its selection is crucial to find the model that best fits the data. Model selection is an old problem, for which key contributions are the AIC, C_p and BIC criteria (Akaike 1973; Mallows 1973; Schwarz 1978) in the context of variable selection, that is, for the discrete problem equivalent to ℓ_0-penalized likelihood. Candidates for selecting the hyperparameter in ℓ_1-penalized log-likelihood models are resampling techniques such as cross validation (Stone 1974) and the bootstrap, an approximate generalized cross validation (Fu 1998) derived for the Lasso (Tibshirani 1995), BIC borrowed from

variable selection (Koenker, Ng, and Portnoy 1994), the empirical Bayes approach (Good 1965) used for ℓ_1 Markov random field smoothing (Sardy and Tseng 2004), and $\mathrm{SL}_1\mathrm{IC}$ (Sardy 2009) based on the universal penalty parameter (Donoho and Johnstone 1994). See Berlinet and Devroye (1994) for a review of bandwidth selection methods for kernel-based density estimates. However, cross validation is computationally intensive while generalized cross validation, originally devised for linear estimators (Craven and Wahba 1979), requires for our problem an ill-defined linearization of the ℓ_1-penalized log-likelihood near the solution. The Bayesian information criterion was intended for variable selection (ℓ_0-penalty), not for ℓ_1-penalized log-likelihood. Empirical Bayes requires maximizing the marginal likelihood of the data with respect to the prior, which is rarely available in a closed form and therefore requires numerical integration tools. So we turn to an extension of the universal penalty parameter in Section 3.1 and an extension of $\mathrm{SL}_1\mathrm{IC}$ to density estimation in Section 3.2.

3.1 Universal penalty parameter

The universal penalty parameter was originally developed in regression for Gaussian wavelets smoothing (Donoho and Johnstone 1994). We review here an important property of the universal penalty parameter, based on which we derive a universal penalty parameter for (4). Suppose $\mathbf{y} = (y_1, \ldots, y_N)$ are Gaussian measurements of $\mathbf{f} = (f(x_1), \ldots, f(x_N))$ at $N = 2^{J+1}$ equispaced locations $x_1, \ldots, x_N$ on $[0, 1]$, and assume the function $f(\cdot)$ can be well represented by a linear combination of approximation $\phi(\cdot)$ and fine scale $\psi(\cdot)$ wavelets. The standard wavelets are multiresolution functions that are locally supported and indexed by a location parameter k and a scale parameter j. A father wavelet $\phi(\cdot)$ such that $\int_0^1 \phi(x)dx = 1$ generates

$p_0 = 2^{j_0}$ approximation wavelets by means of the dilation and translation relation

$$\phi_{j_0,k}(x) = 2^{j_0/2}\phi(2^{j_0}x - k), \quad k = 0, 1, \ldots, 2^{j_0} - 1.$$

Similarly, a mother wavelet $\psi(\cdot)$ such that $\int_0^1 \psi(x)dx = 0$ generates $N - p_0$ fine scale wavelets

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j = j_0, \ldots, J;\ k = 0, 1, \ldots, 2^j - 1.$$

Hence we assume $\mathbf{f} = \Phi\boldsymbol{\alpha} = \Phi_0\boldsymbol{\beta} + \Psi\boldsymbol{\gamma}$, where Φ_0 is the $N \times p_0$ matrix of approximation wavelets, Ψ is the $N \times (N - p_0)$ matrix of fine scale wavelets, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the corresponding coefficients.

Donoho, Johnstone, Hoch, and Stern (1992) use an ℓ_1-penalty on the wavelet coefficients $\beta_{j,k}$ to regularize the log-likelihood and define the estimate as the solution to

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \frac{1}{2}\|\Phi_0\boldsymbol{\beta} + \Psi\boldsymbol{\gamma} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1 \quad = \quad \min_{\boldsymbol{\beta},\mathbf{f}=\Psi\boldsymbol{\gamma}} \frac{1}{2}\|\Phi_0\boldsymbol{\beta} + \mathbf{f} - \mathbf{y}\|_2^2 + \lambda \sum_{j=j_0}^{J} \|\Psi_j^{\mathrm{T}}\mathbf{f}\|, \tag{14}$$

One can show a property similar to Property 1, namely, all coefficient are exactly set to zero (i.e., $\|\hat{\boldsymbol{\gamma}}_\lambda\|_1 = 0$) if and only if $\lambda \geq \lambda_{\mathbf{y}}$, where $\lambda_{\mathbf{y}} = \|\Psi'\mathbf{y}\|_\infty$. Based on this property, the universal penalty parameter λ_N^Ψ is defined as the smallest function of the sample size N such that, when the true underlying function is constant then the above ℓ_1-penalized log-likelihood estimate is also constant with a probability tending to one as $N \to \infty$. A solution $\lambda_N^\Psi = \sigma\sqrt{2\log N}$ is found by controlling the extremal behavior of $\lambda_{\mathbf{Y}}$ when $\mathbf{Y}$ is a vector of independent Gaussian variables with variance σ^2. Importantly the simple universal penalty λ_N^Ψ provides nearly minimax results for a class of loss functions and smoothness classes (Donoho, Johnstone, Kerkyacharian,

and Picard 1995).

Deriving a universal penalty for our TV density estimator is not straightforward. Certainly the successive finite differences $|f_2 - f_1|$, $|f_3 - f_2|$, $|f_4 - f_3|, \ldots$ used by the TV penalty in (4) are reminiscent of the discrete Haar mother wavelets coefficients at level J defined by $|\Psi_{J1}^{\mathrm{T}}\mathbf{f}| = \frac{1}{\sqrt{2}}|f_2 - f_1|, |\Psi_{J2}^{\mathrm{T}}\mathbf{f}| = \frac{1}{\sqrt{2}}|f_4 - f_3|, \ldots$. Yet a major discrepancy is the non disjoint nature of the TV finite differences (for instance f_2 appears in $|f_2 - f_1|$ and $|f_3 - f_2|$) so that f_1 is linked not only to f_2 but also to the last value f_N via all intermediate values. Haar wavelets use only every other finite difference and operate instead finite differences at various scales $j_0, \ldots, J$ (for instance $|\Psi_{(J-1)1}^{\mathrm{T}}\mathbf{f}| = \frac{1}{2}|f_3 + f_4 - f_2 - f_1|$). We propose the following adaptation of the universal penalty to TV regularization. Observe that the property achieved by the universal rule λ_N^{Ψ} with a probability tending to one that all Haar wavelet coefficients γ_{jk} at levels J and $J-1$ are null is true if and only if $f_{4k+1} = f_{4k+2} = f_{4k+3} = f_{4k+4}$ for $k = 0, \ldots, 2^{J-1} - 1$, that is the function is piecewise constant on disjoint intervals of length $2^{J-1}/N$. Hence to extend the universal rule to TV-density estimation, one may select λ_N^{TV} such that the estimate (4) is piecewise constant on blocks of size K according to (6) with a probability tending to one as N goes to infinity when the underlying density is Uniform. This amounts to controlling with λ_N^{TV} the extremal behavior of $\lambda_{\mathbf{x}}$ of Property 1 which sets the estimated density to a piecewise constant density when $\mathbf{x}$ is a $\mathrm{U}(0,1)$ sample of size N. Appendix B provides the derivation of the bound $\lambda_N^{\mathrm{TV}} = \sigma_{N,K}\sqrt{2\log(N/K)}$ with $\sigma_{N,K} = \sqrt{(1 - K/N)K}$ for the choice $K \sim \sqrt{\log N}$ for large N.

3.2 A sparsity ℓ_1 information criterion

The universal penalty is asymptotically minimax but is known to oversmooth in regression when N is small. So we propose to derive an information criterion to select

a penalty based on the sample, which is always bounded above by λ_N^{TV}. Borrowing from Markov random field (Besag 1986; Geman and Geman 1984), the vector $\mathbf{f}$ at the N samples may be interpreted as the realization of a first-order pairwise Laplace Markov random field with improper joint density

$$\pi_{\mathbf{f}}(\mathbf{f} \mid \lambda) = (\lambda/2)^{N-1} \exp\{-\lambda \sum_{i=1}^{N-1} |f_{i+1} - f_i|\},$$

shifted and rescaled to make it a density. The parameter λ is also seen as a random variable with prior distribution $\pi_N(\lambda; \tau_N)$ derived below. Using Bayes theorem to derive the log-posterior distribution of $(\mathbf{f}, \lambda)$, we define the sparsity ℓ_1 information criterion (Sardy 2009) for density estimation as

$$\begin{aligned} \mathrm{SL_1IC}(\mathbf{f}, \lambda) &= -\sum_{i=1}^{N} \log f_i + \{\lambda \|B\mathbf{f}\|_1 - (N-1)\log(\lambda/2)\} \\ &\quad - \log \pi_N(\lambda; \tau_N) \quad \text{s.t.} \quad \mathbf{a}^T\mathbf{f} = 1. \end{aligned} \tag{15}$$

To derive the prior for λ, recall from the previous section and Appendix B that the bound λ_N^{TV} is based on the statistics $\lambda_{\mathbf{x}}$ defined by (13) in Property 1. The statistics $\lambda_{\mathbf{x}}$ is the smallest penalty to achieve piecewise constant density given a Uniform sample $\mathbf{x}$. Its distribution is degenerate as that of many supremum, but results from extreme value theory provide that

$$\varphi_{n_N}(\lambda_{\mathbf{x}}/\sigma_N - d_{n_N}) \longrightarrow_d G_0^{\theta},$$

where $G_0(x) = \exp(-\exp(-x))$ is the Gumbel distribution, $\sigma_N = \sqrt{NL_N(1-L_N)}$, $n_N = N/\sqrt{\log N} = 1/L_N$, $\varphi_{n_N} = \sqrt{2\log n_N}$ and $d_{n_N} = \varphi_{n_N} - (\log\log n_N + \log 4\pi - 2\log 2)/(2\varphi_{n_N})$. The extremal index $\theta \in (0, 1]$ (Hsing, Husler, Hüsler, and Leadbetter 1988; Coles 2001) stems from the dependence of the variables, owing to the correlation

between order statistics. Hence we check that, with the prescribed $\lambda_N^{\mathrm{TV}} = \sigma_N \varphi_{n_N}$, then $\mathrm{P}(\|\mathbf{w}\|_\infty \leq \lambda_N^{\mathrm{TV}}) \doteq 1 - \theta/\sqrt{\log n_N} \overset{N\to\infty}{\longrightarrow} 1$. But the asymptotic distribution also provides an approximate distribution for the smallest finite λ's that fit a piecewise constant density with a probability tending to one from Uniform samples. We consider this distribution for the prior density $\pi_N(\lambda;\tau) = G'(\lambda;\tau)$ with $G(\lambda;\tau) = G_0^\theta(\varphi_{n_N}(\lambda/(\sigma_N\tau) - d_{n_N}))$, where τ is calibrated to fit the true underlying Uniform density (i.e., $f_1 = \cdots = f_N$) for $\lambda = \lambda_N^{\mathrm{TV}}$ with a probability tending to one as the sample size tends to infinity. The calibrated value $\tau_N = \varphi_{n_N}^2/(N-1)$ is found as the approximate root in τ to the first order optimality condition of (15) with respect to λ:

$$\sum_{i=1}^{N-1} |f_{i+1} - f_i| - \frac{N-1}{\lambda} - \frac{G''(\lambda;\tau)}{G'(\lambda;\tau)} = 0 \quad \text{with} \quad \begin{cases} \sum_{i=1}^{N-1} |f_{i+1} - f_i| = 0 \\ \lambda = \lambda_N^{\mathrm{TV}} \end{cases} .$$

Like AIC or BIC, the $\mathrm{SL}_1\mathrm{IC}$ information criterion (15) is minimized alternatively over $\mathbf{f}$ (which is the topic of Section 4 below) and λ (which entails a simple univariate optimization) to provide both an estimate of the density and concomitantly a selection of the hyperparameter. The $\mathrm{SL}_1\mathrm{IC}$ choice of λ is asymptotically minimax in the nonparametric Gaussian sparse regression setting (Sardy 2009), where the estimator has the advantage of having a closed form expression. Here the density estimator cannot be expressed in a closed form, but is the solution of a non-differentiable ℓ_1 penalized likelihood expression (4). This makes the consistency or minimax properties of the proposed estimator harder to establish, and a challenging area of research. In the Gaussian regression setting, Mammen and van de Geer (1997) derive rates of convergence in bounded variation function classes for some optimal penalty λ (unknown in practice) and discuss pointwise limiting distributions. In particular they show that total variation-based estimators adapts to local smoothness.

4 Primal and duality reformulations and a block coordinate relaxation method

4.1 Primal reformulations

Computing the TV-penalized estimate for a given λ in (15) amounts to solving (4), which is not a trivial task owing to the nondifferentiability of the ℓ_1-penalty, high-dimensionality, and the presence of a constraint. Newton-type methods cannot be directly applied. One approach consists in transforming (7) into a linearly constrained problem with *smooth* convex objective function by using a standard trick of replacing $\|\mathbf{u}\|_1$ in the objective with $\mathbf{1}^T(\mathbf{u}^+ + \mathbf{u}^-)$ and replacing $\mathbf{u}$ in the constraint with $\mathbf{u}^+ - \mathbf{u}^-$, subject to $\mathbf{u}^+ \geq \mathbf{0}$, $\mathbf{u}^- \geq \mathbf{0}$. The resulting problem can be solved using an interior-point method (Koenker and Mizera 2006) or other methods; see, e.g., Bertsekas (1999, Chapter 2). However, this significantly increases size of the problem and special care is needed to exploit the sparsity structure of B.

Instead, we show below that (4) can be transformed into a convex program of the special form

$$\min_{\mathbf{v}=(v_1,\ldots,v_n)^T} g(C\mathbf{v} + \mathbf{e}) + h(\mathbf{v}), \tag{16}$$

where $n \geq 1$, $g(f_1, ..., f_N) = -\sum_{i=1}^N \log f_i$, C is an $N \times n$ matrix, $\mathbf{e} \in \Re^N$, and h is a separable convex function (i.e., $h(\mathbf{v}) = \sum_{i=1}^n h_i(v_i)$). Due to the separability of h (even though h may be nondifferentiable), this problem can be solved using a block coordinate relaxation (BCR) method, whereby, at each iteration, a block of coordinates of $\mathbf{v}$ is chosen and the objective function is minimized with respect to these coordinates while the other coordinates are held fixed at their current value (Sardy, Bruce, and Tseng 2000; Tseng 2001). Notice that the TV term $\|B\mathbf{f}\|_1$ in (4) is nondifferentiable and nonseparable (e.g., f_2 appears in two terms: $|f_2 - f_1|$ and

$|f_3 - f_2|$), so the BCR method cannot be directly applied to this problem.

If each block comprises a single coordinate, then the BCR method solves a succession of univariate subproblems until convergence, and is therefore simple to implement (e.g., back-fitting uses the BCR approach). In particular, starting with an initial guess $\mathbf{v}$ satisfying $C\mathbf{v} + \mathbf{e} > \mathbf{0}$, it chooses an $i \in \{1, ..., n\}$ and minimizes $g(C\mathbf{v} + \mathbf{e}) + h(\mathbf{v})$ with respect to v_i while holding the other coordinates of $\mathbf{v}$ fixed to generate a new $\mathbf{v}$.

The rule for choosing successive i is crucial for convergence. The classical *cyclic rule* chooses i in, for example, increasing order $i = 1, \ldots, n$ and then repeats this. The more general *essentially cyclic rule* entails that each $i \in \{1, \ldots, N\}$ is chosen at least once every T ($T \geq N$) successive iterations to allow different orders. The *optimal descent rule* (Sardy, Bruce, and Tseng 2000) chooses an i for which the minimum-magnitude partial derivative of the objective function with respect to v_i, i.e., $\min_{\eta \in \partial h_i(v_i)} \left| \frac{\partial J(\mathbf{v})}{\partial v_i} + \eta \right|$, is the largest, where $J(\mathbf{v}) = g(C\mathbf{v} + \mathbf{e})$. This yields the highest rate of descent and can considerably improves efficiency compared to the essentially cyclic rule.

Our first reformulation involves a change of variables based on the observation that $B\mathbf{f} = \mathbf{v}$, $\mathbf{a}^T\mathbf{f} = 1$ if and only if $\mathbf{f} = C\mathbf{v} + \mathbf{1}/b_N$, where $\mathbf{b} = R^T\mathbf{a}$, R is the upper triangular matrix of ones, and C is the $N \times (N-1)$ matrix with entries $C_{ij} = 1 - b_j/b_N$ if $j \geq i$ and otherwise $C_{ij} = -b_j/b_N$. Then (4) can be transformed into (16) with $\mathbf{e} = \mathbf{1}/b_N$. Global convergence of the BCR method when applied to this transformed problem can be established for the essentially cyclic rule based on Tseng (2001), and for the optimal descent rule based on Sardy, Bruce, and Tseng (2000, Theorem 2). The C matrix is not sparse however, so there is no closed form solution to each univariate subproblem. Consequently, an expensive line search is required at each iteration, a drawback that is avoided by the next reformulation.

4.2 Duality reformulation

Analogous to the iterated dual mode (IDM) algorithm (Sardy and Tseng 2004), we use convex programming duality theory (Rockafellar 1970; Rockafellar 1984) to obtain a dual problem of the form (16), to which we apply the BCR method. This dual approach has the advantage of not requiring an expensive line search at each iteration, as we now see. Let $L(\mathbf{f}, \mathbf{u}, \mathbf{w}, z)$ be the Lagrangian function defined by (8), with Lagrange multipliers $\mathbf{w}$ (for $B\mathbf{f} = \mathbf{u}$) and z (for $\mathbf{a}^T\mathbf{f} = 1$). The minimization (4) is equivalent to

$$\begin{aligned}
& \min_{\mathbf{f},\mathbf{u}} \max_{\mathbf{w},z} L(\mathbf{f}, \mathbf{u}, \mathbf{w}, z) \\
= \; & \max_{\mathbf{w},z} \min_{\mathbf{f},\mathbf{u}} L(\mathbf{f}, \mathbf{u}, \mathbf{w}, z) \\
= \; & \max_{\mathbf{w},z} \left(-z + \min_{\mathbf{f}} \sum_{i=1}^{N} [-\log f_i + f_i\{(B^T\mathbf{w})_i + za_i\}] + \min_{\mathbf{u}} \sum_{i=1}^{N-1} \lambda|u_i| - w_i u_i \right) \\
= \; & N + \max_{\|\mathbf{w}\|_\infty \le \lambda, z} -z + \sum_{i=1}^{N} \log\{(B^T\mathbf{w})_i + za_i\},
\end{aligned}$$

where the first equality uses a strong duality result for monotropic program, i.e., convex program with linear constraints and separable objective function (Rockafellar 1984, Sec. 11D). Dropping the constant terms, the above maximization problem can be rewritten as

$$\min_{\mathbf{w},z} g(B^T\mathbf{w} + z\mathbf{a}) + h(\mathbf{w}, z), \tag{17}$$

where $h(\mathbf{w}, z) = z$ if $\|\mathbf{w}\|_\infty \le \lambda$ and otherwise $h(\mathbf{w}, z) = +\infty$. Clearly (17) is of the form (16). Moreover, h is convex separable, with $h_i(w_i) = 0$ if $|w_i| \le \lambda$ and otherwise $h_i(w_i) = +\infty$, $i = 1, \ldots, N-1$, and $h_N(z) = z$. Convergence to the dual solution in $(\mathbf{w}, z)$ is guaranteed by Theorem 2 below, and the primal solution is then $\hat{f}_i = 1/\{(B^T\mathbf{w})_i + za_i\}$ for $i = 1, \ldots, N$.

Theorem 2. For any initial $(\mathbf{w}, z)$ with $\|\mathbf{w}\|_\infty \leq \lambda$ and $B^T\mathbf{w} + z\mathbf{a} > \mathbf{0}$, the sequence of iterates $(\mathbf{w}, z)$ generated by the dual BCR method, using the essentially cyclic rule, converges to the unique global minimum of (17).

For a proof, see Appendix C.

Thanks to the sparsity of B^T with at most two nonzero entries per column, the univariate subproblem in each w_j has a closed form solution. Consequently, the dual BCR method is efficient, even if programmed in `R`. The *taut string* estimator is faster than *total variation* with BCR as currently programmed however. Note also that the dual reformulation is still applicable for penalty functionals involving higher derivatives such as $\Phi(f) = \int |f''(x)|dx$.

5 Simulation

We perform a Monte Carlo simulation to compare the finite sample properties of four estimators: TV^1 (i.e., piecewise-linear) using universal penalty parameter or SL_1IC, *taut string* with piecewise-linear interpolation (Davies and Kovac 2004), *logspline* (Kooperberg and Stone 2002) and Gaussian *kernel* with global bandwidth (Sheather and Jones 1991) for benchmark.

Similar to Donoho and Johnstone (1994), we use four test densities, some with local features and some smooth ones (see Table 1 and Figure 1). Many measures of performance have been considered to compare estimators; see for instance Berlinet and Devroye (1994) for a comparative study of kernel density estimation. There is no best risk measure, but the L_2 and L_1 risks are complementary measures and the latter corresponds more to the human visual assessment. So we reported both mean integrated squared error (MISE) and mean integrated absolute error. Both L_2 and L_1 risks are calculated using a Riemann sum over a fine grid of 2^{13} points on the

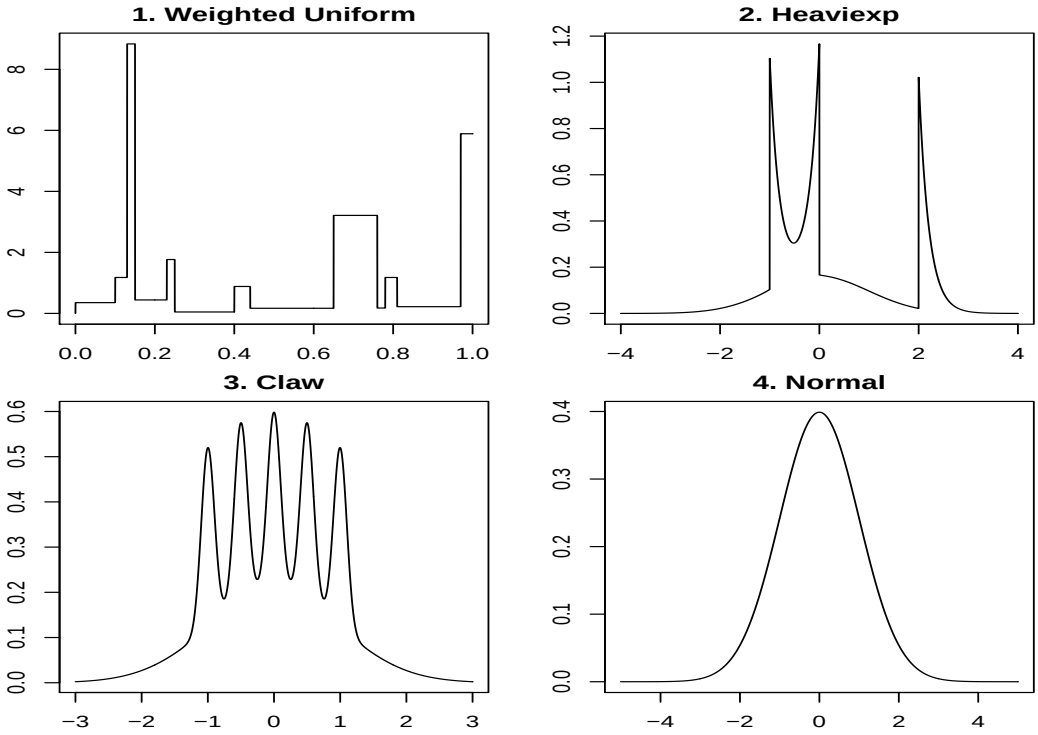

Figure 1: The four test densities used in the Monte Carlo simulation.

Table 1: Densities used in the simulation, their domain, the interval Ω on which the goodness-of-fit measures are calculated on a fine grid, and the number of modes.

	Densities	Domain	Ω	Number of modes
1.	Weighted Uniform[1]	$[0,1]$	$[0,1]$	6
2.	Heaviexp[2]	R	$[-4,4]$	3
3.	Claw †	R	$[-3,3]$	5
4.	Gaussian	R	$[-5,5]$	1

$f^1 = \sum_{i=1}^{13} w_i \mathrm{U}(a_i, b_i) / \sum_{i=1}^{13} w_i$ with $\begin{cases} \mathbf{w} = (1,1,5,1,1,.2,1,1,10,.1,1,1,5) \\ \mathbf{a} = (0,.1,.13,.15,.23,.25,.40,.44,.65,.76,.78,.81,.97) \\ \mathbf{b} = (.1,.13,.15,.23,.25,.40,.44,.65,.76,.78,.81,.97,1) \end{cases}$

$f^2 = \frac{1}{5}\{2 + \mathrm{Exp}(5)\} + \frac{1}{5}\{-\mathrm{Exp}(5)\} + \frac{1}{5}\{-1 + \mathrm{Exp}(5)\} + \frac{2}{5}\mathrm{N}(0,1)$,

† Marron and Wand (1992, #10 p.720)

interval Ω given in Table 1. We also report the number of modes to assess the quality of *taut string*, and compare it with the exact number. Since the standard error on the performance measures decreases as the sample size increases, the average is taken

over $Q \in \{800, 200, 50\}$ samples of respective sizes $N \in \{200, 800, 3200\}$, ranging from small to large to evaluate empirically their relative rates of convergence. For *logspline* we set the maximum number of knots to `maxknots`$= \lfloor\sqrt{N}\rfloor$ and `mind`$= 3$ after discussions with Charles Kooperberg; while this allows more flexibility to fit erratic densities, it also leads to more instability as observed in the second column of table 2 for the Weighted uniform density.

Table 2: Results of the simulation to compare density estimators on the four test functions in terms of MISE (L_2)/MIAE (L_1) and below, in parenthesis, the median of number of modes. The risks are multiplied by one hundred. In **bold**, best linewise.

N	Kernel	Logspline	Taut string	Total variation	
				universal	SL_1IC
1. Weighted Uniform					
200	129/58	110/56	95/46	86/45	**72/40**
	(6)	(4)	(3)	(4)	(4)
800	63/37	57†/39†	35/26	22/21	**19/19**
	(13)	(8†)	(6)	(6)	(7)
3200	29/22	`NA`‡	10/12	5.0/**10**	**4.9/10**
	(25)		(7)	(22)	(23)
2. Heaviexp					
200	9.4/38	9.3/38	9.9/**37**	9.0/39	**8.4/37**
	(6)	(4)	(3)	(2)	(3)
800	5.4/25	4.4/22	**3.1/20**	3.7/23	3.4/21
	(9)	(5)	(3)	(5)	(5)
3200	3.2/16	1.9/11	**1.0/12**	1.2/13	1.1/**12**
	(13)	(4)	(3)	(12)	(13)
3. Claw					
200	4.7/35	5.3/33	6.2/36	**4.0/32**	**4.0/32**
	(2)	(2)	(1)	(2)	(2)
800	**1.2/17**	**1.2/16**	2.2/20	2.0/22	1.9/21
	(9)	(5)	(5)	(6)	(7)
3200	**0.30/8.6**	0.42/9.3	0.67/12	0.59/12	0.58/12
	(10)	(5)	(5)	(13)	(13)
4. Gaussian					
200	**0.39/11**	0.90/16	1.5/22	0.79/17	0.80/17
	(1)	(1)	(1)	(1)	(1)
800	**0.12/6.4**	0.24/8.6	0.57/15	0.34/11	0.34/11
	(1)	(1)	(1)	(3)	(3)
3200	**0.040/3.7**	0.055/4.3	0.22/9.5	0.16/7.5	0.17/7.5
	(1)	(1)	(1)	(9)	(9)

† removing one infinite value; ‡ the code crashed.

We can draw the following conclusions based on the results of Table 2. First *total variation* is mostly better than *taut string* based on L_2 and L_1 risks, while *taut string* estimates the number of modes more accurately than *total variation*. Second *total variation* is better than *kernel* or *logspline* for the first two erratic densities, and worse for smooth densities. Finally, the universal rule is often improved by the sparsity ℓ_1 information criterion for small samples. All these empirical results corroborate the heuristics of the proposed method. In summary *total variation* is recommended when the underlying density is nonsmooth and when a global L_p measure is of interest.

6 Applications with ties

Like many real data, the two data sets we consider have ties due to some rounding. We use these two sets and bootstrapped samples from them to observe the sensitivity of the estimators to ties. Roeder (1990) and Izenman and Sommer (1988) considered the 'galaxy' and '1872 Hidalgo' data to estimate an underlying density using either a Bayesian analysis of mixtures and reversible jump McMC, or a kernel method compared to parametric likelihood ratio tests for mixtures, respectively. The 'galaxy data' consists of the velocities of 82 distant galaxies diverging from our own galaxy (reported with a 10^{-3} precision), and the '1872 Hidalgo data' consists of thickness measurements of 485 stamps printed on different types of paper, with only 62 distinct values due to rounding.

The focus of Roeder (1990) and Izenman and Sommer (1988) was the number of modes. So we consider this criterion, for which *taut string* performs best based on our simulations, to illustrate how *taut string* and *total variation* behave with ties. The two previous studies found $k_1 \in \{4, 5, 6, 7\}$ for the 'galaxy data', and $k_2 = 7$ with a nonparametric method and $k_2 = 3$ with a parametric one for the '1872 Hidalgo

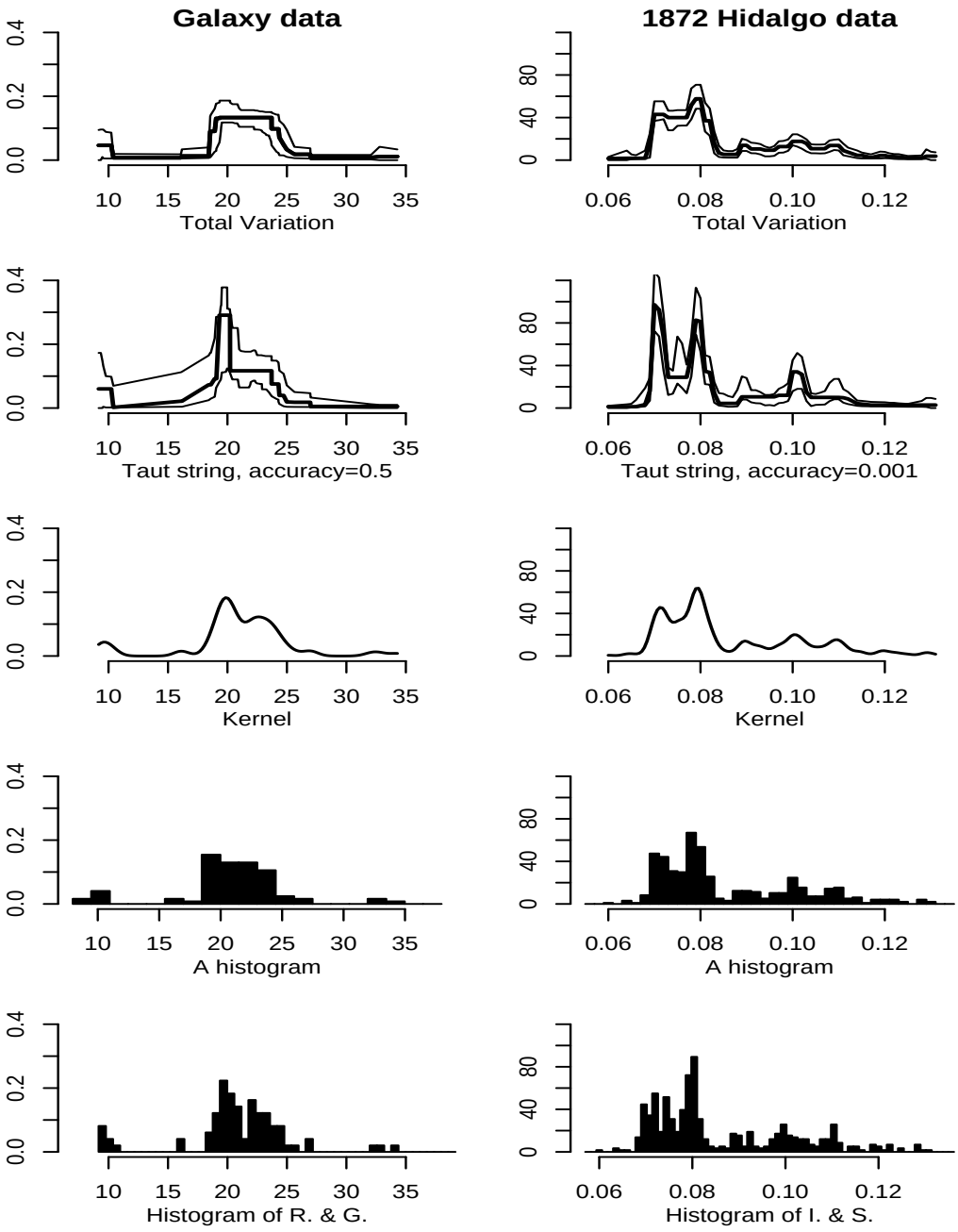

Figure 2: Five density estimates for each data set. From top to bottom: SL_1IC-TV[1], *taut string* (with `accuracy` parameter fixed to 0.5 (left) and 0.001 (right)), Gaussian kernel, and two histograms with different binning. The first two graphs show a pointwise 95% confidence interval estimated by bootstrap.

data'. Comparing these results with $k_1 = 2$ and $k_2 = 7$ obtained with *total variation* (plotted on the first row of Figure 2), and with $k_1 = 2$ and $k_2 = 3$ obtained with *taut string* (plotted on the second row of Figure 2), we see that the number of modes are on the down side with *taut string*. Note that the default value of the `accuracy` parameter of `pmden` (the `R` function for *taut string*) did not give good results for the '1872 Hidalgo data', so we set it to 0.001. We observe also that the histograms plotted in the published analysis (see fifth row of Figure 2) render the false impression of too many modes, due to the chosen bin width and left point values. To illustrate the histogram's sensitivity, we plotted two histograms with different bin width and left point values on the fourth row.

Assuming the underlying densities are smooth enough to consider the bootstrap as a reasonable technique to obtain consistent pointwise confidence intervals, we created more ties by bootstrapping the data. The first two lines of Figure 2 show a pointwise 95% confidence interval. We observe that the bootstrap confidence intervals hints for potential additional modes.

This experience with ties created by rounding or bootstrapping shows that *total variation* handles ties well without the need of an extra parameter. Pursuing this further, we investigate sensitivity to ties by repeating the Monte-Carlo simulation and creating ties by rounding to the second, second, and third decimal places for $N = 200, 800, 3200$ respectively. We report in Table 3 the results for *taut string* (with default parameter values) and *total variation* using the Heaviexp and Gaussian densities. We see that *total variation* is little affected by rounding of the data.

Table 3: Results of the simulation in terms of MISE to illustrate the sensitivity of *taut string* and *total variation* to ties.

N ties	Taut string without/with	SLIC-TV without/with	Taut string without/with	SLIC-TV without/with
	2. Heaviexp		**3. Claw**	
200	9.9/76	8.4/8.6	6.2/36	4.0/4.0
800	3.1/314	3.4/3.9	2.2/118	1.9/1.9
3200	1.0/1.0	1.1/1.2	0.67/0.67	0.58/0.58

7 Conclusions

The good empirical performance of the *total variation* density estimator motivates further theoretical analysis, for instance, proving consistency and minimax results in the space of functions of bounded variation and Besov spaces, and investigating conditions on the underlying functions for the bootstrap to be consistent. An important extension of the method can handle the deconvolution problem, i.e., when data are measured with error, or other inverse problems. The *total variation* estimator can also be generalized to higher dimension, for instance, assuming zeroth-order splines on a Voronoi tessellation of the sample points. Code is available from the first author.

8 Acknowledgments

We thank the Associate Editor and two anonymous referees for their careful reviews, which lead to a significant improvement.

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.

Berlinet, A. and L. Devroye (1994). A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Université de Paris 38*(3), 3–59.

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

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B 48*, 192–236.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag Inc.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 377–403.

Davies, P. L. and A. Kovac (2004). Densities, spectral densities and modality. *The Annals of Statistics 32*, 1093–1136.

Donoho, D. L., I. Johnstone, G. Kerkyacharian, and D. Picard (1995). Wavelet shrinkage: Asymptotia? (with discussion). *Journal of the Royal Statistical Society, Series B 57*(2), 301–369.

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

Donoho, D. L., I. M. Johnstone, J. C. Hoch, and A. S. Stern (1992). Maximum entropy and the nearly black object (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological 54*, 41–81.

Eggermont, P. P. B. and V. N. LaRiccia (1999). Optimal convergence rates for Good's nonparametric maximum likelihood density estimator. *The Annals of Statistics 27*(5), 1600–1615.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of*

Computational and Graphical Statistics 7, 397–416.

Geman, S. and D. Geman (1984). Stochastic relaxation. Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 61*, 721–741.

Good, I. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: M.I.T. Press.

Good, I. (1971). A nonparametric roughness penalty for probability densities. *Nature 229*, 29–30.

Good, I. J. and R. A. Gaskins (1971). Nonparametric roughness penalties for probability densities. *Biometrika 58*, 255–277.

Hall, W. J. and J. A. Wellner (1980). Confidence bands for a survival curve from censored data. *Biometrika 67*, 133–143.

Hsing, T., J. Husler, J. Hüsler, and M. R. Leadbetter (1988). On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields 78*, 97–112.

Izenman, A. J. and C. J. Sommer (1988). Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association 83*, 941–953.

Koenker, R. and I. Mizera (2006). Density estimation by total variation regularization. In *Advances in Statistical Modeling and Inference, Essays in Honor of Kjell A. Doksum*. World Scientific.

Koenker, R., P. Ng, and S. Portnoy (1994). Quantile smoothing splines. *Biometrika 81*, 673–680.

Kooperberg, C. and C. J. Stone (1991). A study of logspline density estimation. *Computational Statistics and Data Analysis 12*, 327–347.

Kooperberg, C. and C. J. Stone (2002). Logspline density estimation with free knots. *Computational Statistics and Data Analysis 12*, 327–347.

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

Mammen, E. and S. van de Geer (1997). Locally adaptive regression splines. *The Annals of Statistics 25*(1), 387–413.

Marron, J. S. and M. P. Wand (1992). Exact mean integrated squared error. *The Annals of Statistics 20*, 712–736.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computation 9*, 363–379.

Penev, S. and L. Dechevsky (1997). On non-negative wavelet-based density estimators. *Journal of Nonparametric Statistics 7*, 365–394.

Pinheiro, A. and B. Vidakovic (1997). Estimating the square root of a density via compactly supported wavelets. *Computational Statistics and Data Analysis 25*, 399–415.

Renaud, O. (2002). Sensitivity and other properties of wavelet regression and density estimators. *Statistica Sinica 12*(4), 1275–1290.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton: Princeton University Press.

Rockafellar, R. T. (1984). *Network Flows and Monotropic Programming*. New-York: Wiley-Interscience; republished by Athena Scientific, Belmont, 1998.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association 85*, 617–624.

Sardy, S. (2009). Adaptive posterior mode estimation of a sparse sequence for model selection. *Scandinavian Journal of Statistics to appear*.

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

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

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

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B, Methodological 53*, 683–690.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics 10*, 795–810.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Stone, C. J. (1990). Large sample inference for logspline model. *Annals of Statistics 18*, 717–741.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions

(with discussion). *Journal of the Royal Statistical Society, Series B 36*, 111–147.

Tibshirani, R. (1995). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 57*, 267–288.

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

Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications 109*, 475–494.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Wiley.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Willett, R. and R. D. Nowak (2003, August). Multiscale Density Estimation. Technical report, Rice University.

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

Paul Tseng, Department of Mathematics, Box 354350, University of Washington, Seattle, WA, 98195-4350, U.S.A.

E-mail: tseng@math.washington.edu

A Two-sided discontinuity at the knot

Suppose $\Omega = [0,1]$, $N = 1$, $0 < x_1 = \frac{1}{2}$. Then (2) with $\Phi = \Phi_{\text{TV}}$ given by (3) reduces to

$$\min_{f_1,u,v\geq 0} -\log(f_1) + \lambda|f_1 - u| + \lambda|f_1 - v| \quad \text{subject to} \quad \frac{u}{2} + \frac{v}{2} = 1,$$

where $f_1 = f(x_1)$, and u, v are the values of the constant functions on the subintervals $[0, x_1)$ and $(x_1, 1]$, respectively. For $\lambda < \frac{1}{2}$, it can be verified that

$$f_1 = \frac{1}{2\lambda}, \quad u = v = 1$$

is the optimal solution (since it satisfies the Karush-Kuhn-Tucker optimality condition for this convex optimization problem). The corresponding f is

$$f(x) = \begin{cases} \frac{1}{2\lambda} & \text{if } x = \frac{1}{2}; \\ 1 & \text{else} \end{cases} \qquad \forall x \in [0, 1],$$

which is 2-sided discontinuous at the knot.

B Derivation of λ_N^{TV}

Let $\mathbf{x} = (x_1, \ldots, x_N)$ be the order statistics of a $\mathrm{U}(0, 1)$ sample, Property 1 asserts the existence of a *finite* $\lambda_{\mathbf{x}}$ for some K such that the estimate $\hat{\mathbf{f}}_{\lambda_{\mathbf{x}}}$ equals a piecewise constant density with K successive identical values. Specifically $\lambda_{\mathbf{x}} = \|\mathbf{w}\|_\infty$, where $\mathbf{w}$ is given by (12). The goal of λ_N^{TV} is to control the extremal behavior of $\lambda_{\mathbf{x}}$. The vector $\mathbf{w}$ can be broken into $n_N = N/K$ nearly independent blocks $\mathbf{w}_i = (w_{i1}, \ldots, w_{i(K-1)})$, each of which converging to a Brownian bridge process on $[0, K/N]$. To see this, consider

$$\mathbf{w}_{1k} \;=\; -N(\frac{x_{k+1} + x_k}{2} - x_1) + \frac{Nk}{K}(\frac{x_{K+1} + x_K}{2} - x_1),$$

and

$$\mathrm{P}(\|\mathbf{w}_1\|_\infty \le \lambda) \;=\; \mathrm{P}(\max_{k=1,\ldots,K-1} |N(\frac{x_{k+1} + x_k}{2} - x_1) - \frac{Nk}{K}(\frac{x_{K+1} + x_K}{2} - x_1)| \le \lambda)$$

$$\doteq \quad \mathrm{P}(\max_{k=1,\ldots,K-1} |\sqrt{N}(x_k - k/N))| \leq \lambda/\sqrt{N}),$$

since $\mathrm{cor}(x_k, x_{k+1}) \stackrel{N\to\infty}{\longrightarrow} 1$ and $Nx_K/K \stackrel{N\to\infty}{\longrightarrow} 1$. The uniform quantile process converges to a Brownian bridge process B on $[0,1]$, so

$$\mathrm{P}(\|\mathbf{w}\|_\infty \leq \lambda/\sqrt{N}) \quad \doteq \quad \left(\mathrm{P}(\sup_{0\leq t\leq L_N} |B(t)| \leq \lambda/\sqrt{N})\right)^{n_N},$$

where $L_N = K/N$. Hall and Wellner (1980, (2.9) p.136) give the explicit formula

$$\begin{aligned}\mathrm{P}(\sup_{0\leq t\leq L} |B(t)| \leq \tilde{\lambda}) \quad =: \quad & G_L(\tilde{\lambda}) = 1 - 2\Phi(-\tilde{\lambda}\{L(1-L)\}^{-1/2}) \\ & +2\sum_{k=1}^{\infty}(-1)^k \exp(-2k^2\tilde{\lambda}^2)(\phi(r(2k+d)) - \phi(r(2k-d))),\end{aligned}$$

where Φ is the Gaussian cdf, $r = \lambda((1-L)/L)^{1/2}$ and $d = (1-L)^{-1}$. This shows that, for L_N small, $\sup_{0\leq t\leq L_N} |B(t)|$ essentially behaves as $\mathrm{N}(0, \{L_N(1-L_N)\})$. Hence $\lambda_N^{\mathrm{TV}}/\sqrt{N}$ must controls the extremal behavior of n_N independent centered Gaussian variables with variance $L_N(1-L_N)$. Choosing $\lambda_N^{\mathrm{TV}}/\sqrt{N}/\sqrt{L_N(1-L_N)} = \sqrt{2\log n_N}$ guarantees that $\mathrm{P}(\|\mathbf{w}\|_\infty \leq \lambda_N^{\mathrm{TV}}) \stackrel{N\to\infty}{\longrightarrow} 1$. Due to the strong dependence between successive values with the TV penalty $+\lambda\sum_{i=1}^{N-1}|f_{i+1} - f_i|$ as opposed to the disjoint multiscale penalties (14) with wavelets, we must choose a small value for L_N; the choice $L_N \sim \sqrt{\log N}/N$ works well in practice to estimate smooth and non-smooth densities.

C Proof of Theorem 2

The dual problem (17) is of the form $\min_{\mathbf{w},z} J(\mathbf{w}, z) + h(\mathbf{w}, z)$, where $J(\mathbf{w}, z) = g(B^T\mathbf{w} + \mathbf{a}z)$. Since $\mathbf{a} > \mathbf{0}$, then $(\mathbf{0}, 1) \in \mathrm{dom}J$ so $\mathrm{dom}J$ is nonempty. Moreover, J is continuously differentiable on $\mathrm{dom}J$, which is an open set. Thus, J satisfies

Assumption 1 in Tseng (2001). Also, g is strictly convex and the columns of B^T and $\mathbf{a}$ are linearly independent, so J is strictly convex. Since h is convex, $J+h$ is strictly convex. Each iterate $(\mathbf{w}, z)$ generated by the BCR method satisfies $\|\mathbf{w}\|_\infty \leq \lambda$. Since $J(\mathbf{w}, z)+h(\mathbf{w}, z) = J(\mathbf{w}, z)+z$ is non-increasing after each iteration, this and the fact that $\log(\cdot)$ tends to infinity sublinearly implies z is bounded. Thus each iterate $(\mathbf{w}, z)$ lies in a compact subset of dom$J$. Then Lemma 3.1 and Theorem 4.1(a) in Tseng (2001) imply that each accumulation point of the iterate sequence is a stationary point of $J + h$. Since $J + h$ is strictly convex, the stationary point is the unique global minimum.

D Proof of Property 3

Consider any $1 \leq i \leq N-1$ with $\frac{1}{a_i} \geq \frac{1}{a_{i+1}}$. Suppose $\hat{f}_{\lambda,i} < \hat{f}_{\lambda,i+1}$ for some $\lambda \geq 0$. We will show that, by increasing f_i and decreasing f_{i+1}, we can obtain a lower objective for (4), thus contradicting $\hat{\mathbf{f}}_\lambda$ being a global minimum of (4). In particular, consider moving from $\hat{\mathbf{f}}_\lambda$ in the direction $\mathbf{d}$ with null components except $d_i = 1$ and $d_{i+1} = -a_i/a_{i+1}$. Then $\mathbf{a}^T\mathbf{d} = 0$, so that $\mathbf{a}^T(\hat{\mathbf{f}}_\lambda + \alpha\mathbf{d}) = 1$ for all $\alpha \in \Re$. Moreover $\hat{\mathbf{f}}_\lambda + \alpha\mathbf{d} > \mathbf{0}$ and $\|B(\hat{\mathbf{f}}_\lambda + \alpha\mathbf{d})\|_1 \leq \|B\hat{\mathbf{f}}_\lambda\|_1$ for all $\alpha > 0$ sufficiently small, and the directional derivative of $-\sum_{i=1}^N \log f_i$ in the direction of $\mathbf{d}$ at $\hat{\mathbf{f}}_\lambda$ is $-\frac{1}{\hat{f}_{\lambda,i}} + \frac{1}{\hat{f}_{\lambda,i+1}}\frac{a_i}{a_{i+1}}$. Since $\frac{1}{a_i} \geq \frac{1}{a_{i+1}} > 0$ and $0 < \hat{f}_{\lambda,i} < \hat{f}_{\lambda,i+1}$, this directional derivative is negative. Thus $\hat{\mathbf{f}}_\lambda + \alpha\mathbf{d}$ improves (4) strictly compared to $\hat{\mathbf{f}}_\lambda$ for all $\alpha > 0$ sufficiently small. This contradicts $\hat{\mathbf{f}}_\lambda$ being a global minimum. □