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

By SYLVAIN SARDY*
Université de Genève and Ecole Polytechnique Fédérale de Lausanne
and PAUL TSENG
University of Washington, U.S.A.

We propose a density estimator based on penalized likelihood and total variation. Driven by a single smoothing parameter, the nonlinear estimator has the properties of being locally adaptive and positive everywhere without a log- or root-transform. For the fast selection of the smoothing parameter we employ the sparsity ℓ_1 information criterion. Furthermore the estimated density has the advantage of being the solution to a convex programming problem; we solve it by an easy-to-implement relaxation algorithm of which we prove convergence. Finally, we compare the finite sample performance of our estimator to existing ones on a Monte Carlo simulation and on two real data sets. The new estimator is particularly efficient at estimating densities with sharp features from large samples, and is stable to the presence of ties (due to rounding or bootstrapping).

Key Words: Convex programming; Penalized likelihood; Sparsity ℓ_1 information criterion; Total variation; Universal prior.

*Address for correspondence: Faculté des sciences de base, Institut de Mathématiques Station 8, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland. Email: Sylvain.Sardy@epfl.ch

1 Introduction

An old problem in statistics (Silverman, 1986; Scott, 1992; Simonoff, 1996) is the estimation of a density function f from a sample of observations $x_1, \ldots, x_M$. Ties might be present due to rounding or bootstrapping: let $x_1, \ldots, x_N$ be the sample with no ties, and let $n_1, \ldots, n_N$ be the number of ties at the order statistics $x_{(1)}, \ldots, x_{(N)}$. Nonparametric methods reduce the possible modeling biases in situations where a parametric model is difficult to guess. Writing the nonparametric log-likelihood function

$$l(f; x_1, \ldots, x_M) = \sum_{i=1}^{N} n_i \log f(x_i), \tag{1}$$

and maximizing it over all densities f with the constraint of integrating to unity leads to the degenerate nonparametric maximum likelihood estimate $\hat{f}(x) = \frac{1}{M} \sum_{i=1}^{N} n_i \delta_{x_i}(x)$, where $\delta_{x_i}(\cdot)$ is the Dirac measure at x_i. It is degenerate in that its total variation is unbounded.

Many nonparametric estimators are based on regularization (Tikhonov, 1963), by adding a penalty or, equivalently, a constraint to (1), to obtain a smoother and more useful estimate. For instance, the histogram estimator is defined as the unique maximum likelihood estimate constrained to be piecewise constant on bins. In a pioneering paper, Good and Gaskins (1971) add a roughness penalty functional $\Phi(f)$ to (1) to define the functional nonparametric penalized likelihood estimate $\hat{f}_\lambda$ as the solution to

$$\max_f \; l(f; x_1, \ldots, x_M) - \lambda \Phi(f) \quad \text{s.t.} \quad \int f(x) dx = 1, \tag{2}$$

("s.t." is short for "subject to"), where $\lambda > 0$ is the smoothing parameter: the estimate tends to the degenerate nonparametric estimate when $\lambda \rightarrow 0$, and to the parametric maximum likelihood estimate in the kernel of the penalty Φ (i.e., the subspace of densities f such that $\Phi(f)$ is minimal) when $\lambda \rightarrow \infty$. 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. They have three important properties. First, the existing penalties intrinsically assume f is differentiable everywhere or the points of nondifferentiability can be ignored–a mathematical assumption which causes practical difficulties when the density is nondifferentiable everywhere, for instance with jumps or peaks. Second, it is often argued that basing the penalty on derivatives of $\sqrt{f}$ or $\log f$ has the advantage of avoiding negative estimates, but we contend that it is redundant to insure positivity. Indeed, the terms $\log f(x_i)$ in (1) are already natural barriers against negative values of $f(x_i)$ at all observations, and f would be positive everywhere if we make an additional mild assumption that f is piecewise monotone between x_is. Third, penalizing smoothness on a square root- or log-scale might cause adverse effects because smoothness at low and high density values is not penalized equally.

Penalized likelihood served as a framework for smoothing splines estimators (Wahba, 1990). O'Sullivan (1988) developed a spline-based estimate to calculate an approximation to the log-density estimate of Silverman (1982). Kooperberg and Stone (1991) derived the logspline estimator which selects spline knots: the location of potential knots are recommended to be near order statistics and their number is automatically selected based on an AIC-like criterion. Silverman (1982) and Stone (1990) derived asymptotic optimal convergence properties under smoothness assumptions.

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 Vidakovic (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 logdensity estimate regularized by a total variation penalty on the first derivative.

In this article, we propose a density estimator based on a total variation penalty, a functional that does not even have first-order differentiability. Consequently, the new estimator defined in Section 2 has the ability to efficiently estimate nonsmooth densities. To select the smoothing parameter, Section 3 derives an information criterion based on the Gumbel prior for the hyperparameter. Section 4 derives and proves convergence of two relaxation algorithms, based on primal and dual transformations, to calculate the estimate. Section 5 investigates the finite sample properties of the new density estimator in comparison with existing ones on a Monte Carlo simulation. We then consider two real data sets in Section 6, and draw some conclusions in Section 7.

2 Total variation penalized likelihood

2.1 Univariate case

We propose to regularize the maximum likelihood estimate (1) by penalizing the log-likelihood function with the total variation of the density. So we choose $\Phi(f) = \mathrm{TV}(f)$ in (2) with

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

where the sup is taken over all possible partitions of the domain of the univariate density. If one assumes that f is absolutely continuous, then total variation matches a more conventional smoothness measure since $\mathrm{TV}(f) = \int |f'(x)|dx$. However total variation (3) is not tractable numerically, because looking at all possible partitions (which includes the histogram bins) is not computationally feasible, even if restricted to a combinatorial problem on a fine grid. So, as O'Sullivan (1988) developed an approximation to calculate the functional estimate of Silverman (1982), we derive an approximation to total variation under the following assumptions.

Since our primary goal is the estimation of the density $f_i = f(x_{(i)})$ at the N unique order statistics and since no information is available in between, we avoid unnecessary variance by assuming piecewise monotone interpolation between midpoints of order statistics, and monotone extrapolation to zero outside the range of the observations. This interpolation scheme is reminiscent of logspline which places knots at order statistics, in the sense that the modeling of the underlying density depends on the sample. We will see however that, for total variation, the "knot" selection has the advantage of being driven by a single smoothing parameter. To avoid any arbitrary specification of the unknown underlying domain and for reasons linked to the universal rule (see Section 3.1), we consider the total variation function on the range $[x_{(1)}, x_{(N)}]$ of the data. With these assumptions, the total variation of f has a simple vector expression in $\mathbf{f} = (f_1, \dots f_N)$:

$$\mathrm{TV}(f) = \sum_{i=1}^{N-1} |f(x_{(i+1)}) - f(x_{(i)})| = \sum_{i=1}^{N-1} |f_{i+1} - f_i|. \tag{4}$$

As a byproduct, piecewise monotone interpolation defines a positive estimate everywhere, provided that all f_i's are positive. We further assume that the interpolation and the extrapolation schemes allow us to write the functional integral constraint in a vector linear form

$$1 = \int f(x)dx = \mathbf{a}'\mathbf{f}, \tag{5}$$

where the weights $\mathbf{a} = \mathbf{a}_{\mathbf{x}}$ are function of the order statistics, but not of $\mathbf{f}$. For instance, we assume in the following that the density is piecewise linear between the $(M-1)$ midpoints, and equals zero outside the range of the observations.

We are now ready to describe our estimator. Given a smoothing parameter λ, the total variation penalized likelihood estimate $\hat{\mathbf{f}}_\lambda$ of the density f at the order statistics solves

$$\min_{\mathbf{f}} - \sum_{i=1}^{N} n_i \log f_i + \lambda \|B\mathbf{f}\|_1, \quad \text{s.t.} \quad \mathbf{a}'\mathbf{f} = 1, \tag{6}$$

where B is the sparse $(N-1)\times N$ 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). Thanks to the log-barrier likelihood and to the monotone interpolation, the estimate is positive everywhere. Moreover, ties are naturally handled in the likelihood part (on the contrary, taut string (Davies and Kovac, 2004) seems unstable when ties are present, as we observe in Section 6). The total variation estimator also has the following properties.

Property 1. The Karush–Kuhn–Tucker first-order optimality conditions of (6) are

$$\begin{aligned} f_i - n_i/\{(B'\mathbf{w})_i + z a_i\} &= 0 \quad \forall i, & (7) \\ -1 + \sum_{i=1}^{N} a_i f_i &= 0, & (8) \\ \|\mathbf{w}\|_\infty &\leq \lambda, & (9) \end{aligned}$$

So the Uniform density $f_i = c$ with $c = 1/(x_{(N)} - x_{(1)})$ for $i = 1, \ldots, N$ (i.e., the function $\mathbf{f}$ in total variation kernel such that $\|B\mathbf{f}\|_1 = 0$ on the range of the data) is the solution to (6) for a *finite* $\lambda_{\mathbf{x}} = \|\mathbf{w}\|_\infty < \infty$, where $(\mathbf{w}, z)$ satisfies (7) to (9), i.e., $w_i = N \sum_{j=1}^{i} a_j - ic$ and $z = N$.
At the other end, as $\lambda \to 0^+$, the estimate converges to the empirical estimate $\hat{f}_{\lambda,i} = \frac{n_i}{a_i M}$, since a continuity argument shows that each accumulation point of $\hat{\mathbf{f}}_\lambda$ is a minimum of (6) with $\lambda = 0$.

Property 2. By strict convexity of the cost function and by linearity of the constraint, any local minimizer of (6) is its unique strict global minimizer. This nice property is similar to the convexity property in O'Sullivan (1988) and contrasts with the multimodality encountered with logspline (Kooperberg and Stone, 1991; Kooperberg and Stone, 2002).

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

Property 4. Under affine transformation of the data $Y = \alpha + \beta X$ with $\beta > 0$, the estimate defined by minimizing (6) satisfies $\hat{f}^X_\lambda(x) = \beta \hat{f}^Y_{\lambda\beta}(\alpha + \beta x)$.

2.2 Multivariate case

If the sample is large enough to allow efficient estimation, nonparametric techniques can be useful to estimate densities of two or more variables. In this section, we sketch a possible extension to higher dimension. Assuming the multivariate density $f \in L^1(U)$, where U is an open subset of R^p $(p \geq 1)$, the more general multivariate definition of total variation

$$\mathrm{TV}(f) = \sup\left\{ \int_U f \mathrm{div}\varphi dx \mid \varphi \in C^1_c(U; R^p), |\varphi| \leq 1 \right\}$$

becomes numerically tractable, if we further assume the function is piecewise constant on a partition of the support U of $f(\cdot)$. The Voronoi tessellation is the natural multivariate extension of univariate midpoint splits used in Section 2.1. The Voronoi polygon assigned to the point X_i consists of all the points in U that are closer to X_i than to any other point $X_j, j \neq i$. Defining Voronoi polygons for the sample $X_1, \ldots, X_N$ results in the Voronoi tessellation. Two points X_i and X_j are said to be Voronoi neighbors if the Voronoi polygons enclosing them share a common edge $E_{i,j}$ of length $\|E_{i,j}\|$. Let ∂_i be the index set of all Voronoi neighbors of X_i. Assuming f is a piecewise constant function on the Voronoi tessellation, its total variation becomes

$$\mathrm{TV}(f) = \sum_{i=1}^{n} \sum_{j \in \partial_i \setminus \{1,\ldots,i-1\}} |f_i - f_j| \cdot \|E_{i,j}\|.$$

The corresponding penalized likelihood is similar to (6), where the components of $\mathbf{a}$ are the areas of the Voronoi polygons and where the ℓ_1 differences are weighted by $\|E_{i,j}\|$. The dual relaxation algorithm of Section 4.2 can therefore be employed to find the unique optimum of the penalized likelihood in the multivariate situation as well.

3 Selection of the smoothing parameter

The smoothing 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 indexing a continuous class of ℓ_1 penalized likelihood models are cross validation (Stone, 1974), 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 the universal rule (Donoho and Johnstone, 1994). 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-based estimator near the solution. The Bayesian information criterion was intended for variable selection (ℓ_0), not for ℓ_1-based penalized likelihood. Finally, 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.

3.1 Sparsity ℓ_1 information criterion

The universal penalty $\lambda_N = \sqrt{2 \log N}$ (Donoho and Johnstone, 1994) originally developed in regression for Gaussian wavelets smoothing (Donoho and Johnstone, 1994) has the property of being near minimax (Donoho, Johnstone, Kerkyacharian, and Picard, 1995) and of controlling the maximum amplitude of i.i.d. Gaussian random variables to fit the underlying constant function with a probability tending to one. Direct extension of the latter property to total variation density estimation defined by (6) would give $\lambda_N^\star = \sqrt{N \log(\log N)/2}$ to control the maximum amplitude of a discretized Brownian bridge (see Appendix A) to fit the underlying Uniform density with a probability tending to one. This bound oversmooths in most applications however.

We employ instead the sparsity ℓ_1 information criterion (Sardy, 2006) based on a Bayesian interpretation of total variation and on the universal prior for λ that we derive below. Borrowing from Markov random field smoothing (Besag, 1986; Geman and Geman, 1984), the vector of true density $\mathbf{f}$ at the N samples can be seen as the realization of a first-order pairwise Laplace Markov random field shifted and rescaled to make it a density. We moreover consider the regularization parameter as a random variable with prior $\pi_\lambda(\lambda;\tau)$ that we derive in Proposition 1 below. Hence using Bayes theorem, the posterior joint distribution of $\mathbf{f}$ and λ leads to the sparsity ℓ_1 information criterion for total variation density estimation

$$\begin{aligned}\mathrm{SL_1IC}(\mathbf{f},\lambda) \quad = \quad & -\sum_{i=1}^{N} n_i \log f_i + \lambda \sum_{i=1}^{N-1} |f_{i+1}-f_i| - (N-1)\log\lambda \\ & -\log \pi_\lambda(\lambda;\tau) \quad \text{s.t.} \quad \mathbf{a}'\mathbf{f}=1. \end{aligned} \tag{10}$$

Minimizing SL$_1$IC over $\mathbf{f}$ and λ provides an estimate of the density as well as a selection of the hyperparameter at once. Derived in Appendix B, the universal prior $\pi_\lambda(\lambda;\tau)$ defined in Proposition 1 below is based on asymptotic considerations linked to Property 1.

Proposition 1 *Let $G_0(x) = \exp(-\exp(-x))$ be the Gumbel distribution and let*

$$G(\lambda;\tau) = G_0[4\{\lambda/\tau - (\log N)/4\}]. \tag{11}$$

Then the universal prior used to estimate the density $\mathbf{f}$ *and the hyperparameter λ based on SL$_1$IC (10) is defined as*

$$\pi_\lambda(\lambda;\tau) = G'(\lambda;\tau_N) \quad \text{with} \quad \tau_N = \log(N\log N)/N. \tag{12}$$

With this prior, SL$_1$IC is both a stable and computationally efficient method to select the regularization parameter, especially by comparison with cross validation.

3.2 Kullback-Leibler V-fold cross validation

Another approach, though more computationally intensive, would be to turn to cross validation. For a given smoothing parameter λ, the Kullback-Leibler information

$$\mathrm{KL}(f,\hat f_\lambda) = \int \log\{f(x)/\hat f_\lambda(x)\}\ f(x)dx \tag{13}$$

measures the quality of the fit between the true but unknown density f and the estimated density $\hat f_\lambda$. This quantity can be estimated by cross validation (CV) or by the bootstrap (see Efron and Tibshirani (1993) for a review). We consider V-fold CV which is computationally feasible as long as V is not too large: it consists of grouping the data set into V randomly selected disjoint sets $\mathbf{x}_v = \{x_i, i \in S_v\}$, $v = 1, \ldots, V$, using $\mathbf{x}_{-v} = \{x_i, i \notin S_v\}$ to estimate f by $\hat{f}_{\lambda,-v}$, and using the interpolation scheme to predict $\{f(x_i)\}_{i \in S_v}$ by $\{\hat{f}_{\lambda,-v}(x_i)\}_{i \in S_v}$. Repeating the same operation V times yields the estimate $\mathrm{CV}(\lambda) = -\sum_{v=1}^{V} \sum_{i \in S_v} \log \hat{f}_{\lambda,-v}(x_i)$ of (13). Calculating this criteria for several λ's, we select the smoothing parameter λ_{CV} which minimizes it.

Van der Laan, Dudoit, and Keles (2004) studied the choice of V and show asymptotic equivalence with a benchmark selection based on the calibration set as long as N/V goes to infinity. This condition is satisfied by 2-fold CV, but rules out leave-one-out CV for which $V = N$. They moreover show asymptotic equivalence with a benchmark selection based on the entire N observations as long as $1/V_N$ converges slowly enough to zero. This rules out 2-fold CV, but is satisfied for instance by $V_N = O(\log N)$.

4 Computation of total variation estimate

Calculating the total variation penalized likelihood estimate is not a trivial task owing to the nondifferentiability, high-dimensionality and the constraint of (6). Using the constraint $\mathbf{a}'\mathbf{f} = 1$, we could express $f_N = (1 - \sum_{i=1}^{N} a_i f_i)/a_N$ and rewrite (6) as an unconstrained minimization problem in $f_1, \ldots, f_{N-1}$. However, the resulting objective function remains nondifferentiable, so the Newton-Raphson method cannot be used. The total variation term is moreover non-separable (e.g., f_2 appears in two terms: $|f_2 - f_1|$ and $|f_3 - f_2|$), so a block coordinate relaxation (BCR) method could not be directly used either (Tseng, 2001). Koenker and Mizera (2006) use interior point methods. Instead we propose a simple BCR method applied to two transformations of the penalized likelihood into an unconstrained problem of the form

$$\min_{\mathbf{v}=(v_1,\ldots,v_n)'} h_0(\mathbf{v}) + h(\mathbf{v}),$$

where h_0 a differentiable convex function and h is a separable nondifferentiable convex function (i.e., $h(\mathbf{v}) = \sum_{i=1}^{n} h_i(v_i)$). The BCR method is simple to implement as it successively solves subproblems of dimension one until convergence, and is computationally efficient using the dual transformation. In particular, starting with an initial guess $\mathbf{v}$, it chooses an $i \in \{1, ..., n\}$ and minimizes $h_0(\mathbf{v}) + h(\mathbf{v})$ with respect to v_i while holding the other components of $\mathbf{v}$ fixed. This is then repeated with the new $\mathbf{v}$ and so on. The rule for choosing i at each iteration is crucial for convergence. In Appendix C, we consider two rules for which we prove convergence of BCR.

4.1 Primal transformation

The obvious transformation detailed in Appendix C is to make a change of variables such that the nondifferentiable part $h(\cdot)$ in the transformed problem

$$\min_{\mathbf{v}} g(C\mathbf{v}) + h(\mathbf{v}), \tag{14}$$

becomes separable. Then Theorem 1 below guarantees convergence of the BCR method. The C matrix is not sparse however, so there is no closed form solution to the convex univariate subproblems. Consequently, the line search required at each iteration slows down the algorithm.

Theorem 1. For any initial $\mathbf{v}$ with $C\mathbf{v} > -\mathbf{1}/b_N$, the sequence of iterates generated by the BCR method, using either the essentially cyclic rule or the optimal descent rule, converges to the unique global minimum of (14).

For the proof see Appendix C.

4.2 Dual transformation

Instead we can use an extension of the iterated dual mode (IDM) algorithm (Sardy and Tseng, 2004) to handle constraints. The second transformation uses results from convex programming duality theory (Rockafellar, 1970; Rockafellar, 1984) and has computational advantages over the primal method. Introducing the Lagrange multipliers $\mathbf{w}$ and z, the minimization (6) is equivalent to

$$\begin{aligned}
& \min_{\mathbf{f},\mathbf{u}} \max_{\mathbf{w},z} -\sum_{i=1}^{N} n_i \log f_i + \lambda \sum_{i=1}^{N-1} |u_i| + \mathbf{w}'(B\mathbf{f} - \mathbf{u}) + z(\mathbf{a}'\mathbf{f} - 1) \\
= \quad & \max_{\mathbf{w},z} -z + \min_{\mathbf{f}} \sum_{i=1}^{N} [-n_i \log f_i + f_i\{(B'\mathbf{w})_i + z a_i\}] + \min_{\mathbf{u}} \sum_{i=1}^{N-1} \lambda |u_i| - w_i u_i \\
= \quad & M - \sum_{i=1}^{N} n_i \log n_i + \max_{\|\mathbf{w}\|_\infty \le \lambda, z} -z + \sum_{i=1}^{N} n_i \log\{(B'\mathbf{w})_i + z a_i\},
\end{aligned}$$

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

$$\min_{\mathbf{w},z} z + g(B'\mathbf{w} + z\mathbf{a}) + h(\mathbf{w}), \tag{15}$$

where $g(v_1, ..., v_N) = -\sum_{i=1}^{N} n_i \log v_i$ and $h(\mathbf{w}) = 0$ if $\|\mathbf{w}\|_\infty \le \lambda$ and $+\infty$ otherwise. The nondifferentiable term $h(\mathbf{w})$ is convex and has a separable form. Notice that $(\mathbf{w}, z)$ needs to satisfy $\|\mathbf{w}\|_\infty \le \lambda$ and $B'\mathbf{w} + z\mathbf{a} > \mathbf{0}$. Convergence to the dual solution in $(\mathbf{w}, z)$ is guaranteed by Theorem 2 below, and the primal solution is then $\hat{f}_i = n_i / \{(B'\mathbf{w})_i + za_i\}$ for $i = 1, \ldots, N$.

Theorem 2. For any initial $(\mathbf{w}, z)$ with $\|\mathbf{w}\|_\infty \le \lambda$ and $B'\mathbf{w} + z\mathbf{a} > \mathbf{0}$, the sequence of iterates generated by the BCR method to (15), using the essentially cyclic rule, converges to the unique global minimum of (14).

For the proof see Appendix D.

Thanks to the sparsity of B' with at most two nonzero entries per column, the univariate subproblems in w_j have a closed form solution. Consequently the dual relaxation algorithm is efficient, even if programmed in a high-level language like `R`. Note that the dual transformation is still applicable for 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: total variation using SL$_1$IC, taut string (Davies and Kovac, 2004), logspline (Kooperberg and Stone, 2002) and rectangular kernel with global bandwidth (Sheather and Jones, 1991) for benchmark.

Table 1: Densities used in the simulation, their domain, the interval Ω on which the mean integrated squared error is calculated on a fine grid.

	Densities	Domain	Ω
1.	Sharp peak †	R	$[0, 12]$
2.	Claw ‡	R	$[-3, 3]$
3.	Weighted Uniform[3]	$[0, 1]$	$[0, 1]$
4.	Heaviexp[4]	R	$[-5, 5]$

† Hansen and Kooperberg (2002), ‡ Marron and Wand (1992, #10 p.720)

$$f^3(x) = \sum_{i=1}^{13} w_i \mathrm{U}(a_i, b_i) / \sum_{i=1}^{13} w_i$$

$$\text{with } \begin{cases} \mathbf{w} = (1, 1, 5, 1, 1, .2, 1, 1, 10, .1, 1, 1, 5) \\ \mathbf{a} = (0, .1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81, .97) \\ \mathbf{b} = (.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81, .97, 1) \end{cases}$$

$$f^4(x) = \tfrac{1}{4}\{2 + \mathrm{Exp}(5)\} + \tfrac{1}{4}\{-\mathrm{Exp}(1)\} + \tfrac{1}{2}\mathrm{N}(0, 1)$$

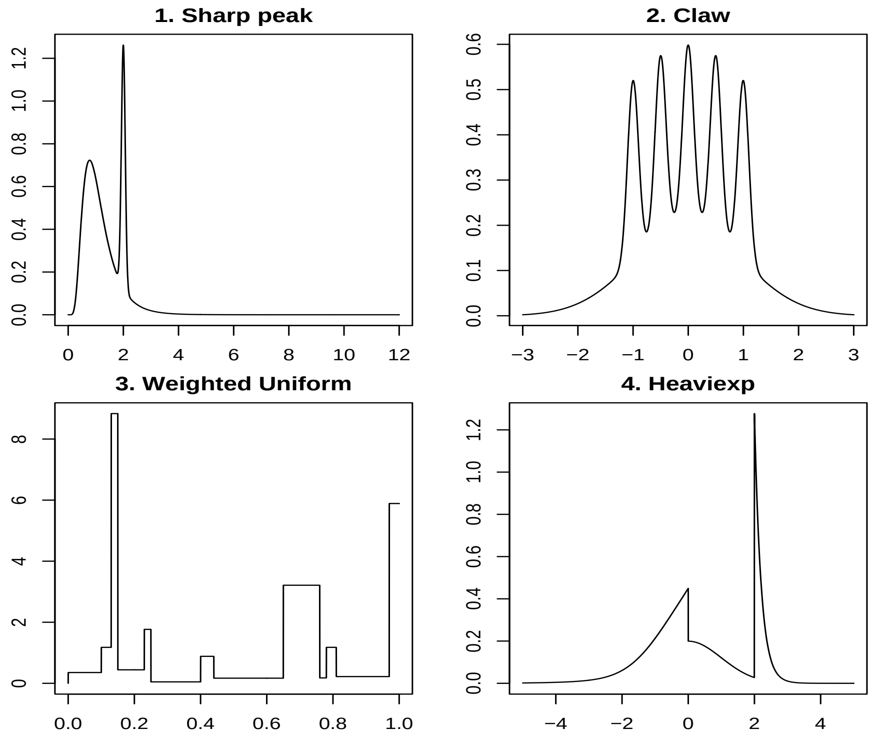

Figure 1: The four test densities used in the Monte Carlo simulation.

Similarly to Donoho and Johnstone (1994), we use four test densities with heterogeneous features (see Table 1 and Figure 1). As in Kooperberg and Stone (2002), the criteria we use to compare estimators is the mean integrated squared error, $\mathrm{MISE}(\hat{f}, f) = \mathrm{E}\int_{\Omega}(\hat{f}(x) - f(x))^2 dx$, approximated by a Riemann sum on a fine grid of 2^{13} points on the interval Ω given in Table 1. Since the standard error on the estimation of the MISE decreases as the sample size increases, the expectation is taken over $Q \in \{1000, 500, 100\}$ samples of respective sizes $M \in \{200, 800, 3200\}$ ranging from small to large to evaluate the relative empirical rates of convergence. We also considered the expected Kullback–Leibler information to compare the estimators, and found correlated results with MISE. Note that for logspline, we had to increase the default value of the maximum number of knots to fit the complexity of the Weighted Uniform density; this revealed occasional numerical problems.

Table 2: Results of the simulation to compare density estimators on the four test functions (average MISE ×100).

N	Kernel	Logspline	Taut string	Total variation
1. Sharp peak				
200	4.9	4.1	5.1	**3.8**
800	1.4	1.2	1.7	**1.4**
3200	0.41	0.38	0.62	**0.56**
2. Claw				
200	5.8	5.3	5.6	**3.4**
800	1.5	1.4	2.1	**1.3**
3200	0.32	0.40	0.70	**0.54**
3. Weighted Uniform				
200	161	89	93	**65**
800	78	55	36	**18**
3200	36	43	9.0	**5.7**
4. Heaviexp				
200	7.8	3.8	7.6	**4.7**
800	4.7	1.8	2.5	**2.0**
3200	2.7	0.95	**0.62**	0.63

NOTE: **in bold**, the best linewise between Taut string and Total variation; underlined, the best linewise between all four estimators. (standard error of the order of the precision reported).

We can draw the following conclusions on the finite sample performance of the estimators considered based on the simulation results of Table 2. First the sparsity ℓ_1 information criterion provides total variation with both a fast and efficient selection of the regularization parameter. Second total variation is overall better than taut string in term of mean integrated squared error; if instead the criterion was the correct number of modes detected, then taut string would perform better than total variation which tends to detect false modes. Finally kernel and logspline are competitive estimators only when the true density does not have sharp features. So total variation is overall best in term of mean integrated squared error.

6 Application

We consider two data sets with multimodality: the 'galaxy data' (Roeder, 1990) consists of the velocities of 82 distant galaxies diverging from our own galaxy, and the '1872 Hidalgo data' (Izenman and Sommer, 1988) consists of thickness measurements of 485 stamps printed on different types of paper.

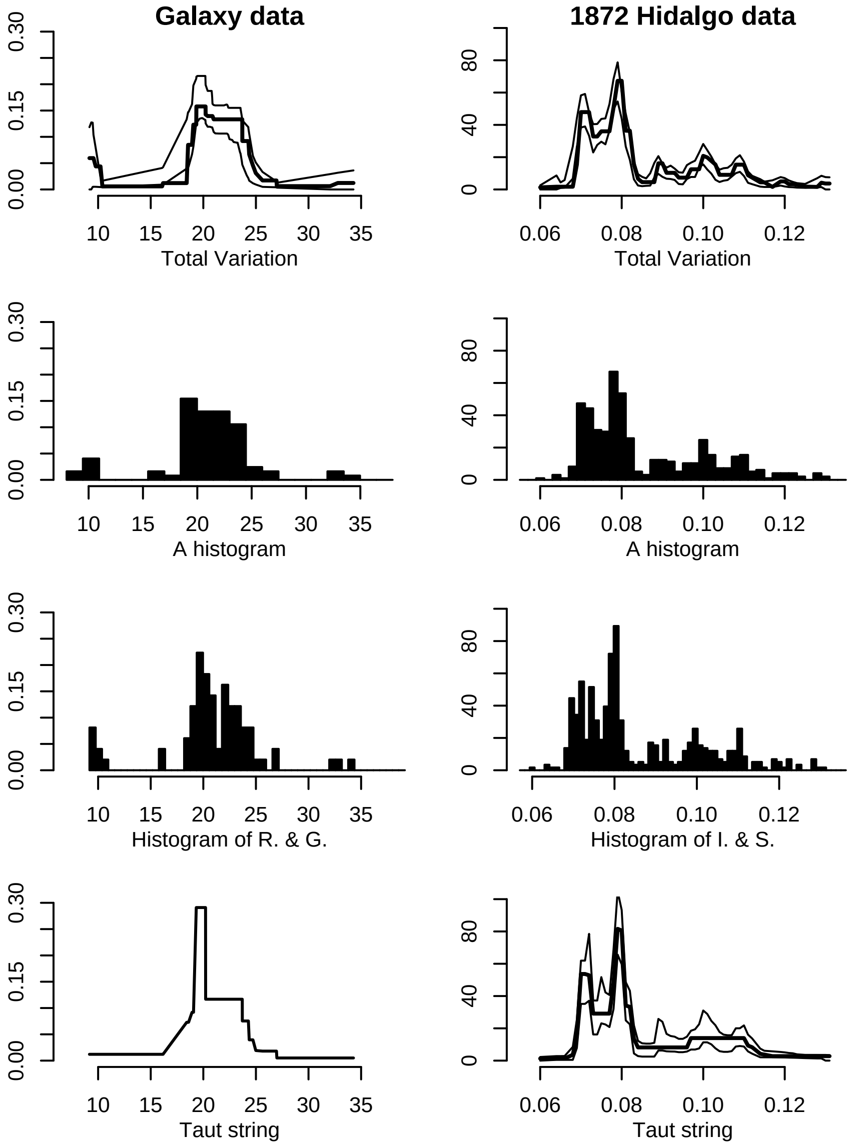

Figure 2: Four density estimates for each data set. From top to bottom: total variation, two histograms with different binning, taut string. The thin lines are bootstrapped pointwise confidence intervals.

Ties are numerous since the precision of the measurements is 0.001. These data sets were respectively used to illustrate a Bayesian analysis of mixtures to test models of varying dimensions using reversible jump McMC, and to compare the kernel method to parametric likelihood ratio tests for mixtures.

The focus of the analysis is the number of modes. The histograms plotted in the published analysis (see third line of Figure 2) are based on a subjective bin width and left point values, and might render the false impression of too many modes. To illustrate the histogram's sensitivity, we plotted two other histograms (with different bin width and left point values) on the second line of Figure 2. The models of Richardson and Green (1997) and Izenman and Sommer (1988) lead to the result that a likely number of modes is $k_1 \in \{4, 5, 6, 7\}$ for the 'galaxy data', and $k_2 = 7$ and $k_2 = 3$ with a nonparametric and parametric approaches for the '1872 Hidalgo data'.

Comparing these results to $k_1 = 3$ and $k_2 = 7$ obtained with total variation plotted on the first line of Figure 2, and $k_1 = 1$ and $k_2 = 3$ obtained with taut string plotted on the fourth line of Figure 2, we see that the estimation of the number of modes is consistent with previous results. Taut string, which does not handle ties naturally, gives a significantly smaller number of modes however. The first and fourth lines of Figure 2 also show a 90% pointwise confidence intervals based on 100 bootstraped replicates, except for taut string that experienced numerical problems for the 'galaxy data' because of the ties; for the '1872 Hidalgo data', the taut string confidence interval shows potential additional modes not revealed by its estimate. On the contrary total variation behaves as well on the bootstrap samples as on the original one.

7 Conclusions

The total variation estimator shows overall better estimating properties than its considered competitors. The efficient estimation of both smooth and nonsmooth densities is especially remarkable considered that only a single regularization parameter has to be selected to accommodate to regions of varying smoothness. The downside is a bumpier look than taut string. The good properties of the total variation estimator are due to the stability of solving of convex program (6) and to the method of selection of λ using the sparsity ℓ_1 information criterion $\mathrm{SL_1IC}$. Furthermore total variation handles ties naturally (as opposed to taut string) which is particularly useful for bootstrapping the procedure, or when data have been truncated as in our two applications. Finally it generalizes to higher dimension as described in

Section 2.2.

8 Acknowledgments

This work was partially funded by grant FNS PP001–106465.

A Bound $\lambda_N^\star$

For a given sample $x_1, \ldots, x_N$, Property 1 states the existence of a *finite* $\lambda_{\mathbf{x}}$ (which depends on the data) such that the estimate $\hat{\mathbf{f}}_{\lambda_{\mathbf{x}}}$ solution to (6) equals the Uniform density $\mathbf{f}_0$. The standard universal rule is defined as the smallest smoothing parameter $\lambda_N^\star$ (which depends only on the sample size) such that the estimate $\hat{\mathbf{f}}_{\lambda_N^\star}$ equals the Uniform density $\mathbf{f}_0$ with a probability tending to one when the size N of the Uniform sample goes to infinity. Without loss of generality, we can restrict to $f_0 = \mathrm{U}[0,1]$ thanks to Property 4. So, given a sample $U_1, \ldots, U_N \overset{\text{i.i.d.}}{\sim} \mathrm{U}[0,1]$, we work on the rescaled variables $X_i = (U_i - U_{(1)})/(U_{(N)} - U_{(1)})$, such that $X_{(1)} = 0$ and $X_{(N)} = 1$ and the density to estimate is $\mathbf{f} = \mathbf{1}$; note that the two samples are asymptotically equivalent. Using the dual formulation of (6) and its KKT first-order optimality conditions (7) to (9), the dual vector $(\mathbf{w}, z)$ is the unique global minimizer of the dual problem (15). So the estimate $\hat{\mathbf{f}}_\lambda$ fits the Uniform (i.e., $\hat{f}_i = 1$) provided λ is at least as large as $\lambda_{\mathbf{x}} = \|\mathbf{w}\|_\infty$, where $\mathbf{w}$ satisfies all the KKT conditions (with $n_i = 1$ with probability one since the Uniform is continuous). Consequently, we have that:

$$\begin{aligned}
\mathrm{P}\left(\hat{\mathbf{f}}_\lambda = \mathbf{1}\right) &= \mathrm{P}\{\|\mathbf{w}\|_\infty \leq \lambda : \; [B' \;\; \mathbf{a}] \begin{pmatrix} \mathbf{w} \\ z \end{pmatrix} = \mathbf{1}\} \\
&= \mathrm{P}\{\max_{i=1,\ldots,N-1} |N \sum_{j=1}^{i} a_j - i| \leq \lambda\}.
\end{aligned}$$

Assuming piecewise linear interpolation, the vector $\mathbf{a}$ satisfying (5) is $a_1 = X_{(2)}/2$, $a_i = (X_{(i+1)} - X_{(i-1)})/2$ for $i = 2, \ldots, N-1$, and $a_N = (1 - X_{(N-1)})/2$. So $\sum_{j=1}^{i} a_j$ has a simple expression and

$$\begin{aligned}
\mathrm{P}(\hat{\mathbf{f}}_\lambda = \mathbf{1}) &= \mathrm{P}(\max_{i=1,\ldots,N-1} |N/2(X_{(i)} + X_{(i+1)}) - i| \leq \lambda) \\
&\overset{N\to\infty}{\longrightarrow} \mathrm{P}(\max_{i=1,\ldots,N} \sqrt{N}|X_{(i)} - i/N| \leq \lambda/\sqrt{N})
\end{aligned}$$

since $\mathrm{cor}(X_{(i)}, X_{(i+1)}) \overset{N\to\infty}{\longrightarrow} 1$. Moreover $\max_{i=1,\ldots,N} \sqrt{N}(X_{(i)} - i/N)$ is an asymptotic pivot, since the uniform quantile process converges to a Brownian bridge W on $[0, 1]$, so

$$\mathrm{P}(\hat{\mathbf{f}}_\lambda = \mathbf{1}) \quad \overset{N\to\infty}{\longrightarrow} \quad \mathrm{P}(\|W\|_\infty \leq \frac{\lambda}{\sqrt{N}}) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2(\frac{\lambda}{\sqrt{N}})^2).$$

Using bounds on a Brownian bridge (Shorack and Wellner, 1986), we find that the bound $\lambda_N^\star := \sqrt{N\log(\log N)/2}$ controls the convergence to one at the rate $O(1/\log N)$. □

B Universal prior

The idea of a universal penalty comes from wavelet smoothing. The first level Haar wavelets for instance, defined by $\frac{1}{\sqrt{2}}(f_{2k} - f_{2k-1})$ for $k = 1, \ldots, N/2$ (we assume N is even), are reminiscent of the successive finite differences $f_{i+1} - f_i$ for $i = 1, \ldots, N$ used by the total variation penalty in (6). However, while the Haar wavelets' finite differences are separable, the ones used for total variation are not separable (for instance f_2 is used in $|f_2 - f_1|$ and $|f_3 - f_2|$). This causes inflation of the bound that controls the maximum amplitude of the Brownian bridge (see Appendix A), and causes oversmoothing. So instead of considering all finite order differences, we derive the universal prior by considering problem (6) with the $(N/2) \times N$ matrix $\tilde{B}$ corresponding to every other finite differences in place of B. Following similar derivations as in Appendix A, the smallest hyperparameter λ that fits a pairwise constant density based on a sample of size N from $f_0 = \mathrm{U}[0, 1]$ (i.e., to satisfy the KKT conditions (7) to (9) with $f_{2k-1} = f_{2k}$ for $k = 1, \ldots, N/2$ and $B = \tilde{B}$) is $\lambda_{\mathbf{X}} = \|\mathbf{W}\|_\infty$ with $W_k = z/2(a_{2k-1} - a_{2k}) = z/4(-X_{(2k-2)} + X_{(2k-1)} + X_{(2k)} - X_{(2k+1)})$ and $z = \sum_{k=1}^{N/2} 2/f_{2k} \approx N$ since $f_0 = \mathrm{U}[0, 1]$. Each $|W_k|$ converges in distribution to an exponential distribution Exp(4) since the density function of W_k is given by (Ramallingam, 1989)

$$\begin{aligned} f_W(w) \quad &= \quad 1_{(-N/4,0]}(w)\ 2(1 + \frac{4w}{N})^{N-1} + 1_{(0,N/4]}(w)\ 2(1 - \frac{4w}{N})^{N-1} \\ &\overset{N\to\infty}{\longrightarrow} \quad 2\exp(-4|w|). \end{aligned}$$

Neglecting the correlation between W_k's, results from extreme value theory guarantee that

$$4(\|\mathbf{W}\|_\infty - \frac{\log N}{4}) \longrightarrow_d G_0(x),$$

where $G_0(x) = \exp(-\exp(-x))$ is the Gumbel distribution. Finally we calibrate the scale parameter τ of the uniform prior (11) such that the universal penalty $\lambda_N = \log(N \log N)/4$ is the root to the first order optimality condition of (10) with respect to λ to fit $f_i = 1$ for all $i = 1, \ldots, N$:

$$\sum_{i=1}^{N-1} |f_{i+1} - f_i| - \frac{N-1}{\lambda} - \frac{\pi'}{\pi}(\lambda; \tau) = 0 \quad \text{with} \quad \left\{ \begin{array}{l} \sum_{i=1}^{N-1} |f_{i+1} - f_i| = 0 \\ \lambda = \lambda_N \end{array} \right. .$$

The approximate root is $\tau_N = 4\lambda_N/N$. □

C Proof of Theorem 1

Letting $u_n = f_n - f_{n+1}$ and $u_N = f_N$, the minimization (6) becomes

$$\min_{\mathbf{u}} - \sum_{i=1}^{N} n_i \log(\sum_{j=i}^{N} u_j) + \lambda \sum_{i=1}^{N-1} |u_i|, \quad \text{s.t.} \quad \mathbf{b}'\mathbf{u} = 1, \tag{16}$$

where $\mathbf{b} = T'\mathbf{a}$ and T is the upper triangular matrix of ones. Since $b_N = \sum_{j=1}^{N} a_j$ and $a_j > 0$, we have that $b_N > 0$. From the equality constraint, we also have that $u_N = (1 - \sum_{i=1}^{N-1} b_i u_i)/b_N$, so letting $\mathbf{v} = (u_1, \ldots, u_{N-1})$ and $C_{ni} = 1(i \geq n) - b_i/b_N$ be the entries of the $N \times (N-1)$ matrix C, the problem writes as (14) where $g(w_1, \ldots, w_N) = -\sum_{i=1}^{N} n_i \log(1/b_N + w_i)$. Since $b_N > 0$, then $\mathbf{0} \in \text{dom} g = \{\mathbf{w} | \mathbf{w} > \mathbf{1}/b_N\}$. So $\text{dom}(g \circ C)$ is also nonempty and open. Moreover, $g \circ C$ is continuously differentiable on its domain. Thus, $g \circ C$ satisfies Assumption 1 in Tseng (2001). Also, g and h are convex functions, so $f = g \circ C + h$ is convex (and hence pseudoconvex). Since h has bounded level sets, then so does f (since $\log(\cdot)$ tends to infinity sublinearly). Thus, the kth iterate $\mathbf{v}^k$ $(k = 0, 1, ...)$ generated by the BCR method is bounded. In fact, they lie in a compact subset of $\text{dom}(g \circ C)$.

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. Lemma 3.1 and Theorem 4.1(a) in Tseng (2001) imply that each accumulation point of $\mathbf{v}^0, \mathbf{v}^1, ...$ is a stationary point of f, the unique global minimum of f by strict convexity.

The *optimal descent rule* (Sardy, Bruce, and Tseng, 2000) chooses an i for which the minimum-magnitude partial derivative of the cost function with respect to v_i, i.e.,

$$\min_{\eta \in \partial h_i(v_i)} \left| \frac{\partial h_0(\mathbf{v})}{\partial v_i} + \eta \right|,$$

is the largest. This yields the highest rate of descent at the current $\mathbf{v}$, and often considerably improves the efficiency of the algorithm. Since $\mathbf{v}^0, \mathbf{v}^1, ...$ lie in a compact subset of $\mathrm{dom}(g \circ C)$, over which $g \circ C$ is continuously differentiable, and $g \circ C$ is strictly convex, the proof of Theorem 2 in Sardy, Bruce, and Tseng (2000) can be applied with little change to yield that each accumulation point of $\mathbf{v}^0, \mathbf{v}^1, ...$ is a stationary point of f. Since f is strictly convex, then the accumulation point is the unique global minimum of f. □

D Proof of Theorem 2

The dual problem (15) is of the form $\min_{\mathbf{w},z} f(\mathbf{w}, z) = f_0(\mathbf{w}, z) + h(\mathbf{w})$, where $f_0(\mathbf{w}, z) = z + g(B'\mathbf{w} + \mathbf{a}z)$. Since $\mathbf{a} > \mathbf{0}$, then $(\mathbf{0}, 1) \in \mathrm{dom} f_0$ so $\mathrm{dom} f_0$ is nonempty. Moreover, f_0 is continuously differentiable on $\mathrm{dom} f_0$, which is an open set. Thus, f_0 satisfies Assumption 1 in Tseng (2001). Also, f_0 and h are convex functions, so f is convex (and hence pseudoconvex). The kth iterate $(\mathbf{w}^k, z^k)$ generated by the BCR method satisfies $\|\mathbf{w}^k\|_\infty \leq \lambda$. Since $f(\mathbf{w}^k, z^k) = f_0(\mathbf{w}^k, z^k)$ is non-increasing with k, this and the fact that $\log(\cdot)$ tends to infinity sublinearly implies z^k is bounded. Thus $(\mathbf{w}^k, z^k)$ lies in a compact subset of $\mathrm{dom} f_0$. Then Lemma 3.1 and Theorem 4.1(a) in Tseng (2001) imply that each accumulation point of $(\mathbf{w}^0, z^0), (\mathbf{w}^1, z^1), \ldots$ is a stationary point of f. □

E Proof of Property 3

Consider any $1 \leq i \leq N-1$ with $\frac{n_i}{a_i} \geq \frac{n_{i+1}}{a_{i+1}}$. 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 cost for (6), thus contradicting $\hat{\mathbf{f}}_\lambda$ being a global minimum of (6). In particular, consider moving from $\hat{\mathbf{f}}_\lambda$ in the direction $\mathbf{d}$ with components

$$d_j = \begin{cases} 1 & \text{if } j = i \\ -a_i/a_{i+1} & \text{if } j = i+1 \\ 0 & \text{else} \end{cases} .$$

Then $\mathbf{a}'\mathbf{d} = 0$, so that $\mathbf{a}'(\hat{\mathbf{f}}_\lambda + \alpha\mathbf{d}) = 1$ for all $\alpha \in \Re$. Moreover $\hat{\mathbf{f}}_\lambda + \alpha\mathbf{d} > \mathbf{0}$ and $\|B(\hat{\mathbf{f}}_\lambda + \alpha\mathbf{d})\|_1 \leq \|B\hat{\mathbf{f}}_\lambda\|_1$ (where the matrix B in defined in Section 4.2) for all $\alpha > 0$ sufficiently small, and the directional derivative of $-\sum_{i=1}^N n_i \log f_i$ in the direction of $\mathbf{d}$ at $\hat{\mathbf{f}}_\lambda$ is $-\frac{n_i}{\hat{f}_{\lambda,i}} + \frac{n_{i+1}}{\hat{f}_{\lambda,i+1}} \frac{a_i}{a_{i+1}}$. Since $\frac{n_i}{a_i} \geq \frac{n_{i+1}}{a_{i+1}} > 0$ 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 (6) strictly compared to $\hat{\mathbf{f}}_\lambda$ for all $\alpha > 0$ sufficiently small. This contradicts $\hat{\mathbf{f}}_\lambda$ being a global minimum. □

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.

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

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

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

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

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap.* Chapman & Hall Ltd.

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

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

Hansen, M. H. and Kooperberg, C. (2002). Spline adaptation in extended linear models (with discussion). *Statistical Science 17*(1), 2–20.

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

Koenker, R. and Mizera, I. (2006). Density estimation by total variation regularization. *preprint*.

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

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

Kooperberg, C. and Stone, C. J. (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.

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

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

Ramallingam, T. (1989). Symbolic computing the exact distributions of L-statistics from a Uniform distribution. *Annals of the Institute of Statistical Mathematics* **41**, 677–681.

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

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (Disc: p758-792) (Corr: 1998V60 p661). *Journal of the Royal Statistical Society, Series B, Methodological* **59**, 731–758.

Rockafellar, R. (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. (2006). Estimating the dimension of parametric and nonparametric ℓ_ν penalized likelihood models for adaptive sparsity. *submitted*.

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

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

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

Shorack, G. and Wellner, J. (1986). *Empirical Processes with Applications to Statistics.* Wiley.

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.

Van der Laan, M. J., Dudoit, S., and Keles, S. (2004). Asymptotic optimality of likelihood based cross-validation. *Statistical Applications in Genetics and Molecular Biology* **3**, Article 4.

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