On the statistical analysis of smoothing by maximizing dirty Markov random field posterior distributions

By SYLVAIN SARDY*

Swiss Federal Institute of Technology, Switzerland

and PAUL TSENG

University of Washington, U.S.A.

We consider Bayesian nonparametric function estimation using a Markov random field prior based on the Laplace distribution. The paper describes efficient methods for finding the exact maximum a posteriori estimate, which handle constraints naturally and avoid the problems posed by non-differentiability of the posterior distribution; the methods also make links to spline and wavelet smoothers, and to a dual posterior distribution. Three automatic smoothing parameter selection procedures are described, by empirical Bayes, by two-fold cross-validation, and through a universal rule for the Laplace prior. Monte Carlo simulation with Gaussian and Poisson responses shows that the new estimator can give better estimates of non-smooth functions than a similar prior based on the Gaussian distribution, or than wavelet-based competitors. Applications are given to spectral density estimation and to Poisson image denoising.

Key Words: Convex programming; Empirical Bayes; ICM; Maximum a posteriori; Nondifferentiable Bayes prior; Total variation.

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

1 Introduction

One common goal of regression in signal processing and related applications is to recover an underlying regression function from noisy measurements, usually taken on an equally-spaced lattice, most often in one or two dimensions. Many nonparametric regression techniques have been developed. The running mean/median, the kernel and nearest neighbor smoothers use a running average of the data in a certain neighborhood. Other techniques assume an expansion of the underlying signal on trigonometric functions, splines or wavelets. This article treats the underlying signal as the realization of a Markov random field (MRF) that, as we will see, has connections with spline and wavelet smoothers. The Bayesian MRF framework, formalized by Geman and Geman (1984) and Besag (1986), is based on two assumptions:

1. The noisy scene $\mathbf{s} = (s_1, \ldots, s_N)$, conditioned on the unknown true scene $\mathbf{x} = (x_1, \ldots, x_N)$, has likelihood and log-likelihood function which we write as $L(\mathbf{x} \mid \mathbf{s})$ and $l(\mathbf{x} \mid \mathbf{s})$ respectively. Independence of the measurements s_n conditioned on the true scene is not assumed.

2. The true scene $\mathbf{x}$ is the realization of a locally dependent MRF with prior density $p(\mathbf{x})$. It is locally dependent in that the conditional distribution of x_n given all other values $\mathbf{x}_{-n}$ depends only on its neighborhood ∂_n,

$$p(x_n \mid \mathbf{x}_{-n}) = p(x_n \mid \mathbf{x}_{\partial_n}). \tag{1}$$

   Under certain conditions (Besag (1974) and references therein), the joint distribution $p(\mathbf{x})$ of the true scene can then be obtained from the conditional distributions by means of the Hammersley–Clifford theorem (Clifford, 1990). As we will see, the choice of a good prior is important.

Bayes' theorem updates the prior distribution $p(\mathbf{x})$ to the posterior distribution $p(\mathbf{x} \mid \mathbf{s}) \propto L(\mathbf{x} \mid \mathbf{s})p(\mathbf{x})$, which leads to two main methods of estimation:

- The Bayesian posterior mean estimate $\hat{\mathbf{x}}_{\mathrm{PMean}} = \mathrm{E}(\mathbf{X} \mid \mathbf{s}) = \int \mathbf{x} \, dP(\mathbf{x} \mid \mathbf{s})$ minimizes *pixelwise* the integrated risk with respect to the L_2 loss function. The high-dimensional integral can be calculated by Markov chain Monte Carlo methods. A related estimate $\hat{\mathbf{x}}_{\mathrm{PMedian}}$ is the pixelwise posterior median that corresponds to an L_1 loss function.

- The posterior mode or maximum a posteriori (MAP) estimate

$$\hat{\mathbf{x}}_{\mathrm{MAP}} \quad = \quad \arg\max_{\mathbf{x}} \; l(\mathbf{x} \mid \mathbf{s}) + \log p(\mathbf{x}) \tag{2}$$

  maximizes *imagewise* the posterior density. It can be regarded as a penalized likelihood estimate. The solution can equivalently be found by minimizing the negative posterior density, a formulation that we use in the sequel. The MAP estimate can possibly be found by Newton, relaxation or simulated annealing methods.

While the likelihood is mainly determined by the noise of the measurement device, the choice of prior distribution $p(\mathbf{x})$ is more subjective. A class of pairwise-difference MRF priors is defined by the conditional densities of the form:

$$p(x_n \mid \mathbf{x}_{\partial_n}; \lambda) \propto \exp\left\{ -\lambda \sum_{i \in \partial_n} \phi(x_n - x_i) \right\}, \tag{3}$$

where ∂_n is the neighborhood of pixel n. For clarity and simplicity of exposition, we consider the first-order neighborhood (left and right for 1D signals and North-South-East-West for 2D signals), but our results apply to higher-order neighborhoods such as the locally quadratic twelve-neighbor prior of Besag and Kooperberg (1995). The smoothing parameter λ in (3), also called the hyper- or regularization parameter, plays a key role as it controls the amount of smoothing: choosing it too large, the estimate is too smooth; choosing it too small, the estimate is too wiggly. Within the class of MRF priors (3), various functions ϕ have been considered, including $\phi_1(u) = |u|^\eta$, $\phi_2(u) = c_1 \log\cosh(c_2 u)$ of Green (1990), and the Huber (1981) function

$$\phi_3(u) = \begin{cases} u^2/2, & |u| \le \tau, \\ \tau \cdot |u| - \tau^2/2, & |u| > \tau, \end{cases}$$

of Kuensch (1994). Though quite different in nature, each of these functions includes the quadratic (or ℓ_2) and the absolute value (or ℓ_1) functions as limiting cases corresponding to Gaussian and Laplace priors. The Gaussian prior has soon been recognized to blur the scene in regions of large intensity gradient. Geman and McClure (1987) proposed $\phi(u) = -1/(1+u^2)$ which is less sensitive to large intensity gradient but is nonconvex. The function $\phi_1(u) = |u|$ advocated by Besag (1989) is convex but not differentiable. The absolute value penalty is also related to the taut string estimate (Davies and Kovac, 2001) for denoising 1D signals and to the local adaptive spline estimate of Mammen and van de Geer (1997). In the signal processing literature, Rudin, Osher, and Fatemi (1992) and Li and Santosa (1996) minimize the total variation functional (which relates to the absolute value penalty) with least squares constraints. In the Kalman filter and general state space models literature (Kitagawa and Gersch, 1996), the Pearson family of densities is commonly used for the process innovations.

The diversity of proposed priors reflects their important role, as they induce various properties of the smoothing. A striking example is given by the following two wavelet estimators. The soft-Waveshrink estimate $\hat{\mathbf{x}}_{\rm MAP} = \Phi\hat{\alpha}_{\rm MAP}$ (Donoho and Johnstone, 1994), where $\hat{\alpha}_{\rm MAP}$ are the estimated wavelet coefficients, minimizes the negative Gaussian log-likelihood penalized by the ℓ_1 norm of the mother wavelet coefficients γ

$$\hat{\alpha}_{\rm MAP} = \arg \min_{\alpha=(\beta,\gamma)} \frac{1}{2}||\mathbf{s} - \Phi\alpha||_2^2 + \lambda||\gamma||_1, \tag{4}$$

where $\mathbf{s}$ is the observed data, $||\gamma||_1$ stems from the Laplace(λ) prior of γ, and the matrix $\Phi = [\Phi_0 \ \Psi]$ built with father wavelets Φ_0 and with mother wavelets Ψ is orthonormal. The soft-Waveshrink estimate (4) has a closed form expression $\hat{\beta} = \Phi_0'\mathbf{s}$ and $\hat{\gamma} = \eta_\lambda^{\rm soft}(\Psi'\mathbf{s})$, where

$$\eta_\lambda^{\rm soft}(x) = {\rm sign}(x)(|x| - \lambda)_+ \tag{5}$$

is the soft thresholding function. Since any raw mother wavelet coefficient $(\Psi'\mathbf{s})_n = x$ between $-\lambda$ and λ is set to zero, then the MAP estimate has the desired 'sparse wavelet representation' property. If, instead of maximizing the posterior distribution (4), the posterior mean or median were calculated, then the estimated wavelet coefficients would all be nonzero with probability one, violating the sparsity property. Abramovich and Silverman (1998) induced sparsity with the posterior median $\hat{\alpha}_{\rm PMedian}$ by instead choosing a prior involving a Dirac mass at zero. The two wavelet estimators $\hat{\alpha}_{\rm MAP}$ (using the Laplace prior) and $\hat{\alpha}_{\rm PMedian}$ (using the Dirac mass at zero) illustrate that the method of estimation and the prior are tightly coupled.

In this article, we study MAP estimation (2) for MRF priors with continuous intensities, with focus on the nondifferentiable MRF Laplace prior with $\phi(u) = |u|$ in (3). In Sections 2, 3 and 4, we propose and prove convergence of methods for computing the MAP estimate for various (non)differentiable MRF priors and likelihood functions, given any smoothing parameter λ. Section 2 discusses issues related to the ICM algorithm, and shows some of its limitations. The new algorithms of Section 3 and 4 extend the scope of ICM, for instance allowing the use of the nondifferentiable Laplace prior, pointing to a connection to spline and wavelet smoothers. In Section 5, we propose three rules for the automatic selection of the smoothing parameter: by empirical Bayes, by two-fold cross validation, and through a universal rule derived for the Laplace prior based on an asymptotic analysis. In Section 6, we use Gaussian and Poisson simulations to evaluate the finite sample efficiency of the MAP MRF-ℓ_1 smoother, and compare it to the MAP MRF-ℓ_2 smoother and to wavelet smoothers for various selection rules of the smoothing parameter. In Section 7, we apply the MAP MRF-ℓ_1 smoother to real data to estimate the spectrum of a chaotic time series in mechanics, and to estimate an erratic astronomy image with Poisson noise. We draw some conclusions in Section 8.

2 ICM revisited

The Newton–Raphson algorithm cannot always be employed to find the MAP estimate defined in (2) because double differentiability of the posterior density function is required. For instance, the Laplace density violates the differentiability condition. An alternative efficient and easy-to-implement algorithm is the celebrated local improvement ICM algorithm of Besag (1986). Although its implementation to maximizing the posterior density (2) with the nondifferentiable Laplace MRF prior looks straightforward from a programming point of view, convergence of ICM is surprisingly not guaranteed, as we will see, even for strictly convex posteriors. ICM may be viewed as a special case of the block coordinate relaxation (BCR) method for solving the minimization problem

$$\min_{\mathbf{x}} \ f(\mathbf{x}).$$

where $f(\mathbf{x}) = -l(\mathbf{x} \mid \mathbf{s}) - \log p(\mathbf{x})$. The BCR method iteratively minimizes $f(\mathbf{x})$ with respect to a block of coordinates of $\mathbf{x}$, with the remaining coordinates held fixed at their current value. The method is described formally below.

BCR method: for a given initial $\mathbf{x}^0$ in the effective domain Ω_f of

$f : \mathbb{R}^{\mathrm{K}} \mapsto \mathbb{R} \cup \{\infty\}$, generate a sequence $\mathbf{x}^\nu \in \Omega_f$, $\nu = 0, 1, ...$

$$\mathbf{x}^{\nu+1} = \arg \min_{x_i = x_i^\nu, i \notin I^\nu} f(\mathbf{x}),$$

where I^ν is a (user-chosen) nonempty subset of $\{1, ..., K\}$.

The idea is to successively solve simple subproblems of dimension $|I^\nu|$. There are various rules for choosing I^ν to ensure convergence. For the popular essentially cyclic rule, there exists a constant $T \geq 1$ such that $I^\nu \cup \cdots \cup I^{\nu+T} = \{1, ..., K\}$ for all ν. The raster scan of Besag (1986) corresponds to $|I^\nu| = 1$ with a varying raster to avoid a directional effect.

Example: Consider the following minimization problem

$$\min_{x_1, x_2} x_1^2 + x_2^2 + |x_1 - x_2|, \tag{6}$$

which corresponds to (2) with Gaussian likelihood, Laplace prior, $N = 2$ observations $s = (0, 0)$, and $\lambda = 1/2$. When applied to this problem, ICM corresponds to the BCR method with $I^\nu = \{1\}$ for ν even and $I^\nu = \{2\}$ for ν odd: ICM alternatively keeps x_2 fixed and solves (6) with respect to x_1, and then keeps x_1 fixed and solves (6) with respect to x_2, until convergence. Using the soft thresholding function (5), these solutions can be expressed in closed form:

$$\begin{aligned} x_1^{\nu+1} &= \begin{cases} x_2^\nu - \eta_{1/2}^{\mathrm{soft}}(x_2^\nu) & \text{if } \nu \text{ is even;} \\ x_1^\nu & \text{if } \nu \text{ is odd,} \end{cases} \\ x_2^{\nu+1} &= \begin{cases} x_2^\nu & \text{if } \nu \text{ is even;} \\ x_1^\nu - \eta_{1/2}^{\mathrm{soft}}(x_1^\nu) & \text{if } \nu \text{ is odd.} \end{cases} \end{aligned}$$

The simplicity of the iterations is a key feature of ICM. However, ICM generally fails to converge to the maximum a posteriori estimate $(0, 0)$. For example, if ICM starts at $\mathbf{x}^0 = (1/2, 1/2)$, it would remain trapped at that point. In general, ICM converges only to a critical point of (6), which is any point in the set $\{-\lambda \leq x_1 = x_2 \leq \lambda\}$. Only $(0, 0)$ is a stationary point and the true solution.

The above bivariate example illustrates a pitfall of the ICM/BCR method when f is nondifferentiable, as it can become trapped at critical points far from a stationary point. This phenomenon has a particularly adverse effect on high-dimensional applications, as was observed by Kuensch (1994). Further illustration of this phenomenon on a realistic denoising problem is presented in Section 6.1. Owing to such phenomenon, it is often perceived that ICM is unsuitable for finding the MAP estimate when the posterior is nondifferentiable. On the other hand, it was shown by Tseng (1993) and Tseng (2001) that convergence of the BCR method to a stationary point can be guaranteed, provided that the nondifferentiable part has a separable structure (see definition below) in addition to some mild coordinatewise convexity assumptions. This suggests that we should transform the problem to have a separable nondifferentiable part. In the next two sections, we describe two such transformations applicable to a general MRF MAP estimation problem (2), with varying neighborhoods, likelihoods and priors, that has the form

$$\min_{\mathbf{x}} \; g(\mathbf{x}) + h(B\mathbf{x}), \tag{7}$$

where

- the matrix $B \in \mathbb{R}^{\mathrm{M}\times\mathrm{N}}$ reflects the neighborhood structure of the locally dependent MRF that can be of first or any higher order, with the total number of local dependencies being the number M of rows in B;
- the negative log-likelihood function $g : \mathbb{R}^{\mathrm{N}} \mapsto \mathbb{R} \cup \{\infty\}$ is assumed to be proper (i.e., finite-valued somewhere);
- the prior function h is continuous separable, i.e., $h(v_1, ..., v_M) = h_1(v_1) + \cdots + h_M(v_M)$, where each $h_k : \mathbb{R} \mapsto \mathbb{R}$ is continuous.

We refer the reader to Appendix A for the precise definitions of mathematical terms used for characterizing, in terms of the prior and likelihood functions chosen, MAP estimation problems encountered in statistical applications.

The MAP estimation problem (7) is quite general and applies to a wide range of estimators. For instance, if B is the identity matrix, then (7) defines the ridge regression estimator for $h_k(x_k) = \lambda x_k^2$ (Hoerl and Kennard, 1970) or the lasso estimator for $h_k(x_k) = \lambda|x_k|$ (Tibshirani, 1995). Thus, the scope of the methods we are developing goes beyond MRF smoothers and applies to parametric regression as well. The following theorem gives general conditions for the ICM/BCR method to converge to a stationary point of the MAP problem when the prior is differentiable.

Theorem 1 (ICM/BCR convergence for differentiable posterior): Suppose that in (7), g is domainwise differentiable, h is differentiable, and $f = g + h \circ B$ is level-bounded and either has coordinatewise unique minimum or is 2-coordinatewise convex. Then the sequence of $\{\mathbf{x}^\nu\}_{\nu=0,1,\ldots}$ generated by applying ICM and, more generally, the BCR method (using the essentially cyclic rule and with $|I^\nu| = 1$ for all ν) to minimize f is bounded and every cluster point is a critical point and hence a stationary point of f.
Proof: See Appendix B.
Note: The assumption in Theorem 1 that f either has coordinatewise unique minimum or is 2-coordinatewise convex is needed. Otherwise, the ICM/BCR method can become trapped far from a stationary point (Powell, 1973).

Theorem 1 applies to any admissible neighborhood structure. The differentiability property of h is satisfied by the Gaussian MRF prior, and by the priors proposed by Kuensch (1994) and Green (1990). It is also satisfied by priors of the form $\phi(u) = |u|^\eta$, $\eta > 1$, but not for $\eta = 1$, although the latter choice, which corresponds to the Laplace prior, endows the corresponding estimator with nice statistical properties (if the exact MAP solution is found). In the next two sections, we describe algorithms for handling priors that are dirty in the sense of being nondifferentiable. Both strategies transform the original problem into one with a nondifferentiable separable structure, either with a change of variables or by means of duality theory for convex programming. Under weaker assumptions than Theorem 1, convergence to the the MAP estimate is proved.

3 Back to ICM/BCR with splines

Suppose that B in (7) has full row rank. By permuting its columns if necessary, we can assume that $B = [B_1 \;\; B_2]$, where B_1 is invertible. Setting $\mathbf{u_1} = B\mathbf{x} = B_1\mathbf{x_1} + B_2\mathbf{x_2}$ gives $\mathbf{x_1} = B_1^{-1}\mathbf{u_1} - B_1^{-1}B_2\mathbf{x_2}$. Then, letting $\mathbf{u} = \left[\begin{smallmatrix}\mathbf{u_1}\\ \mathbf{x_2}\end{smallmatrix}\right]$ and

$$\Phi = \begin{bmatrix} B_1^{-1} & -B_1^{-1}B_2 \\ 0 & I \end{bmatrix}, \tag{8}$$

we can transform the MAP estimation problem (7) into

$$\min_{\mathbf{u}=\left[\begin{smallmatrix}\mathbf{u_1}\\ \mathbf{x_2}\end{smallmatrix}\right]} g(\Phi\mathbf{u}) + h(\mathbf{u_1}). \tag{9}$$

In the transformed problem, the nondifferentiable part $h(\mathbf{u})$ is separable in $\mathbf{u}$, so the BCR method finds the MAP estimate (see Theorem 2 below).

We illustrate this transformation with the full row rank matrix corresponding to univariate first-order pairwise neighborhoods:

$$B = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \cdots & \cdots & \\ & & & 1 & -1 \end{pmatrix} \quad \text{yields} \quad \Phi = \begin{pmatrix} 1 & \cdots & & 1 \\ & 1 & \cdots & 1 \\ & & & \vdots \\ & & & 1 \end{pmatrix} \tag{10}$$

and vice versa. The transformed problem (9) with the matrix Φ in (10) points to an equivalence between the MRF smoother and an associated expansion-based smoother $x(t) = \sum_{n=1}^{N} u_n \phi_n(t)$, where $\phi_n(t) = 1\ (t \geq t_n)$ are the Heaviside basis functions (zero order knot splines) with knots t_n at the sampling locations, and u_n are the corresponding spline coefficients. The coefficient u_N associated with the last column (the intercept) is not penalized, as it is usually done in regression, since the sum in (9), i.e., $h(\mathbf{u_1}) = \sum_{k=1}^{M} h_k(u_k)$, goes up to $M = N-1$. Note that the columns of Φ are not normalized however; doing so would result in an awkwardly *weighted* MRF. Using the Laplace prior, the MAP MRF-ℓ_1 smoother can be seen as an expansion-based estimator on Heaviside basis functions, but, as opposed to the best-subset knot selection technique often adopted with splines, selection is achieved here by an ℓ_1 penalty on the spline coefficients, with the solution $\mathbf{u}$ having more zero entries as the smoothing parameter λ increases. Similarly, Mammen and van de Geer (1997) pointed to the connection between total variation regularization and knot splines estimators. This 'sparse Heaviside representation' obtained with the MAP estimate would not be obtainable from the posterior mean/median estimate, not with an ℓ_2 prior. The sparsity induced by penalties of the form $p_\lambda(|\theta|) = \lambda|\theta|^p$ has been studied for wavelet smoothers by Antoniadis and Fan (2001), who showed that the Laplace ($p = 1$) prior is the only prior that yields a continuous and thresholding solution.

The transformation (10) further points to a connection between the proposed MAP MRF-ℓ_1 smoother and soft-Waveshrink (4) which can be written equivalently as the solution to

$$\min_{\mathbf{x}} \frac{1}{2}||\mathbf{s} - \mathbf{x}||_2^2 + \lambda||\Psi'\mathbf{x}||_1 \tag{11}$$

using the orthogonality of the wavelet matrix $\Phi = [\Phi_0\ \Psi]$. The Haar wavelet, for instance, involves finite differences at various scales and is therefore directly comparable to first-order MRF. The mother wavelets Ψ' can be seen as defining a multilevel MRF. Using smoother wavelet filters than Haar would define higher order Markov stochastic processes with erratic, but smooth, behaviors.

In contrast to the original MAP estimation problem (7), the transformed problem (9) has the advantage that it can be solved by ICM and the BCR method even for the nondifferentiable Laplace prior. The following theorem states the general conditions which guarantee convergence to the original MAP estimate (7) by means of this transformation.

Theorem 2: Suppose g is domainwise differentiable, B has full row rank (so $M \leq N$), h is continuous separable, and $f = g \circ \Phi + h$ is level-bounded and either has coordinatewise unique minimum or is 2-coordinatewise convex, where Φ is given by (8). Then the sequence $\{\mathbf{u}^\nu\}_{\nu=0,1,\ldots}$ generated by applying the BCR method (using the essentially cyclic rule and with $|I^\nu| = 1$ for all ν) to minimize f is bounded and every cluster point $\hat{\mathbf{u}}$ is a critical point and a stationary point of f. Moreover, $\hat{\mathbf{x}} = \Phi\hat{\mathbf{u}}$ is a stationary point of the original MAP problem (7).
Proof: See Appendix B.

Example of Section 2 (cont.): The transformed problem is

$$\min_{\mathbf{u}=\left[\begin{smallmatrix}\mathbf{u_1}\\ \mathbf{x_2}\end{smallmatrix}\right]} (u_1 - u_2)^2 + u_2^2 + |u_1|.$$

Applying the BCR method with $I^\nu = \{1\}$ for ν even yields

$$\begin{aligned} u_1^{\nu+1} &= \begin{cases} \eta_{1/2}^{\text{soft}}(u_2^\nu) & \text{if } \nu \text{ is even;} \\ u_1^\nu & \text{if } \nu \text{ is odd,} \end{cases} \\ u_2^{\nu+1} &= \begin{cases} u_2^\nu & \text{if } \nu \text{ is even;} \\ u_1^\nu/2 & \text{if } \nu \text{ is odd.} \end{cases} \end{aligned}$$

This example satisfies the assumptions of Theorem 2, so that $\mathbf{u}^\nu$ converges to the unique stationary point $\hat{\mathbf{u}} = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$. Moreover, $\Phi\hat{\mathbf{u}} = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$ is a stationary point of the original MAP problem.

Negative log-likelihood functions g like the Gaussian or the exponential satisfy the assumptions of Theorem 2. This is also true for the Poisson likelihood as long as none of the Poisson measurements is zero; if a measurement is zero, we can perturb it to a small positive value (say $\epsilon = 0.01$) to make the negative log-likelihood function domainwise differentiable. The generalized Pareto distribution $\text{GBP}(\sigma, \kappa)$ used in extreme value theory also has a convex domainwise differentiable negative log-likelihood function in $x = \log \sigma$ for a fixed positive shape parameter κ. Letting $\alpha = 1 + 1/\kappa$ and $\beta_n = s_n \kappa$, $g(\mathbf{x}) = \sum_n x_n + \alpha \log(1 + \beta_n \exp(-x_n))$ satisfies the condition of Theorem 2.

The computational efficiency of the BCR method depends on how easy it is to solve (9) in each coordinate u_n, for $n = 1, \dots, N$. The univariate subproblems in u_n may not have a closed form solution; an example is the Poisson likelihood. In general, due to the upper block diagonal structure of Φ, the BCR method is more efficiently implemented in a programming languages like `Fortran` or `C`, rather than high-level ones like `Splus`, `R` or `Matlab`.

4 IDM: a dual maximum a posteriori solution

The variable substitution technique of Section 3 cannot be applied for higher-order neighborhoods and for 2D signals, because the corresponding matrix B is then rank-deficient. However, if the negative log-likelihood function g and the prior function h are convex, we can use duality theory for convex programming to reformulate the original MAP problem (7) as a dual MAP problem as follows. We first write (7) as $\min_{-B\mathbf{x}+\mathbf{v}=0} \; g(\mathbf{x}) + h(\mathbf{v})$. By associating Lagrange multipliers $\mathbf{w}$ with the equality constraints and denoting the *conjugate* likelihood and prior functions (Rockafellar, 1970, p. 104 and Chap. 31)

$$g^*(\mathbf{u}) = \sup_{\mathbf{x}} \mathbf{u}'\mathbf{x} - g(\mathbf{x}) \quad \text{and} \quad h^*(\mathbf{w}) = \sup_{\mathbf{v}} \mathbf{w}'\mathbf{v} - h(\mathbf{v}), \tag{12}$$

we obtain the dual MAP problem

$$\min_{\mathbf{w}} g^*(-B'\mathbf{w}) + h^*(\mathbf{w}). \tag{13}$$

Since h is separable, then h^* is separable with $h^*(\mathbf{w}) = \sum_{k=1}^{M} h_k^*(w_k)$, where h_k^* is the conjugate of h_k. Moreover, each h_k^* is proper, closed convex (Rockafellar, 1970, Thm. 12.2). Thus, in the dual problem, the nondifferentiable part $h^*(\mathbf{w})$ is separable in $\mathbf{w}$, so the BCR method can be effectively applied to find the MAP estimate (see Theorem 3 below). In typical applications, g^* and h^* have an explicit expression for a given likelihood function g and prior distribution h (see the dual Laplace and Gaussian priors h^*, and the dual Gaussian, Poisson and generalized Pareto likelihoods g^* derived below).

The derivation of a dual posterior distribution leads to the new Iterative Dual/Dirty Mode (IDM) method, which, by applying results from Tseng (1993) and Tseng (2001), will find the MAP estimate (7) under suitable assumptions.

Theorem 3 (IDM convergence): Suppose g is a proper closed convex function and h is a convex separable function. Then the sequence $\{\mathbf{w}^\nu\}_{\nu=0,1,\dots}$ generated by the BCR method (using essentially cyclic rule and with $|I^\nu| = 1$ for all ν) to minimize $f = g^* \circ (-B') + h^*$ has the following convergence properties. If g is strictly convex and g^* is finite-valued everywhere, then $\mathbf{x}^\nu = \nabla g^*(-B'\mathbf{w}^\nu)$ is bounded and every cluster point solves (7). If instead g^* is domainwise differentiable and f is level-bounded, then $\{\mathbf{w}^\nu\}_{\nu=0,1,\dots}$ is bounded and every cluster point solves (13), and every cluster point of $\{x^\nu\}_{\nu=0,1,\dots}$ solves (7).
Proof: See Appendix B.

> Example of Section 2 (cont.): The dual problem is $\min_{|w|\le\frac{1}{2}} w^2$. IDM finds the dual solution $w = 0$ in 1 iteration. The solution to the original MAP problem is $-B'w = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$.

We show below that the dual functions have simple forms, and we check whether the assumptions of Theorem 3 are satisfied.

- For the Laplace and the Gaussian priors, we have $h(\mathbf{v}) = \lambda\|\mathbf{v}\|_1$ and $h(\mathbf{v}) = \lambda\|\mathbf{v}\|_2^2$ respectively. Direct calculation yields

$$h^*(\mathbf{w}) = \begin{cases} 0 & \text{if } \|\mathbf{w}\|_\infty \le \lambda \\ \infty & \text{else} \end{cases} \quad \text{and} \quad h^*(\mathbf{w}) = \|\mathbf{w}\|_2^2/(4\lambda), \tag{14}$$

  which are nonnegative-valued and level-bounded.

- For the Gaussian likelihood, we have $g(\mathbf{x}) = \|\mathbf{x} - \mathbf{s}\|_2^2/2$ and

$$g^*(\mathbf{u}) = \mathbf{s}'\mathbf{u} + \|\mathbf{u}\|_2^2/2. \tag{15}$$

  Thus g^* is differentiable and level-bounded. Since h^* for the Laplace and Gaussian priors are nonnegative-valued and level-bounded, this implies $g^* \circ (-B') + h^*$ is level-bounded.

- For the Poisson likelihood, we have $g(\mathbf{x}) = \sum_n (x_n - s_n \log x_n)$ and

$$g^*(\mathbf{u}) = \sum_n (s_n \log s_n - s_n - s_n \log(1 - u_n)).$$

  Thus g^* is domainwise differentiable. Here we assume that $s_n > 0$ for all n, which can be ensured by perturbing zero Poisson observations if necessary. For the Laplace or Gaussian prior, $g^* \circ B' + h^*$ is level-bounded (since logarithm tends to infinity at a sublinear rate).

- For the generalized Pareto distribution $\text{GBP}(\sigma, \kappa)$ discussed in Section 3,

$$g^*(\mathbf{u}) = \begin{cases} \sum_n (u_n - 1)\log\{\beta_n(\alpha + u_n - 1)/(1-u_n)\} & \\ \quad +\alpha \log\{(\alpha + u_n - 1)/\alpha\} & u_n \in [1-\alpha, 1], \\ +\infty & \text{otherwise} \end{cases}$$

  is not finite-valued everywhere. It is also not domainwise differentiable since $\lim_{\mathbf{u}\to(1-\alpha)\mathbf{1}} g^*(\mathbf{u}) = -\alpha \sum_n \log \beta_n$. Thus Theorem 3 cannot be used to infer convergence of IDM here (convergence can still be shown by observing that g^* satisfies assumption (A2) of Tseng (2001)).

IDM has computational advantages similar to ICM while avoiding the pitfalls of ICM. Thanks to the sparse MRF structure of B' in (13), the dual posterior is separable between odd and even indexes for first-order MRF, so that all the odd-index univariate subproblems can be solved and updated together, followed by the even-index ones, until convergence. This property allows IDM to be programmed efficiently in a high-level language like `Splus`, `R` or `Matlab` by vectorizing the univariate subproblems between odd and even indexes, and hence avoiding an expensive `for` loop. Moreover, the separability can provide a closed form solution to the dual univariate subproblems, for the Poisson likelihood for instance. This is an additional advantage of IDM over the variable transformation approach described in the previous section.

5 Selection of the smoothing parameter

Some estimators would be outstanding given the best possible λ, but in practice give a poor estimation due to difficulties in selecting λ. The pseudo-likelihood approach of Besag (1986), though pioneering and well studied, should now be regarded as outdated (Besag, 2001). Cross validation (Stone, 1974) is feasible only if the estimate $\hat{\mathbf{x}}_\lambda$ can be calculated efficiently for several smoothing parameters λ until a nearly optimal one is found. Generalized cross validation requires less computing effort, but requires the estimator to be linear.

In this section we investigate three ways of selecting the smoothing parameter for MRF smoothers: by empirical Bayes, by two-fold cross validation, and by a universal rule derived specifically for the Laplace prior and the sparsity property it infers on the estimate.

5.1 Empirical Bayes approach (EB)

The improper MRF prior $p(\mathbf{x};\lambda) \propto \exp\left\{-\lambda \sum_{n=1}^{N-1} \phi(x_{n+1}-x_n)\right\}$ implies that the increments $\mathbf{u} = B\mathbf{x}$, such that $u_n = x_{n+1} - x_n$, of the MRF process are independent with distribution $p(\mathbf{u};\lambda) \propto \exp\left\{-\lambda \sum_{n=1}^{N-1} \phi(u_n)\right\}$. So define $\mathbf{y} = B\mathbf{s}$, then for certain noise likelihoods its marginal distribution with respect to the prior on $\mathbf{u}$ factorizes into a simple expression. For the Gaussian white noise with variance σ^2 for instance, the marginal distribution $f(\mathbf{y};\lambda,\sigma)$ can be derived from the convolution formula; plugging in an estimate $\hat{\sigma}$ of σ, we can use the marginal distribution to find an empirical Bayes estimate of λ:

$$\lambda_{\mathrm{EB}} = \arg\max_{\lambda>0} f(\lambda;\mathbf{y},\hat{\sigma}), \tag{16}$$

where $\hat{\sigma}$ is, for instance, the median absolute deviation (MAD) of the first order differences rescaled by $1/0.6745$ for Gaussian noise (Donoho and Johnstone, 1995). This approach of maximizing the marginal likelihood can be traced back to the type II maximum likelihood method of Good (1965), and bears the name of empirical Bayes, as is recently used by Johnstone and Silverman (2002) for wavelet smoothing. It is also the standard method for selecting the hyper-parameter of general state space models (Kitagawa and Gersch, 1996).

We now specifically derive the empirical Bayes selection of λ for Gaussian white noise, assuming either a Gaussian or a Laplace MRF prior. For the former, the marginal distribution of $\mathbf{y}$ is the Gaussian distribution $\mathrm{N}(\mathbf{0}, \hat{\sigma}^2 BB' + \frac{1}{2\lambda} I_{N-1})$, and (16) can therefore easily be found. For the latter, however, the multivariate integral expression of the distribution $f(\mathbf{y};\lambda,\sigma)$ cannot be easily calculated. At the cost of losing some efficiency on the selection of λ, we propose an approximate empirical Bayes selection by considering the marginal distributions $g_y(y_j;\lambda,\sigma)$ of each components of $\mathbf{y}$ and select

$$\lambda_{\mathrm{EB}} = \arg\max_{\lambda>0} \prod_{j=1}^{N-1} g_y(\lambda; y_j, \hat{\sigma}).$$

For the Laplace prior, the marginal distribution can be found as

$$\begin{aligned} g_y(y;\lambda,\sigma) &= \frac{\lambda}{2}\Big[\exp\{\lambda(y+\lambda\sigma^2)\}\Phi\{-(y+2\lambda\sigma^2)/(\sqrt{2}\sigma)\} \\ &\quad + \exp\{-\lambda(y-\lambda\sigma^2)\}\{1-\Phi(-(y-2\lambda\sigma^2)/(\sqrt{2}\sigma))\}\Big], \end{aligned} \tag{17}$$

where Φ is the Gaussian c.d.f., and $\sqrt{2}$ stems from the diagonal of BB'.

5.2 Two-fold cross validation (CV_2)

The goal of the computer intensive two-fold cross validation method is to estimate for a given smoothing parameter λ the Kullback-Leibler distance between the likelihood functions that generated the observations and the estimated one. By repeating the operation for several λ's, cross validation selects the one for which the estimated Kullback-Leibler distance is minimum. This can be achieved by folding the data set into two disjoint sets, using one to fit the model and the other to measure its predictive performance, interchanging the two sets, and finally calculating the overall goodness-of-fit criterion. The data set is folded deterministically when the sampling locations are fixed (e.g., odd/even indexes for equispaced design), and randomly otherwise. The advantage of cross validation is that it can essentially be employed for any likelihood.

5.3 Universal rule (U)

Finally, we investigate a selection of λ based on the universal rule for the Laplace MRF prior. This idea of a universal rule, borrowed from thresholding wavelet estimators (Donoho and Johnstone, 1994), entails finding the smallest penalty parameter λ_U such that, under the assumption that the underlying signal $\mathbf{x_0}$ belongs to the kernel of B (e.g., $\mathbf{x_0} = c\mathbf{1}$ for first-order pairwise interactions), the estimate $\hat{\mathbf{x}}_{\lambda_U}$ reproduces the underlying signal $\mathbf{x_0}$ with a probability tending to one as the sample size N goes to infinity. Using the Laplace MRF prior, the estimate is indeed more and more sparse in the Heaviside basis as λ increases, and becomes the constant function for a prescribed finite λ.

Proposition 1: Suppose that the negative log-likelihood g in (7) is closed and strictly convex. Also, suppose that $\mathbf{0} \in \text{int}\Omega_{g^*}$, where g^* is the conjugate of g (12) and $\text{int}\Omega_{g^*}$ is the interior of $\Omega_{g^*} = \{\mathbf{u} : g^*(\mathbf{u}) < \infty\}$. Using the Laplace prior (i.e., $h_k(u) = \lambda|u|$ in (7)), the universal parameter λ_U is defined as the smallest $\lambda(N)$ such that the order of convergence of

$$\mathrm{P}\{\min_{\mathbf{w}_\infty \in \Gamma} \|\mathbf{w}_\infty\|_\infty \leq \lambda(N)\} \stackrel{N\to\infty}{\longrightarrow} 1 \tag{18}$$

is optimal, where $\Gamma = \{\mathbf{w}_\infty : \mathbf{w}_\infty = \arg\min_{\mathbf{w}} g^*(-B'\mathbf{w})\}$ is assumed nonempty.

Solving this asymptotic problem for 1D signals with first-order neighborhood and standard Gaussian white noise, one finds the universal smoothing parameter $\lambda_U = \sqrt{N \log(\log N)/2}$.

Proof: See Appendix C.

6 Simulation

We now investigate through Monte Carlo simulations the finite-sample performance of MRF estimators and wavelet estimators on Gaussian and Poisson data. The goal is not to compare between MRF or wavelets of different order, but rather to fix the order to first-order MRF and to the Haar wavelet (both assume a piecewise constant function), and compare between prior distributions and selection rules for the smoothing parameter. Higher-order MRF and higher-order wavelets assume a smoother underlying scene and may be more appropriate depending on the application. For the selection of λ with MRF, we consider the three rules proposed in Section 5: empirical Bayes (EB), two-fold cross validation (CV_2), and universal rule (U) for the Laplace prior. For wavelet smoothers: the Stein Unbiased Risk Estimate (SURE) rule (Donoho and Johnstone, 1995) and the universal rule (Donoho and Johnstone, 1994).

6.1 Visual illustration

Figure 1 illustrates a few issues discussed in the previous sections on a simulated smoothing example: the piecewise constant `blocks` signal (b) is sampled with Gaussian noise at $N = 256$ equispaced locations (a). Some insights can be gained from the following comparisons of plots:

- (d)-(e): As is expected for denoising a discontinuous function like `blocks`, the superiority of the nonlinear MAP MRF-ℓ_1 estimator (e) over the linear MAP MRF-ℓ_2 estimator (d) illustrates the importance of the prior distribution (the smoothing parameters are selected by CV_2);

- (c)-(e): ICM gets trapped at the critical point plotted in (c) and gives the false impression that the Laplace prior does not provide good results. Finding the exact solution (e) to (2) using IDM shows that the Laplace prior is worth considering.

- (e)-(f): The MAP MRF-ℓ_1 estimate (e) seems better than the soft-Waveshrink estimate (f) on this particular example. The MRF estimator looks promising as quantified by the Monte Carlo simulation below.

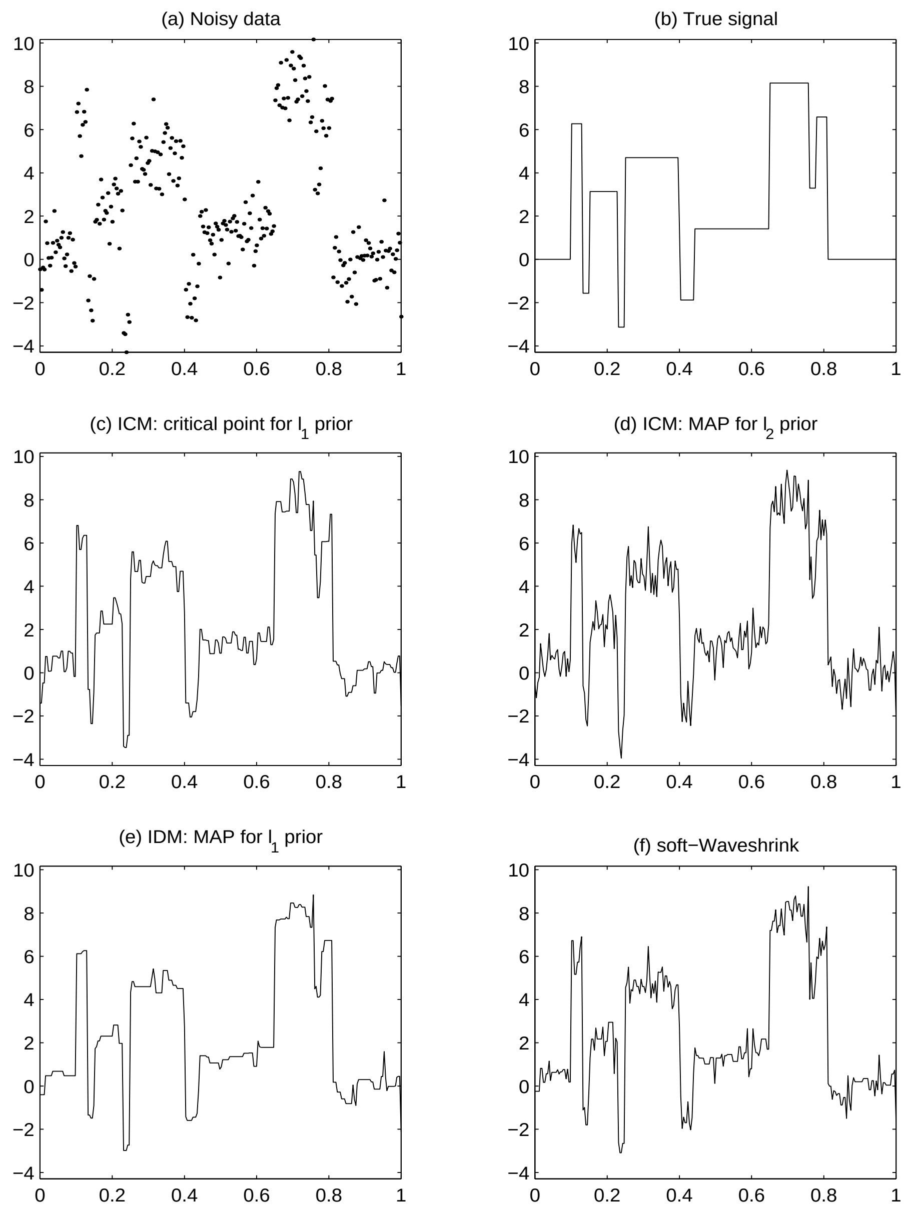

Figure 1: (a) signal sampled with Gaussian noise; (b) underlying `block` signal; (c) ICM finds a critical point instead of the MAP MRF-ℓ_1 estimate; (d) ICM finds the MAP MRF-ℓ_2 estimate; (e) IDM finds the exact MAP MRF-ℓ_1 estimate; (f) wavelet estimate.

6.2 Gaussian simulation

We use four signals often used in the wavelet literature (Donoho and Johnstone, 1994): `blocks`, `bumps`, `heavisine` and `Doppler` rescaled to have a 'standard deviation', defined by $\left\{\int_0^1 (x(t)-\bar{x})^2 dt\right\}^{1/2}$, equal to seven. The Gaussian noise has variance one so that the signal-to-noise ratio is uniquely determined. We repeat the Monte Carlo experiment $m = 25600/N$ times to have the same number of total observations for each sample size $N \in \{256, 1024, 4096\}$. In light

Table 1: Gaussian simulation. Average of observed MSEs ($\times 100$). The best linewise is marked in **bold** (standard error of the order of the precision reported).

	soft-Waveshrink		MAP MRF-ℓ_1			MAP MRF-ℓ_2	
	SURE	U	EB	CV$_2$	U	EB	CV$_2$
blocks							
$N = 256$	44	160	**32**	**30**	370	82	91
$N = 1024$	22	74	15	**7.3**	110	59	64
$N = 4096$	9.8	30	7.8	**2.4**	30	34	36
bumps							
$N = 256$	67	410	**63**	80	2200	95	98
$N = 1024$	51	250	**38**	**41**	1460	76	99
$N = 4096$	32	120	**18**	**19**	590	35	53
heavisine							
$N = 256$	56	130	35	33	120	**26**	**26**
$N = 1024$	30	72	**12**	**12**	47	**12**	**12**
$N = 4096$	15	35	**5.0**	**5.1**	23	5.6	6.1
Doppler							
$N = 256$	83	500	**67**	**66**	980	77	90
$N = 1024$	56	210	48	**41**	440	**38**	61
$N = 4096$	32	94	**16**	**16**	220	**15**	19

of the mean squared error results of Table 1, we draw the following conclusions:

- Comparison between selection rules for the smoothing parameter.
  Comparing the three columns for MAP MRF-ℓ_1, CV$_2$ works best overall. The EB selection is comparable to CV$_2$, but, strikingly, works poorly for `blocks` as the sample size increases. We impute this to the fact that the

Laplace MRF prior does not put enough mass around zero to model no change of function values in a region where `blocks` is constant. Since EB selection relies more on an appropriate prior than CV_2, a more flexible prior is needed for better EB selection of the smoothing parameter(s).

Comparing the two columns for MAP MRF-ℓ_2, the EB selection also works well for the Gaussian prior, and is often better than CV_2, especially for smooth signals for which the Gaussian MRF prior is well suited.

The oversmoothing behavior of the universal threshold often reported with wavelets seems even more pronounced with the MAP MRF-ℓ_1 smoother.

- Comparison between MRF-ℓ_1, MRF-ℓ_2 and wavelet smoothers.
  As is expected, the comparison between the MAP MRF-ℓ_1 and -ℓ_2 estimators is in favor of ℓ_1 for denoising spatially heterogeneous functions (`blocks` and `bumps`) and slightly in favor of the latter for smoother signals (`heavisine` and `Doppler`). We observe that, for a given signal, the most appropriate prior seems to depend on the sample size. This suggests that the prior used should depend not only on the underlying signal, but also on its sampling rate N and sampling design (equispaced or not).

  Because of the close connection (11) between the two MAP estimators, we expect the MAP MRF-ℓ_1 estimator to be competitive with soft-Waveshrink. As for computational efficiency, the advantage goes to soft-Waveshrink whose complexity is $O(N)$ (Mallat, 1989), thanks to a closed form solution to the MAP optimization problem (in the Gaussian case only). As for denoising performance, the MAP MRF-ℓ_1 estimator is superior to soft-Waveshrink based on the mean squared error results of the simulation.

6.3 Poisson simulation

We use the standard Poisson simulation of Timmermann and Nowak (1999). The `blocks` and `bumps` functions are used, but rescaled and shifted to range in $[1/\text{peak}, \text{peak}]$ ($\text{peak} \in \{8, 128\}$) to create abrupt changes and regions of near zero intensities. We repeat the Monte Carlo experiment $m = 25600/N$ times to have the same number of total observations for each sample size $N \in \{512, 1024\}$. Based on the mean squared error results of Table 2, we can compare the performance of the new MAP MRF-ℓ_1 estimator (first column), using CV_2 to select the smoothing parameter, with: the recent ℓ_1-penalized likelihood wavelet estimator (second column) of Sardy, Antoniadis, and Tseng (2004) using Haar and CV_2; soft-Waveshrink applied to approximately Gaussian data (obtained with Anscombe (1948)'s variance stabilizing transformation $2\sqrt{s_n + 3/8}$) using Haar and the minimax rule (third column); the benchmark Bayesian translation invariant Haar-based estimator of Timmermann and Nowak (1999) designed for Poisson measurements (fourth column).

Table 2: Poisson simulation. Average of observed MSEs with the best marked in **bold** (standard error of the order of the precision reported). Estimators from left column to right: MAP MRF-ℓ_1, ℓ_1-penalized likelihood wavelet, soft-Waveshrink applied to gaussianized data by means of Anscombe's transform, the Bayesian estimator of Timmermann and Nowak.

	peak=8				peak=128			
	MRF-ℓ_1	Haar Wavelet			MRF-ℓ_1	Haar Wavelet		
blocks								
$N = 512$	**0.41**	0.81	0.89		**8.2**	19	26	
$N = 1024$	**0.25**	0.62	0.63	0.36	**4.5**	14	18	14
bumps								
$N = 512$	**0.58**	**0.56**	0.84		**8.2**	8.9	37	
$N = 1024$	0.43	0.47	0.61	**0.36**†	**6.6**	8.5	27	10†

† corrected value from Timmermann and Nowak (1999) kindly provided by Anestis Antoniadis, Italia De Feis and Theofanis Sapatinas.

Table 2 shows that the MAP MRF-ℓ_1 estimator is almost always better than its wavelet competitors on that simulation. Deriving an empirical Bayes selection of the smoothing parameter for Poisson data is an open problem which we are studying. Currently, we use two-fold cross validation which would not have been possible without the development of the fast IDM algorithm.

6.4 Conclusions from the simulation

The MAP MRF-ℓ_1 estimator with the EB or CV_2 selection of λ provides a good estimation of smooth and nonsmooth functions. For smoothing `blocks` however, the EB method, which has the advantage of being computationally efficient, is only suboptimal since CV_2 finds a better smoothing parameter. This is due to the fact that the Laplace MRF prior, although more appropriate than

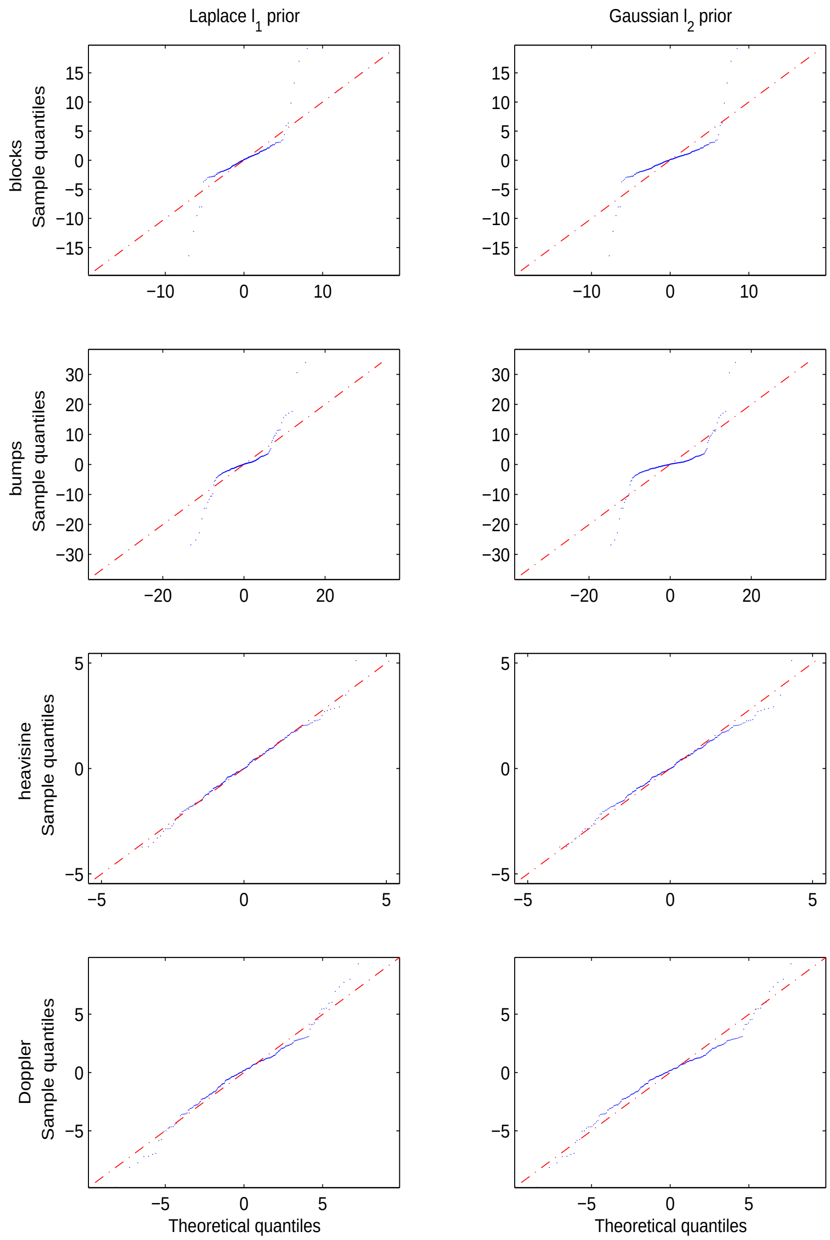

Figure 2: Q-Q plots between the theoretical quantiles of the marginal distributions with the Laplace (left) and Gaussian (right) priors, and the sample quantiles for `blocks`, `bumps`, `heavisine` and `Doppler` for $N = 256$.

the Gaussian MRF prior, is unlikely to have generated the piecewise constant function. To see this on a sample of size $N = 256$, Figure 2 offers diagnostics quantile-quantile plots of $y = B\mathbf{s}$ versus the theoretical quantiles of (17) for the Laplace prior (left) and that of $\mathrm{N}(\mathbf{0}, \hat{\sigma}^2 BB' + \frac{1}{2\lambda} I_{N-1})$ for the Gaussian prior (right), with $\lambda = \lambda_{\mathrm{EB}}$ and $\sigma = \hat{\sigma}$. In light of these plots, the Laplace prior looks more appropriate than the Gaussian prior for the `blocks` and `bumps` functions, and indicates that a power ℓ_η prior with $\eta < 1$ would fit better by putting more mass around zero. A possible avenue to improve the EB approach is therefore to allow for a more flexible prior, as for the recent wavelet smoother of Johnstone and Silverman (2002) that is driven by *two* hyper-parameters.

7 Applications

7.1 Chaotic elastica

A thin beam vibration experiment was designed by Constantine (1999) to generate a chaotic time series with an erratic spectrum. A flexible thin steel beam was mounted vertically to an electromechanical shaker which provided a transverse sinusoidal excitation. The beam tip was placed near two rare earth magnets so as to provide nonlinear buckling forces. The beam was treated with a viscoelastic strip adhered to one side to provide a little damping and form a more distinguishable fractal structure in phase space embeddings. A laser vibrometer recorded the beam tip velocity sampled at 1000Hz and the analog signal was streamed to a National Instruments data acquisition board. The top-left plot of Figure 3 shows the recorded time series of length $T = 2048$.

To estimate the spectrum of the time series, the standard technique consists of first calculating the raw periodogram and then smoothing it. It is known however that the periodogram is badly biased when the true spectrum is not smoothly varying. In that case, a multitaper estimate (Riedel and Sidorenko, 1995) provides a nearly unbiased estimate $\{s_1, \ldots, s_{1024}\}$ of the true spectrum at the cost of introducing dependence. For our application, we choose the $K = 3$ sinusoidal multitaper estimate. The subseries of raw spectrum estimates $\{s_1, s_5, \ldots, s_{1021}\}$ at frequencies $\{f_1, f_5, \ldots, f_{1021}\}$ is known to be approximately uncorrelated with approximate distribution $s_n \sim \frac{x_n}{2K}\chi^2_{2K}$, where x_n is the true spectrum value at frequencies f_n. The same is true for the three other subseries $\{s_2, s_6, \ldots, s_{1022}\}$, $\{s_3, s_7, \ldots, s_{1023}\}$ and $\{s_4, s_8, \ldots, s_{1024}\}$ which can be smoothed separately to estimate the spectrum at all 1024 nominal frequencies.

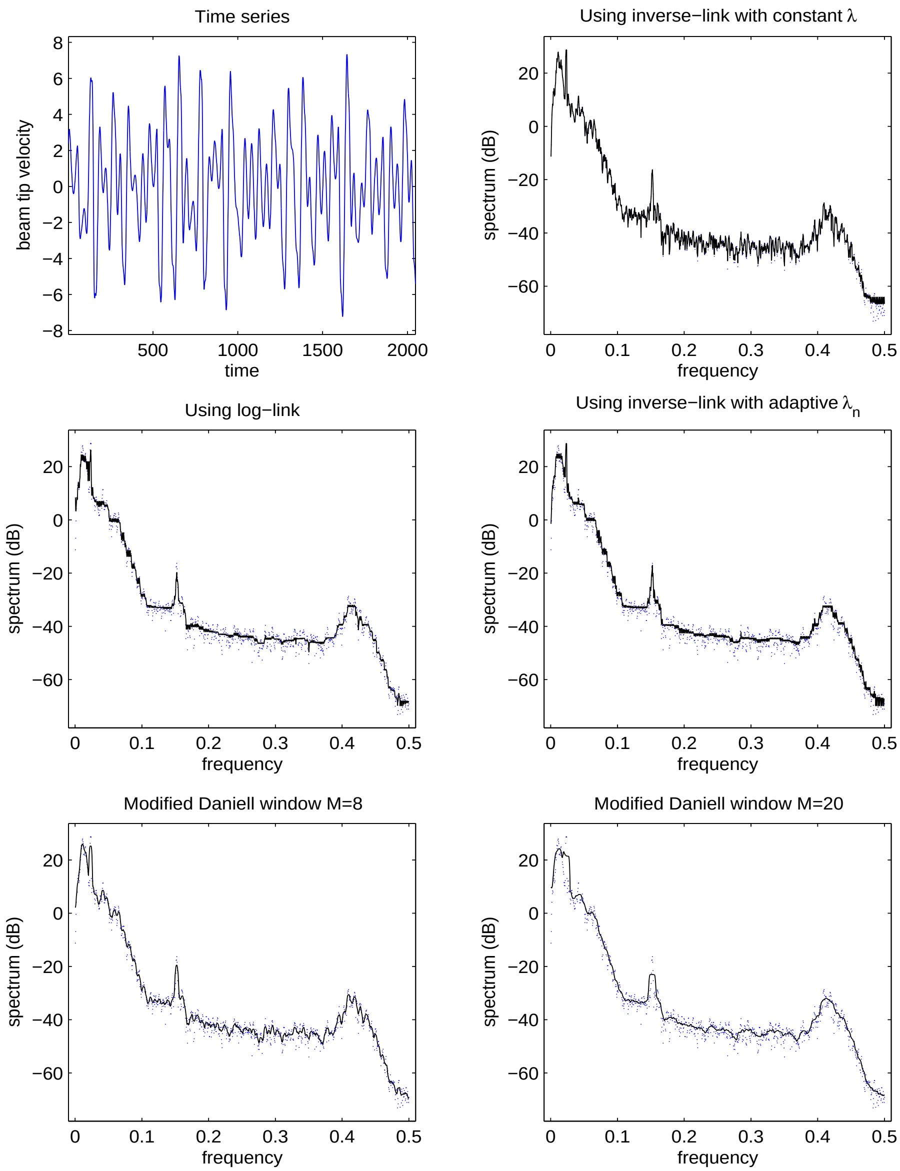

Figure 3: Top left: original time series. In the other five plots, the dots are the multitaper estimates of the spectrum and the lines are the estimates considered. The smoothing parameters are selected by two-fold cross validation.

For any of the four subsamples of N nearly independent components,

$$- \log p(\mathbf{x} \mid \mathbf{s}) = \sum_{n=1}^{N} \log x_n + s_n/x_n + \lambda \sum_{n=1}^{N-1} |x_{n+1} - x_n| \tag{19}$$

is a good approximation to the negative log-posterior distribution of $\mathbf{x}$ given the observations $\mathbf{s}$. It is *nonconvex* in $\mathbf{x}$, however, so IDM cannot be used to find a solution. The conditions of Theorem 2 are also not satisfied. A common, but not innocuous, remedy consists of using a link function $\eta = g(x)$ so that

$$- \log p(\eta \mid \mathbf{s}) = \sum_{n=1}^{N} \log\{g^{-1}(\eta_n)\} + s_n/g^{-1}(\eta_n) + \lambda \sum_{n=1}^{N-1} |\eta_{n+1} - \eta_n| \tag{20}$$

is *convex* in η. The remedy is not innocuous because the ℓ_1-penalties are not on the same scale for the original problem (19) and the transformed problem (20). The change of scale can have dramatic consequences on the estimate using:

1. The inverse-link $\eta = 1/x$. The conditions of Theorem 2 and Theorem 3 apply: the two proposed algorithms solve (20). Neglecting the change of scale in the ℓ_1-penalties between x and η leads to the top right estimate of Figure 3 that oversmooths low amplitudes and undersmooths high amplitudes. Because the inverse link function implies $|\eta_{n+1} - \eta_n| = \frac{|x_{n+1}-x_n|}{x_n x_{n+1}}$, we must choose appropriate λ_n's in $+\lambda \sum_{n=1}^{N-1} \lambda_n |\eta_{n+1} - \eta_n|$ to obtain a solution for (20) similar to that for (19). Comparing the first order optimality conditions suggests choosing $\lambda_n \propto \sqrt{\hat{x}_n \hat{x}_{n+1}}$, where $\hat{x}_n$ is for instance a running median of s_n (see the middle right estimate of Figure 3).

2. The log-link $\eta = \log x$. The conditions of Theorem 2 apply but not that of Theorem 3 (g^* is not domainwise differentiable), so only the primal algorithm solves (20). The ℓ_1-penalties need not adapt to the change of scale when the spectrum is plotted on a log-scale. For a constant λ, the log-link estimate is plotted in the bottom left of Figure 3.

By adjusting for the change of scale, the inverse-link estimate (middle right of Figure 3) is comparable to the log-link estimate (middle left of Figure 3). Without adjustment, the estimate can be dramatically bad (top right of Figure 3). For comparison to a standard spectral estimator, we added two modified Daniell estimates (Percival and Walden, 1993, p. 267) at the bottom of Figure 3. With a bandwidth of $M = 8$ (left), the estimator does not oversmooth the peaks, but undersmooths regions with no apparent features. In contrast, with $M = 20$ (right), the estimator smooths well, but erases two important peaks apparent with the locally adaptive MAP MRF-ℓ_1 estimator.

7.2 Energetic gamma-ray experiment telescope (EGRET)

Gamma-ray astronomy looks at the sky in a high range of energy invisible to the naked eye. A special telescope known as energetic gamma-ray experiment telescope (EGRET) placed aboard a satellite took measurements in the form of photon arrival times, directions and energies. These observations can be binned in space to obtain the image in the top-left image of Figure 4, an image we treat as independent counts from an inhomogeneous Poisson process with unknown intensity μ_n, $n = (i, j)$, $i = 1, \ldots, 360$, $j = 1, \ldots, 180$ (the coordinates are galactic latitude i and longitude j). In practice, the binning of the gamma-rays might cause a mild blurring effect and create some small dependence between measurements. The bright region in the center corresponds to gamma ray emissions from the center of our galaxy, the Milky Way. The counting phenomenon is spatially inhomogeneous with large dark areas contrasting with regions of high counts and a few bright clusters. So the *Laplace* MRF prior or a wavelet representation of the underlying Poisson parameters seems appropriate.

For the MAP MRF-ℓ_1 estimator, we use the identity link function and positivity constraints on the Poisson parameters. We compare its denoising performance to Waveshrink on the preprocessed Anscombe transformed data and to the MAP wavelet-ℓ_1 estimator of Sardy, Antoniadis, and Tseng (2004). The resulting estimates are plotted on a logarithmic scale $\log(\mu + 1)$ in the left column of Figure 4 for low counts in the range $\mu \in (0, 5]$, where the main differences between the three estimators can be observed. The right column looks at the one-dimensional counts at latitude $i = 126$ to illustrate the (in)ability of the estimates to adapt locally. The resulting images show: on the second row, some blurring is present due to the non-Gaussian nature of the transformed low counts; in the right plot, the second bump is oversmoothed; there is only little blurring in the third and fourth row, and the second 1D bump becomes visible especially with the MAP MRF-ℓ_1 estimator in the fourth row.

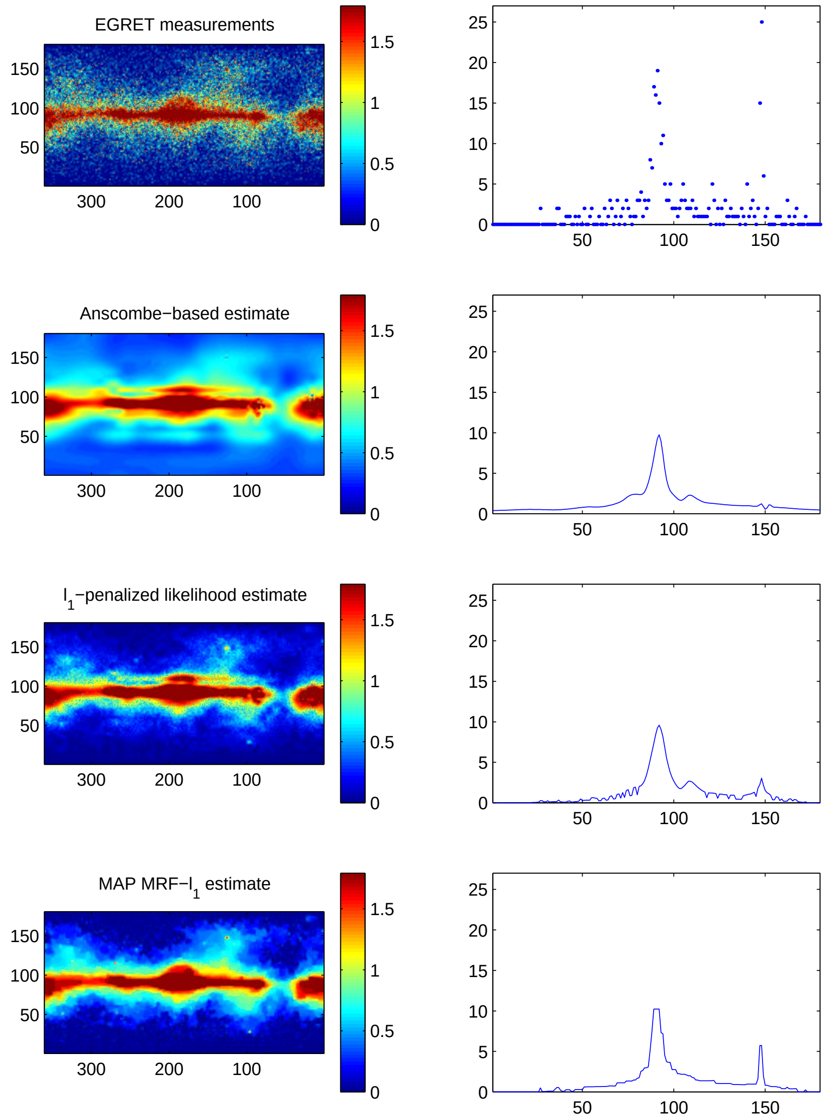

Figure 4: Column-wise: the astronomy image in the range $(0, 5]$ (left) and the one-dimensional intensities at latitude $i = 126$ (right). Row-wise: the raw Poisson measurements; the wavelet estimate using soft-Waveshrink, the universal threshold and Anscombe's transform; the wavelet ℓ_1-penalized likelihood estimate using the Poisson universal thresholds; the MAP MRF-ℓ_1 estimate using the cross validated smoothing parameter $\lambda = 1.05$.

8 Conclusions

The maximum a posteriori (penalized likelihood) estimate using a Laplace Markov random field prior (penalty) provides a good reconstruction of smooth and non-smooth functions. The new dual maximum a posteriori IDM method developed to find the MAP MRF-ℓ_1 estimate is efficient and is applicable on a wide class of likelihood functions and prior distributions. The proposed empirical Bayes selection of the smoothing parameter works well, and can possibly be improved allowing for a more flexible prior.

The above methods do not provide a measure of uncertainty for the estimate. Derivation of confidence interval is a difficult problem because the underlying function is not assumed to be differentiable and because the estimator is nonlinear. The Bayesian paradigm offers a good alternative for modeling uncertainty by calculating credible regions. Using Markov chain Monte Carlo methods to sample from the posterior distribution allows such calculations.

The good denoising performance and computational efficiency of the MAP MRF-ℓ_1 estimator offer promise for its use in nonparametric density estimation, and in more complicated situations such as denoising high-dimensional inverse problems. We are currently working on using MAP MRF-ℓ_1 in brain imaging to solve spatiotemporal inverse problems with spatially or temporally inhomogeneous structures.

9 Acknowledgments

This work was partially supported by the Swiss National Science Foundation and by the U.S. National Science Foundation. We thank Julian Besag for challenging and instructive discussions about a previous version of the paper, Bill Constantine and Donald Percival for providing the chaotic elastica data, Donald Percival for his advice about multitaper estimates, and Eric Kolaczyk for providing help with the astronomy application. We thank Valérie Chavez-Demoulin, Anthony Davison and Marc-Olivier Boldi for comments about the organization of the article. We thank the editor, associate editor and the referees for their valuable comments which substantially improved the article. The first author dedicates this paper to two great advisors: Adele Cutler and Werner Stuetzle.

A Optimization concepts and definitions

For any $f : \mathbb{R}^K \mapsto \mathbb{R} \cup \{\infty\}$, its effective domain is $\Omega_f = \{\mathbf{w} : f(\mathbf{w}) < \infty\}$. For $\mathbf{w} \in \Omega_f$ and any direction $\mathbf{d} \in \mathbb{R}^K$, the (lower) directional derivative at $\mathbf{w}$ is

$$f'(\mathbf{w}; \mathbf{d}) = \lim \inf_{\epsilon \to 0^+} \frac{f(\mathbf{w} + \epsilon \mathbf{d}) - f(\mathbf{w})}{\epsilon}.$$

We say that an $\mathbf{w} \in \Omega_f$ is a stationary point if $f'(\mathbf{w}; \mathbf{d}) \geq 0$ for all $d \in \mathbb{R}^K$. We say that an $\mathbf{w} \in \Omega_f$ is a critical point Auslender (1976, p. 94) if its function value cannot be decreased by changing one variable at a time, i.e.,

$$f(\mathbf{w}) \leq f(w_1, ..., w_{i-1}, \xi, w_{i+1}, ..., w_K) \quad \forall \xi \in \Re,\ 1 \leq i \leq K.$$

If Ω_f is open and f is differentiable on Ω_f, then $\mathbf{w}$ being a stationary point means $\nabla f(\mathbf{w}) = \mathbf{0}$ and is implied by $\mathbf{w}$ being a critical point. If f is convex, then $\mathbf{w}$ being a stationary point means $\mathbf{w}$ is a global minimum point (and hence a critical point). If either f is not convex or f is not differentiable on Ω_f, then a stationary point need not be a critical point (e.g., $f(w) = -w^2$ at $w = 0$) and a critical point need not be a stationary point (e.g., Auslender's example of Section 2). We also recall a few definitions for a function f on $\mathbb{R}^K$:

- f has coordinatewise unique minimum if, for every $\mathbf{w} \in \Omega_f$ and every $1 \leq i \leq K$, the function $\xi \mapsto f(w_1, ..., w_{i-1}, \xi, w_{i+1}, ..., w_K)$ has at most one minimum point.

- f is 2-coordinatewise convex if, for every $\mathbf{w} \in \Omega_f$ and every $1 \leq i < j \leq K$, the function $(\xi, \zeta) \mapsto f(w_1, ..., w_{i-1}, \xi, w_{i+1}, ..., w_{j-1}, \zeta, w_{j+1}, ..., w_K)$ is convex.

- f is level-bounded if the lower level set $\{\mathbf{w} : f(\mathbf{w}) \leq f(\mathbf{w}^0)\}$ is bounded for every $\mathbf{w}^0 \in \Omega_f$.

- f is domainwise differentiable if Ω_f is an open set, and f is differentiable on Ω_f and tends to infinity at any boundary point of Ω_f.

For instance, a strongly convex function is automatically level-bounded and has coordinatewise unique minimum. The converse is not true. For example, $\phi(u) = -1/(1 + u^2)$ of Geman and McClure (1987) is level-bounded and has coordinatewise unique minimum, but it is not strongly convex. A convex function is automatically 2-coordinatewise convex. The converse is not true.

B Algorithmic convergence

Proof of Theorem 1: ICM is a particular case of the BCR method of Section 2 using the essentially cyclic rule and with $|I^\nu| = 1$ for all ν. Since g is domainwise differentiable and h is differentiable, then f is domainwise differentiable (and $\Omega_f = \Omega_g$). Hence f is continuous on Ω_f and tends to infinity at any boundary point of Ω_f, implying each level set of f is closed. Since each level set of f is bounded by assumption, then it is compact. In the case where f is 2-coordinatewise convex, the conclusion follows from Lemma 3.1 and Theorem 4.1(a) of Tseng (2001). In the case where f has coordinatewise unique minimum, the conclusion follows from Lemma 3.1 and Theorem 4.1(c) of Tseng (2001). $\square$

Proof of Theorem 2: The argument is similar to the proof of Theorem 1. Since g is domainwise differentiable and h is continuous, then f is continuous on Ω_f and tends to infinity at any boundary point of Ω_f, implying each level set of f is closed. Since each level set of f is bounded by assumption, then it is compact. When f is 2-coordinatewise convex, the conclusion follows from Lemma 3.1 and Theorem 4.1(a) of Tseng (2001). When f has coordinatewise unique minimum, the conclusion follows from Lemma 3.1 and Theorem 4.1(c) of Tseng (2001), suitably extended to the case of essentially cyclic rule. $\square$

Proof of Theorem 3: If g is strictly convex and g^* is finite-valued everywhere, then g^* is differentiable (Rockafellar, 1970, Thm. 26.3). Since h_k is convex and hence continuous on $\mathbb{R}$, the primal problem (7) satisfies Assumptions A and B in Tseng (1993). By Tseng (1993, Thm. 3.1), the sequence $\{\mathbf{x}^\nu\}_{\nu=0,1,\ldots}$ is bounded and every cluster point solves (7). If instead g^* is domainwise differentiable and f is level-bounded, then it can be argued as in the proof of Theorem 2 that each level set of f is compact. Since f is also convex, it follows from Lemma 3.1 and Theorem 4.1(a) of Tseng (2001) that $\{\mathbf{w}^\nu\}_{\nu=0,1,\ldots}$ is bounded and every cluster point solves (13). Then, $\{\mathbf{x}^\nu\}_{\nu=0,1,\ldots}$ is bounded and, by Rockafellar (1970, Thm. 31.3 and Cor. 31.3.1), every cluster point solves (7). $\square$

C Proof of Proposition 1

Using the dual formulation (13) for the ℓ_1 prior, the conjugate function $h^*(\mathbf{w})$ is given in (14). Since g is proper, closed, and strictly convex, by Rockafellar

(1970, Thm. 26.3), g^* is essentially smooth, meaning that g^* is differentiable on $\text{int}\Omega_{g^*}$ and $\|\nabla g^*(\mathbf{u})\|_2 \to \infty$ as $\mathbf{u}$ approaches a boundary point of Ω_{g^*}. Since $\mathbf{0} \in \text{int}\Omega_{g^*}$, it can be argued that, for a given $\lambda > 0$, the optimization problem $\min_{\mathbf{w}:\|\mathbf{w}\|_\infty \leq \lambda} g^*(-B'\mathbf{w})$ has a solution $\mathbf{w}_\lambda$ with $-B'\mathbf{w}_\lambda \in \text{int}\Omega_{g^*}$. Then the estimate is $\hat{\mathbf{x}}_\lambda = \nabla g^*(-B'\mathbf{w}_\lambda)$. This estimate satisfies $\hat{\mathbf{x}}_\lambda \in \text{Ker}(B)$ provided λ is at least as large as $\lambda_{\min}(\mathbf{s}) = \min_{\mathbf{w}_\infty \in \Gamma} \|\mathbf{w}_\infty\|_\infty$, where $\Gamma = \{\mathbf{w}_\infty : \mathbf{w}_\infty = \arg\min_{\mathbf{w}} g^*(-B'\mathbf{w})\}$ is the set of all the dual solutions for $\lambda = +\infty$. Hence $\{\hat{\mathbf{x}}_\lambda \in \text{Ker}(B)\}$ and $\{\lambda_{\min}(\mathbf{s}) \leq \lambda\}$ are equiprobable. $\square$

We now assume a univariate first-order Laplace MRF prior. For standard Gaussian white noise, the conjugate likelihood function g^* is defined by (15). For $\lambda = +\infty$, the set Γ defined in Proposition 1 is composed of the unique element $\mathbf{w}_\infty := (BB')^{-1}B\mathbf{s}$, where $\mathbf{s} \sim \text{N}(c\mathbf{1}, I_N)$ for some $c \in \mathbb{R}$; correspondingly, the function estimate $\mathbf{x}_\infty = \bar{s}$ is the projection of the noisy signal in the kernel of B spanned by the constant vector $\mathbf{1}$. So the universal parameter is the smallest $\lambda(N)$ such that (18) is satisfied, where $\mathbf{w}_\infty \sim \text{N}(\mathbf{0}, \Sigma)$ with $\Sigma = (BB')^{-1}$. The inverse of BB', a tridiagonal matrix with 2 along the diagonal and -1 above and below the diagonal, has entries $(\Sigma)_{ij} = \min(i,j) - ij/N,\ i,j = 1,\ldots,N-1$, a result that can be checked by multiplying the two matrices to obtain the identity matrix. Hence the distribution of $\mathbf{w}_\infty/\sqrt{N}$ is that of a Brownian bridge W discretized on an equispaced grid $t = i/N,\ i = 0,\ldots,N$ of $[0,1]$, since $\text{Cov}(w_{\infty i}/\sqrt{N}, w_{\infty j}/\sqrt{N}) = t_1 \wedge t_2 - t_1 t_2 \equiv \text{Cov}(W(t_1), W(t_2))$ for $t_1 = i/N$ and $t_2 = j/N$. Using bounds on a Brownian bridge (Shorack and Wellner, 1986), we can majorize the probability of interest

$$\begin{aligned} P\{\|\mathbf{w}_\infty\|_\infty/\sqrt{N} \geq b_N\} &\leq P\{\|W\|_\infty \geq b_N\} \\ &\leq 2P\{\|W^+\|_\infty \geq b_N\} \\ &= 2\exp\{-2b_N^2\}. \end{aligned}$$

Choosing $b_N = \sqrt{\log(\log N)/2}$, the last quantity converges to zero at the rate $O(1/\log N)$. Hence the universal parameter is $\lambda_U = \sqrt{N} b_N$ for $N \geq 3$. $\square$

References

Abramovich, F. Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society,* *Series B* **60**, 725–749.

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35**, 246–254.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations (with discussion). *Journal of the American Statistical Association 96*(455), 939–967.

Auslender, A. (1976). *Optimisation Méthodes Numériques*. Paris: Masson.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

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

Besag, J. (1989). Digital image processing: Towards Bayesian image analysis. *Journal of Applied Statistics* **16**, 395–407.

Besag, J. (2001). Discussion of "Conditionally specified distributions: an introduction" by Arnold, B. C. Castillo, E. and Sarabia, J. M. *Statistical Science* **16**, 249–274.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **84**, 733–746.

Clifford, P. (1990). *Markov random fields in statistics* (G. R. Grimmet and D. J. A. Welsh ed.). In Disorder in Physical Systems: A Volume in Honor of John M. Hammersley. Oxford: Oxford University Press.

Constantine, W. (1999). *Wavelet Techniques for Chaotic and Fractal Dynamics*. Ph. D. thesis, Mechanical Engineering Department, University of Washington.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution (Pkg: P1-65). *The Annals of Statistics 29*(1), 1–48.

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

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90*(432), 1200–1224.

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.

Geman, S. and McClure, D. E. (1987). Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute (Proceedings of the 46th Session)* **52**, Book 4, 5–21.

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

Green, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified em algorithm. *IEEE Transactions on Medical Imaging* **9**, 84–93.

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

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Johnstone, I. and Silverman, B. (2002). Empirical Bayes selection of wavelet thresholds. Technical report, University of Bristol.

Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series.* Springer.

Kuensch, H. R. (1994). Robust priors for smoothing and image restoration. *Annals of the Institute of Statistical Mathematics* **46**, 1–19.

Li, Y. and Santosa, F. (1996). A computational algorithm for minimizing total variation in image restoration. *IEEE Transactions on Image Processing* **5**, 987–995.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

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

Percival, D. B. and Walden, A. T. (1993). *Spectral analysis for physical applications.* Cambridge University Press.

Powell, M. J. D. (1973). On search directions for minimization algorithms. *Mathematical Programming* **4**, 193–201.

Riedel, K. S. and Sidorenko, A. (1995). Minimum bias multiple taper spectral estimation. *IEEE Transactions on Signal Processing* **43**, 188–195.

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

Rudin, L., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D* **60**, 259–268.

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

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

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.

Timmermann, K. E. and Nowak, R. D. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory* **45**, 846–862.

Tseng, P. (1993). Dual coordinate ascent methods for non-strictly convex minimization. *Mathematical Programming* **59**, 231–247.

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