Automatic Smoothing With Wavelets for a Wide Class of Distributions

Sylvain SARDY, Anestis ANTONIADIS, and Paul TSENG

Wavelet-based denoising techniques are well suited to estimate spatially inhomogeneous signals. Waveshrink (Donoho and Johnstone) assumes independent Gaussian errors and equispaced sampling of the signal. Various articles have relaxed some of these assumptions, but a systematic generalization to distributions such as Poisson, binomial, or Bernoulli is missing. We consider a unifying l_1-penalized likelihood approach to regularize the maximum likelihood estimation by adding an l_1 penalty of the wavelet coefficients. Our approach works for all types of wavelets and for a range of noise distributions. We develop both an algorithm to solve the estimation problem and rules to select the smoothing parameter automatically. In particular, using results from Poisson processes, we give an explicit formula for the universal smoothing parameter to denoise Poisson measurements. Simulations show that the procedure is an improvement over other methods. An astronomy example is given.

Key Words: Convex programming; Interior point method; Karush–Kuhn–Tucker conditions; Penalized likelihood; Regularization; Signal denoising; Universal rule.

1. INTRODUCTION

Donoho and Johnstone (1994) introduced a nonparametric smoother called Waveshrink, which uses a set of orthonormal multiresolution basis functions called wavelets. Waveshrink rests on the following two assumptions: the sampling locations are equispaced, either on $[0, 1]$ for one-dimensional signals, or on a lattice of $[0, 1] \times [0, 1]$ for an image; the noise consists of independently and identically distributed Gaussian variables. When these assumptions are met, Waveshrink is an elegant and efficient estimator from both statistical and computational points of view. Its *thresholding* property allows Waveshrink to regularize the least squares problem not only by shrinking the wavelet coefficients toward zero, but by

Sylvain Sardy is TKKKK, Department of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland (E-mail: Sylvain.Sardy@epfl.ch). Anestis Antoniadis is Professor, Laboratoire de Modélisation et Calcul, Université Joseph Fourier, Tour IRMA, B. P. 53, 38041 Grenoble CEDEX 9, France (E-mail: Anestis.Antoniadis@imag.fr). Paul Tseng is Professor, Department of Mathematics, Box 354350, University of Washington, Seattle, WA 98195 (E-mail: tseng@math.washington.edu).

Journal of Computational and Graphical Statistics, Volume 13, Number 2, Pages 1–23
DOI: 10.1198/1061860043399

setting some of them to zero. Thresholding leads to a sparse wavelet representation of the underlying function, a key feature of wavelet-based estimation. Thresholding also provides remarkable theoretical properties, with near minimax results for a wide range of functions (Donoho, Johnstone, Kerkyacharian, and Picard 1995). A difficult problem is the automatic selection of the smoothing parameter. It has been addressed in many ways, including the universal rule of Donoho and Johnstone (1994) that we will elaborate upon.

Since Donoho and Johnstone's 1994 article, Waveshrink has been extended to a wider range of conditions such as dealing with the nonequispaced setting, a situation more common in statistics than in signal processing. This article is concerned with relaxing the Gaussianity of the noise. Most extensions considered so far assume symmetric zero-mean distribution. For instance, Averkamp and Houdré (1996) derived minimax thresholds for Waveshrink for additive noise with non-Gaussian densities; Johnstone and Silverman (1997) considered Gaussian colored noise; and Sardy, Tseng, and Bruce (2001) derived a robust version of Waveshrink for additive long-tailed noise distributions. Borrowing ideas from modulation estimators, Antoniadis and Sapatinas (2001) and Antoniadis, Besbeas, and Sapatinas (2001) developed a wavelet shrinkage methodology for obtaining and assessing smooth estimates for data arising from a (univariate) natural exponential family with quadratic and cubic variance functions. For the Poisson distribution, Kolaczyk (1999a) accounted for non-Gaussianity by proposing a level-dependent thresholding. Although computationally efficient, his adaptation of Waveshrink does not prevent the estimated Poisson means from taking negative values. Kolaczyk (1999b) and Timmermann and Nowak (1999) proposed similar Bayesian approaches: these approaches guarantee the positivity of the Poisson parameters estimates, but have been studied only for the Haar wavelet. Their Bayesian estimation does not satisfy the wavelet sparsity property, however, because the posterior mean estimation is a shrinkage, but not a thresholding, procedure. Finally, the simple approach of Donoho et al. (1995) consists of applying a variance stabilizing transformation (Anscombe 1948) to treat the transformed data as if they were Gaussian. This has the advantage of being computationally efficient, of not being restricted to the Haar wavelets, and of providing a positive estimate with a sparse wavelet representation, but has the drawback of working poorly in regions of small Poisson intensities.

The extension of wavelet-based estimation to a larger class of distributions is not straightforward. This article proposes a penalized likelihood approach that defines a general estimator in a unified way across a large class of distributions. This approach bears similarities with existing methodologies. In parametric regression, the generalized linear model of Nelder and Wedderburn (1972) is also a penalized likelihood approach, constraining the number of nonzero coefficients to be small by searching for a good bias-variance trade-off among subsets of explanatory variables. In nonparametric regression with splines, the penalized likelihood approaches of O'Sullivan, Yandell, and Raynor (1986) and Villalobos and Wahba (1987) constrain the smoothness measured by the quadratic integrated squared second derivatives of the estimated function. In a similar spirit, our estimator penalizes the likelihood with a penalty appropriate for wavelets, namely a nondifferentiable l_1 penalty on the wavelet coefficients.

A key ingredient to the parametric and nonparametric generalized linear models is the *link* function that monotonically maps the linear part of the model to the parameter of interest. We can distinguish two approaches on how to choose the link function:

- The canonical link function is computationally convenient, because it often maps the linear part of the model into the proper domain of the estimand, for example, the inverse of the log-link for the Poisson distribution maps R into R^+. One advantage of this approach is that the corresponding penalized likelihood problem is constraint free. For its simplicity, this approach is often used in parametric generalized linear model. It is also the approach of O'Sullivan, Yandell, and Raynor (1986) with splines.
- Other link functions can be more appropriate in certain applications. To denoise Poisson sampled signals, for instance, it is natural to model the mean of the Poisson process directly as a linear combination of splines or wavelets by using the identity link. The price to pay is that the estimation requires constraints, such as positive Poisson means for instance. Villalobos and Wahba (1987) used this approach with splines.

The flexibility in choosing a link function represents an advantage. Although the log-link is often used in parametric regression with Poisson responses, the identity link is preferable to denoise Poisson signals, as we will illustrate in Sections 5 and 6.

Section 2 lays out the problem and the notation, reviews Waveshrink, and proposes a unified way of generalizing it to a wide class of noise distributions. Section 3 solves the estimation problem with an interior point algorithm. For the automatic selection of the smoothing parameter, Section 4 defines the universal rule for any convex negative log-likelihood distribution. In particular, Section 4 checks that our definition matches the universal rule of Donoho and Johnstone (1994) for the Gaussian distribution. Section 4 also derives explicitly the universal smoothing parameter for denoising Poisson signals. Section 5 uses a Monte Carlo simulation to investigate the small sample behavior of the new estimator. A practical example with gamma-ray measurements is presented in Section 6, and the final section discusses our results, pointing out some further areas of research.

2. l_1-PENALIZED LIKELIHOOD ESTIMATOR

Suppose we have noisy measurements $\mathbf{s} = (s_1, \ldots, s_N)$ of a signal $\mu(\cdot)$ sampled at equispaced locations x_n. Throughout the article, we use the notation $\boldsymbol{\theta}$ to indicate a vector, and θ_n for its nth element with index $n \in \{1, \ldots, N\}$. We assume that the observations s_n are realizations of independent random variables with likelihood function $L(\mu_n; s_n, \sigma)$ parameterized by a parameter $\mu_n = \mu(x_n)$ and a scale parameter σ. Our problem is to estimate $\boldsymbol{\mu}$ taking into account the underlying spatial structure. To model the spatial structure of the signal, expansion-based estimators assume that the underlying signal $\mu(\cdot)$ can be well approximated by a linear expansion on a set of P known functions $\phi_p(\cdot)$,

$$\mu(\cdot) = \sum_{p=1}^{P} \alpha_p \phi_p(\cdot),$$

where P is typically at least as large as the number of observations N. The functions $\phi_p(\cdot)$ can be polynomials indexed by their degrees, smoothing splines indexed by their knots, or trigonometric functions indexed by their frequencies. Waveshrink uses a set of orthonormal wavelets. The standard univariate wavelets are multiresolution functions that are locally supported and indexed by a location parameter k and a scale parameter j. Letting j_0 be a small integer representing the number of low resolution levels, the father wavelet $\phi(\cdot)$ such that $\int_0^1 \phi(x)dx = 1$ generates $p_0 = 2^{j_0}$ approximation wavelets by means of the dilation and translation relation

$$\phi_{j_0,k}(x) = 2^{j_0/2}\phi(2^{j_0}x - k), \quad k = 0, 1, \ldots, 2^{j_0} - 1; \tag{2.1}$$

these capture the coarse features of the signal. Similarly, a mother wavelet $\psi(\cdot)$ such that $\int_0^1 \psi(x)dx = 0$ generates $N - p_0$ fine-scale wavelets

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j = j_0, \ldots, J;\ k = 0, 1, \ldots, 2^j - 1, \tag{2.2}$$

where $J = \log_2 N - 1$. The fine-scale wavelets capture the local features of the signal. We will assume that the true signal $\mu(\cdot)$ can be well approximated by an expansion on N orthonormal wavelets

$$\mu_N(\cdot) = \sum_{k=0}^{2^{j_0}-1} \beta_k \phi_{j_0,k}(\cdot) + \sum_{j=j_0}^{J} \sum_{k=0}^{2^j-1} \gamma_{j,k}\psi_{j,k}(\cdot). \tag{2.3}$$

Having represented $\mu(\cdot)$ as a linear expansion on wavelets, our problem of estimating the parameters $\boldsymbol{\mu}$ has now shifted to estimating the coefficients $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ of its wavelet representation

$$\boldsymbol{\mu} = \boldsymbol{\Phi}\boldsymbol{\alpha} = \begin{bmatrix}\boldsymbol{\Phi}_0 & \boldsymbol{\Psi}\end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix},$$

where $\boldsymbol{\Phi}_0$ is the $N \times p_0$ matrix of approximation wavelets and $\boldsymbol{\Psi}$ is the $N \times (N - p_0)$ matrix of fine-scale wavelets. The maximum likelihood estimator is usually asymptotically optimal in the parametric context, but it has too many degrees of freedom (e.g., too many basis functions) when the parametric dimension is large. Regularization of the maximum likelihood problem consists of adding constraints to decrease the degrees of freedom and the variance of the estimator at the cost of increasing its bias. To control the degrees of freedom, the linear smoothing spline estimator uses a quadratic roughness penalty approach (Green and Silverman 1994). Waveshrink also adds a penalty to the least squares problem that becomes

$$\min_{\boldsymbol{\alpha}=(\boldsymbol{\beta},\boldsymbol{\gamma})} \frac{1}{2}\|\mathbf{s} - \boldsymbol{\Phi}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{j=j_0}^{J} \sum_{k=0}^{2^j-1} |\gamma_{j,k}|^\nu, \tag{2.4}$$

where $\nu \in \{0, 1\}$ (with the convention $0^0 = 0$) and where $\lambda \geq 0$ is the so-called regularization or smoothing parameter. Waveshrink is a nonlinear estimator for $\nu \in \{0, 1\}$.

Hard-Waveshrink ($\nu = 0$) corresponds to best subset variable selection; in general, this combinatorial problem can only be solved approximately with a stepwise search. Soft-Waveshrink ($\nu = 1$) regularizes the least squares problem with an l_1-penalty. Only when the first term in (2.4) is quadratic and when $\mathbf{\Phi}$ is orthonormal, does the optimization problem (2.4) have a closed form solution via the hard- and soft-shrinkage functions (Donoho and Johnstone 1994). For nonorthonormal matrices, matching pursuit (Mallat and Zhang 1993) and basis pursuit (Chen, Donoho, and Saunders 1999) generalize hard- and soft-Waveshrink, respectively. All these estimators assume Gaussian additive noise through the use of the quadratic term. Except for the natural exponential family with quadratic and cubic variance functions (Antoniadis and Sapatinas 2001), the generalization of wavelet-based estimators to other distributions and to possibly nonorthonormal wavelet bases has not yet been undertaken in a systematic way.

We take the following approach: for a noisy signal $\mathbf{s}$ with known log-likelihood function $l(\boldsymbol{\mu}; \mathbf{s}) = \sum_n l(\mu_n; s_n)$, and for a representation of the signal $\boldsymbol{\mu}$ in an orthonormal or overcomplete wavelet basis $\mathbf{\Phi}$, we define the estimate as the solution to the l_1-penalized log-likelihood problem

$$\min_{\boldsymbol{\mu}, \boldsymbol{\alpha}=(\boldsymbol{\beta},\boldsymbol{\gamma})} -l(\boldsymbol{\mu}; \mathbf{s}) + \lambda \|\boldsymbol{\gamma}\|_1 \quad \text{with} \quad \boldsymbol{\mu} = \mathbf{\Phi}\boldsymbol{\alpha} \quad \text{and} \quad \boldsymbol{\mu} \in \mathbf{C}, \tag{2.5}$$

where $\mathbf{C} = C_{s_1} \times \cdots \times C_{s_N}$ and C_s denotes the domain of $l(\cdot; s)$. The l_1 penalty ensures a parsimonious representation in a wavelet basis. We assume that C_s is an interval and that $-l(\cdot\,; s)$ is strictly convex on C_s. This assumption is convenient mathematically because it implies the uniqueness of the solution $\hat{\boldsymbol{\mu}}_\lambda$ to (2.5); the solution is also unique in the wavelet coefficients $\boldsymbol{\alpha}$ if the wavelet matrix $\mathbf{\Phi}$ is orthonormal (but not guaranteed if $\mathbf{\Phi}$ is overcomplete). This assumption might require a reparameterization of the likelihood function by means of a bijective function, the link function of generalized linear models. We henceforth assume that the parameterization is strictly convex. Most canonical links of the exponential family give convexity and conveniently map the estimate in its proper domain, so that no constrained optimization is necessary; this is for instance the approach of O'Sullivan, Yandell, and Raynor (1986) with splines. Other link functions are sometimes more natural, for instance the identity link for the Poisson EGRET image of Section 6. With the identity link, the Poisson log-likelihood is indeed strictly convex, but constraints must then be introduced to maintain the Poisson intensities $\boldsymbol{\mu}$ in $\mathbf{C} = (0, \infty)^N$. Constrained optimization for spline-based estimator is also the approach of Villalobos and Wahba (1987) for estimating probabilities in $[0, 1]$.

3. MAXIMIZING l_1-PENALIZED LIKELIHOOD

The convex programming problem (2.5) that defines our estimator is reminiscent of the basis pursuit estimator of Chen, Donoho, and Saunders (1999) obtained by an interior point method. Problem (2.5) is, however, more complex in that inequality constraints are present, the l_1 penalty in (2.5) only applies to fine-scale coefficients $\boldsymbol{\gamma}$ (not the whole vector $\boldsymbol{\alpha}$ as

in basis pursuit), and the log-likelihood is not quadratic. Consequently the dual problem has additional equality and inequality constraints, which must be handled differently using Lagrange multipliers and additional log-barrier penalties. Special care must also be taken to handle certain log-likelihood functions, such as that corresponding to zero Poisson counts. We derive the dual problem of (2.5) in Section 3.1, and we develop a primal-dual log-barrier interior point method capable of solving (2.5) for a large class of distributions in Section 3.2.

3.1 PRIMAL AND DUAL PROBLEMS

By attaching Lagrange multipliers $\mathbf{y}$ associated with the linear constraints, we can rewrite the primal problem (2.5) as

$$\min_{\boldsymbol{\mu}\in\mathbf{C},\boldsymbol{\alpha}} \max_{\mathbf{y}} -l(\boldsymbol{\mu};\mathbf{s}) + \sum_{p=p_0+1}^{P} \lambda|\gamma_p| + \mathbf{y}'(\boldsymbol{\mu} - \boldsymbol{\Phi}\boldsymbol{\alpha}).$$

Interchanging "max" and "min" yields the dual problem

$$\begin{aligned}
&\max_{\mathbf{y}} \min_{\boldsymbol{\mu}\in\mathbf{C},\boldsymbol{\alpha}} -l(\boldsymbol{\mu};\mathbf{s}) + \mathbf{y}'\boldsymbol{\mu} - \sum_{p=1}^{P} \alpha_p(\mathbf{y}'\boldsymbol{\Phi}_p) + \sum_{p=p_0+1}^{P} \lambda|\gamma_p| \\
&= \max_{\mathbf{y}} \left\{ \sum_n \min_{\mu_n\in C_n} -l(\mu_n; s_n) + y_n\mu_n \right. \\
&\qquad \left. + \sum_{p=1}^{p_0} \min_{\beta_p} -\beta_p(\mathbf{y}'\boldsymbol{\Phi}_p) + \sum_{p=1}^{P-p_0} \min_{\gamma_p} -\gamma_p(\mathbf{y}'\boldsymbol{\Psi}_p) + \lambda|\gamma_p| \right\}. \qquad (3.1)
\end{aligned}$$

Now consider the two univariate functions involved in (3.1):

- $f_1(\mu; y) = -l(\mu; s) + y\mu$ with $\mu \in C_s$. For a given y and s, if f_1 has a minimum at some $\mu \in C_s$, we denote this μ by $\mu_{\min}(y; s)$; it is unique because $-l(\cdot; s)$ is strictly convex. Let K_s denote the set of y for which $\mu_{\min}(y; s)$ is defined. Because $-l(\cdot; s)$ is convex on C_s, then K_s is an interval and $h(y; s) = f_1(\mu_{\min}(y; s); y)$ is concave on K_s (Rockafellar 1984, sec. 8E).
- $f_2(\gamma; c) = c\gamma + \lambda|\gamma|$. For a given c and $\lambda > 0$, f_2 has a minimum at some γ if and only if $-\lambda \le c \le \lambda$, with minimum at $\gamma_{\min} = 0$; for $\lambda = 0$, f_2 has a minimum if and only if $c = 0$.

Denoting $\mathbf{A} = [\boldsymbol{\Psi}, -\boldsymbol{\Psi}]$ and $\mathbf{c} = \lambda\mathbf{1}$, we can rewrite the dual problem (3.1) as

$$\max_{\mathbf{y}} \sum_n h(y_n; s_n) \quad \text{with} \quad \mathbf{A}'\mathbf{y} - \mathbf{c} \le \mathbf{0},\ \boldsymbol{\Phi}_0'\mathbf{y} = \mathbf{0}, \quad \text{and} \quad \mathbf{y} \in \mathbf{K}, \qquad (3.2)$$

where $\mathbf{K} = K_{s_1} \times \cdots \times K_{s_N}$. This is a convex programming problem with linear constraints since $h(\cdot; s)$ is concave. Because the primal cost function is separable and each univariate function $l(\cdot; s)$ is convex on its domain C_s, the duality gap is zero, and a solution of the primal problem yields as byproduct a solution of the dual and conversely (Rockafellar 1984,

Table 1. Negative Log-Likelihood Functions $-l(\mu; s)$, Domain of Observation s, Domain of Parameter of Interest μ, Function $\mu_{\min}(y;s)$ and Domain K_s of Dual Variable y for Some Standard Distributions

Distribution	$-l(\mu; s)$	(s^l, s^u)	C_s	$\mu_{min}(y; s)$	K_s
Gaussian	$(s-\mu)^2/2$	$\mathbb{R}$	$\mathbb{R}$	$s-y$	$\mathbb{R}$
exponential	$-\log\mu + s\mu$	$[0, \infty)$	$(0, \infty)$	$1/(s+y)$	$(-s, \infty)$
Poisson	$\mu - s\log\mu$	$[0, \infty)$	$(0, \infty)$	$s/(1+y)$	$(-1, \infty)$
Bernoulli	$\begin{cases} -\log(1-\mu) \\ -\log\mu \end{cases}$	$\begin{matrix} 0 \\ 1 \end{matrix}$	$(0, 1)$	$\begin{cases} 1+1/y \\ 1/y \end{cases}$	$\begin{cases} y < -1 \\ -y < -1 \end{cases}$

sec. 11D). Solving the primal and the dual problems at once, we can monitor convergence by measuring the gap between them.

Table 1 gives the functions $-l(\mu; s)$, $\mu_{\min}(y; s)$ and their domains for some standard distributions. Because we are applying a Newton-type method (see Section 3.2) $h(y; s)$ must be twice differentiable in y. This occurs if $-l(\mu; s)$ is twice differentiable in μ, if its second derivative is always positive (so $-l(\mu; s)$ is strictly convex) and if its first derivative tends to $-\infty$ and $+\infty$ at the left and right endpoints of its domain (if the endpoints are finite). These conditions hold for the Gaussian and the exponential distributions. They also hold for the Poisson distribution, as long as $s_n \neq 0$ for all n. But, because $-l'(0; s_n = 0) = 1 \neq -\infty$, the conditions do not hold for zero Poisson counts. To cope with N_0 zero counts, the vector $\mathbf{c}$ and the matrix $\mathbf{A}$ must be extended to $\mathbf{c} = (\mathbf{1}_{N_0}, \lambda\mathbf{1})$ and $\mathbf{A} = [\mathbf{A}_0, \mathbf{\Psi}, -\mathbf{\Psi}]$, where $\mathbf{A}_0$ is the $N \times N_0$ matrix whose jth column has a -1 entry on the ith row when s_i is the jth null observation in $\mathbf{s}$. Another possibility is to set zero counts to a small positive value, say $s_n = 0.01$.

3.2 INTERIOR POINT ALGORITHM

The log-barrier subproblem associated with (3.2) is

$$\min_{\mathbf{y}} \sum_{n=1}^{N} -h(y_n; s_n) - \boldsymbol{\eta}'\mathbf{\Phi}_0'\mathbf{y} - \rho \sum_{p=p_0+1}^{P} \log(\lambda - \mathbf{\Psi}_p'\mathbf{y}) - \rho \sum_{p=p_0+1}^{P} \log(\lambda + \mathbf{\Psi}_p'\mathbf{y}),$$

where $\boldsymbol{\eta}$ is a Lagrange multiplier vector associated with the equality constraints in (3.2) and ρ is the log-barrier parameter chosen identically for all the penalty terms. In the case of the Poisson distribution, there is another log-barrier term $-\rho\sum_{n:s_n=0}\log(y_n+1)$ for those y_n that do not already have a barrier implicit in $-h(y_n; s_n)$, when $s_n = 0$. By introducing the slack variables $\mathbf{z} := (\lambda\mathbf{1} - \mathbf{\Psi}'\mathbf{y}, \lambda\mathbf{1} + \mathbf{\Psi}'\mathbf{y})$, the Karush–Kuhn–Tucker conditions for the subproblem can be written as a set of nonlinear equations:

$$\begin{aligned} -\mathbf{A}'\mathbf{y} + \mathbf{c} - \mathbf{z} =: \mathbf{r}_x &= \mathbf{0}, \\ -\mathbf{A}\mathbf{x} + \mu_{\min}(\mathbf{y}; s) + \mathbf{\Phi}_0\boldsymbol{\eta} =: \mathbf{r}_y &= \mathbf{0}, \\ \rho\mathbf{1} - \mathbf{X}\mathbf{z} =: \mathbf{r}_z &= \mathbf{0}, \\ -\mathbf{\Phi}_0'\mathbf{y} =: \mathbf{r}_\eta &= \mathbf{0}, \end{aligned}$$

with $\mathbf{x} > 0$, $\mathbf{z} > 0$ and $\mathbf{X} = \mathrm{diag}(\mathbf{x})$. We take a single Newton step to solve approximately this set of nonlinear equations. The Newton direction $(\Delta\mathbf{x}, \Delta\mathbf{y}, \Delta\mathbf{z}, \Delta\boldsymbol{\eta})$ is obtained by solving the system of linear equations:

$$\begin{aligned} \mathbf{A}'\Delta\mathbf{y} + \Delta\mathbf{z} &= \mathbf{r}_x, \\ \mathbf{A}\Delta\mathbf{x} + P\Delta\mathbf{y} - \boldsymbol{\Phi}_0\Delta\boldsymbol{\eta} &= \mathbf{r}_y, \\ \mathbf{Z}\Delta\mathbf{x} + \mathbf{X}\Delta\mathbf{z} &= \mathbf{r}_z, \\ \boldsymbol{\Phi}_0'\Delta\mathbf{y} &= \mathbf{r}_\eta, \end{aligned}$$

where $\mathbf{Z} = \mathrm{diag}(\mathbf{z})$ and $P = \mathrm{diag}(-\mu'_{\min}(\mathbf{y}; \mathbf{s}))$ has nonnegative diagonals. The dual Newton direction $(\Delta\mathbf{y}, \Delta\boldsymbol{\eta})$ is the solution of

$$\begin{aligned} \mathbf{Q}\Delta\mathbf{y} - \boldsymbol{\Phi}_0\Delta\boldsymbol{\eta} &= \mathbf{r} := \mathbf{r}_y - \mathbf{A}\mathbf{Z}^{-1}(\mathbf{r}_z - \mathbf{X}\mathbf{r}_x), \\ -\boldsymbol{\Phi}_0'\Delta\mathbf{y} &= -\mathbf{r}_\eta, \end{aligned} \tag{3.3}$$

where $\mathbf{Q} = P + \mathbf{A}\mathbf{Z}^{-1}\mathbf{X}\mathbf{A}'$. The dual slack and primal Newton directions are then obtained by $\Delta\mathbf{z} = \mathbf{r}_x - \mathbf{A}'\Delta\mathbf{y}$ and $\Delta\mathbf{x} = \mathbf{Z}^{-1}(\mathbf{r}_z - \mathbf{X}\Delta\mathbf{z})$. Because $\boldsymbol{\Phi}_0$ and $\mathbf{A}$ are typically dense and of high dimension, the system (3.3) cannot be solved using a direct method like Cholesky factorization. Appendix A shows that, by exploiting the fast multiplication $\mathbf{A}$, $\boldsymbol{\Phi}_0$, and $\mathbf{A}'$, $\boldsymbol{\Phi}_0'$, this system of linear equations can be practically solved with variants of the conjugate gradient algorithm, either for an orthonormal wavelet matrix $\boldsymbol{\Phi}$ (leading to the fastest algorithm) or for an overcomplete one. The interior point algorithm requires an initial point $(\mathbf{x}^0, \mathbf{y}^0, \mathbf{z}^0, \boldsymbol{\eta}^0)$ within the feasible domain. Rapid convergence will be obtained if the initial point is close to the optimal solution. The strategy for finding an initial point depends on the distribution. The strategy of Chen, Donoho, and Saunders (1999) for the Gaussian model can be applied, slightly modified, to Poisson and exponential models. For the Poisson case, for instance, we propose the following initial point. Let $\boldsymbol{\alpha} = \boldsymbol{\Phi}'\mathbf{s}$, $\boldsymbol{\alpha}_+ = \max(\boldsymbol{\alpha}, \mathbf{0})$ and $\boldsymbol{\alpha}_- = \max(-\boldsymbol{\alpha}, \mathbf{0})$. Then the primal variables $\mathbf{x}^0 = (\mathbf{0}_{N_0}, \boldsymbol{\alpha}_+, \boldsymbol{\alpha}_-) + .1\mathbf{1}$ are positive (N_0 is the number of $s_n = 0$). Also, let $\boldsymbol{\eta}^0 = -\boldsymbol{\beta}$ ($\boldsymbol{\beta}$ comprises the p_0 approximation coefficients of $\boldsymbol{\alpha}$) and $\mathbf{y} = \mathbf{A}\mathbf{x}^0$ and $\bar{\omega} = 1.1 \max\{\|\mathbf{y}\|_\infty, \|\mathbf{A}'\mathbf{y}\|_\infty / \min(\mathbf{c})\}$. Then the dual variables $\mathbf{y}^0 = \mathbf{y}/\bar{\omega}$ satisfy $\mathbf{y}^0 + \mathbf{1} > \mathbf{0}$ and $\mathbf{A}'\mathbf{y}^0 - \mathbf{c} < \mathbf{0}$, and the dual slack variables $\mathbf{z}^0 = \mathbf{c} - \mathbf{A}'\mathbf{y}^0$ are positive.

Flowchart of the interior point algorithm.

1. Choose an initial point $(\mathbf{x}^0, \mathbf{y}^0, \mathbf{z}^0, \boldsymbol{\eta}^0)$ and the log-barrier parameter ρ;
2. Find Newton directions $(\Delta\mathbf{x}, \Delta\mathbf{y}, \Delta\mathbf{z}, \Delta\boldsymbol{\eta})$ by solving (3.3);
3. Update $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\eta})$ using Newton directions while not violating the inequality constraints, for example, for Poisson:

$$\begin{cases} \mathbf{x}^{\text{new}} = \mathbf{x} + 0.95\nu\Delta\mathbf{x}, \\ \mathbf{y}^{\text{new}} = \mathbf{y} + 0.95\nu\Delta\mathbf{y}, \\ \mathbf{z}^{\text{new}} = \mathbf{z} + 0.95\nu\Delta\mathbf{z}, \\ \boldsymbol{\eta}^{\text{new}} = \boldsymbol{\eta} + 0.95\nu\Delta\boldsymbol{\eta}, \end{cases}$$

where $\nu = \min\{\nu_1, \nu_2, \nu_3\}$ with $\nu_1 = \min_{i:\Delta x_i < 0}\{-x_i/\Delta x_i\}$, $\nu_2 = \min_{i:\Delta y_i < 0}\{-(y_i + 1)/\Delta y_i\}$, and $\nu_3 = \min_{i:\Delta z_i < 0}\{-z_i/\Delta z_i\}$.

Decrease the log-barrier parameter $\rho^{\text{new}} = (1 - \min(\nu, 0.95))\rho$;

4. If convergence criterion (primal and dual feasibility, and duality gap) not met, go to Step 2.

The complexity of interior point algorithms is difficult to assess. Solving (3.3) is the most computer-intensive calculation at each interior point iteration. Chen, Donoho, and Saunders (1999) solved (3.3) by conjugate gradient and studied the time increase as the problem dimension and the desired solution accuracy increase. If the discrete wavelet transform is used, then the pyramid algorithm (Mallat 1989) offered $O(N)$ calculation at each conjugate gradient iteration. If many processors are available, the wavelet transform can be parallelized (Fridman and Manolakos 1997) to speed up the conjugate gradient algorithm. As far as the number of interior point iterations is concerned, we observed that it typically ranges between 5 and 15. For details on the implementation of an interior point algorithm, see Sardy, Bruce, and Tseng (2000). See also the Matlab code provided in Section 8.

4. AUTOMATIC SELECTION OF SMOOTHING PARAMETER

Several automatic rules have been proposed to select the smoothing parameter λ for Waveshrink with Gaussian noise, such as universal and minimax (Donoho and Johnstone 1994), minimizing the Stein unbiased risk estimate (SURE) (Donoho and Johnstone 1995), and minimizing the cross-validation estimate of the mean squared error (Nason 1995). The elegant mathematical derivations for the minimax and SURE selection rules depend heavily on Gaussianity and the orthonormality of the wavelet matrix. Johnstone and Silverman (1997) have extended Waveshrink for data with correlated Gaussian noise. Generalizing the fine minimax and SURE selection rules to other distributions appears difficult.

The universal rule (Donoho and Johnstone 1994) is not based on minimizing an estimate of the mean squared error, but aims to approach the oracular mean squared error within a factor of $2 \log N + 1$. The universal parameter $\lambda_N = \sqrt{2 \log N}$ achieves this property and is nearly minimax for a wide variety of loss functions and smoothness classes (Donoho et al. 1995). Donoho and Johnstone (1994) also showed that, with high probability, the universal parameter will exactly reproduce a zero-constant signal $\boldsymbol{\mu}_0 = \mathbf{0}$ sampled with standard Gaussian white noise $\mathbf{s} = \mathbf{0} + \mathbf{z}$ because

$$\mathrm{P}\left(\max_n |\tilde{z}_n| \le \sqrt{2 \log N}\right) \overset{N \to \infty}{\longrightarrow} 1,$$

where $\tilde{\mathbf{z}} = \boldsymbol{\Phi}'\mathbf{z} \sim \mathrm{N}(\mathbf{0}, I_N)$ is the least squares estimate of the wavelet coefficients. This result can be extended to the situation where the underlying signal is in the range of the approximation wavelets, that is, when $\boldsymbol{\mu}_0 = \boldsymbol{\Phi}_0 \boldsymbol{\beta}_0 + \boldsymbol{\Psi}\mathbf{0}$ for some $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$. In that situation, the universal rule guarantees that, asymptotically, Waveshrink sets all fine-scale wavelet coefficients to zero. Indeed, for $\lambda_N = \sqrt{2 \log N}$, we have

$$\mathrm{P}\left(\hat{\boldsymbol{\gamma}}_{\lambda_N} = \mathbf{0}\right) = \mathrm{P}(\max |\tilde{\mathbf{z}}| \le \lambda_N) \overset{N \to \infty}{\longrightarrow} 1, \tag{4.1}$$

for $\tilde{\mathbf{z}} = \boldsymbol{\Psi}'\mathbf{s} \sim \mathrm{N}(\mathbf{0}, I_{N-p_0})$. In the Gaussian case, the universal parameter is defined as a bound on the maximum of a Gaussian white noise sequence, but its generalization to

other distributions is not straightforward. Proposition 1 shows how the universal rule can be defined for any distribution with a concave and differentiable log-likelihood function, for example, Gaussian, exponential, Poisson and Bernoulli.

Proposition 1. *Suppose that the signal* $\mathbf{s}$ *has a log-likelihood function* $l(\boldsymbol{\mu}_0; \mathbf{s})$ *defined on a domain of the form* $\mathbf{C} = C_{s_1} \times \cdots \times C_{s_N}$, *where* C_s *denotes the domain of* $l(\cdot; s)$. *Suppose that the parameters of interest are a linear combinations of the* p_0 *scaling wavelets only, that is,* $\boldsymbol{\mu}_0 = \boldsymbol{\Phi}_0 \boldsymbol{\beta}_0 + \boldsymbol{\Psi}\mathbf{0}$ *for some* $\boldsymbol{\beta}_0$ *such that* $\boldsymbol{\Phi}_0 \boldsymbol{\beta}_0 \in \mathbf{C}$, *and that the log-likelihood is concave and differentiable in* $\boldsymbol{\mu}$. *Then the universal parameter* λ_N *is defined as the smallest* $\lambda(N)$ *such that*

$$\mathrm{P}\left(\max |\boldsymbol{\Psi}'\mathbf{y}| \le \lambda(N)\right) \stackrel{N\to\infty}{\longrightarrow} 1,$$

where the random vector $\mathbf{y}$ *together with some* $\boldsymbol{\mu} \in \mathbf{C}$ *and* $\boldsymbol{\beta}$ *satisfies*

$$\begin{aligned} -\nabla_{\boldsymbol{\mu}} l(\boldsymbol{\mu}; \mathbf{s}) + \mathbf{y} &= \mathbf{0}, \\ \boldsymbol{\Phi}_0' \mathbf{y} &= \mathbf{0}, \\ \boldsymbol{\Phi}_0 \boldsymbol{\beta} &= \boldsymbol{\mu}. \end{aligned} \tag{4.2}$$

Appendix B gives the proof of this proposition.

For Gaussian noise, one can use Proposition 1 to check that $\sqrt{2\log N}$ is the universal threshold (Donoho and Johnstone 1994). Applying Proposition 1 to the Poisson distribution, the universal parameter is found to be the smallest $\lambda(N)$ such that

$$P(\max |\boldsymbol{\Psi}'(\mathbf{z} - \mathbf{1})| \le \lambda(N)) \stackrel{N\to\infty}{\longrightarrow} 1 \quad \text{subject to} \quad \left\{ \begin{array}{rcl} \mathbf{Z}\boldsymbol{\mu} &=& \mathbf{s}, \\ \boldsymbol{\Phi}_0' \mathbf{z} &=& \boldsymbol{\Phi}_0' \mathbf{1}, \\ \boldsymbol{\Phi}_0 \boldsymbol{\beta} &=& \boldsymbol{\mu}, \end{array} \right. \tag{4.3}$$

where $\mathbf{z} = \mathbf{1} + \mathbf{y}$, $\mathbf{s} \sim$ Poisson$(\boldsymbol{\mu}_0)$, $\mathbf{Z} = \mathrm{diag}(\mathbf{z})$, and $\boldsymbol{\mu}_0 = \boldsymbol{\Phi}_0 \boldsymbol{\beta}_0$ for some $\boldsymbol{\beta}_0$. The asymptotic behavior of $\max |\boldsymbol{\Psi}'(\mathbf{z} - \mathbf{1})|$ constrained to the maximum likelihood conditions in the range of $\boldsymbol{\Phi}_0$ is a difficult problem. In Appendix C, we solve it using results from inhomogeneous continuous-time Poisson process (Andersen, Borgan, Gill, and Keiding 1993). In particular, we prove that, in the Poisson case, the appropriate universal smoothing parameter is level dependent and equal to

$$\lambda_{N,j} = M(\mu_0, \boldsymbol{\Psi}) 2^{j/2} \sqrt{2\log N} / \sqrt{N}, \tag{4.4}$$

where

$$M^2(\mu_0, \boldsymbol{\Psi}) := \max_{u \in [0,1]} \{\psi^2(u)\} \int_0^1 1/\mu_0(s) ds.$$

The constant $M(\mu_0, \boldsymbol{\Psi})$ requires the knowledge of μ_0. In practice $\boldsymbol{\mu}_0$ can be estimated in the range of the approximation wavelets using the estimator proposed by Donoho et al. (1995) for Poisson noise, based on the variance stabilizing transformation of Anscombe (1948). Indeed $\tilde{s}_n = 2\sqrt{s_n + 3/8} \stackrel{\cdot}{\sim} \mathrm{N}(2\sqrt{\mu_n}, 1)$ is approximately Gaussian. So setting

the fine-scale wavelets coefficients of the Waveshrink estimate of $\eta_{0_n} = 2\sqrt{\mu_{0_n}}$ based on the transformed data $\tilde{\mathbf{s}}$, one gets the desired estimate $\hat{\mu}_{0,n} = \{\hat{\eta}_{0,n}/2\}^2$.

A level dependent smoothing parameter for smoothing Poisson signals was also derived by Kolaczyk (1999a) to apply Waveshrink directly to Poisson measurements. Note that $\lambda_{N,j} \propto 2^{j/2}$ amounts to using unnormalized wavelets (see wavelet definition (2.1) and (2.2)). Using unnormalized Haar wavelets for smoothing Poisson data is also the approach of Timmermann and Nowak (1999, sec. II A). Finally, Equation (4.4) is also interesting in the sense that it gives the appropriate scaling $\lambda_{N,j}^{(\mathrm{CV})} = \lambda 2^{j/2}$ for searching the smoothing parameter λ by cross-validation.

5. SIMULATION

We investigate the predictive performance of the l_1-penalized likelihood estimator on Poisson data. The current state-of-the-art method developed specifically for Poisson noise is the Haar-based shift-invariant MMI (multiscale multiplicative innovation) estimator of Timmermann and Nowak (1999). We follow their simulation and generate observations $s_n \sim \mathrm{Poi}(\mu(x_n))$ as the realization of an inhomogeneous Poisson process. We consider two inhomogeneous and erratic intensity functions $\mu(\cdot)$: `blocks` will generate high Poisson counts and `bumps` will generate few counts except at the "bumps" locations. The exact definition of these signals can be found in Donoho and Johnstone (1994). Both functions are rescaled and shifted to range in $[1/\mathrm{peak}, \mathrm{peak}]$ with peak $\in \{8, 128\}$, so as to create a wide range of conditions that can be found in practice, with regions of near zero counts. This situation is different than the "burst-like" Poisson process considered by Kolaczyk (1999a), because no large homogeneous background process is present. For the intensity functions considered, gaussianization of the observations by means of a variance stabilizing transformation such as $2\sqrt{s + 3/8}$ proposed by Anscombe (1948) will not be effective for low counts. So Waveshrink applied to the transformed observations should not perform well.

For the l_1-penalized Poisson likelihood estimator, both the canonical log-link and the identity link can be used. Using the canonical log-link would lead to an undesirable multiplicative model for the Poisson parameters, since

$$\log \mu(x_n) = \sum_{k=0}^{2^{j_0}-1} \beta_k \phi_{j_0,k}(x_n) + \sum_{j=j_0}^{J} \sum_{k=0}^{2^j-1} \gamma_{j,k} \psi_{j,k}(x_n)$$

is equivalent to

$$\mu(x_n) = \prod_{k=0}^{2^{j_0}-1} \tilde{\beta}_k \exp\{\phi_{j_0,k}(x_n)\} \cdot \prod_{j=j_0}^{J} \prod_{k=0}^{2^j-1} \tilde{\gamma}_{j,k} \exp\{\psi_{j,k}(x_n)\},$$

where $\tilde{\beta}_k = \exp \beta_k$ and $\tilde{\gamma}_{j,k} = \exp \gamma_{j,k}$. The identity link is more appropriate, because it models the Poisson intensities directly as a linear expansion on an orthonormal wavelet

basis

$$0 < \mu(x_n) = \sum_{k=0}^{2^{j_0}-1} \beta_k \phi_{j_0,k}(x_n) + \sum_{j=j_0}^{J} \sum_{k=0}^{2^j-1} \gamma_{j,k} \psi_{j,k}(x_n),$$

with a positivity constraint.

For the l_1-penalized Poisson likelihood estimator and Waveshrink-Anscombe, we use an orthonormal wavelet basis with the symmlets of order 4. We also use the Haar wavelet to compare with the simulation results of Timmermann and Nowak (1999). The number of approximation wavelets $p_0 = 2^{j_0}$ is not a crucial issue as the estimator is relatively insensitive to the choice of j_0, both in terms of quality and the time of computation. However, j_0 should not be too large. For instance, $j_0 = 4$ was recommended by Bruce and Gao (1996). In our computation, we choose $j_0 = 3$. For an adaptive selection of wavelet basis and p_0, see Nason (2002). More crucial is the selection of the smoothing parameter. In Section 4, we considered the universal threshold rule. For comparison, we also select the smoothing parameter by two-fold cross-validation (odd/even indexes) for the l_1-penalized Poisson likelihood estimator and the minimax rule (Donoho and Johnstone 1994) for Waveshrink-Anscombe. We sample the signals at N equispaced locations with $N = 512$ and $N = 1{,}024$ to observe the decrease in the mean squared errors criterion. The sample size $N = 1{,}024$ allows comparison with the simulation results of Timmermann and Nowak (1999), although their estimator ought to perform well since it employs a larger set of wavelets, the translation invariant Haar wavelets, using the cycle spinning idea of Coifman and Donoho (1995). Nevertheless, our estimator is a good competitor and could be further improved by using cycle spinning, at the price of becoming computationally more expensive. Note also that the simulation of Timmermann and Nowak (1999) for the `bumps` function had a mistake that Italia De Feis helped us correct.

To visually assess the performance of the estimators, Figures 1 and 2 show, respectively, the `blocks` and `bumps` functions, sampled with Poisson noise, and estimated by the procedures considered. We observe that the l_1 penalized likelihood estimator improves over the Waveshrink-Anscombe estimator mainly in areas of low counts, especially for the `bumps` signal. Note that both estimators remain positive, even in regions of zero Poisson counts.

To fairly assess the performance of the estimators based on the mean squared errors criterion, we repeat the Monte Carlo experiment $m = 25{,}600/N$ times to have the same number of total observations for each N. For each run, we calculate the observed mean squared errors. Figures 3 and 4 present a boxplot summary, and Tables 2 and 3 report the mean of those observed mean squared errors, for peak $= 8$ and peak $= 128$, respectively. The results of the simulation show that both the method of estimation and the selection of the smoothing parameter are crucial to obtain a good estimation. Overall, the l_1-penalized likelihood estimator gives a better prediction than Waveshrink-Anscombe. This is partly due to regions of low counts (sometimes zero counts), where stabilizing the variance does not work. Two-fold cross-validation improves on the universal rule that tends to oversmooth when N is small. The computational cost is larger, however, because the l_1-penalized likelihood equation must be solved for several cross-validated λ's. We note that we encountered

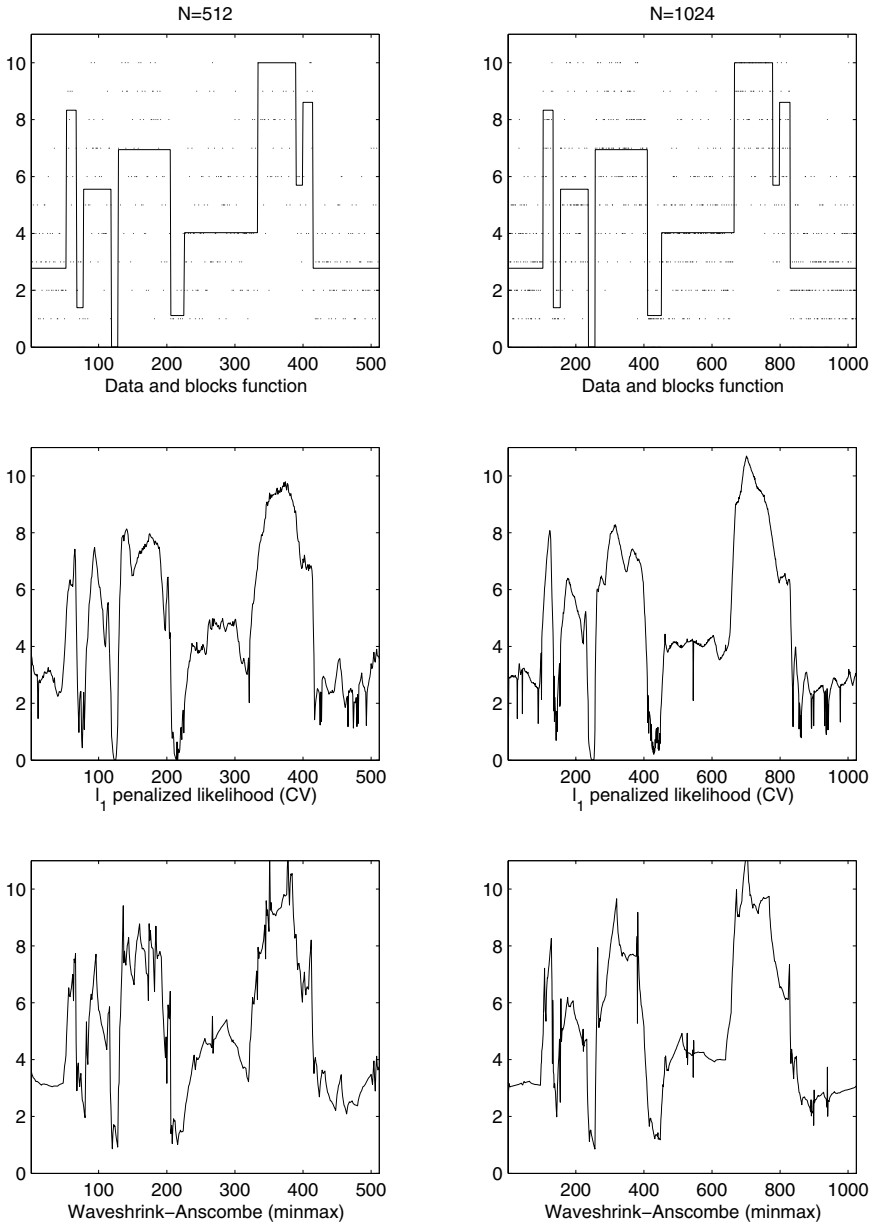

Figure 1. First line: Poisson data (dots) and `blocks` *function (line) for peak = 10. Second line: l_1 penalized likelihood estimate using two-fold cross-validation. Third line: Waveshrink-Anscombe estimate using minimax rule. Left: N = 512, right: N = 1,024.*

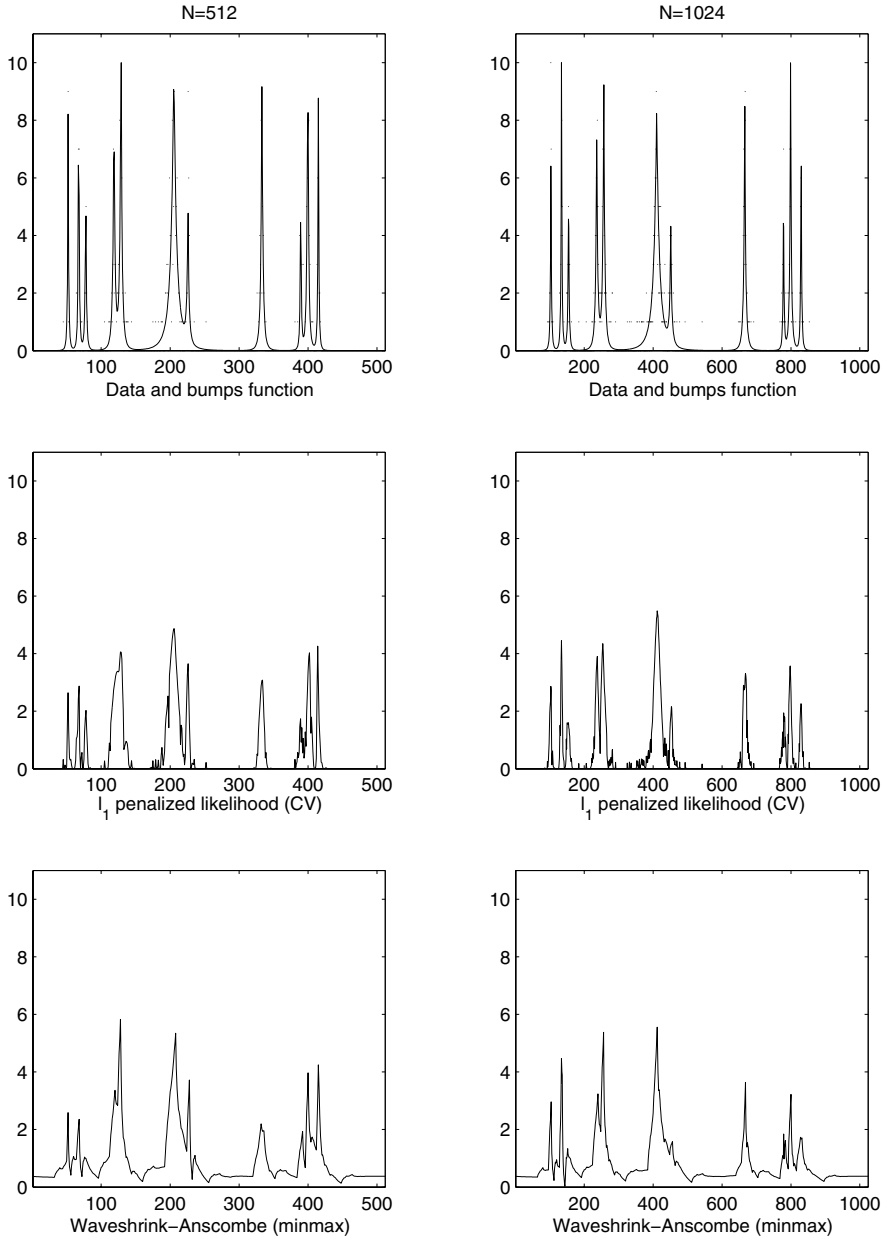

Figure 2. First line: Poisson data (dots) and `bumps` *function (line) for peak = 10. Second line: l_1 penalized likelihood estimate using two-fold cross-validation. Third line: Waveshrink-Anscombe estimate using minimax rule. Left: N = 512, right: N = 1,024.*

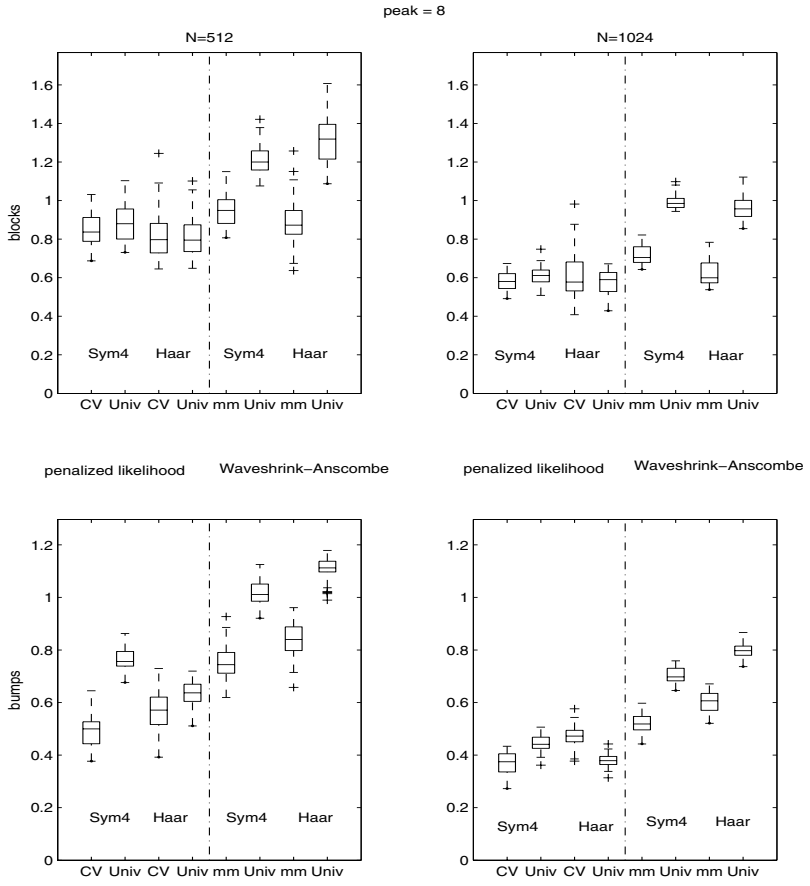

Figure 3. For peak = 8, boxplots of estimated mean squared errors with the symmlets of order 4 or the Haar wavelets. In each panel, left half: l_1-penalized likelihood estimate by cross-validation (CV) and universal rule (Univ); right half: Waveshrink-Anscombe estimate by minimax (mm) and universal rules.

difficulties in choosing the desired solution accuracy for the interior point algorithm over the range of λs tried during the cross validation search. The selection procedure by cross validation can moreover be unstable as we can see in Figure 3 for the `bumps` function and $N = 1{,}024$, by comparing "Univ" with "CV" with the Haar wavelets: the universal rule performs better than cross validation. Nevertheless, two-fold cross-validation overall improves the alleged oversmoothing effect of the universal rule.

6. APPLICATION

Gamma rays are the highest energy form of radiation in the electromagnetic spectrum. Gamma-ray astronomy looks at the sky in that high range of energy invisible to the naked eye. A special, satellite-based telescope, known as the energetic gamma-ray experiment telescope (EGRET), took measurements in the form of photon arrival times, directions, and

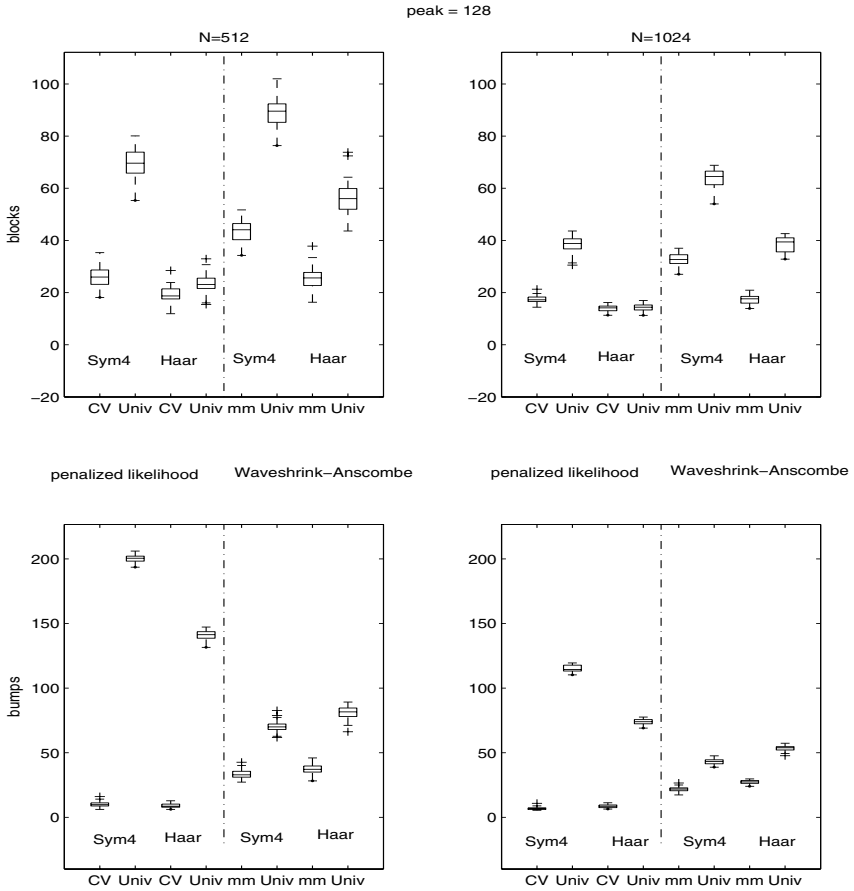

Figure 4. For peak = 128, boxplots of estimated mean squared errors with the symmlets of order 4 or the Haar wavelets. In each panel, left half: l_1-penalized likelihood estimate by cross-validation (CV) and universal rule (Univ); right half: Waveshrink-Anscombe estimate by minimax (mm) and universal rules.

energies. These observations can be binned in space to obtain the image in Figure 5, an image we treat as counts from an inhomogeneous Poisson process with unknown intensity μ_n, $n = (i, j)$, $i = 1, \ldots, 720$, $j = 1, \ldots, 360$. The coordinates used here are galactic latitude and longitude. The particularly 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 (no gamma-ray counts) contrasting with regions of high counts and a few bright clusters. It is therefore advantageous to be able to model the underlying Poisson parameters, not through the log-link, but directly as a positive linear combination of wavelets. We compare the denoising performance of our estimator to that of Waveshrink on the preprocessed Anscombe transformed data and the TIPSH estimator Kolaczyk (1999a). To avoid introducing some asymmetric artifacts we used the orthonormal least asymmetric compactly supported wavelets of order 8 (Daubechies 1994). The resulting estimates are plotted on a logarithmic scale ($\log(\mu + 1)$) in Figure 6. The main difference between the three denoised images can be observed at low counts in the range

Table 2. For Peak = 8, Average of the $m = 25{,}600/N$ Observed Mean Squared Errors (the standard error is of the order of the precision reported)

	l_1-penalized likelihood		*Waveshrink-Anscombe*		*BAYES*
	CV	*universal*	*minimax*	*universal*	
	Symmlet/Haar		*Symmlet/Haar*		*TI-Haar*
blocks					
$N = 512$	0.84/**0.81**	0.89/**0.81**	0.95/0.89	1.22/1.31	
$N = 1{,}024$	0.62/0.62	0.62/0.58	0.74/0.63	1.02/0.96	**0.36**
bumps					
$N = 512$	**0.50**/0.56	0.77/0.63	0.75/0.84	1.0/1.10	
$N = 1{,}024$	**0.37**/0.47	0.44/**0.38**	0.52/0.61	0.70/0.80	**0.36**†

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

$\mu \in (0, 5]$ on a logarithmic scale. The variance stabilizing transformation argument does not apply to regions of small Poisson counts and causes the blurring effect visible in the top right image of Figure 6. The TIPSH estimator specifically developed for Poisson data (bottom left) provides good denoising of the original Poisson image with almost no visible blurring effect. Our l_1 penalized likelihood approach applied to Poisson data (bottom right) gives a denoised image comparable to that obtained with the TIPSH approach. As pointed out by our astronomy colleague, our estimator provides less oversmoothing of small scale features in regions of low counts.

A possible model improvement for this application would consist in taking into account the point spread function of the measurement device that creates dependence between neighboring measurements. This amounts to simply adding the point spread matrix $\mathbf{S}$ to the model $\boldsymbol{\mu} = \tilde{\boldsymbol{\Phi}}\boldsymbol{\alpha} = \mathbf{S}\boldsymbol{\Phi}\boldsymbol{\alpha}$ in (2.5); the interior point algorithm of Section 3.2 applies by replacing $\boldsymbol{\Phi}$ by $\tilde{\boldsymbol{\Phi}}$, and the smoothing parameter can be selected by cross-validation in such a situation. Generalization of the TIPSH approach seems less straightforward. Another issue of interest for this application is to test whether some region of the image emits a significant amount of Gamma rays or if it represents a spurious feature. Getting confidence intervals for the estimate is a difficult problem, however, because the approach is nonparametric

Table 3. For Peak = 128, Average of the $m = 25{,}600/N$ Observed Mean Squared Errors (the standard error is of the order of the precision reported)

	l_1-penalized likelihood		*Waveshrink-Anscombe*		*BAYES*
	CV	*universal*	*minimax*	*universal*	
	Symmlet/Haar		*Symmlet/Haar*		*TI-Haar*
blocks					
$N = 512$	26/**19**	69/23	44/26	89/56	
$N = 1{,}024$	17/**14**	41/**14**	32/18	63/39	**14**
bumps					
$N = 512$	10/**8.9**	200/140	32 /37	70/81	
$N = 1{,}024$	**6.9**/8.5	110/74	21/27	42/53	10†

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

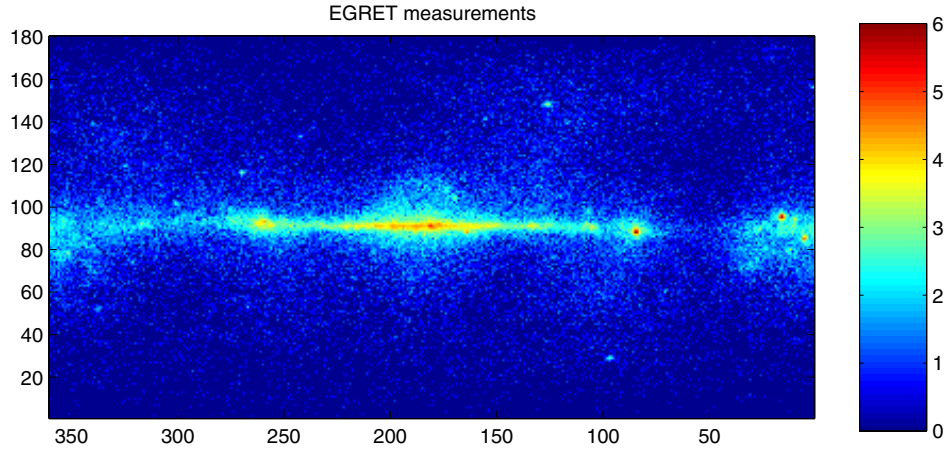

Figure 5. Gamma ray counts on a log scale (log(s + 1)) indexed by galactic latitude and longitude.

and involves a nondifferentiable penalty. The Bayesian paradigm offers an alternative and convenient approach to derive credible regions by sampling the posterior distribution (the penalized likelihood) by means of a Markov chain Monte Carlo technique for instance.

7. CONCLUSION

The wavelet-based estimator developed in this article is defined as the solution to an l_1 penalized likelihood problem (2.5); the solution is unique provided the log-likelihood function is strictly concave in the parameter of interest. Moreover, constraints can be added if the chosen link function does not map the parameter into the domain of the likelihood function. We propose a primal-dual log-barrier interior point algorithm to solve the corresponding convex programming problem (2.5) for a wide class of distributions. As far as the selection of the smoothing parameter is concerned, we define the universal smoothing parameter λ_N for a wide class of distributions, and derive the universal level-dependent smoothing parameter for the Poisson distribution. The universal parameter can be derived similarly for other distributions by appropriate developments in the theory of extremal processes. Using the universal parameter λ_N allows solving the penalized likelihood problem only once for that λ_N. However, λ_N is not data-driven and tends to oversmooth. The universal-based estimation can be improved by using two-fold cross-validation, at the price of solving several penalized likelihood problems until finding an approximate minimum $\hat{\lambda}$ to the cross validation function. Selection of the regularization parameter for ℓ_1 penalized likelihood estimators is a topic of current interest (Tibshirani 1995; Efron, Hastie, T., Johnstone, I., and Tibshirani 2003). It would also be of interest to see how the study of Barron, Birgé, and Massart (1999) on risk bounds for model selection via penalization relates to our l_1-penalized likelihood approach. Finally, deriving a confidence interval for the estimate is an interesting issue not addressed in this article.

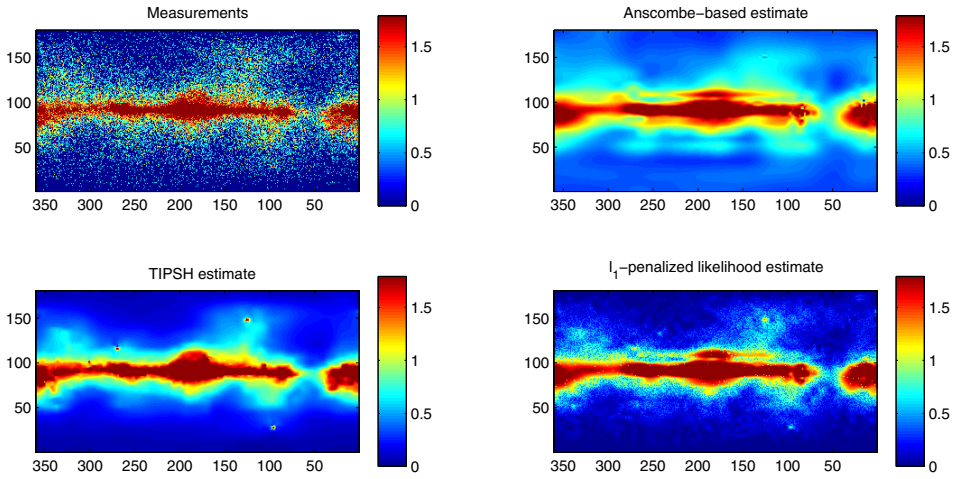

Figure 6. Original counts (top left), the Anscombe-based estimate (top right), the TIPSH estimate (bottom left) and the l_1-penalized likelihood estimate (bottom right). The l_1-penalized likelihood estimate shows the least oversmoothing on small scale features. The images focus on the small scale features by setting all pixel values above 5 to 5, and by using a logarithmic scale. The symmlets of order 8 with $j_0 = 3$ levels were used in each case.

8. SOFTWARE AVAILABILITY

The MatLab code used to generate the figures and the tables, as well as some MatLab functions provided by Dave Dixon for plotting on a log scale the EGRET image of Figure 5 are downloadable at http://statwww.epfl.ch/people/sardy/Tar/l1penlik-10-02.tar.gz.

The current version uses either the MatLab wavelet toolbox or the WaveLab toolbox downloadable at http://www-stat.stanford.edu/~wavelab/.

APPENDIXES

A. CONJUGATE GRADIENT SOLVER

Most of the computational effort at each interior point iteration is spent in computing the dual Newton direction $(\Delta\mathbf{y}, \Delta\boldsymbol{\eta})$ solution to the linear system (3.3). Because multiplication by $\mathbf{A}$, $\boldsymbol{\Phi}_0$ and $\mathbf{A}'$, $\boldsymbol{\Phi}_0'$ is typically fast (on the order of N, $N \log N$, or $N(\log N)^2$ operations depending on the set of wavelets used), the conjugate gradient method is attractive for this computation (Chen, Donoho, and Saunders 1999). The left-hand matrix of (3.3) is not positive definite, however, so we cannot apply the conjugate gradient method to it directly as for basis pursuit.

For an orthonormal wavelet matrix $\boldsymbol{\Phi}$, a reformulation of (3.3) exploits the fact that the columns of $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Psi}$ are mutually orthogonal and together they span $\Re^N$. This implies that $\boldsymbol{\Phi}_0'\Delta\mathbf{y} = \mathbf{r}_\eta$ if and only if $\Delta\mathbf{y} = \boldsymbol{\Phi}_0\mathbf{r}_\eta + \boldsymbol{\Psi}\Delta\mathbf{u}$ for some $\Delta\mathbf{u}$. Then, using this to substitute

for $\Delta\mathbf{y}$ in the first equation of (3.3) and left-multiplying this equation by $\mathbf{\Psi}'$, we obtain

$$\mathbf{\Psi}'\mathbf{Q}\mathbf{\Psi}\Delta\mathbf{u} = \mathbf{\Psi}'(\mathbf{r} - \mathbf{Q}\mathbf{\Phi}_0\mathbf{r}_\eta).$$

Because $\mathbf{\Psi}'\mathbf{\Psi} = I_{N-p_0}$ and $\mathbf{Z}^{-1}\mathbf{X} > 0$, the matrix $\mathbf{\Psi}'\mathbf{Q}\mathbf{\Psi}$ is symmetric and positive definite and so we can apply the conjugate gradient method to this system.

Because the signals are of finite length, border distortion arises. Not all schemes developed to handle borders have a corresponding orthonormal wavelet transform, so the reformulation proposed above is inapplicable. If so, we propose to solve (3.3) with the biconjugate-gradient method (Golub and Van Loan 1996, sec. 10.4.6) or with the conjugate gradient method applied to the least squares reformulation of (3.3), which consists in left-multiplying (3.3) by the transpose of its left-hand matrix (which is nonsingular). This yields the equivalent system

$$\begin{aligned}(\mathbf{Q}^2 + \mathbf{\Phi}_0\mathbf{\Phi}_0')\Delta\mathbf{y} - \mathbf{Q}\mathbf{\Phi}_0\Delta\boldsymbol{\eta} &= \mathbf{Q}\mathbf{r} + \mathbf{\Phi}_0\mathbf{r}_\eta,\\ -\mathbf{\Phi}_0'\mathbf{Q}\Delta\mathbf{y} + \mathbf{\Phi}_0'\mathbf{\Phi}_0\Delta\boldsymbol{\eta} &= -\mathbf{\Phi}_0'\mathbf{r}.\end{aligned}$$

The left-hand matrix is now symmetric positive definite, so we can apply the conjugate gradient method. However this new formulation has the drawback of squaring the condition number of the left-hand matrix.

In practice, the number of (bi-)conjugate gradient iterations required to solve the linear system accurately can become very large as $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \boldsymbol{\eta})$ approaches an optimal solution, thus degrading the performance of the interior point algorithm.

B. PROOF OF THE PROPOSITION

The universal parameter λ_N is defined by the asymptotic condition (4.1) on the event $\{\hat{\boldsymbol{\gamma}}_{\lambda_N} = \mathbf{0}\}$, where $\hat{\boldsymbol{\gamma}}_\lambda$ is the minimum l_1-penalized likelihood estimate for a given λ. The estimate $\hat{\boldsymbol{\gamma}}_\lambda$ is obtained as a solution to (2.5) and satisfies $\hat{\boldsymbol{\gamma}}_\lambda = \mathbf{0}$ provided the penalty parameter λ is large enough, in fact, if it is at least as large as a certain $\lambda_{\min}(\mathbf{s})$. For a given sample $\mathbf{s}$, the smallest possible value $\lambda_{\min}(\mathbf{s})$ can be obtained from the Karush–Kuhn–Tucker conditions for (2.5), which are

$$\begin{aligned}-\nabla_{\boldsymbol{\mu}} l(\boldsymbol{\mu}; \mathbf{s}) + \mathbf{y} &= \mathbf{0},\\ \mathbf{\Psi}'\mathbf{y} &\in [-\lambda\mathbf{1}, \lambda\mathbf{1}],\\ \mathbf{\Phi}_0'\mathbf{y} &= \mathbf{0},\\ \mathbf{\Phi}_0\boldsymbol{\beta} + \mathbf{\Psi}\boldsymbol{\gamma} &= \boldsymbol{\mu},\end{aligned} \tag{2.1}$$

for some $\boldsymbol{\mu} \in \mathbf{C}$ and $\boldsymbol{\beta}$. The smallest λ satisfying the KKT conditions for $\boldsymbol{\gamma} = \mathbf{0}$ in (2.1) is therefore

$$\lambda_{\min}(\mathbf{s}) = \max|\mathbf{\Psi}'\mathbf{y}|.$$

Hence the universal parameter is the smallest $\lambda(N)$ such that

$$\mathrm{P}\left(\hat{\boldsymbol{\gamma}}_{\lambda(N)} = \mathbf{0}\right) = \mathrm{P}\left(\max|\mathbf{\Psi}'\mathbf{y}| \le \lambda(N)\right) \overset{N\to\infty}{\longrightarrow} 1,$$

where $\mathbf{y}$ satisfies the KKT conditions (2.1) for $\boldsymbol{\gamma} = \mathbf{0}$ and for some $\boldsymbol{\mu} \in \mathbf{C}$ and $\boldsymbol{\beta}$. Interestingly these conditions define the maximum likelihood estimate $\boldsymbol{\mu}$ in the range of $\boldsymbol{\Phi}_0$.

C. UNIVERSAL PARAMETER FOR POISSON

Although our Proposition 1 (p. 10) applied to the Poisson distribution (4.3) involves a sequence of independent Poisson variables, an adaptation of these results to Poisson process helps to determine the appropriate asymptotic behavior. The connection between the discrete Poisson model and a Poisson process can be made by letting $S_N(t)$ be an inhomogeneous Poisson process with cumulative intensity function $N\Lambda_N(t)$ defined by

$$\Lambda_N(t) = \frac{1}{N} \sum_{i=1}^{[Nt]} \mu(x_i),$$

for any N; this is a Riemann sum approximating $\int_0^t \mu(s)ds$, $t \in [0, 1]$. Hence the independent variables $s_n = S_N(]x_{n-1}, x_n]) \sim \text{Poisson}(\mu(x_n))$, $n = 1, \ldots, N$, and the observable sequence $\mathbf{s}$ and an observed sample path of the $S_N(\cdot)$ process merge together as $N \to \infty$.

The intensity function of the $S_N(\cdot)$ process is $N\mu(\cdot)$, where $\mu(\cdot)$ has an orthonormal wavelet expansion (2.3) with a maximum resolution controlled level $J' = \log_2(N/\log N) - 1$ suggested by the consideration that follows. The log-likelihood function of the counting process $S_N(\cdot)$ with respect to an homogeneous Poisson process $[0, 1]$ with unit intensity is

$$l(\mu(\cdot)) = \int_0^1 \log\{N\mu(s)\} dS_N(s) - N \int_0^1 \mu(s)ds.$$

If we may interchange the order of differentiation and integration, the vector $\mathbf{U}(\boldsymbol{\alpha})(t)$ of the score processes $U(\alpha_n)(t)$ is

$$U(\alpha_n)(t) = \int_0^t \frac{\partial}{\partial \alpha_n} \log\{N\mu(s)\} dS_N(s) - N \int_0^t \frac{\partial}{\partial \alpha_n} \mu(s)ds,$$

and $\frac{1}{N} U(\alpha_n)(1)$ is the continuous-time analogue of the nth row of $\boldsymbol{\Phi}'(\mathbf{z} - \mathbf{1})$. Hence the solution to Karush–Kuhn–Tucker conditions (4.2) is the solution to $U(\beta_n)(1) = 0$, $n = 0, \ldots, 2^{j_0} - 1$, when $\boldsymbol{\mu} = \boldsymbol{\Phi}_0 \boldsymbol{\beta}_0$ lies in the finite-dimensional approximating space V_{j_0} (i.e., $\boldsymbol{\gamma} = \mathbf{0}$). Moreover if $\mu_0(t) > 0$ on $[0, 1]$ and if $2^{J'} = O(N/\log N)$, then conditions VI.1.1 (A), (B), and (C) of Andersen, Borgan, Gill, and Keiding (1993) are satisfied, so that $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ enjoys the usual properties of a maximum likelihood estimator. Because $2^{J'}/N \to 0$, the estimated score process components $\frac{1}{\sqrt{N}} U(\hat{\boldsymbol{\gamma}}_n)(t)$ with $n = (j, k)$ and $j_0 \le j \le J'$ behave like local square integrable martingales with predictable variation

$$\frac{1}{N} \langle U(\hat{\gamma}_n) \rangle(t) = \int_0^t \psi_{j,k}^2(s)/\hat{\mu}(s)ds.$$

Moreover, using equations (6.1.17) and (6.1.18) of Andersen, Borgan, Gill, and Keiding (1993), the vector $\frac{1}{\sqrt{N}} \mathbf{U}(\hat{\gamma}_n)(1)$ is asymptotically normal with zero mean and covariance

matrix Σ with entries

$$(\Sigma)_{(j,k),(j',k')} = \int_0^1 \psi_{j,k}(s)\psi_{j',k'}(s)/\mu_0(s)ds.$$

Assuming that $1/\mu_0(s)$ is a square integrable function, the diagonal elements of the covariance matrix Σ are bounded independently of k for each resolution level j by

$$\int_0^1 \psi_{j,k}^2(s)/\mu_0(s)ds \le 2^j \left\{ \max_{u\in[0,1]} \{\psi^2(u)\} \int_0^1 1/\mu_0(s)ds \right\} =: 2^j M^2(\mu_0, \mathbf{\Psi}).$$

Using the arguments of Johnstone and Silverman (1997) for correlated Gaussian noise, we have that

$$\mathrm{P}\left(\limsup_{n=(j,k)\ge 2^{j_0}} \frac{1}{\sqrt{N}} |U(\hat{\gamma}_n)(1)| > M(\mu_0, \mathbf{\Psi}) 2^{j/2} \sqrt{2\log N} \right) \overset{N\to\infty}{\longrightarrow} 0.$$

Because the behavior of the rows of $\mathbf{\Phi}'(\mathbf{z}-\mathbf{1})$ at the optimal solution is similar to that of the rows of $\frac{1}{N}U(\hat{\alpha}_n)(1)$, the appropriate universal regularization parameter is level dependent and equal to

$$\lambda_{N,j} = M(\mu_0, \mathbf{\Psi}) 2^{j/2} \sqrt{2\log N}/\sqrt{N}.$$

ACKNOWLEDGMENTS

We thank Dave Dixon and Eric Kolaczyk for providing help with the EGRET image and the TIPSH estimator, and Anthony Davison for discussions about penalized likelihood and for his advice on the organization of the article. We also would like to thank the associate editor and two anonymous referees for their careful and thoughtful reviews. This work was partially supported by the Swiss National Science Foundation.

[Received January 2002. Revised January 2003.]

REFERENCES

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Anscombe, F. J. (1948), "The Transformation of Poisson, Binomial and Negative-Binomial Data," *Biometrika*, 35, 246–254.

Antoniadis, A., Besbeas, P., and Sapatinas, T. (2001), "Wavelet Shrinkage for Natural Exponential Families With Cubic Variance Functions," *Sankhya: The Indian Journal of Statistics, Special Issue on Wavelets*, 63, 1–19.

Antoniadis, A., and Sapatinas, T. (2001), "Wavelet Shrinkage for Natural Exponential Families With Quadratic Variance Functions," *Biometrika*, 88, 805–820.

Averkamp, R., and Houdré, C. (1996), "Wavelet Thresholding for Non (necessarily) Gaussian Noise: A Preliminary Report," technical report, Georgia Institute of Technology, Atlanta.

Barron, A., Birgé, L., and Massart, P. (1999), "Risk Bounds for Model Selection via Penalization," *Probability Theory and Related Fields*, 113, 301–413.

Bruce, A. G., and Gao, H.-Y. (1996), "Understanding WaveShrink: Variance and Bias Estimation," *Biometrika*, 83, 727–745.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1999), "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, 20, 33–61.

Coifman, R. R., and Donoho, D. L. (1995), "Translation–Invariant De–noising," in *Wavelets and Statistics*, eds. A. Antoniadis and G. Oppenheim, New York: Springer-Verlag, pp. 125–150.

Daubechies, I. (1994), *Ten Lectures on Wavelets*, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation via Wavelet Shrinkage," *Biometrika*, 81, 425–455.

——— (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 432, 1200–1224.

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

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2003), "Least Angle Regression" (with discussion), *The Annals of Statistics*, to appear.

Fridman, J., and Manolakos, E. S. (1997), "Discrete Wavelet Transform: Data Dependence Analysis and Synthesis of Distributed Memory and Control Array Architectures," *IEEE Transactions on Signal Processing*, 45, 1291–1308.

Golub, G. H., and Van Loan, C. F. (1996), *Matrix Computations*, Baltimore, MD: Johns Hopkins University Press.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*, London: Chapman and Hall.

Johnstone, I. M., and Silverman, B. W. (1997), "Wavelet Threshold Estimators for Data With Correlated Noise," *Journal of the Royal Statistical Sociey*, Ser. B, 59, 319–351.

Kolaczyk, E. D. (1999a), "Wavelet Shrinkage Estimation of Certain Poisson Intensity Signals Using Corrected Thresholds," *Statistica Sinica*, 9, 119–135.

——— (1999b), "Bayesian Multiscale Models for Poisson Processes," *Journal of the American Statistical Association*, 94, 920–933.

Mallat, S. G. (1989), "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–693.

Mallat, S. G., and Zhang, Z. (1993), "Matching Pursuit in a Time-Frequency Dictionary," *IEEE Transactions on Signal Processing*, 41, 3397–3415.

Nason, G. P. (1995), "Wavelet Function Estimation using Cross-Validation," in *Wavelets and Statistics*, eds. A. Antoniadis and G. Oppenheim, New York: Springer-Verlag, pp. 261–280.

——— (2002), "Choice of Wavelet Smoothness, Primary Resolution and Threshold in Wavelet Shrinkage," *Statistics and Computing*, 12, 219–227.

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Ser. A, 135, 370–384.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J. Jr. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 96–103.

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

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., Tseng, P., and Bruce, A. G. (2001), "Robust Wavelet Denoising," *IEEE Transactions on Signal Processing*, 49, 1146–1152.

Tibshirani, R. (1995), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Sociey*, Ser. 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.

Villalobos, M., and Wahba, G. (1987), "Inequality-Constrained Multivariate Smoothing Splines With Application to the Estimation of Posterior Probabilities," *Journal of the American Statistical Association*, 82, 239–248.