Biometrika (2001), **88,** 3, *pp.* 805–820

Printed in Great Britain

Wavelet shrinkage for natural exponential families with quadratic variance functions

BY ANESTIS ANTONIADIS

Laboratoire IMAG-LMC, University Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France

anestis.antoniadis@imag.fr

AND THEOFANIS SAPATINAS

Department of Mathematics and Statistics, University of Cyprus, P.O. Box 20537, CY 1678 Nicosia, Cyprus

t.sapatinas@ucy.ac.cy

SUMMARY

We propose a wavelet shrinkage methodology for univariate natural exponential families with quadratic variance functions, covering the Gaussian, Poisson, gamma, binomial, negative binomial and generalised hyperbolic secant distributions. Simulation studies for Poisson and binomial data are used to illustrate the usefulness of the proposed methodology, and comparisons are made with other methods available in the literature. We also present applications to datasets arising from high-energy astrophysics and from epidemiology.

Some key words: Crossvalidation mean squared error; Diagonal shrinkage estimation; Modulation estimator; Natural exponential family; Nonparametric regression; Smoothing; Wavelet shrinkage estimation.

1. INTRODUCTION

Wavelet shrinkage estimation has been found to be a powerful tool for the nonparametric estimation of spatially variable phenomena. Most work in this area to date has concentrated primarily on the use of wavelet shrinkage techniques in the nonparametric regression context where the data are modelled as observations of a signal corrupted with additive Gaussian noise. The usual approach is to expand the noisy data in wavelet series, to extract the significant wavelet coefficients by thresholding, and then to invert the wavelet transform of the denoised coefficients. Donoho & Johnstone (1994, 1995, 1998) and Donoho et al. (1995) showed that wavelet shrinkage estimation, with a properly chosen threshold, has various important optimality properties. For recent surveys of relevant research we refer to Antoniadis (1997), Vidakovic (1999) and Abramovich et al. (2000).

If the data are counts, the standard approach is first to pre-process the data using a normalising and variance-stabilising transformation, such as that proposed by Anscombe (1948), and then to apply usual wavelet shrinkage techniques; see Donoho (1993) for an application involving Poisson data. However, this approach has been criticised for frequent oversmoothing or attenuation of fine detail structure in the underlying signal or image, especially in situations involving very low levels of counts.

It is therefore important to develop wavelet shrinkage estimation techniques by directly considering the original, untransformed count data. In the context of data of a Poisson nature, Kolaczyk (1997, 1999a) concentrated on direct applications of the wavelet shrinkage methodology, through appropriate modifications, whilst Kolaczyk (1999b) and Timmermann & Nowak (1999) developed multiscale models using recursive dyadic partitions within a Bayesian framework. For Bernoulli data, Antoniadis & Leblanc (2000) proposed a wavelet shrinkage methodology based on diagonal linear shrinkers.

It is evident that a wavelet shrinkage methodology that could be successfully applied to various types of data would be very useful. Towards this end, we propose a wavelet shrinkage methodology for univariate natural exponential families with quadratic variance functions. The Gaussian, Poisson, gamma, binomial, negative binomial and generalised hyperbolic secant distributions are the only natural exponential families with quadratic variance functions, that is with variance that is a constant or linear or quadratic function of the mean (Morris, 1982, 1983).

In § 2, we briefly review the relevant material on wavelets and the exponential family models. In § 3, we propose a wavelet shrinkage methodology for these models borrowing ideas from modulation estimators that were originally developed for Gaussian data by Beran & Dümbgen (1998). In § 4, simulation studies for Poisson and binomial data are used to illustrate the usefulness of the proposed methodology, and comparisons are made with other methods available in the literature. We also present applications to datasets arising from high-energy astrophysics and epidemiology. The computational algorithms related to wavelet analysis were performed using the Matlab toolbox WaveLab that is freely available from http://www-stat.stanford.edu/software/software.html. The entire study was carried out using the Matlab programming environment.

2. BACKGROUND MATERIAL

2·1. *The wavelet series expansion*

We assume that we are working within an orthonormal basis generated by dilation and translation of a compactly supported scaling function ϕ and a mother wavelet ψ associated with an r-regular multiresolution analysis of $L^2([0,1])$. For simplicity in exposition, we work with periodised wavelet bases on $[0,1]$ (Mallat, 1999, § 7.5.1), letting

$$\phi_{jk}^{\mathrm{per}}(t)=\sum_{l\in\mathbb{Z}}\phi_{jk}(t-l),\quad \psi_{jk}^{\mathrm{per}}(t)=\sum_{l\in\mathbb{Z}}\psi_{jk}(t-l)\quad (t\in[0,1]),$$

where

$$\phi_{jk}(t)=2^{j/2}\phi(2^jt-k),\quad \psi_{jk}(t)=2^{j/2}\psi(2^jt-k).$$

For any $j_0\geqslant 0$, the collection

$$\{\phi_{j_0k}^{\mathrm{per}},\,k=0,1,\ldots,2^{j_0}-1;\,\psi_{j_0k}^{\mathrm{per}},\,j\geqslant j_0\geqslant 0,\,k=0,1,\ldots,2^j-1\}$$

is then an orthonormal basis of $L^2([0,1])$. Despite the poor behaviour of periodic wavelets near the boundaries, where they create high-amplitude wavelet coefficients, they are commonly used because the numerical implementation is particularly simple. Also, as Johnstone (1994) has pointed out, this computational simplification affects only a fixed number of wavelet coefficients at each resolution level and does not affect the qualitative phenomena that we wish to present.

The idea underlying such an approach is to express any function $f\in L^2([0,1])$ in the

form

$$f(t) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k}^{\mathrm{per}}(t) + \sum_{j \geqslant j_0} \sum_{k=0}^{2^{j_0}-1} \beta_{jk}\psi_{jk}^{\mathrm{per}}(t) \quad (j_0 \geqslant 0, t \in [0, 1]),$$

where

$$\alpha_{j_0 k} = \langle f, \phi_{j_0 k}^{\mathrm{per}} \rangle = \int_0^1 f(t)\phi_{j_0 k}^{\mathrm{per}}(t)\, dt \quad (j_0 \geqslant 0, k = 0, 1, \ldots, 2^{j_0} - 1),$$

$$\beta_{jk} = \langle f, \psi_{jk}^{\mathrm{per}} \rangle = \int_0^1 f(t)\psi_{jk}^{\mathrm{per}}(t)\, dt \quad (j \geqslant j_0 \geqslant 0, k = 0, 1, \ldots, 2^j - 1).$$

In statistical settings we are more usually concerned with discretely sampled, rather than continuous, functions, and therefore with the discrete wavelet transform. Given a vector of function values $f = (f(t_1), \ldots, f(t_n))'$ at equally spaced points t_i, the discrete wavelet transform of f is given by $d = Wf$, where d is an $n \times 1$ vector comprising both discrete scaling coefficients, $c_{j_0 k}$, and discrete wavelet coefficients, d_{jk}, and W is an orthogonal $n \times n$ matrix associated with the chosen orthonormal wavelet basis. The $c_{j_0 k}$ and d_{jk} are related to their continuous counterparts $\alpha_{j_0 k}$ and β_{jk}, with an approximation error of order n^{-1}, via the relationship $\alpha_{j_0 k} \doteqdot n^{-\frac{1}{2}} c_{j_0 k}$ and $\beta_{jk} \doteqdot n^{-\frac{1}{2}} d_{jk}$. The factor $n^{-\frac{1}{2}}$ arises because of the difference between the continuous and discrete orthonormality conditions. This $n^{-\frac{1}{2}}$ factor is unfortunate but both the definition of the discrete wavelet transform and the wavelet coefficients are now fixed by convention, from which originates the different notation used to distinguish between the discrete wavelet coefficients and their continuous counterparts. Note that, because of orthogonality of W, the inverse discrete wavelet transform is simply given by $f = W^{\mathrm{T}} d$, where W^{T} denotes the transpose of W.

If $n = 2^J$ for some positive integer J, the discrete wavelet transform and the inverse discrete wavelet transform may be performed through a computationally fast algorithm developed by Mallat (1989) that requires only $O(n)$ operations. In this case, for a given $j_0 \geqslant 0$ and under periodic boundary conditions, the discrete wavelet transform of f results in an n-dimensional vector d comprising both discrete scaling coefficients $c_{j_0 k}$, for $k = 0, \ldots, 2^{j_0} - 1$, and discrete wavelet coefficients d_{jk}, for $j = j_0, \ldots, J - 1$ and $k = 0, \ldots, 2^j - 1$.

For detailed expositions of the mathematical aspects of wavelets we refer to Meyer (1992), Daubechies (1992) and Mallat (1999).

2·2. *Natural exponential families with quadratic variance functions*

This section briefly overviews some material that we shall use in subsequent sections. For a more detailed account we refer to Morris (1982, 1983).

A parametric family of distributions with natural parameter space $\Theta \subset \mathbb{R} = (-\infty, \infty)$ is a univariate natural exponential family if random variables X governed by these distributions satisfy

$$P_\theta(X \in A) = \int_A \exp\{x\theta - \psi(\theta)\}\, dF(x),$$

where $A \subset \mathbb{R}$, F is a Stieltjes measure on $\mathbb{R}$ not depending on $\theta \in \Theta$, the natural parameters, and $\psi(\theta)$ is the cumulant generating function. The random variable X is the natural observation. Exponential families that are not natural exponential families are nonlinear

transformations of natural exponential families. The natural observation X has mean and variance

$$\mu = \psi'(\theta) = E_\theta(X) = \int x\, dF_\theta(x), \quad V(\mu) = \psi''(\theta) = \mathrm{var}_\theta(X) = \int (x-\mu)^2\, dF_\theta(x)$$

and cumulants $C_r(\mu) = \psi^{(r)}(\theta)$ $(r = 1, 2, \ldots)$. The function $V(\mu)$ on its domain $\Omega \equiv \psi'(\Theta)$ is called the variance function of the natural exponential family and characterises the family, but no particular member of the family. Consider now natural exponential families for which

$$V(\mu) = v_0 + v_1\mu + v_2\mu^2, \tag{1}$$

where v_0, v_1 and v_2 are real-valued constants. We shall write

$$X \sim \mathrm{NQ}\{\mu, V(\mu)\}$$

to denote a random variable which follows a natural exponential family with mean μ and quadratic variance function $V(\mu)$ given by (1). It can be shown that exactly six basic types of such distributions exist:

(i) the Gaussian distribution, $N(\mu, \sigma^2)$ with $V(\mu) = \sigma^2$, for $-\infty < \mu < \infty$ and $\sigma^2 > 0$;
(ii) the Poisson distribution, $\mathrm{Po}(\lambda)$ with $\mu = \lambda$ and $V(\mu) = \mu$, for $\lambda > 0$;
(iii) the gamma distribution, $\mathrm{Ga}(r, \lambda)$ with $\mu = r\lambda$ and $V(\mu) = r\lambda^2 = \mu^2/r$, for $r > 0$ and $\lambda > 0$;
(iv) the binomial distribution, $\mathrm{Bi}(r, p)$ with $\mu = rp$ and $V(\mu) = rpq = -\mu^2/r + \mu$, for $r = 1, 2, \ldots, 0 < p < 1$ and $q = 1 - p$;
(v) the negative binomial distribution, $\mathrm{NeBi}(r, p)$ with $\mu = rp/q$ and $V(\mu) = rp/q^2 = \mu^2/r + \mu$, for $r > 0$, $0 < p < 1$ and $q = 1 - p$;
(vi) the generalised hyperbolic secant distribution, which is a continuous distribution with support on $(-\infty, \infty)$ and $\psi(\theta) = -\log\cos\theta$ (Morris, 1982, §§ 4, 5), with $V(\mu) = \mu^2/r + r$, for $r > 0$.

Remark 1. The six basic types of distribution mentioned above can be extended by convolutions, and location and scale changes, all of which preserve both the natural exponential family and the quadratic variance function properties (Morris, 1982, § 3). In other words, let $X_1, \ldots, X_n$ be independent $\mathrm{NQ}\{\mu, V(\mu)\}$ random variables and define

$$Y = \frac{1}{c}\sum_{i=1}^{n}(X_i - b) \quad (-\infty < b < \infty, c > 0).$$

Then $Y \sim \mathrm{NQ}\{\mu^*, V^*(\mu^*)\}$, where

$$\mu^* = E(Y) = \frac{n}{c}(\mu - b), \quad V^*(\mu^*) = \mathrm{var}(Y) = v_0^* + v_1^*\mu^* + v_2^*(\mu^*)^2,$$

with

$$v_0^* = \frac{n}{c^2}V(b), \quad v_1^* = \frac{1}{c}V'(b), \quad v_2^* = \frac{v_2}{n}.$$

These are the key properties for developing our wavelet shrinkage methodology in § 3.

3. Wavelet shrinkage for natural exponential families with quadratic variance functions

3·1. *Introduction*

Let $Y = Y_n = \{Y(t)\}_{t \in T}$ be a random function observed on the set $T = T_n = \{1, \ldots, n\}$. The components $Y(t)$ of Y are assumed to be independent random variables such that

$$Y(t) \sim \text{NQ}[\mu(t), V\{\mu(t)\}] \quad (t \in T), \tag{2}$$

where

$$V\{\mu(t)\} = v_0 + v_1 \mu(t) + v_2 \mu^2(t) \quad (t \in T)$$

for real-valued constants v_0, v_1 and v_2. Working with functions on T rather than vectors in $\mathbb{R}^n$ is convenient for our purposes. We assume hereafter that the mean vector $\mu = \{\mu(t)\}_{t \in T}$ consists of sampled observations at equally spaced points on $[0, 1]$ of an unknown but otherwise smooth function μ that we wish to recover from the data $Y = \{Y(t)\}_{t \in T}$ without assuming any particular parametric form.

Define $\theta = W\mu$ to be the vector of wavelet coefficients corresponding to μ, where W is the $n \times n$ orthogonal matrix associated with the discrete wavelet transform discussed in § 2·1. One reason for using such a transform is that only a few wavelet coefficients are needed to obtain a good approximation of μ, since μ is supposed to be smooth. Squared-error loss is widely used for studying the quality of nonparametric function estimators. In the above context, squared-error loss can be formulated either in the wavelet domain or in the observation domain. By the orthogonality of the wavelet transform, these two quadratic losses are equivalent up to a factor of $n^{\frac{1}{2}}$. In the wavelet domain the loss of any estimator $\check{\theta}$, linear or nonlinear, which depends on $\hat{\theta} = WY$, is defined to be

$$L(\check{\theta}, \theta) = \text{ave}\{(\check{\theta} - \theta)^2\},$$

while the loss in the observation domain is defined to be

$$L(\check{Y}, \mu) = \text{ave}\{(\check{Y} - \mu)^2\} = \text{ave}\{(W^{\text{T}}\check{\theta} - W^{\text{T}}\theta)^2\} = L(W^{\text{T}}\check{\theta}, W^{\text{T}}\theta),$$

where

$$\check{Y} = W^{\text{T}}\check{\theta}, \quad \text{ave}(g) = \frac{1}{n}\sum_{t \in T} g(t),$$

for any $g \in \mathbb{R}^{\text{T}}$, the space of real-valued functions defined on T. The corresponding risk of $\check{\theta}$ is

$$\rho(\check{\theta}, \theta) = E\{L(\check{\theta}, \theta)\} \tag{3}$$

with an associated risk for $\check{Y} = W^{\text{T}}\check{\theta}$ given by

$$\rho(\check{Y}, \mu) = E\{L(\check{Y}, \mu)\} = E\{L(W^{\text{T}}\check{\theta}, \mu)\} = \rho(W^{\text{T}}\check{\theta}, \mu). \tag{4}$$

When the observed function Y is Gaussian with $\text{var}(Y) = \text{var}\{Y(t)\} = \sigma^2$ for all $t \in T$, Donoho & Johnstone (1994) have developed ideal spatial adaptation in the wavelet domain by considering diagonal linear shrinkers $\hat{\theta}_h = HWY$, where $H = \text{diag}(h)$ is the diagonal matrix of order n and $h: T \to [0, 1]$. The ideal coefficients, that yield an ideal risk, are then provided by the oracle estimator of h given by $\tilde{h} = \theta^2/(\theta^2 + \sigma^2)$, the division being componentwise. The oracle estimator cannot be attained in practice since it requires knowledge of the unknown θ. The nonlinear wavelet shrinkage estimators developed by

Donoho & Johnstone (1994, 1995, 1998), however, are surprisingly close to the oracle estimator in case of sparse signals. A simple procedure for estimating the oracle would be to compute unbiased estimators of σ^2 and θ from the data, and then use these estimators to obtain an estimator of $\tilde{h}$. However, such an estimator involves an estimator of the unknown θ and, in general, is less efficient than the one which minimises the estimated risk.

Indeed, Beran & Dümbgen (1998) have shown that one can construct estimators of θ that are asymptotically minimax optimal over a variety of ellipsoids in the parameter space and which take the form $\hat{\theta}_{\hat{h}} = \hat{H}WY$, where $\hat{H} = \mathrm{diag}(\hat{h})$. The function $\hat{h}: T \to [0, 1]$ depends on $\hat{\theta} = WY$ and is chosen to minimise the estimated risk of the linear estimator $\check{\theta} = \hat{\theta}_h = HWY$ over all functions h in a class $\mathscr{H} \subset [0, 1]^T$, the set of functions defined on T and taking values in $[0, 1]$. Such estimators are, by construction, nonlinear and shrink each coordinate towards zero, different coordinates being possibly treated differently. Hereafter, we adopt the terminology of Beran & Dümbgen (1998), so that each function h in a class $\mathscr{H} \subset [0, 1]^T$ will be called a modulator and the estimators $\hat{\theta}_{\hat{h}} = \hat{H}WY$ will be referred to as the modulation estimators.

3·2. *Estimating the best modulator using the crossvalidation mean squared error*

Since the Gaussian distribution is a particular member of our general family, our estimation procedure for estimating θ for the general model will be based on ideas similar to those in § 3·1. Using the division property of natural exponential families with quadratic variance functions, see Remark 1, and a crossvalidation approach similar to that developed by Nowak (1997) in the case where one has more than one independent observations of an unknown signal, we first construct a suitably consistent estimator $\hat{\rho}$ of the risk $\rho(\hat{\theta}_h, \theta)$, with $\hat{\theta}_h = HWY$, and we estimate θ by the modulation estimator $\hat{\theta}_{\hat{h}} = \hat{H}WY$, where $\hat{h}$ is any function in $\mathscr{H}$ that minimises $\hat{\rho}$. By construction, $\hat{\theta}_{\hat{h}}$ is nonlinear.

Let $Z_1, \ldots, Z_p$ be independent random variables with $Z_k = \{Z_k(t)\}_{t \in T}$ such that

$$Z_k(t) \sim \mathrm{NQ}[v(t), R\{v(t)\}] \quad (k = 1, \ldots, p, t \in T),$$

where

$$v(t) = \frac{1}{p}\mu(t), \quad R\{v(t)\} = \frac{1}{p}v_0 + v_1 v(t) + p v_2 v^2(t)$$

for real-valued constants v_0, v_1 and v_2. Then, using the results in § 2·2, we have

$$Y = \sum_{k=1}^{p} Z_k, \quad \bar{Z} = \frac{1}{p}\sum_{k=1}^{p} Z_k = \frac{1}{p}Y.$$

Suppose that instead of observing data Y we observe the pseudo-sample $Z_1, \ldots, Z_p$. Consider now an estimation procedure based on the pseudo-sample producing modulation estimators of the form $\hat{\theta}_h$ depending on a modulator $h = \{h(t)\}_{t \in T}$. Applying this procedure to the pseudo-sample $Z_1, \ldots, Z_p$, without using the jth element, and defining $H = \mathrm{diag}(h)$, leads to a modulation estimator

$$\hat{\theta}_h^{(j)} = HWZ^{(j)},$$

where

$$Z^{(j)} = \frac{1}{p-1}\sum_{\substack{k=1 \\ k \neq j}}^{p} Z_k,$$

with corresponding signal estimator

$$\hat{Z}^{(j)} = W^{\mathrm{T}}\hat{\theta}_h^{(j)} = W^{\mathrm{T}}HWZ^{(j)}.$$

To estimate the best modulator, we first construct a suitable consistent estimator of the risk $\rho(\hat{\theta}_h, \theta)$ based on the crossvalidation mean squared error (Eubank, 1999, p. 43) or its equivalent form, the prediction sum of squares. Let

$$P(h) = \sum_{k=1}^{p} r^{(k)}(h), \tag{5}$$

where $r^{(k)}(h) = \|HWZ^{(k)} - WZ_k\|^2$. For simplicity, we will use hereafter the notation $\hat{\theta} = WY = pW\bar{Z}$, $\hat{\theta}_k = WZ_k$ and $\sigma^2 = \mathrm{var}(\hat{\theta})$. For $t \in T$, let

$$\sigma^2(t) = \frac{p}{p-1}\sum_{k=1}^{p}\left\{\hat{\theta}_k(t) - \frac{1}{p}\hat{\theta}(t)\right\}^2 = \frac{p}{p-1}\sum_{k=1}^{p}\{W(Z_k - \bar{Z})(t)\}^2.$$

Some algebra shows that expression (5) may be written as

$$P(h) = \frac{p}{p-1}\sum_{t\in T}\left[\hat{\sigma}^2(t) - \frac{2}{p}\{1-h(t)\}\hat{\sigma}^2(t) + \frac{1}{p^2}\{1-h(t)\}^2\{\hat{\sigma}^2(t) + (p-1)\hat{\theta}^2(t)\}\right]. \tag{6}$$

As a risk estimator for the risk $\rho(\hat{\theta}_h, \theta)$, the prediction sum of squares criterion leads to an estimator that is biased upwards. Indeed, since the components $Y(t)$ $(t \in T)$ of Y are independent $\mathrm{NQ}[\mu(t), V\{\mu(t)\}]$ and W is orthonormal, it is not difficult to see that, for all $t \in T$, we have

$$E\{\hat{\theta}(t)\} = \theta(t), \quad E\{\hat{\sigma}^2(t)\} = \sum_{l=1}^{n} w_{t,l}^2 V\{\mu(l)\} = \sigma^2(t). \tag{7}$$

Moreover, using (6) and (7) we have

$$\begin{aligned} E\{P(h)\} &= \sum_{t\in T}\left[\frac{1}{p}\{1-h(t)\}^2\theta^2(t) + \frac{1}{p-1}h^2(t)\sigma^2(t) + \sigma^2(t)\right] \\ &= \frac{n}{p}\,\mathrm{ave}\{(1-h)^2\theta^2 + h^2\sigma^2\} + \frac{n}{p(p-1)}\,\mathrm{ave}(h^2\sigma^2) + n\,\mathrm{ave}(\sigma^2), \end{aligned}$$

which implies that

$$E\left\{\frac{p}{n}P(h)\right\} = \rho(\hat{\theta}_h, \theta) + \frac{1}{p-1}\,\mathrm{ave}(h^2\sigma^2) + p\,\mathrm{ave}(\sigma^2)$$

and, therefore $pn^{-1}P(h)$ is an upwardly biased estimator of the risk $\rho(\hat{\theta}_h, \theta)$. A possible correction to this estimator is

$$\begin{aligned} \hat{\rho}(h) &= \frac{p}{n}P(h) - \frac{1}{p-1}\,\mathrm{ave}(h^2\hat{\sigma}^2) - p\,\mathrm{ave}(\hat{\sigma}^2) \\ &= \mathrm{ave}\{(1-h)^2\hat{\theta}^2\} + \mathrm{ave}\{(2h-1)\hat{\sigma}^2\}. \end{aligned} \tag{8}$$

We now follow the approach in Beran & Dümbgen (1998) for studying modulation estimators in an additive Gaussian noise model, but adapting it for our general model. Throughout, C denotes a generic universal real-valued constant which does not depend on n, θ, σ^2 or $\mathscr{H}$, but whose value may be different at different points. Also, let $J(\mathscr{H})$ denote a functional of the uniform covering number of $\mathscr{H}$ (Dudley, 1987).

PROPOSITION 1. *Let $\mathscr{H}$ be any closed subset of $[0,1]^T$ containing 0, let $\tilde{h}$ be a minimiser of $\rho(\hat{\theta}_h, \theta)$ over $h \in \mathscr{H}$ and let $\hat{h}$ minimise $\hat{\rho}(h)$ over $h \in \mathscr{H}$. Then*

$$E\{|\hat{\rho}(\hat{h}) - \rho(\hat{\theta}_{\tilde{h}}, \theta)|\} \leqslant C\left[J(\mathscr{H})\frac{\sqrt{E\{\mathrm{ave}(\hat{\theta}-\theta)^4\}} + \sqrt{\mathrm{ave}(\sigma^2\theta^2)}}{\sqrt{n}} + E\{|\mathrm{ave}(\hat{\sigma}^2 - \sigma^2)|\}\right].$$

Proof. Let $\varepsilon = \hat{\theta} - \theta$ be the vector of residuals and define random functions $S_1 = \varepsilon^2 - \sigma^2$ and $S_2 = \theta \cdot \hat{\theta}$, on T. It is easy to see that

$$\hat{\rho}(h) - \rho(\hat{\theta}_h, \theta) = \mathrm{ave}\{(h^2 - 2h + 1)(S_1 + 2S_2) + (2h-1)S\},$$

where $S = \hat{\sigma}^2 - \sigma^2$. Hence

$$\sup_{h \in \mathscr{H}} |\hat{\rho}(h) - \rho(\hat{\theta}_h, \theta)| \leqslant 4 \sup_{g \in \mathscr{G}} |\mathrm{ave}(gS_1)| + 8 \sup_{g \in \mathscr{G}} |\mathrm{ave}(gS_2)| + |\mathrm{ave}(S)|,$$

where $\mathscr{G} = \{fg : f, g \in \mathscr{H}\}$. The proposition now follows from Lemmas 6.3 and 6.4 of Beran & Dümbgen (1998), by checking that their Theorem 6.1 still holds for nonindependent random vectors. □

Proposition 1 is about convergence of the risk $\hat{\rho}(\hat{h})$. Proposition 2 establishes that $\hat{h}$ and $\tilde{h}$, as well as $\hat{\theta}_{\hat{h}}$ and $\hat{\theta}_{\tilde{h}}$, converge to one another. With the same notation as for the proof of Proposition 1, the proof of Proposition 2 mimics, with the appropriate but obvious modifications, the proof of Theorem 2.2 of Beran & Dümbgen (1998) and it is therefore omitted.

PROPOSITION 2. *Let $\tilde{h}$ and $\hat{h}$ be as defined in Proposition 1. Then*

$$E[\mathrm{ave}\{(\sigma^2 + \theta^2)(\hat{h} - \tilde{h})^2\}] \leqslant CJ(\mathscr{H})\frac{\sqrt{E\{\mathrm{ave}(\hat{\theta}-\theta)^4\}} + \sqrt{\mathrm{ave}(\sigma^2\theta^2)}}{\sqrt{n}} + E\{|\mathrm{ave}(\hat{\sigma}^2 - \sigma^2)|\},$$

$$E[\mathrm{ave}\{(\hat{\theta}_{\tilde{h}} - \hat{\theta}_{\hat{h}})^2\}] \leqslant CJ(\mathscr{H})\frac{\sqrt{E\{\mathrm{ave}(\hat{\theta}-\theta)^4\}}}{\sqrt{n}} + E\{|\mathrm{ave}(\hat{\sigma}^2 - \sigma^2)|\}.$$

Remark 2. It follows from (7) and the strong law of large numbers that $\mathrm{ave}(\hat{\sigma}^2)$ is a consistent estimator of $\mathrm{ave}(\sigma^2)$. Hence, in view of Propositions 1 and 2, a class $\mathscr{H}$ such that $J(\mathscr{H}) = o(n^{\frac{1}{2}})$ together with the boundedness of $E\{\mathrm{ave}(\hat{\theta}-\theta)^4\}$ and $\mathrm{ave}(\sigma^2\theta^2)$ ensure the success of the estimator $\hat{\theta}_{\hat{h}}$. For instance Example 2 of Beran & Dümbgen (1998) shows that, when $\mathscr{H}$ consists of piecewise constant functions on intervals of length $[\log(\log n)]$, where $[x]$ denotes the integer part of x, we indeed get $J(\mathscr{H}) = o(n^{\frac{1}{2}})$. Furthermore, since the components of Y are natural exponential families with quadratic variance functions, boundedness of $E\{\mathrm{ave}(\hat{\theta}-\theta)^4\}$ and $\mathrm{ave}(\sigma^2\theta^2)$ follows if $\mathrm{ave}(\theta^4) < c$, for $c > 0$, which can be interpreted as a smoothness assumption on μ.

Remark 3. A particular consequence of Propositions 1 and 2 is that the estimator of μ derived from our modulation estimator $\hat{\theta}_{\hat{h}}$ attains the optimal mean integrated squared error asymptotic rates $\mathcal{O}(n^{-2s/(2s+1)})$ for a class of submodels for μ, namely the class of functions belonging to an ellipsoid of the Sobolev class W_2^s of smoothness index $s > \frac{1}{2}$. Indeed, in such a case we have $\mathrm{ave}(\theta^4) \leqslant \mathcal{O}(n^{-4s/(2s+1)})$ and, because of the smoothness of μ, it is easy to show that $E\{|\mathrm{ave}(\hat{\sigma}^2 - \sigma^2)|\} \to 0$ at a rate $n^{-2s/(2s+1)}$ as $n \to \infty$. The asymptotic rate of $\hat{\theta}_{\hat{h}}$ is then a direct application of Corollary 2.3 of Beran & Dümbgen (1998).

Our objective now is to choose h to minimise $\hat{\rho}(h)$ defined in (8). The optimum modulation estimator is a multiple Stein estimator similar to that used in Example 2 of Beran

& Dümbgen (1998), see Remark 2, but for practical purposes it is usually adequate to use the simple modulator

$$\hat{h} = \frac{(\hat{\theta}^2 - \hat{\sigma}^2)_+}{\hat{\theta}^2}, \tag{9}$$

where $(u)_+ = \max(u, 0)$ and the division is componentwise. Inspection of (9) shows that, componentwise for each $t \in T$, $\hat{h}(t) = 0$ when $\hat{\theta}^2(t) \leqslant \hat{\sigma}^2(t)$. Hence, the modulator is set to 0 when the signal-to-noise ratio is less than 1. Moreover, componentwise for each $t \in T$, the modulator tends to 1 as the signal-to-noise ratio tends to 1.

The derivation of the modulator in expression (9) relies upon a realisation of a pseudo-sample $Z_1, \ldots, Z_p$ which is not observable. Moreover, one would expect better results when the size p of this pseudo-sample is large. Our purpose now is to show that the limiting, as $p \to \infty$, form of the modulator can be approximated by an expression computed directly from the original data. For each $t \in T$, only $\hat{\sigma}^2(t)$ in (9) depends on the pseudo-sample $Z_1, \ldots, Z_p$. Recall from (7) and the strong law of large numbers that, for each $t \in T$, $\hat{\sigma}^2(t)$ is a consistent estimator of $\sigma^2(t)$ and that $E\{\hat{\sigma}^2(t)\} = \sum_{l=1}^{n} w_{t,l}^2 V\{\mu(l)\}$. By Theorem 3.2 of Morris (1983), which deals with estimation of the variance function in natural exponential families with quadratic variance functions, the uniform minimum variance unbiased estimator of $V\{\mu(l)\}$ is given by $\hat{V}\{\mu(l)\} = V\{Y(l)\}/(1 + v_2)$. If we combine these facts, an intuitively appealing approximation of our modulator, overcoming the fact that the pseudo-sample is never observed, takes the form

$$\hat{h} = \frac{(\hat{\theta}^2 - \tilde{\sigma}^2)_+}{\hat{\theta}^2}, \tag{10}$$

where

$$\tilde{\sigma}^2(t) = \frac{1}{1 + v_2} \sum_{l=1}^{n} w_{t,l}^2 V\{Y(l)\} \quad (t \in T), \tag{11}$$

and therefore our nonlinear modulation estimator is given by

$$\hat{\theta}_{\hat{h}} = \frac{(\hat{\theta}^2 - \tilde{\sigma}^2)_+}{\hat{\theta}^2}\, \hat{\theta}. \tag{12}$$

Remark 4. In the Gaussian case, with $v_0 = \sigma^2$ and $v_1 = v_2 = 0$, the uniform minimum variance unbiased estimator of $V\{\mu(l)\}$ suggested in Theorem 3.2 of Morris (1983) is not defined. For this case we therefore suggest two alternatives. On the one hand, by considering a consistent estimator of σ, such as those given in Beran & Dümbgen (1998, p. 1834), we can obtain a modulation estimator, which is a modified James–Stein estimator, similar to that given in Example 1 of Beran & Dümbgen (1998) based on their bootstrap estimator. Alternatively, by considering a robust estimator of σ, such as the mean absolute deviation of the wavelet coefficients at the finest level, divided by 0·6745, we can use the wavelet shrinkage estimator of Donoho & Johnstone (1994). In the Poisson case, with $v_0 = 0$, $v_1 = 1$ and $v_2 = 0$, our modulator in expression (10) is similar in form to that defined heuristically in Theorem 4.1 of Nowak & Baraniuk (1999).

3·3. *Computation of the modulation estimator*

Expression (12) involves the projection of the variance function estimator $V(Y)/(1 + v_2)$ on to the pointwise square of the wavelet basis functions; see (11). An efficient filter-bank

algorithm for computing such projections for one-dimensional signals can be derived from the diagonal elements of the covariance structure of wavelet coefficients described in the papers of Vannucci & Corradi (1999) or Kovac & Silverman (2000).

Here we use the same notation as in Vannucci & Corradi (1999). Let $\{X(t), t \in \mathbb{R}\}$ be a square-integrable stochastic process and let c_{jk} and d_{jk} respectively be the discrete scaling and discrete wavelet coefficients of $X(t)$ with respect to a wavelet basis. Assume that the variance-covariance matrix of the discrete scaling coefficients $c^{\{j+1\}}$ at level $j+1$ is known and is denoted by $C^{\{j+1\}}$. If we use filter notation and Proposition 1 of Vannucci & Corradi (1999) the following identities hold:

$$D^{\{j\}} = G_{j+1} C^{\{j+1\}} G_{j+1}^{\mathrm{T}}, \quad C^{\{j\}} = H_{j+1} C^{\{j+1\}} H_{j+1}^{\mathrm{T}},$$

where $D^{\{j\}}$ indicates the variance-covariance matrix of the wavelet coefficients $d^{\{j\}}$ at level j. The vectors G_{j+1} and H_{j+1} contain elements corresponding to the quadrature mirror filters associated with the wavelet and scaling coefficients respectively. The two covariance matrices defined above are simply the diagonal blocks of the two-dimensional discrete wavelet transform applied to the matrix $C^{\{j+1\}}$.

In order to derive what we call the squared discrete wavelet transform of a positive signal $s = \{s(t_1), \ldots, s(t_n)\}$, $t_i = i/n$, $n = 2^J$ for some positive $J > 0$, it is therefore sufficient to initialise the $C^{\{J\}}$ matrix at scale J by a diagonal matrix with the signal s on the diagonal and to perform a two-dimensional wavelet transform. By construction the diagonal elements of the resulting matrix project the signal s on the element-wise square of each wavelet basis element. Given the diagonal structure of $C^{\{J\}}$, the complete squared discrete wavelet transform of the signal s requires only $O(n)$ operations and is therefore computationally fast.

4. Applications and comparisons

4·1. *Simulation study: The Poisson case*

Consider Poisson distributed time series data observed on 256 equispaced data points, so that $T = \{t_i = i/n, i = 1, \ldots, n = 256\}$. Three factors of interest were included in the study, namely the morphology of the mean function, the intensity level, i.e. the Poisson rate, and the method of estimation. With practical relevance in mind, we chose two shapes for the underlying mean function, namely that of a burst (Kolaczyk, 1997) given by

$$\mu(t) = A + A_1 I_1(t) + A_2 I_2(t) + A_3 I_3(t),$$

with

$$I_i(t) = \begin{cases} \exp\{-(|t - t_{i,\max}|/\sigma_r)^{\nu}\}, & \text{if } t \leqslant t_{i,\max}, \\ \exp\{-(|t - t_{i,\max}|/\sigma_d)^{\nu}\}, & \text{if } t > t_{i,\max}, \end{cases} \tag{13}$$

and that of a very smooth function (Beran & Dümbgen, 1998) given by

$$\mu(t) = A + 2\{6{\cdot}75 t^2 (1-t)\}^3,$$

where A, A_1, A_2, A_3, $t_{i,\max}$, ν, σ_r and σ_d are given constants. The average intensity level varied over 5, 50 and 200 counts per time point, so as to gain some insight into the behaviour at low, medium and high signal-to-noise ratio of the three estimation procedures considered. Figure 1 shows the two test functions with a medium signal-to-noise ratio.

We applied our wavelet shrinkage method proposed in § 3 and the *P*-splines nonpara-

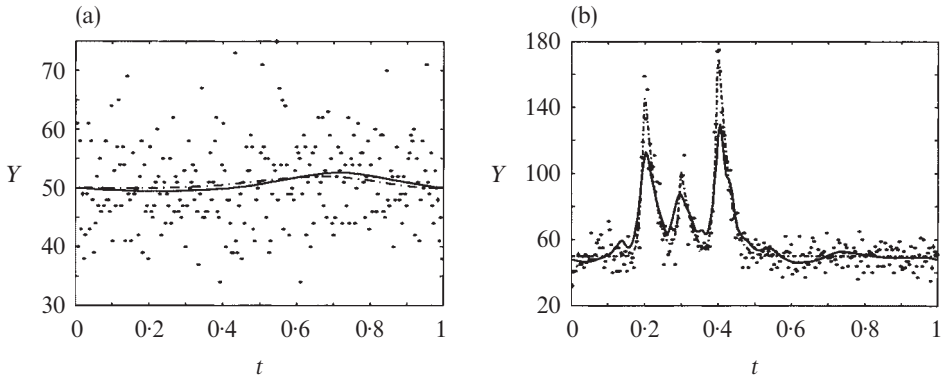

Fig. 1. Test functions for simulation study, dot-dashed, along with estimates from a single trial, using the new method, solid, based on $n = 256$ points per Poisson time series and a medium signal-to-noise ratio intensity level: (a) smooth test function, (b) burst test function.

metric smoothing methodology for generalised linear models proposed by Eilers & Marx (1996), with a smoothing parameter chosen by minimising an AIC criterion. We also applied the wavelet shrinkage method proposed by Donoho (1993) which first pre-processes the data using Anscombe's (1948) square-root transformation, namely $z = 2(y + \frac{3}{8})^{\frac{1}{2}}$, to normalise and variance-stabilise the data, after which the Gaussian thresholds $(2 \log n)^{\frac{1}{2}}$ are used with the usual wavelet shrinkage algorithm. At each of the $6 = 2 \times 3$ combinations of shape and intensity level, estimates were calculated using each of the three methods, over 500 trials. The two wavelet methods were based on Daubechies' nearly symmetric wavelets of order 8. Estimates from a single trial, using our new procedure with a medium signal-to-noise ratio, are shown in Fig. 1. The results of these simulations are shown in Fig. 2.

As one can see from Figs 1 and 2 the new method produced estimates with the lowest mean squared error uniformly across the various combinations of morphology and intensity rate. However, the relative performance of the competing methodologies tended to vary with these factors. For example, for bursts, the *P*-splines procedure fared noticeably worse than both wavelet-based methods, even though we used a spline basis with 40 initial nodes in order to catch the high curvature of the bursts. Finally, despite the polynomial nature of the smooth test function, our new method performs better than the *P*-splines method which is specifically designed to treat such cases. The performance of Donoho's wavelet smoothers in all situations confirms the method's tendency to produce estimates of signals with slightly reduced amplitudes.

4·2. *Simulation study: The binomial case*

We performed another set of simulations with binomial distributed time series data, which has a fully quadratic variance function, observed again on 256 equispaced data points, so that $T = \{t_i = i/n, i = 1, \ldots, n = 256\}$. Simulations were carried out for the burst and smooth functions given in § 4·1, appropriately scaled to produce a positive function bounded above by 1. Various fixed binomial weights were used for these simulations. In order to compare the new procedure with a standard wavelet shrinkage procedure we applied a normalising and variance-stabilising transformation to the data before applying the classical wavelet shrinkage. The transformation consists of pre-processing the data using Anscombe's (1948) arcsine transformation, $z = 2n^{\frac{1}{2}} \arcsin\{(y/n)^{\frac{1}{2}}\}$, after which the

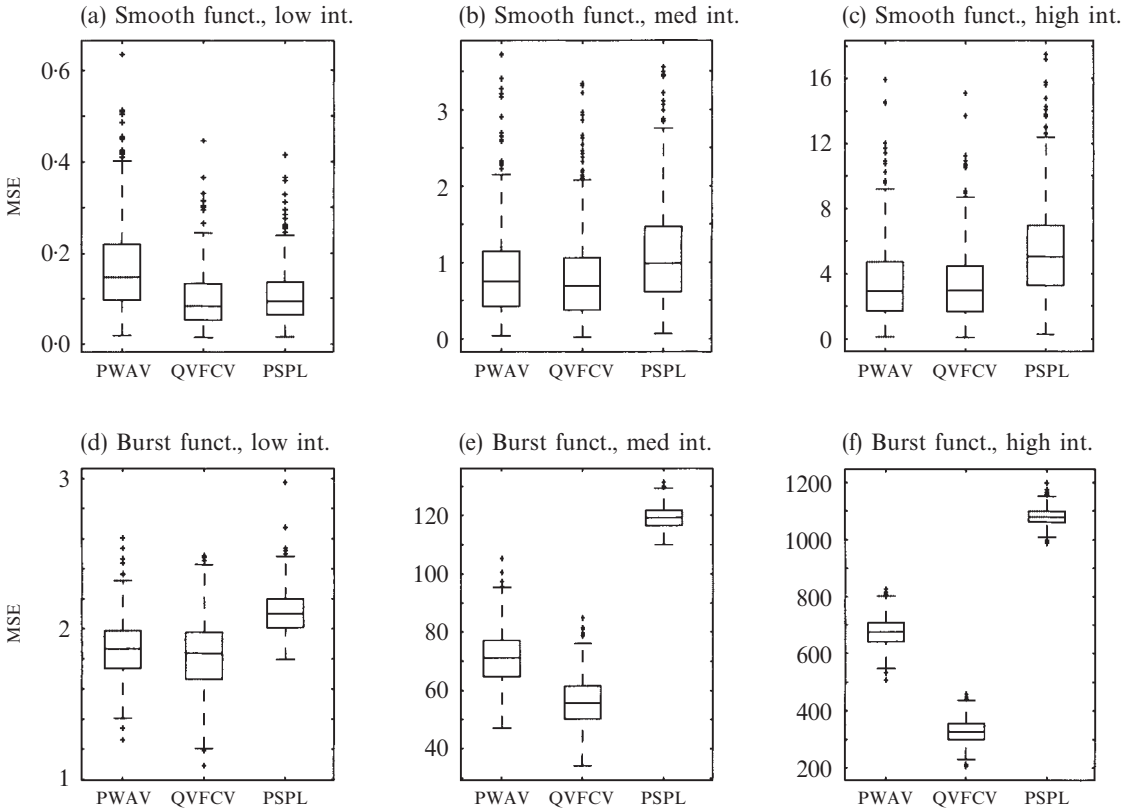

Fig. 2. Boxplots of mean squared error for Poisson simulation results, for the three methods and for all six combinations of signal morphology, smooth and burst test functions, and low, medium and high signal intensity; PWAV, Donoho's wavelet method; QVFCV, the new wavelet method; PSPL, the *P*-splines method.

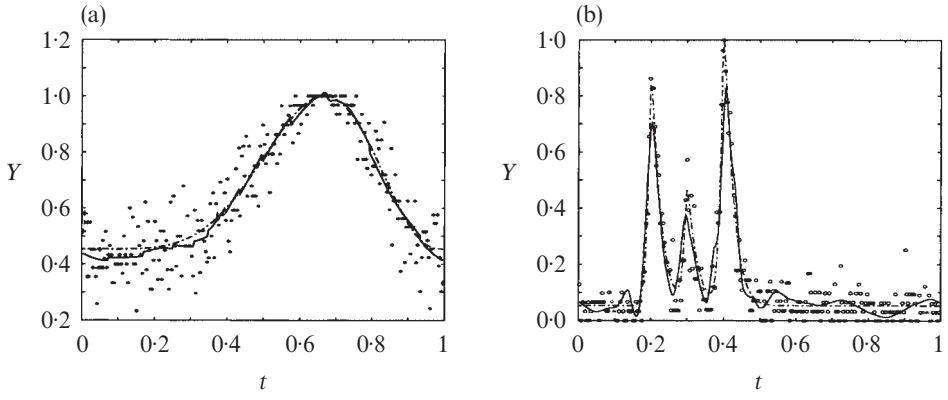

Fig. 3. Test functions for simulation study, dot-dashed, along with estimates from a single trial, using the new method, solid, based on $n = 256$ points per binomial time series and a medium signal-to-noise ratio intensity level: (a) smooth test function, (b) burst test function.

Gaussian thresholds $(2 \log n)^{\frac{1}{2}}$ are used with the usual wavelet shrinkage algorithm. Both wavelet methods were based on Daubechies' nearly symmetric wavelets of order 8. Estimates from a single trial, using our new procedure with a medium signal-to-noise ratio, are shown in Fig. 3. The results of the simulations are shown in Fig. 4.

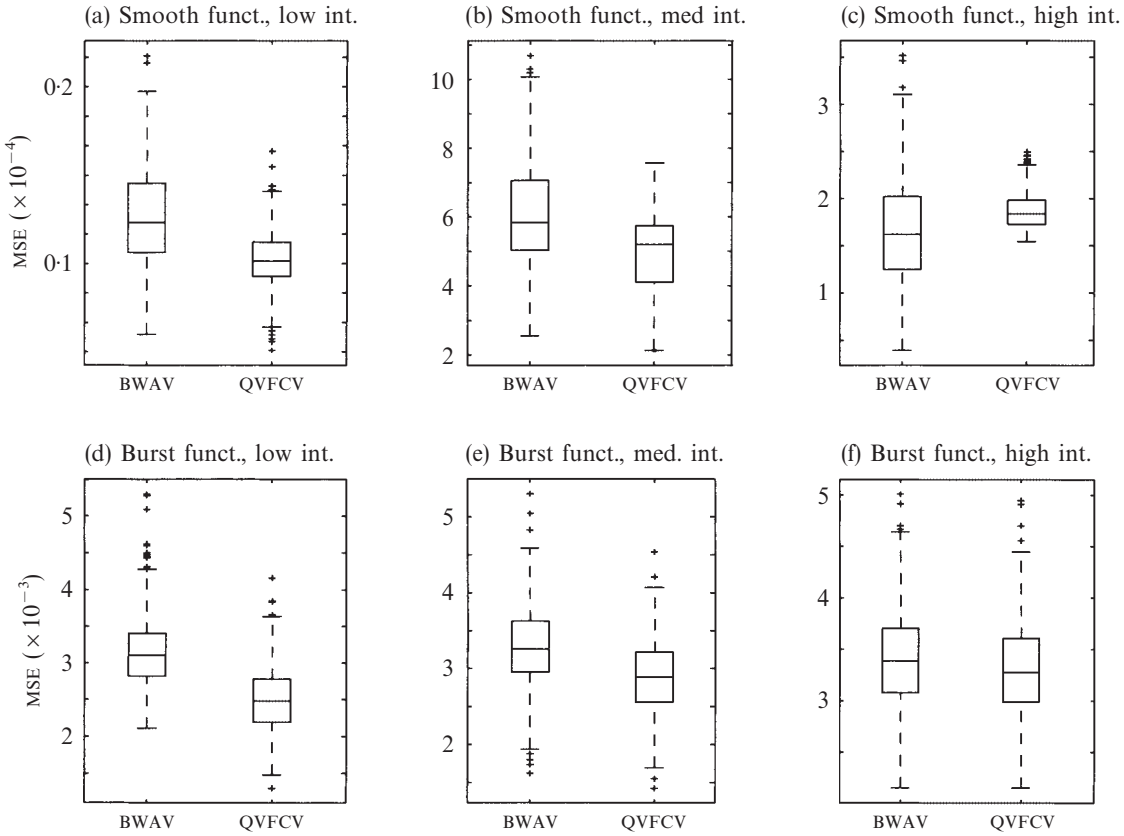

Fig. 4. Boxplots of mean squared error for the binomial simulation results, for the two methods and combinations of signal morphology, smooth and burst test functions, and low, medium and high signal intensity; BWAV, standard wavelet method; QVFCV, the new wavelet method.

The conclusions reached from the simulations are quite similar to those for the Poisson case, although the new method was outperformed in the high-rate setting for the smooth function. Note however that in this case the new method has much less variability.

4·3. *Gamma-ray bursts*

In this section we apply our method to gamma-ray data from instruments, on board NASA's Compton Gamma Ray Observatory (Kolaczyk, 1999b), that record the arrival times of gamma-ray photons corresponding to detected gamma-ray bursts.

Tracking a variable object's changes in brightness, based on photon counting data, is a fundamental problem in astronomy. For example, the importance of activity of galactic and extragalactic objects on time scales at and below the millisecond range led NASA to design its X-ray and gamma-ray observatories to detect individual photons with microsecond timing accuracy. Existing methods do not fully and correctly extract from photon counts the scientifically useful information, which is buried in the fluctuations inherent in the occurrence of discrete, independent events. The time series shown in Fig. 5(a) is constructed, as is typical in practice, by aggregating the photon arrival times into intervals of equal length. It is usually assumed that observational errors for the binned counts are additive and Gaussian. However, counting fluctuations are neither additive nor Gaussian. Indeed, the photon detection phenomenon is close to a Poisson process, the major depar-

ture from this being a lack of independence. In particular, detectors have a dead-time, in that arrival of a photon temporarily inhibits detection of subsequent photons.

Gamma-ray bursts are non-periodic signals, localised in time, that are not part of the global intensity. That is, the intensity of the observed signal is altered by the presence of the bursts and is not of the simple form postulated for the global signal. Bursts can occur randomly, periodically or in any other fashion. Figure 5(a) depicts the binned counts as a function of time, in microseconds, for photon data from the burst called Trigger 0551. The raw data comprise about 29 000 photons. The other curves show the intensity estimates based on the two wavelet procedures used in § 4·1. These crude pulses are then used as initial guesses for a numerical routine that deconvolves overlapping pulses by fitting a parametric model. The initial intensity estimate is very important for encouraging convergence of this fitting procedure to the global optimum.

Note that the two intensity estimates are quite similar. However, the estimate from the new procedure shows distinct visual evidence of an additional pulse at the beginning of the estimated intensity, which was also selected as a statistically significant local maximum by the Bayesian blocks algorithm described in Scargle (1998).

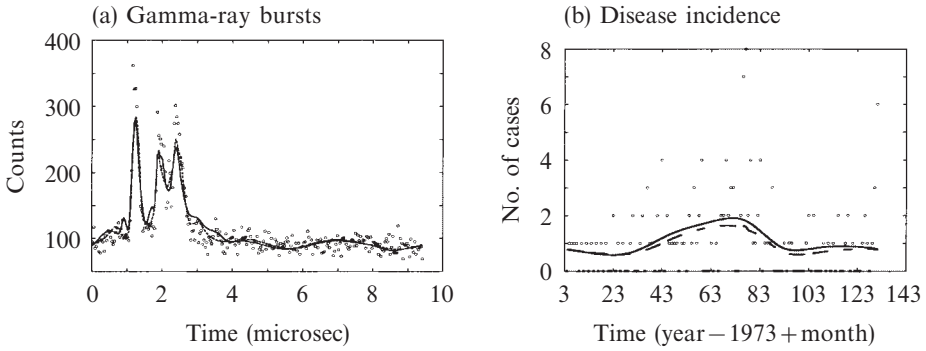

Fig. 5. Estimates of the underlying intensity function based on the new procedure, solid, and Donoho's wavelet method, dashed: (a) the photon detection data from the burst called Trigger 0551, (b) monthly number of cases of poliomyelitis from March 1973 to December 1982.

4·4. *Disease incidence data*

We have applied our estimation procedure to the disease incidence data described in Zeger (1988), which lists the monthly numbers of incidences of poliomyelitis reported by U.S. centres for disease control for the years 1970 to 1982. Of interest is whether or not this record provides evidence of a long-term decrease in the rate of U.S. polio infection during the period from March 1973 to December 1982. We applied both the wavelet procedures used in § 4·1 to estimate the trend in U.S. polio incidence during this period; see Fig. 5(b).

Both procedures produce estimated trends that are very similar and indicate a weak decrease in the rate of polio cases per month, agreeing with the conclusion reached by Zeger (1988) on the same dataset using a parameter-driven extension of log-linear models. This example shows that our new method is quite robust to isolated extreme observations.

Acknowledgement

The authors would like to thank Marina Vannucci of Texas A&M University, who kindly provided her Matlab routine for computing the covariance matrix of the wavelet coefficients. Theofanis Sapatinas would like to thank Anestis Antoniadis for support and excellent hospitality while visiting Grenoble to carry out this work. Helpful comments made by the editor and an anonymous referee are gratefully acknowledged.

References

Abramovich, F., Bailey, T. C. & Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *Statistician* **49**, 1–29.

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika* **35**, 246–54.

Antoniadis, A. (1997). Wavelets in statistics: a review (with Discussion). *J. Ital. Statist. Soc.* **6**, 97–144.

Antoniadis, A. & Leblanc, F. (2000). Nonparametric wavelet regression for binary response. *Statistics* **34**, 183–213.

Beran, R. & Dümbgen, L. (1998). Modulation of estimators and confidence sets. *Ann. Statist.* **26**, 1826–56.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM.

Donoho, D. L. (1993). Non-linear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data. In *Proceedings of Symposia in Applied Mathematics: Different Perspectives on Wavelets*, **47**, Ed. I. Daubechies, pp. 173–205. San Antonio, TX: American Mathematical Society.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.

Donoho, D. L. & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–24.

Donoho, D. L. & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage: asymptopia? (with Discussion). *J. R. Statist. Soc.* B **57**, 301–37.

Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Ann. Prob.* **15**, 1306–26.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statist. Sci.* **11**, 89–121.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd ed. New York: Marcel Dekker.

Johnstone, I. M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics*, **5**, Ed. S. S. Gupta and J. O. Berger, pp. 303–26. New York: Springer-Verlag.

Kolaczyk, E. D. (1997). Non-parametric estimation of Gamma-Ray burst intensities using Haar wavelets. *Astrophys. J.* **483**, 340–9.

Kolaczyk, E. D. (1999a). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sinica* **9**, 119–35.

Kolaczyk, E. D. (1999b). Bayesian multiscale models for Poisson processes. *J. Am. Statist. Assoc.* **94**, 920–33.

Kovac, A. & Silverman, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Am. Statist. Assoc.* **95**, 172–83.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pat. Anal. Mach. Intel.* **11**, 674–93.

Mallat, S. G. (1999). *A Wavelet Tour of Signal Processing*, 2nd ed. San Diego, CA: Academic Press.

Meyer, Y. (1992). *Wavelets and Operators*. Cambridge: Cambridge University Press.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10**, 65–80.

Morris, C. N. (1983). Natural exponential families with quadratic variance functions: statistical theory. *Ann. Statist.* **11**, 515–29.

Nowak, R. D. (1997). Optimal signal estimation using cross-validation. *IEEE Sig. Proces. Lett.* **4**, 23–5.

Nowak, R. D. & Baraniuk, R. G. (1999). Wavelet domain filtering for photon imaging systems. *IEEE Trans. Image Process.* **8**, 666–78.

Scargle, J. D. (1998). Studies in astronomical time series analysis. V. Bayesian blocks, a new method to analyze structure in photon counting data. *Astrophys. J.* **504**, 405–18.

Timmermann, K. E. & Nowak, R. D. (1999). Multiscale modeling and estimation of Poisson processes with applications to photon-limited imaging. *IEEE Trans. Info. Theory* **45**, 846–62.

VANNUCCI, M. & CORRADI, F. (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J. R. Statist. Soc.* B **61**, 971–86.
VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. New York: Wiley.
ZEGER, S. L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621–9.

[*Received January* 2000. *Revised October* 2000]