Sankhyā : The Indian Journal of Statistics
Special issue on Wavelets
2001, Volume 63, Series A, Pt. 3, pp. 309–327

WAVELET SHRINKAGE FOR NATURAL EXPONENTIAL FAMILIES WITH CUBIC VARIANCE FUNCTIONS

By ANESTIS ANTONIADIS
University of Joseph Fourier, France
PANAGIOTIS BESBEAS
University of Kent at Canterbury, England
and
THEOFANIS SAPATINAS
University of Cyprus, Cyprus

SUMMARY. 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. Limited work for applications involving data which are actually counts such as Poisson or Bernoulli data has also been considered. Recently, Antoniadis and Sapatinas (2001) have developed a wavelet shrinkage methodology for obtaining and assessing smooth estimates for complicated data such as those arising from a (univariate) natural exponential family with quadratic variance function (the variance is, at most, a quadratic function of the mean) studied by Morris (1982, 1983). The Gaussian, Poisson, gamma, binomial, negative binomial and generalized hyperbolic secant distributions are the only members of this family.

In this article we show that, subject to certain modifications, the wavelet shrinkage methodology of Antoniadis and Sapatinas (2001) can be extended to the case where the data arise from a (univariate) natural exponential family with cubic variance function (the variance is, at most, a cubic function of the mean) studied by Letac and Mora (1990). Twelve different distributions are the only members of this family. The first six appear in Morris (1982, 1983); most of the other six appear as distributions of the first passage times in the literature, the inverse Gaussian distribution being the most famous example. As an illustration of the proposed wavelet shrinkage methodology, a simulation study for inverse Gaussian data has been conducted.

Paper received April 2000; revised April 2001.

AMS (1991) *subject classification.* 62G05, 62G08

Key words and phrases. Cross-validation mean squared error, diagonal shrinkage estimation, inverse Gaussian distribution, modulation estimators, natural exponential families, 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. Donoho and 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 a recent survey of research in this and other related areas we refer, for example, to Antoniadis (1997), Vidakovic (1999), Abramovich *et al.* (2000) and Antoniadis *et al.* (2001).

In the case where the data are actually counts (such as Poisson data), Donoho (1993) proposed to first pre-process the data using a variance-stabilizing and normalizing transformation, such as the one proposed by Anscombe (1948), and then to apply usual wavelet shrinkage techniques. Despite the seemingly straightforward nature of this approach, it has been criticized for often smoothing away more structure than is tolerable. Typically, the complaint is one of oversmoothing or attenuation of fine detail structure in the underlying object (e.g. signal or image), especially in situations involving very *low levels* of counts. As a result, wavelet shrinkage estimation techniques have been recently developed by directly considering the original, untransformed count data (see Kolaczyk, 1997, 1999a, 1999b, Nowak and Baraniuk, 1999, Timmermann and Nowak, 1999 for Poisson data; Antoniadis and Leblanc, 2000 for Bernoulli data).

Borrowing ideas from *modulation estimators* that were originally developed for Gaussian data by Beran and Dümbgen (1998), Antoniadis and Sapatinas (2001) have recently developed a wavelet shrinkage methodology that has been successfully applied to various types of data. In particular, they have discussed a wavelet shrinkage methodology for (univariate) natural exponential families (NEF) with quadratic variance functions (QVF). The Gaussian, Poisson, gamma, binomial, negative binomial and generalized hyperbolic secant distributions are the only NEF with QVF, i.e., the variance is a polynomial function of the mean with degree less than or equal to 2 for each of these distributions (see Morris, 1982, 1983).

In this article we show that, subject to certain modifications, the methodology of Antoniadis and Sapatinas (2001) can be extended to the case where the data arise from NEF with cubic variance functions (CVF). Twelve different distributions are the only NEF with CVF, i.e., the variance is a polynomial function of the mean with degree less than or equal to 3 for each of these distributions. The first six appear in Morris (1982, 1983) whilst most of the other six appear as distributions of the first passage times in the literature, the inverse Gaussian distribution being the most famous example (see Letac and Mora, 1990); this latter distribution has been used to model the motion of particles in a colloidal suspension under an electric field and in sequential analysis to cite only a few examples (see Seshadri, 1999). The main difference between the results presented here and those obtained by Antoniadis and Sapatinas (2001) is that, while in the NEF-QVF case a closed form estimator is available for the variance function involved in the resulting wavelet estimator, this is not the case for the six extra distributions of the NEF-CVF. The estimation of the variance function is now case specific and it seems that more than two observations at each sampling point are required to get adequate estimates. To overcome this, a further step of binning the data to generate multiple observations has been considered. The results discussed here are, therefore, presented to augment (errors from a larger family of distributions) rather than to supplant the methodology of Antoniadis and Sapatinas (2001).

This article is structured as follows. Section 2 briefly reviews the relevant material on NEF-CVF that we shall need in subsequent sections. In Section 3, we discuss a non-linear wavelet shrinkage methodology for data arising from NEF-CVF. To illustrate the usefulness of the proposed wavelet shrinkage methodology, a simulation study for inverse Gaussian data has been conducted in Section 4. 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 on NEF-CVF Distributions

This section briefly overviews some material that we shall use in subsequent sections. For a more detailed account we refer to Letac and Mora (1990).

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

$$P_\theta(X \in A) = \int_A \exp(x\theta - \psi(\theta))\, dF(x), \qquad (1)$$

with F a Stieltjes measure on $\mathbb{R}$ not depending on $\theta \in \Theta$, the *natural parameters*, and sets $A \subset \mathbb{R}$. The *cumulant generating function* $\psi(\theta)$ gives (1) unit probability. The random variable X is the *natural observation.* Exponential families that are not NEF are nonlinear transformations of NEF. The natural observation X has mean and variance

$$\mu = \psi'(\theta) = \mathbb{E}_\theta(X) = \int x \, dF_\theta(x)$$

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

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* (VF) of the NEF and characterizes the NEF (but no particular members of the NEF).

Consider now NEF which have CVF given by

$$V(\mu) = v_0 + v_1\mu + v_2\mu^2 + v_3\mu^3, \tag{2}$$

where v_0, v_1, v_2 and v_3 are real-valued constants. We shall write

$$X \sim \text{NEF} - \text{CVF}(\mu, V(\mu))$$

to denote a random variable which follows a NEF with mean μ and CVF $V(\mu)$ given by (2). It can be shown that exactly twelve types of NEF-CVF exist. The first six appear in Morris (1982, 1983); the other six are the Abel, Takács, strict arcsine, large arcsine, Kendall-Ressel and inverse Gaussian distributions. Most of these appear as distributions of the first passage times in the literature, the inverse Gaussian distribution being the most famous example. We refer to Table 2 of Letac and Mora (1990) for more details about these latter six distributions. The twelve types of distributions can be extended by convolutions all of which preserve both the NEF and the CVF properties (see Proposition 2.5 in Letac and Mora, 1990). This is the key property for developing the wavelet shrinkage methodology in Section 3.

3. Wavelet Shrinkage for NEF-CVF Distributions

We consider the problem of recovering a signal from independent NEF-CVF observations, which may be formulated as follows. Let $\mathbf{Y} = \mathbf{Y}_n = (Y(t))_{t \in T}$ be a random function observed on the set $T = T_n = \{1, \ldots, n\}$.

The components $Y(t)$ of $\mathbf{Y}$ are assumed to be independent random variables such that

$$Y(t) \sim \mathrm{NEF} - \mathrm{CVF}(\mu(t), V(\mu(t))), \quad t \in T, \tag{3}$$

where

$$V(\mu(t)) = v_0 + v_1\mu(t) + v_2\mu^2(t) + v_3\mu^3(t), \quad t \in T$$

for real-valued constants v_0, v_1, v_2 and v_3. Working with functions on T rather than vectors in $\mathbb{R}^n$ is convenient for our purposes. We assume, hereafter, that the mean vector $\boldsymbol{\mu} = (\mu(t))_{t\in T}$ 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 $\mathbf{Y} = (Y(t))_{t\in T}$ without assuming any particular parametric form.

Define $\boldsymbol{\theta} = W\boldsymbol{\mu}$ to be the vector of wavelet coefficients corresponding to $\boldsymbol{\mu}$, where W is the $n \times n$ orthogonal matrix associated with the *discrete wavelet transform* (DWT) (see, for example, Mallat, 1989). The squared error loss is widely used for studying the quality of nonparametric function estimators. By the orthogonality of the wavelet transform the squared error loss in the wavelet domain is equivalent, up to a $\sqrt{n}$ factor, to the squared error loss in the observation domain. Throughout we work on the wavelet domain; the squared error loss of any estimator (linear or nonlinear) $\breve{\boldsymbol{\theta}}$ (which depends on $\hat{\boldsymbol{\theta}} = W\mathbf{Y}$) and its corresponding risk are respectively defined to be

$$L(\breve{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathrm{ave}[(\breve{\boldsymbol{\theta}} - \boldsymbol{\theta})^2] \quad \text{and} \quad \rho(\breve{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}(L(\breve{\boldsymbol{\theta}}, \boldsymbol{\theta})),$$

where

$$\mathrm{ave}(\mathbf{g}) = \frac{1}{n}\sum_{t\in T} g(t),$$

for any $\mathbf{g} \in \mathbb{R}^T$, the space of real-valued functions defined on T.

When the observed function $\mathbf{Y}$ is Gaussian, with $\mathbb{V}(\mathbf{Y}) = \mathbb{V}(Y(t)) = \sigma^2$ for all $t \in T$, Beran and Dümbgen (1998) have shown that one can construct estimators of $\boldsymbol{\theta}$ by diagonal shrinkage that are (i) asymptotically minimax optimal over a variety of ellipsoids in the parameter space and (ii) sometimes more efficient than the oracle-based estimators introduced by Donoho and Johnstone (1994). These estimators take the form $\hat{H}W\mathbf{Y}$, where $\hat{H} = diag(\hat{\mathbf{h}})$ is the diagonal matrix of order n and $\hat{\mathbf{h}} : T \to [0,1]$ (depends on $\hat{\boldsymbol{\theta}} = W\mathbf{Y}$) is chosen to minimize the estimated risk of the linear estimator $\breve{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathbf{h}} = HW\mathbf{Y}$ over all functions $\mathbf{h}$ in a class $\mathcal{H} \subset [0,1]^T$. Such estimators are, by construction, nonlinear and shrink each coordinate towards zero, different coordinates being possibly treated differently. Adapting the terminology of Beran and Dümbgen (1998), each function $\mathbf{h}$ in a class $\mathcal{H} \subset [0,1]^T$ will be called a *modulator* and the estimators $\hat{\boldsymbol{\theta}}_{\hat{\mathbf{h}}} = \hat{H}W\mathbf{Y}$ will be referred to as the *modulation* estimators.

Since the Gaussian distribution is a particular member of the NEF-CVF family, our estimation procedure developed in Section 3.1 below for estimating $\boldsymbol{\theta}$ for the general NEF-CVF model will based on similar ideas discussed above.

3.1. *Estimating the optimal modulator using the cross-validation mean squared error.* In this section we mostly follow the approach suggested by Antoniadis and Sapatinas (2001) but adapting it to the NEF-CVF case. We will first construct a suitably consistent estimator $\hat{\rho}$ of the risk $\rho(\hat{\boldsymbol{\theta}}_{\mathbf{h}}, \boldsymbol{\theta})$ and we will propose to estimate $\boldsymbol{\theta}$ by the modulation estimator $\hat{\boldsymbol{\theta}}_{\hat{\mathbf{h}}} = \hat{H}W\mathbf{Y}$, where $\hat{\mathbf{h}}$ is any function in $\mathcal{H} \subset [0,1]^T$ that minimizes $\hat{\rho}$. The methodology is actually based on the division property of NEF-CVF mentioned in Section 2, and a cross-validation approach similar to that developed by Nowak (1997) in the case where one has more than one independent observations of an unknown signal.

Suppose that, instead of observing data $\mathbf{Y}$, we observe the *pseudo-sample* $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_p\}$ (i.e. $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$ are independent random variables) with $\mathbf{Z}_k = (Z_k(t))_{t \in T}$ such that

$$Z_k(t) \sim \mathrm{NEF} - \mathrm{CVF}\Big(\nu(t), R(\nu(t))\Big), \quad k = 1, \ldots, p, \quad t \in T,$$

where

$$\nu(t) = \frac{1}{p}\mu(t) \quad \text{and} \quad R(\nu(t)) = \frac{1}{p}v_0 + v_1\nu(t) + pv_2\nu^2(t) + p^2v_3\nu^3(t)$$

for real-valued constants v_0, v_1, v_2 and v_3. Then, using Proposition 2.5 of Letac and Mora (1990),

$$\mathbf{Y} = \sum_{k=1}^{p} \mathbf{Z}_k \quad \text{and} \quad \bar{\mathbf{Z}} = \frac{1}{p}\sum_{k=1}^{p} \mathbf{Z}_k = \frac{1}{p}\mathbf{Y}.$$

Consider now an estimation procedure based on the pseudo-sample $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_p\}$ producing modulation estimators of the form $\hat{\boldsymbol{\theta}}_{\mathbf{h}}$ depending on a modulator $\mathbf{h} = (h(t))_{t \in T}$. Applying this procedure to the pseudo-sample $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_p\}$, without using the jth element, and defining $H = diag(\mathbf{h})$, leads to a modulation estimator

$$\hat{\boldsymbol{\theta}}_{\mathbf{h}}^{(j)} = HW\mathbf{Z}^{(j)} \quad \text{where} \quad \mathbf{Z}^{(j)} = \frac{1}{p-1}\sum_{\substack{k=1 \\ k \neq j}}^{p} \mathbf{Z}_k$$

with corresponding signal estimate

$$\hat{\mathbf{Z}}^{(j)} = W^T \hat{\boldsymbol{\theta}}_{\mathbf{h}}^{(j)}.$$

To estimate the optimal modulator, we shall first construct a suitable consistent estimator of the risk $\rho(\hat{\boldsymbol{\theta}}_{\mathbf{h}}, \boldsymbol{\theta})$ based on the cross-validation mean squared error or its equivalent form, the prediction sum of squares (PRESS) (see, for example, Eubank, 1999, p. 43). Let

$$P(\mathbf{h}) = \sum_{k=1}^{p} \|HW\mathbf{Z}^{(k)} - W\mathbf{Z}_k\|^2. \tag{4}$$

For simplicity, we will use hereafter the following notation

$$\hat{\boldsymbol{\theta}} = W\mathbf{Y} = pW\bar{\mathbf{Z}}, \quad \hat{\boldsymbol{\theta}}_k = W\mathbf{Z}_k \quad \text{and} \quad \boldsymbol{\sigma}^2 = \mathbb{V}(\hat{\boldsymbol{\theta}}).$$

For $t \in T$, let

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

Some algebraic calculations show that expression (4) may be written as

$$\begin{aligned} P(\mathbf{h}) \;\; &= \;\; \frac{p}{p-1} \sum_{t \in T} \Big(\hat{\sigma}^2(t) - 2p^{-1}(1-h(t))\hat{\sigma}^2(t) \\ &\qquad + p^{-2}(1-h(t))^2[\hat{\sigma}^2(t) + (p-1)\hat{\theta}^2(t)] \Big). \end{aligned}$$

Since the components $Y(t)$ of $\mathbf{Y}$ are independent NEF-CVF$(\mu(t), V(\mu(t)))$ and W is orthonormal, it is easily seen that, for all $t \in T$, we have

$$\mathbb{E}(\hat{\theta}(t)) = \theta(t) \quad \text{and} \quad \mathbb{E}(\hat{\sigma}^2(t)) = \sum_{l=1}^{n} w_{t,l}^2 V(\mu(l)) = \sigma^2(t). \tag{5}$$

Moreover,

$$\mathbb{E}(P(\mathbf{h})) = \frac{n}{p} \operatorname{ave}((1-\mathbf{h})^2\boldsymbol{\theta}^2 + \mathbf{h}^2\boldsymbol{\sigma}^2) + \frac{n}{p(p-1)} \operatorname{ave}(\mathbf{h}^2\boldsymbol{\sigma}^2) + n \operatorname{ave}(\boldsymbol{\sigma}^2),$$

which implies that

$$\mathbb{E}\Big(\frac{p}{n}P(\mathbf{h})\Big) = \rho(\hat{\boldsymbol{\theta}}_{\mathbf{h}}, \boldsymbol{\theta}) + \frac{1}{p-1} \operatorname{ave}(\mathbf{h}^2\boldsymbol{\sigma}^2) + p \operatorname{ave}(\boldsymbol{\sigma}^2)$$

and, therefore, $\frac{p}{n}P(\mathbf{h})$ is a biased upwards estimator of the risk $\rho(\hat{\boldsymbol{\theta}}_{\mathbf{h}}, \boldsymbol{\theta})$. A possible correction to this estimator is

$$\begin{aligned}\hat{\rho}(\mathbf{h}) &= \frac{p}{n}P(\mathbf{h}) - \frac{1}{p-1}\,\mathrm{ave}(\mathbf{h}^2\hat{\boldsymbol{\sigma}}^2) - p\,\mathrm{ave}(\hat{\boldsymbol{\sigma}}^2) \\ &= \mathrm{ave}((1-\mathbf{h})^2\hat{\boldsymbol{\theta}}^2) + \mathrm{ave}((2\mathbf{h}-1)\hat{\boldsymbol{\sigma}}^2). \qquad (6)\end{aligned}$$

Throughout C denotes a generic universal real constant which does not depend on n, $\boldsymbol{\theta}$, $\boldsymbol{\sigma}^2$ or $\mathcal{H}$, but whose value may be different in various places. Also, let $J(\mathcal{H})$ be the uniform covering integral associated with the uniform covering number of $\mathcal{H}$ (see, for example, Dudley, 1987, Beran and Dümbgen, 1998). The following two propositions are similar to those appeared in Antoniadis and Sapatinas (2001). The proof of Proposition 1 is based on Lemmas 6.3 and 6.4 of Beran and Dümbgen (1998) while the proof of Proposition 2 is based on Theorem 2.2 of Beran and Dümbgen (1998).

Remark **1** . The proof of Lemma 6.4 of Beran and Dümbgen (1998) assumes random vectors with independent components, which is not anymore true under the present set-up. However, a closer look at the proof of the above mentioned lemma shows that the independence assumption is used to conveniently apply a general theorem (Theorem 6.1) on independent stochastic processes. When we adapt Lemma 6.4 under our setting, and use its proof, the only thing that needs to be checked is the boundedness of the wavelet transform in L^2, which is obviously true for the wavelet basis that we have used. Theorem 6.1 of Beran and Dümbgen (1998) still holds, when the involved processes are linearly transformed by a bounded linear operator in L^2, which is exactly the case here.

Proposition 1 is about convergence of the risk $\hat{\rho}(\hat{\mathbf{h}})$. Proposition 2 establishes that $\hat{\mathbf{h}}$ and $\tilde{\mathbf{h}}$, as well as $\hat{\boldsymbol{\theta}}_{\tilde{\mathbf{h}}}$ and $\hat{\boldsymbol{\theta}}_{\hat{\mathbf{h}}}$, converge to one another.

Proposition **1** . *Let $\mathcal{H}$ be any closed subset of $[0,1]^T$ containing 0, let $\tilde{\mathbf{h}}$ be a minimizer of $\rho(\hat{\boldsymbol{\theta}}_{\mathbf{h}}, \boldsymbol{\theta})$ over $\mathbf{h} \in \mathcal{H}$ and let $\hat{\mathbf{h}}$ be minimize $\hat{\rho}(\mathbf{h})$ over $\mathbf{h} \in \mathcal{H}$. Then*

$$\begin{aligned}&\mathbb{E}(|\hat{\rho}(\hat{\mathbf{h}}) - \rho(\hat{\boldsymbol{\theta}}_{\tilde{\mathbf{h}}}, \boldsymbol{\theta})|) \\ &\leq C\left(J(\mathcal{H})\frac{\sqrt{\mathbb{E}(ave(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^4)} + \sqrt{ave(\boldsymbol{\sigma}^2\boldsymbol{\theta}^2)}}{\sqrt{n}} + \mathbb{E}(|ave(\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2)|)\right).\end{aligned}$$

PROPOSITION **2** . *Let* $\tilde{\mathbf{h}}$ *be a minimizer of* $\rho(\hat{\boldsymbol{\theta}}_{\mathbf{h}}, \boldsymbol{\theta})$ *over* $\mathbf{h} \in \mathcal{H}$ *and let* $\hat{\mathbf{h}}$ *be the minimizer of* $\hat{\rho}(\mathbf{h})$ *over* $\mathbf{h} \in \mathcal{H}$*. Then*

$$\begin{aligned}\mathbb{E}\Big(ave\Big((\boldsymbol{\sigma}^2 + \boldsymbol{\theta}^2)(\hat{\mathbf{h}} - \tilde{\mathbf{h}})^2 \Big)\Big) \ &\leq \ CJ(\mathcal{H})\frac{\sqrt{\mathbb{E}(ave(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^4)} + \sqrt{ave(\boldsymbol{\sigma}^2\boldsymbol{\theta}^2)}}{\sqrt{n}} \\ &\quad + \mathbb{E}(|\, ave(\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2)|),\end{aligned}$$

and

$$\mathbb{E}\Big(ave\Big((\hat{\boldsymbol{\theta}}_{\tilde{\mathbf{h}}} - \hat{\boldsymbol{\theta}}_{\hat{\mathbf{h}}})^2 \Big)\Big) \leq CJ(\mathcal{H})\frac{\sqrt{\mathbb{E}(ave(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^4)}}{\sqrt{n}} + \mathbb{E}(|\, ave(\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2)|).$$

REMARK **2** . It follows from (5) and the Strong Law of Large Numbers that ave$(\hat{\boldsymbol{\sigma}}^2)$ is a consistent estimator of ave$(\boldsymbol{\sigma}^2)$. Hence, in view of Propositions 1 and 2, a class $\mathcal{H}$ such that $J(\mathcal{H}) = o(n^{1/2})$ together with the boundedness of $\mathbb{E}(\text{ave}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^4)$ and ave$(\boldsymbol{\sigma}^2\boldsymbol{\theta}^2)$ ensure the success of the estimator $\hat{\boldsymbol{\theta}}_{\hat{\mathbf{h}}}$. Note that since the components of $\mathbf{Y}$ are NEF-CVF distributed, boundedness of $\mathbb{E}(\text{ave}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^4)$ and ave$(\boldsymbol{\sigma}^2\boldsymbol{\theta}^2)$ follow if ave$(\boldsymbol{\theta}^4) < c$, for some $c > 0$. This 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 $\hat{\boldsymbol{\theta}}_{\hat{\mathbf{h}}}$ 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 $s > 1/2$. Indeed, in such a case we have ave$(\boldsymbol{\theta}^4) \leq \mathcal{O}(n^{-4s/(2s+1)})$ and, because of the smoothness of μ, it is easy to show that $\mathbb{E}(|\text{ave}(\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2)|) \rightarrow 0$ at a rate $n^{-2s/(2s+1)}$ as $n \rightarrow \infty$. The asymptotic rate of $\hat{\theta}_{\hat{h}}$ is then a direct application of Corollary 2.3 of Beran and Dümbgen (1998).

3.2. *Computation of the optimal modulation estimator.* Our objective now is to choose $\mathbf{h}$ to minimize $\hat{\rho}(\mathbf{h})$ defined in (6). As mentioned in Remark 2, a class $\mathcal{H}$ such that $J(\mathcal{H}) = o(n^{1/2})$ ensure the success of the estimator $\hat{\boldsymbol{\theta}}_{\hat{\mathbf{h}}}$. Various examples of modulator classes $\mathcal{H}$ to which Propositions 1 and 2 apply can be now constructed, similar to those given in Examples 1-5 of Beran and Dümbgen (1998) for the Gaussian case.

In what follows, however, we have concentrated on an estimator that is a multiple-Stein shrinkage estimator, similar to the one obtained in Example

2 of Beran and Dümbgen (1998). Let $\mathcal{B} = \mathcal{B}_n$ be a partition of T in intervals of length $[\ln(\ln n)]$, where $[\,x]$ denotes the integer part of x, and consider

$$\mathcal{H} = \left\{ \sum_{B \in \mathcal{B}} 1_B c(B) \,:\; c \in [0,1]^{\mathcal{B}} \right\}, \tag{7}$$

where 1_B is the indicator function of B. Example 2 of Beran and Dümbgen (1998) shows that we indeed get $J(\mathcal{H}) = o(n^{1/2})$. The values of $c(B)$ that define the minimizer, $\hat{\mathbf{h}}$, of $\hat{\rho}(\mathbf{h})$ over all functions $\mathbf{h}$ in the class $\mathcal{H}$ defined in (7) are given by

$$\hat{c}(B) = \frac{\operatorname{ave}(1_B(\hat{\boldsymbol{\theta}}^2 - \hat{\boldsymbol{\sigma}}^2))_+}{\operatorname{ave}(1_B\hat{\boldsymbol{\theta}}^2)}, \tag{8}$$

where $(u)_+ = \max(u, 0)$. Inspection of (8) shows that, for each $t \in B$, $\hat{h}(t) = 0$ when $\operatorname{ave}(1_B\hat{\theta}^2(t)) \leq \operatorname{ave}(1_B\hat{\sigma}^2(t))$. Hence, the modulator is set to 0 when the signal-to-noise ratio is less than 1. Moreover, for each $t \in B$, the modulator tends to 1 as the signal-to-noise ratio tends to 1. (Note that for sample sizes $n \leq 1024$, or even $n \leq 2048$, the partition $\mathcal{B} = \mathcal{B}_n$ considered above is such that only two consecutive sampling points are required to compute the value of the modulator $\hat{\mathbf{h}}$ over B.)

The derivation of the values $\hat{c}(B)$ in (8) that define the optimal modulator rely upon a realization of a pseudo-sample $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_p\}$ which is not observable. Furthermore, one would expect better results when the size p of this pseudo-sample is large. We now show that the limiting (as $p \to \infty$) form of the optimal modulator can be approximated by an expression computed directly from the original data. For each $t \in T$, only $\hat{\sigma}^2(t)$ in (8) depends on the pseudo-sample $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_p\}$. Recall from (5) and the Strong Law of Large Numbers that, for each $t \in T$, $\hat{\boldsymbol{\sigma}}^2(t)$ is a consistent estimator of $\boldsymbol{\sigma}^2(t)$ and that $\mathbb{E}(\hat{\sigma}^2(t)) = \sum_{l=1}^n w_{t,l}^2 V(\mu(l))$. If now, for each $t \in T$, $\widehat{V(\mu(t))}$ is an estimator of $V(\mu(t))$ (such as, for example, a uniform minimum variance unbiased estimator (UMVUE)), an intuitive appealing approximation of the values of $c(B)$ that define our optimal modulator, overcoming the fact that the pseudo-sample is never observed, takes the form

$$\hat{c}(B) = \frac{\operatorname{ave}(1_B(\hat{\boldsymbol{\theta}}^2 - \tilde{\boldsymbol{\sigma}}^2))_+}{\operatorname{ave}(1_B\hat{\boldsymbol{\theta}}^2)},$$

where

$$\tilde{\sigma}^2(t) = \sum_{l=1}^{n} w_{t,l}^2 \widehat{V(\mu(l))}, \quad t \in T. \tag{9}$$

Finally, our optimal (nonlinear) modulation estimator is given by

$$\hat{\boldsymbol{\theta}}_{\hat{\mathbf{h}}} = \sum_{B \in \mathcal{B}} \frac{\operatorname{ave}(1_B(\hat{\boldsymbol{\theta}}^2 - \tilde{\boldsymbol{\sigma}}^2))_+}{\operatorname{ave}(1_B \hat{\boldsymbol{\theta}}^2)} \, 1_B \, \hat{\boldsymbol{\theta}}. \tag{10}$$

We now discuss how to obtain, for each $t \in T$, the UMVUE $\widehat{V(\mu(t))}$ of the variance function $V(\mu(t))$ of a NEF-CVF($\mu(t)$,$V(\mu(t))$) that is needed to obtain $\tilde{\sigma}^2(t)$ given in (9). Kokonendji and Seshadri (1996) have obtained the Rao-Blackwell estimator of the determinant of the variance of a NEF on $\mathbb{R}^d$ based on $(d+1)$ observations. Therefore, the UMVUE of the variance function of a given (univariate) NEF can be easily obtained by applying Theorems 2.2 and 3.3 of Kokonendji and Seshadri (1996) with $d = 1$. However, for the (univariate) NEF-CVF, the UMVUE does not have a closed expression and one has to perform the computation for each type separately (Letac, private communication).

Obviously, in the univariate case, one needs 2 observations for each $t \in T$ and therefore, as Kokonendji and Seshadri (1996) pinpoint, their results in the case of NEF-QVF can be regarded as a partial generalization of the UMVUE obtained by Morris (1983) (one needs 1 observation for each $t \in T$). Therefore, in this case, we suggest the type of the estimator $\widehat{V(\mu(t))}$ and the resulting optimal modulation estimator obtained by Antoniadis and Sapatinas (2001).

For each of the remaining six distributions of a NEF-CVF, we suggest to obtain, for each $t \in T$, a UMVUE $\widehat{V(\mu(t))}$ of the variance function $V(\mu(t))$ by introducing a further step of binning the data to generate multiple observations. More specifically, the data are divided into bins of equal size $L = 2^m$ (for some $m > 0$) and, for each bin, we calculate a UMVUE using a sample of size L. Then, in each bin, we estimate the corresponding variances by the same UMVUE (obtained for that particular bin) and calculate (9) and (10). The binning is, therefore, important and we study its effect on the performance of our optimal modulation estimator by conducting a simulation study using inverse Gaussian data in Section 4 below.

We end this section saying that (10) involves the projection of the variance function estimate $\widehat{V(\mu(t))}$ onto the pointwise square of the wavelet basis functions (see (9)). 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 and Corradi (1999) or Kovac and Silverman (2000). The resulting *squared discrete wavelet transform* (SDWT) of a signal sampled at $n = 2^J$, for some $J > 0$, fixed equidistant points requires only order n operations, so is computationally fast.

4. Simulation Study

The inverse Gaussian distribution is widely used in modelling a variety of phenomena involving first passage times in biology, ecology, reliability and survival analysis, to mention a few. For an excellent monograph on this distribution and its statistical applications, see Seshadri (1999). The probability density function (pdf) of an inverse Gaussian distribution with mean parameter μ and shape parameter λ is given by

$$f(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0; \quad \mu > 0,\ \lambda > 0. \tag{11}$$

The inverse Gaussian distribution is the most famous one among the six extra types of distributions obtained from a NEF-CVF. Thus, in order to gauge the performance of the proposed wavelet shrinkage methodology (we call this procedure CVFCV), we have conducted a simulation study by considering inverse Gaussian distributed times series data, i.e.

$$Y(t) \sim \text{inverse Gaussian}(\mu(t), V(\mu(t))), \quad t \in T,$$

where $V(\mu(t)) = \mu^3(t)/\lambda(t)$. For such data, three factors of interest were included in the study: the morphology of the mean function $\mu(t)$, the bin size L involved in the construction of the UMVUE of the variance $V(\mu(t))$ (as discussed in Section 3.2), and the inverse of coefficient of variation (ICV) $\rho(t) = \lambda(t)/\mu^2(t)$ (i.e the ratio mean/variance).

Motivated by various phenomena of similar nature that are often encountered in practice, the underlying mean function was allowed to take two different shapes: 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

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

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

$$\mu(t) = A + 2[6.75t^2(1-t)]^3,$$

where A, A_1, A_2, A_3, $t_{i,\max}$, ν, σ_r and σ_d are prespecified constants. These functions were used by Antoniadis and Sapatinas (2001) to illustrate their wavelet shrinkage methodology for various types of distributions arising from a NEF-QVF (they called this procedure QVFCV).

The UMVUE of the variance $V(\mu(t)) = \mu^3(t)/\lambda(t)$ of a random variable with pdf (11) for a sample size n greater than or equal to 2 has been obtained by Korwar (1980) and Iwase and Setô (1983). To study the effect of binning in the performance of the resulting modulation estimator for the inverse Gaussian case, we have considered two possibilities: $L = 2$ (sample of size 2) and $L = 4$ (sample of size 4). Using the notation of Iwase and Setô (1983) and replacing n with L, these are given respectively by

$$\hat{k}_2 = \frac{2\bar{X}^3 V}{V\bar{X} + 2} \quad \text{and} \quad \hat{k}_4 = \frac{4\bar{X}^2(1 - \arctan(V\bar{X}/4)^{1/2})}{(V\bar{X})^{1/2}},$$

where

$$\bar{X} = \frac{1}{L}\sum_{i=1}^{L} X_i \quad \text{and} \quad V = \sum_{i=1}^{L}(X_i^{-1} - \bar{X}^{-1}).$$

To gain some insight into the behaviour of the proposed wavelet shrinkage estimation procedure at low, medium and high ICV, $\rho(t)$ admitted the values 1, 3 and 5 per time point. Finally, each mean function was sampled at $N = 256, 512, 1024$ equidistant data points, i.e. $T = \{t_i = i/n,\ i = 1, \ldots, n = N\}$. Figures 1 and 2 show respectively the two mean functions with a medium ICV ($\rho(t) = 3$) for the two bin sizes ($L = 2, 4$) and estimates from a single trial, using the CVFCV procedure based on $N = 256$ equispaced data points. At each of the 18 ($= 2\times3\times3$) design points, and for the 2 ($L = 2, 4$) different bin sizes, estimates were calculated using CVFCV over 500 trials. The results of these simulations are shown in Figures 3 and 4. The method was based on Daubechies' nearly symmetric wavelets of order 8 (see, Daubechies, 1992, p. 195). Random variates from an inverse Gaussian distribution were generated using the method of transformations with multiple roots (see, for example, Michael *et al.*, 1976).

As one can see from these figures, the CVFCV procedure produces estimates with satisfactory mean squared errors uniformly across the various combinations of morphology and ICV for bin size $L = 2$. It is evident that the results are improving considerably as we increase the number of grid points and as we move from low to high ICV. Increasing the bin size ($L = 4$), it has a profound effect on the performance of the CVFCV estimator for the burst function. However, the results are getting slightly better as we increase the number of grid points and as we move from low to high ICV. As it is expected, the binning does not affect substantially the performance of the CVFCV estimator for the smooth function for all combinations of grid designs and ICV.

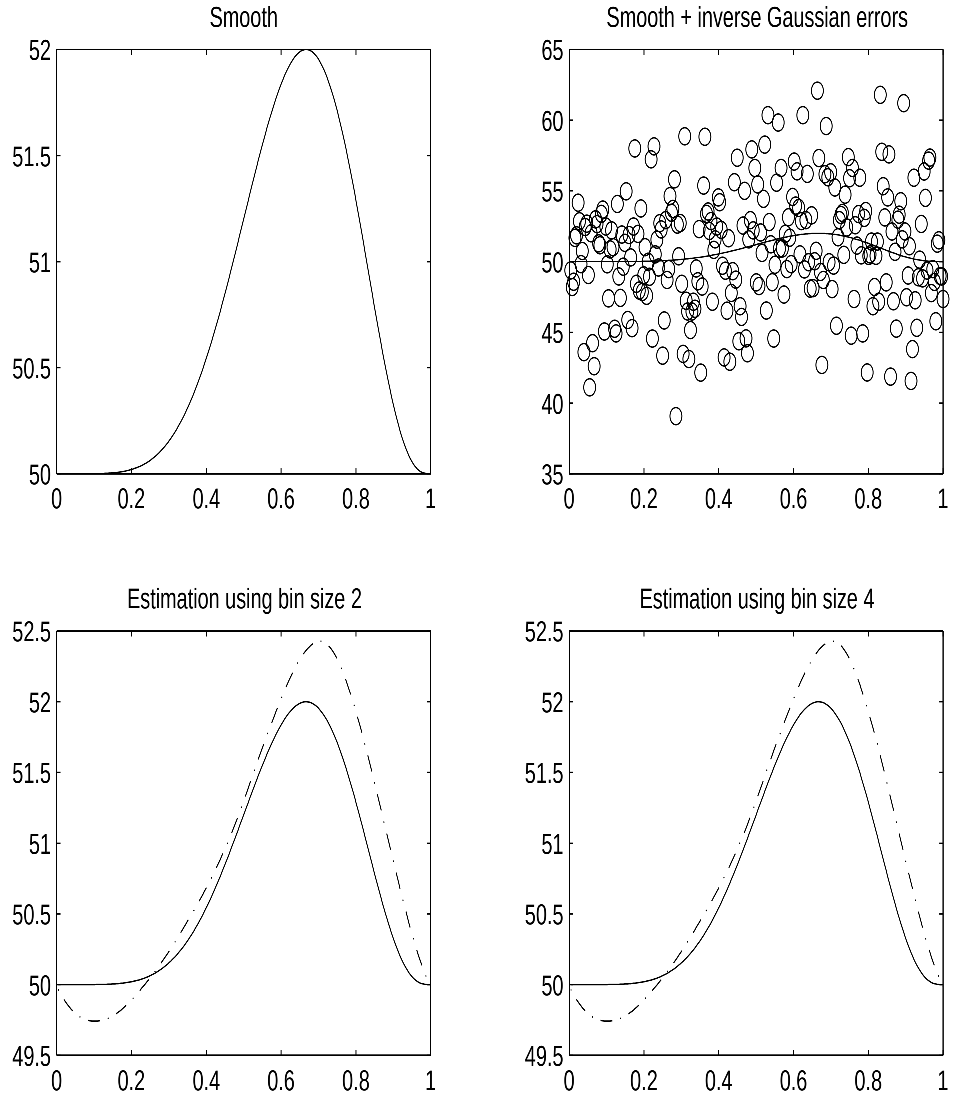

Figure 1. The smooth function (solid), along with estimates from a single trial, using CVFCV (dot-dashed) based on $N = 256$ equispaced data points with a medium ICV ($\rho(t) = 3$) and the two bin sizes ($L = 2, 4$).

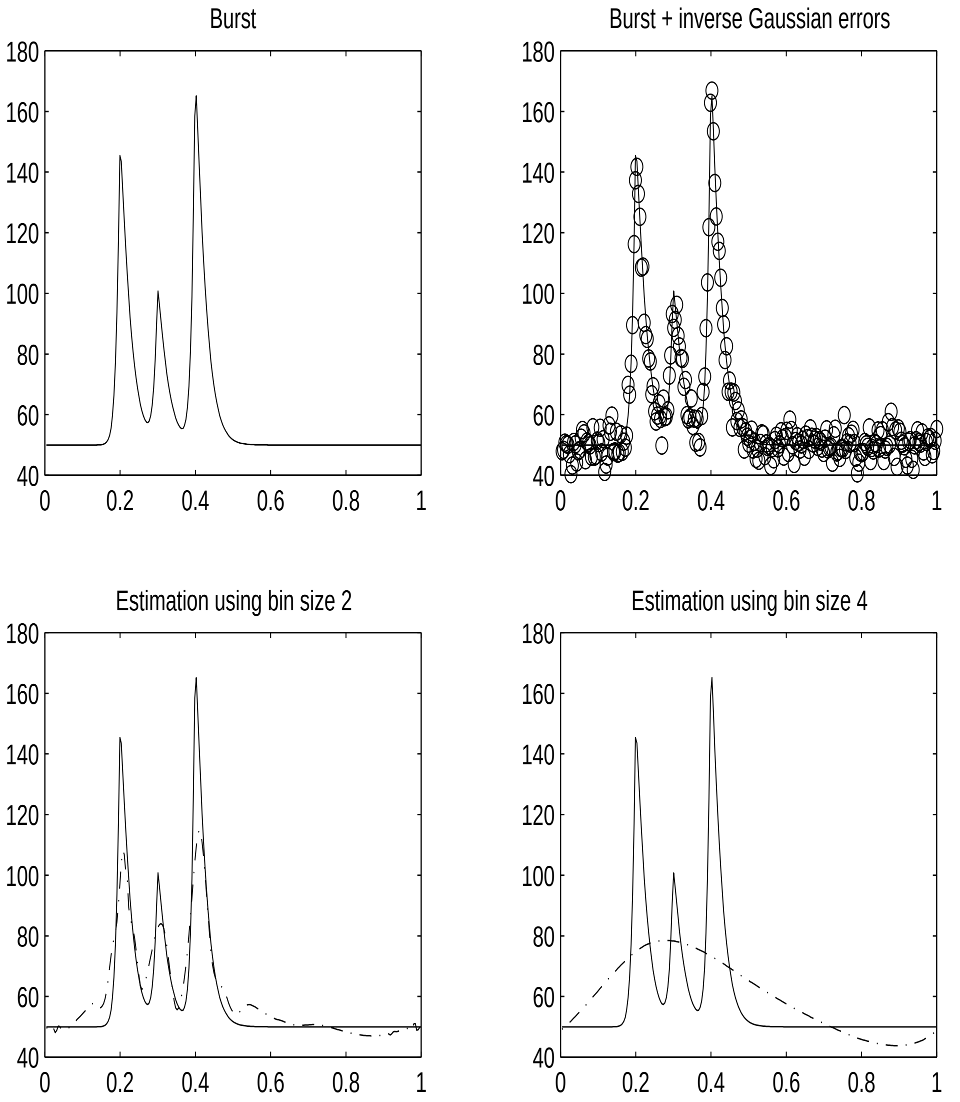

Figure 2. The burst function (solid), along with estimates from a single trial, using CVFCV (dot-dashed) based on $N = 256$ equispaced data points with a medium ICV ($\rho(t) = 3$) and the two bin sizes ($L = 2, 4$).

Acknowledgements. The authors would like to thank Marina Vannucci (Department of Statistics, Texas A&M University, USA) who kindly provided her Matlab routine for computing the variance-covariance matrix of the wavelet coefficients, and Professor Gérard Letac (Laboratoire de

Statistique et Probabilités, Université de Toulouse, France) for constructive comments of the UMVUE of the variance function of a NEF-CVF. We would like also to thank the two anonymous referees and Professor Prakash Patil (the Editor in charge) whose valuable comments and recommendations helped to improve this article.

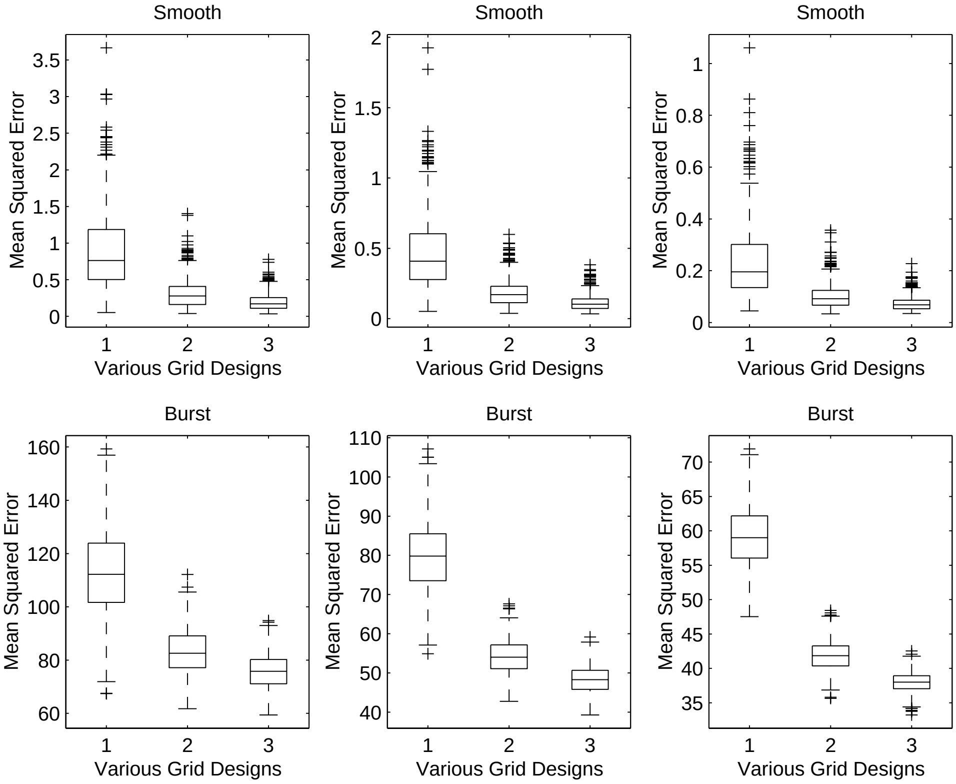

Figure 3. Boxplots of inverse Gaussian simulation results for the smooth function (top) and burst function (bottom). In each panel, there is a triplet of boxplots indicating the mean squared error of the estimates produced by CVFCV over 500 trials using (1) $N = 256$, (2) $N = 512$ and (3) $N = 1024$ equidistant data points. The ICV was taken, in each row as $\rho(t) = 1$ (left panel), $\rho(t) = 3$ (middle panel) and $\rho(t) = 5$ (right panel). All simulations were conducted using bin size $L = 2$.

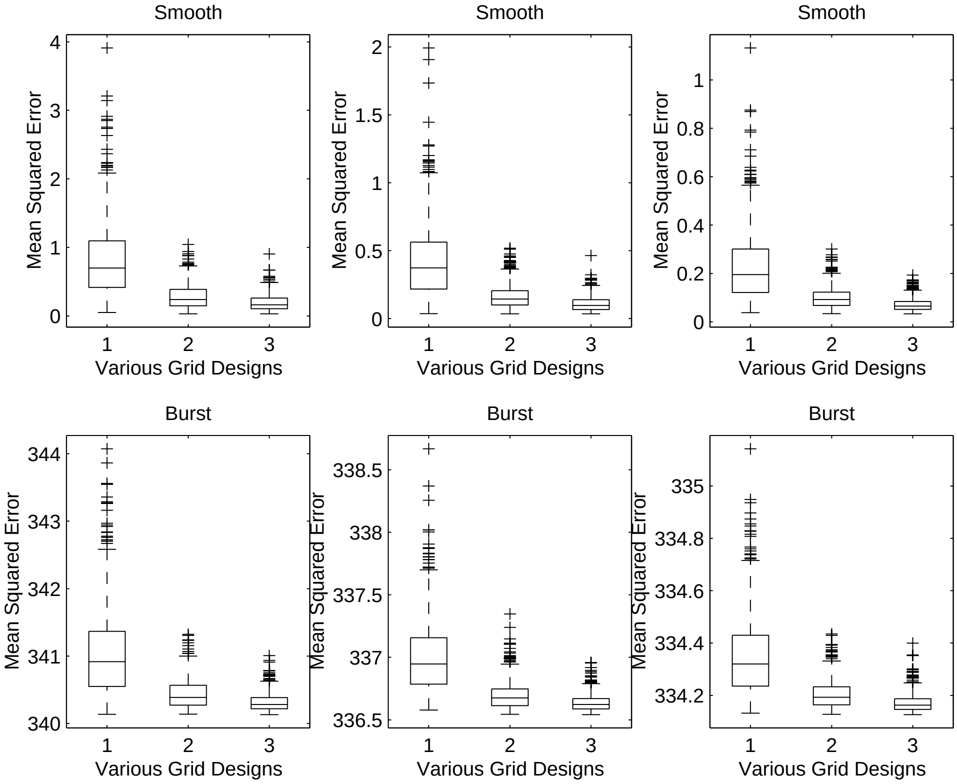

Figure 4. Boxplots of inverse Gaussian simulation results for the smooth function (top) and burst function (bottom). In each panel, there is a triplet of boxplots indicating the mean squared error of the estimates produced by CVFCV over 500 trials using (1) $N = 256$, (2) $N = 512$ and (3) $N = 1024$ equidistant data points. The ICV ratio was taken, in each row as $\rho(t) = 1$ (left panel), $\rho(t) = 3$ (middle panel) and $\rho(t) = 5$ (right panel). All simulations were conducted using bin size $L = 4$.

References

Abramovich, F., Bailey, T.C. and 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**, 246–254.

Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion), *Journal of the Italian Statistical Society*, Series B, **6**, 97–144.

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

Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions, *Biometrika*, **88**, 805–820.

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study, *Journal of Statistical Software*, **6**, Issue 6, 1–83.

Beran, R. and Dümbgen, L. (1998). Modulation of estimators and confidence sets, *Annals of Statistics*, **26**, 1826–1856.

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

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**, Daubechies, I. ed., pp. 173-205, San Antonio: American Mathematical Society.

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

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

Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage, *Annals of Statistics*, **26**, 879–921.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion), *Journal of the Royal Statistical Society*, Series B, **57**, 301–337.

Dudley, R.M. (1987). Universal Donsker classes and metric entropy, *Annals of Probability*, **15**, 1306–1326.

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

Iwase, K. and Setô, N. (1983). Uniformly minimum variance unbiased estimation for the inverse Gaussian distribution, *Journal of the American Statistical Association*, **78**, 660–663.

Kokonendji, C.C. and Seshadri, V. (1996). On the determinant of the second derivative of a Laplace transform, *Annals of Statistics*, **24**, 1813–1827.

Kolaczyk, E.D. (1997). Non-parametric estimation of Gamma-Ray burst intensities using Haar wavelets, *Astrophysical Journal*, **483**, 340–349.

Kolaczyk, E.D. (1999a). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds, *Statistica Sinica*, **9**, 119–135.

Kolaczyk, E.D. (1999b). Bayesian multiscale models for Poisson processes, *Journal of the American Statistical Association*, **94**, 920–933.

Korwar, R.M. (1980). On the uniformly minimum variance unbiased estimators of the variance and its reciprocal of an inverse Gaussian distribution, *Journal of the American Statistical Association*, **75**, 734–735.

Kovac, A. and Silverman, B.W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding, *Journal of the American Statistical Association*, **95**, 172–183.

Letac, G. and Mora, M. (1990). Natural real exponential families with cubic variance functions, *Annals of Statistics*, **18**, 1–37.

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

Michael, J.R., Schucany, W.R. and Haas, R.W. (1976). Generating random variates using transformations with multiple roots, *American Statistician*, **30**, 88–90.

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

Morris, C.N. (1983). Natural exponential families with quadratic variance functions: statistical theory, *Annals of Statistics*, **11**, 515–529.

Nowak, R.D. (1997). Optimal signal estimation using cross-validation, *IEEE Signal Processing Letters*, **4**, 23–25.

Nowak, R.D. and Baraniuk, R.G. (1999). Wavelet domain filtering for photon imaging systems, *IEEE Transactions on Image Processing*, **8**, 666–678.

Seshadri, V. (1999). *The Inverse Gaussian Distribution: Statistical Theory and Applications.* Lecture Notes in Statistics **137**, New York: Springer-Verlag.

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

Vannucci, M. and Corradi, F. (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective, *Journal of the Royal Statistical Society*, Series B, **61**, 971–986.

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

Anestis Antoniadis
Laboratoire IMAG-LMC
University Joseph Fourier
BP 53, 38041 Grenoble Cedex 9
France.
E-mail: Anestis.Antoniadis@imag.fr

Panagiotis Besbeas
Institute of Mathematics and Statistics
University of Kent at Canterbury
Canterbury, Kent CT2 7NF
United Kingdom.
E-mail: P.T.Besbeas@ukc.ac.uk

Theofanis Sapatinas
Department of Mathematics and Statistics
University of Cyprus
P.O. Box 20537
CY 1678 Nicosia
Cyprus.
E-mail: T.Sapatinas@ucy.ac.cy