

# Wavelet Thresholding via a Bayesian Approach

F. Abramovich<sup>1</sup>, T. Sapatinas<sup>2</sup> and B. W. Silverman<sup>2</sup>

<sup>1</sup>Department of Statistics & Operations Research,  
Raymond & Beverly Sackler Faculty of Exact Sciences,  
University of Tel Aviv, Ramat Aviv 69978, Israel.

<sup>2</sup>Department of Mathematics,  
University of Bristol, Bristol BS8 1TW, UK.

Submitted: November 6, 1996

Final revision: October 16, 1997

## Abstract

We discuss a Bayesian formalism which gives rise to a type of wavelet threshold estimation in non-parametric regression. A prior distribution is imposed on the wavelet coefficients of the unknown response function, designed to capture the sparseness of wavelet expansion common to most applications. For the prior specified, the posterior median yields a thresholding procedure. Our prior model for the underlying function can be adjusted to give functions falling in any specific Besov space. We establish a relation between the hyperparameters of the prior model and the parameters of those Besov spaces within which realizations from the prior will fall. Such a relation gives insight into the meaning of the Besov space parameters. Moreover, the established relation makes it possible in principle to incorporate prior knowledge about the function's regularity properties into the prior model for its wavelet coefficients. However, prior knowledge about a function's regularity properties might be hard to elicit; with this in mind, we propose a standard choice of prior hyperparameters that works well in our examples. Several simulated examples are used to illustrate our method, and comparisons are made with other thresholding methods. We also present an application to a data set collected in an anaesthesiological study.

**Keywords:** Adaptive Estimation; Anaesthetics; Bayes Model; Besov Spaces; Non-parametric Regression; Thresholding; Wavelet Transform.

# 1 Introduction

Consider the standard non-parametric regression problem:

$$y_i = g(t_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where  $t_i = i/n$ ,  $\varepsilon_i$  are independent identically distributed normal variables with zero mean and variance  $\sigma^2$ , and we wish to recover the unknown function  $g$  from the noisy data without assuming any particular parametric form. There are several approaches to the non-parametric estimation of the unknown function  $g$  such as spline smoothing, kernel estimation, generalized Fourier series expansion, etc. In this paper we consider wavelet-based estimators of  $g$ . The function  $g$  is expanded in wavelet series in a way similar to the generalized Fourier series approach. The advantage of the wavelet basis is its ‘universality’ – functions from a wide set of function spaces, such as Besov or Triebel spaces, have a parsimonious representation in wavelet series. The usual approach is to expand the noisy data in wavelet series, extract the ‘significant’ wavelet coefficients by thresholding, and then to invert the wavelet transform of the de-noised coefficients. Donoho & Johnstone (1994, 1995), Donoho *et al.* (1995) showed that such wavelet estimators with a properly chosen threshold rule have various important optimality properties. The choice of thresholding rule, therefore, becomes a crucial step in the estimation procedure. Several approaches to thresholding have been introduced in the literature: a minimax approach (Donoho & Johnstone, 1994, 1995), multiple hypothesis testing (Abramovich & Benjamini, 1995, 1996; Ogden & Parzen, 1996a, 1996b), cross-validation (Nason, 1995, 1996; Weyrich & Warhola, 1995). The idea of thresholding has also been studied in the context of correlated errors  $\varepsilon_i$ ; see, for example, Wang (1996), Johnstone & Silverman (1997).

In this paper we consider thresholding within a Bayesian framework. In the Bayesian approach a prior distribution is imposed on the wavelet coefficients of the unknown response function. The prior model is designed to capture the sparseness of wavelet expansion common to most applications. Then, the function is estimated by applying some Bayes rules on the resulting posterior distribution of the wavelet coefficients. The traditional Bayes rule considered in the literature (Vidakovic, 1994; Clyde *et al.*, 1995; Chipman *et al.*, 1997) corresponds to an  $L^2$ -loss (the posterior mean) based on the wavelet coefficients. However, such a rule is not a thresholding rule but a shrinkage. In order to fix terminology, we mention that a *shrinkage rule* is a function that decreases (not necessarily strictly) the absolute values of

the wavelet coefficients, without changing their sign. For a rule to be a *thresholding rule*, it must not only shrink the coefficients towards zero, but must also map actually to zero all coefficients falling in some non-empty interval around zero.

In this paper, instead of the  $L^2$ -loss, we suggest the use of a weighted combination of  $L^1$ -losses based on the wavelet coefficients. These losses correspond to  $L^1$ -losses based on the function and on its derivatives; such losses are natural measures for spatially inhomogeneous functions. The corresponding Bayes rule will be the posterior median and, for a certain prior, yields a thresholding procedure.

The paper is organized as follows: in Section 2 we briefly review the discrete wavelet transform and the relevant aspects of the Donoho-Johnstone work on the non-parametric regression problem. A review of relevant aspects of Besov spaces is also given. In Section 3, we study the problem of wavelet thresholding within a Bayesian approach. In Section 4, we discuss the form for the hyperparameters of the prior model, and demonstrate a relation between Besov space parameters and hyperparameters of the prior model. The implications of this relation for the choice of prior hyperparameters are discussed. In addition, since it may be difficult in practice to elicit prior knowledge of the regularity properties of the function, we also propose a ‘standard’ choice of hyperparameters, which in our experience works well on a range of examples. In Section 5, we provide several simulated examples to illustrate our method, and give comparisons with other thresholding methods. We also present an application to a data set collected in an anaesthesiological study. Some concluding remarks are made in Section 6, and the more technical details and proofs are given in the Appendix.

## 2 Wavelet estimators

### 2.1 Wavelet overview

Wavelet series are generated by dilations and translations of a function  $\psi$ , called the *mother wavelet*:  $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$ ,  $j, k \in \mathbb{Z}$ . For suitable choices of  $\psi$ , the corresponding set of  $\psi_{jk}$  forms an orthonormal basis in  $L^2(\mathbb{R})$ . Examples of mother wavelets with different regularity properties, number of vanishing moments, and compact support can be found in Daubechies (1992). The wavelet series representation of a function  $g \in$

$L^2(\mathbb{R})$  is then:

$$g(t) = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} w_{jk} \psi_{jk}(t),$$

where the wavelet coefficients  $w_{jk}$  are given by:

$$w_{jk} = \int_{\mathbb{R}} g(t) \psi_{jk}(t) dt.$$

Intuitively, the  $\psi_{jk}$  represent “smooth wiggly functions” localized to spatial positions near  $k2^{-j}$  and frequencies near  $2^j$ . In contrast to standard Fourier sine and cosine series, wavelets are local in both frequency/scale (via dilations) and in time (via translations). This localization allows parsimonious representation for a wide set of different functions in wavelet series.

In technical terms this corresponds to the property that, by choosing the mother wavelet with corresponding regularity properties, one can generate an unconditional wavelet basis in a wide set of function spaces, such as Besov (see Section 2.2 below) or Triebel spaces. For a clear and accessible introduction to wavelets we refer to Strang (1993). Jawerth & Sweldens (1994) provide an excellent overview of wavelet-based multiresolution analyses. Meyer (1992) and Daubechies (1992) give detailed expositions of the mathematical aspects of wavelets.

In many practical situations, the functions involved are only defined on a compact set, such as the interval  $[0, 1]$ , and to apply wavelets then requires some modifications. Cohen *et al.* (1993) have obtained the necessary boundary corrections to retain orthonormality. Their wavelets also constitute unconditional bases for the Besov and Triebel spaces on the interval. In later sections, however, we confine attention to periodic functions on  $\mathbb{R}$  with unit period and work in effect with periodic wavelets. In this case, the wavelet coefficients  $w_{jk}$  of the function are actually restricted to the resolution and spatial indices  $j \geq 0$  and  $k = 0, \dots, 2^j - 1$ , respectively; there is also the coarsest scaling coefficient, which is labelled as  $u_{00}$  (see, for example, Daubechies, 1992, Section 9.3). As Johnstone (1994) points out, this computational simplification affects only a fixed number of wavelet coefficients at each resolution level and does not alter the qualitative phenomena we wish to present.

## 2.2 Besov spaces on the interval

In this section, we briefly introduce some relevant aspects of the (inhomogeneous) Besov spaces on the interval that we need further. For a more

detailed study we refer to DeVore & Popov (1988), Triebel (1990), DeVore *et al.* (1992) and Meyer (1992).

Let the  $r$ -th difference of a function  $g$  be:

$$\Delta_h^{(r)} g(t) = \sum_{k=0}^r \binom{r}{k} (-1)^k g(t + kh),$$

and let the  $r$ -th modulus of smoothness of  $g$  in  $L^p[0, 1]$  be:

$$\nu_{r,p}(g; t) = \sup_{h \leq t} \|\Delta_h^{(r)} g\|_{L^p[0, 1- rh]}.$$

Then the Besov seminorm of index  $(s, p, q)$  is defined for  $r > s$ , where  $1 \leq p, q \leq \infty$ , by:

$$|g|_{B_{p,q}^s} = \left\{ \int_0^1 \left( \frac{\nu_{r,p}(g; h)}{h^s} \right)^q \frac{dh}{h} \right\}^{1/q}, \quad \text{if } 1 \leq q < \infty,$$

and by:

$$|g|_{B_{p,\infty}^s} = \sup_{0 < h < 1} \left\{ \frac{\nu_{r,p}(g; h)}{h^s} \right\}.$$

Define the Besov norm as  $\|g\|_{B_{p,q}^s} = \|g\|_{L^p[0,1]} + |g|_{B_{p,q}^s}$ . The Besov space  $B_{p,q}^s$  is then the class of functions  $g : [0, 1] \rightarrow \mathbb{R}$  satisfying  $g \in L^p[0, 1]$  and  $|g|_{B_{p,q}^s} < \infty$ . The parameter  $s$  measures the number of derivatives, where the existence of derivatives is required in an  $L^p$ -sense, while the parameter  $q$  provides a further finer gradation.

The Besov spaces include, in particular, the well-known Sobolev and Hölder spaces of smooth functions  $H^m$  and  $C^s$  ( $B_{2,2}^m$  and  $B_{\infty,\infty}^s$  respectively), but in addition less traditional spaces, like the space of functions of bounded variation, sandwiched between  $B_{1,1}^1$  and  $B_{1,\infty}^1$ . The latter functions are of statistical interest because they allow for better models of spatial inhomogeneity (e.g. Meyer, 1992; Donoho & Johnstone, 1995).

The Besov norm for the function  $g$  is related to a sequence space norm on the wavelet coefficients of the function. As noted in Section 2.1, confining attention to the resolution and spatial indices  $j \geq 0$  and  $k = 0, \dots, 2^j - 1$ ,

respectively, the sequence space norm is given by:

$$\begin{aligned} \|w\|_{b_{p,q}^s} &= |u_{00}| + \left\{ \sum_{j=0}^{\infty} 2^{js'q} \left( \sum_{k=0}^{2^j-1} |w_{jk}|^p \right)^{q/p} \right\}^{1/q}, \quad \text{if } 1 \leq q < \infty, (2) \\ \|w\|_{b_{p,\infty}^s} &= |u_{00}| + \sup_{j \geq 0} \left\{ 2^{js'} \left( \sum_{k=0}^{2^j-1} |w_{jk}|^p \right)^{1/p} \right\}, \end{aligned} \quad (3)$$

where  $s' = s + 1/2 - 1/p$  (see, for example, Donoho *et al.*, 1995).

If the mother wavelet  $\psi$  is of regularity  $r$ , where  $\max(0, 1/p - 1/2) < s < r$ , then we have:

$$K_1 \|g\|_{B_{p,q}^s} \leq \|w\|_{b_{p,q}^s} \leq K_2 \|g\|_{B_{p,q}^s},$$

where  $K_1$  and  $K_2$  are constants, not depending on  $g$  (e.g. Meyer, 1992; Donoho & Johnstone, 1995). Therefore the Besov norm of  $g$  is equivalent to the corresponding sequence space norm (2) or (3). In Section 4.3, we exploit the equivalence of the norms for relating prior information about the function's regularity to hyperparameters of our prior model for the wavelet coefficients  $w_{jk}$ .

In the particular case  $p = q = 1$  the sequence space norm in (2) becomes a weighted sum of the  $|w_{jk}|$  and the corresponding Besov space norm is essentially an  $L^1$ -norm on the derivatives of  $g$  up to order  $s$ . This will provide motivation for the loss function we use in Section 3.

### 2.3 Discrete wavelet transform and thresholding

In practice, given observed discrete data  $\mathbf{y} = (y_1, \dots, y_n)^\top$  from model (1), we may find the vector  $\hat{\mathbf{d}}$  of its sample discrete wavelet coefficients by performing the *discrete wavelet transform* (DWT) of  $\mathbf{y}$ :

$$\hat{\mathbf{d}} = \mathcal{W}\mathbf{y},$$

where  $\mathcal{W}$  is the DWT-matrix with  $(jk, i)$  entry given by  $\sqrt{n} W_{jk,i} \approx \psi_{jk}(i/n) = 2^{j/2} \psi(2^j i/n - k)$ .

The population discrete wavelet coefficients  $d_{jk}$  are defined as the DWT of the vector of function values  $g(t_i)$ ,  $i = 1, \dots, n$ . These are related to the wavelet coefficients  $w_{jk} = \int_{\mathbb{R}} g(t) \psi_{jk}(t) dt$  by  $d_{jk} \approx \sqrt{n} w_{jk}$ . The  $\sqrt{n}$  factor essentially arises from the difference between continuous and discrete orthogonality conditions. Since the definitions of DWT and of the coefficients

$w_{jk}$  are by now standard, this factor cannot be avoided and therefore we use different letters  $d_{jk}$  and  $w_{jk}$  to clarify the distinction.

If  $n = 2^J$  (for some positive integer  $J$ ) then both DWT and inverse DWT are performed by Mallat's (1989) fast algorithm that requires only  $O(n)$  operations and is available in several standard implementations, for example, in the S-plus package `WaveThresh` (Nason, 1993; Nason & Silverman, 1994). The `WaveThresh` package implements a periodized form of the DWT that produces  $(n - 1)$  sample discrete wavelet coefficients  $\hat{d}_{jk}$ ,  $j = 0, \dots, J - 1$ ;  $k = 0, \dots, 2^j - 1$ , and one sample scaling coefficient, which is labelled as  $\hat{c}_{00}$ . Each  $\hat{d}_{jk}$  describes the contribution around spatial location  $2^{-j}k$  and near frequency  $2^j$ , while  $\hat{c}_{00}$  is the sample mean multiplied by  $\sqrt{n}$ . Because of the orthogonality of  $\mathcal{W}$ , the DWT of a white noise is also an array  $\varepsilon_{jk}$  of independent  $N(0, \sigma^2)$  random variables and, hence, equally contaminates the population discrete wavelet coefficients  $d_{jk}$ :

$$\hat{d}_{jk} = d_{jk} + \varepsilon_{jk}, \quad j = 0, \dots, J - 1, \quad k = 0, \dots, 2^j - 1.$$

The next step is to extract those coefficients that really contain information about the unknown function  $g$  and discard the others. This can be done by thresholding the sample discrete wavelet coefficients  $\hat{d}_{jk}$ . The intuitive idea is that the true function  $g$  has a parsimonious wavelet expansion, i.e. only a few 'large'  $\hat{d}_{jk}$  essentially contain real information about  $g$ . If we can decide which ones these are, we can estimate them and set all the others equal to zero.

Donoho and Johnstone (1994, 1995) proposed the *hard* and *soft* thresholding rules:

$$T_{\text{hard}}(\hat{d}_{jk}, \lambda) = \hat{d}_{jk} \mathbf{I}(|\hat{d}_{jk}| > \lambda), \quad (4)$$

$$T_{\text{soft}}(\hat{d}_{jk}, \lambda) = \text{sign}(\hat{d}_{jk}) \max(0, |\hat{d}_{jk}| - \lambda), \quad (5)$$

where  $\lambda \geq 0$  is a threshold parameter and  $\mathbf{I}$  is the usual indicator function. The hard thresholding is a 'keep' or 'kill' rule, while the soft thresholding is a 'shrink' or 'kill' rule. In applications, hard thresholding generally reproduces peak heights and discontinuities better, but at some cost in visual smoothness (Donoho & Johnstone 1994, 1995). By defining  $d_{jk}^{\text{new}} = T_{\text{hard}}(\hat{d}_{jk}, \lambda)$  or  $d_{jk}^{\text{new}} = T_{\text{soft}}(\hat{d}_{jk}, \lambda)$ , one can then reconstruct  $\hat{\mathbf{g}}$  by the inverse DWT:

$$\hat{\mathbf{g}} = \mathcal{W}^T \mathbf{d}^{\text{new}}.$$

The choice of  $\lambda$  is therefore crucial: if the threshold is too small or too large then the wavelet shrinkage estimator will tend to overfit or underfit

the data. Donoho & Johnstone (1994) proposed the *universal* threshold  $\lambda_{DJ} = \sigma \sqrt{2 \log(n)}$  called by them as *VisuShrink*. Despite the simplicity of such a threshold, they showed that the resulting nonlinear wavelet estimator is spatially adaptive and is asymptotically near-minimax within the whole range of Besov spaces. Moreover, Donoho & Johnstone (1997) proved that it asymptotically outperforms any linear estimator (i.e. splines, kernel estimators, truncated Fourier series, etc.) within Besov spaces  $B_{p,q}^s$  with  $p < 2$  that contain spatially inhomogeneous functions. However, the universal threshold depends on the data only through the estimated  $\sigma$  and is otherwise the same for all kinds of functions. It tends to oversmooth in practice, since it does not compromise between signal and noise. In practice, for finite samples, Donoho & Johnstone (1994, 1995) suggest keeping coefficients on the lower ‘coarse’ levels, even if these coefficients do not pass the threshold level. Several *data-driven* thresholding rules have been developed recently. Donoho & Johnstone (1995) proposed a *SureShrink* thresholding which is based on minimizing Stein’s unbiased risk estimate (Stein, 1981) and will usually yield smaller thresholds than the *VisuShrink* method. They have shown that *SureShrink* is also asymptotically near-minimax and the computational effort of the overall procedure is  $O(n \log(n))$ . For a practical demonstration of the advantages of this approach see Johnstone & Silverman (1997). In another development, Nason (1995, 1996) adjusted the well known cross-validation approach for choosing the threshold level. Some possible extensions on Nason’s method are described in Weyrich & Warhola (1995) and Wang (1996). Abramovich & Benjamini (1995, 1996), Ogden & Parzen (1996a, 1996b) considered thresholding as multiple hypotheses testing procedure: for every wavelet coefficient test simultaneously whether it is zero or not. Johnstone & Silverman (1997) have developed a *level-dependent* threshold approach for data with correlated noise, and some of the above approaches can be extended to this case. A Bayesian viewpoint to thresholding was introduced by Vidakovic (1994), Clyde *et al.* (1995), Chipman *et al.* (1997) and will be discussed in detail further in the paper.

### 3 Thresholding within a Bayesian framework

Most of the existing thresholding procedures are essentially minimax and, therefore, they may be ‘too universal’; they do not take into account some specific properties of a concrete  $g$  that we are interested in. A natural way of using the prior belief (knowledge, information) about the unknown  $g$  (say,

its regularity properties) is via a Bayesian approach. Within a Bayesian framework we specify a prior distribution on the population wavelet coefficients. In this section we show that a certain choice of prior model for the population wavelet coefficients implies a Bayesian estimate that produces a thresholding rule, with some features in common with  $T_{\text{hard}}$  and  $T_{\text{soft}}$  given in (4) and (5), respectively.

In this section we work with the sampled white noise model (1) and apply the DWT of Section 2.3. As we have already mentioned, a large variety of different functions allow parsimonious representation in wavelet series where there are only a few non-negligible coefficients present in the expansion. We incorporate this characteristic feature of wavelet bases by placing the following prior on the population discrete wavelet coefficients  $d_{jk}$ :

$$d_{jk} \sim \pi_j N(0, \tau_j^2) + (1 - \pi_j) \delta(0), \quad j = 0, \dots, J - 1; \quad k = 0, \dots, 2^j - 1, \quad (6)$$

where  $0 \leq \pi_j \leq 1$ ,  $\delta(0)$  is a point mass at zero, and  $d_{jk}$  are independent. The hyperparameters  $\pi_j$  and  $\tau_j^2$  have to be specified appropriately (see Section 4 below). Note that we are using the same prior parameters  $\pi_j$  and  $\tau_j^2$  for all coefficients at a given resolution level  $j$ .

According to the prior model (6), every  $d_{jk}$  is either zero with probability  $1 - \pi_j$  or with probability  $\pi_j$  is normally distributed with zero mean and variance  $\tau_j^2$ . The probability  $\pi_j$  gives the proportion of non-zero wavelet coefficients at resolution level  $j$  while the variance  $\tau_j^2$  is a measure of their magnitudes. Clyde *et al.* (1995) use a similar formulation to (6) but with different forms for the hyperparameters  $\pi_j$  and  $\tau_j^2$ . The prior model (6) is an extreme case of a model considered by Chipman *et al.* (1997). Their prior for each  $d_{jk}$  is the mixture of two normal distributions with zero means but different variances for ‘negligible’ and ‘non-negligible’ wavelet coefficients. Subject to the prior (6), the posterior distribution  $d_{jk} | \hat{d}_{jk}$  is also a mixture of a corresponding posterior normal distribution and  $\delta(0)$ . Hence, the posterior cumulative distribution function  $F(d_{jk} | \hat{d}_{jk})$ , letting  $\Phi$  be the standard normal cumulative distribution function, is:

$$F(d_{jk} | \hat{d}_{jk}) = \frac{1}{1 + \omega_{jk}} \Phi \left( \frac{d_{jk} - \hat{d}_{jk} \tau_j^2 / (\sigma^2 + \tau_j^2)}{\sigma \tau_j / \sqrt{\sigma^2 + \tau_j^2}} \right) + \frac{\omega_{jk}}{1 + \omega_{jk}} \mathbf{I}(d_{jk} \geq 0), \quad (7)$$

where the posterior odds ratio for the component at zero is:

$$\omega_{jk} = \frac{1 - \pi_j}{\pi_j} \frac{\sqrt{\tau_j^2 + \sigma^2}}{\sigma} \exp \left( -\frac{\tau_j^2 \hat{d}_{jk}^2}{2\sigma^2(\tau_j^2 + \sigma^2)} \right).$$

The traditional Bayes rule corresponding to the  $L^2$ -loss (the posterior mean) considered in the literature (Vidakovic, 1994; Clyde *et al.*, 1995; Chipman *et al.*, 1997) is not a thresholding rule but a shrinkage. Instead, we suggest the use of any weighted combination of  $L^1$ -losses on the individual wavelet coefficients. Whichever weighted combination is used, the corresponding Bayes rule will be obtained by taking the posterior median of each coefficient. As explained in Section 2.2,  $L^1$ -losses on the estimated function and its derivatives, corresponding to  $B_{1,1}^s$  norms for the function space loss, will be, for all applicable values of  $s$ , equivalent to suitable weighted combinations of  $L^1$ -losses on the wavelet coefficients  $w_{jk}$ . As we shall show below, such  $L^1$  rules (posterior medians) are of thresholding type. Another possible way to obtain a thresholding rule within a Bayesian framework is via hypothesis testing ideas (Vidakovic, 1994).

The posterior cumulative distribution function of  $d_{jk}|\hat{d}_{jk}$  corresponding to (7) has a jump at zero. This fact becomes crucial since solving the equation  $F(d_{jk}|\hat{d}_{jk}) = 0.5$  one gets the posterior median to be zero if  $\omega_{jk} \geq 1$ , and also if:

$$\omega_{jk} < 1 \quad \text{and} \quad 0.5(1 - \omega_{jk}) \leq \Phi \left( -\frac{\hat{d}_{jk}\tau_j}{\sigma\sqrt{\sigma^2 + \tau_j^2}} \right) \leq 0.5(1 + \omega_{jk});$$

it is non-zero otherwise. After straightforward calculus one then has the following closed form:

$$\text{Med}(d_{jk}|\hat{d}_{jk}) = \text{sign}(\hat{d}_{jk}) \max(0, \zeta_{jk}),$$

where:

$$\zeta_{jk} = \frac{\tau_j^2}{\sigma^2 + \tau_j^2} |\hat{d}_{jk}| - \frac{\tau_j\sigma}{\sqrt{\sigma^2 + \tau_j^2}} \Phi^{-1} \left( \frac{1 + \min(\omega_{jk}, 1)}{2} \right). \quad (8)$$

The quantity  $\zeta_{jk}$  is negative for all  $\hat{d}_{jk}$  in some implicitly defined interval  $[-\lambda_j, \lambda_j]$ , and hence  $d_{jk}$  is zero whenever  $|\hat{d}_{jk}|$  falls below the threshold  $\lambda_j$ . The posterior median is therefore a level-dependent thresholding rule with thresholds  $\lambda_j$ . Note that for large  $\hat{d}_{jk}$  the thresholding rule asymptotes to linear shrinkage by a factor of  $\tau_j^2/(\sigma^2 + \tau_j^2)$ , since the second term in (8) becomes negligible as  $|\hat{d}_{jk}| \rightarrow \infty$ . For a plot of the thresholding function for a particular case, see Figure 1.

To complete the prior specification of  $g$ , we place a vague prior on the population scaling coefficient, which is therefore estimated by the sample scaling coefficient  $\hat{c}_{00}$  obtained from the DWT of the data.

## 4 A particular form for the hyperparameters

The hyperparameters  $\pi_j$  and  $\tau_j^2$  of the prior model (6) have to be defined. Different values of hyperparameters will lead to different wavelet estimators, so their proper choice is important. Assume the hyperparameters of the prior model (6) to be of the form:

$$\tau_j^2 = C_1 2^{-\alpha j} \quad \text{and} \quad \pi_j = \min(1, C_2 2^{-\beta j}), \quad j = 0, \dots, J-1, \quad (9)$$

where  $C_1$ ,  $C_2$ ,  $\alpha$ , and  $\beta$  are non-negative constants.

We remark that the universal threshold  $\lambda_{D,J} = \sigma \sqrt{2 \log(n)}$  of Donoho & Johnstone (1994) can be obtained as a particular limiting case of our Bayes rule setting  $\alpha = \beta = 0$  and letting  $C_1 \rightarrow \infty$ ,  $C_2 \rightarrow 0$  as  $n$  increases in such a way that  $\sqrt{C_1}/(C_2 \sigma n) \rightarrow 1$ .

In what follows we discuss the choice of the form (9) and demonstrate a relation between Besov space parameters and hyperparameters of the prior model.

### 4.1 A prior model

Corresponding to the prior model (6) with hyperparameters specified by (9), we consider the following distribution on the infinite sequence of wavelet coefficients  $w_{jk}$ , as defined in Section 2.1:

$$w_{jk} \sim \pi_j N(0, \tau_j^2) + (1 - \pi_j) \delta(0), \quad j \geq 0, \quad k = 0, \dots, 2^j - 1, \quad (10)$$

where  $0 \leq \pi_j \leq 1$ ,  $\delta(0)$  is a point mass at zero, and  $w_{jk}$  are independent. To complete the model we place a vague prior on the scaling coefficient  $u_{00}$ . The hyperparameters of the prior model (10) are assumed to be of the form:

$$\tau_j^2 = C_1^* 2^{-\alpha j} \quad \text{and} \quad \pi_j = \min(1, C_2 2^{-\beta j}), \quad j \geq 0, \quad (11)$$

where  $C_1^* = n^{-1} C_1$ .

It follows from (11) that the prior expected number of non-zero wavelet coefficients on the  $j$ -th level is  $C_2 2^{j(1-\beta)}$ . Appealing to the first Borel-Cantelli lemma, in the case  $\beta > 1$ , the number of non-zero coefficients in the wavelet expansion is finite. In this case the prior model implies that the function is exactly expressed as a finite wavelet expansion. More fruitful and interesting is, however, the case  $0 \leq \beta \leq 1$ . The case  $\beta = 0$

corresponds to the prior belief that all coefficients on all levels have the same probability of being non-zero. This characterises self-similar processes such as white noise or Brownian motion, the overall regularity depending on the value of  $\alpha$ . The case  $\beta = 1$  assumes that the expected number of non-zero wavelet coefficients is the same on each level which is typical, for example, for piecewise polynomial functions, as we shall discuss below. In Section 4.3, we derive an explicit connection between regularity properties of the response function and the hyperparameters (11) in a very general case.

Bayesian simulation has become very popular in recent years. See for example Section 7 of Silverman (1985) for an application of this idea in the curve fitting literature, and many recent papers in Markov chain Monte Carlo. In an earlier paper, Stewart (1979) suggested simulating from *prior* distributions as an aid to the elicitation of prior parameters in the Bayesian paradigm for curve fitting, and our approach is a natural context for the application of this idea. Some realizations from priors for particular values of the hyperparameters  $\alpha$  and  $\beta$  will be given in Figure 2 below.

## 4.2 Some connections with a piecewise polynomial model

To give further intuitive understanding of the model implied by (11) consider a piecewise polynomial function  $g$  (not generated by a wavelet prior). Suppose that there are  $N$  jumps in the  $m$ -th derivative of  $g$ , uniformly located with independent and identically distributed sizes  $h$ , where  $N$  is a random variable with finite mean.

Let the mother wavelet  $\psi$  with a compact support  $[a, b]$  be of regularity  $r > m$  and derive wavelet coefficients of such a piecewise polynomial. Consider a wavelet coefficient  $w_{jk}$ . Then  $w_{jk} = 0$  if there is no jump within  $\text{supp } \psi_{jk} = [2^{-j}(k+a), 2^{-j}(k+b)]$ . For sufficiently large  $j$ , the probability that more than one jump occurs within  $\text{supp } \psi_{jk}$  is negligible, so, after simple calculus, the variance of  $w_{jk}$  conditional on a jump occurring within  $\text{supp } \psi_{jk}$  is given by:

$$\text{Var}(w_{jk} \mid w_{jk} \neq 0) = \frac{E(h^2)}{b-a} 2^{-(2m+1)j} \int_a^b (\Psi^{[m+1]}(u))^2 du,$$

where  $\Psi^{[m+1]}$  is the  $(m+1)$ -fold integral of the mother wavelet  $\psi$ . The probability of a jump occurring within  $\text{supp } \psi_{jk}$  is:

$$P(w_{jk} \neq 0) = (b-a)E(N)2^{-j}.$$

These conditions correspond to the properties of the model (10) with:

$$\alpha = 2m+1, \quad C_1^* = \frac{E(h^2)}{b-a} \int_a^b \left( \Psi^{[m+1]}(u) du \right)^2, \quad \beta = 1, \quad C_2 = (b-a)E(N),$$

though in the piecewise polynomial case the coefficients are not independent and the distribution of  $w_{jk}$  conditioned on  $w_{jk} \neq 0$  is no longer normal. Nevertheless, the connection between piecewise polynomial functions and the prior (10) helps to clarify the intuitive meaning of the constants in (11) in the general case.

### 4.3 A relation between Besov space parameters and hyperparameters of the prior model

In this section, we show that specifying the hyperparameters of the proposed prior model (10), we obtain functions from various Besov spaces. Because of the improper nature of the prior distribution of  $u_{00}$ , we consider the prior distribution of  $g$  conditioned on any given value for  $u_{00}$ . We explore the connections between the parameters  $\alpha$  and  $\beta$  of the prior model (10) with parameters specified by (11) and the Besov space parameters  $s$  and  $p$ . In the Appendix, we study a three parameter prior family that includes (11) as a special case in order to take into account the Besov space parameter  $q$  as well.

Suppose that  $g$  is generated from the prior model (10) with hyperparameters specified by (11). The following theorem establishes necessary and sufficient conditions for  $g$  to fall (with probability one) in any particular Besov space:

**Theorem 1** *Let  $\psi$  be a mother wavelet of regularity  $r$ , where  $\max(0, 1/p - 1/2) < s < r$ ,  $1 \leq p, q \leq \infty$ , and let the wavelet coefficients  $w_{jk}$  of a function  $g$  obey the prior model (10) with  $\tau_j^2 = C_1^* 2^{-\alpha j}$  and  $\pi_j = \min(1, C_2 2^{-\beta j})$ , where  $C_1^*$ ,  $C_2$ ,  $\alpha \geq 0$  and  $0 \leq \beta \leq 1$ . Then, for any fixed value of  $u_{00}$ ,  $g \in B_{p,q}^s$  almost surely if and only if either:*

$$s + 1/2 - \beta/p - \alpha/2 < 0, \tag{12}$$

or:

$$s + 1/2 - \beta/p - \alpha/2 = 0 \text{ and } 0 \leq \beta < 1, 1 \leq p < \infty, q = \infty. \tag{13}$$

**Remark** As we mentioned in Section 4.1, in the case  $\beta > 1$ , the number of non-zero coefficients in the wavelet expansion is finite almost surely. Therefore, with probability one,  $g$  will belong to the same Besov spaces as the mother wavelet  $\psi$ , i.e. those for which  $\max(0, 1/p - 1/2) < s < r$ ,  $1 \leq p, q \leq \infty$ .

Theorem 1 is a particular case of the more general Theorem 2 formulated and proved in the Appendix. On the one hand, Theorem 1 shows how prior knowledge about a specific Besov space can be incorporated into the prior model (10) for the wavelet coefficients by choosing the corresponding hyperparameters of their prior distribution. On the other hand, the above theorem gives insight into the meaning of the Besov space parameters.

Certain priors are important in the derivation of minimax properties of wavelet threshold estimators (see, for example, Donoho & Johnstone, 1994; Johnstone & Silverman, 1997). These priors place a symmetric three point distribution independently on each wavelet coefficient, and give realizations that can be considered as being ‘least favourable’ within particular smoothness classes. Johnstone (1994) investigates various properties of these priors. He presents realizations from these and uses them to illustrate the variety of forms of prior information captured by a family of Besov spaces. For this purpose, the priors we construct may be preferable in producing functions that are typical of particular Besov spaces rather than least favourable with respect to some criterion.

Figure 2 shows realizations with various values of the hyperparameters. It can be seen that, for  $\beta = 1$ , the functions show irregularities in some places with relative smooth behaviour in between. The same is true to a much lesser extent for  $\beta = 0.5$ . For  $\alpha = 1$  there are gross irregularities in the value of the function itself, and for  $\alpha = 2$  these are less marked. The irregularities for  $\alpha = 4$  are not easily visible in the figure, but the first derivative of the realization is similar in character to the corresponding figures for  $\alpha = 2$ . The model  $\alpha = 4, \beta = 0$  is equivalent to an integrated Wiener process which is the prior used to motivate spline smoothing by Kimeldorf & Wahba (1970). The priors we construct can be used to aid our understanding of Besov spaces and norms (see Figure 3). Consider, for example, the case  $s = 1, p = 1$ . Figure 3 demonstrates that realizations (f), (g), (h) and (i) in Figure 2 lie in Besov space  $B_{1,q}^1$ , while realizations (a), (b) and (d) lie outside. Realizations (c) and (e) are on the ‘boundaries’ of ‘just’ in or out; (e) is actually out for  $1 \leq q < \infty$  and is in for  $q = \infty$ . Therefore, from the point of view of Besov norms with  $p = 1$ , realizations (e) and (c) are, roughly speaking, equally irregular. Realization (c) has occasional gross irregularities, and so is more

inhomogeneous, while in (e) the irregularity is more evenly spread. More detailed consideration of  $B_{1,q}^s$  norms in Figure 3 show that the ranking of the realizations in terms of their critical value  $s$  is, from roughest to smoothest, (a), (b) and (d) jointly, (c) and (e) jointly, (f) and (g) jointly, (h), (i).

Now consider the other extreme,  $p = \infty$ . In this case, the realizations within each row of Figure 2 have the same critical value of  $s$ , 0 for the top row, 0.5 for the middle row, and 1.5 for the bottom row. The figure gives a clear demonstration of the way that the  $B_{\infty,q}^s$  norms stress the maximum irregularity. They give quite a different ordering from the case  $p = 1$ . Yet other rankings are yielded by intermediate values of  $p$ .

It can be seen from Figure 3 that for  $\alpha = 0.5$  there is some restriction on the range of  $\beta$  that will give lines that intersect the unshaded part of the figure. However, the case  $\alpha = 0.5$ ,  $\beta = 1$  will give a line with slope  $-1$  and will intersect the horizontal axis at  $p = 4$ . In our subsequent investigation, we shall find that this is a good model in practice and a realization from this model is given in Figure 4. It can be seen that the function is mostly regular but allows for occasional gross irregularities.

In general, the priors discussed in this section can be used to generate a range of functions just on the boundary of membership of any particular Besov space.

#### 4.4 Estimation of the hyperparameters

To apply the proposed Bayesian thresholding procedure in practice, it is necessary first to specify the hyperparameters  $\alpha, \beta, C_1$  and  $C_2$  in (9). Our approach is as follows. The choice of  $\alpha$  and  $\beta$  could be made from prior knowledge about regularity properties of the unknown function making use of the results of Theorem 1. However all but the most fundamentalist Bayesians may find this a daunting prospect, and we investigate the choice further in Section 5.1 below. To estimate  $C_1$  and  $C_2$  we suggest the following procedure.

As we have already mentioned, the set of sample wavelet coefficients  $\hat{d}_{jk}$  contains both ‘non-negligible’ coefficients of the unknown function  $g$  and ‘negligible’ coefficients representing a random noise. Apply the *VisuShrink* threshold  $\lambda_{DJ} = \sigma\sqrt{2\log(n)}$ . When the noise level  $\sigma$  is unknown, it is robustly estimated by the median absolute deviation of the wavelet coefficients at the finest level, divided by 0.6745 (Donoho & Johnstone, 1994). By the construction of this thresholding rule, the probability that even one negligible coefficient will pass the threshold tends to zero (Donoho & Johnstone,

1994), so essentially only non-negligible  $\hat{d}_{jk}$  will survive after thresholding. Suppose that, on level  $j$ , the number of coefficients that pass  $\lambda_{DJ}$  is  $M_j$ , and that the values of these coefficients are  $x_{j1}, \dots, x_{jM_j}$ . Conditioning on the value  $M_j$ , the  $x_{jm}$ ,  $m = 1, \dots, M_j$ , are independent realizations from the tails of the  $N(0, \sigma^2 + \tau_j^2)$  distribution beyond the points  $\pm\sigma\sqrt{2\log(n)}$ . The log likelihood function is therefore, up to a constant:

$$\begin{aligned}
l(\tau_0^2, \dots, \tau_{J-1}^2) = & - \sum_{j=0}^{J-1} M_j \left\{ \frac{1}{2} \log(\sigma^2 + \tau_j^2) - \log \left( \Phi \left[ -\frac{\lambda_{DJ}}{\sqrt{\sigma^2 + \tau_j^2}} \right] \right) \right\} \\
& - \sum_{j=0}^{J-1} \left\{ \frac{1}{2(\sigma^2 + \tau_j^2)} \sum_{m=1}^{M_j} x_{jm}^2 \right\}. \tag{14}
\end{aligned}$$

Substituting  $\tau_j^2 = C_1 2^{-\alpha j}$  and  $\lambda_{DJ} = \sigma\sqrt{2\log(n)}$  we can obtain an estimate of  $C_1$  by a numerical maximization of (14), carrying out a grid search on  $C_1$ .

The parameter  $C_2$  can be chosen by a cognate procedure. We use the numbers  $M_0, \dots, M_{J-1}$  of coefficients passing the threshold to estimate the  $\pi_j$ . Let  $q_j = 2\Phi(-\lambda_{DJ}/\sqrt{\sigma^2 + \tau_j^2})$ , the probability conditional on  $d_{jk} \neq 0$  that  $d_{jk}$  passes the threshold  $\lambda_{DJ}$ . Neglecting the possibility that any  $\hat{d}_{jk}$  corresponding to a zero  $d_{jk}$  passes the threshold  $\lambda_{DJ}$ , the ‘imputed number’ of non-zero  $d_{jk}$  at level  $j$  is  $M_j/q_j$ , and the expected value of  $M_j/q_j$  is  $C_2 2^{(1-\beta)j}$ . Given the value of  $\beta$ , a simple method of moments estimate of  $C_2$  based on the total imputed number of non-zero  $d_{jk}$  is:

$$\begin{aligned}
\hat{C}_2 &= \frac{2^{(1-\beta)} - 1}{2^{(1-\beta)J} - 1} \sum_{j=0}^{J-1} \frac{M_j}{q_j}, \quad \text{if } 0 \leq \beta < 1, \\
\hat{C}_2 &= \frac{1}{J} \sum_{j=0}^{J-1} \frac{M_j}{q_j}, \quad \text{if } \beta = 1.
\end{aligned}$$

## 5 Applications and Comparisons

In this section we first consider simulated examples to illustrate the proposed Bayesian thresholding procedure, which we refer to as *BayesThresh*, and make comparisons with other existing thresholding methods. An application to a data set collected in an anaesthesiological study is then presented.

## 5.1 Simulation study

We consider the four examples of Donoho & Johnstone (1994, 1995) that have become standard tests for wavelet estimators: Blocks, Bumps, Heavisine and Doppler. These functions caricature spatially variable signals arising in imaging, spectroscopy, seismography and other scientific fields.

For each test function, noisy data were generated for 100 replications by corrupting a true function with independent random noise  $\varepsilon_i \sim N(0, \sigma^2)$  at 1024 data points uniformly spaced on  $[0, 1]$ . The values of  $\sigma$  were taken to correspond to values 10, 7, 5 and 3 for the root signal-to-noise ratio (RSNR)  $\{\int_0^1 (g - \bar{g})^2\}^{1/2}/\sigma$ , where  $\bar{g} = \int_0^1 g$ .

We compare *BayesThresh* with several wavelet-based estimators for reconstructing the original functions: *VisuShrink* (Donoho & Johnstone, 1994), *GlobalSure* (a modified version of the *SureShrink* of Donoho & Johnstone, 1995, considered in Nason, 1996), *Cross-validation* (Nason, 1995; 1996) and *False Discovery Rate* (Abramovich & Benjamini, 1995; 1996). Daubechies's least asymmetric wavelet of order 8 (defined through a set of sixteen non-zero coefficients whose numerical values may be found in Daubechies, 1992, Table 6.3, p. 198) was used for all methods. In all the methods, except *BayesThresh*, the soft thresholding (5) was applied, and the wavelet coefficients on the five coarsest levels were not thresholded.

The goodness-of-fit of each estimator was measured by its average mean square error (AMSE) defined as the average over simulated replications  $\hat{g}$  of  $n^{-1} \sum_{i=1}^n (\hat{g}_i - g_i)^2$ . The AMSE's and standard errors over 100 simulations for the different methods appear in Table 1. The simulations show that, in almost all cases, *BayesThresh* ( $\alpha = 0.5$ ,  $\beta = 1$ ) has smaller AMSE with *Cross-validation* usually second, *GlobalSure* third, *False Discovery Rate* fourth and *VisuShrink* fifth in the rankings. In fact, similar results held when we used the hard thresholding (4) instead of the soft, with the exception of the *False Discovery Rate* procedure whose performance is improved substantially, approximately to the level obtained by *GlobalSure*. This is, perhaps, not surprising, since the original idea of the *False Discovery Rate* has a natural interpretation as a hard thresholding procedure (see Abramovich & Benjamini, 1995, 1996).

Within the *BayesThresh* approach, the effect of varying  $\alpha$  and  $\beta$  was investigated. For all four functions, reducing  $\beta$  gave worse results, especially for the Blocks and Doppler functions. This is as might be expected given the irregularity of these functions. The value  $\alpha = 1$  gave very slightly better results for large RSNR, but noticeably worse for smaller RSNR. Larger values

of  $\alpha$  gave poor results, except for the Heavisine example, which is somewhat more regular than the others; even in this case there was no improvement over the case  $\alpha = 0.5$ .

Figures 5, 6, 7 and 8 show the test functions and the reconstructions obtained, with all methods applied to noisy versions of the functions with  $\text{RSNR} = 3$ . It can be seen from these plots that the *BayesThresh* ( $\alpha = 0.5$ ,  $\beta = 1$ ) method generally gives a better reconstruction of the fine scale structure, relative to amount of noise in the smooth parts of the functions. In particular, the *BayesThresh* method gives a better reconstruction of the corners in *Blocks*, the high peaks in *Bumps*, the jumps in *HeaviSine*, and the high frequency parts of *Doppler*.

## 5.2 Inductance plethysmography data

Here, we apply the above thresholding methods to a data set arising from anaesthesiology collected by inductance plethysmography. The recordings were made by the Department of Anaesthesia at the Bristol Royal Infirmary and measure the flow of air during breathing. The same data set has been analysed in Nason (1996). We refer to this latter paper for more details. These signals are intrinsically continuous, and therefore large values of  $\alpha$  in the *BayesThresh* method may be appropriate. Therefore, we consider  $\alpha = 0.5$ , 1 and 2, and estimate  $C_1$  and  $C_2$  as suggested in Section 4.4. The noise level  $\sigma$  is robustly estimated by the median absolute deviation of the wavelet coefficients at the finest level, divided by 0.6745 (Donoho & Johnstone, 1994).

Figures 9 and 10 show a section of plethysmograph recording lasting approximately 80 seconds (4096 data points) together with various reconstructions. The two main sets of regular oscillations correspond to normal breathing. The disturbed behaviour in the centre of the plot where the normal breathing pattern disappears corresponds to the patient vomiting.

Note that *VisuShrink* and *False Discovery Rate* reconstructions remove the noise but tend to attenuate the peaks while *Cross-Validation* and *GlobalSure* procedures retain the sharpness of peaks, but the smooth parts of the curves are still noisy. On the other hand, *BayesThresh* method possesses ‘noise-free’ quality without the attenuation. See, for example, the height of the first peak tabulated in Table 2.

## 6 Concluding remarks

In this paper we have discussed a Bayesian formalism which has given rise to a type of wavelet threshold estimation in non-parametric regression. A prior distribution was imposed on the wavelet coefficients of the unknown response function, designed to capture the sparseness of wavelet expansion common to most applications. For the prior specified, the posterior median yielded a thresholding procedure. Several simulated examples were used to illustrate our method, and comparisons were made with other thresholding methods. We also presented an application to a data set collected in an anaesthesiological study.

Our prior model for the underlying function can be adjusted to give functions falling in any specific Besov space. We have established a relation between the hyperparameters of the prior model and the parameters of those Besov spaces within which realizations from the prior will fall. This makes it possible in principle to incorporate prior knowledge about the function's regularity properties into the prior model for its wavelet coefficients, though in the present state of understanding of Bayesian smoothing methods the 'standard' choice  $\alpha = 0.5$  and  $\beta = 1$  seems to be the best practical approach. As in any applications of Bayesian methods in curve and image processing, there are aspects of our 'genuine' prior knowledge that are not captured in the model. For example, an interesting avenue for future research would be the investigation of the effects of allowing dependence between the wavelet coefficients of the true function. Although the wavelet transform can act as a 'decorrelator' that tends to make each wavelet coefficient statistically independent of all others, it will not completely 'decorrelate' most signals. A recent contribution in this direction is by Crouse *et al.* (1998) who have developed a framework to capture statistical dependencies between wavelet coefficients based on wavelet-domain hidden Markov models.

Another interesting aspect is the estimation of the noise level  $\sigma$ . In our formulation, it is assumed that either  $\sigma$  is known or a reasonably good estimator is available. Where this is not the case, a prior may be put on  $\sigma$ . Clyde *et al.* (1995) deal with this situation and use a Bayesian hierarchical model to define a multiple shrinkage estimator for the wavelet coefficients. They also discuss fast computational implementations through importance sampling and Markov chain Monte Carlo methods. The combination of these ideas with our approaches is an interesting topic for further research.

## Acknowledgments

The financial support of the Engineering and Physical Sciences Research Council, the Israel Academy of Science, and the Royal Society are gratefully acknowledged. This work was partially carried out while BWS was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, California, gratefully supported by NSF grant number SBR-9601236. The inductance plethysmography data was kindly supplied by Andrew Black of the Sir Humphry Davy Department of Anaesthesia, Bristol Royal Infirmary. The *BayesThresh* procedure will be incorporated within the *WaveThresh3* package, which will also contain the inductance plethysmography data. Helpful comments of the reviewers are gratefully acknowledged.

## Appendix: Theoretical details

Our study so far has concentrated on the Besov space parameters  $s$  and  $p$ . In order to take into account the Besov space parameter  $q$  as well, we now introduce a three parameter prior family that includes (11) as a special case. Specifically, we allow a more delicate dependence of the variance parameter  $\tau_j^2$  on the level  $j$  by introducing a third parameter  $\gamma$ , with  $-\infty < \gamma < \infty$ :

$$\tau_j^2 = C_1^* 2^{-\alpha j} j^\gamma.$$

The following theorem extends the results of Theorem 1 to this prior, and contains Theorem 1 as the special case  $\gamma = 0$ :

**Theorem 2** *Let  $\psi$  be a mother wavelet of regularity  $r$ , where  $\max(0, 1/p - 1/2) < s < r$ ,  $1 \leq p, q \leq \infty$ , and let the wavelet coefficients  $w_{jk}$  of a function  $g$  obey the prior model (10) with  $\tau_j^2 = C_1^* 2^{-\alpha j} j^\gamma$  and  $\pi_j = \min(1, C_2 2^{-\beta j})$ , where  $C_1^*$ ,  $C_2$ ,  $\alpha \geq 0$ ,  $0 \leq \beta \leq 1$  and  $\gamma \in \mathbb{R}$ . Then, for any fixed value of  $u_{00}$ ,  $g \in B_{p,q}^s$  almost surely if and only if either:*

$$s + 1/2 - \beta/p - \alpha/2 < 0,$$

or:

$$s + 1/2 - \beta/p - \alpha/2 = 0 \text{ and } \gamma \text{ satisfies the following conditions:}$$

	$p, q < \infty$	$p = \infty, q < \infty$	$p < \infty, q = \infty$	$p, q = \infty$
$0 \leq \beta < 1$	$\gamma < -2/q$	$\gamma < -1 - 2/q$	$\gamma \leq 0$	$\gamma \leq -1$
$\beta = 1$	$\gamma < -2/q$		$\gamma < 0$	

**Proof** Define  $z_j$  to be the vector with elements  $z_{jk}$ , where  $z_{jk} = \tau_j^{-1}w_{jk}$  for  $k = 0, \dots, 2^j - 1$ . We consider the following three cases:

**Case I:**  $0 \leq \beta < 1$ ,  $1 \leq p \leq \infty$ ,  $1 \leq q \leq \infty$ .

For any  $1 \leq p < \infty$ , let  $\nu_p$  be the  $p$ -th absolute moment of the standard normal distribution. We then have  $E\|z_j\|_p^p = C_2 2^{(1-\beta)j} \nu_p$  and  $\text{Var}\|z_j\|_p^p \leq C_2 2^{(1-\beta)j} \nu_{2p}$ . We also define:

$$\delta = \begin{cases} s + 1/2 - \beta/p - \alpha/2 & \text{if } 1 \leq p < \infty \\ s + 1/2 - \alpha/2 & \text{if } p = \infty. \end{cases}$$

Given  $\varepsilon > 0$ , Chebyshev's inequality implies that:

$$\sum_{j=0}^{\infty} P \left\{ \left| 2^{-(1-\beta)j} \|z_j\|_p^p - C_2 \nu_p \right| > \varepsilon \right\} \leq O(1) \varepsilon^{-2} \sum_{j=0}^{\infty} 2^{-(1-\beta)j} < \infty.$$

Appealing to the first Borel-Cantelli lemma, it follows that:

$$2^{-(1-\beta)j} \|z_j\|_p^p \rightarrow C_2 \nu_p \text{ almost surely as } j \rightarrow \infty. \quad (15)$$

By standard extreme value manipulations, using the fact that  $P(|z_{jk}| > u) = 2\pi_j(1 - \Phi(u))$ , where  $\Phi$  is the standard normal distribution function, we also have:

$$j^{-1/2} \|z_j\|_{\infty} \rightarrow \sqrt{2(1-\beta) \log(2)} \text{ almost surely as } j \rightarrow \infty. \quad (16)$$

Hence, in view of (15), (16) and the equivalence of the norms given by (2) and (3), the required conditions that  $g \in B_{p,q}^s$  almost surely will be the finiteness of:

$$\sum_{j=0}^{\infty} 2^{js'q} 2^{(1-\beta)jq/p} \tau_j^q = \sum_{j=0}^{\infty} 2^{j\delta q} j^{\gamma q/2}, \text{ if } 1 \leq p < \infty, 1 \leq q < \infty, \quad (17)$$

$$\sum_{j=0}^{\infty} 2^{js'q} j^{q/2} \tau_j^q = \sum_{j=0}^{\infty} 2^{j\delta q} j^{(\gamma+1)q/2}, \text{ if } p = \infty, 1 \leq q < \infty, \quad (18)$$

$$\sup_{j \geq 0} \{2^{js'} 2^{(1-\beta)j/p} \tau_j\} = \sup_{j \geq 0} \{2^{j\delta} j^{\gamma/2}\}, \text{ if } 1 \leq p < \infty, q = \infty, \quad (19)$$

$$\sup_{j \geq 0} \{2^{js'} j^{1/2} \tau_j\} = \sup_{j \geq 0} \{2^{j\delta} j^{(\gamma+1)/2}\}, \text{ if } p = \infty, q = \infty. \quad (20)$$

In each case, the above expressions will be finite if  $\delta < 0$ , and infinite for  $\delta > 0$ . For  $\delta = 0$ , the expressions will be finite if and only if  $\gamma < -2/q$  in case (17),  $\gamma < -1 - 2/q$  in case (18),  $\gamma \leq 0$  in case (19), or  $\gamma \leq -1$  in case (20).

**Case II:**  $\beta = 1$ ,  $1 \leq p \leq \infty$ ,  $1 \leq q < \infty$ .

The non-zero elements of  $z_j$  consist of  $M_j$  independent standard normal random variables, where  $M_j \sim \text{Binomial}(2^j, C_2 2^{-j})$ . By a standard coupling argument, there is a sequence  $N_j$  of Poisson ( $C_2$ ) random variables such that  $M_j = N_j$  almost surely for all sufficiently large  $j$ . Let  $\xi_j$  be a vector of  $N_j$  independent standard

normal random variables, independently for each  $j$ . Then  $\|\xi_j\|_1$  is a sum of a Poisson number of independent identical  $|N(0, 1)|$  random variables; by standard probability arguments all the moments of  $\|\xi_j\|_1$  are therefore finite since  $Ee^{\|\xi_j\|_1} < \infty$ . Furthermore, for any  $1 \leq p \leq \infty$ ,  $\|\xi_j\|_p \leq \|\xi_j\|_1$  (e.g. Beckenbach & Bellman, 1961, p. 18) and hence  $0 < E\|\xi_j\|_p^q \leq E\|\xi_j\|_1^q < \infty$ .

By the equivalence of the norms given by (2), we now have the result that  $g \in B_{p,q}^s$  almost surely if and only if:

$$\sum_{j=0}^{\infty} 2^{js'q} \tau_j^q \|\xi_j\|_p^q = \sum_{j=0}^{\infty} 2^{j\delta q} j^{\gamma q/2} \|\xi_j\|_p^q < \infty \quad \text{almost surely.} \quad (21)$$

It can be shown from the monotone convergence theorem and the three series theorem (see Karr, 1993, Theorems 4.10 and 7.5 respectively) that, if  $Z_n$  are independent and identically distributed nonnegative random variables with strictly positive finite mean, and  $a_n$  are nonnegative constants, then  $\sum a_n Z_n$  is convergent almost surely if and only if  $\sum a_n$  is convergent. It follows that (21) is equivalent to:

$$\sum_{j=0}^{\infty} 2^{j\delta q} j^{\gamma q/2} < \infty, \quad (22)$$

since the  $\xi_j$  are independent and identically distributed and  $E\|\xi_j\|_p^q$  is finite. The condition (22) is satisfied if and only if either  $\delta < 0$ , or  $\delta = 0$  and  $\gamma < -2/q$ .

**Case III:**  $\beta = 1$ ,  $1 \leq p \leq \infty$ ,  $q = \infty$

By the equivalence of the norms given by (3) and the coupling argument presented previously,  $g \in B_{p,q}^s$  almost surely if and only if:

$$\sup_{j \geq 0} \{2^{js'} \tau_j \|\xi_j\|_p\} = \sup_{j \geq 0} \{2^{j\delta} j^{\gamma/2} \|\xi_j\|_p\} < \infty \quad \text{almost surely.} \quad (23)$$

Appealing to the Borel-Cantelli lemmas, it follows that the condition (23) holds if and only if there exists a constant  $c$  such that:

$$\sum_{j=0}^{\infty} P(2^{j\delta} j^{\gamma/2} \|\xi_j\|_p \geq c) < \infty. \quad (24)$$

After some simple arguments, using the facts that for any  $1 \leq p \leq \infty$ ,  $\|\xi_j\|_p \leq \|\xi_j\|_1$  and  $Ee^{\|\xi_j\|_1} < \infty$ , the condition (24) is satisfied if and only if either  $\delta < 0$ , or  $\delta = 0$  and  $\gamma < 0$ . This completes the proof for this case, and hence we have the theorem.

## References

Abramovich, F. & Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In *Wavelets and*

- Statistics*, Lecture Notes in Statistics **103**, Antoniadis, A. and Oppenheim, G. (Eds.), pp. 5-14, New York: Springer-Verlag.
- Abramovich, F. & Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics & Data Analysis* **22**, 351-361.
- Beckenbach, E.F. & Bellman, R. (1961). *Inequalities*. Berlin: Springer-Verlag.
- Chipman, H.A., Kolaczyk, E.D. & McCulloch, R.E. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association* **92** (to appear).
- Clyde, M., Parmigiani, G. & Vidakovic, B. (1995). Multiple shrinkage and subset selection in wavelets. *Discussion Paper 95-37*, ISDS, Duke University, USA.
- Cohen, A., Daubechies, I., Jawerth, B. & Vial, P. (1993). Multiresolution analysis, wavelets and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris* **316**, 417-421.
- Crouse, M., Nowak, R. & Baraniuk, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing* **46** (To appear).
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM.
- DeVore, R.A., Jawerth, B. & Popov, V. (1992). Compression of wavelet decompositions. *American Journal of Mathematics* **114**, 737-785.
- DeVore, R.A. & Popov, V. (1988). Interpolation of Besov Spaces. *Transactions of the American Mathematical Society* **305**, 397-414.
- Donoho, D.L. & Johnstone, I.M. (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika* **81**, 425-455.
- Donoho, D.L. & Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200-1224.
- Donoho, D.L. & Johnstone, I.M. (1997). Minimax estimation via wavelet shrinkage. *Annals of Statistics* **25**, (In press).

- Donoho, D.L. & Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet Shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **57**, 301-369.
- Jawerth, B & Sweldens, W. (1994). An overview of wavelet based multiresolution analyses. *SIAM Review* **36**, 377-412.
- Johnstone, I.M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics, V*, Gupta, S.S and Berger, J.O. (Eds.), pp. 303-326, New-York: Springer-Verlag.
- Johnstone, I.M. & Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Ser. B*, **59**, 319-351.
- Karr, A.F. (1993). *Probability*. New York: Springer-Verlag.
- Kimeldorf, G. & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics* **41**, 495-502.
- Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **11**, 674-693.
- Meyer, Y. (1992). *Wavelets and Operators*. Cambridge: Cambridge University Press.
- Nason, G.P. (1993). *The WaveThresh Package: Wavelet Transform and Thresholding Software for S*. Available from the StatLib archive.
- Nason, G.P. (1995). Choise of the threshold parameter in wavelet function estimation. In *Wavelets and Statistics*, Lecture Notes in Statistics **103**, Antoniadis, A. and Oppenheim, G. (Eds.), pp. 261-280, New York: Springer-Verlag.
- Nason, G.P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, Ser. B*, **58**, 463-479.
- Nason, G.P. & Silverman, B.W. (1994). The discrete wavelet transform in S. *Journal of Computational & Graphical Statistics* **3**, 163-191.

- Ogden, T. & Parzen, E. (1996a). Data dependent wavelet thresholding in nonparametric regression with change-point applications. *Computational Statistics & Data Analysis* **22**, 53-70.
- Ogden, T. & Parzen, E. (1996b). Change-point approach to data analytic wavelet thresholding. *Statistics & Computing* **6**, 93-99.
- Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **47**, 1-52.
- Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**, 1135-1151.
- Stewart, L. (1979). Multiparameter univariate Bayesian analysis. *Journal of the American Statistical Association* **74**, 684-693.
- Strang, G. (1993). Wavelet transforms versus Fourier transforms. *Bulletin (New Series) of the American Mathematical Society* **28**, 288-305.
- Triebel, H. (1990). *Theory of Function Spaces II*. Basel: Birkhäuser.
- Vidakovic, B. (1994). Non-linear wavelet shrinkage with Bayes rules and Bayes factors. *Discussion Paper 94-24*, ISDS, Duke University, USA.
- Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Annals of Statistics* **24**, 466-484.
- Weyrich, N. & Warhola, G.T. (1995). De-noising using wavelets and cross-validation. In *Approximation Theory, Wavelets and Applications*, NATO ASI Series C, **454**, Singh, S.P. (Ed.), pp. 523-532, Dordrecht: Kluwer.

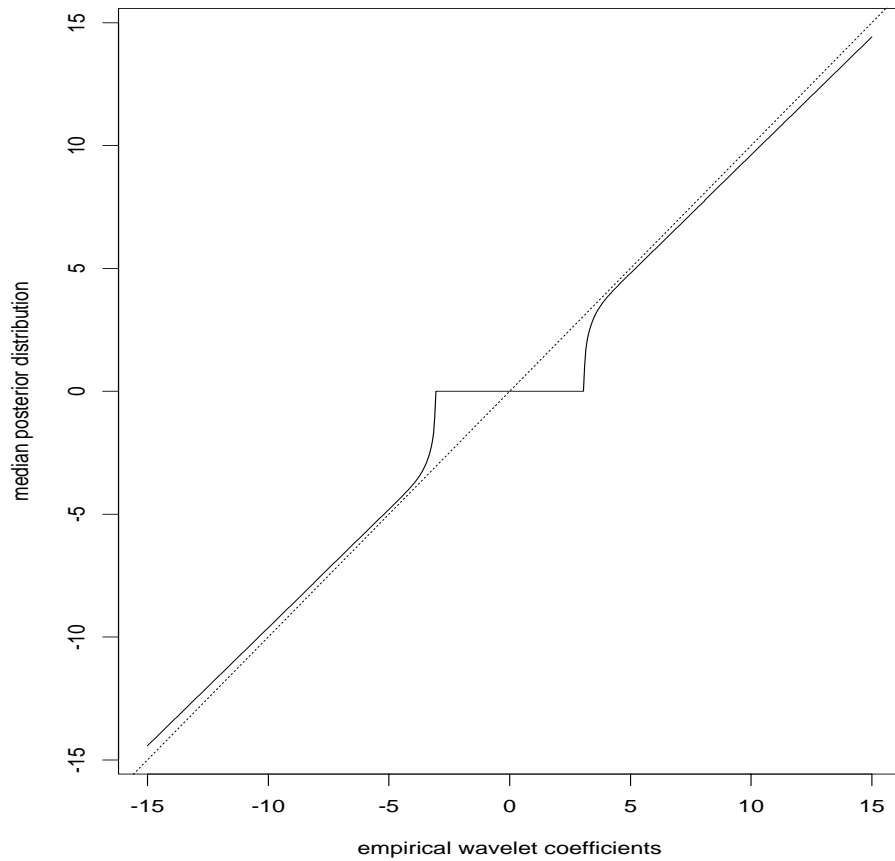


Figure 1: The median of posterior distribution (solid line) as function of the empirical wavelet coefficients. Hyperparameters were chosen as  $\tau^2 = 25$ ,  $\pi = 0.05$ , while  $\sigma$  was fixed at 1. (The dotted line shows the diagonal.)

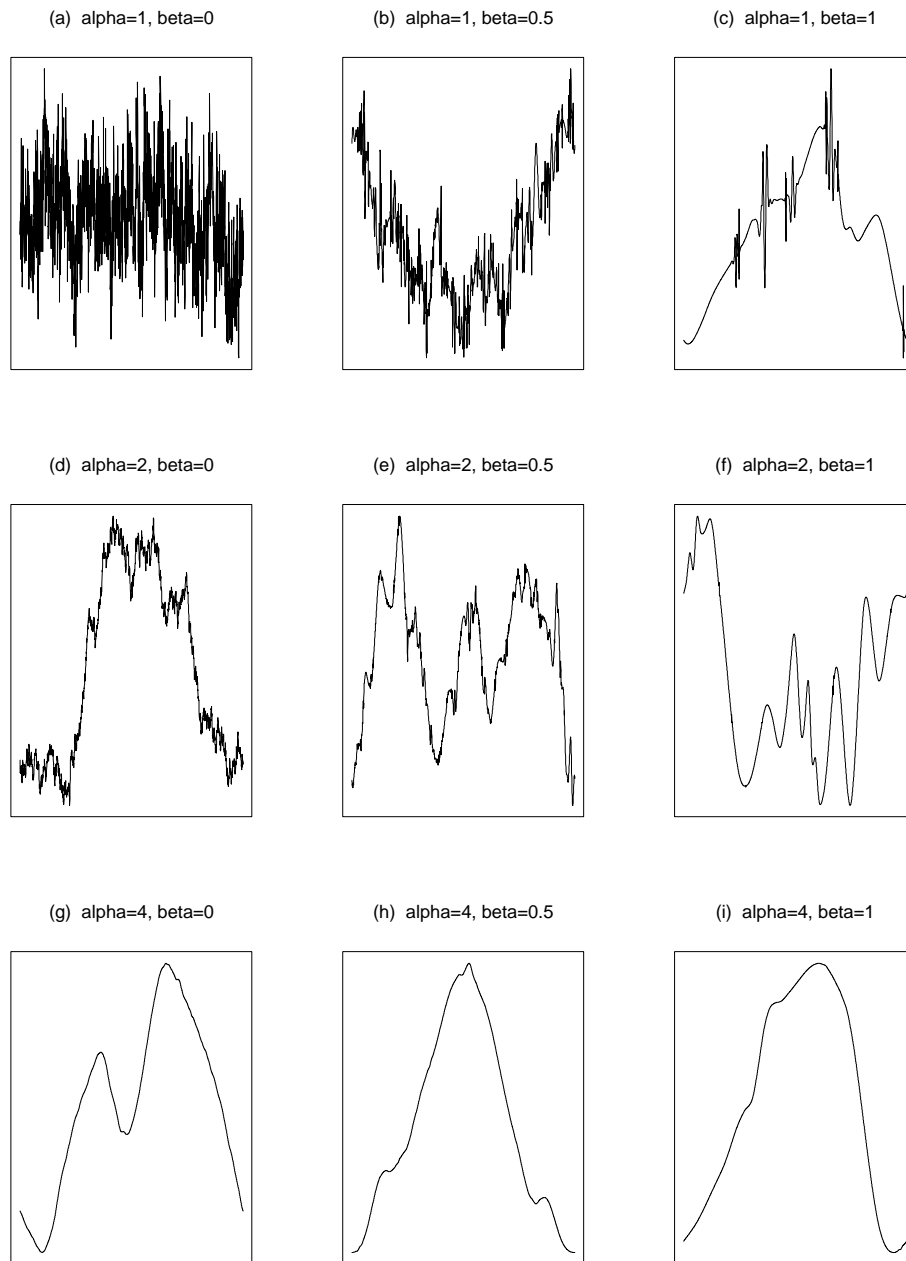


Figure 2: Realizations with various values of the hyperparameters  $\alpha$  and  $\beta$ , with  $n = 2048$ ,  $C_1 = 1$  and  $C_2 = 2$ . 27

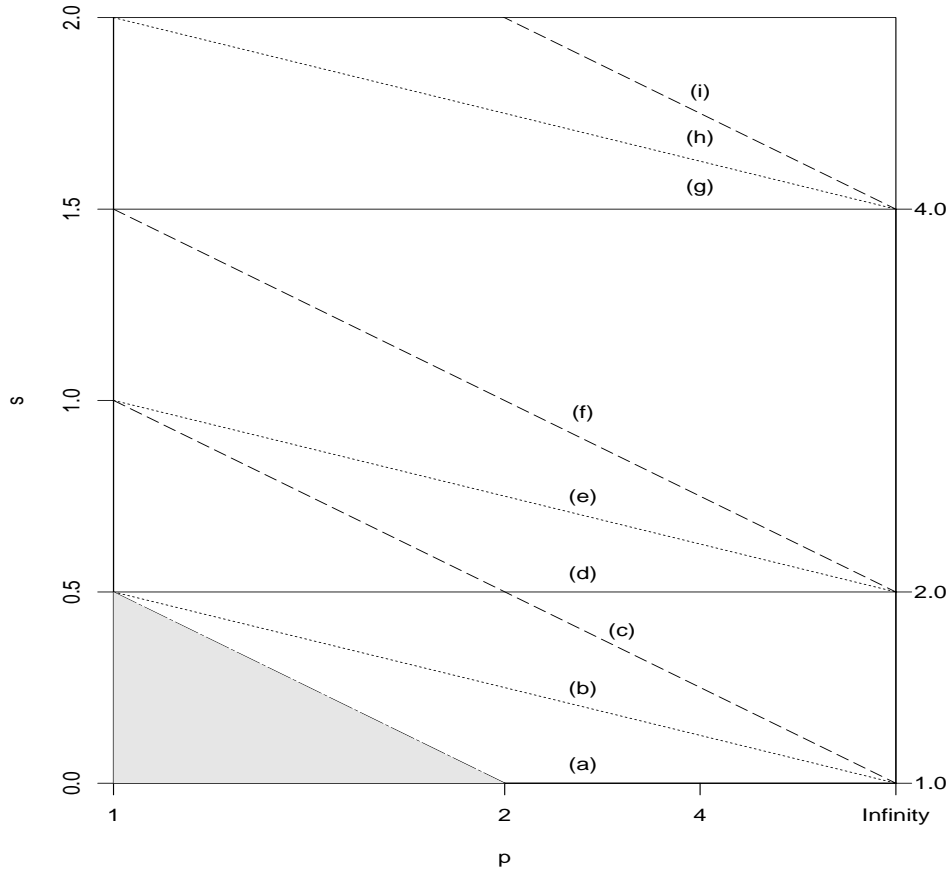


Figure 3: The critical values of the Besov space parameters  $s$  and  $p$  for certain values of  $\alpha$  and  $\beta$  for any values of  $1 \leq q \leq \infty$ . The values of  $\alpha = 1, 2$  and  $4$  are indicated on the right hand axis. The values of  $\beta$  are indicated by the line style (solid:  $\beta = 0$ ; dotted:  $\beta = 0.5$ ; dashed:  $\beta = 1$ ). For each value  $(\alpha, \beta)$ , realizations lie in all Besov spaces with parameter values below the line plotted. They also lie in spaces on the critical line if  $0 \leq \beta < 1$ ,  $1 \leq p < \infty$  and  $q = \infty$ . The shaded region represents the range of  $(p, s)$  excluded by the conditions of Theorem 1. The  $p$ -axis is transformed to be linear in  $(1 - 1/p)$ . Labels (a)-(i) correspond to the realizations given in Figure 2.

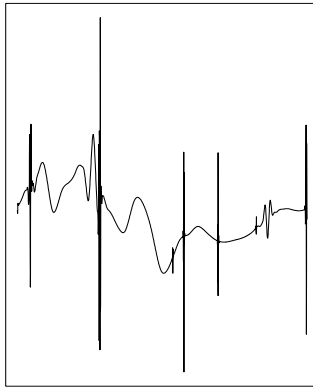


Figure 4: A realization with hyperparameters  $\alpha = 0.5$ ,  $\beta = 1$ , with  $n = 2048$ ,  $C_1 = 1$  and  $C_2 = 2$ .

Method	RSNR	Blocks	Bumps	HeaviSine	Doppler
<i>BayesThresh</i> $\alpha = 0.5, \beta = 1$	10	0.22 (0.002)	0.25 (0.002)	0.06 (0.001)	0.09 (0.001)
	7	0.38 (0.003)	0.45 (0.004)	0.10 (0.001)	0.16 (0.003)
	5	0.67 (0.008)	0.74 (0.006)	0.15 (0.002)	0.30 (0.004)
	3	1.60 (0.014)	1.73 (0.019)	0.30 (0.002)	0.69 (0.009)
<i>Cross-Validation</i>	10	0.23 (0.002)	0.25 (0.002)	0.06 (0.001)	0.11 (0.001)
	7	0.41 (0.003)	0.46 (0.003)	0.10 (0.014)	0.21 (0.002)
	5	0.72 (0.006)	0.84 (0.005)	0.16 (0.003)	0.39 (0.004)
	3	1.68 (0.013)	2.08 (0.016)	0.32 (0.005)	0.91 (0.005)
<i>GlobalSure</i>	10	0.25 (0.002)	0.29 (0.003)	0.08 (0.007)	0.11 (0.001)
	7	0.42 (0.003)	0.48 (0.004)	0.12 (0.001)	0.21 (0.002)
	5	0.82 (0.009)	0.92 (0.008)	0.18 (0.002)	0.59 (0.009)
	3	3.32 (0.047)	3.31 (0.031)	0.32 (0.004)	1.73 (0.022)
<i>False Discovery Rate</i>	10	0.55 (0.005)	0.69 (0.006)	0.08 (0.008)	0.22 (0.003)
	7	0.96 (0.008)	1.23 (0.011)	0.12 (0.001)	0.39 (0.005)
	5	1.58 (0.015)	2.08 (0.022)	0.17 (0.003)	0.65 (0.006)
	3	3.15 (0.025)	4.68 (0.043)	0.31 (0.004)	1.35 (0.015)
<i>VisuShrink</i>	10	0.77 (0.006)	1.04 (0.009)	0.08 (0.007)	0.27 (0.002)
	7	1.29 (0.012)	1.77 (0.017)	0.12 (0.001)	0.47 (0.005)
	5	2.08 (0.016)	2.99 (0.028)	0.17 (0.002)	0.77 (0.009)
	3	3.69 (0.024)	6.21 (0.057)	0.32 (0.004)	1.55 (0.015)

Table 1: Average mean square error for *BayesThresh*, *VisuShrink*, *GlobalSure*, *Cross-Validation* and *False Discovery Rate* estimators, using various test functions, for various levels of the root signal-to-noise ratio. (Standard errors are given in parentheses.)

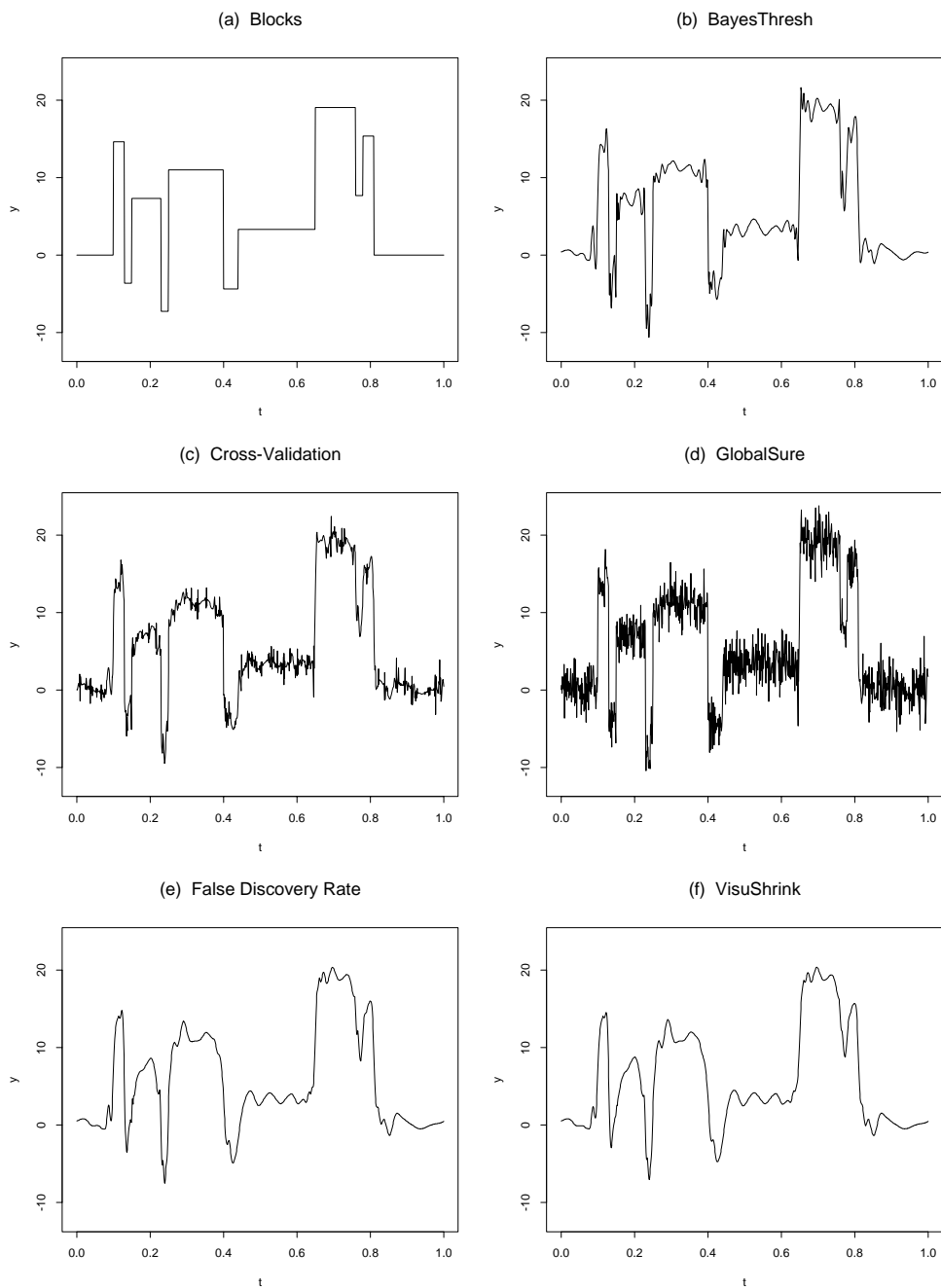


Figure 5: (a) Blocks function; various reconstructions based on 1024 equally-spaced values of the function with the addition of independent  $N(0, \sigma^2)$  noise with  $\sigma = 7/3$  (RSNR = 3): (b) *BayesThresh* ( $\alpha = 0.5$ ,  $\beta = 1$ ); (c) *Cross-Validation*; (d) *GlobalSure*; (e) *False Discovery Rate*; (f) *VisuShrink*.

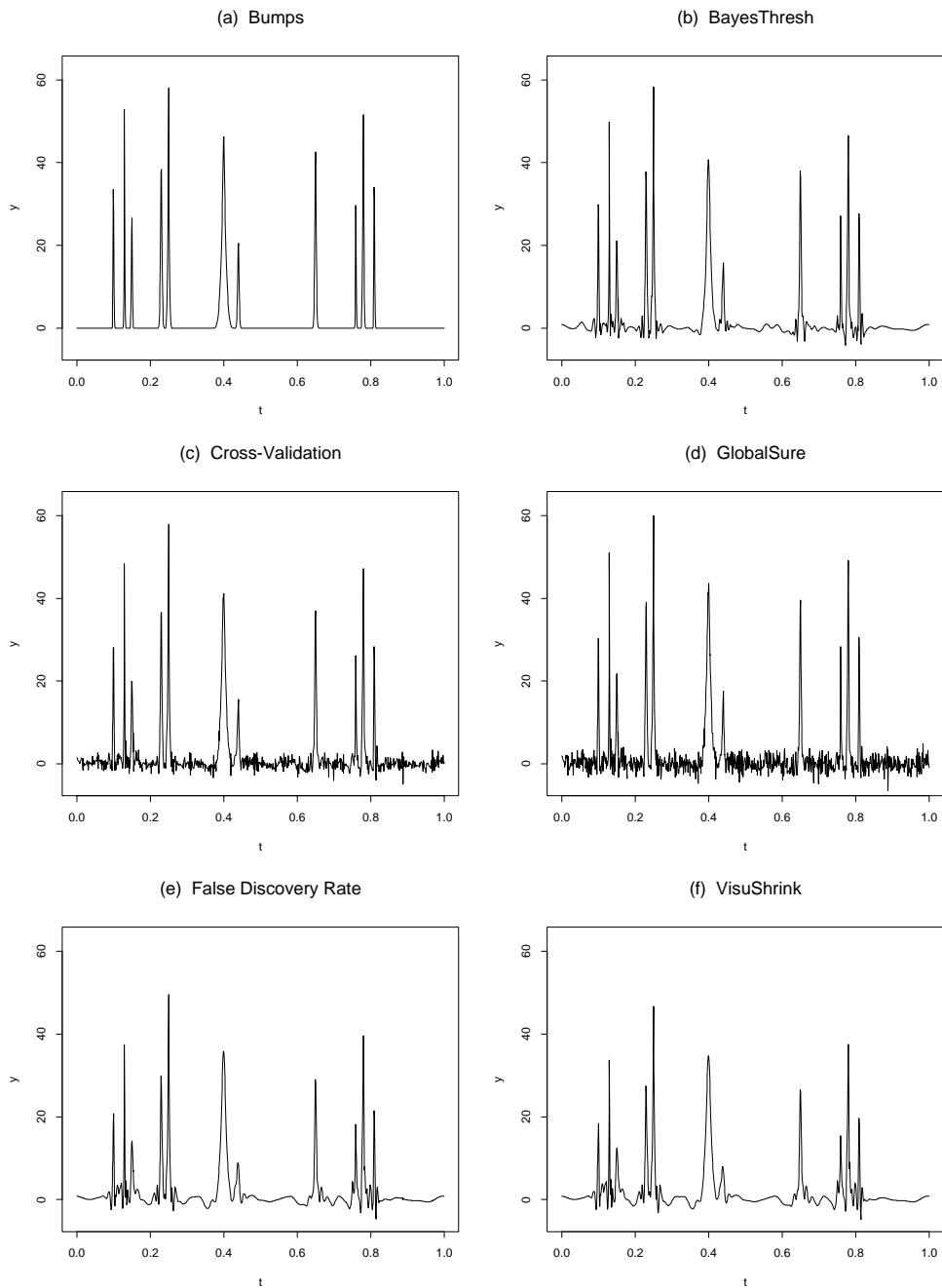


Figure 6: (a) Bumps function; various reconstructions based on 1024 equally-spaced values of the function with the addition of independent  $N(0, \sigma^2)$  noise with  $\sigma = 7/3$  (RSNR = 3): (b) *BayesThresh* ( $\alpha = 0.5$ ,  $\beta = 1$ ); (c) *Cross-Validation*; (d) *GlobalSure*; (e) *False Discovery Rate*; (f) *VisuShrink*.

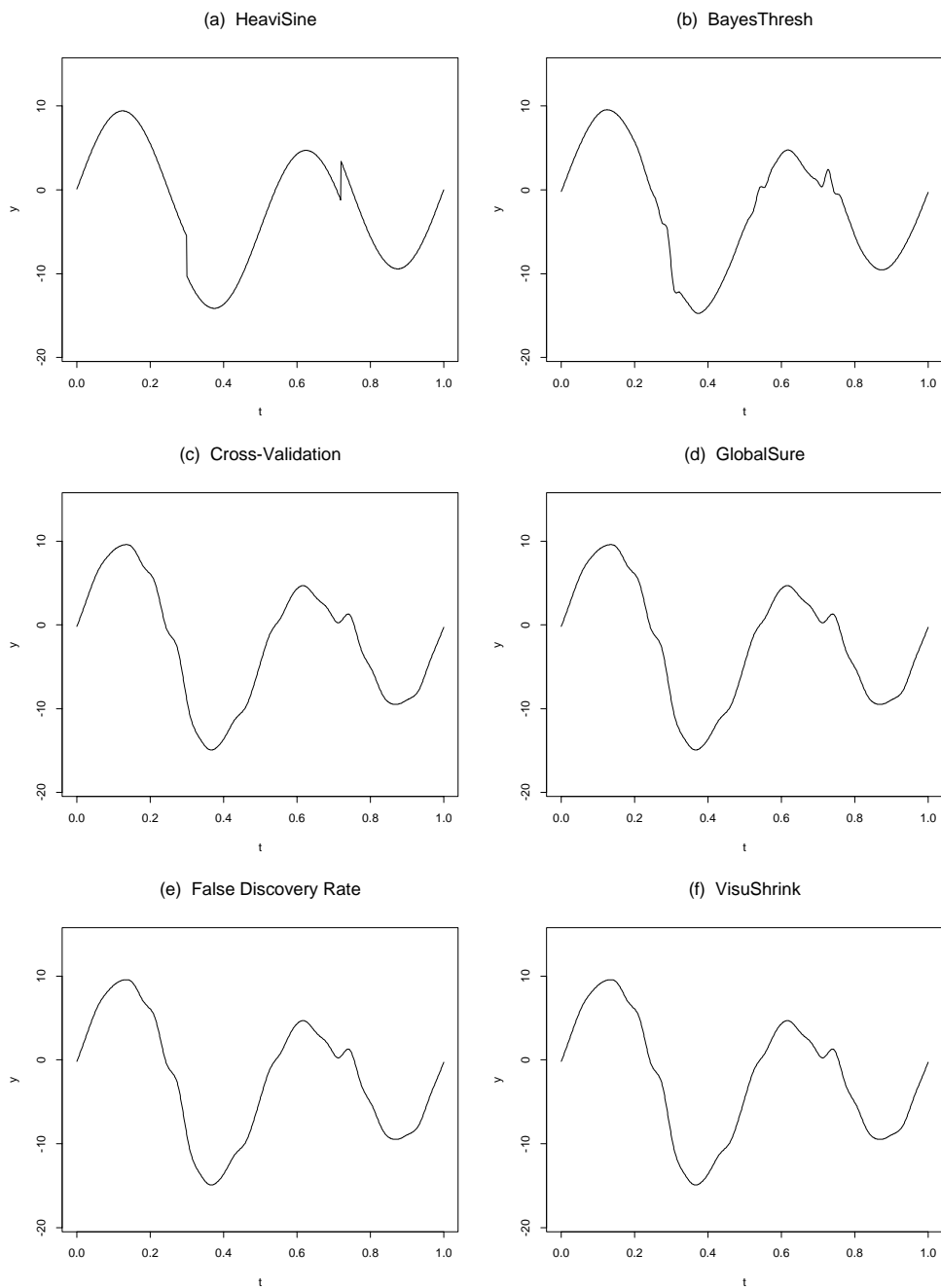


Figure 7: (a) HeaviSine function; various reconstructions based on 1024 equally-spaced values of the function with the addition of independent  $N(0, \sigma^2)$  noise with  $\sigma = 7/3$  (RSNR = 3): (b) *BayesThresh* ( $\alpha = 0.5$ ,  $\beta = 1$ ); (c) *Cross-Validation*; (d) *GlobalSure*; (e) *False Discovery Rate*; (f) *VisuShrink*.

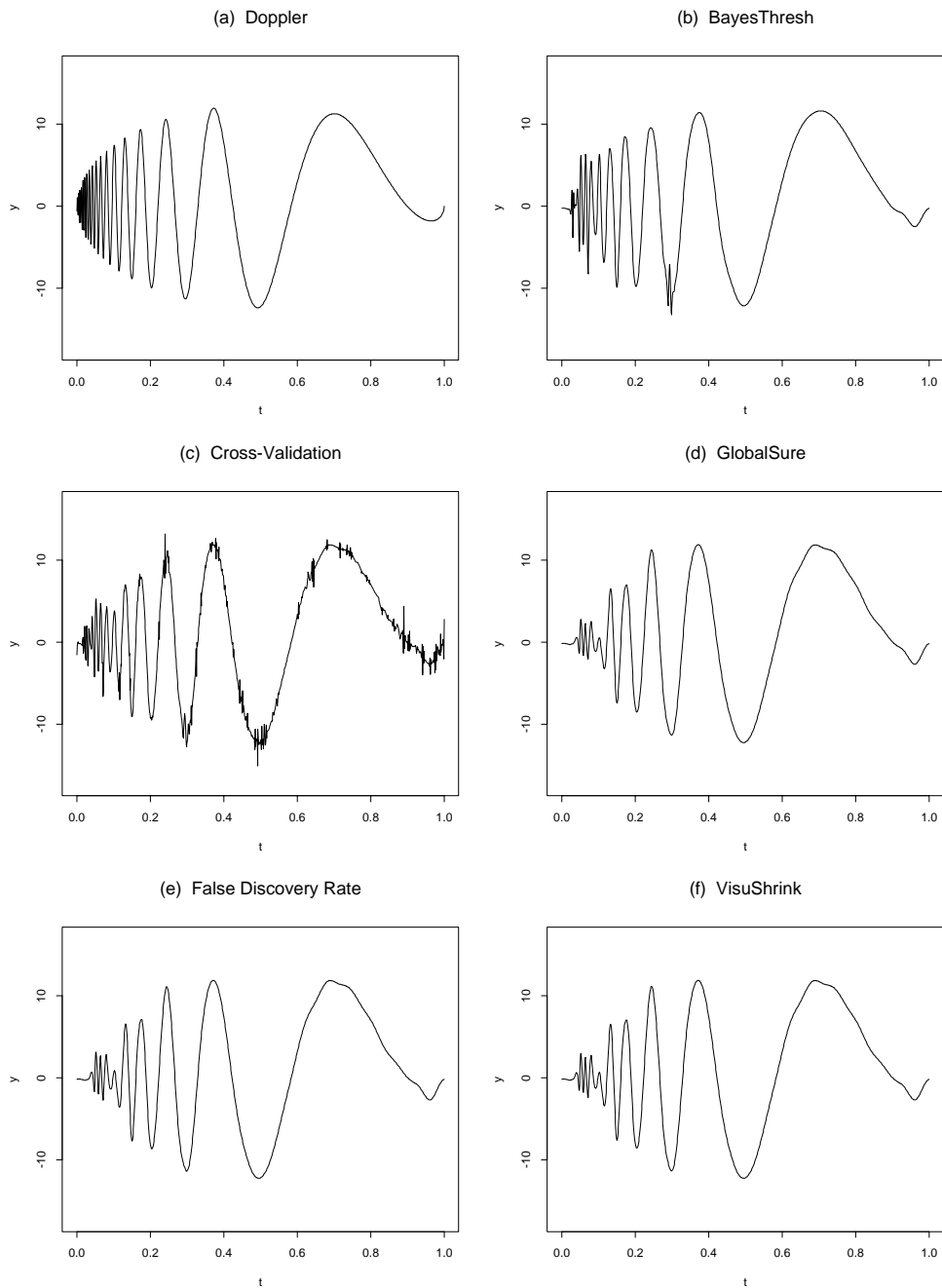


Figure 8: (a) Doppler function; various reconstructions based on 1024 equally-spaced values of the function with the addition of independent  $N(0, \sigma^2)$  noise with  $\sigma = 7/3$  (RSNR = 3): (b) *BayesThresh* ( $\alpha = 0.5$ ,  $\beta = 1$ ); (c) *Cross-Validation*; (d) *GlobalSure*; (e) *False Discovery Rate*; (f) *VisuShrink*.

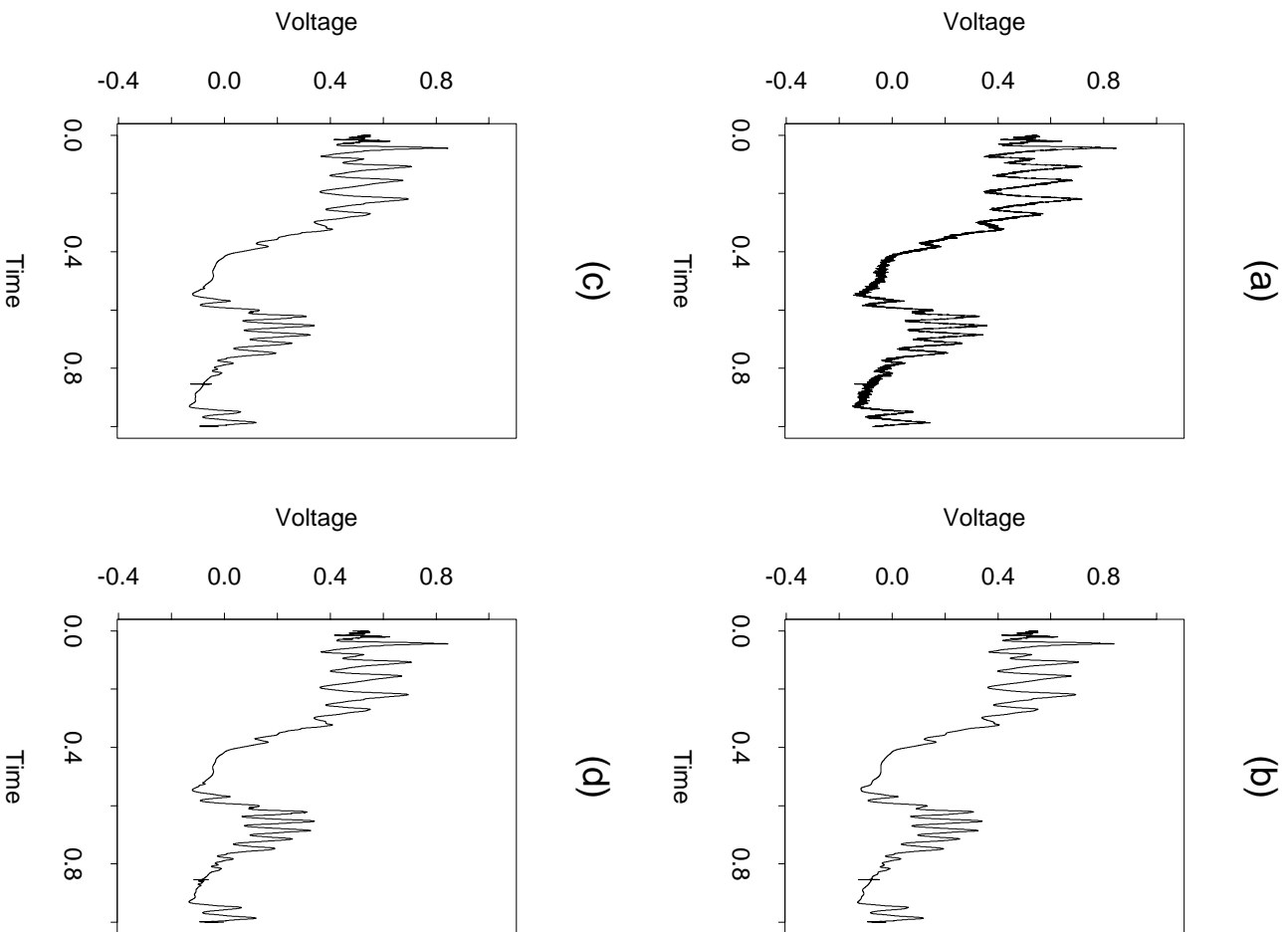


Figure 9: (a) Section of an inductance plethysmograph recording and the curve estimates obtained by: (b) *BayesThresh* ( $\alpha = 0.5, \beta = 1$ ); (c) *BayesThresh* ( $\alpha = 1, \beta = 1$ ); (d) *BayesThresh* ( $\alpha = 2, \beta = 1$ ). It can be seen that the choice of  $\alpha$  does not have an appreciable effect.

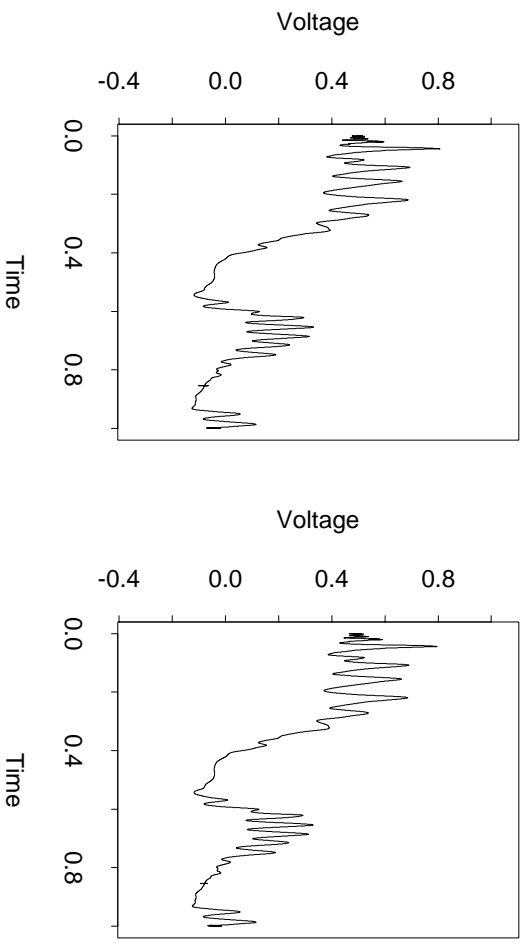
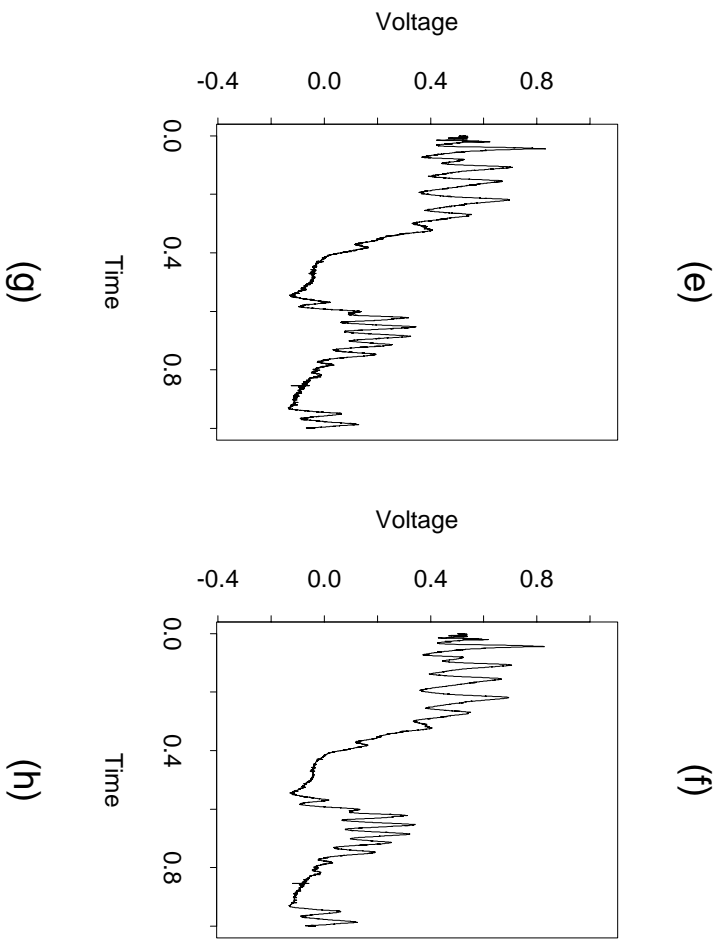


Figure 10: Curve estimates of the inductance plethysmograph recording given in Figure 8 (a), obtained by: (a) *GlobalSure*; (b) *GlobalSure*; (c) *False Discovery Rate*; (d) *VisuShrink*. It can be seen that *Cross-Validation* and *GlobalSure* do not eliminate high frequency noise while *False Discovery Rate* and *VisuShrink* attenuate the peaks (see Table 2).

	Highest peak value
<i>Data</i>	0.847
<i>BayesThresh</i> ( $\alpha = 2, \beta = 1$ )	0.845
<i>BayesThresh</i> ( $\alpha = 1, \beta = 1$ )	0.836
<i>BayesThresh</i> ( $\alpha = 0.5, \beta = 1$ )	0.835
<i>Cross-Validation</i>	0.835
<i>GlobalSure</i>	0.828
<i>False Discovery Rate</i>	0.806
<i>VisuShrink</i>	0.796

Table 2: Highest peak value (first peak) in the curves shown in Figures 8 and 9 for the inductance plethysmography data.