Bayesian Approach To Wavelet Decomposition and Shrinkage

Felix Abramovich
Theofanis Sapatinas

ABSTRACT We consider Bayesian approach to wavelet decomposition. We show how prior knowledge about a function's regularity can be incorporated into a prior model for its wavelet coefficients by establishing a relationship between the hyperparameters of the proposed model and the parameters of those Besov spaces within which realizations from the prior will fall. Such a relation may be seen as giving insight into the meaning of the Besov space parameters themselves. Furthermore, we consider Bayesian wavelet-based function estimation that gives rise to different types of wavelet shrinkage in non-parametric regression. Finally, we discuss an extension of the proposed Bayesian model by considering random functions generated by an overcomplete wavelet dictionary.

1 Introduction

Consider the standard non-parametric regression problem:

$$y_i = g(t_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{1.1}$$

and suppose we wish to recover the unknown function g from additive noise ϵ_i given noisy data y_i at discrete points $t_i = i/n$. Only very general assumptions about g are made like that g belongs to a certain class of functions.

One of the basic techniques in non-parametric regression and signal processing is the generalized Fourier series approach. An unknown response function g is expanded in some orthogonal basis $\{\psi_j\}$:

$$g(t) = \sum_j \omega_j \psi_j(t),$$

where $\omega_j = \langle g, \psi_j \rangle$. The key point for the efficiency of such an approach is obviously a proper choice of a basis. A 'good' basis should allow *parsimonious* expansion for a wide variety of possible responses using a relatively small number of basis functions. The original signal then is represented by the set of its few generalized Fourier coefficients ω_j with high accuracy. Such parsimonious representations may be the key to understanding the basic features of a signal, detecting its regularity, compression, etc. It also plays a crucial role in the non-parametric regression problem (1.1) - instead of estimating g directly, we estimate its generalized Fourier coefficients ω_j and the resulting estimate is $\hat{g}(t) = \sum_j \hat{\omega}_j \psi_j(t)$. In a way, we transfer the original non-parametric problem to an infinitely parametric one.

For a fixed basis, assuming g to belong to a specific class of possible responses, it implicitly or explicitly yields corresponding assumptions on its generalized Fourier coefficients ω_j. Bayesian approach seems only natural to exhibit these assumptions through putting a prior model on ω_j. Examples of Bayesian orthogonal series estimators based on different bases and priors are well known in the literature. Wahba (1983) proposed a Bayesian model for spline smoothing estimation. It turns out that her prior model for the unknown response function is equivalent to placing a certain prior on its Fourier sine and cosine coefficients. Silverman (1985) obtained similar results for B-spline basis. Steinberg (1990) presented a Bayesian model for the coefficients of a function's expansion in a power series of Hermite polynomials.

Here we discuss Bayesian estimators using orthogonal wavelet series. In Section 2, we show how prior knowledge about a function's regularity can be incorporated into a prior model for its wavelet coefficients. A relationship between the hyperparameters of the proposed model and the parameters of those Besov spaces within which realizations from the prior will fall is established. Such a relation may be seen as giving insight into the meaning of the Besov space parameters themselves. Furthermore, in Section 3, we discuss Bayesian wavelet-based estimation that gives rise to different types of wavelet shrinkage in non-parametric regression. In, particular, for the prior specified, we show that a posterior median is a *bona fide* thresholding rule. Finally, in Section 4, we discuss an extension of the proposed Bayesian model by considering random functions generated by an overcomplete wavelet dictionary.

2 Wavelets and Besov spaces

2.1 Wavelet series

Orthogonal wavelet series in $L^2(\mathbb{R})$ are generated by dilations and translations of a mother wavelet ψ: $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$, $j, k \in \mathbb{Z}$. 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. In later sections, however, we confine attention to periodic functions on $\mathbb{R}$ with unit period and work in effect with periodic wavelets (see, for example, Daubechies, 1992, Section 9.3). The wavelet coefficients w_{jk} of the function are then actually restricted to the resolution and spatial indices $j \geq 0$ and $k = 0, \ldots, 2^j - 1$ respectively, and the function can be expanded in the orthogonal wavelet series as:

$$g(t) = c_0\phi(t) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} w_{jk}\psi_{jk}(t),$$

where:

$$c_0 = \int g(t)\phi(t)dt \ , \quad w_{jk} \ = \int g(t)\psi_{jk}(t)dt.$$

The function ϕ called the *scaling function* or the *father* wavelet (see any standard text on wavelets).

Wavelets are local in both time (via translations) and frequency/scale (via dilations) domains. This localization allows parsimonious representation for a wide set of different functions in wavelet series – by choosing the scaling function and the mother wavelet with corresponding regularity properties, one can generate an unconditional wavelet basis in a wide set of function spaces, such as Besov spaces (see Section 2.3 below). For detailed comprehensive expositions of the mathematical aspects of wavelets we refer, for example, to Meyer (1992) and Wojtaszczyk (1997).

We shall assume that the scaling function ϕ and the mother wavelet ψ correspond to an r-regular multiresolution analysis, for some integer $r > 0$ (see, for example, Daubechies, 1992). This will imply that ϕ and ψ are members of the Hölder space C^r, and that ψ has vanishing moments up to order r. For examples of mother wavelets with various regularity properties, and with compact support, see Daubechies (1992).

2.2 Prior model

As we have already mentioned, a large variety of different functions allow parsimonious representation in wavelet series where only a few non-negligible coefficients are present in the expansion. To capture this characteristic feature of wavelet bases, Abramovich, Sapatinas & Silverman (1998a) suggested to place the prior on ω_{jk} of the following form:

$$\omega_{jk} \sim \pi_j N(0, \tau_j^2) + (1 - \pi_j)\delta(0), \quad j \geq 0; \quad k = 0, \ldots, 2^j - 1, \qquad (1.2)$$

where $0 \leq \pi_j \leq 1$, $\delta(0)$ is a point mass at zero, and ω_{jk} are independent. To complete the model a vague prior is placed on the scaling coefficient c_0.

According to the prior model (1.11), every ω_{jk} is either zero with probability $1 - \pi_j$ or with probability π_j is normally distributed with zero mean and variance τ_j^2. The probability π_j gives the proportion of non-zero wavelet coefficients at resolution level j while the variance τ_j^2 is a measure of their magnitudes. Note that the prior parameters π_j and τ_j^2 are the same for all coefficients at a given resolution level j.

The hyperparameters of the prior model (1.2) are assumed to be of the form:

$$\tau_j^2 = c_1 2^{-\alpha j} \qquad \text{and} \qquad \pi_j = \min(1, c_2 2^{-\beta j}),\ \ j \geq 0, \tag{1.3}$$

where c_1, c_2, α, and β are non-negative constants. Some intuitive understanding of the model implied by (1.2), (1.3) can be found in Abramovich, Sapatinas & Silverman (1998a, Section 4.2).

It is interesting to compare the priors (1.2) with the three point 'least favourable' priors of the form:

$$\omega_{jk} \sim \frac{\pi_j}{2}\delta(\mu_j) + \frac{\pi_j}{2}\delta(-\mu_j) + (1-\pi_j)\delta(0) \tag{1.4}$$

used for derivation minimax wavelet estimators. The expressions for π_j and μ_j are given in Donoho & Johnstone (1994) and Johnstone (1994).

Clyde, Parmigiani & Vidakovic (1998) use a similar formulation to (1.2) but with different forms for the hyperparameters π_j and τ_j^2. The prior model (1.2) is also an extreme case of that of Chipman, Kolaczyk & McCulloch (1997). Their prior for each ω_{jk} is the mixture of two normal distributions with zero means but different variances for 'negligible' and 'non-negligible' wavelet coefficients.

2.3 Besov spaces on the interval

Before establishing a relation between the hyperparameters of the prior model (1.2) and the parameters of those Besov spaces within which realizations from the prior will fall, we introduce a brief review of some relevant aspects of the theory of the (inhomogeneous) Besov spaces on the interval that we exploit further. For a more detailed study we refer, for example, to DeVore & Popov (1988), DeVore, Jawerth & Popov (1992), Meyer (1992) and Wojtaszczyk (1997).

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:

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

and by:

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

The Besov space $B^s_{p,q}$ on $[0,1]$ is the class of functions $g : [0,1] \to \mathbb{R}$ for which $g \in L^p[0,1]$ and $|g|_{B^s_{p,q}} < \infty$. The Besov norm is defined then as:

$$\begin{aligned} ||g||_{B^s_{p,q}} &= ||g||_{L^p[0,1]} + |g|_{B^s_{p,q}}, \quad 1 \leq q < \infty, \\ ||g||_{B^s_{p,\infty}} &= ||g||_{L^p[0,1]} + |g|_{B^s_{p,\infty}}. \end{aligned}$$

The (not necessarily integer) 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 ($B^m_{2,2}$) and Hölder ($B^s_{\infty,\infty}$) spaces of smooth functions, but in addition less traditional spaces, like the space of functions of bounded variation, sandwiched between $B^1_{1,1}$ and $B^1_{1,\infty}$. The latter functions are of statistical interest because they allow for better models of spatial inhomogeneity (see, for example, Meyer, 1992; Donoho & Johnstone, 1995).

For $j \geq 0$, define w_j to be the vector of wavelet coefficients w_{jk}, $k = 0, 1, \ldots, 2^j - 1$, as defined in Section 2.1. The Besov norm of g is equivalent to the corresponding sequence space norm:

$$||w||_{b^s_{p,q}} = |c_0| + \left\{ \sum_{j=0}^{\infty} 2^{js'q} ||w_j||_p^q \right\}^{1/q}, \quad \text{if} \quad 1 \leq q < \infty, \qquad (1.5)$$

$$||w||_{b^s_{p,\infty}} = |c_0| + \sup_{j \geq 0} \left\{ 2^{js'} ||w_j||_p \right\}, \qquad (1.6)$$

where $s' = s + 1/2 - 1/p$ (see, for example, Meyer, 1992; Donoho *et al.*, 1995).

In Section 2.4, we exploit this equivalence of the norms for relating prior information about the function's regularity to the hyperparameters of our prior model for the wavelet coefficients w_{jk}.

In the particular case $p = q = 1$ the sequence space norm in (1.5) 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.

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

In this section we demonstrate how knowledge about regularity properties of an unknown response function can be incorporated into the prior model (1.2) for its wavelet coefficients by specifying the hyperparameters of the prior. We explore the connections between the parameters α and β in (1.2) of the prior model (1.11) and the Besov space parameters s and p.

Note first that it follows from (1.3) that the prior expected number of non-zero wavelet coefficients on the j-th level is $C_2 2^{j(1-\beta)}$. Then, appealing to the first Borel-Cantelli lemma, in the case $\beta > 1$, the number of non-zero coefficients in the wavelet expansion is finite almost surely and, hence, with probability one, g will belong to the same Besov spaces as the mother wavelet ψ, i.e. those for which $\max(0, 1/p - 1/2) < s < r,\ 1 \leq p, q \leq \infty$.

More fruitful and interesting is, therefore, 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 α. 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 (see Abramovich, Sapatinas & Silverman, 1998a for details). In general, for the case $0 \leq \beta \leq 1$, the resulting random functions are fractal (rough) (see, Wang, 1997).

Suppose that g is generated from the prior model (1.2) with hyperparameters specified by (1.3). Because of the improper nature of the prior distribution of c_0, we consider the prior distribution of g conditioned on any given value for c_0. The following theorem, proved in Abramovich, Sapatinas & Silverman (1998a), establishes necessary and sufficient conditions for g to fall (with probability one) in any particular Besov space.

Theorem 1 *(Abramovich, Sapatinas & Silverman, 1998a). Let ψ be a mother wavelet that corresponds to an r-regular multiresolution analysis. Consider constants s, p and q such that* $\max(0, 1/p - 1/2) < s < r$, $1 \leq p, q \leq \infty$. *Let the wavelet coefficients w_{jk} of a function g obey the prior model* (1.2) *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 $g \in B^s_{p,q}$ almost surely if and only if either:

$$s + 1/2 - \beta/p - \alpha/2 < 0, \tag{1.7}$$

or:

$$s + 1/2 - \beta/p - \alpha/2 = 0 \ \ \textit{and} \ \ 0 \leq \beta < 1,\ 1 \leq p < \infty,\ q = \infty. \tag{1.8}$$

Remark 2.1. The result of Theorem 1 is true for all values of the Besov space parameter q. This should not be surprising due to the embedding properties of Besov spaces (see, for example, Peetre, 1975). To give some insight on the role of q, Abramovich, Sapatinas & Silverman (1998a) considered a more delicate dependence of the variance τ_j^2 on the level j by adding a third hyperparameter $\gamma \in \mathbb{R} : \tau_j^2 = c_1 2^{-\alpha j} j^{\gamma}$, and extended the results of Theorem 1 for this case (see their Theorem 2).

Theorem 1 essentially includes several important aspects. It shows how prior knowledge about a function's regularity (measured by a Besov space membership) can be incorporated into the prior model (1.2) for its wavelet coefficients by choosing the corresponding hyperparameters of their prior distribution. It may also be seen as giving insight into the meaning of the Besov space parameters themselves and, in a way, attempts to 'exorcise' these 'devilish' spaces for statisticians ('besov' is the literal Russian translation of 'devilish'!). Finally, unlike 'least favourable' realizations implied by the three-point prior (1.4), the priors (1.2) may be preferable to generate 'typical' functions of particular Besov spaces (see, Abramovich, Sapatinas & Silverman, 1998a, Section 4.3) for Bayesian simulation procedures that have become very popular in recent years.

3 Bayesian wavelet estimators

3.1 Wavelet-based thresholding procedure

Before discussing Bayesian wavelet estimators, we review some basic aspects of the wavelet-based thresholding procedure. Recall that according to the original model (1.1), the unknown response function $g(t), t \in [0,1]$ corrupted by 'white' noise is observable at n discrete points $t_i = i/n$:

$$y_i = g(t_i) + \epsilon_i, \qquad i = 1, \dots, n,$$

where ϵ_i are independent and identically distributed normal variables with zero mean and variance σ^2.

Given observed discrete data $\mathbf{y} = (y_1, \dots, y_n)^{\mathrm{T}}$, 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 orthogonal DWT-matrix with (jk, i) entry given by:

$$\sqrt{n}\, W_{jk,i} \approx \psi_{jk}(i/n) = 2^{j/2}\psi(2^j i/n - k).$$

As usual, we assume that $n = 2^J$ for some positive integer J. Then the DWT yields $(n-1)$ sample discrete wavelet coefficients $\hat{d}_{jk}$, $j = 0, \dots, J-1$; $k = 0, \dots, 2^j - 1$, and one sample scaling coefficient $\hat{c}_0$, which is the sample mean $\bar{y}$ multiplied by $\sqrt{n}$.

Both DWT and inverse DWT are performed by Mallat's (1989) fast algorithm that requires only $O(n)$ operations. Due to the orthogonality of $\mathcal{W}$, the DWT of a white noise is also an array ε_{jk} of independent $N(0, \sigma^2)$, so

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

where the discrete wavelet coefficients d_{jk} are the DWT of the vector of discrete function values $(g(t_1), ..., g(t_n))^{\mathrm{T}}$ and are related to the 'theoretical' wavelet coefficients $w_{jk} = \int g(t)\psi_{jk}(t)\ dt$ by $d_{jk} \approx \sqrt{n}\ w_{jk}$. The $\sqrt{n}$ factor essentially arises from the difference Between continuous and discrete orthogonality conditions. This factor cannot be avoided and, therefore, we use different letters d_{jk} and w_{jk} to clarify the distinction.

As we have discussed before, wavelets allow parsimonious representation for a wide variety of functions so it is reasonable to assume that only a few 'large' $\hat{d}_{jk}$ really contain information about the unknown function g, while the 'small' coefficients are attributed to the noise. The extraction of those 'significant' coefficients can be naturally done by thresholding $\hat{d}_{jk}$'s:

$$\begin{aligned} \hat{d}^{\star}_{jk} &= \hat{d}_{jk}\ \mathrm{I}(|\hat{d}_{jk}| > \lambda) \quad \text{(hard thresholding)} && (1.9) \\ \hat{d}^{\star}_{jk} &= \mathrm{sign}(\hat{d}_{jk})\ \max(0, |\hat{d}_{jk}| - \lambda) \quad \text{(soft thresholding),} && (1.10) \end{aligned}$$

where $\lambda \geq 0$ is a threshold value. The hard thresholding is a 'keep' or 'kill' rule, while the soft thresholding is a 'shrink' or 'kill' rule. The resulting coefficients $\hat{d}^{\star}_{jk}$ are then used for selective reconstruction of an estimate by the inverse DWT:

$$\hat{\mathbf{g}} = \mathcal{W}^{\mathrm{T}}\hat{\mathbf{d}}^{\star}$$

The choice of λ is obviously crucial: small/large threshold values will produce estimates that tend to overfit/underfit the data. Donoho & Johnstone (1994) proposed the *universal* threshold $\lambda_{DJ} = \sigma\sqrt{2\log n}$. Despite the 'triviality' of such a threshold, they showed that the resulting wavelet estimator is asymptotically near-minimax among all estimators within the whole range of Besov spaces. Wang (1996) and Johnstone & Silverman (1997) studied corresponding universal thresholds for the case of 'coloured' noise. Abramovich & Silverman (1998) derived universal thresholds for wavelet estimators based on *indirect* data in inverse problems. However, the universal threshold essentially 'ignores' the data and, hence, it is not 'tuned' to the specific problem at hand.

Several *data-adaptive* thresholding rules have been developed recently. Donoho & Johnstone (1995) proposed the *SureShrink* thresholding rule which is based on minimizing the Stein's unbiased risk estimate (Stein, 1981). Abramovich & Benjamini (1996), Ogden & Parzen (1996a, 1996b) considered thresholding as multiple hypotheses testing procedure. Nason (1996), Jansen, Malfait & Bultheel (1997) adjusted the well known cross-validation approach for choosing λ.

Bayesian approaches to thresholding were recently explored by Chipman, Kolaczyk & McCulloch (1997), Abramovich, Sapatinas & Silverman (1998a), Clyde & George (1998), Clyde, Pargimiani & Vidakovic (1998), Crouse, Nowak & Baraniuk (1998), Johnstone & Silverman (1998) and Vidakovic (1998) among others, and some of them will be discussed in detail below.

3.2 Bayesian wavelet shrinkage

In this section we discuss a Bayesian formalism that leads to different types of wavelet shrinkage estimators.

For the discrete wavelet coefficients d_{jk}, the corresponding prior model will be:

$$d_{jk} \sim \pi_j N(0, \tau_j^2) + (1 - \pi_j)\delta(0), \quad j = 0, \ldots, J-1; \quad k = 0, \ldots, 2^j - 1, \tag{1.11}$$

where the hyperparameters are of the form:

$$\tau_j^2 = C_1 2^{-\alpha j} \quad \text{and} \quad \pi_j = \min(1, C_2 2^{-\beta j}), \quad j = 0, \ldots, J-1, \tag{1.12}$$

with $C_1 = nc_1, C_2 = c_2$.

Subject to the prior (1.11), the posterior distribution $d_{jk}|\hat{d}_{jk}$ is also a mixture of a corresponding posterior normal distribution and $\delta(0)$. Letting Φ be the standard normal cumulative distribution function, the posterior cumulative distribution function $F(d_{jk}|\hat{d}_{jk})$ is:

$$F(d_{jk} \mid \hat{d}_{jk}) = \frac{1}{1+\eta_{jk}} \Phi \left(\frac{d_{jk} - \hat{d}_{jk}\tau_j^2/(\sigma^2 + \tau_j^2)}{\sigma\tau_j/\sqrt{\sigma^2 + \tau_j^2}} \right) + \frac{\eta_{jk}}{1+\eta_{jk}} \mathrm{I}(\hat{d}_{jk} \geq 0), \tag{1.13}$$

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

$$\eta_{jk} = \frac{1-\pi_j}{\pi_j} \frac{\sqrt{\tau_j^2 + \sigma^2}}{\sigma} \exp\left(-\frac{\tau_j^2 \hat{d}_{jk}^2}{2\sigma^2(\tau_j^2 + \sigma^2)} \right). \tag{1.14}$$

Different losses lead to different Bayesian rules. The traditional Bayes rule usually considered in the literature (see, for example, Chipman, Kolaczyk & McCullagh, 1997; Clyde, Pargimiani & Vidakovic, 1998; Vidakovic, 1998) corresponds to the L^2-loss and yields the posterior mean. Using (1.13) and (1.14), we then have:

$$E(d_{jk} \mid \hat{d}_{jk}) = \frac{1}{1+\eta_{jk}} \frac{\tau_j^2}{\tau_j^2 + \sigma^2} \hat{d}_{jk}. \tag{1.15}$$

Obviously, such a rule is never a thresholding rule but a (nonlinear) smoothing shrinkage. Instead, Abramovich, Sapatinas & Silverman (1998a) suggested the use of the posterior median that corresponds to the L^1-loss and leads to a *bona fide* thresholding rule. To fix terminology, a *shrinkage* rule shrinks wavelet coefficients towards zero, whilst a *thresholding* rule in addition sets actually to zero all coefficients below a certain threshold. As explained in Section 2.3, 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 ω_{jk}. Thus, whichever weighted combination is used, the corresponding Bayes rule will be obtained by taking the posterior median of each wavelet coefficient. By following Abramovich, Sapatinas & Silverman (1998a), one gets the following closed form for the posterior medians:

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

where:

$$\zeta_{jk} = \frac{\tau_j^2}{\sigma^2 + \tau_j^2} |\hat{d}_{jk}| - \frac{\tau_j \sigma}{\sqrt{\sigma^2 + \tau_j^2}} \Phi^{-1} \left(\frac{1 + \min(\eta_{jk}, 1)}{2} \right). \qquad (1.16)$$

The quantity ζ_{jk} is negative for all $\hat{d}_{jk}$ in some implicitly defined interval $[-\lambda_j, \lambda_j]$, and hence $\mathrm{Med}(d_{jk}|\hat{d}_{jk})$ is zero whenever $|\hat{d}_{jk}|$ falls below the threshold λ_j. The posterior median is therefore a level-dependent 'kill' or 'shrink' thresholding rule with thresholds λ_j.

Abramovich, Sapatinas & Silverman (1998a) called this Bayesian thresholding procedure *BayesThresh*. Note that, unlike soft thresholding (1.10), extent of shrinkage in *BayesThresh* depends on $|\hat{d}_{jk}|$: the larger $|\hat{d}_{jk}|$, the less it is shrinked. For large $\hat{d}_{jk}$ the *BayesThresh* asymptotes to linear shrinkage by a factor of $\tau_j^2/(\sigma^2 + \tau_j^2)$, since the second term in (1.16) becomes negligible as $|\hat{d}_{jk}| \to \infty$.

Remark 3.1. The universal threshold $\lambda_{DJ} = \sigma\sqrt{2\log n}$ of Donoho & Johnstone (1994) can be also obtained as a particular limiting case of *BayesThresh* rule setting $\alpha = \beta = 0$ and letting $C_1 \to \infty$, $C_2 \to 0$ as n increases in such a way that $\sqrt{C_1}/(C_2\sigma n) \to 1$. Such a peculiar prior is a direct consequence of its 'least favourable' nature.

Another way to obtain a *bona fide* thresholding rule within a Bayesian framework is via a hypothesis testing approach (see, Vidakovic, 1998). The idea is simple: after observing $\hat{d}_{jk}$, test the hypothesis $H_0 : d_{jk} = 0$ against a two-sided alternative $H_1 : d_{jk} \neq 0$. If the hypothesis H_0 is rejected, d_{jk} is estimated by $\hat{d}_{jk}$, otherwise $d_{jk} = 0$. Such a procedure essentially mimics the hard thresholding rule:

$$\hat{d}^{\star}_{jk} = \hat{d}_{jk} I(\eta_{jk} < 1), \qquad (1.17)$$

where:

$$\eta_{jk} = P(H_0 \mid \hat{d}_{jk}) / P(H_1 \mid \hat{d}_{jk})$$

is the posterior odds ratio. Vidakovic (1998) called this thresholding rule Bayes factor (*BF*) thresholding since the posterior odds ratio is obtained by multiplying the Bayes factor with the prior odds ratio. Thus, a wavelet coefficient $\hat{d}_{jk}$ will be thresholded if the corresponding posterior odds ratio $\eta_{jk} > 1$ and will be kept as it is otherwise, where η_{jk} for our prior model (1.11), (1.12) is given by (1.14).

To compare *BayesThresh* and *BF*, note that *BF* is always a 'keep' or 'kill' hard thresholding, whilst *BayesThresh* is a 'shrink' or 'kill' thresholding, where extend of shrinkage depends on the absolute values of the wavelet coefficients. In addition, *BF* thresholds $\hat{d}_{jk}$ if the corresponding $\eta_{jk} > 1$. One can verify from (1.16) that *BayesThresh* will 'kill' those $\hat{d}_{jk}$, whose:

$$\eta_{jk} > 1 - 2\Phi\left(-\frac{\tau_j|\hat{d}_{jk}|}{\sigma\sqrt{\sigma^2+\tau_j^2}}\right)$$

and, hence, will threshold more coefficients. Figure 1 shows the different Bayesian rules for some choices of the hyperparameters.

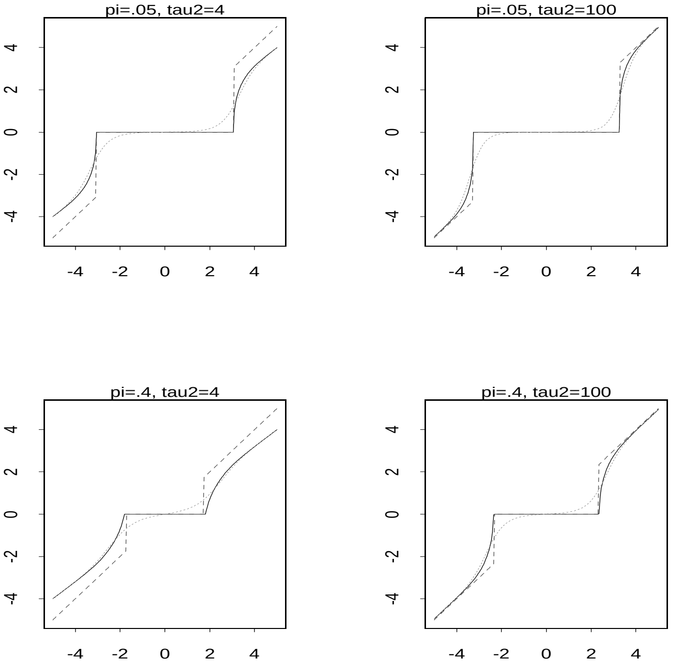

FIGURE 1. The posterior medians (solid lines), the posterior means (dotted lines) and the *BF* (dashed lines) rules as functions of the empirical wavelet coefficients for some choices of the hyperparameters π and τ^2, while σ was fixed at 1.

3.3 Estimation of the hyperparameters

To apply the Bayesian wavelet-based methods discussed in Section 3.2 in practice, it is necessary first to specify the hyperparameters α, β, C_1 and C_2 in (1.12). Ideally, the choice of α and β should be made from prior knowledge about regularity properties of the unknown function making use of the results of Theorem 1. Some practical issues for the choice of α and β have been investigated by Abramovich, Sapatinas & Silverman (1998a) and will be briefly discussed further in Section 3.4 below.

In what follows, we assume that α and β have been chosen in advance (or they are known quantities), and estimate C_1 and C_2 by the following procedure in the spirit of empirical Bayes as suggested by Abramovich, Sapatinas & Silverman (1998a).

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 universal threshold of Donoho & Johnstone (1994) $\lambda_{DJ} = \sigma\sqrt{2\log n}$ described above in Section 3.1. Donoho & Johnstone (1994) showed that the probability that even one negligible coefficient will pass the threshold value λ_{DJ} tends to zero, so essentially only non-negligible $\hat{d}_{jk}$ will survive after universal thresholding. Suppose that, on level j, the number of coefficients that pass λ_{DJ} is M_j, and that the values of these coefficients are $x_{j1}, \ldots, x_{jM_j}$. Conditioning on the value M_j, the x_{jm}, $m = 1, \ldots, M_j$, 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:

$$\begin{aligned} l(\tau_0^2, \ldots, \tau_{J-1}^2) = & - \sum_{j=0}^{J-1} M_j \left\{ \frac{1}{2}\log(\sigma^2 + \tau_j^2) - \log\left(\Phi\left[-\frac{\lambda_{DJ}}{\sqrt{\sigma^2 + \tau_j^2}} \right] \right) \right\} \\ & - \sum_{j=0}^{J-1} \left\{ \frac{1}{2(\sigma^2 + \tau_j^2)} \sum_{m=1}^{M_j} x_{jm}^2 \right\}. \end{aligned} \tag{1.18}$$

Substituting $\tau_j^2 = C_1 2^{-\alpha j}$ and $\lambda_{DJ} = \sigma\sqrt{2\log n}$, and given the values of α and σ, we can obtain an estimate of C_1 by a numerical maximization of (1.18).

The parameter C_2 can be chosen by a cognate procedure. We use the numbers $M_0, \ldots, M_{J-1}$ of coefficients passing the threshold to estimate the π_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 λ_{DJ}. Neglecting the possibility that any $\hat{d}_{jk}$ corresponding to a zero d_{jk} passes the threshold λ_{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 β, a simple method of moments estimate of

C_2 based on the total imputed number of non-zero d_{jk} is:

$$\hat{C}_2 = \frac{2^{(1-\beta)}-1}{2^{(1-\beta)J}-1}\sum_{j=0}^{J-1}\frac{M_j}{q_j}, \quad \text{if } 0 \le \beta < 1,$$

$$\hat{C}_2 = \frac{1}{J}\sum_{j=0}^{J-1}\frac{M_j}{q_j}, \quad \text{if } \beta = 1.$$

Note also that if the noise level σ is unknown, it is usual in practice to estimate it robustly by the median absolute deviation of the wavelet coefficients at the finest level, $\hat{d}_{J-1,k} : k = 0, 1, \ldots, 2^{J-1} - 1$, divided by 0.6745 (see, Donoho & Johnstone, 1994). Alternatively, one can adapt a fully Bayesian approach by placing a prior on σ^2 and considering a hierarchical Bayesian model (see, for example, Clyde, Pargimiani & Vidakovic, 1998; Vidakovic, 1998).

We point out that empirical Bayes approaches (conditional maximum likelihood, marginal maximum likelihood) for estimating the hyperparameters π_j and τ_j^2, at each resolution level j separately, in the general form (1.11) have been recently considered by Clyde & George (1998), Johnstone & Silverman (1998).

3.4 Simulations

Abramovich, Sapatinas & Silverman (1998a) performed a comprehensive simulation study to compare *BayesThresh* procedure with standard non-Bayesian thresholding rules. They considered the ‘Blocks’, ‘Bumps’, ‘Heavisine’ and ‘Doppler’ test functions of Donoho & Johnstone (1994) that caricature spatially variable signals arising in diverse scientific fields and have become standard tests for wavelet estimators.

The results showed that, for various signal-to-noise ratios for all test functions, *BayesThresh* compares favourably with its non-Bayesian counterparts (see, Abramovich, Sapatinas & Silverman, 1998a for details). In particular, for $\alpha = 0.5$ and $\beta = 1$, *BayesThresh* outperformed all non-Bayesian estimators in almost all cases in terms of the mean square error. Abramovich, Sapatinas & Silverman (1998a) suggested to make it a ‘standard default’ choice for prior hyperparameters when prior knowledge about a function’s regularity properties is difficult to elicit.

We have continued the study of Abramovich, Sapatinas & Silverman (1998a). Using the same test functions and for the same signal-to-noise ratios, we have compared different Bayesian wavelet procedures discussed in Section 3.2: posterior means (1.15), posterior medians (*BayesThresh*) and Bayes Factor (1.17) using the ‘standard’ choice $\alpha = 0.5$ and $\beta = 1$. The three Bayesian methods yield quite similar results. Usually, posterior means have a smaller mean square error, *BayesThresh* second with *BF* very close to it. All of them outperformed their non-Bayesian competitors in all cases.

4 Stochastic expansions in an overcomplete wavelet dictionary

4.1 From bases to dictionaries

In recent years there has been growing interest in the atomic decomposition of functions in overcomplete dictionaries (see, for example, Mallat & Zhang, 1993; Davis, Mallat & Zhang, 1994; Chen, Donoho & Saunders, 1999). Every basis is essentially only a *minimal* necessary dictionary needed to represent a large variety of different functions. Such 'miserly' representation usually causes poor adaptivity (Mallat & Zhang, 1993). The use of *overcomplete* dictionaries increases the adaptivity of the representation, because one can choose now the most suitable one among many available. One can see an interesting analogy with colours. Theoretically, every other colour can be generated by combining three basic colours (green, red and blue) in corresponding proportions. However, a painter would definitely prefer to use the whole available palette (overcomplete dictionary) to get the hues he needs!

In mathematical terms, an atomic decomposition of a function (signal) g is an expression of g as a superposition of a parametric collection of waveforms $(\psi_\lambda)_{\lambda\in\Lambda}$:

$$g(t) = \sum_{\lambda\in\Lambda} \omega_\lambda \psi_\lambda(t).$$

The collection of waveforms $(\psi_\lambda)_{\lambda\in\Lambda}$ is called a *dictionary*, and the waveforms ψ_λ are called *atoms*.

Here, we naturally focus on wavelet dictionaries. The atoms of a wavelet dictionary $\mathcal{D}_\Lambda = \{\psi_\lambda : \lambda \in \Lambda\}$ with the set Λ of indices $\lambda = (a, b)$ are translations and dilations of a single mother wavelet and are of the form:

$$\psi_\lambda(t) = a^{1/2}\psi(a(t-b)), \quad a \geq 1, \quad 0 \leq b \leq 1.$$

In particular, for the orthonormal wavelet dictionary $\Lambda = \{(2^j, k2^{-j}),\ j \geq 0,\ k = 0, ..., 2^j - 1\}$. *Overcomplete* wavelet dictionaries are obtained by sampling indices more finely. An important example of overcomplete wavelet dictionaries is the non-decimated (or stationary or translation-invariant) wavelet dictionary (see, for example, Coifman & Donoho, 1995; Nason & Silverman, 1995). Atomic decompositions in overcomplete dictionaries are obviously nonunique and one may think about choosing the 'best' possible representation among many available (see, for example, Mallat & Zhang, 1993; Davis, Mallat & Zhang, 1994; Chen, Donoho & Saunders, 1999).

4.2 Prior model

Consider the overcomplete wavelet dictionary where the scales and dilations of wavelet atoms ψ_λ are not dyadic constraints any longer, but arbitrary. To extend the prior model (1.2) for orthonormal wavelet bases, Abramovich, Sapatinas & Silverman (1998b) modelled the set of the locations of wavelet atoms and their magnitudes as being sampled from a certain marked Poisson process.

More specifically, let the set Λ of indices $\lambda = (a, b)$ be sampled from a Poisson process S on $[1, \infty) \times [0, 1]$ with intensity $\mu(\lambda)$. Conditional on S, the corresponding coefficients ω_λ are assumed to be independent normal random variables:

$$\omega_\lambda \mid S \sim N(0, \tau^2(\lambda)). \tag{1.19}$$

To complete the model it is assumed that both the variance $\tau^2(\lambda)$ and the intensity $\mu(\lambda)$ depend on the scale a only, and are of the form:

$$\tau_a^2 \propto a^{-\delta} \qquad \text{and} \qquad \mu_a \propto a^{-\zeta}, \quad a \geq 1, \tag{1.20}$$

where δ, $\zeta \geq 0$, with $\delta + \zeta > 0$.

The intuitive basis of the proposed model is an extension of the notion that the orthogonal wavelet series representation of an unknown function is sparse. The parameter ζ controls the relative rarity of 'fine-scale' wavelet atoms in the function, while the parameter δ controls the size of the contribution of these atoms when they appear. For example, if ζ is small and δ is large, there will be a considerable number of 'fine-scale' atoms but these will each have fairly low contribution, so one might expect the functions to be reasonably smooth and homogeneous. On the other hand, if ζ is large and δ is small, there will be occasional large 'fine-scale' effects in the functions.

4.3 Regularity properties of random functions

Consider now a random function g generated by the wavelet dictionary $(\psi_\lambda)_{\lambda \in \Lambda}$:

$$g(t) = \sum_{\lambda \in \Lambda} \omega_\lambda \psi_\lambda(t), \tag{1.21}$$

where the random locations λ of atoms and their random magnitudes ω_λ obey the prior (1.19), (1.20). The following Theorem 2 proved in Abramovich, Sapatinas & Silverman (1998b) establishes a relation between the hyperparameters ζ and δ of the prior and the parameters s and p of those Besov spaces within which g will fall (with probability one), extending thus Theorem 1 for orthonormal wavelet bases.

Note that, for $\zeta > 1$, the intensity $\mu_a \propto a^{-\zeta}$ is integrable over the range of λ for which ψ_λ has support intersecting $[0, 1]$. Therefore, the number of relevant terms in the atomic decomposition (1.21) is finite almost surely and, hence, with probability one, g will belong to the same Besov spaces as the mother wavelet ψ, namely those for which $\max(0, 1/p - 1/2) < s < r$, $1 \leq p \leq \infty$, $1 \leq q \leq \infty$. The more interesting case is again $0 \leq \zeta \leq 1$.

Theorem 2 *(Abramovich, Sapatinas & Silverman, 1998b). Let ψ be a compactly supported mother wavelet that corresponds to an r-regular multiresolution analysis. Consider constants s, p and q such that* $\max(0, 1/p - 1/2) < s < r$, $1 \leq p, q \leq \infty$. *Consider a function g as defined in (1.21), with the conditional variances $\tau_a^2 \propto a^{-\delta}$ and the intensity of the Poisson process $\mu_a \propto a^{-\zeta}$. Assume that $\delta \geq 0$, $0 \leq \zeta \leq 1$, and that $\delta + \zeta > 0$. Assume also that the wavelets are sufficiently regular that $\delta < 2r + 2\rho - 1$.*

Then $g \in B^s_{p,q}$ almost surely if and only if

$$s + 1/2 - \zeta/p - \delta/2 < 0. \tag{1.22}$$

Theorem 2 establishes a sufficient and necessary condition for realizations to fall in a particular Besov space. It shows that the function's smoothness measured by the parameter s depends both on the intensity of 'fine-scale' atoms (via ζ) and their magnitudes (via δ). The parameter p can be seen as 'discouraging inhomogeneity', in that the larger the value of p the more emphasis is placed on the parameter δ. For large δ, no matter how many 'fine-scale' atoms there are, they each make a relatively low contribution. On the other hand, if p is small, then there is a trade-off where large weights on 'fine-scale' atoms (small δ) can be tolerated if the corresponding atoms are relatively rare (large ζ).

Theorem 2 makes it possible in principle to incorporate prior knowledge about a function's regularity into a prior model for its atomic wavelet representation. The models considered in this section show how Bayesian ideas can be extended to a broader range of wavelet models, freed from the dyadic positions and scales considered in the classical case. The algorithmic details, probably involving modern Bayesian computational methods, have yet to be worked out in detail and are an interesting subject for future research. The improvement to 'standard' wavelet methods obtained by moving from the discrete (decimated) wavelet transform to the non-decimated wavelet transform (see, for example, Coifman & Donoho, 1995; Nason & Silverman, 1995; Lang *et al.*, 1996; Johnstone & Silverman, 1997) suggest that a Bayesian approach based on a general atomic decomposition may result in yet better performing wavelet shrinkage estimators.

Acknowledgements

This work was started when Dr Theofanis Sapatinas was a Research Associate in the School of Mathematics, University of Bristol, supported by EPSRC grant GR/K70236.

References

Abramovich, F. & Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis* **22**, 351-361.

Abramovich, F. & Silverman, B.W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115-129.

Abramovich, F., Sapatinas, T. & Silverman, B.W. (1998a). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society*, Ser. B **60**, 725-749.

Abramovich, F., Sapatinas, T. & Silverman, B.W. (1998b). Stochastic expansions in an overcomplete wavelet dictionary. *Probability Theory and Related Fields* (under invited revision).

Chen, S.S.B., Donoho, D.L. & Saunders, M.A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, **20**, 33-61.

Chipman, H.A., Kolaczyk, E.D. & McCulloch, R.E. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association* **92**, 1413-1421.

Clyde, M. & George, E.I. (1998). Robust empirical Bayes estimation in wavelets. *Discussion Paper* **98-21**, Institute of Statistics and Decision Sciences, Duke University, USA.

Clyde, M., Parmigiani, G. & Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391-401.

Cohen, A., Daubechies, I., Jawerth, B. & Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis* **1**, 54-81.

Coifman, R.R. & Donoho, D.L. (1995). Translation-invariant de-noising. In *Wavelets and Statistics*, Lecture Notes in Statistics **103**, Antoniadis, A. and Oppenheim, G. (Eds.), pp. 125-150, New York: Springer-Verlag.

Coifman, R.R. & Wickerhauser, M.V. (1992). Entropy-based algorithms for best-basis selection. *IEEE Transactions on Information Theory* **38**, 713-718.

Crouse, M., Nowak, R. & Baraniuk, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing* **46**, 886-902.

Daubechies, I. (1988). Time-frequency localization operators: a geometric phase space approach. *IEEE Transactions on Information Theory* **34**, 605-612.

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

Davis, G., Mallat, S.G. & Zhang, Z. (1994). Adaptive time-frequency approximations with matching pursuit. In *Wavelets: Theory, Algorithms, and Applications*, Chui, C.K., Montefusco, L. and Puccio, L. (Eds.), pp. 271-293, San Diego: Academic Press.

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.

Jansen, M., Malfait, M. & Bultheel, A. (1997). Generalized cross validation for wavelet thresholding. *Signal Processing* **56**, 33-44.

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.

Johnstone, I.M. & Silverman, B.W. (1998). Empirical Bayes approaches to mixture problems and wavelet regression. *Technical Report*, Department of Mathematics, University of Bristol, UK.

Lang, M., Guo, H., Odegard, J.E., Burrus, C.S. & Wells Jr, R.O. (1996). Noise reduction using an undecimated discrete wavelet transform. *IEEE Signal Processing Letters* **3**, 10-12.

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

Mallat, S.G. & Zhang, Z. (1993). Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing* **41**, 3397-3415.

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

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 and Graphical Statistics* **3**, 163-191.

Nason, G.P. & Silverman, B.W. (1995). The stationary wavelet transform and some statistical applications. In *Wavelets and Statistics*, Lecture Notes in Statistics **103**, Antoniadis, A. and Oppenheim, G. (Eds.), pp. 281-300, New York: Springer-Verlag.

Ogden, T. & Parzen, E. (1996a). Data dependent wavelet thresholding in nonparametric regression with change-point applications. *Computational Statistics and Data Analysis* **22**, 53-70.

Ogden, T. & Parzen, E. (1996b). Change-point approach to data analytic wavelet thresholding. *Statistics and Computing* **6**, 93-99.

Peetre, J. (1975). *New Thoughts on Besov Spaces.* Durham: Duke University Press.

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.

Steinberg, D.M. (1990). A Bayesian approach to flexible modeling of multivariate response for functions. *Journal of Multivariate Analysis* **34**, 157-172.

Vidakovic, B. (1998). Non-linear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association* **93**, 173-179.

Wahba, G. (1983). Bayesian ‘confidence intervals’ for the cross-validated smoothing spline. *Journal of the Royal Statistical Society*, Ser. B **45**, 133-150.

Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Annals of Statistics* **24**, 466-484.

Wang, Y. (1997). Fractal function estimation via wavelet shrinkage. *Journal of the Royal Statistical Society*, Ser. B **59**, 603-612.

Wojtaszczyk, P. (1997). *A Mathematical Introduction to Wavelets.* Cambridge: Cambridge University Press.