BAYESIAN WAVELET SHRINKAGE FOR NONPARAMETRIC MIXED-EFFECTS MODELS

Su-Yun Huang and Henry Horng-Shing Lu

Academia Sinica, Taipei, and National Chiao Tung University

Abstract

The main purpose of this article is to study the wavelet shrinkage method from a Bayesian viewpoint. Nonparametric mixed-effects models are proposed and used for interpretation of the Bayesian structure. Bayes and empirical Bayes estimation are discussed. The latter is shown to have the Gauss-Markov type optimality (i.e., BLUP), to be equivalent to a method of regularization estimator (MORE), and to be minimax in a certain class. Characterization of prior and posterior regularity is discussed. The smoothness of posterior estimators is controlled via prior parameters. Computational issues including the use of generalized cross validation are discussed, and examples are presented.

Key words and phrases: Nonparametric regression, Bayesian regression, Gauss-Markov estimation, best linear unbiased prediction (BLUP), wavelet shrinkage, Besov spaces, Sobolev regularization, generalized cross validation.

1 Introduction

A typical nonparametric regression problem is

$$y_i = f(t_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n, \quad t_i \in [0, 1], \tag{1}$$

where $t_1, \ldots, t_n$ are design points, $\sigma > 0$ is a noise level, $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$ are random errors with zero means and a known positive definite covariance matrix R. The function f is to be recovered based on the observations $Y = (y_1, \ldots, y_n)^T$. Wavelets have been applied to the estimation of f when f is spatially varying. The main purpose of this article is to study the wavelet method from a Bayesian viewpoint. The Bayesian formulation here is conceptually inspired by the work of Parzen (1961), as well as by the work of Kimeldorf

[1]Su-Yun Huang, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C. Email: syhuang@stat.sinica.edu.tw.

[2]Henry Horng-Shing Lu, Institute of Statistics, College of Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 30050, Taiwan. Email: hslu@stat.nctu.edu.tw.

and Wahba (1970, 1971), and Wahba (1978, 1990). The equivalence to regularization and the minimaxity property for Bayesian wavelet shrinkage is influenced by the work of Li (1982). Our Bayesian setup and viewpoints are different from those in Vidakovic (1998a), Clyde, Parmigiani and Vidakovic (1998), Chipman, Kolaczyk and McCulloch (1997), and Abramovich, Sapatinas and Silverman (1998). The setup here has some common structure with spline models and our modelling makes it easy to incorporate prior information on smoothness, to obtain optimality results, and to relate them to the method of regularization. A review of current research involving Bayesian inference in wavelet nonparametric problems is in Vidakovic (1998b).

This article is organized as follows. In Section 2, nonparametric mixed-effects models (NPMEM) are proposed and used for interpretation of Bayesian structure. Characterization of prior regularity is discussed. The relationship between the prior parameters and Besov spaces is explored. In Section 3, the Bayes and empirical Bayes estimators are derived. Some optimality properties are obtained. The posterior space is investigated and is shown to have higher regularity (i.e. smoother) than the prior space. The empirical Bayes estimator derived in Section 3 is a shrinkage estimation. In Section 4, it is shown that such shrinkage estimator is asymptotically equivalent to a diagonal shrinkage wavelet estimator. The methodology developed in Section 3 provides an oracle for guiding the adaptive diagonal shrinkage estimation. When parameter values are not available, adaptive estimations are necessary. In Section 5, we discuss adaptive variants and some computational issues. Examples using generalized cross validation (GCV) for choices of parameters are presented. Concluding remarks are in Section 6. A brief description of the GCV is in Appendix–A. All proofs are in Appendix–B.

2 The prior model–NPMEM

It is assumed that

$$f(t) = \sum_{k=1}^{m} \beta_k \phi_k(t) + \delta\ Z(t),\ t \in [0,1], \tag{2}$$

where $\phi_k(t)$'s are known linearly independent functions, β_k's are fixed but unknown coefficients, and $Z(t)$ is a stochastic process with zero mean and covariance kernel $EZ(s)Z(t) = \mathcal{W}(s,t)$. The covariance kernel is assumed to satisfy the conditions: $\int_0^1 \mathcal{W}(t,t)dt < \infty$ and $\int_0^1 \int_0^1 \mathcal{W}^2(s,t)dsdt < \infty$. The first condition ensures that the sample path of $Z(t)$, and hence $f(t)$, is in $L_2[0,1]$ *a.s.* The second condition ensures that $\mathcal{W}$ has an eigenfunction-eigenvalue decomposition by the Hilbert-Schmidt theorem (Reed and Simon (1972)). That is, there exist functions ψ_ν, $\nu = 1, 2, \ldots$, and a sequence of numbers $\lambda_1 \geq \lambda_2 \ldots \geq 0$ such that $\mathcal{W}(s,t) = \sum_\nu \lambda_\nu \psi_\nu(s)\psi_\nu(t)$. Thus, the process $Z(t)$ has the so-called Karhunen-Loève representation, $Z(t) \sim \sum_\nu \gamma_\nu \psi_\nu(t)$, where '$\sim$' means 'equal in distribution' and γ_ν, $\nu = 1, 2, \ldots$, is a sequence of uncorrelated random variables with zero means and variances $\lambda_1, \lambda_2, \ldots$ These random coefficients, γ_ν, are called random effects. Together with the fixed effects, β_k, the resulting model is a mixed-effects model. It is usually assumed that the closure of linear span by $\{\phi_k, \psi_\nu\}_{k,\nu}$ is the whole space of $L_2[0,1]$. This is the reason for using the terminology 'nonparametric' in front of 'mixed-effects model'.

The prior model (2) is flexible by choice of different bases of ϕ's and ψ's for modelling fixed effects and random effects respectively. Results for a general model are discussed in Huang and Lu (2000). We would like to point out here that, given a process $Z(t)$, one cannot freely choose basis of ψ's to admit the representation $Z(t) \sim \sum_\nu \gamma_\nu \psi_\nu(t)$. Instead, we think of expanding an underlying function f with respect to a choice of basis functions ϕ's and ψ's. Coefficients for ϕ's are modelled as fixed effects, which usually reflect the main features of f. Coefficients for ψ's are modelled as random effects, which usually reflect fine features of f. These random coefficients are assumed uncorrelated. We focus on prior models based on choices of wavelet bases. For references on construction and regularity of functions with random coefficients see Holschneider (1995) and Wornell (1996). See also Cramér and Leadbetter (1967) for sample function (sample path) properties including the Karhunen-Loève representation for stochastic processes.

From now on and throughout this work, a system of wavelets is used as our choice of basis. Using conventional notation, $\{\phi_{j,k}(t)\}$ and $\{\psi_{\ell,k}(t)\}$ denote scaling functions and wavelets on the interval respectively. For the scaling functions $\{\phi_{j,k}(t)\}$, j is a fixed resolution level and k is in a finite index set $\mathcal{I}_j$. Denote the size of $\mathcal{I}_j$ by m, which is of order $O(2^j)$. The wavelets are $\{\psi_{\ell,k}(t)\}$, for each $\ell \geq j$ and k in a finite index set $\mathcal{J}_\ell$, with number of elements of order $O(2^\ell)$. These scaling functions and wavelets are assumed to retain orthonormality. It is also assumed that this system has regularity $r > 0$ and ψ's have $N > r$ vanishing moments. The properties of regularity and vanishing moments ensure that these ϕ's and ψ's form an unconditional basis for Besov spaces $B^s_{p,q}$ for $0 < s < r$ and $1 \leq p, q \leq \infty$. The construction of wavelets on an interval can be found in Cohen, Daubechies, Jawerth and Vial (1993) and Cohen, Daubechies and Vial (1993).

Based on the choice of wavelet basis, the prior model becomes:

$$f(t) = \sum_{k=1}^{m} \beta_k \phi_{j,k}(t) + \delta \; Z(t), \quad Z(t) \sim \sum_{\ell \geq j} \sum_{k \in \mathcal{J}_\ell} \gamma_{\ell,k} \; \psi_{\ell,k}(t), \tag{3}$$

where $\gamma_{\ell,k}$ are uncorrelated random variables with zero means and $E(\gamma^2_{\ell,k}) = \lambda_\ell$.

The definition of Besov space is summarized below. For $0 < b \leq 1$ and $1 \leq p, q \leq \infty$, $J^b_{p,q}$ is defined by

$$J^b_{p,q}(f) = \begin{cases} \left(\int_{h=0}^{1/(1+[b])} h^{-qb-1} \|\Delta_h^{1+[b]} f\|^q_{L_p[0,1-(1+[b])h]} \; dh \right)^{1/q}, & q < \infty, \\ \sup_{h \in [0,1/(1+[b])]} h^{-b} \|\Delta_h^{1+[b]} f\|_{L_p[0,1-(1+[b])h]}, & q = \infty, \end{cases}$$

where Δ^k_h is the kth order difference and $[b]$ is the greatest integer less than or equal to b. When $s > 0$ with $s = a + b$ for a positive integer a and $0 < b \leq 1$, the Besov space $B^s_{p,q}$ is the collection of functions f such that $f, f^{(1)}, \ldots, f^{(a)} \in L_p[0,1]$ and $J^b_{p,q}(f^{(a)}) < \infty$. This space $B^s_{p,q}$ is equipped with the norm $\|f\|_{B^s_{p,q}} = \|f\|_{L_p} + \sum_{k=1}^a J^b_{p,q}(f^{(k)})$.

Expand f in terms of the wavelet basis: $f = \sum_k \beta_{0,k} \; \phi_{0,k} + \sum_{j \geq 0} \sum_k \gamma_{j,k} \; \psi_{j,k}$. Let $\|\beta_{0,\cdot}\|_{\ell_p} = (\sum_k \beta^p_{0,k})^{1/p}$ and $\|\gamma_{j,\cdot}\|_{\ell_p} = (\sum_k \gamma^p_{j,k})^{1/p}$ with the usual modification for $p = \infty$. The norm $\|f\|_{B^s_{p,q}}$ is equivalent to the sequence norm given by

$$\|\beta_{0,\cdot}\|_{\ell_p} + \left\{ \sum_{j \geq 0} \left(2^{j(s+1/2-1/p)} \|\gamma_{j,\cdot}\|_{\ell_p} \right)^q \right\}^{1/q},$$

with the usual modification for $q = \infty$. Note that $B^s_{2,2} = W^s_2$ for $s > 0$. For references, see Bergh and Löfström (1976), Holschneider (1995) and Meyer (1992). In this article, the L_2 norm is used as a standard measurement of function discrepancy unless otherwise specified.

The relation between the prior parameters and the Besov spaces is characterized below. The prior model can be regarded as a nonparametric prior over the space $B^{s-1/2}_{2,\infty}$ with scale coefficients as fixed effects and wavelet coefficients as random effects. In a later section, it is shown that the 'posterior' of $f(t)$ is in $W^s_2 = B^s_{2,2}$, a smoother space than the prior space $B^{s-1/2}_{2,\infty}$. The smoothness regularity of posterior estimators can be controlled via the prior parameters λ_ℓ.

Theorem 2.1. *Suppose that* $\sup_{\ell \geq j} E(\gamma^4_{\ell,k}) < \infty$. *The prior sample path of* $f(t)$ *is in* $B^{s-1/2}_{2,\infty}$ *a.s. if and only if* $\lambda_\ell = O(2^{-2\ell s})$, *where* $s > 1/2$.

Consider the functional class consisting of sample paths of prior model (2) with a common upper bound $\limsup_{\ell\to\infty} 2^{2\ell s}\lambda_\ell \leq C$ imposed upon prior parameters. Let π denote the induced prior probability measure. By the proof for Theorem 2.1 in Appendix B, we see that $\|f\|_{B^{s-1/2}_{2,\infty}} \leq C$ almost surely. That is, $\pi\{\ \|f\|_{B^{s-1/2}_{2,\infty}} \leq C\ \} = 1$. It indicates that we are working over a compact prior functional class.

Corollary 2.2. *Suppose that* $\sup_{\ell \geq j} E(\gamma^4_{\ell,k}) < \infty$. *The prior sample path of* $f(t)$ *is in* $B^{s'-1/2}_{2,q}$ *a.s. for all* $1/2 < s' < s$, $1 \leq q \leq \infty$, *if* $\lambda_\ell = O(2^{-2\ell s})$.

Theorem 2.3. *Assume that* $Z(t)$ *is a Gaussian process. For* $1 \leq p \leq \infty$, *the prior sample path of* $f(t)$ *is in* $B^{s-1/2}_{p,\infty}$ *a.s. if and only if* $\lambda_\ell = O(2^{-2\ell s})$, *where* $s > 1/2$.

Again, with the common upper bound $\limsup_{\ell\to\infty} 2^{2\ell s}\lambda_\ell \leq C$, we have that $\pi\{\ \|f\|_{B^{s-1/2}_{p,\infty}} \leq C\ \} = 1$.

Corollary 2.4. *Assume that* $Z(t)$ *is a Gaussian process. The prior sample path of* $f(t)$ *is in* $B^{s'-1/2}_{p,q}$ *a.s. for all* $1/2 < s' < s$, $1 \leq p, q \leq \infty$, *if* $\lambda_\ell = O(2^{-2\ell s})$.

Corollary 2.2 and Corollary 2.4 can be obtained by embedding theorems: $B^s_{p,q} \subset B^{s^-}_{p,q^+}$ for $s \geq s^- > 0$ and $1 \leq q \leq q^+ \leq \infty$. The space $B^{s-1/2}_{p,\infty}$ is a more precise prior space than $B^{s'-1/2}_{p,q}$ is. Corollary 2.2 and Corollary 2.4 can be compared with Theorems 1 and 2 in Abramovich, Sapatinas and Silverman (1998). Note that Theorem 2.1 and Corollary 2.2 in this article do not require distribution specification, only a few moment conditions are assumed.

3 The BLUP

If ϵ and $Z(t)$ are assumed to be Gaussian processes, the posterior mean of $f(t)$ given the observations Y can be calculated straightforwardly as

$$\begin{aligned} & E(f(t)|Y) = \mu(t) + \delta\ E\{Z(t)\ |Y\} \\ = \ & \mu(t) + (\delta^2/\sigma^2)\ w^T(t)\ M^{-1}(Y-\mu), \end{aligned} \tag{4}$$

where $\mu(t) = \sum_{k=1}^m \beta_k \phi_{j,k}(t)$, $w(t) = [\mathcal{W}(t,t_1), \ldots, \mathcal{W}(t,t_n)]^T$, $M = R + (\delta^2/\sigma^2)W$, $\mu = [\,\mu(t_1), \ldots, \mu(t_n)\,]^T$, and W is an $n \times n$ matrix with the (i,j)-th entry given by $\mathcal{W}(t_i,t_j)$. The posterior mean (4) is the Bayes rule under squared error loss and the Gaussian assumptions. Without the Gaussian assumptions, (4) is linear Bayes in the sense that it minimizes the

squared error loss among all linear rules of the form $a_0(t)+\sum_{i=1}^n a_i(t)\ y_i$, for some functions, $a_i(t),\ i=0,\ 1,\ 2,\ \ldots,\ n$, (Parzen (1961)).

If the coefficients β are not known, one needs to estimate them from the data Y. This is known as the empirical Bayes approach. A generalized least squares estimate is proposed to estimate β here,

$$\hat{\beta} = (X^T M^{-1} X)^{-1} X^T M^{-1} Y, \tag{5}$$

where X is the design matrix for fixed effects. The empirical Bayes estimator becomes

$$\hat{f}(t) = \hat{\mu}(t) + (\delta^2/\sigma^2) w^T(t)\ M^{-1}(Y - \hat{\mu}), \tag{6}$$

where $\hat{\mu}(t) = \sum_{k=1}^m \hat{\beta}_k \phi_{j,k}(t)$ and $\hat{\mu} = [\hat{\mu}(t_1), \ldots, \hat{\mu}(t_n)]^T$. The estimator (6) turns out to be the BLUP as shown in Theorem 3.1 below.

According to the conventional terminology, estimators of random effects are *predictors* and estimators of fixed effects are *estimators*. If there is no ambiguity, estimators or predictors are used without distinction. It is noteworthy that Gaussian assumptions are not required in Theorem 3.1.

Definition. *A predictor $\hat{f}(t)$ is the best linear unbiased predictor (BLUP) of $f(t)$ if and only if (i) $\hat{f}(t)$ is linear in Y, (ii) $\hat{f}(t)$ is unbiased in the sense that $E\hat{f}(t) = Ef(t) = \sum_{k=1}^m \beta_k \phi_{j,k}(t)$ for all $t \in [0,1]$ and all $\beta \in R^m$, (iii) $\hat{f}(t)$ has the minimum mean squared error, among all linear unbiased estimators $\tilde{f}(t)$, i.e., $E(\hat{f}(t) - f(t))^2 \le E(\tilde{f}(t) - f(t))^2$ for all $t \in [0,1]$ and all $\beta \in R^m$.*

Theorem 3.1. *Assume the data model (1) and the prior model (3) with fixed effects $\sum_{k=1}^m \beta_k \phi_{j,k}(t)$ and random-effects covariance*

$$\mathcal{W}(s,t) = \sum_{\ell \ge j} \sum_{k \in \mathcal{J}_\ell} \lambda_\ell \psi_{\ell,k}(s) \psi_{\ell,k}(t).$$

Assume the prior parameters λ_ℓ are known. Then $\hat{f}(t)$ in (6) is the unique BLUP of $f(t)$.

In a parametric mixed-effects model, the model is $Y = X\beta + Z\gamma + \epsilon$, where X and Z are known design matrices, β denotes fixed effects, and γ denotes random effects. The BLUP was investigated by Harville (1976). A recent review was in Robinson (1991). As for nonparametric models, there was pioneering work by Kimeldorf and Wahba ((1970) and (1971)), and Wahba (1978). The BLUP was called either the minimum variance linear unbiased estimation or the Gauss-Markov estimation in those papers. In Section 7 of Kimeldorf and Wahba (1971), they considered the model given by (1) and (2) with random errors $\epsilon \sim N(0, R)$, random coefficients $\beta \sim N(0, I)$, and an independent zero mean Gaussian process $Z(t)$ with a known covariance kernel. Their definition of unbiasedness was conditioned on a fixed β and so is ours. However, their variance was calculated by averaging over values of β according to its distribution $N(0, I)$, while ours is obtained by conditioning on a fixed but arbitrary β. Thus the minimum variance result here is more general. In Wahba (1978), $\beta \sim N(0, \xi I)$ and ξ went to infinity. For fixed σ^2, δ^2, and ξ, let $E_\xi(f(t)|Y)$ denote the posterior mean of $f(t)$ given Y. Under Gaussian assumptions, the posterior mean $E_\xi(f(t)|Y)$ is the BLUP in the sense that both the unbiasedness and the mean squared error are averaged over values of β according to its distribution $N(0, \xi I)$, instead of conditioning on a fixed value of β. By letting $\xi \to \infty$, the resulting estimate $\lim_{\xi \to \infty} E_\xi(f(t)|Y)$ is BLUP at *design* points (Speed 1991). Our results extend to non-design points.

The BLUP $\hat{f}(t)$ is a shrinkage estimator, which shrinks the data toward $\hat{\mu}(t)$, where $\hat{\mu}(t)$ is the generalized least squares fit of data to the low dimensional subspace spanned by $\{\phi_{j,k}(t)\}_{k=1}^{m}$. Let H_0 be the linear subspace of $L_2[0,1]$ spanned by $\{\phi_{j,k}(t)\}_k$ with a fixed resolution level j and k in the corresponding finite index set $\mathcal{I}_j$. Let $H_{\mathcal{W}}$ be the closure (not the L_2 closure, but rather the closure with respect to the norm given below) of the linear space spanned by $\{\psi_{\ell,k}(t)\}_{\ell,k}$ with $\ell \geq j$ and k in the corresponding finite index set $\mathcal{J}_\ell$ for each fixed ℓ. The norm in $H_{\mathcal{W}}$ is defined by $\| \sum_{\ell \geq j} \sum_{k \in \mathcal{J}_\ell} \gamma_{\ell,k}\ \psi_{\ell,k} \|^2_{H_{\mathcal{W}}} = \sum_{\ell \geq j} \sum_{k \in \mathcal{J}_\ell} \gamma^2_{\ell,k}/\lambda_\ell$, where $0/0$ is defined to be zero. The BLUP in (6) is also equivalent to a penalized least squares estimator in (7) of the next theorem.

Theorem 3.2. *The BLUP $\hat{f}(t)$ in (6) can be obtained as the solution to the following minimization problem:*

$$\min_{f \in H_0 \oplus H_{\mathcal{W}}} \ n^{-1}(Y-F)^T R^{-1}(Y-F) + \lambda \|f\|^2_{H_{\mathcal{W}}} \tag{7}$$

with $\lambda = \frac{\sigma^2}{n\delta^2}$ and $F = (f(t_1), \cdots, f(t_n))^T$.

The penalized least squares estimation in (7) is a method of regularization based on a Sobolev norm, $\| \cdot \|_{H_{\mathcal{W}}}$. This is similar in spirit to a Sobolev regularization proposed in Section 5.4 of Amato and Vuza (1997).

The Bayes/linear Bayes estimator in (4) and its empirical Bayes version in (6) are smooth random functions. It is interesting to know their regularity and compare this with the prior regularity given in Section 2. It is referred to the posterior regularity.

Theorem 3.3. *When λ_ℓ is of order $O(2^{-2\ell s})$ as $\ell \to \infty$, the posterior space $H_0 \oplus H_{\mathcal{W}}$ is simply $W_2^s[0,1]$.*

The space H_0 is the prior space as well as the posterior space for fixed effects. The space $H_{\mathcal{W}}$ is the posterior space for random effects. The method of regularization in (7) penalizes random effects. The penalty on random effects is magnified as the resolution level gets finer. Such penalty discourages high frequency wavelet coefficients and diminishes high frequency fluctuation to get smoother posterior curves than the prior curves. Unlike other Bayesian approaches based on some prior distribution assumption, the approach here (BLUP and its equivalent MORE) does not require us to specify any parametric form of distribution, but only the first two moments. Furthermore, one obtains the following theorem as an immediate result of Theorem 2.2 of Li (1982).

Theorem 3.4. *For f in the functional class*

$$W_2^s(\delta) = \left\{ f \in W_2^s[0,1], s > 0, \ \ \textit{and} \ \ \|f\|^2_{H_{\mathcal{W}}} \leq \delta \right\},$$

the estimator $\hat{f}(t)$ in (6), which is the same as the estimator in (7), is the minimax linear estimator of $f(t)$ under mean squared error.

In the literature of nonparametric function estimation, minimaxity and convergence rate results have been obtained for a wide variety of cases. It is known that the minimax rate for estimating a function f at a point $f(t)$ with squared error loss is $O(n^{-2s/(2s+1)})$ for f in a Sobolev ball $W_2^s(\delta)$ and $\hat{f}(t)$ linear in observations (Stone (1980), Ibragimov and Has'minskii (1981)). The restriction to linear estimators for obtaining the minimaxity rate can be relaxed to estimators taking values in a Besov ball of $B^s_{2,\infty}$ using ideas and techniques developed in Kerkyacharian and Picard (1993).

4 Asymptotic equivalence

It is shown below that the BLUP $\hat{f}$ in Theorems 3.1 and 3.2 is asymptotically equivalent to a diagonal shrinkage (DS) estimator.

Theorem 4.1. *If the random variables ϵ are uncorrelated with common variance σ^2, the design points $\{t_i\}_{i=1}^n$ are uniformly distributed, and $\lambda \to 0$, $n\lambda \to \infty$ as $n \to \infty$, then conditional on f,*

$$\hat{f}(t) = \hat{f}_L(t) + \hat{f}_{res,DS}(t) + O\left(\frac{1}{n\lambda}\right) \quad a.s. \text{ for a fixed } t \in (0,1), \tag{8}$$

where

$$\hat{f}_L(t) = \sum_{k=1}^{m} \hat{\beta}_k^\star \phi_{j,k}(t), \;\; \hat{\beta}_k^\star = n^{-1} \sum_{i=1}^{n} \phi_{j,k}(t_i) y_i, \tag{9}$$

$$\hat{f}_{res,DS}(t) = \sum_{\ell \geq j} \sum_{k \in \mathcal{J}_\ell} \frac{\lambda_\ell}{\lambda_\ell + \lambda} \hat{\gamma}_{\ell,k} \; \psi_{\ell,k}(t), \tag{10}$$

with $\hat{\gamma}_{\ell,k} = n^{-1} \sum_{i=1}^{n} \psi_{\ell,k}(t_i) y_i$.

The notation $\hat{\beta}^\star$ is used in the above theorem to distinguish it from its asymptotically equivalent form–the generalized least squares estimator $\hat{\beta}$ in (5). This asymptotic equivalence is useful in computational implementation. The fast algorithm of discrete wavelet transform can be applied to get the estimates $\hat{\beta}^\star$ and $\hat{\gamma}_{\ell,k}$ in linear complexity $O(n)$.

5 Adaptive Variants

When the parameter values for σ, δ and the λ_ℓ are not available, adaptive estimates are necessary. The methodology developed in previous sections provide an oracle for guiding adaptive diagonal shrinkage estimation. We name the estimator (6) the BLUPWAVE. One approach in obtaining the BLUPWAVE is to solve the optimization problem in (7). The other adaptive and computationally economical alternatives are suggested below, based on the asymptotic equivalence in Theorem 4.1 and the result in Proposition 5.1. The resulting adaptive variants are nonlinear.

Proposition 5.1. *Under the same conditions in Theorem 4.1, for fixed k, ℓ we have $E(\hat{\gamma}_{\ell,k}^2) = \delta^2 \lambda_\ell + \sigma^2/n + O(n^{-2})$ as $n \to \infty$.*

Note that $\frac{\lambda_\ell}{\lambda_\ell + \lambda} = \frac{\lambda_\ell}{\lambda_\ell + \sigma^2/(n\delta^2)} = 1 - \frac{\sigma^2/n}{\delta^2 \lambda_\ell + \sigma^2/n}$. By Proposition 5.1, an adaptive version of the BLUPWAVE is

$$\hat{f}_{\text{BLUPWAVE}}(t) = \sum_{k=1}^{m} \hat{\beta}_k^\star \phi_{j,k}(t) + \sum_{\ell \geq j} \sum_{k \in \mathcal{J}_\ell} \left(1 - \frac{\sigma^2/n}{\hat{\gamma}_{\ell,k}^2}\right)_+ \hat{\gamma}_{\ell,k} \; \psi_{\ell,k}(t). \tag{11}$$

The positive sign in (11) is to assure that the shrinkage ratio estimate is nonnegative. Thus, the resulting rule is a thresholding rule that combines the shrinkage and the keep-or-kill

rule. Often σ^2 is not known, we consider a completely data-driven procedure

$$\hat{f}^{\rm GCV}_{\rm BLUPWAVE}(t) = \sum_{k=1}^{m} \hat{\beta}^{\star}_{k}\phi_{j,k}(t) + \sum_{\ell \geq j}\sum_{k\in\mathcal{J}_\ell}\left(1 - \frac{cn^{-1}}{\hat{\gamma}^2_{\ell,k}}\right)_{+} \hat{\gamma}_{\ell,k}\ \psi_{\ell,k}(t) \tag{12}$$

with c selected by GCV. A brief description of the GCV procedure is in Appendix A.

The computational cost of $\hat{f}^{\rm GCV}_{\rm BLUPWAVE}$ is low. First, the fast Mallat's pyramid algorithm can be applied to get the estimates of scale and wavelet coefficients, $\hat{\beta}^{\star}_{k}$ and $\hat{\gamma}_{\ell,k}$, in linear complexity. Then, these estimated coefficients can be applied to obtain the GCV scores and to select c accordingly. With c selected, the fast algorithm can be applied to perform the reconstruction in linear complexity. Four test examples taken from Donoho and Johnstone (1994), Blocks, Bumps, HeaviSine and Doppler, are studied to investigate the finite sample performance. The most nearly symmetric Daubechies wavelets with 8 vanishing moments are used. The periodic wavelet basis over $[0, 1]$ is applied to these four examples. The computation is based on the WaveLab package for MATLAB (Buckheit, Chen, Donoho, Johnstone and Scargle (1995)). The primary resolution scale is $j = 5$ and the remaining fine scales are all included up to the finest possible resolution. The averages and standard errors (in parentheses) of average squared errors (ASEs), $ASE = n^{-1}\sum_{i=1}^{n}\left\{\hat{f}^{\rm GCV}_{\rm BLUPWAVE}(t_i) - f(t_i)\right\}^2$, for 100 replications with various sample sizes and root signal-to-noise ratios (RSNRs) are reported in Table 1. In order to compare with Table 4 in Donoho and Johnstone (1994) and Table 1 in Abramovich, Sapatinas and Silverman (1998), the noise levels are standardized so that $\sigma = 1$ when RSNR=7, as those tables do.

Referring to Table 4 in Donoho and Johnstone (1994), our results are better than the VisuShrink in all four examples and they are also better than those shrinkage estimators with optimal threshold λ^*_n in examples of Blocks, Bumps and Doppler.

When compared with BayesThresh in Table 1 of Abramovich, Sapatinas and Silverman (1998) (wherein $n = 1024$), our results in Table 1 are often better. See Table 2 for comparison. The symbol '+' means that our results are better, blank means about the same and '-' means our results are worse. However, one evident advantage of our estimator is that it is totally data-driven, while their prior parameters α and β are assigned *a priori*.

Displayed in Figure 1 are reconstructions for these four examples with cases which are close to the averages of ASEs in Table 1 and whose RSNR=7. The noises are effectively removed and the heights are preserved in the reconstruction.

6 Concluding Discussion

Starting with nonparametric mixed-effects models, we use Bayesian and empirical Bayesian formulation in discussing the prior and posterior spaces. Also with the help of Bayesian and empirical Bayesian techniques, we obtain the shrinkage estimation in (6). The estimation method (6) is shown to have the Gauss-Markov type optimality, to be equivalent to a regularization method, and to be linear minimax for a certain class. The estimator in (6) is not practical for computation. Thus its asymptotic equivalent is derived in Theorem 4.1. Adaptive variant by GCV along with a simulation study are discussed in Section 5. This method is simple and computationally economical.

When preparing this article, the authors found a related technical report by Huang and Cressie (2000). Independently, they discussed the deterministic/stochastic wavelet decomposition, which is the same setup as the nonparametric mixed-effects model in this article. Under Gaussian assumptions they also derived the Bayes estimate. They used a different technique based on the normal probability plot for the empirical estimate of the Bayes estimate. Their simulation results also confirmed the advantage of Bayesian estimates. To sum up, either from the perspective of nonparametric mixed-effects models or from that of deterministic/stochastic wavelet decomposition, the Gauss-Markov type predictor and its adaptive variants are useful for the recovery of signals.

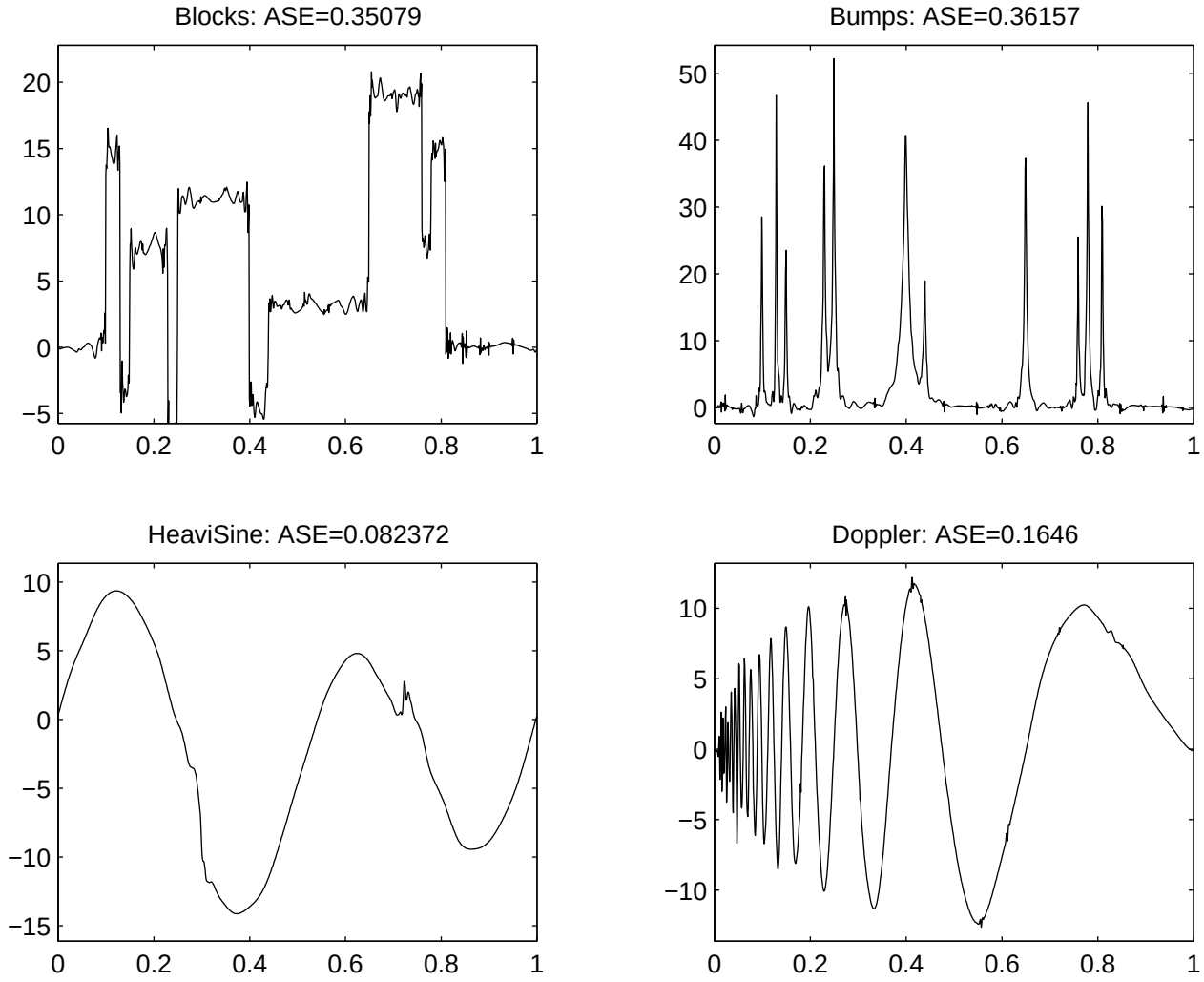
Table 1. Averages
and standard errors (in parentheses) of ASEs from 100 replications of BLUPWAVE by GCV.

RSNR	n	**Blocks**	**Bumps**	**HeaviSine**	**Doppler**
10	256	0.3388 (0.00670)	0.3830 (0.00577)	0.1346 (0.00279)	0.2036 (0.00307)
	512	0.2451 (0.00329)	0.2666 (0.00315)	0.0906 (0.00179)	0.1233 (0.00182)
	1024	0.1878 (0.00180)	0.1868 (0.00178)	0.0524 (0.00088)	0.0834 (0.00112)
	2048	0.1191 (0.00093)	0.1270 (0.00086)	0.0356 (0.00047)	0.0510 (0.00071)
	4096	0.0772 (0.00047)	0.0783 (0.00048)	0.0218 (0.00035)	0.0261 (0.00034)
	8192	0.0480 (0.00028)	0.0466 (0.00029)	0.0136 (0.00016)	0.0148 (0.00016)
7	256	0.6903 (0.01254)	0.7200 (0.01168)	0.2396 (0.00442)	0.3995 (0.00617)
	512	0.4831 (0.00683)	0.5285 (0.00675)	0.1669 (0.00307)	0.2478 (0.00371)
	1024	0.3533 (0.00341)	0.3514 (0.00326)	0.0904 (0.00145)	0.1643 (0.00213)
	2048	0.2302 (0.00201)	0.2443 (0.00183)	0.0613 (0.00099)	0.0987 (0.00127)
	4096	0.1470 (0.00096)	0.1547 (0.00123)	0.0391 (0.00058)	0.0539 (0.00076)
	8192	0.0929 (0.00053)	0.0893 (0.00055)	0.0257 (0.00032)	0.0289 (0.00030)
5	256	1.2492 (0.02050)	1.3348 (0.02275)	0.3895 (0.00703)	0.7776 (0.01293)
	512	0.9058 (0.01159)	0.9913 (0.01239)	0.2586 (0.00365)	0.4737 (0.00787)
	1024	0.6423 (0.00645)	0.6434 (0.00595)	0.1442 (0.00271)	0.3051 (0.00444)
	2048	0.4203 (0.00365)	0.4369 (0.00370)	0.0963 (0.00175)	0.1793 (0.00222)
	4096	0.2760 (0.00208)	0.2824 (0.00230)	0.0642 (0.00098)	0.1036 (0.00134)
	8192	0.1750 (0.00104)	0.1648 (0.00118)	0.0428 (0.00055)	0.0544 (0.00057)
3	256	3.2561 (0.06112)	3.4741 (0.06931)	0.8724 (0.02073)	2.0865 (0.03432)
	512	2.2178 (0.03368)	2.5254 (0.03287)	0.5115 (0.01328)	1.2395 (0.02177)
	1024	1.6077 (0.01664)	1.6455 (0.01610)	0.3137 (0.00693)	0.7253 (0.01096)
	2048	0.9989 (0.00976)	1.0635 (0.00849)	0.2066 (0.00433)	0.4329 (0.00573)
	4096	0.7126 (0.00491)	0.6844 (0.00525)	0.1410 (0.00282)	0.2609 (0.00351)
	8192	0.4467 (0.00274)	0.4127 (0.00259)	0.0885 (0.00134)	0.1462 (0.00189)

Table 2. Comparison with BayesThresh

RSNR	n	**Blocks**	**Bumps**	**HeaviSine**	**Doppler**
10	1024	+	+	+	+
7	1024	+	+	+	
5	1024	+	+		
3	1024	-	+	-	-

Figure 1: The reconstructions of BLUPWAVE and ASEs by the GCV method when $n = 1024$, RSNR=7, and σ is unknown.



7 Appendix A. Generalized Cross Validation

This appendix gives a brief description of a data-driven selection method via GCV for the threshold parameter c in the BLUPWAVE. The CV and GCV methods for hard and soft wavelet thresholding are studied by Weyrich and Warhola (1995), Jansen, Malfait and Bultheel (1997). Parallel to their ideas, we derive a GCV method in our context. Suppose $n = 2^{J+1}$. Let $y = (y_1, \ldots, y_n)^t$ and $w = (w_1, \ldots, w_n)^t = \sqrt{n}\ (\hat{\beta}^*_1, \ldots, \hat{\beta}^*_{2^j}, \hat{\gamma}_{j,1}, \ldots, \hat{\gamma}_{J,2^J})^t$, the empirical wavelet coefficients. The empirical wavelet coefficients w, by discrete wavelet transform, can be represented by an orthogonal matrix W. That is, $w = Wy$.

Let $\hat{y}$ be the estimates of $\hat{f}^{\rm GCV}_{\rm BLUPWAVE}$ evaluated at design points $t_i = i/n,\ i = 1, \ldots, n$. Then $\hat{y}$ can be represented as $\hat{y} = W^{-1}D_c\ w = W^{-1}D_cWy$, where D_c is an $n \times n$ diagonal matrix $D_c = \text{diag}\{d_{ii}\}$, $i = 1, \ldots, n$, with

$$d_{ii} \quad = \quad \begin{cases} 1, & \text{for } \ i = 1, \ldots, 2^j, \\ 0, & \text{for } \ i = 2^j + 1, \ldots, n \ \text{ and } \ w_i^2 \le c, \\ 1 - \frac{c}{w_i^2}, & \text{for } \ i = 2^j + 1, \ldots, n \ \text{ and } \ w_i^2 > c. \end{cases} \tag{13}$$

Note that the elements of D_c depend on the signal w, thus D_c is a nonlinear mapping. Define the influence matrix by $A = W^{-1}D_cW$ and the differential influence matrix by $A^{dif}_{i,j} = \partial\hat{y}_i/\partial y_j$ entrywise. It can be shown that the differential matrix A^{dif} is given by $A^{dif} = W^{-1}D^{dif}W$, where $D^{dif}_{i,j} = \partial(D_cw)_i/\partial w_j$. We then have $tr(A^{dif}) = tr(D^{dif})$. Let $w_c = D_cw$. The generalized cross validation score is

$$\begin{aligned} GCV(c) \quad &= \quad \frac{n^{-1}\|y - \hat{y}\|^2}{\left[n^{-1}tr(I - A^{dif})\right]^2} = \frac{n^{-1}\|w - w_c\|^2}{\left[n^{-1}(trI - trA^{dif})\right]^2} \\ &= \quad \frac{n\sum_{i=2^j+1}^{n}(1 - d_{ii})^2 w_i^2}{\left[\sum_{i=2^j+1, d_{ii}=0}^{n}\ 1 - \sum_{i=2^j+1, d_{ii}\neq 0}^{n}(c/w_i^2)\right]^2} \end{aligned}$$

The computation of a GCV curve is very fast because it is based on the wavelet coefficients already generated in the construction of BLUPWAVE. No more forward and inverse discrete wavelet transforms are needed for GCV curves.

8 Appendix B. Proofs

Proof of Theorem 2.1: Recall that $Z(t) \sim \sum_{\ell \ge j}\sum_{k\in\mathcal{J}_\ell}\gamma_{\ell,k}\ \psi_{\ell,k}(t)$, where $\gamma_{\ell,k}$'s are uncorrelated random variables with zero means and $E\gamma^2_{\ell,k} = \lambda_\ell$. The sample path of $Z(t)$ is in $B^{s-1/2}_{2,\infty}$ *a.s.* if and only if the wavelet coefficients satisfy the constraint

$$\sup_{\ell \ge j} 2^{2\ell(s-1/2)} \sum_{k\in\mathcal{J}_\ell}\gamma^2_{\ell,k} < \infty \ \ a.s.$$

The number of elements in $\mathcal{J}_\ell$ is of order $O(2^\ell)$. By the Strong Law of Large Numbers for uncorrelated random variables whose second moments have a common upper bound (Theorem 5.1.2. in Chung (1974)), $\limsup_{\ell\to\infty} 2^{2\ell s}\frac{\sum_{k\in\mathcal{J}_\ell}\gamma^2_{\ell,k}}{2^\ell} = \limsup_{\ell\to\infty} 2^{2\ell s}\lambda_\ell \quad a.s.$

Thus, $\sup_{\ell \geq j} 2^{2\ell(s-1/2)} \sum_{k \in \mathcal{J}_\ell} \gamma_{\ell,k}^2 < \infty$ *a.s.* if and only if $\sup_{\ell \geq j} 2^{2\ell s} \lambda_\ell < \infty$. Therefore, the sample path of $Z(t)$, as well as the sample path of $f(t)$, are in the space $B_{2,\infty}^{s-1/2}$ *a.s.* if and only if λ_ℓ is of order $O(2^{-2\ell s})$. Q.E.D.

Proof of Theorem 2.3: THE CASE $1 \leq p < \infty$. The sample path of $Z(t)$ is in $B_{p,\infty}^{s-1/2}$ *a.s.* if and only if the wavelet coefficients satisfy the condition $\sup_{\ell \geq j} 2^{p\ell(s-1/p)} \sum_{k \in \mathcal{J}_\ell} |\gamma_{\ell,k}|^p < \infty$ *a.s.* By the Strong Law of Large Numbers,

$$\limsup_{\ell \to \infty} 2^{p\ell s} \frac{\sum_{k \in \mathcal{J}_\ell} |\gamma_{\ell,k}|^p}{2^\ell} = \limsup_{\ell \to \infty} 2^{p\ell s} \lambda_\ell^{p/2} \quad a.s.$$

Thus, $\sup_{\ell \geq j} 2^{p\ell(s-1/p)} \sum_{k \in \mathcal{J}_\ell} \gamma_{\ell,k}^p < \infty$ *a.s.* if and only if $\sup_{\ell \geq j} 2^{p\ell s} \lambda_\ell^{p/2} < \infty$. Therefore, the sample path of $Z(t)$ (and the sample path of $f(t)$) is in the space $B_{p,\infty}^{s-1/2}$ *a.s.* if and only if λ_ℓ is of order $O(2^{-2\ell s})$.

THE CASE $p = \infty$. The sample path of $Z(t)$ is in $B_{\infty,\infty}^{s-1/2}$ *a.s.* if and only if the wavelet coefficients satisfy the condition $\limsup_\ell \left(2^{\ell s} \max_{k \in \mathcal{J}_\ell} |\gamma_{\ell,k}|\right) < \infty$ *a.s.* That is, if and only if, for any $K > 0$,

$$P\{2^{\ell s} \max_{k \in \mathcal{J}_\ell} |\gamma_{\ell,k}| > K \ \ i.o.\} = 0. \tag{14}$$

By the Borel-Cantelli Lemma, (14) holds if $\sum_{\ell=j}^\infty P\{2^{\ell s} \max_{k \in \mathcal{J}_\ell} |\gamma_{\ell,k}| > K\}$ is finite. For every fixed resolution level ℓ and constant $K > 0$, let $A_k^{(\ell,K)}$ denote the set where $|\gamma_{\ell,k}| > 2^{-\ell s} K$, but $|\gamma_{\ell,k'}| \leq 2^{-\ell s} K$ for $k' < k$, where $k', k \in \mathcal{J}_\ell$. We have

$$\begin{aligned}
& \sum_{\ell=j}^\infty P\{2^{\ell s} \max_{k \in \mathcal{J}_\ell} |\gamma_{\ell,k}| > K\} = \sum_{\ell=j}^\infty \sum_{k \in \mathcal{J}_\ell} P\{A_k^{(\ell,K)}\} = \sum_{\ell=j}^\infty \sum_{k \in \mathcal{J}_\ell} \int_{A_k^{(\ell,K)}} dP \\
\leq & \sum_{\ell=j}^\infty 2^{2\ell s} K^{-2} \sum_{k \in \mathcal{J}_\ell} \int_{A_k^{(\ell,K)}} |\gamma_{\ell,k}|^2 dP = \sum_{\ell=j}^\infty 2^{2\ell s} K^{-2} \int_{\cup_{k \in \mathcal{J}_\ell} A_k^{(\ell,K)}} |\gamma_{\ell,1}|^2 dP \\
\leq & \sum_{\ell=j}^\infty 2^{2\ell s} K^{-2} E|\gamma_{\ell,1}|^2 = \sum_{\ell=j}^\infty 2^{2\ell s} \lambda_\ell / K^2,
\end{aligned}$$

which is finite if and only if λ_ℓ is of order $O(2^{-2\ell s})$. Q.E.D.

Proof of Theorem 3.1: (i) It is clear that $\hat{f}(t)$ is linear in Y. (ii) By the unbiasedness of the generalized least squares estimate, $E\hat{\mu}(t) = \mu(t)$ and $E(Y - \hat{\mu}) = 0$. Thus, $E\hat{f}(t) = Ef(t) = \sum_{k=1}^m \beta_k \phi_{j,k}(t)$ holds for all $t \in [0,1]$ and $\beta \in R^m$. (iii) Let L_t be an n-vector depending on t and let $L_t^T Y$ be an arbitrary linear unbiased estimate of $f(t)$. A necessary and sufficient condition for $L_t^T Y$ being unbiased is $L_t^T X = \Phi^T(t)$, where $\Phi(t) = (\phi_{j,1}(t), \phi_{j,2}(t), \ldots, \phi_{j,m}(t))^T$. The BLUP for $f(t)$ is the solution $L_t Y$ to the minimization problem:

$$\min_{L_t} E(L_t^T Y - f(t))^2 \ \text{ for } \ L_t \in \mathcal{I}_t \equiv \{L_t \in R^n : L_t^T X = \Phi^T(t)\}.$$

The following lemma can be proved straightforwardly and the proof is omitted.

Lemma A. Let A be an $n \times n$ positive definite matrix, X be an $n \times m$ $(m < n)$ full rank matrix, L be an n-vector and u be an m-vector. The unique solution to the minimization problem $\min_L L^T A L$, $L \in \{L \in R^n : L^T X = u^T\}$ is given by $L = A^{-1} X (X^T A^{-1} X)^{-1} u$.

Let $Z_n = (Z(t_1), \ldots, Z(t_n))^T$, $L_{t,\star} = L_t - (\delta^2/\sigma^2)M^{-1}w(t)$ and $\Phi_{j,\star}(t) = \Phi(t) - (\delta^2/\sigma^2)X^T M^{-1}w(t)$. Then one has

$$\begin{aligned} & \min_{L_t \in \mathcal{I}_t} E(L_t^T Y - f(t))^2 = \min_{L_t \in \mathcal{I}_t} E(L_t^T(\delta\ Z_n + \sigma\epsilon) - \delta\ Z(t))^2 \\ = & \min_{L_t \in \mathcal{I}_t}\ \sigma^2 L_t^T M L_t - 2\delta^2 L_t^T w(t) + \delta^2 \mathcal{W}(t,t) \\ = & \min_{L_t \in \mathcal{I}_t}\ \sigma^2 L_{t,\star}^T M L_{t,\star} + \text{some constant}. \end{aligned}$$

It is observed that $L_t^T X = \Phi^T(t)$ holds if and only if $L_{t,\star}^T X = \Phi_{j,\star}^T(t)$ holds. Therefore, $\min_{L_t \in \mathcal{I}_t} L_{t,\star}^T M L_{t,\star} = \min_{L_{t,\star} \in \mathcal{I}_{t,\star}} L_{t,\star}^T M L_{t,\star}$, where $\mathcal{I}_{t,\star} \equiv \{L_{t,\star} \in R^n : L_{t,\star}^T X = \Phi_{j,\star}^T(t)\}$. By Lemma A, the unique solution is given by $L_{t,\star} = M^{-1}X(X^T M^{-1}X)^{-1}\ \Phi_{j,\star}(t)$ or, equivalently, by $L_t = (\delta^2/\sigma^2)M^{-1}w(t) + M^{-1}X(X^T M^{-1}X)^{-1}\ \Phi_{j,\star}(t)$. It is then straightforward to check that $L_t^T Y = \hat{\mu}(t) + (\delta^2/\sigma^2)w^T(t)M^{-1}(Y - \hat{\mu})$. Q.E.D.

Proof of Theorem 3.2: We establish this theorem by showing that it is a special case of Theorem 3.1 of Huang and Lu (2000) with $\lambda = \alpha/n = \sigma^2/(n\delta^2)$. The regression model (1) together with the prior model (3) in this paper is a special case of the model given by (1.1) and (1.2) in Huang and Lu (2000). To apply their Theorem 3.1, we have to check the following assumptions.

Identifiability condition. The scaling functions $\{\phi_{j,k}\}_{k=1}^m$ used in this paper are linearly independent, so the identifiability condition is met.

Boundedness condition I. Evaluation on a reproducing kernel Hilbert space is a bounded linear functional and this condition is met.

Boundedness condition II. For $g \in H_0 \oplus H_{\mathcal{W}}$, define $L(g) = n^{-1}Y^T R^{-1}G$, where $G = (g(t_1), \ldots, g(t_n))^T$. For an n-vector v, let $\|v\|_2^2 = n^{-1}\sum_{j=1}^n v_j^2$. Then

$$\begin{aligned} \|L\| &= \sup_{g \in H_0 \oplus H_{\mathcal{W}}} \frac{|Y^T R^{-1}G|/n}{\|g\|_{H_0 \oplus H_{\mathcal{W}}}} \le \sup_{g \in H_0 \oplus H_{\mathcal{W}}} \frac{\|R^{-1}Y\|_2 \cdot \|G\|_2}{\|g\|_{H_0 \oplus H_{\mathcal{W}}}} \\ &\le \|R^{-1}Y\|_2 \sup_{g \in H_0 \oplus H_{\mathcal{W}}} \frac{\left(n^{-1}\sum_{j=1}^n g^2(t_j)\right)^{1/2}}{\|g\|_{H_0 \oplus H_{\mathcal{W}}}}. \end{aligned}$$

For $g \in H_0 \oplus H_{\mathcal{W}}$, let $l_{t_j}(g) = g(t_j)$ denote evaluation at t_j. As $H_0 \oplus H_{\mathcal{W}}$ is a reproducing kernel Hilbert space, we have $\|l_{t_j}\| = \sup_{g \in H_0 \oplus H_{\mathcal{W}}}(|g(t_j)|/\|g\|_{H_0 \oplus H_{\mathcal{W}}}) < \infty$. Thus

$$\sup_{g \in H_0 \oplus H_{\mathcal{W}}} \frac{\left(n^{-1}\sum_{j=1}^n g^2(t_j)\right)^{1/2}}{\|g\|_{H_0 \oplus H_{\mathcal{W}}}} \le \left(n^{-1}\sum_{j=1}^n \|l_{t_j}\|^2\right)^{1/2}.$$

Therefore, $\|L\| < \infty$ for an arbitrary realization of Y. Q.E.D.

Proof of Theorem 3.3: By the characterization theorem of Sobolev spaces (Mallat (1989) or Meyer (1992)) and the definition of $H_{\mathcal{W}}$-norm, we have that the posterior space $H_0 \oplus H_{\mathcal{W}}$ is simply the Sobolev space $W_2^s[0,1]$, when $\lambda_\ell = O(2^{-2\ell s})$. Q.E.D.

In what follows, '$\sum_{\ell,k}$' means '$\sum_{\ell \ge j}\sum_{k \in \mathcal{J}_\ell}$'. The symbol $\mathbf{O}_{n\times m}(\delta_n)$ stands for a sequence of $n \times m$ (m can be a fixed or an n-dependent index) matrices whose (i,j)-th

entry is uniformly of order $O(\delta_n)$, as $n \to \infty$. When there is no ambiguity, the subscript $n \times m$ is suppressed. Let $W_{\ell,k}$ be the $n \times n$ matrix with its (i,j)-th entry given by $\mathcal{W}_{\ell,k}(t_i,t_j) = \psi_{\ell,k}(t_i)\psi_{\ell,k}(t_j)$. Then the W matrix defined in Section 3 has the decomposition $W = \sum_{\ell,k} \lambda_\ell W_{\ell,k}$. Let

$$\mathcal{W}^{[p]}(s,t) = \int_{[0,1]^{p-1}} \mathcal{W}(s,u_1)\mathcal{W}(u_1,u_2)\cdots\mathcal{W}(u_{p-1},t)\ du_1\ du_2\cdots du_{p-1},$$

then we have the decomposition $\mathcal{W}^{[p]}(s,t) = \sum_{\ell,k} \lambda_\ell^p \psi_{\ell,k}(s)\psi_{\ell,k}(t)$. Let $W^{[p]}$ be the $n \times n$ matrix with its (i,j)-th entry given by $\mathcal{W}^{[p]}(t_i,t_j)$, then the matrix $W^{[p]}$ has the decomposition $W^{[p]} = \sum_{\ell,k} \lambda_\ell^p W_{\ell,k}$. We remind the reader that the symbol $\mathcal{W}$ stands for kernel function and the symbol W for matrix.

The following lemma is needed for Theorem 4.1.

Lemma B.
(i) $n^{-p}X^TW^p = \mathbf{O}(n^{-p})$,
(ii) $M^{-1} = \left\{I + \sum_{p=1}^{\infty}(-1)^p \left(\frac{W}{n\lambda}\right)^p\right\} = \left\{I - \sum_{\ell,k} \frac{\lambda_\ell W_{\ell,k}}{n(\lambda_\ell+\lambda)}\right\} + \mathbf{O}\left(\frac{1}{n^2\lambda^2}\right)$, and
(iii) $X^TM^{-1}(I - X(X^TX)^{-1}X^T) = \mathbf{O}\left(\frac{1}{n\lambda}\right)$.

Proof of Lemma B: (i) The (i,j)-th entry of a matrix A is $A(i,j)$. For a fixed entry, say the (i,j)-th entry, and for integer $p \geq 1$, we have

$$\begin{aligned} n^{-p}(X^TW^p)(i,j) &= n^{-p}\sum_{d_1=1}^{n}\cdots\sum_{d_p=1}^{n} X^T(i,d_1)W(d_1,d_2)\cdots W(d_p,j) \\ &= \int_{[0,1]^p} \phi_i(u_1)\mathcal{W}(u_1,u_2)\cdots\mathcal{W}(u_p,t_j)\ du_1\ du_2\cdots du_p + O(n^{-p}) \\ &= O(n^{-p}), \ \text{ uniformly in } i \text{ and } j. \end{aligned}$$

That is, for integer $p \geq 1$, $n^{-p}X^TW^p = \mathbf{O}(n^{-p})$.

(ii) Recall that $M = I + (\delta^2/\sigma^2)W$. For a fixed (i,j)-th entry, we have

$$\begin{aligned} n^{-p+1}W^p(i,j) &= n^{-p+1}\sum_{d_1=1}^{n}\cdots\sum_{d_{p-1}=1}^{n} W(i,d_1)W(d_1,d_2)\cdots W(d_{p-1},j) \\ &= \int_{[0,1]^{p-1}} \mathcal{W}(t_i,u_1)\mathcal{W}(u_1,u_2)\cdots\mathcal{W}(u_{p-1},t_j)\ du_1\ du_2\cdots du_{p-1} + O(n^{-p+1}) \\ &= W^{[p]}(i,j) + O(n^{-p+1}) \ \text{ uniformly in } i \text{ and } j. \end{aligned}$$

Therefore,

$$\begin{aligned} M^{-1} = \{I + (n\lambda)^{-1}W\}^{-1} &= \left\{I + \sum_{p=1}^{\infty}(-n\lambda)^{-p}W^p\right\} \\ &= I - \frac{W}{\lambda\ n} + \sum_{p=2}^{\infty}\frac{W^{[p]} + \mathbf{O}\left(n^{-p+1}\right)}{(-\lambda)^p\ n} \\ &= I - \frac{\sum_{\ell,k}\lambda_\ell W_{\ell,k}}{\lambda\ n} + \sum_{p=2}^{\infty}\frac{\sum_{\ell,k}\lambda_\ell^p W_{\ell,k} + \mathbf{O}\left(n^{-p+1}\right)}{(-\lambda)^p\ n} \end{aligned}$$

$$
\begin{aligned}
&= I - \frac{\sum_{\ell,k} \lambda_\ell W_{\ell,k}}{\lambda\ n} + \sum_{\ell,k} n^{-1} W_{\ell,k} \left\{ \sum_{p=2}^{\infty} (-\lambda_\ell/\lambda)^p \right\} + \mathbf{O}\left(n^{-2}\lambda^{-2}\right) \\
&= I - \frac{\sum_{\ell,k} \lambda_\ell W_{\ell,k}}{\lambda\ n} + \sum_{\ell,k} n^{-1} W_{\ell,k} \left\{ \frac{\lambda_\ell^2}{\lambda(\lambda+\lambda_\ell)} \right\} + \mathbf{O}\left(n^{-2}\lambda^{-2}\right) \\
&= I - \sum_{\ell,k} \frac{\lambda_\ell W_{\ell,k}}{n(\lambda_\ell+\lambda)} + \mathbf{O}\left(\frac{1}{n^2\lambda^2}\right).
\end{aligned}
$$

(Here we implicitly assume that $\lambda_\ell/\lambda < 1$, which is eventually true as $n \to \infty$. The finitely many $\lambda_\ell/\lambda \geq 1$ will not affect the result of Theorem 4.1.)

(iii)Observe that

$$
\begin{aligned}
& X^T M^{-1}(I - X(X^TX)^{-1}X^T) = X^T\{I + \sum_{p=1}^{\infty}(-n\lambda)^{-p}W^p\}(I - X(X^TX)^{-1}X^T) \\
= \ & X^T\{\sum_{p=1}^{\infty}(-n\lambda)^{-p}W^p\}(I - X(X^TX)^{-1}X^T) = X^T\{\sum_{p=1}^{\infty}(-n\lambda)^{-p}W^p\} = \mathbf{O}\left(\frac{1}{n\lambda}\right),
\end{aligned}
$$

as $n^{-p}X^TW^p = \mathbf{O}(n^{-p})$.

Proof of Theorem 4.1: First we show that

$$
\hat{\mu}(t) = \hat{f}_L(t) + O\left(\frac{1}{n\lambda}\right) \quad a.s. \tag{15}
$$

One has

$$
\begin{aligned}
& \hat{\mu}(t) = \Phi^T(t)(X^TM^{-1}X)^{-1}X^TM^{-1}Y \\
= \ & \Phi^T(t)(X^TM^{-1}X)^{-1}X^TM^{-1}X(X^TX)^{-1}X^TY \\
& +\Phi^T(t)(X^TM^{-1}X)^{-1}X^TM^{-1}\{I - X(X^TX)^{-1}X^T\}Y \\
= \ & \Phi^T(t)(X^TX)^{-1}X^TY + \Phi^T(t)(X^TM^{-1}X)^{-1}\mathbf{O}\left((n\lambda)^{-1}\right)Y.
\end{aligned}
$$

As $(n^{-1}X^TX)^{-1} = I + \mathbf{O}(n^{-1})$, $\Phi^T(t)(X^TX)^{-1}X^TY = \hat{f}_L(t) + O\left(n^{-1}\right)$ $a.s.$ Also note that

$$
\begin{aligned}
n^{-1}X^TM^{-1}X &= n^{-1}X^TX + \sum_{p=1}^{\infty}\frac{(-1)^pX^TW^pX}{n^{p+1}\lambda^p} \\
&= I + \mathbf{O}\left(\frac{1}{n}\right) + \mathbf{O}\left(\frac{1}{n\lambda}\right) = I + \mathbf{O}\left(\frac{1}{n\lambda}\right);
\end{aligned} \tag{16}
$$

then

$$
\Phi^T(t)(X^TM^{-1}X)^{-1}\mathbf{O}\left((n\lambda)^{-1}\right)Y = \mathbf{O}\left(\frac{1}{n\lambda}\right). \tag{17}
$$

By (16) and (17), we have (15).

Next, we show that $(\delta^2/\sigma^2)w^T(t)M^{-1}(Y - X\hat{\beta}) = \hat{f}_{res,DS}(t) + O(1/n\lambda)$ *a.s.* Let $w_{\ell,k}(t) = [\mathcal{W}_{\ell,k}(t,t_1),\ldots,\mathcal{W}_{\ell,k}(t,t_n)]^T$. By Lemma B (ii),

$$(\delta^2/\sigma^2)w^T(t)M^{-1}(Y - X\hat{\beta}) = (\delta^2/\sigma^2)w^T(t)\left\{I + \sum_{p=1}^{\infty}(-n\lambda)^{-p}W^p\right\}(Y - X\hat{\beta})$$

$$= \frac{1}{n\lambda}\sum_{\ell,k}\lambda_\ell w_{\ell,k}^T(t)\left\{I - \sum_{\ell,k}\frac{\lambda_\ell W_{\ell,k}}{n(\lambda_\ell+\lambda)} + \mathbf{O}\left(\frac{1}{n^2\lambda^2}\right)\right\}(Y - X\hat{\beta})$$

$$= \frac{1}{n\lambda}\left\{\sum_{\ell,k}\lambda_\ell w_{\ell,k}^T(t) - \sum_{\ell,k}\frac{\lambda_\ell^2 w_{\ell,k}^T(t)W_{\ell,k}}{n(\lambda_\ell+\lambda)} + \mathbf{O}\left(\frac{1}{n^2\lambda^2}\right)\right\}(Y - X\hat{\beta})$$

$$= \frac{1}{n\lambda}\left\{\sum_{\ell,k}\lambda_\ell w_{\ell,k}^T(t) - \sum_{\ell,k}\frac{\lambda_\ell^2 w_{\ell,k}^T(t)}{(\lambda_\ell+\lambda)} + \mathbf{O}\left(\frac{1}{n\lambda}\right) + \mathbf{O}\left(\frac{1}{n^2\lambda^2}\right)\right\}(Y - X\hat{\beta})$$

$$= \frac{1}{n\lambda}\left\{\sum_{\ell,k}\frac{\lambda_\ell\lambda\ w_{\ell,k}^T(t)}{(\lambda_\ell+\lambda)} + \mathbf{O}\left(\frac{1}{n\lambda}\right)\right\}(Y - X\hat{\beta}).$$

Notice that an explicit expression for $\mathbf{O}\left(\frac{1}{n\lambda}\right)$ is

$$\sum_{\ell,k}\frac{\lambda_\ell^2 w_{\ell,k}^T(t)}{(\lambda_\ell+\lambda)} - \sum_{\ell,k}\frac{\lambda_\ell^2 w_{\ell,k}^T(t)W_{\ell,k}}{n(\lambda_\ell+\lambda)}$$

and we have the order

$$\frac{1}{n\lambda}\left(\sum_{\ell,k}\frac{\lambda_\ell^2 w_{\ell,k}^T(t)}{(\lambda_\ell+\lambda)} - \sum_{\ell,k}\frac{\lambda_\ell^2 w_{\ell,k}^T(t)W_{\ell,k}}{n(\lambda_\ell+\lambda)}\right)(Y - X\hat{\beta}) = \mathbf{O}\left(\frac{1}{n\lambda}\right).$$

Therefore,

$$(\delta^2/\sigma^2)w^T(t)M^{-1}(Y - X\hat{\beta}) = \sum_{\ell,k}\frac{\lambda_\ell w_{\ell,k}^T(t)(Y - X\hat{\beta})}{n(\lambda_\ell+\lambda)} + O\left(\frac{1}{n\lambda}\right) \quad a.s.$$

$$= \sum_{\ell,k}\frac{\lambda_\ell w_{\ell,k}^T(t)Y}{n(\lambda_\ell+\lambda)} + O\left(\frac{1}{n\lambda}\right) \quad a.s. = \hat{f}_{res,DS}(t) + O\left(\frac{1}{n\lambda}\right) \quad a.s.$$

Q.E.D.

Proof of Proposition 5.1: One has

$$\left|n^{-2}\sum_{i,j=1}^{n}\mathcal{W}_{\ell,k}(t_i,t_j)\mu(t_i)\mu(t_j)\right| = O(n^{-2}),$$

$$n^{-2}\delta^2\sum_{i,j=1}^{n}\mathcal{W}_{\ell,k}(t_i,t_j)\mathcal{W}_{\ell',k'}(t_i,t_j)$$

$$= \int_0^1 \int_0^1 \delta^2 \mathcal{W}_{\ell,k}(s,t)\mathcal{W}_{\ell',k'}(s,t)\ ds\ dt + O(n^{-2}) = O(n^{-2}),\ \ (\ell,k) \neq (\ell',k')$$

$$n^{-2}\delta^2 \sum_{i,j=1}^{n} \mathcal{W}^2_{\ell,k}(t_i,t_j) = \int_0^1 \int_0^1 \delta^2 \mathcal{W}^2_{\ell,k}(s,t)\ ds\ dt + O(n^{-2}) = \delta^2 + O(n^{-2}),$$

$$n^{-2}\sigma^2 \sum_{i=1}^{n} \mathcal{W}_{\ell,k}(t_i,t_i) = n^{-1}\sigma^2 \left(\int_0^1 \mathcal{W}_{\ell,k}(t,t)dt + O(n^{-1}) \right) = n^{-1}\sigma^2 + O(n^{-2}).$$

Therefore, for fixed (ℓ,k),

$$E\ \hat{\gamma}^2_{\ell,k} = n^{-2} \sum_{i,j=1}^{n} \mathcal{W}_{\ell,k}(t_i,t_j) E\{y_i\ y_j\}$$

$$= n^{-2} \left\{ \sum_{i,j=1}^{n} \mathcal{W}_{\ell,k}(t_i,t_j) \Big[\mu(t_i)\mu(t_j) + \delta^2 EZ(t_i)Z(t_j) + \sigma^2 E\epsilon_i\epsilon_j \Big] \right\}$$

$$= O(n^{-2}) + n^{-2}\delta^2 \sum_{(\ell',k')\neq(\ell,k)} \sum_{i,j=1}^{n} \lambda_{\ell'} \mathcal{W}_{\ell,k}(t_i,t_j)\mathcal{W}_{\ell',k'}(t_i,t_j)$$

$$+ n^{-2}\delta^2 \sum_{i,j=1}^{n} \lambda_\ell \mathcal{W}^2_{\ell,k}(t_i,t_j) + n^{-2}\sigma^2 \sum_{i=1}^{n} \mathcal{W}_{\ell,k}(t_i,t_i) = \delta^2\lambda_\ell + \sigma^2/n + O(n^{-2}).$$

Q.E.D.

Acknowledgments. The authors would like to thank Y.-C. Tung, H.-Y. Chang and F.-J. Lin for their help in simulation studies. The authors also thank the Associate Editor and two referees for valuable comments. This research was partially supported by the National Science Council, Taiwan, R.O.C.

References

1. Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc.*, B 60, 725–749.
2. Amato, U. and Vuza, D. T. (1997). Besov regularization, thresholding and wavelets for smoothing data. Technical Report, Instituto per Applicazioni della Matematica, CNR, Italy.
3. Bergh, J. and Löfström, J. (1976). *Interpolation Spaces: An Introduction.* Springer-Verlag, New York.
4. Buckheit, J., Chen, S., Donoho, D. L., Johnstone, I. M. and Scargle, J. (1995). About WaveLab. Technical Report, Department of Statistics, Stanford University.
5. Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* 92, 1413–1421.
6. Chung, K. L. (1974). *A Course in Probability Theory.* (2nd Edition). Academic Press, Boston.
7. Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* 85, 391–401.
8. Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1993). Multiresolution analysis, wavelets and fast algorithms on an interval. *C. R. Acad. Sci. Paris series I* 316, 417–421.
9. Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comp. Harmonic Anal.* 1, 54–81.

10. Cramér, H. and Leadbetter, M. R. (1967). *Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications.* John Wiley, New York.
11. Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425–455.
12. Harville, D. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Ann. Statist.* 4, 384–395.
13. Holschneider, M. (1995). *Wavelets: An Analysis Tool.* Oxford Science Publications, Oxford.
14. Huang, H.-C., and Cressie, N. (2000). Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics*, to appear.
15. Huang, S.Y. and Lu, H. H.-S. (2000). Extended Gauss-Markov theorem for nonparametric mixed-effects models. *J. Multi. Anal.*, to appear.
16. Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation Asymptotic Theory.* Springer-Verlag, New York.
17. Jansen, M., Malfait, M. and Bultheel, A. (1997). Generalized cross validation for wavelet thresholding. *Signal Processing* 56, 33–44.
18. Kerkyacharian, G. and Picard, D. (1993). Density estimation by kernel and wavelets methods: optimality of Besov spaces. *Statist. Probab. Lett.* 18, 327–336.
19. Kimeldorf, G. and Wahba, G. (1970). A correspondence between Bayesian estimation in stochastic processes and smoothing by splines. *Ann. Math. Statist.* 41, 495–502.
20. Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Analysis Appl.* 33, 82–95.
21. Li, K.-C. (1982). Minimaxity of the method of regularization on stochastic processes. *Ann. Statist.* 10, 937–942.
22. Mallat, S. G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Trans. Amer. Math. Soc.* 315, 69–87.
23. Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.
24. Parzen, E. (1961). An approach to time series analysis. *Ann. Math. Statist.* 32, 951–989.
25. Reed, M. and Simon, B. (1972). *Methods of Modern Mathematical Physics: I. Functional Analysis.* Academic Press, Boston.
26. Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statist. Sci.* 6, 15–51.
27. Speed, T. (1991). Comment on "That BLUP is a good thing: the estimation of random effects." by Robinson, G.K. (1991). *Statist. Sci.* 6, 15–51.
28. Stone, C. J. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* 8, 1348–1360.
29. Vidakovic, B. (1998a). Non-linear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.* 93, 173–179.
30. Vidakovic, B. (1998b). Wavelet-based nonparametric Bayes methods. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds. D. Dey, P. Müller and D. Sinha), 133–155. Springer-Verlag, New York.
31. Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc.* B 40, 364–372.
32. Wahba, G. (1990). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia.
33. Weyrich, N. and Warhola, G.T. (1995). De-noising using wavelets and cross validation. In *Approximation Theory, Wavelets and Applications.* (S.P. Singh ed.), 523–532. Kluwer, New York.
34. Wornell, G. W. (1996). *Signal Processing with Fractals: A Wavelet Based Approach.* Prentice Hall, New Jersey.