Empirical Bayes approaches to mixture problems and wavelet regression

Iain M. Johnstone and Bernard W. Silverman*

Stanford University and University of Bristol

Abstract

We consider model selection in a hierarchical Bayes formulation of the sparse normal linear model in which individual variables have, independently, an unknown prior probability of being included in the model. The focus is on orthogonal designs, which are of particular importance in nonparametric regression via wavelet shrinkage. Empirical Bayes estimates of hyperparameters are easily obtained via the EM algorithm, and this approach is contrasted with a recent conditional likelihood proposal. Our model selection approach yields a straightforward method for data dependent threshold selection in wavelet regression. Performance on standard test sets and data examples is encouraging, especially if a translation invariant form of the estimator is used. Since the method produces separate threshold estimates on each wavelet resolution level, it also comfortably handles stationary correlated error structures.

1 Introduction

Consider the nonparametric regression problem where we have observations at 2^m regularly spaced points t_i of some unknown function f subject to noise:

$$X_i = f(t_i) + \epsilon_i \tag{1}$$

where the ϵ_i are independent $N(0, \sigma^2)$ random variables. The standard wavelet-based approaches to the estimation of f proceed by taking the discrete wavelet transform of the data X_i, processing the resulting coefficients to remove noise, and then transforming back to obtain the estimate.

1.1 Bayesian approaches to wavelet regression

The underlying notion behind wavelet methods is that the unknown function has an economical wavelet expression, in that f is, or is well approximated by, a function with a relatively small proportion of nonzero wavelet coefficients. This notion is naturally expressed by a Bayesian model for the wavelet coefficients of f. Write d_{jk} for the elements of the discrete wavelet transform (DWT) of the vector of values $f(t_i)$ and $\hat{d}_{jk}$ for the DWT of the observed data X_i.

*This work was partly carried out while BWS was a Fellow at the Center for Advanced Study in the Behavioral Sciences supported by NSF grant SBR-9601236. IMJ was supported in part by NIH grant R01 CA72028

Clyde et al. (1998) and Abramovich et al. (1998) have considered a particular mixture prior for this problem. Under this prior, the d_{jk} are independently distributed with

$$d_{jk} \sim (1 - \pi_j)\delta(0) + \pi_j N(0, \tau_j^2), \tag{2}$$

a mixture of an atom of probability at zero and a normal distribution with variance τ_j^2. The parameters of the distribution (2) depend on the level j of the coefficient in the transform.

Abramovich et al. (1998) investigated the use of the posterior median of d_{jk} as a summary of the posterior distribution. This is a true thresholding rule, in that for $|\hat{d}_{jk}|$ less than some threshold, the point estimate of d_{jk} will be exactly zero. In the wavelet context, the coefficient-wise posterior median corresponds to a point estimate of the posterior distribution under a family of loss function equivalent to L^1 norms on the function and its derivatives. Such L^1 losses are in any case more natural if one wishes to allow for the possibility of inhomogeneous functions, one of the aims of the wavelet approach.

A related prior was considered by Chipman et al. (1997); for a full survey of work in this area, see Vidakovic (1998).

1.2 Specifying the prior parameters

How should the parameters in the prior be chosen? In the existing literature, the parameters are either chosen directly by reference to prior information about f, or by a combination of prior information and data-based criteria. For example, the BayesThresh approach of Abramovich et al. (1998) assumes that

$$\tau_j^2 = C_1 2^{-\alpha j}, \qquad \pi_j = \min(1, C_2 2^{\beta} j).$$

The hyperparameters α and β are then specified by consideration of the likely smoothness properties of the function, or by using a standard choice of values, for example $\alpha = 0.5$ and $\beta = 1$. The parameters C_1 and C_2 are then chosen by arguments related to the method of moments; see their paper for further details. Though BayesThresh gives good results, it clearly invites the possibility of a more systematic approach to the choice of the hyperparameters.

In the present paper, we take an empirical Bayes (or marginal maximum likelihood) approach, in order to get a completely data-based method of choosing the prior parameters. Within the Bayesian formulation set out above, wavelet regression at a single resolution level j is a special case of a Bayesian model selection problem considered by George & Foster (1997) among others. In the simplest case, this problem is as follows.

Suppose that $Y = (Y_1, \ldots, Y_n)$ are observations satisfying

$$Y_i = \theta_i + \epsilon_i \tag{3}$$

where the ϵ_i are independent $N(0, \sigma^2)$ random variables. It is supposed that the unknown coefficients θ_i are mostly zero, but some of them may be nonzero, and with this in mind it is of interest to estimate the θ_i on the basis of the observed data. In the model selection context, the nonzero θ_i correspond to parameters that actually enter the model. The connection with wavelet regression is natural: the Y_i might be the sample wavelet coefficients at a particular level, and these are noisy observations of a sequence of population wavelet coefficients θ_i which are mostly zero.

1.3 Summary of the paper

In Section 2 the sparse Gaussian mixture model will be reviewed. We develop an empirical Bayes approach based on a metric inequality, equivalent to an EM algorithm. In Section 3 the criterion proposed by George & Foster (1997) will be explored in more detail. The criterion is based on an approximation to the marginal likelihood, and we shall see that it may lead to biased results in the large sample context, which is particularly relevant to wavelet regression. We then return to wavelet regression in Section 4, reporting simulation results showing that the method performs well. As with most wavelet methods, an adaptation of the method to the translation-invariant wavelet transform provides considerable improvement, and this is explored in Section 5. In both these sections the performance of the method on a data set collected in an anesthesiological study is explored. In Section 6 the extension of the method to the case of correlated data is considered. Excellent results are obtained both for simulated data and for a data set constructed in a neurophysiological context. Section 7 concludes with remarks on data-based thresholding.

2 Bayesian model selection

2.1 The Bayesian model

Suppose we have data of the form (3). The variance σ^2 will be assumed known, and until further notice, we assume $\sigma^2 = 1$, without loss of generality. To formalize the notion that most of the θ_i are zero, George & Foster (1997) considered a hierarchical Bayes model for the situation where only an unknown subset of terms enter a model. Let γ_i be a sequence of independent Bernoulli(w) random variables, such that each γ_i is 1 if the corresponding θ_i actually enters the model. Define

$$\begin{aligned} \theta_i &= 0 \text{ if } \gamma_i = 0 \\ \theta_i &\sim N(0, C) \text{ if } \gamma_i = 1. \end{aligned} \tag{4}$$

The marginal prior distribution of θ_i is a mixture distribution precisely of the same form as (2):

$$\theta_i \sim (1-w)\delta(0) + wN(0, C) \tag{5}$$

Given the hyperparameters w and C, the posterior distribution of θ_i given Y_i is a mixture of an atom of probability at zero and a general normal distribution. If a point estimate $\hat{\theta}_i$ is required, most authors consider the posterior mean, but as we have already pointed out, the posterior median has desirable properties, particularly in the wavelet context.

Consider the likelihood of the parameters w and C given the data Y_i. Since the data will have a mixture distribution of the form $(1-w)N(0,1) + wN(0, 1+C)$ the likelihood will be of the form

$$L(C, w|Y) \propto \prod_{i=1}^{n} [(1-w)\exp(-\tfrac{1}{2}Y_i^2) + w(1+C)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}Y_i^2/(1+C)\}]. \tag{6}$$

Rather than maximize this mixture likelihood directly, George & Foster (1997) suggested an approximation which could be maximized quickly. It was study of this approximation which initially stimulated our interest, and we return to it in Section 3 below.

2.2 Using a metric inequality

Here we return to direct maximization of the likelihood L to obtain marginal maximum likelihood (MML) estimates of w and C. The iteration we describe is an instance of the EM algorithm (Dempster et al. (1977)) but the present mixture setting permits an alternative derivation, which we present briefly, together with remarks about the way the steps are actually carried out.

We bring in the binary entropy function

$$H(\xi) = -\xi \log \xi - (1-\xi)\log(1-\xi) \tag{7}$$

which has conjugate

$$\log(1+e^{-x}) = \sup_{0\le\xi\le 1} \{H(\xi) - \xi x\}, \tag{8}$$

the maximum being attained at $\xi = 1/(1+\exp x)$. Up to a constant depending only on the data, the log of the likelihood (6) can therefore be written as

$$l(C,w) = n\log(1-w) + \sum_{i=1}^{n} \log[1+\exp\{-h(C,w,Y_i)\}] \tag{9}$$

where

$$h(C,w,Y) = -\operatorname{logit} w + \tfrac{1}{2}\log(1+C) - \tfrac{1}{2}CY^2/(1+C),$$

and $\operatorname{logit} w = \log w - \log(1-w)$. Define

$$l^\dagger(C,w,\xi_1,\ldots,\xi_n) = n\log(1-w) + \sum_{i=1}^{n}[H(\xi_i) - \xi_i h(C,w,Y_i)].$$

It follows from (8) that if we maximize $l^\dagger$ over all its arguments, we will have found the maximum of the original full likelihood function.

A strategy that works well is to optimize alternately over the ξ_i and over (C,w). In the next section we show that this yields an EM algorithm. Let $\phi_{\sigma^2}(y)$ denote the Gaussian density with mean zero and variance σ^2. For given C and w, simple calculus shows that the global maximum over ξ_i in $[0,1]$ is obtained by setting

$$\xi_i = \frac{w\phi_{1+C}(Y_i)}{(1-w)\phi_1(Y_i) + w\phi_{1+C}(Y_i)}, \tag{10}$$

and that for fixed ξ the global maximum over w in $[0,1]$ and C in $[0,\infty)$ is obtained by setting

$$w = n^{-1}\sum_{i=1}^{n}\xi_i, \qquad \text{and} \qquad C = \left(\frac{\sum_{i=1}^n \xi_i Y_i^2}{\sum_{i=1}^n \xi_i} - 1\right)_+ . \tag{11}$$

We can be sure that these are global maxima for the following reasons: For fixed C and w, $l^\dagger$ is clearly a sum of concave functions of the individual ξ_i each of whose unique maxima is given by solving the equation $H'(\xi_i) = h(C,w,Y_i)$, which leads to (10) after some manipulation. For fixed ξ_i, the function $l^\dagger$ can be rewritten as $Q+R$, with $R(\xi,Y)$ not involving C and w, and

$$Q = (n - \textstyle\sum \xi_i)\log(1-w) + \sum \xi_i \log w - \tfrac{1}{2}\left\{\sum \xi_i \log(1+C) + \frac{\sum \xi_i Y_i^2}{1+C}\right\}.$$

Table 1: Results of applying the MML criterion to samples of various sizes with true values $w = 0.1$ and $C = 10$. One hundred replications were used for each sample size.

Sample size	$\hat{w}$ Mean	SD	$\hat{C}$ Mean	SD
50	0.168	0.134	9.5	10.8
100	0.122	0.085	11.2	7.9
200	0.114	0.046	9.0	4.7
500	0.106	0.029	9.9	2.8
1000	0.105	0.019	9.8	2
2000	0.100	0.014	10.1	1.4

This is a sum of a concave function of w, with maximum given in (11), and a function of C that is not concave; however we have

$$\frac{\partial l^\dagger}{\partial C} = \tfrac{1}{2}(1+C)^{-2}\{\sum \xi_i Y_i^2 - (1+C)\sum \xi_i\}, \tag{12}$$

the product of a strictly positive quantity and a linearly decreasing function of C. It follows at once from (12) that the global maximum of $l^\dagger$ over C is given in (11).

2.3 Connections with the EM algorithm

Here is the connection with the EM algorithm. If we regard as missing data the variables γ_i that take the value zero if $Y_i \sim N(0,1)$ and one if $Y_i \sim N(0, 1+C)$, then the ξ_i as given by equation (10) are the posterior expected values of γ_i given the data and the current values of the hyperparameters w and C. The function Q is just the expected complete data log-likelihood for (C, w) given the data and given the previous iterate's estimate of (C, w) as encoded in the posterior expected values ξ_i.

Either directly, or from standard EM convergence results (see e.g. McLachlan & Krishnan (1997, Section 3.4.4)), it follows that successive iterations increase the objective function $l^\dagger$, or Q, and that all the limit points of the iterations on (C, w) are stationary points. No difficulties with convergence to local maxima were encountered in practice. We have found that good results are obtained by initializing $\xi_i = 0.99$ if $|Y_i| > 2.5$ and $\xi_i = 0.01$ otherwise, and then beginning the iteration by applying (11).

The classical theory of maximum likelihood demonstrates that the MML approach is consistent, and this can be verified by a simple asymptotic calculation, showing that if we replace the sums in (11) by their expected values, then the true values of w and C yield a stationary point of the iteration. To check how the asymptotic values are approached in finite samples, a small simulation study was carried out. For a number of sample sizes n, one hundred replications were carried out in which n model points θ_i and sample points Y_i were generated from the model with these true parameter values, and the estimates obtained by the MML method were found. In Table 1 we present the results of the MML approach, implemented using this choice of initial conditions and then iterating the alternating maximization algorithm to convergence. The consistency of the approach is borne out by the simulations, though there is evidence of mild upward bias in w in relatively small samples.

3 The conditional maximum likelihood approach

Rather than maximize the likelihood (6) directly, George & Foster (1997) suggest the use of an approximation, replacing the sum in each of the product terms by the larger of the two summands:

$$L^*(C, w|Y) = \prod_{i=1}^{n} \max[(1-w)\exp(-\tfrac{1}{2}Y_i^2), w(1+C)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}Y_i^2/(1+C)\}]. \qquad (13)$$

This is the likelihood obtained by considering the maximum likelihood choice of the sequence γ_i and then conditioning on these values. George and Foster show that it may be maximized very quickly in practice. Following Clyde & George (1998), who develop the idea further in the wavelet context, we term it the *conditional maximum likelihood* (CML) approach. In this section, we study the CML approach in detail, and show that it may lead to inconsistent estimators.

To investigate its properties, define

$$\tau(C, w) = \{(1+C)/C\}\{\log(1+C) - 2\,\mathrm{logit}\, w\}. \qquad (14)$$

We can then rewrite (13), up to a factor involving the data y only, as

$$L^*(C, w|Y) \propto (1-w)^n \prod_{i=1}^{n} \max\left\{1, \exp[\tfrac{1}{2}C\{Y_i^2 - \tau(C,w)\}/(1+C)]\right\}. \qquad (15)$$

The maximum in (15) will be greater than 1 if and only if $Y_i^2 > \tau(C, w)$. Taking the logarithm and multiplying by 2, and neglecting the constant of proportionality depending only on the data, the maximization of (15) is therefore equivalent to that of maximizing

$$l^*(C, w|Y) = 2n\log(1-w) + 2\sum \max\{0, \tfrac{1}{2}C\{Y_i^2 - \tau(C,w)\}/(1+C)\}$$

In order to investigate the large sample behavior of the CML approach, suppose that the θ are drawn from a mixture distribution (4) with true parameters w_0 and C_0 and that $Y = \theta + \epsilon$ where ϵ has a standard normal distribution independently of θ. For any positive threshold τ,

$$P(Y^2 > \tau) = 2(1-w_0)\tilde{\Phi}(\tau^{1/2}) + 2w_0\tilde{\Phi}\{\tau^{1/2}/(1+C_0)\} \qquad (16)$$

$$EY^2 I[Y^2 > \tau] = 2(1-w_0)\tilde{\Phi}_2(\tau^{1/2}) + 2w_0(1+C_0)\tilde{\Phi}_2\{\tau^{1/2}/(1+C_0)\} \qquad (17)$$

where $\tilde{\Phi} = 1 - \Phi$, and $\tilde{\Phi}_2(t) = t\phi(t) + \tilde{\Phi}(t)$. We denote (16) and (17) by $p_i(\tau|C_0, w_0)$ for $i = 1, 2$ respectively.

We proceed as in the classical analysis of maximum likelihood estimation. For large n, for each C and w the quantity $n^{-1}l^*(C, w)$ will converge to its expectation $\lambda^*(C, w|C_0, w_0)$, given by

$$\lambda^*(C, w|C_0, w_0) = 2\log(1-w) + \{2\,\mathrm{logit}\, w - \log(1+C)\}p_1\{\tau(C,w)|C_0, w_0\} + \{C/(1+C)\}p_2\{\tau(C,w)|C_0, w_0\} \qquad (18)$$

where $\tau(C, w)$ is given by (14) above. Assuming sufficient regularity to allow the maximization over C and w to be interchanged with the limit over n, the large sample values of the estimates of C and w will be the maximizers C^* and w^* of $\lambda^*(C, w)$. These cannot

Table 2: Results of applying the CML criterion to a samples of various sizes with true values $w = 0.1$ and $C = 10$, the same samples as in Table 1. One hundred replications were used for each sample size. The number of replications for which the algorithm broke down and gave the estimate $\hat{w} = 1$ is shown. The means and standard deviations are only worked out from those realizations where the estimate of w is not equal to 1.

Sample size	number with $\hat{w} = 1$	$\hat{w}$ Mean	SD	$\hat{C}$ Mean	SD
50	44	0.043	0.029	32.1	16.3
100	20	0.037	0.02	29.7	11.7
200	12	0.033	0.015	26.1	8.7
500	1	0.035	0.011	25.9	4.9
1000	0	0.035	0.006	25.3	3.3
2000	0	0.034	0.005	25.7	2.2

be found analytically, but are easily found by a numerical maximization in any particular case.

In the corresponding calculation for maximum likelihood, the large sample values of the parameter estimates can be shown to be equal to the true parameter values. However, in the case of the function $\lambda^*(C, w|C_0, w_0)$ this will not in general be the case, and C^* and w^* can be quite different from C_0 and w_0. For example, if $w_0 = 0.1$ and $C_0 = 10$, then the large sample values of the estimates are $w^* = 0.0344$ and $C^* = 25.5$.

We return to the simulation study of Table 1. Using the same simulated data replications, the CML algorithm yields results given in Table 2. For relatively large sample sizes, the estimates clearly cluster around their asymptotic values; when the sample size is small, there is some chance that the method chooses the value $w = 1$, which corresponds to a model in which the marginal distribution of the observations is a single normal distribution. It is clear from these results that, apart from the cases where the CML method breaks down completely and yields the value $w = 1$, the performance is in line with what the asymptotic theory would predict; note that the variances are approximately inversely proportional to sample size.

4 Application to wavelet regression

4.1 The MML paradigm in the wavelet context

As already explained in Sections 1.1 and 1.2, the model selection approach can be applied to wavelet regression. The MML method we have set out is used in this context by regarding the wavelet coefficients as being modeled by a mixture model of the kind investigated above, with values of w and C that depend on the level of the transform. The MML algorithm is therefore applied to each level of the transform separately. Assuming that the original noise is independent, the variance σ^2 can, as is conventional, be estimated from the median of the squared coefficients at the highest level. Given the values of w and C level by level, we can then estimate the wavelet coefficients by a Bayesian approach; following the BayesThresh procedure of Section 1.2, we use the posterior median which yields estimates most of whose wavelet coefficients are exactly zero.

Table 3: Mean square error of the MML procedure as compared to the BayesThresh method, for each of the test functions of Donoho & Johnstone (1994) sampled at 1024 points, with various values of root signal to noise. The simulations are based on 100 replications, with standard errors given in brackets.

	RSNR	Blocks	Bumps	Heavisine	Doppler
MML	10	0.190 (0.002)	0.238 (0.002)	0.052 (0.001)	0.084 (0.001)
	7	0.340 (0.003)	0.437 (0.003)	0.088 (0.001)	0.161 (0.002)
	5	0.607 (0.007)	0.733 (0.007)	0.135 (0.002)	0.297 (0.004)
	3	1.529 (0.017)	1.755 (0.018)	0.271 (0.005)	0.698 (0.010)
BayesThresh	10	0.214 (0.002)	0.249 (0.002)	0.061 (0.001)	0.085 (0.001)
	7	0.376 (0.004)	0.440 (0.003)	0.102 (0.001)	0.160 (0.002)
	5	0.680 (0.008)	0.732 (0.006)	0.147 (0.003)	0.298 (0.004)
	3	1.694 (0.018)	1.732 (0.018)	0.300 (0.006)	0.681 (0.010)

4.2 A simulation comparison

In our first simulation comparison, we use as a benchmark the BayesThresh procedure with the standard choice of parameters $\alpha = 0.5$, $\beta = 1$. A detailed, and favorable, comparison between that method and several other methods is given by Abramovich et al. (1998). The simulation consisted of 100 replications at each of four noise levels for each of the four standard test functions defined by Donoho & Johnstone (1994), with $2^m = 1024$. The results are shown in Table 3. The noise level is specified in terms of root signal-to-noise ratio (RSNR), the ratio of the standard deviation of the function values to the standard deviation σ of the noise.

The basic message of the table is that the MML method performs better than BayesThresh in the case of the blocks and heavisine functions, and the two methods are equally good for the bumps and doppler functions. Although the improvements are not vast (typically 10% of the mean square error) they do provide support for the use of the MML approach, which is in any case avoids some of the *ad hoc* features of the choice of prior in the BayesThresh procedure. For the very noisy case where the RSNR is 3, the MML method performs marginally worse for the bumps and doppler functions, but there are still relatively larger improvements for the other two functions. Some intuition as to the reason for this behavior can be obtained by examining the hyperparameter estimates level by level for the two methods. Inspection of a few particular cases suggests, as one might expect, that the parameters as estimated by the two methods are less discrepant for the bumps and Doppler functions, especially at the middle levels of the transform.

4.3 The inductance plethysmography data

The method was also applied to the inductance plethysmography data described in Abramovich et al. (1998), and shown in Figure 1. In that paper it was found that adjusting the first BayesThresh parameter from the standard choice $\alpha = 0.5$ to $\alpha = 2$ gave somewhat better results in terms of allowing through the highest peak values while eliminating high frequency noise. The height of the first peak value yielded by the MML method is 0.8446, identical to 4dp to that yielded by the BayesThresh method with $\alpha = 2$, and the treatment of the high frequency noise is very similar. In terms of summed absolute difference, the MML curve is about four times as far from the BayesThresh ($\alpha = 0.5$) curve as from the

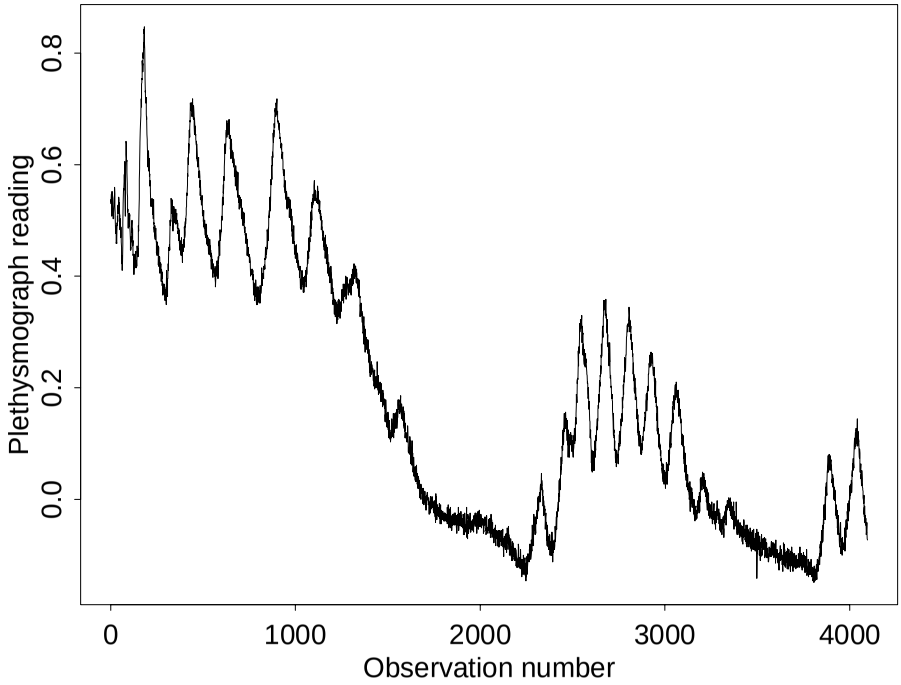

Figure 1: The inductance plethysmography data

BayesThresh ($\alpha = 2$) curve.

In contrast, when the MML method is compared to the two BayesThresh methods for a data set drawn from the Donoho-Johnstone blocks function with RSNR equal to 7, the situation is reversed: the MML function is closer to the BayesThresh ($\alpha = 0.5$) estimate by a factor of over 5. Not surprisingly, the MML approach adapts to the smoothness of the unknown function, as expressed in terms of the relative properties of its wavelet expansion at different levels, in a way that the BayesThresh method is not able to do. Further details of the results are given in the next section, where another method will be introduced into the comparison.

5 Using the translation-invariant transform

5.1 An as-if-independent MML approach

It is by now well recognized that the translation-invariant wavelet transform (Coifman & Donoho 1995) in general gives much better results than the conventional transform applied with a fixed origin. In order to extend our method to the translation-invariant case, we proceed as follows. At each level, the translation-invariant transform gives a sequence of 2^m values that are not actually independent. Nevertheless, we construct an as-if-independent likelihood function for the mixture hyperparameters at each level, and use the alternating maximization procedure to give estimates of these parameters. The estimates are then used to give individual posterior medians of each of the coefficients of the stationary transform; a reconstruction of the function is then found by the average basis approach. We can relate this procedure to the one described in Section 4. Once the hyperparameters have

Table 4: Mean square errors of the translation-invariant marginal maximum likelihood procedure, for each of the test functions of Donoho & Johnstone (1994) sampled at 1024 points, with various values of root signal to noise. The simulations are based on the same 100 replications as in Table 3, with standard errors given in brackets.

RSNR	Blocks	Bumps	Heavisine	Doppler
10	0.106 (0.0011)	0.138 (0.0012)	0.034 (0.0005)	0.051 (0.0007)
7	0.204 (0.0023)	0.256 (0.0020)	0.064 (0.0011)	0.093 (0.0012)
5	0.400 (0.0051)	0.439 (0.0039)	0.107(0.0019)	0.165 (0.0024)
3	1.094 (0.013)	1.067 (0.011)	0.221 (0.004)	0.427 (0.007)

been estimated, the translation-invariant method gives the result of applying the standard method at every possible choice of time origin, and then averaging over the position of the time origin.

Notice that the prior model for the standard wavelet expansion will not be exactly consistent between different positions of the origin. If a random function has independent wavelet coefficients with mixture distribution (4) for one position of the origin, then its wavelet decomposition with respect to another choice will not have this form at all levels, in general. (The distribution *will* be same at those levels where the shift between the two origins is a multiple of the scale, because the wavelet coefficients at those levels will be the same for both decompositions.) The prior distributions cannot be generated from a single underlying prior model for the curve, so the translation-invariant procedure should, strictly speaking, be seen as a separate modeling of the prior information at each position of the origin.

Using an as-if-independent likelihood at each level to choose the hyperparameters is reminiscent of the *independence estimating equation* approach of Liang & Zeger (1986) to parameter fitting in the marginal distribution of a sequence of identically distributed but non-independent observations. Their paper was concerned with observations with generalized linear model dependence on the parameters and covariates. The use of independence estimating equations in our context is a method of combining the separate problems of choosing the prior into a single problem at each level, by maximizing the average of the log likelihoods over the different positions of the origin.

A simulation study was carried out using exactly the same random realizations as for the simulation above. The results are shown in Table 4. The improvements in mean integrated square error over the fixed basis MML method are substantial, typically around 40%.

5.2 Illustration on the plethysmography data

To continue the comparison based on the plethysmography data, see Figures 2 to 5. In Figure 2 the translation-invariant method of this section (termed TI-MML in the figure) is compared with the adjusted BayesThresh($\alpha = 2$) and the MML method. The differences between the methods are very difficult to see, and all three methods are good at extracting even sharp features of the data while eliminating noise. Figure 3 compares the methods in the region of the first peak. The MML and adjusted BayesThresh methods are virtually identical, while the TI-MML method removes a small amount of very local variation in the region of the peak. Figures 4 and 5 consider the region near time 3500. The large fluctuation is present in all three curves, but is smaller in the TI-MML curve; examination of the section of the original data plotted in Figure 5 shows a single observation at time 3498 that is a

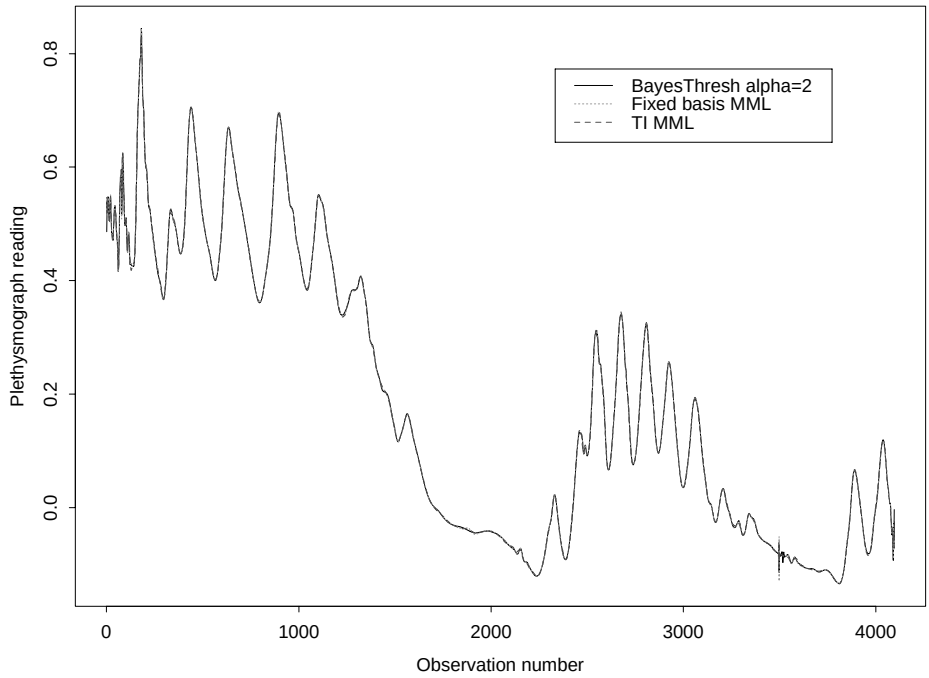

Figure 2: Three methods of wavelet smoothing applied to the inductance plethysmography data

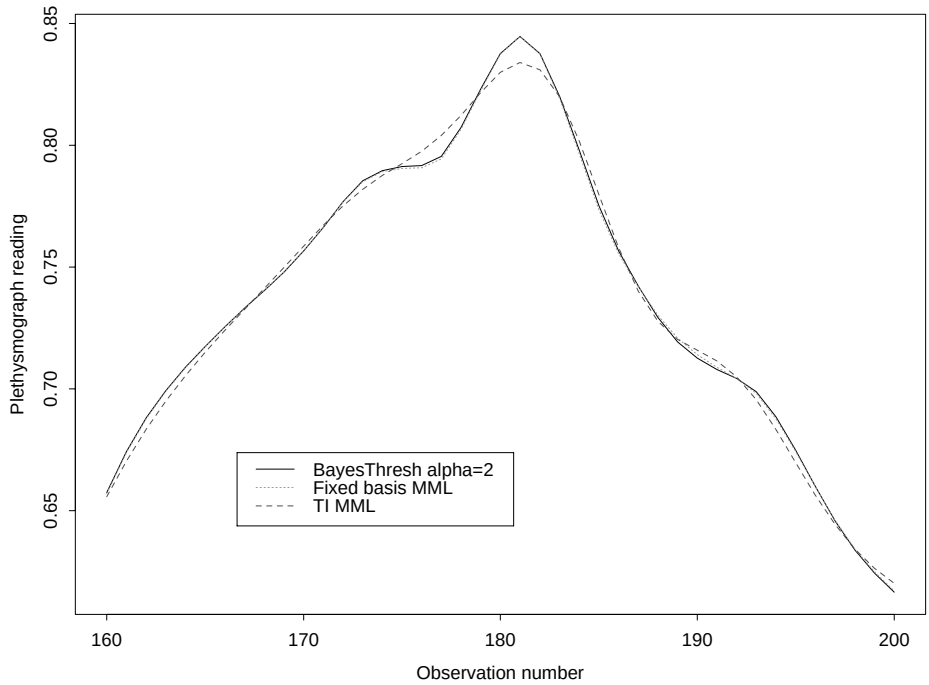

Figure 3: The comparison of Figure 2 for a short sequence of the plethysmography data, in the region of the first peak

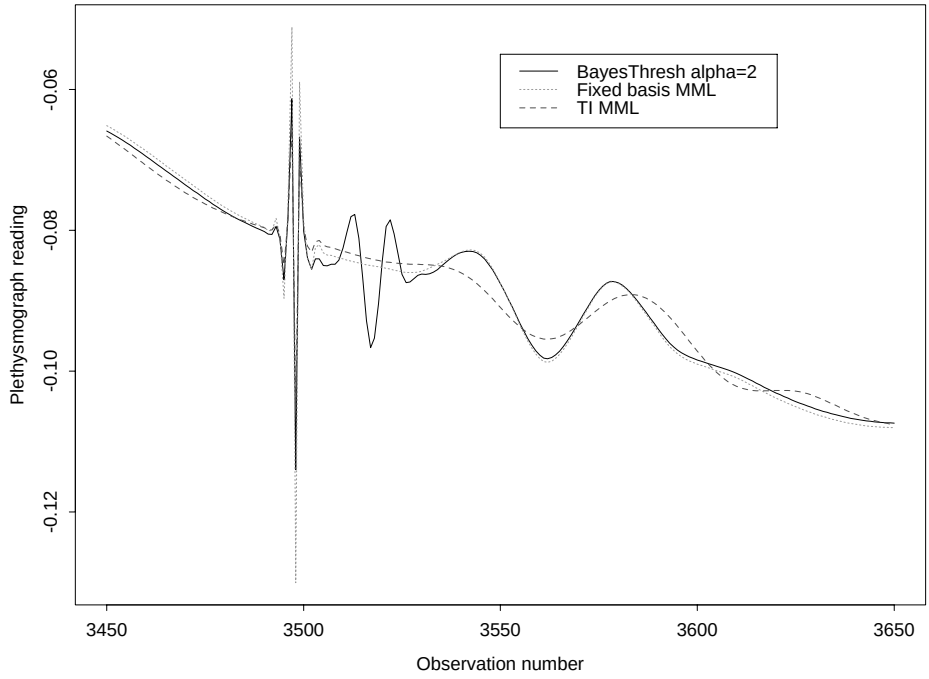

Figure 4: The comparison of Figure 2 for a short sequence of the plethysmography data, near time 3500

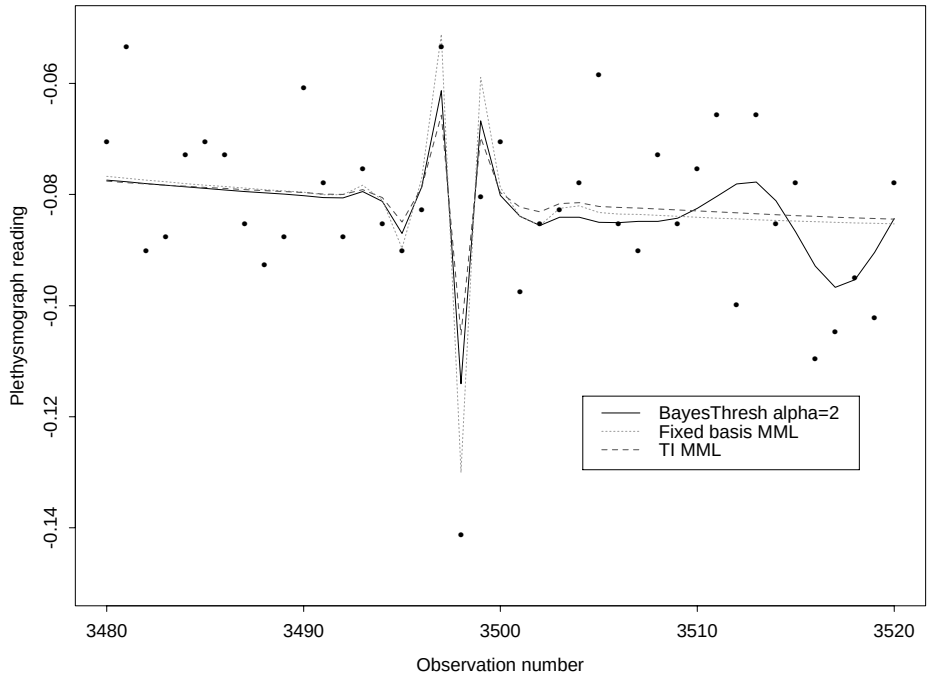

Figure 5: A very short sequence of the plethysmography data near time 3500, showing the three wavelet-based methods and the original data

possible outlier. In the interval from 3510 onwards, both the adjusted BayesThresh and the MML methods have moderately high frequency behavior that is smoothed out by the TI-MML method. The main point here is that the Empirical Bayes method automatically produces an estimator similar to that obtained after subjective tinkering with BayesThresh.

6 Correlated data

6.1 A level-dependent MML approach

The errors for the plethysmography data show very little evidence of correlation. (For example, an autocorrelation function of the first differences of the data shows a correlation at lag 1 very close to -0.5, and no appreciable correlation at lags 2 or greater.) However, the procedure we have set out extends immediately to correlated data, using ideas developed by Johnstone & Silverman (1997) The natural extension of the method is to treat each level separately, with its own estimate of the error variance. There is obviously a problem with the lowest levels. This has not much to do with the translation-invariant prescription. Rather, it can be explained by considering what happens if the true value of w is greater than about 0.5 and the variance is estimated from the data by the median absolute deviation, by estimating the standard deviation at level j by

$$\hat{\sigma}_j = 1.4826\,\mathrm{MAD}\{d^j\} \tag{19}$$

where d^j is the vector of wavelet coefficients at level j and MAD denotes the median absolute deviation.

In the case where w is large we will tend to overestimate the variance substantially, and then to estimate $C = 1$. These values do not identify the mixture at all, but estimate all the parameters to be zero. The estimate of σ will confuse signal and noise, because σ will be estimated from observations that contain signal rather than pure noise. To avoid this difficulty in the level-dependent case we can either use the same estimate of σ for all the lower levels, or else set $w = 1$ for the lower levels. The latter procedure corresponds to the standard 'primary resolution level' approach (see, for example, Hall & Nason (1997)) to wavelet smoothing, where at the lowest J_{keep} levels the sample wavelet coefficients are used directly as the estimates. The value J_{keep} is called the primary resolution level. A third possibility is to make use of the notion that the true error variance does not grow too quickly as one proceeds down the levels of the transform. Because the variances at very high levels can naturally be small, it seems best not to apply this prescription to very high levels. For example we modify the estimate of the standard deviation by recursively setting

$$\tilde{\sigma}_j = \min\{\hat{\sigma}_j, \alpha\tilde{\sigma}_{j+1}\} \tag{20}$$

for $j = m-4, m-3, \ldots, 1$, setting $\tilde{\sigma}_i = \hat{\sigma}_i$ for the three top levels $j = m-3, m-2, m-1$. The multiple α controls the rate at which the standard deviation is allowed to increase. The value $\alpha = \sqrt{2}$ is conservative, because it is an upper bound on the ratio that can obtained from a stationary fractional Brownian motion process, and seems to give good results.

6.2 The ion channel data

An important problem in molecular physiology is the detection and measurement of the picoamp currents that flow through the single membrane channels that control movement in and out of cells. In Johnstone & Silverman (1997), where further details on the data may

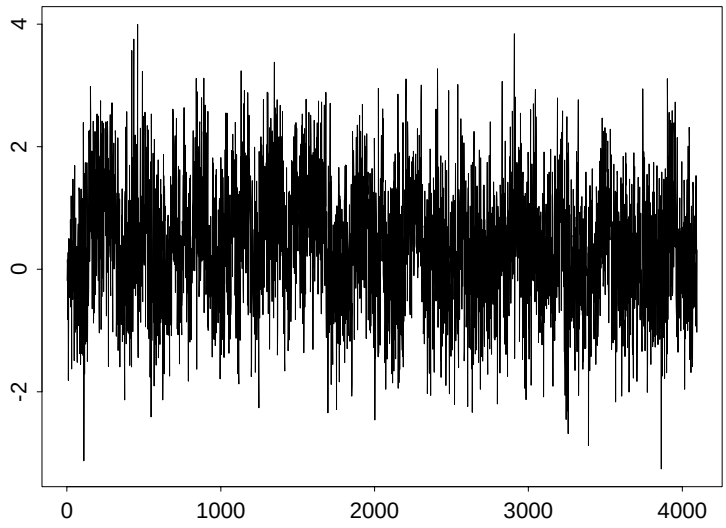

Figure 6: An extract of length 4096 from the data file provided by Eisenberg and Levis

be found, we analyzed some data generated by physiologists R. Eisenberg and R. Levis to represent most of the relevant challenges in processing such single channel data.

Figure 6 shows an extract of length 4096 from the data file of 100,000 points provided by Eisenberg and Levis. The data consists of a step function switching between values 0 ("off") and 1 ("on") at random, with average period about 125 points, and measured in the presence of significant additive, correlated noise.

For the ion channel data, it turns out that choosing a high primary resolution level is a good strategy. On a sample data set, it was found that choosing the primary resolution level such that we only subject the top three levels to the TI MML prescription works well. In Figure 7 we illustrate the effect of this approach on a single sequence of length 2^{15} from the data; the first 6000 time points are shown in the figure. The relatively noisy curve, when thresholded above or below 0.5, gives an excellent estimate of the true signal shown as a solid plot. In only processing the top three levels, we have let through substantial amounts of noise, but since we are only interested in whether the true value is 0 or 1, this does not affect the result deleteriously. In classical curve estimation terms, we prefer a moderate amount of variance rather than bias, because the thresholding at 0.5 will remove this variance.

A more systematic evaluation is obtained by applying the method to 10 successive segments of length 4096 from the original data record, as was done for the methods studied in Johnstone & Silverman (1997). Counting the number of errors in classifying time instants as "on" or "off" yields an average counting error of 83.8. This dominates all the methods considered in our earlier paper and is statistically indistinguishable from the result of 82.2 using Eisenberg and Levis's special purpose algorithm; the paired sample t-statistic for the difference between the two methods on the ten segments considered is 0.97.

6.3 Doppler example

We now turn to the Doppler example illustrated in Johnstone & Silverman (1997, Figure 3). Realizations of two different noise processes were added in turn to an evaluation of the

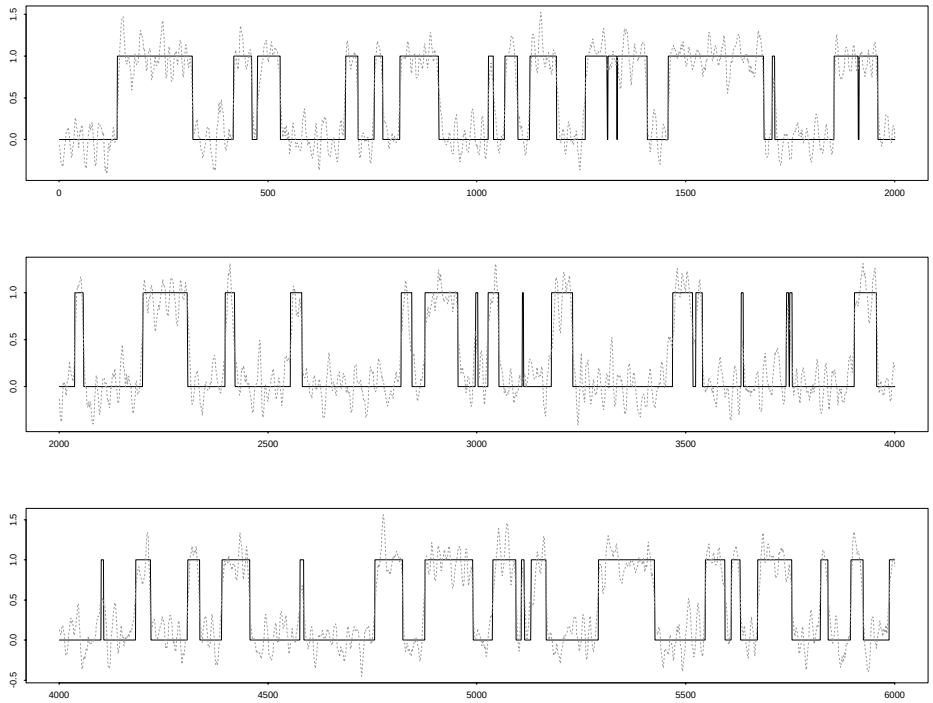

Figure 7: The true ion channel signal, and the curve obtained by applying the TI-MML approach to the top three levels only. Rounding the curve off to the nearest integer gives an excellent estimate of the original signal.

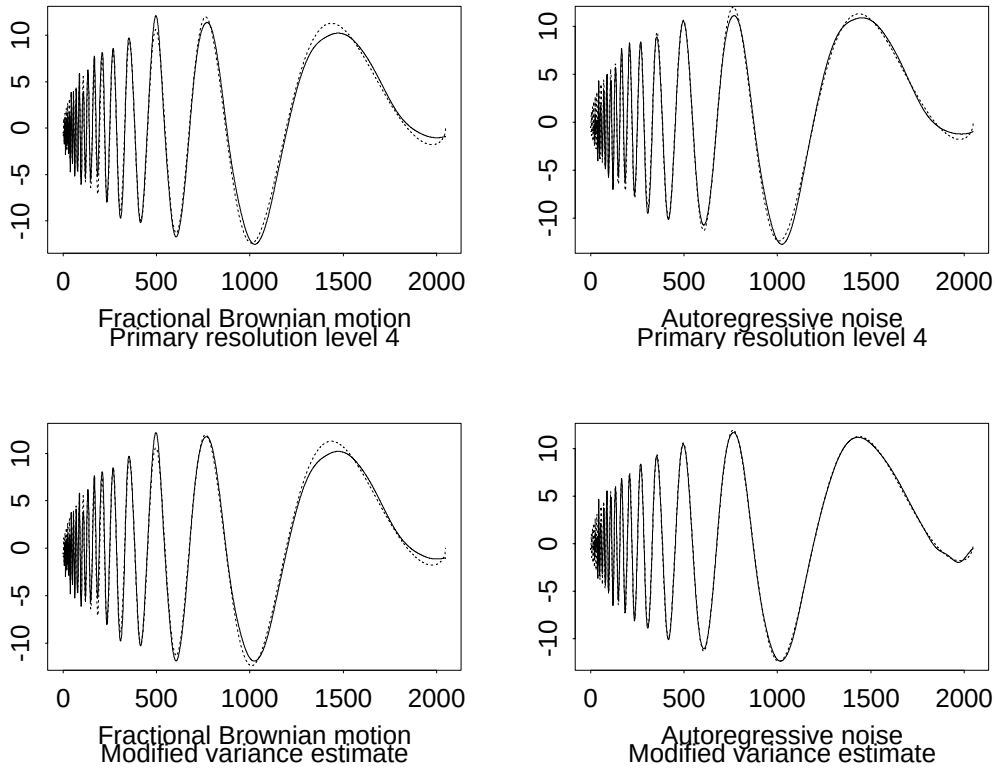

Figure 8: In each case the dotted curve is the true doppler curve and the solid curve is the estimate. In the left column the noise process is fractional Brownian motion, while in the right it is AR(2). In the upper two figures the estimate is obtained by the level-dependent MML method with primary resolution level 4. In the lower two figures all levels are processed, but the variance estimates are constrained as in (20).

underlying function at 2048 points. The first process is a fractional Brownian motion with power spectrum proportional to $1/f^{0.9}$; this was simulated by working in the frequency domain and then applying an inverse discrete Fourier transform. The second process is an AR(2) process y_j satisfying $y_j = \frac{4}{3}y_{j-1} - \frac{8}{9}y_{j-2} + z_j$, where z_j is a white noise process. In both cases, the noise process was scaled to have variance 1. Two methods of reconstruction were used. The level-dependent paradigm was applied with primary resolution level equal to 4, but not the translation-invariant option. The other approach was to process all levels of the wavelet transform, but to modify the variance estimates as described in (20) and the subsequent discussion. The results are shown in Figure 8. It can be seen that there is very little to choose between the results using a primary resolution level and a modified variance estimate.

7 Concluding remarks

The methods we have developed are attractive algorithmically in that they give calculations that are essentially linear in the amount of data being considered, and they work well on both simulated and real data. Software (based on SPLUS) is available from the authors.

Because the posterior median of the particular prior we use gives a thresholding rule, our empirical Bayes estimates of the hyperparameters can, of course, equally be regarded as generating data-based selections of thresholds. The essential property of the prior in this regard is the atom of probability at zero; one might also investigate extensions to priors where the nonzero part of the mixture is more general than a normal distribution.

One impetus for data-based threshold choices comes from certain shortcomings of constant threshold rules: high settings, such as the standard universal threshold $\sigma\sqrt{2\log n}$ of Donoho & Johnstone (1994) and Donoho et al. (1995) can have significant bias when the signal contains a reasonably large number of nonzero wavelet coefficients, while low values (such as 2σ) can pass through noise in sparse cases.

Another automatic choice is based on Stein's unbiased estimate of mean squared error (SURE); see Donoho & Johnstone (1995). While this has some attractive properties, it can be unstable in sparse cases in practice. Another approach is cross-validation; see, for example, Nason (1996) and Jansen et al. (1997). While a comparison of cross-validation and SURE is beyond the scope of this paper, it seems at least possible, bearing in mind results in the bandwidth selection literature (Haerdle et al. 1988, Hall & Johnstone 1992) that their behavior is quite similar. Preliminary investigations, to be reported elsewhere, suggest that the empirical Bayes threshold estimator studied in this paper seems happily to avoid the instabilities of both these methods.

References

Abramovich, F., Sapatinas, T. & Silverman, B. W. (1998), 'Wavelet thresholding via a bayesian approach', *J. Royal Statistical Society, Series B* **60**, 725–749.

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

Clyde, M. & George, E. I. (1998), Robust empirical bayes estimation in wavelets, Technical report, Institute of Statistics and Decision Sciences, Duke University.

Clyde, M., Parmigiani, G. & Vidakovic, B. (1998), 'Multiple shrinkage and subset selection in wavelets', *Biometrika* **85**, 391–403.

Coifman, R. R. & Donoho, D. L. (1995), Translation-invariant de-noising, *in* A. Antoniadis, ed., 'Wavelets and Statistics', Springer Verlag Lecture Notes.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the em algorithm (with discussion)', *J. Royal Statistical Society, Series B* **39**, 1–38.

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

Donoho, D. L. & Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *J. Amer. Statist. Assoc.* **90**, 1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995), 'Wavelet shrinkage: Asymptopia? (with discussion)', *Journal of the Royal Statistical Society, Series B* **57**, 301–369.

George, E. I. & Foster, D. P. (1997), Empirical bayes variable selection, Technical report, University of Texas at Austin.

Haerdle, W., Hall, P. & Marron, J. S. (1988), 'How far are automatically chosen regression smoothing parameters from their optimum?', *Journal of the American Statistical Association* **83**, 86–95.

Hall, P. & Johnstone, I. (1992), 'Empirical functionals and efficient smoothing parameter selection (with discussion)', *Journal of the Royal Statistical Society, Series B* **54**, 475–530.

Hall, P. & Nason, G. P. (1997), 'On choosing a non-integer resolution level when using wavelet methods.', *Statist. Probab. Lett.* **34**, 5–11.

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

Johnstone, I. M. & Silverman, B. W. (1997), 'Wavelet threshold estimators for data with correlated noise', *Journal of the Royal Statistical Society, Series B.* **59**, 319–351.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

McLachlan, G. J. & Krishnan, T. (1997), *The EM Algorithm and Extensions*, Wiley.

Nason, G. P. (1996), 'Wavelet shrinkage using cross-validation', *Journal of the Royal Statistical Society, Series B* **58**, 463–479.

Vidakovic, B. (1998), Wavelet-based nonparametric bayes methods, Technical report, ISDS, Duke University.