

Empirical Bayes Estimation in Wavelet Nonparametric Regression

Merlise A. Clyde
Edward I. George

ABSTRACT

Bayesian methods based on hierarchical mixture models have demonstrated excellent mean squared error properties in constructing data dependent shrinkage estimators in wavelets, however, subjective elicitation of the hyperparameters is challenging. In this chapter we use an Empirical Bayes approach to estimate the hyperparameters for each level of the wavelet decomposition, bypassing the usual difficulty of hyperparameter specification in the hierarchical model. The EB approach is computationally competitive with standard methods and offers improved MSE performance over several Bayes and classical estimators in a wide variety of examples.

1 Introduction

Wavelet shrinkage has become an increasingly popular method for compression and denoising of data in the context of signal and image processing as well as nonparametric regression (Donoho and Johnstone (1994, 1995)). The nonparametric regression model can be specified as

$$Y_i = f_i + \epsilon_i$$

where f_i represents the underlying unknown mean function and ϵ_i are independent $N(0, \sigma^2)$ random errors, representing additive white noise. In the wavelet domain, this can be equivalently expressed as

$$D_{jk} = \beta_{jk} + \epsilon_{jk} \tag{1}$$

where D_{jk} represent the elements of the data after applying the discrete wavelet transformation (DWT) and β_{jk} represent the wavelet coefficients of the function f ; the double indices reflect the multiresolution decomposition in the wavelet domain. Wavelet shrinkage estimation proceeds by estimating the β_{jk} by some shrinkage procedure, and then transforming the estimated coefficients back to the original domain by applying the inverse discrete wavelet transformation to obtain an estimate of the function f .

Bayesian methods, which offer coherent data-dependent shrinkage, have exhibited excellent integrated mean squared error properties in several studies (Abramovich et al. 1998, Chipman et al. 1997, Clyde et al. 1998) for estimation of f . The above Bayesian methods involve taking the standard linear model (1) with independent normal errors and embedding it in a conjugate hierarchical mixture model that takes into account that some wavelet coefficients will be zero or close to zero. The multiresolution decomposition suggests a natural grouping of wavelet coefficients by level which is reflected in specifying the distribution for β_{jk} conditional on the level j . Clyde et al. (1998) use a hierarchical model that expresses the belief that some of the wavelet coefficients β_{jk} are zero

$$\beta_{jk}|\gamma_{jk} \sim N(0, c_j\gamma_{jk}\sigma^2) \quad (2)$$

$$\gamma_{jk} \sim \text{Bernoulli}(\omega_j) \quad (3)$$

through the indicator variable γ_{jk} that determines if the coefficient is non-zero ($\gamma_{jk} = 1$), arising from a normal distribution with variance $c_j\sigma^2$, or degenerate at zero ($\gamma_{jk} = 0$). In the next stage of the hierarchy, the indicator variables γ_{jk} have independent Bernoulli distributions with $P(\gamma_{jk} = 1) = \omega_j$, for some fixed hyperparameter ω_j . The hyperparameter ω_j reflects the expected fraction of non-zero wavelet coefficients at level j . By collapsing these two stages, the prior distribution for β_{jk} can be equivalently represented as a two point mixture distribution,

$$\beta_{jk} \sim (1 - \omega_j)\delta(0) + \omega_j N(0, c_j\sigma^2)$$

where $\delta(0)$ represents a point-mass at 0. Chipman et al. (1997) consider a similar prior, but replace the point-mass at zero by a normal distribution that is tightly distributed around zero as in George and McCulloch (1993).

Because of the conditional independence structure in the prior distributions, the β_{jk} s are *a posteriori* conditionally independent,

$$p(\beta_{jk}|\gamma_{jk}, Y) \sim N\left(\gamma_{jk}\frac{c_j}{1+c_j}D_{jk}, \gamma_{jk}\sigma^2\frac{c_j}{1+c_j}\right).$$

Threshold estimators, where some of the coefficients are set to zero, can be obtained by selecting the highest posterior probability model, $\hat{\gamma}$, and using the posterior mean conditional on $\hat{\gamma}$,

$$E(\beta_{jk}|\hat{\gamma}, Y) = \hat{\gamma}_{jk}\frac{c_j}{1+c_j}D_{jk}. \quad (4)$$

(Clyde and George 1998). The posterior median (Abramovich et al. 1998) is another thresholding estimator.

An alternative shrinkage estimator is based on the posterior mean under Bayesian model averaging which takes into account uncertainty about γ_{jk} ,

$$E(\beta_{jk}|Y) = \pi(\gamma_{jk} = 1|Y)\frac{c_j}{1+c_j}D_{jk}, \quad (5)$$

(Clyde et al. 1998). Given Y , the γ_{jk} are independently distributed as Bernoulli random variables with

$$\pi(\gamma_{jk} = 1|Y) = \frac{O_{jk}}{1 + O_{jk}}, \quad (6)$$

where O_{jk} is the posterior odds that $\gamma_{jk} = 1$,

$$O_{jk} = (1 + c_j)^{-1/2} \left(\frac{\omega_j}{1 - \omega_j} \right) \exp \left\{ \frac{1}{2} \left(\frac{D_{jk}}{\sigma} \right)^2 \left(\frac{c_j}{1 + c_j} \right) \right\}. \quad (7)$$

While Bayesian methods are very flexible in the range of shrinkage patterns they can produce, subjective elicitation of the hyperparameters ω_j and c_j at each level j , is a difficult task. Clyde et al. (1998) used ideas of George and Foster (1997) to specify the prior hyperparameters so that the highest posterior model corresponds to the model selected using a classical model selection criterion. This, in effect, requires that one either elicit utilities/losses for model selection, which can be as difficult as specifying the prior hyperparameters, or use default choices such as AIC (Akaike 1973), BIC (Schwartz 1978), or RIC (Foster and George 1994). Abramovich et al. (1998) establish a relationship between the prior hyperparameters and Besov space parameters (α, β) which allows them to take into account the likely smoothness and regularity properties of the function. They assume the hyperparameters have the following structure

$$\begin{aligned} c_j &= C_1 (2^{-j})^\alpha \\ \omega_j &= \min(1, C_2 (2^j)^\beta) \end{aligned}$$

where C_1 and C_2 are additional hyperparameters (see also the chapter by Abramovich and Sapatinas in this volume). Noting that it is often difficult to elicit prior information about the smoothness of the function, they suggest default choices for α and β and use method of moments estimators for C_1 and C_2 .

Because of the difficulties of subjective elicitation, lack of knowledge about the function, and concern that a default prior may be at odds with the data, many of the proposed Bayesian methods use some form of data-dependent prior combined with assumptions about how the hyperparameters are related by level (Abramovich et al. 1998, Chipman et al. 1997, Yau and Kohn 1999). Rather than imposing any structure on the hyperparameters, Clyde and George (1998) and Johnstone and Silverman (1998) take an Empirical Bayes (EB) approach and estimate the hyperparameters in the prior distribution based on the marginal distribution of the data. These EB procedures not only bypass the difficulty of specifying the hyperparameters in the prior distributions, but are also very competitive with other wavelet shrinkage methods on computational grounds. In this chapter, we review these Empirical Bayes approaches and show how they can be used to construct both thresholding and shrinkage estimators for wavelet nonparametric regression.

2 Empirical Bayes

In an Empirical Bayes analysis, one would estimate the hyperparameters of the hierarchical model by some estimation procedure, commonly method of moments or maximum likelihood, and then proceed with the posterior analysis for the parameters of interest by treating the estimated hyperparameters as if they were known *a priori*. In the hierarchical model given by (1), (2), and (3), the unknown hyperparameters are σ^2 , c_j and ω_j . Many papers have considered estimating σ^2 using the MAD estimate,

$$\hat{\sigma} = \text{Median}(|D_{1k}|)/0.6745$$

(Donoho et al. 1995) using the wavelet coefficients at the finest level of resolution. We will first consider estimation of c_j and ω_j conditional on using the MAD estimate of σ , and then later proceed with joint estimation of σ in addition to c_j and ω_j by maximum likelihood estimation. Maximum likelihood estimates can be found by either direct maximization of the marginal likelihood (Clyde and George 1998) or by using the EM algorithm with an augmented likelihood (Johnstone and Silverman 1998).

2.1 Direct Maximum Likelihood Estimation of c_j and ω_j

Given an estimate for σ , such as the MAD estimate, c_j and ω_j can be estimated via maximum likelihood estimation using the marginal distribution of the data at level j . Marginalizing over β_{jk} and γ_{jk} , and conditioning on c_j , ω_j , and σ , the observations D_{jk} are independently distributed as a mixture of two normal components. The log likelihood \mathcal{L} for c_j and ω_j is

$$\begin{aligned} \mathcal{L}(c_j, \omega_j) &= \sum_k \log [\omega_j \phi(D_{jk}; 0, \sqrt{1+c_j}\sigma) + (1-\omega_j)\phi(D_{jk}; 0, \sigma)] \\ &= \text{constant} + \sum_k \log \left[1 + \omega_j \left((1+c_j)^{-\frac{1}{2}} e^{\frac{1}{2} \frac{D_{jk}^2}{\sigma^2} \frac{c_j}{1+c_j}} - 1 \right) \right] \end{aligned} \quad (8)$$

where $\phi(x; \mu, \sigma)$ denotes the normal density evaluated at the point x with mean μ and standard deviation σ . This form does not lead to closed form solutions for the maximum likelihood estimates \hat{c}_j and $\hat{\omega}_j$, and numerical methods must be used to obtain the MLEs. Clyde and George (1998) used nonlinear Gauss-Seidel iteration (see Thisted 1988, pp. 187-188). This involves solving the single variable optimization problem to first find \hat{c}_j as function of ω_j and then finding $\hat{\omega}_j$ using the estimate of \hat{c}_j . One cycles through these two optimization problems, successively substituting the current estimate until convergence is achieved. Any popular root finding algorithm may be used to solve the single variable equations. If the Hessian is positive definite for all values of c_j and ω_j (excluding the boundaries), then if the algorithm converges the solution is the global maximum.

2.2 Maximum Likelihood Estimation using the EM Algorithm

Johnstone and Silverman (1998) use an EM algorithm to find the MLE, based on a derivation that introduces an entropy function to create a modified likelihood, where the global maximum of the modified likelihood function is the global MLE of the marginal likelihood. This approach is equivalent to the general EM algorithm given by Neal and Hinton (1998). The EM algorithm in exponential family problems is particularly simple to implement (Dempster, Laird and Rubin 1977, Tanner 1996), and we present this alternative derivation. To implement the EM algorithm, we consider the likelihood given D and the latent variable γ , rather than the marginal likelihood (8). The log likelihood for the “augmented” or “complete” data, $X = (D, \gamma)$,

$$\begin{aligned} \mathcal{L}(c_j, \omega_j | D, \gamma) &= \left[\log \left(\frac{\omega_j}{1 - \omega_j} \right) - \frac{1}{2} \log(1 + c_j) \right] \sum_k \gamma_{jk} \\ &\quad - \frac{1}{2} (1 + c_j)^{-1} \sum_k \gamma_{jk} D_{jk}^2 / \sigma^2 \\ &\quad + n_j \log(1 - \omega_j) - \frac{1}{2} \sum_k (1 - \gamma_{jk}) \frac{D_{jk}^2}{\sigma^2} \end{aligned} \quad (9)$$

belongs to a regular exponential family of the form $a(\theta)^T b(X) + c(\theta) + d(X)$ where $\theta = (c_j, \omega_j)$, $a(\theta)$ is the vector of natural parameters and $b(X)$ is the vector of sufficient statistics, $b(X) = (\sum_k \gamma_{jk}, \sum_k (\gamma_{jk} D_{jk}^2 / \sigma^2))^T$.

Because of the exponential family form, the E-step of the EM algorithm consists of computing the expectation of the sufficient statistics with respect to the distribution of γ of given D

$$E[b(X) | D, c_j^{(i)}, \omega_j^{(i)}] = \left(\sum_k \hat{\gamma}_{jk}^{(i)}, \sum_k \hat{\gamma}_{jk}^{(i)} D_{jk}^2 / \sigma^2 \right)^T = \hat{b}^{(i)}(X)$$

where

$$\hat{\gamma}_{jk}^{(i)} = \frac{O_{jk}^{(i)}}{1 + O_{jk}^{(i)}}$$

is the posterior mean of γ_{jk} and $O_{jk}^{(i)}$ is the posterior odds (7) evaluated using the current estimates $\hat{c}_j^{(i)}$ and $\hat{\omega}_j^{(i)}$.

The M-step consists of maximizing $c(\theta) + a(\theta)^T \hat{b}^{(i)}(X)$, resulting in the solution

$$\hat{c}_j^{(i+1)} = \max \left(0, \frac{\sum_k \hat{\gamma}_{jk}^{(i)} D_{jk}^2}{\sum_k \hat{\gamma}_{jk}^{(i)} \sigma^2} - 1 \right) \quad (10)$$

$$\hat{\omega}_j^{(i+1)} = \frac{\sum_k \hat{\gamma}_{jk}^{(i)}}{n_j}. \quad (11)$$

If the estimates are in the interior of the parameter space, because the augmented likelihood belongs to a regular exponential family, the solutions for c_j and ω_j are the unique global solutions (conditional on $\hat{\gamma}_{jk}$) which follows from standard exponential family theory. The E and M steps are repeated until the estimates converge, and yield a stationary point of the marginal likelihood (8). Because the convergence rate of the EM algorithm is linear (Dempster, Laird and Rubin 1977), the Gauss-Seidel algorithm applied to (8) may be faster. As in the Gauss-Seidel algorithm, this results in a global solution if and only if the marginal likelihood is unimodal. In practice, however, we have noticed little difference in performance between the two approaches or difficulties with convergence.

2.3 Maximum Likelihood Estimation of σ^2

Rather than using the MAD estimate for σ , the augmented data likelihoods (9) at each level j can be combined to construct a complete data likelihood for estimating σ^2 through the EM algorithm. This complete data likelihood is still in a regular exponential family. The sufficient statistics for σ involve the same terms as in estimating c_j , so the E-step only involves the expectation of γ_{jk} . The M-step for estimating σ^2 has solution

$$\sigma^{2(i+1)} = \frac{1}{N} \sum_{j,k} \left(D_{jk}^2 - \frac{c_j}{1+c_j} \hat{\gamma}_{jk}^{(i)} D_{jk}^2 \right)$$

while the M-steps for c_j and ω_j are the same as before. The M-steps for σ^2 and c_j now involve iterative solutions. This approach takes full advantage of the data at all levels to construct an estimate of σ^2 , unlike the MAD estimate.

2.4 Conditional Likelihood Estimates

Clyde and George (1998) also consider a conditional likelihood approximation to the full likelihood, which yields rapidly computable analytic expressions for \hat{c}_j and $\hat{\omega}_j$. This can be viewed as taking the augmented likelihood (9) and evaluating it at the mode for γ_{jk} , rather than using the posterior mean, as in the EM algorithm. At level j , consider models γ where $q_j = \sum_k \gamma_{jk}$ is the number of nonzero wavelet coefficients. For fixed j , let $D_{j(k)}^2$ denote the sorted values (in decreasing order) of D_{jk}^2 . Then the most likely model with q_j nonzero components, $\gamma^*(q_j)$, corresponds to $\gamma_{j(k)}^* = 1$ if $k \leq q_j$, and 0 otherwise, based on assigning the q_j largest D_{jk}^2 values to the mixture component representing signal, and the remaining to the noise component. For each value of q_j , the values of c_j and ω_j that maximize the

conditional log likelihood are

$$\begin{aligned} c_j(q_j) &= \max \left\{ 0, \frac{\sum_k \gamma_{j(k)}^*(q_j) D_{j(k)}^2}{\sum_k \gamma_{j(k)}^*(q_j) \sigma^2} - 1 \right\} \\ &= \max \left\{ 0, \frac{\sum_{k \leq q_j} D_{j(k)}^2}{q_j \sigma^2} - 1 \right\} \end{aligned} \quad (12)$$

$$\omega_j(q_j) = \frac{\sum_k \gamma_{j(k)}^*}{n_j} = \frac{q_j}{n_j}. \quad (13)$$

It is straightforward to find the \hat{q}_j that maximizes the conditional likelihood, and the corresponding \hat{c}_j and $\hat{\omega}_j$ that yield the largest mode. Note that $L(c_j, \omega_j \mid \hat{\gamma}(\hat{q}_j))$ may be thought of as a profile likelihood approximation to $L(c_j, \omega_j)$. The conditional maximum likelihood estimates \hat{c}_j and $\hat{\omega}_j$ are alternative EB estimates which can be rapidly computed.

2.5 Comparing the Hyperparameter Estimators

Figure 1 illustrates profile likelihood plots of the marginal log likelihood, $\mathcal{L}(\hat{c}_j(\omega_j), \omega_j)$ based on (8), (left column) and the conditional log likelihood, $\mathcal{L}(\hat{c}_j(\omega_j), \omega_j \mid \hat{\gamma}(q_j(\omega_j)))$ based on (9), (right column) as a function of ω for a wavelet decomposition with 7 levels. To construct the profile likelihoods, $\hat{c}_j(\omega_j)$ is the MLE of c_j obtained by fixing ω_j in the marginal and conditional log likelihoods respectively. The corresponding marginal and conditional maximum likelihood estimates are given in Table 1.1. Although comparison of the marginal and conditional maximum likelihood estimates in Table 1.1 shows relatively close agreement, there is a suggestion of systematic bias in the conditional estimates, with a slight underestimation of ω_j and overestimation of c_j . Also, for the finest level of resolution there is a bimodality in the conditional loglikelihood, in which case we cannot distinguish between noise and signal. For cases like this in practice we find that the likelihood is extremely flat with estimates near the boundary with $\hat{\omega}_j \approx 0$ or $\hat{c}_j \approx 0$. As the posterior mean is approximately the same under both cases, this has not resulted in any serious bias for estimation in our experience. By comparing the EM and conditional MLE estimators (10) to (12) and (11) to (13), one sees that the estimators have the exact same form, but the EM estimates are evaluated with γ_{jk} at the posterior mean while the conditional estimates are evaluated with γ_{jk} at the posterior mode. One can see that in general the conditional and marginal maximum likelihood estimates will not agree, unless the posterior distribution of γ_{jk} is degenerate at 1 or 0, in which case the expected values and the modes for γ_{jk} will coincide. This difference will not disappear, even as the number of coefficients grows asymptotically, as posterior model probabilities will not necessarily converge to 0 or 1 asymptotically. For the coarser levels, with predominantly large coefficients (in absolute value) the posterior mean of

Level	MMLE ω	CMLE ω	MMLE c	CMLE c
s	0.85	0.81	1477.5	1557.7
6	0.71	0.63	660.5	760.5
5	0.62	0.53	313.6	367.4
4	0.39	0.36	391.8	431.4
3	0.21	0.18	197.7	238.7
2	0.08	0.07	72.5	93.9
1	0.04	0.03	21.5	35.5

TABLE 1.1. Maximum likelihood estimates of c_j and ω_j from the marginal (MMLE) and conditional likelihoods (CMLE).

γ_{jk} is often close to 1, resulting in less bias. The difference between the conditional and marginal estimators will be the most extreme if the posterior means of all the γ_{jk} equal one half. Fortunately in wavelets, a good basis should result in posterior model probabilities being close to zero or one, reducing the potential for bias. To understand the effect of the bias on shrinkage, note that the posterior model probabilities are nonlinear functions of c_j and ω_j , and it is the linear shrinkage in the form $c_j/(1 + c_j)$ and the multiple shrinkage through the posterior model probabilities that is critical in determining the posterior mean. As we will see later in the simulation study, these two errors appear to cancel each other for estimating the posterior mean.

2.6 Empirical Bayes Estimators

The EB estimates of σ^2 , c_j and ω_j are now used in the hierarchical model as if they had been fixed in advanced and are used to construct Bayesian estimators of the wavelet coefficients. A threshold estimator is obtained by first selecting the highest posterior probability model $\hat{\gamma}$, where $\hat{\gamma}_{jk} = 1$ if $\hat{\pi}(\gamma_{jk} = 1|Y) \geq 0.5$ and is zero otherwise, and $\hat{\pi}(\gamma_{jk} = 1|Y)$ is obtained by inserting the EB estimates into (6). The conditional posterior mean $E(\beta_{jk}|\hat{\gamma}, Y)$ in (4) is given by

$$\hat{E}(\beta_{jk}|\hat{\gamma}, Y) = \hat{\gamma}_{jk} \frac{\hat{c}_j}{1 + \hat{c}_j} D_{jk}. \quad (14)$$

This model selection shrinkage estimator thresholds the data by setting $\hat{\beta}_{jk} = 0$ whenever $\hat{\gamma}_{jk} = 0$. This is useful for compression problems where dimension reduction and elimination of negligible coefficients is important.

Alternatively, one might use the EB estimates to estimate the overall posterior mean, $E(\beta_{jk}|Y)$ in (5), which yields

$$\hat{E}(\beta_{jk}|Y) = \hat{\pi}(\gamma_{jk} = 1|Y) \frac{\hat{c}_j}{1 + \hat{c}_j} D_{jk}. \quad (15)$$

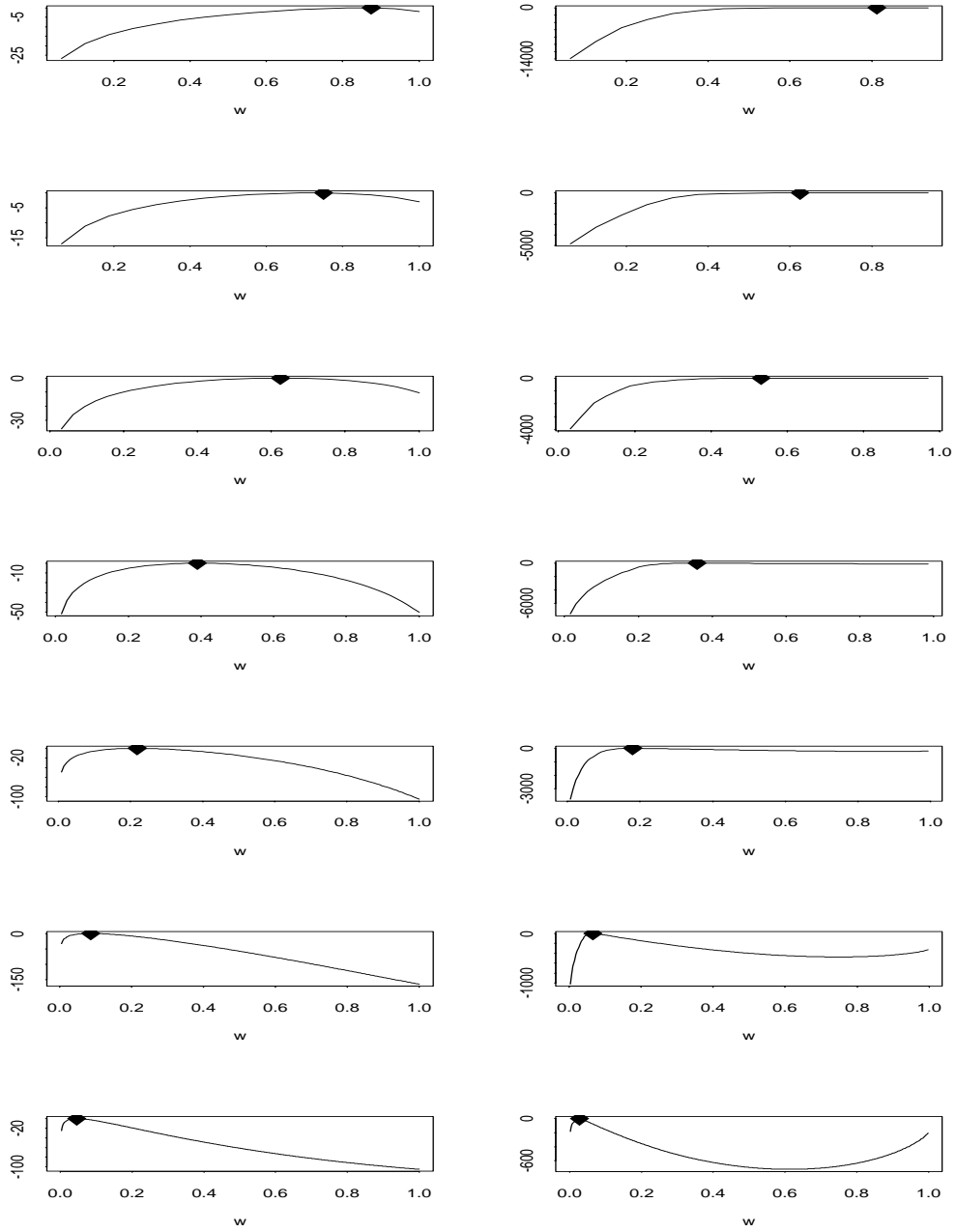


FIGURE 1. Marginal loglikelihood, $\log L(\hat{c}_j(\omega_j), \omega_j)$, (left column) and conditional loglikelihood $\log L(\hat{c}_j(\omega_j), \omega_j \mid \hat{\gamma}(q_j(\omega_j)))$, (right column) as a function of ω_j , for $j = s, 6, \dots, 1$. The triangles represent the location of the maximum.

This multiple shrinkage estimator (George 1986, Clyde et al. 1998) corresponds to Bayesian model averaging. In contrast to the thresholding behavior induced by $\hat{\gamma}_{jk}$ in (14), (15) includes an additional shrinkage factor $\hat{\pi}(\gamma_{jk} = 1|Y)$ to compensate for model uncertainty and appears to offer improved performance (Clyde and George 1998). Finally, note that both of the EB estimators (14) and (15) are fully automatic, as opposed to (4) and (5) which require hyperparameter specification.

3 Simulations

We compared the EB estimators to several existing shrinkage strategies: HARD: Hard thresholding with the universal rule (Donoho and Johnstone 1994) and SURE: SureShrink adaptive shrinkage rule as implemented in S+Wavelets, based on Donoho and Johnstone’s (1995) Sureshrink procedure, and RIC, which fixes the hyperparameters so that $c_j \equiv 1048561$ and $\omega_j \equiv 0.50$ corresponding to the Risk Inflation Criterion of Foster and George (1994). We used the four test functions “blocks”, “bumps”, “doppler”, “heavisine”, proposed by Donoho and Johnstone, and generated 100 samples of each function with $N = 1024$ and $\sigma = 1$. The signal-to-noise ratio $\text{SNR} = 7$ and the wavelet bases are chosen to match Donoho and Johnstone (1995). We evaluated the performance based on the average mean squared error (MSE) from the 100 simulations as

$$\text{MSE} = \frac{1}{100} \sum_{l=1}^{100} \sum_{i=1}^N \frac{(f_i - \hat{f}_i^l)^2}{N}$$

where f_i is the true signal and \hat{f}_i^l is the estimate of the function from simulation l .

Table 1.2 presents the average MSEs and standard deviations from the simulation study. We compared the EB model averaging estimator (15) with the marginal MLE of c_j and ω_j and the conditional MLE estimates using the MAD estimate of σ (the first two columns respectively) to the joint MLE of σ^2 , c_j and ω_j (column 3). The results indicate that all three EB estimators are superior to HARD, SURE, and RIC in this setting. Interestingly, performance is hardly affected, if at all, by using the conditional EB estimates instead of the marginal EB estimates. Apparently, the individual biases of the estimates of c and ω discussed in Section 2.5 have little effect. Using the MLE EB estimate of σ^2 is generally more efficient than the robust MAD estimate, as one would expect since it is a function of all of the data. However, the EM algorithm for estimation of σ^2 , c_j , and ω_j often took much longer to converge (sometimes more than 50 iterations), than the EM algorithm for c_j and ω_j with the MAD estimate of σ^2 .

Figure 2 shows the distribution of the maximum likelihood estimates of c_j and ω_j for the four test functions from the 100 simulations using

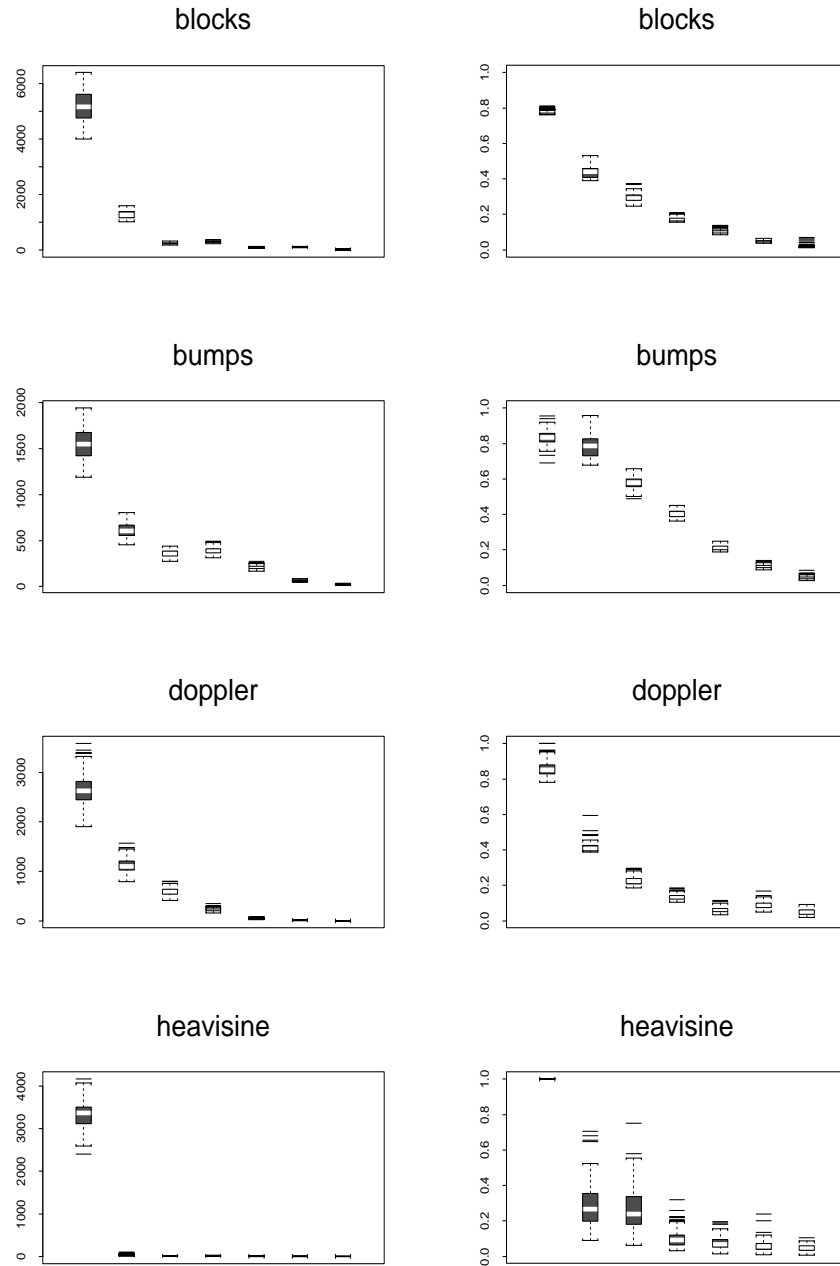


FIGURE 2. Distribution of the EB estimates of c_j (left column) and ω_j (right column) by level from the 100 simulations for the four functions.

Function	EB CML	EB MML	EB MML (all)	RIC	HARD	SURE
blocks	0.100 (0.002)	0.102 (0.002)	0.101 (0.002)	0.106 (0.002)	0.142 (0.002)	0.213 (0.002)
bumps	0.323 (0.003)	0.310 (0.003)	0.308 (0.003)	0.342 (0.003)	0.433 (0.0040)	0.382 (0.003)
doppler	0.147 (0.002)	0.144 (0.002)	0.143 (0.002)	0.152 (0.002)	0.208 (0.003)	0.204 (0.002)
heavisine	0.084 (0.002)	0.082 (0.001)	0.083 (0.001)	0.096 (0.002)	0.109 (0.002)	0.118 (0.002)

TABLE 1.2. Average mean squared error and (standard deviations) from 100 simulations for each function based on $\text{SNR} = 7$. The EB estimates are based on the posterior mean using the conditional MLE (CML) and the marginal MLE (MML) of c_j and ω_j using the MAD estimate of σ^2 , and the joint MLE of σ^2 , c_j , and ω_j (MML all). Values in bold indicate the estimator with the minimum average MSE.

the estimates from the joint estimation of c_j , ω_j and σ^2 . The variation of the estimates across the different levels is striking, revealing strong decay in both ω_j and c_j from top to bottom, but very different rates across functions. Although such decay might be roughly anticipated using a fixed hyperparameter Bayes setup with subjective prior inputs, it is very difficult to pre-specify the appropriate magnitude and rate of decay. Indeed, such fixed Bayes estimators did not perform as well as the EB estimators in Clyde and George (1998).

4 Discussion

In this chapter, we have discussed Empirical Bayes methods for wavelet estimation. Embedding the wavelet setup in a hierarchical normal mixture model, we considered conditional and marginal likelihood estimates of the unknown hyperparameters for each wavelet level. We then obtained shrinkage and threshold estimators based on posterior means under the estimated prior distributions. When applied to a variety of simulated examples, these shrinkage estimators performed better than current methods including fixed hyperparameter Bayes estimators. Johnstone and Silverman (1998) obtain similar results using the posterior median as an estimator.

Clyde and George (1998) consider extensions of the normal hierarchical model to include scale mixtures of normals. This allows for robustness to outliers through the use of heavy tailed error distributions such as the Student- t or power exponential distribution (Box and Tiao 1973). The EB approach yields robust estimators that are computationally competitive with classical methods (order N). The hierarchical Student- t EB estimates

are superior across a wide variety of situations.

An explanation for the improved performance of the EB estimators is that they allow for wide variation of hyperparameter values across different wavelet levels, yielding flexible shrinkage patterns. One could also achieve this by elaborating our hierarchical setup to include prior distributions on all the hyperparameters. If the computational issues for this approach could be simplified, this would be a promising competitor to our methods, and would provide improved estimates of the posterior variances over the naive EB approach that ignores uncertainty in the hyperparameter estimates.

Clyde and George (1998) found that the EB estimates can be very sensitive to the choice of σ . When prior information or data are available, the EB approach can easily incorporate prior information about the noise level σ . Some additional improvement could be made by placing a prior distribution on σ^2 and using MCMC methods, but with additional computational cost. The EB methods can also be adapted to the case of correlated noise, by replacing σ with a level dependent estimate σ_j as in Johnstone and Silverman (1997).

Finally, another avenue for future research in this area is performance evaluation. Any simulation evaluation such as ours is necessarily limited to one part of the overall parameter space. Although our Bayesian estimators do not appear to offer oracle or risk inflation like minimax guarantees (Donoho and Johnstone 1994; Foster and George 1994), it would be worthwhile to investigate regions of worst performance. In this vein, we would expect the EB estimators offer more robustness than fixed hyperparameter Bayes estimators. Another interesting, but difficult, direction would be asymptotic evaluation of the EB procedures. This is complicated by the fact that the model dimension is always increasing with the sample size. While the marginal EB estimates of c_j and ω_j appear to be asymptotically consistent (as n_j goes to infinity), this is not necessarily the case with the conditional EB estimates (Johnstone and Silverman 1998). But, even if the hyperparameter estimates are consistent, the posterior model probabilities do not necessarily converge to 0 or 1 asymptotically (particularly when c_j is small), thus model selection will not generally be asymptotically consistent in the wavelet context.

References

- Abramovich, F., Sapatinas, T., and Silverman, B.W. (1998). "Wavelet Thresholding via a Bayesian Approach," *Journal of the Royal Statistical Society, Series B*, 60, 725-749.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Eds B.N. Petrov and F. Csaki, pp. 267-81. Budapest:

Akademia Kiado.

- Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*, Wiley, NY.
- Chipman, H., Kolaczyk, E., and McCulloch, R. (1997). “Adaptive Bayesian Wavelet Shrinkage”, *Journal of the American Statistical Association*, 92, 1413-1421.
- Clyde, M. and George, E.I. (1998). “Robust Empirical Bayes Estimation in Wavelets”, ISDS Discussion Paper 98-21/<http://www.isds.duke.edu/>
- Clyde, M., Parmigiani, G., Vidakovic, B. (1998). “Multiple Shrinkage and Subset Selection in Wavelets,” *Biometrika*, 85, 391-402.
- Dempster, A.P. Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- Donoho, D.L., and Johnstone, I.M., (1994). “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, 81, 425-256.
- Donoho, D. and Johnstone, I. (1995). “Adapting to Unknown Smoothness via Wavelet Shrinkage,” *Journal of the American Statistical Association*, 90, 1200-1224.
- Donoho, D., Johnstone, I., Kerkycharian, G., and Picard, D. (1995). “Wavelet shrinkage: Asymptopia?” *Journal of the Royal Statistical Society, Series B*, 57, 301-369.
- Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics* 22, 1947–75.
- George, E.I. (1986). “Minimax multiple shrinkage estimation,” *Annals of Statistics*, 14, 188-205.
- George, E.I. and Foster, D.P. (1997). “Empirical Bayes Variable Selection”, Tech Report, University of Texas at Austin.
- George, E.I. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88, 881–89.
- Johnstone, I.M. and Silverman, B.W. (1997). “Wavelet Threshold Estimators for Data with Correlated Noise,” *Journal of the Royal Statistical Society, Series B*, 59, 319-351.
- Johnstone, I.M. and Silverman, B.W. (1998). “Empirical Bayes approaches to mixture problems and wavelet regression”, Technical report. Department of Mathematics, University of Bristol.

- Neal, R. M. and Hinton, G. E. (1998) “A view of the EM algorithm that justifies incremental, sparse, and other variants”, in M. I. Jordan (editor) *Learning in Graphical Models*, Dordrecht: Kluwer Academic Publishers, pages 355-368.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.
- Tanner, M.A. (1996). Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. New York: Springer, 3rd Edition. Chapter 4, pages 64-89.
- Yau, P. and Kohn, R. (1999). “Wavelet Nonparametric Regression Using Basis Averaging”. To appear in *Bayesian Inference in Wavelet Based Models* eds P. Müller and B. Vidakovic. Springer-Verlag.