



Flexible Empirical Bayes Estimation for Wavelets

Merlise Clyde; Edward I. George

Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 62, No. 4.
(2000), pp. 681-698.

Stable URL:

<http://links.jstor.org/sici?sici=1369-7412%282000%2962%3A4%3C681%3AFEBEFW%3E2.0.CO%3B2-B>

Journal of the Royal Statistical Society. Series B (Statistical Methodology) is currently published by Royal Statistical Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/rss.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

Flexible empirical Bayes estimation for wavelets

Merlise Clyde

Duke University, Durham, USA

and Edward I. George

University of Texas, Austin, USA

[Received September 1998. Final revision March 2000]

Summary. Wavelet shrinkage estimation is an increasingly popular method for signal denoising and compression. Although Bayes estimators can provide excellent mean-squared error (MSE) properties, the selection of an effective prior is a difficult task. To address this problem, we propose empirical Bayes (EB) prior selection methods for various error distributions including the normal and the heavier-tailed Student *t*-distributions. Under such EB prior distributions, we obtain threshold shrinkage estimators based on model selection, and multiple-shrinkage estimators based on model averaging. These EB estimators are seen to be computationally competitive with standard classical thresholding methods, and to be robust to outliers in both the data and wavelet domains. Simulated and real examples are used to illustrate the flexibility and improved MSE performance of these methods in a wide variety of settings.

Keywords: Bayesian model averaging; EM algorithm; Hierarchical models; Model selection; Multiple shrinkage; Orthogonal regression; Outliers; Robustness; Thresholding

1. Introduction

Wavelets are families of orthonormal basis functions that provide efficient representations for wide classes of functions, representations that are increasingly useful for signal denoising and compression. For this and motivated by frequentist considerations, Donoho and Johnstone (1994, 1995) proposed a variety of wavelet shrinkage procedures for estimating a function observed with normal error. For this set-up, recent papers such as Abramovich *et al.* (1998), Chipman *et al.* (1997) and Clyde *et al.* (1998) have proposed Bayesian shrinkage estimators that, for a variety of functions, yield an improved performance over the earlier shrinkage methods. Although such improvements are available over broad prior classes, the selection of the most effective prior can be especially difficult when little is known about the unknown function.

In this paper, we propose empirical Bayes (EB) methods to resolve the prior selection problem for Bayesian estimation of wavelet coefficients. We develop these methods not only for the conventional normal error model but also for heavier-tailed error models such as Student *t*-distributions that can better accommodate extreme values and outliers. To handle these different error distributions, we begin with a hierarchical Bayesian wavelet model that uses a scale mixture of normals for the error distribution (Section 2). We then propose EB methods that identify EB priors by using the data to estimate the hyperparameters and the

Address for correspondence: Merlise Clyde, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251, USA.

E-mail: clyde@isds.duke.edu

error variance of this hierarchical model (Section 3). Using these EB priors, we obtain threshold estimators based on model selection, and multiple-shrinkage (MS) estimators based on model averaging. These EB procedures not only bypass the difficulty of specifying the hyperparameters in the prior distributions but are also competitive with other shrinkage methods on computational grounds. These new EB estimators are seen to offer an improved performance over previous methods in simulation studies (Section 4) and are illustrated on real data (Section 5).

2. Statistical model

Suppose that $Y = (Y_1, \dots, Y_n)'$ are n equally spaced noisy observations of an unknown function $f = (f_1, f_2, \dots, f_n)'$ of interest, where n is a power of 2. Let $f = W\beta$ be an orthogonal wavelet decomposition of f , where W is an $n \times n$ orthonormal matrix and β is the $n \times 1$ vector of wavelet coefficients for f . Let $D = W'Y$ be the discrete wavelet transform (DWT) of Y , so that D is the $n \times 1$ vector of empirical or observed wavelet coefficients. For this set-up, we assume that the model for the data in the wavelet domain is

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

where ϵ_{jk} represents an additive error term. The doubly indexed subscripts reflect the multi-resolution nature of the wavelet decomposition; j indexes scales or levels of resolution and k indexes locations.

A standard assumption is that the ϵ_{jk} are independent and normally distributed with mean 0 and variance σ^2 . Although this is reasonable in many cases, extreme values and potential outliers may be more realistically modelled by heavy-tailed error distributions. To extend the error distribution beyond the standard normal model, we replace it by a scale mixture of normals

$$\epsilon_{jk} | \lambda_{jk} \sim N(0, \sigma^2 / \lambda_{jk}), \tag{2}$$

$$\lambda_{jk} \sim h \tag{3}$$

where the ϵ_{jk} are conditionally independent, the λ_{jk} are independent and $h(\cdot)$ is a scale mixing distribution on $(0, \infty)$. Scale mixtures of normals have been widely used in robustness studies and in outlier analysis (Andrews and Mallows, 1974; West, 1984, 1987; O'Hagan, 1979, 1988) and include as special cases the normal, Laplace, exponential power and Student t -distributions. The normal model is obtained when $\lambda_{jk} \equiv 1$, whereas independent errors with the t_ν -distribution (the Student t -distribution with ν degrees of freedom) are obtained when h is a gamma distribution,

$$\lambda_{jk} \sim \text{gamma}(\nu/2, 2/\nu) \tag{4}$$

with mean 1. For notational convenience, we shall use $p_h(x; 0, \sigma)$ to denote the marginal density of X when $X|\lambda \sim N(0, \sigma^2|\lambda)$ and $\lambda \sim h$.

For non-degenerate h , independent scale mixtures of normal errors in the wavelet domain correspond to a scale mixture of normal errors in the data domain, though a different mixture. To see this, note that given Λ , the diagonal matrix with diagonal terms λ_{jk} , the data domain errors are normal with covariance matrix $\sigma^2 W\Lambda^{-1}W'$. Mixing over Λ yields a different scale mixture with uncorrelated but typically dependent errors. (The unconditional covariance matrix, $\sigma^2 WE_h(\Lambda^{-1})W'$, is proportional to the identity matrix as the λ_{jk} are independent and identically distributed.) In particular, if the wavelet domain errors have

independent t_ν -distributions, then the data domain errors will be linear combinations of those t_ν random variables.

Although the specification of independent t_ν -distributions in the data domain may seem simpler than using other heavy-tailed error distributions, exact calculations in the wavelet domain become intractable and require computationally intensive methods. Furthermore, if the only available information about the process is that the errors appear to be uncorrelated with a heavy-tailed distribution (in either domain), then approximating the error distribution in the wavelet domain with independent t_ν -distributions, perhaps with level-dependent ν , seems reasonable and is more appropriate than the standard normal model. At the very least, this model provides a useful approximation which, as we show, yields computationally tractable EB estimators for non-normal error distributions that are robust to outliers.

2.1. Hierarchical model

The corner-stone of our Bayesian approach is an extension of the hierarchical normal mixture prior used by Clyde *et al.* (1998). Based on the natural multilevel grouping of the wavelet coefficients, this distribution for the β_{jk} at level j is

$$\beta_{jk} | \lambda_{jk}^*, \gamma_{jk} \sim N(0, \sigma^2 c_j \gamma_{jk} / \lambda_{jk}^*), \tag{5}$$

$$\lambda_{jk}^* \sim h^*, \tag{6}$$

$$\gamma_{jk} \sim \text{Bernoulli}(\omega_j) \tag{7}$$

where the β_{jk} are conditionally independent, the λ_{jk}^* and γ_{jk} are independent, and $h^*(\cdot)$ is a scale mixing distribution on $(0, \infty)$. We also assume that the β_{jk} and the ϵ_{jk} are conditionally independent given the λ_{jk}^* and the λ_{jk} .

When the indicator variable $\gamma_{jk} = 0$, the wavelet coefficient $\beta_{jk} = 0$; when $\gamma_{jk} = 1$, β_{jk} is conditionally normal with a variance $\sigma^2 c_j / \lambda_{jk}^*$. There are two fixed hyperparameters at each level: the expected fraction ω_j of non-zero wavelet coefficients at level j and a fixed scaling factor $c_j > 0$ for these non-zero coefficients. As with the error distribution, the introduction of random scale coefficients λ_{jk}^* allows for a heavy-tailed distribution of the non-zero β_{jk} . As will be seen in the next section, the effect of the choice of h^* on posterior probabilities is strongly influenced by the choice of h for the error distribution.

The final specification issue concerns the dependence between the λ_{jk}^* and the λ_{jk} . Because we do not expect a relationship between the signal and noise components, the most reasonable choice would be to specify these as independent, in which case the β_{jk} and the ϵ_{jk} would be unconditionally independent. However, as will be seen, the EB calculations under this independent λ specification become analytically intractable and must be carried out numerically.

To ease the computational burden substantially, it may be useful to consider the alternative specification obtained by forcing $\lambda_{jk}^* = \lambda_{jk}$ but maintaining the independence of the λ_{jk} . Under this common λ specification, each β_{jk} and ϵ_{jk} pair are still uncorrelated (because of their conditional independence) but become unconditionally dependent through λ_{jk} . As is illustrated in Fig. 1 in the next section, this dependence appears to have only a mild effect on posterior quantities of interest. Because the resulting posterior calculations are faster than the fully independent $(\lambda_{jk}, \lambda_{jk}^*)$ model and yield predictions that compare favourably with competing methods, this common λ specification is at least a useful approximation to the full independence specification.

2.2. Posterior distributions

The γ_{jk} are *a posteriori* independent Bernoulli random variables, as a result of the conditional independence structure in the prior distributions and orthogonality of the DWT. The posterior probability that a coefficient is non-zero, $\pi(\gamma_{jk} = 1|Y)$, can be expressed as a function of the posterior odds O_{jk} ,

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

$$O_{jk} = \left(\frac{\omega_j}{1 - \omega_j} \right) \frac{m_1(D_{jk}|\gamma_{jk} = 1)}{m_0(D_{jk}|\gamma_{jk} = 0)} \tag{8}$$

where m_0 and m_1 are the marginal distributions of the data D_{jk} when $\gamma_{jk} = 0$ and $\gamma_{jk} = 1$ respectively. The first of these is a standard normal scale mixture

$$m_0(D_{jk}|\gamma_{jk} = 0) = \int \left(\frac{\lambda_{jk}}{2\pi\sigma^2} \right)^{1/2} \exp \left(-\frac{1}{2} \frac{D_{jk}^2 \lambda_{jk}}{\sigma^2} \right) h(d\lambda_{jk}) = p_h(D_{jk}; 0, \sigma), \tag{9}$$

which can be computed in closed form in many cases of interest. The second of these is

$$m_1(D_{jk}|\gamma_{jk} = 1) = \int \int \frac{1}{\sqrt{(2\pi\sigma^2)}} \left(\frac{c_j}{\lambda_{jk}^*} + \frac{1}{\lambda_{jk}} \right)^{-1/2}$$

$$\times \exp \left\{ -\frac{1}{2} D_{jk}^2 \left(\frac{\sigma^2 c_j}{\lambda_{jk}^*} + \frac{\sigma^2}{\lambda_{jk}} \right)^{-1} \right\} h^*(d\lambda_{jk}^*) h(d\lambda_{jk}) \tag{10}$$

when λ_{jk}^* and λ_{jk} are independent, and

$$m_1(D_{jk}|\gamma_{jk} = 1) = \int \left\{ \frac{\lambda_{jk}}{2\pi\sigma^2(1 + c_j)} \right\}^{1/2} \exp \left\{ -\frac{1}{2} \frac{D_{jk}^2 \lambda_{jk}}{\sigma^2(1 + c_j)} \right\} h(d\lambda_{jk})$$

$$= p_h\{D_{jk}; 0, \sigma\sqrt{(1 + c_j)}\} \tag{11}$$

when $\lambda_{jk}^* = \lambda_{jk}$.

Table 1 summarizes several special cases to which we turn our attention. The normal-normal model for which the β_{jk} and ϵ_{jk} have independent normal distributions is obtained when $\lambda_{jk}^* \equiv \lambda_{jk} \equiv 1$ and yields a closed form expression for the posterior odds O_{jk} of $\gamma_{jk} = 1$:

Table 1. Joint distributions of β_{jk} and λ_{jk}

Label	λ_{jk}^* and λ_{jk} distributions	β_{jk} and ϵ_{jk} distributions	β_{jk} and ϵ_{jk} dependence
Normal-normal	$\lambda_{jk}^* = \lambda_{jk} \equiv 1$	Normal prior Normal errors	Independent
Normal-t	$\lambda_{jk}^* \equiv 1$	Normal prior	Independent
t-t	$\lambda_{jk}^* \sim \text{gamma}(\nu/2, 2/\nu)$ $\lambda_{jk} \sim \text{gamma}(\nu/2, 2/\nu)$	t_ν -errors t_ν -prior	Independent
Cauchy-t	$\lambda_{jk}^* \sim \text{gamma}(\nu/2, 2/\nu)$ $\lambda_{jk}^* \sim \text{gamma}(1/2, 2)$	t_ν -errors Cauchy prior	Independent
Bivariate t	$\lambda_{jk}^* \sim \text{gamma}(\nu/2, 2/\nu)$ $\lambda_{jk}^* = \lambda_{jk}$ $\lambda_{jk} \sim \text{gamma}(\nu/2, 2/\nu)$	t_ν -errors t_ν -prior t_ν -errors	Uncorrelated but dependent

$$O_{jk} = (1 + c_j)^{-1/2} \left(\frac{\omega_j}{1 - \omega_j} \right) \exp \left\{ \frac{1}{2} \left(\frac{D_{jk}}{\sigma} \right)^2 \frac{c_j}{1 + c_j} \right\}. \tag{12}$$

Next, to allow for heavy-tailed error distributions, we consider several specifications under which the errors have the t_ν -distribution corresponding to λ_{jk} independent and identically distributed (IID) gamma($\nu/2, 2/\nu$). The first of these is the normal- t model obtained by $\lambda_{jk}^* \equiv 1$, where the β_{jk} are normal and independent of the ϵ_{jk} . Although this provides robustness to outliers through the likelihood, the tails of the error distribution dominate the tails of the prior. This leads to posterior distributions that classify extremely large values of $|D_{jk}|$ as noise rather than signal, which may not be desirable. Posterior computations for this case must be done by numerical integration.

From a Bayesian prior robustness perspective, the tails of the prior should typically be at least as flat as the tails of the likelihood (Fan and Berger, 1990). When the errors have a t_ν -distribution, this is achieved when the β_{jk} are IID t_{ν^*} with $\nu^* \leq \nu$ and are independent of the ϵ_{jk} . This corresponds to taking λ_{jk}^* IID gamma($\nu^*/2, 2/\nu^*$) with $\nu^* \leq \nu$, independently of the λ_{jk} . We focus on the two special cases of this independent t specification listed in Table 1: the t - t model with equal degrees of freedom, $\nu^* = \nu$, and the Cauchy- t model with $\nu^* = 1$. The marginal distribution in equation (10) can be obtained by first integrating out λ_{jk} and λ_{jk}^* ,

$$m_1(D_{jk} | \gamma_{jk} = 1) = \int \frac{K_\nu}{\sigma \{1 + (D_{jk} - \beta_{jk})^2 / \nu \sigma^2\}^{(\nu+1)/2}} \frac{K_{\nu^*}}{\sigma \sqrt{c_j} (1 + \beta_{jk}^2 / \nu^* c_j \sigma^2)^{(\nu^*+1)/2}} d\beta_{jk} \tag{13}$$

reducing equation (10) to a one-dimensional integral over β_{jk} of the product of the two t -densities (K_ν and K_{ν^*} are the normalizing constants of the t_ν - and t_{ν^*} -densities respectively). Furthermore, for cases where ν and ν^* are both odd integers, Fan and Berger (1990) have shown that exact closed form expressions for equation (13) and the posterior mean of β_{jk} can be obtained. Although the marginal distribution in such cases can be expressed in closed form (the lengthy expressions are omitted), it is a complicated function of c_j and σ^2 . As a result, numerical methods are required to obtain the corresponding EB estimates discussed in Section 3.

Finally, a prior set-up for t_ν -distribution errors, which lessens the above computational burdens, is the bivariate t model where the $(\beta_{jk}, \epsilon_{jk})$ pairs have independent bivariate t_ν -distributions corresponding to $\lambda_{jk}^* = \lambda_{jk} \sim \text{gamma}(\nu/2, 2/\nu)$. The marginal distributions m_0 and m_1 are both rescaled t_ν -distributions, and the odds expression simplifies to

$$O_{jk} = (1 + c_j)^{-1/2} \left(\frac{\omega_j}{1 - \omega_j} \right) \left\{ \frac{1 + (1/\nu) D_{jk}^2 / \sigma^2 (1 + c_j)}{1 + (1/\nu) D_{jk}^2 / \sigma^2} \right\}^{-(\nu+1)/2}. \tag{14}$$

Although β_{jk} and ϵ_{jk} are dependent here, they remain uncorrelated. As will be seen later, posterior probabilities under the bivariate t model closely approximate posterior probabilities under the t - t model but can be computed in less than a seventh of the time.

Fig. 1 shows the posterior probability $\pi(\gamma_{jk} = 1 | Y)$ as a function of $|t_{jk}| = |D_{jk}|/\sigma$ for various choices of c_j under each of the priors listed in Table 1 for $\nu = 5$. (Numerical integration was required to compute the curves for the normal- t model.) Under the normal-normal model when $c_j > 0$, $\pi(\gamma_{jk} = 1 | Y)$ increases rapidly in $|t_{jk}|$ to 1, thus treating extreme values as signal. In sharp contrast, under the normal- t model, $\pi(\gamma_{jk} = 1 | Y)$ eventually decreases in $|t_{jk}|$, causing it to treat extreme values as noise. This occurs because the tails of the t_ν -distribution eventually dominate the normal prior. This behaviour, which is more

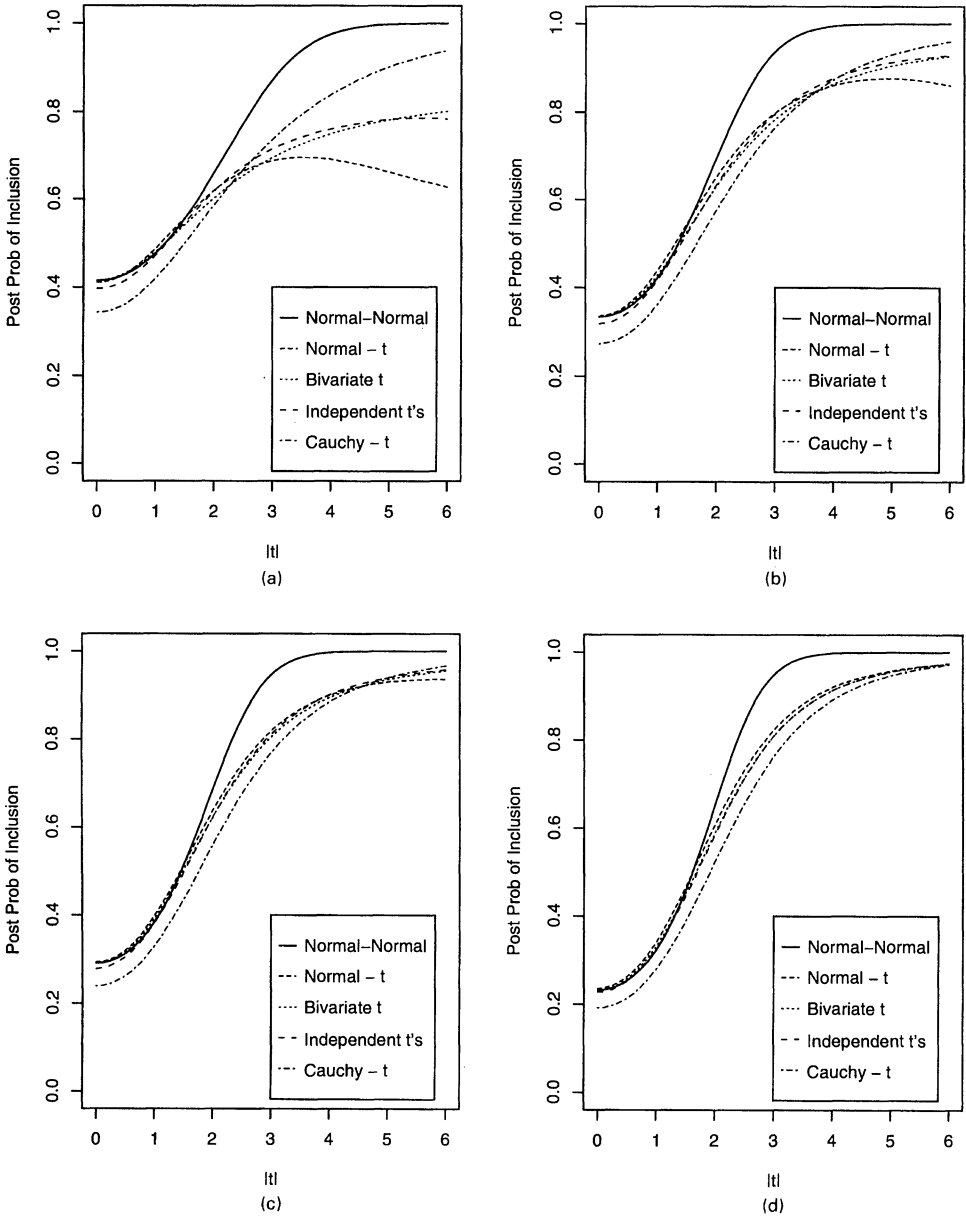


Fig. 1. Posterior probability of inclusion, $\pi(\gamma_{jk} = 1|Y)$, as a function of $|t_{jk}| = |D_{jk}|/\sigma$ under the specifications listed in Table 1 (in all cases $\omega = 0.5$): (a) $c = 1, 5$ degrees of freedom; (b) $c = 3, 5$ degrees of freedom; (c) $c = 5, 5$ degrees of freedom; (d) $c = 10, 5$ degrees of freedom

evident when c_j is small, is undesirable for functions that have a few very large coefficients at the finest level because the normal-t model would tend to treat them as outliers.

The Cauchy-t curve corresponds to an independent t specification with a Cauchy prior ($\nu^* = 1$) and a t_5 error distribution. The Cauchy prior is the easiest of the independent t priors to handle computationally and has desirable robustness properties. Like the normal-

normal model, $\pi(\gamma_{jk} = 1|Y)$ under the Cauchy- t eventually approaches 1 for large $|D_{jk}|$. However, it does so more slowly, exhibiting greater uncertainty that an observation represents signal.

The $\pi(\gamma_{jk} = 1|Y)$ curves for both the bivariate t and the t - t model (with $\nu^* = \nu = 5$) are very similar, especially for large c_j . Providing a compromise between the normal-normal and the normal- t models, both add robustness while avoiding the normal- t non-monotonicity phenomenon. For the bivariate t model, the posterior odds O_{jk} have a limit of $\{\omega_j/(1 + \omega_j)\}(1 + c_j)^{\nu_j/2}$, as $|t_{jk}|$ goes to ∞ . For large c_j , this behaves effectively like the normal model. However, compared with both the normal-normal and the Cauchy- t models, for small c_j (where it is difficult to distinguish between signal and noise), there is more uncertainty about whether a coefficient is non-zero, which does not disappear for large $|D_{jk}|$. Because of the availability of simple closed form posterior expressions, the bivariate t model can be seen as an attractive approximation to the t - t model, providing a reasonable compromise between robust estimation and computational ease. Because of the similarity of the bivariate t and the t - t models, we shall focus our attention on the normal-normal, bivariate t and the Cauchy- t models.

3. Empirical Bayes methods

The implementation of the Bayes hierarchical model (5)–(7) with fixed hyperparameter values requires the specification of c_j and ω_j at each level j of the wavelet decomposition. Unfortunately, a meaningful prespecification of these is difficult at best. Although one’s prior beliefs might reflect an ordering of the c_j and ω_j by level, it is difficult subjectively to elicit additional information. A fully automatic EB approach that estimates c_j , ω_j and σ^2 from the data is appealing because it avoids this difficulty and the possibility of specifying values which are in conflict with the data.

The EB approach that we propose is based on using maximum marginal likelihood (MML) estimates of the hyperparameters and extends the approach of George and Foster (2000) for the normal linear model. Because closed form EB estimates of c_j and ω_j are generally unavailable, numerical methods must be used. For this purpose, we consider several computationally attractive alternatives: direct numerical maximization, EM algorithms and a conditional maximum likelihood (CML) approximation.

3.1. Direct maximum likelihood estimation

Marginally, the observations D_{jk} given c_j , ω_j and σ^2 are independently distributed as a mixture of m_0 as in equation (9) and m_1 as given by equation (11) or (13), with respective weights ω_j and $1 - \omega_j$. The log-likelihood \mathcal{L} for $c = (c_1, \dots, c_J)'$, $\omega = (\omega_1, \dots, \omega_J)'$ and σ^2 is then

$$\mathcal{L}(c, \omega, \sigma^2) = \sum_j \left[\sum_k \log\{\omega_j m_1(D_{jk}; c_j, \sigma) + (1 - \omega_j) m_0(D_{jk}; \sigma)\} \right]. \tag{15}$$

Although closed form solutions for the MML estimates \hat{c}_j and $\hat{\omega}_j$ are generally unavailable, standard iterative methods can be used for direct maximization of the log-likelihood (15). In particular, we have found non-linear Gauss-Seidel iteration (Thisted (1988), pages 187–188) to work well. This entails iterating between finding the maximizing \hat{c}_j given $(\hat{\omega}_j, \hat{\sigma}^2)$, the maximizing $\hat{\omega}_j$ given $(\hat{c}_j, \hat{\sigma}^2)$ for each j and the maximizing $\hat{\sigma}^2$ given $(\hat{c}_1, \hat{\omega}_1), \dots, (\hat{c}_J, \hat{\omega}_J)$. As the Hessian of the log-likelihood is in general not positive definite for all values of c_j , ω_j and σ^2 , convergence may only be to a local mode. However, we have achieved reasonable success

using the median absolute deviation (MAD) estimate $\hat{\sigma} = \text{median}(|D_{1k}|)/0.6745$ (Donoho *et al.*, 1995) as an initial value for σ , and $\hat{\omega}_j$ equal to the number of observed wavelet coefficients that exceed $\sqrt{2 \log(n)}\hat{\sigma}$ based on hard thresholding.

3.2. Maximum likelihood estimation using the EM algorithm

For the normal-normal model with a fixed value for σ^2 , Johnstone and Silverman (1998) have recently proposed an EM algorithm based on a metric inequality. To implement an EM algorithm here, we consider the augmented log-likelihood given D and latent or ‘missing’ variables λ and γ , rather than the marginal likelihood in equation (15). Unless otherwise noted, we shall now restrict attention to common λ models such as the normal-normal and the bivariate t models where $\lambda_{jk}^* = \lambda_{jk}$, which leads to closed form iterative solutions for the parameters.

Let $\tau_j^{-1} = (1 + c_j)\sigma^2$, where τ_j corresponds to the unknown precision parameter for the data D_{jk} given $\gamma_{jk} = 1$ and let $\phi = 1/\sigma^2$, the precision when $\gamma_{jk} = 0$. The log-likelihood for $\tau = (\tau_1, \dots, \tau_J)'$, ω and ϕ based on the ‘augmented’ or ‘completed’ data, $X = (D, \lambda, \gamma)$, is

$$\begin{aligned} \mathcal{L}(\omega, \tau, \phi|D, \lambda, \gamma) = & \sum_j \left\{ n_j \log(1 - \omega_j) + \log\left(\frac{\omega_j}{1 - \omega_j}\right) \sum_k \gamma_{jk} \right\} \\ & + \frac{1}{2} \sum_j \left\{ \log(\tau_j) \sum_k \gamma_{jk} - \tau_j \sum_k D_{jk}^2 \lambda_{jk} \gamma_{jk} \right\} \\ & + \frac{1}{2} \log(\phi) \sum_{jk} (1 - \gamma_{jk}) - \frac{1}{2} \phi \sum_{j,k} D_{jk}^2 \lambda_{jk} (1 - \gamma_{jk}) + \sum_{jk} \log\{h(\lambda_{jk})\} \quad (16) \end{aligned}$$

and belongs to a regular exponential family of the form $a(\theta)' b(X) + c(\theta) + d(X)$ where $\theta = (\omega, \tau, \phi)$, $a(\theta)$ is the $(2J + 1)$ -dimensional vector of natural parameters and $b(X)$ is the $(2J + 1)$ -vector of sufficient statistics with components $(\sum_k \gamma_{jk}, \sum_k D_{jk}^2 \lambda_{jk} \gamma_{jk})$ for $j = 1, \dots, J$ and $\sum_{jk} D_{jk}^2 \lambda_{jk} (1 - \gamma_{jk})$.

The E-step of the EM algorithm consists of computing the expectations of the sufficient statistics with respect to the distribution of (λ, γ) given D, c and ω . The posterior mean of γ_{jk} is

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

where $O_{jk}^{(i)}$ are the posterior odds (8) evaluated using the current estimates $\hat{c}_j^{(i)}$ and $\hat{\omega}_j^{(i)}$. For the bivariate t model, where $\lambda_{jk} \sim \text{gamma}(\nu_j/2, 2/\nu_j)$, the posterior distribution of λ_{jk} given γ_{jk} is gamma, with

$$\begin{aligned} E(\lambda_{jk} \gamma_{jk}|D) &= \frac{\nu_j + 1}{\nu_j + D_{jk}^2 \hat{\phi}^{(i)}} \hat{\gamma}_{jk}^{(i)} = [\hat{\lambda}_{jk} \hat{\gamma}_{jk}]^{(i)}, \\ E(\lambda_{jk} (1 - \gamma_{jk})|D) &= \frac{\nu_j + 1}{\nu_j + D_{jk}^2 \hat{\phi}^{(i)}} (1 - \hat{\gamma}_{jk}^{(i)}) = [\hat{\lambda}_{jk} (1 - \hat{\gamma}_{jk})]^{(i)}. \end{aligned}$$

The M-step consists of maximizing the augmented likelihood with the latent data now replaced by their posterior expected values, resulting in

$$\hat{\sigma}^{2^{(i+1)}} = \frac{\sum_{jk} D_{jk}^2 [\hat{\lambda}_{jk}(1 - \hat{\gamma}_{jk})]^{(i)}}{n - \sum_{jk} \hat{\gamma}_{jk}^{(i)}}, \tag{17}$$

$$\hat{\omega}_j^{(i+1)} = \frac{\sum_k \hat{\gamma}_{jk}^{(i)}}{n_j}, \tag{18}$$

$$\hat{c}_j^{(i+1)} = \max \left(0, \frac{\sum_k D_{jk}^2 [\hat{\lambda}_{jk} \hat{\gamma}_{jk}]^{(i)}}{\hat{\sigma}^{2^{(i+1)}} \sum_k \hat{\gamma}_{jk}^{(i)}} - 1 \right). \tag{19}$$

The EM algorithm for the normal-normal model is obtained by simply setting $\hat{\lambda}_{jk} \equiv 1$ in the above.

If the parameter estimates are in the interior of the parameter space, the solutions above are the unique solutions (conditional on the values of latent data) because the complete data belong to a regular exponential family. The E- and M-steps are repeated until the estimates converge and yield a stationary point of the marginal likelihood (15). However, as in the case of direct maximization of the marginal likelihood using Gauss-Seidel or other methods, the EM algorithm may converge to a local mode. Because the convergence rate of the standard EM algorithm is linear (Dempster *et al.*, 1977), the direct maximization methods described above may result in faster convergence. However, the iterative solutions are available in closed form for the common scale λ_{jk} model and provide some insight into the problem and connections to the CML estimates of George and Foster (2000). The M-step estimates of σ^2 and w_j have natural interpretations; the estimate of σ^2 is the ratio of the posterior expected error sum of squares to the posterior expected degrees of freedom, and the estimate of w_j is the posterior expected fraction of non-zero coefficients.

Although it is not necessary to introduce λ_{jk} as a latent variable, this does permit closed form iterative expressions for the hyperparameter estimates of the bivariate t model. For models where the complete data are not from a regular exponential family, such as in the Cauchy- t model, numerical optimization is required to carry out the M-step for c_j and σ^2 , although $\hat{\omega}_j$ has the same form as in equation (18). In practice, we have noticed very little difference between maximum likelihood estimates (MLEs) using the EM algorithm or the direct maximization approach for the normal-normal, bivariate t or Cauchy- t models.

3.3. Conditional likelihood approximations

For the normal-normal model, the CML approach of George and Foster (2000) provides a fast alternative to direct maximization and EM methods, and can be viewed as taking the complete data likelihood (16) and evaluating it at the mode for γ , rather than using the posterior mean, as in the EM algorithm. Let $q_j = \sum_k \gamma_{jk}$ denote the number of non-zero wavelet coefficients at level j . For fixed j , let $D_{j(k)}^2$ denote the sorted values (in decreasing order) of $(D_{jk}/\hat{\sigma})^2$, where $\hat{\sigma}^2$ corresponds to the MAD estimate of σ^2 . The most likely model with q_j non-zero components at level j , $\tilde{\gamma}(q_j)$, has elements $\tilde{\gamma}_{j(k)} = 1$ if $k \leq q_j$ and 0 otherwise.

For each value of q_j , the values of c_j and ω_j that maximize the conditional log-likelihood are

$$\omega_j(q_j) = \frac{\sum_k \tilde{\gamma}_{j(k)}(q_j)}{n_j} = \frac{q_j}{n_j},$$

$$c_j(q_j) = \max \left\{ 0, \frac{\sum_k D_{j(k)}^2 \tilde{\gamma}_{j(k)}(q_j)}{\sum_k \tilde{\gamma}_{j(k)}(q_j) \hat{\sigma}^2} - 1 \right\}. \tag{20}$$

The CML estimators have the same form as the EM MML estimators and are the same when the posterior distribution for γ_{jk} is degenerate at 1 or 0, where we have perfect classification of the observation into noise or signal. For large values of c_j , there is often a very good separation of the signal and noise, resulting in little bias of the CML estimators.

The difference between the MML and the CML estimates will be the most extreme when the posterior mean of γ_{jk} is 0.5 and when c_j is small. The simulation studies described later suggest, however, that the EB estimators using the conditional estimates have better mean-squared error (MSE) performance than earlier shrinkage estimators. Although this approach does not readily extend to the case of t -errors without further approximations, the normal-normal model solutions can be used as starting values for iterative methods such as the Gauss-Seidel or EM algorithms.

3.4. Empirical Bayes estimators

We consider two types of EB estimators: threshold shrinkage (TS) estimators based on model selection and MS estimators based on model averaging. The EB estimators are obtained as posterior means, treating the MML hyperparameter estimates as if they were fixed in advance. The β_{jk} under the common λ model are *a posteriori* conditionally independent,

$$p(\beta_{jk} | \gamma_{jk}, \lambda_{jk}, Y) \sim N \left(\gamma_{jk} \frac{c_j}{1 + c_j} D_{jk}, \frac{\gamma_{jk}}{\lambda_{jk}} \sigma^2 \frac{c_j}{1 + c_j} \right). \tag{21}$$

Expressions for the posterior mean of β_{jk} given $\gamma_{jk} = 1$ under the Cauchy- t model can be obtained from equations in Fan and Berger (1990). Unless noted otherwise, all results below hold for the Cauchy- t model, with $\{c_j/(1 + c_j)\}D_{jk}$ replaced by $E[\beta_{jk} | \gamma_{jk} = 1, Y]$ computed under the Cauchy- t model.

3.4.1. Threshold shrinkage estimators

TS estimators are obtained as posterior means conditionally on the highest posterior probability model or $\gamma = (\gamma_{11}, \dots, \gamma_{JK})$. (As in the variable selection context, a ‘model’ can be identified with γ .) As a consequence of the conditional independence structure, the posterior probability of γ is of the form

$$p(\gamma | D) = \prod_{jk} \pi(\gamma_{jk} = 1 | D_{jk})^{\gamma_{jk}} \{1 - \pi(\gamma_{jk} = 1 | D_{jk})\}^{(1-\gamma_{jk})}, \tag{22}$$

a product of independent Bernoulli probabilities. Because of the product structure, the highest posterior probability model, $\hat{\gamma} = (\hat{\gamma}_{11}, \dots, \hat{\gamma}_{JK})$, is obtained by setting $\hat{\gamma}_{jk} = 1$ if $\pi(\gamma_{jk} = 1 | D_{jk}) \geq 0.5$ and $\hat{\gamma}_{jk} = 0$ otherwise. The posterior mean under $\hat{\gamma}$ for the common λ model is then

$$E(\beta_{jk} | \hat{\gamma}, Y) = \hat{\gamma}_{jk} \frac{\hat{c}_j}{1 + \hat{c}_j} D_{jk}. \tag{23}$$

This TS estimator thresholds the data by setting $\hat{\beta}_{jk} = 0$ whenever $\hat{\gamma}_{jk} = 0$ and then shrinks the remaining coefficients. This is useful for compression problems where dimension reduction and elimination of negligible coefficients is important. Because the inverse DWT (IDWT) is linear, the posterior mean of f under model $\hat{\gamma}$ is obtained by applying the IDWT to the posterior mean of β under model $\hat{\gamma}$.

3.4.2. Multiple-shrinkage estimators

Instead of conditioning on a single model as above, an estimator of β_{jk} which incorporates uncertainty about γ_{jk} is the unconditional posterior mean

$$E(\beta_{jk}|Y) = \pi(\gamma_{jk} = 1|Y) \frac{\hat{c}_j}{1 + \hat{c}_j} D_{jk}. \tag{24}$$

Note that $\pi(\gamma_{jk} = 1|Y) = \pi(\gamma_{jk} = 1|D_{jk})$ is given by expression (8), where the marginals m_0 and m_1 are based on the EB hyperparameter estimates \hat{c}_j and $\hat{\omega}_j$. For specifications such as the normal-normal and the bivariate t , the resulting closed form expressions for $\pi(\gamma_{jk} = 1|Y)$ and the EB MS estimators can be computed quickly once we have the hyperparameter estimates. In contrast, the MS estimator for the Cauchy- t model takes approximately 30 times longer (not including the extra time that is necessary to obtain the hyperparameter estimates) for $n = 1024$. The posterior mean of f is obtained by applying the IDWT to the MS estimator of β .

The estimator (24) is an MS estimator (George, 1986; Clyde *et al.*, 1998) and corresponds to model averaging. Because of the form of the conditional distribution (21), the usual posterior weighted sum of conditional expectations reduces here to the simple form (24) that inserts $\pi(\gamma_{jk} = 1|Y)$ as an additional shrinkage factor. As opposed to the TS estimator, which sets negligible coefficients to 0, the MS estimator provides non-zero estimates for all coefficients for which $\hat{\pi}(\gamma_{jk} = 1|D_{jk}) \neq 0$.

For problems where thresholding rather than shrinkage is desired, it is worth noting that the threshold estimator (23) minimizes the squared error loss for estimation or prediction under model selection (J. O. Berger, personal communication). We provide a proof for the case of wavelet regression. Let \hat{f}_γ denote the posterior mean for f under model γ corresponding to equation (23), and let \hat{f} denote the unconditional posterior mean corresponding to equation (24); \hat{f} is the Bayes rule under squared error loss for the estimation of f . The best single model γ^* for estimating f under posterior expected squared error loss is the model such that \hat{f}_γ is closest to \hat{f} or that minimizes

$$(\hat{f}_\gamma - \hat{f})'(\hat{f}_\gamma - \hat{f}).$$

Because W is orthonormal, this can be equivalently expressed in the wavelet domain by

$$(\hat{f}_\gamma - \hat{f})' W W' (\hat{f}_\gamma - \hat{f}) = \sum_{jk} \{\gamma_{jk} - \pi(\gamma_{jk} = 1|Y)\}^2 E(\beta_{jk}|Y, \gamma_{jk} = 1)^2.$$

The last expression is minimized by the model with $\gamma_{jk}^* = 1$ if $\pi(\gamma_{jk} = 1|Y) \geq 0.5$ and with $\gamma_{jk}^* = 0$ if $\pi(\gamma_{jk} = 1|Y) < 0.5$, which is equivalent to the highest posterior model when the posterior distribution for γ has the form given by equation (22) and is the median probability model as defined by Berger.

Although the MS estimator tends to outperform the TS estimator in terms of the squared error loss, sometimes the improvements are small. Such gains may not compensate for the costs of including additional terms, as might occur, for example, in compression problems.

We shall see in the next section that both equations (23) and (24) appear to offer improved frequentist risk performance over several classical estimators.

4. Simulations

To investigate the practical potential of our EB method, we evaluated the performance of the EB estimators under the normal-normal, the bivariate t and Cauchy- t specifications, using noisy versions of the four test functions ‘blocks’, ‘bumps’, ‘doppler’ and ‘heavisine’ proposed by Donoho and Johnstone (1994). In every case, the functions were scaled so that the signal-to-noise ratio was 7 (the ratio of the standard deviation of the function values f_i to the standard deviation of the noise). We compared this performance with that of several existing shrinkage strategies: HARD, hard thresholding with the universal rule, SOFT, soft thresholding with the universal rule, and SURE, the ‘SureShrink’ adaptive shrinkage rule as implemented in S+Wavelets (Donoho and Johnstone, 1994, 1995). The wavelet bases were chosen to match those used by Donoho and Johnstone (1995). To assess robustness, we considered both normal and heavy-tailed distributions for the noise. The performance was evaluated by the average MSE over 100 simulations for each problem,

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

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

4.1. Normal errors

For our first set of simulations, we repeatedly added $n = 1024$ independent $N(0, 1)$ errors to each of the four test functions. For the normal-normal specification, we first compared the TS estimator in equation (23) and the MS estimator in equation (24) with HARD thresholding. We also compared the TS and MS estimators using EB hyperparameter estimates based on the exact E MML and on the approximate A CML discussed in Section 3.

Fig. 2 presents box plots of MSEs of these five estimators, and shows that all four EB estimators are superior to HARD in this setting. Of the EB estimators, MS is superior to TS for both exact and approximate hyperparameter estimates. However, estimates based on the exact marginal likelihood are at most slightly better than those based on the rapidly computable approximate conditional likelihood.

Restricting attention now to MS estimators using exact marginal likelihood hyperparameter estimates, Fig. 3 compares the EB estimators for the normal-normal, the bivariate t_5 (with 5 degrees of freedom) and the Cauchy- t_5 specifications, labelled EB N, EB T5 and EB C5, with the SURE, HARD and SOFT estimators. The EB N estimator exhibits excellent MSE performance compared with the other methods. However, EB T5 performs nearly as well, demonstrating only a small loss of efficiency, and is equal to or better than HARD, SURE or SOFT.

Fig. 4 shows the distribution of the MLEs of $\sqrt{c_j}$ and ω_j under the normal model for the four test functions. The variation in the estimates across levels is striking, revealing strong decay in both ω_j and c_j but different rates across functions. Although such decay might be roughly expected by using a fixed hyperparameter Bayes set-up with subjective prior inputs, it is difficult to prespecify the appropriate magnitude and rate of decay. The MSE performance comparisons with the fixed hyperparameter Bayes estimators of Clyde *et al.* (1998) again favoured our EB estimators.

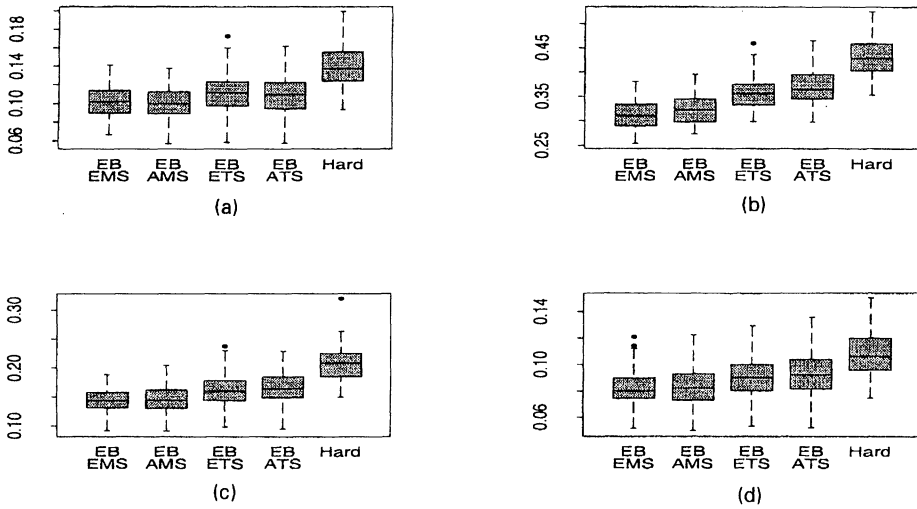


Fig. 2. MSEs for the EB TS and MS estimators using exact E and approximate A hyperparameter estimates and the HARD estimator: (a) blocks; (b) bumps; (c) doppler; (d) heavisine

4.2. Student *t*-errors in the wavelet domain

For our second set of simulations, we repeatedly added $n = 1024$ independent t_5 -errors to the true wavelet coefficients for each of the four test functions. Fig. 5 compares MSEs for the MS EB estimators EB N, EB T5 and EB C5 with those for the SURE, HARD and SOFT estimators.

Comparing Fig. 5 with Fig. 3 reveals that the MSE performance of all the estimators tends to be worse under heavy-tailed error distributions (recall that the signal-to-noise ratios were the same across simulations), and the number of ‘MSE outliers’ increases dramatically. Of the three EB estimators, the performance of EB N worsened the most, indicating a stronger sensitivity to extreme errors. Compared with all the estimators, EB T5 and EB C5 performed

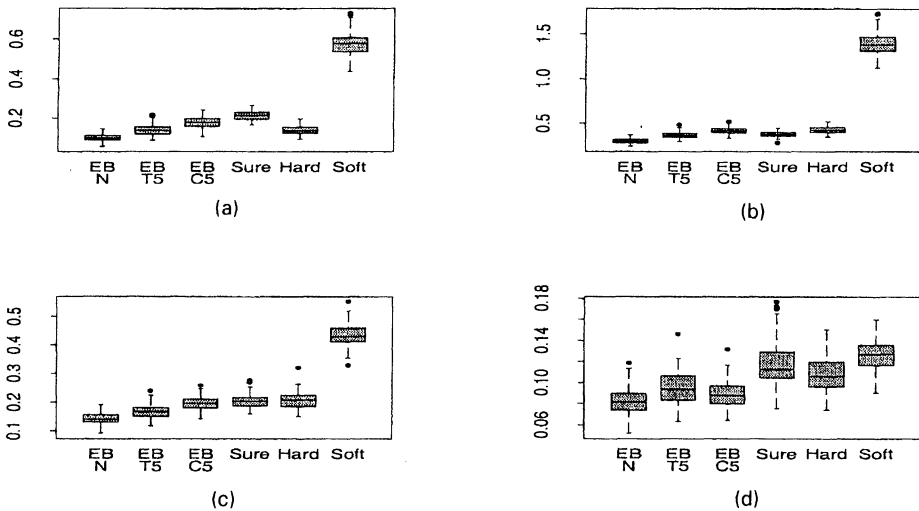


Fig. 3. MSE comparisons for data simulated with normal errors (the EB estimators are all based on MS): (a) blocks; (b) bumps; (c) doppler; (d) heavisine

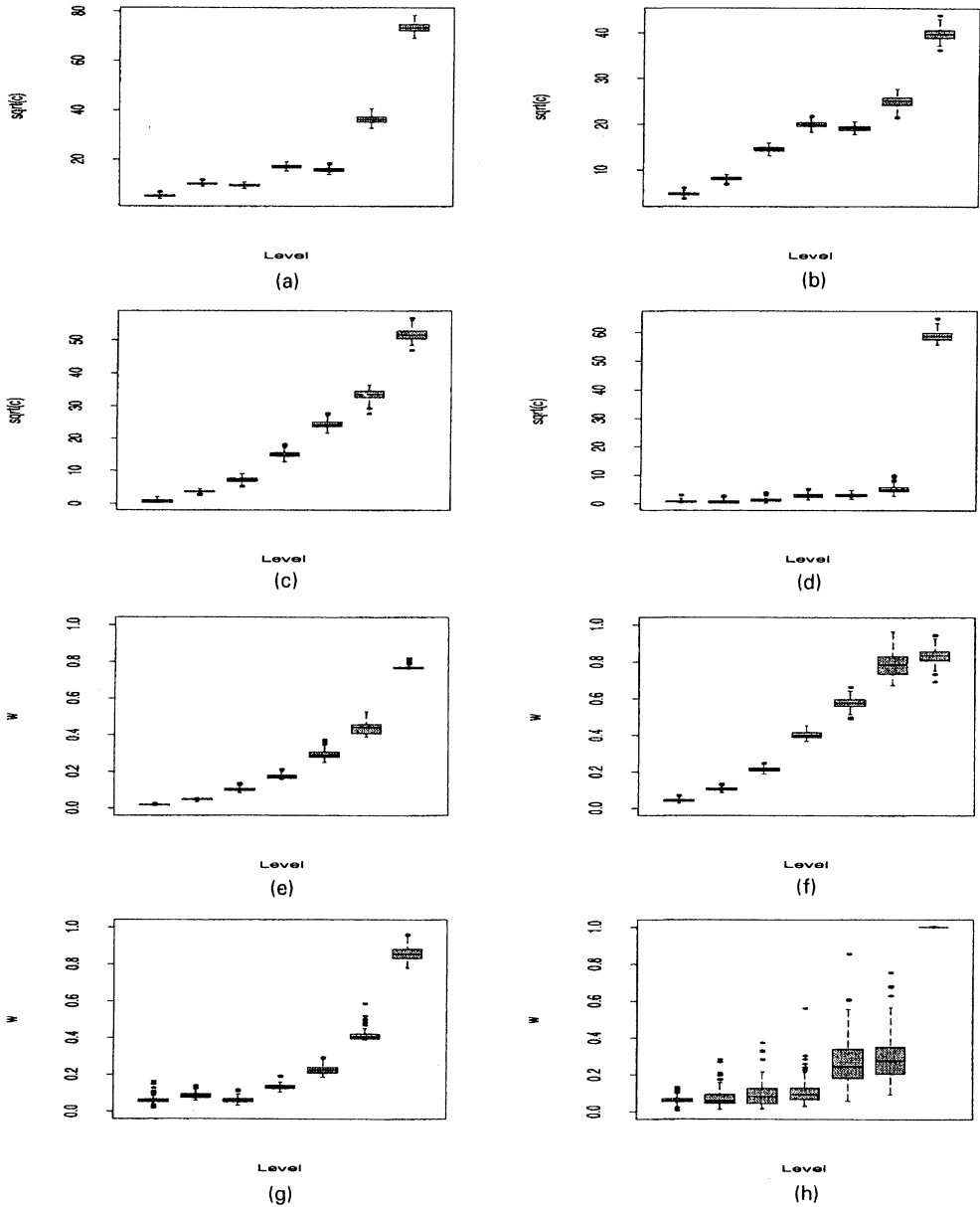


Fig. 4. EB estimators of $\sqrt{c_j}$ and w_j for data simulated with normal errors: (a), (e) blocks; (b), (f) bumps; (c), (g) doppler; (d), (h) heavisine

well, offering substantial gains in efficiency for all functions; the one exception is SOFT in heavisine. The average performances of EB T5 and EB C5 are similar, although computations for EB T5 can be carried out in a fraction of the time for EB C5.

We also compared the results with the robust smoother cleaner (RSC) transform developed by Bruce *et al.* (1994), which is a fast wavelet decomposition that prevents outliers and outlier patches from leaking into the wavelet coefficients at coarse levels. The MSE performance of

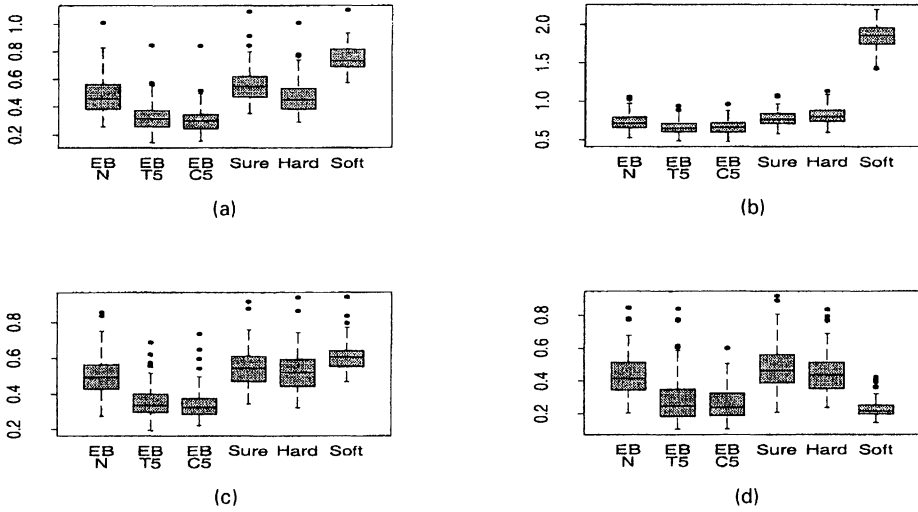


Fig. 5. MSE comparisons for data simulated with t_5 -errors in the wavelet domain: (a) blocks; (b) bumps; (c) doppler; (d) heavisine

the RSC transform in blocks, bumps and doppler is much poorer than any of the other methods with average MSEs of 5.37, 47.13 and 8.7 respectively (the box plots are not included because of the scale). For heavisine, the average MSE is 0.25, which is comparable with that of EB C5. Visual inspections of reconstructions suggested that although the RSC transform effectively removed outliers it often mistook signals for outliers, resulting in oversmoothing, particularly in bumps.

4.3. Student t -errors in the data domain

In our final set of simulations, we assessed the performance of EB N, EB T5 and EB C5 under the more realistic assumption of independent t -errors in the data domain. To do this we repeatedly added $n = 1024$ independent t_5 -errors directly to each of the four test functions, rather than to the wavelet coefficients as in the previous section, but still maintaining a signal-to-noise ratio of 7.

Here, the wavelet transformation induces dependent, though uncorrelated, errors in the wavelet domain that are linear combinations of t random variables. Thus, the wavelet domain errors are more normally distributed, especially at the coarser levels, where the linear combinations typically involve more terms. Although a hierarchical Bayesian analysis under the exact wavelet domain error distribution is computationally impractical, one might consider a modification of the EB T estimator to allow for increasing degrees of freedom at the coarser levels to account for the tendency towards normality. To explore this approach, we considered an additional estimator, labelled EB TM (where M is for ‘matched’). The degrees of freedom for EB TM were matched to the error process by applying the wavelet transform to independent t_5 -errors. For each level, we selected the degrees of freedom that maximized the marginal distribution of the data, and then repeated this procedure 100 times. The average degrees of freedom for the seven levels were 5, 6, 6, 7, 7, 7 and 8 (after rounding) for the ‘haar’, ‘d6’ and ‘s16’ wavelet transforms, suggesting that, despite the averaging due to the DWT, heavy tails can still be an issue at the coarsest levels.

Fig. 6 compares the MSEs of the four EB estimators EB N, EB T5, EB TM and EB C5

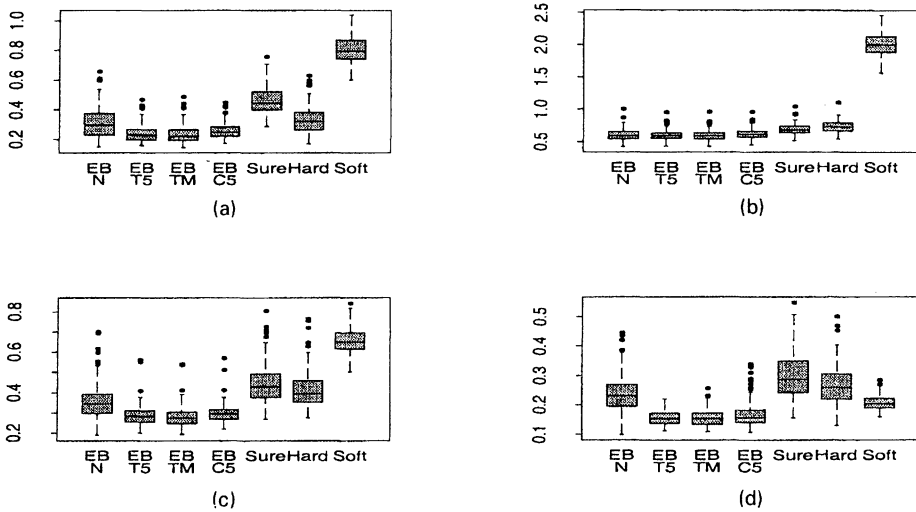


Fig. 6. MSE comparisons for data simulated with t_5 -errors in the data domain: (a) blocks; (b) bumps; (c) doppler; (d) heavisine

with those for the SURE, HARD and SOFT estimators in this setting. A comparison with Figs 3 and 5 shows that the MSEs here are somewhat smaller than under t -errors but still much larger than under normal errors. Because of the wavelet transform, extreme values in the time domain have a smaller effect in the wavelet domain, but apparently the tendency towards normality in the wavelet domain is not especially pronounced. All three EB estimators based on heavy-tailed error distributions exhibit substantial gains in efficiency over the standard methods, with EB TM performing only marginally better than EB T5 or EB C5. With the exception of the SOFT estimator for the heavisine function, EB N offers an improvement over SURE, HARD and SOFT. As in the previous section, the RSC transform performed poorly, obtaining larger MSEs than all other methods for blocks, bumps and doppler (the average MSEs are 5.36, 47.46 and 8.73 respectively), with the exception of heavisine, where it was comparable with EB N, but still not as good as EB T5, EB TM or EB C5.

5. GLINT example

To observe their performance on a real data set, we applied our EB estimators to the GLINT data of Bruce and Gao (1994). These data are radar reflections or glints measured in degrees for a rotating airplane model and are subject to errors that can be as large as 150° in absolute value. The true signal is supposed to be a low frequency oscillation about zero with possible level shifts, but the glint spikes are highly non-Gaussian (Bruce and Gao, 1994).

We used an estimate of $\sigma = 29$ so that 3 standard errors contain almost all the glint spikes in the original signal, based on information from Bruce and Gao (1994). In Fig. 7, we compare the EB estimates under the normal-normal and bivariate t models with hard thresholding and the RSC wavelet transform of Bruce *et al.* (1994). In this example, shrinkage via hard thresholding has not removed the glint spikes and Bruce and Gao (1994) viewed the resulting estimate of the signal as quite poor. The RSC wavelet transform was designed to remove outlier patches from the wavelet decomposition and results in a much improved, smoother reconstruction than HARD thresholding (Bruce and Gao, 1994).

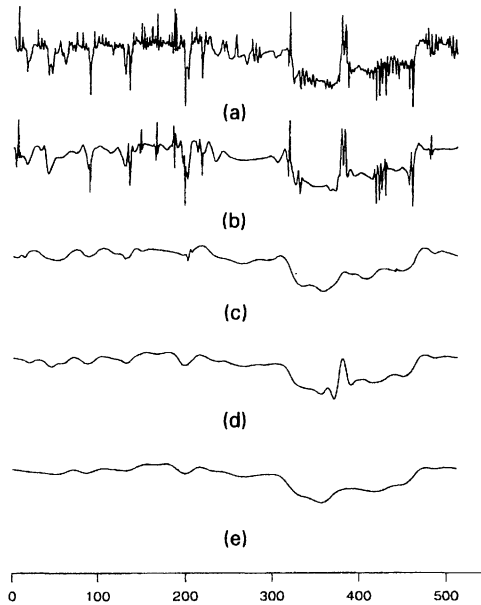


Fig. 7. Shrinkage estimates for the GLINT data: (a) original data; (b) hard thresholding; (c) RSC transform; (d) normal EB shrinkage; (e) Student t EB shrinkage

Whereas the EB normal estimate may still retain a spurious glint spike around $t = 400$, the EB shrinkage estimates under the bivariate t model with 3 degrees of freedom have effectively removed the outliers or non-Gaussian glint spikes in the original signal. Unlike the RSC wavelet transform, the EB approach is completely model based and can incorporate prior information about the process.

6. Discussion

We have proposed robust empirical Bayes methods for wavelet estimation. Embedding the wavelet set-up in a hierarchical normal scale mixture model, we obtained MLEs of the unknown hyperparameters for each wavelet level, as well as an estimate of σ^2 , using data from all levels. 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 parameter Bayes estimators. The EB N model provides an excellent adaptive default prior for most cases, whereas the EB T and EB TM approaches provide a balance between robustness when heavy-tailed errors are a concern and computational ease over independent scale models in either the wavelet or original domain.

An explanation for the improved performance of the EB estimators is that they allow for a wide variation in hyperparameter values across different wavelet levels, yielding flexible shrinkage patterns. The Bayesian hierarchical model can also incorporate prior uncertainty about the degrees of freedom, which leads to a posterior distribution for the degrees of freedom at each level. This avenue of research may lead to additional improvements over the current EB T or EB TM estimators in the case of heavy-tailed errors in the data domain.

It would be straightforward to use our EB prior distributions and resulting posterior distributions to estimate the posterior median, as in Abramovich *et al.* (1998), for robust L_1

EB estimators. Analogously to Johnstone and Silverman (1997), the EB methods can be adapted to the case of correlated noise, by replacing the MLE of σ with a level-dependent MLE $\hat{\sigma}_j$. Because sample sizes at coarser levels are small, there may be substantial uncertainty in the EB estimators of c_j , w_j and σ_j . In cases with small signal-to-noise ratios, improved estimation can be achieved by using Gibbs sampling (Clyde *et al.*, 1998) combined with Rao–Blackwellized estimators. If the computational issues for this approach could be simplified, this would be a promising competitor to our methods and would provide improved estimates over a naïve EB approach that ignores uncertainty in the hyperparameter estimates.

Acknowledgements

This work was supported by National Science Foundation grants DMS-96.26135, DMS-97.33013 and DMS-98.03756, and Texas Higher Education Coordinating Board Advanced Research Program grants 003658.130, 003658.452 and 003658.690.

References

- Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc. B*, **60**, 725–749.
- Andrews, D. F. and Mallows, C. L. (1974) Scale mixtures of normal distributions. *J. R. Statist. Soc. B*, **36**, 99–102.
- Bruce, A., Donoho, D., Gao, H.-Y. and Martin, D. (1994) Denoising and robust nonlinear wavelet analysis. *SPIE Proc.*, **2242**, 325–336.
- Bruce, A. and Gao, H.-Y. (1994) *S+Wavelets, Users Manual*. Seattle: StatSci.
- Chipman, H., Kolaczyk, E. and McCulloch, R. (1997) Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Ass.*, **92**, 1413–1421.
- Clyde, M., Parmigiani, G. and 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 from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, **39**, 1–38.
- Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.
- (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.
- Donoho, D., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia (with discussion)? *J. R. Statist. Soc. B*, **57**, 301–369.
- Fan, T.-H. and Berger, J. O. (1990) Exact convolutions of t distributions with applications to Bayesian inference for a normal mean with t prior distributions. *J. Statist. Computn Simuln*, **36**, 209–228.
- George, E. I. (1986) Minimax multiple shrinkage estimation. *Ann. Statist.*, **14**, 188–205.
- George, E. I. and Foster, D. P. (2000) Calibration and empirical Bayes variable selection. *Biometrika*, to be published.
- Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc. B*, **59**, 319–351.
- (1998) Empirical Bayes approaches to mixture problems and wavelet regression. *Technical Report*. University of Bristol, Bristol.
- O’Hagan, A. (1979) On outlier rejection phenomena in Bayes inference. *J. R. Statist. Soc. B*, **41**, 358–367.
- (1988) Modelling with heavy tails. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 345–359. Oxford: Clarendon.
- Thisted, R. A. (1988) *Elements of Statistical Computing*. New York: Chapman and Hall.
- West, M. (1984) Outlier models and prior distributions in Bayesian linear regression. *J. R. Statist. Soc. B*, **46**, 431–439.
- (1987) On scale mixtures of normal distributions. *Biometrika*, **74**, 646–648.