Statistica Sinica **10**(2000), 109-128

WAVELET ESTIMATION USING BAYESIAN BASIS SELECTION AND BASIS AVERAGING

Robert Kohn, J. S. Marron* and Paul Yau*

*University of New South Wales and * University of North Carolina, Chapel Hill*

Abstract: Wavelet shrinkage methods are widely recognized as a useful tool for nonparametric regression and signal recovery, while Bayesian approaches to choosing the shrinkage method in wavelet smoothing are known to be effective. In this paper we extend the Bayesian methodology to include choice among wavelet bases (and the Fourier basis), and averaging of the regression function estimates over different bases. This results in improved function estimates.

Key words and phrases: Empirical Bayes, Fourier series, nonparametric regression.

1. Introduction

Wavelet methods are an interesting approach to nonparametric regression because they provide a new type of orthogonal series estimator which handles spatially varying smoothness more efficiently than classical linear methods. Mathematical definitions for wavelet methods are given in Section 2, with detailed discussions given in Donoho and Johnstone (1995) and Donoho, Johnstone, Kerkyacherian and Picard (1995).

The first wavelet nonparametric curve estimators use simple methods of shrinkage, but many alternative shrinkage functions are possible. In particular, Chipman, Kolaczyk and McCulloch (1997) and Clyde, Parmigiani and Vidakovic (1997) propose Bayesian shrinkage methods and show that they compare favorably to the earlier ones.

While wavelet bases are impressive in terms of adapting to a wide range of regression functions, no single wavelet basis is uniformly best. This suggests that improved performance can be obtained by automatic choice of wavelet basis or averaging over estimates for different bases. Section 4 extends the Bayesian frameworks of Chipman, Kolaczyk and McCulloch (1997) and Clyde, Parmigiani and Vidakovic (1997) to include data driven choice of the basis and weighted averaging over the estimates produced for different bases, using the posterior probabilities of the bases as weights. The resulting estimators are shown in Section 5 to perform well in comparison to single basis methods. In addition to providing important new types of wavelet estimators, our work shows that basis choice is a serious issue, and it is not enough to rely on the good overall performance properties of the best wavelet bases.

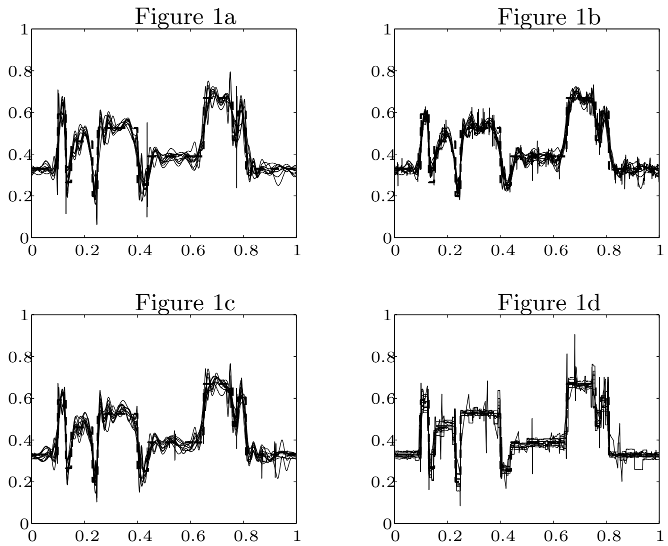

Figure 1. Target regression curve, shown as the heavy dashed line, with 10 simulated estimates (from the Donoho and Johnstone Blocks example with high noise). Figures 1a and 1b are the simple hard and soft thresholded estimators. Figure 1c is the Chipman, et al. estimator. Figures 1a–1c use the Symmlet 8 basis. Figure 1d is the empirical Bayes basis selection estimator.

Figure 1 shows how basis selection improves on the single basis estimators when the regression function is piecewise constant. Figures 1a–1c give the function estimates obtained with hard and soft thresholding and the empirical Bayes estimator of Chipman, Kolaczyk and McCulloch (1997); all three estimators use the Symmlet 8 basis. The hard thresholding method has spurious spikes which disappear for soft thresholding, but at the cost of excessive rounding of corners. The empirical Bayes estimator gives noticeably better performance, but there is still a disappointing rounding of the corners. The relatively poor performance of the single basis estimators is more a weakness of the basis than of the shrinkage method because the Symmlet 8 basis functions are smooth and it is difficult to reconstruct the sharp corners from high noise data. However, the Haar basis of step functions gives a much better reconstruction of this signal, although this basis is inappropriate for smooth signals. Figure 1d shows that our basis selection method improves on the estimators using just the Symmlet 8 basis by choosing from a suite of bases, which includes the Haar. In this example, the basis selection method usually chooses the Haar basis, which results in much better reconstruction. There is some chance of mis-estimating the correct basis, which is shown in Figure 1d by one estimate based on the Daubechies 4 basis. The reconstruction is not perfect because the noise level is moderately high here, but it is much better than for the single basis methods.

2. Wavelet Basics

Wavelet shrinkage is considered in the context of nonparametric regression, where it is desired to recover the smooth curve $m(x)$ from noisy observations

$$y_i = m(x_i) + \varepsilon_i,\ \ i = 1, \ldots, n,$$

with $E(\varepsilon_i) = 0$. We make the following technical assumptions:
A1 The x_i are equally spaced with $x_i = (i - 1/2)/n$.
A2 The sample size n is a power of two with $n = 2^k$.
A3 The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent $N(0, \sigma^2)$.

These assumptions can be weakened, although there is some cost in doing so in terms of insight, computational complexity, and tractability of analysis.

Orthogonal series methods for recovering the signal vector $\boldsymbol{m} = (m(x_1), \ldots, m(x_n))^t$ from the data vector $\boldsymbol{y} = (y_1, \ldots, y_n)^t$ are understood easily via spectral representation. That is, write $\boldsymbol{m}$ as the linear combination

$$\boldsymbol{m} = \sum_{i=1}^{n} \beta_i \boldsymbol{\psi}_i = \boldsymbol{\Psi\beta} \tag{2.1}$$

of vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ that form an orthonormal basis of n-dimensional Euclidean space. The matrix $\boldsymbol{\Psi}$ has ith column $\boldsymbol{\psi}_i$ and is orthonormal, and $\boldsymbol{\beta} = \boldsymbol{\Psi}^t\boldsymbol{m}$. Let $\boldsymbol{w} = \boldsymbol{\Psi}^t\boldsymbol{y}$ and $\boldsymbol{e} = \boldsymbol{\Psi\varepsilon}$. Then $\boldsymbol{w} = \boldsymbol{\beta} + \boldsymbol{e}$, with $\boldsymbol{e} \sim N(0, \sigma^2 I)$.

The transformed observation vector $\boldsymbol{w}$ is an unbiassed estimator of $\boldsymbol{\beta}$ and is useful for estimating $\boldsymbol{\beta}$ when most of the power of the signal, as measured by the sum of squares $\boldsymbol{m}^t\boldsymbol{m}$, is captured by a few of the β_i. In signal processing terms this is called good signal compression. For example, this happens using the Fourier basis when the signal $\boldsymbol{m}$ is smooth and periodic. In this case, most of the β_i in (2.1) can be set to 0; this entails damping most of the noise while retaining most of the signal, and thus results in good estimates of $\boldsymbol{m}$.

More generally, we consider the class of shrinkage estimators of $\boldsymbol{m}$,

$$\widehat{\boldsymbol{m}} = \sum_{i=1}^{n} \widehat{\beta}_i \boldsymbol{\psi}_i, \tag{2.2}$$

where $\widehat{\beta}_i = \eta_i(w_i)$ and $\eta_i(w_i)$ shrinks w_i towards 0 or even sets $\widehat{\beta}_i$ to 0 for small w_i.

The large amount of current interest in wavelets is due to their ability to do effective signal compression for both smooth and periodic signals, as well as for signals that are smooth in most locations, but have some points of nonsmoothness. Wavelet bases effectively compress signals with smoothness varying by location because they are adaptive in terms of both scale (this concept is the same as frequency in Fourier analysis) and location. Because of this dual ability to adapt, wavelet bases are most conveniently represented using the following double indexing notation: $j(=j_0,\ldots,\log_2(n/2))$ indexes scale, and $k(=0,\ldots,2^j-1)$ indexes location. The parameter j_0 is chosen as small as possible for the given basis. This index system has a simple correspondence to the indices $i=2^{j_0}+1,\ldots,n$ used above via

$$i \leftrightarrow (j,k) \text{ as } i = 2^j + k + 1.$$

These two index systems are used interchangeably in the rest of the paper, with the choice of indexing made in terms of convenience. While the Fourier basis does not have a structure requiring double indexing, it is useful to organize it in this way for the empirical Bayes estimator.

It is also useful to write the wavelet bases in terms of father wavelets $\boldsymbol{\psi}_i$, which are the $i=1,\ldots,2^{j_0}$ elements of the basis, and the mother wavelets $\boldsymbol{\psi}_{j,k}$, which correspond to $i=2^{j_0}+1,\ldots,n$. The wavelet version of the spectral representation (2.1) becomes

$$\boldsymbol{m} = \sum_{i=1}^{2^{j_0}} \beta_i \boldsymbol{\psi}_i + \sum_{j=j_0}^{\log_2(\frac{n}{2})} \sum_{k=0}^{2^j-1} \beta_{j,k} \boldsymbol{\psi}_{j,k}.$$

The shrinkage estimator (2.2) becomes

$$\boldsymbol{m} = \sum_{i=1}^{2^{j_0}} \widehat{\boldsymbol{\beta}}_i \boldsymbol{\psi}_i + \sum_{j=j_0}^{\log_2(\frac{n}{2})} \sum_{k=0}^{2^j-1} \widehat{\beta}_{j,k} \boldsymbol{\psi}_{j,k},$$

where $\hat{\beta}_i = w_i$ and $\hat{\beta}_{j,k} = \eta_j(w_{j,k})$. Note that the shrinkage functions $\eta_j, j \geq j_0$, depend on the level j, but not on the location k. The father terms, $w_i, i = 1,\ldots,2^{j_0}$, are typically left unthresholded because they represent low frequency terms that usually contain important components of the signal. Thus the parameter j_0 controls how many terms are of this type. The hard thresholding estimator studied by Donoho and Johnstone (1994) has a shrinkage function that can be expressed as

$$\eta_{H,j}(w) = \begin{cases} w \text{ for } |w| \geq \lambda\sigma \\ 0 \text{ for } |w| < \lambda\sigma \end{cases}, \text{ for } j \geq j_0, \tag{2.3}$$

for a given value of the threshold λ. The function η_H is plotted in Figure 2 and shows that η_H zeros small values of w and leaves large values of w untouched.

The soft thresholding estimator proposed by Donoho and Johnstone (1994) has the shrinkage function

$$\eta_{S,j}(w) = \begin{cases} sgn(w)\cdot(|w|-\lambda\sigma) & \text{for } |w| \geq \lambda\sigma \\ 0 & \text{for } |w| < \lambda\sigma \end{cases}, \text{ for } j \geq j_0, \qquad (2.4)$$

and is again independent of k. The function η_S is plotted in Figure 2 and corresponds to moving the w terms $\lambda\sigma$ units towards the origin. We write the hard and soft thresholding estimators as $\widehat{\boldsymbol{m}}_H$ and $\widehat{\boldsymbol{m}}_S$.

The hard and soft thresholding estimators considered in all examples in this paper have $j_0 = 5$, with $n = 1024$. This choice of j_0 gives good overall performance in our simulations.

The error standard deviation σ is usually unknown, but a good estimator of σ is obtained by using a robust scale estimate based on the highest frequency terms $w_{j,k}$, as suggested by Donoho and Johnstone (1994). Following their suggestion, our scale estimate is based on the median absolute deviation from the median (normalized by dividing by the corresponding standard normal term), which we write as $\widehat{\sigma}$.

We consider two choices of the threshold value λ. Donoho and Johnstone (1994) proposed the denoising threshold,

$$\lambda_D = \sqrt{2\ln n}, \qquad (2.5)$$

which corresponds to the largest size of Gaussian pure noise terms, and is well suited for use with η_H. For η_S a somewhat smaller threshold value is more appropriate, which results in more terms in the model to counter the shrinkage effect. Donoho and Johnstone (1994) proposed a minimax optimal value, called λ_{MO}.

Detailed analyses, and additional insights, about hard and soft thresholding are in Marron, Adak, Johnstone, Neumann and Patil (1998). A number of variations on these thresholding schemes are suggested in the literature, but the choice among them is unclear.

3. Bayesian Shrinkage

3.1. Introduction

We consider two Bayesian estimators of the vector $\boldsymbol{m}$ based on wavelets. The first is an empirical Bayes estimator in which the prior is determined from the data. The second is what we call a "calibrated" Bayes estimator because the parameters of the prior are determined by calibrating the posterior distribution against some variable selection criterion. Our empirical Bayes approach is almost identical to that of Chipman, Kolaczyk and McCulloch (1997), while the calibrated prior approach is based on the work of George and Foster (1997) and is similar to that used by Clyde, Parmigiani and Vidakovic (1997).

For both approaches, let $\gamma_i = 1$ if the basis vector $\boldsymbol{\psi}_i$ is included in the model (2.1) and let $\gamma_i = 0$ if it is not. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$. For a given $\boldsymbol{\gamma}$ we have from (2.1) that $\boldsymbol{m}$ is a linear function of those vectors $\boldsymbol{\psi}_i$ for which $\gamma_i = 1$. Let $\pi_i = p(\gamma_i = 1)$ be the prior probability that $\boldsymbol{\psi}_i$ is included in the model and let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$.

3.2. Estimation using empirical (data based) priors

For $\gamma_i = 1$ the prior for β_i is $N(0, c_i^2)$. For wavelets we use the double index notation and choose $c_i = c_{jk}$ and $\pi_i = \pi_{jk}$ to depend on the scale j only, so that

$$\begin{aligned} c_j &= ||\boldsymbol{\psi}_{j,k}||_1 \times \max\nolimits_{i=1,\ldots,n} |y_i|/3, \\ \pi_j &= \#\{k : |y_{j,k}| > \sqrt{2\log n}\}/2^j, \end{aligned} \tag{3.1}$$

where $||\boldsymbol{\psi}_{j,k}||_1$ is the L_1 norm of the basis vectors at scale j; we note that all $\boldsymbol{\psi}_{j,k}$ have the same norm for a given level j. This prior is essentially that of Chipman, Kolaczyk and McCulloch (1997) and a motivation for the choice of c_j and π_j in (3.1) is given in Appendix 1. We use the posterior mean of β_i as its estimate and note the w_i are independent given σ^2; this means that $E(\beta_i|\boldsymbol{y}, \sigma^2, \pi_i, c_i) = E(\beta_i|w_i, \sigma^2, \pi_i, c_i)$. We can show that

$$E(\beta_i|w_i, \sigma^2, \pi_i, c_i) = \eta_{EB,i}(w_i), \tag{3.2}$$

where

$$\eta_{EB,i}(w) = \frac{\pi_i \sigma c_i^2 w}{\pi_i \sigma (c_i^2 + \sigma^2) + (1 - \pi_i)(c_i^2 + \sigma^2)^{3/2} \exp\left(-\frac{w^2}{2\sigma^2}\left(\frac{c_i^2}{c_i^2+\sigma^2}\right)\right)}. \tag{3.3}$$

We use the notation $\eta_{EB,i}$ to signify that the empirical Bayes estimator works in the same way as the threshold functions (2.3) and (2.4). This is demonstrated in Figure 2 which plots the shrinkage function (3.3) for several values of the hyperparameter c_i. For moderate values of c_i, the function $\eta_{EB,i}$ behaves like the hard thresholding function η_H; it essentially zeros out small coefficients and keeps large coefficients unchanged. The key difference between η_{EB} and η_H occurs at intermediate values of w; in this range η_{EB} is a smoother function of w, i.e., small changes in w create small changes in $\eta_{EB,i}(w)$. An attractive feature of the Bayesian approach is that it provides a natural set of choices for the shrinkage function through the choice of prior.

For the Fourier basis we found it convenient to estimate the c_j and π_j by grouping as for the wavelet bases, even though the Fourier basis does not have a structure that requires double indexing. We are investigating other methods of obtaining an empirical Bayes prior for the Fourier basis that is also suitable for basis selection and basis averaging.

The empirical Bayes estimator (3.3) assumes that the error variance σ^2 is known. In practice it is usually unknown and needs to be estimated from the data. We follow Chipman, Kolaczyk and McCulloch (1997) and plug in an estimate $\widehat{\sigma}^2$ for σ^2 as we did for the hard and soft thresholding estimators. We write this estimator as $\widehat{\boldsymbol{m}}_{EB}$.

Alternatively, we can place a prior on σ^2, for example the noninformative prior $p(\sigma^2) \propto 1/\sigma^2$, and obtain the posterior mean estimate of $\boldsymbol{\beta}$ with σ^2 integrated out. However, it is now necessary to use a Markov chain Monte Carlo method (see Clyde, Parmigiani and Vidakovic) to estimate the posterior means because a closed form expression is not available for them. The disadvantage of using Markov chain Monte Carlo is that it is appreciably slower than using the plug-in approach.

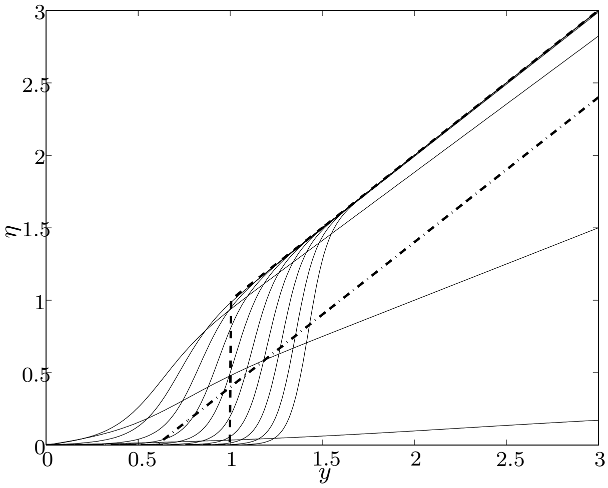

Figure 2. Shrinkakge functions for $\sigma = 1$, hard η_H with $\lambda = 1$ shown as the thick black dashed line, soft η_S with $\lambda = \lambda_{MO}/\lambda_H$ (for $n = 1204$, to make visually comparable with η_H) shown as the thick dot-dashed line, and the empirical Bayes $\eta_{EB,i}$ for $c_i = 4^l$, $l = -1, 0, 1, \ldots, 10$, $\pi_i = 0.5$ and $\sigma = 1$ shown as the thin lines.

3.3. Estimation using calibrated priors

George and Foster (1997) emphasize the importance of the choice of priors in variable selection in regression. They show how the choice of unknown parameters in the prior can be calibrated against frequentist variable selection criteria and F-statistics. We follow their approach, but use posterior odds ratios to calibrate the parameters of the prior. A related approach based on F-values is given by Clyde, Parmigiani and Vidakovic (1997).

For $\gamma_i = 1$ we take the prior for $\beta_i \sim N(0, \sigma^2 c_i^2)$. The posterior odds ratio that $\gamma_i = 1$ vs $\gamma_i = 0$, for given values of σ^2, c_i and π_i, is

$$O_i = \frac{p(\gamma_i = 1 | w_i, \sigma^2, c_i, \pi_i)}{p(\gamma_i = 0 | w_i, \sigma^2, c_i, \pi_i)};$$

O_i can be expressed as

$$O_i = \frac{1}{\sqrt{1+c_i^2}} \exp\left(\frac{w_i^2 c_i^2}{2\sigma^2(1+c_i^2)}\right) \frac{\pi_i}{1-\pi_i}.$$

A Bayesian equivalent of hard thresholding is to select $\gamma_i = 1$ if $O_i > 1$, and to set $\gamma_i = 0$ otherwise. With some algebra we obtain that $O_i > 1$ is equivalent to

$$\left(\frac{w_i}{\sigma}\right)^2 > \frac{1+c_i^2}{c_i^2}\left\{\log(1+c_i^2) + 2\log\left(\frac{\pi_i}{1-\pi_i}\right)\right\}.$$

One way of choosing c_i and π_i is to equate the Bayesian decision to hard thresholding, in which case we take

$$\left(\frac{w_i}{\sigma}\right)^2 > \frac{1+c_i^2}{c_i^2}\left\{\log(1+c_i^2) + 2\log\left(\frac{\pi_i}{1-\pi_i}\right)\right\} = 2\log n. \tag{3.4}$$

Because c_i and π_i cannot be determined simultaneously from (3.4), we take $\pi_i = 0.5$ without loss of generality, giving $c_i \approx n$.

We now give a more general strategy for choosing the prior. Take $\pi_i = 0.5$ and $c_i = c$ for all i. Let

$$Q = \prod_{i=1}^{n} p(O_i < 1 | \sigma, c).$$

Then Q is the probability that $O_i < 1$ simultaneously for all i, that is, Q is the probability that we take the regression function $\boldsymbol{m}$ identically equal to zero under the Bayesian decision rule. It is not difficult to show that if $\boldsymbol{m}$ is identically equal to zero then

$$Q = \{2\Phi(t_c) - 1\}^n, \tag{3.5}$$

where $t_c = \frac{1+c^2}{c^2}\log(1+c^2)$ and Φ is the standard normal cdf.

If we take $Q = 0.8$ then $t_c \approx 2\log n$ which gives the hard thresholding rule. However, we can set Q at any level and solve for c. Experimentally, we found that setting $Q = 0.9$ or even 0.98 works quite well. Let

$$c_Q = \exp\left(0.5 \times \Phi^{-1}\{(Q^{1/n} + 1)/2\}\right)$$

be the (approximate) solution of (3.5) for a given Q. Using $c = c_Q$ and $\pi_i = 0.5$, the posterior mean of β_i can be obtained as in Section 3.2. For a given value of Q we write the calibrated Bayes estimator of $\boldsymbol{m}$ as

$$\widehat{\boldsymbol{m}}_{CB}^{Q} = E(\boldsymbol{m}|\boldsymbol{y}, c_Q, \sigma^2).$$

If the error variance σ^2 is unknown, then a plug-in estimate is used as in Section 3.2.

4. Basis Selection and Basis Averaging

In addition to giving natural shrinkage functions, as demonstrated in the last section, the Bayesian approach provides a simple method for choosing between different wavelet bases or averaging the curve estimate over different bases. Here we consider a choice among the Symmlet 8, Daubechies 4, Haar and Fourier bases.

Let B denote a generic basis and assign it prior probability $p(B)$, taken to be uniform over these four bases (i.e., $p(B) = \frac{1}{4}$) in all examples here. The resulting posterior probability is

$$p(B|\boldsymbol{y}, \sigma^2) \propto p(\boldsymbol{w}|\sigma^2, B)p(B)$$

and

$$p(\boldsymbol{w}|\sigma^2, B) = \prod_{i=1}^{n} p(w_i|\sigma^2, B) = \exp\left(\sum_{i=1}^{n} \ln p(w_i|\sigma^2, B\right), \tag{4.1}$$

because the w_i are independent conditional on σ^2. The expression on the right in (4.1) is computationally more stable than the expression in the middle. Each of the densities $p(w_i|\sigma^2, B)$ is evaluated directly using

$$p(w_i|\sigma^2, B) = p(w_i|\sigma^2, B, \gamma_i = 0)(1 - \pi_i) + p(w_i|\sigma^2, B, \gamma_i = 1)\pi_i.$$

We assume that if σ^2 is unknown, then its estimate is plugged in. If a prior distribution is put on the unknown σ^2 then it is necessary to use a Markov chain Monte Carlo method such as Gibbs sampling to compute $p(\boldsymbol{w}|B)$ with σ^2 integrated out because the w_i are independent only conditional on σ^2. In our experiments, the results using the Gibbs sampler were essentially the same, in the sense that the main comparisons came out the same way, as those obtained using a plugged in estimate of σ^2. Although some fast algorithms are available to carry out Gibbs sampling when σ^2 is unknown, the computation time is substantially less when using a plugged-in estimate.

Using the posterior probabilities for each basis we can form two estimators of $\boldsymbol{m}$. The first is the estimator using the most probable basis as determined by the posterior probabilities. The second is a weighted average of the posterior mean estimates over the four bases, using the posterior probabilities as weights. We use the notation $\widehat{\boldsymbol{m}}_{EB,B}$ and $\widehat{\boldsymbol{m}}_{EB,A}$ for the empirical Bayes estimator using the best (most probable) basis and the basis averaged estimator, respectively, where

$$\widehat{\boldsymbol{m}}_{EB,A} = \sum_{j=1}^{n} \widehat{\boldsymbol{m}}_{EB}^{j} p(B_j|\boldsymbol{w})$$

and $\widehat{\boldsymbol{m}}_{EB}^{j}$ is the posterior mean estimate of $\boldsymbol{m}$ for basis B_j. We use similar notation for the calibrated Bayes estimator. That is, for a given Q, $\widehat{\boldsymbol{m}}_{CB,B}^{Q}$ is the posterior mean estimate using the best basis and $\widehat{\boldsymbol{m}}_{EB,A}^{Q}$ is the basis averaged estimator.

5. Simulation Comparison

5.1. Introduction

This section reports the results of a simulation study comparing the performance of single basis estimators, both Bayesian and non-Bayesian, with estimators using basis selection and basis averaging. We use the Symmlet 8 basis for those experiments in which it is necessary to choose a single "best" basis because it is a good all-round performer as shown by the simulation results in Section 5.4, which is in accordance with wavelet folklore about these bases.

Figure 3 plots the twelve target curves used in the simulation. The first ten are used by Marron, Adak, Johnstone, Neumann and Patil (1998), the 11th is the zero function representing no mean structure, and the 12th is a mixture of two Gaussian densities representing a smooth function which requires a variable bandwidth estimator. This choice of targets induces a weighting scheme which has an impact on the simulation results reported below, but we feel the induced weights are reasonable for assessing the estimation methods. The design points are $n = 1024$ equally spaced points on $[0, 1]$, and independent Gaussian noise is added to yield the simulated observations $y_i = m(x_i) + \varepsilon_i$. Two noise levels are used; $\sigma = 0.02$ (low noise), and $\sigma = 0.1$ (high noise). Visual impression of these two noise levels is given by the lower right panels of Figure 3. The low noise is close to that often used in the wavelet examples of Donoho and Johnstone.

The performance of an estimator $\widehat{m}$ is conveniently summarized by the Average Squared Error

$$ASE = \frac{1}{n} \sum_{i=1}^{n} \{\widehat{m}(x_i) - m(x_i)\}^2 .$$

For each setting, the various estimators and their ASE values were computed for each of 1000 replicate pseudo data sets. These ASE's are summarized by their average ASE ($AASE$) over the 1000 replicates. The error criterion ASE is known to be somewhat different from visual impression, so we tried the visual error criterion of Marron and Tsybakov (1995) in some of these cases, but the main lessons are similar.

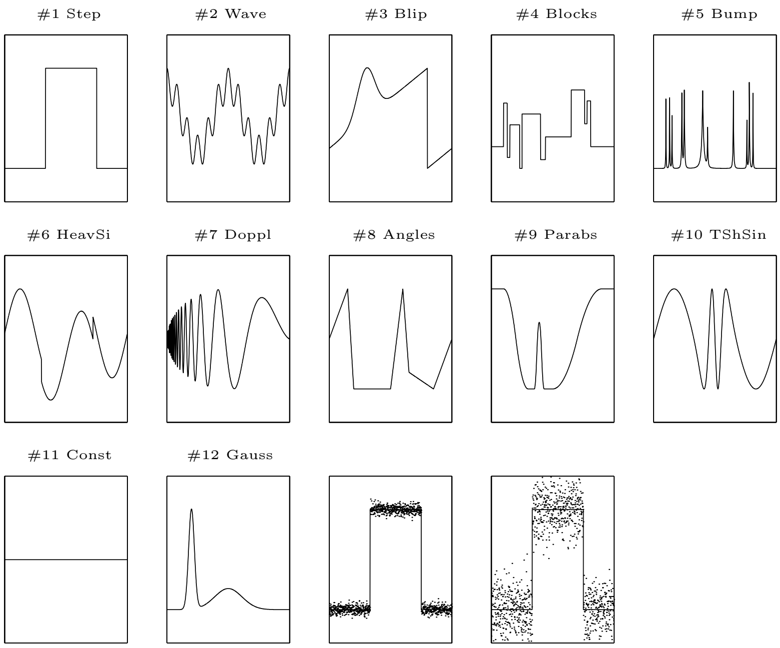

Figure 3. 12 target regression curves, plus step with $n = 1024$ pseudo observations show low, $\sigma = 0.02$, and high, $\sigma = 0.1$, Gaussian noise.

Because of the relatively large number of estimators studied, overall summarization can obscure the main ideas. Hence we present our results in several parts, designed to address three issues of interest. Section 5.2 studies whether Bayesian methods are worthwhile. Section 5.3 studies the gain available from basis selection and basis averaging. Section 5.4 compares the performance of the empirical Bayes approach to the calibrated Bayes approach for the four bases.

For each part, and for each of the 24 settings (12 target curves and 2 noise levels), the estimators in that part are compared through the proportion $(AASE/AASE_{\text{best}}) - 1$, where $AASE_{\text{best}}$ is the estimator having smallest AASE for that part and setting. This proportion is the main entry of each of the tables in Appendix 2. The accompanying entries in parentheses in each table reflect the simulation uncertainty by giving an estimate of the standard errors of each main entry in the table. To keep the tables manageable, we give the standard error estimates for the high noise case only. The estimates for the low noise case are similar qualitatively. Figures 4 to 6 summarize the tables. The figures show the performance of each estimator relative to others being studied in that part, for each of the 24 settings, by classifying the corresponding table entries into one of the following five categories.

1. Best: either the best overall $AASE$, or else within Monte Carlo variability (defined as 2 standard errors) of the best.
2. Excellent: $AASE$ within 10% of the best for that setting.
3. Very Good: $AASE$ between 10% and 20% of the best for that setting.
4. Acceptable: $AASE$ farther than 20% from the best for that setting, but less than 100% from the best.
5. Poor: $AASE$ more than 100% from the best.

The figures show the number of times (over the 24 settings) the estimator falls into each category, relative to the other members in that group.

5.2. Fixed basis Bayesian methods

Figure 4 compares some earlier wavelet estimators with the empirical Bayes estimator, and the calibrated Bayes estimators taking $Q = 0.90$ and 0.98. All estimators use the Symmlet 8 basis. The figure summarizes Table 1 in Appendix 2.

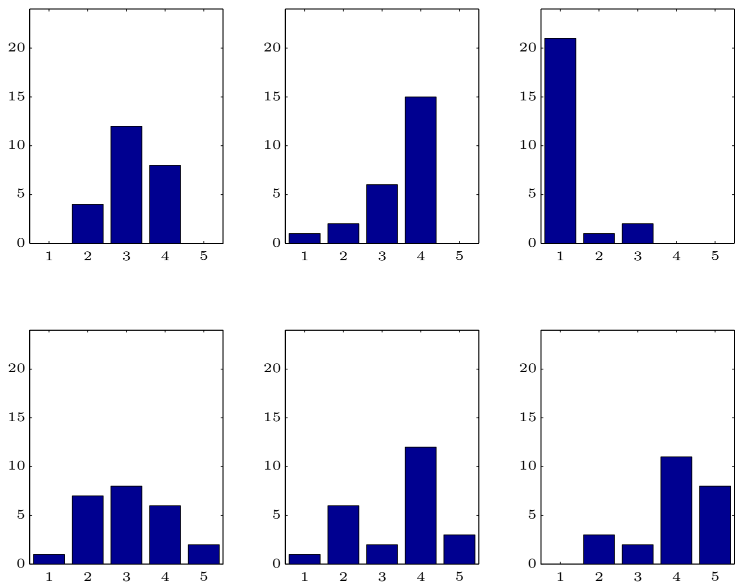

Figure 4. Comparison of classical and Bayesian methods using the Symmlet 8 basis. EB is empirical Bayes, CB is calibrated Bayes. 1 is best and 5 is worst.

Figure 4 shows that the empirical Bayes method is the best of the 6 estimators and that generally the Bayesian estimators perform favourably relative to the other estimators. The favourable performance of the Bayesian estimators is consistent with the results of Chipman, Kolaczyk and McCulloch (1997) and Clyde, Parmigiani and Vidakovic (1997). As observed in Marron, Adak, Johnstone, Neumann and Patil (1998), hard thresholding is somewhat better overall than soft thresholding in the sense of $AASE$. We were surprised by the poor performance of Sure Shrink, but do not believe there is a problem with our implementation, since we used the SUREShrink function SUREThresh.m in WaveLab (http://playfair.stanford.edu/∼wavelab).

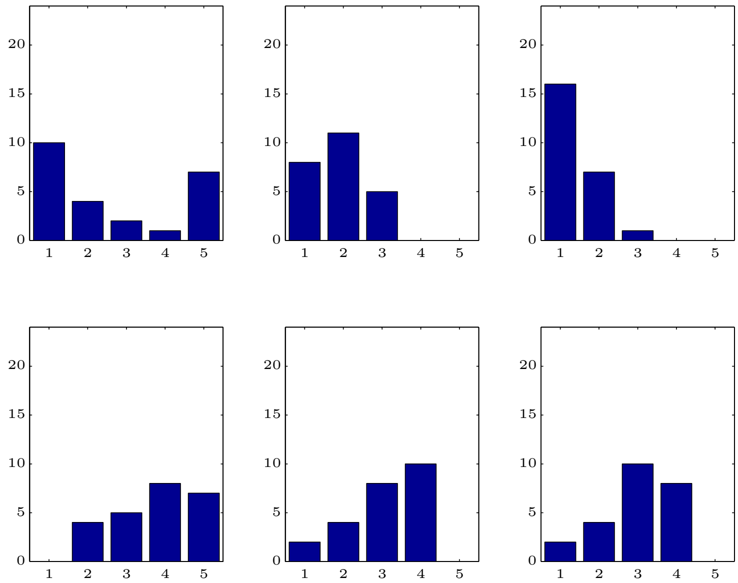

Figure 5. Comparison of the empirical Bayes estimator using the Symmlet 8 basis and the empirical Bayes estimators using basis selection and basis averaging.

5.3. Basis selection

Figure 5 and Table 2 compare the empirical Bayes and calibrated Bayes estimators using the Symmlet 8 basis with the corresponding estimators using basis selection and basis averaging. For the calibrated Bayes estimator we use $Q = 0.9$ because it performs better overall than $Q = 0.98$ as evident from Figures 4 and 6 and Tables 1, 3 and 4. The empirical Bayes method with basis averaging performs best overall with empirical Bayes with basis selection second. The empirical Bayes estimator using the Symmlet 8 basis does well for those functions for which it is the correct basis, but it can perform poorly for those functions for which it is not. In particular, Figure 5 and Table 2 emphasize the big gains made by basis selection and basis averaging.

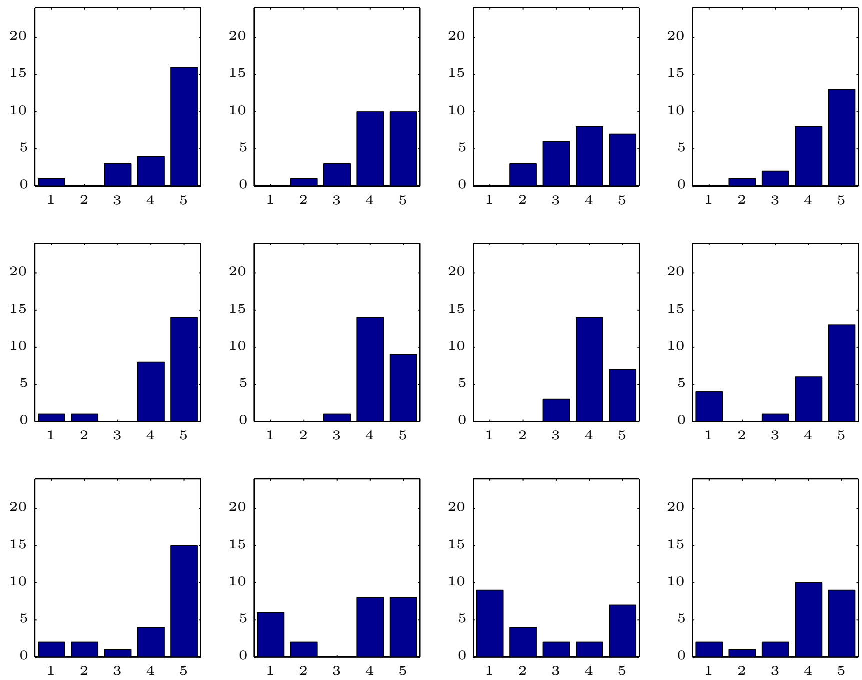

Figure 6. Comparison of empirical Bayes (EB) and calibrated (CB) Bayes approaches for individual bases.

5.4. Empirical Bayes vs calibrated Bayes for individual bases

Figure 6 studies the performance of empirical Bayes and calibrated Bayes for the four different bases. The figure summarizes the results in Tables 3 and 4 of Appendix 2. Rows in Figure 6 correspond to different Bayesian approaches and columns correspond to different bases. Comparing across rows shows that the empirical Bayes approach is generally somewhat better than the calibrated Bayes approach, and the calibrated Bayes approach with $Q = 0.90$ is generally a little better than the calibrated Bayes approach with $Q = 0.98$. This is why the calibrated Bayes method with $Q = 0.90$ is chosen in Figure 5. An exception is the Fourier basis where the calibrated Bayes estimator with $Q = 0.98$ outperforms the calibrated Bayes estimator with $Q = 0.90$. Comparison of columns, especially in the bottom row, confirms the conventional ideas about wavelet bases: the Symmlet 8 gives generally solid all-around performance. But perhaps less well known is our result that, in quite a few of these examples, the Symmlet 8 basis can be much worse than other bases.

Acknowledgement

We thank the Associate Editor and the referees for helping to improve the presentation of the paper. Robert Kohn was partially supported by an Australian Research Council Grant.

Appendix

1. Motivating the empirical Bayes prior

To make the paper self contained we follow Chipman, Kolaczyk and McCulloch (1997) and motivate the empirical Bayes prior in (3.1). To motivate the choice of c_j, we note that by Hölder's inequality

$$|\beta_{j,k}| \leq \|\psi_{j,k}\|_1 \cdot \max_{i=1,\ldots,n} |m_i| \, .$$

A rough approximation to $\max_{i=1,\ldots,n} |m_i|$ is $\max_{i=1,\ldots,n} |y_i|$. In the spirit of the mean plus or minus three standard deviations capturing most of the mass of the Gaussian distribution, $3c_j$ is an approximate bound on $|\beta_{j,k}|$ making $N(0, c_j^2)$ a reasonable prior for $\beta_{j,k}$. The probability π_j is the empirical probability of a term at scale j being larger than λ_D, the denoising threshold defined at (2.5). The Bayesian shrinkage estimator can be viewed as an update of the hard thresholded method. A slight difference with the hard and soft thresholded estimators is that for this Bayesian estimator, thresholding is applied to all coefficients, including the father coefficients (although this typically does not make a difference, since usually $\pi_j = 1$ for the important large scale coefficients). We experimented with other values, but found these to be quite effective. Chipman, Kolaczyk and McCulloch (1997) also include another hyperparameter τ, but we have not included it because it usually had no important effect unless set to extreme values which resulted in poorer performance.

2. Tables of Results

Tables 1 to 4 present, for each estimator and each setting, $(AASE - AASE_{\text{best}})/AASE_{\text{best}}$; $AASE_{\text{best}}$ is the $AASE$ for that estimator having smallest $AASE$ for that setting within a table. The numbers in brackets are the corresponding standard errors. To keep the size of the tables manageable, we present the standard errors only for the high noise case. The stanadard errors for the low noise case are qualitatively similar. Tables 3 and 4 present the results of Section 5.4 and should be viewed as one table.

Table 1. Comparison of Bayesian and non-Bayesian estimators using the Symmlet 8 basis. The numbers in brackets are standard error estimates for the high noise case.

	CB, Q=0.90, Symm8	CB, Q=0.98, Symm8	Emp. Bayes, Symm8	hard thresh Symm8	soft thresh Symm8	sure shrink Symm8
step, lo	0.0582	0.1824	0.0000	0.0877	0.5524	3.0952
step, hi	0.1877	0.3429	0.0000	0.0161	0.0306	0.0912
	(0.0052)	(0.0056)	(0.0045)	(0.0045)	(0.0042)	(0.0048)
wave, lo	0.4771	0.8519	0.0000	0.3392	1.3339	9.2133
wave, hi	0.1021	0.0842	0.0000	0.0846	0.0655	0.1839
	(0.0040)	(0.0034)	(0.0040)	(0.0058)	(0.0044)	(0.0066)
blip, lo	0.0014	0.1399	0.0000	0.0406	0.5059	1.5220
blip, hi	0.2238	0.4181	0.0000	0.1658	0.1903	0.2555
	(0.0088)	(0.0103)	(0.0068)	(0.0079)	(0.0060)	(0.0073)
bloc, lo	0.1414	0.3493	0.0000	0.2529	0.7371	6.1121
bloc, hi	0.2877	0.4550	0.0000	0.1646	0.0550	0.2148
	(0.0042)	(0.0048)	(0.0033)	(0.0035)	(0.0029)	(0.0065)
bump, lo	0.1453	0.3515	0.0000	0.2094	0.6682	2.6104
bump, hi	0.1934	0.3336	0.0000	0.1672	0.0155	0.0370
	(0.0036)	(0.0039)	(0.0032)	(0.0033)	(0.0026)	(0.0039)
hvsi, lo	0.1778	0.3475	0.1097	0.0263	0.0000	0.0387
hvsi, hi	0.0991	0.1018	0.0000	0.2306	0.2413	0.4141
	(0.0064)	(0.0054)	(0.0061)	(0.0092)	(0.0070)	(0.0102)
dopp, lo	0.0021	0.1245	0.0000	0.1101	0.3682	1.0172
dopp, hi	0.1208	0.2416	0.0000	0.2678	0.3635	0.4275
	(0.0054)	(0.0062)	(0.0046)	(0.0056)	(0.0055)	(0.0076)
angl, lo	0.2204	0.3385	0.0000	0.0032	0.0335	0.2981
angl, hi	0.3664	0.5085	0.0119	0.0000	0.0263	0.1921
	(0.0100)	(0.0119)	(0.0071)	(0.0081)	(0.0062)	(0.0091)
para, lo	0.1513	0.2597	0.0000	0.0428	0.3136	0.6968
para, hi	0.1708	0.2363	0.0000	0.3988	0.3129	0.4172
	(0.0072)	(0.0071)	(0.0069)	(0.0085)	(0.0071)	(0.0097)
tshs, lo	0.1575	0.2187	0.0000	0.1242	0.4297	0.7106
tshs, hi	0.2428	0.3423	0.0000	0.1619	0.1753	0.3475
	(0.0097)	(0.0107)	(0.0078)	(0.0091)	(0.0071)	(0.0102)
cons, lo	0.2951	0.0041	0.0000	3.0086	3.2318	4.1586
cons, hi	0.2921	0.0000	0.1422	3.0922	3.3220	4.2559
	(0.0234)	(0.0171)	(0.0209)	(0.0407)	(0.0311)	(0.0453)
gaus, lo	0.1409	0.1745	0.0000	0.1026	0.4256	0.6343
gaus, hi	0.1554	0.1373	0.0000	0.1984	0.2268	0.4122
	(0.0066)	(0.0053)	(0.0066)	(0.0096)	(0.0074)	(0.0108)

Table 2. Comparison of Bayesian estimators using the Symmlet 8 basis, basis selection and basis averaging. The numbers in brackets are standard error estimates for the high noise case.

	Emp. Bayes, Symm8	Emp. Bayes, Basis sel.	Emp. Bayes, Basis avg.	CB, Q=0.90, Symm8	CB, Q=0.90, Basis sel.	CB, Q=0.90, Basis avg.
step, lo	10.0771	0.0000	0.0000	10.7214	0.1477	0.1477
step, hi	3.3981	0.0997	0.0997	4.2234	0.0000	0.0000
	(0.0200)	(0.0133)	(0.0133)	(0.0230)	(0.0124)	(0.0124)
wave, lo	4.9807	0.0000	0.0000	7.8342	0.0351	0.0351
wave, hi	9.2726	0.0000	0.0000	10.3211	0.1848	0.1848
	(0.0409)	(0.0200)	(0.0200)	(0.0409)	(0.0306)	(0.0306)
blip, lo	0.0779	0.0022	0.0000	0.0793	0.0785	0.0656
blip, hi	0.0000	0.0809	0.0253	0.2238	0.3219	0.2502
	(0.0068)	(0.0069)	(0.0069)	(0.0088)	(0.0087)	(0.0090)
bloc, lo	2.9481	0.0000	0.0000	3.5062	0.1054	0.1054
bloc, hi	0.2295	0.0116	0.0000	0.5832	0.2162	0.2030
	(0.0041)	(0.0049)	(0.0046)	(0.0052)	(0.0072)	(0.0067)
bump, lo	0.1097	0.0000	0.0000	0.2710	0.1838	0.1836
bump, hi	0.0460	0.0150	0.0000	0.2483	0.2333	0.2266
	(0.0033)	(0.0038)	(0.0038)	(0.0037)	(0.0045)	(0.0044)
hvsi, lo	0.0001	0.0004	0.0000	0.0615	0.0615	0.0615
hvsi, hi	0.0000	0.0870	0.0615	0.0991	0.4621	0.4563
	(0.0061)	(0.0066)	(0.0067)	(0.0064)	(0.0099)	(0.0100)
dopp, lo	0.0000	0.0000	0.0000	0.0021	0.0021	0.0021
dopp, hi	0.0000	0.0000	0.0000	0.1208	0.1208	0.1208
	(0.0046)	(0.0046)	(0.0046)	(0.0054)	(0.0054)	(0.0054)
angl, lo	0.0551	0.0406	0.0000	0.2876	0.2089	0.1658
angl, hi	0.0000	0.1100	0.0375	0.3503	0.4903	0.4032
	(0.0070)	(0.0077)	(0.0081)	(0.0099)	(0.0099)	(0.0111)
para, lo	0.0000	0.0018	0.0007	0.1513	0.1513	0.1513
para, hi	0.0000	0.0104	0.0063	0.1708	0.1736	0.1725
	(0.0069)	(0.0073)	(0.0072)	(0.0072)	(0.0074)	(0.0073)
tshs, lo	0.0000	0.1559	0.1254	0.1575	0.1654	0.1648
tshs, hi	0.1771	0.1097	0.0546	0.4629	0.0004	0.0000
	(0.0091)	(0.0106)	(0.0108)	(0.0114)	(0.0099)	(0.0099)
cons, lo	2.2641	0.1226	0.0000	3.2274	0.7810	0.6822
cons, hi	1.6149	0.1610	0.0000	1.9583	0.2404	0.1772
	(0.0479)	(0.0405)	(0.0354)	(0.0536)	(0.0491)	(0.0462)
gaus, lo	0.0000	0.0554	0.0363	0.1409	0.2649	0.2173
gaus, hi	0.0437	0.0603	0.0000	0.2059	0.2841	0.2116
	(0.0068)	(0.0081)	(0.0082)	(0.0068)	(0.0105)	(0.0110)

Table 3. Comparison of single basis Bayesian estimators. The numbers in brackets are the standard error estimates for the high noise case. The table is continued in table 4.

	CB, Q=0.90, Haar	CB, Q=0.90, Daub4	CB, Q=0.90, Symm8	CB, Q=0.90, Fourier	CB, Q=0.98, Haar	CB, Q=0.98, Daub4
step, lo	0.1906	4.5474	11.1594	118.2406	0.0000	5.1432
step, hi	0.0000	2.1775	4.2234	9.8014	0.0490	2.5934
	(0.0124)	(0.0197)	(0.0230)	(0.0290)	(0.0126)	(0.0221)
wave, lo	50.1419	25.7619	7.8414	0.0359	55.9800	28.8413
wave, hi	50.8575	33.0783	13.1985	0.4859	57.4023	36.2924
	(0.1618)	(0.1127)	(0.0512)	(0.0383)	(0.1809)	(0.1298)
blip, lo	1.6256	0.0632	0.0813	12.0845	1.9453	0.1863
blip, hi	0.4848	0.2413	0.2238	2.4523	0.6255	0.4107
	(0.0074)	(0.0078)	(0.0088)	(0.0097)	(0.0081)	(0.0084)
bloc, lo	0.1054	1.9624	3.5062	14.1679	0.2589	2.4679
bloc, hi	0.1763	0.7051	0.6027	0.8690	0.3530	0.9610
	(0.0047)	(0.0059)	(0.0053)	(0.0063)	(0.0052)	(0.0065)
bump, lo	0.5322	0.1838	0.2715	2.7888	0.7795	0.3904
bump, hi	0.4834	0.2351	0.2657	0.8487	0.6822	0.4011
	(0.0045)	(0.0040)	(0.0038)	(0.0044)	(0.0049)	(0.0046)
hvsi, lo	2.1987	0.1013	0.1797	1.5860	2.5466	0.2817
hvsi, hi	2.6043	0.8429	0.0991	0.4644	3.0835	0.9059
	(0.0161)	(0.0072)	(0.0064)	(0.0099)	(0.0180)	(0.0067)
dopp, lo	3.7246	0.8232	0.0021	4.0631	4.4628	1.0769
dopp, hi	2.3685	1.3441	0.1208	2.6409	2.8030	1.6365
	(0.0096)	(0.0075)	(0.0054)	(0.0103)	(0.0110)	(0.0083)
angl, lo	2.9786	0.1796	0.3341	0.5063	3.5721	0.3181
angl, hi	1.6537	0.3251	0.3503	0.5114	1.9194	0.4595
	(0.0115)	(0.0091)	(0.0099)	(0.0094)	(0.0127)	(0.0098)
para, lo	3.6651	0.6136	0.1513	0.4892	4.2784	0.7593
para, hi	1.8548	0.6147	0.1708	1.6217	2.2170	0.7843
	(0.0138)	(0.0099)	(0.0072)	(0.0167)	(0.0157)	(0.0109)
tshs, lo	9.7622	2.0998	0.1575	0.1654	11.0479	2.5843
tshs, hi	5.2259	2.0632	0.7526	0.1974	5.8689	2.5355
	(0.0211)	(0.0201)	(0.0136)	(0.0119)	(0.0242)	(0.0230)
cons, lo	1.7828	1.7851	4.8834	1.2977	0.4937	0.4851
cons, hi	1.4115	1.4562	4.2627	1.1156	0.3172	0.3104
	(0.0758)	(0.0752)	(0.0953)	(0.0863)	(0.0467)	(0.0466)
gaus, lo	3.9090	0.6801	0.1409	0.3836	4.5741	0.8343
gaus, hi	0.8357	0.2698	0.2362	1.0253	0.9173	0.3476
	(0.0102)	(0.0106)	(0.0070)	(0.0147)	(0.0123)	(0.0117)

Table 4. Comparison of single basis Bayesian estimators. The numbers in brackets are standard error estimates for the high noise case. This table is a continuation of table 3

	CB, Q=0.98, Symm8	CB, Q=0.98, Fourier	Emp. Bayes, Haar	Emp. Bayes, Daub4	Emp. Bayes, Symm8	Emp. Bayes, Fourier
step, lo	12.5864	132.8428	0.0374	4.7550	10.4910	115.9567
step, hi	4.9062	10.8879	0.0997	1.8549	3.3981	8.3799
	(0.0245)	(0.0326)	(0.0133)	(0.0183)	(0.0200)	(0.0248)
wave, lo	10.0849	0.0000	40.2423	19.8853	4.9856	0.0008
wave, hi	12.9688	0.0000	33.6587	24.1176	11.8836	0.2542
	(0.0440)	(0.0265)	(0.1228)	(0.1069)	(0.0513)	(0.0251)
blip, lo	0.2308	13.6217	1.0504	0.0000	0.0798	11.4359
blip, hi	0.4181	2.7788	0.2795	0.0346	0.0000	1.8427
	(0.0103)	(0.0110)	(0.0076)	(0.0061)	(0.0068)	(0.0085)
bloc, lo	4.3273	16.0946	0.0000	1.7203	2.9481	13.6969
bloc, hi	0.8110	1.1306	0.0000	0.2167	0.2446	0.4785
	(0.0060)	(0.0076)	(0.0042)	(0.0047)	(0.0041)	(0.0047)
bump, lo	0.5004	3.3952	0.3609	0.0000	0.1102	2.6429
bump, hi	0.4145	1.0452	0.1623	0.0000	0.0606	0.5844
	(0.0041)	(0.0048)	(0.0038)	(0.0034)	(0.0034)	(0.0039)
hvsi, lo	0.3497	1.9347	1.7985	0.0000	0.1115	1.3660
hvsi, hi	0.1018	0.6011	1.5799	0.5243	0.0000	0.1313
	(0.0054)	(0.0107)	(0.0110)	(0.0077)	(0.0061)	(0.0061)
dopp, lo	0.1245	4.8131	2.8549	0.6490	0.0000	3.7428
dopp, hi	0.2416	3.1320	1.5719	0.8470	0.0000	1.8523
	(0.0062)	(0.0121)	(0.0072)	(0.0065)	(0.0046)	(0.0076)
angl, lo	0.4632	0.6458	2.3965	0.0000	0.0932	0.4103
angl, hi	0.4907	0.6759	1.1405	0.0679	0.0000	0.0705
	(0.0118)	(0.0100)	(0.0097)	(0.0078)	(0.0070)	(0.0075)
para, lo	0.2597	0.6400	3.1199	0.4730	0.0000	0.3513
para, hi	0.2363	2.1644	1.1755	0.2208	0.0000	0.6852
	(0.0071)	(0.0203)	(0.0105)	(0.0087)	(0.0069)	(0.0115)
tshs, lo	0.2187	0.1975	8.0715	1.3652	0.0000	0.1473
tshs, hi	0.8928	0.2469	3.7925	1.0481	0.4101	0.0000
	(0.0151)	(0.0122)	(0.0168)	(0.0130)	(0.0110)	(0.0099)
cons, lo	3.5615	0.0000	0.8219	0.7351	3.5427	0.2964
cons, hi	3.0729	0.0000	1.2241	1.1693	3.6519	0.4805
	(0.0697)	(0.0580)	(0.0733)	(0.0728)	(0.0852)	(0.0596)
gaus, lo	0.1745	0.4274	3.4715	0.5461	0.0000	0.3495
gaus, hi	0.2169	1.3331	0.6860	0.0000	0.0700	0.4148
	(0.0057)	(0.0178)	(0.0085)	(0.0088)	(0.0070)	(0.0101)

References

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

Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92**, 1413-1421.

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

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

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

Donoho, D. L., Johnstone, I. M., Kerkyacherian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 301-370.

George, E. I. and Foster, D. (1997). Calibration and empirical Bayes variable selection. Preprint.

Marron, J. S. and Tsybakov, A. B. (1995). Visual error criteria for qualitative smoothing. *J. Amer. Statist. Assoc.* **90**, 499-507.

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. and Patil, P. (1998). Exact risk analysis of wavelet regression. *J. Comput. Graphical Statist.* To appear.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.* **93**, 173-179.

Australian Graduate School of Management, University of New South Wales, PO Box 1, Kensington NSW 2033, Australia.

E-mail: r.kohn@unsw.edu.au

E-mail: marron@stat.unc.edu

(Received March 1997; accepted March 1999)