



## Multiple Shrinkage and Subset Selection in Wavelets

Merlise Clyde; Giovanni Parmigiani; Brani Vidakovic

*Biometrika*, Vol. 85, No. 2 (Jun., 1998), 391-401.

Stable URL:

<http://links.jstor.org/sici?sici=0006-3444%28199806%2985%3A2%3C391%3AMSASSI%3E2.0.CO%3B2-3>

*Biometrika* is currently published by Biometrika Trust.

---

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/bio.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 creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

## Multiple shrinkage and subset selection in wavelets

BY MERLISE CLYDE, GIOVANNI PARMIGIANI AND BRANI VIDAKOVIC

*Institute of Statistics and Decision Sciences, Duke University, Durham,  
North Carolina 27708-0251, U.S.A.*

clyde@stat.duke.edu gp@stat.duke.edu brani@stat.duke.edu

### SUMMARY

This paper discusses Bayesian methods for multiple shrinkage estimation in wavelets. Wavelets are used in applications for data denoising, via shrinkage of the coefficients towards zero, and for data compression, by shrinkage and setting small coefficients to zero. We approach wavelet shrinkage by using Bayesian hierarchical models, assigning a positive prior probability to the wavelet coefficients being zero. The resulting estimator for the wavelet coefficients is a multiple shrinkage estimator that exhibits a wide variety of nonlinear patterns. We discuss fast computational implementations, with a focus on easy-to-compute analytic approximations as well as importance sampling and Markov chain Monte Carlo methods. Multiple shrinkage estimators prove to have excellent mean squared error performance in reconstructing standard test functions. We demonstrate this in simulated test examples, comparing various implementations of multiple shrinkage to commonly-used shrinkage rules. Finally, we illustrate our approach with an application to the so-called ‘glint’ data.

*Some key words:* Gibbs sampling; Importance sampling; Model averaging.

### 1. INTRODUCTION

Wavelets are families of basis or basis-like functions that can be used to approximate other functions. They combine powerful properties such as orthonormality, compact support, varying degrees of smoothness, localisation in time and scale, i.e. frequency, and fast implementation. Daubechies and Mallat sparked interest in wavelets in the statistics community when they connected wavelets with discrete data processing (Daubechies, 1988; Mallat, 1989). Donoho, Johnstone et al. showed that wavelet shrinkage had desirable statistical optimality properties in problems concerning elimination of noise (Donoho & Johnstone, 1994, 1995; Donoho et al., 1995).

The wavelet regression model can be described as follows. Suppose the function  $f(\cdot)$  is sampled at  $N$  equally spaced points,  $f = \{f(x_1), \dots, f(x_N)\}'$ , but is observed with additive white noise  $\varepsilon$ . If  $Y = (Y_1, \dots, Y_N)'$  is the vector of observations, then the model is  $Y = f + \varepsilon$ , or equivalently in a wavelet regression form,  $Y = W\beta + \varepsilon$ , where  $f = W\beta$ ,  $\beta$  is the discrete wavelet transformation of  $f$ , and  $W'$  is the  $N \times N$  orthogonal matrix corresponding to the discrete wavelet transformation. The least squares estimator of  $\beta$  is  $\hat{\beta} = W'Y$  and, in fact, is the discrete wavelet transformation of  $Y$ . In practice, there is no need to compute  $W$ , as there are fast algorithms to compute the empirical wavelet coefficients,  $\hat{\beta}$ , directly in only  $O(N)$  operations.

Shrinkage of empirical wavelet coefficients is particularly effective when many of the

coefficients represent noise rather than signal. Wavelet shrinkage can be described by a three-step procedure: (i) noisy observations  $Y$  are transformed by a discrete wavelet transformation to obtain  $\hat{\beta}$ , (ii) the coefficients  $\hat{\beta}$  are shrunk towards zero or possibly set to 0, (iii) the shrunk coefficients are returned to the  $Y$  domain by the inverse discrete wavelet transformation. The resulting vector is a wavelet shrinkage estimator  $\hat{f}$  of the unknown vector  $f$ .

In this paper, we present methods for multiple shrinkage of wavelet coefficients based on a Bayesian hierarchical model. Our main focus is on incorporating model uncertainty about which of the coefficients are zero. The resulting posterior mean of the coefficients combines linear shrinkage to the prior mean, zero, with additional nonlinear shrinkage to zero resulting from averaging over models where the coefficient is zero.

As the posterior probability that a coefficient is zero is typically not available in closed form, unless the error variance is known, Bayesian model averaging for constructing multiple shrinkage estimators is more computationally intensive compared to many of the alternative approaches. We present two analytical approximations that can be calculated in closed form, and compare these with Monte Carlo methods for obtaining posterior means and variances.

In § 2, we discuss the Bayesian hierarchical regression model. In § 3 we introduce the multiple shrinkage estimators and analytical approximations for calculating approximate posterior means and variances. These approximations are used to implement importance sampling and Gibbs sampling algorithms. These lead to more accurate estimates of posterior means and variances, but at an additional computational cost. In § 4, we compare our multiple shrinkage estimators to conventional alternatives, using simulation with standard test functions. We consider both normal and Student- $t$  errors. In § 5, we present an application illustrating the importance of prior information in signal denoising.

## 2. BAYESIAN HIERARCHICAL MODEL

The statistical model is described by the following distributional assumptions. First, we assume that the errors  $\varepsilon_i$  are independent normal with mean 0 and variance  $\sigma^2$  so that

$$Y|\beta, \sigma^2 \sim N(W\beta, \sigma^2 I_N).$$

This assumption is standard in many wavelet applications (Donoho & Johnstone, 1995). In § 4, we explore the robustness of our estimators to that assumption. We introduce the  $N$ -dimensional vector  $\gamma$ , which is a sequence of binary random variables, to represent which elements of  $\beta$  are zero. We will identify individual elements of vectors such as  $\beta$ ,  $\hat{\beta}$  and  $\gamma$  by the indices  $j$  and  $k$  in subscripts  $jk$ , corresponding to basis elements  $\psi_{jk}$ . The  $\psi_{jk}$  are dilations at level or scale  $j$  and location translations by  $k2^j$  of the mother wavelet function  $\psi$ . In the prior distribution, the  $\beta_{jk}$ 's given  $\gamma$  and  $\sigma$  are independently distributed as

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

which is normal with variance  $c_{jk}\sigma^2$  when  $\gamma_{jk} = 1$ , and degenerate at zero when  $\gamma_{jk} = 0$ . We use a conjugate distribution for  $\sigma^2$ ,

$$\lambda v/\sigma^2 \sim \chi_v^2,$$

where  $\lambda$  and  $v$  are fixed hyperparameters. The choice  $v = \lambda = 0$  is commonly adopted to represent lack of prior knowledge. We use this in the simulation studies in § 4. In § 5, we

illustrate a case in which prior information is used to determine appropriate values of  $\gamma$  and  $\lambda$ .

Finally, the  $\gamma_{jk}$  are independently distributed as Bernoulli random variables,

$$\gamma_{jk} \sim \text{Ber}(\theta_{jk}).$$

It can be useful to set  $\theta_{jk} = \theta_j$  and  $c_{jk} = c_j$ , so that the hyperparameters are constant within level  $j$ , resulting in monotonic shrinkage of coefficients within each level. E. I. George and D. Foster, in a technical report from the University of Texas at Austin, propose a method for calibrating  $c$  so that, conditional on  $\sigma$ , model selection based on posterior model probabilities corresponds to classical model selection methods. The value of  $c$  is chosen by solving

$$F(c, \theta) = \frac{1+c}{c} \left\{ \log(1+c) + 2 \log\left(\frac{1-\theta}{\theta}\right) \right\},$$

where the value of  $F$  depends on the model selection criterion. For example,  $F(c, \theta) = \log N$  corresponds to model selection by the BIC, or Bayesian Information Criterion (Akaike, 1978), and  $F(c, \theta) = 2 \log N$  corresponds to the Risk Inflation Criterion (Foster & George, 1994; Donoho & Johnstone, 1994). The latter choice corresponds to selecting a model where all the  $|t\text{-statistics}|$  are greater than  $(2 \log N)^{\frac{1}{2}}$ , as in universal thresholding. This can be applied separately for each level.

Mixtures of normals and point masses have been used in Bayesian variable selection, as reviewed in George & McCulloch (1997). A related approach is George & McCulloch's (1993) stochastic search variable selection, where the prior distribution on the coefficients is a mixture of two normal components, one concentrated around zero and the other suitably dispersed. Chipman, Kolaczyk & McCulloch (1997) apply this method to wavelets.

### 3. POSTERIOR INFERENCE

#### 3.1. The multiple shrinkage estimator

The posterior mean of  $\beta_{jk}$  conditional on  $\gamma$  is  $E(\beta_{jk} | \gamma, Y) = \gamma_{jk} \hat{\beta}_{jk} / (1 + c_{jk}^{-1})$ . In turn, the posterior distribution of  $\gamma$  is

$$\pi(\gamma | Y) = q(\gamma) / \sum_{\text{all } \gamma'} q(\gamma'), \tag{1}$$

where

$$q(\gamma) = \left[ \prod_{jk} \left\{ \frac{\theta_{jk}}{1-\theta_{jk}} (1+c_{jk})^{-\frac{1}{2}} \right\}^{\gamma_{jk}} \right] \left\{ \lambda v + Y'Y - \sum_{jk} \gamma_{jk} \hat{\beta}_{jk}^2 / (1+c_{jk}^{-1}) \right\}^{-(N+v)/2}.$$

Posterior probabilities in (1) can be used for model selection under various loss functions.

As an alternative to selecting a  $\gamma$  corresponding to the 'best' model, we can make inferences by averaging over all possible  $\gamma$ 's. The posterior mean of  $\beta_{jk}$  is

$$E(\beta_{jk} | Y) = \sum_{\gamma} \pi(\gamma | Y) E(\beta_{jk} | Y, \gamma) = E(\gamma_{jk} | Y) \hat{\beta}_{jk} / (1 + c_{jk}^{-1}). \tag{2}$$

This is a multiple shrinkage estimator, in which regression coefficients are shrunk linearly towards their respective prior mean, zero in this case, and are further nonlinearly shrunk

towards zero as a result of uncertainty about whether or not the  $\gamma_{jk}$  is zero. The use of Bayesian model averaging in other contexts has resulted in improved predictive performance (Draper, 1995; Raftery, Madigan & Volinski, 1996). George (1986) discusses desirable Bayes and minimax properties of multiple shrinkage estimators in a general context. The appeal of this shrinkage approach is its flexibility and adaptation to the observed data. The difficulty in implementing this in real time is that the expectations,  $E(\gamma_{jk}|Y)$ , cannot be calculated analytically. We present four approaches for approximating the posterior means and variances.

3.2. Analytic approximation given  $\sigma$ : Method S

Closed-form approximations for the posterior mean and variance can be obtained based on conditioning on  $\sigma$ . Given  $Y$  and  $\sigma$ , the  $\gamma_{jk}$  are independent Bernoulli random variables with parameters

$$p_{jk}(\sigma) = \frac{a_{jk}(Y, \sigma)}{1 + a_{jk}(Y, \sigma)}, \quad a_{jk}(Y, \sigma) = (1 + c_{jk})^{-\frac{1}{2}} \left( \frac{\theta_{jk}}{1 - \theta_{jk}} \right) \exp\left(\frac{1}{2} \frac{S_{jk}^2}{\sigma^2}\right), \tag{3}$$

where  $S_{jk}^2 = \hat{\beta}_{jk}^2 / (1 + c_{jk}^{-1})$ . The conditional posterior mean is

$$E(\beta_{jk}|Y, \sigma) = p_{jk}(\sigma) \hat{\beta}_{jk} / (1 + c_{jk}^{-1}), \tag{4}$$

which can be computed in real time. As  $\sigma$  is typically unknown, we can use an estimate, such as  $\hat{\sigma} = \text{median}(|\hat{\beta}_{1k}|) / 0.6745$ , proposed by Donoho et al. (1995).

The conditional posterior variance of  $\beta$  given  $\sigma$  is available in closed form as

$$\text{var}(\beta_{jk}|Y, \sigma) = \frac{p_{jk}(\sigma)}{1 + c_{jk}^{-1}} \sigma^2 + \frac{p_{jk}(\sigma)\{1 - p_{jk}(\sigma)\}}{(1 + c_{jk}^{-1})^2} \hat{\beta}_{jk}^2, \tag{5}$$

and can be used to obtain an approximation of the posterior variance by evaluating (5) at  $\hat{\sigma}$ . Let  $V$  be the diagonal matrix with  $\text{var}(\beta_{jk}|Y, \hat{\sigma})$  on the diagonal and zero elsewhere. The covariance terms are all zero because of the prior independence of the  $\beta_{jk}$ 's and orthogonality of  $W$ . The posterior covariance matrix of  $f$  given  $Y$  and  $\sigma$  is  $(W'VW) = W'(W'V)'$ . An efficient method to compute the variance of  $\hat{f}$  is suggested by Chipman et al. (1997). This involves applying the inverse discrete wavelet transformation to the  $N$  columns of  $V$  and then successively to the  $N$  columns of  $(W'V)'$ , requiring  $2N$  inverse wavelet transformations.

3.3. Analytic approximations: Method A

If the posterior distribution for  $\sigma$  is not concentrated, the estimator in (4) may not be very accurate. An alternative approach is to approximate (1) by a model of independence,

$$\tilde{\pi}(y|y) = \prod_{jk} p_{jk}^{\gamma_{jk}} (1 - p_{jk})^{1 - \gamma_{jk}}, \tag{6}$$

where

$$p_{jk} = \frac{\theta_{jk} \exp(-\frac{1}{2} \log(1 + c_{jk}) + \frac{1}{2}(N + v)L \sum_{l=1}^{l^*} [(S_{jk}^2)^l / \{l(v\lambda + Y'Y)^l\}])}{1 - \theta_{jk} + \theta_{jk} \exp(-\frac{1}{2} \log(1 + c_{jk}) + \frac{1}{2}(N + v)L \sum_{l=1}^{l^*} [(S_{jk}^2)^l / \{l(v\lambda + Y'Y)^l\}])} \tag{7}$$

and  $S_{jk}^2 = \hat{\beta}_{jk}^2 / (1 + c_{jk}^{-1})$ . The quantity  $L$  is a calibration constant defined by

$$L = \frac{\log(v\lambda + Y'Y) - \log(v\lambda + Y'Y - \sum_{jk} S_{jk}^2)}{\sum_{l=1}^{l^*} [\sum_{jk} (S_{jk}^2)^l / \{l(v\lambda + Y'Y)^l\}]}$$

The approximation is based on a Taylor series expansion of the logarithm of the posterior model probabilities (Clyde, DeSimone & Parmigiani, 1996). The  $p_{jk}$ 's can be used to obtain a direct approximation to the multiple shrinkage estimator,

$$E(\beta_{jk} | Y) \approx p_{jk} \hat{\beta}_{jk} / (1 + c_{jk}^{-1}), \tag{8}$$

which takes into account uncertainty in  $\sigma$ .

3.4. Importance sampling: Method I

Importance sampling is based on generating a sample of  $\gamma$ 's from (6), by drawing each  $\gamma_{jk}$  as an independent Bernoulli random variable with  $\text{pr}(\gamma_{jk} = 1) = p_{jk}$ . An importance sampling estimator for  $\beta$  is then

$$E(\beta | Y) \approx \sum w(\gamma) \gamma_{jk} \hat{\beta}_{jk} / (1 + c_{jk}^{-1}), \tag{9}$$

where  $w(\gamma)$  is the importance sampling weight,

$$w(\gamma) = \frac{n(\gamma)q(\gamma)/\hat{\pi}(\gamma)}{\sum_{\gamma'} n(\gamma')q(\gamma')/\hat{\pi}(\gamma')},$$

and  $n(\gamma)$  is the number of times model  $\gamma$  is sampled. All sums are over the set of sampled models. Alternatively, one can use weights  $\hat{\pi}$  based on renormalising the  $q(\gamma)$ 's obtained from the sampled models (Clyde, DeSimone & Parmigiani, 1996):

$$w(\gamma) = \hat{\pi} \equiv q(\gamma) / \sum_{\gamma'} q(\gamma'). \tag{10}$$

This is preferable if the variance of the importance sampling weights is large.

3.5. Gibbs sampling: Method G

In the specific context of wavelets and other orthogonal bases, we can implement an efficient block Gibbs sampler that alleviates some of the slow mixing problems noted in Markov chain Monte Carlo samplers that generate from the full conditional distributions of  $\beta$ ,  $\gamma$  and  $\sigma$ . We obtain samples from the joint posterior distribution of  $\sigma$  and  $\gamma$  by drawing from the full conditional distributions in two blocks, as follows. First, given  $\gamma$  and  $Y$ ,  $\sigma^2$  has an inverse Gamma distribution, or,

$$\sigma^2 | \gamma, Y \sim \left( \lambda v + Y'Y - \sum \gamma_{jk} S_{jk}^2 \right) / \chi_{v+N}^2.$$

Secondly, conditional on  $\sigma$ , the distribution of  $\gamma$  factorises as a product of independent Bernoulli random variables with probabilities  $p_{jk}(\sigma)$  using (3). This has an advantage over other Markov chain Monte Carlo methods that sample from (1) in that the  $\gamma_{jk}$ 's are conditionally independent, reducing the computations to generate  $\gamma$ . The main advantage is that the posterior expectation of  $\beta$  given  $Y$  can be estimated using a Rao-Blackwellised estimator based on (4),

$$E(\beta_{jk} | Y) \approx \frac{1}{M} \sum_{m=1}^M p_{jk}(\sigma_m) \hat{\beta}_{jk} / (1 + c_{jk}^{-1}), \tag{11}$$

where  $\sigma_m$  ( $m = 1, \dots, M$ ) are the sampled values.



universal thresholding gives the same result as maximising the posterior model probability under priors (b) and (g), given  $\sigma = \hat{\sigma}$ . Thus the comparison between universal thresholding and multiple shrinkage using priors (b) and (g) with method S can be seen as a comparison between model selection and model averaging.

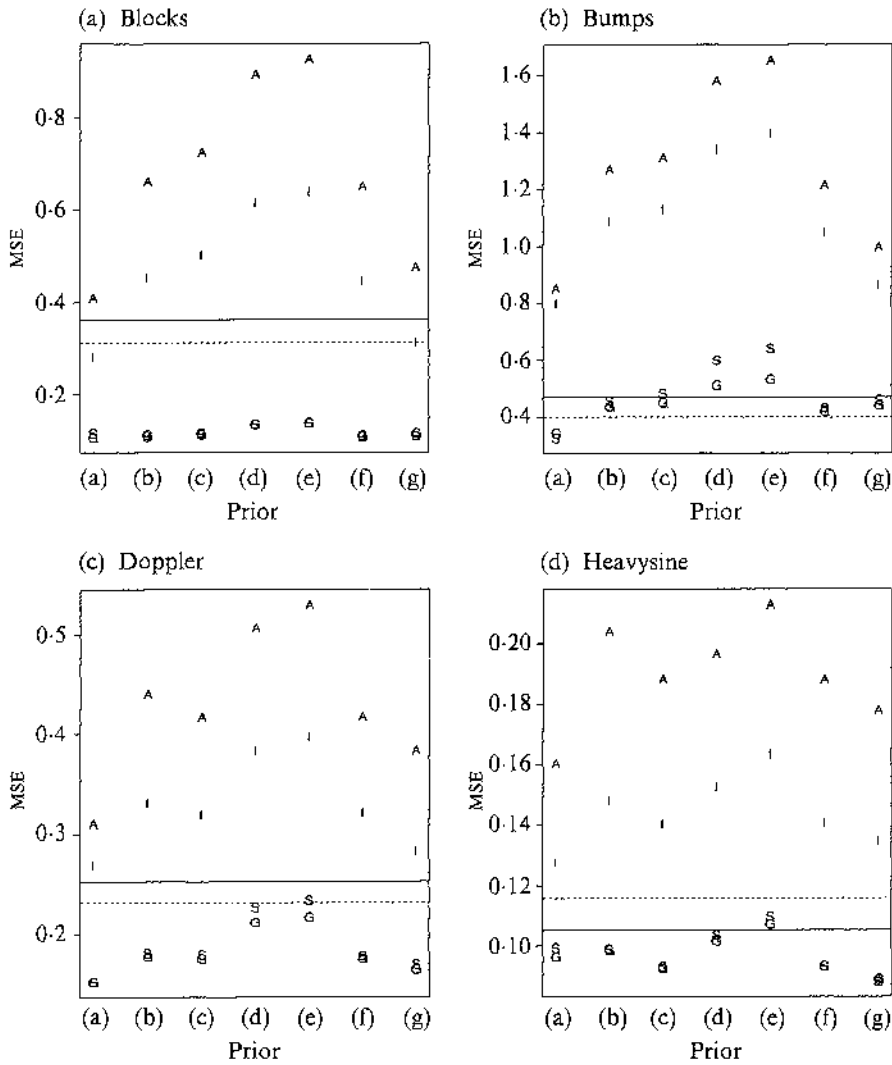


Fig. 1. Comparison of algorithms for the four test functions for methods G, S, I and A. The horizontal lines correspond to the thresholding rules: the continuous line is 'Hard', the dotted line is 'Soft', the dashed line is 'Sure'.

Figure 1 summarises the results of the simulations. The performances of the Gibbs sampler G and the analytical approximation S are very close to each other and consistently outperform all alternatives. Accounting for uncertainty about  $\sigma$  does not appear to be essential at this level of SNR and  $N$ , because the posterior distribution of  $\sigma$  is very concentrated. The performance of the importance sampling method I and the related analytical approximation A is consistently worse than the performance of G and S in this scenario. Inspection of the importance sampling weights reveals a high variance. A further factor contributing to the better performance of G compared to I is the Rao-Blackwellisation

estimator (11) for posterior probabilities. Since the model space is large, containing  $2^{1024}$  models, convergence of the importance sampling estimates is slow compared to the Rao–Blackwellised estimator. Posterior probabilities may be estimated as zero, resulting in hard thresholding.

In addition we generated data from the same functions using an  $\text{SNR} = 3$  and Student- $t$  errors with 5 degrees of freedom to generate data with outliers. For brevity, only results for prior (c) will be discussed. All methods show a higher MSE compared to normal errors with the same SNR, but their relative standings remain roughly unchanged, with the multiple shrinkage estimators doing better overall than the standard approaches. For both ‘heavisine’ and ‘doppler’, the Gibbs estimator G had the overall best MSE, 0.125 and 0.234, respectively compared to Soft, which had the best MSE of the standard methods with 0.131 and 0.364, respectively. In ‘blocks’, G still did better than the best standard method, 0.312 for G versus 0.446 for Sure, but both S and I had better mean squared errors than G. For ‘bumps’ the analytical approximation was the only method that did better than Sure, 0.551 compared to 0.568. However, S, G and I all did better than Hard or Soft.

### 5. GLINT DATA EXAMPLE

To illustrate further features of the multiple shrinkage approach proposed here we used the glint data presented by Bruce & Gao (1994, Ch. 5, p. 2). The data series includes 512 equally spaced observations. The true signal is a low frequency oscillation about zero, resulting from rotating an aircraft model. As the model rotates, the centre of mass will shift slightly. The measurements, given in angles, are subject to large errors, and can be off the truth by as much as  $150^\circ$ .

Throughout, we used the least asymmetric Daubechies 8-tap mother wavelet ‘s8’, which is the default in S + Wavelets. This basis provides a sensible balance between the compactness of support and smoothness of the wavelet. In Fig. 2 we illustrate the fit resulting from universal thresholding. The result may not be satisfactory if it is believed that the true signal should be smoother and that many of the large spikes represent noise. In order to get more realistic estimates, Bruce & Gao (1994) used simple smoothing techniques before applying wavelet shrinkage.

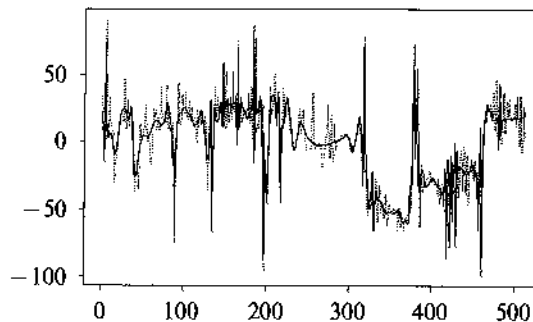


Fig. 2. Glint data, dotted line, and fitted signal, solid line, based on the universal thresholding rule.

In our multiple shrinkage estimator, we can incorporate the prior information about the noise. For the multiple shrinkage estimator, the prior hyperparameters for  $\sigma$  are given by  $\lambda = 1000$  and  $\nu = 22$ , reflecting the belief that errors can be as large as 100. The choice of  $c$  is based on the George and Foster risk inflation criterion rule discussed in § 2, leading

to  $c = 262\,130.5$ . The prior distribution for  $\gamma$  was based on  $\theta_j = 2^{j-8}$  for  $j = 1, \dots, 6$  and  $\theta_7 = 0.9999$ , so that coefficients from the smooth space are included with a probability near one. This prior distribution penalises inclusion of coefficients at the finest level of detail,  $j = 1$ .

Figure 3 shows the resulting posterior estimates from methods A, I, S and G. The two Monte Carlo approaches, G and I, were run for 50 000 iterations, although estimates of the posterior means stabilise much earlier. There was little qualitative difference between the estimates based on I and G, which are both computationally more intensive than method A. Method A, however, results in less shrinkage than the other two. The analytical approximation S gives a fit that includes many more coefficients than that of the other three methods, primarily because it does not incorporate any of the prior information or uncertainty about  $\sigma$ , resulting in a fit more similar to hard thresholding. The other methods all take into account that  $\sigma$  is unknown. Ignoring uncertainty about  $\sigma$  also leads to much narrower prediction intervals for S compared to G; see Fig. 3(b) and (c).

Finally, Fig. 4 considers the shrinkage curves resulting from two prior specifications; the first prior being the one previously considered. The second prior is based on selecting  $c$  based on the BIC criterion with a uniform prior distribution for  $\gamma$ ,  $\theta_{jk} = \frac{1}{2}$ . This leads to  $c = 3.92$ . The prior distribution for  $\sigma$  is the same as in the first prior. The shrinkage curves display nonlinearities from multiple shrinkage. The non-monotonicities in the overall

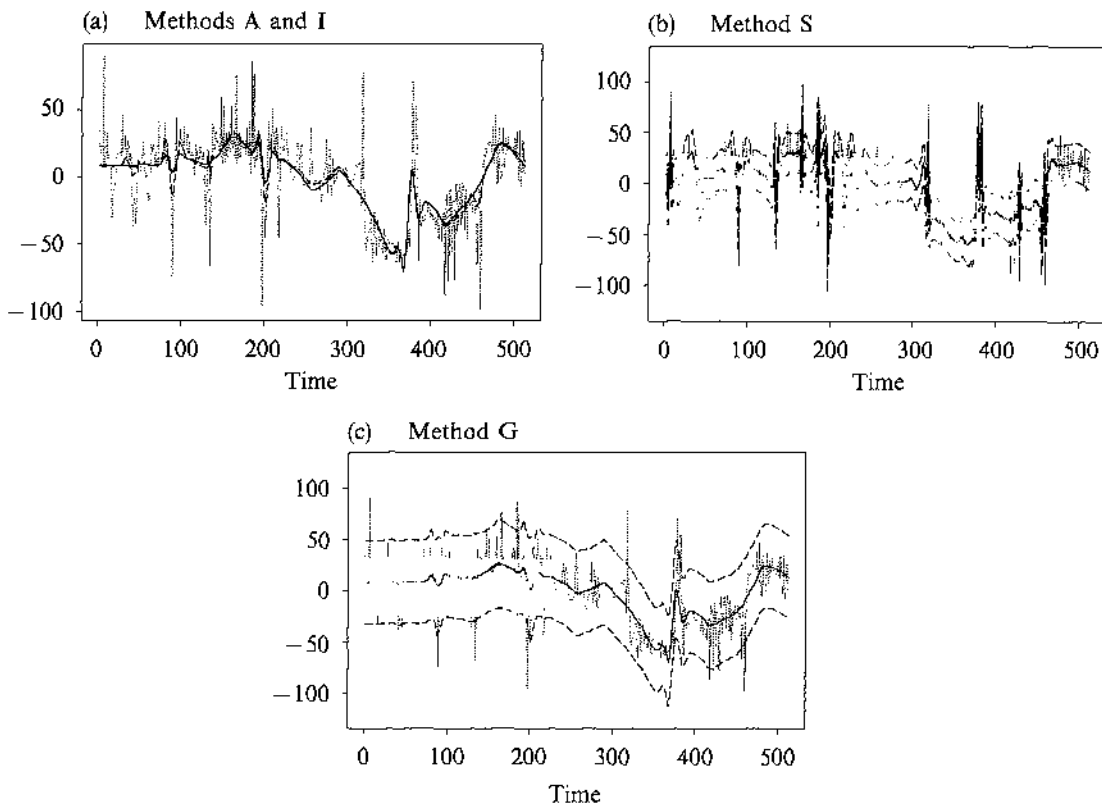


Fig. 3. Glint data. (a) Estimates resulting from the analytic approximation A, dashed line, and the importance sampling, solid line. (b) Posterior mean, solid line, and prediction intervals about  $\hat{f} \pm$  two standard errors, dashed lines, using the analytic approximation S. (c) Posterior mean, solid line, and prediction intervals about  $\hat{f} \pm$  two standard errors, dashed lines, using Gibbs sampling.

shrinkage curve for the first prior arise from the level dependent prior distribution on  $\gamma$ , which allows a stronger shrinkage at higher level of detail. Within a specific level, however, shrinkage is monotonic.

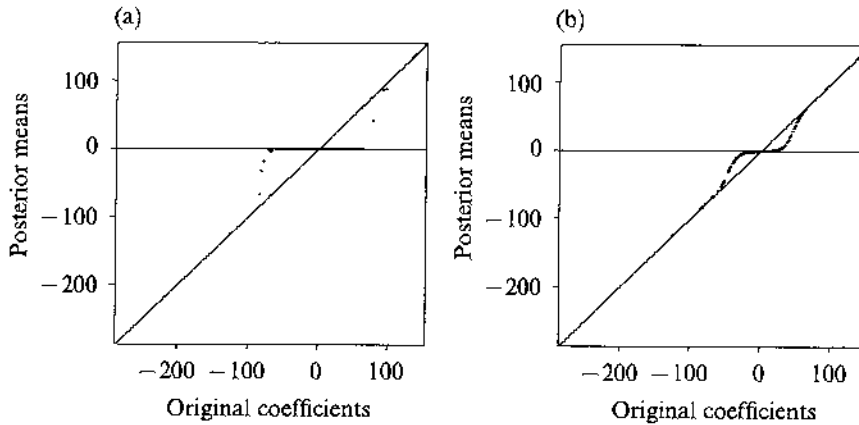


Fig. 4. Glint data. Effect of the choice of prior distributions on shrinkage. (a) Prior previously considered, and (b) BIC prior.

## 6. DISCUSSION

We conclude by discussing some potential developments of our approach. The model can be extended from a single fixed basis to include wavelet packets and other over-complete libraries that can produce a variety of orthonormal bases. For example, the wavelet packet library, or packet table, contains many different orthonormal bases. Wavelet packets are often preferred to wavelets for representing signals that exhibit oscillatory or periodic behaviour. One can add the choice of basis to the model hierarchy and use the stochastic selection methods described to select the best subspace from the collection of orthonormal bases or to provide multiple shrinkage estimators, where one is now additionally averaging over different bases. Non-white noise is a problem in many applications. Modelling the error terms as Student- $t$  errors or other heavy tailed distributions can be achieved by using scale mixtures of normals and extending the hierarchical model. While computationally more intensive, Gibbs sampling and model averaging may provide performance superior to the current alternatives.

## ACKNOWLEDGEMENT

Our work has been partially supported by the U.S. National Science Foundation. We thank Dean Foster and Ed George for helpful discussions and for kindly sharing unpublished results. We thank Andrew Bruce and Hong-ye Gao from StatSci, who provided background information about the glint dataset.

## REFERENCES

- AKAIKE, H. (1978). A new look at the Bayes procedure. *Biometrika* **65**, 53–9.  
 BRUCE, A. & GAO, H-Y. (1994). *S + Wavelets, Users Manual*. Seattle: StatSci.

- CHIPMAN, H., KOLACZYK, E. & MCCULLOCH, R. (1997). Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Assoc.* **92**, 1413–21.
- CLYDE, M., DESIMONE, H. & PARMIGIANI, G. (1996). Prediction via orthogonalized model mixing. *J. Am. Statist. Assoc.* **91**, 1197–208.
- DAUBECHIES, I. (1988). Orthonormal bases of compactly supported wavelets. *Commun. Pure Appl. Math.* **41**, 909–96.
- DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.
- DONOHO, D. & JOHNSTONE, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.
- DONOHO, D. & JOHNSTONE, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–24.
- DONOHO, D., JOHNSTONE, I., KERKYACHARIAN, G. & PICARD, D. (1995). Wavelet shrinkage: Asymptopia? (with Discussion). *J. R. Statist. Soc. B* **57**, 301–69.
- DRAPER, D. (1995). Assessment and propagation of model uncertainty (with Discussion). *J. R. Statist. Soc. B* **57**, 45–98.
- FOSTER, D. & GEORGE, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–75.
- GEORGE, E. (1986). Combining minimax shrinkage estimators. *J. Am. Statist. Assoc.* **81**, 437–45.
- GEORGE, E. I. & MCCULLOCH, R. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–9.
- GEORGE, E. I. & MCCULLOCH, R. (1997). Fast Bayes variable selection. *Statist. Sinica* **7**, 339–74.
- MALLAT, S. (1989). A theory for multi-resolution signal decomposition: the wavelet representation. *IEEE Trans. Pat. Anal. Mach. Intel.* **11**, 674–93.
- RAFTERY, A. E., MADIGAN, D. M. & VOLINSKY, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance (with Discussion). In *Bayesian Statistics 5*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 323–50. Oxford: Oxford University Press.

[Received November 1995. Revised September 1997]