Bayesian Wavelet analysis with a model complexity prior

C C Holmes D G T Denison

May 1998

Abstract

In this paper Bayesian methods for the selection and shrinkage of wavelet coefficients are considered. The wavelet coefficient are assigned a zero mean normal prior with precision matrix $\boldsymbol{\lambda}$. This allows for a more general model than has been previously studied and uncertainty in $\boldsymbol{\lambda}$ can be controlled through a hyperprior. It is argued that model interpretability is improved by foresaking the conjugate Wishart prior $\boldsymbol{\lambda}$ in favour of a non-conjugate hyperprior that penalises the degrees of freedom of the model. The degrees of freedom provides a single intuitive measure of model complexity that accommodates both the number of coefficients and the extent of the shrinkage on them. We compare three different sampling strategies and comment on their performance.
Keywords: degrees of freedom, shrinkage, smoothing, wavelets, variable selection.

1 Introduction

Wavelet analysis has quickly established itself as a standard method for the analysis and smoothing of time series. Of particular importance is the application to the denoising and compression of signals (e.g. Bruce and Gao, 1996).

In this paper we consider wavelets within a Bayesian framework. The main motivation is to generalise previous work (Clyde, Parmagiani and Vidavokic, 1998; Müller and Vidakovic, 1995) and to make prior assumptions that can be compared to conventional model selection criteria. In addition we compare the relative performance of three well-used sampling strategies and assess their relative merits.

Bayesian analysis begins with the elicitation of prior distributions on the wavelets coefficients that reflect subjective beliefs about the probable values of the coefficients. This infers prior beliefs about the nature of the underlying signal of interest. The most common prior, and the one we consider here, is a zero mean normal 'shrinkage' prior on the vector of wavelet coeffificients $\boldsymbol{\beta}$ with some precision matrix $\boldsymbol{\lambda}$. A suitable value of $\boldsymbol{\lambda}$ is rarely known *a priori* and setting it can prove problematic as $\boldsymbol{\lambda}$ ultimately controls the complexity, and hence smoothness, of the wavelet reconstruction.

We use a hierarchical Bayesian approach which accomodates uncertainty in $\boldsymbol{\lambda}$ by including a hyperprior. We refrain from using the conjugate Wishart prior $\boldsymbol{\lambda}$ because this suffers from a lack of interpretability, effectively diluting uncertainty on $\boldsymbol{\lambda}$ into uncertainty on this hyper-hyper prior. Instead we assign a non-conjugate prior distribution over $\boldsymbol{\lambda}$, motivated by the the degrees of freedom of the wavelet network. We choose this as it offers a natural vehicle for expressing prior knowledge about the underlying function.

Many classical model choice criteria are defined in terms of degrees of freedom. We show that these can be used to aid prior elicitation and leads to the most probable wavelet networks being given by the network that minimises the respective classical model choice criteria.

The rest of the paper is as follows. In section 2 we briefly overview wavelet decomposition and some classical shrinkage and selection procedures. In section 3 we present the Bayesian analysis where it is shown that the degrees of freedom presents a natural measure of model complexity. Section 4 discusses different methods for sampling from the model space. Section 5 presents some results and in Section 6 we give a final discussion.

2 Classical Wavelet Analysis

2.1 Wavelet bases

Regression is concerned with modelling the dependence of the response variable Y on the predictor variable X with the data given at a number of locations $\mathcal{D} = (y_i, x_i)_{i=1}^N$. We wish to estimate the true regression function f, which is modelled by the relationship

$$Y = f(X) + \epsilon,$$

where ϵ is commonly taken to be Gaussian white noise. Note that, without loss of generality, we shall assume that the $x_i \in [0,1]$ and we wish to estimate the regression function only over this interval.

Wavelet methods estimate f via the wavelet expansion

$$\widehat{f} = \beta_0 \chi\{x \in [0,1]\} + \sum_{j=0}^{n-1} \sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x)$$

where the β_{jk} are the wavelet coefficients with corresponding wavelet basis functions ψ_{jk} and n is chosen so that $N = 2^n$, and $\chi(\cdot)$ is the usual indicator function.

The wavelet basis functions are constructed by translating and scaling (dilating) what is referred to as the *mother* wavelet, ψ. Thus,

$$\psi_{jk} = 2^{-j/2}\psi(2^{-j}x - k),$$

where j is the dilation number and k the translation number. As bases with the same value of j are just translations of each other j is often referred to as the resolution level.

The mother wavelet is chosen so that the ψ_{jk} are all orthogonal and have compact support as it is these features that give wavelet bases their good properties. The simplest wavelet that satisfies these condition is the well known Haar wavelet which takes the value 1 for $x \in [0, \frac{1}{2}]$, -1 for $x \in (\frac{1}{2}, 1]$ and zero otherwise. It was only recently discovered that other mother wavelets also produced bases with the required properties.

2.2 Wavelet shrinkage

Nonparametric regression using wavelets is undertaken in three basic steps: (i) apply a discrete wavelet transform (DWT) to the responses; (ii) 'shrink' the empirical wavelet coefficients and (iii) invert the DWT to get $\widehat{f}$, the estimate to f. Steps (i) and (iii) are straightforward to carry out but step (ii) is more problematic. This step is crucial as it is the way the shrinkage of the coefficients takes place that determines the properties of $\widehat{f}$. As is common in regression a balance has to be struck between the bias and variance of the estimate; that is between oversmoothing and undersmoothing $\widehat{f}$.

The most widely used shrinkage procedures involve determining a *threshold* value (or values) and then applying a shrinkage function to the coefficients using these thresholds. Examples of threshold values are the universal and minimax thresholds and the most commonly used threshold functions are the hard and soft rules. All of these are discussed in more detail in Donoho and Johnstone (1994).

3 Bayesian Wavelet Analysis

Bayesian wavelet analysis is particulary straightforward due to the linear nature of the wavelet model. We can draw on standard results from the Bayes linear model (Lindley and Smith, 1972) in our analysis. Under the assumption of Gaussian white noise we have

$$\begin{aligned} Y &= N(W\boldsymbol{\beta}, \sigma^2 I) \\ \boldsymbol{\beta} &= N(\mathbf{0}, \boldsymbol{\lambda}^{-1}) \end{aligned}$$

where $\sigma^2 I$ is the variance of the noise process, $\boldsymbol{\lambda}$ is the prior precision matrix of $\boldsymbol{\beta} = (\beta_{jk})$ and W is the design matrix of the using the wavelet basis functions (ψ_{jk}). Using standard results (Bernardo and Smith, 1994) we know that $p(\boldsymbol{\beta}|Y) \sim N(\widehat{\boldsymbol{\beta}}, \boldsymbol{V})$ where the posterior variance $\boldsymbol{V}$ is given by

$$\boldsymbol{V} = (\sigma^{-2} W'W + \boldsymbol{\lambda})^{-1}, \tag{1}$$

and the posterior mean $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\sigma^{-2} W'W + \boldsymbol{\lambda})^{-1} W' \sigma^{-2} Y. \tag{2}$$

These expressions simplfy because $W'W = I$ due to the orthonormality of the wavelet transform.

Although the method we propose is more general, it is helpful to consider the special case where we assign independent priors on the coefficients. This is completely justified due to the orthoganility of the wavelet basis functions and will be adopted in the rest of the paper. Hence, we choose $\boldsymbol{\lambda} =$ diag $(\lambda_1, \ldots, \lambda_N)$. This reveals the effect of each the prior precision component λ_i, which is to shrink the classical least squares estimate β_i^* by $(1 + \sigma^2\lambda_i)^{-1}$.

This also shows the importance of $\boldsymbol{\lambda}$ to the Bayesian wavelet model. The conjugate Wishart prior for $\boldsymbol{\lambda}$ does little to aid our elicitation of prior knowledge. Significantly, our true beliefs lie on the complexity and smoothness of the underlying signal that we are trying to reconstruct which is infered by a prior on $\boldsymbol{\beta}$, rather than the values of the parameters *per se*. This motivates us to examine further the relationship between $\boldsymbol{\lambda}$ and the model fit. First note that the mean of the posterior predictive density is $\widehat{f} = W\widehat{\boldsymbol{\beta}}$ which, from (2), can be written as $\widehat{f} = SY$ where S is the smoothing matrix given by $S = \sigma^{-2}W(\sigma^{-2}I + \boldsymbol{\lambda})^{-1}W'$.

The degrees of freedom of a linear smoother is given by the trace of the smoothing matrix (Hastie and Tibshirani, 1990). and we use this as a natural measure of model complexity. Thus, for our model the degrees of freedom is given by

$$DF = \sum_{i=1}^{N} \frac{1}{1 + \sigma^2\lambda_i}.$$

As we have a prior preference for smooth models with low degrees of freedom a natural choice for the prior on λ_i is

$$p(\lambda_i|\sigma^2) \propto \exp(-c\sum_{i=1}^{N}(1 + \sigma^2\lambda_i)^{-1}) \tag{3}$$

for some constant c which determines how much we penalise model complexity. This prior is chosen because, after analytically integrating out the wavelet coefficients $\boldsymbol{\beta}$, we find that

$$\text{Log Model Probability} = \text{Log Marginal Likelihood } - (c \times DF) \tag{4}$$

which has the form of many classical model choice criteria allowing us to choose c to correspond with these methods. The four criteria we shall compare in this paper are

1. $c = 0$. The uniform prior results in the ratio of model probabilities being given by the Bayes factor, see Kass and Raftery (1995).

2. $c = 1$. Akaike's Information Criteria (AIC), (Akaike, 1974).

3. $c = 1/2 \log n$. Bayesian Information Criteria (BIC), (Schwartz, 1978).

4. $c = \log n$. The Risk Inflation Criteria (RIC), (Foster and George, 1995).

Unfortunately, the prior in (3) is improper so we must constrain $\boldsymbol{\lambda}$ to some large range. In this paper we took $10^{-10} < \lambda_i < 10^{10}$. Over this range each λ_i results in effectively no shrinkage on the associated wavelet coefficient, up to the coeffificient being effectively removed from the model.

Note that we do not attempt to perform level dependent shrinkage because this appears to be inconsistent with our knowledge that under the wavelet decomposition, the noise enters additively across all coefficients (sec 6.2, Bruce and Gao, 1996).

4 Three Samplers

The model probability is given by the marginal likelihood and a prior on the degrees of freedom of the model. In this section we consider three methods of sampling from the posterior model space. The first approach is to fix λ_i small for all i so that little shrinkage is performed, and then use a varying dimensional reversible jump sampler (Green, 1995) to sample appropriate models by adding or removing wavelet bases. We refer to this as the "selection" approach. The second method is to use the full model and sample $\boldsymbol{\lambda}$ using a random walk Metropolis-Hastings (MH) algorithm (Hastings, 1970): this is termed the "shrinkage" method. Finally we can construct a hybrid sampler based on both the varying dimension and MH samplers which can be considered as a "shrink and select" process.

All three methods are easy to implement. For the varying dimension sampler we take $\lambda_i = 10^{-5}$. At each time step we randomly select one of the N possible wavelets. If the wavelet is currently in the model then we propose to remove it. If it is absent then we propose to add it in. The full Bayesian model, uses a normal kernel to propose changes to $\log \lambda_i$ subject to the constraint that $10^{-10} < \lambda_i < 10^{10}$. The variance of the proposal density is chosen during a pre-sample stage to obtain acceptance rates of around 30%. The hybrid method atempts to change the dimension of the model at each step with probability 0.2 otherwise it proposes a MH update to $\boldsymbol{\lambda}$ in the current model. If a proposal is made to add a wavelet then the value of the new λ_i is drawn from the prior using a simple rejection sampler (Ripley, 1987).

For the first and second methods where either the dimension or the precision is sampled (but not both) the acceptance ratio for proposing a move from model $\mathcal{M}$ to model $\mathcal{M}'$ is just

$$\alpha = \min \left(1, \frac{|\boldsymbol{\lambda}'|^{1/2}|\boldsymbol{V}'|^{1/2}a'}{|\boldsymbol{\lambda}|^{1/2}|\boldsymbol{V}|^{1/2}a} \times \frac{p'}{p}\right) \tag{5}$$

where $|\boldsymbol{V}|$ is the determinant of $\boldsymbol{V}$ given in (1), and a is given by

$$a = \exp\left(-\frac{1}{2\sigma^2}[\tau_2 + y'y - \widehat{\boldsymbol{\beta}}\boldsymbol{V}\widehat{\boldsymbol{\beta}}]\right) \tag{6}$$

where τ_2 comes from the prior on the noise precision σ^{-2}, assigned a Gamma, $\Gamma(\tau_1, \tau_2)$ prior and p in (5) is taken from the prior and is given by

$$p = \exp(-c\sum_{i=1}^{k}(1+\sigma^2\lambda_i)^{-1})) \tag{7}$$

where k is the number of wavelet coefficients in the model $\mathcal{M}$ with prior p.

For the hybrid method, the *constant* dimension sampling uses (5) as an acceptance ratio. However, for the *varying* dimension proposal, the prior term p'/p in (5) is one as we are drawing the new value of λ_i from the prior.

The terms in (6), (7) are conditioned on the noise variance σ^2 which is drawn at each iteration using a Gibbs step as we know the full conditional is just $p(\sigma^{-2}|y) =$ Ga $(n/2 + \tau_1, y'y - \tau_2 + y'y - \widehat{\boldsymbol{\beta}}\boldsymbol{V}\widehat{\boldsymbol{\beta}})$ where n is the number of data points (Bernardo and Smith, 1994).

5 Simulations

The three sampling methods and the four default priors were tested on two standard wavelet data sets. These are the "bumps" and "heavisine" functions investigated by Donoho and Johnstone (1994). We generated 1024 datapoints from each function and following the approach of Donoho and Johnstone (1994) Gaussian white noise was added to make the signal to noise ratio 7.

Each sampling method and prior were tested for five runs with different random seeds. Each run consisted of a burn in period of 200,000 iterations with a following sampling period of 300,000 of which every seventh sample was retained.

Some posterior mean statistics are given in Tables 1-2 for the various default priors and for each of the three sampling methods: (a) selection; (b) selection and shrinkage and (c) shrinkage. In Fig. 1 we display the resulting estimates for each method which demonstrates the smoothness obtained under models (b) and (c) and the problems associated with selection using a common shrinkage parameter. For comparison, the classical waveshrink procedure in *S plus* was run using universal thresholding. This gave ASE scores of 0.138 and 1.257 for the Heavisine and Bumps data respectively and, 0.127 and 0.614 with the minimax thresholding rule.

In Fig. 2 we show the posteior mean shrinkage of each the coefficients for the full Bayes model (c) using RIC. The shrinkage is clearly nonlinear and demonstrates that the large coefficients are hardly shrunk while many of the small ones are effectively set to zero. This method strikes a balance between the hard and soft thresholding rules of Donoho and Johnstone (1994).

6 Discussion

We suggest that the degrees of freedom offers an explicit measure of model complexity that can be used to aid our prior elicitation of model hyperparameters. The degrees of freedom captures the effect of both the number of coefficients in the model and the extent of the prior "shrinkage" on them. This allows us to perform both selection and generalised shrinkage on each coefficient together.

This use of a full precision matrix $\boldsymbol{\lambda}$ with individual precision parameters λ_i for each wavelet coefficient produces a richer and more general model structure. The independence of the priors, with diagonal $\boldsymbol{\lambda}$, is appropriate for the assumption of i.i.d noise in an othogonal basis set.

When storage is not an issue it seems to us that there is little justification, in a Bayesian analysis, for removing coeffcients from the model. The orthogonality of the wavelet transform means that working with the full model holds only minor computational overheads. The gains are that no information is discarded. It was shown in Figure 2 that the full model with full adaptive precision matrix $\boldsymbol{\lambda}$ can effectivly perform coefficient selection.

References

Akaike, H (1974) A new look at statistical model identification. *IEEE Trans. Automatic Control*, 19, 716-727.

Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory.* New York: Wiley.

Bruce, A. and Gao, H-Y. (1996) *Applied Wavelet Analysis with S-Plus.* New York: Springer-Verlag.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998) Multiple shrinkage and subset selection in wavelets. *Biometrika*, **85**, 391-402.

Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaption by wavelet shrinkage. *Biometrika*, **81**, 425-55.

Foster, D and George, E.I. (1994) The risk inflation criterion for multiple regression. *Annals of Statistics* 22:1947-1975.

Green, P.J. (1995) Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models.* London: Chapman and Hall.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. B* **34**, 1-41.

Müller, P. and Vidakovic, B (1995) Bayesian inference with wavelets: Density estimation. Discussion Paper, 95-33, ISDS, Duke University.

Ripley, B.D. (1987) *Stochastic Simulation.* New York: Wiley.

Schwartz, G. (1978) Estimating the dimension of a model. *Annals of Statistics.* 6, 461-464.

Prior	ASE	DF	No. Wavelets
RIC (a)	0.169	50.29	50.51
(b)	0.072	29.67	496.95
(c)	0.073	33.95	1024
BIC (a)	0.182	63.57	63.88
(b)	0.072	35.82	494.86
(c)	0.073	45.20	1024
AIC (a)	0.193	107.32	107.91
(b)	0.078	51.15	443.5
(c)	0.092	81.10	1024
ML (a)	0.198	163.44	164.48
(b)	0.087	62.67	381.12
(c)	0.114	119.54	1024

Table 1: Results for Heavisine data. ASE refers to avarege squared error of the reconstruction against the true function. The posterior mean degrees of freedom (DF) for the models along with the mean number of wavelet coefficients are shown in the third and forth columns.

Prior	ASE	DF	No. Wavelets
RIC (a)	0.330	142.80	143.37
(b)	0.319	128.01	574.42
(c)	0.319	133.42	1024
BIC (a)	0.331	154.57	155.32
(b)	0.320	136.05	568.83
(c)	0.311	147.82	1024
AIC (a)	0.332	194.02	196.28
(b)	0.306	154.37	514.77
(c)	0.304	185.95	1024
ML (a)	0.333	247.24	249.09
(b)	0.307	164.92	453.19
(c)	0.306	224.87	1024

Table 2: Results for Bumps data. ASE refers to avarege squared error of the reconstruction against the true function. The posterior mean degrees of freedom (DF) for the models along with the mean number of wavelet coefficients are shown in the third and forth columns.

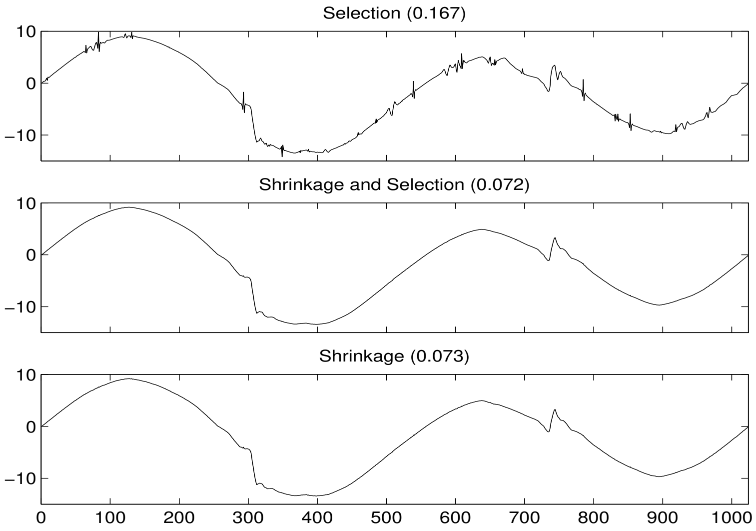

Figure 1: Estimates for the heavisine data using the RIC prior.

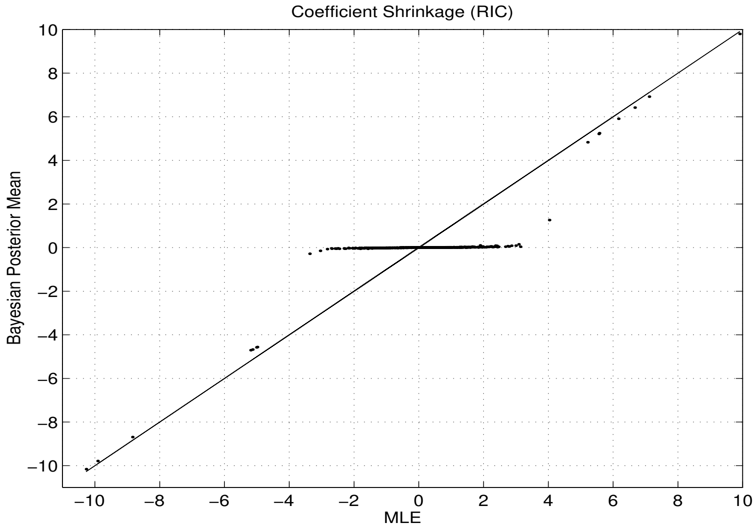

Figure 2: Shrinkage using the RIC prior on the full model.