Journal of Modern Applied Statistical Methods

Volume 3 | Issue 1 Article 20

5-1-2004

A Visually Adaptive Bayesian Model In Wavelet Regression

Dongfeng Wu
Mississippi State University, dw183@ra.msstate.edu

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 200-212

1538 – 9472/04/$95.00

A Visually Adaptive Bayesian Model In Wavelet Regression

Dongfeng Wu
Department of Mathematics and Statistics
Mississippi State University

The implementation of a Bayesian approach to wavelet regression that corresponds to the human visual system is examined. Most existing research in this area assumes non-informative priors, that is, a prior with mean zero. A new way is offered to implement prior information that mimics a visual inspection of noisy data, to obtain a first impression about the shape of the function that results in a prior with non-zero mean. This visually adaptive Bayesian (VAB) prior has a simple structure, intuitive interpretation, and is easy to implement. Skorohod topology is suggested as a more appropriate measure in signal recovering than the commonly used mean-squared error.

Key words: Wavelet regression, wavelet shrinkage, optimal, Skorohod topology, uniform distance, mean-squared error

Introduction

Wavelets unify many ideas from the fields of applied mathematics, signal processing, and physics (see Daubechies 1992). Wavelets are families of basis functions that can be used to approximate other functions, with powerful properties such as orthonormality, compact support, localization in time and scale, etc. Daubechies (1988) and Mallat (1989) encouraged the use of wavelets in the mathematical sciences, while Donoho and Johnstone (1994, 1995) popularized wavelets in the statistics community.

Some of the uses of wavelets for statistical problems have been developed by Donoho and Johnstone (1993, 1994) and Nason (1994) and are available in the S+ package. More recent work includes the block thresholding method of Cai (1999), which achieves adaptivity, spatial adaptivity and computational efficiency simultaneously.

The author wishes to thank David V. Hinkley for suggestions regarding this study. Most of the numerical work was conducted using *wavethresh* (Nason, 1994, Version 3). Contact the author at dw183@ra.msstate.edu

When fitting wavelet-based models, shrinkage of the empirical wavelet coefficients is an effective tool for denoising the data. Shrinkage of the empirical wavelet coefficients works best in problems where the underlying set of the true coefficients of f is sparse. One natural way to obtain the shrinkage estimates of the true coefficients is via Bayesian methods.

An appealing and simple model (ABWS) using the posterior mean has been proposed by Chipman, Kolaczyk, and McCulloch (1997) who assume that an accurate estimate of the noise level σ is available. A more complete Bayesian approach that captures the uncertainty about the noise level σ was proposed by Clyde, Parmigiani, and Vidakovic(1998). Abramovich, Sapatinas and Silverman (1998) proposed the posterior median method, with almost the same set up as Clyde et.al., but using the posterior medians to estimate the true coefficients. Huang and Cressie (1999) proposed a normal prior with non-zero means for wavelet coefficients, and estimated the hyper-parameters of the prior covariance by a pseudo maximum likelihood method.

A different prior structure with non-zero means is offered. The model is simple, combining a normal prior with non-zero mean and a point mass. Explanations are provided for each hyper-parameter in addition to a specific way to choose the prior parameters.

Methodology

The Bayesian model

Suppose the function f is sampled at $n = 2^J$ equally spaced points, but is observed with additive white noise,

$$y_i = f(i/n) + \sigma z_i, \qquad i=0, 1,\ldots, n-1, \tag{1}$$

where z_i, $i = 0,1,\ldots ,n-1$, are iid standard normal random variables, and σ is unknown. Equivalently this observation model can be expressed in wavelet regression form,

$$v_{j,k} = w_{j,k} + \sigma z_{j,k}, \quad j = 0, \ldots, J-1, \quad k= 0, \ldots, 2^j, \tag{2}$$

where $v_{j,k}$s are the discrete wavelet coefficients of noisy observation y; $w_{j,k}$s are the discrete wavelet coefficients of f; and $z_{j,k}$s are still iid N(0,1) random variables.

In the Bayesian approach, a prior distribution is placed on the coefficients, and some particular prior distributions that are designed to capture the sparseness common to most wavelet applications are proposed. Most of the published works in this area have a common characteristic, that is, a prior distribution is designed such that some of the mass is concentrated on values close to zero or just being zero, while the rest of the mass is spread to accommodate the possibility of large coefficients.

Then, the posterior means or the posterior medians are used as the estimates of the true coefficients. Though appealing, this framework assumes that all of the coefficients have the same prior in each level, with zero mean, which overlooks the facts that certain coefficients are significantly departs from zero. The overall shape of the curve gives us more useful information, and accommodation of this information will ease the procedure to denoise, and hence, recover the curve.

Inspired by the work of Chipman et al. (1997), Clyde et al. (1998), and Abramovich et al. (1998), and assuming that a good estimate of the obtained noise level σ, the following prior model is proposed:

$$w_{j,k} \mid \gamma_{j,k} \sim \gamma_{j,k} N(a_{j,k}, \tau_j^2) + (1-\gamma_{j,k})\delta(0) \tag{3}$$

In this prior model, the coefficients are mutually independent, and modeled as a mixture of a normal distribution and a point mass at zero. The innovation is that assumed is that the normal prior has non-zero mean $a_{j,k}$ for each coefficient $w_{j,k.}$ Also, a really small variance τ_j depends on each level j, so that each coefficient has a different prior associated with it.

This idea comes from the observation that when coefficients are changed in a small scale in each level, the function estimate won't change much, and it won't affect our visual perspective either. This means that each coefficient can change around its true value in a small scale, called its safety range, without any deleterious effects. This is captured in the form of $N(a_{j,k}, \tau_j^2)$, where $a_{j,k}$ is the prior information on the true value of the coefficient, and τ_j is the allowable perturbation on level j, so that the estimate would be close to the true function.

A point mass at zero is assumed based on the belief that the coefficients are sparse. This simple form of prior modeling has intuitive interpretations and captures the few big spikes in the coefficients. Empirical evidence shows that if $a_{j,k} = w_{j,k}$, $\forall j = 0, \ldots, J-1, k= 0, \ldots, 2^j$, the "recovered estimate" $\tilde{f}$ is a slight shift from the true f.

The mixture parameter $\gamma_{j,k}$ has its own prior distribution given by

$$\gamma_{j,k} \sim Bernoulli(p_{j,k}), \tag{4}$$

The prior parameters $a_{j,k}$, τ_j, $p_{j,k}$ need to be decided. A different prior is assigned for each individual coefficient, though in each level the coefficients share a common prior variance τ_j, which reflects the perturbation in level j.

Once data are observed, the wavelet coefficients of the signal y are distributed as

$$v_{j,k} \mid w_{j,k}, \sigma^2 \sim N(w_{j,k}, \sigma^2). \tag{5}$$

The posterior distribution on the (unobserved) true value of $w_{j,k}$, and use its expected value as the estimate. Then the inverse wavelet is applied transformation to get $\hat{f}$.

The Prior Parameters

In this section details are given on how to choose the values for each of the prior parameters. This prior seems more intuitive, and computer simulation demonstrates that it works well.

The intuitive meaning of $a_{j,k}$ is the prior mean of each coefficient. The value of a specific coefficient is not necessarily zero, but is determined by the overall shape of the signal; in other words, it is related to the first impression of the data. The *Universal* thresholding method is used to get the value $a_{j,k}$ for each coefficient. The *Universal* threshold value is generally bigger than all the other methods, and gives the overall shape of the data. Suppose a sound estimate exists of σ, say $\hat{\sigma}$, then for each level $j = 1, \ldots J$, let $t_j = \hat{\sigma}\sqrt{2\log(2^j)}$ according to the *Universal* rule, then

$$\begin{aligned} a_{j,k} &= T_{soft}(v_{j,k}, t_j) \\ &= \operatorname{sgn}(v_{j,k})(|v_{j,k}| - t_j) I(|v_{j,k}| > t_j) \end{aligned} \tag{6}$$

This process mimics a visual inspection of the noisy data whereby the first impression about the shape of the function is obtained. Using the threshold value as the empirical prior information of $a_{j,k}$ makes sense. Because this estimate is close to the true curve, only small perturbations are allowed, so the τ_j will be a small number compares to the scale in the same level j. It is believed that this τ_j is largely connected with the scales of the coefficients in the same level. Chosen was τ_j = 10% M_j based on previous empirical experience, where $M_j = \max_{0 \le k \le 2^j - 1}\{|v_{j,k}|\}$.

Usually, for a smaller signal-to-noise ratio, a bigger percentage is chosen to obtain τ_j; and a bigger signal-to-noise ratio means a smaller percentage to obtain τ_j. As for $p_{j,k}$, the probability that one specific coefficient is non-zero, also depends on the scales of the coefficients in that level. If $v_{j,k}$ is comparatively large, it is more likely that $w_{j,k} \neq 0$, and choose was $p_{j,k} = |a_{j,k} / M_j|$, which is the ratio of the absolute value of that coefficient over the largest one in that level. Now, the prior parameters for each coefficient are given.

In practice, the noise level σ is unknown and must be replaced by an estimate $\hat{\sigma}$. Used here is the slope estimate in Wu (2002), defined by

$$\hat{\sigma} = \frac{v_{(0.75n)} - v_{(0.25n)}}{z_{0.75} - z_{0.25}} \approx \frac{IQR}{0.6745 * 2}, \tag{7}$$

where $v_{(k)}$s are the order statistics of the highest level wavelet coefficients, $z_{0.75}$ and $z_{0.25}$ are the quantiles of the standard Normal distribution; n is the total number of coefficients in the highest level *J-1*, IQR is the inter-quartile range of the observed coefficients. Simulation studies show that this estimation is accurate in the applications (Wu, 2002).

Posterior Distribution of the Coefficients

Based on this model, it is derived that the posterior mean and variance of $w_{j,k}$ given the observation of noisy date Y, where $w_{j,k}$, $v_{j,k}$, $\gamma_{j,k}$, $a_{j,k}$, $p_{j,k}$, τ_j are simplified as w, v, γ, a, p, τ.

$$\begin{aligned} & E(w|v)\gamma E_{\gamma|v}[E(w|v,\gamma)] \\ &= P(\gamma = 1|v)E(w|v,\gamma = 1) \\ &+ P(\gamma = 0|v)E(w|v,\gamma = 0) \\ &= P(\gamma = 1|v)E(w|v,\gamma = 1). \end{aligned} \tag{8}$$

Because

$$w|v,\gamma = 1 \sim N\left(\frac{a\sigma^2 + v\tau^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right), \tag{9}$$

this implies

$$E(w|v,\gamma = 1) = \frac{a\sigma^2 + v\tau^2}{\sigma^2 + \tau^2}. \tag{10}$$

Because

$$P(\gamma=1|v) = \frac{p\pi(v|\gamma=1)}{p\pi(v|\gamma=1)+(1-p)\pi(v|\gamma=0)} \equiv \frac{O}{O+1}, \quad (11)$$

where

$$O = p\pi(v|\gamma=1)/[(1-p)\pi(v|\gamma=0)], \quad (12)$$

and because

$$\pi(v|\gamma=1) \sim N(a, \sigma^2+\tau^2), \quad (13)$$

$$\pi(v|\gamma=0) \sim N(0, \sigma^2), \quad (14)$$

when plugged into (12), the following

$$O = \frac{p}{1-p}\sqrt{\frac{\sigma^2}{\tau^2+\sigma^2}}\exp\left\{\frac{v^2}{2\sigma^2} - \frac{(v-a)^2}{2(\sigma^2+\tau^2)}\right\}, \quad (15)$$

and

$$E(w|v) = \frac{O}{O+1}\frac{a\sigma^2+v\tau^2}{\sigma^2+\tau^2}. \quad (16)$$

This is, the posterior mean of the coefficient. Then, apply the inverse wavelet transformation to obtain the function.

The posterior variance of a coefficient can be calculated similarly,

$$\begin{aligned}\mathrm{var}(w|v) &= E_{\gamma|v}[\mathrm{var}(w|v,\gamma)] + \mathrm{var}_{\gamma|v}[E(w|v,\gamma] \\ &= E_{\gamma|v}[\mathrm{var}(w|v,\gamma)] + E_{\gamma|v}[E(w|v,\gamma)^2] \\ &\quad -[E_{\gamma|v}E(w|v,\gamma)]^2\end{aligned} \quad (17)$$

where

$$\begin{aligned}E_{\gamma|v}[\mathrm{var}(w|v,\gamma)] &= P(\gamma=1|v)\,\mathrm{var}(w|v,\gamma=1) \\ &= \frac{O}{O+1}\cdot\frac{\sigma^2\tau^2}{\sigma^2+\tau^2},\end{aligned} \quad (18)$$

$$\begin{aligned}E_{\gamma|v}[E(w|v,\gamma)^2] &= P(\gamma=1|v)E(w|v,\gamma=1)^2 \\ &= \frac{O}{O+1}\cdot\left(\frac{a\sigma^2+v\tau^2}{\sigma^2+\tau^2}\right)^2,\end{aligned} \quad (19)$$

and

$$[E_{\gamma|v}E(w|v,\gamma)]^2 = [E(w|v)]^2 = \left(\frac{O}{O+1}\cdot\frac{a\sigma^2+v\tau^2}{\sigma^2+\tau^2}\right)^2. \quad (20)$$

Hence,

$$\mathrm{var}(w|v) = \frac{O}{O+1}\cdot\frac{\sigma^2\tau^2}{\sigma^2+\tau^2} + \frac{O}{(O+1)^2}\cdot\left(\frac{a\sigma^2+v\tau^2}{\sigma^2+\tau^2}\right)^2. \quad (21)$$

Results

Presented are some simulation results of different shrinkage methods. For estimation of *f*, the usual L_2 norm is used to evaluate performance. Let $f=\{f(x_i)\}_{i=1}^n$ and $\hat{f} = \{\hat{f}(x_i)\}_{i=1}^n$ be the vectors of true and estimated function values where x_i are equally sampled. Performance is measured by the average mean-squared error

$$R(\hat{f}-f) = \frac{1}{n}\left\|\hat{f}-f\right\|_{2,n}^2 = \frac{1}{n}\sum_{i=1}^n\left[\hat{f}(x_i)-f(x_i)\right]^2. \quad (22)$$

A smaller $R(\hat{f}, f)$ means a better estimation.

The optimal thresholding value is the value *t* that minimizes

$$M(t) = \sum_{i=1}^n\left[\hat{f}_t(x_i)-f(x_i)\right]^2 = \sum_{j,k}\left(\hat{w}_{j,k}-w_{j,k}\right)^2, \quad (23)$$

where $\hat{f}_t$ is the *t*-threshold estimator using soft-thresholding. The optimal value is an ideal that is not available in a practical problem because *f* is unknown; however, it is a benchmark.

To simplify the presentation, the following abbreviations are used for the several thresholding methods, as follows:

OPT: the level-dependent optimal thresholding method. ABWS: the adaptive Bayesian wavelet shrinkage method in Chipman et al. (1997).

MethodS: the multiple shrinkage MethodS in Clyde et al. (1998). VAB: the visually Adaptive Bayesian method presented here.

Eight testing functions were used as in Figure 1. The add iid N(0, σ^2) noise to each function to generate 1000 simulated noisy data sets, and run the ABWS, multiple shrinkage (MethodS) and the new method in Section 4 on these data sets. The parameters θ and c in MethodS is θ =(0.90, 0.90, 0.90, 0.90, 0.90, 0.50, 0.50, 0.50, 0.50, 0.05) and c=1048561 according to Clyde et al. (1998). The resulting L_2 deviations from the true function are summarized in Table 1.

In these eight simulations, ABWS performs best in the PIECEWISE polynomial and CORNER case, method S performs best in HEAVISINE and BUMPS, and our new VAB method performs best in the remaining four cases. In fact, in the case of BUMPS and SMOOTH signal, the performances of method S and our method are very close to each other; in the case of CORNER, the performances of ABWS and method S are very close to each other. Notice that in the case of DOPPLER, our VAB method slightly outperformed the level-dependent optimal soft-thresholding. There are a few other cases in which Bayes shrinkage is very close to the optimal soft-thresholding, such as, ABWS in the PIECEWISE polynomial case, VAB in the SMOOTH signal and CHIRP case, method S and ABWS in the CORNER case.

Simulation examples are plotted in Figures 2-9. In each figure, upper left is the noisy data; upper right is the signal recovered by ABWS, with real signal in dotted line; lower left is the signal recovered by method S with real signal in dotted line; lower right is the signal recovered by VAB, with real signal in dotted line.

An inspection of Figures 2-9 reveals some facts. ABWS tends to over-smooth the data, sometimes this over-smooth will cause a big departure from the original signal, as in the case of CHIRP and DOPPLER. MethodS and VAB both capture the coarse shape of the curve very effectively.

The L_2 norm might not be an appropriate value to measure performance. It is easy to find two estimates $\hat{f}_1$ and $\hat{f}_2$, such that $\left\|\hat{f}_1 - f\right\|_2 < \left\|\hat{f}_2 - f\right\|_2$, but visually $\hat{f}_2$ is preferred. It is not uncommon in our simulation study, because only a slight left or right shift of f will lead to this result.

This created a motivation to do more investigation to determine a measure that better reflects our visual system. Clearly distance plays a very important role in pattern recognition. Many books and papers on pattern recognition try to define picture similarity without success. In fact it is not understood what is truly meant by cognitive similarity. That is the underlying intuition. However, it was found that Skorohod topology might be a good choice.

Let $D[0,1] = \{f;\ f{:}[0,1] \to R^1$, with properties 1) to 3)$\}$, where properties 1) to 3) are defined as follows:

$$\begin{aligned} &1)\ \lim_{u \downarrow t} f(u) = f(t+) = f(t), \forall 0 \le t < 1, \\ &2)\ \lim_{u \uparrow t} f(u) = f(t-), \forall 0 < t \le 1, \\ &3)\ f(1-) = f(1). \end{aligned} \tag{24}$$

Denote $\Lambda = \{\lambda; \lambda : [0,1] \mapsto [0,1],$ is a 1-1 monotone continuous mapping$\}$, and denote $\Lambda_\varepsilon = \{\lambda \in \Lambda; \sup_{t\in[0,1]} |\lambda(t) - t| \le \varepsilon\}$, then for any $f,g \in D[0,1]$, define

$$Sk(f,g) = \inf\left\{\varepsilon > 0; \exists \lambda \in \Lambda_{\varepsilon,} \sup_{t\in[0,1]} \left|f(t) - g(\lambda(t))\right| \le \varepsilon\right\}. \tag{25}$$

The Skorohod distance considers the distance between two functions after translating or revolving them, and describes the similarity of functions very well. For details, see Billingsley (1968).

The Skorohod distance is more reasonable in describing the difference between broken functions by considering the uniform distance between two functions after doing a monotone continuous lengthening or shortening to the independent variables of the functions. It introduces a certain level of invariance to distortions and translations.

Table 1: Results of 1,000 simulations using L_2 – the mean-squared error.

Applications	σ	$(1/1000)\sum_{i=1}^{1000} \|\|\hat{f}_i - f\|\|_2^2$			
		ABWS	methodS	new method	OPT
SMOOTH	0.1	38.81(10.30)	9.90(3.23)	9.41(2.94)	7.14(2.24)
PIECEWISE	0.1	9.09(1.97)	12.36(2.39)	12.23(2.39)	8.76(1.64)
CHIRP	0.1	748.2(55.3)	45.48(5.96)	36.47(4.31)	35.92(2.83)
CORNER	0.1	11.23(2.26)	11.65(2.23)	12.26(2.29)	9.41(1.74)
BLOCKS	0.2	930.4(103.)	187.3(21.5)	154.1(16.0)	137.7(10.4)
BUMPS	0.3	1325.(118.)	360.3(35.5)	360.6(37.4)	275.8(20.1)
DOPPLER	0.2	1075.(140.)	106.1(14.8)	99.83(11.7)	101.1(9.32)
HEAVISINE	0.3	409.8(73.2)	97.09(13.8)	102.4(14.0)	82.42(11.3)

The unit is 10^{-4}; standard deviations are inside ().

Table 2: Uniform distance of the same 1,000 simulations.

Applications	σ	$(1/1000)\sum_{i=1}^{1000} d(\hat{f}_i, f)$			
		ABWS	methodS	new method	OPT
SMOOTH	0.1	15.56(2.35)	9.41(4.07)	8.39(1.79)	7.21(1.45)
PIECEWISE	0.1	28.00(2.57)	25.50(3.10)	24.09(1.91)	23.37(2.66)
CHIRP	0.1	94.49(1.63)	51.29(9.12)	40.70(5.61)	35.73(4.61)
CORNER	0.1	22.42(4.44)	18.62(3.98)	21.84(3.72)	17.58(3.35)
BLOCKS	0.2	179.7(8.13)	92.75(16.0)	77.93(11.0)	70.08(9.46)
BUMPS	0.3	346.9(12.9)	125.3(18.8)	150.2(21.3)	107.0(14.8)
DOPPLER	0.2	136.8(6.20)	83.94(14.6)	77.11(11.6)	64.56(8.67)
HEAVISINE	0.3	103.8(5.64)	104.5(13.1)	94.42(8.66)	89.81(12.6)

The unit is 10^{-2}; standard deviations are inside ().

Figure 1: **The test functions**

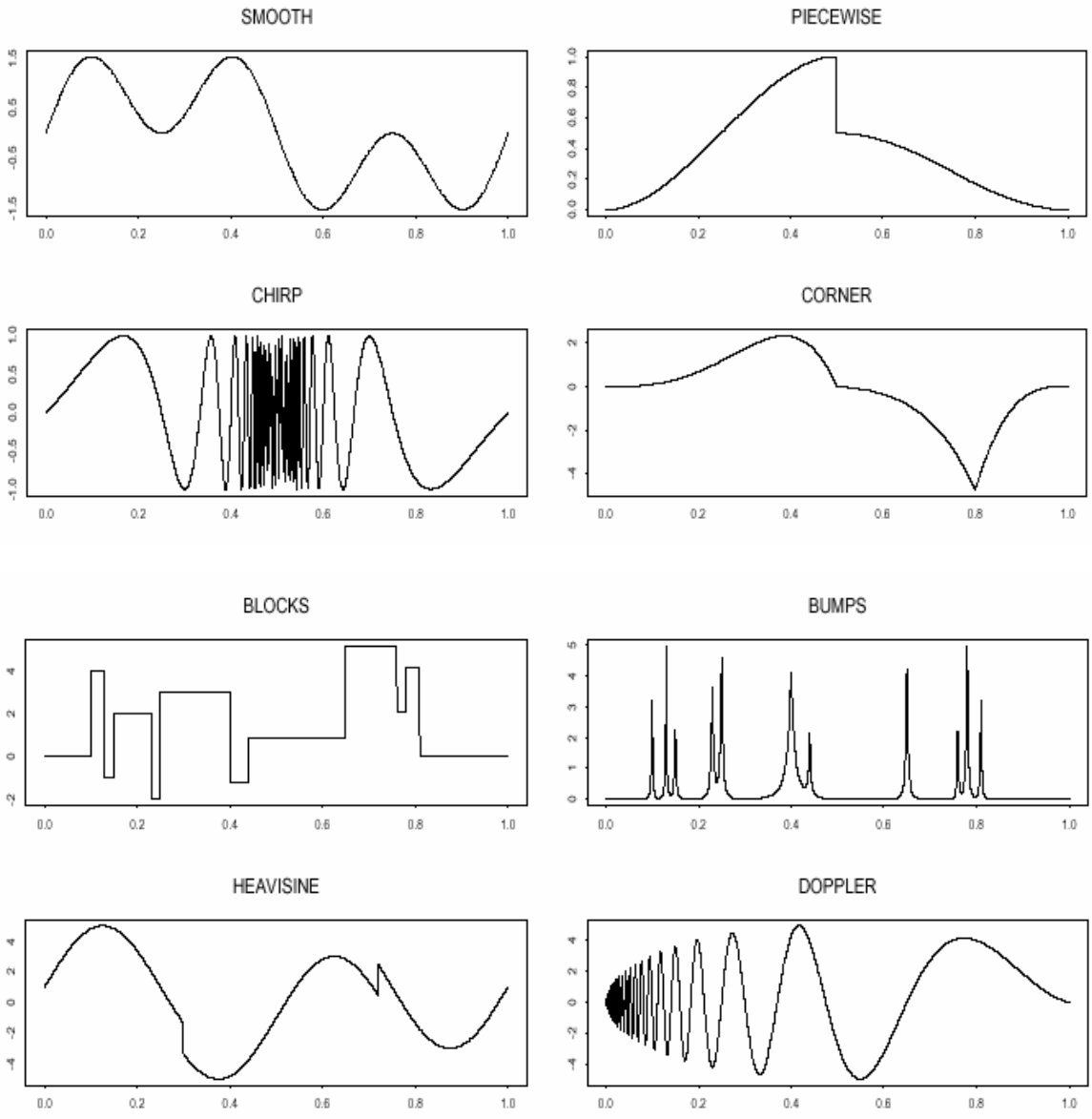


Figure 2: Smooth signal data, with $\sigma = 0.1$

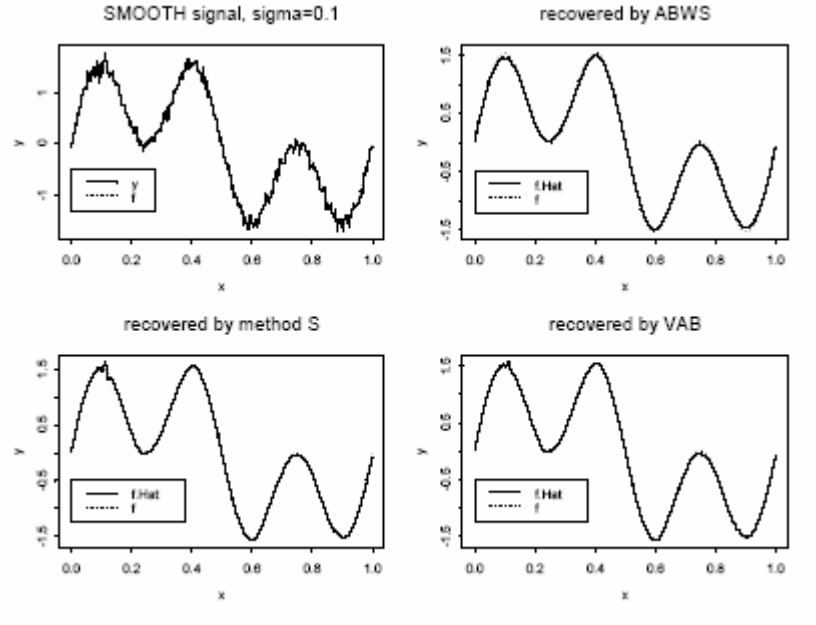


Figure 3: Piecewise polynomial data, with $\sigma = 0.1$.

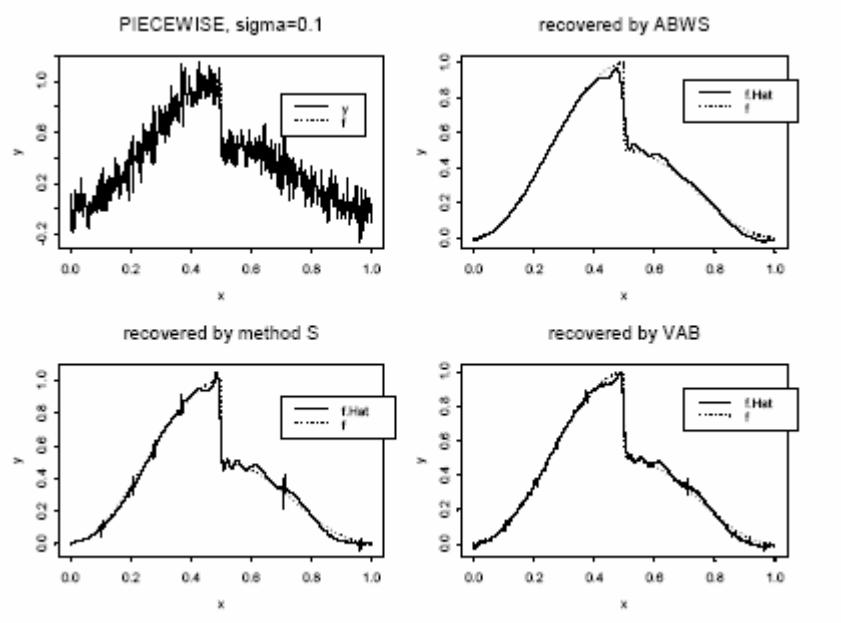


Figure 4: Chirp data, with $\sigma = 0.1$.

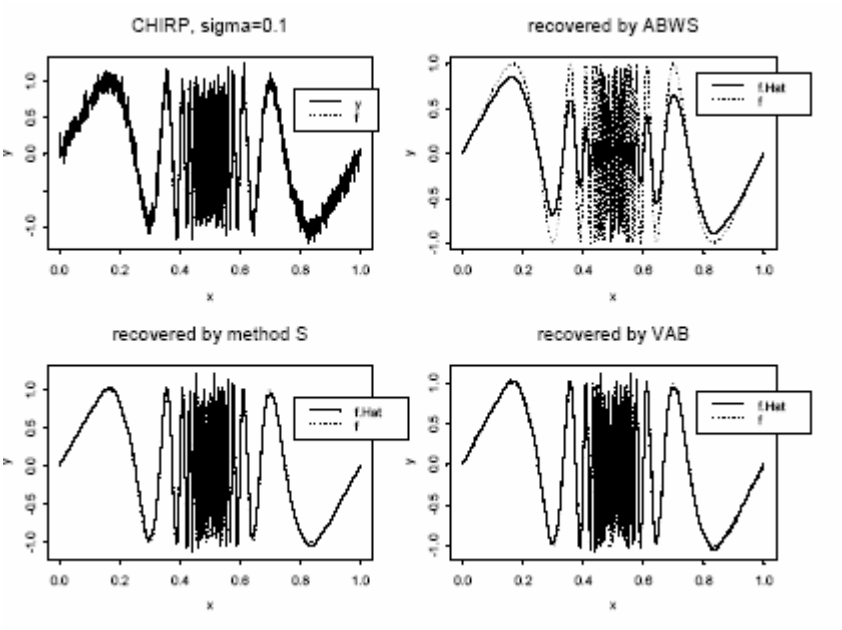


Figure 5: Corner data, with $\sigma = 0.1$.

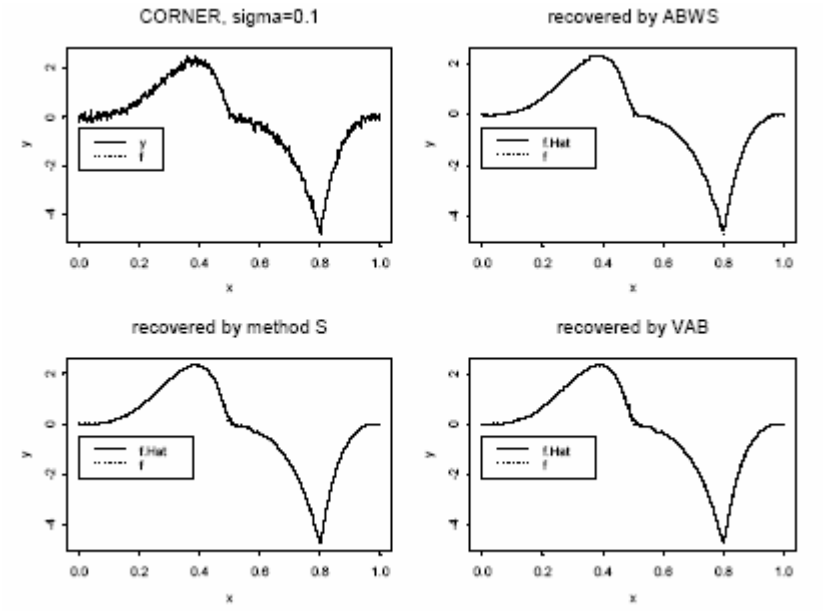


Figure 6: Blocks data, with $\sigma = 0.2$.

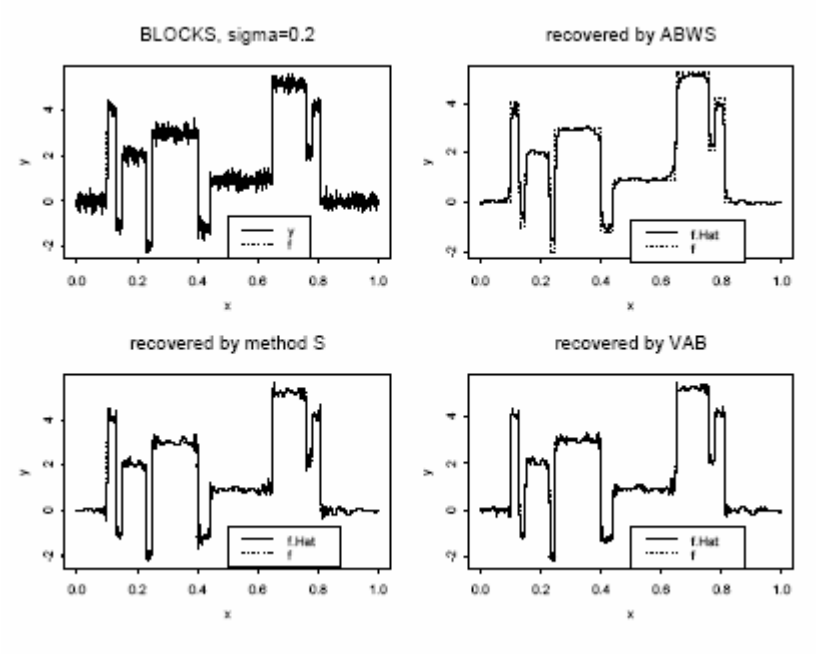


Figure 7: Bumps data, with $\sigma = 0.3$.

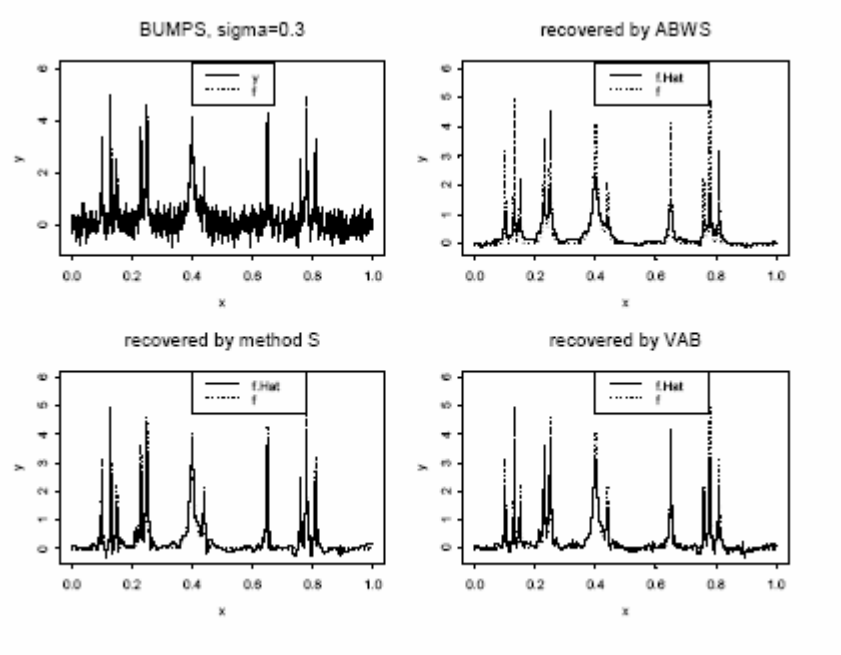


Figure 8: Doppler data, with σ = 0.2.

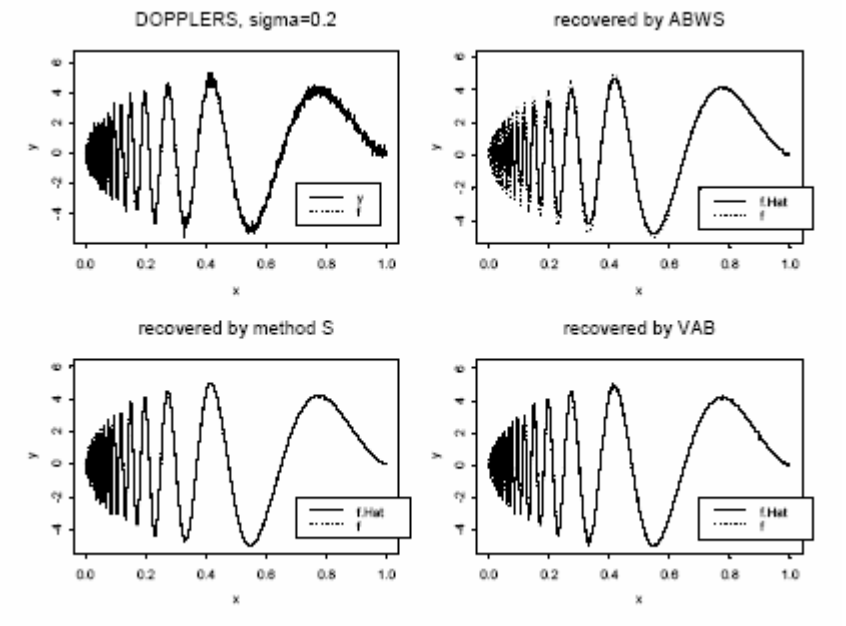
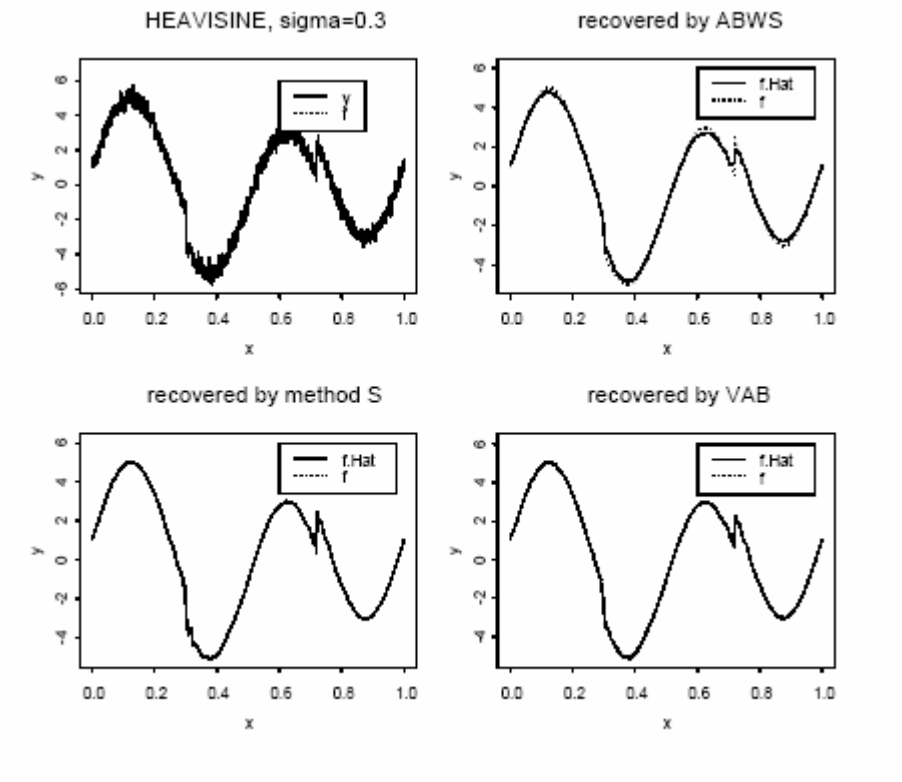


Figure 9: Heavisine data, with σ = 0.3.



It is well known that C [0,1] $\subset$ D [0,1] (Billingsley 1968), which means that the uniform topology is equivalent to the Skorohod topology for continuous functions. It is easy to show that in discrete cases such as in computer simulation, the uniform topology is equivalent to the Skorohod topology, where the uniform topology is defined as

$$d(f,g) = \sup_{0<x<1} |f(x) - g(x)|. \qquad (26)$$

Convergence in the uniform topology implies convergence in the L_2 norm, but convergence in L_2 norm can not guarantee convergence in the uniform topology. In this sense, Uniform topology seems to be a better candidate to serve as the measurement of the performance.

Table 2 summarizes the uniform topology in the same simulation study. Notice that in the case of PIECEWISE polynomial, CORNER and HEAVISINE, the pedigree of the uniform topology and the L_2 are very controversial. Our visual impression seems to prefer the uniform topology. In the other cases, the two measurements are compatible.

Conclusion

This article presents and implements a new VAB method to recover signals from noisy data. The VAB method was compared with existing Bayesian methods. The results support the notion that many methods are serviceable when iid Normal noise are added.

The appealing part of this model is that it can capture the few big spikes in the coefficients effectively, thereby preserving the coarse shape of the picture. The simplicity of the model is also an advantage. Compared with other prior models, VAB uses less CPU time. In simulation studies, VAB performs best in four out of the eight cases when using the mean-squared error, and it performs best in six out of the eight cases studied when using the uniform distance.

References

Abramovich, F., Sapatinas, T., & Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of Royal Statistical Society, 60 (Series B)*, 725-749.

Billingsley, P. (1968). *Convergence of probability measures*, NY: Wiley.

Bruce, A., & Gao, H. (1996a). *Applied wavelet analysis with S-Plus*. Berlin: Springer.

Bruce, A., & Gao, H. (1996b). Understanding WaveShrink: Variance and bias estimation, *Biometrika*, *83*, 727-745.

Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Annals of Statistics*, 27, 898-924.

Chipman, Kolaczyk, & McCulloch (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, *92*, 1413-1421.

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

Clyde, M., DeSimone, H., & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association, 91, 1197-208.*

Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

Daubechies, I. (Ed.) (1993). *Different Perspectives on Wavelets*. Applied Mathematics Society.

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association, 90(432)*, 1200-1224.

Donoho, D. L., & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics, 26*, 879-921.

Gao, H., & Bruce, A. (1997), WaveShrink with firm shrinkage. *Statistica Sinica, 7*, 855-874.

Huang, H. C., & Cressie, N. (1999). Empirical Bayesian spatial prediction using wavelets, *Bayesian inference in wavelet-based models*, Springer-Verlag, New York.

Johnstone, I. M., & Silverman, B. M. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, 59 (Serial B)*, 319-351.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11(7)*,674-693.

Nason, G. P., & Silverman B. W. (1994). The discrete wavelet transform in S. *Journal of Computational and Graphical statistics*, *3*, 163-191.

Nason, G. P. (1994). Wavelet regression by cross-validation. Technical Report 447.

Nason, G. P. (1995). Choice of the threshold parameter in wavelet function estimation: *Wavelets and statistics*, Lecture Notes in Statistics *103*, 261-280. NY: Springer-Verlag.

Strang, G. (1993). Wavelet transforms versus Fourier transforms. *Bulletin (New Series) of the American Mathematical Society*, *28* (2),288-305.

Wu, D. (2002). NORM thresholding method in wavelet regression. *Journal of Statistical Computation and Simulation*, *72* (No.3), 233-246.