BAYESIAN FUNCTION ESTIMATION USING CONTINUOUS WAVELET DICTIONARIES

Jen-Hwa Chu[1], Merlise A. Clyde[2] and Feng Liang[3]

[1]*Harvard Medical School,* [2]*Duke University and* [3]*University of Illinois at Urbana-Champaign*

Supplementary Material

S1. RJ-MCMC Algorithm

We follow the general framework by Green (1995) and Denison, Mallick, and Smith (1998) and include three types of moves in the RJ-MCMC algorithm: birth step (add a wavelet), death step (delete a wavelet), and update step (move a wavelet). Suppose the current model has K wavelet elements, as in Green (1995), the probabilities for birth, death, and update steps are chosen to be

$$\begin{aligned} p_b &= c\min\{1, p(K+1)/p(K)\}, \\ p_d &= c\min\{1, p(K)/p(K+1)\}, \\ p_u &= 1 - p_b - p_d, \end{aligned}$$

where $c < 0.5$ is some constant and $p(K)$ is the negative binomial prior distribution over K

$$p(k|s,q) = \binom{s+k-1}{k} q^s (1-q)^K \quad k = 0, 1, \ldots \tag{S1.1}$$

where $s = \alpha_\gamma$ and $q = \beta_\gamma/(\beta_\gamma + c(a_0, a_1, \zeta))$ with

$$c(a_0, a_1, \zeta) = \frac{1 - (a_0/a_1)^{\zeta-1}}{(\zeta-1)a_0^{\zeta-1}}, \text{if } \zeta \neq 1; \quad \log\frac{a_1}{a_0}, \text{ otherwise.}$$

Denison et al. (1998) recommended using $c = 0.4$, leading to a larger proportion of birth and death steps relative to update steps. We found that using $c = 0.05$, which leads to about 90% update steps, resulted in more efficient mixing of the Markov chain. Because we are in a continuous dictionary, update steps actually lead to the creation of new dictionary elements by local moves, but keep the model dimension fixed.

S1.1 Birth Step

In a birth step, we propose to add a new wavelet element $(a_{K+1}, b_{K+1}, \beta_{K+1})$ from some proposal distribution distribution $q(a, b, \beta)$. Because of the local nature of wavelets, we utilize

information in the residuals in proposing the location b of new wavelets. We construct a novel proposal distribution for the location parameter b of the new wavelet function to be added, which is a mixture of point masses on the data points with weights that depend on the current residuals and uniform on $[0, 1]$. In particular, the proposal distribution for the location b in a birth step at iteration $t + 1$ is

$$q(b_{K+1}) = \pi \sum_{i=1}^{n} \delta_{x_i}(b_{K+1}) v_i + (1 - \pi), \quad 0 < \pi < 1, \tag{S1.2}$$

where

$$v_i = \frac{|Y_i - \hat{f}^{(t)}(x_i)|}{\sum_{j=1}^{n} |Y_j - \hat{f}^{(t)}(x_j)|}$$

is proportional to the magnitude of the residual from the model fit at iteration t and

$$\hat{f}^{(t)}(x) = \sum_{k=0}^{K^{(t)}} \beta^{(t)}_{\lambda_k^{(t)}} \psi_{\lambda_k^{(t)}}(x) \tag{S1.3}$$

is the estimate of $f(x)$ using the current values of β_λ and λ at iteration t. Since the prior for b is also a mixture of point masses at data points and a uniform distribution, the proposal distribution and prior distributions are absolutely continuous with respect to each other which is a necessary condition for the transition kernel to be reversible.

For the remaining parameters, we propose the scale a_{k+1} from the prior distribution

$$p(a) \propto a^{-\zeta}, \quad a_0 \leq a \leq a_1 \text{ and } \zeta > 0. \tag{S1.4}$$

and conditional on the location and scale, the coefficient β_λ is proposed from the conditional posterior distribution obtained under the Gaussian prior distribution. The proposal distribution $q(\beta \mid a, b)$ for the coefficient β_{K+1} is

$$\beta_{K+1} \mid a_{1:K+1}, b_{1:K+1}, \beta_{1:K}, \mathbf{Y} \sim \mathsf{N}(\hat{\beta}, \hat{\sigma}^2_\beta) \tag{S1.5}$$

where

$$\hat{\sigma}^2_\beta = \left[\frac{1}{ca^{-\delta}} + \frac{\boldsymbol{\psi}'_{\boldsymbol{\lambda}_{K+1}} \boldsymbol{\psi}_{\boldsymbol{\lambda}_{K+1}}}{\sigma^2} \right]^{-1}, \quad \hat{\beta} = \frac{\hat{\sigma}^2_\beta}{\sigma^2} \boldsymbol{\psi}_{\boldsymbol{\lambda}_{K+1}}{}'(\mathbf{Y} - \hat{\mathbf{f}}),$$

and $\boldsymbol{\psi}_{\boldsymbol{\lambda}_{K+1}}$ is the vector of length n of the proposed wavelet evaluated at the observed data and $\hat{\mathbf{f}}$ is the vector of the estimated function values given the current parameters. This independent proposal distribution for β can improve the acceptance rate not only for normal prior distributions (where it is the conditional posterior distribution), but also for heavy-tailed prior distributions where the posterior for β does not have a closed-form density.

S1.2 Death Step

For the death step, we remove the kth wavelet (and associate parameters) with a probability

inversely proportional to the current wavelet coefficients:

$$q(k \mid K) = \frac{1/|\beta_k|}{\sum_{i=1}^{K}(1/|\beta_i|)}, \tag{S1.6}$$

so that small magnitude coefficients are more likely to be removed.

S1.3. Update Step

For the update step, we randomly pick an index k to update from the uniform distribution on $\{1, \dots K\}$. We propose a new scale $\tilde{a}_k$ using a Gaussian random walk proposal centered at the current value and with variance σ_a^2. The proposal for the update step of the location b is

$$q(\tilde{b}_k \mid b_k) = \delta_{b_k}(\tilde{b}_k)u_k + \mathsf{N}(\tilde{b}_k; b_k, \sigma_b^2)(1 - u_k) \tag{S1.7}$$

where

$$u_k = \begin{cases} 1 & \text{if } b_k \text{ is a data point} \\ 0 & \text{otherwise.} \end{cases}$$

If b_k is at a data point x_i then we do not change the location; this is necessary to ensure that the proposal and prior distributions are absolutely continuous with respect to each other. Otherwise we update the location using a Gaussian random walk step centered at the current location b_k with variance σ_b^2. We set $\sigma_a^2 = \sigma_b^2 = 0.001$ in the examples, which led to approximately a 30-40% acceptance rate. We propose a new wavelet coefficient $\tilde{\beta}_k$ from the full conditional distribution obtained under the normal prior distribution in (S1.5) using the proposed scale $\tilde{a}_k$ and location. $\tilde{b}_k$. Through updating the index λ of a dictionary element ψ_λ we are able to "smoothly" transition from one dictionary element to a new one.

S1.4. Fixed Dimensional Parameters

Updating σ^2 is a straightforward Gibbs update via a conjugate gamma distribution,

$$1/\sigma^2 \sim \mathsf{G}(n/2, \mathsf{SSE}/2)$$

where $\mathsf{SSE} = \sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2$ in the rate parameter of the gamma distribution.

S1.5. Acceptance Ratio

For each step, the acceptance ratio can be calculated by

$$\mathsf{LR} \times \text{prior ratio} \times \text{proposal ratio}$$

where

$$\mathsf{LR} = \frac{\mathsf{N}(\mathbf{Y}; \tilde{f}(x), \sigma^2 I)}{\mathsf{N}(\mathbf{Y}; \hat{f}(x), \sigma^2 I)} \tag{S1.8}$$

denotes the likelihood ratio and $\hat{f}(x)$ and $\tilde{f}(x)$ denote the estimates for the current model and for the proposed model, respectively. The prior distribution over K and the triplets $\{\beta_k, a_k, b_k\}_{k=1}^{K}$,

factorized as

$$p(K)\prod_{k=1}^{K} p(a_k, b_k, \beta_k).$$

Simple calculations reveal that the acceptance ratio is

$$\mathsf{LR} \times (K+1) \times \frac{p(a_{K+1}, b_{K+1}, \beta_{K+1})}{q(a_{K+1}, b_{K+1}, \beta_{K+1})} \times \frac{1/|\beta_{K+1}|}{\sum_{k=1}^{K+1}(1/|\beta_k|)}$$

for the birth step,

$$\mathsf{LR} \times \frac{1}{K} \times \frac{q(a_k, b_k, \beta_k)}{p(\beta_k, a_k, b_k)} \times \left[\frac{1/|\beta_K|}{\sum_{k=1}^{K}(1/|\beta_k|)}\right]^{-1}$$

for the death step, and

$$\mathsf{LR} \times \frac{p(\tilde{\beta}_k, \tilde{a}_k, \tilde{b}_k)}{p(\beta_k, a_k, b_k)} \times \frac{q(\beta_k \mid a_k, b_k)}{q(\tilde{\beta}_k \mid a_k, b_k)}$$

for the update step, where $q(\beta \mid a, b)$ is given in equation (S1.5).

The RJ-MCMC algorithm goes as follows: Start with $K = 0$. Repeat the following steps until convergence and then repeat for an additional T iterations for inference:

1. Generate a $\mathsf{Unif}(0,1)$ random number u,
   (a) If $u < p_b(K)$, perform the birth step.
   (b) If $p_b(K) < u < p_b(K) + p_d(K)$, perform the death step.
   (c) If $u > p_b(K) + p_d(K)$, perform the update step.
2. Update σ^2 by a Gibbs step:

$$1/\sigma^2 \sim \mathsf{G}(n/2, \mathsf{SSE}/2).$$

Although we use the null model in initialization for computational convenience, the algorithm can have any other model as a starting point. In our studies, we found very little difference in results due to starting models.

Standard convergence diagnostic methods, such as Gelman and Rubin (1992), do not apply for assessing convergence of the joint posterior distribution since we using a trans-dimensional sampler. Instead we look at K and the mean squared error, which have a coherent interpretation throughout the model space (Brooks and Giudici, 2000). The trace plots and the Gelman-Rubin shrink factor for K and mean squared error suggest that convergence usually occurs within 500,000 MCMC iterations.

S2. Illustration

Examples of the estimated functions in the high noise scenario are shown in Figure 1. The estimates using the CWD appear to be slightly smoother than those of EBayes, although we note that both methods appear to give good estimates. The CWD methods do appear to capture more of the high frequency oscillations of the `doppler` function. In the high noise

scenario, however, both methods miss the discontinuity in the heavisine function around $x = 0.7$ for this simulated example.

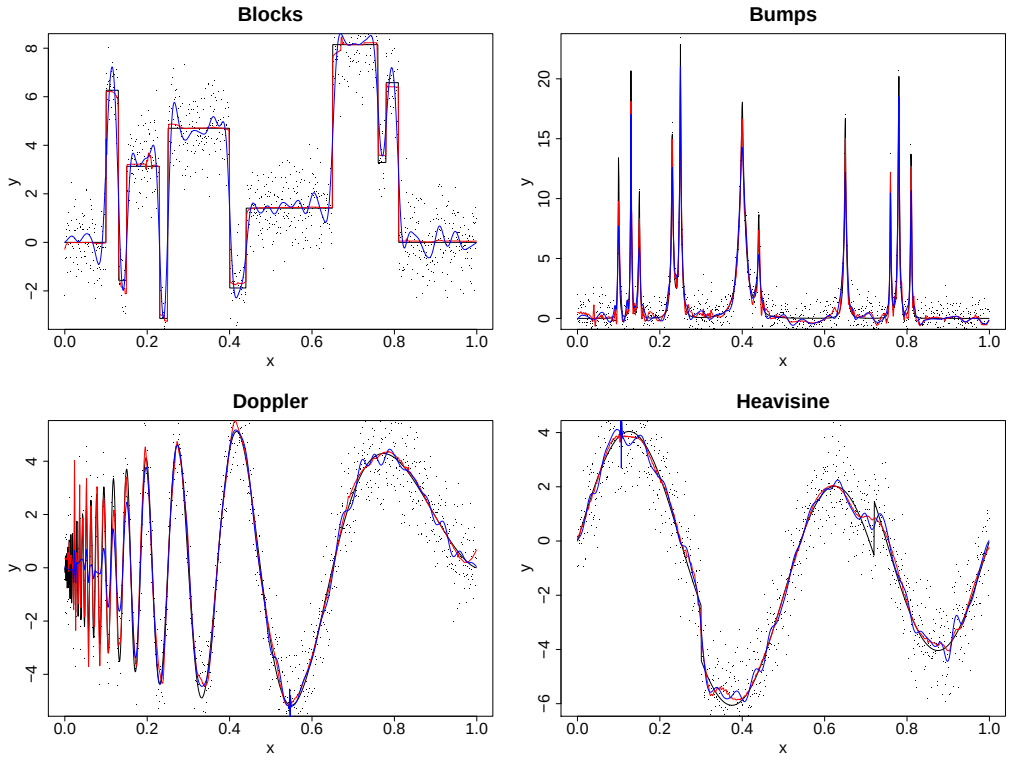

Figure 1: Some examples for the estimated functions: True function (black), CWD fit (red), EBayes (blue) and data points with SNR=3.

Acknowledgment

The authors would like to thank Brani Vidakovic for suggesting the Daubechies-Lagarias algorithm for evaluating continuous wavelets. The authors would also like to thank the Associate Editor and anonymous referees for extremely helpful comments and suggestions. The authors acknowledge support of the National Science Foundation (NSF) through grants DMS–0342172 DMS–0422400, and DMS–0406115. Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF.

References

S. Brooks and P. Giudici. Markov chain monte carlo convergence assessment via two-way analysis of variance. *Journal of Computational and Graphical Statistics*, 9(2):266–285, 2000.

D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. Automatic Bayesian curve fitting. *J. R. Statist. Soc. B*, 60:333–350, 1998.

A. Gelman and D. P. Rubin. Inference from iterative simualtion using mutliple sequences. *Statistical Science*, 7:457–511, 1992.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.