Biometrika (2011), **98**, 1, *pp.* 1–5

 Advance Access publication on 30 June 2011

Printed in Great Britain

Supplementary material for "Bayesian graphical models for multivariate functional data"

BY H. ZHU, D. B. DUNSON

Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708 U.S.A.

hz52@stat.duke.edu dunson@stat.duke.edu

AND N. STRAWN

Department of Mathematics, Duke University, Box 90320, Durham, NC 27708 U.S.A.

nstrawn@math.duke.edu

1. DETAILS OF THE ALGORITHMS

Algorithm 1

Step 0. Set an initial decomposable graph G and the prior parameters c_0, δ, U^*.

Step 1. With probability $1-p$, propose $\widetilde{G} \mid G \sim p(\widetilde{G} \mid G)$ by randomly adding or deleting an edge (each with probability 0.5) in the space of decomposable graphs, and accept the new $\widetilde{G}$ with probability

$$\alpha = \min\left\{1, \frac{p(\widetilde{G} \mid \{\mathrm{c}_i\})p(G \mid \widetilde{G})}{p(G \mid \{\mathrm{c}_i\})p(\widetilde{G} \mid G)}\right\},$$

where

$$\frac{p(\widetilde{G} \mid \{\mathrm{c}_i\})}{p(G \mid \{\mathrm{c}_i\})} = \frac{p(\{\mathrm{c}_i\} \mid \mathrm{c}_0, \widetilde{G})}{p(\{\mathrm{c}_i\} \mid \mathrm{c}_0, G)} \cdot \frac{p(\widetilde{G})}{p(G)}.$$

For the case of adding (i.e. $\widetilde{G}$ has one more edge than G), there are two cases. Case (1), the two nodes (denoted as k, l) being connected belong to two different connected components. Here a connected component is defined as a cluster of nodes that are connected so that for any node in the cluster there is a route from one node to another. In this case, the likelihood ratio takes the form:

$$\frac{p(\{\mathrm{c}_i\} \mid \mathrm{c}_0, \widetilde{G})}{p(\{\mathrm{c}_i\} \mid \mathbf{c}_0, G)} = \frac{|U^*_{k,l}|^{(\delta+d_{k,l}-1)/2}}{|U^*_k|^{(\delta+d_k-1)/2}|U^*_l|^{(\delta+d_l-1)/2}} \cdot \frac{|\widetilde{U}^*_k|^{(\widetilde{\delta}+d_k-1)/2}|\widetilde{U}^*_l|^{(\widetilde{\delta}+d_l-1)/2}}{|\widetilde{U}^*_{k,l}|^{(\widetilde{\delta}+d_{k,l}-1)/2}}$$
$$\cdot \frac{\Gamma_{d_{k,l}}(\frac{\widetilde{\delta}+d_{k,l}-1}{2})}{\Gamma_{d_{k,l}}(\frac{\delta+d_{k,l}-1}{2})} \cdot \frac{\Gamma_{d_k}(\frac{\delta+d_k-1}{2})}{\Gamma_{d_k}(\frac{\widetilde{\delta}+d_k-1}{2})} \cdot \frac{\Gamma_{d_l}(\frac{\delta+d_l-1}{2})}{\Gamma_{d_l}(\frac{\widetilde{\delta}+d_l-1}{2})},$$

where U^*_k, U^*_l and $U^*_{k,l}$ are sub-matrices of U^* that corresponding to corresponding functional components, and $\Gamma_d(a) = \pi^{d(d-1)/2}\prod_{i=0}^{d-1}\Gamma(a - i/2)$. Here d_k, d_l and $d_{k,l}$ are the size of the corresponding sub-matrices. Case (2), the two nodes k, l being connected belong to the same connected components. The decomposability implies that after connecting, k, l lie in the same clique, denoted as C_q. Denote $S_q = C_q/\{k, l\}$, $C_{q_1} = C_q/k$, $C_{q_2} = C_q/l$ and $D = \{k, l\}$, we can write $U^*_{C_q}$ in the form of

$$\begin{pmatrix} U^*_{S_q} & U^*_{S_q,D} \\ U^*_{D,S_q} & U^*_D \end{pmatrix}.$$

Then the likelihood ratio takes the form

$$\frac{p(\{\mathrm{c}_i\} \mid \mathrm{c}_0, \widetilde{G})}{p(\{\mathrm{c}_i\} \mid \mathrm{c}_0, G)} = \frac{|U^*_{C_q}|^{(\delta+d_{C_q}-1)/2}|U^*_{S_q}|^{(\delta+d_{S_q}-1)/2}}{|U^*_{C_{q2}}|^{(\delta+d_{C_{q_2}}-1)/2}|U^*_{C_{q_1}}|^{(\delta+d_{C_{q_1}}-1)/2}} \cdot \frac{|\widetilde{U}^*_{C_{q_2}}|^{(\widetilde{\delta}+d_{C_{q_2}}-1)/2}|\widetilde{U}^*_{C_{q_1}}|^{(\widetilde{\delta}+d_{C_{q_1}}-1)/2}}{|\widetilde{U}^*_{C_q}|^{(\widetilde{\delta}+d_{C_q}-1)/2}|\widetilde{U}^*_{S_q}|^{(\widetilde{\delta}+d_{S_q}-1)/2}}$$
$$\cdot \frac{\Gamma_{d_{C_q}}(\frac{\widetilde{\delta}+d_{C_q}-1}{2})}{\Gamma_{d_{C_q}}(\frac{\delta+d_{C_q}-1}{2})} \cdot \frac{\Gamma_{d_{S_q}}(\frac{\widetilde{\delta}+d_{S_q}-1}{2})}{\Gamma_{d_{S_q}}(\frac{\delta+d_{S_q}-1}{2})} \cdot \frac{\Gamma_{d_{C_{q_2}}}(\frac{\delta+d_{C_{q_2}}-1}{2})}{\Gamma_{d_{C_{q_2}}}(\frac{\widetilde{\delta}+d_{C_{q_2}}-1}{2})} \cdot \frac{\Gamma_{d_{C_{q_1}}}(\frac{\delta+d_{C_{q_1}}-1}{2})}{\Gamma_{d_{C_{q_1}}}(\frac{\widetilde{\delta}+d_{C_{q_1}}-1}{2})}.$$

If using independent Bernoulli priors (with parameter r) for the edges included in G, $p(\widetilde{G})/p(G) = r/(1-r)$. The proposal ratio $p(\widetilde{G} \mid G)/p(\widetilde{G} \mid G) = (p(p-1)/2 -$

$n_e)/(n_e+1)$, for n_e the number of edges in G. The likelihood ratio for the case of deleting is simply the inverse of that for the case of adding.

With probability p, propose $\widetilde{G} \sim$ Unif, a (discrete) uniform distribution supported on the set of all decomposable graphs, and accept the proposal with probability

$$\alpha = \min\left\{1, \frac{p(\widetilde{G} \mid \{\mathrm{c}_i\})}{p(G \mid \{\mathrm{c}_i\})}\right\}.$$

Repeat Step 1 for a large number of iterations.

2. SET MODEL PARAMETERS

Several parameters need to be determined before applying Algorithm 1 and 2. The truncation parameters $\{K_j\}$ can be determined using some approximation criteria. For example, if using FPC, we can control the fraction of variables explained (FVE) to be above certain threshold. Other criteria such as AIC, BIC and cross validations will also work. The degrees of freedom δ of the BHIW prior of Q^* is chosen as a positive integer. Smaller values of δ imply larger variances so that the prior is more "vague". For the scale matrix U^* of the BHIW prior, we determine its value by first decomposing $U^* = ZR^*Z$, where $Z = \mathrm{diag}\{\tau\}$ is the marginal standard deviation of the basis coefficients. If using FPC analysis, τ can be taken as the square root of the eigenvalues. In other cases, we suggest to choose τ to be proportional to the (marginal) sample standard deviation, from the empirical Bayes perspective. The pattern of R^* can be hard to determine. We set $R^* = \mathbf{I}$ in our simulations and real data application.

Other priors, like the Hyper-inverse Wishart g-prior of Carvalho & Scott (2009) would also be good options. In Algorithm 2, one also needs to determine the noise variance Λ, whose value would influence of the identification of Q^*. In this work, we have assumed additive white noise. Any orthogonal basis transform of Gaussian white noise is still white noise. The variance of the white noise in the frequency domain equals the corresponding variance in the time domain

up to a scale parameter, which is approximately $T_j/(n_j - 1)$. Therefore, we can estimate the white noise variance by firstly applying a localized linear smoother to the function, and then computing the sample variances of the residuals. This variance can then be transformed to the frequency domain. If using FPC analysis, the PACE algorithm of Yao et al. (2005) can be directly applied to compute the noise variances and eigen-basis, even for extremely sparse data. For the initial values $\{\mathrm{c}_i\}$ for Algorithm 2, one can simply set $\mathrm{c}_i = \mathrm{d}_i$. If the data are centered in a pre-processing step, one can set c_0 to be the zero vector, otherwise, using the sample mean of the estimated basis coefficients.

3. METHODS FOR IMPROVING MIXING

Although the dimension of the problem has been significantly reduced using orthogonal basis expansion, the curse of dimensionality may still present because including more than one "features" for each node (function) essentially make the posterior distribution more "spiker". When the chain stays in a local mode, the local proposals which are based on one edge difference may be hardly accepted due to the spiky nature of the density. Although the mixture proposal of the SW sampler helps the chain to jump out of a local mode, it may still take a long time to get a good proposal especially when the heavier tailed proposal is taken to be the uniform. One good improvement is to use an annealing/tempering method, in which a temperature parameter is used to flatten the posterior distribution so that acceptance rate can be increased. Note that if using annealing or parallel tempering, the algorithm is not MCMC any more. One alternative that retains the MCMC algorithm is the SW sampler with tempering (SWt) proposed by unpublished work of Guan and Stephens; in which case the heavy tailed proposal in the SW sampler is replaced by a tempered version of the posterior distribution. In SWt, two chains need to be run in parallel, one is for sampling from the original posterior distribution, the other is for sampling from the

tempered distribution (called exploring chain) whose samples will be used as the proposals in the main sampler. Our algorithm 1 can be adapted using the annealing method and the SWt method as the graph G is the only parameters involved. As a hybrid MCMC algorithm, Algorithm 2 can not be directly adapted for annealing, nor for the SWt. However, one can always apply these techniques in the burn-in period of Algorithm 2 which is an ad hoc way to improve mixing.

4. More results for simulation 2

A plot of the noisy data was shown in the panel (a) of Figure 1, with its smooth estimates shown in Panel (b). The posterior estimate of the time domain correlation was plotted in panel (c), which corresponds to the true correlation plotted in the bottom left of Figure 2 in the main text. The trace plot of the conditional log posterior densities of the graph was shown in panel (d).

References

Carvalho, C. M. & Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 497–512.

Yao, F., Müller, H. G. & Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.* **100**, 577–590.

[*Received January* 2011. *Revised June* 2011]

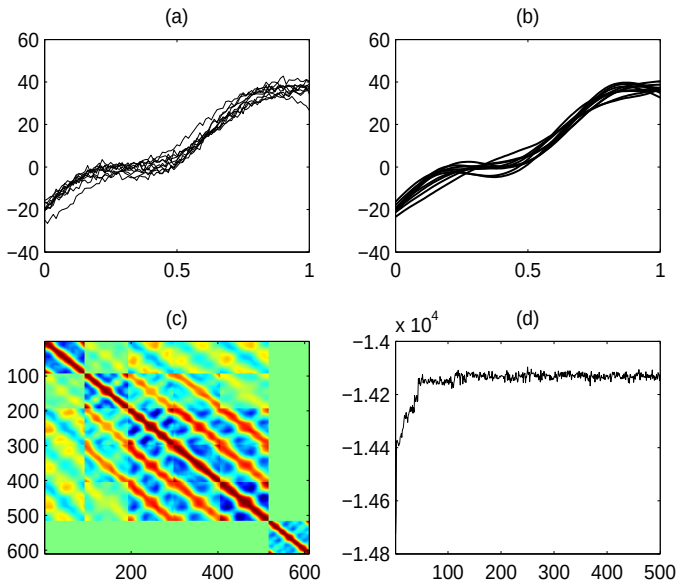

Fig. 1. Results for Simulation 2. (a): The plot of raw data for the first 10 samples of functional component 1. (b): The posterior mean estimate of $f_{i1}(t)$ corresponding to the curves in (a). (c): the posterior mean estimate of the time domain correlation matrix. (d): The trace plot of the log posterior densities of the first 500 samples.