Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Supplementary Materials for "Bayesian Graphical Models for Multivariate Functional Data"

Hongxiao Zhu HONGXIAO@VT.EDU
Department of Statistics
Virginia Tech
Blacksburg, VA 24061, USA

Nate Strawn NATE.STRAWN@GEORGETOWN.EDU
Department of Mathematics and Statistics
Georgetown University
Washington D.C. 20057, USA

David B. Dunson DUNSON@DUKE.EDU
Department of Statistical Science
Duke University
Durham NC 27708, USA

Editor: Leslie Pack Kaelbling

1. More details of Algorithm 1

Step 0. Choose an initial decomposable graph G and the prior parameters c_0, δ, $\mathbf{U}$.

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

$$\alpha = \min\left\{1, \frac{p(\widetilde{G} \mid \{\mathbf{c}_i^M\}, \mathbf{c}_0^M)\; p(G \mid \widetilde{G})}{p(G \mid \{\mathbf{c}_i^M\}, \mathbf{c}_0^M)\; p(\widetilde{G} \mid G)}\right\}.$$

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:

$$\begin{aligned}\frac{p(\{\mathbf{c}_i^M\} \mid \mathbf{c}_0^M, \widetilde{G})}{p(\{\mathbf{c}_i^M\} \mid \mathbf{c}_0^M, G)} =& \frac{|\mathbf{U}_{k,l}|^{(\delta+d_{k,l}-1)/2}}{|\mathbf{U}_{k,k}|^{(\delta+d_{k,k}-1)/2}|\mathbf{U}_{l,l}|^{(\delta+d_{l,l}-1)/2}} \\ &\times \frac{|\widetilde{\mathbf{U}}_{k,k}|^{(\widetilde{\delta}+d_{k,k}-1)/2}|\widetilde{\mathbf{U}}_{l,l}|^{(\widetilde{\delta}+d_{l,l}-1)/2}}{|\widetilde{\mathbf{U}}_{k,l}|^{(\widetilde{\delta}+d_{k,l}-1)/2}} \\ &\times \frac{\Gamma_{d_{k,l}}(\frac{\widetilde{\delta}+d_{k,l}-1}{2})}{\Gamma_{d_{k,l}}(\frac{\delta+d_{k,l}-1}{2})} \frac{\Gamma_{d_{k,k}}(\frac{\delta+d_{k,k}-1}{2})}{\Gamma_{d_{k,k}}(\frac{\widetilde{\delta}+d_{k,k}-1}{2})} \frac{\Gamma_{d_{l,l}}(\frac{\delta+d_{l,l}-1}{2})}{\Gamma_{d_{l,l}}(\frac{\widetilde{\delta}+d_{l,l}-1}{2})},\end{aligned}$$

where $\mathbf{U}_{k,k}$, $\mathbf{U}_{l,l}$ and $\mathbf{U}_{k,l}$ are sub-matrices of $\mathbf{U}$ associated with corresponding functional components, and $\Gamma_d(a) = \pi^{d(d-1)/2} \prod_{i=0}^{d-1} \Gamma(a - i/2)$. Here $d_{k,k}$, $d_{l,l}$ 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 $S_q = C_q \setminus \{k, l\}$, $C_{q_1} = C_q \setminus k$, $C_{q_2} = C_q \setminus l$ and $D = \{k, l\}$, we can write $\mathbf{U}_{C_q}$ in the form of

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

Then the likelihood ratio takes the form

$$\begin{aligned} \frac{p(\{\mathbf{c}_i^M\} \mid \mathbf{c}_0^M, \widetilde{G})}{p(\{\mathbf{c}_i^M\} \mid \mathbf{c}_0^M, G)} =& \frac{|\mathbf{U}_{C_q}|^{(\delta+d_{C_q}-1)/2} |\mathbf{U}_{S_q}|^{(\delta+d_{S_q}-1)/2}}{|\mathbf{U}_{C_{q_2}}|^{(\delta+d_{C_{q_2}}-1)/2} |\mathbf{U}_{C_{q_1}}|^{(\delta+d_{C_{q_1}}-1)/2}} \\ &\times \frac{|\widetilde{\mathbf{U}}_{C_{q_2}}|^{(\widetilde{\delta}+d_{C_{q_2}}-1)/2} |\widetilde{\mathbf{U}}_{C_{q_1}}|^{(\widetilde{\delta}+d_{C_{q_1}}-1)/2}}{|\widetilde{\mathbf{U}}_{C_q}|^{(\widetilde{\delta}+d_{C_q}-1)/2} |\widetilde{\mathbf{U}}_{S_q}|^{(\widetilde{\delta}+d_{S_q}-1)/2}} \\ &\times \frac{\Gamma_{d_{C_q}}(\frac{\widetilde{\delta}+d_{C_q}-1}{2})}{\Gamma_{d_{C_q}}(\frac{\delta+d_{C_q}-1}{2})} \frac{\Gamma_{d_{S_q}}(\frac{\widetilde{\delta}+d_{S_q}-1}{2})}{\Gamma_{d_{S_q}}(\frac{\delta+d_{S_q}-1}{2})} \frac{\Gamma_{d_{C_{q_2}}}(\frac{\delta+d_{C_{q_2}}-1}{2})}{\Gamma_{d_{C_{q_2}}}(\frac{\widetilde{\delta}+d_{C_{q_2}}-1}{2})} \\ &\times \frac{\Gamma_{d_{C_{q_1}}}(\frac{\delta+d_{C_{q_1}}-1}{2})}{\Gamma_{d_{C_{q_1}}}(\frac{\widetilde{\delta}+d_{C_{q_1}}-1}{2})}. \end{aligned}$$

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 $p(\widetilde{G} \mid G)/p(\widetilde{G} \mid G) = (p(p-1)/2 - n_e)/(n_e + 1)$, with 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

$$\alpha = \min \left\{ 1, \frac{p(\widetilde{G} \mid \{\mathbf{c}_i^M\}, \mathbf{c}_0^M)}{p(G \mid \{\mathbf{c}_i^M\}, \mathbf{c}_0^M)} \right\}.$$

Repeat step 1 for a large number of iterations until convergence is achieved.

2. More details on setting model parameters

Several parameters need to be determined before applying Algorithm 1 or 2. The truncation parameters $\{M_j\}$ can be determined using some approximation criteria as discussed in the paper. The degrees of freedom δ of the HIWP_G prior of $\mathbf{Q}_{\mathcal{C}}$ is chosen as a positive integer. Smaller values of δ imply larger variances so that the prior is more "vague." For the scale matrix $\mathbf{U}$ of the HIWP_G prior, we determine its value by first decomposing $\mathbf{U} = \mathbf{ZRZ}$, where $\mathbf{Z} = \mathrm{diag}\{\boldsymbol{\tau}\}$ is the marginal standard deviation of the basis coefficients. If using FPC analysis, $\boldsymbol{\tau}$ can be taken as the square root of the eigenvalues. In other cases, we suggest to choose $\boldsymbol{\tau}$ to be proportional to the (marginal) sample standard deviation, from the empirical Bayes perspective. The pattern of $\mathbf{R}$ can be hard to determine. We set $\mathbf{R} = \mathbf{I}$ in our simulations and real data application. Other priors, like the Hyper-inverse Wishart g-prior of Carvalho and Scott (2009), would also be good options. In Algorithm 2, one also needs to determine the noise variance $\boldsymbol{\Lambda}$, whose value would influence the identification of $\mathbf{Q}_{\mathcal{C}}$. 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 $|T_j|/(|\mathbf{t}_j| - 1)$, where $|T_j|$ is the length of T_j and $|\mathbf{t}_j|$ is the number of grid points on T_j. 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 eigenbasis, even for sparse functional data. For the initial values $\{\mathbf{c}_i^M\}$ in Algorithm 2, one can simply set $\mathbf{c}_i^M = \mathbf{d}_i$. If the data are centered in a pre-processing step, one can set $\mathbf{c}_0^M$ to be the zero vector; otherwise, one can use the sample mean of the estimated basis coefficients.

3. Methods for improving mixing

Even though the small-world sampler in the MCMC Algorithms 1 and 2 helps improve mixing, as the number of vertices p and the truncation parameters $\{M_j\}_{j=1}^p$ increase, the Metropolis-Hastings step may suffer low acceptance rate, causing slow convergence. More advanced Monte Carlo strategies, such as parallel tempering Liu (2008), may be adopted to further improve mixing. Another alternative is the Small-world MCMC with Tempering algorithm proposed by Guan and Stephens (http://arxiv.org/abs/1211.4675), in which the heavy tailed proposal in the small-world sampler is replaced by a tempered version of the posterior distribution.

4. More results for simulation 2

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

References

C. M. Carvalho and J. G. Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512, 2009.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2008.

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

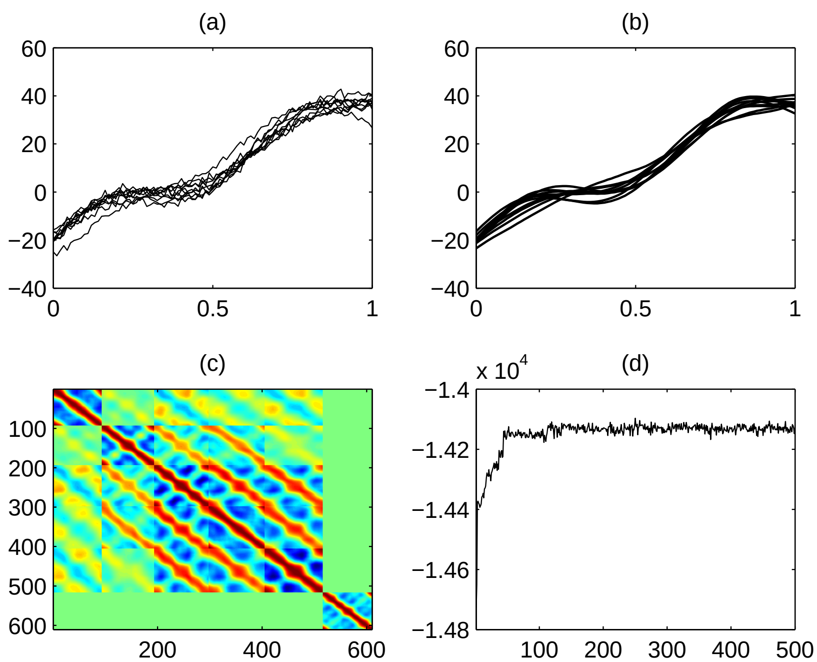

Figure 1: Results for Simulation 2. (a): The plot of raw data for the first 10 samples of functional component 1. (b): The posterior mean estimate of $f_{i1}(t)$ corresponding to the curves in (a). (c): the posterior mean estimate of the data domain correlation matrix. (d): The trace plot of the log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