

# Appendices to “To Center or Not to Center: That is Not the Question — An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency”

Yaming Yu

Department of Statistics, University of California, Irvine

Xiao-Li Meng

Department of Statistics, Harvard University

June 14, 2011

## A Auxiliary Material for Section 3

### A.1 Details of the MCMC Steps in the Poisson Time Series Example

Step 1: This step consists of  $T$  substeps, one for each  $t = 1, \dots, T$ .

For substep  $t$ , the target conditional density is

$$p(\xi_t | \xi_{t-1}, \xi_{t+1}, \beta, \rho, \delta, Y) \propto \exp\{-(\xi_t - \mu_t)^2 / (2\sigma_t^2) - d_t e^{X_t \beta + \xi_t}\},$$

where  $\mu_t = (Y_t \delta^2 + (\xi_{t-1} + \xi_{t+1})\rho) / (1 + \rho^2)$ ,  $\sigma_t^2 = \delta^2 / (1 + \rho^2)$ , if  $t \neq 1$  and  $t \neq T$ , and  $\mu_1 = Y_1 \delta^2 + \xi_2 \rho$ ,  $\mu_T = Y_T \delta^2 + \xi_{T-1} \rho$ ,  $\sigma_1^2 = \sigma_T^2 = \delta^2$ .

Define  $l_t(\xi_t) = \log p(\xi_t | \xi_{t-1}, \xi_{t+1}, \beta, \rho, \delta, Y)$ . First, use Newton-Raphson to locate the mode  $x$  of  $l_t$  and then calculate  $l_t''(x)$ . Then draw  $t_5$  according to a  $t$  distribution with 5 degrees of freedom, and propose  $\xi_t^{new} = x + t_5 / \sqrt{-l_t''(x)}$ . Draw a uniform random number  $u$  between 0 and 1. Accept  $\xi_t^{new}$  if

$$u \leq \exp\{l_t(\xi_t^{new}) - l_t(\xi_t^{old}) - h(\xi_t^{new}) + h(\xi_t^{old})\},$$

where  $h(\cdot)$  is the log density of  $t_5$  centered at  $x$  with scale  $1/\sqrt{-l_t''(x)}$ .

Step 2<sub>A</sub>: The target conditional density is

$$p(\beta | \xi, \rho, \delta, Y) \propto \exp\left\{\sum_t (Y_t X_t \beta - d_t e^{X_t \beta + \xi_t})\right\}.$$

Define  $l(\beta) = \log p(\beta | \xi, \rho, \delta, Y)$ . First, use Newton-Raphson to locate the mode  $\hat{\beta}$  of  $l$ , and compute  $I = -\partial^2 l(\hat{\beta}) / \partial \beta \partial \beta^\top$ . (This amounts to fitting a Poisson GLM and computing

the MLE and the observed Fisher information.) Draw  $T_5$  according to a p-variate  $t_5$ , and propose  $\beta^{new} = \hat{\beta} + I^{-1/2}T_5$ . Draw  $u \sim U(0, 1)$ , and accept  $\beta^{new}$  if

$$u \leq \exp\{l(\beta^{new}) - l(\beta^{old}) - H(\beta^{new}) + H(\beta^{old})\},$$

where  $H(\cdot)$  is the log density of  $T_5$  centered at  $\hat{\beta}$  and with scale  $I^{-1/2}$ .

Step 2<sub>S</sub>: The target conditional density  $p(\beta|\eta, \rho, \delta, Y)$  is multivariate normal.

Let  $Z^\top = (Z_1^\top, \dots, Z_T^\top)$ , where  $Z_1 = \sqrt{1 - \rho^2}X_1$  and  $Z_t = X_t - \rho X_{t-1}$ ,  $t \geq 2$ . Let  $\tilde{\eta} = (\sqrt{1 - \rho^2}\eta_1, \eta_2 - \rho\eta_1, \eta_3 - \rho\eta_2, \dots, \eta_T - \rho\eta_{T-1})^\top$ . Compute  $\hat{\beta} = (Z^\top Z)^{-1}Z^\top \tilde{\eta}$ , and then draw  $\beta^{new} \sim N_p(\hat{\beta}, (Z^\top Z)^{-1}\delta^2)$ . Set  $\xi_t^{new} = \eta_t - X_t\beta^{new}$ .

Step 3<sub>S</sub>: The target conditional density is

$$p(\rho, \delta|\beta, \xi, Y) \propto \delta^{-T} \exp\left\{-\frac{1}{2\delta^2} \left[ (1 - \rho^2)\xi_1^2 + \sum_{t=2}^T (\xi_t - \rho\xi_{t-1})^2 \right]\right\}.$$

Compute  $\hat{\rho} = \sum_{t=2}^T \xi_t \xi_{t-1} / \sum_{t=2}^{T-1} \xi_t^2$  and  $\hat{\delta}^2 = (1 - \hat{\rho}^2)\xi_1^2 + \sum_{t=2}^T (\xi_t - \hat{\rho}\xi_{t-1})^2$ . Draw  $\delta_{new}^2 = \hat{\delta}^2 / \chi_{T-2}^2$ , and  $\rho_{new} \sim N(\hat{\rho}, \delta_{new}^2 / \sum_{t=2}^{T-1} \xi_t^2)$ , where  $\chi_{T-2}^2$  is a  $\chi^2$  random variable with  $T - 2$  degrees of freedom. Accept  $\delta_{new}^2$  and  $\rho_{new}$  if  $-0.99 \leq \rho_{new} \leq 0.99$ .

Step 3<sub>A</sub>: The target conditional density is

$$p(\rho, \delta|\beta, \kappa, Y) \propto (1 - \rho^2)^{-1/2} \exp\left\{\sum (\xi_t Y_t - d_t e^{\xi_t + X_t \beta})\right\},$$

where  $\xi$  and  $\kappa$  are related by  $\kappa_1 = \sqrt{1 - \rho^2}\xi_1/\delta$ , and  $\kappa_t = (\xi_t - \rho\xi_{t-1})/\delta$ ,  $t \geq 2$ . Define  $l(\rho, \delta) = \log p(\rho, \delta|\beta, \kappa, Y)$ .

Propose a random-walk type move,  $(\rho, \delta) \rightarrow (\rho^{new}, \delta^{new})$ , by setting  $\rho^{new} = \rho + s_1 u_1$  and  $\delta^{new} = \delta \exp\{s_2 u_2\}$ , where  $u_1, u_2$  are i.i.d Uniform $(-1/2, 1/2)$ , and  $s_1, s_2$  are suitable step sizes (which may be tuned adaptively during the burn-in period). Draw a uniform random number  $u_0$  between 0 and 1, and accept  $(\rho^{new}, \delta^{new})$  if  $-0.99 \leq \rho^{new} \leq 0.99$  and

$$u_0 \leq \exp\{l(\rho^{new}, \delta^{new}) - l(\rho, \delta) + s_2 u_2\}.$$

Repeat the entire procedure several times to achieve a reasonable acceptance rate. Keep  $\xi$  updated via (3.7).

Step 3'<sub>A</sub>: Same as Step 3<sub>A</sub>, except that we fix  $\delta$ , i.e., we set  $s_2 = 0$ .

Step 3''<sub>A</sub>: Same as Step 3<sub>A</sub>, except that we fix  $\rho$ , i.e., we set  $s_1 = 0$ .

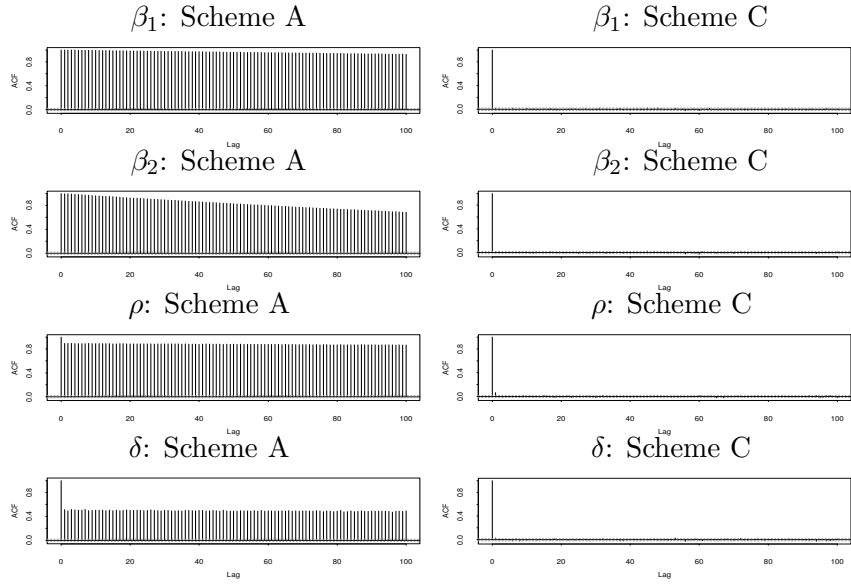


Figure A.1: Comparing Scheme A with Scheme C on DATA1. Autocorrelations of the Monte Carlo draws (excluding the burn-in period) are displayed.

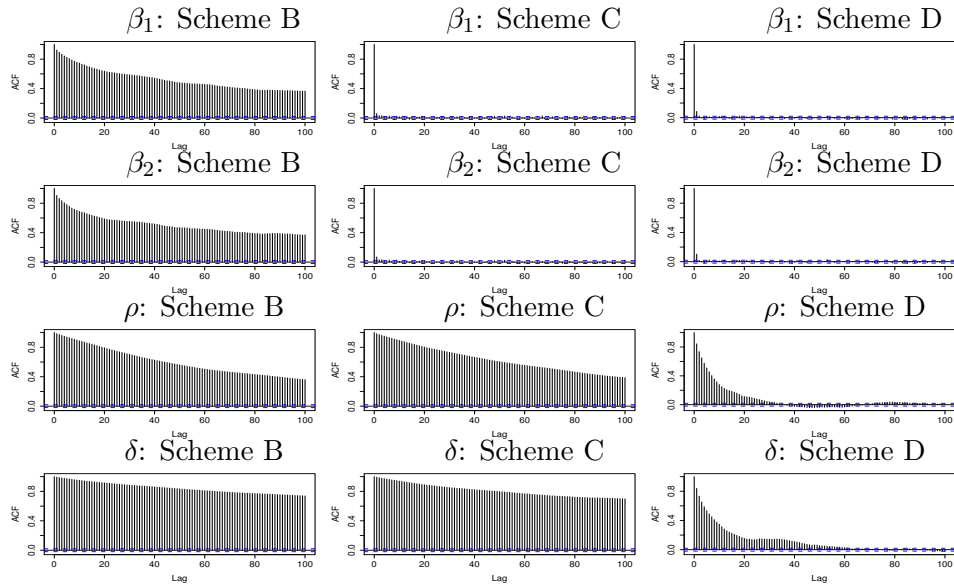


Figure A.2: Autocorrelations under Schemes B, C, and D on DATA2, after the burn-in period.

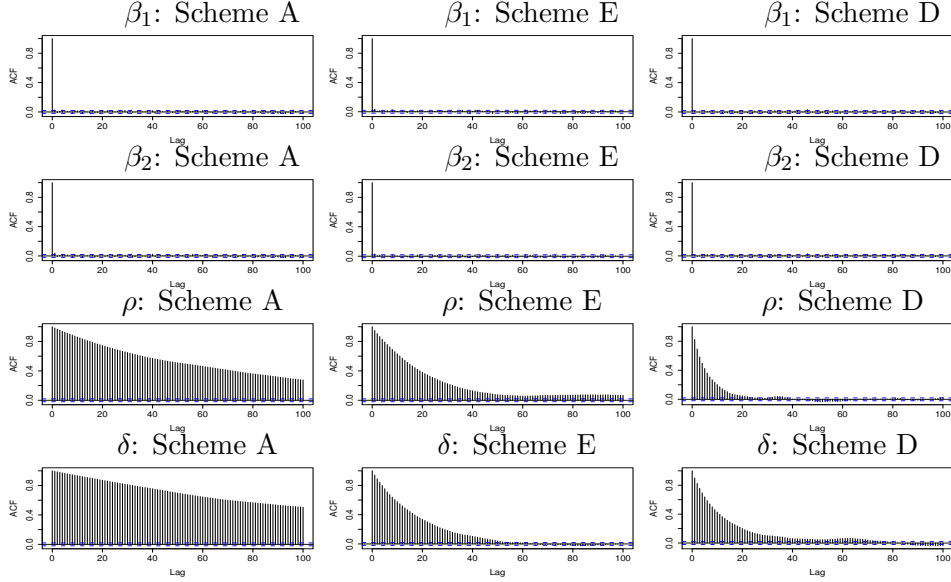


Figure A.3: Autocorrelations under Schemes A, D and E on the Chandra X-ray data, after the burn-in period.

## A.2 Autocorrelations Plots of the Monte Carlo Draws (Fig. A.1 – Fig. A.3)

# B Auxiliary Material for Section 5

## B.1 A Reducible Chain as a Result of Combining Two Transition Kernels (Fig. B.1)

## B.2 Proof of Lemma 1

*Proof.* Let  $X$  be  $\mathcal{A}_1$ -measurable and  $Z$  be  $\mathcal{A}_2$ -measurable such that  $0 < V[X - E(X|\mathcal{M})] < \infty$  and  $0 < V[Z - E(Z|\mathcal{M})] < \infty$ . Write  $X - E(X|\mathcal{M}) = X_0 + X_\perp$  with  $X_0 = E(X|\mathcal{N}) - E(X|\mathcal{M})$  and  $X_\perp = X - E(X|\mathcal{N})$ , and similarly for  $Z - E(Z|\mathcal{N})$ . Then  $X_0$  and  $Z_0$  are projections onto  $\mathcal{N}$  because  $\mathcal{M} \subset \mathcal{N}$ , and hence they both are  $\mathcal{N}$ -measurable, and

$$\begin{aligned} \text{Cov}(X_0, X_\perp) &= \text{Cov}(X_0, Z_\perp) = \text{Cov}(Z_0, Z_\perp) = \text{Cov}(Z_0, X_\perp) = 0, \\ V(X_0 + X_\perp) &= V(X_0) + V(X_\perp), \quad V(Z_0 + Z_\perp) = V(Z_0) + V(Z_\perp). \end{aligned}$$

Consequently,

$$\begin{aligned} \text{Cov}(X_0 + X_\perp, Z_0 + Z_\perp) &= \text{Cov}(X_0, Z_0) + \text{Cov}(X_\perp, Z_\perp) \\ &\leq \sqrt{V(X_0)V(Z_0)} + \mathcal{R}_{\mathcal{N}}(\mathcal{A}_1, \mathcal{A}_2)\sqrt{V(X_\perp)V(Z_\perp)}, \end{aligned} \tag{B.1}$$

by the definition of  $\mathcal{R}_{\mathcal{N}}(\mathcal{A}_1, \mathcal{A}_2)$  as in (5.4). It follows that

$$\text{Corr}(X_0 + X_\perp, Z_0 + Z_\perp) \leq R_X R_Z + \mathcal{R}_{\mathcal{N}}(\mathcal{A}_1, \mathcal{A}_2)\sqrt{(1 - R_X^2)(1 - R_Z^2)}, \tag{B.2}$$

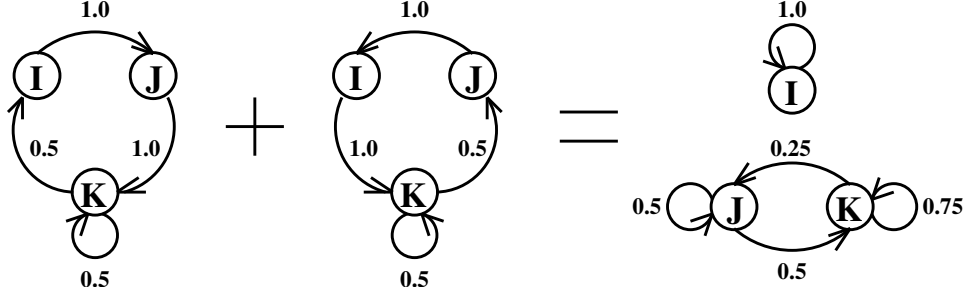


Figure B.1: Alternating two irreducible Markov chains gives a reducible chain. The state space is  $\Omega = \{I, J, K\}$ , and the target distribution is  $\pi = (1/4, 1/4, 1/2)$ . Left: transition probability specification (numbers on the arrows) of one chain. Middle: transition probabilities of a second chain with the same stationary distribution. Right: transition probabilities of the combined chain, which becomes reducible even though the original two chains are irreducible.

where (with a bit of abuse of notation)

$$R_X = \sqrt{\frac{V(X_0)}{V(X_0 + X_\perp)}} \quad \text{and} \quad R_Z = \sqrt{\frac{V(Z_0)}{V(Z_0 + Z_\perp)}},$$

where, without lose of generality, we have assumed  $V(X_0) > 0$  and  $V(Z_0) > 0$ . By the simple inequality  $\sqrt{(1 - R_X^2)(1 - R_Z^2)} \leq 1 - R_X R_Z$ , the right hand side of (B.2) is dominated by

$$\mathcal{R}_{\mathcal{N}}(\mathcal{A}_1, \mathcal{A}_2) + [1 - \mathcal{R}_{\mathcal{N}}(\mathcal{A}_1, \mathcal{A}_2)]R_X R_Z.$$

Noting  $X_0 + X_\perp = X - E(X|\mathcal{M})$  and  $X_0 = X_{\mathcal{N}} - E(X_{\mathcal{N}}|\mathcal{M})$ , where  $X_{\mathcal{N}} \equiv E(X|\mathcal{N})$ , we have

$$R_X = \frac{\text{Cov}(X_0 + X_\perp, X_0)}{\sqrt{V(X_0 + X_\perp)V(X_0)}} \leq \mathcal{R}_{\mathcal{M}}(\mathcal{A}_1, \mathcal{N}).$$

Similarly  $R_Z \leq \mathcal{R}_{\mathcal{M}}(\mathcal{A}_2, \mathcal{N})$ . The claim then follows.  $\square$

### B.3 Proof of Theorem 1

*Proof.* With the change of notation  $Y_{mis,1} = Y_{mis}$ ,  $Y_{mis,2} = \tilde{Y}_{mis}$ , each iteration of GIS as defined in Section 2.3 can be represented by a directed graph, as in (2.13),

$$\theta^{(t)} \longrightarrow Y_{mis,1}^{(t)} \longrightarrow Y_{mis,2}^{(t+1)} \longrightarrow \theta^{(t+1)}.$$

That is,  $\theta^{(t)}$  and  $Y_{mis,2}^{(t+1)}$  are conditionally independent given  $Y_{mis,1}^{(t)}$ , etc. Let us focus on the marginal chain  $\{\theta^{(t)}\}$  and bound its spectral radius  $r_{1\&2}$ :

$$r_{1\&2} \leq \mathcal{R}(\theta^{(t)}, \theta^{(t+1)}) \leq \mathcal{R}(\theta^{(t)}, Y_{mis,1}^{(t)}) \mathcal{R}(Y_{mis,1}^{(t)}, Y_{mis,2}^{(t+1)}) \mathcal{R}(Y_{mis,2}^{(t+1)}, \theta^{(t+1)}), \quad (\text{B.3})$$

where we apply (5.6) twice for the last inequality. Under stationarity  $\mathcal{R}(\theta^{(t)}, Y_{mis,1}^{(t)})$  is the maximal correlation between  $\theta$  and  $Y_{mis,1}$  in their joint posterior distribution, and likewise for  $\mathcal{R}(Y_{mis,2}^{(t+1)}, \theta^{(t+1)})$ .

They are related to  $r_1$  and  $r_2$ , the convergence rates of the two ordinary DA schemes via (see Liu *et al.* 1994, 1995)

$$r_1 = \mathcal{R}^2(Y_{mis,1}^{(t)}, \theta^{(t)}), \quad \text{and} \quad r_2 = \mathcal{R}^2(\theta^{(t+1)}, Y_{mis,2}^{(t+1)}).$$

Under stationarity, the distribution of  $\{Y_{mis,1}^{(t)}, Y_{mis,2}^{(t+1)}\}$  is simply the joint posterior of  $\{Y_{mis,1}, Y_{mis,2}\}$  (with  $\theta$  integrated out). Hence Theorem 1 follows from (B.3).  $\square$

## B.4 Proof of Theorem 2

*Proof.* We use the same notation as in the proof of Theorem 1. Letting  $\mathcal{A} = \sigma(Y_{mis,1}^{(t)}) \cap \sigma(Y_{mis,2}^{(t+1)})$ , and applying Lemma 1, we get

$$\begin{aligned} r_{1\&2} &\leq \mathcal{R}(\theta^{(t)}, \theta^{(t+1)}) \\ &\leq \mathcal{R}_{\mathcal{A}}(\theta^{(t)}, \theta^{(t+1)}) + (1 - \mathcal{R}_{\mathcal{A}}(\theta^{(t)}, \theta^{(t+1)}))\mathcal{R}(\theta^{(t)}, \mathcal{A})\mathcal{R}(\mathcal{A}, \theta^{(t+1)}) \\ &= \mathcal{R}^2(\theta, \mathcal{N}) + (1 - \mathcal{R}^2(\theta, \mathcal{N}))\mathcal{R}_{\mathcal{A}}(\theta^{(t)}, \theta^{(t+1)}). \end{aligned}$$

In the last equality, we have used the fact that under stationarity,  $\mathcal{R}(\mathcal{A}, \theta^{(t+1)}) = \mathcal{R}(\theta^{(t)}, \mathcal{A}) = \mathcal{R}(\theta, \mathcal{N})$ .

Letting  $\mathcal{B} = \sigma(Y_{mis,1}^{(t)})$ , and applying Lemma 1 again, we get

$$\begin{aligned} \mathcal{R}_{\mathcal{A}}(\theta^{(t)}, \theta^{(t+1)}) &\leq \mathcal{R}_{\mathcal{B}}(\theta^{(t)}, \theta^{(t+1)}) + (1 - \mathcal{R}_{\mathcal{B}}(\theta^{(t)}, \theta^{(t+1)}))\mathcal{R}_{\mathcal{A}}(\theta^{(t)}, \mathcal{B})\mathcal{R}_{\mathcal{A}}(\mathcal{B}, \theta^{(t+1)}) \\ &= \mathcal{R}_{\mathcal{A}}(\theta^{(t)}, Y_{mis,1}^{(t)})\mathcal{R}_{\mathcal{A}}(Y_{mis,1}^{(t)}, \theta^{(t+1)}), \end{aligned}$$

because  $\mathcal{R}_{\mathcal{B}}(\theta^{(t)}, \theta^{(t+1)}) = 0$  by conditional independence. Similarly, by taking  $\mathcal{B} = \sigma(Y_{mis,2}^{(t+1)})$  and applying Lemma 1, we conclude

$$\mathcal{R}_{\mathcal{A}}(Y_{mis,1}^{(t)}, \theta^{(t+1)}) \leq \mathcal{R}_{\mathcal{A}}(Y_{mis,1}^{(t)}, Y_{mis,2}^{(t+1)})\mathcal{R}_{\mathcal{A}}(Y_{mis,2}^{(t+1)}, \theta^{(t+1)}).$$

Theorem 2 then follows from these three inequalities because, under stationarity, it is easy to show that  $\mathcal{R}_{\mathcal{A}}(\theta^{(t)}, Y_{mis,1}^{(t)}) = \mathcal{R}_{\mathcal{N}}(\theta, Y_{mis,1})$ ,  $\mathcal{R}_{\mathcal{A}}(Y_{mis,2}^{(t+1)}, \theta^{(t+1)}) = \mathcal{R}_{\mathcal{N}}(Y_{mis,2}, \theta)$ , and  $\mathcal{R}_{\mathcal{A}}(Y_{mis,1}^{(t)}, Y_{mis,2}^{(t+1)}) = \mathcal{R}_{\mathcal{N}}(Y_{mis,1}, Y_{mis,2})$ .  $\square$

## B.5 Proof of Theorem 3

*Proof.* To prove (5.11), we start by taking  $X = \theta^{(t)} = \{\theta_1^{(t)}, \dots, \theta_J^{(t)}\}$ ,  $Z = \theta^{(t+1)} = \{\theta_1^{(t+1)}, \dots, \theta_J^{(t+1)}\}$ , and  $Y = \theta_1^{(t+1)}$ , all with respect to the joint stationary distribution  $\{\theta^{(t)}, \theta^{(t+1)}\}$ . We then apply the following version of the key inequality (5.8)

$$\mathcal{S}_W(X, Z) \geq \mathcal{S}_Y(X, Z)[\mathcal{S}_W(X, Y) + \mathcal{S}_W(Y, Z) - \mathcal{S}_W(X, Y)\mathcal{S}_W(Y, Z)], \quad (\text{B.4})$$

where  $W$  is also a part of  $\{\theta^{(t)}, \theta^{(t+1)}\}$  such that  $\sigma(W) \subset \sigma(Y)$ . Since  $Y$  is a part of  $Z$  and hence  $\mathcal{S}(Y, Z) = 0$ , we first take a trivial  $W = 0$  in (B.4) to arrive at

$$\mathcal{S}_{CIS} \equiv \mathcal{S}(\theta^{(t)}, \theta^{(t+1)}) \geq \mathcal{S}_{\theta_1^{(t+1)}}(\theta^{(t)}, \theta^{(t+1)})\mathcal{S}(\theta^{(t)}, \theta_1^{(t+1)}). \quad (\text{B.5})$$

Keeping the same  $X$  and  $Z$ , but now taking  $Y = \theta_{\leq 2}^{(t+1)}$  and  $W = \theta_1^{(t+1)}$ , we apply (B.4) again to obtain

$$\mathcal{S}_{\theta_1^{(t+1)}}(\theta^{(t)}, \theta^{(t+1)}) \geq \mathcal{S}_{\theta_{\leq 2}^{(t+1)}}(\theta^{(t)}, \theta^{(t+1)}) \mathcal{S}_{\theta_1^{(t+1)}}(\theta^{(t)}, \theta_{\leq 2}^{(t+1)}). \quad (\text{B.6})$$

Combining (B.5) and (B.6), we see

$$\mathcal{S}_{CIS} \geq \mathcal{S}_{\theta_{\leq 3}^{(t+1)}}(\theta^{(t)}, \theta^{(t+1)}) \prod_{j=1}^2 \mathcal{S}_{\theta_{\leq j}^{(t+1)}}(\theta^{(t)}, \theta_{\leq j}^{(t+1)}). \quad (\text{B.7})$$

We continue the above argument by taking  $Y = \theta_{\leq k}^{(t+1)}$  and  $W = \theta_{< k}^{(t+1)}$ , and applying (B.4) to  $\mathcal{S}_{\theta_{< k}^{(t+1)}}(\theta^{(t)}, \theta^{(t+1)})$  for  $k = 3, \dots, J-1$ , to reach

$$\mathcal{S}_{CIS} \geq \prod_{j=1}^J \mathcal{S}_{\theta_{\leq j}^{(t+1)}}(\theta^{(t)}, \theta_{\leq j}^{(t+1)}). \quad (\text{B.8})$$

To further factor each term on the right hand side of (B.8), let us first take  $X = \theta^{(t)}$ ,  $Z = \theta_{\leq j}^{(t+1)}$ ,  $Y = (\theta_{> j}^{(t)}, \theta_{< j}^{(t+1)})$  and  $W = \theta_{< k}^{(t+1)}$ , and apply (B.4) again. Note that as long as  $j < J$ ,  $\mathcal{S}_{\theta_{< j}^{(t+1)}}(X, Y) = 0$ , and hence by (B.4), we have

$$\mathcal{S}_{\theta_{< j}^{(t+1)}}(\theta^{(t)}, \theta_{\leq j}^{(t+1)}) \geq \mathcal{S}_Y(\theta^{(t)}, \theta_{\leq j}^{(t+1)}) \mathcal{S}_{\theta_{< j}^{(t+1)}}(Y, \theta_{\leq j}^{(t+1)}), \quad j = 1, \dots, J-1. \quad (\text{B.9})$$

To deal with the first term on the right-hand side of (B.9), we need the fact that if  $X_1$  and  $Z_2$  are conditionally independent given  $\{X_2, X_3, Z_1\}$ , then

$$\mathcal{S}_{(Z_1, X_3)}((X_1, X_2, X_3), Z_2) \geq \mathcal{S}_{(Z_1, X_3)}((Z_1, X_2, X_3), Z_2). \quad (\text{B.10})$$

If we let  $X_1 = \theta_{< j}^{(t)}$ ,  $X_2 = \theta_j^{(t)}$ ,  $X_3 = \theta_{> j}^{(t)}$ ,  $Z_1 = \theta_{< j}^{(t+1)}$ , and  $Z_2 = \theta_{\leq j}^{(t+1)}$ , then we can apply (B.10) to  $\mathcal{S}_Y(\theta^{(t)}, \theta_{\leq j}^{(t+1)})$  because by construction,  $\theta_j^{(t+1)}$  and hence  $\theta_{\leq j}^{(t+1)}$  is independent of  $\theta_{< j}^{(t)}$  when *conditional* on  $\theta^{(t+\frac{i-1}{J})}$ , the output of the CIS sampler just before the  $j$ th component is updated, which is exactly  $(\theta_{< j}^{(t+1)}, \theta_j^{(t)}, \theta_{> j}^{(t)}) \equiv \{Z_1, X_2, X_3\}$ . Thus

$$\mathcal{S}_Y(\theta^{(t)}, \theta_{\leq j}^{(t+1)}) \geq \mathcal{S}_Y(\theta^{(t+\frac{i-1}{J})}, \theta_{\leq j}^{(t+1)}) \geq \mathcal{S}_j, \quad j = 1, \dots, J-1, \quad (\text{B.11})$$

where  $\mathcal{S}_j$  is defined by (5.10). The last inequality in (B.11) is due to the easily verifiable inequality  $\mathcal{S}_{\mathcal{M}}(\mathcal{A}_1, \mathcal{A}_2) \geq \mathcal{S}_{\mathcal{M}}(\mathcal{A}_1, \sigma(\mathcal{A}_2 \cup \mathcal{M}))$ , and the fact that  $\sigma(Y) = \sigma_{j-1} \cap \sigma_j$  and  $\sigma(\sigma(\theta_{\leq j}^{(t+1)}) \cup \sigma(Y)) = \sigma_j$ .

For  $j = J$ , (B.10) still applies as long as we take  $X_3 = \theta_{> J}^{(t)} = 0$ . It then becomes

$$\mathcal{S}_{\theta_{< J}^{(t+1)}}(\theta^{(t)}, \theta_{\leq J}^{(t+1)}) \geq \mathcal{S}_{\theta_{< J}^{(t+1)}}((\theta_{< J}^{(t+1)}, \theta_J^{(t)}, \theta_{\leq J}^{(t+1)}) = \mathcal{S}_J. \quad (\text{B.12})$$

Combining (B.9), (B.11) and (B.12) leads to

$$\mathcal{S}_{CIS} \geq \left( \prod_{j=1}^J \mathcal{S}_j \right) \left[ \prod_{j=1}^{J-1} \mathcal{S}_{\theta_{< j}^{(t+1)}}((\theta_{> j}^{(t)}, \theta_{< j}^{(t+1)}), \theta_{\leq j}^{(t+1)}) \right]. \quad (\text{B.13})$$

To show that  $\tilde{\mathcal{S}}_G$ , the second product on the right hand side of (B.13), is completely determined by  $\pi$ , we note that  $(\theta_{<j}^{(t+1)}, \theta_j^{(t+1)}, \theta_{>j}^{(t)})$  is simply  $\theta^{(t+\frac{j}{J})}$  in (2.27), which follows  $\pi$  assuming the CIS chain is stationary. We can write  $\tilde{\mathcal{S}}_G$  as in (5.12) because  $\sigma(\theta_{<j}^{(t+1)}) = \sigma_{j-1} \cap \sigma_J$  and  $\sigma(\theta_{>j}^{(t)}, \theta_{<j}^{(t+1)}) = \sigma_{j-1} \cap \sigma_j$ ,  $j = 1, \dots, J$ .

To show  $\mathcal{S}_G \geq \tilde{\mathcal{S}}_G$ , we note that, stochastically, drawing  $\theta_j^{(t+1)}$  directly from its full conditional is the same as having  $Y_{mis,j}$  and  $\tilde{Y}_{mis,j}$  conditionally independent given  $\theta_{>j}^{(t)}$  and  $\theta_{<j}^{(t+1)}$ . Hence  $\mathcal{S}_G \geq \tilde{\mathcal{S}}_G$  is a special case of (B.13) with  $\mathcal{S}_j = 1$  for all  $j = 1, \dots, J$ .

To show  $\mathcal{S}_G = \tilde{\mathcal{S}}_G$  when  $J = 2$ , we first note that when  $J = 2$ , we have  $\mathcal{S}_G = 1 - \mathcal{R}(\theta_1, \theta_2)$ , where the MCC calculation is with respect to  $\pi$  (see Liu *et al.*, 1994). However, by the definition of  $\tilde{\mathcal{S}}_G$ , we also have  $\mathcal{S}(\theta_1, \theta_2) = 1 - \mathcal{R}(\theta_1, \theta_2)$ , and the claim follows.  $\square$

## B.6 Proof of Theorem 4

*Proof.* Because  $Y_{mis}$  is sufficient for  $\theta$ , we can write  $p(Y_{obs}|Y_{mis}, \theta)$  as  $g(Y_{obs}; Y_{mis})$ . Similarly, because  $\tilde{Y}_{mis}$  is ancillary, we can write  $p(\tilde{Y}_{mis}|\theta)$  as  $f(\tilde{Y}_{mis})$ . Then the joint posterior density of  $(\theta, \tilde{Y}_{mis})$ , with respect to the joint product measure of the Haar measure  $H(\cdot)$  for  $\theta$  and Lebesgue measure for  $\tilde{Y}_{mis}$ , is

$$\begin{aligned} p(\theta, \tilde{Y}_{mis}|Y_{obs}) &\propto p(Y_{obs}|\tilde{Y}_{mis}, \theta)p(\tilde{Y}_{mis}|\theta)p_0(\theta) \\ &\propto p(Y_{obs}|Y_{mis} = M_\theta^{-1}(\tilde{Y}_{mis}), \theta)p(\tilde{Y}_{mis}|\theta)p_0(\theta) \\ &\propto g(Y_{obs}; M_\theta^{-1}(\tilde{Y}_{mis}))f(\tilde{Y}_{mis})p_0(\theta). \end{aligned} \quad (\text{B.14})$$

Hence the conditional draw at Step  $2_A$  of the interwoven scheme is

$$\theta|(\tilde{Y}_{mis}, Y_{obs}) \sim p_0(\theta)g(Y_{obs}; M_\theta^{-1}(\tilde{Y}_{mis})). \quad (\text{B.15})$$

Noting (B.14) and  $Y_{mis} = M_\theta^{-1}(\tilde{Y}_{mis})$ , the joint posterior of  $(\theta, Y_{mis})$  is

$$p(\theta, Y_{mis}|Y_{obs}) \propto p_0(\theta)f(M_\theta(Y_{mis}))g(Y_{obs}; Y_{mis})J(\theta, Y_{mis}),$$

where  $J(\theta, Y_{mis}) = |\det[\partial M(Y_{mis}; \theta)/\partial Y_{mis}]|$ . Hence the conditional draw at Step  $2_S$  of the interwoven scheme is

$$\theta|(Y_{mis}, Y_{obs}) \sim p_0(\theta)f(M_\theta(Y_{mis}))J(\theta, Y_{mis}). \quad (\text{B.16})$$

Consider the PX-DA algorithm specified by the Theorem. According to Liu and Wu (1999), when Condition C1 is satisfied we can equivalently implement the optimal PX-DA algorithm (with the uniform prior density on  $\alpha$  with respect to the Haar measure) as follows:

(1) Set  $\alpha = e$  (identity element of the group). Draw  $Y_{mis}|(\theta, Y_{obs})$ , which is the same as Step 1 of ASIS. Let  $z = Y_{mis}$ .

(2) Draw  $(\alpha, \theta)|(Y_{mis}^\alpha = z, Y_{obs})$  jointly. This can be accomplished by drawing  $\alpha|(z, Y_{obs})$  and then  $\theta|(\alpha, z, Y_{obs})$ . We first observe that the joint posterior of  $(\alpha, \theta)$  can be expressed as

$$p(\alpha, \theta|Y_{mis}^\alpha, Y_{obs}) \propto p(Y_{obs}|Y_{mis}^\alpha, \alpha, \theta)p(Y_{mis}^\alpha|\alpha, \theta)p_0(\theta)p(\alpha). \quad (\text{B.17})$$



Since  $Y_{mis}^\alpha = M_\alpha(Y_{mis})$ , we have

$$\begin{aligned} p(Y_{obs}|Y_{mis}^\alpha, \alpha, \theta) &= p(Y_{obs}|Y_{mis} = M_\alpha^{-1}(Y_{mis}^\alpha), \alpha, \theta) \\ &= g(Y_{obs}; M_\alpha^{-1}(Y_{mis}^\alpha)). \end{aligned} \quad (\text{B.18})$$

But we also have  $Y_{mis}^\alpha = M_\alpha(Y_{mis}) = M_\alpha(M_\theta^{-1}(\tilde{Y}_{mis})) = M_{\alpha\theta^{-1}}(\tilde{Y}_{mis})$ . Therefore we may obtain  $p(Y_{mis}^\alpha|\theta, \alpha)$  via  $p(\tilde{Y}_{mis}|\theta) = f(\tilde{Y}_{mis})$ . That is,

$$p(Y_{mis}^\alpha|\theta, \alpha) \propto f(M_{\theta\alpha^{-1}}(Y_{mis}^\alpha))J(\theta \cdot \alpha^{-1}, Y_{mis}^\alpha). \quad (\text{B.19})$$

Substituting (B.18–B.19) into (B.17) and noting  $p(\alpha) \propto 1$ , we have

$$p(\alpha, \theta|z, Y_{obs}) \propto p_0(\theta)f(M_{\theta\alpha^{-1}}(z))g(Y_{obs}; M_{\alpha^{-1}}(z))J(\theta \cdot \alpha^{-1}, z),$$

where  $z$  is used as a shorthand for  $Y_{mis}^\alpha$ . Now integrate out  $\theta$ :

$$\begin{aligned} p(\alpha|z, Y_{obs}) &\propto g(Y_{obs}; M_{\alpha^{-1}}(z)) \int p_0(\theta)f(M_{\theta\alpha^{-1}}(z))J(\theta \cdot \alpha^{-1}, z) H(d\theta) \\ (\text{letting } \theta' = \theta \cdot \alpha^{-1}) &\propto g(Y_{obs}; M_{\alpha^{-1}}(z)) \int p_0(\theta' \cdot \alpha)f(M_{\theta'}(z))J(\theta', z) H(d\theta') \\ (\text{by Condition C2}) &\propto g(Y_{obs}; M_{\alpha^{-1}}(z)) \int p_0(\theta')p_0(\alpha)f(M_{\theta'}(z))J(\theta', z) H(d\theta') \\ &\propto g(Y_{obs}; M_{\alpha^{-1}}(z))p_0(\alpha). \end{aligned} \quad (\text{B.20})$$

On the other hand

$$\begin{aligned} p(\theta|\alpha, z, Y_{obs}) &\propto p_0(\theta)f(M_{\theta\alpha^{-1}}(z))J(\theta \cdot \alpha^{-1}, z) \\ &\propto p_0(\theta)f(M_\theta(M_\alpha^{-1}(z)))J(\theta, M_\alpha^{-1}(z)), \end{aligned}$$

which matches equation (B.16), i.e.,  $p(\theta|Y_{mis}, Y_{obs})$ , for  $Y_{mis} = M_\alpha^{-1}(z)$ .

In summary, when the current iterate is  $\theta^{(t)}$ , the steps of PX-DA are

Step 1. Same as Step 1 of ASIS.

Step 2a. Let  $z = Y_{mis}$ , and draw  $\alpha|(z, Y_{obs})$  according to (B.20).

Step 2b. Let  $z' = M_\alpha^{-1}(z)$ , and draw  $\theta^{(t+1)} \sim p(\theta|Y_{mis} = z', Y_{obs})$ .

Put  $\alpha' = \theta^{(t)} \cdot \alpha$ . Based on (B.20), Step 2a is equivalent to drawing  $\alpha'$  according to

$$\begin{aligned} p(\alpha'|z, Y_{obs}) &\propto g(Y_{obs}; M_{\alpha'^{-1}\theta^{(t)}}(z))p_0([\theta^{(t)}]^{-1} \cdot \alpha') \\ &\propto g(Y_{obs}; M_{\alpha'^{-1}}(w))p_0(\alpha'), \end{aligned} \quad (\text{B.21})$$

where  $w = M_{\theta^{(t)}}(z)$ . Note this  $w$  is the same as  $\tilde{Y}_{mis}$  in Step 2<sub>A</sub> of ASIS, because  $z = Y_{mis}$ . Observe that (B.21) matches  $p(\theta|\tilde{Y}_{mis} = w, Y_{obs})$  of (B.15) when we equate  $\theta$  with  $\alpha'$ . Therefore if we correspond  $\alpha'$  with  $\theta^{(t+.5)}$ , which is the output of Step 2<sub>A</sub> of ASIS, then Step 2a is the same as Step 2<sub>A</sub>. Furthermore, with  $\alpha' = \theta^{(t+.5)}$ , in Step 2b  $z' = M_\alpha^{-1}(z) = M_{\alpha'}^{-1}(w) = Y_{mis}$ , and we can draw an exact correspondence between Step 2b of PX-DA and Step 2<sub>S</sub> of ASIS as well. (Note here the Step 2<sub>A</sub> and Step 2<sub>S</sub> are in the reversed order, if we match the notation with that for GIS, as defined in Section 2.2; but recall the order does not affect the validity.)  $\square$

## C Auxiliary Material for Section 6

The following example illustrates both the relevance and the limitations of Theorem 4. Consider the univariate  $t$  model, a well known model for which PX-DA can be applied. We observe  $Y_{obs} = (y_1, \dots, y_n)$ , where

$$y_i \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2/q_i), \quad q_i \stackrel{\text{i.i.d.}}{\sim} \chi_\nu^2/\nu.$$

The parameters are  $\theta = (\mu, \sigma)$  and the missing data are  $q = (q_1, \dots, q_n)^\top$ . The degree of freedom  $\nu$  is assumed known. Assume the standard flat prior on  $(\mu, \log(\sigma))$ . By introducing a parameter  $\alpha$ , this model can be expanded into

$$y_i \stackrel{\text{i.i.d.}}{\sim} N(\mu, \alpha\sigma^2/w_i), \quad w_i \stackrel{\text{i.i.d.}}{\sim} \alpha\chi_\nu^2/\nu,$$

where  $w_i = \alpha q_i$ . Each iteration of the optimal PX-DA algorithm (see Liu and Wu 1999, and Meng and van Dyk 1999) can be written compactly as

- (1) Draw  $q_i \sim \chi_{\nu+1}^2 / [(y_i - \mu)^2/\sigma^2 + \nu]$ , independently for  $i = 1, \dots, n$ ;
- (2) Compute  $\hat{\mu} = \sum_{i=1}^n q_i y_i / \sum_{i=1}^n q_i$ , and then draw

$$\sigma^2 \sim \left[ \sum_{i=1}^n q_i (y_i - \hat{\mu})^2 \right] / \chi_{n-1}^2, \quad \mu \sim N \left[ \hat{\mu}, \sigma^2 / \sum_{i=1}^n q_i \right];$$

- (3) Redraw  $\sigma^2 \sim \sigma^2 \chi_{n\nu}^2 / (\nu \sum_{i=1}^n q_i)$ .

These three steps are simply the following conditional draws under the original model:

- (1)  $q | (\mu, \sigma, Y_{obs})$ ;
- (2)  $(\mu, \sigma) | (q, Y_{obs})$ ;
- (3)  $\sigma | (\mu, z, Y_{obs})$ , where  $z = (z_1, \dots, z_n)^\top = (q_1/\sigma^2, \dots, q_n/\sigma^2)^\top$ .

In other words, Step 1 draws  $q$ , the missing data, given  $\theta = (\mu, \sigma)$ ; Step 2 draws  $\theta$  given  $q$ , which is an AA for  $\theta$ ; and Step 3 draws  $\sigma$  given  $z$ , which is an SA for  $\sigma$ . If we focus on  $\sigma$ , ignoring the part for  $\mu$ , then the above algorithm is exactly an ASIS sampler for  $\sigma$ ; just as Theorem 4 claims, it coincides with the optimal PX-DA algorithm. However, because  $z$  is not an SA for  $\mu$ , this scheme does not correspond to an ASIS for  $(\mu, \sigma)$  as a whole. This suggests that there may be a generalization of Theorem 4 that deals with a form of conditional ASIS. Such results would also shed light on optimality properties of CIS, or reveal an even better formulation.