

Convergence analysis of data augmentation algorithms for Bayesian robust multivariate linear regression with incomplete data

Haoxiang Li^{a,*}, Qian Qin^a, Galin L. Jones^a

^a*School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA*

Abstract

Gaussian mixtures are commonly used for modeling heavy-tailed error distributions in robust linear regression. Combining the likelihood of a multivariate robust linear regression model with a standard improper prior distribution yields an analytically intractable posterior distribution that can be sampled using a data augmentation algorithm. When the response matrix has missing entries, there are unique challenges to the application and analysis of the convergence properties of the algorithm. Conditions for geometric ergodicity are provided when the incomplete data have a “monotone” structure. In the absence of a monotone structure, an intermediate imputation step is necessary for implementing the algorithm. In this case, we provide sufficient conditions for the algorithm to be Harris ergodic. Finally, we show that, when there is a monotone structure and intermediate imputation is unnecessary, intermediate imputation slows the convergence of the underlying Monte Carlo Markov chain, while post hoc imputation does not. An R package for the data augmentation algorithm is provided.

Keywords: Drift and minorization, Geometric ergodicity, Markov chain Monte Carlo, Missing data imputation.

2020 MSC: Primary 60J05, Secondary 62F15

1. Introduction

Markov chain Monte Carlo (MCMC) algorithms are frequently used for sampling from intractable probability distributions in Bayesian statistics. We study the convergence properties of a class of MCMC algorithms designed for Bayesian multivariate linear regression with heavy-tailed errors. Contrary to most existing works in the area of MCMC convergence analysis, we focus on scenarios where the data set is incomplete. The existence of missing data substantially complicates the analysis.

Gaussian mixtures are suitable for constructing heavy-tailed error distributions in robust linear regression models [4, 13, 33]. In a Bayesian setting where a simple improper prior is used, the mixture representation facilitates a data augmentation (DA) MCMC algorithm, proposed by Liu [15], that can be used to sample from the posterior distribution of the regression coefficients and error scatter matrix. When there are no missing data, this algorithm is geometrically ergodic under regularity conditions [9, 29]. Loosely speaking, geometric ergodicity means that the Markov chain converges to the posterior distribution at a geometric, or exponential rate. Geometric ergodicity is important because it guarantees the existence of a central limit theorem for ergodic averages, which is in turn crucial for assessing the accuracy of Monte Carlo estimators [2, 5, 10–12, 31, 32].

The current work studies Liu’s DA algorithm [15] when the response matrix has missing entries. An incomplete data set brings unique challenges to the implementation and analysis of the DA algorithm. For instance, establishing posterior propriety becomes much more difficult than when there are no missing values. If the posterior measure is not proper, the DA algorithm will produce nonsensical results [8]. However, if one can show that the algorithm is geometrically ergodic, then the posterior distribution is guaranteed to be proper.

When the missing data have a certain “monotone” structure, the DA algorithm can be carried out without an intermediate step to impute the missing data. In this case, we establish geometric ergodicity under conditions on the error

*Corresponding author. Email address: li001294@umn.edu

distribution and the amount of incomplete response components. Roughly speaking, when the mixing distribution in the Gaussian mixture representation of the error distribution places little mass near the origin and the number of incomplete components is not too large, the DA algorithm is geometrically ergodic. The conditions are satisfied by many mixing distributions, including distributions with finite support, log-normal, generalized inverse Gaussian, and inverse Gamma or Fréchet with shape parameter greater than $d/2$, where d is the dimension of the response variable. Some Gamma, Weibull, and F distributions also satisfy the conditions. A post hoc imputation step can be added to fill in the missing values and the convergence properties of the DA algorithm will be unaffected.

When the missing data do not possess a monotone structure, some missing entries need to be imputed to implement the DA algorithm. This results in a data augmentation algorithm with an intermediate (as opposed to post hoc) imputation step, which we call a DAI algorithm for short. We provide sufficient conditions for the DAI algorithm to be Harris ergodic. Harris ergodicity is weaker than geometric ergodicity, but it guarantees posterior propriety as well as the existence of a law of large numbers for ergodic averages [17, Theorem 17.1.7].

When the missing data have a monotone structure, both the DA (with or without post hoc imputation) and DAI algorithms can be applied. However, we show that the DA algorithm converges in L^2 at least as fast as the DAI algorithm.

Our key strategy is to draw a connection from cases where the data set is incomplete to the standard case where the data set is fully observed. This allows for an analysis of the former using tools built for the latter.

Finally, we provide an R package `Bayesianrobust` that implements the DA and DAI algorithms. The R package is available from <https://github.com/haoxiangliu/Bayesianrobust>. While the algorithms were proposed by Liu [15], we do not know of software that implements them.

The remainder is organized as follows. Section 2 recounts the Bayesian robust linear regression model with incomplete data. In Section 3, we describe the DA and DAI algorithms. Our main results are in Section 4, where we provide conditions for geometric ergodicity of the DA algorithm and Harris ergodicity of the DAI algorithm. The section also contains a comparison between the DA and DAI algorithms in terms of L^2 convergence rate. Section 5 presents a numerical experiment. We conclude our paper in Section 6.

2. Robust Linear Regression with Incomplete Data

Let (Y_i, \mathbf{x}_i) , $i \in \{1, \dots, n\}$, be n independent data points, where $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^\top$ is a $p \times 1$ vector of known predictors, and $Y_i = (Y_{i,1}, \dots, Y_{i,d})^\top$ is a $d \times 1$ random vector. Consider the multivariate linear regression model

$$Y_i = \mathbf{B}^\top \mathbf{x}_i + \boldsymbol{\Sigma}^{1/2} \boldsymbol{\varepsilon}_i, \quad i \in \{1, \dots, n\},$$

where \mathbf{B} is a $p \times d$ matrix of unknown regression coefficients, $\boldsymbol{\Sigma}$ is a $d \times d$ unknown positive definite scatter matrix, and $\boldsymbol{\varepsilon}_i$, $i \in \{1, \dots, n\}$, are $d \times 1$ iid random errors. Let \mathbf{Y} be the $n \times d$ response matrix whose i th row is Y_i^\top , and let \mathbf{x} be the $n \times p$ design matrix whose i th row is \mathbf{x}_i^\top .

To allow for error distributions with potentially heavy tails, assume that the distribution of each $\boldsymbol{\varepsilon}_i$ is described by a scale mixture of multivariate normal densities, which takes the form

$$f_{\text{err}}(\boldsymbol{\varepsilon}) = \int_0^\infty \frac{w^{d/2}}{(2\pi)^{d/2}} \exp\left(-\frac{w}{2} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}\right) P_{\text{mix}}(dw), \quad \boldsymbol{\varepsilon} \in \mathbb{R}^d,$$

where $P_{\text{mix}}(\cdot)$ is a probability measure on $(0, \infty)$ referred to as the mixing distribution. Gaussian mixtures constitute a variety of error distributions, and are widely used for robust regression. For instance, when $P_{\text{mix}}(\cdot)$ corresponds to the $\text{Gamma}(v/2, v/2)$ distribution for some $v > 0$, i.e., $P_{\text{mix}}(\cdot)$ has a density with respect to the Lebesgue measure given by

$$p_{\text{mix}}(w) \propto w^{v/2-1} \exp(-vw/2), \quad w > 0,$$

the errors follow the multivariate t distribution with v degrees of freedom, which has density function

$$f_{\text{err}}(\boldsymbol{\varepsilon}) \propto (1 + \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}/v)^{-(d+v)/2}, \quad \boldsymbol{\varepsilon} \in \mathbb{R}^d.$$

See Section 4.2 for more examples.

Consider a Bayesian setting. Throughout, we use $(\boldsymbol{\beta}, \boldsymbol{\varsigma})$ to denote values of $(\mathbf{B}, \boldsymbol{\Sigma})$. Assume that $(\mathbf{B}, \boldsymbol{\Sigma})$ has the following prior density:

$$p_{\text{prior}}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \propto |\boldsymbol{\varsigma}|^{-(m+1)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\varsigma}^{-1} \mathbf{a}) \right\}, \quad \boldsymbol{\beta} \in \mathbb{R}^{p \times d}, \quad \boldsymbol{\varsigma} \in S_+^{d \times d}, \quad (1)$$

where $m \in \mathbb{R}$, $\mathbf{a} \in S_+^{d \times d}$, and $S_+^{d \times d}$ is the convex cone of $d \times d$ (symmetric) positive semi-definite real matrices. Equation (1) defines a class of commonly used default priors [3, 15]. For instance, the independence Jeffrey's prior corresponds to $m = d$ and $\mathbf{a} = \mathbf{0}$. Denote by P_{prior} the measure associated with p_{prior} . The model of interest can then be written in the hierarchical form:

$$\mathbf{Y}_i | \mathbf{W}, \mathbf{B}, \boldsymbol{\Sigma} \stackrel{\text{ind}}{\sim} \text{N}_d(\mathbf{B}^\top \mathbf{x}_i, \mathbf{W}_i^{-1} \boldsymbol{\Sigma}), \quad \mathbf{W}_i | \mathbf{B}, \boldsymbol{\Sigma} \stackrel{\text{ind}}{\sim} P_{\text{mix}}(\cdot), \quad i \in \{1, \dots, n\}; \quad \mathbf{B}, \boldsymbol{\Sigma} \sim P_{\text{prior}}(\cdot),$$

where N_d denotes d -variate normal distributions, and $\mathbf{W} := (\mathbf{W}_1, \dots, \mathbf{W}_n)^\top$ is a vector of latent random variables. The joint measure of $(\mathbf{Y}, \mathbf{W}, \mathbf{B}, \boldsymbol{\Sigma})$ is given by

$$P_{\text{joint}}(d\mathbf{y}, d\mathbf{w}, d\boldsymbol{\beta}, d\boldsymbol{\varsigma}) = p_{\text{joint}}(\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\varsigma}) d\mathbf{y} P_{\text{mix}}(d\mathbf{w}) d\boldsymbol{\beta} d\boldsymbol{\varsigma},$$

where, for $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top \in \mathbb{R}^{n \times d}$, $\mathbf{w} = (w_1, \dots, w_n)^\top \in (0, \infty)^n$, $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$, and $\boldsymbol{\varsigma} \in S_+^{d \times d}$,

$$p_{\text{joint}}(\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\varsigma}) \propto p_{\text{prior}}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \prod_{i=1}^n \frac{w_i^{d/2}}{|\boldsymbol{\varsigma}|^{1/2}} \exp \left\{ -\frac{1}{2} w_i (\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^\top \boldsymbol{\varsigma}^{-1} (\mathbf{y}_i - \boldsymbol{\beta}^\top \mathbf{x}_i) \right\}. \quad (2)$$

Throughout, conditional distributions concerning $(\mathbf{Y}, \mathbf{W}, \mathbf{B}, \boldsymbol{\Sigma})$ are to be uniquely defined through the density p_{joint} .

We assume that \mathbf{x} is fully observed, but \mathbf{Y} may contain missing values. The missing structure can be described by an $n \times d$ random matrix $\mathbf{K} = (K_{i,j})_{i=1}^n_{j=1}^d$, with $K_{i,j} = 1$ indicating $Y_{i,j}$ is observed, and $K_{i,j} = 0$ indicating $Y_{i,j}$ is missing. For a realized response matrix $\mathbf{y} = (\mathbf{y}_{i,j})_{i=1}^n_{j=1}^d \in \mathbb{R}^{n \times d}$ and a realized missing structure $\mathbf{k} = (k_{i,j})_{i=1}^n_{j=1}^d \in \{0, 1\}^{n \times d}$, let $\mathbf{y}_{(k)}$ be the array $(y_{i,j})_{(i,j) \in A(\mathbf{k})}$, where $A(\mathbf{k}) = \{(i, j) : k_{i,j} = 1\}$. Then $\mathbf{Y}_{(k)}$ gives the observable portion of \mathbf{Y} . In practice, instead of observing \mathbf{Y} , one sees $(\mathbf{K}, \mathbf{Y}_{(k)})$. Throughout, we assume that the missing data mechanism is ignorable, which means the following.

Definition 1. The missing data mechanism is ignorable if, for any $\mathbf{k} \in \{0, 1\}^{n \times d}$ and almost every $\mathbf{y} \in \mathbb{R}^{n \times d}$, the posterior distribution of $(\mathbf{B}, \boldsymbol{\Sigma})$ given $(\mathbf{K}, \mathbf{Y}_{(k)}) = (\mathbf{k}, \mathbf{y}_{(k)})$ is the same as the conditional distribution of $(\mathbf{B}, \boldsymbol{\Sigma})$ given $\mathbf{Y}_{(k)} = \mathbf{y}_{(k)}$.

If, given $(\mathbf{B}, \boldsymbol{\Sigma}) = (\boldsymbol{\beta}, \boldsymbol{\varsigma})$, the distribution of \mathbf{K} does not depend on $(\boldsymbol{\beta}, \boldsymbol{\varsigma})$, and the data are ‘‘realized missing at random’’ [30, Theorem 2], which holds if \mathbf{K} is independent of \mathbf{Y} , then the missing data mechanism is ignorable. From here on, fix a realized missing structure $\mathbf{k} = (k_{i,j})_{i=1}^n_{j=1}^d \in \{0, 1\}^{n \times d}$, and only condition on $\mathbf{Y}_{(k)}$ when studying posterior distributions. Without loss of generality, assume that each row of \mathbf{k} contains at least one nonzero element, i.e., for each i , \mathbf{Y}_i has at least one observed entry.

To write down the exact form of the posterior, we introduce some additional notation. Suppose that given $i \in \{1, \dots, n\}$, $k_{i,j} = 1$ if and only if $j \in \{j_1, \dots, j_{d_i}\}$ for some $d_i \in \{1, \dots, d\}$, where $j_1 < \dots < j_{d_i}$. Given i , let $\mathbf{c}_{i,(k)}$ be the $d_i \times d$ matrix that satisfies the following: For $\ell \in \{1, \dots, d_i\}$, in the ℓ th row of $\mathbf{c}_{i,(k)}$, all elements except the j_ℓ th one are 0, while the j_ℓ th one is 1. For instance, if $d = 4$, $d_i = 2$, $j_1 = 1$, $j_2 = 3$, then

$$\mathbf{c}_{i,(k)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Then $\mathbf{Y}_{i,(k)} := \mathbf{c}_{i,(k)} \mathbf{Y}_i$ is a vector consisting of the observed components of \mathbf{Y}_i if $\mathbf{K} = \mathbf{k}$, and we can write $\mathbf{Y}_{(k)}$ as $(\mathbf{Y}_{i,(k)})_{i=1}^n$. For a realized value of \mathbf{Y} , say $\mathbf{y} = (\mathbf{y}_{i,j})_{i=1}^n_{j=1}^d \in \mathbb{R}^{n \times d}$, denote by \mathbf{y}_i the i th row of \mathbf{y} transposed, let $\mathbf{y}_{i,(k)} = \mathbf{c}_{i,(k)} \mathbf{y}_i$, and let $\mathbf{y}_{(k)} = (\mathbf{y}_{i,(k)})_{i=1}^n$. Based on (2), the posterior density of $(\mathbf{B}, \boldsymbol{\Sigma})$ given $\mathbf{Y}_{(k)} = \mathbf{y}_{(k)}$ has the following form:

$$\pi_k(\boldsymbol{\beta}, \boldsymbol{\varsigma} | \mathbf{y}_{(k)}) \propto p_{\text{prior}}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \prod_{i=1}^n \int_0^\infty \frac{w_i^{d_i/2}}{|\mathbf{c}_{i,(k)} \boldsymbol{\varsigma} \mathbf{c}_{i,(k)}^\top|^{1/2}} \exp \left(-\frac{r_{i,(k)} w_i}{2} \right) P_{\text{mix}}(dw_i),$$

| | | | | |
|-------------|-------------|-----------------|----------|-----------------|
| Pattern 1 | $y_{1,1}$ | $y_{1,2}$ | \cdots | $y_{1,d}$ |
| | \vdots | \vdots | \ddots | \vdots |
| | $y_{n_1,1}$ | $y_{n_1,2}$ | \cdots | $y_{n_1,d}$ |
| Pattern 2 | | $y_{n_1+1,2}$ | \cdots | $y_{n_1+1,d}$ |
| | | \vdots | \ddots | \vdots |
| | | $y_{n_1+n_2,2}$ | \cdots | $y_{n_1+n_2,d}$ |
| \vdots | | | \ddots | \vdots |
| Pattern d | | | | $y_{n-n_d+1,d}$ |
| | | | | \vdots |
| | | | | $y_{n,d}$ |

Fig. 1: The observed response under a monotone structure. For a realized response matrix $\mathbf{y} = (y_{i,j})_{i=1}^n_{j=1}^d \in \mathbb{R}^{n \times d}$, a realized missing structure $\mathbf{k} = (k_{i,j})_{i=1}^n_{j=1}^d \in \{0, 1\}^{n \times d}$ is monotone if the observed response, i.e., the elements in $\mathbf{y}_{(\mathbf{k})}$, can be arranged in the format above.

where, for $i \in \{1, \dots, n\}$,

$$r_{i,(k)} = (\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \boldsymbol{\beta}^\top \mathbf{x}_i)^\top (\mathbf{c}_{i,(k)} \boldsymbol{\Sigma} \mathbf{c}_{i,(k)}^\top)^{-1} (\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \boldsymbol{\beta}^\top \mathbf{x}_i). \quad (3)$$

Remark 1. Because the prior distribution is not proper, $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$ may not be a proper probability density. As a side product of our convergence analysis, we give sufficient conditions for posterior propriety in Section 4.

The posterior density $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$ is almost always intractable in the sense that it is hard to calculate its features such as expectation and quantiles. Liu [15] proposed a data augmentation (DA) algorithm, or two-component Gibbs sampler, that can be used to sample from this distribution. This is an MCMC algorithm that simulates a Markov chain $(\mathbf{B}(t), \boldsymbol{\Sigma}(t))_{t=0}^\infty$ that is reversible with respect to the posterior. In the next section, we describe the algorithm in detail.

3. The DA and DAI Algorithms

3.1. Missing structures

To adequately describe the DA and DAI algorithms, we need to introduce some concepts regarding missing structure matrices.

A realized missing structure $\mathbf{k} = (k_{i,j})_{i=1}^n_{j=1}^d \in \{0, 1\}^{n \times d}$ is said to be monotone if the following conditions hold:

- (i) If $k_{i,j} = 1$ for some $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, d\}$, then $k_{i',j'} = 1$ whenever $i' \leq i$ and $j' \geq j$.
- (ii) $k_{i,d} = 1$ for $i \in \{1, \dots, n\}$.

The monotone missing structure occurs, for example, in studies with attrition, where units drop out prior to the end of the study and do not return [14, Example 1.6]. In practice, there are cases where the observed missing structure can be re-arranged to become monotone by permuting the rows and columns of the response matrix. If there are no missing data, the corresponding missing structure is monotone.

Let $\mathbf{k} = (k_{i,j})_{i=1}^n_{j=1}^d$ be monotone. Then, for a realized response matrix $\mathbf{y} = (y_{i,j})_{i=1}^n_{j=1}^d \in \mathbb{R}^{n \times d}$, the elements in $\mathbf{y}_{(\mathbf{k})}$ can be arranged as in Fig. 1. We say that the i th observation belongs to pattern ℓ for some $\ell \in \{1, \dots, d\}$ if $k_{i,j} = 1$ for $j \geq \ell$ and $k_{i,j} = 0$ for $j < \ell$. There are d possible patterns. For $\ell \in \{1, \dots, d\}$, denote by $n_\ell(\mathbf{k})$ the number of observations that belong to pattern ℓ . Let $\mathbf{y}_{(\mathbf{k},\ell)}$ be the $[\sum_{j=1}^\ell n_j(\mathbf{k}) \times (d - \ell + 1)]$ matrix whose i th row is $(y_{i,\ell}, \dots, y_{i,d})$, and let $\mathbf{x}_{(\mathbf{k},\ell)}$ be the submatrix of \mathbf{x} formed by the first $\sum_{j=1}^\ell n_j(\mathbf{k})$ rows of \mathbf{x} . Denote by $\mathbf{y}_{(\mathbf{k},\ell)} : \mathbf{x}_{(\mathbf{k},\ell)}$ the matrix formed by attaching $\mathbf{x}_{(\mathbf{k},\ell)}$ to the right of $\mathbf{y}_{(\mathbf{k},\ell)}$. We say that Condition (4) holds for the pair $(\mathbf{k}, \mathbf{y}_{(\mathbf{k})})$ if

$$r(\mathbf{y}_{(\mathbf{k},\ell)} : \mathbf{x}_{(\mathbf{k},\ell)}) = p + d - \ell + 1, \quad \sum_{j=1}^\ell n_j(\mathbf{k}) > p + d - m + \ell - 1, \quad \ell \in \{1, \dots, d\}. \quad (4)$$

Here, $r(\cdot)$ returns the rank of a matrix. This condition is crucial for the DA algorithm to be well-defined and implementable.

3.2. The DA algorithm

Fix a realized response $\mathbf{y} \in \mathbb{R}^{n \times d}$ and a realized missing structure $\mathbf{k} \in \{0, 1\}^{n \times d}$. Given the current state $(\mathbf{B}(t), \mathbf{\Sigma}(t)) = (\boldsymbol{\beta}, \boldsymbol{\varsigma})$, the DA algorithm for sampling from $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$ draws the next state $(\mathbf{B}(t+1), \mathbf{\Sigma}(t+1))$ using the following steps.

1. *I step.* Draw $\mathbf{W}^* = (W_1^*, \dots, W_n^*)$ from the conditional distribution of \mathbf{W} given $(\mathbf{B}, \mathbf{\Sigma}, \mathbf{Y}_{(\mathbf{k})}) = (\boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(\mathbf{k})})$. Call the sampled value \mathbf{w} .
2. *P step.* Draw $(\mathbf{B}(t+1), \mathbf{\Sigma}(t+1))$ from the conditional distribution of $(\mathbf{B}, \mathbf{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(\mathbf{k})}) = (\mathbf{w}, \mathbf{y}_{(\mathbf{k})})$.

This algorithm simulates a Markov chain $(\mathbf{B}(t), \mathbf{\Sigma}(t))_{t=0}^{\infty}$ that is reversible with respect to $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$.

For $i \in \{1, \dots, n\}$, let d_i be the number of nonzero entries in the i th row of \mathbf{k} . (Note: if \mathbf{k} is monotone, and the i th observation belongs to some pattern ℓ , then $d_i = d - \ell + 1$.) Then the conditional distribution of $\mathbf{W} = (W_1, \dots, W_n)$ given $(\mathbf{B}, \mathbf{\Sigma}, \mathbf{Y}_{(\mathbf{k})}) = (\boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(\mathbf{k})})$ is

$$P_{\mathbf{k}}(d\mathbf{w} \mid \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(\mathbf{k})}) \propto \prod_{i=1}^n w_i^{d_i/2} \exp\left(-\frac{r_{i,(\mathbf{k})} w_i}{2}\right) P_{\text{mix}}(dw_i), \quad (5)$$

where $r_{i,(\mathbf{k})}$ is defined in (3) for $i \in \{1, \dots, n\}$. To ensure that this conditional distribution is always proper, we assume throughout that

$$\int_0^{\infty} w^{d/2} P_{\text{mix}}(dw) < \infty. \quad (6)$$

The conditional distribution $P_{\mathbf{k}}(\cdot \mid \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(\mathbf{k})})$ corresponds to n independent univariate random variables. For most commonly used mixing distributions $P_{\text{mix}}(\cdot)$, this is not difficult to sample. For instance, if $P_{\text{mix}}(\cdot)$ is a Gamma distribution, $P_{\mathbf{k}}(\cdot \mid \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(\mathbf{k})})$ is the product of n Gamma distributions.

The conditional distribution of $(\mathbf{B}, \mathbf{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(\mathbf{k})}) = (\mathbf{w}, \mathbf{y}_{(\mathbf{k})})$ is

$$p_{\mathbf{k}}(\boldsymbol{\beta}, \boldsymbol{\varsigma} \mid \mathbf{w}, \mathbf{y}_{(\mathbf{k})}) \propto p_{\text{prior}}(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \prod_{i=1}^n \frac{1}{|\mathbf{c}_{i,(\mathbf{k})} \boldsymbol{\varsigma} \mathbf{c}_{i,(\mathbf{k})}^{\top}|^{1/2}} \exp\left(-\frac{r_{i,(\mathbf{k})} w_i}{2}\right), \quad (7)$$

where, for $i \in \{1, \dots, n\}$, $r_{i,(\mathbf{k})}$ is defined in (3). Under general missing structures, the conditional distribution $p_{\mathbf{k}}(\cdot \mid \mathbf{w}, \mathbf{y}_{(\mathbf{k})})$ is not always tractable. In fact, since an improper prior is used, this conditional is possibly improper. Liu [15] provided a method for sampling from this distribution when \mathbf{k} is monotone. When Condition (4) holds for $(\mathbf{k}, \mathbf{y}_{(\mathbf{k})})$, the conditional of $(\mathbf{B}, \mathbf{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(\mathbf{k})}) = (\mathbf{w}, \mathbf{y}_{(\mathbf{k})})$ is proper for any $\mathbf{w} = (w_1, \dots, w_n) \in (0, \infty)^n$. The monotone missing structure allows for a factorization of this conditional distribution after appropriate transformations. Similar techniques are commonly used in missing data literature [14, Chapter 7]. Then the conditional distribution $p_{\mathbf{k}}(\cdot \mid \mathbf{w}, \mathbf{y}_{(\mathbf{k})})$ can be sampled using chi-square and normal distributions. The method is intricate, so we relegate the details to Appendix A.

Let $\mathbf{k}_0 \in \{0, 1\}^{n \times d}$ be the missing structure that corresponds to a completely observable response, i.e., all elements of \mathbf{k}_0 are 1. Denote by $\mathbf{Y}_{(\mathbf{k}_0 - \mathbf{k})}$ the missing parts in the response. Sometimes, we are interested in the posterior distribution of $\mathbf{Y}_{(\mathbf{k}_0 - \mathbf{k})}$ given $\mathbf{Y}_{(\mathbf{k})} = \mathbf{y}_{(\mathbf{k})}$, which takes the following form:

$$P_{\mathbf{Y}_{(\mathbf{k}_0 - \mathbf{k})}}(d\mathbf{z} \mid \mathbf{k}, \mathbf{y}_{(\mathbf{k})}) \propto d\mathbf{z} \int_{S_+^{d \times d}} \int_{\mathbb{R}^{p \times d}} \int_{(0, \infty)^n} p_{\text{joint}}(\mathbf{y}^{\mathbf{z}}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\varsigma}) P_{\text{mix}}(d\mathbf{w}) d\boldsymbol{\beta} d\boldsymbol{\varsigma},$$

where p_{joint} is given in (2), and $\mathbf{y}^{\mathbf{z}}$ is a realized value of \mathbf{Y} such that $\mathbf{y}_{(\mathbf{k})}^{\mathbf{z}} = \mathbf{y}_{(\mathbf{k})}$ and $\mathbf{y}_{(\mathbf{k}_0 - \mathbf{k})}^{\mathbf{z}} = \mathbf{z}$. To sample from $P_{\mathbf{Y}_{(\mathbf{k}_0 - \mathbf{k})}}(\cdot \mid \mathbf{k}, \mathbf{y}_{(\mathbf{k})})$, one can add a post hoc imputation step at the end of each DA iteration.

3. *Post hoc imputation step.* Draw $\mathbf{Z}(t+1)$ from the conditional distribution of $\mathbf{Y}_{(\mathbf{k}_0 - \mathbf{k})}$ given $(\mathbf{Y}_{(\mathbf{k})}, \mathbf{W}, \mathbf{B}, \mathbf{\Sigma}) = (\mathbf{y}_{(\mathbf{k})}, \mathbf{w}, \boldsymbol{\beta}^*, \boldsymbol{\varsigma}^*)$, where $(\boldsymbol{\beta}^*, \boldsymbol{\varsigma}^*)$ is the sampled value of $(\mathbf{B}(t+1), \mathbf{\Sigma}(t+1))$.

Recall that, given $(\mathbf{W}, \mathbf{B}, \mathbf{\Sigma}) = (\mathbf{w}, \boldsymbol{\beta}^*, \boldsymbol{\varsigma}^*)$, \mathbf{Y} has a multivariate normal distribution. Then the conditional distribution of $\mathbf{Y}_{(\mathbf{k}_0 - \mathbf{k})}$ given $(\mathbf{Y}_{(\mathbf{k})}, \mathbf{W}, \mathbf{B}, \mathbf{\Sigma}) = (\mathbf{y}_{(\mathbf{k})}, \mathbf{w}, \boldsymbol{\beta}^*, \boldsymbol{\varsigma}^*)$ is also (univariate or multivariate) normal, and $(\mathbf{B}(t), \mathbf{\Sigma}(t), \mathbf{Z}(t))_{t=0}^{\infty}$ is a Markov chain whose stationary distribution is the posterior distribution of $(\mathbf{B}, \mathbf{\Sigma}, \mathbf{Y}_{(\mathbf{k}_0 - \mathbf{k})})$ given $\mathbf{Y}_{(\mathbf{k})} = \mathbf{y}_{(\mathbf{k})}$.

The imputation step is post hoc because it can be implemented at the end of the whole simulation, as long as the value of \mathbf{w} is recorded in the I step of each iteration. Post hoc imputation is possible because the I and P steps do not rely on the value of $\mathbf{Z}(t)$. Naturally, post hoc imputation does not affect the convergence properties of the Markov chain. Indeed, a standard argument [22, Theorem 1] shows that $(\mathbf{B}(t), \mathbf{\Sigma}(t), \mathbf{Z}(t))_{t=0}^{\infty}$ is geometrically ergodic if and only if $(\mathbf{B}(t), \mathbf{\Sigma}(t))_{t=0}^{\infty}$ is geometrically ergodic. Moreover, the proof of Theorem 1 in Roberts and Rosenthal [22] implies that the L^2 convergence rate of the two chains, as defined in Section 4.1, are the same. Thus, when studying the convergence properties of the DA algorithm, we can restrict our attention to $(\mathbf{B}(t), \mathbf{\Sigma}(t))_{t=0}^{\infty}$ instead of $(\mathbf{B}(t), \mathbf{\Sigma}(t), \mathbf{Z}(t))_{t=0}^{\infty}$ even if there is post hoc imputation.

3.3. The DAI algorithm

For two realized missing structures $\mathbf{k} = (k_{i,j}) \in \{0, 1\}^{n \times d}$ and $\mathbf{k}' = (k'_{i,j}) \in \{0, 1\}^{n \times d}$, write $\mathbf{k} < \mathbf{k}'$ if (i) $\mathbf{k} \neq \mathbf{k}'$ and (ii) $k'_{i,j} = 1$ whenever $k_{i,j} = 1$. That is, $\mathbf{k} < \mathbf{k}'$ if the observed response entries under structure \mathbf{k} is a proper subset of those under \mathbf{k}' . If $\mathbf{k} < \mathbf{k}'$, then $\mathbf{k}' - \mathbf{k}$ is a missing structure matrix that gives the entries that are missing under \mathbf{k} , but not under \mathbf{k}' . In other words, when $\mathbf{k} < \mathbf{k}'$, $\mathbf{Y}_{(\mathbf{k}' - \mathbf{k})}$ is observed under \mathbf{k}' , but not under \mathbf{k} .

Fix a realized response $\mathbf{y} \in \mathbb{R}^{n \times d}$ and missing structure \mathbf{k} . One may imagine that \mathbf{k} is not monotone, and the DA algorithm cannot be implemented efficiently. The DAI algorithm [15] overcomes non-monotonicity by imputing the value of $\mathbf{Y}_{(\mathbf{k}' - \mathbf{k})}$ in the I step, where \mathbf{k}' is a monotone realized missing structure such that $\mathbf{k} < \mathbf{k}'$ chosen by the user. We now describe the details of this algorithm.

The DAI algorithm associated with the triplet $(\mathbf{y}, \mathbf{k}, \mathbf{k}')$ simulates a Markov chain $(\mathbf{B}(t), \mathbf{\Sigma}(t))_{t=0}^{\infty}$ that is reversible with respect to $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$. Given the current state $(\mathbf{B}(t), \mathbf{\Sigma}(t)) = (\boldsymbol{\beta}, \boldsymbol{\varsigma})$, the DAI algorithm draws the next state $(\mathbf{B}(t+1), \mathbf{\Sigma}(t+1))$ using the following steps.

1. *I1 step.* Draw $\mathbf{W}^* = (W_1^*, \dots, W_n^*)$ from the conditional distribution of \mathbf{W} given $(\mathbf{B}, \mathbf{\Sigma}, \mathbf{Y}_{(\mathbf{k})}) = (\boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(\mathbf{k})})$, whose exact form is given in (5). Call the sampled value \mathbf{w} .
2. *I2 step.* Draw $\mathbf{Z}(t+1)$ from the conditional distribution of $\mathbf{Y}_{(\mathbf{k}' - \mathbf{k})}$ given $(\mathbf{Y}_{(\mathbf{k})}, \mathbf{W}, \mathbf{B}, \mathbf{\Sigma}) = (\mathbf{y}_{(\mathbf{k})}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\varsigma})$. Call the sampled value \mathbf{z} .
3. *P step.* Draw $(\mathbf{B}(t+1), \mathbf{\Sigma}(t+1))$ from the conditional distribution of $(\mathbf{B}, \mathbf{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(\mathbf{k})}, \mathbf{Y}_{(\mathbf{k}' - \mathbf{k})}) = (\mathbf{w}, \mathbf{y}_{(\mathbf{k})}, \mathbf{z})$.

Recall that, given $(\mathbf{W}, \mathbf{B}, \mathbf{\Sigma}) = (\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\varsigma})$, \mathbf{Y} has a multivariate normal distribution. Then in the I2 step, the conditional distribution of $\mathbf{Y}_{(\mathbf{k}' - \mathbf{k})}$ given $(\mathbf{Y}_{(\mathbf{k})}, \mathbf{W}, \mathbf{B}, \mathbf{\Sigma}) = (\mathbf{y}_{(\mathbf{k})}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\varsigma})$ is (univariate or multivariate) normal. In the P step, one is in fact sampling from the conditional distribution of $(\mathbf{B}, \mathbf{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(\mathbf{k}')}) = (\mathbf{w}, \mathbf{y}_{(\mathbf{k}')}^*)$, where \mathbf{y}^* is a realized value of \mathbf{Y} such that $\mathbf{y}_{(\mathbf{k})}^* = \mathbf{y}_{(\mathbf{k})}$ and $\mathbf{y}_{(\mathbf{k}' - \mathbf{k})}^* = \mathbf{z}$. To ensure that this step can be implemented efficiently, we need two conditions: (i) \mathbf{k}' is monotone, and (ii) Condition (4) holds for $(\mathbf{k}', \mathbf{y}_{(\mathbf{k}')}^*)$. If these two conditions hold, then one can use methods in Appendix A to implement the P step. Of course, in each step of the DAI algorithm, $\mathbf{z} = \mathbf{y}_{(\mathbf{k}' - \mathbf{k})}^*$ changes. Despite this, due to the following result, which is easy to verify, we do not have to check (ii) in every step.

Proposition 1. *Let \mathbf{k}' be a monotone missing structure such that $\mathbf{k} < \mathbf{k}'$, and let $\mathbf{y}^* \in \mathbb{R}^{n \times d}$ be such that $\mathbf{y}_{(\mathbf{k})}^* = \mathbf{y}_{(\mathbf{k})}$. Suppose that there exists a monotone structure \mathbf{k}'' such that $\mathbf{k}'' < \mathbf{k}$, and that Condition (4) holds for $(\mathbf{k}'', \mathbf{y}_{(\mathbf{k}'')}^*)$. Then Condition (4) holds for $(\mathbf{k}', \mathbf{y}_{(\mathbf{k}')}^*)$ regardless of the value of $\mathbf{y}_{(\mathbf{k}' - \mathbf{k})}^*$.*

The DAI algorithm can be used to impute the missing values of \mathbf{Y} . Indeed, $(\mathbf{B}(t), \mathbf{\Sigma}(t), \mathbf{Z}(t))_{t=0}^{\infty}$ is a Markov chain whose stationary distribution is the posterior distribution of $(\mathbf{B}, \mathbf{\Sigma}, \mathbf{Y}_{(\mathbf{k}' - \mathbf{k})})$ given $\mathbf{Y}_{(\mathbf{k})} = \mathbf{y}_{(\mathbf{k})}$. This is similar to the DA algorithm with post hoc imputation described in Section 3.2, especially if all the entries of \mathbf{k}' are 1. Compared to the DA algorithm with post hoc imputation, the DAI algorithm can be implemented even when \mathbf{k} is not monotone. However, the intermediate imputation step I2 cannot be performed after the whole simulation process since the P step of DAI relies on the value of $\mathbf{Z}(t+1)$.

The Markov chain $(\mathbf{B}(t), \mathbf{\Sigma}(t), \mathbf{Z}(t))_{t=0}^{\infty}$ is geometrically ergodic if and only if $(\mathbf{B}(t), \mathbf{\Sigma}(t))_{t=0}^{\infty}$ is geometrically ergodic [22, Theorem 1]. The proof of Theorem 1 in Roberts and Rosenthal [22] shows that the two chains have the same L^2 convergence rate. Thus, when studying the convergence properties of the DAI algorithm, we can restrict our attention to $(\mathbf{B}(t), \mathbf{\Sigma}(t))_{t=0}^{\infty}$ instead of $(\mathbf{B}(t), \mathbf{\Sigma}(t), \mathbf{Z}(t))_{t=0}^{\infty}$ even if we care about imputing missing data.

4. Convergence Analysis

4.1. Preliminaries

We start by reviewing some general concepts regarding the convergence properties of Markov chains. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. Consider a Markov chain $(X(t))_{t=0}^{\infty}$ whose state space is $(\mathcal{X}, \mathcal{F})$, and let $K : \mathcal{X} \times \mathcal{F} \rightarrow [0, 1]$ be its transition kernel. For a signed measure μ and a measurable function f on $(\mathcal{X}, \mathcal{F})$, let

$$\mu f = \int_{\mathcal{X}} f(x) \mu(dx),$$

and K can act on μ and f as follows:

$$\mu K(A) = \int_{\mathcal{X}} K(x, A) \mu(dx), \quad A \in \mathcal{F}, \quad Kf(x) = \int_{\mathcal{X}} f(y) K(x, dy), \quad x \in \mathcal{X},$$

assuming that the integrals are well-defined. Suppose that the chain has a stationary distribution π , i.e., $\pi K(\cdot) = \pi(\cdot)$. The chain is said to be reversible if, for $A_1, A_2 \in \mathcal{F}$,

$$\int_{A_1 \times A_2} \pi(dx) K(x, dy) = \int_{A_2 \times A_1} \pi(dx) K(x, dy).$$

For two probability measures μ_1 and μ_2 on $(\mathcal{X}, \mathcal{F})$, denote the total variation (TV) distance between the two probability measures by $\|\mu_1(\cdot) - \mu_2(\cdot)\|_{TV}$. For $t \in \mathbb{Z}_+$, let $K^t : \mathcal{X} \times \mathcal{F} \rightarrow [0, 1]$ be the t -step transition kernel of the chain, so that $K^1 = K$, and $\mu K^{t+1}(\cdot) = (\mu K^t)K(\cdot)$ for any probability measure μ on $(\mathcal{X}, \mathcal{F})$. A ϕ -irreducible aperiodic Markov chain with stationary distribution π is Harris ergodic if and only if $\lim_{t \rightarrow \infty} \|K^t(x, \cdot) - \pi(\cdot)\|_{TV} = 0$, for all $x \in \mathcal{X}$ [17, 24]. The chain is said to be geometrically ergodic if

$$\|K^t(x, \cdot) - \pi(\cdot)\|_{TV} \leq M(x)\rho^t, \quad x \in \mathcal{X}, \quad t \in \mathbb{Z}^+, \quad (8)$$

for some $\rho \in [0, 1)$ and $M : \mathcal{X} \rightarrow [0, \infty)$. As mentioned in the Introduction, Harris ergodicity guarantees a law of large numbers [17, Theorem 17.1.7], and geometric ergodicity guarantees a central limit theorem [5, 10, 11].

Another commonly used distance between probability measures is the L^2 distance. Let $L^2(\pi)$ be the set of measurable functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that

$$\pi f^2 := \int_{\mathcal{X}} f^2(x) \pi(dx) < \infty.$$

The L^2 distance between two probability measures μ_1 and μ_2 on $(\mathcal{X}, \mathcal{F})$ is

$$\|\mu_1(\cdot) - \mu_2(\cdot)\|_2 = \sup_{f \in L^2(\pi)} |\mu_1 f - \mu_2 f|.$$

Denote by $L_*^2(\pi)$ the set of probability measures μ on $(\mathcal{X}, \mathcal{F})$ such that μ is absolutely continuous to π and that $d\mu/d\pi$, the density of μ with respect to π , is in $L^2(\pi)$. We say the chain is L^2 geometrically ergodic if

$$\|\mu K^t(\cdot) - \pi(\cdot)\|_2 \leq C(\mu)\rho^t, \quad \mu \in L_*^2(\pi), \quad t \in \mathbb{Z}^+, \quad (9)$$

for some $\rho \in [0, 1)$ and $C : L_*^2(\pi) \rightarrow [0, \infty)$. For a ϕ -irreducible reversible Markov chain, geometric ergodicity implies L^2 geometric ergodicity [21, Theorem 2.1 and Remark 2.3]. See also [25] Theorem 2 and [20] Section 3.3.

The infimum of $\rho \in [0, 1]$ such that (9) holds for some $C : L_*^2(\pi) \rightarrow [0, \infty)$ is called the L^2 convergence rate of the chain. The smaller this rate is, the faster the convergence.

4.2. Geometric ergodicity of the DA algorithm

Define three classes of mixing distributions based on their behaviors near the origin. These classes were first examined in an analysis of the DA algorithm when the response matrix is fully observed [9]. We say that the mixing distribution $P_{\text{mix}}(\cdot)$ is zero near the origin if there exists $\theta > 0$ such that $\int_0^\theta P_{\text{mix}}(dw) = 0$. Assume now that $P_{\text{mix}}(\cdot)$ admits a density function $p_{\text{mix}}(\cdot) : (0, \infty) \rightarrow [0, \infty)$ with respect to the Lebesgue measure. If there exists $c > -1$, such that $\lim_{w \rightarrow 0} p_{\text{mix}}(w)/w^c < \infty$, we say that $P_{\text{mix}}(\cdot)$ is polynomial near the origin with power c . The mixing distribution $P_{\text{mix}}(\cdot)$ is faster than polynomial near the origin if, for all $c > 0$, there exists $\kappa_c > 0$ such that the ratio $p_{\text{mix}}(w)/w^c$ is strictly increasing on $(0, \kappa_c)$.

Most commonly used mixing distributions fall into one of these three classes. Examples will be given after we state the main result of this section.

Again, fix an observed missing structure \mathbf{k} and observed response $\mathbf{y}_{(k)}$. Let d_i be the number of nonzero elements in the i th row of \mathbf{k} . Recall that n is the number of observations, p is the number of predictors, d is the dimension of the responses, and m is a parameter in the prior distribution (1).

Theorem 1. *Consider the DA algorithm targeting $\pi_{\mathbf{k}}(\cdot | \mathbf{y}_{(k)})$, as described in Section 3.2. Suppose that (6) holds, and that the conditional distribution of $(\mathbf{B}, \mathbf{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$ is proper for every $\mathbf{w} = (w_1, \dots, w_n)^\top \in (0, \infty)^n$. If any one of the following conditions holds, then the posterior $\pi_{\mathbf{k}}(\cdot | \mathbf{y}_{(k)})$ is proper and the underlying Markov chain is geometrically ergodic.*

- (i) $P_{\text{mix}}(\cdot)$ is zero near the origin;
- (ii) $P_{\text{mix}}(\cdot)$ is faster than polynomial near the origin;
- (iii) $P_{\text{mix}}(\cdot)$ is polynomial near the origin with power $c > c_1$, where

$$c_1 = \frac{n - p + m - \min\{d_1, \dots, d_n\}}{2}.$$

Remark 2. Recall that when \mathbf{k} is monotone, the conditional distribution of $(\mathbf{B}, \mathbf{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$ is proper for every $\mathbf{w} = (w_1, \dots, w_n)^\top \in (0, \infty)^n$ if Condition (4) holds for $(\mathbf{k}, \mathbf{y}_{(k)})$.

Remark 3. When $P_{\text{mix}}(\cdot)$ has a density function with respect to the Lebesgue measure and \mathbf{Y} is fully observed, Theorem 1 reduces to an existing result [9, Theorem 1].

The proof of Theorem 1 is in Appendix B. In what follows, we list commonly used mixing distributions that fall into the three categories in Theorem 1. We also check whether each mixing distribution satisfies (6).

When the mixing distribution $P_{\text{mix}}(\cdot)$ is discrete with finite support, the mixing distribution is zero near the origin. This is the case when errors follow finite mixtures of Gaussian. Obviously, (6) holds in this case.

A Pareto(a, b) distribution has density $p(w | a, b) \propto w^{-b-1}$, $w \in [a, \infty)$, where $a > 0, b > 0$. It is zero near the origin as the support is $[a, \infty)$. Condition (6) holds if $b > d/2$.

A generalized inverse Gaussian distribution GIG(a, b, q) with density

$$p(w | a, b, q) \propto w^{q-1} \exp\left(-\frac{aw + b/w}{2}\right),$$

where $a > 0, b > 0, q \in \mathbb{R}$, is faster than polynomial near the origin. Condition (6) holds for any GIG distribution. When the mixing distribution is GIG, the distribution of the error is called Generalized Hyperbolic [1].

The density of an inverse gamma distribution IG(a, b) is $p(w | a, b) \propto w^{-a-1} \exp(-b/w)$, where $a > 0, b > 0$. IG(a, b) is faster than polynomial near the origin. Condition (6) is satisfied if $a > d/2$.

For $\mu \in \mathbb{R}$ and $v > 0$, a Log-normal(μ, v) distribution has density

$$p(w | \mu, v) \propto \frac{1}{w} \exp\left\{-\frac{(\ln w - \mu)^2}{2v^2}\right\}.$$

This distribution is faster than polynomial near the origin and (6) holds.

A Fréchet distribution with the shape $\alpha > 0$ and scale $s > 0$ is given by

$$p(w | \alpha, s) \propto w^{-(1+\alpha)} \exp\{-(s/w)^\alpha\}.$$

It is faster than polynomial near the origin. Moreover, (6) holds whenever $\alpha > d/2$.

A Gamma(a, b) distribution has density $p(w | a, b) \propto w^{a-1} \exp(-bw)$, where $a > 0$ and $b > 0$. Gamma(a, b) is polynomial near the origin with power $c = a - 1$. The power $c > c_1$ if

$$a > \frac{n - p + m - \min\{d_1, \dots, d_n\}}{2} + 1,$$

where c_1 is given in Theorem 1. Condition (6) always holds for Gamma distributions. In particular, when $a = b = v/2$, the error has multivariate t distribution with degrees of freedom v , and $c > c_1$ if

$$v > n - p + m - \min\{d_1, \dots, d_n\} + 2.$$

A Beta(a, b) distribution has density $p(w | a, b) \propto w^{a-1} (1-w)^{b-1}$, $w \in (0, 1)$, where $a > 0$ and $b > 0$. When $b = 1$, the error is called multivariate slash distribution [13, 26]. Beta(a, b) is polynomial near the origin with power $c = a - 1$. Condition (6) always holds for Beta distributions.

A Weibull(a, b) distribution has density $p(w | a, b) \propto w^{a-1} \exp\{-(w/b)^a\}$, where $a > 0$ and $b > 0$. Weibull(a, b) is polynomial near the origin with power $c = a - 1$. Condition (6) always holds for Weibull distributions.

An $F(a, b)$ distribution has density $p(w | a, b) \propto w^{a/2-1} (aw + b)^{-(a+b)/2}$, where $a > 0$ and $b > 0$. $F(a, b)$ is polynomial near the origin with power $c = a/2 - 1$. The power $c > c_1$ if

$$a > n - p + m - \min\{d_1, \dots, d_n\} + 2.$$

Condition (6) is satisfied if $b > d$.

4.3. Harris ergodicity of the DAI algorithm

Now we give sufficient conditions for the DAI algorithm to be Harris ergodic.

Theorem 2. Assume that (6) holds. Let \mathbf{k} be a missing structure. Suppose that there is a missing structure \mathbf{k}' and a realized value of $\mathbf{Y}_{(\mathbf{k}')}$ denoted by \mathbf{z} such that $\mathbf{k}' < \mathbf{k}$, and that $\pi_{\mathbf{k}'}(\cdot | \mathbf{z})$, the conditional density of $(\mathbf{B}, \mathbf{\Sigma})$ given $\mathbf{Y}_{(\mathbf{k}')} = \mathbf{z}$, is proper. Then, for Lebesgue almost every \mathbf{z}' in the range of $\mathbf{Y}_{(\mathbf{k}-\mathbf{k}')}$, the posterior density $\pi_{\mathbf{k}}(\cdot | \mathbf{y}_{(\mathbf{k})})$, where \mathbf{y} satisfies $\mathbf{y}_{(\mathbf{k}')} = \mathbf{z}$ and $\mathbf{y}_{(\mathbf{k}-\mathbf{k}')} = \mathbf{z}'$, is proper, and any DAI chain targeting this posterior is Harris ergodic.

Proof. The posterior density $\pi_{\mathbf{k}}(\cdot | \mathbf{y}_{(\mathbf{k})})$ can be regarded as the posterior density of $(\mathbf{\beta}, \mathbf{\Sigma})$ given $\mathbf{Y}_{(\mathbf{k}-\mathbf{k}')} = \mathbf{z}'$ when the prior density is $\pi_{\mathbf{k}'}(\cdot | \mathbf{z})$. Since $\pi_{\mathbf{k}'}(\cdot | \mathbf{z})$ is assumed to be proper, this posterior is proper for almost every possible value of \mathbf{z}' .

Under (6) and posterior propriety, a DAI chain targeting the posterior is Harris ergodic [24, Theorem 6 and Corollary 13]. \square

By Theorem 1, a posterior density $\pi_{\mathbf{k}'}(\cdot | \mathbf{y}_{(\mathbf{k}')})$ is proper if each of the following conditions holds:

- \mathbf{k}' is monotone, and Condition (4) holds for $(\mathbf{k}', \mathbf{y}_{(\mathbf{k}')})$;
- $P_{\text{mix}}(\cdot)$ satisfies (6);
- $P_{\text{mix}}(\cdot)$ is either zero near the origin, or faster than polynomial near the origin, or polynomial near the origin with power $c > c_1$, where

$$c_1 = \frac{n - p + m - \min\{d'_1, \dots, d'_n\}}{2},$$

and for $i \in \{1, \dots, n\}$, d'_i is the number of nonzero entries in the i th row of \mathbf{k}' .

By Theorem 2, under (6), to ensure Harris ergodicity of a DAI algorithm targeting $\pi_{\mathbf{k}}(\cdot | \mathbf{y}_{(\mathbf{k})})$ in an almost sure sense, one only needs to find a missing structure $\mathbf{k}' < \mathbf{k}$ that satisfies the conditions above. In theory, when such a \mathbf{k}' exists, it is still possible that the posterior $\pi_{\mathbf{k}}(\cdot | \mathbf{y}_{(\mathbf{k})})$ is improper – even though it only happens on a zero measure set [3].

4.4. Comparison between the DA and DAI algorithms

Again, fix a missing structure \mathbf{k} and a realized response \mathbf{y} , and let d_i be the number of nonzero elements in the i th row of \mathbf{k} . Assume that there is at least one missing entry, i.e., $d_i < d$ for some $i \in \{1, \dots, n\}$. Assume also that $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$ is proper. In principle, one can either use the DA algorithm (with or without post hoc imputation) or the DAI algorithm associated with $(\mathbf{y}, \mathbf{k}, \mathbf{k}')$ where $\mathbf{k} < \mathbf{k}'$ to sample from the posterior. In this subsection, we compare the two algorithms in terms of their L^2 convergence rates. This comparison is important when \mathbf{k} is monotone, and Conditions (4) and (6) hold. In this case, both algorithms can be efficiently implemented.

Let S and T be two random elements. Denote by π the marginal distribution of S . A generic data augmentation algorithm for sampling from π simulates a Markov chain $(S(t))_{t=0}^{\infty}$ that is reversible to π . Given the current state $S(t)$, the next state $S(t+1)$ is generated through the following procedure.

1. *I step.* Draw T^* from the conditional distribution of T given $S = S(t)$. Call the observed value t^* .
2. *P step.* Draw $S(t+1)$ from the conditional distribution of S given $T = t^*$.

The DA and DAI algorithms for Bayesian robust linear regression are special cases of the above method. Indeed, let S_1 , T_1 , and T_2 be three random elements such that the joint distribution of S_1 , T_1 , and T_2 is the conditional joint distribution of (\mathbf{B}, Σ) , \mathbf{W} , and $\mathbf{Y}_{(\mathbf{k}'-\mathbf{k})}$ given $\mathbf{Y}_{(\mathbf{k})} = \mathbf{y}_{(\mathbf{k})}$. Then, taking $S = S_1$ and $T = T_1$ in the generic algorithm yields the DA algorithm; taking $S = S_1$ and $T = (T_1, T_2)$ yields the DAI algorithm associated with $(\mathbf{y}, \mathbf{k}, \mathbf{k}')$.

The L^2 convergence rate of the generic data augmentation chain is precisely the squared maximal correlation between S and T [16]. The maximal correlation between S and T is

$$\gamma(S, T) := \sup \text{corr}[f(S), g(T)],$$

where corr means linear correlation, and the supremum is taken over real functions f and g such that the variances of $f(S)$ and $g(T)$ are finite. Evidently,

$$\gamma(S_1, T_1) \leq \gamma(S_1, (T_1, T_2)).$$

We then have the following result.

Theorem 3. *Suppose that $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$ is proper, and that the conditional distributions in the DA and DAI algorithms are well-defined. Then, the L^2 convergence rate of the DA chain targeting $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$ is at least as small as that of any DAI chain targeting $\pi_{\mathbf{k}}(\cdot \mid \mathbf{y}_{(\mathbf{k})})$.*

Recall that a smaller convergence rate means faster convergence. Thus, when computation time is not considered and the observed missing structure is monotone, the DA algorithm is faster than the DAI algorithm. In this case, imputation of missing data, if needed, should be performed in a post hoc rather than intermediate manner. In Section 5 we use numerical experiments to show that this appears to be the case even after computation cost is taken into account.

5. Numerical Experiment

We compare the performance of the DA and DAI algorithms using simulated data. All simulations are implemented through the `Bayesianrobust` R package.

Suppose that we have $n = 50$ observations in a study. Assume that for each observation, the response has $d = 2$ components, while the predictor has the form $\mathbf{x}_i = (1, x_i)^\top$ where $x_i \in \mathbb{R}$. We generate the x_i s using independent normal distributions. The response matrix \mathbf{y} is generated according to the robust linear regression model, with the mixing distribution being $\text{Gamma}(1, 1)$. The simulated data set is fixed throughout the section.

On the modeling side, we consider an independence Jeffreys prior with $m = d = 2$ and $\mathbf{a} = \mathbf{0}$ (see (1)) and three mixing distributions:

- G: The mixing distribution is $\text{Gamma}(2, 2)$. The error is t distribution with degrees of freedom 4. By (5), in the I step of the DA algorithm, one draws n independent Gamma random variables.
- GIG: The mixing distribution is $\text{GIG}(1, 1, -0.5)$. The error is generalized hyperbolic. By (5), in the I step of the DA algorithm, one draws n independent generalized inverse Gaussian random variables.

- P: The mixing distribution is the point mass at 1. The error is multivariate normal distribution. In this case, the posterior can be exactly sampled by the DA algorithm.

We study three realized missing structures. Under these missing structures, the response matrix \mathbf{y} has, respectively, 45, 40, and 35 rows fully observed. The other rows all have only the second entry observed. It is clear that all three missing structures are monotone.

In total, we consider nine combinations of mixing distributions and missing structures. We apply both the DA and DAI algorithms in each scenario. In the I2 step of the DAI algorithm (see Section 3.3), \mathbf{k}' is taken to be the $n \times d$ matrix whose entries are all 1, i.e., \mathbf{k}' corresponds to the case that the response matrix is fully observed. At the end of each iteration of the DA algorithm, a post hoc imputation step is performed, so both algorithms impute all missing response entries.

Consider estimating the regression coefficients \mathbf{B} and scatter matrix $\mathbf{\Sigma}$ using the posterior mean computed via the DA or DAI algorithm. We compare the efficiency of the two algorithms based on the effective sample size (ESS) of each component separately and all components jointly. The ESS is estimated by the `mcmcse` R package [6]. At a given simulation length N , the univariate ESS of an MCMC estimator is defined to be N times the posterior variance divided by the asymptotic variance of the estimator [7]. For a multivariate Markov chain on $\mathbb{R}^{\tilde{p}}$, where \tilde{p} is the dimension of the Markov chain, let $\mathbf{\Lambda}$ be the posterior covariance matrix and $\mathbf{\Phi}$ be the asymptotic covariance matrix of the estimator. The multivariate ESS of an MCMC estimator is defined to be $N(|\mathbf{\Lambda}|/|\mathbf{\Phi}|)^{1/\tilde{p}}$ [32]. To account for computation cost, we also consider the ESS per minute, $\text{ESSpm} = \text{ESS}/t_N$, where t_N is the number of minutes needed to run N iterations of the algorithm. The simulation length is set to be $N = 30,000$ without burn-in, and the initial values are the ordinary least squares estimates using the observations that belong to pattern 1, i.e., the observations without missing elements.

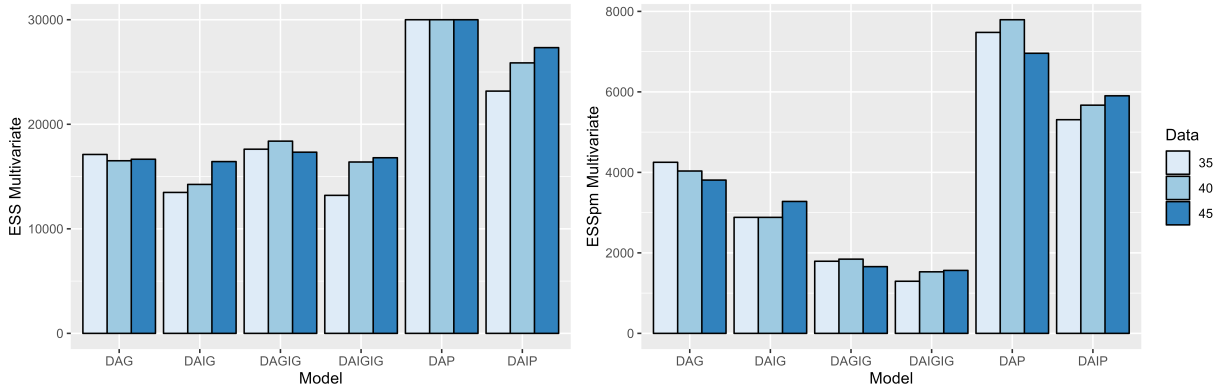


Fig. 2: The ESS and ESS per minute (ESSpm) of all components jointly in different scenarios. The x axis lists the mixing distributions and samplers used, and the legend labels the missing structures by the number of fully observed rows.

Fig. 2 gives the ESS and ESSpm of all components jointly for each of the $3 \times 3 \times 2$ combinations of mixing distributions, missing structures, and MCMC algorithms. The ESS and ESSpm of each component separately are included in Table C.1 and C.2 in Appendix C. We see that the DA algorithm gives a larger ESS compared to the DAI algorithm. For the DAI algorithm, the ESS is lower when there are more missing data. Similar trends appear when we consider the ESSpm. In short, imputation of missing responses values slows an algorithm down. This is consistent with our theoretical results. (Strictly speaking, our theoretical results concern convergence rate, not effective sample size, but for data augmentation chains which are reversible and positive, these two concepts are closely related. See, e.g., [28], Section 3.)

In addition to the experiment above, we consider another series of mixing distributions, namely Gamma(ν, ν) with $\nu = 2, 8, 25$. The resultant error distributions are multivariate t distributions with 2ν degrees of freedom. Applying the DA and DAI algorithms to these models, we obtain Fig. 3, the ESS and ESSpm of all components jointly. The ESS and ESSpm of each component separately are in Table C.3 and C.4 in Appendix C. Our simulation shows that when the model assumes that the error distribution has a lighter tail, the MCMC algorithms tend to be more efficient.

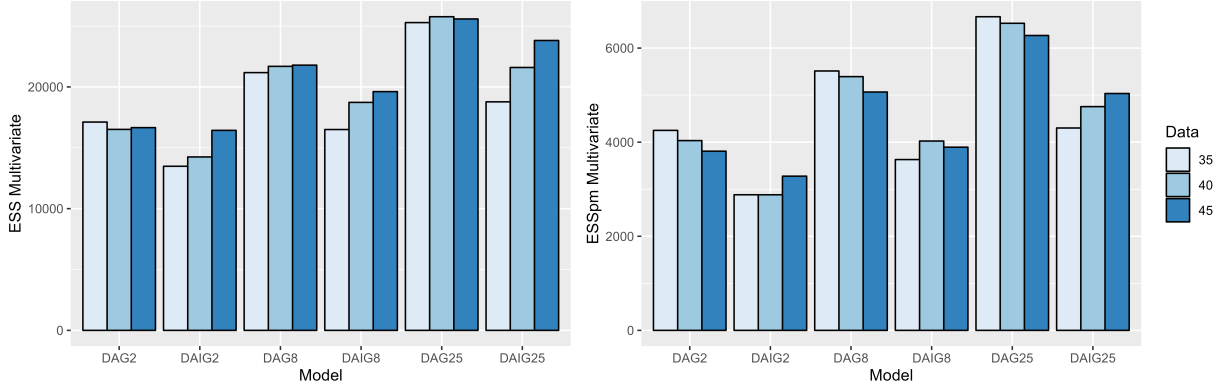


Fig. 3: The ESS and ESS per minute (ESSpm) of all components jointly in different scenarios. The x axis lists the Gamma(v, v) mixing distributions and samplers used, and the legend labels the missing structures by the number of fully observed rows.

6. Discussion

We conduct convergence analysis for data augmentation algorithms used in Bayesian robust multivariate linear regression with incomplete data. The algorithms were first proposed by Liu [15]. But, previously, little was known about its theoretical properties when the response matrix contains missing values. We consider two versions of the algorithm, DA and DAI. The DA algorithm can only be applied when the missing structure is monotone, whereas the DAI algorithm can be implemented for an arbitrary missing structure. We establish geometric ergodicity of the DA algorithm under simple conditions. For the DAI algorithm, we give conditions for Harris ergodicity. We compare the L^2 convergence rates of the DA and DAI algorithms. The L^2 convergence rate of the DA algorithm is at least as small as a corresponding DAI algorithm. A numerical study is provided. Under monotone missing structures, the DA algorithm outperforms the DAI computationally and theoretically.

When the responses are completely observed, Qin and Hobert [19] in 2018 provided conditions for the DA Markov chains to be trace-class, i.e., compact with summable eigenvalues, which is a property stronger than geometric ergodicity. It is reasonable to speculate that the trace-class property may persist even in scenarios where the data are incomplete. However, establishing this, if feasible at all, demands a nuanced and intricate analysis. This is an intriguing topic for future studies.

Acknowledgments

The authors thank the Editor, Associate Editor and referees for their helpful suggestions. The first and second authors were supported by the National Science Foundation [grant number DMS-2112887], and the third author was supported by the National Science Foundation [grant number DMS-2152746].

Appendix A. Details of the DA algorithm

The DA algorithm consists of two steps, the I step and the P step. The I step is shown in Section 3.2. In this section, we describe the P step when the missing structure \mathbf{k} is monotone and Condition (4) holds.

Recall from Section 3.1 that the monotone missing structure \mathbf{k} has d possible patterns. For $\ell \in \{1, \dots, d\}$, $n_{\ell(k)}$ is the number of observations that belong to pattern ℓ . $\mathbf{x}_{(k,\ell)}$ and $\mathbf{y}_{(k,\ell)}$ are submatrices of \mathbf{x} and \mathbf{y} respectively defined in Section 3.1. Define an $n \times n$ diagonal matrix $\lambda_\ell = \text{diag}(w_1, \dots, w_{\sum_{j=1}^{\ell} n_{j(k)}}, 0, \dots, 0)$. That is, the first $\sum_{j=1}^{\ell} n_{j(k)}$ diagonal entries of λ_ℓ are $w_1, \dots, w_{\sum_{j=1}^{\ell} n_{j(k)}}$, listed in order, and all the other entries are 0. Similarly, let $\lambda'_\ell = \text{diag}(w_1, \dots, w_{\sum_{j=1}^{\ell} n_{j(k)}})$. Denote $\hat{\beta}_\ell$ as the weighted least squares estimates of the regression coefficient of $\mathbf{y}_{(k,\ell)}$

on $\mathbf{x}_{(k,\ell)}$ with weight $(w_1, \dots, w_{\sum_{j=1}^{\ell} n_{j(k)}})$, and s_ℓ as the corresponding weighted total sum of residual squares and cross-products matrix, i.e.,

$$\hat{\boldsymbol{\beta}}_\ell = (\mathbf{x}^\top \boldsymbol{\lambda}_\ell \mathbf{x})^{-1} \mathbf{x}_{(k,\ell)}^\top \boldsymbol{\lambda}'_\ell \mathbf{y}_{(k,\ell)}, \quad s_\ell = (\mathbf{y}_{(k,\ell)} - \mathbf{x}_{(k,\ell)} \hat{\boldsymbol{\beta}}_\ell)^\top \boldsymbol{\lambda}'_\ell (\mathbf{y}_{(k,\ell)} - \mathbf{x}_{(k,\ell)} \hat{\boldsymbol{\beta}}_\ell).$$

Let \mathbf{a}_ℓ be the lower right $(d - \ell + 1) \times (d - \ell + 1)$ submatrix of the positive semi-definite matrix \mathbf{a} given in the prior (1).

P step. Draw $(\mathbf{B}(t+1), \boldsymbol{\Sigma}(t+1))$ from the conditional distribution of $(\mathbf{B}, \boldsymbol{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$ using the following procedure.

1. Draw $\boldsymbol{\Sigma}(t+1)$ given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$:

For $\ell \in \{1, \dots, d\}$, let $\mathbf{c}_\ell = \mathbf{a}_\ell + s_\ell$, and let \mathbf{e}_ℓ be the lower triangular Cholesky factor of \mathbf{c}_ℓ^{-1} (so that $\mathbf{c}_\ell^{-1} = \mathbf{e}_\ell \mathbf{e}_\ell^\top$).

Draw a sequence of random vectors $\mathbf{F}_1, \dots, \mathbf{F}_d$ such that $\mathbf{F}_\ell = (F_{\ell,\ell}, \dots, F_{d,\ell})^\top$ is $(d - \ell + 1)$ dimensional, and that

- (a) $F_{ij} \sim N(0, 1)$ for $1 \leq j < i \leq d$,
- (b) $F_{ii}^2 \sim \chi^2(df_i)$ for $i \in \{1, \dots, d\}$, where $df_i = (\sum_{j=1}^i n_j) - i + m - p - d + 1$,
- (c) F_{ij} are independent for all $1 \leq j \leq i \leq d$.

Denote the sampled values by f_1, \dots, f_d .

Let $\mathbf{h}_\ell = \mathbf{e}_\ell f_\ell$ for $\ell \in \{1, \dots, d\}$, and let \mathbf{h} be a $d \times d$ lower triangular matrix with its lower triangular non-zero part formed by columns $\mathbf{h}_1, \dots, \mathbf{h}_d$. Then $\boldsymbol{\varsigma} = (\mathbf{h} \mathbf{h}^\top)^{-1}$ serves as a sampled value of $\boldsymbol{\Sigma}(t+1)$.

2. Draw $\mathbf{B}(t+1)$ given $(\boldsymbol{\Sigma}(t+1), \mathbf{W}, \mathbf{Y}_{(k)}) = (\boldsymbol{\varsigma}, \mathbf{w}, \mathbf{y}_{(k)})$:

Independently, draw p -dimensional standard normal random vectors $\mathbf{Z}_1, \dots, \mathbf{Z}_d$, and call the sampled values z_1, \dots, z_d . For $\ell \in \{1, \dots, d\}$, let \mathbf{u}_ℓ be the lower triangular Cholesky factor of $(\mathbf{x}^\top \boldsymbol{\lambda}_\ell \mathbf{x})^{-1}$. Then

$$(\mathbf{u}_1 z_1, \dots, \mathbf{u}_d z_d) \mathbf{h}^{-1} + (\hat{\boldsymbol{\beta}}_1 \mathbf{h}_1, \dots, \hat{\boldsymbol{\beta}}_d \mathbf{h}_d) \mathbf{h}^{-1}$$

serves as a sampled value of $\mathbf{B}(t+1)$.

Let us consider a special case. Recall that $\mathbf{k}_0 \in \{0, 1\}^{n \times d}$ is the missing structure that corresponds to a completely observable response, i.e., all elements of \mathbf{k}_0 are 1. Then $\mathbf{Y}_{(\mathbf{k}_0)} = \mathbf{Y}$, $\mathbf{y}_{(\mathbf{k}_0)} = \mathbf{y}$, and \mathbf{k}_0 is monotone. Let $\mathbf{y} : \mathbf{x}$ be the matrix obtained by attaching \mathbf{x} to the right of \mathbf{y} . Then Condition (4) for $(\mathbf{k}_0, \mathbf{y}_{(\mathbf{k}_0)})$ is equivalent to

$$r(\mathbf{y} : \mathbf{x}) = p + d, \quad n > p + 2d - m - 1. \quad (\text{A.1})$$

Note that if there is some monotone \mathbf{k} such that Condition (4) holds for $(\mathbf{k}, \mathbf{y}_{(\mathbf{k})})$, then (A.1) necessarily holds. Under (A.1), the conditional distribution of $(\mathbf{B}, \boldsymbol{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}) = (\mathbf{w}, \mathbf{y})$ is proper and rather simple. We say $\mathbf{Z} \sim N_{p,d}(\boldsymbol{\mu}, \mathbf{u}, \mathbf{v})$ for $\boldsymbol{\mu} \in \mathbb{R}^{p \times d}$, $\mathbf{u} \in S_+^{p \times p}$, and $\mathbf{v} \in S_+^{d \times d}$ if \mathbf{Z} is a $p \times d$ random matrix associated with the probability density function

$$\frac{\exp\left[-\frac{1}{2} \text{tr}\left\{\mathbf{v}^{-1}(\mathbf{z} - \boldsymbol{\mu})^T \mathbf{u}^{-1}(\mathbf{z} - \boldsymbol{\mu})\right\}\right]}{(2\pi)^{pd/2} |\mathbf{v}|^{p/2} |\mathbf{u}|^{d/2}}, \quad \mathbf{z} \in \mathbb{R}^{p \times d}.$$

$N_{p,d}$ is called a matrix normal distribution. We say $\mathbf{Z} \sim \text{IW}_d(\mathbf{v}, \boldsymbol{\psi})$ for $\mathbf{v} > d - 1$ and $\boldsymbol{\psi} \in S_+^{d \times d}$ if \mathbf{Z} is a $d \times d$ random matrix associated with the probability density function

$$\frac{|\boldsymbol{\psi}|^{v/2}}{2^{vd/2} \Gamma_d(v/2)} |\mathbf{z}|^{-(v+d+1)/2} \exp\left\{-\frac{1}{2} \text{tr}(\boldsymbol{\psi} \mathbf{z}^{-1})\right\}, \quad \mathbf{z} \in S_+^{d \times d},$$

where $\text{tr}(\cdot)$ returns the trace of a matrix, and $\Gamma_d(\cdot)$ is a multivariate gamma function. IW_d is called an inverse Wishart distribution, and it is well-known that a random matrix follows the $\text{IW}_d(\mathbf{v}, \boldsymbol{\psi})$ distribution if and only if its inverse follows the Wishart distribution $\text{W}_d(\mathbf{v}, \boldsymbol{\psi}^{-1})$, which has density

$$\frac{|\boldsymbol{\psi}|^{v/2}}{2^{vd/2} \Gamma_d(v/2)} |\mathbf{z}|^{(v-d-1)/2} \exp\left\{-\frac{1}{2} \text{tr}(\boldsymbol{\psi} \mathbf{z})\right\}, \quad \mathbf{z} \in S_+^{d \times d}.$$

For $\mathbf{w} = (w_1, \dots, w_n) \in (0, \infty)^n$, let $\boldsymbol{\lambda} = \text{diag}(w_1, \dots, w_n)$, i.e., the diagonal matrix whose diagonal elements are w_1, \dots, w_n , listed in order. Let

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top \boldsymbol{\lambda} \mathbf{x})^{-1} \mathbf{x}^\top \boldsymbol{\lambda} \mathbf{y}, \quad \mathbf{s} = (\mathbf{y} - \mathbf{x} \hat{\boldsymbol{\beta}})^\top \boldsymbol{\lambda} (\mathbf{y} - \mathbf{x} \hat{\boldsymbol{\beta}}),$$

and $\boldsymbol{\xi} = (\mathbf{x}^\top \boldsymbol{\lambda} \mathbf{x})^{-1}$. Then it is easy to see from (2) that

$$\boldsymbol{\Sigma} \mid (\mathbf{W}, \mathbf{Y}) = (\mathbf{w}, \mathbf{y}) \sim \text{IW}_d(n - p + m - d, \mathbf{s} + \mathbf{a}), \quad \mathbf{B} \mid (\boldsymbol{\Sigma}, \mathbf{W}, \mathbf{Y}) = (\boldsymbol{\varsigma}, \mathbf{w}, \mathbf{y}) \sim N_{p,d}(\hat{\boldsymbol{\beta}}, \boldsymbol{\xi}, \boldsymbol{\varsigma}).$$

Appendix B. Proof of Theorem 1

We prove posterior propriety and geometric ergodicity by establishing drift and minorization (d&m) conditions, which we now describe. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. Consider a Markov chain $(X(t))_{t=0}^{\infty}$ whose state space is $(\mathcal{X}, \mathcal{F})$, and let $K : \mathcal{X} \times \mathcal{F} \rightarrow [0, 1]$ be its transition kernel. We say that $(X(t))$ satisfies a d&m condition with drift function $V : \mathcal{X} \rightarrow [0, \infty)$, minorization measure $\nu : \mathcal{F} \rightarrow [0, 1]$, and parameters $(\eta, \varrho, \delta, \epsilon) \in \mathbb{R}^4$ if each of the following two conditions holds:

Drift condition: For $x \in \mathcal{X}$,

$$KV(x) := \int_{\mathcal{X}} V(x') K(x, dx') \leq \eta V(x) + \varrho,$$

where $\eta < 1$ and $\varrho < \infty$. Note that $KV(x)$ can be interpreted as the conditional expectation of $V(X(t+1))$ given $X(t) = x$, where t can be any non-negative integer.

Minorization condition: Let ν be a probability measure, $\epsilon > 0$, and $\delta > 2\varrho/(1-\eta)$. Moreover, whenever $V(x) < \delta$,

$$K(x, A) \geq \epsilon \nu(A) \text{ for each } A \in \mathcal{F}.$$

It is well-known that if a d&m condition holds, then the Markov chain has a proper stationary distribution, and it is geometrically ergodic [11, 17, 23, 27].

We begin by establishing a minorization condition for the DA algorithm associated with a realized response \mathbf{y} and missing structure \mathbf{k} . As before, d_i will be used to denote the number of nonzero elements in the i th row of \mathbf{k} . Let $(\mathbf{B}(t), \mathbf{\Sigma}(t))_{t=0}^{\infty}$ be the underlying Markov chain, and denote its Markov transition kernel by K . Set $\mathcal{X} = \mathbb{R}^{p \times d} \times S_+^{d \times d}$, and let \mathcal{F} be the usual Borel algebra associated with \mathcal{X} . Define a drift function

$$V(\boldsymbol{\beta}, \boldsymbol{\varsigma}) = \sum_{i=1}^n r_{i,(\mathbf{k})},$$

where, for $i \in \{1, \dots, n\}$,

$$r_{i,(\mathbf{k})} = (\mathbf{y}_{i,(\mathbf{k})} - \mathbf{c}_{i,(\mathbf{k})} \boldsymbol{\beta}^\top \mathbf{x}_i)^\top (\mathbf{c}_{i,(\mathbf{k})} \boldsymbol{\varsigma} \mathbf{c}_{i,(\mathbf{k})}^\top)^{-1} (\mathbf{y}_{i,(\mathbf{k})} - \mathbf{c}_{i,(\mathbf{k})} \boldsymbol{\beta}^\top \mathbf{x}_i).$$

Then the following holds.

Lemma 1. *For any $\delta > 0$, there exist a probability measure $\nu : \mathcal{F} \rightarrow [0, 1]$ and $\epsilon > 0$ such that whenever $V(\boldsymbol{\beta}, \boldsymbol{\varsigma}) < \delta$,*

$$K((\boldsymbol{\beta}, \boldsymbol{\varsigma}), A) > \epsilon \nu(A) \text{ for each } A \in \mathcal{F}.$$

Proof. One can write

$$K((\boldsymbol{\beta}, \boldsymbol{\varsigma}), A) = \int_{(0, \infty)^n} Q_k(A | \mathbf{w}, \mathbf{y}_{(k)}) P_k(d\mathbf{w} | \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)}),$$

where $P_k(\cdot | \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)})$ is the conditional distribution of \mathbf{W} given $(\mathbf{B}, \mathbf{\Sigma}, \mathbf{Y}_{(k)}) = (\boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)})$, and $Q_k(\cdot | \mathbf{w}, \mathbf{y}_{(k)})$ is the conditional distribution of $(\boldsymbol{\beta}, \mathbf{\Sigma})$ given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$, both derived from (2). As stated in Section 3.2,

$$P_k(d\mathbf{w} | \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)}) = \prod_{i=1}^n \frac{w_i^{d_i/2} \exp(-r_{i,(\mathbf{k})} w_i/2) P_{\text{mix}}(dw_i)}{\int_0^\infty w_i^{d_i/2} \exp(-r_{i,(\mathbf{k})} w_i/2) P_{\text{mix}}(dw_i)}, \quad \mathbf{w} = (w_1, \dots, w_n)^\top \in (0, \infty)^n.$$

Assume that $V(\boldsymbol{\beta}, \boldsymbol{\varsigma}) < \delta$ for some $\delta > 0$. Then $r_{i,(\mathbf{k})} < \delta$ for each i . It follows that, for any measurable $A' \in (0, \infty)^n$,

$$P_k(A' | \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)}) \geq \epsilon \int_{A'} \prod_{i=1}^n \frac{w_i^{d_i/2} \exp(-\delta w_i/2) P_{\text{mix}}(dw_i)}{\int_0^\infty w_i^{d_i/2} \exp(-\delta w_i/2) P_{\text{mix}}(dw_i)},$$

where

$$\epsilon = \left\{ \frac{\int_0^\infty w^{d_i/2} \exp(-\delta w/2) P_{\text{mix}}(dw)}{\int_0^\infty w^{d_i/2} P_{\text{mix}}(dw)} \right\}^n.$$

Thus, for each $A \in \mathcal{F}$,

$$K((\beta, \varsigma), A) \geq \epsilon \nu(A),$$

where $\nu(\cdot)$ is a probability measure given by

$$\nu(A) = \int_{(0, \infty)^n} Q_k(A \mid \mathbf{w}, \mathbf{y}_{(k)}) \prod_{i=1}^n \frac{w_i^{d_i/2} \exp(-\delta w_i/2) P_{\text{mix}}(dw_i)}{\int_0^\infty w_i^{d_i/2} \exp(-\delta w_i/2) P_{\text{mix}}(dw_i)}.$$

□

To establish d&m, it remains to show the following.

Lemma 2. Suppose that (6) holds, and that the conditional distribution of (\mathbf{B}, Σ) given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$ is proper for every $\mathbf{w} = (w_1, \dots, w_n)^\top \in (0, \infty)^n$. If any one of the following conditions holds, then there exist $\eta < 1$ and $\varrho < \infty$ such that, for $(\beta, \varsigma) \in \mathcal{X}$,

$$KV(\beta, \varsigma) < \eta V(\beta, \varsigma) + \varrho.$$

- (i) $P_{\text{mix}}(\cdot)$ is zero near the origin;
- (ii) $P_{\text{mix}}(\cdot)$ is faster than polynomial near the origin;
- (iii) $P_{\text{mix}}(\cdot)$ is polynomial near the origin with power $c > c_1$, where

$$c_1 = \frac{n - p + m - \min\{d_1, \dots, d_n\}}{2}.$$

Proof. We will prove the result when \mathbf{k} contains at least one vanishing element. When there are no missing data under \mathbf{k} , the result is proved by Hobert et al. [9]. (To be absolutely precise, Hobert et al. [9] assumed that $P_{\text{mix}}(\cdot)$ is absolutely continuous with respect to the Lebesgue measure in the “zero near the origin” case, but their proof can be easily extended to the case where absolute continuity is absent.)

For $\beta \in \mathbb{R}^{p \times d}$ and $\varsigma \in S_+^{d \times d}$, let $P_k(\cdot \mid \beta, \varsigma, \mathbf{y}_{(k)})$ be the conditional distribution of \mathbf{W} given $(\mathbf{B}, \Sigma, \mathbf{Y}_{(k)}) = (\beta, \varsigma, \mathbf{y}_{(k)})$. For $\mathbf{w} \in (0, \infty)^n$, let $Q_k(\cdot \mid \mathbf{w}, \mathbf{y}_{(k)})$ be the conditional distribution of (\mathbf{B}, Σ) given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$. Then

$$KV(\beta, \varsigma) = \int_{(0, \infty)^n} \int_{\mathbb{R}^{p \times d} \times S_+^{d \times d}} V(\beta^*, \varsigma^*) Q_k(d\beta^*, d\varsigma^* \mid \mathbf{w}, \mathbf{y}_{(k)}) P_k(d\mathbf{w} \mid \beta, \varsigma, \mathbf{y}_{(k)}). \quad (\text{B.1})$$

Let $\mathbf{k}_0 \in \{0, 1\}^{n \times d}$ be a matrix such that all its entries are 1. In other words, $\mathbf{K} = \mathbf{k}_0$ if there are no data missing. $\mathbf{Y}_{(k_0-k)}$ denotes the unobservable entries of \mathbf{Y} under the missing structure \mathbf{k} . By our assumptions, given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$, $(\mathbf{B}, \Sigma, \mathbf{Y}_{(k_0-k)})$ has a proper conditional distribution. For $\mathbf{w} \in (0, \infty)^n$, denote by $Q_{1,k}(\cdot \mid \mathbf{w}, \mathbf{y}_{(k)})$ the conditional distribution of $\mathbf{Y}_{(k_0-k)}$ given $(\mathbf{W}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{y}_{(k)})$. For $\mathbf{w} \in (0, \infty)^n$ and a realized value of $\mathbf{Y}_{(k_0-k)}$, say, $\mathbf{z} \in \mathbb{R}^{\sum_{i=1}^n (d-d_i)}$, let $Q_{2,k}(\cdot \mid \mathbf{w}, \mathbf{z}, \mathbf{y}_{(k)})$ be the conditional distribution of (\mathbf{B}, Σ) given $(\mathbf{W}, \mathbf{Y}_{(k_0-k)}, \mathbf{Y}_{(k)}) = (\mathbf{w}, \mathbf{z}, \mathbf{y}_{(k)})$. We now describe this distribution (see Appendix A). Let \mathbf{y}^* be a realized value of \mathbf{Y} such that $\mathbf{y}_{(k)}^* = \mathbf{y}_{(k)}$. Let $\lambda = \text{diag}(w_1, \dots, w_n)$, where w_1, \dots, w_n are the components of \mathbf{w} . Let

$$\hat{\beta} = (\mathbf{x}^\top \lambda \mathbf{x})^{-1} \mathbf{x}^\top \lambda \mathbf{y}^*, \quad \mathbf{s} = (\mathbf{y}^* - \mathbf{x} \hat{\beta})^\top \lambda (\mathbf{y}^* - \mathbf{x} \hat{\beta}),$$

and $\xi = (\mathbf{x}^\top \lambda \mathbf{x})^{-1}$. Then $Q_{2,k}(\cdot \mid \omega, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)})$ is given by

$$\Sigma \mid (\mathbf{W}, \mathbf{Y}) = (\mathbf{w}, \mathbf{y}^*) \sim \text{IW}_d(n - p + m - d, \mathbf{s} + \mathbf{a}), \quad \mathbf{B} \mid (\Sigma, \mathbf{W}, \mathbf{Y}) = (\varsigma, \mathbf{w}, \mathbf{y}^*) \sim \text{N}_{p,d}(\hat{\beta}, \xi, \varsigma).$$

It is easy to verify that Q_k , $Q_{1,k}$, and $Q_{2,k}$ are connected through the following tower property:

$$Q_k(\cdot \mid \mathbf{w}, \mathbf{y}_{(k)}) = \int_{\mathbb{R}^{\sum_{i=1}^n (d-d_i)}} Q_{2,k}(\cdot \mid \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)}) Q_{1,k}(d\mathbf{y}_{(k_0-k)}^* \mid \mathbf{w}, \mathbf{y}_{(k)}). \quad (\text{B.2})$$

In light of (B.1) and (B.2), let us first investigate the integral

$$\begin{aligned} & \int_{\mathbb{R}^{p \times d} \times S_+^{d \times d}} V(\beta^*, \varsigma^*) Q_{2,k}(d\beta^*, d\varsigma^* \mid \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)}) \\ &= \int_{S_+^{d \times d}} \int_{\mathbb{R}^{p \times d}} V(\beta^*, \varsigma^*) Q_{4,k}(d\beta^* \mid \varsigma^*, \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)}) Q_{3,k}(d\varsigma^* \mid \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)}), \end{aligned} \quad (\text{B.3})$$

where $Q_{3,k}(\cdot \mid \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)})$ is the $\text{IW}_d(n-p+m-d, s+\mathbf{a})$ distribution, and $Q_{4,k}(\cdot \mid \boldsymbol{\varsigma}^*, \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)})$ is the $N_{p,d}(\hat{\boldsymbol{\beta}}, \boldsymbol{\xi}, \boldsymbol{\varsigma}^*)$ distribution. It follows from basic properties of matrix normal distributions that

$$\begin{aligned} & \int_{\mathbb{R}^{p \times d}} V(\boldsymbol{\beta}^*, \boldsymbol{\varsigma}^*) Q_{4,k}(\mathrm{d}\boldsymbol{\beta}^* \mid \boldsymbol{\varsigma}^*, \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)}) \\ &= \sum_{i=1}^n \left\{ \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right)^\top \left(\mathbf{c}_{i,(k)} \boldsymbol{\varsigma}^* \mathbf{c}_{i,(k)}^\top \right)^{-1} \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right) + d_i \mathbf{x}_i^\top \boldsymbol{\xi} \mathbf{x}_i \right\}. \end{aligned}$$

If $\boldsymbol{\Sigma}^*$ is a random matrix that follows the $\text{IW}_d(n-p+m-d, s+\mathbf{a})$ distribution, then $(\mathbf{c}_{i,(k)} \boldsymbol{\Sigma}^* \mathbf{c}_{i,(k)}^\top)^{-1}$ follows the $W_{d_i}(n-p+m-2d+d_i, [\mathbf{c}_{i,(k)}(s+\mathbf{a})\mathbf{c}_{i,(k)}^\top]^{-1})$ distribution [18, Theorem 3.2.11]. Then by basic properties of Wishart distributions,

$$\begin{aligned} & \int_{S_{+}^{d \times d}} \int_{\mathbb{R}^{p \times d}} V(\boldsymbol{\beta}^*, \boldsymbol{\varsigma}^*) Q_{4,k}(\mathrm{d}\boldsymbol{\beta}^* \mid \boldsymbol{\varsigma}^*, \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)}) Q_{3,k}(\mathrm{d}\boldsymbol{\varsigma}^* \mid \mathbf{w}, \mathbf{y}_{(k_0-k)}^*, \mathbf{y}_{(k)}) \\ &= \sum_{i=1}^n (n-p+m-2d+d_i) \left[\left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right)^\top \left\{ \mathbf{c}_{i,(k)}(s+\mathbf{a})\mathbf{c}_{i,(k)}^\top \right\}^{-1} \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right) \right] + \sum_{i=1}^n d_i \mathbf{x}_i^\top \boldsymbol{\xi} \mathbf{x}_i. \end{aligned} \quad (\text{B.4})$$

For $i \in \{1, \dots, n\}$, denote by \mathbf{y}_i^* the i th row of \mathbf{y}^* , and note that $\mathbf{c}_{i,(k)} \mathbf{y}_i^* = \mathbf{y}_{i,(k)}$. It follows from Lemma 3, which is stated right after this proof, that, for $i \in \{1, \dots, n\}$,

$$\mathbf{x}_i^\top \boldsymbol{\xi} \mathbf{x}_i = \mathbf{x}_i^\top \left(\sum_{j=1}^n w_j \mathbf{x}_j \mathbf{x}_j^\top \right)^{-1} \mathbf{x}_i \leq \frac{1}{w_i},$$

and

$$\begin{aligned} & \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right)^\top \left\{ \mathbf{c}_{i,(k)}(s+\mathbf{a})\mathbf{c}_{i,(k)}^\top \right\}^{-1} \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right) \\ &= \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right)^\top \left[\mathbf{c}_{i,(k)} \left\{ \sum_{j=1}^n w_j (\mathbf{y}_j^* - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_j) (\mathbf{y}_j^* - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_j)^\top \right\} \mathbf{c}_{i,(k)}^\top + \mathbf{c}_{i,(k)} \mathbf{a} \mathbf{c}_{i,(k)}^\top \right]^{-1} \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right) \\ &= \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right)^\top \left\{ \sum_{j=1}^n w_j (\mathbf{y}_{j,(k)} - \mathbf{c}_{j,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_j) (\mathbf{y}_{j,(k)} - \mathbf{c}_{j,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_j)^\top + \mathbf{c}_{i,(k)} \mathbf{a} \mathbf{c}_{i,(k)}^\top \right\}^{-1} \left(\mathbf{y}_{i,(k)} - \mathbf{c}_{i,(k)} \hat{\boldsymbol{\beta}}^\top \mathbf{x}_i \right) \leq \frac{1}{w_i}. \end{aligned}$$

Notice that $1/w_i$ is a constant, and does not depend on $\mathbf{y}_{(k_0-k)}^*$. It then follows from (B.1) to (B.4) that

$$KV(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \leq \int_{(0,\infty)^n} \left(\sum_{i=1}^n \frac{n-p+m-2d+2d_i}{w_i} \right) P_k(\mathrm{d}\mathbf{w} \mid \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)}) \leq \int_{(0,\infty)^n} \left(\sum_{i=1}^n \frac{n-p+m}{w_i} \right) P_k(\mathrm{d}\mathbf{w} \mid \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)}),$$

where $P_k(\mathrm{d}\mathbf{w} \mid \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)})$ is given by (5). By Lemma 4 below, there exist $\eta' < 1/(n-p+m)$ and $\varrho' < \infty$ such that, for $(\boldsymbol{\beta}, \boldsymbol{\varsigma}) \in \mathcal{X}$,

$$\int_{(0,\infty)^n} \left(\sum_{i=1}^n \frac{n-p+m}{w_i} \right) P_k(\mathrm{d}\mathbf{w} \mid \boldsymbol{\beta}, \boldsymbol{\varsigma}, \mathbf{y}_{(k)}) \leq \sum_{i=1}^n (n-p+m)(\eta' r_{i,(k)} + \varrho').$$

Then desired result holds with $\eta = (n-p+m)\eta' < 1$ and $\varrho = n(n-p+m)\varrho' < \infty$. \square

Lemma 3 ([29, Lemma 3]). *Let $\boldsymbol{\varphi}$ be a positive definite matrix, and \mathbf{v} be a vector, such that the matrix $\boldsymbol{\varphi} - \mathbf{v}\mathbf{v}^\top$ is positive semidefinite, then $\mathbf{v}^\top \boldsymbol{\varphi}^{-1} \mathbf{v} \leq 1$.*

Lemma 4. *Suppose that (6) holds and that $P_{\text{mix}}(\cdot)$ is either zero near the origin, or faster than polynomial near the origin, or polynomial near the origin with power $c > c_1$, where*

$$c_1 = \frac{n-p+m-\min\{d_1, \dots, d_n\}}{2}.$$

Then there exist $\eta \in [0, 1/(n - p + m))$ and $\varrho < \infty$, such that for all $\tilde{d} \in [\min\{d_1, \dots, d_n\}, d]$ and $\tilde{r} \in [0, \infty)$,

$$\frac{\int_0^\infty w^{(\tilde{d}-2)/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)}{\int_0^\infty w^{\tilde{d}/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)} \leq \eta \tilde{r} + \varrho.$$

Proof. We will make use of existing results [9].

When $P_{\text{mix}}(\cdot)$ is zero near the origin, there exists $\theta > 0$, such that $\int_0^\theta P_{\text{mix}}(dw) = 0$. Then, for all $\tilde{d} \in [\min\{d_1, \dots, d_n\}, d]$ and $\tilde{r} \in [0, \infty)$,

$$\frac{\int_0^\infty w^{(\tilde{d}-2)/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)}{\int_0^\infty w^{\tilde{d}/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)} = \frac{\int_\theta^\infty (1/w) w^{\tilde{d}/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)}{\int_\theta^\infty w^{\tilde{d}/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)} \leq \frac{(1/\theta) \int_\theta^\infty w^{\tilde{d}/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)}{\int_\theta^\infty w^{\tilde{d}/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)} = 1/\theta.$$

When $P_{\text{mix}}(\cdot)$ is polynomial near the origin or faster than polynomial near the origin, recall that $p_{\text{mix}}(\cdot)$ is the density function of $P_{\text{mix}}(\cdot)$ with respect to the Lebesgue measure. Let $S(p_{\text{mix}})$ be the set of $\eta \in [0, \infty)$ such that there exists $\varrho_\eta \in \mathbb{R}$, satisfying

$$\frac{\int_0^\infty w^{-1/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)}{\int_0^\infty w^{1/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)} \leq \eta \tilde{r} + \varrho_\eta$$

for all $\tilde{r} \in [0, \infty)$. If $S(p_{\text{mix}})$ is non-empty, let $\eta_{p_{\text{mix}}} = \inf S(p_{\text{mix}})$; otherwise, set $\eta_{p_{\text{mix}}} = \infty$.

Consider a Gamma $(\alpha, 1)$ mixing distribution with density

$$p_{G(\alpha)}(w) = \{\Gamma_1(\alpha)\}^{-1} w^{\alpha-1} \exp(-w).$$

It is easy to see that if $\alpha > 1/2$, for all $\tilde{r} \in [0, \infty)$,

$$\frac{\int_0^\infty w^{-1/2} \exp(-\tilde{r}w/2) p_{G(\alpha)}(w) dw}{\int_0^\infty w^{1/2} \exp(-\tilde{r}w/2) p_{G(\alpha)}(w) dw} = \frac{1}{2\alpha - 1} \tilde{r} + \frac{2}{2\alpha - 1}.$$

Therefore, if $\alpha > 1/2$, $\eta_{p_{G(\alpha)}} = 1/(2\alpha - 1)$.

Let $p_{\text{mix}}^{\tilde{d}}(\cdot)$ be a mixing density proportional to $w^{(\tilde{d}-1)/2} p_{\text{mix}}(\cdot)$. Note that

$$\frac{\int_0^\infty w^{(\tilde{d}-2)/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)}{\int_0^\infty w^{\tilde{d}/2} \exp(-\tilde{r}w/2) P_{\text{mix}}(dw)} = \frac{\int_0^\infty w^{-1/2} \exp(-\tilde{r}w/2) p_{\text{mix}}^{\tilde{d}}(w) dw}{\int_0^\infty w^{1/2} \exp(-\tilde{r}w/2) p_{\text{mix}}^{\tilde{d}}(w) dw}.$$

Therefore, it suffices to prove that $\eta_{p_{\text{mix}}^{\tilde{d}}} < 1/(n - p + m)$.

When $P_{\text{mix}}(\cdot)$ is polynomial near the origin with power $c > c_1$, the distribution associated with $p_{\text{mix}}^{\tilde{d}}(\cdot)$ is polynomial near the origin with power $c + (\tilde{d} - 1)/2$. Let $p_G^{\tilde{d}}(\cdot)$ be a mixing density following Gamma $(c + (\tilde{d} + 1)/2, 1)$. We have

$$\lim_{w \rightarrow 0} \frac{p_{\text{mix}}^{\tilde{d}}(w)}{p_G^{\tilde{d}}(w)} \in (0, \infty).$$

By Lemma 1 in Hobert et al. [9], for all $\tilde{d} \in [\min\{d_1, \dots, d_n\}, d]$,

$$\eta_{p_{\text{mix}}^{\tilde{d}}} = \eta_{p_G^{\tilde{d}}} = \frac{1}{2c + \tilde{d}} < \frac{1}{n - p + m}.$$

When $P_{\text{mix}}(\cdot)$ is faster than polynomial near the origin, again let $p_{\text{mix}}^{\tilde{d}}(\cdot)$ be a mixing density proportional to $w^{(\tilde{d}-1)/2} p_{\text{mix}}(\cdot)$. The distribution associated with $p_{\text{mix}}^{\tilde{d}}(\cdot)$ is faster than polynomial near the origin. Let $\alpha > 1/2$. Recall that $p_{G(\alpha)}(w)$ is the density of the Gamma $(\alpha, 1)$ mixing distribution. By the definition of faster than polynomial near the origin, there exists $\kappa_{\alpha-1} > 0$, such that $p_{\text{mix}}^{\tilde{d}}(w)/w^{\alpha-1}$ is strictly increasing on $(0, \kappa_{\alpha-1})$. Thus, $p_{\text{mix}}^{\tilde{d}}(w)/p_{G(\alpha)}(w)$ is strictly increasing on $(0, \kappa_{\alpha-1})$. By Lemma 2 in Hobert et al. [9],

$$\eta_{p_{\text{mix}}^{\tilde{d}}} \leq \eta_{p_{G(\alpha)}} = \frac{1}{2\alpha - 1}.$$

Since α can be any value larger than $1/2$, $\eta_{p_{\text{mix}}^{\tilde{d}}} = 0$, for all $\tilde{d} \in [\min\{d_1, \dots, d_n\}, d]$.

□

Appendix C. The ESS and ESSpm of Each Component in Section 5

Table C.1: The ESS of each component in different scenarios. The model column lists the mixing distributions and samplers used, and the data column labels the missing structures by the number of fully observed rows. The B_{11} , B_{21} , B_{12} , and B_{22} columns store the ESS of each component of the regression coefficients B in different scenarios. The Σ_{11} , Σ_{21} , and Σ_{22} columns store the ESS of each component of the scatter matrix Σ .

| Model | Data | B_{11} | B_{21} | B_{12} | B_{22} | Σ_{11} | Σ_{21} | Σ_{22} |
|--------|------|----------|----------|----------|----------|---------------|---------------|---------------|
| DAG | 45 | 19332 | 17450 | 19098 | 16447 | 10290 | 11646 | 10528 |
| DAG | 40 | 20001 | 18399 | 17549 | 19992 | 10315 | 12061 | 10714 |
| DAG | 35 | 16960 | 18459 | 16786 | 16967 | 11605 | 12442 | 11551 |
| DAIG | 45 | 16471 | 19242 | 17976 | 19443 | 9970 | 10723 | 9888 |
| DAIG | 40 | 14292 | 15528 | 18045 | 17066 | 8335 | 11155 | 9897 |
| DAIG | 35 | 13330 | 17612 | 17745 | 16203 | 9676 | 8988 | 10064 |
| DAGIG | 45 | 16843 | 17429 | 14810 | 15401 | 13796 | 13232 | 13422 |
| DAGIG | 40 | 20661 | 20371 | 21888 | 14722 | 13375 | 14279 | 12765 |
| DAGIG | 35 | 17729 | 15738 | 17347 | 15574 | 15619 | 12788 | 12722 |
| DAIGIG | 45 | 18267 | 15193 | 17571 | 15148 | 11803 | 12759 | 13323 |
| DAIGIG | 40 | 16540 | 18311 | 16763 | 17924 | 14042 | 14860 | 14749 |
| DAIGIG | 35 | 12246 | 14077 | 18134 | 15717 | 8973 | 9506 | 12939 |
| DAP | 45 | 30000 | 30000 | 30000 | 30000 | 30000 | 29724 | 30000 |
| DAP | 40 | 30000 | 30000 | 30000 | 30000 | 30000 | 30000 | 30438 |
| DAP | 35 | 30000 | 30000 | 29216 | 30000 | 30000 | 30000 | 30000 |
| DAIP | 45 | 22822 | 26402 | 30000 | 30000 | 25435 | 29601 | 30000 |
| DAIP | 40 | 20145 | 23817 | 30000 | 30000 | 20160 | 26305 | 30000 |
| DAIP | 35 | 13894 | 22094 | 30000 | 30000 | 16530 | 13943 | 30000 |

Table C.2: The ESS per minute (ESSpm) of each component in different scenarios. The model column lists the mixing distributions and samplers used, and the data column labels the missing structures by the number of fully observed rows. The B_{11} , B_{21} , B_{12} , and B_{22} columns store the ESSpm of each component of the regression coefficients B in different scenarios. The Σ_{11} , Σ_{21} , and Σ_{22} columns store the ESSpm of each component of the scatter matrix Σ .

| Model | Data | B_{11} | B_{21} | B_{12} | B_{22} | Σ_{11} | Σ_{21} | Σ_{22} |
|--------|------|----------|----------|----------|----------|---------------|---------------|---------------|
| DAG | 45 | 4421 | 3990 | 4367 | 3761 | 2353 | 2663 | 2407 |
| DAG | 40 | 4886 | 4495 | 4287 | 4884 | 2520 | 2947 | 2618 |
| DAG | 35 | 4213 | 4585 | 4170 | 4215 | 2883 | 3090 | 2869 |
| DAIG | 45 | 3286 | 3838 | 3586 | 3878 | 1989 | 2139 | 1972 |
| DAIG | 40 | 2891 | 3141 | 3651 | 3452 | 1686 | 2257 | 2002 |
| DAIG | 35 | 2850 | 3766 | 3794 | 3464 | 2069 | 1922 | 2152 |
| DAGIG | 45 | 1611 | 1667 | 1417 | 1473 | 1320 | 1266 | 1284 |
| DAGIG | 40 | 2072 | 2043 | 2195 | 1477 | 1341 | 1432 | 1280 |
| DAGIG | 35 | 1803 | 1601 | 1765 | 1584 | 1589 | 1301 | 1294 |
| DAIGIG | 45 | 1701 | 1415 | 1636 | 1410 | 1099 | 1188 | 1240 |
| DAIGIG | 40 | 1543 | 1709 | 1564 | 1672 | 1310 | 1387 | 1376 |
| DAIGIG | 35 | 1201 | 1381 | 1779 | 1542 | 880 | 932 | 1269 |
| DAP | 45 | 6959 | 6959 | 6959 | 6959 | 6959 | 6895 | 6959 |
| DAP | 40 | 7793 | 7793 | 7793 | 7793 | 7793 | 7793 | 7907 |
| DAP | 35 | 7478 | 7478 | 7282 | 7478 | 7478 | 7478 | 7478 |
| DAIP | 45 | 4929 | 5702 | 6479 | 6479 | 5493 | 6393 | 6479 |
| DAIP | 40 | 4415 | 5219 | 6574 | 6574 | 4418 | 5765 | 6574 |
| DAIP | 35 | 3183 | 5062 | 6873 | 6873 | 3787 | 3194 | 6873 |

Table C.3: The ESS of each component in different scenarios. The model column lists the Gamma(ν, ν) mixing distributions and samplers used, and the data column labels the missing structures by the number of fully observed rows. The \mathbf{B}_{11} , \mathbf{B}_{21} , \mathbf{B}_{12} , and \mathbf{B}_{22} columns store the ESS of each component of the regression coefficients \mathbf{B} in different scenarios. The $\mathbf{\Sigma}_{11}$, $\mathbf{\Sigma}_{21}$, and $\mathbf{\Sigma}_{22}$ columns store the ESS of each component of the scatter matrix $\mathbf{\Sigma}$.

| Model | Data | \mathbf{B}_{11} | \mathbf{B}_{21} | \mathbf{B}_{12} | \mathbf{B}_{22} | $\mathbf{\Sigma}_{11}$ | $\mathbf{\Sigma}_{21}$ | $\mathbf{\Sigma}_{22}$ |
|--------|------|-------------------|-------------------|-------------------|-------------------|------------------------|------------------------|------------------------|
| DAG2 | 45 | 19332 | 17450 | 19098 | 16447 | 10290 | 11646 | 10528 |
| DAG2 | 40 | 20001 | 18399 | 17549 | 19992 | 10315 | 12061 | 10714 |
| DAG2 | 35 | 16960 | 18459 | 16786 | 16967 | 11605 | 12442 | 11551 |
| DAIG2 | 45 | 16471 | 19242 | 17976 | 19443 | 9970 | 10723 | 9888 |
| DAIG2 | 40 | 14292 | 15528 | 18045 | 17066 | 8335 | 11155 | 9897 |
| DAIG2 | 35 | 13330 | 17612 | 17745 | 16203 | 9676 | 8988 | 10064 |
| DAG8 | 45 | 26359 | 23824 | 25032 | 25504 | 16537 | 19326 | 16185 |
| DAG8 | 40 | 26122 | 26707 | 25362 | 26588 | 14952 | 19118 | 15616 |
| DAG8 | 35 | 25470 | 24347 | 24819 | 25501 | 18445 | 16390 | 16759 |
| DAIG8 | 45 | 20631 | 23024 | 25552 | 26183 | 13685 | 15991 | 15078 |
| DAIG8 | 40 | 18122 | 21684 | 23664 | 25824 | 13061 | 18082 | 16806 |
| DAIG8 | 35 | 14837 | 22029 | 24967 | 26256 | 11553 | 11712 | 15336 |
| DAG25 | 45 | 29041 | 29055 | 27919 | 28865 | 21494 | 22572 | 23345 |
| DAG25 | 40 | 28018 | 29302 | 28548 | 29081 | 23303 | 21042 | 23013 |
| DAG25 | 35 | 29310 | 29162 | 29052 | 30000 | 18884 | 16623 | 20118 |
| DAIG25 | 45 | 23968 | 24454 | 27999 | 29108 | 17254 | 20737 | 22495 |
| DAIG25 | 40 | 18613 | 23362 | 28295 | 28963 | 15049 | 18725 | 21078 |
| DAIG25 | 35 | 17014 | 20338 | 28073 | 29150 | 10450 | 9397 | 22567 |

Table C.4: The ESS per minute (ESSpm) of each component in different scenarios. The model column lists the Gamma(ν, ν) mixing distributions and samplers used, and the data column labels the missing structures by the number of fully observed rows. The \mathbf{B}_{11} , \mathbf{B}_{21} , \mathbf{B}_{12} , and \mathbf{B}_{22} columns store the ESSpm of each component of the regression coefficients \mathbf{B} in different scenarios. The $\mathbf{\Sigma}_{11}$, $\mathbf{\Sigma}_{21}$, and $\mathbf{\Sigma}_{22}$ columns store the ESSpm of each component of the scatter matrix $\mathbf{\Sigma}$.

| Model | Data | \mathbf{B}_{11} | \mathbf{B}_{21} | \mathbf{B}_{12} | \mathbf{B}_{22} | $\mathbf{\Sigma}_{11}$ | $\mathbf{\Sigma}_{21}$ | $\mathbf{\Sigma}_{22}$ |
|--------|------|-------------------|-------------------|-------------------|-------------------|------------------------|------------------------|------------------------|
| DAG2 | 45 | 4421 | 3990 | 4367 | 3761 | 2353 | 2663 | 2407 |
| DAG2 | 40 | 4886 | 4495 | 4287 | 4884 | 2520 | 2947 | 2618 |
| DAG2 | 35 | 4213 | 4585 | 4170 | 4215 | 2883 | 3090 | 2869 |
| DAIG2 | 45 | 3286 | 3838 | 3586 | 3878 | 1989 | 2139 | 1972 |
| DAIG2 | 40 | 2891 | 3141 | 3651 | 3452 | 1686 | 2257 | 2002 |
| DAIG2 | 35 | 2850 | 3766 | 3794 | 3464 | 2069 | 1922 | 2152 |
| DAG8 | 45 | 6129 | 5539 | 5820 | 5930 | 3845 | 4494 | 3763 |
| DAG8 | 40 | 6495 | 6641 | 6306 | 6611 | 3718 | 4754 | 3883 |
| DAG8 | 35 | 6632 | 6339 | 6462 | 6640 | 4803 | 4268 | 4364 |
| DAIG8 | 45 | 4095 | 4570 | 5072 | 5197 | 2717 | 3174 | 2993 |
| DAIG8 | 40 | 3894 | 4659 | 5085 | 5549 | 2806 | 3885 | 3611 |
| DAIG8 | 35 | 3265 | 4847 | 5494 | 5777 | 2542 | 2577 | 3375 |
| DAG25 | 45 | 7115 | 7118 | 6840 | 7072 | 5266 | 5530 | 5719 |
| DAG25 | 40 | 7097 | 7423 | 7232 | 7367 | 5903 | 5330 | 5830 |
| DAG25 | 35 | 7728 | 7689 | 7660 | 7910 | 4979 | 4383 | 5304 |
| DAIG25 | 45 | 5067 | 5170 | 5919 | 6154 | 3648 | 4384 | 4755 |
| DAIG25 | 40 | 4099 | 5144 | 6231 | 6378 | 3314 | 4123 | 4642 |
| DAIG25 | 35 | 3898 | 4660 | 6432 | 6679 | 2394 | 2153 | 5170 |

References

- [1] O. Barndorff-Nielsen, J. Kent, M. Sørensen, Normal variance-mean mixtures and z distributions, *International Statistical Review/Revue Internationale de Statistique* 50 (1982) 145–159.
- [2] K. S. Chan, C. J. Geyer, Discussion: Markov chains for exploring posterior distributions, *The Annals of Statistics* 22 (1994) 1747–1758.
- [3] C. Fernández, M. F. Steel, Multivariate Student-t regression models: Pitfalls and inference, *Biometrika* 86 (1999) 153–167.
- [4] C. Fernandez, M. F. Steel, Bayesian regression analysis with scale mixtures of normals, *Econometric Theory* 16 (2000) 80–101.
- [5] J. M. Flegal, M. Haran, G. L. Jones, Markov chain Monte Carlo: Can we trust the third significant figure?, *Statistical Science* 23 (2008) 250–260.
- [6] J. M. Flegal, J. Hughes, D. Vats, N. Dai, K. Gupta, U. Maji, *mcmcse: Monte Carlo Standard Errors for MCMC*, Riverside, CA, and Kanpur, India, 2021. R package version 1.5-0.
- [7] L. Gong, J. M. Flegal, A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo, *Journal of Computational and Graphical Statistics* 25 (2016) 684–700.
- [8] J. P. Hobert, G. Casella, The effect of improper priors on Gibbs sampling in hierarchical linear mixed models, *Journal of the American Statistical Association* 91 (1996) 1461–1473.
- [9] J. P. Hobert, Y. J. Jung, K. Khare, Q. Qin, Convergence analysis of MCMC algorithms for Bayesian multivariate linear regression with non-Gaussian errors, *Scandinavian Journal of Statistics* 45 (2018) 513–533.
- [10] G. L. Jones, On the Markov chain central limit theorem, *Probability Surveys* 1 (2004) 299–320.
- [11] G. L. Jones, J. P. Hobert, Honest exploration of intractable probability distributions via Markov chain Monte Carlo, *Statistical Science* 16 (2001) 312–334.
- [12] G. L. Jones, Q. Qin, Markov chain Monte Carlo in practice, *Annual Review of Statistics and Its Application* 9 (2021).
- [13] K. Lange, J. S. Sinsheimer, Normal/independent distributions and their applications in robust regression, *Journal of Computational and Graphical Statistics* 2 (1993) 175–198.
- [14] R. Little, D. Rubin, *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ, 2019.
- [15] C. Liu, Bayesian robust multivariate linear regression with incomplete data, *Journal of the American Statistical Association* 91 (1996) 1219–1227.
- [16] J. S. Liu, W. H. Wong, A. Kong, Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes, *Biometrika* 81 (1994) 27–40.
- [17] S. P. Meyn, R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer Science & Business Media, 2005.
- [18] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, Hoboken, NJ, 2009.
- [19] Q. Qin, J. P. Hobert, Trace-class Monte Carlo Markov chains for Bayesian multivariate linear regression with non-Gaussian errors, *Journal of Multivariate Analysis* 166 (2018) 335–345.
- [20] Q. Qin, G. L. Jones, Convergence rates of two-component MCMC samplers, *Bernoulli* 28 (2022) 859–885.
- [21] G. Roberts, J. Rosenthal, Geometric ergodicity and hybrid Markov chains, *Electronic Communications in Probability* 2 (1997) 13–25.
- [22] G. O. Roberts, J. S. Rosenthal, Markov chains and de-initializing processes, *Scandinavian Journal of Statistics* 28 (2001) 489–504.
- [23] G. O. Roberts, J. S. Rosenthal, General state space Markov chains and MCMC algorithms, *Probability Surveys* 1 (2004) 20–71.
- [24] G. O. Roberts, J. S. Rosenthal, Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains, *The Annals of Applied Probability* 16 (2006) 2123–2139.
- [25] G. O. Roberts, R. L. Tweedie, Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains, *Journal of Applied Probability* 38 (2001) 37–41.
- [26] W. H. Rogers, J. W. Tukey, Understanding some long-tailed symmetrical distributions, *Statistica Neerlandica* 26 (1972) 211–226.
- [27] J. S. Rosenthal, Minorization conditions and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association* 90 (1995) 558–566.
- [28] J. S. Rosenthal, Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms, *Journal of the American Statistical Association* 98 (2003) 169–177.
- [29] V. Roy, J. P. Hobert, On Monte Carlo methods for Bayesian multivariate regression models with heavy-tailed errors, *Journal of Multivariate Analysis* 101 (2010) 1190–1202.
- [30] S. Seaman, J. Galati, D. Jackson, J. Carlin, What is meant by “missing at random”?, *Statistical Science* 28 (2013) 257–268.
- [31] D. Vats, J. M. Flegal, G. L. Jones, Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo, *Bernoulli* 24 (2018) 1860–1909.
- [32] D. Vats, J. M. Flegal, G. L. Jones, Multivariate output analysis for Markov chain Monte Carlo, *Biometrika* 106 (2019) 321–337.
- [33] A. Zellner, Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms, *Journal of the American Statistical Association* 71 (1976) 400–405.