A Note on Ising Network Analysis with Missing Data

Siliang Zhang and Yunxiao Chen

Abstract

The Ising model has become a popular psychometric model for analyzing item response data. The statistical inference of the Ising model is typically carried out via a pseudo-likelihood, as the standard likelihood approach suffers from a high computational cost when there are many variables (i.e., items). Unfortunately, the presence of missing values can hinder the use of pseudo-likelihood, and a listwise deletion approach for missing data treatment may introduce a substantial bias into the estimation and sometimes yield misleading interpretations. This paper proposes a conditional Bayesian framework for Ising network analysis with missing data, which integrates a pseudo-likelihood approach with iterative data imputation. An asymptotic theory is established for the method. Furthermore, a computationally efficient Pólya-Gamma data augmentation procedure is proposed to streamline the sampling of model parameters. The method's performance is shown through simulations and a real-world application to data on major depressive and generalized anxiety disorders from the National Epidemiological Survey on Alcohol and Related Conditions (NESARC).

KEYWORDS: Ising model, iterative imputation, full conditional specification, network psychometrics, mental health disorders, major depressive disorder, generalized anxiety disorder

1 Introduction

Recent years have witnessed the emergence of network psychometrics (van der Maas et al., 2006; Borsboom, 2008; Marsman & Rhemtulla, 2022), a family of statistical graphical models and related inference procedures, for analyzing and interpreting the dependence structure in psychometric data. These models embed psychometric items as nodes in an undirected or directed network (i.e., graph) and visualize their interrelationships through the network

edges, which represent certain probabilistic conditional dependencies. Network psychometric methods concern the learning of the network structure. They have been developed under various settings, including undirected graphical models for cross-sectional data (Epskamp, Waldorp, et al., 2018; Burger et al., 2022), directed networks for longitudinal data (Gile & Handcock, 2017; Borsboom et al., 2021; Ryan et al., 2022), and extended networks with latent variables for time-series data or panel data (Epskamp, 2020). These methods have received wide applications in education (Sweet et al., 2013; Willcox & Huang, 2017; Koponen et al., 2019; Siew, 2020; Simon de Blas et al., 2021), psychology (Burgess & Hitch, 1999; Van Der Maas et al., 2017; Fried et al., 2017; Epskamp, Borsboom, & Fried, 2018; Borsboom et al., 2021), and health sciences (Luke & Harris, 2007; Brunson & Laubenbacher, 2018; Mkhitaryan et al., 2019; Kohler et al., 2022).

Analyzing cross-sectional binary item response data with the Ising model (Ising, 1925) is common in network psychometric analysis. This analysis is typically performed based on a conditional likelihood (Besag, 1974) because the standard likelihood function is computationally infeasible when involving many variables. In this direction, Bayesian and frequentist methods have been developed, where sparsity-inducing priors or penalties are combined with the conditional likelihood for learning a sparse network structure (Yuan & Lin, 2007; Mazumder & Hastie, 2012; Van Borkulo et al., 2014; Epskamp & Fried, 2018; Li et al., 2019; Marsman et al., 2022). Besides, the Ising model is shown to be closely related to Item Response Theory (IRT) models (Holland, 1990; Anderson & Yu, 2007). The log-multiplicative association models (Anderson & Yu, 2007), which are special cases of the Ising model, can be used as item response theory models and yield very similar results as IRT models. Furthermore, the Ising model and the conditional likelihood have been used for modeling the local dependence structure in locally dependent IRT models (Ip, 2002; Chen et al., 2018).

Due to its construction, the conditional likelihood does not naturally handle data with missing values, despite the omnipresence of missing data in psychometric data. To deal with missing values in an Ising network analysis, listwise deletion (Haslbeck & Fried, 2017; Fried et al., 2020) and single imputation (e.g., Huisman, 2009; Armour et al., 2017; Lin et al., 2020) are typically performed, which arguably may not be the best practice. In particular, it is well-known that listwise deletion is statistically inefficient and requires the Missing Completely At Random (MCAR) assumption (Little & Rubin, 2019) to ensure consistent estimation. Moreover, a naïve imputation procedure, such as mode imputation, likely introduces bias

into parameter estimation. A sophisticated imputation procedure must be developed to ensure statistical validity and computational efficiency.

In this note, we propose an iterative procedure for learning an Ising network. The proposed procedure combines iterative imputation via Full Conditional Specification (FCS; Liu et al., 2014; van Buuren, 2018) and Bayesian estimation of the Ising network. We show that the FCS leads to estimation consistency when the conditional models are chosen to take logistic forms. In terms of computation, we propose a joint Pólya-Gamma augmentation procedure by extending the Pólya-Gamma augmentation procedure for logistic regression (Polson et al., 2013). It allows us to efficiently sample parameters of the Ising model. Simulations are conducted to compare the proposed procedure with estimations based on the listwise deletion and single imputation. Finally, the proposed procedure and a complete-case analysis are applied to study the network of Major Depressive Disorder (MDD) and Generalised Anxiety Disorders (GAD) based on data from the National Epidemiological Survey on Alcohol and Related Conditions (NESARC; Grant et al., 2003). Both analyses suggest that the symptoms are densely connected within each mental health disorder while only loosely connected between the two disorders. However, the two methods estimate a strong edge to be of opposite signs, leading to substantially different interpretations. A close scrutiny of the item content and the data missingness mechanism suggests that the result from the proposed method is more sensible.

2 Proposed Method

2.1 Ising Model

Consider a respondent answering J binary items. Let $\mathbf{Y} = (Y_1, ..., Y_J)^\top \in \{0, 1\}^J$ be a binary random vector representing the respondent's responses. We say $\mathbf{Y}$ follows an Ising model if its probability mass function satisfies

$$P(\mathbf{Y} = \boldsymbol{y} \mid \mathbf{S}) = \frac{1}{c(\mathbf{S})} \exp\left[\frac{1}{2}\boldsymbol{y}^\top \mathbf{S}\boldsymbol{y}\right] = \frac{1}{c(\mathbf{S})} \exp\left[\sum_{j=1}^{J} s_{jj} y_j/2 + \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} s_{jk} y_j y_k\right], \qquad (1)$$

where $\mathbf{S} = (s_{ij})_{J\times J}$ is a J by J symmetric matrix that contains parameters of the Ising model and

$$c(\mathbf{S}) = \sum_{\boldsymbol{y}\in\{0,1\}^J} \exp\left[\sum_{j=1}^{J} s_{jj}y_j/2 + \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} s_{jk}y_jy_k\right]$$

is a normalizing constant. The parameter matrix $\mathbf{S}$ encodes a network with the J items being the nodes. More specifically, an edge is present between nodes i and j if and only if the corresponding entry s_{ij} is nonzero. If an edge exists between nodes i and j, then Y_i and Y_j are conditionally dependent given the rest of the variables. Otherwise, the two variables are conditionally independent.

In Ising network analysis, the goal is to estimate the parameter matrix $\mathbf{S}$. The standard likelihood function is computationally intensive when J is large, as it requires computing a normalizing constant $c(\mathbf{S})$ which involves a summation of all the 2^J response patterns. To address this computational issue, Besag (1975) proposed a conditional likelihood which is obtained by aggregating the conditional distributions of Y_j given $\mathbf{Y}_{-j} = (Y_1, ..., Y_{j-1}, Y_{j+1}, ..., Y_J)^\top$, for $j = 1, ..., J$, where the conditional distribution of Y_j given $\mathbf{Y}_{-j}$ takes a logistic regression form. More precisely, the conditional likelihood with one observation $\boldsymbol{y}$ is defined as

$$p^*(\boldsymbol{y} \mid \mathbf{S}) = \prod_{j=1}^{J} p(y_j \mid \boldsymbol{y}_{-j}, \mathbf{S}) = \prod_{j=1}^{J} \frac{\exp\left[(s_{jj}/2 + \sum_{k\neq j} s_{jk}y_k)y_j\right]}{1 + \exp\left(s_{jj}/2 + \sum_{k\neq j} s_{jk}y_k\right)}. \tag{2}$$

A disadvantage of the conditional likelihood is that it requires a fully observed dataset because missing values cannot be straightforwardly marginalized out from (2). In what follows, we discuss how missing data can be treated in the conditional likelihood.

2.2 Proposed Method

Consider a dataset with N observations. Let $\Omega_j \subset \{1, ..., N\}$ be the subset of observations whose data on item j are missing. For each observation i and item j, y_{ij} denotes the observed response if $i \notin \Omega_j$, and otherwise, y_{ij} is missing. Thus, the observed data include Ω_j and y_{ij}, for $i \in \{1, ..., N\}\backslash\Omega_j$ and $j = 1, ..., J$.

The proposed procedure iterates between two steps – (1) imputing the missing values of y_{ij} for $i \in \Omega_j$, $j = 1, ..., J$, achieved via a full conditional specification, and (2) sampling

the posterior distribution of $\mathbf{S}$ given the most recently imputed data. Let t be the current iteration number. Further, let $\boldsymbol{y}_i^{(t-1)} = (y_{i1}^{(t-1)}, ..., y_{iJ}^{(t-1)})^\top, i = 1, ..., N$, be imputed data from the previous iteration, where $y_{ij}^{(t-1)} = y_{ij}$ for $i \notin \Omega_j$ and $y_{ij}^{(t-1)}$ is imputed in the $(t-1)$th iteration for $i \in \Omega_j$. For the tth iteration, the imputation and sampling steps are described as follows.

Imputation. We initialize the imputation in the tth iteration by setting $\boldsymbol{y}_i^{(t,0)} = \boldsymbol{y}_i^{(t-1)}$. Then, we run a loop over all the items, $j = 1, ..., J$. In step j of the loop, we impute y_{ij} for all $i \in \Omega_j$, given the most recently imputed data, denoted by $\boldsymbol{y}_i^{(t,j-1)}$, $i = 1, ..., N$. We then obtain $\boldsymbol{y}_i^{(t,j)}$ by updating $\boldsymbol{y}_i^{(t,j-1)}$ with the imputed values of y_{ij}.

The imputation of each variable j is based on the conditional distribution of Y_j given $\mathbf{Y}_{-j}$. Under the Ising model, this conditional distribution takes a logistic regression form. For computational reasons to be discussed in the sequel, we introduce an auxiliary parameter vector $\boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jJ})^\top$ as coefficients in the logistic regression, instead of directly using $\mathbf{S}$ from the previous iteration to sample the missing y_{ij}s. Unlike the constraint of $s_{ij} = s_{ji}$ in the symmetric matrix $\mathbf{S}$, no constraints are imposed on $\boldsymbol{\beta}_j$, $j = 1, ..., J$, which makes the sampling computationally efficient; see discussions in Section 2.4. The imputation of variable j consists of the following two steps:

1. Sample auxiliary parameter vector $\boldsymbol{\beta}_j^{(t)}$ from the posterior distribution

$$p^{(t,j)}(\boldsymbol{\beta}_j) \propto \pi_j(\boldsymbol{\beta}_j) \prod_{i=1}^{N} \frac{\exp\left[(\beta_{jj}/2 + \sum_{k \neq j} \beta_{jk} y_{ik}^{(t,j-1)}) y_{ij}^{(t,j-1)}\right]}{1 + \exp(\beta_{jj}/2 + \sum_{k \neq j} \beta_{jk} y_{ik}^{(t,j-1)})}, \tag{3}$$

   where $\pi_j(\boldsymbol{\beta}_j)$ is the prior distribution for the auxiliary parameters $\boldsymbol{\beta}_j$.

2. Sample $y_{ij}^{(t)}$ for each $i \in \Omega_j$ from a Bernoulli distribution with success probability

$$\frac{\exp(\beta_{jj}^{(t)}/2 + \sum_{k \neq j} \beta_{jk}^{(t)} y_{ik}^{(t,j-1)})}{1 + \exp(\beta_{jj}^{(t)}/2 + \sum_{k \neq j} \beta_{jk}^{(t)} y_{ik}^{(t,j-1)})}. \tag{4}$$

After these two steps, we obtain $\boldsymbol{y}_i^{(t,j)}$ by updating the jth element of $\boldsymbol{y}_i^{(t,j-1)}$ with $y_{ij}^{(t)}$, for $i \in \Omega_j$. We emphasize that only the missing values are updated. For $i \notin \Omega_j$, the jth element of $\boldsymbol{y}_i^{(t,j)}$ is always the observed value of y_{ij}. After the loop over all the items, we set

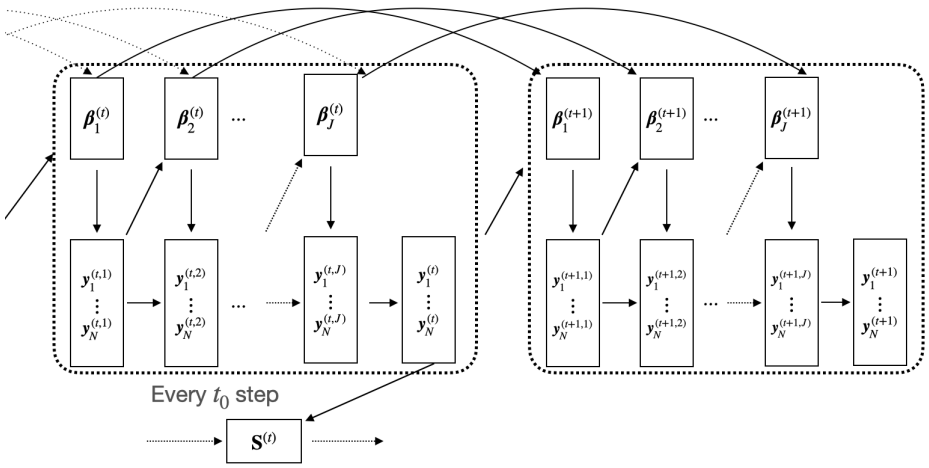


Figure 1: Flow chart of the updating rule for the proposed method.

$\boldsymbol{y}_i^{(t)} = \boldsymbol{y}_i^{(t,J)}$ as the output from this imputation step.

Sampling S. Given the most recently imputed data $\boldsymbol{y}_i^{(t)}$, $i = 1, ..., N$, update $\mathbf{S}^{(t)}$ by sampling from the pseudo-posterior distribution

$$p(\mathbf{S} \mid \boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_N^{(t)}) \propto \pi(\mathbf{S}) \prod_{i=1}^{N} p^*(\boldsymbol{y}_i^{(t)} \mid \mathbf{S}), \tag{5}$$

where $\pi(\mathbf{S})$ is the prior distribution for the Ising parameter matrix $\mathbf{S}$ and recall that $\prod_{i=1}^{N} p^*(\boldsymbol{y}_i^{(t)} \mid \mathbf{S})$ is the conditional likelihood.

Figure 1 visualizes the steps performed in the proposed method. Note that it is unnecessary to sample the parameter matrix $\mathbf{S}$ during the burn-in period and in every iteration after the burn-in period; thus, we employ a thinning step after the burn-in period. This is done to both decrease computational cost and reduce the auto-correlation in the imputed data. Moreover, we outline the proposed algorithm in Algorithm 1. The final estimate of $\mathbf{S}$ is obtained by averaging all the $\mathbf{S}^{(t)}$ obtained after the burn-in period. The computational details, including the sampling of auxiliary parameters and Ising parameter matrix and discussions of the computational complexity, are given in Section 2.4.

2.3 Statistical Consistency

As our method is not a standard Bayesian inference procedure, we provide an asymptotic theory under the frequentist setting to justify its validity. In particular, we show that the $\mathbf{S}$

Algorithm 1: Ising Network Analysis with Iterative Imputation

Data: Given observed data, initial values for the Ising model parameters $\mathbf{S}^{(0)}$, randomly imputed missing data $\boldsymbol{y}_1^{(0)}, \ldots, \boldsymbol{y}_N^{(0)}$, MCMC length T, burn-in size T_0, thinning steps size t_0. Let auxiliary parameters $\boldsymbol{\beta}_j^{(0)} = \boldsymbol{s}_j^{(0)}, j = 1, \ldots, J$.

for *each iteration* $t = 1$ *to* T **do**

  for *each* $j = 1$ *to* J **do**

    Sample auxiliary parameter vector $\boldsymbol{\beta}_j^{(t)}$ from $p^{(t,j)}(\boldsymbol{\beta}_j)$.

    Sample $y_{ij}^{(t)}$ for each $i \in \Omega_j$ from the Bernoulli distribution given in (4).

  end

  if $t > T_0$ *and* t *is a multiple of* t_0 **then**

    Sample $\mathbf{S}^{(t)}$ from $p(\mathbf{S} \mid \boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_N^{(t)})$ given in (5).

  end

end

Output: $\hat{\mathbf{S}} = \frac{1}{M - M_0} \sum_{t \in \{(M_0+1)t_0, \ldots, Mt_0\}} \mathbf{S}^{(t)}$, where $M = \lfloor T/t_0 \rfloor$ and $M_0 = \lfloor T_0/t_0 \rfloor$.

parameter sampled from the pseudo-posterior distribution converges to the true parameter $\mathbf{S}_0$, under the assumptions that the Ising model is correctly specified and the data are Missing At Random (MAR; Little & Rubin, 2019).

Consider one observation with a complete data vector $\mathbf{Y} = (Y_1, ..., Y_J)^\top$. Further, let $\mathbf{Z} = (Z_1, ..., Z_J)^\top$ be a vector of missing indicators, where $Z_{ij} = 1$ if Y_{ij} is observed and $Z_{ij} = 0$ otherwise. We further let $\mathbf{Y}_{obs} = \{Y_j : Z_j = 1, j = 1, ..., J\}$ and $\mathbf{Y}_{mis} = \{Y_j : Z_j = 0, j = 1, ..., J\}$ be the observed and missing entries of $\mathbf{Y}$, respectively. Consider the joint distribution of observable data $(\mathbf{Y}_{obs}, \mathbf{Z})$, taking the form

$$P(\mathbf{Y}_{obs} = \boldsymbol{y}_{obs}, \mathbf{Z} = \mathbf{z} \mid \mathbf{S}, \boldsymbol{\phi}) = \sum_{y_j : z_j = 0} \left(\exp\left(\boldsymbol{y}^\top \mathbf{S} \boldsymbol{y} / 2 \right) / c(\mathbf{S}) \right) q(\mathbf{z} \mid \boldsymbol{y}, \boldsymbol{\phi}), \tag{6}$$

where $\exp\left(\boldsymbol{y}^\top \mathbf{S} \boldsymbol{y} / 2 \right) / c(\mathbf{S})$ is the distribution of $\mathbf{Y} = \boldsymbol{y}$ under the Ising model, $q(\mathbf{z} \mid \boldsymbol{y}, \boldsymbol{\phi})$ denotes the conditional probability of $\mathbf{Z} = \mathbf{z}$ given $\mathbf{Y} = \boldsymbol{y}$, and $\boldsymbol{\phi}$ denotes the unknown parameters of this distribution. The MAR assumption, also known as the ignorable missingness assumption, means that the conditional distribution $q(\mathbf{z} \mid \boldsymbol{y}, \boldsymbol{\phi})$ depends on $\boldsymbol{y}$ only through the observed entries, i.e., $q(\mathbf{z} \mid \boldsymbol{y}, \boldsymbol{\phi}) = q(\mathbf{z} \mid \boldsymbol{y}_{obs}, \boldsymbol{\phi})$. In that case, (6) can be factorized as

$$P(\mathbf{Y}_{obs} = \boldsymbol{y}_{obs}, \mathbf{Z} = \mathbf{z} \mid \mathbf{S}, \boldsymbol{\phi}) = q(\mathbf{z} \mid \boldsymbol{y}_{obs}, \boldsymbol{\phi}) \times \left(\sum_{y_j : z_j = 0} \exp\left(\boldsymbol{y}^\top \mathbf{S} \boldsymbol{y} / 2 \right) / c(\mathbf{S}) \right). \tag{7}$$

Consequently, the inference of $\mathbf{S}$ does not depend on the unknown distribution $q(\mathbf{z} \mid \boldsymbol{y}, \boldsymbol{\phi})$.

As shown in Liu et al. (2014), the MAR assumption, together with additional regularity conditions, ensures that the iterative imputation of the missing responses converges to the imputation distribution under a standard Bayesian procedure as the number of iterations and the sample size N go to infinity. A key to this convergence result is the compatibility of the conditional models in the imputation step – the logistic regression models are compatible with the Ising model as a joint distribution. The validity of the imputed samples further ensures the consistency of the estimated Ising parameter matrix. We summarize this result in Theorem 1.

Theorem 1. *Assume the following assumptions hold: 1) The Markov chain for missing data, generated by the iterative imputation algorithm Algorithm 1, is positive Harris recurrent and thus admits a unique stationary distribution; 2) The missing data process is ignorable; 3) A regularity condition holds for prior distributions of Ising model parameters and auxiliary parameters, as detailed in Appendix A. Let $\pi_N^*(\mathbf{S})$ be the posterior density of $\mathbf{S}$ implied by the stationary distribution of the proposed method. Given the true parameters $\mathbf{S}_0$ for the Ising model and any $\varepsilon > 0$, we have $\pi_N^*(\mathbf{S})$ concentrates at $\mathbf{S}_0$,*

$$\int_{B_\varepsilon(\mathbf{S}_0)} \pi_N^*(\mathbf{S}) d\mathbf{S} \to 1, \tag{8}$$

in probability as $N \to \infty$. $B_\varepsilon(\mathbf{S}_0) = \{\mathbf{S} : \|\mathbf{S} - \mathbf{S}_0\| < \varepsilon\}$ is the open ball of radius ε at $\mathbf{S}_0$.

We note that the regularity condition on the prior distributions holds for the normal priors adopted in the current paper; see Section 2.4 for the specification of the priors and Appendix A.1 for the verification of the condition under the normal priors.

2.4 Computational Details

In what follows, we discuss the specification of the prior distributions and the sampling of auxiliary parameters $\boldsymbol{\beta}_j$ and Ising model parameters $\mathbf{S}$.

Sampling $\boldsymbol{\beta}_j$. We set independent mean-zero normal priors for entries of $\boldsymbol{\beta}_j$. For the intercept parameter β_{jj}, we use a weakly informative prior by setting the variance to 100. For the slope parameters β_{jk}, $k \neq j$, we set a more informative prior by setting the variance

to be 1, given that these parameters correspond to the off-diagonal entries of $\mathbf{S}$, which are sparse and typically do not take extreme values. The sampling of the auxiliary parameters $\boldsymbol{\beta}_j$, following (3), is essentially a standard Bayesian logistic regression problem. We achieve it by a Markov chain Monte Carlo (MCMC) sampler called the Pólya-Gamma sampler (Polson et al., 2013).

To obtain $\boldsymbol{\beta}_j^{(t)}$, this sampler starts with $\boldsymbol{\beta}_j^{(t-1)}$ from the previous step. It constructs an MCMC transition kernel by a data argumentation trick. More precisely, the following two steps are performed.

1. Given $\boldsymbol{\beta}^{(t-1)}$, independently sample N augmentation variables, each from a Pólya-Gamma distribution (Barndorff-Nielsen et al., 1982).

2. Given the N augmentation variables, sample $\boldsymbol{\beta}^{(t)}$ from a J-variate normal distribution.

The details of these two steps are given in Appendix B.1, including the forms of the Pólya-Gamma distributions and the mean and covariance matrix of the J-variate normal distribution. We choose the Pólya-Gamma sampler because it is very easy to construct and computationally efficient. It is much easier to implement than Metropolis-Hastings samplers which often need tuning to achieve good performance.

We comment on the computational complexity of the sampling of $\boldsymbol{\beta}_j$. The sampling from the Pólya-Gamma distribution has a complexity $O(NJ)$, and the sampling from the J-variate normal distribution has a complexity of $O(NJ^2) + O(J^3)$. Consequently, a loop of all the $\boldsymbol{\beta}_j, j = 1, ..., J$, has a complexity of $O((N + J)J^3)$.

Sampling S. Since $\mathbf{S}$ is a symmetric matrix, we reparametrize it by vectorizing its off-diagonal entries (including the diagonal entries). Specifically, the reparameterization is done by half-vectorization, denoted by $\boldsymbol{\alpha} = \text{vech}(\mathbf{S}) = (s_{11}, ..., s_{J1}, s_{22}, ..., s_{J2}, ..., s_{JJ})^\top \in \mathbb{R}^{J(J+1)/2}$. It is easy to see that $\text{vech}(\cdot)$ is a one-to-one mapping between $\mathbb{R}^{J(J+1)/2}$ and $J \times J$ symmetric matrices. Therefore, we impose a prior distribution on $\boldsymbol{\alpha}$ and sample $\boldsymbol{\alpha}^{(t)}$ in the tth iteration when $\mathbf{S}$ is sampled. Then we let $\mathbf{S}^{(t)} = \text{vech}^{-1}(\boldsymbol{\alpha}^{(t)})$.

Recall that a thinning step is performed, and t_0 is the gap between two samples of $\mathbf{S}$. Let t be a multiple of t_0 and $\boldsymbol{\alpha}^{(t-t_0)} = \text{vech}(\mathbf{S}^{(t-t_0)})$ be previous value of $\boldsymbol{\alpha}$. The sampling of $\boldsymbol{\alpha}^{(t)}$ is also achieved by a Pólya-Gamma sampler, which involves the following two steps similar to the sampling of $\boldsymbol{\beta}_j$.

1. Given $\boldsymbol{\alpha}^{(t-t_0)}$, independently sample NJ augmentation variables, each from a Pólya-Gamma distribution.

2. Given the NJ augmentation variables, sample $\boldsymbol{\alpha}^{(t)}$ from a $J(J+1)/2$-variate normal distribution.

The details of these two steps are given in Appendix B.2. We note that the computational complexity of sampling the NJ augmentation variables is $O(NJ^2)$, and that of sampling $\boldsymbol{\alpha}^{(t)}$ is $O(NJ^5) + O(J^6)$, resulting in an overall complexity $O((N+J)J^5)$. Comparing the complexities of the imputation and sampling $\mathbf{S}$ steps, we notice that the latter is computationally much more intensive. This is the reason why we choose to impute data by introducing auxiliary parameters $\boldsymbol{\beta_j}$s rather than using Ising network parameters $\mathbf{S}$ so that the iterative imputation mixes much faster in terms of the computation time. In addition, we only sample $\mathbf{S}$ every t_0 iterations for a reasonably large t_0 to avoid a high computational cost and also reduce the auto-correlation between the imputed data.

We remark that Marsman et al. (2022) considered a similar Ising network analysis problem based on fully observed data, in which they proposed a Bayesian inference approach based on a spike-and-slab prior to learning $\mathbf{S}$. Their Bayesian inference is also based on a Pólya-Gamma sampler. However, they combined Gibbs sampling with a Pólya-Gamma sampler, updating one parameter in $\mathbf{S}$ at a time. This Gibbs scheme often mixes slower than the joint update of $\mathbf{S}$ as in the proposed method and, thus, is computationally less efficient. The proposed Pólya-Gamma sampler may be integrated into the framework of Marsman et al. (2022) to improve computational efficiency.

3 Numerical Experiments

We illustrate the proposed method and show its power via two simulation studies and a real-world data application.

3.1 Simulation

Study I: A six-node case. We generate data from an Ising model with $J = 6$ variables. Missing values are generated under an MAR setting that is not MCAR. The proposed method is then compared with Bayesian inference based on (1) listwise deletion and (2) a single

imputation, where the single imputation is based on the imputed data from the Tth iteration of Algorithm 1, recalling that T_0 is the burn-in size.

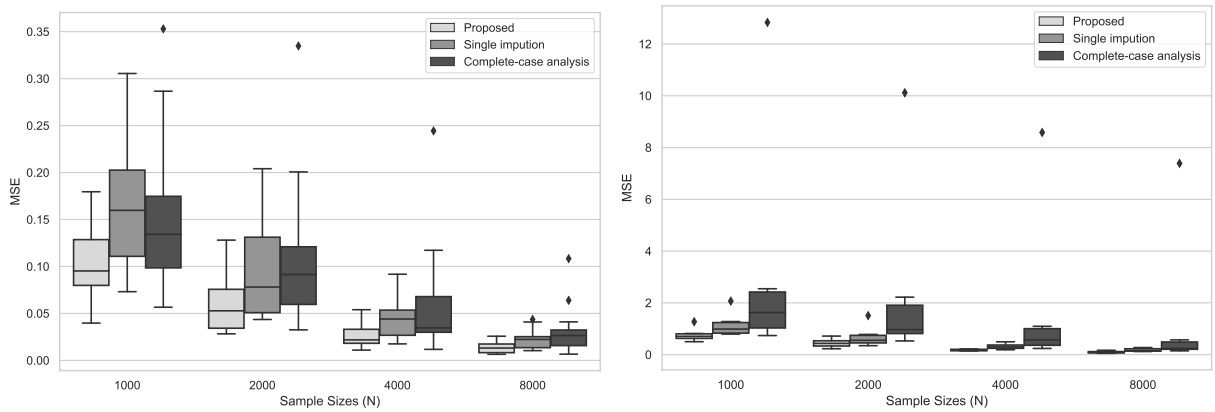
We configure the true parameter matrix $\mathbf{S}_0$ as follows. Since $\mathbf{S}_0$ is a symmetric matrix, we only need to specify its upper triangular matrix and then the diagonal entries. For the upper triangular entries (i.e., $s_{jl},\ j < l$), we randomly assign 50% of them to zero to introduce a moderately sparse setting. In addition, the nonzero parameters are then generated by sampling from a uniform distribution over the set $[-1, -0.4] \cup [0.4, 1]$. The intercept parameters $s_{jj}, j = 1, \ldots, J$ are set to zero. The true parameter values are given in Appendix C.1. Missing data are simulated by masking particular elements under an MAR mechanism. In particular, we have $z_{i6} = 1$, so that the sixth variable is always observed. We further allow the missingness probabilities of the first five variables (i.e., $z_{ij} = 0, j = 1, \ldots, 5$) to depend on the values of y_{i6}. The specific settings on $p(z_{ij} = 0 \mid y_{i6}), j = 1, \ldots, 5$ are detailed in Appendix C.2. Data are generated following the aforementioned Ising model and MAR mechanism for four different sample sizes, $N = 1,000, 2,000, 4,000$, and $8,000$, respectively. For each sample size, 50 independent replications are created.

Three methods are compared – the proposed method, Bayesian inference with a single imputation, and Bayesian inference based on complete cases from listwise deletion. The Bayesian inference for complete data is performed by sampling parameters from the posterior implied by the pseudo-likelihood and a normal prior, which is a special case of the proposed method without iterative imputation steps. All these methods shared the same initial values $s_{jl}^{(0)} \sim U(-0.1, 0.1), 1 \leqslant j \leqslant l \leqslant J$. For our proposed method, we set the length of the Markov Chain Monte Carlo (MCMC) iterations to $T = 5,000$ and a burn-in size of $T_0 = 1,000$. This setup leads to an effective total of 400 MCMC samples for the Ising parameter matrix $\mathbf{S}$. Notably, identical MCMC length and burn-in configuration are applied during parameters inference in the single imputation and complete-case analyses.

Figure 2 gives the plots for the mean squared errors (MSE) of the estimated edge parameters and intercept parameters under different sample sizes and for different methods. The MSE for each parameter s_{jl} is defined as

$$\frac{1}{50} \sum_{k=1}^{50} (\hat{s}_{k,jl} - s_{0,jl})^2. \tag{9}$$

Here, $\hat{s}_{k,jl}$ denotes the estimated value from the kth replication while $s_{0,jl}$ refers to the true

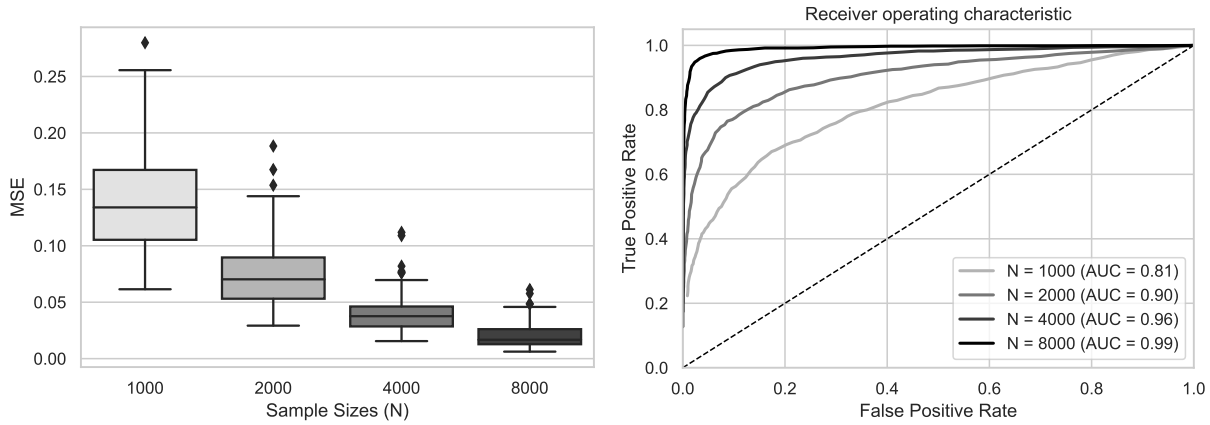


(a) (b)

Figure 2: Panel (a): Boxplots of MSEs for edge parameters s_{jl}. Panel (b): Boxplots of MSEs for intercept parameters s_{jj}.

value. Each box in panel (a) corresponds to the 15 edge parameters, and each box in panel (b) corresponds to the 6 intercept parameters. We notice that the listwise deletion procedure introduces biases into the edge and intercept estimation, resulting in the MSEs for certain parameters not decaying toward zero as the sample size grows. Additionally, both the proposed method and the single imputation method offer accurate parameter estimation, with MSEs decaying toward zero as the sample size increases. Notably, the proposed method is substantially more accurate than the single imputation method, suggesting that aggregating over multiple imputed datasets improves the estimation accuracy. Furthermore, for smaller sample sizes, the complete-case analysis seems to yield slightly more accurate estimates of the edge parameters than the single imputation method.

A fifteen-node example. We further simulate a fifteen-node Ising model to demonstrate the performance of the proposed method in terms of parameter estimation and edge selection. Similar to the six-node scenario, we generate model parameters by randomly setting 70% of the edge parameters s_{jl} to zero to create a sparse network. The true parameter values can be found in Appendix C.3. To simulate missing data, we implement the MCAR mechanism and randomly label 50% of the data entries as missing. Data are generated for four sample sizes of N = 1,000, 2,000, 4,000, and 8,000, following the specified Ising model parameters and MCAR mechanism. For each sample size, 50 independent replications are generated. Algorithm 1 is applied to these datasets, where a random starting point is used as in the six-node example. We set the MCMC iterations length to T = 5,000 and a burn-in size T_0 =



(a) (b)

Figure 3: Panel (a): Boxplots of MSEs for edge parameters. Panel (b): The ROC curves for edge selection and the corresponding AUCs.

1,000.

The resulting MSEs for edge parameter estimation under various sample sizes are displayed in Figure 3(a), where each box corresponds to the MSEs for 105 edge parameters. As we can see, the MSEs decrease toward zero as the sample size increases. Furthermore, by employing a hard thresholding step after edge parameter estimation, a receiver operating characteristic (ROC) curve is created for edge selection under each setting, and the corresponding Area Under the Curve (AUC) is calculated to evaluate the performance of the proposed method. That is, given a hard threshold τ, the True Positive Rate (TPR) and False Positive Rate (FPR) are calculated as

$$\mathrm{TPR}(\tau) = \frac{\sum_{k=1}^{50}\sum_{j<l}\mathbf{1}_{\{|\hat{s}_{k,jl}|>\tau \text{ and } s_{0,jl}\neq 0\}}}{50\times\sum_{j<l}\mathbf{1}_{\{s_{0,jl}\neq 0\}}}, \quad \mathrm{FPR}(\tau) = \frac{\sum_{k=1}^{50}\sum_{j<l}\mathbf{1}_{\{|\hat{s}_{k,jl}|>\tau \text{ and } s_{0,jl}= 0\}}}{50\times\sum_{j<l}\mathbf{1}_{\{s_{0,jl}= 0\}}}. \quad (10)$$

A ROC curve is obtained by varying the value of τ. The ROC curves and the corresponding AUC values are given in Figure 3(b).

3.2 A Real Data Application

We analyze the dataset for the 2001-2002 National Epidemiological Survey of Alcohol and Related Conditions (NESARC), which offers valuable insights into alcohol consumption and associated issues in the U.S. population (Grant et al., 2003). The dataset consists of 43,093 civilian non-institutionalized individuals aged 18 and older. In this analysis, we focus on

two specific sections of the survey that concern two highly prevalent mental health disorders – Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD). Because MDD and GAD have high symptom overlap (Hettema, 2008) and often co-occur (Hasin et al., 2005), it is important to perform a joint analysis of the symptoms of the two mental health disorders and study their separation. In particular, Blanco et al. (2014) performed factor analysis based on the same data and found that the two mental health disorders have distinct latent structures. We reanalyze the data, hoping to gain some insights from the network perspective of the two mental health disorders.

Following Blanco et al. (2014), we consider data with nine items measuring MDD and six items measuring GAD. These items are designed according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (American Psychiatric Association, 2000). These items ask the participants if they have recently experienced certain symptoms; see Table 1 for their short descriptions. After eliminating samples with entirely absent values across the 15 items, a total of 42,230 cases remain in the dataset. Note that any "Unknown" responses in the original data are converted into missing values. The dataset exhibits a significant degree of missingness, with only 2,412 complete cases for the 15 items, representing approximately 6% of the total cases. Specifically, the missing rate for each item is given in Table 1. Importantly, items D1 and D2 function as screening items and, thus, have a very low missing rate. The respondents did not need to answer items D3-D9 if the responses to D1 and D2 were "No" or "Unknown", resulting in high missing rates for these items. This pattern suggests that the missing data in this study is not MCAR. The GAD items A1-A6 also have a screening item, which results in the high missing rates in A1 through A6. Following the treatment in Blanco et al. (2014), these screening items are not included in the current analysis.

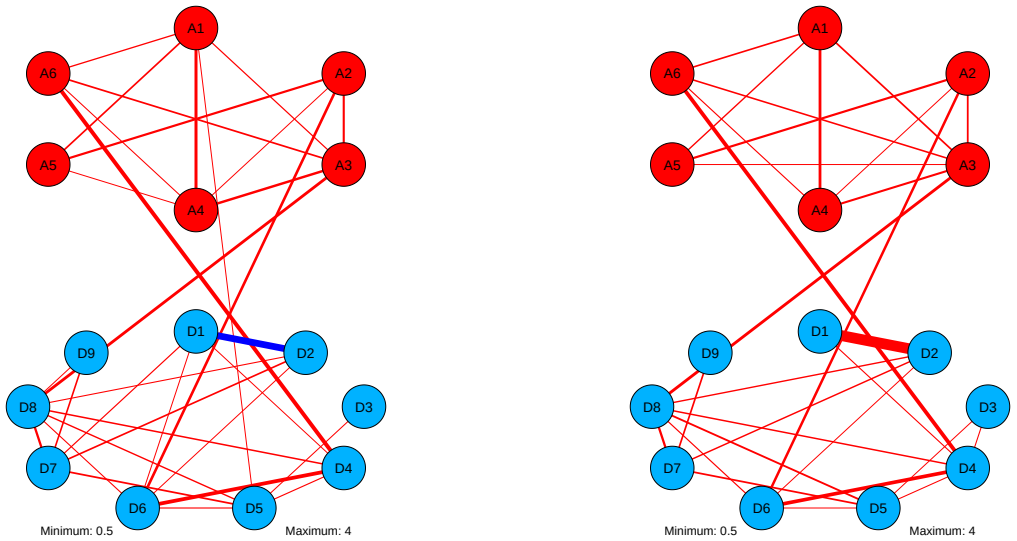
We apply the proposed method and the complete-case analysis to the data. For each method, 10 MCMC chains with random starting values are used, each having 10,000 MCMC iterations and a burn-in size 5,000. The Gelman-Rubin statistics are always below 1.018, confirming the satisfactory convergence of all 120 parameters for both methods. The estimated network structures for MDD and GAD items are presented in Figure 4, where an edge is shown between two variables when the estimated parameter has an absolute value greater than the hard threshold of 0.5. This hard threshold is chosen to ensure a clear visualization of the network. The nine MDD items are shown as blue nodes at the bottom, and the six

MDD Item Description	
D1 (0.1%): Depressed mood	D5 (68.5%): Psychomotor agitation/retardation
D2 (0.2%): Diminished interest	D6 (68.0%): Fatigue/loss of energy
D3 (68.5%): Weight loss or gain	D7 (67.9%): Feelings of guilt
D4 (67.9%): Insomnia or hypersomnia	D8 (67.9%): Diminished concentration
	D9 (67.7%): Recurrent thoughts of death
GAD Item Description	
A1 (91.8%): Restlessness	A4 (91.8%): Irritability
A2 (91.9%): Easily fatigued	A5 (91.9%): Muscle tension
A3 (91.8%): Difficulty concentrating	A6 (91.8%): Sleep disturbance

Table 1: Descriptions of MDD and GAD items and their missing rates.

GAD items are represented shown as red nodes at the top. The edges are colored red and blue, which represent positive and negative parameter estimates, respectively. In addition, the line thickness of the edges indicates their magnitude. A clear difference between the two methods is the edge between D1 "depressed mood most of the day, nearly every day," and D2 "markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day", which are two screening questions in the survey that all the participants responded to. The estimated parameter for this edge has a large absolute value under each of the two methods, but the estimated parameter is negative in the complete-case analysis, while it is positive according to the proposed method. Given that data are not MCAR and considering the content of the two items, we believe that the estimate from the proposed method is more sensible. Furthermore, we also see that the complete-case analysis yields more edges than the proposed method; for example, the edges of A4-A5, A1-D5, D1-D6, D1-D7, D3-D4, and D8-D9 appear in the estimated network from the complete-case analysis but not in that of the proposed method. They are likely false positives due to the higher estimation variance of the complete-case analysis, where the high variance is due to the relatively small sample size.

Finally, our analysis shows that the symptoms of each mental health disorder tend to densely connect with each other in the Ising network, while the symptoms are only loosely but positively connected between the two mental health disorders. The edges between the two mental health disorders identify the overlapping symptoms, including "D4: Insomnia or hypersomnia" and "A6: Sleep disturbance", "A2: Easily fatigued" and "D6: Fatigue/loss



(a) (b)

Figure 4: Estimated network structure for MDD and GAD. Panel (a): Complete-case analysis. Panel (b) Proposed method.

of energy", and "A3: Difficulty concentrating" and "D8: Diminished concentration". These results suggest that MDD and GAD are two well-separated mental health disorders, despite their high symptom overlap and frequent co-occurrence. This result confirms the conclusion of Blanco et al. (2014) that GAD and MDD are closely related but different nosological entities.

4 Concluding Remarks

In this paper, we propose a new method for Ising network analysis in the presence of missing data. The proposed method integrates iterative imputation into a Bayesian inference procedure based on conditional likelihood. An asymptotic theory is established that guarantees the consistency of the proposed estimator. Furthermore, a Pólya-Gamma machinery is proposed for the sampling of Ising model parameters, which yields efficient computation. The power of the proposed method is further shown via simulations and a real-data application. An R package has been developed that will be made publicly available upon the acceptance of the paper.

The current work has several limitations that require future theoretical and method-

ological developments. First, we did not investigate using sparsity-inducing priors to better explore the Ising network structure when it is sparse. We believe that the proposed method, including the iterative imputation and the Pólya-Gamma machinery, can be adapted when we replace the normal prior with the spike-and-slab prior considered in Marsman et al. (2022). This adaptation can be done by adding some Gibbs sampling steps. Second, from the frequentist perspective, the asymptotic normality of the Bayesian estimator remains to be established through a result in a similar form as the Bernstein-von-Mises theorem. This analysis is challenging and is left for future investigation. Finally, the computational cost for the Ising network analysis is still quite high, as discussed in Section 2.4. When sparsity-inducing priors are incorporated, more scalable algorithms may be developed, following recent advances in high-dimensional Bayesian model selection (e.g., Narisetty et al., 2018).

Acknowledgement

The research was supported in part by the Shanghai Science and Technology Committee Rising-Star Program (22YF1411100).

Appendix

A Technical proofs

A.1 A lemma for imputation consistency

Following the derivation of Section 2.3 in the main text, under ignorable missingness assumption, the posterior distribution for $\mathbf{S}$ satisfies $\pi_N(\mathbf{S}) \propto p(\boldsymbol{y}_{obs} \mid \mathbf{S})\pi(\mathbf{S})$. Under the same Bayesian model, one can impute the missing values from the posterior predictive distribution. That is, the posterior predictive distribution for $\boldsymbol{y}_{i,mis}$, $i = 1, ..., N$, takes the form

$$p_N(\boldsymbol{y}_{1,mis}, ..., \boldsymbol{y}_{N,mis}) = \int \pi_N(\mathbf{S}) \left(\prod_{i=1}^{N} \frac{\exp\left(\boldsymbol{y}_i^\top \mathbf{S} \boldsymbol{y}_i / 2\right)}{(c(\mathbf{S}) p_i(\boldsymbol{y}_{i,obs} \mid \mathbf{S}))} \right) d\mathbf{S},$$

where $p_i(\boldsymbol{y}_{i,obs} \mid \mathbf{S}) = \sum_{y_{ij}: z_{ij}=0} \exp\left(\boldsymbol{y}_i^\top \mathbf{S} \boldsymbol{y}_i / 2\right) / c(\mathbf{S})$. Further, suppose that the Algorithm 1 converges to a stationary distribution, and let $p_N^*(\boldsymbol{y}_{1,mis}, ..., \boldsymbol{y}_{N,mis})$ be the implied posterior predictive distribution given the observed data. Then we show in Lemma S1, which is an adaptation of Theorem 1 of Liu et al. (2014), suggests that $p_N(\boldsymbol{y}_{1,mis}, ..., \boldsymbol{y}_{N,mis})$ and $p_N^*(\boldsymbol{y}_{1,mis}, ..., \boldsymbol{y}_{N,mis})$ converge to each other in the total variation sense.

Lemma S1. *Assume the following assumptions hold: 1) The Markov chain for missing data, generated by the iterative imputation algorithm Algorithm 1, is positive Harris recurrent and thus admits a unique stationary distribution denoted by p_N^*; 2) The missing data process is ignorable; 3) A regularity condition holds for prior distributions of Ising model parameters and auxiliary parameters, as detailed in Assumption 1. Then the implied posterior predictive distribution p_N^* is consistent with the true posterior predictive distribution, p_N, i.e.,*

$$d_{TV}\left(p_N^*, p_N\right) = \max_{\boldsymbol{y}_{1,mis}, ..., \boldsymbol{y}_{N,mis}} |p_N^*(\boldsymbol{y}_{1,mis}, ..., \boldsymbol{y}_{N,mis}) - p_N(\boldsymbol{y}_{1,mis}, ..., \boldsymbol{y}_{N,mis})| \to 0, \qquad \text{(S1)}$$

in probability as $N \to \infty$.

To prove Lemma S1, we start by define a Gibbs sampling process for the joint Ising model, as outlined in Algorithm 2. This algorithm is constructed for the theoretical purposes since the step of sampling $\mathbf{S}$ is intractable. The aim of our proof is to validate the posterior predictive distribution of missing data given observed data p_N^*, implied by the Algorithm 1,

Algorithm 2: Ising model Gibbs chain

Data: Given observed data, initial values for the Ising model parameters $\mathbf{S}^{(0)}$, randomly imputed missing data $\boldsymbol{y}_1^{(0)}, \ldots, \boldsymbol{y}_N^{(0)}$, MCMC length T, burn-in size T_0.

for *iteration* t *in range* 1 *to* T **do**

  for *each* $j = 1$ *to* J **do**

    Sample $\mathbf{S}^{(t)}$ from $p(\mathbf{S} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) \propto \pi(\mathbf{S}) \prod_{i=1}^{N} \exp(\boldsymbol{y}_i^\top \mathbf{S} \boldsymbol{y}_i / 2) / c(\mathbf{S})$.

    Sample $y_{ij}^{(t)}$ for each $i \in \Omega_j$ from the Bernoulli distribution $p(y_{ij} \mid \boldsymbol{y}_{i,-j}, \mathbf{S}^{(t)})$ and update $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ with the sampled values.

  end

end

Result: $\hat{\mathbf{S}} = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \mathbf{S}^{(t)}$.

converges in total variation to the true posterior predictive distribution p_N. We first establish that p_N^* in fact converges to the stationary distribution of the Gibbs chain, denoted as p'_N, implied by Algorithm 2. By corroborating the convergence of p'_N and p_N, the proof of Lemma S1 is thereby completed. In the following proof, we reparameterize $\boldsymbol{\alpha} = \text{vech}(\mathbf{S})$ for convenience. We define the following for the proof.

Definition S1.

- *We denotes $\mathcal{Y}$ the data matrix with N samples and J variables, $\mathcal{Y}_j$ the jth column and $\mathcal{Y}_{-j}$ the remaining $j-1$ columns.*

- *Define $A_N = \{\mathcal{Y} \mid \|\hat{\boldsymbol{\alpha}}(\mathcal{Y})\| \leqslant \gamma\}$, where $\hat{\boldsymbol{\alpha}}(\mathcal{Y})$ is the complete-data maximum likelihood estimator, where γ can be sufficiently large so that*

$$\begin{aligned} &p_N^*(A_N) \to 1, \text{ and} \\ &p'_N(A_N) \to 1, \end{aligned} \tag{S2}$$

  in probability as $N \to \infty$.

- *Let*

$$K(\omega, d\omega') = pr(\mathcal{Y}_{mis}^{(k+1)} \in d\omega' \mid \mathcal{Y}_{mis}^{(k)} = \omega) \tag{S3}$$

  be the transition kernels for the missing data chain, which depend on $\mathcal{Y}_{obs}$.

- *Let $K^*(\omega, d\omega')$ and $K'(\omega, d\omega')$ be the transition kernels for the missing data chains from Algorithm 1 and Algorithm 2, respectively.*

- *We further define the transition kernels conditional on A_N by*

$$\tilde{K}(\omega, B) = \frac{K(\omega, B \cap A_N)}{K(\omega, A_N)}. \tag{S4}$$

So we have $\tilde{K}^*(\omega, \cdot), \tilde{K}'(\omega, \cdot)$ are two transition kernels for the missing data chains conditional on A_N. And let $\tilde{p}_N^*$, $\tilde{p}_N'$ be their stationary distributions, respectively.

- *Define $\|\mu\|_1 = \sup_{|h| \leqslant 1} \int h(x)\mu(dx)$.*

Assumption 1 (A regularity condition for priors)**.** *Let $\hat{\boldsymbol{\alpha}}(\mathcal{Y})$ be the complete data maximum likelihood estimator and $A_N = \{\mathcal{Y} : \|\hat{\boldsymbol{\alpha}}(\mathcal{Y})\| \leqslant \gamma\}$. Since the logistic models are with Ising model, we also have map $\boldsymbol{\beta}_j = T_j(\boldsymbol{\alpha}), j = 1, \ldots, J$. Let $\pi_j(\boldsymbol{\beta}_j)$ and $\pi(\boldsymbol{\alpha})$ be prior distributions. Further define $\boldsymbol{\beta}_j^* = T_j^*(\boldsymbol{\alpha})$ such that $\tilde{T}_j(\boldsymbol{\alpha}) = \{T_j(\boldsymbol{\alpha}), T_j^*(\boldsymbol{\alpha})\}$ is a one-to-one invertible map ($\boldsymbol{\beta}_j^*$ can be $\boldsymbol{\alpha} \backslash \boldsymbol{\beta}_j$). Define*

$$\pi_j^*(\boldsymbol{\beta}_j, \boldsymbol{\beta}_j^*) = \det(\partial \tilde{T}_j / \partial \boldsymbol{\alpha})^{-1} \pi(\tilde{T}_j^{-1}(\boldsymbol{\beta}_j, \boldsymbol{\beta}_j^*)).$$

Let $L_j(\boldsymbol{\beta}_j) = \pi_j(\boldsymbol{\beta}_j)/\pi_{j,\mathcal{Y}_{-j}}(\boldsymbol{\beta}_j)$, where

$$\begin{aligned} \pi_{j,\mathcal{Y}_{-j}}(\boldsymbol{\beta}_j) &= \int p(\mathcal{Y}_{-j} \mid \boldsymbol{\beta}_j, \boldsymbol{\beta}_j^*) \pi_j^*(\boldsymbol{\beta}_j, \boldsymbol{\beta}_j^*) d\boldsymbol{\beta}_j^* \\ &= \int \sum_{y_{1j}, \ldots, y_{Nj}} p(\mathcal{Y}_j, \mathcal{Y}_{-j} \mid \boldsymbol{\beta}_j, \boldsymbol{\beta}_j^*) \pi_j^*(\boldsymbol{\beta}_j, \boldsymbol{\beta}_j^*) d\boldsymbol{\beta}_j^*. \end{aligned} \tag{S5}$$

The assumption requires that on the set A_N,

$$\sup_{\|\boldsymbol{\beta}_j\| < \gamma} \partial \log L_j(\boldsymbol{\beta}_j) < \infty.$$

We remark that the above assumption holds for the Ising model with the normal priors adopted in the current paper. Specifically, $\pi_j(\boldsymbol{\beta}_j)$ and $\pi(\boldsymbol{\alpha})$ are J-variate and $J(J+1)/2$-variate normal distributions, respectively. Moreover, π_j^* can also be a $J(J+1)/2$ normal distribution. Since on A_N, $L_j(\boldsymbol{\beta}_j)$ is a continuously differentiable function defined in $\mathbb{R}^{J(J+1)/2}$, we have then it is Lipschitz on any compact set in $\mathbb{R}^{J(J+1)/2}$. That is, on A_N, $\sup_{\|\boldsymbol{\beta}_j\| < \gamma} \partial \log L_j(\boldsymbol{\beta}_j) = \sup_{\|\boldsymbol{\beta}_j\| < \gamma} \left[\partial \log \pi_j(\boldsymbol{\beta}_j) - \partial \log \pi_{j,\mathcal{Y}_{-j}}(\boldsymbol{\beta}_j)\right] < \infty$.

Proof of Lemma S1. According to the assumptions, the Markov chain for the missing data

produced by the Gibbs sampling procedure Algorithm 2 is positive Harris recurrent and thus admit a unique stationary distribution p'_N. We verify the conditions holds. First, on A_N, the Fisher information of the Ising model has a lower bound of ϵn for some ϵ. So according to proposition 1 of Liu et al. (2014), we have $\|K^*(\omega,\cdot) - K'(\omega,\cdot)\|_1 \to 0$ uniformly for $\omega \in A_N$, that is,

$$\lim_{N\to\infty} \|\tilde{K}^*(\omega,\cdot) - \tilde{K}'(\omega,\cdot)\|_1 = 0. \tag{S6}$$

According to the standard bound for Markov chain convergence rates, there exists a common starting value $\omega \in C$ and a bound r_k such that (ii) of Lemma 2 holds. Then Lemma 2 gives us

$$d_{TV}(\tilde{p}^*_N, \tilde{p}'_N) \to 0, \tag{S7}$$

Further combining with conclusions in Lemma 1 in Liu et al. (2014) that $d_{TV}(p^*_N, \tilde{p}^*_N) \to 0$, and $d_{TV}(p'_N, \tilde{p}'_N) \to 0$, we have the convergence of iterative imputation of compatible models,

$$d_{TV}(p^*_N, p'_N) \to 0, \tag{S8}$$

in probability as $N \to \infty$. Next, based on the construction of the Gibbs sampling procedure Algorithm 2, we have the sequence converges to the target distribution, that is,

$$d_{TV}(p'_N, p_N) \to 0, \tag{S9}$$

in probability as $N \to \infty$. Based on (S8) and (S9), we have

$$\begin{aligned} d_{TV}(p^*_N, p_N) &= \sup_{A\in\mathcal{F}} |p^*_N(A) - p_N(A)| \\ &\leqslant \sup_{A\in\mathcal{F}} |p^*_N(A) - p'_N(A)| + \sup_{A\in\mathcal{F}} |p'_N(A) - p_N(A)| \to 0, \end{aligned} \tag{S10}$$

in probability as $N \to \infty$. □

Remark 1. *Lemma S1 emphasizes the consistency of the proposed iterative imputation process. It implies that the implied posterior predictive distribution gradually converges to the posterior predictive distribution under standard Bayesian inference, underscoring the validity of the iterative imputation.*

A.2 Proof of Theorem 1

We will use $\boldsymbol{\alpha}$, the half-vectorization of $\mathbf{S}$ in the proof. Denote $\pi_N^*(\boldsymbol{\alpha})$ the posterior density of $\boldsymbol{\alpha}$ implied by the stationary distribution of the proposed method. Let $\mathcal{Y}$ be the data matrix with $\mathcal{Y}_{mis}$ and $\mathcal{Y}_{obs}$ being the missing and observed parts, respectively. We have

$$\begin{aligned}
&\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \pi_N^*(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \\
&= \int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \left[\sum_{\mathcal{Y}_{mis}} p^*(\mathcal{Y}_{mis}, \mathcal{Y}_{obs} \mid \boldsymbol{\alpha}) p^*(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs}) \right] \pi(\boldsymbol{\alpha}) / c_N d\boldsymbol{\alpha} \\
&= \sum_{\mathcal{Y}_{mis}} \left[\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} p^*(\mathcal{Y}_{mis}, \mathcal{Y}_{obs} \mid \boldsymbol{\alpha}) \pi(\boldsymbol{\alpha}) / c_N d\boldsymbol{\alpha} \right] p^*(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs}) \\
&= \sum_{\mathcal{Y}_{mis}} \left[\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \exp(-N f_N(\boldsymbol{\alpha})) \pi(\boldsymbol{\alpha}) / c_N d\boldsymbol{\alpha} \right] p^*(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs}),
\end{aligned} \tag{S11}$$

where $f_N(\boldsymbol{\alpha}) = -\frac{1}{N} \sum_{i=1}^N \log p^*(\boldsymbol{y}_i \mid \boldsymbol{\alpha})$ given in (2), $c_N = \int \exp(-N f_N(\boldsymbol{\alpha})) \pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha}$. Further let $\pi_N(\boldsymbol{\alpha}) = \sum_{\mathcal{Y}_{mis}} \left[\exp(-N f_N(\boldsymbol{\alpha})) \pi(\boldsymbol{\alpha}) / c_N \right] p(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs})$, we have

$$\begin{aligned}
&\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \pi_N(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \\
&= \sum_{\mathcal{Y}_{mis}} \left[\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \exp(-N f_N(\boldsymbol{\alpha})) \pi(\boldsymbol{\alpha}) / c_N d\boldsymbol{\alpha} \right] p(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs}).
\end{aligned} \tag{S12}$$

Let $\Theta \in \mathbb{R}^{J(J+1)/2}, E \subset \Theta$ be open and bounded. It can be veried that: 1) f_N have continuous third derivatives; 2) $f_N \to f$ pointwise for some f; 3) $f''(\boldsymbol{\alpha}_0)$ is positive definite; 4) $f'''(\boldsymbol{\alpha}_0)$ is uniformly bounded on E; 5) each f_N is convex and $f'(\boldsymbol{\alpha}_0) = 0$. Then, according to the generalized posterior concentration theorem (see Theorem 5, Miller, 2021), we have for any $\varepsilon > 0$,

$$\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \exp(-N f_N(\boldsymbol{\alpha})) \pi(\boldsymbol{\alpha}) / c_N d\boldsymbol{\alpha} \to 1 \tag{S13}$$

in probability as $N \to \infty$. Consequently,

$$\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \pi_N(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \to 1, \tag{S14}$$

in probability as $N \to \infty$. Finally, by employing the convergence of imputation from Lemma S1,

specifically

$$d_{TV}(p^*(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs}) - p(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs})) \to 0$$

in probability as $N \to \infty$, we arrive at

$$\sum_{\mathcal{Y}_{mis}} \left[\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \exp(-N f_N(\boldsymbol{\alpha})) \pi(\boldsymbol{\alpha}) / c_N d\boldsymbol{\alpha} \right] (p^*(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs}) - p(\mathcal{Y}_{mis} \mid \mathcal{Y}_{obs})) \to 0 \tag{S15}$$

in probability as $N \to \infty$. This conclude the proof, given that

$$\int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \pi_N^*(\boldsymbol{\alpha}) d\boldsymbol{\alpha} = \int_{B_\varepsilon(\boldsymbol{\alpha}_0)} \pi_N(\boldsymbol{\alpha}) d\boldsymbol{\alpha} + \int_{B_\varepsilon(\boldsymbol{\alpha}_0)} (\pi_N^*(\boldsymbol{\alpha}) - \pi_N(\boldsymbol{\alpha})) d\boldsymbol{\alpha}, \tag{S16}$$

where the first term converges to 1 (i.e., (S14)) and the second term converges to 0 (i.e., (S15)) in probability as $N \to \infty$.

B Computation details for sampling $\boldsymbol{\beta}_j$ and S

Upon observation of the logistic form presented in the conditional distribution when sampling the auxiliary parameters $\boldsymbol{\beta}_j$, we employ the Pólya-Gamma for effective sampling. Denote a random variable ω follows the Pólya-Gamma distribution $PG(b, c), b > 0, c \in \mathbb{R}$ with parameters $b > 0$ and $c \in \mathbb{R}$ if it is a weighted sum of independent Gamma random variables

$$\omega = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}, \tag{S17}$$

where $g_k \sim \Gamma(b, 1)$, which is the Gamma distribution with shape and rate parameters as b and 1, respectively.

B.1 Derivation of posterior distribution of $\boldsymbol{\beta}_j$

By introducing Pólya-Gamma latent variables $\omega_{ij} \sim PG(1, 0), i = 1, \ldots, N$, we establish a connection between the logistic form and the normal distribution. We rephrase the jth conditional distribution $p(y_{ij} \mid \boldsymbol{y}_{i,-j}, \boldsymbol{\beta}_j)$ by the following equation,

$$\frac{\exp(\phi_{ij})^{y_{ij}}}{1 + \exp(\phi_{ij})} = 2^{-1} \exp(\kappa_{ij} \phi_{ij}) \mathbb{E}_{\omega_{ij}} \left[\exp(-\omega_{ij} \phi_{ij}^2 / 2) \right], \tag{S18}$$

where $\kappa_{ij} = y_{ij} - 1/2$, $\omega_{ij} \sim PG(1,0)$, $\omega_{ij} \mid \phi_{ij} \sim PG(1,\phi_{ij})$, $\phi_{ij} = \beta_{jj}/2 + \sum_{k\neq j}\beta_{jk}y_{ik}$.

Denote $\mathcal{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^\top$, $\mathcal{Y}_j$ as the jth column of $\mathcal{Y}$, and $\mathcal{Y}_{-j}$ the remaining $j-1$ columns. Given $\boldsymbol{\beta}_j$, sample N augmentation variables $\omega_{ij}, i = 1, \ldots, N$, each from a Pólya-Gamma distribution

$$\omega_{ij} \mid \boldsymbol{\beta}_j, \boldsymbol{y}_i \sim PG(1, \beta_{jj}/2 + \sum_{k\neq j}\beta_{jk}y_{ik}), \tag{S19}$$

based on eq. (S18). Moreover, for the jth variable, we have

$$\begin{aligned} p(\mathcal{Y}_j \mid \mathcal{Y}_{-j}, \boldsymbol{\beta}_j, \boldsymbol{\omega}_j) &= \prod_{i=1}^{N} p(y_{ij} \mid \boldsymbol{y}_{i,-j}, \boldsymbol{\beta}_j, \omega_{ij}) \\ &= \prod_{i=1}^{N} 2^{-1}\exp(\kappa_{ij}(\beta_{jj}/2 + \sum_{k\neq j}\beta_{jk}y_{ik}))\exp(-\omega_{ij}(\beta_{jj}/2 + \sum_{k\neq j}\beta_{jk}y_{ik})^2/2) \\ &\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}_j^\top(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)^\top D_{\omega_j}(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)\boldsymbol{\beta}_j - 2\boldsymbol{\kappa}_j^\top(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)\boldsymbol{\beta}_j\right)\right], \end{aligned} \tag{S20}$$

where $\boldsymbol{\kappa}_j = (\kappa_{1j}, \ldots, \kappa_{Nj})^\top, \kappa_{ij} = y_{ij} - 1/2, D_{\omega_j} = \mathrm{diag}(\boldsymbol{\omega}_j)$. We further have the following conditional distribution for $\boldsymbol{\beta}_j$

$$\begin{aligned} p(\boldsymbol{\beta}_j \mid \mathcal{Y}, \boldsymbol{\omega}_j) &\propto p(\mathcal{Y}_j \mid \mathcal{Y}_{-j}, \boldsymbol{\beta}_j, \boldsymbol{\omega}_j)\pi_j(\boldsymbol{\beta}_j) \\ &\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta}_j^\top(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)^\top D_{\omega_j}(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)\boldsymbol{\beta}_j - 2\boldsymbol{\kappa}_j^\top(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)\boldsymbol{\beta}_j\right) - \frac{1}{2}\boldsymbol{\beta}_j^\top D_{\beta_j}\boldsymbol{\beta}_j\right] \\ &= \exp\left[-\frac{1}{2}(\boldsymbol{\beta}_j - \boldsymbol{\mu}_{\beta_j})^\top\Sigma_{\beta_j}^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\mu}_{\beta_j})\right], \end{aligned} \tag{S21}$$

where $\Sigma_{\beta_j} = \left[(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)^\top D_{\omega_j}(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top) + D_{\beta_j}\right]^{-1}$, $\boldsymbol{\mu}_{\beta_j} = \Sigma_{\beta_j}(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)^\top\boldsymbol{\kappa}_j$. Here, $\boldsymbol{e}_j$ is a J-dimensional vector with the jth element be one and all others be zeros, $D_{\omega_j} = \mathrm{diag}(\boldsymbol{\omega}_j)$, $D_{\beta_j} = \mathrm{diag}(\boldsymbol{\tau}_j)$, where $\tau_{jl} = \sigma_1^{-2}$, for $l \neq j$ and $\tau_{jj} = \sigma_2^{-2}$. A weak informative prior on the intercept parameter by letting $\sigma_2^2 > \sigma_1^2$.

To summarize, the introduced Pólya-Gamma latent variables ω_{ij} establish a connection between the logistic form and the normal distribution that lead to a normal form of the posterior $\boldsymbol{\beta}_j$, i.e.,

$$\begin{aligned} \boldsymbol{\beta}_j \mid \mathcal{Y}, \boldsymbol{\omega}_j &\sim N(\boldsymbol{\mu}_{\beta_j}, \Sigma_{\beta_j}), \\ \Sigma_{\beta_j} &= \left[(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)^\top D_{\omega_j}(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top) + D_{\beta_j}\right]^{-1}, \\ \boldsymbol{\mu}_{\beta_j} &= \Sigma_{\beta_j}(\mathcal{Y} - \boldsymbol{\kappa}_j\boldsymbol{e}_j^\top)^\top\boldsymbol{\kappa}_j. \end{aligned} \tag{S22}$$

B.2 Derivation of posterior distribution of S

Observe that there is constraint between edge parameters $s_{jk} = s_{jk}$ for $j \neq k$, which is reflected by the symmetry of matrix $\mathbf{S}$. It is sufficient to examine the lower triangular elements of $\mathbf{S}$ while adhering to the equality constraint. To impose such symmetric constraint on $\mathbf{S}$, we reparameterize $\mathbf{S}$ by $\boldsymbol{\alpha} = \text{vech}(\mathbf{S})$, which is the half-vectorization of $\mathbf{S}$. Specifically,

$$\boldsymbol{\alpha} = (s_{11}, \ldots, s_{J1}, s_{22}, \ldots, s_{J2}, \ldots, s_{J-1,J-1}, s_{J,J-1}, s_{J,J})^\top = (\alpha_1, \ldots, \alpha_{J(J+1)/2})^\top. \tag{S23}$$

To establish a relationship between $\mathbf{S}$ and $\boldsymbol{\alpha}$, we first define the following equation,

$$\boldsymbol{s}_j = E_j \text{vec}(\mathbf{S}). \tag{S24}$$

In this equation, $\text{vec}(\mathbf{S}) = (\boldsymbol{s}_1^\top, \ldots, \boldsymbol{s}_J^\top)^\top$ represents the vectorization of the matrix $\mathbf{S}$. The matrix $E_j = (0_J, \ldots, I_J, \ldots, 0_J)$ is a $J \times J^2$ matrix, where the jth row block is the identity matrix I_J and all other row blocks are zero matrices. Next, we can express $\text{vec}(\mathbf{S})$ as follows,

$$\text{vec}(\mathbf{S}) = D_J \boldsymbol{\alpha}, \tag{S25}$$

where D_J is a $J^2 \times J(J+1)/2$ duplication matrix, which can be explicitly defined as,

$$D_J^\top = \sum_{i \geqslant j} u_{ij} (\text{vec} T_{ij})^\top. \tag{S26}$$

Here, u_{ij} is a unit vector of order $J(J+1)/2$ with ones in the position $(j-1)J + i - j(j-1)/2$ and zeros elsewhere. The matrix T_{ij} is a $J \times J$ matrix with ones in position (i, j) and (j, i) and zeros in all other positions. By combining equations (S24) and (S25), we obtain,

$$\boldsymbol{s}_j = T_j \boldsymbol{\alpha}, \tag{S27}$$

where $T_j = E_j D_J$ is a $J \times J(J+1)/2$ transformation matrix.

Given $\boldsymbol{\alpha}$, we first sample NJ augmentation variables $\Omega = (\omega_{ij})_{N \times J}$ from

$$\omega_{ij} \mid \mathcal{Y}, \boldsymbol{\alpha} \sim PG\left(1, \sigma^2_{\omega,ij}\right),$$

where $\sigma^2_{\omega,ij}$ is the (i,j)th entry of $\Sigma_\omega = \mathcal{Y}\mathbf{S} - (\mathcal{Y} - \frac{1}{2}\mathbf{1}_N\mathbf{1}_J^\top) \circ \mathbf{1}_N \text{diag}(\mathbf{S})^\top$. $\mathbf{S} = \text{vech}^{-1}(\boldsymbol{\alpha})$, $\text{diag}(\mathbf{S})$ is the diagonal vector of the matrix $\mathbf{S}$, and $\circ$ is the Hadamard product. Furthermore, given the above transformation, the sampling of $\mathbf{S}$ can be done instead by sampling $\boldsymbol{\alpha}$ from its posterior with a similar Pólya-Gamma augmentation procedure. Specifically, the pseudo likelihood with Pólya-Gamma augmentation is

$$\begin{aligned} p^*(\mathcal{Y} \mid \mathbf{S}, \Omega) &= \prod_{j=1}^{J} p(\mathcal{Y}_j \mid \mathcal{Y}_{-j}, \boldsymbol{s}_j, \boldsymbol{\omega}_j) \\ &\propto \exp\left[-\frac{1}{2} \sum_{j=1}^{J} \left(\boldsymbol{s}_j^\top (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top)^\top D_{\omega_j} (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) \boldsymbol{s}_j - 2\boldsymbol{\kappa}_j^\top (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) \boldsymbol{s}_j \right) \right]. \end{aligned} \tag{S28}$$

Then we have the posterior of $\mathbf{S}$

$$\begin{aligned} &p(\mathbf{S} \mid \mathcal{Y}, \Omega) \propto p^*(\mathcal{Y} \mid \mathbf{S}, \Omega) \pi(\mathbf{S}) \\ &\propto \exp\left[-\frac{1}{2} \sum_{j=1}^{J} \left(\boldsymbol{\beta}_j^\top (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top)^\top D_{\omega_j} (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) \boldsymbol{\beta}_j - 2\boldsymbol{\kappa}_j^\top (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) \boldsymbol{\beta}_j \right) - \frac{1}{2} \sum_{j=1}^{J} \boldsymbol{\beta}_j^\top D_{s_j} \boldsymbol{\beta}_j \right] \\ &= \exp\left\{ -\frac{1}{2} \sum_{j=1}^{J} \left[\boldsymbol{\beta}_j^\top ((\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top)^\top D_{\omega_j} (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) + D_{s_j}) \boldsymbol{\beta}_j - 2\boldsymbol{\kappa}_j^\top (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) \boldsymbol{\beta}_j \right] \right\}. \end{aligned} \tag{S29}$$

Plugging (S27) into (S29) we have,

$$\begin{aligned} &p(\boldsymbol{\alpha} \mid \mathcal{Y}, \Omega) \\ &\quad \propto \exp\left\{ -\frac{1}{2} \sum_{j=1}^{J} \left[\boldsymbol{\alpha}^\top T_j^\top ((\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top)^\top D_{\omega_j} (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) + D_{s_j}) T_j \boldsymbol{\alpha} - 2\boldsymbol{\kappa}_j^\top (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) T_j \boldsymbol{\alpha} \right] \right\} \\ &\quad \propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\alpha}^\top \left(\sum_{j=1}^{J} T_j^\top ((\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top)^\top D_{\omega_j} (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) + D_{s_j}) T_j \right) \boldsymbol{\alpha} \right. \right. \\ &\qquad \left. \left. - 2 \left(\sum_{j=1}^{J} ((\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) T_j)^\top \boldsymbol{\kappa}_j \right)^\top \boldsymbol{\alpha} \right] \right\} \\ &\quad \propto \exp\left[-\frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha)^\top \Sigma_\alpha^{-1} (\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha) \right], \end{aligned} \tag{S30}$$

where

$$\Sigma_\alpha = \left[\sum_{j=1}^{J} T_j^\top ((\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top)^\top D_{\omega_j} (\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) + D_{s_j}) T_j\right]^{-1}, \boldsymbol{\mu}_\alpha = \Sigma_\alpha \left[\sum_{j=1}^{J} ((\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) T_j)^\top \boldsymbol{\kappa}_j\right], \tag{S31}$$

which can be further simplified as below.

In summary, the posterior of $\boldsymbol{\alpha}$ is

$$\begin{aligned} &\boldsymbol{\alpha} \mid \mathcal{Y}, D_\omega \sim N(\boldsymbol{\mu}_\alpha, \Sigma_\alpha), \\ &\Sigma_\alpha = \left[M^\top D_\omega M + T^\top D_S T\right]^{-1}, \\ &\boldsymbol{\mu}_\alpha = \Sigma_\alpha M^\top \boldsymbol{\kappa}, \end{aligned} \tag{S32}$$

where $\mathcal{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^\top$, $M = ([(\mathcal{Y} - \boldsymbol{\kappa}_1 \boldsymbol{e}_1^\top) T_1]^\top, \ldots, [(\mathcal{Y} - \boldsymbol{\kappa}_j \boldsymbol{e}_j^\top) T_J]^\top)^\top$, $D_\omega = \mathrm{diag}(\boldsymbol{\omega})$, $\boldsymbol{\omega} = (\omega_{11}, \ldots, \omega_{N1}, \omega_{12}, \ldots, \omega_{NJ})^\top$, $T = (T_1^\top, \ldots, T_J^\top)^\top$, $\boldsymbol{\kappa} = (\boldsymbol{\kappa}_1^\top, \ldots, \boldsymbol{\kappa}_J^\top)^\top$, and $D_S = \mathrm{diag}(\boldsymbol{\tau})$, where $\boldsymbol{\tau} = (\tau_{11}, \ldots, \tau_{J1}, \tau_{12}, \ldots, \tau_{JJ})^\top$, $\tau_{jl} = \sigma_1^{-2}$, for $l \neq j$ and $\tau_{jj} = \sigma_2^{-2}$.

Instead of conventional matrix inversion for calculating Σ_α, we use Cholesky decomposition of a symmetric positive definite (SPD) matrix, which offers enhanced efficiency and numerical stability (Blake, 2015). More precisely, we start by performing the Cholesky decomposition of $\Sigma_\alpha^{-1} = LL^\top$, and then proceed to solve two triangular systems: i) $LY = I$, and ii) $L^\top X = Y$, thus deriving $X = \Sigma_\alpha$.

C Detailed settings for simulations

C.1 True parameters used in the six-node example

S	node 1	node 2	node 3	node 4	node 5	node 6
node 1	0.00000	-0.73672	0.00000	0.00000	0.76924	0.79083
node 2	-0.73672	0.00000	-0.40826	0.00000	0.00000	0.63832
node 3	0.00000	-0.40826	0.00000	0.61893	0.00000	0.00000
node 4	0.00000	0.00000	0.61893	0.00000	0.00000	0.00000
node 5	0.76924	0.00000	0.00000	0.00000	0.00000	0.74086
node 6	0.79083	0.63832	0.00000	0.00000	0.74086	0.00000

Table 2: The true parameters used in the six-node Ising model simulation.

C.2 MAR settings in the six-node example

$p(z_{ij} = 0 \mid y_{i6})$	z_{i1}	z_{i2}	z_{i3}	z_{i4}	z_{i5}
$y_{i6} = 1$	0.3	0.8	0.3	0.7	0.2
$y_{i6} = 0$	0.8	0.5	0.7	0.4	0.9

Table 3: The probabilities of missingness for the first five variables conditioning on the sixth variable.

C.3 True parameters used in the fifteen-node example

S	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-														
2	-0.96	-													
3	-0.96	0.50	-												
4	1.00	0.50	-	-											
5	0.48	-	-	-	-										
6	0.95	-	0.95	-	0.47	-									
7	-	-	-	0.55	-0.92	-	-								
8	-	-	-	-	-	0.98	0.74	-							
9	-0.41	-	-0.54	-	-	-	-	-	-						
10	-0.47	-	-0.47	-	-	-	-0.83	-	-0.41	-					
11	-0.74	-	-0.74	0.52	-0.55	0.85	-	0.75	-0.98	-	-				
12	-0.54	-	0.77	-	-	0.41	-	-	-	-0.74	-	-			
13	-	-	-	-	-	-0.80	0.56	-	-	-0.78	-	-	-		
14	-	-	-	-	-	-	0.95	-	-	-	-	-	-0.96	-	
15	-0.97	-	-	-	-0.97	-	-	0.78	-	-	-	-	-	-	-

Table 4: The true parameters used in the 15-node Ising model simulation.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR* (4th, Text Revision ed.). Washington, DC: American Psychiatric Association.

Anderson, C. J., & Yu, H.-T. (2007). Log-multiplicative association models as item response models. *Psychometrika*, *72*(1), 5–23.

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans. *Journal of Anxiety Disorders*, *45*, 49–59.

Barndorff-Nielsen, O., Kent, J., & Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review/Revue Internationale de Statistique*, 145–159.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 192–225.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *24*(3), 179–195.

Blake, E. (2015). *Fast and accurate symmetric positive definite matrix inverse using Cholesky decomposition.* Retrieved from `https://uk.mathworks.com/matlabcentral/fileexchange/34511-fast-and-accurate-symmetric-positive-definite-matrix-inverse-using-cholesky-decomposition` (Accessed on June 11, 2023)

Blanco, C., Rubio, J. M., Wall, M., Secades-Villa, R., Beesdo-Baum, K., & Wang, S. (2014). The latent structure and comorbidity patterns of generalized anxiety disorder and major depressive disorder: a national study. *Depression and Anxiety*, *31*(3), 214–222.

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*(9), 1089–1108.

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., ... Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 58.

Brunson, J. C., & Laubenbacher, R. C. (2018). Applications of network analysis to routinely collected health care data: a systematic review. *Journal of the American Medical Informatics Association*, *25*(2), 210–221.

Burger, J., Isvoranu, A. M., Lunansky, G., Haslbeck, J. M. B., Epskamp, S., Hoekstra, R. H. A., ... Blanken, T. F. (2022). Reporting standards for psychological network analyses in cross-sectional data. *Psychological Methods*.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*(3), 551.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Robust measurement via a fused latent and graphical item response theory model. *Psychometrika*, *83*, 538–562.

Epskamp, S. (2020). Psychometric network models from time-series and panel data. *Psychometrika*, *85*(1), 206–231.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*, 195–212.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617.

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480.

Fried, E. I., van Borkulo, C. D., Cramer, A. O., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*, 1–10.

Fried, E. I., Von Stockert, S., Haslbeck, J., Lamers, F., Schoevers, R., & Penninx, B. (2020). Using network analysis to examine links between individual depressive symptoms, inflammatory markers, and covariates. *Psychological Medicine*, *50*(16), 2682–2690.

Gile, K. J., & Handcock, M. S. (2017). Analysis of networks with missing data with application to the national longitudinal study of adolescent health. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(3), 501.

Grant, B. F., Moore, T., Shepard, J., & Kaplan, K. (2003). Source and accuracy statement: Wave 1 national epidemiologic survey on alcohol and related conditions (NESARC). *Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism*, *52*.

Hasin, D. S., Goodwin, R. D., Stinson, F. S., & Grant, B. F. (2005). Epidemiology of major depressive disorder: results from the national epidemiologic survey on alcoholism and related conditions. *Archives of General Psychiatry*, *62*(10), 1097–1106.

Haslbeck, J. M., & Fried, E. I. (2017). How predictable are symptoms in psychopathological networks? A reanalysis of 18 published datasets. *Psychological Medicine*, *47*(16), 2767–2776.

Hettema, J. M. (2008). The nosologic relationship between generalized anxiety disorder and major depression. *Depression and Anxiety*, *25*(4), 300–316.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, *55*, 5–18.

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, *10*(1), 1–29.

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, *67*, 367–386.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, *31*(1), 253–258.

Kohler, K., Jankowski, M. D., Bashford, T., Goyal, D. G., Habermann, E. B., & Walker, L. E. (2022). Using network analysis to model the effects of the SARS Cov2 pandemic on acute patient care within a healthcare system. *Scientific Reports*, *12*(1), 10050.

Koponen, M., Asikainen, M. A., Viholainen, A., & Hirvonen, P. E. (2019). Using network analysis methods to investigate how future teachers conceptualize the links between the domains of teacher knowledge. *Teaching and Teacher Education*, *79*, 137–152.

Li, Y., Craig, B. A., & Bhadra, A. (2019). The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, *28*(3), 747–757.

Lin, S.-Y., Fried, E. I., & Eaton, N. R. (2020). The association of life stress with substance use symptoms: A network analysis and replication. *Journal of Abnormal Psychology*, *129*(2), 204.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., & Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, *101*(1), 155–173.

Luke, D. A., & Harris, J. K. (2007). Network analysis in public health: history, methods, and applications. *Annual Review of Public Health*, *28*, 69–93.

Marsman, M., Huth, K., Waldorp, L., & Ntzoufras, I. (2022). Objective Bayesian edge screening and structure selection for Ising networks. *Psychometrika*, *87*(1), 47–82.

Marsman, M., & Rhemtulla, M. (2022). Guest editors' introduction to the special issue "network psychometrics in action": Methodological innovations inspired by empirical problems. *Psychometrika*, *87*(1), 1–11.

Mazumder, R., & Hastie, T. (2012). The graphical Lasso: New insights and alternatives. *Electronic Journal of Statistics*, *6*, 2125.

Miller, J. W. (2021). Asymptotic normality, concentration, and coverage of generalized posteriors. *The Journal of Machine Learning Research*, *22*(1), 7598–7650.

Mkhitaryan, S., Crutzen, R., Steenaart, E., & de Vries, N. K. (2019). Network approach in health behavior research: how can we explore new questions? *Health Psychology and Behavioral Medicine*, *7*(1), 362–384.

Narisetty, N. N., Shen, J., & He, X. (2018). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association*.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, *108*(504), 1339–1349.

Ryan, O., Bringmann, L. F., & Schuurman, N. K. (2022). The challenge of generating causal hypotheses using network models. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(6), 953–970.

Siew, C. S. (2020). Applications of network science to education research: Quantifying knowledge and the development of expertise through network analysis. *Education Sciences*, *10*(4), 101.

Simon de Blas, C., Gomez Gonzalez, D., & Criado Herrero, R. (2021, 03). Network analysis: An indispensable tool for curricula design. A real case-study of the degree on mathematics at the URJC in Spain. *PLOS ONE*, *16*(3), 1-21.

Sweet, T. M., Thomas, A. C., & Junker, B. W. (2013). Hierarchical network models for education research: Hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, *38*(3), 295–318.

Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, *4*(1), 1–10.

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (Second ed.). Boca Raton: Chapman and Hall/CRC.

Van Der Maas, H. L., Kan, K.-J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence*, *5*(2), 16.

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842–861.

Willcox, K. E., & Huang, L. (2017). Network models for mapping educational data. *Design Science*, *3*, e18.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*(1), 19–35.