CrossMark

A novel Bayesian approach for latent variable modeling from mixed data with missing values

Ruifei Cui[1] · Ioan Gabriel Bucur[1] · Perry Groot[1] · Tom Heskes[1]

Abstract
We consider the problem of learning parameters of latent variable models from mixed (continuous and ordinal) data with missing values. We propose a novel Bayesian Gaussian copula factor (BGCF) approach that is proven to be consistent when the data are missing completely at random (MCAR) and that is empirically quite robust when the data are missing at random, a less restrictive assumption than MCAR. In simulations, BGCF substantially outperforms two state-of-the-art alternative approaches. An illustration on the 'Holzinger & Swineford 1939' dataset indicates that BGCF is favorable over the so-called robust maximum likelihood.

Keywords Latent variables · Gaussian copula factor model · Parameter learning · Mixed data · Missing values

1 Introduction

In psychology, social sciences, and many other fields, researchers are usually interested in "latent" variables that cannot be measured directly, e.g., depression, anxiety, or intelligence. To get a grip on these latent concepts, one commonly used strategy is to construct a measurement model for such a latent variable, in the sense that domain experts design multiple "items" or "questions" that are considered to be indicators of the latent variable. For exploring evidence of construct validity in theory-based instrument construction, confirmatory factor analysis (CFA) has been widely studied (Jöreskog 1969; Castro et al. 2015; Li 2016). In CFA, researchers start with several hypothesized latent variable models that are then fitted to the data individually, after which the one that fits the data best is picked to explain the observed phenomenon. In this process, the fundamental task is to learn the parameters of a hypothesized model from observed data,

which is the focus of this paper. For convenience, we simply refer to these hypothesized latent variable models as CFA models from now on.

The most common method for parameter estimation in CFA models is maximum likelihood (ML), because of its attractive statistical properties (consistency, asymptotic normality, and efficiency). The ML method, however, relies on the assumption that observed variables follow a multivariate normal distribution (Jöreskog 1969). When the normality assumption is not deemed empirically tenable, ML may not only reduce the accuracy of parameter estimates, but may also yield misleading conclusions drawn from empirical data (Li 2016). To this end, a robust version of ML was introduced for CFA models when the normality assumption is slightly or moderately violated (Kaplan 2008), but still requires the observations to be continuous. In the real world, the indicator data in questionnaires are usually measured on an ordinal scale (resulting in a bunch of ordered categorical variables, or simply ordinal variables) (Poon and Wang 2012), in which neither normality nor continuity is plausible (Lubke and Muthén 2004). In this case, Item Response Theory (IRT) models (Embretson and Reise 2013) are widely used, in which a mathematical item response function is applied to link an item to its corresponding latent trait. However, the likelihood of the observed ordinal random vector does not have closed-form and is considerably complex due to the presence a multi-dimensional integral, so that learning the model given just the ordinal observa-

✉ Ruifei Cui
  r.cui@science.ru.nl

  Ioan Gabriel Bucur
  g.bucur@cs.ru.nl

  Perry Groot
  perry.groot@science.ru.nl

  Tom Heskes
  t.heskes@science.ru.nl

[1] Radboud University Nijmegen, Nijmegen, Netherlands

tions is typically intractable especially when the number of latent variables and the number of categories of the observed variables are large. Another class of methods designed for ordinal observations is the diagonally weighted least squares (DWLS), which has been suggested to be superior to the ML method and is usually considered to be preferable over other methods (Barendse et al. 2015; Li 2016). Various implementations of DWLS are available in popular softwares or packages, e.g., LISREL (Jöreskog 2005), M*plus* (Muthén 2010), **lavaan** (Rosseel 2012) and **OpenMx** (Boker et al. 2011)

However, there are two major issues that the existing approaches do not consider. One is the mixture of continuous and ordinal data. As we mentioned above, ordinal variables are omnipresent in questionnaires, whereas sensor data are usually continuous. Therefore, a more realistic case in real applications is mixed continuous and ordinal data. A second important issue concerns missing values. In practice, all branches of experimental science are plagued by missing values (Little and Rubin 1987), e.g., failure of sensors, or unwillingness to answer certain questions in a survey. A straightforward idea in this case is to combine missing values techniques with existing parameter estimation approaches, e.g., performing listwise-deletion or pairwise-deletion first on the original data and then applying DWLS to learn parameters of a CFA model. However, such deletion methods are only consistent when the data are *missing completely at random* (MCAR), which is a rather strong assumption (Rubin 1976), and cannot transfer the sampling variability incurred by missing values to follow-up studies. The two modern missing data techniques, maximum likelihood and multiple imputation, are valid under a less restrictive assumption, *missing at random* (MAR) (Schafer and Graham 2002), but they require the data to be multivariate normal.

Therefore, there is a strong demand for an approach that is not only valid under MAR but also works for mixed continuous and ordinal data. For this purpose, we propose a novel Bayesian Gaussian copula factor (BGCF) approach, in which a Gibbs sampler is used to draw pseudo Gaussian data in a latent space restricted by the observed data (unrestricted if that value is missing) and draw posterior samples of parameters given the pseudo data, iteratively. We prove that this approach is consistent under MCAR and empirically show that it works quite well under MAR.

The rest of this paper is organized as follows. Section 2 reviews background knowledge and related work. Section 3 gives the definition of a Gaussian copula factor model and presents our novel inference procedure for this model. Section 4 compares our BGCF approach with two alternative approaches on simulated data, and Sect. 5 gives an illustration on the 'Holzinger & Swineford 1939' dataset. Section 6 concludes this paper and provides some discussion.

2 Background

This section reviews basic missingness mechanisms and related work on parameter estimation in CFA models.

2.1 Missingness mechanism

Following Rubin (1976), let $Y = (y_{ij}) \in \mathbb{R}^{n \times p}$ be a data matrix with the rows representing independent samples, and $R = (r_{ij}) \in \{0, 1\}^{n \times p}$ be a matrix of indicators, where $r_{ij} = 1$ if y_{ij} was observed and $r_{ij} = 0$ otherwise. Y consists of two parts, Y_{obs} and Y_{miss}, representing observed and missing elements in Y, respectively. When the missingness does not depend on the data, i.e., $P(R|Y, \theta) = P(R|\theta)$ with θ denoting unknown parameters, the data are said to be *missing completely at random* (MCAR), which is a special case of a more realistic assumption called *missing at random* (MAR). MAR allows the dependency between missingness and observed values, i.e., $P(R|Y, \theta) = P(R|Y_{\text{obs}}, \theta)$. For example, all people in a group are required to take a blood pressure test at time point 1, while only those whose values at time point 1 lie in the abnormal range need to take the test at time point 2. This results in some missing values at time point 2 that are MAR.

2.2 Parameter estimation in CFA models

When the observations follow a multivariate normal distribution, maximum likelihood (ML) is the mostly-used method. It is equivalent to minimizing the discrepancy function F_{ML} (Jöreskog 1969):

$$F_{\text{ML}} = \ln|\Sigma(\theta)| + \text{trace}[S\Sigma^{-1}(\theta)] - \ln|S| - p,$$

where θ is the vector of model parameters, $\Sigma(\theta)$ is the model-implied covariance matrix, S is the sample covariance matrix, and p is the number of observed variables in the model. When the normality assumption is violated either slightly or moderately, robust ML (MLR) offers an alternative. Here, parameter estimates are still obtained using the asymptotically unbiased ML estimator, but standard errors are statistically corrected to enhance the robustness of ML against departures from normality (Kaplan 2008; Muthén 2010). Another method for continuous nonnormal data is the so-called asymptotically distribution free method, which is a weighted least squares (WLS) method using the inverse of the asymptotic covariance matrix of the sample variances and covariances as a weight matrix (Browne 1984).

When the observed data are on ordinal scales, Muthén (1984) proposed a three-stage approach. It assumes that a normal latent variable x^* underlies an observed ordinal variable x, i.e.,

$$x = m, \text{ if } \tau_{m-1} < x^* < \tau_m, \tag{1}$$

where $m \ (= 1, 2, \ldots, c)$ denotes the observed values of x, τ_m are thresholds ($-\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_c = +\infty$), and c is the number of categories. The thresholds and polychoric correlations are estimated from the bivariate contingency table in the first two stages (Olsson 1979; Jöreskog 2005). Parameter estimates and the associated standard errors are then obtained by minimizing the weighted least squares fit function F_{WLS}:

$$F_{\text{WLS}} = [s - \sigma(\theta)]^{\text{T}} W^{-1} [s - \sigma(\theta)],$$

where θ is the vector of model parameters, $\sigma(\theta)$ is the model-implied vector containing the nonredundant vectorized elements of $\Sigma(\theta)$, s is the vector containing the estimated polychoric correlations, and the weight matrix W is the asymptotic covariance matrix of the polychoric correlations. A mathematically simple form of the WLS estimator, the unweighted least squares (ULS), arises when the matrix W is replaced with the identity matrix I. Another variant of WLS is the diagonally weighted least squares (DWLS), in which only the diagonal elements of W are used in the fit function (Muthén et al. 1997; Muthén 2010), i.e.,

$$F_{\text{DWLS}} = [s - \sigma(\theta)]^{\text{T}} W_{\text{D}}^{-1} [s - \sigma(\theta)],$$

where $W_{\text{D}}^{-1} = \text{diag}(W)$ is the diagonal weight matrix. Various recent simulation studies have shown that DWLS is favorable compared to WLS, ULS, as well as the ML-based methods for ordinal data (Barendse et al. 2015; Li 2016).

3 Method

In this section, we introduce the Gaussian copula factor model and propose a Bayesian inference procedure for this model. Then, we theoretically analyze the identifiability and prove the consistency of our procedure.

3.1 Gaussian copula factor model

Definition 1 (*Gaussian copula factor model*) Consider a latent random (factor) vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^{\text{T}}$, a response random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^{\text{T}}$ and an observed random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^{\text{T}}$, satisfying

$$\boldsymbol{\eta} \sim \mathcal{N}(0, C), \tag{2}$$
$$\boldsymbol{Z} = \Lambda \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{3}$$
$$Y_j = F_j^{-1}\big(\Phi\big[Z_j / \sigma(Z_j)\big]\big), \quad \forall j = 1, \ldots, p, \tag{4}$$

with C a correlation matrix over factors, $\Lambda = (\lambda_{ij})$ a $p \times k$ matrix of factor loadings ($k \leq p$), $\boldsymbol{\epsilon} \sim \mathcal{N}(0, D)$ residuals
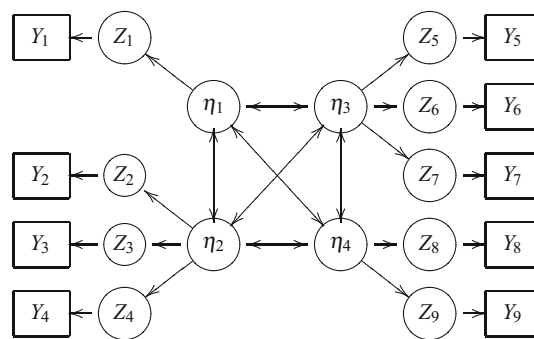

Fig. 1 Gaussian copula factor model

with $D = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, $\sigma(Z_j)$ the standard deviation of Z_j, $\Phi(\cdot)$ the cumulative distribution function (CDF) of the standard Gaussian, and $F_j^{-1}(t) = \inf\{x : F_j(x) \geq t\}$ the pseudo-inverse of a CDF $F_j(\cdot)$. Then, this model is called a *Gaussian copula factor model*.

The model is also defined in Murray et al. (2013), but the authors restrict the factors to be independent of each other while we allow for their interactions. Our model is a combination of a Gaussian factor model (from $\boldsymbol{\eta}$ to $\boldsymbol{Z}$) and a Gaussian copula model (from $\boldsymbol{Z}$ to $\boldsymbol{Y}$). The factor model allows us to grasp the latent concepts that are measured by multiple indicators. The copula model provides a good way to conduct multivariate data analysis for two reasons. First, it raises the theoretical framework in which multivariate associations can be modeled separately from the univariate distributions of the observed variables (Nelsen 2007). Especially, when we use a Gaussian copula, the multivariate associations are uniquely determined by the covariance matrix because of the elliptically symmetric joint density, which makes the dependency analysis very simple. Second, the use of copulas is advocated to model multivariate distributions involving diverse types of variables, say binary, ordinal, and continuous (Dobra and Lenkoski 2011). A variable Y_j that takes a finite number of ordinal values $\{1, 2, \ldots, c\}$ with $c \geq 2$, is incorporated into our model by introducing a latent Gaussian variable Z_j, which complies with the well-known standard assumption for an ordinal variable (Muthén 1984) (see Eq. 1). Figure 1 shows an example of the model. Note that we allow the special case of a factor having a single indicator, e.g., $\eta_1 \to Z_1 \to Y_1$, because this allows us to incorporate other (explicit) variables (such as age and income) into our model. In this special case, we set $\lambda_{11} = 1$ and $\epsilon_1 = 0$, thus $Y_1 = F_1^{-1}(\Phi[\eta_1])$.

In the typical design for questionnaires, one tries to get a grip on a latent concept through a particular set of well-designed questions (Martínez-Torres 2006; Byrne 2013), which implies that a factor (latent concept) in our model is connected to multiple indicators (questions) while an indicator is only used to measure a single factor, as shown in Fig. 1. This kind of measurement model is called a *pure measure-*

ment model (Definition 8 in Silva et al. (2006)). Throughout this paper, we assume that all measurement models are pure, which indicates that there is only a single non-zero entry in each row of the factor loadings matrix Λ. This inductive bias about the sparsity pattern of Λ is fully motivated by the typical design of a measurement model.

In what follows, we transform the Gaussian copula factor model into an equivalent model that is used for inference in the next subsection. We consider an integrated $(p + k)$-dimensional random vector $X = (Z^T, \eta^T)^T$, which is still multivariate Gaussian, and obtain its covariance matrix

$$\Sigma = \begin{bmatrix} \Lambda C \Lambda^T + D & \Lambda C \\ C \Lambda^T & C \end{bmatrix}, \tag{5}$$

and precision matrix

$$\Omega = \Sigma^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}\Lambda \\ -\Lambda^T D^{-1} & C^{-1} + \Lambda^T D^{-1}\Lambda \end{bmatrix}. \tag{6}$$

Since D is diagonal and Λ only has one non-zero entry per row, Ω contains many intrinsic zeros. The sparsity pattern of such $\Omega = (\omega_{ij})$ can be represented by an undirected graph $G = (V, E)$, where $(i, j) \notin E$ whenever $\omega_{ij} = 0$ by construction. Then, a Gaussian copula factor model can be transformed into an equivalent model controlled by a single precision matrix Ω, which in turn is constrained by G, i.e., $P(X|C, \Lambda, D) = P(X|\Omega_G)$.

Definition 2 (*G-Wishart distribution*) Given an undirected graph $G = (V, E)$, a zero-constrained random matrix Ω has a G-Wishart distribution, if its density function is

$$p(\Omega|G) = \frac{|\Omega|^{(\nu-2)/2}}{I_G(\nu, \Psi)} \exp\left[-\frac{1}{2}\text{trace}(\Psi\Omega) \right] \mathbb{1}_{\Omega \in M^+(G)},$$

with $M^+(G)$ the space of symmetric positive definite matrices with off-diagonal elements $\omega_{ij} = 0$ whenever $(i, j) \notin E$, ν the number of degrees of freedom, Ψ a scale matrix, $I_G(\nu, \Psi)$ the normalizing constant, and $\mathbb{1}$ the indicator function (Roverato 2002).

The G-Wishart distribution is the conjugate prior of precision matrices Ω that are constrained by a graph G (Roverato 2002). That is, given the G-Wishart prior, i.e., $P(\Omega|G) = \mathscr{W}_G(\nu_0, \Psi_0)$ and data $X = (x_1, \ldots, x_n)^T$ drawn from $\mathscr{N}(0, \Omega^{-1})$, the posterior for Ω is another G-Wishart distribution:

$$P(\Omega|G, X) = \mathscr{W}_G(\nu_0 + n, \Psi_0 + X^T X). \tag{7}$$

When the graph G is fully connected, the G-Wishart distribution reduces to a Wishart distribution (Murphy 2007). Placing a G-Wishart prior on Ω is equivalent to placing an

inverse-Wishart on C, a product of multivariate normals on Λ, and an inverse-gamma on the diagonal elements of D. With a diagonal scale matrix Ψ_0 and the number of degrees of freedom ν_0 equal to the dimension of X plus one, the implied marginal densities between any pair of variables are uniformly distributed between $[-1, 1]$ (Barnard et al. 2000).

3.2 Inference for Gaussian copula factor model

We first introduce the inference procedure for complete mixed data and incomplete Gaussian data, respectively, based on which the procedure for mixed data with missing values is then derived. From this point on, we use S to denote the correlation matrix over the response vector Z.

3.2.1 Mixed data without missing values

For a Gaussian copula model, Hoff (2007) proposed a likelihood that only concerns the ranks among observations, which is derived as follows. Since the transformation $Y_j = F_j^{-1}(\Phi[Z_j])$ is non-decreasing, observing $y_j = (y_{1,j}, \ldots, y_{n,j})^T$ implies a partial ordering on $z_j = (z_{1,j}, \ldots, z_{n,j})^T$, i.e., z_j lies in the space restricted by y_j:

$$\mathscr{D}(y_j) = \left\{ z_j \in \mathbb{R}^n : y_{i,j} < y_{k,j} \Rightarrow z_{i,j} < z_{k,j} \right\}.$$

Therefore, observing Y suggests that Z must be in

$$\mathscr{D}(Y) = \{ Z \in \mathbb{R}^{n \times p} : z_j \in \mathscr{D}(y_j), \forall j = 1, \ldots, p \}.$$

Taking the occurrence of this event as the data, one can compute the following likelihood Hoff (2007)

$$P(Z \in \mathscr{D}(Y)|S, F_1, \ldots, F_p) = P(Z \in \mathscr{D}(Y)|S).$$

Following the same argumentation, the likelihood in our Gaussian copula factor model reads

$$P(Z \in \mathscr{D}(Y)|\eta, \Omega, F_1, \ldots, F_p) = P(Z \in \mathscr{D}(Y)|\eta, \Omega),$$

which is independent of the margins F_j.

For the Gaussian copula factor model, inference for the precision matrix Ω of the vector $X = (Z^T, \eta^T)^T$ can now proceed via construction of a Markov chain having its stationary distribution equal to $P(Z, \eta, \Omega|Z \in \mathscr{D}(Y), G)$, where we ignore the values for η and Z in our samples. The prior graph G is uniquely determined by the sparsity pattern of the loading matrix $\Lambda = (\lambda_{ij})$ and the residual matrix D (see Eq. 6), which in turn is uniquely decided by the pure measurement models. The Markov chain can be constructed by iterating the following three steps:

1. **Sample Z**: $Z \sim P(Z|\eta, Z \in \mathscr{D}(Y), \Omega)$;
   Since each coordinate Z_j directly depends on only one factor, i.e., η_q such that $\lambda_{jq} \neq 0$, we can sample each of them independently through $Z_j \sim P(Z_j|\eta_q, z_j \in \mathscr{D}(y_j), \Omega)$.
2. **Sample η**: $\eta \sim P(\eta|Z, \Omega)$;
3. **Sample Ω**: $\Omega \sim P(\Omega|Z, \eta, G)$.

3.2.2 Gaussian data with missing values

Suppose that we have Gaussian data Z consisting of two parts, Z_{obs} and Z_{miss}, denoting observed and missing values in Z, respectively. The inference for the correlation matrix of Z in this case can be done via the so-called data augmentation technique that is also a Markov chain Monte Carlo procedure and has been proven to be consistent under MAR (Schafer 1997). This approach iterates the following two steps to impute missing values (Step 1) and draw correlation matrix samples from the posterior (Step 2):

1. $Z_{miss} \sim P(Z_{miss}|Z_{obs}, S)$;
2. $S \sim P(S|Z_{obs}, Z_{miss})$.

3.2.3 Mixed data with missing values

For the most general case of mixed data with missing values, we combine the procedures of Sects. 3.2.1 and 3.2.2 into the following four-step inference procedure:

1. $Z_{obs} \sim P(Z_{obs}|\eta, Z_{obs} \in \mathscr{D}(Y_{obs}), \Omega)$;
2. $Z_{miss} \sim P(Z_{miss}|\eta, Z_{obs}, \Omega)$;
3. $\eta \sim P(\eta|Z_{obs}, Z_{miss}, \Omega)$;
4. $\Omega \sim P(\Omega|Z_{obs}, Z_{miss}, \eta, G)$.

A Gibbs sampler that achieves this Markov chain is summarized in Algorithm 1 and implemented in R.[1] Note that we put Step 1 and Step 2 together in the actual implementation since they share some common computations (lines 2–4). The difference between the two steps is that the values in Step 1 are drawn from a space restricted by the observed data (lines 5–13), while the values in Step 2 are drawn from an unrestricted space (lines 14–17). Another important point is that we need to relocate the data such that the mean of each coordinate of Z is zero (line 20). This is necessary for the algorithm to be sound because the mean may shift when missing values depend on the observed data (MAR).

By iterating the steps in Algorithm 1, we can draw correlation matrix samples over the integrated random vector X, denoted by $\{\Sigma^{(1)}, \ldots, \Sigma^{(m)}\}$. The mean over all the samples is a natural estimate of the true Σ, i.e.,

---

Algorithm 1 Gibbs sampler for Gaussian copula factor model with missing values

---

Require: Prior graph G, observed data Y.
  # Step 1 and **Step 2**:
1: **for** $j \in \{1, \ldots, p\}$ **do**
2: q = factor index of Z_j
3: $a = \Sigma_{[j,q+p]}/\Sigma_{[q+p,q+p]}$
4: $\sigma_j^2 = \Sigma_{[j,j]} - a \times \Sigma_{[q+p,j]}$
   # Step 1: $Z_{obs} \sim P(Z_{obs}|\eta, Z_{obs} \in \mathscr{D}(Y_{obs}), \Omega)$
5: **for** $y \in$ unique$\{y_{1,j}, \ldots, y_{n,j}\}$ **do**
6:   $z_l = \max\{z_{i,j} : y_{i,j} < y\}$
7:   $z_u = \min\{z_{i,j} : y < y_{i,j}\}$
8:   **for** i such that $y_{i,j} = y$ **do**
9:    $\mu_{i,j} = \eta_{[i,q]} \times a$
10:    $u_{i,j} \sim \mathscr{U}\left(\Phi\left[\frac{z_l - \mu_{i,j}}{\sigma_j}\right], \Phi\left[\frac{z_u - \mu_{i,j}}{\sigma_j}\right]\right)$
11:    $z_{i,j} = \mu_{i,j} + \sigma_j \times \Phi^{-1}(u_{i,j})$
12:   **end for**
13: **end for**
   # Step 2: $Z_{miss} \sim P(Z_{miss}|\eta, Z_{obs}, \Omega)$
14: **for** i such that $y_{i,j} \in Y_{miss}$ **do**
15:   $\mu_{i,j} = \eta_{[i,q]} \times a$
16:   $z_{i,j} \sim \mathscr{N}(\mu_{i,j}, \sigma_j^2)$
17: **end for**
18: **end for**
19: $Z = (Z_{obs}, Z_{miss})$
20: $Z = (Z^T - \mu)^T$, with μ the mean vector of Z
   # Step 3: $\eta \sim P(\eta|Z, \Omega)$
21: $A = \Sigma_{[\eta,Z]}\Sigma_{[Z,Z]}^{-1}$
22: $B = \Sigma_{[\eta,\eta]} - A\Sigma_{[Z,\eta]}$
23: **for** $i \in \{1, \ldots, n\}$ **do**
24: $\mu_i = (Z_{[i,:]}A^T)^T$
25: $\eta_{[i,:]} \sim \mathscr{N}(\mu_i, B)$
26: **end for**
27: $\eta_{[:,j]} = \eta_{[:,j]} \times \text{sign}(\text{Cov}[\eta_{[:,j]}, Z_{[:,f(j)]}]), \forall j$, where $f(j)$ is the index of the first indicator of η_j.
   # Step 4: $\Omega \sim P(\Omega|Z, \eta, G)$
28: $X = (Z, \eta)$
29: $\Omega \sim \mathscr{W}_G(\nu_0 + n, \Psi_0 + X^T X)$
30: $\Sigma = \Omega^{-1}$
31: $\Sigma_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}, \forall i, j$

$$\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} \Sigma^{(i)}. \tag{8}$$

Based on Eqs. (5) and (8), we obtain estimates of the parameters of interests:

$$\hat{C} = \hat{\Sigma}_{[\eta,\eta]};$$
$$\hat{\Lambda} = \hat{\Sigma}_{[Z,\eta]}\hat{C}^{-1};$$
$$\hat{D} = \hat{S} - \hat{\Lambda}\hat{C}\hat{\Lambda}^T, \quad \text{with} \quad \hat{S} = \hat{\Sigma}_{[Z,Z]}. \tag{9}$$

We refer to this procedure as a *Bayesian Gaussian copula factor approach* (BGCF).

3.2.4 Discussion on prior specification

For the default choice of the prior G-Wishart distribution, we set the degrees of freedom $\nu_0 = \dim(X) + 1$ and the

---

[1] The code including those used in simulations and real-world applications is provided in https://github.com/cuiruifei/CopulaFactorModel.

scale matrix $\Psi_0 = \epsilon\mathbb{1}$ in the limit $\epsilon \downarrow 0$, where $\dim(X)$ is the dimension of the integrated random vector X and $\mathbb{1}$ is the identity matrix. This specification results in a non-informative prior, in the sense that the posterior only depends on the data and the prior is ignorable. We recall Eq. (7) and take the posterior expectation as an example. The expectation of the covariance matrix is

$$\mathbb{E}\left(\Sigma\right) = \mathbb{E}\left(\Omega^{-1}\right) = \frac{\Psi_0 + X^\mathrm{T}X}{\nu_0 + n - \dim(X) - 1} = \frac{\Psi_0 + X^\mathrm{T}X}{n},$$

which reduces to the maximum likelihood estimate in the limit $\epsilon \downarrow 0$. In the actual implementation, we simply set $\Psi_0 = \mathbb{1}$, which is accurate enough when the sample size is not too small. In the case of a very small data size, it is needed to make Ψ_0 smaller than the identity matrix.

To incorporate prior knowledge into the inference procedure, our model enjoys some flexibility. As mentioned in Sect. 3.1, placing a G-Wishart prior on Ω is equivalent to placing an inverse-Wishart on C, a product of multivariate normals on Λ, and an inverse-gamma on the diagonal elements of D. Therefore, one could choose one's favorite informative priors on C, Λ, and D separately, and then derive the resulting G-Wishart prior on Ω. While the inverse-Wishart and inverse-gamma distributions have been criticized as unreliable when the variances are close to zero (Schuurman et al. 2016), our model does not suffer from this issue. This is because in our model the response variables (i.e., the Z variables) depend only on the ranks of the observed data, and in our sampling process we always set the variances of the response variables and latent variables to one, which is scale-invariant to the observed data.

One limitation of the current inference procedure is that one has to choose the prior on C from the inverse-Wishart family, on Λ from the normal family, and on D from the inverse-gamma family in order to keep the conjugacy, so that one can enjoy the fast and concise inference. When the prior is chosen from other families, sampling Ω from the posterior distribution (Step 4 in Algorithm 1) is no longer straightforward. In this case, a different strategy like the Metropolis-Hastings algorithm might be needed to implement our Step 4.

3.3 Theoretical analysis

3.3.1 Identifiability of C

Without additional constraints, C is non-identifiable (Anderson and Rubin 1956). More precisely, given a decomposable matrix $S = \Lambda C\Lambda^\mathrm{T} + D$, we can always replace Λ with ΛU and C with $U^{-1}CU^{-\mathrm{T}}$ to obtain an equivalent decomposition $S = (\Lambda U)(U^{-1}CU^{-\mathrm{T}})(U^\mathrm{T}\Lambda^\mathrm{T}) + D$, where U is a

$k \times k$ invertible matrix. Since Λ only has one non-zero entry per row in our model, U can only be diagonal to ensure that ΛU has the same sparsity pattern as Λ (see Lemma 1 in "Appendix"). Thus, from the same S, we get a class of solutions for C, i.e., $U^{-1}CU^{-1}$, where U can be any invertible diagonal matrix. In order to get a unique solution for C, we impose two sufficient identifying conditions: 1) restrict C to be a correlation matrix; 2) force the first non-zero entry in each column of Λ to be positive. See Lemma 2 in "Appendix" for proof. Condition 1 is implemented via line 31 in Algorithm 1. As for the second condition, we force the covariance between a factor and its first indicator to be positive (line 27), which is equivalent to Condition 2. Note that these conditions are not unique; one could choose one's favorite conditions to identify C, e.g., setting the first loading to 1 for each factor. The reason for our choice of conditions is to keep it consistent with our model definition where C is a correlation matrix.

3.3.2 Identifiability of Λ and D

Under the two conditions for identifying C, factor loadings Λ and residual variances D are also identified except for the case in which there exists one factor that is independent of all the others and this factor only has two indicators. For such a factor, we have 4 free parameters (2 loadings, 2 residuals) while we only have 3 available equations (2 variances, 1 covariance), which yields an underdetermined system. See Lemmas 3 and 4 in "Appendix" for detailed analysis. Once this happens, one could put additional constraints to guarantee a unique solution, e.g., by setting the variance of the first residual to zero. However, we would recommend to leave such an independent factor out (especially in association analysis) or study it separately from the other factors.

Under sufficient conditions for identifying C, Λ, and D, our BGCF approach is consistent even with MCAR missing values. This is shown in Theorem 1, whose proof is provided in "Appendix".

Theorem 1 (Consistency of the BGCF approach) *Let $Y_n = (y_1, \ldots, y_n)^\mathrm{T}$ be independent observations drawn from a Gaussian copula factor model. If Y_n is complete (no missing data) or contains missing values that are missing completely at random, then*

$$\lim_{n\to\infty} P\big(\hat{C}_n = C_0\big) = 1,$$
$$\lim_{n\to\infty} P\big(\hat{\Lambda}_n = \Lambda_0\big) = 1,$$
$$\lim_{n\to\infty} P\big(\hat{D}_n = D_0\big) = 1,$$

where $\hat{C}_n$, $\hat{\Lambda}_n$, and $\hat{D}_n$ are parameters learned by BGCF, while C_0, Λ_0, and D_0 are the true ones.

4 Simulation study

In this section, we compare our BGCF approach with alternative approaches via simulations.

4.1 Setup

4.1.1 Model specification

Following typical simulation studies on CFA models in the literature (Yang-Wallentin et al. 2010; Li 2016), we consider a correlated 4-factor model in our study. Each factor is measured by 4 indicators, since Marsh et al. (1998) concluded that the accuracy of parameter estimates appeared to be optimal when the number of indicators per factor was four and marginally improved as the number increased. The interfactor correlations (off-diagonal elements of the correlation matrix C over factors) are randomly drawn from [0.2, 0.4], which is considered a reasonable and empirical range in the applied literature (Li 2016). For the ease of reproducibility, we construct our C as follows.

```
set.seed(12345)
C <- matrix(runif(4^2, 0.2, 0.4), ncol=4)
C <- (C*lower.tri(C)) + t(C*lower.tri(C))
diag(C) <- 1
```

In the majority of empirical research and simulation studies (DiStefano 2002), reported standardized factor loadings range from 0.4 to 0.9. For facilitating interpretability and again reproducibility, each factor loading is set to 0.7. Each corresponding residual variance is then automatically set to 0.51 under a standardized solution in the population model, as done in Li (2016).

4.1.2 Data generation

Given the specified model, one can generate data in the response space (the Z in Definition 1) via Eqs. (2) and (3). When the observed data (the Y in Definition 1) are ordinal, we discretize the corresponding margins into the desired number of categories. When the observed data are nonparanormal, we set the $F_j(\cdot)$ in Eq. (4) to the CDF of a χ^2-distribution with degrees of freedom df. The reason for choosing a χ^2-distribution is that we can easily use df to control the extent of non-normality: a higher df implies a distribution closer to a Gaussian. To fill in a certain percentage β of missing values (we only consider MAR), we follow the procedure in Kolar and Xing (2012), i.e., for $j = 1, \ldots, \lfloor p/2 \rfloor$, $i = 1, \ldots, n$: $y_{i,2*j}$ is missing if $z_{i,2*j-1} < \Phi^{-1}(2*\beta)$.

4.1.3 Evaluation metrics

We use average relative bias (ARB) and root mean squared error (RMSE) to examine the parameter estimates, which are defined as

$$\text{ARB} = \frac{1}{r}\sum_{i=1}^{r}\frac{\hat{\theta}_i - \theta_i}{\theta_i}, \quad \text{RMSE} = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(\hat{\theta}_i - \theta_i)^2},$$

where $\hat{\theta}_i$ and θ_i represent the estimated and true values, respectively. An ARB value less than 5% is interpreted as a *trivial* bias, between 5 and 10% as a *moderate* bias, and greater than 10% as a *substantial* bias (Curran et al. 1996). Note that ARB describes an overall picture of average bias, that is, summing up bias in a positive and a negative direction together. A smaller absolute value of ARB indicates better performance on average.
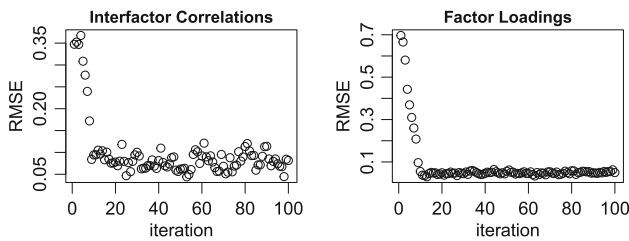
4.2 Ordinal data without missing values

In this subsection, we consider ordinal complete data since this matches the assumptions of the diagonally weighted least squares (DWLS) method, in which we set the number of ordinal categories to be 4. We also incorporate the robust maximum likelihood (MLR) as an alternative approach, which was shown to be empirically tenable when the number of categories is more than 5 (Rhemtulla et al. 2012; Li 2016). See Sect. 2 for details of the two approaches.

Before conducting comparisons, we first check the convergence property of the Gibbs sampler used in our BGCF approach. We randomly generate a dataset of sample size $n = 500$. With this dataset, we run our Gibbs sampler five times independently (with different starting values), in which we collect 2000 successive samples for each chain. Table 1 shows the Potential Scale Reduction Factor (PSRF) (Gelman and Rubin 1992) with 95% upper confidence limit (within the parentheses) of the 6 interfactor correlations and 16 factor loadings over the 5 chains. From Table 1, we see quite a good convergence of the Gibbs sampler. Figure 2 shows the RMSE of estimated interfactor correlations (left panel) and factor loadings (right panel) over the first 100 iterations for the first chain. We see that the sampler converges very fast, in which the burn-in period is only around 10. More experiments done for different numbers of categories and different random datasets show that the burn-in is less than 20 on the whole across various conditions. Figure 3 shows the autocorrelation function of Gibbs samples in the first chain, where we randomly select 3 interfactor correlations and 3 factor loadings, respectively. We see that the autocorrelations almost disappear with a lag of 10. Based on these results, in our following simulations, we just run one chain, in which we set the burn-in to 50, thin the chain with step-size 10, and collect 100 independent Gibbs samples.[2]
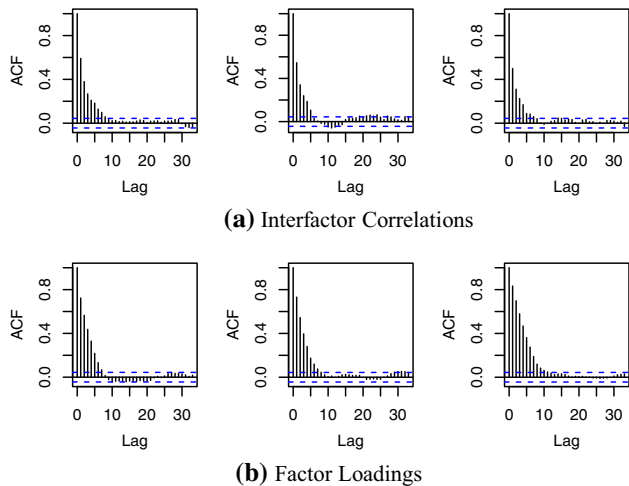
---

[2] Note that the parameter values used here to specify the Gibbs sampler are based on some empirical results. These values can be treated

Table 1 Potential Scale Reduction Factor (PSRF) with 95% upper confidence limit of the 6 interfactor correlations and 16 factor loadings over 5 chains

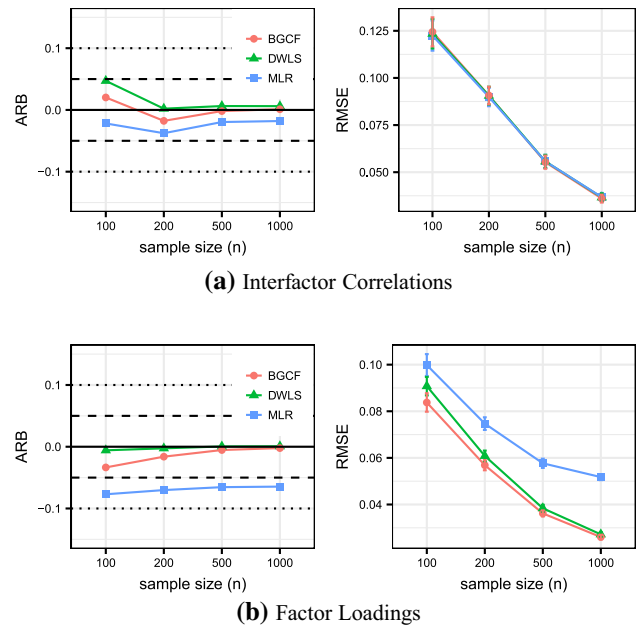|	PSRF		PSRF		PSRF
C_{12}	1.00 (1.00)	λ_1	1.01 (1.02)	λ_9	1.01 (1.02)
C_{13}	1.00 (1.01)	λ_2	1.00 (1.01)	λ_{10}	1.00 (1.01)
C_{14}	1.00 (1.01)	λ_3	1.01 (1.02)	λ_{11}	1.00 (1.00)
C_{23}	1.00 (1.01)	λ_4	1.00 (1.00)	λ_{12}	1.00 (1.00)
C_{24}	1.00 (1.01)	λ_5	1.00 (1.00)	λ_{13}	1.00 (1.01)
C_{34}	1.00 (1.00)	λ_6	1.01 (1.03)	λ_{14}	1.02 (1.05)
		λ_7	1.02 (1.06)	λ_{15}	1.00 (1.00)
		λ_8	1.01 (1.03)	λ_{16}	1.01 (1.02)

Fig. 2 Convergence property of our Gibbs sampler over 100 iterations. Left panel: RMSE of interfactor correlations; Right panel: RMSE of factor loadings

Fig. 3 Autocorrelation function (ACF) of Gibbs samples for **a** randomly select three out of six interfactor correlations, and **b** randomly select three out of sixteen factor loadings

Now we evaluate the three involved approaches. Figure 4 shows the performance of BGCF, DWLS, and MLR over different sample sizes $n \in \{100, 200, 500, 1000\}$, providing

Fig. 4 Results obtained by the Bayesian Gaussian copula factor (BGCF) approach, the diagonally weighted least squares (DWLS), and the robust maximum likelihood (MLR) on complete ordinal data (4 categories) over different sample sizes, showing the mean of ARB (left panel) and the mean of RMSE with 95% confidence interval (right panel) over 100 experiments for **a** interfactor correlations and **b** factor loadings, where dashed lines and dotted lines in left panels denote $\pm 5\%$ and $\pm 10\%$ bias, respectively
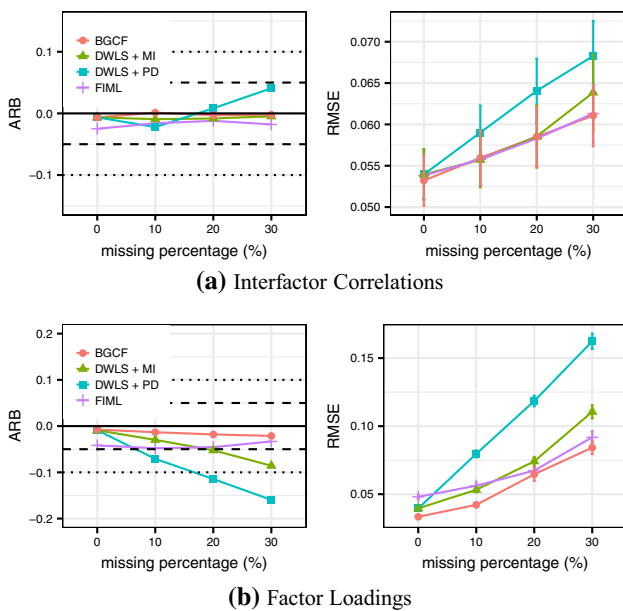
the mean of ARB (left panel) and the mean of RMSE with 95% confidence interval (right panel) over 100 experiments. From Fig. 4a, interfactor correlations are, on average, trivially biased (within two dashed lines) for all the three methods that in turn give indistinguishable RMSE regardless of sample sizes. From Fig. 4b, MLR moderately underestimates the factor loadings and performs worse than DWLS w.r.t. RMSE especially for a larger sample size, which confirms the conclusion in previous studies (Barendse et al. 2015; Li 2016).

4.3 Mixed data with missing values

In this subsection, we consider mixed nonparanormal and ordinal data with missing values, since some latent variables in real-world applications are measured by sensors that usually produce continuous but not necessarily Gaussian data. The 8 indicators of the first 2 factors (4 per factor) are transformed into a χ^2-distribution with $df = 8$, which yields a slightly nonnormal distribution (skewness is 1, excess kurtosis is 1.5) (Li 2016). The 8 indicators of the last 2 factors are discretized into ordinal with 4 categories.

One alternative approach in such cases is DWLS with pairwise-deletion (DWLS + PD), in which heterogeneous correlations (Pearson correlations between numeric variables, polyserial correlations between numeric and ordi-

---

Footnote 2 continued

as default choice, but we recommend to retest the convergence for a specific real-world problem and make the best choice. If this is difficult to do, one could just choose a larger value than the current one to stay in a safe condition since the larger the better for all these parameters.

(a) Interfactor Correlations

(b) Factor Loadings

Fig. 5 Results for $n = 500$ obtained by BGCF, DWLS + PD (pairwise deletion), DWLS + MI (multiple imputation), and the full information maximum likelihood (FIML) on mixed nonparanormal ($df = 8$) and ordinal (4 categories) data with different percentages of missing values, for the same experiments as in Fig. 4

nal variables, and polychoric correlations between ordinal variables) are first computed based on pairwise complete observations, and then DWLS is used to estimate model parameters. A second alternative concerns DWLS with multiple imputation (DWLS + MI), where we choose 20 imputed datasets for the follow-up study.[3] Specifically, we use the R package **mice** (Buuren and Groothuis-Oudshoorn 2010), in which the default imputation method "predictive mean matching" is applied. A third alternative is the full information maximum likelihood (FIML) (Arbuckle 1996; Rosseel 2012), which first applies an EM algorithm to impute missing values and then uses MLR to learn model parameters.

Figure 5 shows the performance of BGCF, DWLS + PD, DWLS + MI, and FIML for $n = 500$ over different percentages of missing values $\beta \in \{0\%, 10\%, 20\%, 30\%\}$. First, despite a good performance with complete data ($\beta = 0\%$) DWLS + PD deteriorates significantly with an increasing percent of missing values especially for factor loadings. DWLS + MI works better than DWLS + PD, but still does not perform well when there are more missing values. Second, our BGCF approach overall outperforms FIML: indistinguishable for interfactor correlations but better for factor loadings.

Two more experiments are provided in "Appendix". One concerns incomplete ordinal data with different numbers of

---

[3] The overall recommendations are to use 20 imputations to have proper estimated coefficients, and use 100 imputations to have proper estimated coefficients and standard errors.

categories, showing that BGCF is favorable over the alternatives for learning factor loadings. Another one considers incomplete nonparanormal data with different extents of deviation from a Gaussian, which indicates that FIML is rather sensitive to the deviation and only performs well for a slightly nonnormal distribution, while the deviation has no influence on BGCF at all. See "Appendix" for more details.

5 Application to real-world data

In this section, we illustrate our approach on the 'Holzinger & Swineford 1939' dataset (Holzinger and Swineford 1939), a classic dataset widely used in the literature and publicly available in the R package **lavaan** (Rosseel 2012). The data consists of mental ability test scores of 301 students, in which we focus on 9 out of the original 26 tests as done in Rosseel (2012). A latent variable model that is often proposed to explore these 9 variables is a correlated 3-factor model shown in Fig. 6, where we rename the observed variables to "Y1, Y2, …, Y9" for simplicity in visualization and to keep it identical to our definition of observed variables (Definition 1). The interpretation of these variables is given in the following list.

- Y1: Visual perception;
- Y2: Cubes;
- Y3: Lozenges;
- Y4: Paragraph comprehension;
- Y5: Sentence completion;
- Y6: Word meaning;
- Y7: Speeded addition;
- Y8: Speeded counting of dots;
- Y9: Speeded discrimination straight and curved capitals.

The summary of the 9 variables in this dataset is provided in Table 2, showing the number of unique values, skewness, and (excess) kurtosis for each variable (this dataset contains no missing values). From the column of unique values, we notice that the data are approximately continuous. The average of 'absolute skewness' and 'absolute excess kurtosis' over the 9 variables are around 0.40 and 0.54, respectively, which is considered to be slightly nonnormal (Li 2016). Therefore, we choose MLR as the alternative to be compared with our BGCF approach, since these conditions match the assumptions of MLR.

We run our Bayesian Gaussian copula factor approach on this dataset. The learned parameter estimates are shown in Fig. 6, in which interfactor correlations are on the bidirected edges, factor loadings are in the directed edges, and unique variance for each variable is around the self-referring arrows. The parameters learned by the MLR approach are not shown

Fig. 6 Path diagram for the Holzinger & Swineford data, in which latent variables are in ovals while observed variables are in squares, bidirected edges between latent variables denote correlation coefficients (interfactor correlations), directed edges denote factor loadings, and self-referring arrows denote residual variance, respectively. The edge weights in the graph are the model parameters learned by our BGCF approach
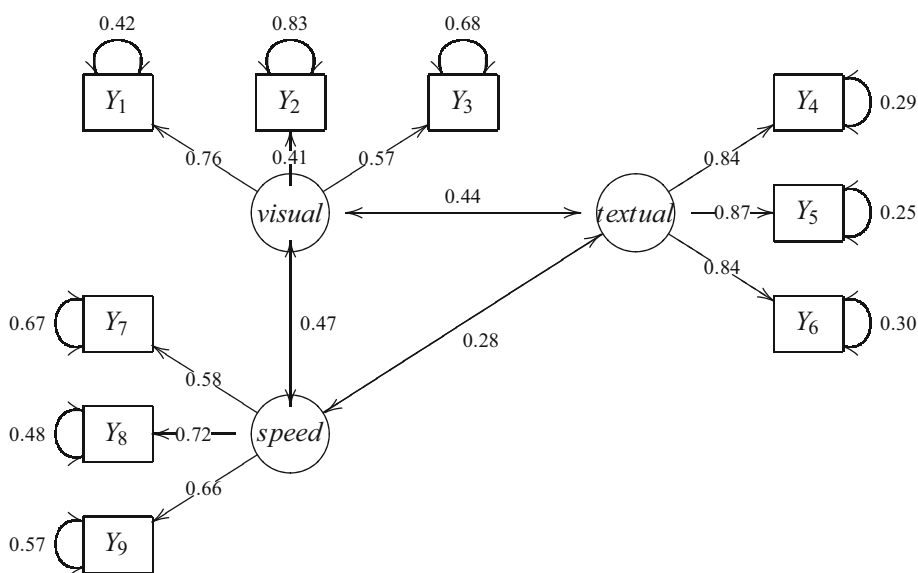
Table 2 The number of unique values, skewness, and (excess) kurtosis of each variable in the 'HolzingerSwineford1939' dataset

Variables	Unique values	Skewness	Kurtosis
Y1	35	− 0.26	0.33
Y2	25	0.47	0.35
Y3	35	0.39	− 0.89
Y4	20	0.27	0.10
Y5	25	− 0.35	− 0.54
Y6	40	0.86	0.84
Y7	97	0.25	− 0.29
Y8	84	0.53	1.20
Y9	129	0.20	0.31

here, since we do not know the ground truth so that it is hard to conduct a comparison between the two approaches.

In order to compare the BGCF approach with MLR quantitatively, we consider answering the question: "What is the value of Y_j when we observe the values of the other variables, denoted by $Y_{\setminus j}$, given the population model structure in Fig. 6?"

This is a regression problem but with additional constraints to obey the population model structure. The difference from a traditional regression problem is that we should learn the regression coefficients from the model-implied covariance matrix rather than the sample covariance matrix over observed variables.

– For MLR, we first learn the model parameters on the training set, from which we extract the linear regression intercept and coefficients of Y_j on $Y_{\setminus j}$. Then, we predict the value of Y_j based on the values of $Y_{\setminus j}$. See Algorithm 2 for pseudo code of this procedure.

– For BGCF, we first estimate the correlation matrix $\hat{S}$ over response variables (the $\mathbf{Z}$ in Definition 1) and the empirical CDF $\hat{F}_j$ of Y_j on the training set. Then we draw latent Gaussian data Z_j given $\hat{S}$ and $Y_{\setminus j}$, i.e., $P(Z_j|\hat{S}, \mathbf{Z}_{\setminus j} \in \mathscr{D}(Y_{\setminus j}))$. Lastly, we obtain the value of Y_j from Z_j via $\hat{F}_j$, i.e., $Y_j = \hat{F}_j^{-1}(\Phi[Z_j])$. See Algorithm 3 for pseudo code of this procedure. Note that we iterate the prediction stage (lines 7–8) for multiple times in the actual implementation to get multiple solutions to $Y_j^{(new)}$, then the average over these solutions is taken as the final predicted value of $Y_j^{(new)}$. This idea is quite similar to multiple imputation.

Algorithm 2 Pseudo code of MLR for regression.

1: **Input:** $Y^{(train)}$ and $Y_{\setminus j}^{(new)}$.
2: **Output:** $Y_j^{(new)}$.
3: **Training Stage:**
4: Fit the model using MLR on $Y^{(train)}$;
5: Extract the model-implied covariance matrix from the fitted model, denoted by $\hat{S}$;
6: Extract regression coefficients $\boldsymbol{b}$ of Y_j on $Y_{\setminus j}$ from $\hat{S}$, that is, $\boldsymbol{b} = \hat{S}_{[\setminus j, \setminus j]}^{-1} \hat{S}_{[\setminus j, j]}$;
7: Obtain the regression intercept b_0, that is, $b_0 = \mathbb{E}(Y_j^{(train)}) - \boldsymbol{b} \cdot \mathbb{E}(Y_{\setminus j}^{(train)})$.
8: **Prediction Stage:**
9: $Y_j^{(new)} = b_0 + \boldsymbol{b} \cdot Y_{\setminus j}^{(new)}$.

The mean squared error (MSE) is used to evaluate the prediction accuracy, where we repeat a tenfold cross validation for 10 times (thus 100 MSE estimates totally). Also, we take Y_j as the outcome variable alternately while treating the others as predictors (thus 9 tasks totally). Figure 7 provides the results of BGCF and MLR for all the 9 tasks, showing the mean of MSE with a standard error represented by error bars over the 100 estimates. We see that BGCF outperforms MLR

Fig. 7 MSE obtained by BGCF and MLR when we take each Y_j as outcome variable (the others as predictors) alternately, showing the mean over 100 experiments (10 times tenfold cross validation) with error bars representing a standard error

---
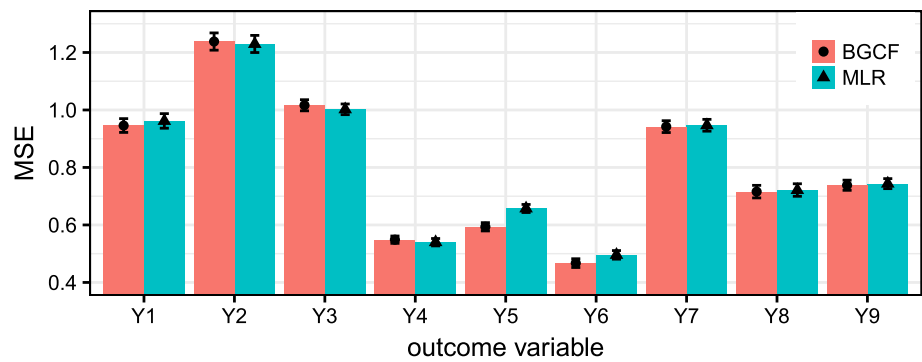
Algorithm 3 Pseudo code of BGCF for regression.

1: **Input:** $Y^{(train)}$ and $Y^{(new)}_{\setminus j}$.
2: **Output:** $Y^{(new)}_j$.
3: **Training Stage:**
4: Apply BGCF to learn the correlation matrix over response variables, i.e., $\hat{S} = \hat{\Sigma}_{[\mathbf{Z},\mathbf{Z}]}$;
5: Learn the empirical cumulative distribution function of Y_j, denoted by $\hat{F}_j$.
6: **Prediction Stage:**
7: Sample $Z^{(new)}_j$ from $P(Z^{(new)}_j | \hat{S}, \mathbf{Z}_{\setminus j} \in \mathscr{D}(Y_{\setminus j}))$;
8: Obtain $Y^{(new)}_j$, i.e., $Y^{(new)}_j = \hat{F}^{-1}_j \big(\Phi[Z^{(new)}_j]\big)$.

---

for Tasks 5 and 6 although they perform indistinguishably for the other tasks. The advantage of BGCF over MLR is encouraging, considering that the experimental conditions match the assumptions of MLR. More experiments are done (not shown) after we make the data moderately or substantially nonnormal, suggesting that BGCF is significantly favorable to MLR, as expected.

6 Summary and discussion

In this paper, we proposed a novel Bayesian Gaussian copula factor (BGCF) approach for learning parameters of CFA models that can handle mixed continuous and ordinal data with missing values. We analyzed the separate identifiability of interfactor correlations C, factor loadings Λ, and residual variances D, since different researchers may care about different parameters. For instance, it is sufficient to identify C for researchers interested in learning causal relations among latent variables (Silva and Scheines 2006; Silva et al. 2006; Cui et al. 2016), with no need to worry about additional conditions to identify Λ and D. Under sufficient identification conditions, we proved that our approach is consistent for MCAR data and empirically showed that it works quite well for MAR data.

In the experiments, our approach outperforms DWLS even under the assumptions of DWLS. Apparently, the approximations inherent in DWLS, such as the use of the polychoric

correlation and its asymptotic covariance, incur a small loss in accuracy compared to an integral approach like the BGCF. When the data follow from a more complicated distribution and contain missing values, the advantage of BGCF over its competitors becomes more prominent. Another highlight of our approach is that the Gibbs sampler converges quite fast, where the burn-in period is rather short. To further reduce the time complexity, a potential optimization of the sampling process is available (Kalaitzis and Silva 2013).

There are various generalizations to our inference approach. While our focus in this paper is on the correlated k-factor models, it is straightforward to extent the current procedure to other class of latent models that are often considered in CFA, such as bi-factor models and second-order models, by simply adjusting the sparsity structure of the prior graph G.

Also, one may consider models with impure measurement indicators, e.g., a model with an indicator measuring multiple factors (cross-loadings) or a model with residual covariances (Bollen 1989), which can be easily solved with BGCF by changing the sparsity pattern of Λ and D. However, two critical issues might arise in this case: the non-identification problems due to a large number of parameters and the slow convergence problem of MCMC algorithms because of dependencies in D. The first issue can be solved by introducing strongly-informative priors (Muthén and Asparouhov 2012), e.g., putting small-variance priors on all cross-loadings. The caveat here is that one needs to choose such priors very carefully to reach a good balance between incorporating correct information and avoiding non-identification. See Muthén and Asparouhov (2012) for more details about the choice of priors on cross-loadings and correlated residuals. Once having the priors on C, Λ, and D, one can derive the prior on Ω. The second issue can be alleviated via the parameter expansion technique (Ghosh and Dunson 2009; Merkle and Rosseel 2018), in which the residual covariance matrix is decomposed into a couple of simple components through some phantom latent variables, resulting in an equivalent model called a *working model*. Then,

our inference procedure can proceed based on the working model.

It is possible to extend the current approach to multiple groups to accommodate cross-national research or by incorporating a multilevel structure, although this is not quite straightforward. Then, one might not be able to draw the precision matrix directly from a *G*-Wishart (Step 4 in Algorithm 1) since different groups may have different *C* and *D* while they share the same Λ. However, this step can be implemented by drawing *C*, Λ, and *D* separately.

Another line of future work is to analyze standard errors and confidence intervals while this paper concentrates on the accuracy of parameter estimates. Our conjecture is that BGCF is still favorable because it naturally transfers the extra variability incurred by missing values to the posterior Gibbs samples: we indeed observed a growing variance of the posterior distribution with the increase of missing values in our simulations. On top of the posterior distribution, one could conduct further studies, e.g., causal discovery over latent factors (Silva et al. 2006; Cui et al. 2018), regression analysis (as we did in Sect. 5), or other machine learning tasks. Instead of using a Gaussian copula, some other choices of copulas are available to model advanced properties in the data such as tail dependence and tail asymmetry (Krupskii and Joe 2013, 2015).

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

Appendix A: Proof of Theorem 1

Theorem 1 (Consistency of the BGCF approach) *Let $Y_n = (y_1, \ldots, y_n)^{\mathrm{T}}$ be independent observations drawn from a Gaussian copula factor model. If Y_n is complete (no missing data) or contains missing values that are missing completely at random, then*

$$\lim_{n\to\infty} P\big(\hat{C}_n = C_0\big) = 1,$$
$$\lim_{n\to\infty} P\big(\hat{\Lambda}_n = \Lambda_0\big) = 1,$$

$$\lim_{n\to\infty} P\big(\hat{D}_n = D_0\big) = 1,$$

where $\hat{C}_n$, $\hat{\Lambda}_n$, and $\hat{D}_n$ are parameters learned by BGCF, while C_0, Λ_0, and D_0 are the true ones.

Proof If $S = \Lambda C \Lambda^{\mathrm{T}} + D$ is the response vector's covariance matrix, then its correlation matrix is $\widetilde{S} = V^{-\frac{1}{2}} S V^{-\frac{1}{2}} = V^{-\frac{1}{2}} \Lambda C \Lambda^{\mathrm{T}} V^{-\frac{1}{2}} + V^{-\frac{1}{2}} D V^{-\frac{1}{2}} = \widetilde{\Lambda} C \widetilde{\Lambda}^{\mathrm{T}} + \widetilde{D}$, where *V* is a diagonal matrix containing the diagonal entries of *S*. We make use of Theorem 1 from Murray et al. (2013) to show the consistency of $\widetilde{S}$. Our factor-analytic prior puts positive probability density almost everywhere on the set of correlation matrices that have a *k*-factor decomposition. Then, by applying Theorem 1 in Murray et al. (2013), we obtain the consistency of the posterior distribution on the response vector's correlation matrix for complete data, i.e.,

$$\lim_{n\to\infty} \Pi(\widetilde{S} \in \mathcal{V}(\widetilde{S}_0)|Z_n \in \mathcal{D}(Y_n)) = 1 \text{ a.s. } \forall \, \mathcal{V}(\widetilde{S}_0), \quad (10)$$

where $\mathcal{D}(Y_n)$ is the space restricted by observed data, and $\mathcal{V}(\widetilde{S}_0)$ is a neighborhood of the true parameter $\widetilde{S}_0$. When the data contain missing values that are completely at random (MCAR), we can also directly obtain the consistency of $\widetilde{S}$ by again using Theorem 1 in Murray et al. (2013), with an additional observation that the estimation of ordinary and polychoric/polyserial correlations from pairwise complete data is still consistent under MCAR. That is to say, the consistency shown in Eq. (10) also holds for data with MCAR missing values.

From this point on, to simplify notation, we will omit adding the tilde to refer to the rescaled matrices $\widetilde{S}$, $\widetilde{\Lambda}$, and $\widetilde{D}$. Thus, *S* from now on refers to the correlation matrix of the response vector. Λ and *D* refer to the scaled factor loadings and noise variance, respectively.

The Gibbs sampler underlying the BGCF approach has the posterior of Σ (the correlation matrix of the integrated vector *X*) as its stationary distribution. Σ contains *S*, the correlation matrix of the response random vector, in the upper left block and *C* in the lower right block. Here, *C* is the correlation matrix of factors, which implicitly depends on the *Gaussian copula factor model* from Definition 1 of the main paper via the formula $S = \Lambda C \Lambda^{\mathrm{T}} + D$. In order to render this decomposition identifiable, we need to put constraints on *C*, Λ, *D*. Otherwise, we can always replace Λ with ΛU and *C* with $U^{-1} C U^{-1}$, where *U* is any $k \times k$ invertible matrix, to obtain the equivalent decomposition $S = (\Lambda U)(U^{-1} C U^{-T})(U^{\mathrm{T}} \Lambda^{\mathrm{T}}) + D$. However, we have assumed that Λ follows a particular sparsity structure in which there is only a single non-zero entry for each row. This assumption restricts the space of equivalent solutions, since any ΛU has to follow the same sparsity structure as Λ. More explicitly, ΛU maintains the same sparsity pattern if and only if *U* is a diagonal matrix (Lemma 1).

By decomposing S, we get a class of solutions for C and Λ, i.e., $U^{-1}CU^{-1}$ and ΛU, where U can be any invertible diagonal matrix. In order to get a unique solution for C, we impose two identifying conditions: (1) we restrict C to be a correlation matrix; (2) we force the first non-zero entry in each column of Λ to be positive. These conditions are sufficient for identifying C uniquely (Lemma 2). We point out that these sufficient conditions are not unique. For example, one could replace the two conditions with restricting the first non-zero entry in each column of Λ to be one. The reason for our choice of conditions is to keep it consistent with our model definition where C is a correlation matrix. Under the two conditions for identifying C, factor loadings Λ and residual variances D are also identified except for the case in which there exists one factor that is independent of all the others and this factor only has two indicators. For such a factor, we have 4 free parameters (2 loadings, 2 residuals), while we only have 3 available equations (2 variances, 1 covariance), which yields an underdetermined system. Therefore, the identifiability of Λ and D relies on the observation that a factor has a single or at least three indicators if it is independent of all the others. See Lemmas 3 and 4 for detailed analysis.

Now, given the consistency of S and the unique smooth map from S to C, Λ, and D, we obtain the consistency of the posterior mean of the parameter C, Λ, and D, which concludes our proof. □

Lemma 1 *If $\Lambda = (\lambda_{ij})$ is a $p \times k$ factor loading matrix with only a single non-zero entry for each row, then ΛU will have the same sparsity pattern if and only if $U = (u_{ij})$ is diagonal.*

Proof ($\Rightarrow$) We prove the direct statement by contradiction. We assume that U has an off-diagonal entry that is not equal to zero. We arbitrarily choose that entry to be $u_{rs}, r, s \in \{1, 2, \dots, k\}, r \neq s$. Due to the particular sparsity pattern, we have chosen for Λ, there exists $q \in \{1, 2, \dots, p\}$ such that $\lambda_{qr} \neq 0$ and $\lambda_{qs} = 0$, i.e., the unique factor corresponding to the response Z_q is η_r. However, we have $(\Lambda U)_{qs} = \lambda_{qr} u_{rs} \neq 0$, which means (ΛU) has a different sparsity pattern from Λ. We have reached a contradiction, therefore U is diagonal.

($\Leftarrow$) If U is diagonal, i.e., $U = \text{diag}(u_1, u_2, \dots, u_k)$, then $(\Lambda U)_{ij} = \lambda_{ij} u_j$. This means that $(\Lambda U)_{ij} = 0 \iff \lambda_{ij} u_j = 0 \iff \lambda_{ij} = 0$, so the sparsity pattern is preserved. □

Lemma 2 (Identifiability of C) *Given the factor structure defined in Sect. 3 of the main paper, we can uniquely recover C from $S = \Lambda C \Lambda^T + D$ if (1) we constrain C to be a correlation matrix; (2) we force the first element in each column of Λ to be positive.*

Proof Here, we assume that the model has the stated factor structure, i.e., that there is some Λ, C, and D such that $S =$

$\Lambda C \Lambda^T + D$. We then show that our chosen restrictions are sufficient for identification using an argument similar to that in Anderson and Rubin (1956).

The decomposition $S = \Lambda C \Lambda^T + D$ constitutes a system of $\frac{p(p+1)}{2}$ equations:

$$
\begin{aligned}
s_{ii} &= \lambda_{if(i)}^2 + d_{ii} \\
s_{ij} &= c_{f(i)f(j)}\lambda_{if(i)}\lambda_{jf(j)} , \ i < j,
\end{aligned}
\tag{11}
$$

where $S = (s_{ij}), \Lambda = (\lambda_{ij}), C = (c_{ij}), D = (d_{ij})$, and $f : \{1, 2, \dots, p\} \to \{1, 2, \dots, k\}$ is the map from a response variable to its corresponding factor. Looking at the equation system in (11), we notice that each factor correlation term $c_{qr}, q \neq r$, appears only in the equations corresponding to response variables indexed by i and j such that $f(i) = q$ and $f(j) = r$ or vice versa. This suggests that we can restrict our analysis to submodels that include only two factors by considering the submatrices of S, Λ, C, D that only involve those two factors. To be more precise, the idea is to look only at the equations corresponding to the submatrix $S_{f^{-1}(q)f^{-1}(r)}$, where f^{-1} is the preimage of $\{1, 2, \dots, k\}$ under f. Indeed, we will show that we can identify each individual correlation term corresponding to pairs of factors only by looking at these submatrices. Any information concerning the correlation term provided by the other equations is then redundant.

Let us then consider an arbitrary pair of factors in our model and the corresponding submatrices of Λ, C, D, and S (the case of a single factor is trivial). In order to simplify notation, we will also use Λ, C, D, and S to refer to these submatrices. We also re-index the two factors involved to η_1 and η_2 for simplicity. In order to recover the correlation between a pair of factors from S, we have to analyze three separate cases to cover all the bases (see Fig. 8 for examples concerning each case):

1. The two factors are not correlated, i.e., $c_{12} = 0$ (there are no restrictions on the number of response variables that the factors can have).
2. The two factors are correlated, i.e., $c_{12} \neq 0$, and each has a single response, which implies that $Z_1 = \eta_1$ and $Z_2 = \eta_2$.
3. The two factors are correlated, i.e., $c_{12} \neq 0$, but at least one of them has at least two responses.

Case 1 If the two factors are not correlated (see the example in the left panel of Fig. 8), this fact will be reflected in the matrix S. More specifically, the off-diagonal blocks in S, which correspond to the covariance between the responses of one factor and the responses of the other factor, will be set to zero. If we notice this zero pattern in S, we can immediately determine that $c_{12} = 0$.
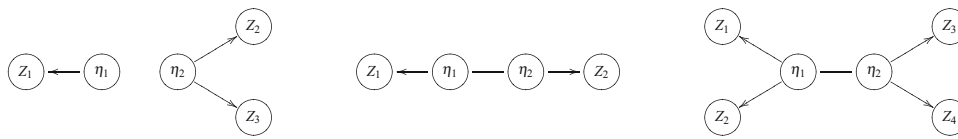
Fig. 8 Left panel: *Case 1* ($c_{12} = 0$); Middle panel: *Case 2* ($c_{12} \neq 0$ and only one response per factor); Right panel: *Case 3* ($c_{12} \neq 0$ and at least one factor has multiple responses)

Case 2 If the two factors are correlated and each factor has a single associated response (see the middle panel of Fig. 8), the model reduces to a Gaussian Copula model. Then, we directly get $c_{12} = s_{12}$ since we have put the constraints $Z = \eta$ if η has a single indicator Z.

Case 3 If at least one of the factors (w.l.o.g., η_1) is allowed to have more than one response (see the example in the right panel of Fig. 8), we arbitrarily choose two of these responses. We also require one response variable corresponding to the other factor (η_2). We use λ_{i1}, λ_{j1}, and λ_{l2} to denote the loadings of these response variables, where $i, j, l \in \{1, 2, \ldots, p\}$. From Eq. (11) we have:

$$s_{ij} = \lambda_{i1}\lambda_{j1}$$
$$s_{il} = c_{12}\lambda_{i1}\lambda_{l2}$$
$$s_{jl} = c_{12}\lambda_{j1}\lambda_{l2}.$$

Since we are in the case in which $c_{12} \neq 0$, which automatically implies that $s_{jl} \neq 0$, we can divide the last two equations to obtain $\frac{s_{il}}{s_{jl}} = \frac{\lambda_{i1}}{\lambda_{j1}}$. We then multiply the result with the first equation to get $\frac{s_{ij}s_{il}}{s_{jl}} = \lambda_{i1}^2$. Without loss of generality, we can say that λ_{i1} is the first entry in the first column of Λ, which means that $\lambda_{i1} > 0$. This means that we have uniquely recovered λ_{i1} and λ_{j1}.

We can also assume without loss of generality that λ_{l2} is the first entry in the second column of Λ, so $\lambda_{l2} > 0$. If η_2 has at least two responses, we use a similar argument to the one before to uniquely recover λ_{l2}. We can then use the above equations to get c_{12}. If η_2 has only one response, then $d_{ll} = 0$, which means that $s_{ll} = \lambda_{l2}^2$, so again λ_{l2} is uniquely recoverable and we can obtain c_{12} from the equations above.

Thus, we have shown that we can correctly determine c_{qr} only from $S_{f^{-1}(q)f^{-1}(r)}$ in all three cases. By applying this approach to all pairs of factors, we can uniquely recover all pairwise correlations. This means that, given our constraints, we can uniquely identify C from the decomposition of S. □

Lemma 3 (Identifiability of Λ) *Given the factor structure defined in Sect. 3 of the main paper, we can uniquely recover Λ from $S = \Lambda C \Lambda^{\mathrm{T}} + D$ if (1) we constrain C to be a correlation matrix; (2) we force the first element in each column of Λ to be positive; (3) when a factor is independent of all the others, it has either a single or at least three indicators.*

**Fig. 9** A factor model with three indicators
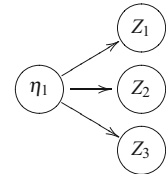
Proof Compared to identifying C, we need to consider another case in which there is only one factor or there exists one factor that is independent of all the others (the former can be treated as a special case of the latter). When such a factor only has a single indicator, e.g., η_1 in the left panel of Fig. 8, we directly identify $d_{11} = 0$ because of the constraint $Z_1 = \eta_1$. When the factor has two indicators, e.g., η_2 in the left panel of Fig. 8, we have four free parameters ($\lambda_{22}, \lambda_{32}, d_{22}$, and d_{33}) while we can only construct three equations from S (s_{22}, s_{33}, and s_{23}), which cannot give us a unique solution. Now we turn to the three-indicator case, as shown in Fig. 9. From Eq. (11) we have:

$$s_{12} = \lambda_{11}\lambda_{21}$$
$$s_{13} = \lambda_{11}\lambda_{31}$$
$$s_{23} = \lambda_{21}\lambda_{31}.$$

We then have $\frac{s_{12}s_{13}}{s_{23}} = \lambda_{11}^2$, which has a unique solution for λ_{11} together with the second constraint $\lambda_{11} > 0$, after which we can naturally get the solutions to λ_{21} and λ_{31}. For the other cases, the proof follows the same line of reasoning as Lemma 2. □

Lemma 4 (Identifiability of D) *Given the factor structure defined in Sect. 3 of the main paper, we can uniquely recover D from $S = \Lambda C \Lambda^{\mathrm{T}} + D$ if (1) we constrain C to be a correlation matrix; (2) when a factor is independent of all the others, it has either a single or at least three indicators.*

Proof We conduct our analysis case by case. For the case where a factor has a single indicator, we trivially set $d_{ii} = 0$. For the case in Fig. 9, it is straightforward to get $d_{11} = s_{11} - \lambda_{11}^2$ from $\frac{s_{12}s_{13}}{s_{23}} = \lambda_{11}^2$ (the same for d_{22} and d_{33}). Another case we need to consider is Case 3 in Fig. 8, where we have $\frac{s_{ij}s_{il}}{s_{jl}} = \lambda_{i1}^2$ (see analysis in Lemma 2), based on which we obtain $d_{ii} = s_{ii} - \lambda_{i1}^2$. By applying this approach to all single factors or pairs of factors, we can uniquely recover all elements of D. □

(a) Interfactor Correlations

(b) Factor Loadings
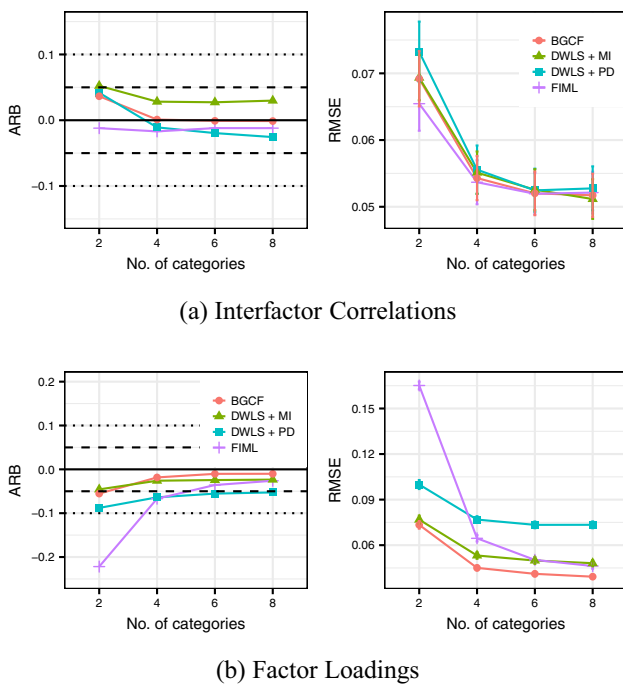
Fig. 10 Results for $n = 500$ and $\beta = 10\%$ obtained by BGCF, DWLS+ PD, DWLS + MI, and FIML on ordinal data with different numbers of categories, showing the mean of ARB (left panel) and the mean of RMSE with 95% confidence interval (right panel) over 100 experiments for **a** interfactor correlations and **b** factor loadings, where dashed lines and dotted lines in left panels denote $\pm 5\%$ and $\pm 10\%$ bias, respectively

(a) Interfactor Correlations

(b) Factor Loadings
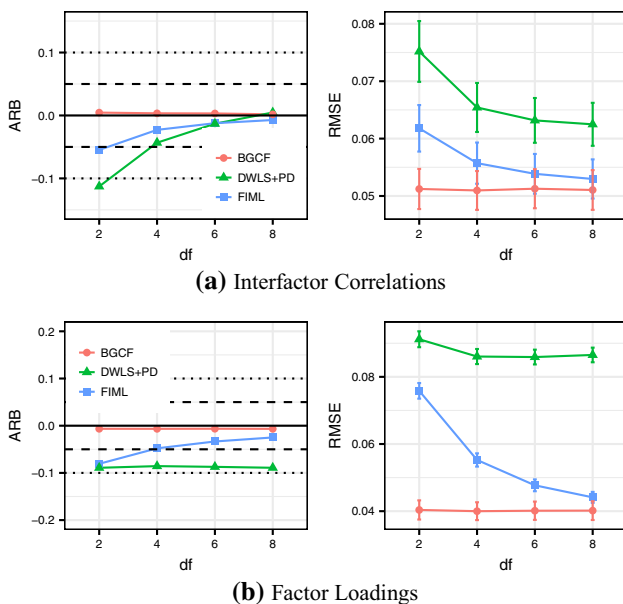
Fig. 11 Results for $n = 500$ and $\beta = 10\%$ obtained by BGCF, DWLS with PD, and FIML on nonparanormal data with different extents of non-normality, for the same experiments as in Fig. 10

Appendix B: Extended simulation study

This section continues the experiments in Sect. 4 of the main paper, in order to check the influence of the number of cat-

egories for ordinal data and the extent of non-normality for nonparanormal data.

B1: Ordinal data with different numbers of categories

In this subsection, we consider ordinal data with various numbers of categories $c \in \{2, 4, 6, 8\}$, in which the sample size and missing values percentage are set to $n = 500$ and $\beta = 10\%$, respectively. Figure 10 shows the results obtained by BGCF (Bayesian Gaussian copula factor), DWLS + PD (diagonally weighted least squares with pairwise deletion), DWLS+MI (diagonally weighted least squares with multiple imputation) and FIML (full information maximum likelihood), providing the mean of ARB (average relative bias) and the mean of RMSE (root mean squared error) with 95% confidence interval over 100 experiments for (a) interfactor correlations and (b) factor loadings. From Fig. 10a, although DWLS+MI has very similar behavior as BGCF w.r.t. RMSE, BGCF is less biased especially when there are more categories. From Fig. 10b, BGCF outperforms all the three alternative approaches w.r.t. both ARB and RMSE.

B2: Nonparanormal data with different extents of non-normality

In this subsection, we consider nonparanormal data, in which we use the degrees of freedom df of a χ^2-distribution to control the extent of non-normality (see Section 4.1.2 of the main paper for details). The sample size and missing values percentage are set to $n = 500$ and $\beta = 10\%$, respectively, while the degrees of freedom varies $df \in \{2, 4, 6, 8\}$.

Figure 11 shows the results obtained by BGCF, DWLS + PD, and FIML, providing the mean of ARB (left panel) and the mean of RMSE with 95% confidence interval (right panel) over 100 experiments for (a) interfactor correlations and (b) factor loadings. We do not include DWLS + MI in this experiment because it becomes approximately the same with FIML for fully continuous data. The major conclusion drawn here is that, while a nonparanormal transformation has no effect on our BGCF approach, FIML is quite sensitive to the extent of non-normality, especially for factor loadings.

References

Anderson, T.W., Rubin, H.: Statistical inference in factor analysis. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry, University of California Press, Berkeley, CA, pp. 111–150 (1956)

Arbuckle, J.L.: Full information estimation in the presence of incomplete data. In: Marcoulides, G.A., Schumacker, R.E. (eds.)

Advanced Structural Equation Modeling: Issues and Techniques, vol. 243, p. 277. Lawrence Erlbaum Associates, Mahwah (1996)

Barendse, M., Oort, F., Timmerman, M.: Using exploratory factor analysis to determine the dimensionality of discrete responses. Struct. Equ. Model. **22**(1), 87–101 (2015)

Barnard, J., McCulloch, R., Meng, X.L.: Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Stat. Sin. **10**, 1281–1311 (2000)

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., et al.: Openmx: an open source extended structural equation modeling framework. Psychometrika **76**(2), 306–317 (2011)

Bollen, K.: Structural Equations with Latent Variables. Wiley, New York (1989)

Browne, M.W.: Asymptotically distribution-free methods for the analysis of covariance structures. Br. J. Math. Stat. Psychol. **37**(1), 62–83 (1984)

Buuren, S.V., Groothuis-Oudshoorn, K.: mice: multivariate imputation by chained equations in R. J. Stat. Softw. **45**, 1–68 (2010)

Byrne, B.M.: Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming. Routledge, London (2013)

Castro, L.M., Costa, D.R., Prates, M.O., Lachos, V.H.: Likelihood-based inference for Tobit confirmatory factor analysis using the multivariate Student-t distribution. Stat. Comput. **25**(6), 1163–1183 (2015)

Cui, R., Groot, P., Heskes, T.: Copula PC algorithm for causal discovery from mixed data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp. 377–392 (2016)

Cui, R., Groot, P., Heskes, T.: Learning causal structure from mixed data with missing values using Gaussian copula models. Stat. Comput. (2018). https://doi.org/10.1007/s11222-018-9810-x

Curran, P.J., West, S.G., Finch, J.F.: The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychol. Methods **1**(1), 16 (1996)

DiStefano, C.: The impact of categorization with confirmatory factor analysis. Struct. Equ. Model. **9**(3), 327–346 (2002)

Dobra, A., Lenkoski, A., et al.: Copula Gaussian graphical models and their application to modeling functional disability data. Ann. Appl. Stat. **5**(2A), 969–993 (2011)

Embretson, S.E., Reise, S.P.: Item Response Theory. Psychology Press, Hove (2013)

Gelman, A., Rubin, D.B., et al.: Inference from iterative simulation using multiple sequences. Stat. Sci. **7**(4), 457–472 (1992)

Ghosh, J., Dunson, D.B.: Default prior distributions and efficient posterior computation in bayesian factor analysis. J. Comput. Graph. Stat. **18**(2), 306–320 (2009)

Hoff, P.D.: Extending the rank likelihood for semiparametric copula estimation. Ann. Stat. **1**, 265–283 (2007)

Holzinger, K.J., Swineford, F.: A study in factor analysis: the stability of a bi-factor solution. Suppl. Educ. Monogr. **48**, 468–469 (1939)

Jöreskog, K.G.: A general approach to confirmatory maximum likelihood factor analysis. Psychometrika **34**(2), 183–202 (1969)

Jöreskog, K.G.: Structural Equation Modeling with Ordinal Variables Using LISREL. Technical Report. Scientific Software International Inc, Lincolnwood, IL (2005)

Kalaitzis, A., Silva, R.: Flexible sampling of discrete data correlations without the marginal distributions. In: Advances in Neural Information Processing Systems, pp. 2517–2525 (2013)

Kaplan, D.: Structural Equation Modeling: Foundations and Extensions, vol. 10. Sage Publications, Thousand Oaks (2008)

Kolar, M., Xing, E.P.: Estimating sparse precision matrices from data with missing values. In: International Conference on Machine Learning (2012)

Krupskii, P., Joe, H.: Factor copula models for multivariate data. J. Multivar. Anal. **120**, 85–101 (2013)

Krupskii, P., Joe, H.: Structured factor copula models: theory, inference and computation. J. Multivar. Anal. **138**, 53–73 (2015)

Li, C.H.: Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. Behav. Res. Methods **48**(3), 936–949 (2016)

Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, Hoboken (1987)

Lubke, G.H., Muthén, B.O.: Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. Struct. Equ. Model. **11**(4), 514–534 (2004)

Marsh, H.W., Hau, K.T., Balla, J.R., Grayson, D.: Is more ever too much? The number of indicators per factor in confirmatory factor analysis. Multivar. Behav. Res. **33**(2), 181–220 (1998)

Martínez-Torres, M.R.: A procedure to design a structural and measurement model of intellectual capital: an exploratory study. Inf. Manag. **43**(5), 617–626 (2006)

Merkle, E.C., Rosseel, Y.: blavaan: Bayesian structural equation models via parameter expansion. J. Stat. Softw. **85**(4), 1–30 (2018)

Murphy, K.P.: Conjugate Bayesian analysis of the Gaussian distribution. def **1**(2), 16 (2007)

Murray, J.S., Dunson, D.B., Carin, L., Lucas, J.E.: Bayesian Gaussian copula factor models for mixed data. J. Am. Stat. Assoc. **108**(502), 656–665 (2013)

Muthén, B.: A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika **49**(1), 115–132 (1984)

Muthén, B., Asparouhov, T.: Bayesian structural equation modeling: a more flexible representation of substantive theory. Psychol. Methods **17**(3), 313 (2012)

Muthén, B., du Toit, S., Spisic, D.: Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Psychometrika (1997)

Muthén, L.: Mplus User's Guide. Muthén & Muthén, Los Angeles (2010)

Nelsen, R.B.: An Introduction to Copulas. Springer, Berlin (2007)

Olsson, U.: Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika **44**(4), 443–460 (1979)

Poon, W.Y., Wang, H.B.: Latent variable models with ordinal categorical covariates. Stat. Comput. **22**(5), 1135–1154 (2012)

Rhemtulla, M., Brosseau-Liard, P.É., Savalei, V.: When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol. Methods **17**(3), 354 (2012)

Rosseel, Y.: lavaan: an R package for structural equation modeling. J. Stat. Softw. **48**(2), 1–36 (2012)

Roverato, A.: Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. Scan. J. Stat. **29**(3), 391–411 (2002)

Rubin, D.B.: Inference and missing data. Biometrika **63**, 581–592 (1976)

Schafer, J.L.: Analysis of Incomplete Multivariate Data. CRC Press, Boca Raton (1997)

Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods **7**(2), 147 (2002)

Schuurman, N., Grasman, R., Hamaker, E.: A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. Multivar. Behav. Res. **51**(2–3), 185–206 (2016)

Silva, R., Scheines, R.: Bayesian learning of measurement and structural models. In: International Conference on Machine Learning, pp 825–832 (2006)

Silva, R., Scheines, R., Glymour, C., Spirtes, P.: Learning the structure of linear latent variable models. J. Mach. Learn. Res. **7**(Feb), 191–246 (2006)

Yang-Wallentin, F., Jöreskog, K.G., Luo, H.: Confirmatory factor analysis of ordinal variables with misspecified models. Struct. Equ. Model. **17**(3), 392–423 (2010)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.