Asymptotically Exact and Fast Gaussian Copula Models for Imputation of Mixed Data Types

Benjamin Christoffersen   BENCHR@KTH.SE
Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Sweden

Mark Clements   MARK.CLEMENTS@KI.SE
Keith Humphreys   KEITH.HUMPHREYS@KI.SE
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden

Hedvig Kjellström   HEDVIG@KTH.SE
Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Sweden

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Missing values with mixed data types is a common problem in a large number of machine learning applications such as processing of surveys and in different medical applications. Recently, Gaussian copula models have been suggested as a means of performing imputation of missing values using a probabilistic framework. While the present Gaussian copula models have shown to yield state of the art performance, they have two limitations: they are based on an approximation that is fast but may be imprecise and they do not support unordered multinomial variables. We address the first limitation using direct and arbitrarily precise approximations both for model estimation and imputation by using randomized quasi-Monte Carlo procedures. The method we provide has lower errors for the estimated model parameters and the imputed values, compared to previously proposed methods. We also extend the previous Gaussian copula models to include unordered multinomial variables in addition to the present support of ordinal, binary, and continuous variables.
Keywords: Imputation; Gaussian Copulas; Quasi-Monte Carlo

1. Introduction

Data sets in medical applications, surveys, user ratings, etc. are becoming larger which brings possibilities for machine learning applications. However, these larger data sets often have missing values. Therefore, imputation of missing values becomes increasingly important as a preprocessing step to many applications. Zhao and Udell (2020a,b); Landgrebe et al. (2020) describe a method for imputation for continuous, binary, and ordinal variables using Gaussian copulas which yields state of the art performance. They address the important issue of performing imputation for data sets containing variables such as ratings in reviews, integer scales of how severe the spread of a tumor is (medical data), and rank variables in surveys, which need to be analyzed in combination with continuous variables such as age, income, etc. using a Gaussian copula. This is straightforward since it is often easy to describe the marginal distribution of each variable, few assumptions are made about the marginals, and use of such a probabilistic framework allows for construction of confidence

intervals for the imputed values similar to the measure developed by Zhao and Udell (2020b). Lastly, the Gaussian copulas have computational advantages.

However, there are three open issues. Firstly, Zhao and Udell (2020a) use an approximate expectation maximization (AEM) algorithm in a way that is related to the work by Guo et al. (2015). While the approximation is fast, the approximation may yield inefficient and possibly even biased results. Secondly, like inference functions for margins (IFM) for fully parametric models, they use a two-stage estimation method that may be inefficient in some cases. Finally, their model does not support multinomial variables.[1] Thus, for many data sets, their method cannot be used.

This paper makes two main contributions:

1. The method we provide gives an asymptotically exact and fast approximation of the log marginal likelihood, the derivatives, and the quantities needed to perform imputation. Moreover, we estimate some of the parameters in the marginal distributions jointly with the copula parameters, instead of using an IFM like method, which increases the efficiency.

   Our method provides conditional probabilities for missing binary, ordinal, and multinomial variables with arbitrary precision. The AEM method cannot provide such approximations which are key for multiple imputation. This contribution is mainly described in Section 4 and partly in Section 5.

2. Our Gaussian copula model supports multinomial variables in addition to binary, ordinal, and continuous variables. Thus, our method is applicable to a large number of data sets with mixed data types. This is further described in Section 5.

We will cover the Gaussian copula model in Section 3 but start with a review of related work in Section 2. Our alternative imputation and estimation method are shown Section 4 along with experimental results where we can compare with the AEM by Zhao and Udell (2020a). The extension to multinomial variables along with experimental results are shown in Section 5. Two technical and experimental sections are used to emphasize the two main and distinct contributions of this paper. Our implementation is publicly available as the mdgc package which is available on the Comprehensive R Archive Network (CRAN) and is provided along with the script to produce the results in the supplementary material.

2. Related Work

Zhao and Udell (2020a) method was based on earlier work on Gaussian copula models (D. Hoff, 2007; Liu et al., 2009; Murray et al., 2013; Fan et al., 2017; Feng and Ning, 2019; Cui et al., 2019). These models can be seen as a generalization of the linear mixed model in the sense that we no longer assume that the marginals for the continuous variables are normal distributed. Instead we assume that they have been transformed with bijective transformations and, thus, allow for greater flexibility for the marginal distributions. Secondly, there are binary and ordinal variables that are created by cutting the latent variables into

---

1. We will write multinomial when we refer to unordered multinomial and ordinal when we refer to ordered multinomial.

bins. The resulting model can be estimated with semiparametric methods. However, unlike in the linear mixed model, there is no closed form solution for the likelihood.

Previous work on related models has focused on model estimation rather than imputation. Moreover, some methods are only able to estimate the model with complete data. The Markov chain Monte Carlo (MCMC) method suggested by D. Hoff (2007) is an exception but it can be very slow. Thus, Zhao and Udell (2020a) propose to estimate the model parameters with a frequentist approach. Their estimation method and imputation method is based on an expectation maximization (EM) algorithm which requires evaluation of moments of the truncated multivariate normal distribution (TMVN). They approximate the moments using an approximation similar to the one suggested by Guo et al. (2015). This approximation is fast and, as Zhao and Udell (2020a) show, it yields superior single imputation performance, in their examples, compared with state of the art non-parametric methods such as the random forest based missForest (Stekhoven and Buehlmann, 2012) and the principal components based imputeFAMD (Audigier et al., 2014; Josse and Husson, 2016).

Although Zhao and Udell (2020b) state that direct maximum likelihood is hard to optimize because the likelihood involves Gaussian integrals, moderately precise and, importantly, fast randomized quasi-Monte Carlo (RQMC) procedures have been developed by Genz (1992); Hajivassiliou et al. (1996); Genz and Bretz (2002). These methods are also easy to generalize to related quantities like derivatives and conditional means or probabilities for missing values. Having arbitrarily precise procedures are important as they can yield more efficient estimators of the model parameters, which the researcher may be interested in per se, and can potentially improve the imputation.

3. Gaussian Copula Models

We will cover the model that Zhao and Udell (2020a) use, and thereafter provide our extension to include multinomial variables and alternative methods for estimation and imputation. $\boldsymbol{X}_i$ denotes the vector with K variables for observation $i = 1, \ldots, n$ where some entries may be missing. The X_{ij}'s for $j \in \mathcal{C} \subseteq \{1, \ldots K\}$ are continuous, the entries with $j \in \mathcal{O} \subseteq \{1, \ldots K\}$, with $\mathcal{O} \cap \mathcal{C} = \emptyset$, are ordinal, and the entries with $j \in \mathcal{B} = \{1, \ldots K\} \setminus \mathcal{O} \cup \mathcal{C}$ are binary. For now, we assume that all entries are observed with value $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Let $\Phi^{(k)}(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the k-dimensional multivariate normal distribution cumulative density function (CDF) with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ over the box with limits at $\boldsymbol{a}$ and $\boldsymbol{b}$ given by

$$\Phi^{(k)}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} \phi^{(k)}(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{d}\boldsymbol{u}$$

where $\phi^{(k)}$ is the corresponding density function. We omit the mean vector and covariance matrix in the standard case, omit the superscript in the univariate case, and write $\Phi^{(k)}(\boldsymbol{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ when $\boldsymbol{a} = (-\infty, \ldots, -\infty)^\top$. Zhao and Udell (2020a,b) use a Gaussian copula model where it is assumed that there is a K-dimensional latent variable $\boldsymbol{Z}_i \in \mathbb{R}^K$ such that

$\boldsymbol{Z}_i \sim N^{(K)}(\mathbf{0}_K, \boldsymbol{\Sigma})$ where $\mathbf{0}_k$ is a vector with k zeros and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \sigma_{12} & \cdots & & \sigma_{1K} \\ \sigma_{21} & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \sigma_{K-1,K} \\ \sigma_{K2} & \cdots & \sigma_{K,K-1} & & 1 \end{pmatrix}.$$

The relation to the observed outcomes is

$$X_{ij} = f_j(Z_{ij}), \quad j \in \mathcal{C}$$

$$X_{ij} = \begin{cases} 1 & Z_{ij} > -\Phi^{-1}(p_j) \\ 0 & \text{otherwise} \end{cases}, \quad j \in \mathcal{B} \tag{1}$$

$$X_{ij} = k \Leftrightarrow \alpha_{jk} < Z_{ij} \le \alpha_{j,k+1}, \quad j \in \mathcal{O} \tag{2}$$

where $k = 0, \ldots m_j - 1$ in Equation (2), f_j is a given unknown bijective function if the jth variable is continuous, p_j is the unknown marginal probability of $X_{ij} = 1$ if the jth variable is binary, m_j is the number of categories of the jth variable if it is ordinal, and $\alpha_{j0}, \ldots, \alpha_{jm_j}$ are bounds for the jth variable if it is ordinal with $\alpha_{j0} = -\infty$ and $\alpha_{jm_j} = \infty$.

The interpretation of the models is that the continuous variables are transformed normal distributed variables, the binary variables are thresholded normal distributed variables, and the ordinal variables are normal distributed variables which are cut into bins. The flexibility of the copula model is due to the few assumptions about the f_j's for the continuous variables. Thus, the marginal distribution for each variable can be very complex. The parametric assumption is the particular copula we use. Other copulas can be used but the Gaussian copula has computational advantages which we extensively use in Sections 4 and 5.

Zhao and Udell (2020a,b) fit the model by first estimating the marginals, i.e. estimating the f_j's for the continuous variables using rescaled empirical CDFs, estimating the p_j's for the binary variables, and estimating the borders $\alpha_{j1}, \ldots, \alpha_{j,m_j-1}$ for the ordinal variables. They then estimate $\boldsymbol{\Sigma}$ conditional on the marginal distributions. This two-stage method is often referred to as IFM in the fully parametric case without the continuous variables.

Joe (2005) shows that IFM has a high relative efficiency compared with the maximum likelihood estimate (MLE) of all parameters (that is, joint estimation of the marginals and $\boldsymbol{\Sigma}$) in related models. However, the efficiency tends to decrease as the number of variables increase or when the dependence is high. Therefore, we perform joint estimation of some of the marginal distribution parameters which Zhao and Udell (2020a,b) fix in the second step of IFM. In particular, we can let Z_{ij} for a binary variable have a non-zero mean given by $\mu_j = \Phi^{-1}(p_j)$ and assume that $X_{ij} = 1$ if $Z_{ij} > 0$. We then jointly estimate μ_j and $\boldsymbol{\Sigma}$ for $j \in \mathcal{B}$. We still use a two-step procedure where we fix the borders for the ordinal variables (the αs) and estimate the f_j's non-parametrically.[2]

Let $\boldsymbol{X}_{i\mathcal{I}} = (x_{l_1}, \ldots, x_{l_k})^\top$ where $\mathcal{I} = \{l_1, \ldots, l_k\}$ and let $\hat{f}_j$ be the estimate of f_j. Let $\hat{\boldsymbol{z}}_{i\mathcal{C}}$ be a vector with $\hat{z}_{ij} = \hat{f}_j^{-1}(x_{ij})$. Then the log marginal likelihood conditional on

---

2. It is also possible to jointly estimate the borders for the ordinal variables and the f_j's if we parameterize them. We discuss this further in the discussion in Section 6.

the estimated marginals and disregarding the determinant of the Jacobian matrix of the transformation which does not depend on $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$ is $l(\mathbf{\Sigma}, \boldsymbol{\mu}) = \sum_{i=1}^{n} l_i(\mathbf{\Sigma}, \boldsymbol{\mu})$ where

$$l_i(\mathbf{\Sigma}, \boldsymbol{\mu}) = \log \phi^{(|\mathcal{C}|)}(\widehat{\boldsymbol{z}}_{i\mathcal{C}}; \mathbf{0}, \mathbf{\Sigma}_{\mathcal{C}\mathcal{C}}) + \log \Phi^{(|\mathcal{B}\cup\mathcal{O}|)}\bigg(\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{\mu}_{\mathcal{B}\cup\mathcal{O},\mathcal{C}} + \mathbf{\Sigma}_{\mathcal{B}\cup\mathcal{O},\mathcal{C}} \mathbf{\Sigma}_{\mathcal{C}\mathcal{C}}^{-1} \widehat{\boldsymbol{z}}_{i\mathcal{C}}, \tag{3}$$

$$\mathbf{\Sigma}_{\mathcal{B}\cup\mathcal{O},\mathcal{B}\cup\mathcal{O}} - \mathbf{\Sigma}_{\mathcal{B}\cup\mathcal{O},\mathcal{C}} \mathbf{\Sigma}_{\mathcal{C},\mathcal{C}}^{-1} \mathbf{\Sigma}_{\mathcal{C},\mathcal{B}\cup\mathcal{O}} \bigg).$$

$\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ depend on the estimated borders for the ordinal variables along with the observed variables, $\boldsymbol{x}_{\mathcal{B}\cup\mathcal{O}}$, as explained by Zhao and Udell (2020a) and in supplementary material S1, and $\boldsymbol{\mu}_{\mathcal{B}\cup\mathcal{O},\mathcal{C}}$ contains the possibly non-zero means for the binary variables. The $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ entries for the binary (and later multinomial variables) are $-\infty$ and 0 or 0 and ∞, respectively. See supplementary material S1 for further details and derivations.

4. New Estimation and Imputation Method

The main computational burden in evaluating the log marginal likelihood in Equation (3) is to approximate the $|\mathcal{B} \cup \mathcal{O}|$ dimensional CDF. Zhao and Udell (2020b) state that direct optimization is hard because of the CDF and, therefore, use an approximation of the type suggested by Guo et al. (2015) in an AEM algorithm. However, Genz (1992); Genz and Bretz (2002, 2009) show that the CDF $\Phi^{(k)}(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\omega}, \mathbf{\Omega})$ can be approximated quickly using importance sampling using the importance distribution with density h given by

$$h(\boldsymbol{u}) = \prod_{j=1}^{k} \begin{cases} \phi(u_j)/w_j(\boldsymbol{u}) & \hat{a}_j(\boldsymbol{u}_{1:(j-1)}) < u_j < \hat{b}_j(\boldsymbol{u}_{1:(j-1)}) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$w_j(\boldsymbol{u}) = \Phi(\hat{b}_j(\boldsymbol{u}_{1:(j-1)})) - \Phi(\hat{a}_j(\boldsymbol{u}_{1:(j-1)}))$$

$$\hat{b}_j(\boldsymbol{x}) = \begin{cases} o_{11}^{-1}(b_1 - \omega_1) & j = 1 \\ o_{jj}^{-1}(b_j - \omega_j - \boldsymbol{o}_{1:(j-1),j}^{\top}\boldsymbol{x}) & j > 1 \end{cases}$$

$$\hat{a}_j(\boldsymbol{x}) = \begin{cases} o_{11}^{-1}(a_1 - \omega_1) & j = 1 \\ o_{jj}^{-1}(a_j - \omega_j - \boldsymbol{o}_{1:(j-1),j}^{\top}\boldsymbol{x}) & j > 1 \end{cases}$$

where $\mathbf{O}$ is a Cholesky decomposition of $\mathbf{\Omega}$ such that $\mathbf{O}^{\top}\mathbf{O} = \mathbf{\Omega}$. Thus, we use a method, which is a Monte Carlo (MC) approximation of

$$\Phi^{(k)}(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\omega}, \mathbf{\Omega}) = \int h(\boldsymbol{u}) \prod_{j=1}^{k} w_j(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}.$$

Genz (1992); Genz and Bretz (2002) use a heuristic variable reordering to reduce the variance of the estimator at a fixed number of samples and use RQMC to get a better bound on the error. The advantage of using RQMC is that the error is bounded by $\mathcal{O}\left(s^{-1+\epsilon}\right)$, or more precisely $\mathcal{O}\left(s^{-1}(\log s)^l\right)$ for some $l \le k$, where s is the number of samples (Caflisch, 1998). This is in contrast to the $\mathcal{O}\left(s^{-1/2}\right)$ bound of MC methods.

Gradient approximations with respect to the mean and covariance matrix of the log of the CDF can be written as

$$\frac{\int \boldsymbol{g}(\mathbf{O}^\top \boldsymbol{u} + \boldsymbol{\omega}; \boldsymbol{\omega}, \boldsymbol{\Omega})h(\boldsymbol{u})\prod_{j=1}^k w_j(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}}{\Phi^{(k)}(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\omega}, \boldsymbol{\Omega})} \tag{5}$$

for a given function $\boldsymbol{g}$ as described by Hajivassiliou et al. (1996) for the Geweke, Hajivassiliou, and Keane (GHK) simulator they use. This is the expectation of $\boldsymbol{g}(\boldsymbol{X})$ where $\boldsymbol{X}$ follows a TMVN with location parameter $\boldsymbol{\omega}$, scale parameter $\boldsymbol{\Omega}$, and truncated such that $a_j < X_j < b_j$ for $j = 1, \ldots, k$. We have rewritten the Fortran code by Genz and Bretz (2002); Genz et al. (2020) in C++ to also provide an approximation to the numerator in Equation (5). Details of the method are provided in supplementary material S5. Standard applications of the chain rule can then be used to get an approximation of the gradient of the log marginal likelihood in Equation (3), once a gradient approximation of the CDF is implemented. Details are provided in supplementary material S4. The computational complexity of all our approximations are $\mathcal{O}\left(nK^3\right)$ at a fixed number of RQMC samples like the AEM method.

Masarotto and Varin (2012) and Nikoloulopoulos (2016) use MC or RQMC forms of the importance sampler developed by Genz and Bretz (2002) to estimate Gaussian copula models.[3] However, they all compute the derivatives with finite difference which scales very poorly in the number of variables. To see this, both our approximation of $\Phi^{(k)}$ and the gradient with respect to the mean and covariance matrix have $\mathcal{O}\left(k^3\right)$ computational complexity as shown in supplementary material S5. There are $k(k+3)/2$ parameters so computing the gradient with finite difference has $\mathcal{O}\left(k^5\right)$ complexity in addition to being less precise. Thus, implementation of the gradient is crucial for fast estimation.

The model can be estimated by using a log Cholesky decomposition (Pinheiro and Bates, 1996) of $\boldsymbol{\Sigma}$. Stochastic gradient descent methods are easy to apply, because of the n independent log marginal likelihood terms, if one re-scales $\boldsymbol{\Sigma}$ to be a correlation matrix between each iteration, as Zhao and Udell (2020a) do. We have tried ADAM (Kingma and Ba, 2015) and stochastic variance reduced gradient (SVRG) (Johnson and Zhang, 2013). The latter seems to work well with an appropriate learning rate. We have also implemented an augmented Lagrangian method to avoid the ad hoc re-scaling. On average, the augmented Lagrangian method tends to provide slightly better estimates of $\boldsymbol{\Sigma}$ than SVRG. We however omit this from our comparisons because it is slower.

For the imputations, we use the conditional means for the latent variables (the Z_{ij}'s) corresponding to missing continuous variables (a missing variable with $j \in \mathcal{C}$) and map back using $\hat{f}_j$'s. This is similar to Zhao and Udell (2020a), but our method can yield an arbitrarily precise approximation of the conditional means. We use either conditional probabilities or medians for each of the binary, ordinal, and later multinomial variables. These quantities are given by a suitable choice of $\boldsymbol{g}$ in Equation (5).[4]

---

3. Masarotto and Varin (2012) state that the GHK simulator is different from the approximation developed by Genz and Bretz (2002). They are equivalent although Genz and Bretz (2002) use RQMC and a heuristic variable reordering which drastically reduces the variance in some cases (Genz and Bretz, 2009). Both are improvements in most cases.

4. The means can be computed by using the identity function and the conditional probabilities can be computed with a function which returns an one-hot vector which has a one in the category that the sampled $\boldsymbol{z}_i$ implies. The conditional median for the ordinal variables can be computed from the conditional probabilities.

Table 1: Mean classification error, RMSE, and SMAE each with plus or minus two standard errors for the first simulation study. The error rows are the mean classification errors for the ordinal variables and the mean RMSEs of the continuous variables. The last two columns show the mean computation times in seconds and relative errors for the correlation matrix estimates. The best result in each comparison is in bold.

Metric	Method	Binary	Ordinal	Continuous	Time	Relative Σ error
Error	RQMC (our)	$\mathbf{0.2429 \pm 0.0009}$	$\mathbf{0.582 \pm 0.001}$	$\mathbf{0.741 \pm 0.002}$	$\mathbf{184 \pm 4}$	$\mathbf{0.0812 \pm 0.0004}$
	Median (our)		0.602 ± 0.001			
	AEM (ZU)	0.2502 ± 0.0009	0.615 ± 0.001	0.750 ± 0.002	412 ± 2	0.1172 ± 0.0006
SMAE	RQMC (our)	$\mathbf{0.4861 \pm 0.0018}$	0.694 ± 0.003	$\mathbf{0.705 \pm 0.002}$		
	Median (our)		$\mathbf{0.650 \pm 0.002}$			
	AEM (ZU)	0.5008 ± 0.0018	0.659 ± 0.002	0.710 ± 0.002		

Thus, we are also able to approximate the quantities needed for imputations with the new C++ code for CDF approximation. Zhao and Udell (2020a) transform back their approximate conditional means for the binary and ordinal variables. Their method cannot directly be used to provide conditional probabilities. Finally, as Zhao and Udell (2020a,b) do, we assume that data is missing completely at random and leave handling of data which is missing at random for future work.

As Equation (5) is the expectation of $\boldsymbol{g}(\boldsymbol{X})$ where $\boldsymbol{X}$ follows a TMVN, one could directly sample from the TMVN to avoid the separate computation of the denominator and numerator. However, sampling from a TMVN is hard. Interestingly, Botev (2017) has recently extended the work by Genz (1992); Genz and Bretz (2002) to sample from a TMVN using an accept-reject sampling schema based on a minimax tilting method with a tilted version of Equation (4). It will be interesting to apply the importance sampler or accept-reject sampling schema by Botev (2017) in our application but this is beyond the scope of the present paper.

4.1. Application of the New Method

We now address how our new estimation method and imputation method compare with the AEM based method of Zhao and Udell (2020a), when (i) the model is correctly specified, and (ii) when the methods are applied to observational data. We start with (i). An argument for using the AEM method is that it is fast even when there are moderately many variables (moderately large K) and many observations (large n). We draw $n = 10000$ observations with $K = 60$ variables of which 20 each are continuous, binary, and ordinal. As in Zhao and Udell (2020a), we use five equally likely categories for the ordinal variables and generate 100 data sets where we mask each variable independently at random with a 30 percent chance. For each data set, we sample $\boldsymbol{\Sigma}$ from a Wishart distribution with K degrees of freedom with an identity matrix as the scale matrix and scale the sampled matrix such that it is a correlation matrix. Thus, the methods are compared on different correlation matrices. As in Zhao and Udell (2020a), we let the continuous variables have standard exponential (marginal) distributions. Throughout the paper, we use a batch size 100, make at most 50 passes through the data, and use a learning rate of 0.001 with our SVRG implementation. However, it is perceivable that the configuration can be tuned to each data set.
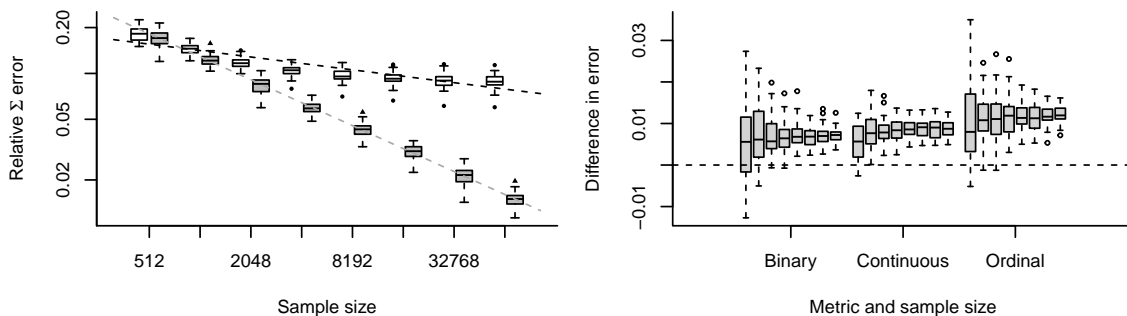
Figure 1: The left plot shows the relative error of the covariance matrix versus the sample size on the log-log scale. The gray boxes are the RQMC method and the white boxes are the AEM method. Regression lines are added on the log-log scale. The right plot shows the difference in mean RMSE and classification error for the three types of data for increasing sample sizes (increasing from left to right). We use the median method for the ordinal outcomes with RQMC as this is the closest to the AEM method.

We use three different metrics to measure the performance of the methods. The first two are also used by Zhao and Udell (2020a). The user may be interested in $\boldsymbol{\Sigma}$ per se and, thus, we use the relative error $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F / \|\boldsymbol{\Sigma}\|_F$, where $\widehat{\boldsymbol{\Sigma}}$ is an estimated correlation matrix and $\|\cdot\|_F$ is the Frobenius norm. We use the scaled mean absolute error (SMAE) given by $\sum_{i \in \mathcal{H}_j} \|\widehat{X}_{ij} - X_{ij}\|_1 / \sum_{i \in \mathcal{H}_j} \|\tilde{X}_j - X_{ij}\|_1$ where $\mathcal{H}_j$ is the set of indices with missing entries for variable j, $\widehat{X}_{ij}$ is an imputed value, and $\tilde{X}_j$ is the median estimate based on the observed values. Finally, we use the classification error for the binary and ordinal variables and the root mean square error (RMSE) for the continuous variables.

Table 1 summaries the results of the simulation study. The RQMC rows are our method which performs imputation by assigning the category with the highest conditional probability. The "Median" rows also use our estimation method but selects the conditional median category similar to the AEM algorithm which back transforms an approximate mean. Our new method performs better on all metrics. Selecting the category with the highest conditional probability leads to a lower classification error but higher SMAE. The computation time of our method is competitive although we stress that the AEM implementation does not support computation in parallel.[5] Importantly, the average relative error for the correlation matrix is much lower.

For our second evaluation, we compare the methods on six data sets. We use four of the data sets where Zhao and Udell (2020a) show that the AEM method yields superior single imputation performance compared to state of the art non-parametric methods. These are

---

5. Our simulations are run on a Laptop with a Intel® Core™ i7-8750H CPU, our code is compiled with GCC 10.1.0, and we use four threads. We use between 500 and 10000 samples per log marginal likelihood term evaluation and 2000 to 20000 samples for the imputations in all examples in this article. This gave reliable estimate for similar sized data sets in preliminary experiments. The RQMC approximations are stopped early if the relative error is small. The method by Zhao and Udell (2020a) is installed from https://github.com/udellgroup/mixedgcImp.

Table 2: Error estimates (see Table 1) with four different data sets used by Zhao and Udell (2020a) and the two largest data sets from the catdata package. The average computation times in seconds are shown in the last column.

Data set	Metric	Method	Binary	Ordinal	Continuous	Time
Rent	Error	RQMC (our)	**0.1212 ± 0.0008**	**0.3702 ± 0.0039**	0.747 ± 0.004	**31.67 ± 0.20**
		Median (our)		0.3710 ± 0.0041		
		AEM (ZU)	0.1237 ± 0.0008	0.3716 ± 0.0039	**0.746 ± 0.004**	35.41 ± 0.65
	SMAE	RQMC (our)	**0.9721 ± 0.0048**	0.5102 ± 0.0053	**0.701 ± 0.003**	
		Median (our)		**0.5095 ± 0.0054**		
		AEM (ZU)	0.9974 ± 0.0016	0.5107 ± 0.0052	0.706 ± 0.002	
Medcare	Error	RQMC (our)	**0.3418 ± 0.0016**	**0.4398 ± 0.0010**	**0.984 ± 0.006**	**6.74 ± 0.27**
		Median (our)		0.4661 ± 0.0010		
		AEM (ZU)	0.3466 ± 0.0017	0.4647 ± 0.0010	0.987 ± 0.006	22.63 ± 0.11
	SMAE	RQMC (our)	**0.8020 ± 0.0046**	0.9823 ± 0.0011	**0.951 ± 0.001**	
		Median (our)		**0.9639 ± 0.0013**		
		AEM (ZU)	0.8138 ± 0.0038	0.9641 ± 0.0010	0.952 ± 0.001	
ESL	Error	RQMC (our)		0.4911 ± 0.0036		**1.80 ± 0.31**
		Median (our)		**0.4908 ± 0.0036**		
		AEM (ZU)		0.4919 ± 0.0035		2.05 ± 0.04
	SMAE	RQMC (our)		0.5309 ± 0.0053		
		Median (our)		**0.5238 ± 0.0053**		
		AEM (ZU)		0.5254 ± 0.0052		
LEV	Error	RQMC (our)		**0.6337 ± 0.0023**		**2.20 ± 0.33**
		Median (our)		0.6476 ± 0.0021		
		AEM (ZU)		0.6484 ± 0.0021		5.40 ± 0.07
	SMAE	RQMC (our)		0.9637 ± 0.0067		
		Median (our)		**0.8696 ± 0.0028**		
		AEM (ZU)		0.8714 ± 0.0027		
GBSG	Error	RQMC (our)	0.2895 ± 0.0031	0.3610 ± 0.0044	**0.921 ± 0.008**	**2.64 ± 0.35**
		Median (our)		0.3604 ± 0.0044		
		AEM (ZU)	**0.2889 ± 0.0031**	**0.3572 ± 0.0046**	0.925 ± 0.008	7.13 ± 0.15
	SMAE	RQMC (our)	0.7276 ± 0.0084	1.0207 ± 0.0060	0.879 ± 0.003	
		Median (our)		1.0189 ± 0.0061		
		AEM (ZU)	**0.7257 ± 0.0082**	**1.0085 ± 0.0041**	**0.878 ± 0.002**	
TIPS	Error	RQMC (our)	**0.2857 ± 0.0058**	**0.4133 ± 0.0072**	0.786 ± 0.016	1.67 ± 0.17
		Median (our)		0.4414 ± 0.0088		
		AEM (ZU)	0.2874 ± 0.0062	0.4489 ± 0.0091	**0.784 ± 0.016**	**1.56 ± 0.04**
	SMAE	RQMC (our)	**0.8241 ± 0.0142**	0.8007 ± 0.0174	0.763 ± 0.008	
		Median (our)		**0.7786 ± 0.0154**		
		AEM (ZU)	0.8284 ± 0.0139	0.7881 ± 0.0155	**0.761 ± 0.008**	

the lecturers evaluation (LEV) data set, employee selection (ESL) data set, German breast cancer study group (GBSG) data set, and restaurant tips (TIPS). We also add the two largest data sets from the catdata package (Fox et al., 2020). This is the medcare data set containing the number of physician office visits and the rent in Munich data set.[6] Both data sets have variables of all types (binary, ordinal, and continuous). We generate 100 data sets for each of the six where we independently at random mask each variable with a 30 percent chance and standardize all continuous variables such that the RMSE is not dominated by a few variables. The first four data sets have between 244 and 1000 observations while the latter two have 2053 and 4406 observations. Further details about all data sets in this paper can be found in supplementary material S6.

Table 2 summarises the results. With respect to the standard errors there are only noticeable differences in the average errors for the two larger data sets where our method is preferred. A reason for there seemingly being no differences for the smaller data sets can be that the AEM method tends to select a correlation matrix close to a diagonal matrix. The average differences between the Frobenius norm of the estimated correlation matrix

---

6. We remove the rent per square meter because of the deterministic relationship with rent and the living space and we remove the municipality code as it is a multinomial variable.

of the two methods ranged from 0.0123 to 0.691 over the six data sets. This might lead to bias-variance trade-off.

To explore this further, we have conducted a new simulation study with $K = 15$ variables and 5 of each type of outcome to reduce the computation time. We use $n = 512, 1024, \ldots, 65536$ observations with 50 data sets for each n. The results are summarized in Figure 1 which shows that the AEM method does not improve in terms of the relative error of the estimated covariance matrix (or at best very slowly) for large sample sizes. This makes it hard to believe that the AEM method will converge, which implies that it must be biased for some sample sizes. This is troubling if the researcher is interested in the covariance matrix, as in Figure 6 in Zhao and Udell (2020a). In contrast, our proposed method has a regression slope of -0.5006 consistent with the typical $\sqrt{n}$ convergence rate of maximum likelihood estimators. Figure 1 also shows that the gap in imputation error increases as a function of the sample size.

To summarize, we find large improvements when the model is correctly specified and only improvements or identical results with observational data sets. There is a big difference in the average error for the estimated correlation matrix for the correctly specified model which is very important for researchers who are interested in the correlation matrix per se. Importantly, the differences in the error of the two methods increases as the number of samples increase which is consistent with the observational data. This suggests that the RQMC is preferable when there are many observations unlike the data sets presented here which have at most 4406 observations.

5. Adding Multinomial Variables

We extend the model in this section to also support multinomial variables. To build up the intuition, we first show the formulation with a simple model with one binary variable $X_1 \in \{0, 1\}$ and one continuous variable X_2. Suppose that

$$\boldsymbol{Z} \sim N^{(3)} \left(\begin{pmatrix} 0 \\ \mu_2 \\ 0 \end{pmatrix}, \begin{pmatrix} \xi & 0 & 0 \\ 0 & 1 & \sigma_{23} \\ 0 & \sigma_{32} & 1 \end{pmatrix} \right) \tag{6}$$

$$X_1 = j - 1 \Leftrightarrow Z_j \geq \max(Z_1, Z_2) \qquad j \in \{1, 2\} \tag{7}$$
$$X_2 = f(Z_3)$$

for a bijective function f and $\mu_2 = \Phi^{-1}(p_1)$. Using that $\boldsymbol{Z}_{1:2} \mid Z_3 = z \sim N(\boldsymbol{m}, \mathbf{S})$ with $\boldsymbol{m} = (0, \mu_2 + \sigma_{23}z)^\top$ and

$$\mathbf{S} = \begin{pmatrix} \xi & 0 \\ 0 & 1 - \sigma_{23}^2 \end{pmatrix},$$

we get

$$P(X_1 = 1 \mid Z_3 = z) = \Phi(0; -m_2, \xi + s_{22}) = \Phi \left(\frac{\mu_2 + \sigma_{23}z}{\sqrt{1 + \xi - \sigma_{23}^2}} \right)$$

where the result is attained using the identity in supplementary material S2. This is the Gaussian copula model we have been working with if we let $\xi \to 0^+$ which effectively means

that $Z_1 = 0.$[7] In practice, we derive the formulas we need and work with $\xi = 0$ and we only work with one latent variable for binary variables as Zhao and Udell (2020a).

Now, consider the extension to the scenario with one multinomial variable $X_1 \in \{0, \cdots, G-1\}$, and L continuous variables. Assume that $\boldsymbol{Z} \sim N^{(L+G)}(\boldsymbol{\mu}, \boldsymbol{\Psi})$ where

$$\boldsymbol{\Psi} = \begin{pmatrix} \xi & \boldsymbol{0}_{G-1}^\top & \boldsymbol{0}_L^\top \\ \boldsymbol{0}_{G-1} & \boldsymbol{\Psi}_{2:G2:G} & \boldsymbol{\Psi}_{2:G(-1:G)} \\ \boldsymbol{0}_L & \boldsymbol{\Psi}_{(-1:G)2:G} & \boldsymbol{\Psi}_{(-1:G)(-1:G)} \end{pmatrix}$$

$$\boldsymbol{\Psi}_{2:G2:G} = \begin{pmatrix} 1 & \sigma_{23} & \cdots & \sigma_{2K} \\ \sigma_{32} & \sigma_{33} & \ddots & \sigma_{3G} \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{G2} & \sigma_{G3} & \cdots & \sigma_{GG} \end{pmatrix}$$

$$X_1 = j - 1 \Leftrightarrow Z_j \geq \max(Z_1, \ldots, Z_G)$$

$\boldsymbol{X}_{(-1)} = \boldsymbol{f}^{-1}(\boldsymbol{Z}_{(-1:G)})$, $j \in \{1, \ldots, G\}$, $\boldsymbol{\mu} = (0, \boldsymbol{\mu}_{2:G}^\top, \boldsymbol{0}_L^\top)^\top$, $\boldsymbol{\Psi}_{(-1:G)2:G} \in \mathbb{R}^{L \times (G-1)}$ is a dense matrix, and $\boldsymbol{\Psi}_{(-1:G)(-1:G)} \in \mathbb{R}^{L \times L}$ is restricted to have ones in the diagonal.[8] This parameterization is used by Bunch (1991), without the continuous variables. We have G conditional probabilities given by

$$P(X_1 = j - 1 \mid \boldsymbol{Z}_{(-1:G)} = \boldsymbol{z}) = \int \phi(u; m_j, s_{jj}) \Phi^{(G-1)} \Big($$

$$\boldsymbol{1}_{G-1}u; \big(\boldsymbol{m}_{(-j)} + \mathbf{S}_{(-j)j}\frac{u - m_j}{s_{jj}}\big), \big(\mathbf{S}_{(-j)(-j)} - s_{jj}^{-1}\mathbf{S}_{(-j)j}\mathbf{S}_{j(-j)}\big)\Big) \, \mathrm{d}u$$

which equals

$$\Phi^{(G-1)} \Big(\boldsymbol{0}_{G-1}; \boldsymbol{m}_{(-j)} - \boldsymbol{1}_{G-1}m_j, \mathbf{S}_{(-j)(-j)} + s_{jj}\boldsymbol{1}_{G-1}\boldsymbol{1}_{G-1}^\top - \boldsymbol{1}_{G-1}\mathbf{S}_{j(-j)} - \mathbf{S}_{(-j)j}\boldsymbol{1}_{G-1}^\top\Big)$$

where $\boldsymbol{Z}_{1:G} \mid \boldsymbol{Z}_{(-1:G)} = \boldsymbol{z} \sim N(\boldsymbol{m}, \mathbf{S})$,

$$\boldsymbol{m} = \boldsymbol{\mu}_{1:G} + \boldsymbol{\Sigma}_{1:G(-1:G)}\boldsymbol{\Sigma}_{(-1:G)(-1:G)}^{-1}\boldsymbol{z}$$
$$\mathbf{S} = \boldsymbol{\Sigma}_{1:G,1:G} - \boldsymbol{\Sigma}_{1:G(-1:G)}\boldsymbol{\Sigma}_{(-1:G)(-1:G)}^{-1}\boldsymbol{\Sigma}_{(-1:G)1:G},$$

$\boldsymbol{1}_k$ is a vector with k ones, and $(-j)$ or $(-\mathcal{J})$ in a subscript implies all elements but the jth or those indices in $\mathcal{J}$ for vectors or all but those indices in the rows or columns for matrices.

5.1. New Log Marginal Likelihood

We can now turn to the general case with all types including multiple multinomial variables and show the log marginal likelihood term for each observation. Suppose that the entries $X_{ij} \in \{0, 1, \ldots m_j - 1\}$ with $j \in \mathcal{M} \subseteq \{1, \ldots, K\}$ are multinomial variables with m_j

---

7. To see this, compare Equation (7) with $Z_1 = 0$ with Equation (1) and use $\mu_2 = \Phi^{-1}(p_1)$.
8. We arbitrarily select the first category to be the reference and the corresponding $\boldsymbol{Z}$ entry is uncorrelated with all other $\boldsymbol{Z}$ entries.

categories with $\mathcal{M} \cap \mathcal{C} = \emptyset$, $\mathcal{M} \cap \mathcal{O} = \emptyset$, and $\mathcal{B} = \{1, \ldots, K\} \setminus \mathcal{C} \cup \mathcal{O} \cup \mathcal{M}$. The observed variables, the $\boldsymbol{X}_i$'s, are ordered such that the multinomial variables are first without loss of generality and let $c_{\mathcal{M}} = |\mathcal{M}|$ and $c_{\setminus\mathcal{M}} = K - c_{\mathcal{M}}$. The model is then $\boldsymbol{Z}_i \sim N^{(W)}(\boldsymbol{\mu}, \boldsymbol{\Psi})$ where $W = c_{\setminus\mathcal{M}} + \sum_{j=1}^{c_{\mathcal{M}}} m_j$, $\boldsymbol{\Psi}$ is given in supplementary material S3, and $|\mathcal{B}| + \sum_{j=1}^{c_{\mathcal{M}}} m_j - c_{\mathcal{M}}$ entries of $\boldsymbol{\mu}$ are non-zero. One can permute $\boldsymbol{\Psi}$ with a permutation matrix $\mathbf{P}$ such that:

$$\mathbf{P}\boldsymbol{\Psi}\mathbf{P}^\top = \begin{pmatrix} \xi\mathbf{I}_{c_{\mathcal{M}}} & \mathbf{0}_{c_{\mathcal{M}}\times(W-c_{\mathcal{M}})} \\ \mathbf{0}_{(W-c_{\mathcal{M}})\times c_{\mathcal{M}}} & \boldsymbol{\Sigma} \end{pmatrix} \tag{8}$$

where $\mathbf{I}_k$ is the k dimensional identity matrix, $\mathbf{0}_{k\times l}$ is a $k \times l$ matrix with zeros, and $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix where all elements are free except for K of the diagonal entries which are restricted to be one. Thus, we parameterize $\boldsymbol{\Psi}$ in terms of a log Cholesky decomposition of $\boldsymbol{\Sigma}$. As before, the restriction in the diagonal of $\boldsymbol{\Sigma}$ is either applied after each stochastic gradient iteration by scaling the rows and columns, similar to what Zhao and Udell (2020a) do, or by using a constrained nonlinear optimization method.

We now turn to the new expression of the log marginal likelihood. Let

$$\mathbf{D} = \begin{pmatrix} \mathbf{1}_{m_1-1} & \mathbf{0}_{m_1-1} & \cdots & \mathbf{0}_{m_1-1} \\ \mathbf{0}_{m_2-1} & \mathbf{1}_{m_2-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0}_{m_{c_{\mathcal{M}-1}-1}} \\ \mathbf{0}_{m_{c_{\mathcal{M}}}-1} & \cdots & \mathbf{0}_{m_{c_{\mathcal{M}}}-1} & \mathbf{1}_{m_{c_{\mathcal{M}}}-1} \\ \mathbf{0}_{|\mathcal{O}\cup\mathcal{B}|} & \cdot & \mathbf{0}_{|\mathcal{O}\cup\mathcal{B}|} & \mathbf{0}_{|\mathcal{O}\cup\mathcal{B}|} \end{pmatrix}.$$

Furthermore, we let $\mathcal{I}_i$ be the set with the indices of the latent variables belonging to each of the observed categories for the multinomial variables for observation i. That is $\mathcal{I}_i = \{k = 1, \ldots, c_{\mathcal{M}} : X_{ik} + 1 + \sum_{l=1}^{k-1} m_l\}$ where $\sum_{l=1}^{0} m_l = 0$ by definition.

Let $\tilde{\mathcal{C}}$ be the indices of the latent variables corresponding to the observed continuous variables and similarly define $\tilde{\mathcal{O}}$, $\tilde{\mathcal{B}}$, and $\tilde{\mathcal{M}}$. Using

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_{\tilde{\mathcal{B}}\cup\tilde{\mathcal{O}}\cup\tilde{\mathcal{M}}} + \boldsymbol{\Psi}_{\tilde{\mathcal{B}}\cup\tilde{\mathcal{O}}\cup\tilde{\mathcal{M}},\tilde{\mathcal{C}}} \boldsymbol{\Psi}_{\tilde{\mathcal{C}}\tilde{\mathcal{C}}}^{-1} \widehat{\boldsymbol{z}}_{i\tilde{\mathcal{C}}} \tag{9}$$

$$\bar{\mathbf{S}} = \boldsymbol{\Psi}_{\tilde{\mathcal{B}}\cup\tilde{\mathcal{O}}\cup\tilde{\mathcal{M}},\tilde{\mathcal{B}}\cup\tilde{\mathcal{O}}\cup\tilde{\mathcal{M}}} - \boldsymbol{\Psi}_{\tilde{\mathcal{B}}\cup\tilde{\mathcal{O}}\cup\tilde{\mathcal{M}},\tilde{\mathcal{C}}} \boldsymbol{\Psi}_{\tilde{\mathcal{C}},\tilde{\mathcal{C}}}^{-1} \boldsymbol{\Psi}_{\tilde{\mathcal{C}},\tilde{\mathcal{B}}\cup\tilde{\mathcal{O}}\cup\tilde{\mathcal{M}}}, \tag{10}$$

and applying the identity in supplementary material S2 yields the log marginal likelihood

$$l_i(\boldsymbol{\Psi}, \boldsymbol{\mu}) = \log \phi^{(|\mathcal{C}|)}(\widehat{\boldsymbol{z}}_{i\tilde{\mathcal{C}}}; \mathbf{0}, \boldsymbol{\Psi}_{\tilde{\mathcal{C}}\tilde{\mathcal{C}}}) + \log \int \phi^{(c_{\mathcal{M}})}(\boldsymbol{u}; \bar{\boldsymbol{\mu}}_{\mathcal{I}_i}, \bar{\mathbf{S}}_{\mathcal{I}_i\mathcal{I}_i}) \Phi^{(W)}\Big(\boldsymbol{a}_i, \boldsymbol{b}_i; \tag{11}$$

$$\big(\bar{\boldsymbol{\mu}}_{(-\mathcal{I}_i)} - \mathbf{D}\boldsymbol{u} + \mathbf{S}_{(-\mathcal{I}_i)\mathcal{I}_i}\bar{\mathbf{S}}_{\mathcal{I}_i\mathcal{I}_i}^{-1}(\boldsymbol{u} - \bar{\boldsymbol{\mu}}_{\mathcal{I}_i})\big), \big(\bar{\mathbf{S}}_{(-\mathcal{I}_i)(-\mathcal{I}_i)} - \bar{\mathbf{S}}_{(-\mathcal{I}_i)\mathcal{I}_i}\bar{\mathbf{S}}_{\mathcal{I}_i\mathcal{I}_i}^{-1}\bar{\mathbf{S}}_{\mathcal{I}_i(-\mathcal{I}_i)}\big)\Big) \mathrm{d}\boldsymbol{u}$$

$$= \log \phi^{(|\mathcal{C}|)}(\widehat{\boldsymbol{z}}_{i\mathcal{C}}; \mathbf{0}, \boldsymbol{\Psi}_{\tilde{\mathcal{C}}\tilde{\mathcal{C}}}) + \log \Phi^{(W)}\Big($$

$$\boldsymbol{a}_i, \boldsymbol{b}_i; \big(\bar{\boldsymbol{\mu}}_{(-\mathcal{I}_i)} - \mathbf{D}\bar{\boldsymbol{\mu}}_{\mathcal{I}_i}\big), \big(\bar{\mathbf{S}}_{(-\mathcal{I}_i)(-\mathcal{I}_i)} + \mathbf{D}\bar{\mathbf{S}}_{\mathcal{I}_i\mathcal{I}_i}\mathbf{D}^\top - \mathbf{D}\bar{\mathbf{S}}_{\mathcal{I}_i(-\mathcal{I}_i)} - \bar{\mathbf{S}}_{(-\mathcal{I}_i)\mathcal{I}_i}\mathbf{D}^\top\big)\Big).$$

A missing X_{ij} implies that the corresponding $\boldsymbol{Z}_i$ entry or entries are unrestricted. Thus, they do not contribute anything to the log marginal likelihood in Equation (3) and (11) after integrating them out and can be omitted. Notice that the non-zero mean terms, some of the terms of $\boldsymbol{\mu}_{\tilde{\mathcal{B}}\cup\tilde{\mathcal{O}}\cup\tilde{\mathcal{M}}}$, enter linearly in Equation (9). This makes it easy to jointly optimize the mean and covariance matrix, rather than fixing the mean as suggested by Zhao and Udell (2020a), as we have a gradient approximation of the mean of the CDF. Thus, we can easily estimate the mean vector for the binary and multinomial variables together with $\boldsymbol{\Sigma}$.

Table 3: Average RMSEs and classification errors on simulated data sets and three observational data sets. The average computation times in seconds are given in the last column.

Data set	Method	Binary	Ordinal	Continuous	Multinomial	Time
Simulation	RQMC (our)	**0.3051 ± 0.0064**	**0.647 ± 0.006**	**0.883 ± 0.013**	**0.535 ± 0.005**	47.05 ± 0.99
	missForest	0.3451 ± 0.0065	0.692 ± 0.006	0.949 ± 0.011	0.603 ± 0.006	**8.72 ± 0.25**
	imputeFAMD	0.3573 ± 0.0074	0.701 ± 0.006	0.929 ± 0.011	0.613 ± 0.007	81.19 ± 0.38
Cholesterol	RQMC (our)	0.3009 ± 0.0012	**0.679 ± 0.002**		**0.538 ± 0.002**	10.64 ± 0.83
	missForest	0.4022 ± 0.0046	0.696 ± 0.003		0.700 ± 0.008	**2.88 ± 0.12**
	imputeFAMD	**0.2994 ± 0.0011**	0.691 ± 0.002		0.564 ± 0.003	5.61 ± 0.27
Rent	RQMC (our)	**0.1096 ± 0.0007**	**0.370 ± 0.004**	**0.731 ± 0.003**	**0.890 ± 0.002**	173.55 ± 0.54
	missForest	0.1408 ± 0.0009	0.389 ± 0.004	0.757 ± 0.004	0.903 ± 0.002	**19.07 ± 0.70**
	imputeFAMD	0.1121 ± 0.0008	0.537 ± 0.005	0.777 ± 0.003	0.959 ± 0.002	102.97 ± 2.24
Colon	RQMC (our)	**0.2217 ± 0.0014**	**0.223 ± 0.002**	1.029 ± 0.008	0.659 ± 0.004	28.73 ± 0.91
	missForest	0.2666 ± 0.0025	0.394 ± 0.005	1.060 ± 0.007	**0.614 ± 0.004**	**11.50 ± 0.44**
	imputeFAMD	0.2275 ± 0.0014	0.223 ± 0.002	**1.017 ± 0.007**	0.671 ± 0.004	51.42 ± 0.98

5.2. Application

We compare our method with other top performing single imputation methods in this section. The methods we compare with are missForest and the PCA-like imputeFAMD. We are particularly interested in (i) the improvements from using our method for a correctly specified model and (ii) the performance of all three methods on observational data. We start with (i) which will give an indication of the improvements we may expect when the data generating process is approximately a Gaussian copula.

For (i) and (ii) we generate 100 data sets and within each there is a 30 percent chance that a variable is missing (missing independently at random). The imputeFAMD has a tuning parameter which is the number of components. We estimate this using five-fold cross validation and include the cross validation time in the estimation time. The only tuning parameter for missForest is there number trees where more trees is better. We use 100 trees as Stekhoven and Buehlmann (2012). In our simulation study (i) we simulate data from a Gaussian copula in a similar manner to that described in Section 4.1. That is, we simulate $\mathbf{\Sigma}$ in $\mathbf{\Psi}$ in Equation (8) from a Wishart distribution like in the previous simulation. We have 8 variables per observation with 2 of each type. The ordinal and multinomial variables have five equally likely categories and there is a total 2000 observations.

The result of the simulation study is shown in the "Simulation" rows in Table 3.[9] Our method performs much better for all four types of variables. The average computation times for the simulated and observational data show that our method is the slowest of the three.[10] However, it is not orders of magnitude slower.

We include three observational data sets: the rent data set, used also in Section 4.1, but this time with the multinomial area code, the cholesterol data from a US survey from the survey package (Lumley, 2020), and data from one of the first successful trials of adjuvant chemotherapy for colon cancer from the survival package (Therneau, 2020). Our method

---

9. We do not include the SMAE as there is no obvious way to compute this for the multinomial variables, the integer values assigned to each category of the ordinal variables is arbitrary, and it favors the Gaussian copula models for the ordinal variables as the other methods treat ordinal variables as multinomial variables (especially if we use the conditional median similar to what Zhao and Udell (2020a) do as shown in Table 1).

10. We use four threads with missForest and imputeFAMD does not support computation in parallel.

often performs best or is close to the best performing method. It is expected that our method is not always the best performing method as we do make a parametric assumption with the particular copula we use. It is, however, encouraging that the data generating processes seem well approximated with the Gaussian copula for all three data sets.

6. Conclusion and Discussion

We have shown that direct maximum likelihood estimation of the model used by Zhao and Udell (2020a,b) using RQMC is feasible and provides improvements compared with the approximation that they use. Our RQMC method yielded lower errors for the imputed values, and estimated model parameters closer to the true values in our simulation study. Importantly, the error of our method decreases faster as a function of the number of samples. We have also extended the model to support multinomial variables which increases the applicability of the method.

6.1. Future Work

Our RQMC method can be extended to provide e.g. an arbitrarily precise approximation of the quantiles of the latent variables conditional on the observed data and estimated model parameters. A major advantage of having precise quantile approximations is that the user can get a quantile estimate at a given level which is close to the nominal level if the data generating process is well approximated by the Gaussian copula. These may provide substantial improvements over the lower bounds and approximate confidence intervals based on Gaussian approximations, that Zhao and Udell (2020b) suggest using, since the conditional distribution of the latent variables may be very non-Gaussian.

We could replace the non-parametric density estimator with a flexible parametric transformation like those used in transformation models (Hothorn et al., 2018). The model would be fully parametric model and it would be possible to do joint optimization over all parameters. This could increase the efficiency and subsequently the performance. Our RQMC could be extended to use the recently developed minimax tilting method by Botev (2017) which could decrease the computation time at a fixed variance of the estimator. Finally, our method can easily be used for multiple imputation (to, importantly, account for imputation uncertainty), since it is a probabilistic framework.

References

Vincent Audigier, François Husson, and Julie Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10 (1):5–26, 2014.

Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1): 125–148, 2017.

David S. Bunch. Estimability in the multinomial probit model. *Transportation Research Part B: Methodological*, 25(1):1 – 12, 1991.

Russel E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49, 1998.

Ruifei Cui, Ioan Gabriel Bucur, Perry Groot, and Tom Heskes. A novel Bayesian approach for latent variable modeling from mixed data with missing values. *Statistics and Computing*, 29(5):977–993, 2019.

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

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):405–421, 2017.

Huijie Feng and Yang Ning. High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 654–663. PMLR, 16–18 Apr 2019.

John Fox, Sanford Weisberg, and Brad Price. *carData: Companion to Applied Regression Data Sets*, 2020. R package version 3.0-4.

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992.

Alan Genz and Frank Bretz. Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, 11(4):950–971, 2002.

Alan Genz and Frank Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg, 2009. ISBN 978-3-642-01688-2.

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2020. R package version 1.0-12.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Graphical models for ordinal data. *Journal of Computational and Graphical Statistics*, 24(1):183–204, 2015.

Vassilis Hajivassiliou, Daniel McFadden, and Paul Ruud. Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results. *Journal of Econometrics*, 72(1):85 – 134, 1996. ISSN 0304-4076.

Torsten Hothorn, Lisa Möst, and Peter Bühlmann. Most likely transformations. *Scandinavian Journal of Statistics*, 45(1):110–134, 2018.

Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401 – 419, 2005. ISSN 0047-259X.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 315–323. Curran Associates, Inc., 2013.

Julie Josse and François Husson. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Eric Landgrebe, Yuxuan Zhao, and Madeleine Udell. Online mixed missing value imputation using Gaussian copula. In *ICML Workshop on the Art of Learning with Missing Values (Artemiss)*, 2020.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(80):2295–2328, 2009.

Thomas Lumley. survey: analysis of complex survey samples, 2020. R package version 4.0.

Guido Masarotto and Cristiano Varin. Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6(none):1517 – 1549, 2012.

Jared S. Murray, David B. Dunson, Lawrence Carin, and Joseph E. Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108 (502):656–665, 2013.

Aristidis K. Nikoloulopoulos. Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses. *Stochastic Environmental Research and Risk Assessment*, 30(2):493–505, 2016.

J.C. Pinheiro and D.M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6:289–296, 1996.

Daniel J. Stekhoven and Peter Buehlmann. Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Terry M Therneau. *A Package for Survival Analysis in R*, 2020. R package version 3.2-7.

Yuxuan Zhao and Madeleine Udell. Missing value imputation for mixed data via Gaussian copula. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 636–646, New York, NY, USA, 2020a. Association for Computing Machinery.

Yuxuan Zhao and Madeleine Udell. Matrix completion with quantified uncertainty through low rank Gaussian copula. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.