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

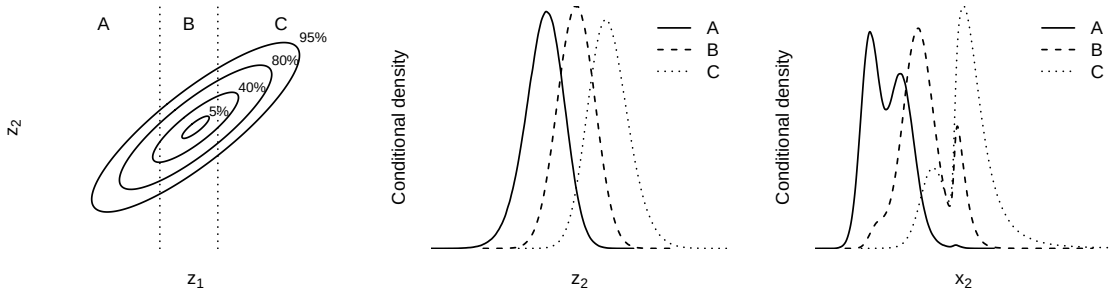

Figure 1: Illustration of the relationship between an ordinal and a continuous variable. The left plot shows contours of the joint distribution of the latent variables along with breakpoints for the ordinal variable X_1. The plot in the middle shows the conditional density of the latent variable Z_2 given the ordinal variable X_1. The right plot shows the conditional density of the continuous variable $X_2 = f_2(Z_2)$ given the ordinal variable X_1.

S1. Model Details and Derivation

The model will be explained in greater detail in this section. An ordinal variable X_{ij} with m_j categories and marginal probabilities $p_{j1}, \ldots, p_{jm_j}$ can be viewed as a normal distributed variable which is cut into m_j bins with breakpoints at $\alpha_{j0} = -\infty$ and $\Phi(\alpha_{jk}) = \sum_{l=1}^{k} p_{jl}$ for $k > 0$. Let Z_{ij} denote the latent normal distributed variable such that $X_{ij} = k \Leftrightarrow Z_{ij} \in (\alpha_{1,k-1}, \alpha_{1k})$. We assume that an ordinal variable X_{ij} is related to a continuous variable $X_{ij'}$ by assuming that $X_{ij'} = f_{j'}(Z_{ij'})$ where $f_{j'}$ is some bijective function and $(Z_{ij'}, Z_{ij})$ are jointly normal distributed. Such a relationship is illustrated in Figure 1.

Extensions to multiple ordinal variables are done by assuming that their latent variables are jointly normal distributed. Binary variables are a special case of ordinal variables where $m_j = 2$. Few assumptions are made about the f_js for the continuous variables yielding a very flexible model for the marginal distributions of the continuous variables. However, assumptions are made on the dependence between the variables through the particular copula we use.

Next, we derive the marginal likelihood which is partly shown in Equation (3). Without loss of generality, let the continuous variables have the last indices. Let $\boldsymbol{f}_{\mathcal{C}}^{-1}(\boldsymbol{x}) = (f_j^{-1}(x_j))_{j \in \mathcal{C}}$ be the inverse $\boldsymbol{f}_{\mathcal{C}}(\boldsymbol{z}) = (f_j(z_j))_{j \in \mathcal{C}}$. Then the density of $\boldsymbol{X}_i$ at $\boldsymbol{x}$ is

$$\begin{aligned} \exp \tilde{l}_i(\boldsymbol{\Sigma}, \boldsymbol{\mu}) = \phi^{(|\mathcal{C}|)}(\boldsymbol{f}_{\mathcal{C}}^{-1}(\boldsymbol{x}_{\mathcal{C}}); \boldsymbol{0}, \boldsymbol{\Sigma}_{\mathcal{C}\mathcal{C}}) |\det(\nabla \boldsymbol{f}_{\mathcal{C}}^{-1})(\boldsymbol{x}_{\mathcal{C}})| \\ \cdot P\left(\boldsymbol{Z}_{\mathcal{B}\cup\mathcal{O}} \in \mathcal{V}(\boldsymbol{x}_{\mathcal{B}\cup\mathcal{O}}) \middle| \boldsymbol{Z}_{\mathcal{C}} = \boldsymbol{f}_{\mathcal{C}}^{-1}(\boldsymbol{x}_{\mathcal{C}})\right) \end{aligned} \tag{1}$$

where

$$\mathcal{V}(\boldsymbol{u}) = \left\{ \boldsymbol{v} \in \mathbb{R}^{|\mathcal{B}\cup\mathcal{O}|} : u_j = k \Rightarrow \alpha_{j,k-1} < v_j \leq \alpha_{jk} \right\}$$

and $\det(\nabla \cdot)$ is the Jacobian determinant. Equation (3) follows by removing the determinant factor, $|\det(\nabla \boldsymbol{f}_{\mathcal{C}}^{-1})(\boldsymbol{x}_{\mathcal{C}})|$, which does not depend on $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$ and by using the special case of the conditional probability of $\boldsymbol{Z}_{\mathcal{B}\cup\mathcal{O}} \in \mathcal{V}(\boldsymbol{x}_{\mathcal{B}\cup\mathcal{O}})$.

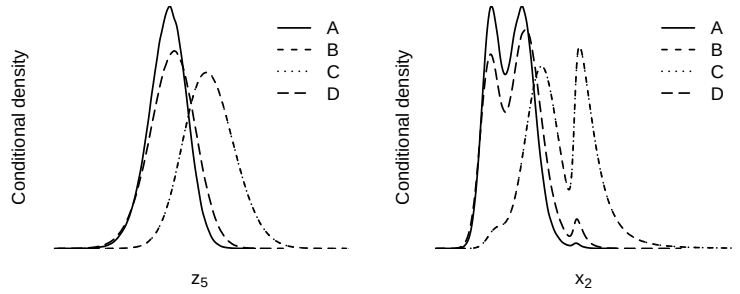

Figure 2: Conditional densities of a continuous variable and its latent variable given the value of a four level multinomial outcome. Two of the conditional densities coincide because the correlation with the latent variable for the continuous outcome and mean is the same for the corresponding latent variable for two of the multinomial categories. The transformation, f_2, is the same as in Figure 1.

Each ordinal variable has $m_j - 1$ free parameters to be estimated. As shown in Section 3, the free parameter for each binary variable is easily estimated jointly with the correlation matrix by parameterizing it in terms of the mean. The $m_j - 1$ free parameters for the ordinal variables with $m_j > 2$ can also be estimated jointly with the correlation matrix but this is not done in this paper. One can use a parametric form of f_j for the continuous variable which will allow for joint estimation of the transformation as mentioned in Section 6.1. Though, this will require that one keeps the determinant factor in Equation (1).

We end with an illustrative example of multinomial variables. Figure 2 shows the conditional distribution of a continuous variable given the value of a multinomial variable. As explained in Section 5, the first latent variable for a multinomial variable is fixed to zero and the remaining variables are allowed to have non-zero means and be correlated with other latent variables. Thus, the continuous variables can be associated with a multinomial variable and have the same conditional density for some outcomes of a multinomial variable as shown in Figure 2. This is not possible with the parameterization of an ordinal variable because of the assumed normal distribution of the latent variables and the monotone relationship between the latent variable of an ordinal variable and the observed category.

S2. Multivariate Normal CDF Identity

In this section, we show an identity which we will use repeatably. Let

$$\begin{pmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \end{pmatrix} \sim N^{(k_1+k_2)} \left(\begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Xi}_{11} & \boldsymbol{\Xi}_{12} \\ \boldsymbol{\Xi}_{21} & \boldsymbol{\Xi}_{22} \end{pmatrix} \right)$$

where $\boldsymbol{V}_1 \in \mathbb{R}^{k_1}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{k_2}$, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are mean vectors for $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$, respectively, and $\boldsymbol{\Xi}$ is a covariance matrix where the sub matrices have k_1 or k_2 rows and columns. Then the

joint density of $\boldsymbol{V}_1 = \boldsymbol{v}_1$ and $\boldsymbol{a} < \boldsymbol{V}_2 < \boldsymbol{b}$ (a box constraint on $\boldsymbol{V}_2$) is

$$\begin{aligned}
&\phi^{(k_1)}(\boldsymbol{v}_1; \boldsymbol{\xi}_1, \boldsymbol{\Xi}_{11}) \, \mathrm{P}\left(\boldsymbol{a} < \boldsymbol{V}_2 < \boldsymbol{b} \,|\, \boldsymbol{V}_1 = \boldsymbol{v}_1\right) \\
&\quad = \phi^{(k_1)}(\boldsymbol{v}_1; \boldsymbol{\xi}_1, \boldsymbol{\Xi}_{11}) \Phi^{(k_2)}\Big(\boldsymbol{a}, \boldsymbol{b}; \left(\boldsymbol{\xi}_2 + \boldsymbol{\Xi}_{21}\boldsymbol{\Xi}_{11}^{-1}(\boldsymbol{v}_1 - \boldsymbol{\xi}_1)\right), \left(\boldsymbol{\Xi}_{22} - \boldsymbol{\Xi}_{21}\boldsymbol{\Xi}_{11}^{-1}\boldsymbol{\Xi}_{12}\right)\Big)
\end{aligned}$$

and the marginal for $\mathrm{P}(\boldsymbol{a} < \boldsymbol{V}_2 < \boldsymbol{b})$ is

$$\begin{aligned}
&\mathrm{P}(\boldsymbol{a} < \boldsymbol{V}_2 < \boldsymbol{b}) \\
&\quad = \Phi^{(k_2)}(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\xi}_2, \boldsymbol{\Xi}_{22}) \\
&\quad = \int \phi^{(k_1)}(\boldsymbol{v}_1; \boldsymbol{\xi}_1, \boldsymbol{\Xi}_{11}) \, \mathrm{P}\left(\boldsymbol{a} < \boldsymbol{V}_2 < \boldsymbol{b} \,|\, \boldsymbol{V}_1 = \boldsymbol{v}_1\right) \mathrm{d}\boldsymbol{v}_1.
\end{aligned}$$

Next, we can define

$$\begin{aligned}
\mathbf{Z} &= \boldsymbol{\Xi}_{21}\boldsymbol{\Xi}_{11}^{-1} \\
\boldsymbol{c} &= \boldsymbol{\xi}_2 - \boldsymbol{\Xi}_{21}\boldsymbol{\Xi}_{11}^{-1}\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 - \mathbf{Z}\boldsymbol{\xi}_1 \\
\boldsymbol{\Omega} &= \boldsymbol{\Xi}_{22} - \boldsymbol{\Xi}_{21}\boldsymbol{\Xi}_{11}^{-1}\boldsymbol{\Xi}_{12} = \boldsymbol{\Xi}_{22} - \mathbf{Z}\boldsymbol{\Xi}_{11}\mathbf{Z}^\top
\end{aligned}$$

which we can use to show that

$$\int \phi^{(k_1)}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi^{(k_2)}(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{c} + \mathbf{Z}\boldsymbol{x}, \boldsymbol{\Omega}) \, \mathrm{d}\boldsymbol{x} = \Phi^{(k_2)}(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{c} + \mathbf{Z}\boldsymbol{\mu}, \boldsymbol{\Omega} + \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^\top). \tag{2}$$

S3. Parameterization for the General Model with More than One Multinomial Variable

Following the notation in Section 5.1, we can generalize the covariance matrix for $\boldsymbol{Z}$ in Section 5 to include all variable types including multiple multinomial variables by letting $\boldsymbol{\Psi}$ be:

$$\begin{aligned}
\boldsymbol{\Psi} &= \begin{pmatrix} \boldsymbol{\Psi}^{(11)} & \dots & \boldsymbol{\Psi}^{(1l)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Psi}^{(l1)} & \dots & \boldsymbol{\Psi}^{(ll)} \end{pmatrix} \\
\boldsymbol{\Psi}^{(kk)} &= \begin{pmatrix} \xi & 0 & 0 & \cdots & 0 \\ 0 & 1 & \psi_{23}^{(kk)} & \cdots & \psi_{2m_k}^{(kk)} \\ 0 & \psi_{32}^{(kk)} & \psi_{33}^{(kk)} & \ddots & \psi_{3m_k}^{(kk)} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \psi_{m_k2}^{(kk)} & \psi_{m_k3}^{(kk)} & \dots & \psi_{m_km_k}^{(kk)} \end{pmatrix} \\
\boldsymbol{\Psi}^{(kk')} &= \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & \psi_{22}^{(kk')} & \dots & \psi_{2m_{k'}}^{(kk')} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \psi_{m_k2}^{(kk')} & \dots & \psi_{m_km_{k'}}^{(kk')} \end{pmatrix}
\end{aligned}$$

$$
\begin{aligned}
\boldsymbol{\Psi}^{(kl)} &= \begin{pmatrix} 0 & \cdots & 0 \\ \psi_{21}^{(kl)} & \cdots & \psi_{2c_{\setminus\mathcal{M}}}^{(kl)} \\ \vdots & \ddots & \vdots \\ \psi_{m_k1}^{(kl)} & \cdots & \psi_{m_kc_{\setminus\mathcal{M}}}^{(kl)} \end{pmatrix} \\
\boldsymbol{\Psi}^{(ll)} &= \\
&\begin{pmatrix} 1 & \psi_{12}^{(ll)} & \cdots & \psi_{1c_{\setminus\mathcal{M}}}^{(ll)} \\ \psi_{21}^{(ll)} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \psi_{c_{\setminus\mathcal{M}}-1,c_{\setminus\mathcal{M}}}^{(ll)} \\ \psi_{c_{\setminus\mathcal{M}}1}^{(ll)} & \cdots & \psi_{c_{\setminus\mathcal{M}},c_{\setminus\mathcal{M}}-1}^{(ll)} & 1 \end{pmatrix}
\end{aligned}
$$

for $k, k' \in \{1, \ldots, c_{\mathcal{M}}\} \wedge k' \neq k$ and $l = c_{\mathcal{M}} + 1$. The k^{th} $\boldsymbol{\Psi}^{(kk)}$ in the diagonal for $k = 1, \ldots, c_{\mathcal{M}}$ is for the latent variables for the k^{th} multinomial variable. The last $\boldsymbol{\Psi}^{(ll)}$ block is for the latent variables for the binary, continuous, and ordinal variables.

S4. Derivative Approximations

We show the gradient of the log marginal likelihood with respect to the covariance matrix, $\boldsymbol{\Psi}$, in this section. Without loss of generality, suppose that covariance matrix $\boldsymbol{\Psi}$ is permutated such that the first indices are for the latent variables for the continuous variables, the next $|\mathcal{M}|$ indices are the latent variables corresponding to the observed categories of the multinomial variables, and the final indices are the remaining latent variables for the multinomial, ordinal, and binary variables. That is,

$$
\begin{aligned}
\mathcal{R}_i &= \tilde{\mathcal{M}} \cup \tilde{\mathcal{O}} \cup \tilde{\mathcal{B}} \setminus \mathcal{I}_i \\
\boldsymbol{\Psi} &= \begin{pmatrix} \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}} & \boldsymbol{\Psi}_{\mathcal{C}\mathcal{I}_i} & \boldsymbol{\Psi}_{\mathcal{C}\mathcal{R}_i} \\ \boldsymbol{\Psi}_{\mathcal{I}_i\mathcal{C}} & \boldsymbol{\Psi}_{\mathcal{I}_i\mathcal{I}_i} & \boldsymbol{\Psi}_{\mathcal{I}_i\mathcal{R}_i} \\ \boldsymbol{\Psi}_{\mathcal{R}_i\mathcal{C}} & \boldsymbol{\Psi}_{\mathcal{R}_i\mathcal{I}_i} & \boldsymbol{\Psi}_{\mathcal{R}_i\mathcal{R}_i} \end{pmatrix} \\
\bar{\boldsymbol{\mu}} &= (\bar{\boldsymbol{\mu}}_{1:c_{\mathcal{M}}}, \bar{\boldsymbol{\mu}}_{(-1:c_{\mathcal{M}})})^\top \\
\bar{\mathbf{S}} &= \begin{pmatrix} \bar{\mathbf{S}}_{1:c_{\mathcal{M}}1:c_{\mathcal{M}}} & \bar{\mathbf{S}}_{1:c_{\mathcal{M}}(-1:c_{\mathcal{M}})} \\ \bar{\mathbf{S}}_{(-1:c_{\mathcal{M}})1:c_{\mathcal{M}}} & \bar{\mathbf{S}}_{(-1:c_{\mathcal{M}})(-1:c_{\mathcal{M}})} \end{pmatrix}
\end{aligned}
$$

where $\bar{\boldsymbol{\mu}}$ and $\bar{\mathbf{S}}$ are given in Equation (9) and (10). It follows that the log marginal likelihood in Equation (11) is

$$
\begin{aligned}
l_i &= \log \phi^{(|\mathcal{C}|)}(\hat{\boldsymbol{z}}_{i\mathcal{C}}; \mathbf{0}, \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}) + \log \Phi^{(c_{\setminus\mathcal{M}} + \sum_{j=1}^{c_{\mathcal{M}}}(m_j - 1))}(\boldsymbol{a}_i, \boldsymbol{b}_i; \boldsymbol{m}, \mathbf{M}) \\
\boldsymbol{m} &= \bar{\boldsymbol{\mu}}_{(-1:c_{\mathcal{M}})} - \mathbf{D}\bar{\boldsymbol{\mu}}_{1:c_{\mathcal{M}}} \\
\mathbf{M} &= \bar{\mathbf{S}}_{(-1:c_{\mathcal{M}})(-1:c_{\mathcal{M}})} + \mathbf{D}\bar{\mathbf{S}}_{1:c_{\mathcal{M}}1:c_{\mathcal{M}}}\mathbf{D}^\top - \mathbf{D}\bar{\mathbf{S}}_{1:c_{\mathcal{M}}(-1:c_{\mathcal{M}})} - \bar{\mathbf{S}}_{(-1:c_{\mathcal{M}})1:c_{\mathcal{M}}}\mathbf{D}^\top.
\end{aligned}
$$

For a matrix $\mathbf{A} \in \mathbb{R}^{g \times g}$, we let

$$
\mathbf{A}' = \begin{pmatrix} \partial l_i / \partial a_{11} & \cdots & \partial l_i / \partial a_{1g} \\ \vdots & \ddots & \vdots \\ \partial l_i / \partial a_{g1} & \ldots & \partial l_i / \partial a_{gg} \end{pmatrix}.
$$

It then follows that

$$\begin{aligned}
\bar{\mathbf{S}}' &= \begin{pmatrix} \mathbf{D}^\top \mathbf{M}' \mathbf{D} & -\mathbf{D}^\top \mathbf{M}' \\ -\mathbf{M}' \mathbf{D} & \mathbf{M}' \end{pmatrix} \\
\boldsymbol{\Psi}'_{\mathcal{I}_i \cup \mathcal{R}_i, \mathcal{I}_i \cup \mathcal{R}_i} &= \bar{\mathbf{S}}' \\
\bar{\boldsymbol{\mu}}' &= \begin{pmatrix} \mathbf{D}^\top \boldsymbol{m}' \\ \boldsymbol{m}' \end{pmatrix} \\
\boldsymbol{\Psi}'_{\mathcal{C}\mathcal{C}} &= \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} \boldsymbol{\Psi}_{\mathcal{C}, \mathcal{I}_i \cup \mathcal{R}_i} \bar{\mathbf{S}}' \boldsymbol{\Psi}_{\mathcal{I}_i \cup \mathcal{R}_i, \mathcal{C}} \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} - \frac{1}{2} \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} \boldsymbol{\Psi}_{\mathcal{C}, \mathcal{I}_i \cup \mathcal{R}_i} \bar{\boldsymbol{\mu}}' \widehat{\boldsymbol{z}}_{i\mathcal{C}}^\top \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} \\
&\quad - \frac{1}{2} \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} \widehat{\boldsymbol{z}}_{i\mathcal{C}} \bar{\boldsymbol{\mu}}'^\top \boldsymbol{\Psi}_{\mathcal{I}_i \cup \mathcal{R}_i, \mathcal{C}} \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} - \frac{1}{2} \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} + \frac{1}{2} \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} \widehat{\boldsymbol{z}}_{i\mathcal{C}} \widehat{\boldsymbol{z}}_{i\mathcal{C}}^\top \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} \\
\boldsymbol{\Psi}'_{\mathcal{I}_i \cup \mathcal{R}_i, \mathcal{C}} &= -\bar{\mathbf{S}}' \boldsymbol{\Psi}_{\mathcal{I}_i \cup \mathcal{R}_i, \mathcal{C}} \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1} + \frac{1}{2} \bar{\boldsymbol{\mu}}' \widehat{\boldsymbol{z}}_{i\mathcal{C}}^\top \boldsymbol{\Psi}_{\mathcal{C}\mathcal{C}}^{-1}.
\end{aligned}$$

Thus, we can get an approximation of all the derivatives we need if we have an approximation of $\boldsymbol{m}'$ and $\mathbf{M}'$. Details of our approximation of the latter quantities are provided in Section 4 and in the next section.

S5. Quasi-Monte Carlo Procedure

Algorithm 1 shows pseudocode for the method we use to approximate the intractable integrals in the log likelihood, the gradient of the log likelihood, and the quantities used to perform the imputation. The algorithm is $\mathcal{O}\left(k^3\right)$ because of the Choleksy decomposition but the primary bottleneck for practical problems is the loop. For small to moderate k (say $k < 50$), the computation time spent evaluating Φ and Φ^{-1} is substantial. For moderate to large k, the dot product in $\hat{a}_j$ and $\hat{b}_j$ takes a relatively larger part of the computation time.

The dot product in $\hat{a}_j$ and $\hat{b}_j$ does not take full advantage of single instruction, multiple data (SIMD) instructions on modern CPUs. Thus, we found substantial reductions in computation time by simultaneously processing multiple draws. An adaptive randomized quasi-Monte Carlo (RQMC) method can be used if we run algorithm 1 using multiple randomized quasi-random sequences in parallel as Genz and Bretz (2002). This allows us to get an estimate of the error which can be used to stop early if the error is less than a user specified threshold. We use the Fortran code written by Genz and Bretz (2002) to simultaneously compute the Cholesky decomposition and find the permutation of the variables. The permutation is based on a heuristic to reduce the variance when we approximate the likelihood with $\boldsymbol{g}(\boldsymbol{x}) = 1$.

As for the gradient, let

$$\tilde{L}(\boldsymbol{\omega}, \boldsymbol{\Omega}) = \int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} \phi^{(k)}(\boldsymbol{u}; \boldsymbol{\omega}, \boldsymbol{\Omega}) \, \mathrm{d}\boldsymbol{u}.$$

Input: Lower and upper bounds, $\boldsymbol{a} \in \mathbb{R}^k$ and $\boldsymbol{b} \in \mathbb{R}^k$, mean vector $\boldsymbol{\omega} \in \mathbb{R}^k$, covariance matrix $\boldsymbol{\Omega} \in \mathbb{R}^{k \times k}$, number of samples S, function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^H$, and procedure to generate a k dimensional quasi-random sequence or pseudorandom numbers in $(0,1)^k$

Output: Approximation of $\int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} \boldsymbol{g}(\boldsymbol{u}) \phi^{(k)}(\boldsymbol{u}; \boldsymbol{\omega}, \boldsymbol{\Psi}) \, \mathrm{d}\boldsymbol{u}$

Compute the Cholesky decomposition $\mathbf{O}^\top \mathbf{O} = \boldsymbol{\Sigma}$ and set $\boldsymbol{r} = \mathbf{0}_H$

for $j = 1$ **to** k **do**

  Set $a_j \leftarrow o_{jj}^{-1}(a_j - \omega_j)$, $b_j \leftarrow o_{jj}^{-1}(b_j - \omega_j)$, and $\boldsymbol{o}_{1:j,j} \leftarrow o_{jj}^{-1} \boldsymbol{o}_{1:j,j}$

end

for $s = 1$ **to** S **do**

  Draw the next $\boldsymbol{u} \in (0,1)^k$ and set $w = 1$

  for $j = 1$ **to** k **do**

    Compute

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

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

    Set $w \leftarrow w \cdot (\Phi(\hat{b}_j) - \Phi(\hat{a}_j))$ and $x_j = \Phi^{-1}(\Phi(\hat{a}_j) + u_j(\Phi(\hat{b}_j) - \Phi(\hat{a}_j)))$

  end

  Update the mean estimator, $\boldsymbol{r} \leftarrow \boldsymbol{r} + s^{-1}(w\boldsymbol{g}(\mathbf{O}^\top \boldsymbol{x} + \boldsymbol{\omega}) - \boldsymbol{r})$

end

return $\boldsymbol{r}$

Algorithm 1: (Quasi) Monte Carlo (MC) procedure to approximate integrals that are needed to estimate the model and impute the missing values.

Then the gradient of the likelihood is given

$$\begin{aligned}
\nabla_{\boldsymbol{\omega}} \tilde{L}(\boldsymbol{\omega}, \boldsymbol{\Omega}) &= \boldsymbol{\Omega}^{-1} \int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} (\boldsymbol{u} - \boldsymbol{\omega}) \phi^{(k)}(\boldsymbol{u}; \boldsymbol{\omega}, \boldsymbol{\Omega}) \, \mathrm{d}\boldsymbol{u} \\
&= \mathbf{O}^{-1} \int \boldsymbol{u} h(\boldsymbol{u}) \prod_{j=1}^{k} w_j(\boldsymbol{u}) \, \mathrm{d}u \\
\nabla_{\boldsymbol{\Omega}} \tilde{L}(\boldsymbol{\omega}, \boldsymbol{\Omega}) &= \\
&\frac{1}{2} \left(\boldsymbol{\Omega}^{-1} \left(\int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} (\boldsymbol{u} - \boldsymbol{\omega})(\boldsymbol{u} - \boldsymbol{\omega})^\top \phi^{(k)}(\boldsymbol{u}; \boldsymbol{\omega}, \boldsymbol{\Omega}) \, \mathrm{d}\boldsymbol{u} \right) \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1} \tilde{L}(\boldsymbol{\omega}, \boldsymbol{\Omega}) \right) \\
&= \frac{1}{2} \left(\mathbf{O}^{-1} \left(\int \boldsymbol{u}\boldsymbol{u}^\top h(\boldsymbol{u}) \prod_{j=1}^{k} w_j(\boldsymbol{u}) \, \mathrm{d}u \right) \mathbf{O}^{-\top} - \boldsymbol{\Omega}^{-1} \tilde{L}(\boldsymbol{\omega}, \boldsymbol{\Omega}) \right)
\end{aligned}$$

with $\nabla_{\boldsymbol{\Omega}} = (\partial / \partial \Omega_{ij})_{i,j=1,\dots,k}$ and h is the importance distribution's density function in Equation (4). The needed choice of $\boldsymbol{g}$ can be seen from the two integrals above but it can

Table 1: Summary statistics for each observational data set. The first column is the number observations. The other columns indicates the number of variables of each type except for the "levels" columns which show the maximum number of categories for the ordinal and multinomial variables, respectively.

	# Observations	Binary	Continuous	Ordinal	Levels	Multinomial	Levels
Medcare	4406	2	2	4	19	0	0
Rent	2053	7	3	1	6	1	25
ESL	488	0	0	5	10	0	0
LEV	1000	0	0	5	5	0	0
GBSG	686	3	6	1	3	0	0
TIPS	244	3	2	2	6	0	0
Cholesterol	7846	2	0	1	4	1	4
Colon	1776	5	2	2	4	1	3

also be seen that one can simplify the expressions by working with $\boldsymbol{g}(\mathbf{O}^\top \boldsymbol{x} + \boldsymbol{\omega})$ in Algorithm 1.

S6. Data Sets

We will briefly describe the data sets we have used in this section. We have removed incomplete observations from each data set. We removed the potentially right censored survival time along with the censoring status from the colon data set. Such variables can be properly handled by extending our imputation method to support censored variables, which at present are not supported. This data set is not used by Zhao and Udell (2020) and, therefore, there are not any results to compare with.

In contrast, we keep the potentially right censored survival time and the censoring status for the German breast cancer study group (GBSG) data set as do Zhao and Udell (2020). Thus, our results can be compared with their results.

Table 1 summarises the number of variables of each type, the maximum number of categories for the ordinal and multinomial variables, and the numbers of observations in the complete data sets.

References

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.

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, 2020. Association for Computing Machinery.