Nonparametric Copula Models for Mixed Data with Informative Missingness

Joseph Feldman* and Daniel R. Kowal†
Department of Statistics, Rice University

Abstract

Modern datasets commonly feature both substantial missingness and variables of mixed data types, which present significant challenges for estimation and inference. Complete case analysis, which proceeds using only the observations with fully-observed variables, is often severely biased, while model-based imputation of missing values is limited by the ability of the model to capture complex dependencies and accommodate mixed data types. To address these challenges, we develop a novel Bayesian mixture copula for joint and nonparametric modelling of count, continuous, ordinal, and unordered categorical variables, and deploy this model for inference, prediction, and imputation of missing data. Most uniquely, we introduce a new and efficient strategy for marginal distribution estimation, which eliminates the need to specify any marginal models yet delivers strong posterior consistency for both the marginal distributions and the copula parameters even in the presence of informative missingness (i.e., missingness-at-random). Extensive simulation studies demonstrate exceptional modeling and imputation capabilities relative to competing methods, especially with mixed data types, complex missingness mechanisms, and nonlinear dependencies. We conclude with a data analysis that highlights how improper treatment of missing data can distort a statistical analysis, and how the proposed approach offers a resolution.

Keywords: Bayesian inference, Factor models, Imputation, Mixture models

*Corresponding author email: jrfl11@rice.edu. An R package implementing the proposed approach is available on the author’s github page, found at https://github.com/jfeldman396/GMCImpute
†Research was sponsored by the Army Research Office (W911NF-20-1-0184), the National Institute of Environmental Health Sciences of the National Institutes of Health (R01ES028819), and the National Science Foundation (SES-2214726). The content, views, and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, the National Institutes of Health, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.
1 Introduction

Missing data are ever-present in modern statistics and data analysis. The sources of missingness are vast and varied: participant non-response in surveys (Rubin, 1976), participant attrition in longitudinal studies (Gustavson et al., 2012), linking multiple data sources (Reiter, 2012), or errors in the data collection process all contribute to missingness. Any statistic meant to be computed on a fully-observed sample of data—including frequentist estimators and Bayesian posterior distributions—must be modified carefully in the presence of missing data. At the broadest level, the goal remains to infer an unknown population quantity $Q$, and specifically to provide accurate point estimates and precise uncertainty quantification for $Q$; here, we focus on the additional challenges and implications of abundant missingness.

The primary challenge occurs when the missingness is informative, which is made precise below. As an illustrative example, consider correlated bivariate data $\{(Y_{i1}, Y_{i2})\}_{i=1}^{n}$ and suppose that the missingness in $Y_{2}$ is determined by the value of $Y_{1}$. Figure 1 shows the impact of informative missingness: the empirical cumulative distribution function (ECDF) of $Y_{2}$ is severely biased, which implicates functionals of this term—including traditional statistics as well as popular Bayesian semiparametric copula models discussed subsequently (Hoff, 2007; Murray et al., 2013; Feldman and Kowal, 2022). Thus, without some accounting for missingness, many frequentist and Bayesian statistics would be biased, and may lead to inaccurate predictions, incorrect conclusions, and suboptimal decision-making.

Figure 1: Bivariate data $\{(Y_{i1}, Y_{i2})\}_{i=1}^{n}$ with informative missingness (left) and the corresponding true and empirical cumulative distribution function (ECDF) for $Y_{2}$ (right). The missing data severely biases the ECDF, which impacts functionals of this term—including traditional statistics as well as Bayesian semiparametric copula models.
This simple example highlights the pitfalls of complete case analysis, which proceeds using only the observations for which all variables are observed. Yet complete case analysis is widely used in practice: it is the simplest strategy for handling missing data and merely requires (i) subsetting the data to the complete cases and (ii) proceeding using unaltered frequentist or Bayesian methods. Under informative missingness, complete case analysis cannot avoid the biases such as those in Figure 1. Crucially, these effects are compounded for more complex dependencies, multivariate and mixed data, and intricate marginal distributions (see Table 1 and Figure 2). These issues are explored in detail in Sections 5–6.

Informative missingness is made precise by considering a likelihood for all observable data and missingness indicators, \( p(Y, R | \theta, \phi) = p(Y | \theta) \ p(R | Y, \phi) \), where \( Y = (Y_{ij}) \) denotes the observable data, \( R = (R_{ij}) \) denotes whether \( Y_{ij} \) was observed \( (R_{ij} = 0) \) or not \( (R_{ij} = 1) \), \( \theta \) are the usual model parameters (for \( Y \)) and \( \phi \) are the distinct parameters of the missingness mechanism (e.g., Rubin, 2004). When the missingness model \( p(R | Y, \phi) = p(R | \phi) \) does not depend on \( Y \), then the data are missing-completely-at-random (MCAR), so the missingness is uninformative and complete case analysis is valid. By contrast, informative missingness occurs when the missingness model depends on either the observed data \( Y^{\text{obs}} = \{Y_{ij} : R_{ij} = 0\} \) or the missing data \( Y^{\text{mis}} = \{Y_{ij} : R_{ij} = 1\} \), or both. Figure 1 features missing-at-random (MAR): the missingness model \( p(R | Y, \phi) = p(R | Y^{\text{obs}}, \phi) \) depends on the observed data, and introduces significant problems for complete case analysis. The more extreme setting is missing-not-at-random (MNAR), where the missingness mechanism depends on both \( Y^{\text{obs}} \) and \( Y^{\text{mis}} \). Because our analysis considers a moderate to large number of informative variables, we focus on MAR, which remains a widely applicable and challenging scenario. Note that MCAR and MAR are often referred to as ignorable missingness (Rubin, 2004); we tend to avoid this terminology because MAR is actually disruptive for the class of models considered subsequently, and thus requires careful consideration.

Revisiting Figure 1, it is clear that a complete case statistic \( \hat{Q}(Y^{\text{obs}}_2) \) will be inadequate relative to a full data statistic \( \hat{Q}(Y_2) = \hat{Q}(Y^{\text{obs}}_2, Y^{\text{mis}}_2) \) when the missingness is informative. Further, the complete case version \( \hat{Q}(Y^{\text{obs}}_2) \) uses a smaller sample size, and therefore sacrifices some precision. Thus, it is appealing to consider imputation of \( Y^{\text{mis}} \) based on a model for \( Y \), for example using a posterior predictive distribution \( p(Y^{\text{mis}} | Y^{\text{obs}}) \). This approach may be conceptualized by repeatedly simulating \( y^{\text{mis}} \sim p(Y^{\text{mis}} | Y^{\text{obs}}) \) and computing...
\( \hat{Q}(Y^{obs}, y^{mis}) \). If the underlying model adequately captures the dependencies in \( Y \), then this strategy can eliminate the bias of complete case analysis while providing uncertainty quantification for the unobserved data. We emphasize that, although the posterior distribution \( p(\theta \mid Y^{obs}) \) is sufficient for inference on \( \theta \) under MAR (under some assumptions), there are many interesting and important statistics \( \hat{Q}(Y) = \hat{Q}(Y^{obs}, Y^{mis}) \) computed on the full dataset, and these require careful accounting of the missingness mechanism.

Model-based imputation is ultimately limited by the adequacy of the model to (i) capture complex dependencies in the data and (ii) accommodate variables of mixed data types. Our motivating example comes from a collection of variables in the National Health and Nutrition Examination Survey (NHANES). As highlighted in Table 1, these variables include count, continuous, ordinal, and unordered categorical variables, with missingness as high as 43% for some variables. Notably, these variables include self-reported mental health—which displays complex and discrete marginal distributional features (Figure 2)—along with demographic and socioeconomic variables, alcohol and drug use variables, and health-related variables. As demonstrated in Section 6, complete case analysis is unsatisfactory or, in some cases, misleading. Thus, model-based imputation for these data is essential, and requires the ability to handle mixed data types with possibly complex multivariate dependencies and nonstandard marginal distributions.

![Marginal Distribution of DMHNG](image)

**Figure 2:** The marginal distribution of days of self-reported poor mental health (DMHNG) from the NHANES data, which is the response variable of interest in our real data analysis. Discreteness, boundedness, heaping, and zero-inflation combine to make modeling difficult.

The literature on imputation is quite robust, yet limited in its ability to address these critical challenges. Imputation strategies were introduced by Rubin (1976); see Murray (2018)
| Variable                      | Values                        | % Missing |
|-------------------------------|-------------------------------|-----------|
| **Response variable:**        |                               |           |
| DaysMentHlthNotGood (DMHNG)  | {0, 1, …, 30}                 | 14%       |
| **Demographic and socioeconomic variables:** |                               |           |
| Gender                       | Male, Female                  | 0         |
| Age (years)                  | {18, …, 80}                   | 0         |
| Race*                        | White, Black, Hispanic, Other | 0         |
| Education Level*             | < HS, = HS, > HS              | 5%        |
| Family Income* (FI)          | Low, Middle, High             | 4%        |
| Uninsured*                   | Yes, No                       | 0.2%      |
| **Alcohol and drug use variables:** |                               |           |
| HeavyDrinker                 | Yes, No                       | 29%       |
| UseNicotine                  | Yes, No                       | 15%       |
| UsedMarijuana                | Yes, No                       | 43%       |
| UsedHardDrug                 | Yes, No                       | 30%       |
| **Health-related variables:** |                               |           |
| Body Mass Index (BMI, kg/m²) | [13.4, 81.2]                  | 6%        |
| HasHighBP (BPQ020 at link)   | Yes, No                       | .1%       |
| HasHighChol (BPQ080 at link) | Yes, No                       | 6%        |
| HasDiabetes*                 | Yes, No                       | 0.08%     |

Table 1: Variables in the analysis dataset with hyperlinks to the online NHANES descriptions. Annotated variables (∗) include minor modifications (e.g., collapsed categories) from the original NHANES variables.

for a thorough review. In particular, multiple imputation advocates the generation and combination of multiple “completed” datasets, with the goal of providing point estimation and uncertainty quantification that accounts for the additional uncertainty due to missing values (Rubin, 2004).

Broadly, there are two main frameworks for imputation. The first, fully conditional specification (FCS), imputes missing values by (i) specifying a univariate regression model for each variable in the dataset conditional on all other variables and (ii) using each regression model to impute (separately) the missing values for each variable (Van Buuren and Oudshoorn, 1999; Raghunathan et al., 2001). This approach offers several advantages: it is amenable to mixed data types, allows customization of each univariate model to increase flexibility (Burgette and Reiter, 2010; Tang and Ishwaran, 2017), and is implemented is freely available software (Van Buuren and Groothuis-Oudshoorn, 2011). However, FCS in general does not specify a valid joint distribution for the data and is difficult to tune in high
dimensions. Perhaps most important, FCS often cannot capture complex relationships in
the data (Murray and Reiter, 2016), which we confirm in Section 5.

The second main approach constructs a joint distribution for all variables in the dataset
and then imputes missing values from the (posterior) predictive distribution. Thus, Bayesian
models are particularly attractive, and especially nonparametric Bayesian models due to
their distributional flexibility (Dunson and Xing, 2009; Manrique-Vallier and Reiter, 2013,
2014; Murray and Reiter, 2016; Manrique-Vallier and Reiter, 2017; DeYoreo et al., 2017).
However, few of these approaches are capable of jointly modeling count, continuous, ordinal,
and unordered categorical variables, and most are limited to a small number of variables.

Copula models offer a potential solution: they combine arbitrary marginal distributions
with a mechanism to model joint dependencies, and can handle certain mixed data types
(Joe, 2014). Zhao and Udell (2020a,b) deployed Gaussian copula models for imputation
of MCAR data with continuous and ordinal variables, but did not consider informative
missingness or other data types. Pitt et al. (2006) specified parametric families for count
and continuous variables within a Bayesian Gaussian copula model. However, parametric
specification of marginal distributions is restrictive and time-consuming, especially when
there are complex marginal distributions (Figure 2) and many variables to consider (Table 1).
Hoff (2007) partially resolved this issue for count, continuous, and ordinal variables using
the extended rank-likelihood (RL) for Gaussian copula estimation, which was extended to
higher dimensions using factor models in Murray et al. (2013). The RL features a rank-
based approximation to the likelihood for semiparametric inference, whereby the Gaussian
copula (correlation) parameters are inferred using only the ranks of the observed data. Most
recently, Feldman and Kowal (2022) introduced the extended rank-probit likelihood (RPL) to
include count, continuous, ordinal, and now unordered categorical variables. Most uniquely,
the R(P)L delivers inference for the copula parameters without requiring any estimation or
model specification of the marginal distributions, which is a substantial simplification.

Despite these advantages, semiparametric Bayesian copula models have two glaring short-
comings in the presence of missing data. First, these models prominently feature the ECDF
to define the posterior predictive distribution. Specifically, posterior predictive variates are
sampled by repeatedly (i) drawing a latent Gaussian variable under the model and (ii) ap-
plying the inverse ECDF. As highlighted in Figure 1, the ECDF is significantly flawed under
informative missingness. Thus, the resulting posterior predictive imputations will produce inaccurate estimation and uncertainty quantification for \( Q \)—even if the joint dependencies are well-modeled by the Gaussian copula. Clearly, these models cannot be relied upon for prediction or imputation with MAR (or MNAR) data.

Second, Gaussian copula models only specify linear associations on the latent scale. As such, they lack the ability to capture complex and nonlinear dependencies and interactions, which we demonstrate empirically in Section 5. Gaussian mixture copulas (Tewari et al., 2011; Rajan and Bhattacharya, 2016) offer some additional distributional flexibility, but these methods are highly parameterized, less robust than rank-based methods, and limited to certain data types.

To resolve these limitations, we develop a novel Bayesian mixture copula model for joint and nonparametric modelling of count, continuous, ordinal, and unordered categorical variables. The model features a rank-based likelihood paired with a latent mixture of factor models that is designed to provide robust, parsimonious, and flexible characterization of complex dependencies among mixed data types. Most uniquely, we propose a novel margin adjustment strategy that eliminates the reliance on the ECDF in the posterior predictive distribution of rank-based copula models (Hoff, 2007; Murray et al., 2013; Feldman and Kowal, 2022). Crucially, the margin adjustment requires no additional modeling assumptions or parameters and yields computationally efficient, empirically accurate, and consistent marginal distribution estimation, even in the presence of informative missingness. The margin adjustment is also central in our derivation of posterior consistency for the copula (correlation) parameters, for which previous results are no longer applicable under MAR (Murray et al., 2013). The importance of these modeling choices is highlighted using both simulated and real data, which decisively show that the proposed imputation strategy offers significant improvements over competing methods. These gains are especially pronounced in the presence of informative missingness and nonlinear dependencies.

This paper is organized as follows. In Section 2, we introduce copula models and the extended rank-probit likelihood as a means for specifying a valid joint model for count, continuous, ordinal and unordered categorical data. In Section 3 we develop the margin adjustment for consistent estimation of margins and multivariate dependence under copula models for data with informative missingness. In Section 4 we introduce a novel Gaussian
mixture copula for estimation of complex multivariate and mixed joint distributions and extend the margin adjustment to this model. We then apply our proposed approach in Section 5 with two simulation studies and a real data example in Section 6. We conclude in Section 7. Supplementary material includes proofs of all results, details on the computations, additional simulation results, and an R package that implements the proposed approach.

2 Copula Models for Mixed Data

The first objective is to develop a joint Bayesian model for mixed (count, continuous, ordinal, and unordered categorical) data types. This joint model will be used to generate posterior predictive draws for the missing data, thereby allowing estimation and uncertainty quantification of \( \hat{\mathcal{Q}}(\mathbf{Y}) = \hat{\mathcal{Q}}(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}) \). The class of candidate models we consider are copulas, which link arbitrary marginal distributions with a mechanism to model joint dependencies. By Sklar’s Theorem (Sklar, 1959), the joint cumulative distribution function (CDF) \( F \) of a \( p \)-dimensional random vector \((Y_1, \ldots, Y_p)\) can be expressed through the univariate marginal CDFs \( \{F_j\}_{j=1}^p \) and a copula \( C \):

\[
F(y_1, \ldots, y_p) = C\{F_1(y_1), \ldots, F_p(y_p)\},
\]

where \( C \) is any multivariate CDF on the \( p \)-dimensional unit hypercube. In this work, we use the Gaussian copula as the building block for innovation:

\[
F(y_1, \ldots, y_p) = \Phi_p[\Phi^{-1}\{F_1(y_1)\}, \ldots \Phi^{-1}\{F_p(y_p)\}],
\]

where \( \Phi_p \) is the CDF of a \( p \)-dimensional Gaussian random vector with mean zero and correlation matrix \( C \) and \( \Phi \) is the univariate standard normal CDF. A more convenient data-generating representation of (2) is given by

\[
z \sim N_p(\mathbf{0}, C), \quad z = (z_1, \ldots, z_p)'
\]

\[
y_j = F_j^{-1}\{\Phi(z_j)\}, \quad j = 1, \ldots, p.
\]
As such, the joint distribution of \((Y_1, \ldots, Y_p)\) under the Gaussian copula is parameterized by the copula correlation \(C\), which encodes the dependence structure, and the univariate marginal distributions \(\{F_j\}_{j=1}^p\). Note that the transformation (4) is specific to the Gaussian copula via \(\Phi\); our subsequent generalizations will modify this term appropriately, as well as the latent data model (3).

Bayesian inference for the Gaussian copula model requires prior distributions for the unknown \(C\) and \(\{F_j\}\). Given posterior samples of \(C\) and \(\{F_j\}\), posterior predictive simulations for the missing data are generated by drawing from (3)–(4), i.e., simulating \(\tilde{z}_i \sim N_p(0, C)\) and setting \(Y_{ij}^{\text{mis}} = F_j^{-1}\{\Phi(\tilde{z}_{ij})\}\) for each missing observation \(i\) in component \(j\). This algorithm highlights the mutual importance of the copula (correlation) parameters \(C\), the Gaussian copula assumption in (2), and the margins \(\{F_j\}\)—each of which we explore and generalize in subsequent sections.

### 2.1 The Extended Rank Probit Likelihood

The Gaussian copula model has several critical limitations. First, the correlation structure \(C\) captures only latent linear dependencies, and thus may not be suitable for more complex relationships (see Section 4). Second, the link (4) is not well-defined for unordered categorical variables. Lastly, the margins \(F_j\) must be specified either parametrically, which is restrictive and time-consuming when \(p\) is moderate or large, or nonparametrically, which requires careful consideration of the resulting computational burdens. For continuous variables, a convenient semiparametric estimation strategy is to infer \(C\) using the transformed data \(z_{ij} = \Phi^{-1}\{\hat{F}_j(y_{ij})\}\), which substitutes the ECDF \(\hat{F}_j\) into (4). However, for discrete variables, this strategy is unsuccessful: the transformation only changes the sample space and not the distribution of the data, so the maximum likelihood estimator for \(C\) is inconsistent (Hoff, 2007). Similar problems arise for Bayesian inference. Thus, additional modifications are needed to provide reasonable and valid joint models for mixed data types—even for completely-observed data.

One such approach is given by the extended rank-probit likelihood (RPL) of Feldman and Kowal (2022), which generalized Hoff (2007) and Murray et al. (2013) to accommodate these mixed data types. Let \(Y = \{y_{ij}\}_{i=1}^n\) with \(y_i = (y_{i1}, \ldots, y_{ip})'\) be comprised of \(p\) variables, \(q < p\) of which are unordered categorical variables with \(k_1, \ldots, k_q\) unique levels, respectively.
Similar to the data-generating representation of the Gaussian copula model, the RPL introduces a latent Gaussian data matrix \( Z = \{ z_i \}_{i=1}^n \) of equal dimension to observed data \( Y \). The RPL then pairs a rank-based likelihood for the continuous, count, and ordinal variables with an alternative latent construction for the unordered categorical variables.

First, consider the \( r = p - q \) continuous, count, and ordinal variables. Building upon the link in (4), it is clear that the non-decreasing nature of the transformation implies that the ordering in the data is preserved on the latent scale for (3): for variable \( j \) and observations \( i \) and \( k \), we know that \( y_{ij} < y_{kj} \implies z_{ij} < z_{kj} \). This ordering is preserved for each of these \( r \) variables in the dataset when the following event occurs:

\[
D(Y^r) := \{ Z^r \in \mathbb{R}^{n \times r} : \max\{z_{kj} : y_{kj} < y_{ij}\} < z_{ij} < \min\{z_{kj} : y_{ij} < y_{kj}\}\} \quad (5)
\]

and \( Z^r \) is the subset of the latent data corresponding to these \( r \) variables. Rank likelihood methods focus on the information in the ordering event \( D(Y^r) \), which is inspired by (4) but crucially does not directly depend on the marginal distributions \( \{F_j\}_{j=1}^r \).

For the \( q \) unordered categorical variables in the data set, the RPL links the latent data with the observed values using a diagonal orthant representation (Johndrow et al., 2013). For each categorical variable \( Y_c \) with \( k_c \) levels, we encode the variable numerically using a vector of \( k_c \) binary variables \( \gamma_c \) such that if \( y_{ic} = m \implies \{\gamma_{icm} = 1 \cap \gamma_{icl} = 0, l \neq m\} \), i.e., only the \( m \)th component is one and the others are all zero. Unlike other categorical variable representations (Chib and Greenberg, 1998), this approach does not necessitate the selection of a base level. The binary variables \( \gamma_c \) are then linked to the latent data via \( \{\gamma_{icm} = 1 \cap \gamma_{icl} = 0, l \neq m\} \implies \{z_{icm} > 0 \cap z_{icl} < 0, l \neq m\} \), i.e., \( y_{ic} = m \) implies that only the \( m \)th component is positive and the others are all negative. Aggregating this representation across all \( q \) unordered categorical variables, the observed categorical memberships satisfy the event

\[
D'(Y^q) := \bigcup_{c=1}^q \{ Z^{kc} : \gamma_{ij} = 1 \implies z_{ij} > 0 \cap \{z_{i\ell} < 0\}_{\ell \neq j}\} \quad (6)
\]

This model for unordered categorical variables is often preferable for ordinal variables as well, especially when the number of ordinal levels is small (Feldman and Kowal, 2022).

The RPL combines the rank-based event (5) and categorical variable event (6) to arrive
at $\mathcal{E}(Y) := \mathcal{D}(Y^r) \cup \mathcal{D}'(Y^q)$. Letting $\theta$ denote the copula (correlation) parameters, the full data likelihood for the RPL is decomposed as

$$p(Y | \theta, \{F_j\}_{j=1}^r) = p\{Y, Z \in \mathcal{E}(Y) | \theta, \{F_j\}_{j=1}^r\} = \prod_{i}\left[ p\{Z \in \mathcal{E}(Y) | \theta\} \right] \cdot \prod_{i,j} p\{Y_i | Z \in \mathcal{E}(Y), \theta, \{F_j\}_{j=1}^r\}$$

where the equivalence in (7) arises because observation of $Y$ implies that $Z$ satisfies the event $\mathcal{E}(Y)$, with the RPL augmented $Z$ of dimension $n \times p^*$, where $p^* = r + \sum_{c=1}^{q} k_c$. Rank likelihood methods argue that the first term in (8) contains most of the information about $\theta$, and indeed Murray et al. (2013) showed that using $p\{Z \in \mathcal{E}(Y) | \theta\}$ to approximate the full likelihood (7) is sufficient to obtain strong posterior consistency for $\theta$. Thus, the RPL targets the rank-based posterior distribution

$$p\{\theta | Z \in \mathcal{E}(Y)\} \propto p\{Z \in \mathcal{E}(Y) | \theta\} \cdot p(\theta)$$

where $Y$ may include continuous, count, ordinal, and unordered categorical variables.

A key feature of the RPL posterior (9) is that it does not directly involve the marginal CDFs $\{F_j\}$. Although this feature is a useful reduction for posterior inference on $\theta$, consideration of $\{F_j\}$ is necessary for the posterior predictive distribution—and thus for missing data imputation. The common strategy for posterior predictive simulation in Bayesian rank-based copula models proceeds as in the Gaussian copula case: given posterior samples of $\theta$ from (9), sample $[\tilde{z}_i | \theta]$ from (3) and set $Y_{ij}^{mis} = \hat{F}_j^{-1}\{\Phi(\tilde{z}_{ij})\}$. Modifications for the unordered categorical variables are available using the sample proportions of each category. Thus, despite the absence of the marginal CDFs in the posterior (9), some estimate of the marginal CDF is required for posterior predictive simulation and missing data imputation. Under informative missingness, the default use of ECDFs and marginal proportions (Hoff, 2007; Murray et al., 2013; Feldman and Kowal, 2022) is highly problematic (see Figure 1 and Sections 5–6), and urgently requires modification.
3 The Margin Adjustment

To eliminate reliance on the ECDFs for posterior predictive sampling and imputation—while still maintaining the beneficial structure of the RPL Bayesian copula model—we propose a new strategy called the margin adjustment. Most notably, the margin adjustment does not require any additional modeling assumptions or parameters and provides computationally efficient and consistent marginal distribution estimation, even in the presence of informative missingness. Equally important, the margin adjustment automates estimation of \( \{F_j\} \), i.e., it does not require individualized specification of each (parametric or nonparametric) marginal distribution. By comparison, parametric specification of each \( \{F_j\} \)—which, when learned jointly with \( \theta \), can potentially avoid the type of biases observed in Figure 1—is arduous and computationally intensive when \( p \) is moderate or large. Although parametric models offer useful simplifications in many settings, that is not the case here: such an approach requires (i) a suitable choice of each \( \{F_j\} \), which can be challenging for complex marginal distributions (Figure 2), and (ii) significant modifications to posterior sampling algorithms, which often require complex sampling steps (Pitt et al., 2006). Crucially, the margin adjustment does not affect posterior inference for \( \theta \) via (9), so the efficient posterior sampling algorithms for Bayesian rank-based copula models (Hoff, 2007; Murray et al., 2013; Feldman and Kowal, 2022) remain unchanged.

3.1 The Margin Adjustment: Derivation and Theory

We derive the margin adjustment under general conditions, but later show that this approach is applicable to any valid copula model, including the Gaussian copula and the Gaussian mixture copula. The key insight of the margin adjustment is that the combination of the RPL rank constraints (5)–(6) and the latent data model (3) are sufficient to infer the marginal distributions \( \{F_j\}_{j=1}^r \) with strong theoretical guarantees.

First, consider count, continuous, or ordinal variables. While the latent data model (3) characterizes the multivariate dependencies, the RPL enforces consistent orderings among \( Z_j \) and \( Y_j \) for each numeric variable \( j \). Therefore, the order statistics of \( Z_j \) and \( Y_j \) must match, and thus, upon ordering both \( Z_j \) and \( Y_j \), the position of \( Z_j^n(x) = \max\{Z_{ij} : Y_{ij} \leq x\} \) among \( \{Z_{ij}\}_{i=1}^n \) will be identical to the maximum position of \( x \) among \( \{Y_{ij}\}_{i=1}^n \) for any \( x \).
Then, informally, if \( F_j(x) = \tau \), \( Z^n_j(x) \) will approach the \( \tau \)th quantile under the marginal latent data model. Thus, we propose the following marginal distribution estimator:

\[
\tilde{F}_j(x) = G_j\{Z^n_j(x)\},
\]

where \( G_j \) is the marginal distribution for \( Z_j \) induced by the latent data model. For the Gaussian copula (3), we simply have \( G_j = \Phi \), \( j = 1, \ldots, r \); see Section 4 for the Gaussian mixture copula.

The margin adjustment (10) is a function of the latent data \( Z \), and therefore inherits a posterior distribution under the RPL Bayesian copula model. Notably, the latent data \( Z \) are sampled from their joint posterior distribution as part of the MCMC algorithm for the copula parameters \( \theta \) (see the supplement), and thus posterior inference for each \( \tilde{F}_j \) is readily available. The additional cost to compute (10) is minimal.

**Theorem 1.** Suppose \( \{Z_i\}_{i=1}^n \overset{iid}{\sim} F_Z \) and \( \{Y_i\}_{i=1}^n \overset{iid}{\sim} F_Y \), where \( F_Z \) is continuous and both \( \{Z_i\}_{i=1}^n \) and \( \{Y_i\}_{i=1}^n \) are ordered (non-decreasing). Defining \( Z^n(x) := \max\{Z_i : Y_i \leq x, i \in \{1, \ldots, n\}\} \), the margin adjustment satisfies \( \tilde{F}(x) := F_Z\{Z^n(x)\} \overset{a.s.}{\rightarrow} F_Y(x) \) for all \( x \).

The more challenging setting occurs when data are missing and the missingness is informative. In particular, the ECDF is no longer consistent under MAR or MNAR (see Figure 1), which undermines both prediction and imputation. We prove that a suitable modification of (10) remains consistent under MAR or MCAR. For simplicity, we demonstrate our result for \( p = 2 \) variables, one of which is MAR.

**Theorem 2.** Suppose \( Z_i = (Z_{i1}, Z_{i2}) \overset{iid}{\sim} G \) and \( Y_i = (Y_{i1}, Y_{i2}) \overset{iid}{\sim} F \) for \( i = 1, \ldots, n \), where \( G \) is continuous with marginal distributions \( G_1, G_2 \) and \( Y_{ij} = F_j^{-1}\{G_j(Z_{ij})\} \) for \( j = 1, 2 \) with marginal distributions \( F_1, F_2 \). Suppose that \( Y_1 \) is missing-at-random and \( Y_2 \) is completely observed. Define \( Z^n_1(x) := \max\{Z_{i1} : Y_{i1}^{obs} \leq x\} \). Then the margin adjustment satisfies \( \tilde{F}_1(x) := G_1\{Z^n_1(x)\} \overset{a.s.}{\rightarrow} F_1(x) \) for all \( x \).
Under MAR, the missingness mechanism for $Y_1$ depends only on $Y_{i1}^{obs}$ and $Y_{i2}$. Generalizations beyond the bivariate case are straightforward, but must maintain ignorability. The theorem also applies for discrete $Y_j$, where $F_j^{-1}$ maps quantile intervals defined by the left and right limits of the step function $F_j$ to elements in the support of $Y_j$.

For binary and unordered categorical variables, estimation of the marginal distributions requires estimation of the category proportions rather than a CDF. The probit link (6) demonstrates that these quantities depend on the parameter $\theta$ governing the dependence among the latent variables $Z$. As such, consistent estimation of $\theta$ under MAR is sufficient for consistent estimation of the category proportions, which is proved subsequently.

### 3.2 Strong Posterior Consistency with Informative Missingness

We now establish the asymptotic properties of the posterior distribution of the Gaussian copula (correlation) parameter $\theta$ under the RPL with informative missing data, and demonstrate how this posterior consistency naturally extends to the margin adjustment (10) as well as categorical proportions. These results expand upon Murray et al. (2013), who proved posterior consistency of $\theta$ under the RL, which omits unordered categorical variables. Crucially, the Murray et al. (2013) proof relied upon almost sure convergence of the ECDF $\hat{F}_j$ to $F_j$, which maintains under MCAR but not MAR (or MNAR). Thus, their result requires modifications to accommodate unordered categorical variables and MAR.

**Theorem 3.** Suppose $Y$ is comprised of $r$ numeric variables and $q = p - r$ unordered categorical variables, each with a potentially unique number of levels $k_c, c \in \{1, \ldots, q\}$. Let $p^* = r + \sum_{c=1}^q k_c$, and $\Pi$ be a prior distribution on the space of all $p^* \times p^*$ positive semi-definite correlation matrices $\theta$ with corresponding density $\pi(\theta)$ with respect to a measure $\nu$. Suppose $\pi(\theta) > 0$ almost everywhere with respect to $\nu$ and that $F_1, \ldots, F_r$ are the true marginal CDFs for the $r$ numeric variables. Furthermore, assume that the mississnes mechanism is ignorable. Then, for $\theta_0$ a.e. $[\nu]$ and any neighborhood $A$ of $\theta_0$, we have that

$$\lim_{n \to \infty} \Pi\{\theta \in A \mid Z \in \mathcal{E}(Y^{obs})\} = 1 \ a.s \ [G_{\theta_0, F_1, \ldots, F_r}^{\infty}]$$

where $G_{\theta_0, F_1, \ldots, F_r}^{\infty}$ is the data generating distribution of $\{y_i\}_{i=1}^{\infty}$ under the Gaussian copula and $Y^{obs}$ is the observed component of the data.
The strong posterior consistency of \( \theta \) also yields posterior consistency for both the margin adjustment and unordered categorical proportions. First, we consider continuous, count, and ordinal variables with the margin adjustment, which also leverages Theorems 1 and 2.

**Corollary 1.** Under the conditions of Theorem 3, define \( \tilde{F}_j \) as in (10) with \( G_j = \Phi \) for each \( j \in \{1, \ldots, r\} \). Then for any neighborhood \( A \) of \( F_j(x) \),

\[
\lim_{n \to \infty} \Pi \{ \tilde{F}_j(x) \in A \mid Z \in \mathcal{E}(Y_{\text{obs}}) \} = 1 \text{ a.s } [G_{\theta_0, F_1, \ldots, F_r}]
\]

for all \( x \).

For categorical proportions, the RPL links the latent data model to a multinomial distribution for each categorical variable through a continuous function of copula parameters. Thus, strong posterior consistency follows immediately.

**Corollary 2.** Under the conditions of Theorem 3, let \( p_{c\ell}^c = P(Y_c = c_\ell), c \in \{r+1, \ldots, p\}, \ell \in \{1, \ldots, k_c\} \). Then, for any neighborhood \( A \) of \( p_{c\ell}^c \),

\[
\lim_{n \to \infty} \Pi \{ p_{c\ell}^c \in A \mid Z \in \mathcal{E}(Y_{\text{obs}}) \} = 1 \text{ a.s } [G_{\theta_0, F_1, \ldots, F_r}]
\]

These results are powerful: the RPL Gaussian copula with the margin adjustment delivers fully Bayesian inference with strong posterior consistency for both the marginal distributions and the copula parameters. Notably, these results apply for mixed (count, continuous, ordinal, and unordered categorical) data types and in the presence of informative (MAR or MCAR) missingness. Thus, this framework provides a promising approach for inference and imputation with missing data.

### 4 Gaussian Mixture Copulas via Latent Factors

Although we have established theoretical guarantees for the RPL and the margin adjustment under a Gaussian copula model—including for mixed (count, continuous, ordinal, and unordered categorical) variables and missing data—the Gaussian copula is fundamentally limited by the simplicity of the latent data model (3). In particular, this model only captures linear associations on the latent scale through \( C \). As such, it may not be sufficiently
powerful to capture nonlinearities and interactions on the observed scale (see Section 5),
which is essential for imputation under complex dependencies. However, generalizations of
the latent data model (3) must carefully consider computational scalability, model parsimony,
and suitable adaptations to the margin adjustment.

To build an imputation model capable of adapting to unanticipated features in data, we
develop a novel Gaussian mixture copula (GMC) based on latent factor mixtures. The model
is designed to deliver nonparametric distributional flexibility that can capture complex data
shapes, while the factor structure maintains model parsimony and computational scalability
in higher dimensions. This latter feature offers important benefits relative to existing GMC
models (Tewari et al., 2011; Rajan and Bhattacharya, 2016).

The GMC extends the Gaussian copula by replacing the latent data model (3) with a
finite mixture:
\[
z \sim \sum_{h=1}^{H} \pi_h N_p(\alpha_h, C_h)
\] (12)
where the marginal distribution of the \( j \)th component is
\[
z_j \sim \sum_{h=1}^{H} \pi_h N(\{\alpha_h\}_j, \{C_h\}_{jj}).
\]
Thus, the GMC is defined on \( u \in [0,1]^p \) as
\[
C_{GMC}(u) = \Psi(\psi_1^{-1}(u_1), \ldots, \psi_p^{-1}(u_p))
\] (13)
where \( \Psi = \sum_{h=1}^{H} \pi_h \Phi_p(\alpha_h, C_h) \) and \( \psi_j = \sum_{h=1}^{H} \pi_h \Phi(\{\alpha_h\}_j, \{C_h\}_{jj}) \). For completeness, we
show that (13) is indeed a valid copula.

**Theorem 4.** \( C_{GMC} \) defines a valid copula.

Although the GMC (12) provides greater representational ability than the Gaussian copula (3),
especially for nonlinearities and interactions, the GMC modeling and computational
capabilities are limited in higher dimensions. In particular, the GMC is parametrized by
\( \{\alpha_h, C_h\}_{h=1}^{H} \), which contains many parameters when \( p \) is moderate or large. Further, Gaussian
mixture models tend to over-cluster when \( p \) is large, which results in more clusters—and
thus more parameters—than necessary. Instead, we apply our mixture on lower-dimensional
latent factors \( \eta \in \mathbb{R}^k \) with \( k \ll p \):
\[
\eta_i \sim \sum_{h=1}^{H} \pi_h N_k(\mu_h, \Delta_h), \quad z_i | \eta_i \sim N_p(\Lambda \eta_i, \Sigma)
\] (14)
where $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, $\Lambda$ is a $p \times k$ dimensional matrix of factor loadings, and $\eta_i$ is a $k$-dimensional vector of latent factors. The latent factor mixture model (14) induces a mixture model (12) for $Z$ through marginalization over $\eta$, and specifically with $\alpha_h = \Lambda \mu_h$ and $C_h = \Lambda \Delta_h \Lambda' + \Sigma$ for $h = 1, \ldots, H$. Thus, the latent data $Z$ is still endowed with a flexible mixture model, but the clustering is directed to a lower-dimensional space. Chandra et al. (2020) recently applied this strategy for continuous data and demonstrated how it alleviates the curse of dimensionality in model-based clustering—i.e., as $p$ grows, the number of nonempty clusters trivially tends toward $n$—but to our knowledge this approach has not been deployed for copula models or mixed data types.

The remaining challenge lies in Bayesian modeling of the finite mixture on $\eta$. In practice, the number of latent clusters $H$ will be unknown and should be determined based on the data. Thus, we propose a Dirichlet process (DP) to allow $H \to \infty$, and specifically using a stick-breaking process for the mixing weights $\{\pi_h\}$ (Ishwaran and James, 2001):

$$\pi_h = V_h \prod_{l<h} (1 - V_l), \quad V_l \overset{iid}{\sim} \text{Beta}(1, \alpha),$$

along with the hyperprior $\alpha \sim \text{Gamma}(a, b)$. For computational convenience, we implemented a truncated DP (Ishwaran and James, 2002) that resembles (14), where now $H$ is a conservative upper bound. The componentwise mean and covariance are assigned a Normal-Inverse Wishart prior, $(\mu_h, \Delta_h) \sim \text{NIW}(\mu_0, \delta^2 I, \kappa_0, \nu_0)$, while the diagonal elements of $\Sigma$ are assigned $\sigma_j^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$. Lastly, we apply a global-local shrinkage prior for the loadings matrix $\Lambda = \{\lambda_{jh}\}$ that encourages columnwise shrinkage for rank selection (Bhattacharya and Dunson, 2011): $\lambda_{jh} \sim N(0, \phi_{jh}^{-1} \tau_h^{-1})$ with local scale parameters $\phi_{jh} \sim \text{Gamma}(\nu_\phi/2, \nu_\phi/2)$ and global scale parameters $\tau_h = \prod_{l=1}^H \delta_l^\tau$, with $\delta_1^\tau \sim \text{Gamma}(a_1, 1)$ and $\delta_l^\tau \sim \text{Gamma}(a_2, 1)$, $l \geq 2$, $a_2 \geq 1$. By design, this ordered shrinkage prior reduces sensitivity to the choice of $k$, provided $k$ is sufficiently large.

The Gibbs sampling algorithm under the RPL (9) for the GMC (14) is detailed in the supplement. The Gibbs sampler alternates between the full conditionals for the latent data, $[Z \mid Y^{obs}, \theta]$, which features univariate truncated Gaussian draws for the elements of $Z^{obs}$ and univariate Gaussian draws for the elements of $Z^{mis}$, and the GMC parameters $[\theta \mid Y^{obs}, Z] = [\theta \mid Z]$, which uses standard sampling steps for Gaussian factor or mixture models. By
specifying our GMC using covariances rather than correlations, we avoid unwieldy constraints and the accompanying computational challenges, but can still access the implied model-based correlation matrices with minimal additional costs.

Despite the complexity of the GMC (14) relative to the Gaussian copula (3), the margin adjustment (10) is easy to apply and efficient to compute, even in the presence of missingness. For each posterior draw of \( \mathbf{Z} \) and \( \boldsymbol{\theta} \), we compute \( \tilde{F}_j(x) = \psi_j\{Z_j^*(x)\} \) for \( j = 1, \ldots, r \), where \( \psi_j \) depends on \( \boldsymbol{\theta} \) and \( Z_j^* \) is defined in Theorem 2 to account for missing data. Since each \( \tilde{F}_j \) is a step function with jumps determined by the observed values \( Y_j^{obs} \), we fit a monotone interpolating spline to \( (x, \tilde{F}_j(x)) \) for all \( x \in Y_j^{obs} \). Crucially, the interpolating spline preserves \( \tilde{F}_j \) at the observed data values and expands the support of the data-generating process beyond only the observed values \( Y_j^{obs} \), which is important for both count and continuous variables. Lastly, we enforce compact support for each variable \( j \) using reasonable values \( \ell_j \) and \( u_j \) such that \( \tilde{F}_j(\ell_j) = 0 \) and \( \tilde{F}_j(u_j) = 1 \), which ensures that each draw of \( \tilde{F}_j \) is a valid CDF. By applying this process for each posterior draw of \( \mathbf{Z} \) and \( \boldsymbol{\theta} \), we obtain a posterior distribution for the margin adjustment \( \tilde{F}_j \).

Using the posterior draws of \( \mathbf{Z}^{mis} \), \( \boldsymbol{\theta} \), and \( \tilde{F}_j \), we deliver model-based imputation:

\[
Y_{ij}^{mis} = \tilde{F}_j^{-1}\{\psi_j(Z_{ij}^{mis})\}.
\] (16)

Unlike existing rank-based copula models, this imputation strategy does not depend on the ECDF, which is biased under MAR; and unlike parametric copula models, the margin adjustment does not require specification and diagnostics for each marginal model for \( F_j \). Instead, the margin adjustment provides consistent estimation and inference even under MAR—and does so with no additional model assumptions and minimal computing cost.

This represents our proposed model and imputation strategy: the Gaussian mixture copula paired with the margin adjustment (GMC-MA), with posterior predictive draws generated from (16).
5 Simulation Studies

We present two detailed simulation studies. First, we generate fully synthetic data with nonlinear dependencies and subject to MAR, and assess the proposed GMC-MA approach for marginal distribution estimation and imputation (Section 5.1). Second, we generate hybrid synthetic data that uses NHANES variables as covariates within a synthetic regression model subject to MAR, and evaluate the GMC-MA for estimation and inference for the regression coefficients (Section 5.2). In each case we consider mixed data types and demonstrate the misleading consequences of complete case analysis.

5.1 Mixed Data Types, Nonlinearity, and Informative Missingness

In the first simulation study, we evaluate (i) the impact of informative missingness (MAR) on marginal distribution estimation via the ECDF—and show how the margin adjustment corrects the resulting biases—and (ii) assess whether the proposed GMC-MA is capable of accurate imputation under nonlinear dependencies. We generate synthetic datasets with continuous, count, and binary variables ($p = 3$) related nonlinearly as follows:

\begin{align*}
Y_1 & \sim N(0, 1) \\
Y_2 | Y_1 = y_1 & \sim \text{Poisson}(5|y_1|) \\
Y_3 | Y_2 = y_2, Y_1 = y_1 & \sim \text{Bernoulli}\{\Phi(-0.5 + y_{2\text{scale}})\}
\end{align*}

for $n \in \{500, 1000, 2000\}$, where $y_{2\text{scale}}$ is the centered and scaled version of $y_2$. Next, we introduce missingness using a MAR mechanism that links the missingness in both $Y_2$ and $Y_3$ with the observed value of $Y_1$:

\begin{align*}
R_j | Y_1 = y_1 & \sim \text{Bernoulli}\{\Phi(-0.5 + \beta|y_1|)\}, \quad j = 2, 3
\end{align*}

where $\beta$ determines both the amount of missingness and the impact of $Y_1$ on the missingness for each of $Y_2$ and $Y_3$. We consider $\beta \in \{0.5, 1\}$, with the lower value resulting in approximately 30% complete cases and 50% of each variable missing, and the higher value yielding approximately 20% complete cases with 60% marginal missingness. For $\beta = 1$, the missing-
ness mechanism removed nearly all $Y_3 = 1$, so we randomly selected 5% of the ground truth observations to remain untouched by $R$ to maintain a more reasonable setting.

We highlight the challenging nonlinearities and missingness under this data-generating mechanism ($n = 2000$, $\beta = 0.5$) with a single simulated dataset in Figure 3. Compared to the full dataset, the complete cases omit larger values of $Y_2$ and many instances of $Y_3 = 1$. To visualize the comparative imputation methods, we provide a single imputed dataset from the proposed GMC-MA and compare it to Hoff (2007) using the \texttt{sbgcop} package in R, which uses a single component Gaussian copula with the ECDF for posterior predictive simulations. Clearly, the GMC-MA offers significant improvements in both capturing the nonlinear relationship between $Y_1$ and $Y_2$ and correctly imputing additional $Y_3 = 1$ values when $|Y_1|$ is large. These results are confirmed with additional realizations in the supplement.

We emphasize that the proposed approach does not specify the true nonlinear relationships or the true marginal distributions in the data-generating process (17); we only input the observed values and the data type (continuous, count, and binary). Thus, it is clear that the GMC-MA is highly favorable for detecting nonlinearities, especially compared to the single component version, which leads to significant gains in imputation accuracy.

Next, we evaluate the margin adjustment, and specifically seek to assess whether it corrects the biases of the ECDF in the presence of informative missingness (e.g., Figure 1). We focus on the marginal distribution for $Y_2$, which is a count variable subject to MAR. For each combination of $(n, \beta)$, we compute the ECDF of $Y_2$ prior to removing missingness, which we treat as the ground truth (black points); the ECDF computed on the observed data $Y_2^{obs}$ (red points); and posterior draws (gray lines) and the posterior expectation (triangles) based on the margin adjustment under the GMC model. Posterior inference uses the estimators described in Section 4 and the Gibbs sampler from the supplement, which we run for 20,000 iterations. Trace plots of the draws of the marginal distribution functions for $Y_1$ and $Y_2$ indicate that the MCMC algorithm converges after about 5,000 samples, but as a conservative measure, we discard the first 15,000 iterations as a burn-in.

The results are presented in Figure 4. Most notably, the ECDF on the observed data is badly biased, especially for larger $\beta$, and this bias is unresolved for larger $n$. Crucially, previous work on Bayesian semiparametric (rank-based) copula models has relied exclusively on the ECDF for imputation (Hoff, 2007; Murray et al., 2013; Feldman and Kowal, 2022).
Figure 3: A simulated dataset without missingness (top left) and the complete cases after applying the MAR mechanism (top right) with $\beta = 0.5$. The proposed approach (bottom left) is significantly better than the Gaussian copula (bottom right) at capturing the challenging nonlinear relationship between $Y_1$ and $Y_2$ and correctly imputing additional $Y_3 = 1$ values (blue) when $|Y_1|$ is large.

The use of the ECDF places imputed values of $Y_2$ on the incorrect scale, which significantly limits the ability of these existing methods for model-based imputation under MAR. By comparison, the posterior distribution for the margin adjustment concentrates quickly around the ground truth as $n$ grows, while the point estimates of the marginal distribution are highly accurate. Notably, these favorable results occur even with considerably more bias from the missingness mechanism ($\beta = 1$, bottom row). These results suggest that Corollary 1 may be applicable more broadly, including for GMCs. Finally, we note that the interpolation strategy for the marginal distribution is effective: several values in the support of $Y_2$ are unobserved in $Y_2^{\text{obs}}$, yet the margin adjustment remains accurate for these cumulative probabilities.
Figure 4: Estimation and inference for the marginal distribution of $Y_2$ under MAR with varying $(n, \beta)$. The ECDF of $Y_{2 \text{obs}}$ (red points) deviates significantly from the ECDF of $Y_2$ prior to removing missingness (black points). The posterior draws (gray lines) and posterior mean (triangles) from the margin adjustment show that the proposed approach is highly accurate even under MAR, with the posterior distribution concentrating around the ground truth as $n$ grows.

Analogous results for binary $Y_3$ are presented in the supplement, and similarly demonstrate the exceptional performance of the margin adjustment. In conjunction, these results show that the proposed GMC-MA approach is capable of capturing nonlinear dependencies among mixed data types and accurately recovering mixed marginal distributions, even in the presence of abundant and informative (MAR) missingness.

5.2 Imputation for Regression Analysis using Hybrid Simulations

In the second simulation study, we study the impacts of imputation within the broader context of a regression analysis, and include comparisons with popular non-Bayesian alternatives for multiple imputation. To incorporate the challenges of real-world data analysis while maintaining partial control over the data-generating process, we use hybrid synthetic
data. First, we select three variables from the 2011-2012 NHANES data (see Table 1): a categorical variable (Family Income (FI)), a count variable (Age), and a continuous variable (BMI). Both Age and BMI are centered and scaled, while FI has three levels: Low, Middle, and High. Next, for each of the $n = 2434$ complete NHANES observations, we generate a continuous response variable using a Gaussian linear model with an FI:BMI interaction:

$$\text{New}_i \mid - \sim N(\alpha + \text{Middle}_i \beta_{\text{Middle}} + \text{High}_i \beta_{\text{High}} + \text{Age}_i \beta_{\text{Age}} + \text{BMI}_i \beta_{\text{BMI}} + \text{Middle:BMI}_i \beta_{\text{Middle:BMI}} + \text{High:BMI}_i \beta_{\text{High:BMI}}, \sigma^2).$$

We set $\sigma^2$ based on a prespecified signal-to-noise-ratio, $\text{SNR} = \text{var}(X\beta_{\text{true}})/\sigma^2$, where $\beta_{\text{true}} = (\alpha, \beta_{\text{Middle}}, \beta_{\text{High}}, \beta_{\text{Age}}, \beta_{\text{BMI}}, \beta_{\text{Middle:BMI}}, \beta_{\text{High:BMI}}) = (1, 1, 2, 0.5, -2, 2, 4)$ is the true coefficient vector and $X$ is the observed data matrix. We consider low (SNR = 1) and high (SNR = 3) ratios. Finally, we introduce MAR for each variable in (19) with the exception of BMI:

$$R_{ij} \mid - \sim \text{Bernoulli}\{\Phi(-0.7 + \text{BMI}_i + \omega_{ij})\}$$

where $\omega_{ij}$ are standard Gaussian innovations with $\text{Corr}(\omega_{ij}, \omega_{ij'}) = 0.3$ and $\text{Corr}(\omega_{ij}, \omega_{i'j}) = 0$. Thus, (20) introduces correlated and data-dependent (via BMI) patterns of missingness across both the response variable and the covariates. The missingness mechanism (20) is applied to all but 300 observations, which are left completely observed. This MAR mechanism yields about 49% complete cases, while the remaining $n = 2434$ observations have at least one missing value. Each variable has approximately 30% missing values. We repeat this data-generating process to create 100 hybrid synthetic datasets.

Since the missingness in New, Age, and FI is linked to BMI, complete case analysis is at risk of significant bias. To illustrate this point, we compute ordinary least squares estimators ($\hat{\beta}_{\text{CC}}$) and standard errors ($\hat{\sigma}_{\text{CC}}$) for the regression coefficients using only the completely-observed data, and average these values across simulations. Table 2 compares $\beta_{\text{true}}$ with these point estimates, and includes $1.96\hat{\sigma}_{\text{CC}}$ to emphasize the implications for inference (e.g., via confidence intervals). For all variables besides Age, the complete case analysis yields estimates and inference that depart significantly from the ground truth. Thus, alternative estimation and inference techniques are required, and specifically ones that can properly account for the informative (MAR) missingness. Note that these results are for the high
SNR case (SNR = 3), but the results are similar for SNR = 1.

|       | Intercept | Middle | High | BMI | Age | Middle:BMI | High:BMI |
|-------|-----------|--------|------|-----|-----|------------|----------|
| $\beta_{true}$ | 1 | 1 | 2 | -2 | 0.5 | 2 | 4 |
| $\hat{\beta}_{CC}$ | 1.77 | 0.19 | 0.39 | -1.50 | 0.51 | 1.50 | 3.01 |
| (1.96$\hat{\sigma}_{CC}$) | (0.10) | (0.12) | (0.16) | (0.09) | (0.05) | (0.11) | (0.17) |

Table 2: Complete case coefficient estimates and (scaled) standard errors averaged across simulations. The complete case analysis is severely biased for all variables except Age.

For evaluations and comparisons among imputation methods, we generate $m = 20$ multiple imputations for each hybrid synthetic dataset using several distinct approaches. First, we use the proposed approach to generate posterior predictive samples under the RPL-based GMC with the margin adjustment from Section 4 (GMC-MA). Next, we use the same model and posterior draws, but replace the margin adjustment with the ECDF (GMC-ECDF). By comparing GMC-MA and GMC-ECDF, we can isolate the impact of the margin adjustment. These imputations are based on the Gibbs sampler (see the supplementary material) run for 20,000 iterations, with the first 10,000 discarded as a burn-in and the imputations computed every 500th sample to achieve $m = 20$. Among non-Bayesian methods, we create multiple imputations from the popular algorithm MICE (multiple imputation using chained equations; Van Buuren and Oudshoorn, 1999) under default settings in the R package mice. In addition to the default MICE algorithm, which employs linear models with main effects for each variable, we include a modified version that features classification and regression trees for each variable (MICE-CART), which is better suited to capture interactions (Burgette and Reiter, 2010). These MICE approaches represent the FCS strategy for imputation.

For each completed dataset, we fit the regression model (19) using ordinary least squares and use the combining rules from Rubin (2004) to create point estimates and 99% confidence intervals. The estimates are summarized in Figure 5 via the absolute bias for each point estimate and coverage rates and widths for each interval estimate, averaged across 100 simulations. In this highly challenging scenario, the proposed GMC-MA imputations consistently provide the most accurate point estimates (smallest absolute bias), the most well-calibrated intervals (largest coverage rates), and among the most precise inference (smallest interval widths). The competing methods all suffer from significant undercoverage, and often are much less precise (wider intervals). Clearly, the margin adjustment is crucial: the GMC-
ECDF intervals do not provide close to the nominal coverage—despite using the same underlying model as the GMC-MA—due to the bias in the ECDF under informative missingness. This result supports the insight gained in Theorem 3, which suggests that marginal distributions play an important role in multivariate inference under copula models. As expected, the GMC-MA intervals are slightly wider than the GMC-ECDF intervals: the former account for the uncertainty in the marginal distributions via the posterior distribution, while the latter treat the marginal distributions as fixed (at the ECDFs).

Figure 5: Absolute bias (left), interval coverage rates (center), and interval widths (right) for point and 99% interval estimates computed under each imputation method. The GMC-MA approach consistently provides the most accurate point estimates (small absolute bias), the most well-calibrated intervals (large coverage rates), and highly precise inference (small interval widths).

The default MICE approach performs quite poorly: the point estimates are the least accurate and the interval estimates provide less than 5% coverage for all variables except Age. MICE-CART offers some improvements, but still lags in estimation accuracy and the intervals are not close to the nominal coverage, and substantially wider. Further, significant coverage gaps remain in both SNR settings for the interaction terms. As Burgette and Reiter (2010) note, one potential disadvantage of MICE-CART is the decreased efficiency when a parametric imputation model is suitable, such as (19). In the supplement, we highlight instances of this inefficiency across multiple imputations; the CART imputations often misclassify FI, which creates significant problems for estimating the interaction effects.
6 Real Data Application

6.1 Setting and goals

The 2011-2012 National Health and Nutrition Experimentation Survey (NHANES) asks the question, “For how many days during the past 30 days was your mental health not good?” The responses can be linked to other demographic and behavioral variables included in the questionnaire, enabling important insights into self-reported mental health. Of particular importance is the identification of key associative behaviors for at-risk individuals, i.e., those individuals with exceptionally high values of self-reported poor mental health. More broadly, mental health indicators are proxies for quality of life, depression, and risk for self-harm (Horwitz and Scheid, 1999).

We study the association between self-reported marijuana use (UsedMarijuana), gender, race, and high levels of self-reported poor mental health (DMHNG). However, there are several significant challenges for this analysis. First, the data are subject to substantial missingness (Table 1), especially for the variables of interest. In particular, UsedMarijuana is not asked of any individual older than 59, and is over 40% missing. Thus, a complete case analysis of the association between UsedMarijuana and DMHNG—among other variables—would omit all individuals older than 59, and potentially bias the results. Fortunately, Age is recorded, which implies that MAR is reasonable for these missing values.

Second, the NHANES variables are of mixed data types (Table 1), while DMHNG exhibits a complex marginal distribution (Figure 2) characterized by discreteness, zero-inflation, boundedness, and heaping. This distributional complexity introduces challenges for regression analysis (Kowal and Wu, 2021). To focus on at-risk individuals with large DMHNG values, we study the upper quantiles of DMHNG and their association with key variables of interest. However, the sample quantiles of this discrete variable DMHNG do not satisfy asymptotic normality, and thus are ill-suited for traditional multiple imputation (Rubin, 2004).

Instead, we provide posterior predictive uncertainty quantification for our target statistics \( \hat{Q} \). The analysis data set features \( n = 5856 \) observations of \( p = 15 \) variables subject to missingness (Table 1). We include all of these variables in our GMC-MA model (Section 4), which helps justify the MAR assumption. The model is estimated by running the Gibbs sampler (see the supplementary material) for 20,000 iterations and discarding the first 10,000
iterations as a burn-in. Inference is conducted by generating 500 posterior predictive datasets \( \{ \tilde{Y}_{ij} \} \) of size \( n \times p \) and computing our summary statistics \( \hat{Q}(\tilde{Y}) \) on each of these predictive datasets, which delivers posterior predictive inference for \( \hat{Q} \). Crucially, this observation-driven inference leverages the GMC-MA model to capture challenging marginal and joint distributions across mixed data types and in the presence of missingness, and simultaneously accounts for the joint uncertainties in the model parameters, the missing data, and the replicability for a new dataset of the same size. This latter source of uncertainty is similar to a sampling distribution, but instead conditions on the observed data and delivers posterior (predictive) uncertainty quantification via the GMC-MA model. By comparison, statistics computed on the imputed data \( \{ Y^{obs}, Y^{mis} \} \) only incorporate uncertainty from \( \theta \) and \( Y^{mis} \), which limits the generalizability of the inference and conclusions.

Lastly, we consider the sampling design in the NHANES survey. The NHANES sampling design includes certain oversampled subgroups, most notably stratified by race. Thus, we stratify our analysis by race (and gender), and apply our methods to these subgroups, which avoids the need to reweight for population-level inference that aggregates across all strata.

### 6.2 Checking Calibration

Our aim is to quantify the extent to which model-based inferences change between a complete case (CC) analysis and a full data analysis that accounts for the missing values. To this end, we fit the GMC-MA model on both the CC data and the full data, which serves two important purposes. First, by fitting the model on the CC data, we can compute posterior predictive diagnostics to assess whether the model is adequate for these data. Posterior predictive diagnostics compare the posterior predictive distribution of a statistic \( \hat{Q}(\tilde{Y}) \) to the observed value \( \hat{Q}(Y) \). However, \( \hat{Q}(Y) \) is unavailable for the full dataset due to abundant missingness. Thus, the CC diagnostics are the best available option. Second, by comparing the GMC-MA model output from the CC and full datasets, we can assess the impact of missingness on the analysis, and in particular whether the CC analysis is biased or misleading.

The CC data is created by dropping any observation with missing values, yielding a data set of size \( n_{CC} = 2434 \), compared to \( n = 5856 \) for the full dataset. We fit the GMC-MA on both the CC and the full datasets as described above, and generate predictive datasets...
of size \( n_{CC} \times p \) and \( n \times p \), respectively. For our statistics \( \hat{Q} \), we compute three quantities stratified by race, gender, and marijuana use: the empirical distribution of \( \text{DMHNG} \), which is useful for posterior predictive diagnostics, and the 75th and 90th sample quantiles of \( \text{DMHNG} \), which target the at-risk individuals within each stratum. The GMC-MA fit the full dataset must account for the additional uncertainty due to the missing observations, but also benefits from a much larger sample size. Thus, we expect that the model fit to the full dataset will yield posterior (predictive) distributions with less spread than those from the CC dataset. Differences in location, however, suggest a bias due to missing data.

We compare the posterior predictive samples of the empirical distribution of \( \text{DMHNG} \) from the CC and full dataset fits in Figure 6, and include the ECDF computed on the CC data. Specifically, we compute the ECDF on each posterior predictive dataset \( \{\tilde{Y}_{ij}\} \)—stratified by race, gender, and marijuana use—for the CC and full datasets, and report the pointwise 95\% highest posterior density (HPD) intervals. If we were to omit the stratification, then these posterior predictive ECDFs would converge to the margin adjustment \( \tilde{F}_j \) as the size of the posterior predictive sample \( \{\tilde{Y}_{ij}\} \) diverged to infinity. Thus, the stratification incorporates features of the joint distribution, while the uncertainty quantification depends on the sample size, as in traditional posterior predictive diagnostics.

First, we observe that the ECDF on the CC data falls in the center of the 95\% HPD intervals from the GMC-MA (CC) fit, which confirms that the proposed model adequately captures the challenging distributional features of \( \text{DMHNG} \), even upon stratification by race, gender, and marijuana use. In particular, the GMC-MA (CC) fit accurately describes the zero-inflation (the value at \( \text{DMHNG} = 0 \)), heaping (the large jumps around \( \text{DMHNG} \in \{7, 10, 14, 15, 20\} \)), and boundedness at 30 (the lower interval converges to one at \( \text{DMHNG} = 30 \)). Thus, the GMC-MA performs exceptionally well for these nuanced features of the joint distribution. Similar results for males are presented in the supplementary material.

Next, we compare the fitted GMC-MA models on the CC and full datasets. Most notably, the GMC-MA fit to the full dataset has substantially narrower 95\% HPD intervals, and often disagrees with both the ECDF and the GMC-MA (CC) output. In particular, expanding from the CC data to the full dataset not only dramatically narrows but also shifts the predictive ECDF intervals, especially for white females. For some strata, the predictive ECDF from the GMC-MA actually excludes the ECDF fit to the CC data, while the GMC-
Figure 6: Posterior predictive summaries for models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). For each race-gender-marijuana use stratum, we compare the 95% HPD intervals for the posterior predictive ECDFs, and include the ECDF on the CC data for reference. The GMC-MA (CC) output is well-calibrated to the observed data. By comparison, the GMC-MA fit to the full dataset produces intervals that are narrower and shifted, which suggests that the missingness mechanism is informative—and that CC analysis is unreliable.

MA and the GMC-MA (CC) intervals do not fully overlap. Because the GMC-MA (CC) output broadly agrees with the empirical version, we argue that these discrepancies are not due to model misspecification, but rather due to the significant impacts of informative missingness. These results confirm our expectations based on the simulation results (Section 5) and suggest that a CC analysis of these data is unreliable.

6.3 Associating Marijuana Use with Self-Reported Mental Health

To investigate the associations between at-risk self-reported mental health and race, gender, and marijuana use, we compute the posterior predictive statistics $\hat{Q}(\tilde{Y})$ for the 75th (see the supplement) and 90th quantile of $\text{DMHNG}$, stratified by gender-race-marijuana use. These quantities are computed for both GMC-MA on the full dataset and GMC-MA (CC), and provided posterior predictive uncertainty quantification for these statistics, which is summarized using the posterior median and 95% HPD intervals.

Figure 7 summarizes these point and interval estimates for the 90th quantile of $\text{DMHNG}$
for each stratum. Across all strata, there are several intervals from the GMC-MA (CC) fit with substantial overlap between marijuana users and non-users; yet many of these intervals become well-separated under the full dataset analysis with GMC-MA. Thus, CC analysis fails to identify certain strong, significant, and adverse associations between UsedMarijuana and larger values of DMHNG, which are detected clearly under the full dataset analysis.

Figure 7: Posterior predictive medians and 95% HPD intervals for the predictive 90th quantiles of DMHNG by race-gender-marijuana use and comparing models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). The CC analysis produces wider intervals with more overlap between marijuana users and non-users across all strata, which dilutes the strong, significant, and adverse effects detected by GMC-MA fit to the full dataset.

This point is further emphasized by Figure 8, which reports the difference between the posterior predictive medians of $\hat{Q}(\tilde{Y})$ for marijuana users and non-users in each stratum. The estimated differences are positive for all strata: the 90th quantile of DMHNG is greater for marijuana users than non-users. Notably, all points lie above the 45 degree line, which implies that these estimated differences are consistently larger for the GMC-MA on the full dataset compared to GMC-MA (CC). Thus, accounting for missingness not only impacts the spread of these posterior predictive distributions, but also shifts the location. And, most critically, the CC analysis attenuates the estimated differences between marijuana users and non-users across all strata. This attenuation is nontrivial: for example, the estimated difference in the 90th quantile of DMHNG between marijuana users and non-users for white males is 10 for the GMC-MA on the full dataset, but only 5 for GMC-MA (CC). Similar trends are observed for the 75th quantile, although the discrepancies are less pronounced (see the supplementary
material).

Figure 8: Difference in posterior predictive medians for the predictive 90th quantiles of DMHNG between marijuana users and non-users and comparing models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). The CC point estimates attenuate the differences between marijuana users and non-users across all strata, which dilutes the strong, significant, and adverse effects detected by GMC-MA fit to the full dataset.

These results emphasize the serious risks posed by CC analysis, which can produce biased or misleading conclusions. The implications are particularly important for mental health studies: accurate estimation of the relationship between specific behaviors or attributes and proxies for at-risk individuals is vital. By using the GMC-MA, we were able to perform full dataset analysis—despite abundant missingness, mixed data types, and complex marginal and joint distributions—and highlight the specific limitations of CC analysis.

7 Conclusion

We proposed a nonparametric copula model for mixed (count, continuous, ordinal, and unordered categorical) data types subject to informative missingness, in particular missing-at-random (MAR) data. The model features a latent mixture of factor models to induce a nonlinear and scalable Gaussian mixture copula model. We employed the rank-probit likelihood for posterior inference, which circumvents the need to specify marginal distributions yet maintains strong posterior consistency for the parameters of the underlying copula model. A central innovation was the introduction and theoretical analysis of the margin adjustment, which delivers consistent inference for each marginal distribution under rank-based copula
models with no further modeling assumptions and minimal additional computing cost. The margin adjustment eliminates any reliance on the ECDF for prediction and imputation, which is the default approach in rank-based copula models yet can be severely biased under MAR. Using carefully-designed simulation studies, we demonstrated the significant improvements in imputation and marginal distribution estimation for the proposed approach relative to state-of-the-art alternatives, especially in the presence of nonlinear dependencies, mixed data types, and informative missingness. We applied our model and imputation strategies to self-reported mental health data from NHANES and demonstrated the pitfalls of complete case analysis—and showed how the proposed approach may resolve these issues.

There are numerous interesting directions for future work. First, the proposed Gaussian mixture copula model and the margin adjustment apply not only to imputation, but also to prediction. Our strong theoretical results suggest that the proposed framework may prove useful for posterior prediction of multivariate and mixed data, especially in the presence of missingness. Similarly, our theoretical analyses suggest that rank-probit likelihoods—along with the margin adjustment—may apply more broadly to non-Gaussian copula models. Such developments would broaden the applicability of Bayesian inference for copula models, while simultaneously eliminating the need to specify models for each marginal distribution yet potentially delivering consistent inference for these margins. Finally, an important and challenging extension is to adapt the proposed framework for missing-not-at-random (MNAR) data. For instance, the latent factor mixture (14) is an attractive option for parsimonious joint modeling of the missingness mechanism and the observed data in a low-rank, shared parameter model (e.g., Creemers et al., 2010).
Supplement to
“Nonparametric Copula Models for Mixed Data with Informative Missingness”

A Figures

All figures referenced in this supplement are available at its conclusion.

B Proofs

Theorem 1. Suppose \( \{Z_i\}_{i=1}^{n} \overset{iid}{\sim} F_Z \) and \( \{Y_i\}_{i=1}^{n} \overset{iid}{\sim} F_Y \), where \( F_Z \) is continuous and both \( \{Z_i\}_{i=1}^{n} \) and \( \{Y_i\}_{i=1}^{n} \) are ordered (non-decreasing). Defining \( Z^n(x) := \max\{Z_i : Y_i \leq x, i \in \{1, \ldots, n\}\} \), the margin adjustment satisfies \( \tilde{F}(x) := F_Z\{Z^n(x)\} \overset{a.s.}{\rightarrow} F_Y(x) \) for all \( x \).

Proof. Fix \( \tau = F_Y(x) \) for a given \( x \) and let \( S \) denote the position of \( Z^n(x) \), which is also the maximum position of \( x \) due to the ordering of \( \{Z_i\}_{i=1}^{n} \) and \( \{Y_i\}_{i=1}^{n} \). By the Glivenko-Cantelli Theorem, \( S/n = n^{-1} \sum_{i=1}^{n} I\{Y_i \leq x\} \overset{a.s.}{\rightarrow} \tau \). Now, consider the random variable \( U_i = F_Z(Z_i) \), which is uniform on \((0, 1)\) with \( \{U_i\}_{i=1}^{n} \) ordered. Therefore, the \( S \)th order statistic of \( \{U_i\}_{i=1}^{n} \) satisfies \( U^{(S)} = F_Z\{Z^n(x)\} \). It is well known that \( U^{(S)} \sim \text{Beta}(S, n-S+1) \), with \( E[U^{(S)}] = S/(n+1) \) and \( V[U^{(S)}] = Sn/\{(S+n)^2(S+n+1)\} < n^{-1} \). As such, \( V[U^{(S)}] \to 0 \) as \( n \to \infty \), so \( U^{(S)} \) converges in distribution to a degenerate random variable with point mass at the limit of its expectation: \( S/(n+1) \overset{a.s.}{\rightarrow} \tau \). Since \( \tau \) is fixed, this also implies that \( U^{(S)} \overset{D}{\to} \tau \). Finally, observe that the sequence \( Z^n(x) \) is monotone in \( n \), which implies that the sequence \( U^{(S)} \) is also monotone in \( n \). Coupling monotonicity and convergence in probability, we have that \( U^{(S)} = F_Z\{Z^n(x)\} \overset{a.s.}{\rightarrow} \tau = F_Y(x) \). \( \square \)

Theorem 2. Suppose \( Z_i = (Z_{i1}, Z_{i2}) \overset{iid}{\sim} G \) and \( Y_i = (Y_{i1}, Y_{i2}) \overset{iid}{\sim} F \) for \( i = 1, \ldots, n \), where \( G \) is continuous with marginal distributions \( G_1, G_2 \) and \( Y_{ij} = F_j^{-1}\{G_j(Z_{ij})\} \) for \( j = 1, 2 \) with marginal distributions \( F_1, F_2 \). Suppose that \( Y_1 \) is missing-at-random and \( Y_2 \) is completely observed. Define \( Z^n(x) := \max\{Z_{i1} : Y_{i1}^{obs} \leq x\} \). Then the margin adjustment satisfies \( \tilde{F}_1(x) := G_1\{Z^n(x)\} \overset{a.s.}{\rightarrow} F_1(x) \) for all \( x \).

Proof. For this proof, we will use upper case letters to denote random variables, lower case letters for observed data, and bold face for vectors. Probabilities are given by \( P(x) \), where
the subscript refers to the respective (marginal, conditional, or joint) distribution. The proof will show that $Z_n^n(x) \overset{a.s.}{\to} G^{-1}_1\{F_1(x)\}$ as $n \to \infty$, which yields the stated result via the continuous mapping theorem.

First, suppose that $Y_1$ and $Y_2$ are continuous; the discrete case is addressed subsequently. Given that $Y$ is a component-wise monotone transformation of $Z$, the joint distribution $F$ may be expressed in terms of $G$:

$$F(y_1, y_2) = G[G^{-1}_1\{F_1(y_1)\}, G^{-1}_2\{F_2(y_2)\}].$$  \hspace{1cm} (21)

Similarly, the conditional probability $P(Y_1 \leq x \mid Y_2 = y)$ can be expressed in terms of $Z$:

$$P(Y_1 \leq x \mid Y_2 = y) = P[Z_1 \leq G^{-1}_1\{F_1(x)\} \mid Z_2 = G^{-1}_2\{F_2(y)\}]$$

$$\hspace{2cm} = \int_{-\infty}^{G^{-1}_1\{F_1(x)\}} \frac{g(z, G^{-1}_2\{F_2(y)\})}{g_2[G^{-1}_2\{F_2(y)\}]} \, dz$$  \hspace{1cm} (22)

where $g$ and $g_2$ are the density functions of $G$ and $G_2$, respectively.

Now, consider the sequence of probabilities $P_{Z_1|R_1=0}\{Z_1 \leq Z^n_1(x)\}$, i.e., the marginal probability that $Z_1$ is less than $Z^n_1(x)$ given that $Y_1$ is observed ($R_1 = 0$). This sequence is monotone increasing because $Z^n_1(x)$ is monotone increasing, and clearly bounded above by one. Thus, it converges almost surely to its limit by the monotone convergence theorem. Therefore, $Z^n_1(x)$ must also converge almost surely to a limit, which we will label $Z^{\infty}_1(x)$.

The conditional probability is equivalently

$$P_{Z_1|R_1=0}\{Z_1 \leq Z^n_1(x)\} = E_{Z_2|R_1=0}[P_{Z_1|Z_2,R_1=0}\{Z_1 \leq Z^n_1(x)\}]$$  \hspace{1cm} (23)

where the expectation is taken with respect to the distribution of $Z_2$ given that $Y_1$ is observed (i.e., not missing). Because $R$ is missing-at-random, $Z_1$ is conditionally independent of $R_1$ given $Z_2$, so

$$P_{Z_1|Z_2,R_1=0}\{Z_1 \leq Z^n_1(x)\} = P_{Z_1|Z_2}\{Z_1 \leq Z^n_1(x)\} = \int_{-\infty}^{Z^n_1(x)} \frac{g(z, Z_2)}{g_2(Z_2)} \, dz$$  \hspace{1cm} (24)

and thus

$$P_{Z_1|R_1=0}\{Z_1 \leq Z^n_1(x)\} = E_{Z_2|R_1=0}[P_{Z_1|Z_2}\{Z_1 \leq Z^n_1(x)\}]$$  \hspace{1cm} (25)
Denoting (25) by \( h\{Z_1^n(x)\} \), note that (24)–(25) imply that \( h \) is a continuous function. Consequently, we can write the limit of \( P_{Z_1|R_1=0}\{Z_1 \leq Z_1^n(x)\} \) explicitly, yielding that

\[
P_{Z_1|R_1=0}\{Z_1 \leq Z_1^n(x)\} = h\{Z_1^n(x)\} \overset{a.s.}{\to} h\{Z_1^\infty(x)\}. \tag{26}
\]

Next, consider an application of Theorem 1 to the observed data. Upon ordering \( Z_{i1}^{obs} \) and \( Y_{i1}^{obs} \), \( Z_1^n(x) \) and the maximum position of \( x \) will have the same position because \( Y_1 \) is a monotone transformation of \( Z_1 \). By Theorem 1, this implies that if \( x \) is the \( \tau \)th quantile under the distribution of observed \( [Y_1 \mid R_1 = 0] \), then \( Z_1^n(x) \) will converge to the \( \tau \)th quantile under the distribution of \( Z_1 \) corresponding to observed \( [Y_1 \mid R_1 = 0] \):

\[
P_{Z_1|R_1=0}\{Z_1 \leq Z_1^n(x)\} \overset{a.s.}{\to} P_{Y_1|R_1=0}(Y_1 \leq x). \tag{27}
\]

We can also re-write \( P_{Y_1|R_1=0}(Y_1 \leq x) \) in terms of \( Z_1 \) and \( Z_2 \) by (22) and (25):

\[
P_{Y_1|R_1=0}(Y_1 \leq x) = P_{Z_1|R_1=0}[Z_1 \leq G_1^{-1}\{F_1(x)\}] \tag{28}
\]

\[
= E_{Z_2|R_1=0}(P_{Z_1|Z_2,R_1=0}[Z_1 \leq G_1^{-1}\{F_1(x)\}]) \tag{29}
\]

\[
= E_{Z_2|R_1=0}(P_{Z_1|Z_2}[Z_1 \leq G_1^{-1}\{F_1(x)\}]) \tag{30}
\]

\[
= h[G_1^{-1}\{F_1(x)\}]. \tag{31}
\]

Finally, we see the equivalence between (26) and (31) holds if and only if \( Z_1^\infty(x) = G_1^{-1}\{F_1(x)\} \), which implies that \( Z_1^n(x) \overset{a.s.}{\to} G_1^{-1}\{F_1(x)\} \). Once again, an application of the continuous mapping theorem demonstrates the consistency of the MA estimator at \( x \).

When \( Y_1 \) or \( Y_2 \) is discrete, the proof requires only minor modifications. First, the conditional probability \( P(Y_1 \leq x \mid Y_2 = y) \) can still be written in terms of \( Z \). Specifically, \( G_2^{-1}\{F_2(y)\} \) maps \( y \) to the interval \( (G_2^{-1}\{F_2^-(y)\}, G_2^{-1}\{F_2(y)\}] \) where \( F_2^-(y) \) is the left limit of \( F_2 \) at \( y \) (Zhao and Udell, 2020b). Therefore, for discrete \( Y_2 \), \( P(Y_1 \leq x \mid Y_2 = y) \) is now

\[
P(Y_1 \leq x \mid Y_2 = y) = P[Z_1 \leq G_1^{-1}\{F_1(x)\} \mid Z_2 \in (G_2^{-1}\{F_2^-(y)\}, G_2^{-1}\{F_2(y)\}]]
\]

\[
= \int_{-\infty}^{G_1^{-1}(F_1(x))} \int_{G_2^{-1}(F_2^-(y))}^{G_2^{-1}(F_2(y))} g(z_1, z_2) \, dz_2 \, dz_1
\]
If $Y_1$ is discrete, the event $Y_1 \leq x$ is equivalent to $Z_1 \leq G^{-1}\{F_1(x)\}$, where $F_1(x)$ is the right limit of $F_1$ at $x$. Therefore, the argument does not change, and the rest of the proof follows identically.

**Theorem 3.** Suppose $Y$ is comprised of $r$ numeric variables and $q = p - r$ unordered categorical variables, each with a potentially unique number of levels $k_c, c \in \{1, \ldots, q\}$. Let $p^* = r + \sum_{c=1}^q k_c$, and $\Pi$ be a prior distribution on the space of all $p^* \times p^*$ positive semi-definite correlation matrices $\theta$ with corresponding density $\pi(\theta)$ with respect to a measure $\nu$. Suppose $\pi(\theta) > 0$ almost everywhere with respect to $\nu$ and that $F_1, \ldots, F_r$ are the true marginal CDFs for the $r$ numeric variables. Furthermore, assume that the missigness mechanism is ignorable. Then, for $\theta_0$ a.e. $[\nu]$ and any neighborhood $A$ of $\theta_0$, we have that

$$\lim_{n \to \infty} \Pi\{\theta \in A \mid Z \in \mathcal{E}(Y^{obs})\} = 1 \text{ a.s } [G^\infty_{\theta_0,F_1,\ldots,F_r}]$$

where $G^\infty_{\theta_0,F_1,\ldots,F_r}$ is the data generating distribution of $\{y_i\}_{i=1}^\infty$ under the Gaussian copula and $Y^{obs}$ is the observed component of the data.

**Proof.** Because the missingness mechanism is ignorable, the priors for the parameters governing the data generating process for $Y$ and the missingness mechanism $R$ are independent. Therefore, we can utilize the variant of Doob’s theorem to prove the result, as was done in Murray et al. (2013).

**Doob’s Theorem.** (Gu and Ghosal, 2009) Let $X_i$ be observations whose distributions depend on a parameter $\theta$, both taking values in Polish spaces. Assume $\theta \sim \Pi$ and $X_i \mid \theta \sim P_\theta$. Let $X_N$ be the $\sigma$-field generated by $X_1, \ldots, X_N$, and $X_\infty = \sigma(\bigcup_{i=1}^\infty X_i)$. If there exists a $X_\infty$ measurable function $f$ such that for $(\omega, \theta) \in \Omega^\infty \times \Theta$, $\theta = f(\omega)$ a.e. $[P^\infty_\theta \times \Pi]$ then the posterior is strongly consistent at $\theta$ for almost every $\theta$ $[\Pi]$.

For the proof, it suffices to establish the existence of a consistent estimator $\hat{\theta}$ of $\theta_0$, the data generating Gaussian copula correlation matrix, that is measurable with respect to the $\sigma$-field generated by the sequence $\{\mathcal{E}(Y^{obs}_n)\}_{n=1}^\infty$, where $n$ indexes the sample size.

Suppose $Y_n$ is comprised of $n_1 > 1$ complete cases without any missing values ($Y^{CC}$) and $n_2$ cases with missing values for at least one variable ($Y^{inc}$) such that $n_1 + n_2 = n$. Next, suppose that $Y_n$ is comprised of $q$ unordered categorical variables, each with a potentially
unique number of levels $k_1, \ldots, k_q$, and $r = p - q$ numeric variables which may be discrete or continuous. Augment $Y_n$ with the full binarization of categorical variables in Equation (6) such that that data has $p^* = r + \sum_{c=1}^q k_c$ columns.

For each observation $i$ in $Y_{CC}$, for each variable $j \in \{1, \ldots, p^*\}$ with $y_{ij} = x$, consider the margin adjustment (10) with $Z_n(x) = \max \{Z_{ij} : Y_{ij}^{obs} \leq x\}$. Let $T_{nij} = \sum_{i=1}^{n} I\{z_{ij} \leq Z_n^j(x)\}$, with $T_n(Y^{obs}_i) = (T_{n1}, \ldots, T_{nip^*})$ and $T_n(Y^{obs}_{n}) = \{T_n(Y^{obs}_i)\}_{i=1}^{n}$. Note that $Z_n^j(x)$ is an element of $E(Y^{obs}_{n})$, the extended rank-probit likelihood induced by $Y^{obs}_{n}$. Therefore, $T_{nij}$ is measurable with respect to $E(Y^{obs}_{n})$. Hence, any function that is measurable with respect to $\mathcal{T}_N$, the $\sigma$-algebra generated by the sequence $\{T_n(Y^{obs}_{n})\}_{n=1}^{N}$, is also measurable with respect to the $\sigma$-algebra generated by the corresponding sequence of extended rank-probit likelihoods induced by $Y^{obs}_{n}$. Consequently, as in Murray et al. (2013), we exclusively work with $\mathcal{T}_N$.

Now, $Z_n^j(x)$ is a random variable, and hence $T_{nij}$ is a random variable. We work with its expectation under the true data generating model. Define $\hat{U}_{nij} = E[T_{nij}]/(n + 1)$ and $\hat{U}_{ni} = (\hat{U}_{n1}, \ldots, \hat{U}_{nip^*})$. Then,

$$\hat{U}_{nij} = \frac{1}{n + 1} E \sum_{l=1}^{n} I\{z_{lj} \leq Z_n^j(x)\}$$

$$= \frac{1}{n + 1} \sum_{l=1}^{n} P\{z_{lj} \leq Z_n^j(x)\} \quad (33)$$

By the SLLN and Theorem 2, (33) converges almost surely to

$$P\{z_j \leq \Phi^{-1}\{F_j(x)\}\} = F_j(x) \quad (34)$$

Therefore, we have that $\hat{U}_{nij} \xrightarrow{a.s.} U_{ij}$, where $U_{ij} = F_j(x)$, the cumulative marginal probability of $x$ for variable $j$ under the true distribution $F_j$. Consequently, $\hat{U}_{ni} \xrightarrow{a.s.} U_i = (U_1, \ldots, U_{p^*})$, and $U_i$ is $\mathcal{T}_\infty$ measurable.

Note that we have augmented each of the $q$ categorical variables to their associated binary representation. As Murray et al. (2013) note, for $j$ associated with a categorical sub-level (along with $j$ associated with discrete numerical variables), this means that $U_{ij}$ is merely a re-
labeling of $y_{ij}$, with marginal “cumulative probabilities” of the binary indicator. Therefore, through the RPL data augmentation of unordered categorical variables, we have introduced ‘pseudo’ distribution functions for each sub-level. Consequently, $U_i$ is sample from a Gaussian copula model with correlation $\theta_0$, where continuous margins are Uniform[0, 1] while the discrete margins and unordered categorical sub-levels are completely specified.

Therefore, without exception in the presence of unordered categorical variables, we may apply the argument of Murray et al. (2013), which states that $U_i$ is a sample from a Gaussian copula with correlation matrix $\theta_0$, and the existence of a consistent estimator follows from their proof. Specifically, the problem reduces to estimating polychoric/polyserial correlations with fixed margins, where $U_i$ is a regular parametric family admitting a consistent estimator of $\theta_0$.

**Corollary 1.** Under the conditions of Theorem 3, define $\tilde{F}_j$ as in (10) with $G_j = \Phi$ for each $j \in \{1, \ldots, r\}$. Then for any neighborhood $\mathcal{A}$ of $F_j(x)$,

$$
\lim_{n \to \infty} \Pi\{\tilde{F}_j(x) \in \mathcal{A} \mid Z \in \mathcal{E}(Y^{obs})\} = 1 \ a.s \ [G_{\theta_0, F_1, \ldots, F_r}^\infty]
$$

for all $x$.

**Proof.** This result follows from an application of Doob’s Theorem presented above and Theorem 2. Specifically, $\tilde{F}_j(x)$ is measureable with respect to the $\sigma$-field generated by the sequence \{\mathcal{E}(Y^{obs})_m\}_m=1^n, since $Z^n_j(x)$ is an element of $\mathcal{E}(Y^{obs})$, the set of orderings on $Z^{obs}_n$ consistent with the orderings on $Y^{obs}_n$. Since $\tilde{F}_j(x)$ is a strongly consistent estimator of $F_j(x)$ by Theorem 2, it follows that $F_j(x)$ is $\{\mathcal{E}(Y^{obs})_n\}_n=1^\infty$ measurable. Therefore, we have shown that the consistent estimator $\tilde{F}_j(x)$ of $F_j(x)$ is measureable with respect to the $\sigma$-field generated by the sequence $\{\mathcal{E}(Y^{obs})_n\}_n=1^\infty$ and thus, the posterior of $\tilde{F}_j(x)$ is strongly consistent at $F_j(x)$. This applies for each continuous, count, or ordinal variable $j = 1, \ldots, r$.

**Corollary 2.** Under the conditions of Theorem 3, let $p^c_{\ell} = P(Y_c = c_{\ell}), c \in \{r+1, \ldots, p\}, \ell \in \{1, \ldots, k_c\}$. Then, for any neighborhood $\mathcal{A}$ of $p^c_{\ell}$,

$$
\lim_{n \to \infty} \Pi\{p^c_{\ell} \in \mathcal{A} \mid Z \in \mathcal{E}(Y^{obs})\} = 1 \ a.s \ [G_{\theta_0, F_1, \ldots, F_r}^\infty]
$$
**Proof.** The result follows from an immediate application of the continuous mapping theorem. Define $p^{c,\ell} = P(z_c > 0, z_{c-\ell} < 0)$ under the RPL Gaussian copula. This quantity is a functional of the strongly consistent posterior of $\theta$ under the conditions in Theorem 3. Therefore, the continuous mapping theorem preserves the strong posterior consistency of $p^{c,\ell}$ at $p^{c,\ell}_0$.

**Theorem 4.** $\mathbb{C}_{GMC}$ defines a valid copula

**Proof.** To prove that $\mathbb{C}_{GMC}$ defines a valid copula, we verify that it satisfies the following three properties:

1. $\mathbb{C}_{GMC}(u_1, \ldots, u_p)$ is non-decreasing in each component $j \in \{1, \ldots, p\}$

Let $j \in \{1, \ldots, p\}$ be arbitrary and consider $u_{j_1} < u_{j_2}$. Define $z_{j_1} = \psi_j^{-1}(u_{j_1})$ and $z_{j_2} = \psi_j^{-1}(u_{j_2})$. Because $\psi_j$ is a valid continuous distribution function, it is monotone, and therefore $z_{j_1} < z_{j_2}$.

Consider the ratio

$$
\frac{\mathbb{C}_{GMC}(u_1, \ldots, u_{j_1} \ldots, u_p)}{\mathbb{C}_{GMC}(u_1, \ldots, u_{j_2} \ldots, u_p)} = \frac{\Psi\{\psi_1^{-1}(u_1), \ldots, z_{i_1}, \ldots, \psi_p^{-1}(u_p)\}}{\Psi\{\psi_1^{-1}(u_1), \ldots, z_{j_2}, \ldots, \psi_p^{-1}(u_p)\}}
$$

$$
= \sum_{h=1}^H \pi_h \frac{\Phi\{\{\psi_1^{-1}(u_1), \ldots, z_{j_1}, \ldots, \psi_p^{-1}(u_p)\}; \alpha_h, \sigma_h^2\}}{\Phi\{\{\psi_1^{-1}(u_1), \ldots, z_{j_2}, \ldots, \psi_p^{-1}(u_p)\}; \alpha_h, \sigma_h^2\}}
$$

By the properties of multivariate Gaussian random vectors, the sum simplifies to

$$
\sum_{h=1}^H \pi_h \frac{\Phi\{z_{j_1}; \{\psi_i^{-1}(u_i)\}_{i \neq j}, \alpha_h^*, \sigma_h^{*2}\}}{\Phi\{z_{j_2}; \{\psi_i^{-1}(u_i)\}_{i \neq j}, \alpha_h^*, \sigma_h^{*2}\}} < \sum_{h=1}^H \pi_h = 1 \quad (35)
$$

$$
\Rightarrow \mathbb{C}_{GMC}(u_1, \ldots, u_{i_1} \ldots, u_p) < \mathbb{C}_{GMC}(u_1, \ldots, u_{i_2} \ldots, u_p) \quad (36)
$$

where $\alpha_h^*, \sigma_h^{*2}$ are the conditional mean and variance of the Gaussian random variable obtained by conditioning on $\{\psi_i^{-1}(u_i)\}_{i \neq j}$ for cluster $h$. The inequality is due to the fact univariate Gaussian distribution functions are strictly monotone, implying that the ratio inside the sum in (35) is strictly less than 1 for each component $h$.

2. For any $j \in \{1, \ldots, p\}$, $\mathbb{C}_{GMC}(u_1 = 1, \ldots, u_j = u, \ldots, u_p = 1) = u$
Note that $\forall j \in \{1, \ldots, p\}$, $\psi^{-1}_j(1) = \sum_{h=1}^{H} \pi_h \Phi^{-1}\{1; (\alpha_h)_j, (C_h)_{jj}\} = \infty$.

Using the above result, it is simple to see that

$$C_{GMC}(u_1 = 1, \ldots, u_j = u, \ldots, u_p = 1) = \Psi\{\infty, \ldots, \psi^{-1}_j(u), \ldots, \infty\}$$

$$= \psi_j\{\psi^{-1}_j(u)\}$$

$$= u$$

3. For $a_j < b_j$, $a_j, b_j \in [0, 1]$, $j = 1, \ldots, p$, $C_{GMC}(u_1 \in [a_1, b_1], \ldots, u_p \in [a_p, b_p]) \geq 0$

$$C_{GMC}(u_1 \in [a_1, b_1], \ldots, u_p \in [a_p, b_p]) = C_{GMC}(u_1 \leq b_1, \ldots, u_p \leq b_p) - C_{GMC}(u_1 \leq a_1, \ldots, u_p \leq a_p)$$

$$= \Psi\{\psi^{-1}_1(b_1), \ldots, \psi^{-1}_p(b_p)\} - \Psi\{\psi^{-1}_1(a_1), \ldots, \psi^{-1}_p(a_p)\}$$

$$\geq 0$$ (By 1.)

\[\square\]

C  Gibbs Sampling Algorithm

Bayesian estimation of the GMC-MA in the presence of missing data alternates sampling model parameters from their marginal posteriors conditional on complete latent data, and then sampling latent data corresponding to $Y^{mis}$ given $Z^{obs}$ and model parameters. Two aspects of our model simplify this task. First, the margin adjustments $\{\tilde{F}_j\}_{j=1}^p$ are functionals of posterior samples of GMC parameters, and second, GMC parameters depend only latent $Z = (Z^{mis}, Z^{obs})$ through the RPL. Conjugate priors for GMC parameters allow for simple Gibb’s sampling steps, while the factor model 14 allows for independence among the components of $z_i$ conditional on $\eta_i$. Consequently, the sampling of $Z^{mis}$ is quite efficient, as the predictive distribution for each component is conditionally univariate normal.

The algorithm is broken down in to five blocks for simplicity. In each, $z_i$ is assumed complete, meaning that components corresponding to missing values in $y_i$ ($z^{mis}_i$) components have been sampled.

1. **Sample Cluster Specific Parameters** For each cluster $1, \ldots, H$
• $c_i \sim \text{Multinomial}(p)$, where $p = (p_1, \ldots, p_H)$ and $p_h \propto \pi_h \psi(c_i \mid \eta; \mu_h, \Delta_h)$

• $V_h \sim \text{Beta} (1 + n_h, \alpha_0 + \sum_{v=h+1}^H n_v), h = 1, \ldots, H - 1, n_h = \sum_{i=1}^N 1(c_i = h)$

• $\Delta_h \sim \text{IW}(\nu_{post}, \Psi_{post}), \nu_{post} = \nu_0 + n_h, \Psi_{post} = I_k + S_h + \frac{\kappa_{0n}^2}{\kappa_0 + n_h} T_h, S_h = \sum_{i:c_i=h}(\eta_i - \overline{\eta}) (\eta_i - \overline{\eta})^T, T_h = (\mu_0 - \overline{\eta})(\mu_0 - \overline{\eta})^T, \overline{\eta} = \eta_h^1 \sum_{i:c_i=h} \eta_i$

• $\mu_h \sim N(\mu_{post}, \kappa_{post}^{-1} \Delta_h), \mu_{post} = \frac{\kappa_{0n} \eta_0 + n_h \eta_h}{\kappa_0 + n_h}, \kappa_{post} = n_h + \kappa_0$

• $\alpha_0 \sim \text{Gamma}(a_0 + H - 1, b_0 - \sum_{h=1}^H \log(V_h))$

2. **Sample Factor Model Parameters**

• $\eta_i \mid c_i = h, \sim N_k((\Delta_h^{-1} + (\Lambda^T \Sigma^{-1} \Lambda)^{-1})^{-1} (\Lambda^T \Sigma^{-1} z_i + \Delta_h^{-1} \mu_h), (\Delta_h^{-1} + (\Lambda^T \Sigma^{-1} \Lambda)^{-1})^{-1})$

• $\lambda_{jh} \sim N((D_j^{-1} + \sigma_j^{-2} \eta^T \eta)^{-1} \eta^T \sigma_j^{-2} z_j), \text{where } D_j^{-1} = \text{diag}(\phi_{j1} \tau_1, \ldots, \phi_{jk} \tau_k), \ z_j = (z_{1j}, \ldots, z_{nj})^T, \text{and } \eta = (\eta_{1j}, \ldots, \eta_{nj})^T, \text{for } j = 1, \ldots, p$

• $\sigma_j^{-2} \sim \text{Gamma}(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2} \sum_{i=1}^n (z_{ij} - (\lambda_j^T \eta_{ij}))^2, \text{for } j = 1, \ldots, p$

• $\phi_{jh} \sim \text{Gamma}(\frac{\nu+1}{2}, \frac{\nu + \eta_h \lambda_j^2}{2}), \text{for } j = 1, \ldots, p, \ h = 1, \ldots, k$

• $\delta_1 \sim \text{Gamma}(a_1 + \frac{p_k}{2}, 1 + \frac{1}{2} \sum_{l=1}^k \tau_{l1}^{(1)} \sum_{j=1}^p \phi_{jl} \lambda_{jl}^2), \text{and for } h \geq 2$

  • $\delta_h \sim \text{Gamma}(a_1 + \frac{p(k-h+1)}{2}, 1 + \frac{1}{2} \sum_{l=1}^k \tau_{lh}^{(h)} \sum_{j=1}^p \phi_{jl} \lambda_{jl}^2), \text{where } \tau_{lh} = \prod_{t=1,t\neq h}^l \delta_t, \text{for } h = 1, \ldots, k$

3. **Re-sample $Z_{ij}^{obs}, Z_{ij}^{mis}$**

Given the conditional independence among the components of $Z_i$ given $\eta_i$, components of $Z$ corresponding to observed data points are sampled column-by-column, consistent with the ordering induced by the RPL. For components of $Z_i$ associated with missing values, no ordering is imposed, and only the diagonal orthant restriction for categorical variables is enforced.

• **Missing categorical/binary data:** If $j$ corresponds to one of the levels of categorical variable $q$ with $k_q$ levels, $Z_{ij}^{mis}$ and the other associated levels of $q$ must be sampled consistently with the diagonal orthant set restriction of the RPL. That is, one component of the vector $Z_{k_q}$ must be positive while the others negative. To ensure this condition is met, we first calculate the predictive probability that
\( Y_{ij}^{mis} \) assumes the \( j \)th level for each \( j \) among the \( k_q \) levels, which is equivalent to

\[
P(Z_{ij}^{mis} > 0 \cap Z_{i\ell-j}^{mis} < 0 \mid -) \propto 1 - \Phi(0; \sum_{h=1}^{k} \lambda_{jh} \eta_{hk}, \sigma_j^2) \prod_{\ell \in \{c_1, \ldots, c_{k_q}\}, \ell \neq j} \Phi(0; \sum_{h=1}^{k} \lambda_{th} \eta_{hk}, \sigma_\ell^2), j \in \{c_1, \ldots, c_{k_q}\}
\]

Then, we sample the level of \( Y_{ij}^{mis} \) using these probabilities, with the resulting classification used in the re-sampling of \( Z_{ij}^{mis} \) under the RPL. Let \( \text{TN}(\mu, \sigma^2, a, b) \) denote a truncated univariate normal with mean \( \mu \), variance \( \sigma^2 \), lower truncation \( a \), and upper truncation \( b \). The re-sampling step for \( Z_{ij}^{mis} \) is given by

\[
z_{ij}^{mis} \sim \begin{cases} 
\text{TN}(\sum_{h=1}^{k} \lambda_{jh} \eta_{hk}, \sigma_j^2, 0, \infty), & y_{ij}^{mis} = 1 \\
\text{TN}(\sum_{h=1}^{k} \lambda_{jh} \eta_{hk}, \sigma_j^2, -\infty, 0), & y_{ij}^{mis} = 0
\end{cases}
\]

If \( j \) is binary, the probability of one level versus the other is instead given by

\[
P(z_{ij}^{mis} > 0 \mid c_i = h, -) = 1 - \Phi_h(0; \sum_{h=1}^{k} \lambda_{jh} \eta_{hk}, \sigma_j^2), \text{ but the re-sampling step 38 remains the same}
\]

- **Missing numeric data:** In this case, latent \( Z_{ij}^{mis} \) is sampled from the unrestricted univariate Gaussian

\[
Z_{ij}^{mis} \mid - \sim N(\sum_{h=1}^{k} \lambda_{jh} \eta_{hk}, \sigma_j^2)
\]

- **Observed data:** for each column, sample \( Z_{ij}^{obs} \) from a truncated normal, with lower and upper bounds for each observation specified by the RPL:

\[
Z_{ij}^{obs} \mid - \sim \text{TN}(\sum_{h=1}^{k} \lambda_{jh} \eta_{hk}, \sigma_j, z_{ij}^{obs}, z_{ij}^{obs})
\]

For ordinal, count, and continuous variables, the truncation limits are \( z_{ij}^{obs} = \max\{z_{ij}^{obs} : y_{obs-i-j} < y_{ij}^{obs}\} \), and \( z_{ij}^{obs} = \min\{z_{ij}^{obs} : y_{obs-i-j} > y_{ij}^{obs}\} \), where \( z_{ij}^{obs} = z_{ij}^{obs} \setminus z_{ij}^{obs} \). For columns corresponding to categorical levels, the upper and lower
truncation limits are

\[
\begin{align*}
\tilde{Z}_{ij}^{\text{obs}} &= \begin{cases} 
0, & Y_{ij}^{\text{obs}} = 1, \\
-\infty, & Y_{ij}^{\text{obs}} = 0.
\end{cases} \\
\tilde{Z}_{ij}^{\text{mis}} &= \begin{cases} 
\infty, & Y_{ij}^{\text{obs}} = 1, \\
0, & Y_{ij}^{\text{obs}} = 0.
\end{cases}
\end{align*}
\] (41)

4. Sample \( \tilde{F}_j \)

For each unique \( x \in Y_j^{\text{obs}} \), we first find \( Z_{ij}^n(x) = \max\{Z_{ij}^{\text{obs}} : Y_{ij}^{\text{obs}} \leq x\} \), and compute

\[
\tilde{F}_j^n(x) = \psi_j \{Z_j^n(x)\}
\]

Where \( \psi_j \) is a function of the current draw of GMC parameters. To estimate \( \tilde{F}_j \) across unobserved values, we then fit a monotone interpolating spline to \( \{x, \tilde{F}_j^n(x)\} \) as described in Section 4, and use this estimate to approximate \( \tilde{F}_j(x') \) for \( x' / \in Y_j^{\text{obs}} \).

The smoothing step in the sampling of \( \tilde{F}_j \) is crucial in multiple imputation, as the transformation \( Y_{ij}^{\text{mis}} = \tilde{F}_j^{-1}(Z_{ij}^{\text{mis}}) \) provides realizations that may assume values across the entire support of variable \( j \), instead of only values that were observed.

D Hyperparameter Tuning

Throughout our simulated and real data studies, we find that the default model specification given in Section 4 requires little hyperparameter tuning. For hyperparameter values governing the multiplicative gamma process prior on \( \Lambda \), see Feldman and Kowal (2022) for recommended values.

Otherwise, there are three parameters that we vary to bolster model performance. The first is the dimension of latent \( \eta \). For all of our studies, we use a default value of \( \lceil 0.7p^* \rceil \), where \( p^* \) is the dimension of the augmented data matrix under the RPL. Next, we modify the scaling constant, \( \delta \), from the normal-inverse Wishart prior specified for the cluster specific
components from the mixture model on $\eta_i$. Recall the following hierarchical model structure

$$z_i \sim N(\Lambda \eta_i, \Sigma)$$

$$\eta_i \sim \sum_{h=1}^{H} \pi_h N_k(\mu_h, \Delta_h)$$

$$(\mu_h, \Delta_h) \sim \text{NIW}(\mu_0, \delta^2 I_k, \kappa_0, \nu_0)$$

In the simulation studies and real data analysis, we found that $\delta$ impacts model fit through the number of clusters that are discovered. Though we recommend a default value of $\delta = 10$, we find that generally, decreasing $\delta$ has the effect of increasing the number of clusters discovered. As such, we use $\delta = 5$ in the second simulation study as a lack of separability in the hybrid data reduces the stability of the repeated model fits with $\delta = 10$. Posterior predictive sampling can be employed to tune this parameter.

The other parameter tuned is the number of unique values present among each numeric variable for re-sampling to occur under the RPL data augmentation mentioned in Section C. For continuous variables, the number of unique levels observed in the data will be large, perhaps approaching $n$ or the number of rows in the sample data. As such, re-sampling columns of $Z$ associated with such variables above could be computationally intensive, as it would necessitate looping through all unique values of the continuous variable at each iteration of the MCMC.

Instead of enduring this expense, we instead choose an upper bound for the number of unique levels that a particular variable may have to engage in the re-sampling in step 3 of our MCMC algorithm for copula estimation. Any latent columns corresponding to observed variables having more than this number of unique levels are not re-sampled within the MCMC. Instead, they are initialized through a “pseudo-data” transformation, which results from scaling the column to have mean zero and unit variance. This is done to maintain the orderings and marginal properties of each continuous variable, like skew and multi-modality, which effectively improves the model and margin estimation while placing latent variables on a common scale. In our applications, we choose this threshold to be 350.
Section 6 utilizes posterior predictive inference to highlight discrepancies between a complete case analysis and one that accounts for potentially informative missing data. We include here the algorithm developed to produce the posterior predictive data sets used for this analysis.

The procedure is facilitated by the conditional independence implied by the factor model developed in Section 4. The algorithm begins with ordinary Gaussian mixture sampling steps for sampling of predictive $\tilde{\eta}_i$, which in turn enables sampling of predictive $\tilde{z}_i$. We then link each component of $\tilde{z}_i$ with $\tilde{F}_j^{-1}\{\psi_j(\tilde{z}_{ij})\}$ for numeric variables, or with the categorical link mentioned in Section C.

\begin{algorithm}
\caption{Simulation of a posterior predictive data set of size $n$}
\begin{algorithmic}
\Input One sample of GMC posterior parameters $\theta = (\{\pi\}_{h=1}^H, \Lambda, \Sigma, \{\alpha_h\}_{h=1}^H, \{\Delta_h\}_{h=1}^H, \{\tilde{F}_j\}_{j=1}^r)$
\Output One posterior predictive data set of size $n$, $\{\tilde{y}_i\}_{i=1}^n$
\For{$i$ in $1, \ldots, n$}
\State Sample cluster membership $c_i \sim \text{multinomial}(\{\pi\}_{h=1}^H)$
\State Sample latent factor $\tilde{\eta}_i \mid c_i = h \sim N(\alpha_h, \Delta_h)$
\For{$j$ in $1, \ldots, p^*$}
\If{$j$ corresponds to a binary variable}
\State Sample $\tilde{z}_{ij} \sim N(\sum_{h=1}^k \lambda_{jh} \tilde{\eta}_h, \sigma_j^2)$
\State $\tilde{y}_{ij} = 1 \iff \tilde{z}_{ij} > 0$
\EndIf
\If{$j$ corresponds to a categorical variable with $q$ levels indexed by $\{c_1, \ldots, c_q\}$}
\State Sample the predictive categorical level from the associated $q$-dimensional multinomial, where:
\State \begin{equation}
P(\tilde{y}_{ij} = 1 \mid -) = P(\tilde{z}_{ij} > 0 \cap \tilde{z}_{ik} \forall k \neq j < 0 \mid -) \label{eq:E.1}
\end{equation}
\State \begin{equation}
\alpha 1 - \Phi(b; \sum_{h=1}^k \lambda_{jh} \tilde{\eta}_h, \sigma_j^2) \prod_{\ell \in \{c_1, \ldots, c_q\}, \ell \neq j} \Phi(b; \sum_{h=1}^k \lambda_{jh} \tilde{\eta}_h, \sigma_j^2) \label{eq:E.2}
\end{equation}
\State \endIf
\EndIf
\EndIf
\EndFor
\end{algorithmic}
\end{algorithm}

45
F Further Simulation Results

F.1 Mixed Data Types, Nonlinearity, and Informative Missingness

As mentioned in Section 5.1, we also estimate the posterior distribution of the probability of a positive indicator for binary variable \( X_3 \). Under Model (14) and the RPL, this is simply the probability that latent \( Z_3 \) is greater than zero. This quantity is computed as \( 1 - \sum_{h=1}^{H} \pi_{s}^{h} \Phi\{0; (\Lambda_{s}^{s} \mu_{s}^{h})_{3}, (\Lambda_{s}^{s} \Lambda_{s}^{s} \Lambda_{s}^{s} + \Sigma^{s})_{33}\} \), where the superscript \( s \) denotes the \( s \)th posterior sample of model parameters.

To evaluate the proposed model, we compare the posterior probability of \( X_3 \) to a “ground truth” value of 0.335, which is the empirical probability of a positive indicator upon simulating 10,000,000 observations under the data generating model. In Figure 9, we plot the posterior probability of a positive indicator for each \((n, \beta)\) combination. In both plots, we use 10,000 posterior samples for inference, and see that the distributions contract around the ground truth value of 0.335 as the sample size increases, with expected precision loss due the amount of missing data caused by varying \( \beta \). This result further develops the potential for the categorical margin adjustment to be extended to more general copula models. For both \( \beta \) settings, the missingness mechanism badly biases the empirical estimate of the probability of a positive \( X_3 \); for \( \beta = 0.5 \), this probability is on average 0.26, while for \( \beta = 1 \), this probability is 0.23. Like numeric margins, we see posterior inference for binary proportions under the proposed approach correcting the bias caused by missing data, supporting the insight of Corollary 2.

Next, we include analogous plots to Figure 3 in the main paper for each additional \((n, \beta)\) combination in Figures 10-14. In each imputation procedure, the proposed approach is able to model non-linearity in the data, whereas the Gaussian copula (Hoff, 2007) is ineffective. Notice the consistency with which GMC-MA imputations capture specific features in the data, from the curvature in the relationship between \( Y_1 \) and \( Y_2 \) to the enhanced probability that \( Y_3 = 1 \) for large values of both \( Y_1 \) and \( Y_2 \), regardless of sample size and the amount of missingness. In addition, the margin adjustment relieves reliance on the ECDF for multiple imputation, yielding much more broad support in realized values of missing \( Y_2 \).
F.2 Imputation for Regression Analysis using Hybrid Simulations

Highlighted in Section 5.2, MICE-CART yields substantial coverage gaps in estimation of the interaction terms in the regression model of interest. We hypothesize that this is due to model inefficiency of CART when a parametric model is suitable. To visualize this, we plot the interaction between BMI, FI and New using the 10th completed data set under MICE with CART from several iterations of the repeated simulation study in Figure 15. We include corresponding ground truth data sets without missing values for comparison in the top row of the figure.

The first-order linear interaction model is clear, but MICE-CART is unable to model the differing linear slopes by family income. For instance, in the bottom-left panel, a group of individuals with high BMI and low values for New are classified as having high family income. However, in the ground truth data sets, there is a strong positive association between New and BMI for individuals with high family income. Clearly, MICE-CART is producing implausible imputed values, demonstrating the inadequacy of this method when simpler models suffice.

G Real Data Application

In Section 6 of the main paper, we check model calibration of the GMC-MA for complete case females. We include the same visual checks highlighted in Figure 16 for complete case males with similar conclusions. In each stratum, there is substantial overlap with the posterior predictive inference under the GMC-MA (CC) fit and ECDF estimates, which subsequently fails in certain cases for the full data fit. This result is consistent with what is presented in Section 6, which supports the notion that missing data may yield a biased complete case analysis.

To complete the information mentioned in the main paper, we also compare point estimates and uncertainty using the posterior predictive distribution of the 75th sample quantile in Figures 17-18, as was done in Figures 7-8 in the main paper. As expected, the differences between the full and complete case fits are not as pronounced, owing to fact that for most strata, between 70 and 90% of individuals have ≤ 10 DMHNG. However, several discrepancies still do arise, and some intervals that substantially overlap in the CC fit are much more clearly separated on the full data set.
Figure 9: Posterior Probability of a positive indicator for $X_3$: As $n$ increases the uncertainty decreases and the distribution concentrates around the ground truth (dotted line)
Figure 10: $n = 500, \beta = 0.5$
Figure 11: $n = 1000, \beta = 0.5$
Figure 12: $n = 500, \beta = 1$
Figure 13: $n = 1000, \beta = 1$
Figure 14: $n = 2000, \beta = 1$
Figure 15: MICE-CART completed data sets compared to ground truth from the 1st (left panel), 50th (middle panel) and 100th (right panel) iterations of the repeated simulation study from Section 5.2. MICE with CART is unable to capture the interactive relationship between \texttt{New}, \texttt{Age}, and \texttt{BMI}, as demonstrated by misclassified \texttt{FI} at the tails of \texttt{New} and \texttt{BMI}. These classifications leverage the regression fit, yielding inaccurate estimates and uncertainty quantification for the regression model of interest.
Figure 16: Posterior predictive summaries for models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). Among males, for each male race-gender-marijuana use stratum, we compare the 95% HPD intervals for the posterior predictive ECDFs, and include the ECDF on the CC data for reference. The GMC-MA (CC) output is well-calibrated to the observed data. By comparison, the GMC-MA fit to the full dataset produces intervals that are narrower and shifted, which suggests that the missingness mechanism is informative—and that CC analysis is unreliable.

Figure 17: Posterior predictive medians and 95% HPD intervals for the predictive 75th quantiles of DMHNG by race-gender-marijuana use and comparing models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). The CC analysis produces wider intervals with more overlap between marijuana users and non-users across all strata, which dilutes the strong, significant, and adverse effects detected by GMC-MA fit to the full dataset.
Figure 18: Difference in posterior predictive medians for the predictive 75th quantiles of DMHNG between marijuana users and non-users and comparing models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). The CC point estimates attenuate the differences between marijuana users and non-users across nearly all strata, which dilutes the strong, significant, and adverse effects detected by GMC-MA fit to the full dataset.
References

Bhattacharya, A. and Dunson, D. B. (2011). Sparse bayesian infinite factor models. *Biometrika*, pages 291–306.

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, 172(9):1070–1076.

Chandra, N. K., Canale, A., and Dunson, D. B. (2020). Escaping the curse of dimensionality in bayesian model based clustering. *arXiv preprint arXiv:2006.02700*.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M. G. (2010). A sensitivity analysis for shared-parameter models for incomplete longitudinal outcomes. *Biometrical Journal*, 52(1):111–125.

DeYoreo, M., Reiter, J. P., and Hillygus, D. S. (2017). Bayesian mixture models with focused clustering for mixed ordinal and nominal data. *Bayesian Analysis*, 12(3):679–703.

Dunson, D. B. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Feldman, J. and Kowal, D. R. (2022). Bayesian data synthesis and the utility-risk trade-off for mixed epidemiological data. *The Annals of Applied Statistics*, 16(4):2577–2602.

Gu, J. and Ghosal, S. (2009). Bayesian roc curve estimation under binormality using a rank likelihood. *Journal of statistical planning and inference*, 139(6):2076–2083.

Gustavson, K., von Soest, T., Karevold, E., and Røysamb, E. (2012). Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study and a monte carlo simulation study. *BMC public health*, 12(1):1–11.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283.
Horwitz, A. V. and Scheid, T. L. (1999). A handbook for the study of mental health: Social contexts, theories, and systems. Cambridge University Press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association, 96(453):161–173.

Ishwaran, H. and James, L. F. (2002). Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information. Journal of Computational and Graphical statistics, 11(3):508–532.

Joe, H. (2014). Dependence modeling with copulas. CRC press.

Johndrow, J., Dunson, D., and Lum, K. (2013). Diagonal orthant multinomial probit models. In Artificial Intelligence and Statistics, pages 29–38. PMLR.

Kowal, D. R. and Wu, B. (2021). Semiparametric count data regression for self-reported mental health. Biometrics.

Manrique-Vallier, D. and Reiter, J. P. (2013). Bayesian multiple imputation for large-scale categorical data with structural zeros.

Manrique-Vallier, D. and Reiter, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. Journal of Computational and Graphical Statistics, 23(4):1061–1079.

Manrique-Vallier, D. and Reiter, J. P. (2017). Bayesian simultaneous edit and imputation for multivariate categorical data. Journal of the American Statistical Association, 112(520):1708–1719.

Murray, J. S. (2018). Multiple imputation: a review of practical and theoretical findings. Statistical Science, 33(2):142–159.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). Bayesian gaussian copula factor models for mixed data. Journal of the American Statistical Association, 108(502):656–665.

58
Murray, J. S. and Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient bayesian inference for gaussian copula regression models. *Biometrika*, 93(3):537–554.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P., et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96.

Rajan, V. and Bhattacharya, S. (2016). Dependency clustering of mixed data with gaussian mixture copulas. In *IJCAI*, pages 1967–1973.

Reiter, J. P. (2012). Bayesian finite population imputation for data fusion. *Statistica Sinica*, pages 795–811.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

Tang, F. and Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377.

Tewari, A., Giering, M. J., and Raghunathan, A. (2011). Parametric characterization of multimodal distributions with non-gaussian modes. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 286–292. IEEE.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

Van Buuren, S. and Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. Leiden: TNO.
Zhao, Y. and Udell, M. (2020a). Matrix completion with quantified uncertainty through low rank gaussian copula. *Advances in Neural Information Processing Systems, 33*:20977–20988.

Zhao, Y. and Udell, M. (2020b). Missing value imputation for mixed data via gaussian copula. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 636–646.