Journal of Machine Learning Research 25 (2024) 1-50 Submitted 4/23; Revised 2/24; Published 5/24

Nonparametric Copula Models for Multivariate, Mixed, and Missing Data

Joseph Feldman JOSEPH.FELDMAN@DUKE.EDU
Department of Statistical Science
Duke University, Durham, NC 27708-0251, USA

Daniel R. Kowal DAN.KOWAL@CORNELL.EDU
Department of Statistics and Data Science
Cornell University, Ithaca, NY 14853-2601, USA
Department of Statistics
Rice University, Houston, TX 77251-1892, USA

Editor: Debdeep Pati

Abstract

Modern data sets commonly feature both substantial missingness and many 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 among (possibly many) variables of mixed data types. To address these challenges, we develop a novel Bayesian mixture copula for joint and nonparametric modeling of multivariate 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 computationally efficient strategy for marginal distribution estimation that eliminates the need to specify any marginal models yet delivers posterior consistency for each marginal distribution and the copula parameters under 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 nonparametrics, Bayesian inference, Factor models, Imputation, Mixture models

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 among many variables of mixed data types.

When confronted with missing data, there are two options for analysis. The first is to proceed using only observations for which all variables are observed. However, this *complete case* (CC) analysis, while common in practice, is highly problematic in many settings. CC analysis often substantially decreases the sample size, leading to imprecise and underpowered analysis. More critically, CC analysis can introduce various and significant forms of bias. Consider a sample of correlated bivariate data $\{(Y_{i1}, Y_{i2})\}_{i=1}^n$, and suppose that the missingness in Y_1 is determined by the value of Y_2, which is fully observed (missingness-at-random; see below). Figure 1 shows the potential impacts of a CC analysis: the empirical cumulative distribution function (ECDF) of Y_1 is severely biased, which implicates inference on $Q(Y_1)$ as well as popular Bayesian semiparametric copula models discussed subsequently (Hoff, 2007; Murray et al., 2013; Cui et al., 2019; Feldman and Kowal, 2022).

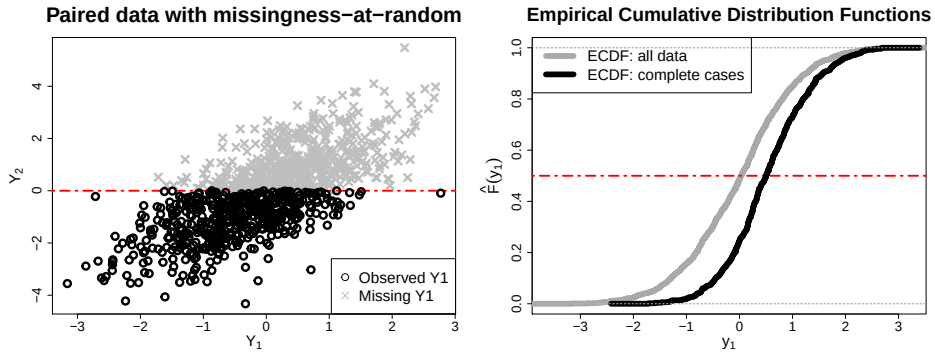

Figure 1: Bivariate data $\{(Y_{i1}, Y_{i2})\}_{i=1}^n$ with missing-at-random missingness (left) and the corresponding true and empirical cumulative distribution function (ECDF) for Y_1 (right). The missing data severely biases the ECDF, which impacts functionals of this term—including traditional statistics as well as Bayesian semiparametric copula models.

The second option, which we pursue here, is *imputation* of missing values. Informally, a statistical model is fit to the observed data and then used to repeatedly simulate the missing values, thus forming many completed data sets. Then, estimates $\hat{Q}$ are computed on each completed data set, and combined to produce point estimates and uncertainty quantification for Q. If the model adequately captures the features of the data, we can expect the inference based on an imputation procedure to correct the shortcomings of a CC analysis.

The specification of an imputation model is made precise by considering a joint model for all data $\boldsymbol{Y} = (Y_{ij})$ and binary missingness variables $\boldsymbol{R} = (R_{ij})$, where $R_{ij} = 1$ indicates that Y_{ij} is missing, and $R_{ij} = 0$ means that Y_{ij} is observed. Let $\boldsymbol{Y}^{obs} = (Y_{ij} : R_{ij} = 0)$ denote the observed data and $\boldsymbol{Y}^{mis} = (Y_{ij} : R_{ij} = 1)$ the missing values. We assume that this model is indexed by distinct parameters $\boldsymbol{\theta}$ for $\boldsymbol{Y}$ and $\boldsymbol{\phi}$ for $\boldsymbol{R}$, with joint likelihood

$$p(\boldsymbol{R}, \boldsymbol{Y}^{obs} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \int p(\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis} \mid \boldsymbol{\theta})\; p(\boldsymbol{R} \mid \boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}, \boldsymbol{\phi}) d\boldsymbol{Y}^{mis} \tag{1}$$

We focus on *missingness-at-random* (MAR), which allows the missingness mechanism to depend on the observed (but not missing) data: $p(\boldsymbol{R} \mid \boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}, \boldsymbol{\phi}) = p(\boldsymbol{R} \mid \boldsymbol{Y}^{obs}, \boldsymbol{\phi})$ (Rubin, 1976). In this case the missingness is *ignorable*, and the model specified on the observed data $p(\boldsymbol{Y}^{obs} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis} \mid \boldsymbol{\theta}) d\boldsymbol{Y}^{mis}$ may be used for imputation. A stronger assumption is *missing-completely-at-random* (MCAR), $p(\boldsymbol{R} \mid \boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis}, \boldsymbol{\phi}) = p(\boldsymbol{R} \mid \boldsymbol{\phi})$, which is a special case of MAR.

There are several important considerations for MAR. First, CC analysis is strongly inadvisable (see Figure 1), and thus imputation is needed in general. Second, MAR is most likely satisfied when $\boldsymbol{Y}^{obs}$ contains many potentially informative variables (Little, 2021). Thus, MAR demands a model capable of accommodating multiple variables, possibly of mixed types. Finally, the suitability of MAR in practice depends on the adequacy of the assumed model. In aggregate, MAR necessitates a model for multivariate and mixed data that can adapt to complex marginal and joint distributional features.

Our motivating example comes from a collection of variables (see Table 1) in the National Health and Nutrition Examination Survey (NHANES). These variables include count, continuous, ordinal, and unordered categorical variables, with missingness as high as 43% for some variables and missing values for each data type. 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 with intricate multivariate relationships. Most importantly, CC analysis is unsatisfactory or misleading for these data (see Section 7). Thus, an imputation model is required—and in particular one capable of accommodating many variables of mixed types with intricate distributional features.

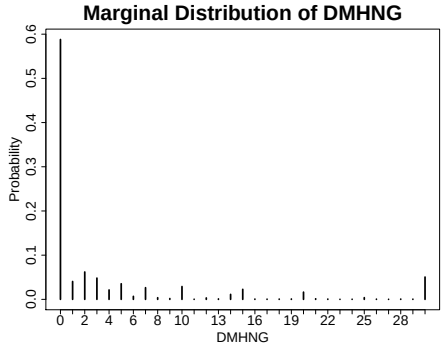

Figure 2: The marginal distribution of days of self-reported poor mental health (`DMHNG`) from the NHANES data, which is the response variable of interest in our real data analysis. Discreteness, boundedness, heaping, and zero-inflation combine to make modeling difficult.

The literature on imputation models is extensive, yet limited in its ability to address these critical challenges; see Murray (2018) for a thorough review. 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 data set conditional on all other variables and (ii) using each regression model to impute (separately) the missing values for each variable (Raghunathan et al., 2001; Van Buuren, 2007). This approach offers several advantages: it is amenable to mixed data types, allows

Variable	Values	% Missing
Response variable:		
DaysMentHlthNotGood (DMHNG)	$\{0, 1, \ldots, 30\}$	14%
Demographic and socioeconomic variables:		
Gender	Male, Female	0
Age (years)	$\{18, \ldots, 80\}$	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/m^2)	$[13.4, 81.2]$	6%
HasHighBP (BPQ020 at link)	Yes, No	0.1%
HasHighChol (BPQ080 at link)	Yes, No	6%
HasDiabetes*	Yes, No	0.08%

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

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 does not guarantee a valid joint distribution for the data, which is especially problematic for Bayesian inference, and is difficult to tune in high dimensions, since it requires a separate model fit for each variable. Perhaps most important, FCS often cannot capture complex multivariate relationships in the data (Murray and Reiter, 2016), which we confirm in Section 6.

The second main approach constructs a joint distribution for all variables in the data set and then imputes missing values from the (posterior) predictive distribution. Bayesian nonparametric models are particularly attractive, including for imputation of multiple categorical (Dunson and Xing, 2009; Manrique-Vallier and Reiter, 2014, 2017), ordinal (Kottas et al., 2005; DeYoreo et al., 2017), or categorical and continuous variables (Murray and Reiter, 2016; Roy et al., 2018). Related, Taddy and Kottas (2010) proposed a multivariate mixture model with kernel components specific to each data type, but did not consider imputation. These existing approaches have several limitations. First, they do not simultaneously accommodate categorical, continuous, count, and ordinal variables. Second, they often require careful model specification for each variable, which is arduous in moderate to high dimensions. Finally, the accompanying MCMC samplers are typically complex and

computationally intensive. To our knowledge, there is no publicly available software for imputation based on these methods, which limits their practical utility.

Copula models offer a potential avenue to estimate a joint model on mixed data types: they combine arbitrary marginal distributions with a mechanism to model joint dependencies (Joe, 2014). Zhao and Udell (2020a,b) deployed frequentist Gaussian copula models for imputation of MCAR data with continuous and ordinal variables, but did not consider MAR missingness or other data types. Pitt et al. (2006) specified parametric families for count and continuous variables within a Bayesian Gaussian copula model which could be extended for imputation. 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) and deployed for imputation in Cui et al. (2019). The RL uses 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. 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 that facilitates high-dimensional imputation.

Despite these advantages, semiparametric Bayesian copula models have two glaring shortcomings in the presence of missing data. First, these models do not estimate the marginal distributions and thus do not provide a data generating process for prediction or imputation. Instead, the default approach is to fix each margin at its ECDF and generate posterior predictive variates by repeatedly (i) drawing a latent Gaussian variable under the model and (ii) applying the inverse ECDF. However, the ECDF is significantly flawed under MAR (see Figure 1), so 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. We demonstrate this limitation in Section 6.2, and conclude that clearly, these models cannot be relied upon for prediction or imputation with MAR data.

Second, Gaussian copula models only specify linear associations on the latent scale. As such, they cannot capture complex and nonlinear dependencies and interactions, which we demonstrate empirically in Section 6.1. Gaussian mixture copulas (Tewari et al., 2011; Rajan and Bhattacharya, 2016) offer some additional distributional flexibility, but 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 modeling and imputation 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. A primary innovation in this work is the introduction and theoretical justification for the *margin adjustment*, which eliminates the reliance on the ECDF in the posterior predictive distribution of rank-based copula models. The margin adjustment features several key properties:

1. It requires no specification of any marginal models, no additional assumptions, and no additional parameters;

2. It delivers posterior consistency and posterior uncertainty quantification for each marginal distribution, even under MAR;

3. It is computationally scalable and empirically accurate for estimation and imputation.

The importance of these features is highlighted using both simulated and real data, which decisively show that the proposed imputation strategy offers significant improvements over competing methods, especially in the presence of data MAR and nonlinear dependencies.

This paper is organized as follows. Section 2 introduces Bayesian copula models for mixed data types. In Section 3, we define and study the margin adjustment. Section 4 describes our novel Gaussian mixture copula, with extensions in Section 5 for unordered categorical variables. We apply our proposed approach in Section 6 with two simulation studies and a real data example in Section 7. We conclude in Section 8. Supplementary material includes proofs of all results, details on the computations, additional simulation results, and an `R`[1] package that implements the proposed approach.

2. Bayesian Copula Models for Mixed Data Types

2.1 The Gaussian Copula

Our first objective is to develop a Bayesian model for multivariate and mixed data. This model will be used to generate posterior predictive draws of the missing data, thereby allowing estimation and uncertainty quantification of arbitrary Q through imputation. Consider the Gaussian copula, which models the p-dimensional vector $\boldsymbol{y} = (y_1, \ldots, y_p)$ using

$$\boldsymbol{z} \sim N_p(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\theta}}), \quad \boldsymbol{z} = (z_1, \ldots, z_p)^T \tag{2}$$

$$y_j = F_j^{-1}\{\Phi(z_j)\}, \quad j = 1, \ldots, p. \tag{3}$$

The Gaussian copula links the univariate marginal distributions $\{F_j\}_{j=1}^p$ for each component of $\boldsymbol{y}$ with a multivariate model for latent Gaussian data $\boldsymbol{z}$ governed by correlation matrix $\boldsymbol{C}_{\boldsymbol{\theta}}$, which is indexed by parameters $\boldsymbol{\theta}$. Thus, each F_j describes the marginal features of y_j while $\boldsymbol{C}_{\boldsymbol{\theta}}$ encodes the dependencies among $\boldsymbol{y}$. Model (2)-(3) implies the joint CDF for $\boldsymbol{y}$ is $F(y_1, \ldots, y_p) = \Phi_p[\Phi^{-1}\{F_1(y_1)\}, \ldots \Phi^{-1}\{F_p(y_p)\}]$, where Φ_p is the CDF of a p-dimensional Gaussian random vector with mean zero and correlation matrix $\boldsymbol{C}_{\boldsymbol{\theta}}$ and Φ is the univariate standard normal CDF.

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

1. An `R` package implementing the proposed approach is available on the author's GitHub page, found at `https://github.com/jfeldman396/GMCImpute`

The Gaussian copula model has several critical limitations. First, the margins F_j must be specified either parametrically (Pitt et al., 2006), which is restrictive and time-consuming when p is moderate or large, or nonparametrically, which significantly increases the computational burdens. Alternatively, fixing margins at their empirical estimates (Hoff, 2007; Murray et al., 2013; Feldman and Kowal, 2022) may introduce significant bias into the posterior predictive distribution in the presence of missing data (Figure 1) and fails to account for the uncertainty about these parameters. Second, the correlation matrix $\boldsymbol{C}_{\boldsymbol{\theta}}$ captures only latent linear dependencies, and thus may not be suitable for more complex relationships. Lastly, the link (3) is not well-defined for unordered categorical variables (Feldman and Kowal, 2022), and estimation of the Gaussian copula with discrete Y_j is problematic (Hoff, 2007). Thus, modifications are needed to provide valid joint models for mixed data types, with particular focus on modeling flexibility and computational scalability.

2.2 Semiparametric Copula Models for Mixed Continuous, Count, and Ordinal Data

One approach that bypasses the need to specify individual marginal models for continuous, count, and ordinal variables is the extended rank-likelihood (RL; Hoff (2007)). Let $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n$ with $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^T$ contain p numeric (continuous, count, or ordinal) variables; modifications for unordered categorical variables are discussed in Section 5. The non-decreasing link in (3) implies a partial ordering on the latent scale for (2): 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 variable in the data set when the event $\mathcal{D}(\boldsymbol{Y}) := \{\boldsymbol{Z} \in \mathbb{R}^{n\times p} : \max\{z_{kj} : y_{kj} < y_{ij}\} < z_{ij} < \min\{z_{kj} : y_{ij} < y_{kj}\}\}$ occurs.

The RL is derived by first expressing the Gaussian copula likelihood in terms of $\mathcal{D}(\boldsymbol{Y})$:

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \{F_j\}_{j=1}^p) = p\{\boldsymbol{Y}, \boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}) \mid \boldsymbol{\theta}, \{F_j\}_{j=1}^p\} \tag{4}$$

$$= p\{\boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}) \mid \boldsymbol{\theta}\}\; p\{\boldsymbol{Y} \mid \boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}), \boldsymbol{\theta}, \{F_j\}_{j=1}^p\}. \tag{5}$$

The decomposition (4)-(5) is made possible since the event $\boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y})$ does not depend on the marginal distributions $\{F_j\}_{j=1}^p$ and must occur with observation of $\boldsymbol{Y}$. Hoff (2007) argued that the left term in (5) should contain most of the information about $\boldsymbol{\theta}$, and Murray et al. (2013) showed that it is indeed sufficient for posterior consistency for $\boldsymbol{\theta}$. Thus, the RL enables semiparametric inference on $\boldsymbol{\theta}$ (and $\boldsymbol{C}_{\boldsymbol{\theta}}$) by targeting the posterior

$$p\{\boldsymbol{\theta} \mid \boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y})\} \propto p\{\boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}) \mid \boldsymbol{\theta}\}\; p(\boldsymbol{\theta}). \tag{6}$$

Notably, (6) does not include the marginal CDFs $\{F_j\}$. When priority lies in inference for $\boldsymbol{C}_{\boldsymbol{\theta}}$, which contains substantive information on multivariate relationships in the data, this artifact is particularly convenient for model specification and computation (see Algorithm 1). However, consideration of $\{F_j\}$ is necessary for the posterior predictive distribution—and thus for missing data imputation. We address this challenge in Section 3.

2.3 Estimating Semiparametric Copulas with Missing Data

With missing data, we only have access to the observed ranks. Thus, we modify the RL posterior appropriately:

$$p\{\boldsymbol{\theta} \mid \boldsymbol{Z}^{obs} \in \mathcal{D}(\boldsymbol{Y}^{obs})\} \propto \int p\{\boldsymbol{Z}^{obs} \in \mathcal{D}(\boldsymbol{Y}^{obs}), \boldsymbol{Z}^{mis} \mid \boldsymbol{\theta}\} d\boldsymbol{Z}^{mis}\ p(\boldsymbol{\theta}) \tag{7}$$

where $\boldsymbol{Z}^{obs} = (Z_{ij} : R_{ij} = 0)$ and $\boldsymbol{Z}^{mis} = (Z_{ij} : R_{ij} = 1)$. Though (7) non-standard, it is relatively simple to construct a Gibbs sampling algorithm to sample from this distribution. The sampler (Algorithm 1) alternates between drawing from $[(\boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}) \mid \boldsymbol{Y}^{obs}, \boldsymbol{\theta}]$ and $[\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs}, \boldsymbol{Z}]$. The first step features univariate truncated normal draws for each Z_{ij}^{obs} based on (2) and the RL, while prediction of Z_{ij}^{mis} is unrestricted by any observed ordering constraints. Next, $[\boldsymbol{\theta} \mid \boldsymbol{Y}^{obs}, \boldsymbol{Z}] = [\boldsymbol{\theta} \mid \boldsymbol{Z}]$ is drawn from a posterior which features a multivariate Gaussian likelihood, and thus sampling is straightforward for many choices of priors $p(\boldsymbol{\theta})$.

Algorithm 1 Bayesian RL Gaussian Copula Gibbs Sampler with Missing Data

Require: prior $p(\boldsymbol{\theta})$

- **Step 1**: Sample $(\boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}) \mid \boldsymbol{\theta}$

  for $Z_{ij} \in \boldsymbol{Z}^{obs}$ **do**
  Compute $z_\ell = \max\{z_{kj}^{obs} : y_{kj}^{obs} < y_{ij}^{obs}\}$ and $z_u = \min\{z_{kj}^{obs} : y_{ij}^{obs} < y_{kj}^{obs}\}, k \neq i$
  Sample $Z_{ij} \sim \text{Normal}(\mu_{ij}, \sigma_j^2)\mathbb{1}(z_\ell, z_u)$
  for $Z_{ij} \in \boldsymbol{Z}^{mis}$ **do**
  Sample $Z_{ij} \sim \text{Normal}(\mu_{ij}, \sigma_j^2)$
  where $\mu_{ij} = (\boldsymbol{C}_{\boldsymbol{\theta}})_{j-j}(\boldsymbol{C}_{\boldsymbol{\theta}}^{-1})_{-j-j}\boldsymbol{Z}_{i-j}$, $\sigma_j^2 = (\boldsymbol{C}_{\boldsymbol{\theta}})_{jj} - (\boldsymbol{C}_{\boldsymbol{\theta}})_{j-j}(\boldsymbol{C}_{\boldsymbol{\theta}}^{-1})_{-j-j}(\boldsymbol{C}_{\boldsymbol{\theta}})_{-jj}$

- **Step 2**: Sample $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}, \boldsymbol{Y}^{obs}) = p(\boldsymbol{\theta} \mid \boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis})$

  where $p(\boldsymbol{\theta} \mid \boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}) \propto N_p\{(\boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}); \boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\theta}}\}p(\boldsymbol{\theta})$

A remarkable feature of Algorithm 1 is the absence of any marginals $\{F_j\}$. While this is advantageous for posterior inference on $\boldsymbol{\theta}$ — it removes the need to specify or estimate any marginal distributions — the margins are in fact necessary for prediction and imputation as predictive samples of $\boldsymbol{Z}^{mis}$ need be transformed to $\boldsymbol{Y}^{mis}$. The default semiparametric procedure is to fix each $\{F_j\}$ at the ECDF (Hoff, 2007; Murray et al., 2013; Cui et al., 2019; Feldman and Kowal, 2022), and the accompanying imputation step would apply Algorithm 1 and compute $\hat{F}_j^{-1}\{\Phi(z_{ij}^{mis})\}$. However, the ECDF does not account for the uncertainty about each F_j. More critically, the ECDF is at risk for significant bias under MAR (see Figure 1 and Sections 6-7), which will lead to inaccurate predictions and imputations — even if $\boldsymbol{\theta}$ is inferred correctly.

3. The Margin Adjustment

To eliminate reliance on the ECDFs for posterior predictive sampling and imputation—while still maintaining the beneficial structure of the Bayesian RL copula model—we propose

a new strategy called the *margin adjustment*. The margin adjustment does not require any additional modeling assumptions or parameters, is fully automated (i.e., individual specification of marginal models for each F_j is not needed), and provides computationally efficient and consistent posterior inference for the margins $\{F_j\}$, even in the presence of data MAR. The margin adjustment does not impact posterior inference for $\boldsymbol{\theta}$, and thus Algorithm 1 is unchanged.

3.1 Derivation and Theory

The key insight of the margin adjustment is that the combination of the RL rank constraints (4)-(5) and the latent data model (2) are sufficient to infer the marginal distributions $\{F_j\}_{j=1}^p$ with strong theoretical guarantees. Under the RL, Z_j is a non-decreasing (and unknown) transformation of Y_j for each j. Thus, upon ordering both $\{Z_{ij}\}_{i=1}^n$ and $\{Y_{ij}\}_{i=1}^n$, the position of $\max\{Z_{ij} : Y_{ij} \leq x\}$ among $\{Z_{ij}\}_{i=1}^n$ will be identical to the position of $\max\{Y_{ij} : Y_{ij} \leq x\}$ among $\{Y_{ij}\}_{i=1}^n$ for any x greater than the minimum of $\{Y_{ij}\}_{i=1}^n$. For any x below this value, the set $\{Y_{ij} \leq x\}$ will be empty with probability 1. Thus, we define

$$Z_j^n(x) = \max[\{Z_{ij} : Y_{ij} \leq x\} \cup \{Z_{ij} : Y_{ij} = \min(\{Y_{ij}\}_{i=1}^n)\}, i \in \{1, \ldots, n\}]. \tag{8}$$

Informally, if $F_j(x) = \tau$, then $Z_j^n(x)$ will approximate the τth quantile under the marginal latent data model. This motivates the following marginal distribution estimator:

$$\tilde{F}_j(x) = G_j\{Z_j^n(x)\}, \tag{9}$$

where G_j is the marginal distribution for Z_j induced by the latent data model under the copula. Importantly, the margin adjustment is compatible with *any* rank-based copula model. All that is required is the multivariate model for $\boldsymbol{Z}$, which induces marginals G_j. Under the Gaussian copula (2)-(3), G_j is the standard normal CDF Φ; modifications for the Gaussian mixture copula are available in Section 4.

More formally, Theorem 1 provides a general setting in which a continuous random variable Z may be used to infer the distribution function of $Y = h(Z)$ with almost sure convergence, where h is any monotone increasing function. Our primary example is the RL posterior, where h ensures the necessary ordering across realizations of (Z, Y).

Theorem 1 *Suppose* $\{Z_i\}_{i=1}^n \overset{i.i.d}{\sim} F_Z$ *and* $\{Y_i\}_{i=1}^n = \{h(Z_i)\}_{i=1}^n \sim F_Y$ *, where* F_Z *is continuous and* h *is a monotone increasing function. Defining* $Z^n(x)$ *as* (8)*, the margin adjustment satisfies* $\tilde{F}(x) \coloneqq F_Z\{Z^n(x)\} \overset{a.s.}{\to} F_Y(x)$ *for all* $x \in \mathbb{R}$.

The more challenging setting occurs when data are MAR. In the presence of missing data, we modify (8) appropriately:

$$Z_j^n(x) = \max[\{Z_{ij}^{obs} : Y_i^{obs} \leq x\} \cup \{Z_{ij}^{obs} : Y_{ij}^{obs} = \min(Y_{ij}^{obs})\}, i \in \{1, \ldots, n\}]. \tag{10}$$

Crucially, with this modification, the margin adjustment remains consistent under MAR. For simplicity, we demonstrate our result for $p = 2$ variables, one of which is MAR.

Theorem 2 *Suppose* $\{\boldsymbol{Z}_i\}_{i=1}^n = \{(Z_{i1}, Z_{i2})\}_{i=1}^n \overset{i.i.d}{\sim} G$*, where* G *is continuous with marginal distributions* G_1, G_2*, and* $\{\boldsymbol{Y}_i\}_{i=1}^n = \{(Y_{i1}, Y_{i2})\}_{i=1}^n = [(F_1^{-1}\{G_1(Z_{i1})\}, F_2^{-1}\{G_2(Z_{i2})\})]_{i=1}^n$

has joint distribution function F with marginal distributions F_1, F_2. Suppose Y_2 is completely observed and Y_1 is MAR. The margin adjustment satisfies $\tilde{F}_1(x) := G_1\{Z_1^n(x)\} \overset{a.s.}{\to} F_1(x)$ for all $x \in \mathbb{R}$.

Generalizations beyond the bivariate case are straightforward. 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. Notably, the consistency in Theorem 2 is *not* valid for the ECDF under MAR (see Figure 1), which undermines any methods that rely on the ECDF for prediction and imputation (Hoff, 2007; Murray et al., 2013; Cui et al., 2019; Feldman and Kowal, 2022).

3.2 Bayesian Estimation and Imputation

The margin adjustment (9) is a function of the latent data $\boldsymbol{Z}$, and thus inherits a posterior distribution under the Bayesian RL copula model. Under Algorithm 1, $\boldsymbol{Z}^{obs}$ is sampled from its joint posterior, and so the margin adjustment may be integrated in any Bayesian RL copula model to provide margin estimation and uncertainty quantification. For clarity, we outline the procedure to obtain posterior samples of the margin adjustment for each $j \in \{1, \ldots, p\}$ under the RL Gaussian copula in Algorithm 2; modifications for the Gaussian mixture copula are available in Section 4.

Algorithm 2 The Margin Adjustment Sampling under the Bayesian RL Gaussian Copula

Require: One posterior sample of $\boldsymbol{Z}^{obs}$ from Algorithm 1
Return: One posterior sample of $\tilde{F}_j(x)$, $j \in \{1, \ldots, p\}$
for $j \in \{1, \ldots, p\}$ and any x **do**
  Compute $Z_j^n(x)$ as (10)
  Compute $\tilde{F}_j(x) = \Phi\{Z_j^n(x)\}$

We provide a visualization of Algorithm 2 for a single variable Y_j in Figure 3. Notably, posterior sampling for the margin adjustment is seamlessly incorporated into Algorithm 1 with minimal computational expense. The procedure is augmented with an efficient post-processing of posterior samples of $\boldsymbol{Z}^{obs}$, which in turn provides posterior inference for each marginal.

In practice, we compute the margin adjustment for each unique $x \in \{Y_{ij}^{obs}\}_{i=1}^n$. Since the resulting $\tilde{F}_j$ is a step function with jumps determined by these observed values, we then fit a monotone interpolating spline to $\{x, \tilde{F}_j(x)\}$ with pre-specified upper and lower bounds. These bounds may be available through domain knowledge (e.g., age cannot be negative and is typically less than or equal to 110) or they can be fixed with reasonable heuristics. For instance, when known bounds are not available, one solution is to add/subtract a constant from the observed upper/lower bounds of each variable. The interpolating spline preserves $\tilde{F}_j$ at the observed data values but expands the support of the data-generating process beyond only those observed values, which is important for imputation.

Most important, we apply the margin adjustment to deliver model-based imputation. We demonstrate this using the RL Gaussian copula in Algorithm 3, but once again this algorithm may be generalized to any RL copula model. Thus, the margin adjustment

Algorithm 3 Bayesian RL Gaussian Copula Imputation using the Margin Adjustment

Require: One posterior sample of $\boldsymbol{Z}^{mis}$ from Algorithm 1 and $\{\tilde{F}_j\}_{j=1}^p$ from Algorithm 2
Return: One completed data set $\boldsymbol{Y} = (\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis})$
for $z_{ij}^{mis} \in \boldsymbol{Z}^{mis}$ **do**
  Set $y_{ij}^{mis} = \tilde{F}_j^{-1}\{\Phi(z_{ij}^{mis})\}$

replaces the default ECDF for imputation with an estimator that remains consistent in the presence of MAR. This benefit is explored empirically in Sections 6-7, and yields substantial improvements in prediction and imputation inference.

3.3 Strong Posterior Consistency under MAR

We now establish the asymptotic properties of the posterior distribution of the Gaussian copula correlation parameter $\boldsymbol{C}_{\boldsymbol{\theta}}$ under the RL with ignorable missing data, and demonstrate how this posterior consistency extends to the margin adjustment (9). First, we adapt the result in Murray et al. (2013), which established posterior consistency of $\boldsymbol{C}_{\boldsymbol{\theta}}$ under the RL without missingness for mixed continuous, count, and ordinal data types. However, their proof relied upon the almost sure convergence of the ECDF $\hat{F}_j$ to F_j, which is not maintained under MAR.

Theorem 3 *Suppose* $\{\boldsymbol{Y}_i\}_{i=1}^n \overset{i.i.d}{\sim} G^{\infty}_{\boldsymbol{C}_0,F_1,\ldots,F_p}$, *where* $G^{\infty}_{\boldsymbol{C}_0,F_1,\ldots,F_p}$ *is the Gaussian copula for the joint distribution of p-dimensional* $\boldsymbol{Y}$ *with true copula parameters* $\boldsymbol{C}_0$ *and true marginal CDFs* $F_1,\ldots,F_p$. *Let* Π *be a prior distribution on the space of all* $p \times p$ *positive semi-definite correlation matrices* $\boldsymbol{C}_{\boldsymbol{\theta}}$ *with corresponding density* $\pi(\boldsymbol{C}_{\boldsymbol{\theta}})$ *with respect to a measure* ν. *Suppose* $\pi(\boldsymbol{C}_{\boldsymbol{\theta}}) > 0$ *almost everywhere with respect to* ν *and assume that the missingness is ignorable. Then, for* $\boldsymbol{C}_0$ *a.e.* $[\nu]$ *and any neighborhood* $\mathcal{A}$ *of* $\boldsymbol{C}_0$, *we have that* $\lim_{n\to\infty} \Pi\{\boldsymbol{C}_{\boldsymbol{\theta}} \in \mathcal{A} \mid \boldsymbol{Z}_n^{obs} \in \mathcal{D}(\boldsymbol{Y}_n^{obs})\} = 1$ *a.s.* $[G^{\infty}_{\boldsymbol{C}_0,F_1,\ldots,F_p}]$.

In conjunction with Theorems 1–3, the strong posterior consistency of $\boldsymbol{C}_{\boldsymbol{\theta}}$ also yields posterior consistency for the margin adjustment.

Corollary 4 *Under the conditions of Theorem 3, define* $\tilde{F}_j$ *as in* (9) *with* $Z_j^n(x)$ *as* (10) *and* $G_j = \Phi$ *for each* $j \in \{1,\ldots,p\}$. *Then for any* $x \in \mathbb{R}$ *and any neighborhood* $\mathcal{A}$ *of* $F_j(x)$ $\lim_{n\to\infty} \Pi\{\tilde{F}_j(x) \in \mathcal{A} \mid \boldsymbol{Z}_n^{obs} \in \mathcal{D}(\boldsymbol{Y}_n^{obs})\} = 1$ *a.s* $[G^{\infty}_{\boldsymbol{C}_0,F_1,\ldots,F_p}]$.

These results are powerful: the RL 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 continuous, count, and ordinal variables with MAR data.

4. Gaussian Mixture Copulas via Latent Factors

Although we have established theoretical guarantees for the RL and the margin adjustment under a Gaussian copula model—including for mixed (count, continuous, ordinal) variables and missing data—the Gaussian copula only captures linear associations on the

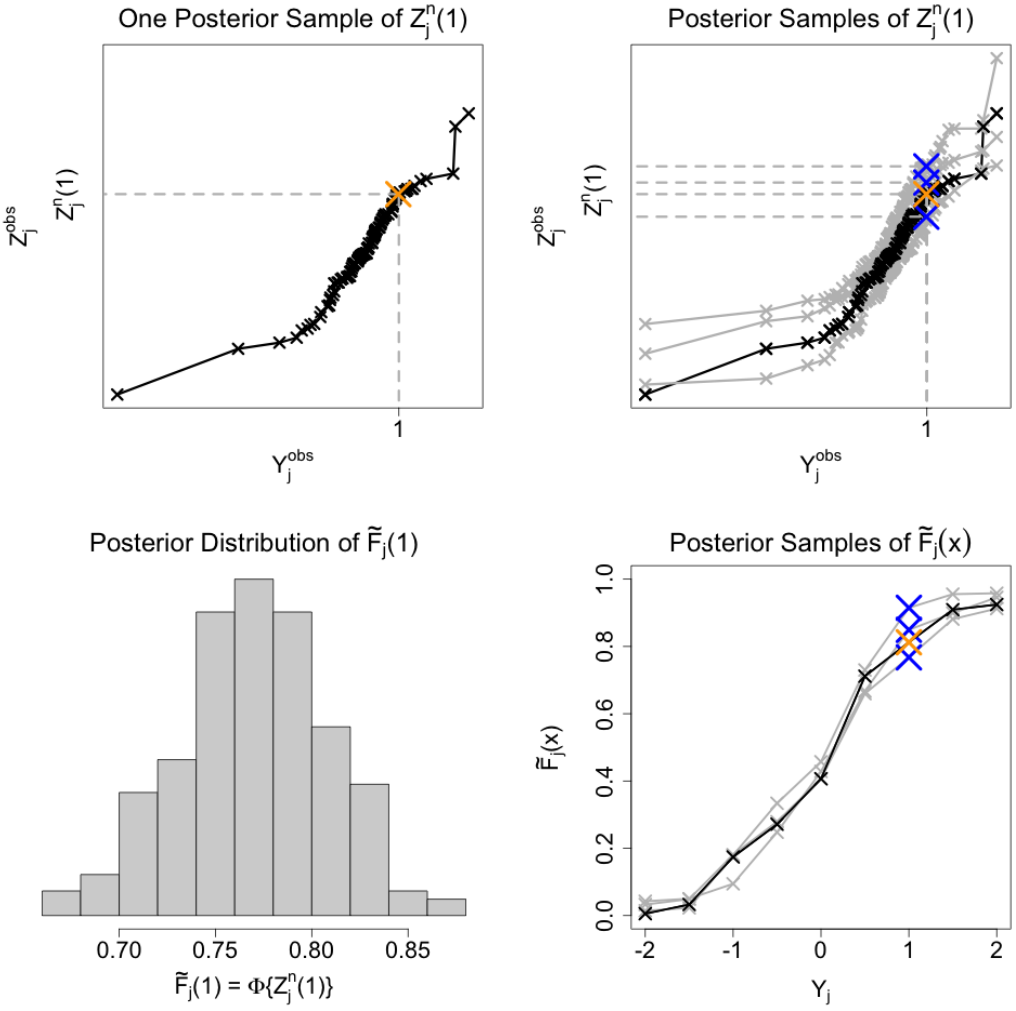

Figure 3: Visualizing the margin adjustment for a single variable Y_j under the RL Gaussian copula. In the top-left panel, we show the monotone relationship between Y_j^{obs} and Z_j^{obs} for one realization of $\boldsymbol{Z}^{obs}$ from Algorithm 1. We highlight $Z_j^n(1)$ (orange cross) which is computed as (10) with $x = 1$. In the top-right panel, we plot additional posterior realizations of Z_j^{obs} (gray curves), which yield multiple posterior samples of $Z_j^n(1)$ (blue crosses). The bottom-left panel plots the posterior distribution of the margin adjustment (9) applied to posterior samples of $Z_j^n(1)$ accumulated across iterations of Algorithm 1. In the bottom-right panel, we show the margin adjustment along an evenly spaced grid of $x \in [-2, 2]$ for each realization of Z_j^{obs} directly above, which provides posterior inference for F_j.

latent scale through latent correlation matrix $\boldsymbol{C}_{\boldsymbol{\theta}}$. As such, it may not be sufficiently powerful to capture nonlinearities and interactions on the observed scale (see Section 6), which is vital to justify MAR and for imputation under complex dependencies. However, generalizations of the Gaussian copula must carefully consider computational scalability, model parsimony, and suitable adaptations of the margin adjustment.

To build an imputation model capable of adapting to unanticipated features in data, we develop a novel Gaussian mixture copula (GMC). When paired with the RL and margin adjustment, the resulting copula model is fully nonparametric, and may be used to impute variables of arbitrary type. The GMC extends the Gaussian copula by replacing the latent data model (2) with a finite mixture:

$$\boldsymbol{z} \sim \sum_{h=1}^{H} \pi_h N_p(\boldsymbol{\alpha}_h, \boldsymbol{C}_h) \tag{11}$$

where the marginal distribution of the jth component is $z_j \sim \sum_{h=1}^{H} \pi_h N(\{\boldsymbol{\alpha}_h\}_j, \{\boldsymbol{C}_h\}_{jj})$. The multivariate mixture on the latent data can be combined with the observed data marginals to define the GMC. For completeness, we show that is indeed a valid copula.

Theorem 5 *Let* $\mathbb{C}_{GMC}(\boldsymbol{u}) = \Psi(\psi_1^{-1}\{F_1(y_1)\}, \ldots, \psi_p^{-1}\{F_p(y_p)\})$, *where* $\Psi = \sum_{h=1}^{H} \pi_h \Phi_p(\boldsymbol{\alpha}_h, \boldsymbol{C}_h)$, $\psi_j = \sum_{h=1}^{H} \pi_h \Phi(\{\boldsymbol{\alpha}_h\}_j, \{\boldsymbol{C}_h\}_{jj})$, *and* $\{F_j\}_{j=1}^{p}$ *are the marginals of* $\{Y_j\}_{j=1}^{p}$. *Then* $\mathbb{C}_{GMC}$ *defines a valid copula.*

The data generating representation of $\mathbb{C}_{GMC}$ simulates $\boldsymbol{z}$ from (11) and links the realization to the observed scale via $y_j = F_j^{-1}\{\psi_j(z_j)\}$.

Although the GMC latent data model (11) provides greater representational ability than the Gaussian copula (2), especially for nonlinearities and interactions, the GMC modeling and computational capabilities are limited in higher dimensions. In particular, the GMC is parameterized by $\boldsymbol{\theta} = \{\pi_h, \boldsymbol{\alpha}_h, \boldsymbol{C}_h\}_{h=1}^{H}$, 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 $\boldsymbol{\eta} \in \mathbb{R}^k$ with $k \ll p$:

$$\boldsymbol{\eta}_i \sim \sum_{h=1}^{H} \pi_h N_k(\boldsymbol{\mu}_h, \boldsymbol{\Delta}_h), \quad \boldsymbol{z}_i \mid \boldsymbol{\eta}_i \sim N_p(\boldsymbol{\Lambda}\boldsymbol{\eta}_i, \boldsymbol{\Sigma}) \tag{12}$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, $\boldsymbol{\Lambda}$ is a $p \times k$ dimensional matrix of factor loadings, and $\boldsymbol{\eta}_i$ is a k-dimensional vector of latent factors. The latent factor mixture model (12) induces a mixture model (11) for $\boldsymbol{Z}$ through marginalization over $\boldsymbol{\eta}$, and specifically with $\boldsymbol{\alpha}_h = \boldsymbol{\Lambda}\boldsymbol{\mu}_h$ and $\boldsymbol{C}_h = \boldsymbol{\Lambda}\boldsymbol{\Delta}_h\boldsymbol{\Lambda}^T + \boldsymbol{\Sigma}$ for $h = 1, \ldots, H$. Thus, the latent data $\boldsymbol{Z}$ are still endowed with a flexible mixture model, but the clustering is directed to a lower-dimensional space. This feature offers important benefits relative to existing GMCs (Tewari et al., 2011; Rajan and Bhattacharya, 2016), namely, that it affords parsimonious estimation of the dependence structure among moderate to high-dimensional data. Chandra et al. (2023) recently applied this strategy for clustering of 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 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 $\boldsymbol{\eta}$. 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 use a stick-breaking process for the mixing weights $\{\pi_h\}$ (Ishwaran and James, 2001): $\pi_h = V_h \prod_{l<h}(1 - V_l)$

with $V_\ell \overset{i.i.d}{\sim} \text{Beta}(1, \alpha_\pi)$ and $\alpha_\pi \sim \text{Gamma}(a_\alpha, b_\alpha)$. For computational convenience, we implement a truncated DP (Ishwaran and James, 2002) that resembles (12), where now H is a conservative upper bound. The component-wise mean and covariance are assigned a Normal-Inverse Wishart prior, $(\boldsymbol{\mu}_h, \boldsymbol{\Delta}_h) \sim \text{NIW}(\boldsymbol{\mu}_0, \delta^2 \boldsymbol{I}_k, \kappa_0, \nu_0)$, while the diagonal elements of $\boldsymbol{\Sigma}$ are assigned $\sigma_j^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$. Lastly, we apply a global-local shrinkage prior for the loadings matrix $\boldsymbol{\Lambda} = \{\lambda_{jt}\}$ that encourages columnwise shrinkage for rank selection, which reduces sensitivity to the choice of k (Bhattacharya and Dunson, 2011).

Our approach for Bayesian inference, prediction and imputation combines the GMC with the RL and margin adjustment (GMC-MA). Despite the complexity of the GMC-MA, only minor modifications are needed for Algorithms 1-3. A particular convenience comes from (12): conditional on $\boldsymbol{\theta}$ under the mixture of factor models, $Z_{ij} \sim N(\sum_{t=1}^k \lambda_{jt}\eta_{it}, \sigma_j^2)$, which implies the components of $\boldsymbol{Z}_i$ are independent univariate Gaussian. Therefore, even in the presence of missing data, the core elements of Algorithm 1 are the same: each Z_{ij}^{obs} is sampled from a truncated Gaussian, the posterior of Z_{ij}^{mis} is once again unrestricted, and sampling of $\boldsymbol{\theta}$ involves standard steps for Gaussian factor and mixture models; details are provided in the supplementary material. For the margin adjustment, we simply modify Algorithm 2 to use $G_j = \psi_j$. Finally, imputations are similarly generated by modifying Algorithm 3 to use ψ_j in place of Φ.

5. Extensions for Unordered Categorical Variables

We incorporate unordered categorical variables via the extended rank-probit likelihood (RPL) (Feldman and Kowal, 2022), which generalizes the RL. Suppose $\boldsymbol{Y} = \boldsymbol{Y}^r \cup \boldsymbol{Y}^q$, where r indexes the numeric variables, q indexes the unordered categorical variables and $p = r + q$. For each categorical variable $\boldsymbol{Y}_c$ with k_c levels, when $y_{ic} = m$, the RPL encodes a vector of k_c binary variables $\boldsymbol{\gamma}_c$ with the corresponding latent data restriction $\{\gamma_{ic_m} = 1 \cap \gamma_{ic_l} = 0, l \neq m\} \implies \{z_{ic_m} > 0 \cap z_{ic_l} < 0, l \neq m\}$, i.e., $y_{ic} = m$ implies that only the mth component is positive and the others are all negative. This representation avoids the need to select reference groups. Aggregating this representation across all q unordered categorical variables, the observed categorical memberships must satisfy the event $\mathcal{D}'(\boldsymbol{Y}^q) := \cup_{c=1}^q \{\boldsymbol{Z}^{n\times k_c} : \gamma_{ij} = 1 \implies z_{ij} > 0 \cap \{z_{i\ell} < 0\}_{\ell \neq j}\}$. This representation is also recommended for ordinal variables with few levels (Feldman and Kowal, 2022).

The RPL joins the rank event on the r numeric variables with the probit-style representation of the q categorical variables to define $\mathcal{E}(\boldsymbol{Y}) = \mathcal{D}(\boldsymbol{Y}^r) \cup \mathcal{D}'(\boldsymbol{Y}^q)$, which substitutes for $\mathcal{D}(\boldsymbol{Y})$ in (4)-(5) for joint modeling of continuous, count, ordinal and now unordered categorical variables. Imputation for unordered categorical variables is carried out by estimating the multinomial probabilities of each level. These quantities are easily computed with posterior samples of $\boldsymbol{\theta}$ and the latent data model, as the probability that an observation assumes a particular level is given by the probability that the corresponding latent component is positive and all others are negative; the supplement provides the Gibbs sampling and imputation algorithm for the GMC-MA under the RPL.

6. Simulation Studies

6.1 Mixed Data, Nonlinearity, and MAR

In the first simulation study, we evaluate (i) the impact of 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. The latter objective aims to highlight the benefit of the latent mixture model (11) over the single component Gaussian copula ($H = 1$).

We generate mixed data sets with nonlinear dependencies by simulating

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

for $n \in \{500, 1000, 2000\}$, where $y_{2_{scale}}$ is the centered and scaled version of y_2. Next, we introduce missingness using the MAR mechanism

$$R_j \mid Y_1 = y_1 \sim \text{Bernoulli}\{\Phi(-0.5 + \beta|y_1|)\}$$

for $j = 2, 3$, which links the missingness in both Y_2 and Y_3 with the observed value of Y_1. The parameter β 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.

Figure 4: Simulated data sets without missingness (left column) and the complete cases after applying the MAR mechanism (left-middle) with $n = 1000, \beta = 0.5$. The proposed approach (right-middle) is significantly better than the Gaussian copula (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.

We highlight the challenging nonlinearities and missingness under this data-generating mechanism ($n = 1000, \beta = 0.5$) with example simulated data in Figure 4. Compared to the full data set, 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 data set from the proposed GMC-MA and compare it to Hoff (2007) using the `sbgcop` package in `R`, which uses a single component Gaussian copula with the ECDF for posterior predictive simulations (similar results are obtained for additional realizations and simulation settings in the supplement).

Across the simulation settings, the GMC-MA tends to discover two clusters in the data. Clearly, this leads to significant improvements in detecting non-linearity: it is capable of capturing the complex relationship between Y_1 and Y_2 and correctly imputing additional $Y_3 = 1$ values when $|Y_1|$ is large. We emphasize that GMC-MA does not leverage any aspect of the true data-generating process beyond the variable data type.

Next, we evaluate the margin adjustment, and specifically seek to assess whether it corrects the biases of the ECDF in the presence of data MAR. In Figure 5, we focus on the marginal distribution for Y_2, which is a count variable subject to MAR. 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) of $\tilde{F}_2$ under the GMC-MA. Posterior inference uses the estimators described in Section 4 and the Gibbs sampler from the supplement, which we run for 5,000 iterations. Trace plots of the draws of the marginal distribution function $\tilde{F}_2$ indicate that the MCMC algorithm converges after about 1,500 samples, which we discard as a burn-in. We display the results for both missingness settings ($\beta = 0.5, 1$).

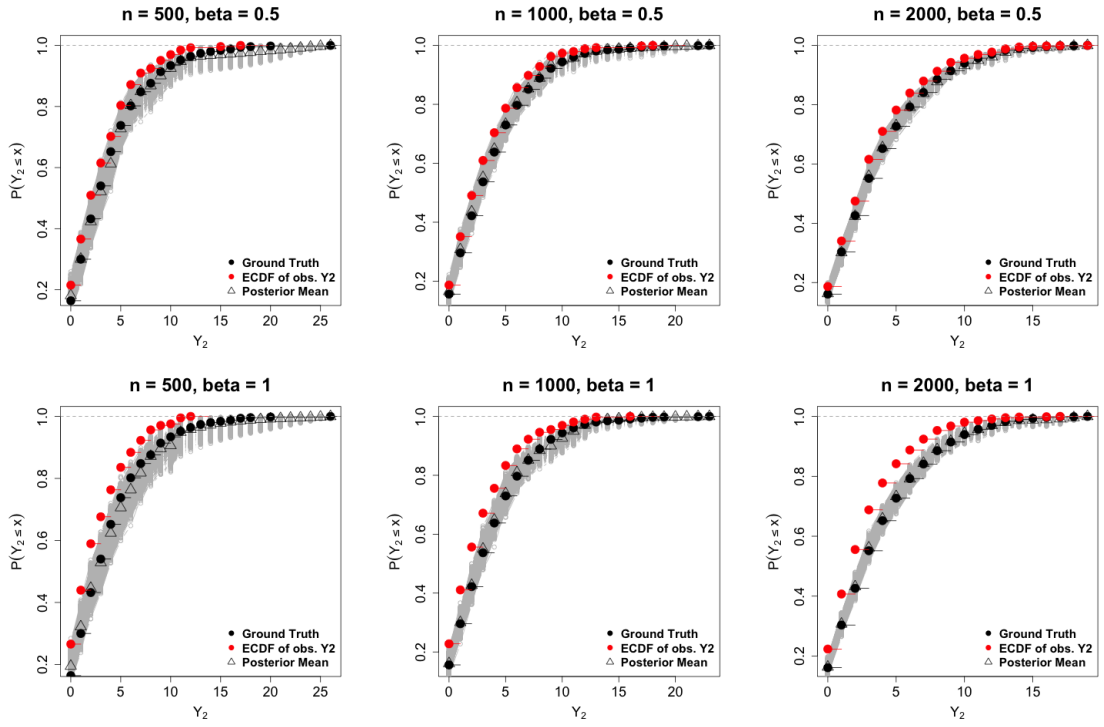

Figure 5: Estimation and inference for the marginal distribution of Y_2 using the margin adjustment under MAR with varying n. The ECDF of Y_2^{obs} (red points) deviates significantly from the ECDF of Y_2 prior to removing missingness (black points). The posterior draws (gray lines) and posterior mean (triangles) from the margin adjustment show that the proposed approach is highly accurate, even under severe MAR which is present with $\beta = 1$ (bottom row).

Most notably, the ECDF on the observed data is badly biased, and this flawed estimator would be used for imputation under default semiparametric (rank-based) copula models (Hoff, 2007; Murray et al., 2013; Cui et al., 2019; Feldman and Kowal, 2022). 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. These results suggest that Corollary 4 may be applicable more broadly, including for GMCs. In addition, the contraction of the margin adjustment around the ground truth is minimally affected by the proportion of missing data. 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^{obs}, yet the margin adjustment remains accurate for these cumulative probabilities. Analogous results for binary Y_3 are presented in the supplement, and demonstrate the exceptional performance of the GMC-MA for modeling MAR and mixed data.

6.2 Imputation for Regression Analysis

In the second simulation study, we study the impacts of imputation within the broader context of a regression analysis, and include comparisons with Bayesian nonparametric models and 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 (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 `New` using a Gaussian linear model with an `FI:BMI` interaction,

$$\texttt{New}_i \mid - \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}_{true}, \sigma^2)$$

where $\boldsymbol{\beta}_{true}$ is defined in Table 2 , and set σ^2 via the signal-to-noise-ratio $\text{SNR} = \text{var}(\boldsymbol{X}\boldsymbol{\beta}_{true})/\sigma^2 \in \{1, 3\}$. Finally, we introduce MAR for each variable j with the exception of `BMI`,

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

where ω_{ij} are Gaussian with $\text{Corr}(\omega_{ij}, \omega_{ij'}) = 0.3$ and mean -0.2. Thus, $\boldsymbol{R}$ introduces correlated and data-dependent (via `BMI`) patterns of missingness across both the response variable and the covariates. The missingness mechanism $\boldsymbol{R}$ is applied to all but 300 observations, and yields on average 49% complete cases. We repeat this data-generating process to create 100 hybrid synthetic data sets.

Since the missingness in `New`, `Age`, and `FI` is linked to `BMI`, CC analysis is at risk of significant bias. We illustrate this point in Table 2, where we compute ordinary least squares estimators ($\hat{\boldsymbol{\beta}}_{CC}$) and standard confidence bounds ($1.96\hat{\sigma}_{CC}$) for the regression coefficients using only the completely-observed data. For all variables besides `Age`, CC 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 MAR missingness.

For evaluations and comparisons among imputation methods, we generate $m = 20$ multiple imputations for each hybrid synthetic data set using several distinct approaches. First, we use the proposed GMC-MA approach to generate posterior samples and imputations. We employ the margin adjustment for `Age`, `New`, and `BMI`. Following the suggestions of Feldman and Kowal (2022), we treat `FI` as an unordered categorical variable. Next, we use

	Intercept	Middle	High	BMI	Age	Middle:BMI	High:BMI
$\boldsymbol{\beta}_{true}$	1	1	2	-2	0.5	2	4
$\hat{\boldsymbol{\beta}}_{CC}$	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 standard confidence bounds averaged across simulations. The CC analysis is severely biased for all variables except `Age`.

the same model and posterior draws, but replace the margin adjustment with the ECDF (GMC-ECDF). This comparison isolates the downstream impact of biased margin estimates for prediction under copula models. Specifically, improvements in imputation accuracy under the proposed approach demonstrate that the benefits of the margin adjustment extend more broadly to multivariate inference, and also the accumulating risk of using the ECDF for imputation with MAR data. These imputations are based on the Gibbs sampler (see the supplementary material) run for 10,000 iterations, with the first 5,000 discarded as a burn-in and the imputations computed every 50th sample to achieve $m = 20$.

To compare our approach to a Bayesian nonparametric alternative, we estimate the Gaussian mixture of factor models (12) on $\boldsymbol{Y}^{obs}$ (instead of $\boldsymbol{Z}^{obs}$), using the same priors and hyperpriors discussed in Section 4. This model is closely related to Chandra et al. (2023), and provides a flexible model for multivariate data. This model treats each variable as continuous, and offers an opportunity to evaluate the gains of employing a rank-based copula model for mixed variable types. To convert `FI` to a numeric variable, its levels are relabeled $(\texttt{low}, \texttt{middle}, \texttt{high}) = (1, 2, 3)$ which captures the ordinal properties of the variable, and we center and scale the observed values of `BMI`, `Age`, and `New`. Imputations of `Age` and `FI` under the Gaussian mixture model are rounded (GM-RND) to preserve these variables' discreteness in the completed data sets.

Among FCS (and non-Bayesian) methods, we create multiple imputations from the popular algorithm MICE (multiple imputation using chained equations; Van Buuren and Groothuis-Oudshoorn, 2011) 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).

For each completed data set, we fit a linear regression model for `New` (using the correct covariates and interactions) and use the combining rules from Rubin (2004) to create point estimates and 99% confidence intervals. The results are summarized in Figure 6 for SNR = 1 via the absolute bias for each point estimate, and the 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 results are similar for SNR = 3, and available in the supplement.

Clearly, the margin adjustment is crucial: the GMC-ECDF intervals do not provide close to the nominal coverage for several coefficients—despite using the same underlying model as the GMC-MA—due to the bias in the ECDF under MAR. As expected, the

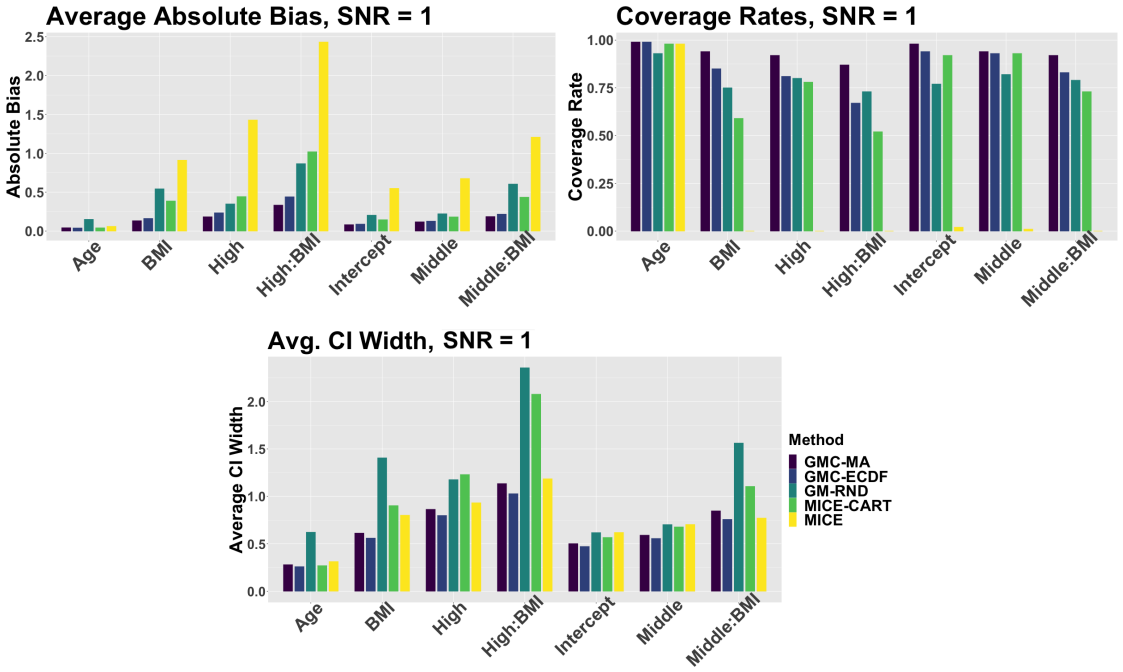

Figure 6: Absolute bias (left), interval coverage rates (right), and interval widths (center) for point and 99% interval estimates computed under each imputation method. The GMC-MA approach consistently provides the most accurate point estimates (small absolute bias), the most well-calibrated intervals (large coverage rates), and highly precise inference (small interval widths). Similar results for SNR = 3 are presented in the supplement.

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).

GM-RND provides mostly unsatisfactory results, evidenced by high average interval widths and poor coverage rates. The Gaussian mixture fit to the observed data clearly does not have the distributional flexibility to model mixed data types, even with a helpful rounding step. The RPL is specifically designed for this purpose, which yields significant improvement in modeling and imputation.

The default MICE approach also 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 while 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. 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.

7. Real Data Application

7.1 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, as 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, CC 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 suggests that MAR may be reasonable for these missing values. To strengthen this assumption, we include all variables in Table 1 in our GMC-MA model.

Second, the data set contains many variables of mixed data types, with $n = 5856$ observations of $p = 15$ variables (Table 1). The marginal distributions are nontrivial: `DMHNG` exhibits discreteness, zero-inflation, boundedness, and heaping (Figure 2), which are challenging for regression analysis (Kowal and Wu, 2023). To focus on at-risk individuals with high `DMHNG` values, we study the upper quantiles of `DMHNG` and the association with key variables of interest. However, the sample quantiles of this discrete variable do not satisfy asymptotic normality, and thus are ill-suited for traditional multiple imputation (Rubin, 2004). Instead, we propose an alternative and general strategy for uncertainty quantification based on the posterior predictive distribution. Specifically, we generate 500 posterior predictive data sets $\{\tilde{Y}_{ij}\}$ of size $n \times p$ and compute our summary statistics $\hat{Q}(\tilde{\boldsymbol{Y}})$ on each of these predictive data sets, which delivers posterior predictive inference for Q. This observation-driven inference leverages the GMC-MA model to capture challenging marginal and joint distributions across mixed data types in the presence of missingness, and simultaneously accounts for the joint uncertainty arising from (i) model parameters, (ii) missing data, and (iii) the replicability for a new data set of the same size. By comparison, statistics computed on the imputed data $(\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{mis})$ only incorporate uncertainty from $\boldsymbol{\theta}$ and $\boldsymbol{Y}^{mis}$, 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 over-sampled subgroups, most notably stratified by race. We stratify our analysis by race (and gender) and compute these quantities for each subgroup, which avoids the need to re-weight for population-level inference that aggregates across all strata.

7.2 Checking Calibration

Our aim is to quantify the extent to which model-based inferences change between 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. The CC data is created by dropping any observation with missing values, yielding a data set of size $n_{CC} = 2434$.

To understand the impact of the missingness, we rely on posterior predictive diagnostics. Posterior predictive diagnostics compare the posterior predictive distribution of a statistic $\hat{Q}(\tilde{\boldsymbol{Y}})$ to the observed value $\hat{Q}(\boldsymbol{Y})$. However, $\hat{Q}(\boldsymbol{Y})$ is unavailable for the full data set due to abundant missingness. Thus, the CC diagnostics are the best available option. Next, by comparing the GMC-MA model output from the CC and full data sets, we can assess the impact of missingness on the analysis, and in particular whether the CC analysis is biased or misleading.

Using the GMC-MA fits to both the CC and the full data sets, we generate predictive data sets 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 `DMHNG`, which is useful for posterior predictive diagnostics, and the 75th and 90th sample quantiles of `DMHNG`, which target the at-risk individuals within each stratum. The GMC-MA fit to the full data set must account for the additional uncertainty due to the missing observations, but also benefits from a much larger sample size. Differences in location for these posterior (predictive) distributions, however, would suggest bias due to missing data. Such discrepancies may be expected, as GMC-MA fit to the full and CC data sets discover distinct clusters—the former yields six versus five in the latter.

We compare the posterior predictive samples of the empirical distribution of `DMHNG` from the CC and full data set fits in Figure 7, and overlay the ECDF computed on all available cases in each strata. Specifically, we compute the ECDF on each posterior predictive data set $\{\tilde{Y}_{ij}\}$—stratified by race, gender, and marijuana use—for the CC and full data sets, and report the pointwise 95% highest posterior density (HPD) intervals.

The ECDF on the CC data falls within the 95% HPD intervals from the GMC-MA (CC) fit. Thus, the model accurately describes the challenging features of `DMHNG`: zero-inflation, heaping (the large jumps around `DMHNG` $\in \{7, 10, 14, 15, 20\}$), and boundedness at 30 (the lower interval converges to one at `DMHNG` $= 30$). These results are stratified by race, gender, and marijuana use, and thus evaluate the *joint* distribution. Similar results for the male-race-marijuana use strata are presented in the supplementary material.

Next, we compare the fitted GMC-MA models on the CC and full data sets. Most notably, the GMC-MA fit to the full data set has substantially narrower 95% HPD intervals, and often *shifts* the predictive intervals. For some strata, the predictive ECDF from the GMC-MA actually excludes the ECDF fit to the CC data (i.e. for white females) while the GMC-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 missing data. These results confirm our expectations based on the simulation results (Section 6) and suggest that a CC analysis of these data is unreliable.

7.3 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 $\hat{Q}(\tilde{\boldsymbol{Y}})$ for the 75th (see the supplement) and 90th quantile of `DMHNG`, stratified by gender-race-marijuana

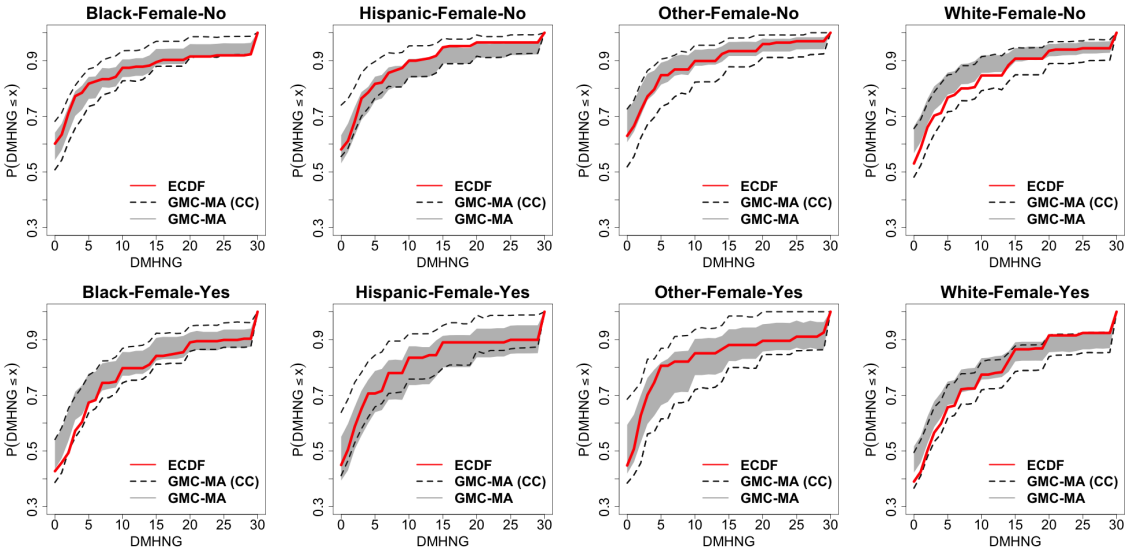

Figure 7: Posterior predictive summaries for models fit to the CC data set (GMC-MA (CC)) and the full data set (GMC-MA). Among women, for each race-marijuana use stratum, we compare the 95% HPD intervals for the posterior predictive ECDFs, and include the ECDF on the CC data for reference. As such, each panel is to be evaluated individually for model calibration and the impact of missingness. The GMC-MA (CC) output is well-calibrated to the observed data. By comparison, the GMC-MA fit to the full data set produces intervals that are narrower and shifted, which provides evidence of MAR—and that CC analysis is unreliable.

use. These quantities are computed for both GMC-MA on the full data set and GMC-MA (CC), and provide posterior predictive uncertainty quantification, which is summarized using the posterior median and 95% HPD intervals.

Figure 8 summarizes these point and interval estimates for the 90th quantile of `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 data set analysis with GMC-MA. The point estimates (posterior predictive medians) are similarly attenuated for the CC analysis, which is evident in Figure 9. Specifically, consider the difference in predictive medians between marijuana users and non-users in Figure 8. The estimated differences are positive for all strata: the 90th quantile of `DMHNG` is greater for marijuana users than non-users. However, these estimated differences are consistently larger for the GMC-MA on the full data set compared to GMC-MA (CC). 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 data set, but only 5 for GMC-MA (CC), which is highlighted in Figure 9. Similar trends are observed for the 75th quantile, although the discrepancies are less pronounced (see the supplementary material). Thus, CC analysis fails to identify certain strong, significant, and adverse associations between marijuana use and larger values of `DMHNG`, which are detected clearly under the full data set analysis.

These results emphasize the serious risks posed by CC analysis, which can produce biased or misleading conclusions. The implications are important for mental health studies:

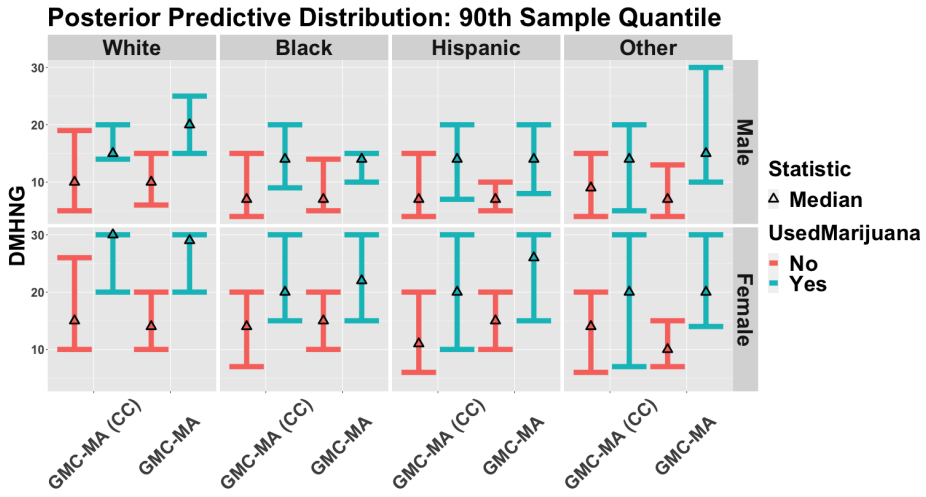

Figure 8: 95% HPD intervals for the predictive 90th quantiles of DMHNG by race-gender-marijuana use and comparing models fit to the complete case (CC) data set (GMC-MA (CC)) and the full data set (GMC-MA). The CC analysis produces wider intervals with more overlap between marijuana users and non-users across all strata, which dilutes the strong, significant, and adverse effects detected by GMC-MA fit to the full data set.

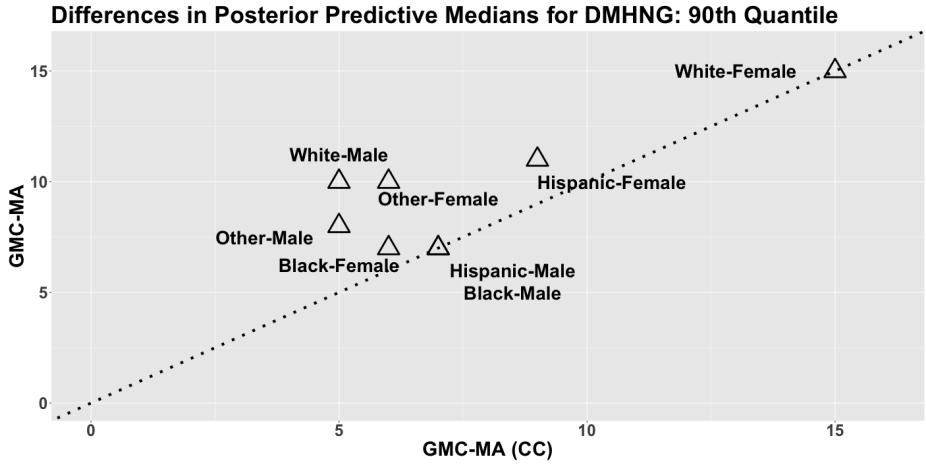

Figure 9: Difference in posterior predictive medians for the predictive 90th quantiles of DMHNG between marijuana users and non-users and comparing models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). The CC point estimates attenuate the differences between marijuana users and non-users across all strata, which dilutes the strong, significant, and adverse effects detected by GMC-MA fit to the full dataset.

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 data set analysis—despite abundant missingness, mixed data types, and complex marginal and joint distributions—and highlight the specific limitations of CC analysis.

8. Conclusion

We proposed a nonparametric copula model for mixed (count, continuous, ordinal, and unordered categorical) data types subject to values missing-at-random. The model featured 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. We applied our model and imputation strategies to self-reported mental health data and demonstrated the pitfalls of complete case analysis—and showed how the proposed approach may resolve these issues.

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. Carefully-designed simulation studies showed 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 MAR missingness. With its computational simplicity and desirable theoretical properties, we recommend integrating the margin adjustment in place of the ECDF for prediction and imputation under rank-based copula models.

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. Therefore, understanding the contraction rates of rank-likelihood posteriors, including as a function of the proportion of missing observations, is an important open question to provide implementation guidelines under increasing amounts of missingness. Similarly, our analyses suggest that rank-likelihoods and the margin adjustment may be applied more broadly for imputation using non-Gaussian copula models such as the rank-likelihood vine copula of Tekumalla et al. (2017). Such developments would broaden the applicability of Bayesian inference for copula models in the presence of missing data, while simultaneously eliminating the need to specify models for each marginal distribution and potentially delivering consistent inference for these margins. Finally, an important and challenging extension is to adapt the proposed framework for missingness-not-at-random. The latent factor mixture (12) 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).

Acknowledgments

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.

Appendix A. Proofs

Theorem 1 *Suppose* $\{Z_i\}_{i=1}^n \overset{i.i.d}{\sim} F_Z$ *and* $\{Y_i\}_{i=1}^n = \{h(Z_i)\}_{i=1}^n \sim F_Y$ *, where* F_Z *is continuous and* h *is a monotone increasing function. Defining* $Z^n(x)$ *as* (8)*, the margin adjustment satisfies* $\tilde{F}(x) \coloneqq F_Z\{Z^n(x)\} \overset{a.s.}{\to} F_Y(x)$ *for all* $x \in \mathbb{R}$*.*

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

For any x such that $\tau = F_Y(x) = 0$, $S = 1$ since the position of $Z^n(x)$ is now the same as $\min(Y_i)$. Applying the same argument above, the sequence of first order statistics of a uniform random variable will converge in probability to a degenerate random variable with point mass at 0. In addition, the sequence is monotone decreasing, which maintains the almost sure convergence. Thus, the almost sure convergence holds for any $x \in \mathbb{R}$. □

Theorem 2 *Suppose* $\{\boldsymbol{Z}_i\}_{i=1}^n = \{(Z_{i1}, Z_{i2})\}_{i=1}^n \overset{i.i.d}{\sim} G$*, where* G *is continuous with marginal distributions* G_1, G_2*, and* $\{\boldsymbol{Y}_i\}_{i=1}^n = \{(Y_{i1}, Y_{i2})\}_{i=1}^n = [(F_1^{-1}\{G_1(Z_{i1})\}, F_2^{-1}\{G_2(Z_{i2})\})]_{i=1}^n$ *has joint distribution function* F *with marginal distributions* F_1, F_2*. Suppose* Y_2 *is completely observed and* Y_1 *is MAR. The margin adjustment satisfies* $\tilde{F}_1(x) \coloneqq G_1\{Z_1^n(x)\} \overset{a.s.}{\to} F_1(x)$ *for all* $x \in \mathbb{R}$

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 $P_\cdot(x)$, where the subscript refers to the respective (marginal, conditional, or joint) distribution. The proof will show that $Z_1^n(x) \overset{a.s.}{\to} G_1^{-1}\{F_1(x)\}$ as $n \to \infty$, which yields the stated result via the continuous mapping theorem.

Note that because G_1, G_2, F_1, F_2 are non-decreasing, we have that $y_{ij}^{obs} < y_{lj}^{obs} \implies z_{ij}^{obs} < z_{lj}^{obs}$, $j = 1, 2$, $\forall l \neq i$ – i.e. the orderings between Y_j^{obs} and Z_j^{obs} are consistent. Next, suppose that Y_1 and Y_2 are continuous; the discrete case is addressed subsequently. Given that $\boldsymbol{Y}$ is a component-wise monotone transformation of $\boldsymbol{Z}$, the joint distribution F may be expressed in terms of G:

$$F(y_1, y_2) = G[G_1^{-1}\{F_1(y_1)\}, G_2^{-1}\{F_2(y_2)\}]. \tag{13}$$

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

$$\begin{aligned} P(Y_1 \leq x \mid Y_2 = y) &= P[Z_1 \leq G_1^{-1}\{F_1(x)\} \mid Z_2 = G_2^{-1}\{F_2(y)\}] \\ &= \int_{-\infty}^{G_1^{-1}\{F_1(x)\}} \frac{g[z, G_2^{-1}\{F_2(y)\}]}{g_2[G_2^{-1}\{F_2(y)\}]}\, dz \end{aligned} \tag{14}$$

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

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

The conditional probability $P_{Z_1|R_1=0}\{Z_1 \leq Z_1^n(x)\}$ is equivalently

$$P_{Z_1|R_1=0}\{Z_1 \leq Z_1^n(x)\} = E_{Z_2|R_1=0}[P_{Z_1|Z_2,R_1=0}\{Z_1 \leq Z_1^n(x)\}] \tag{15}$$

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

$$P_{Z_1|Z_2,R_1=0}\{Z_1 \leq Z_1^n(x)\} = P_{Z_1|Z_2}\{Z_1 \leq Z_1^n(x)\} = \int_{-\infty}^{Z_1^n(x)} \frac{g(z, Z_2)}{g_2(Z_2)}\, dz \tag{16}$$

and thus

$$P_{Z_1|R_1=0}\{Z_1 \leq Z_1^n(x)\} = E_{Z_2|R_1=0}[P_{Z_1|Z_2}\{Z_1 \leq Z_1^n(x)\}]. \tag{17}$$

Denoting (17) by $h\{Z_1^n(x)\}$, note that (16)–(17) imply that h is a continuous function. Consequently, we can write the limit of $P_{Z_1|R_1=1}\{Z_1 \leq Z_1^n(x)\}$ explicitly, yielding that

$$P_{Z_1|R_1=0}\{Z_1 \leq Z_1^n(x)\} = h\{Z_1^n(x)\} \overset{a.s.}{\to} h\{Z_1^\infty(x)\}. \tag{18}$$

Next, consider an application of Theorem 1 to the *observed* data. By construction, $Z_1^n(x)$ and the maximum position of $Y_1 \leq x$ 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 τth quantile under the distribution of observed $[Y_1 \mid R_1 = 0]$, then $Z_1^n(x)$ will converge to the τth quantile under the distribution of Z_1 corresponding to observed $[Y_1 \mid R_1 = 0]$:

$$P_{Z_1|R_1=0}\{Z_1 \leq Z_1^n(x)\} \overset{a.s.}{\to} P_{Y_1|R_1=1}(Y_1 \leq x). \tag{19}$$

We can also re-write $P_{Y_1|R_1=0}(Y_1 \leq x)$ in terms of Z_1 and Z_2 by (14) and (17):

$$\begin{aligned} P_{Y_1|R_1=0}(Y_1 \leq x) &= P_{Z_1|R_1=0}[Z_1 \leq G_1^{-1}\{F_1(x)\}] && (20) \\ &= E_{Z_2|R_1=0}(P_{Z_1|Z_2,R_1=0}[Z_1 \leq G_1^{-1}\{F_1(x)\}]) && (21) \\ &= E_{Z_2|R_1=0}(P_{Z_1|Z_2}[Z_1 \leq G_1^{-1}\{F_1(x)\}]) && (22) \\ &= h[G_1^{-1}\{F_1(x)\}]. && (23) \end{aligned}$$

Finally, we see the equivalence between (18) and (23) 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 $\boldsymbol{Z}$. 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

$$\begin{aligned} P(Y_1 \leq x \mid Y_2 = y) &= P[Z_1 \leq G_1^{-1}\{F_1(x)\} \mid Z_2 \in (G_2^{-1}\{F_2^-(y)\}, G_2^{-1}\{F_2(y)\}]] \\ &= \int_{-\infty}^{G_1^{-1}\{F_1(x)\}} \int_{G_2^{-1}\{F_2^-(y)\}}^{G_2^{-1}\{F_2(y)\}} \frac{g(z_1, z_2)}{g_2(z_2)} \, dz_2 \, dz_1 \end{aligned}$$

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.

To extend this result to p-dimensions, we first extend the structure of the joint distributions for $\boldsymbol{Z}$ and $\boldsymbol{Y}$ into p dimensions, where G, F are p-dimensional distribution functions with marginals $\{G_j\}_{j=1}^p$ and $\{Y_j\}_{j=1}^p$. As in (13)-(14), joint and conditional distributions for $\boldsymbol{Y}$ can similarly be expressed in terms of $\boldsymbol{Z}$.

We then partition $\boldsymbol{Y}$ into $(\boldsymbol{Y}^{comp}, \boldsymbol{Y}^{part})$, where $Y_j \in \boldsymbol{Y}^{part} \implies \sum_{i=1}^n R_{ij} > 0$ i.e. a variable is partially observed if it has at least one missing value in the sample. Y^{comp} is comprised of all variables which are completely observed, i.e. $Y_j \in \boldsymbol{Y}^{comp} \implies \sum_{i=1}^n R_{ij} = 0$. We assume that all variables in $\boldsymbol{Y}^{part}$ are MAR, which implies that $\boldsymbol{Y}^{comp}$ is non-empty. Then, for any $Y_j \in \boldsymbol{Y}^{part}$ and x, we define $Z_j^n(x)$ as in (10) and consider the sequence of probabilities $P_{Z_j|R_j=0}\{Z_j \leq Z_j^n(x)\}$. Defining $\boldsymbol{Z}_{-j}$ to be $\boldsymbol{Z} \setminus Z_j$, the rest of the proof follows identically by noticing that $P_{Z_j|R_j=0}\{Z_j \leq Z_j^n(x)\} = E_{\boldsymbol{Z}_{-j}|R_j=0}[P_{Z_j|\boldsymbol{Z}_{-j},R_j=0}\{Z_j \leq Z_j^n(x)\}]$ converges almost surely to a limit since $Z_j^n(x)$ is monotone and bounded, while Theorem 1 implies that $P_{Z_j|R_j=0}\{Z_j \leq Z_j^n(x)\} \overset{a.s.}{\to} P_{Y_j|R_j=0}(Y_j \leq x) = h[G_j^{-1}\{F_j(x)\}]$. □

Theorem 3 *Suppose* $\{\boldsymbol{Y}_i\}_{i=1}^n \overset{i.i.d}{\sim} G^\infty_{\boldsymbol{C}_0,F_1,\ldots,F_p}$, *where* $G^\infty_{\boldsymbol{C}_0,F_1,\ldots,F_p}$ *is the Gaussian copula for the joint distribution of p-dimensional* $\boldsymbol{Y}$ *with true copula parameters* $\boldsymbol{C}_0$ *and true marginal CDFs* $F_1, \ldots, F_p$. *Let* Π *be a prior distribution on the space of all* $p \times p$ *positive semi-definite correlation matrices* $\boldsymbol{C}_{\boldsymbol{\theta}}$ *with corresponding density* $\pi(\boldsymbol{C}_{\boldsymbol{\theta}})$ *with respect to a measure* ν. *Suppose* $\pi(\boldsymbol{C}_{\boldsymbol{\theta}}) > 0$ *almost everywhere with respect to* ν *and assume that the missingness is ignorable. Then, for* $\boldsymbol{C}_0$ *a.e.* $[\nu]$ *and any neighborhood* $\mathcal{A}$ *of* $\boldsymbol{C}_0$, *we have that* $\lim_{n\to\infty} \Pi\{\boldsymbol{C}_{\boldsymbol{\theta}} \in \mathcal{A} \mid \boldsymbol{Z}_n^{obs} \in \mathcal{D}(\boldsymbol{Y}_n^{obs})\} = 1$ *a.s* $[G^\infty_{\boldsymbol{C}_0,F_1,\ldots,F_p}]$.

Proof. Because the missingness mechanism is ignorable, the priors for the parameters governing the data generating process for $\boldsymbol{Y}$ and the missingness mechanism $\boldsymbol{R}$ are independent (Rubin, 1976). Therefore, we can utilize the following 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* $\boldsymbol{\theta}$, *both taking values in Polish spaces. Assume* $\boldsymbol{\theta} \sim \Pi$ *and* $X_i \mid \boldsymbol{\theta} \sim P_{\boldsymbol{\theta}}$. *Let* $\mathcal{X}_N$ *be the* σ*-field generated by* $X_1, \ldots, X_N$, *and* $\mathcal{X}_\infty = \sigma(\bigcup_i^\infty \mathcal{X}_i)$. *If there exists a* $\mathcal{X}_\infty$

measureable function f such that for $(\boldsymbol{\omega}, \boldsymbol{\theta}) \in \Omega^{\infty} \times \Theta$, $\boldsymbol{\theta} = f(\boldsymbol{\omega})$ a.e. $[P_{\boldsymbol{\theta}}^{\infty} \times \Pi]$ then the posterior is strongly consistent at $\boldsymbol{\theta}$ for almost every $\boldsymbol{\theta}$ $[\Pi]$.

To adopt Doobs Theorem to this setting, note the data generating process for $\boldsymbol{Y}_n$

$$\boldsymbol{z}_i \overset{i.i.d}{\sim} N_p(\boldsymbol{0}, \boldsymbol{C}_0) \tag{24}$$

$$y_{ij} = F_j^{-1}\{\Phi(z_{ij})\} \tag{25}$$

implies each $\boldsymbol{Z}_n^{obs}$, generated i.i.d by a probability distribution indexed by $\boldsymbol{C}_0$, must satisfy the event $\mathcal{D}(\boldsymbol{Y}_n^{obs})$ where n indexes the sample size. Therefore, it suffices to establish the existence of a consistent estimator $\boldsymbol{C}_{\boldsymbol{\theta}}$ of $\boldsymbol{C}_0$, the data generating Gaussian copula correlation matrix, that is measureable with respect to the sigma-field generated by the sequence $\{\boldsymbol{Z}_n^{obs} \in \mathcal{D}(\boldsymbol{Y}_n^{obs})\}_{n=1}^{\infty}$, where n indexes the sample size.

Suppose $\boldsymbol{Y}_n$ is comprised of $n_1 > 1$ complete cases without any missing values ($\boldsymbol{Y}^{CC}$) and n_2 cases with missing values for at least one variable ($\boldsymbol{Y}^{inc}$) such that $n_1 + n_2 = n$. For each observation i in $\boldsymbol{Y}^{CC}$, for each variable $j \in \{1, \dots, p\}$ with $y_{ij} = x$, consider $Z_j^n(x)$ as in (10). Let $T_{nij} = \sum_{i=1}^{n} \mathbb{1}\{z_{ij} \leq Z_j^n(x)\}$, with $\boldsymbol{T}_{ni}(\boldsymbol{Y}_i^{obs}) = (T_{ni1}, \dots, T_{nip})$ and $\boldsymbol{T}_n(\boldsymbol{Y}_n^{obs}) = \{\boldsymbol{T}_{ni}(\boldsymbol{Y}_i^{obs})\}_{i=1}^{n}$. As noted in Murray et al. (2013), the information contained $\boldsymbol{T}_n(\boldsymbol{Y}_n^{obs})$ is also contained in $\boldsymbol{Z}_n^{obs}$. Namely, $\boldsymbol{T}_n(\boldsymbol{Y}_n^{obs})$ may be extracted from the boundary conditions of $\boldsymbol{Z}_n^{obs} \in \mathcal{D}(\boldsymbol{Y}_n^{obs})$. Therefore, any function measureable with respect to $\mathscr{T}_n$, the sigma-algebra generated by the sequence $\{\boldsymbol{T}_n(\boldsymbol{Y}_m^{obs})\}_{m=1}^{n}$ is also measureable with respect to the sigma algebra generated by the corresponding sequence $\{\boldsymbol{Z}_m^{obs} \in \mathcal{D}(\boldsymbol{Y}_m^{obs})\}_{m=1}^{n}$. Consequently, as in Murray et al. (2013), we exclusively work with $\mathscr{T}_n$.

Now, $Z_j^n(x)$ 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 $\hat{\boldsymbol{U}}_{ni} = (\hat{U}_{ni1}, \dots, \hat{U}_{nip})$. Then,

$$\hat{U}_{nij} = \frac{1}{n+1} E \sum_{l=1}^{n} \mathbb{1}\{z_{lj} \leq Z_j^n(x)\} \tag{26}$$

$$= \frac{1}{n+1} \sum_{l=1}^{n} P\{z_{lj} \leq Z_j^n(x)\} \tag{27}$$

By the strong law of large numbers and Theorem 2, (27) converges almost surely to

$$P\left[z_j \leq \Phi^{-1}\{F_j(x)\}\right] = F_j(x) \tag{28}$$

since $\Phi\{Z_j^n(x)\} \overset{a.s.}{\to} F_j(x)$. Therefore, we have that $\hat{U}_{nij} \overset{a.s.}{\to} U_{ij}$, where $U_{ij} = F_j(x)$, the cumulative marginal probability of x for variable j under the true distribution F_j. Consequently, $\hat{\boldsymbol{U}}_{ni} \overset{a.s.}{\to} \boldsymbol{U}_i = (U_1, \dots, U_p)$, and $\boldsymbol{U}_i$ is $\mathscr{T}_{\infty}$ measureable.

Therefore, $\boldsymbol{U}_i$ is a sample from a Gaussian copula with correlation matrix $\boldsymbol{C}_0$ where the continuous margins are Uniform$[0, 1]$ while the discrete margins are merely re-labeled with their ground-truth cumulative probabilities. We then may apply the argument of Murray et al. (2013) which establishes the existence of a consistent estimator of $\boldsymbol{C}_0$ which is a

function of $\boldsymbol{U}_i$ and thus $\mathscr{T}_\infty$ measureable. Specifically, the problem reduces to estimating polychoric/polyserial correlations with fixed margins, where $\boldsymbol{U}_i$ is a regular parametric family admitting a sequence of consistent estimators of $\boldsymbol{C}_0$, for instance using the estimators of Olsson (1979) and Olsson et al. (1982). □

Corollary 4 *Under the conditions of Theorem 3, define* $\tilde{F}_j$ *as in* (9) *with* $Z_j^n(x)$ *as* (10) *and* $G_j = \Phi$ *for each* $j \in \{1, \ldots, p\}$. *Then for any* $x \in \mathbb{R}$ *and any neighborhood* $\mathcal{A}$ *of* $F_j(x)$ $\lim_{n\to\infty} \Pi\{\tilde{F}_j(x) \in \mathcal{A} \mid \boldsymbol{Z}_n^{obs} \in \mathcal{D}(\boldsymbol{Y}_n^{obs})\} = 1$ *a.s* $[G^\infty_{\boldsymbol{C}_0, F_1, \ldots, F_p}]$.

Proof. This result follows from an application of Doob's Theorem presented above and Theorem 2. To apply Doob's theorem, consider the marginal distributions of $\{Z_j\}_{j=1}^p$ under the data generating copula model. Specifically, for any $j \in \{1, \ldots, p\}$, each Z_{ij}^{obs} is standard normal, but restricted to fall in the subset of the real line determined by the right and left limits of the true marginal distribution F_j evaluated at the realized value of Y_{ij}^{obs}. Then, defining the marginal event $\mathcal{D}(Y_{j_n}^{obs}) \coloneqq \{Z^{n\times 1} : y_{lj}^{obs} < y_{kj}^{obs} \implies z_{lj}^{obs} > z_{kj}^{obs}, k \neq l\}$ and $Z_{j_n}^{obs}$ the latent vector corresponding to $Y_{j_n}^{obs}$, it must be the case that $Z_{j_n}^{obs} \in \mathcal{D}(Y_{j_n}^{obs})$. In addition, the sigma-field generated by the sequence $\{Z_{j_m} \in \mathcal{D}(Y_{j_m}^{obs})\}_{m=1}^n$ is a sub sigma-field of the sigma-field generated by the sequence $\{\boldsymbol{Z}_m^{obs} \in \mathcal{D}(\boldsymbol{Y}_m^{obs})\}_{m=1}^n$, and hence any function measureable with respect to the former is also measureable with respect to the latter.

Therefore, it suffices to demonstrate the existence of a strongly consistent estimator of $F_j(x)$ that is measureable with respect to the sigma-field generated by the sequence of $\{Z_{j_n} \in \mathcal{D}(Y_{j_n}^{obs})\}_{n=1}^\infty$. For any n, $Z_j^n(x)$ is clearly an element of $Z_{j_n}^{obs} \in \mathcal{D}(Y_{j_n}^{obs})$. As such, $\tilde{F}_j(x) = \Phi\{Z_j^n(x)\}$ is measureable with respect to sigma-field generated by the sequence $\{Z_{j_m} \in \mathcal{D}(Y_{j_m}^{obs})\}_{m=1}^n$. By Theorem 2 and because the data are generated from a Gaussian copula under the conditions in Theorem 3 with true marginals $\{F_j\}_{j=1}^p$, it follows that $\tilde{F}_j(x)$ is a strongly consistent estimator of $F_j(x)$ and hence $F_j(x)$ is measureable with respect to the sigma-field generated by the sequence of $\{Z_{j_n} \in \mathcal{D}(Y_{j_n}^{obs})\}_{n=1}^\infty$. Thus, the posterior of $\tilde{F}_j(x)$ is strongly consistent at $F_j(x)$. This applies for each continuous and count variable $j = 1, \ldots, p$. □

Theorem 5 *Let* $\mathbb{C}_{GMC}(\boldsymbol{u}) = \Psi(\psi_1^{-1}\{F_1(y_1)\}, \ldots, \psi_p^{-1}\{F_p(y_p)\})$, *where* $\Psi = \sum_{h=1}^H \pi_h \Phi_p(\boldsymbol{\alpha}_h, \boldsymbol{C}_h)$, $\psi_j = \sum_{h=1}^H \pi_h \Phi(\{\boldsymbol{\alpha}_h\}_j, \{\boldsymbol{C}_h\}_{jj})$, *and* $\{F_j\}_{j=1}^p$ *are the marginals of* $\{Y_j\}_{j=1}^p$. *Then,* $\mathbb{C}_{GMC}$ *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 $u_{j1} < u_{j2}$. Define $z_{j1} = \psi_j^{-1}(u_{j1})$ and $z_{j2} = \psi_j^{-1}(u_{j2})$. Because ψ_j is a valid continuous distribution function, it is monotone, and therefore $z_{j1} < z_{j2}$.

Consider the ratio

$$\begin{aligned}\frac{\mathbb{C}_{GMC}(u_1,\ldots,u_{j1}\ldots,u_p)}{\mathbb{C}_{GMC}(u_1,\ldots,u_{j2}\ldots,u_p)} &= \frac{\Psi\{\psi_1^{-1}(u_1),\ldots,z_{i1}\ldots,\psi_p^{-1}(u_p)\}}{\Psi\{\psi_1^{-1}(u_1),\ldots,z_{j2}\ldots,\psi_p^{-1}(u_p)\}} \\ &= \sum_{h=1}^{H}\pi_h \frac{\Phi\left[\{\psi_1^{-1}(u_1),\ldots,z_{j1},\ldots,\psi_p^{-1}(u_p)\};\boldsymbol{\alpha}_h,\boldsymbol{C}_h\right]}{\Phi\left[\{\psi_1^{-1}(u_1),\ldots,z_{j2},\ldots,\psi_p^{-1}(u_p)\};\boldsymbol{\alpha}_h,\boldsymbol{C}_h\right]}\end{aligned}$$

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

$$\sum_{h=1}^{H}\pi_h \frac{\Phi\left[z_{j1};\{\psi_l^{-1}(u_l)\}_{l\neq j},\alpha_h^*,\sigma_h^{2^*}\right]}{\Phi\left[z_{j2};\{\psi_l^{-1}(u_l)\}_{l\neq ij},\alpha_h^*,\sigma_h^{2^*}\right]} < \sum_{h=1}^{H}\pi_h = 1 \tag{29}$$

$$\implies \mathbb{C}_{GMC}(u_1,\ldots,u_{i1}\ldots,u_p) < \mathbb{C}_{GMC}(u_1,\ldots,u_{i2}\ldots,u_p) \tag{30}$$

where $\alpha_h^*, \sigma_h^{2^*}$ are the conditional mean and variance of the Gaussian random variable obtained by conditioning on $\{\psi_l^{-1}(u_{;})\}_{l\neq j}$ 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 (29) 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 $\forall j \in \{1,\ldots,p\}$, $\psi_j^{-1}(1) = \sum_{h=1}^{H}\pi_h\Phi^{-1}\{1;(\boldsymbol{\alpha}_h)_j,(\boldsymbol{C}_h)_{jj}\} = \infty$.

   Using the above result, it is simple to see that

$$\begin{aligned}\mathbb{C}_{GMC}(u_1 = 1,\ldots,u_j = u,\ldots,u_p = 1) &= \Psi\{\infty,\ldots,\psi_j^{-1}(u),\ldots,\infty\} \\ &= \psi_j\{\psi_j^{-1}(u)\} \\ &= u\end{aligned}$$

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

$$\begin{aligned}\mathbb{C}_{GMC}(u_1 \in [a_1,b_1],\ldots,u_p \in [a_p,b_p]) &= \mathbb{C}_{GMC}(u_1 \leq b_1,\ldots,u_p \leq b_p) - \mathbb{C}_{GMC}(u_1 \leq a_1,\ldots,u_p \leq a_p) \\ &= \Psi\{\psi_1^{-1}(b_1),\ldots,\psi_p^{-1}(b_p)\} - \Psi\{\psi_1^{-1}(a_1),\ldots,\psi_p^{-1}(a_p)\} \\ &\geq 0 \text{ (By 1.)}\end{aligned}$$

□

Appendix B. Bayesian RPL Gaussian Copula Sampling with Unordered Categorical Variables

Section 2.3 outlines the sampling algorithm for the Bayesian RL Gaussian copula with missing data for $\boldsymbol{Y}$ comprised of numeric variables. To incorporate unordered categorical in to model, one simple modification is required to Step 1 of Algorithm 1. To see this, the probit event $\mathcal{D}'(\boldsymbol{Y}^q)$ dictates that for any categorical variable Y_c with k_c levels, if $y_{ic} = m$, the k_c dimensional latent data vector must satisfy the event $z_{i_{k_m}} > 0 \cap z_{i_{k_\ell}} < 0, \ell \neq m$. Therefore, the upper bounds for each $\boldsymbol{Z}_{ij}^{obs}$ are pre-specified at either 0 or ∞, while the

lower bounds are similarly pre-specified at $-\infty$ or 0 for any j corresponding to a categorical level. If the indicator $\gamma_{ij} = 1$, then $z_u = \infty, z_\ell = 0$. On the other hand, for $\gamma_{ij} = 0$ then $z_u = 0, z_\ell = -\infty$. The sampling step for Z_{ij}^{mis} first calculates the predictive probability that $Y_{ic}^{mis} = m$, and then samples the corresponding latent vector with identical upper and lower truncation bounds as the observed case depending on the predictive level. This sampling augments the latent data matrix to $p^* = r + \sum_{c=1}^{q} k_c$ columns. Step 3 in Section C.2 contains details on this computation under the GMC-MA.

Appendix C. Model Specification, Gibbs Sampling, and Scalability

C.1 Global-local shrinkage priors

In the main paper, we mention the use of global-local shrinkage priors for the parameters of the factor-loading matrix $\boldsymbol{\Lambda} = \{\lambda_{jt}\}$. The following prior encourages columnwise shrinkage for rank selection (Bhattacharya and Dunson, 2011): $\lambda_{jt} \sim N(0, \phi_{jt}^{-1}\tau_t^{-1})$ with local scale parameters $\phi_{jt} \sim \text{Gamma}(\nu_\phi/2, \nu_\phi/2)$ and global scale parameters $\tau_t = \prod_{w=1}^{t} \delta_w$, with $\delta_1 \sim \text{Gamma}(a_1, 1)$ and $\delta_t \sim \text{Gamma}(a_2, 1),\ t \geq 2,\ a_2 \geq 1$. By design, this ordered shrinkage prior reduces sensitivity to the choice of k, provided k is sufficiently large. Throughout our simulation studies and real data analysis, we set $a_1 = 2, a_2 = 3, \nu_\phi = 3$.

C.2 Gibbs Sampling for the RPL GMC-MA

Bayesian estimation of the GMC-MA under the RPL 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 $\boldsymbol{Y}^{mis}$ given $\boldsymbol{Z}^{obs}$ and model parameters. Two aspects of our model simplify this task. First, the margin adjustments $\{\tilde{F}_j\}_{j=1}^r$ are functionals of posterior samples of GMC parameters, and second, GMC parameters depend only latent $\boldsymbol{Z} = (\boldsymbol{Z}^{mis}, \boldsymbol{Z}^{obs})$ through the RPL. Conjugate priors for GMC parameters allow for simple Gibb's sampling steps, while the factor model 12 allows for independence among the components of $\boldsymbol{Z}_i$ conditional on $\boldsymbol{\eta}_i$. Consequently, the sampling of $\boldsymbol{Z}^{mis}$ is quite efficient, as the predictive distribution for each component is conditionally univariate normal.

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

1. **Sample Cluster-Specific Parameters** For each cluster $1, \ldots, H$
   - $c_i \mid - \sim \text{Multinomial}(\boldsymbol{p})$, where $\boldsymbol{p} = (p_1, \ldots, p_H)$ and $p_h \propto \pi_h \psi_k(\boldsymbol{\eta}_i; \boldsymbol{\mu}_h, \boldsymbol{\Delta}_h)$
   - $V_h \mid - \sim \text{Beta}\ (1 + n_h, \alpha_\pi + \sum_{v=h+1}^{H} n_v)$, $h = 1, \ldots, H-1$, $n_h = \sum_{i=1}^{N} \mathbb{I}(c_i = h)$
   - $\boldsymbol{\Delta}_h \mid - \sim \text{IW}(\nu_{post}, \Psi_{post})$, $\nu_{post} = \nu_0 + n_h$, $\Psi_{post} = \boldsymbol{I}_k + \boldsymbol{S}_h + \frac{\kappa_0 n_h}{\kappa_0 + n_h}\boldsymbol{T}_h$, $\boldsymbol{S}_h = \sum_{i:c_i=h}(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_h)(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_h)^T, \boldsymbol{T}_h = (\boldsymbol{\mu}_0 - \bar{\boldsymbol{\eta}}_h)(\boldsymbol{\mu}_0 - \bar{\boldsymbol{\eta}}_h)^T, \bar{\boldsymbol{\eta}}_h = n_h^{-1}\sum_{i:c_i=h}\boldsymbol{\eta}_i$
   - $\boldsymbol{\mu}_h \mid - \sim N_k(\boldsymbol{\mu}_{post}, \kappa_{post}^{-1}\boldsymbol{\Delta}_h)$, $\boldsymbol{\mu}_{post} = \frac{\kappa_0\boldsymbol{\mu}_0 + n_h\bar{\boldsymbol{\eta}}_h}{\kappa_0 + n_h}$, $\kappa_{post} = n_h + \kappa_0$
   - $\alpha_\pi \mid - \sim \text{Gamma}(a_\alpha + H - 1, b_\alpha - \sum_{h=1}^{H-1}\log(V_h))$
2. **Sample Factor Model Parameters**

- $\boldsymbol{\eta}_i \mid c_i = h, - \sim N_k((\boldsymbol{\Delta}_h^{-1} + (\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1})^{-1}(\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_i + \boldsymbol{\Delta}_h^{-1}\boldsymbol{\mu}_h), (\boldsymbol{\Delta}_h^{-1} + (\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1})^{-1})$
- $\boldsymbol{\lambda}_{j,-} \mid - \sim N_k((\boldsymbol{D}_j^{-1} + \sigma_j^{-2}\boldsymbol{\eta}^T\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^T\sigma_j^{-2}\boldsymbol{z}_j, (\boldsymbol{D}_j^{-1} + \sigma_j^{-2}\boldsymbol{\eta}^T\boldsymbol{\eta})^{-1})$, where $\boldsymbol{D}_j^{-1} = diag(\phi_{j1}\tau_1, \ldots, \phi_{jk}\tau_{jk})$, $\boldsymbol{z}_j = (z_{1j}, \ldots z_{nj})^T$, and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots \boldsymbol{\eta}_n)^T$, for $j = 1, \ldots, p$
- $\sigma_j^{-2} \mid - \sim \text{Gamma}(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^n \sum_{t=1}^k \{z_{ij} - (\lambda_{jt}\eta_{it})\}^2$, for $j = 1, \ldots, p$
- $\phi_{jt} \mid - \sim \text{Gamma}(\frac{\nu+1}{2}, \frac{\nu+\tau_h\lambda_{jt}^2}{2})$, for $j = 1, \ldots, p,\ t = 1, \ldots, k$
- $\delta_1 \mid - \sim \text{Gamma}(a_1 + \frac{pk}{2}, 1 + \frac{1}{2}\sum_{t=1}^k \tau_t^{(1)} \sum_{j=1}^p \phi_{jt}\lambda_{jt}^2)$, and for $v \geq 2$
  $\delta_t \mid - \sim \text{Gamma}(a_1 + \frac{p(k-t+1)}{2}, 1 + \frac{1}{2}\sum_{t=v}^k \tau_t^{(v)} \sum_{j=1}^p \phi_{jt}\lambda_{jt}^2)$, where $\tau_t^{(v)} = \prod_{w=1, w\neq v}^t \delta_w$, for $v = 1, \ldots, k$

3. **Re-sample** $Z_{ij}^{obs}, Z_{ij}^{mis}$
Given the conditional independence among the components of $\boldsymbol{Z}_i$ given $\boldsymbol{\eta}_i$, components of $\boldsymbol{Z}$ corresponding to observed data points are sampled column-by-column, consistent with the ordering induced by the RPL. For components of $\boldsymbol{Z}_i$ 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 $\boldsymbol{Z}_{k_q}$ must be positive while the others negative. To ensure this condition is met, we first calculate the predictive probability that Y_{iq}^{mis} assumes the jth level for each j among the k_q levels, which is equivalent to

$$P(Z_{ij}^{mis} > 0 \cap \{Z_{i\ell}^{mis}\}_{\ell \in \{c_1, \ldots, c_{k_q}\},\ \ell \neq j} < 0 \mid -) \propto \tag{31}$$
$$1 - \Phi(0; \sum_{t=1}^k \lambda_{jt}\eta_{it}, \sigma_j^2) \prod_{\ell \in \{c_1, \ldots, c_{k_q}\}, \ell \neq j} \Phi(0; \sum_{t=1}^k \lambda_{\ell t}\eta_{it}, \sigma_\ell^2), j \in \{c_1, \ldots, c_{k_q}\}$$

  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 μ, variance σ^2, lower truncation a, and upper truncation b. The re-sampling step for Z_{ij}^{mis} is given by

$$z_{ij}^{mis} \sim \begin{cases} \text{TN}(\sum_{t=1}^k \lambda_{jt}\eta_{it}, \sigma_j^2, 0, \infty), & y_{ij}^{mis} = 1 \\ \text{TN}(\sum_{t=1}^k \lambda_{jt}\eta_{it}, \sigma_j^2, -\infty, 0), & y_{ij}^{mis} = 0 \end{cases} \tag{32}$$

  If j is binary, the probability of one level versus the other is instead given by $P(z_{ij}^{mis} > 0 \mid c_i = h, -) = 1 - \Phi(0; \sum_{t=1}^k \lambda_{jt}\eta_{it}, \sigma_j^2)$, but the re-sampling step 32 remains the same
- *Missing numeric data:* In this case, latent Z_{ij}^{mis} is sampled from the unrestricted univariate Gaussian

$$Z_{ij}^{mis} \mid - \sim N(\sum_{t=1}^k \lambda_{jt}\eta_{it}, \sigma_j^2) \tag{33}$$

- *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:

$$Z_{ij}^{obs} \mid - \sim \text{TN}(\sum_{t=1}^{k} \lambda_{jt}\eta_{it}, \sigma_j, z_{ij}^{\ell}, z_{ij}^{u}) \tag{34}$$

For ordinal, count, and continuous variables, the truncation limits are $z_{ij}^{\ell} = \max\{z_{kj}^{obs} : y_{kj} < y_{ij}^{obs}, k = 1, \ldots, n,\ k \neq i\}$, and $z_{ij}^{u} = \min\{z_{kj}^{obs} : y_{kj}^{obs} > y_{ij}^{obs}, k = 1, \ldots, n,\ k \neq i\}$. For columns corresponding to categorical levels, the upper and lower truncation limits are

$$z_{ij}^{\ell} = \begin{cases} 0, & Y_{ij}^{obs} = 1 \\ -\infty, & Y_{ij}^{obs} = 0 \end{cases}, \qquad z_{ij}^{u} = \begin{cases} \infty, & Y_{ij}^{obs} = 1 \\ 0, & Y_{ij}^{obs} = 0 \end{cases} \tag{35}$$

4. **Sample** $\tilde{F}_j$
For each unique $x \in \boldsymbol{Y}_j^{obs}$, we first find $Z_j^n(x) = \max\{Z_{ij}^{obs} : Y_{ij}^{obs} \leq x\}$, and compute

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

Where ψ_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 $\{x, \tilde{F}_j(x)\}_{x \in \boldsymbol{Y}_j^{obs}}$ as described in Section 4, and use this estimate to approximate $\tilde{F}_j(x')$ for $x' \notin \boldsymbol{Y}_j^{obs}$.

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

C.3 Run Time Considerations

The run time of the Gibbs sampling algorithm depends on both the number of observations in the data set as well as the number of variables for which the margin adjustment is computed. The bulk of computational expense in the detailed Algorithm presented in Section C.2 is in Steps 3 and 4: in the former, each component of the latent data matrix is sampled from its conditional posterior, truncated to upper and lower bounds consistent with rank restrictions on the observed scale. In Step 4, the margin adjustment requires computing the cut-point $Z_j^n(x)$ for each observed x in column j.

In the NHANES example, the augemented data set has $p = 22$ and $n = 5856$ in the full data case. We ran the Gibbs sampler in Appendix C.2 for 20,000 iterations, and applied the margin adjustment for each unique value among the numeric columns. The total run time for this process was just over an hour. For our simulation examples in Section 6, there were generally many fewer variables. As such, run times were generally much quicker – between 1 and 3 minutes with $p = 3$ and $n \in \{500, 1000, 2000\}$ in the first exercise, and around 5 minutes when increasing p in the second example. All experiments were run locally on a 2023 Macbook Pro with 32 GB of memory.

As a general guideline, the GMC-MA may be estimated on thousands of observations and dozens of variables. To provide further scalability, we note that Step 3 (and Step 4)

of the Gibbs sampling algorithm may utilize parallel computing across p: the factor model implies conditional independence between the columns of the latent data matrix, and so each column may be sampled in parallel. This technique may also be extended to the margin adjustment, since it is merely a post-processing step given posterior samples of $\boldsymbol{Z}$. Though we do not pursue this strategy in the main paper, we anticipate significant computational gains applying these modifications, enabling estimation on higher dimensional data sets. We leave this to further research.

C.4 Monitoring Convergence

The GMC-MA is a highly parameterized model due to the latent mixture. However, we find it simple to monitor convergence of the Gibbs sampler C.2 through examination of the margin adjustment for each variable, since it is computed as a posterior functional of the mixture copula parameters. In general, this is done by examining trace plots of $\psi_j\{Z_j^n(x)\}$ across j and unique values x of Y_j. The mixing of these quantities may be sensitive to n, p, and the proportion of missing data. In all of our simulated and real data examples, trace plots of usually indicated convergence within 1,000 iterations of the sampler, although much more conservative burn-in periods were used for imputation.

Appendix 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.

Specifically, there are three parameters that we vary to bolster model performance. The first is the dimension of latent $\boldsymbol{\eta}$. 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, δ, from the normal-inverse Wishart prior specified for the cluster specific components from the mixture model on $\boldsymbol{\eta}_i$. Recall the following hierarchical model structure

$$\begin{aligned} \boldsymbol{z}_i &\sim N_p(\boldsymbol{\Lambda}\boldsymbol{\eta}_i, \boldsymbol{\Sigma}) \\ \boldsymbol{\eta}_i &\sim \sum_{h=1}^{H} \pi_h N_k(\boldsymbol{\mu}_h, \boldsymbol{\Delta}_h) \\ (\boldsymbol{\mu}_h, \boldsymbol{\Delta}_h) &\sim \text{NIW}(\boldsymbol{\mu}_0, \delta^2 \boldsymbol{I}_k, \kappa_0 = 0.001, \nu_0 = k+2) \end{aligned}$$

In the simulation studies and real data analysis, we found that δ 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 δ 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 diagnostics— for instance checking marginal and multivariate properties of posterior predictive data sets created via the sampling algorithm in Section E —may be used 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 n. As such, re-sampling columns of $\boldsymbol{Z}$ 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 variables having more than this number of unique levels are not re-sampled within the MCMC. Instead, corresponding $\boldsymbol{Z}^{obs}$ is fixed by scaling the column to have mean zero and unit variance, which maintains consistent orderings with $\boldsymbol{Y}^{obs}$. In our applications, we choose this threshold to be 350. We emphasize that this is carried out for computational efficiency and did not affect the performance of the GMC-MA in any of the simulations or real data analysis.

Appendix E. Posterior Predictive Sampling Algorithm

Section 7 utilizes posterior predictive inference to highlight discrepancies between a complete case analysis and one that accounts for potentially MAR missing data. We include Algorithm 4 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 $\tilde{\boldsymbol{\eta}}_i$, which in turn enables sampling of predictive $\tilde{\boldsymbol{z}}_i$. We then link each component of $\tilde{\boldsymbol{z}}_i$ with $\tilde{F}_j^{-1}\{\psi_j(\tilde{z}_{ij})\}$ for numeric variables, or with the categorical link mentioned in Section C.

Appendix F. Further Simulation Results

F.1 Mixed Data Types, Nonlinearity, and MAR

As mentioned in Section 6.1, we also estimate the posterior distribution of the probability of a positive indicator for binary variable X_3. Under Model (12) and the RPL, this is simply the probability that latent Z_3 is greater than zero. This quantity is computed as $1 - \sum_{h=1}^{H} \pi_h^s \Phi\{0; (\boldsymbol{\Lambda}^s \boldsymbol{\mu}_h^s)_3, (\boldsymbol{\Lambda}^s \boldsymbol{\Delta}_h^s \boldsymbol{\Lambda}^{Ts} + \boldsymbol{\Sigma}^s)_{33}\}$, where the superscript s denotes the sth 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 10, we plot the posterior probability of a positive indicator for each (n, β) combination. In both plots, we use 5,000 posterior samples for inference and discard the first 1,500 as burn-in. It is evident that the posterior 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 β. For both β 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.

Algorithm 4 Simulation of a posterior predictive data set of size n under the GMC-MA

Input: One sample of GMC posterior parameters $\boldsymbol{\theta} = (\{\pi_h\}_{h=1}^H, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \{\boldsymbol{\alpha}_h\}_{h=1}^H, \{\boldsymbol{\Delta}_h\}_{h=1}^H, \{\tilde{F}_j\}_{j=1}^r)$

Output: One posterior predictive data set of size n, $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^n$

for i in $1, \ldots, n$ **do**
  Sample cluster membership $\tilde{c}_i \sim \text{multinomial}(\{\pi_h\}_{h=1}^H)$
  Sample latent factor $\tilde{\boldsymbol{\eta}}_i \mid \tilde{c}_i = h \sim N_k(\boldsymbol{\alpha}_h, \boldsymbol{\Delta}_h)$
  for j in $1, \ldots, p^*$ **do**
    if j corresponds to a binary variable **then**
      Sample $\tilde{z}_{ij} \sim N(\sum_{t=1}^k \lambda_{jt}\tilde{\eta}_{it}, \sigma_j^2)$
      Set $\tilde{y}_{ij} = \begin{cases} 1 & \tilde{z}_{ij} > 0 \\ 0 & \tilde{z}_{ij} \leq 0 \end{cases}$
    if j corresponds to a categorical variable with q levels indexed by $\{c_1, \ldots, c_{k_q}\}$ **then**
      Sample the predictive categorical level from the associated q-dimensional multinomial, where:

$$P(\tilde{Y}_{ij} = 1 \mid -) = P(\tilde{Z}_{ij} > 0 \cap \{\tilde{Z}_{i\ell}\}_{\ell \in \{c_1,\ldots,c_{k_q}\},\ \ell \neq j} < 0 \mid -) \tag{36}$$

$$\propto 1 - \Phi(0; \sum_{t=1}^k \lambda_{jt}\tilde{\eta}_{it}, \sigma_j^2) \prod_{\ell \in \{c_1,\ldots,c_{k_q}\}, \ell \neq j} \Phi(0; \sum_{t=1}^k \lambda_{\ell t}\tilde{\eta}_{it}, \sigma_\ell^2) \tag{37}$$

    if j corresponds to a numeric variable **then**
      Sample $\tilde{z}_{ij} \sim N(\sum_{t=1}^k \lambda_{jt}\tilde{\eta}_{it}, \sigma_j^2)$
      Transform $\tilde{y}_{ij} = \tilde{F}_j^{-1}\{\psi_j(\tilde{z}_{ij})\}$

Next, we include analogous plots to Figure 4 in the main paper for each additional (n, β) combination in Figures 13-15. 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

For completeness, we include in Figure 16 simulation results for SNR = 3, noting similarly exceptional performance of the GMC-MA in terms of point estimation, interval width, and interval calibration. In addition, the shortcomings of the competitors are once again apparent

Also highlighted in Section 6.2, MICE-CART yields substantial coverage gaps in estimation of the interaction terms in the regression model of interest. We hypothesize that this is

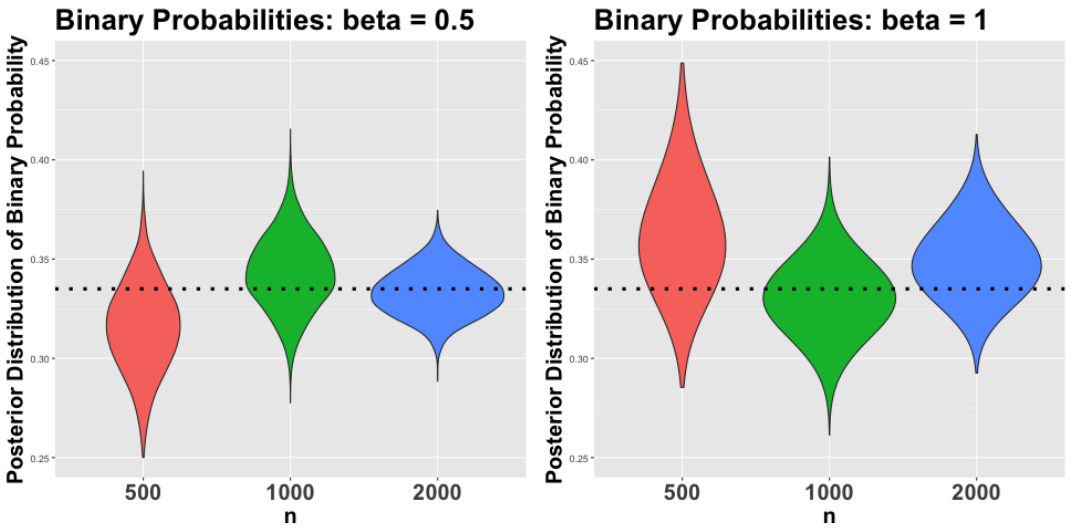

Figure 10: Posterior probabilities of a positive indicator for X_3: As n increases the uncertainty decreases and the distribution concentrates around the ground truth (dotted line)

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 17. 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.

Finally, we also present imputations under GM-RND, the Bayesian nonparametric competitor. As mentioned in the main text, the continuous treatment of ordinal variable `FI` clearly results in poor imputations, which affect the downstream regression analysis.

Appendix G. Real Data Application

In Section 7 of the main paper, we check model calibration of the GMC-MA among females. We complete the information presented in Figure 7 and include the same visual checks highlighted in Figure 18 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 7, which supports the notion that missing data may yield a biased complete case analysis.

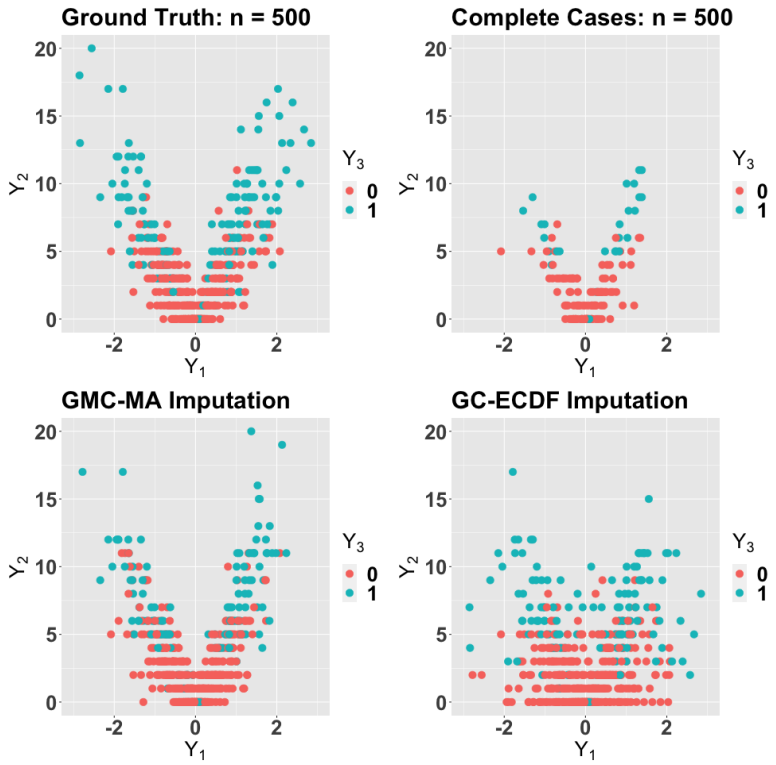

Figure 11: The same plot as Figure 4 in the main text, but for the simulation setting $n = 500, \beta = 0.5$. Here, the simulated data set without missingness is in the top-left, while complete cases are in the top-right. The proposed approach (bottom-left) is significantly better than the Gaussian copula (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.

In addition, we also include point estimates and uncertainty using the posterior predictive distribution of the 75th sample quantile in Figures 19-20, as was done in Figures 8-9 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.

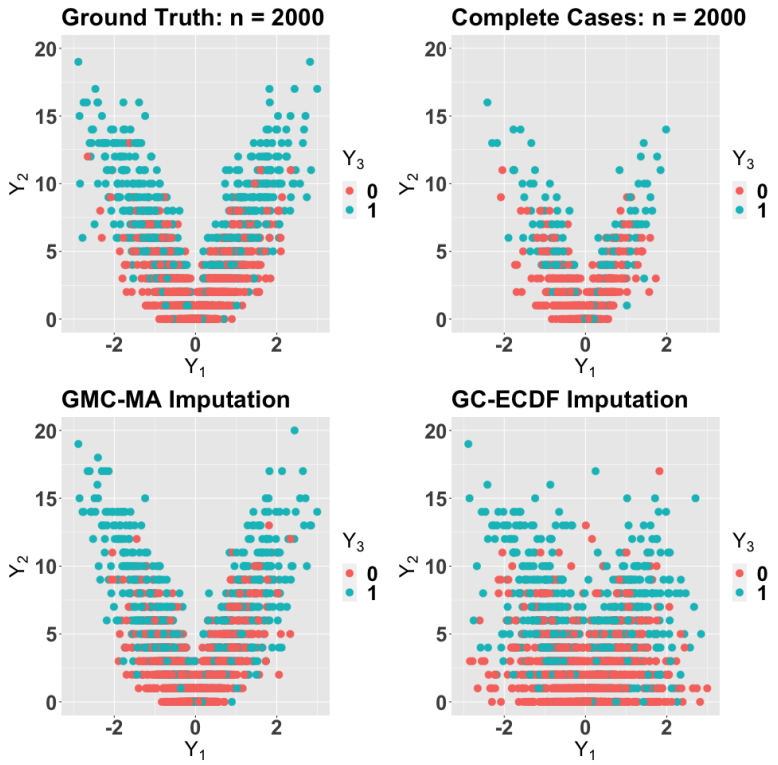

Figure 12: The same plot as Figure 4 in the main text, but for the simulation setting $n = 2000, \beta = 0.5$. Here, the simulated data set without missingness is in the top-left, while complete cases are in the top-right. The proposed approach (bottom-left) is significantly better than the Gaussian 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.

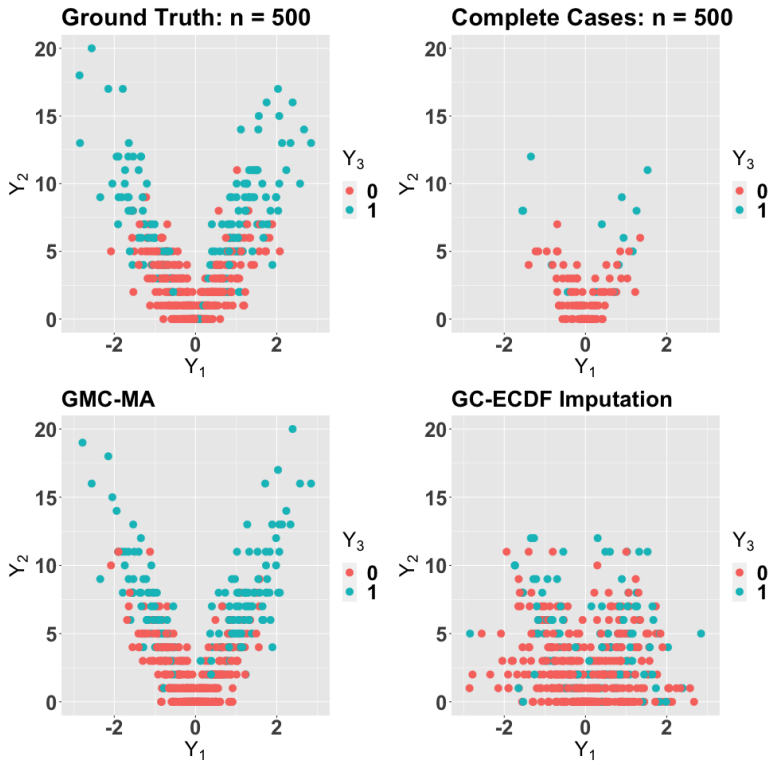

Figure 13: The same plot as Figure 4 in the main text, but for the simulation setting $n = 500, \beta = 1$. Here, the simulated data set without missingness is in the top-left, while complete cases are in the top-right. The proposed approach (bottom-left) is significantly better than the Gaussian copula (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.

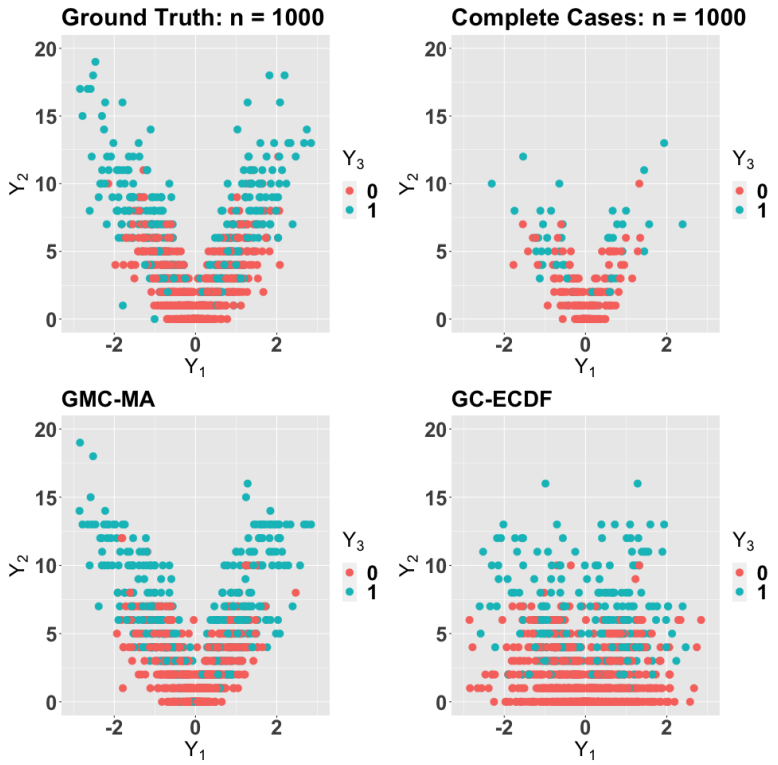

Figure 14: The same plot as Figure 4 in the main text, but for the simulation setting $n = 1000, \beta = 1$. Here, the simulated data set without missingness is in the top-left, while complete cases are in the top-right. The proposed approach (bottom-left) is significantly better than the Gaussian copula (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.

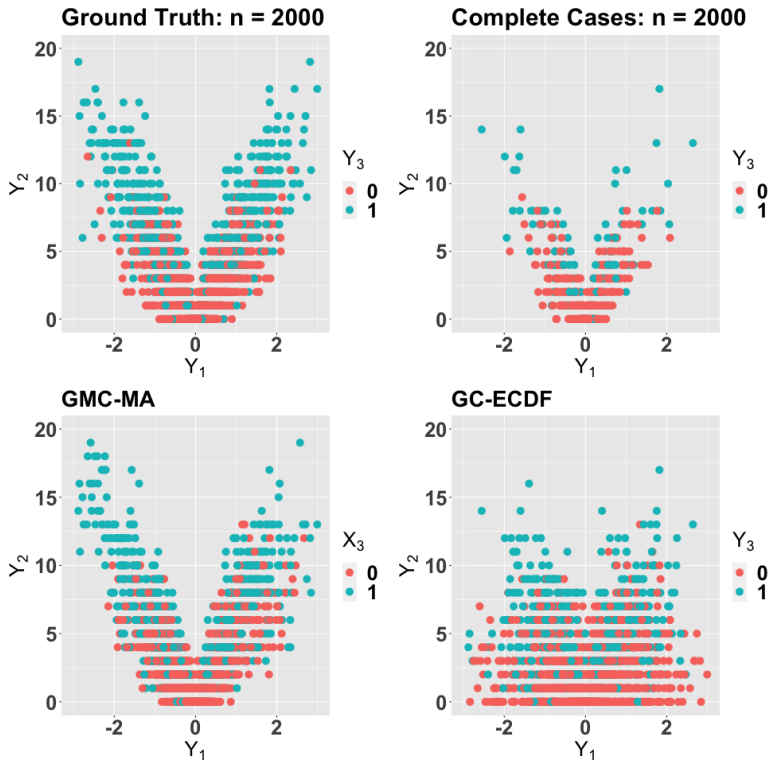

Figure 15: The same plot as Figure 4 in the main text, but for the simulation setting $n = 2000, \beta = 1$. Here, the simulated data set without missingness is in the top-left, while complete cases are in the top-right. The proposed approach (bottom-left) is significantly better than the Gaussian copula (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.

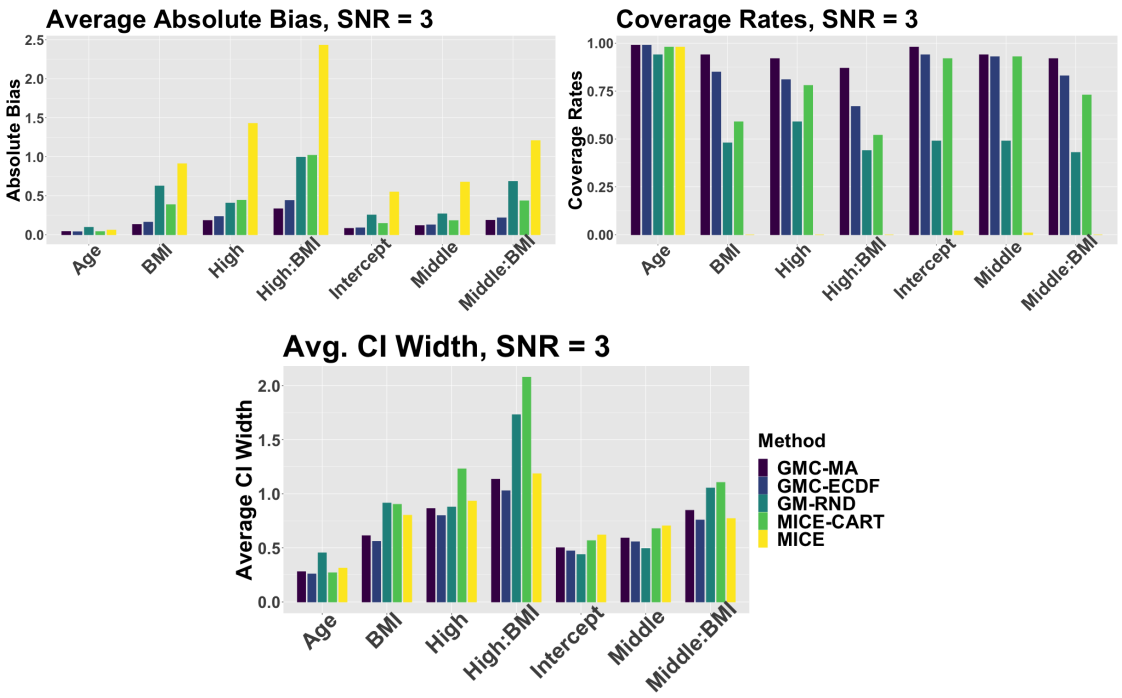

Figure 16: SNR = 3 absolute bias (top left), interval coverage rates (top right), and interval widths (center) for point and 99% interval estimates computed under each imputation method. The GMC-MA approach consistently provides the most accurate point estimates (small absolute bias), the most well-calibrated intervals (large coverage rates), and highly precise inference (small interval widths).

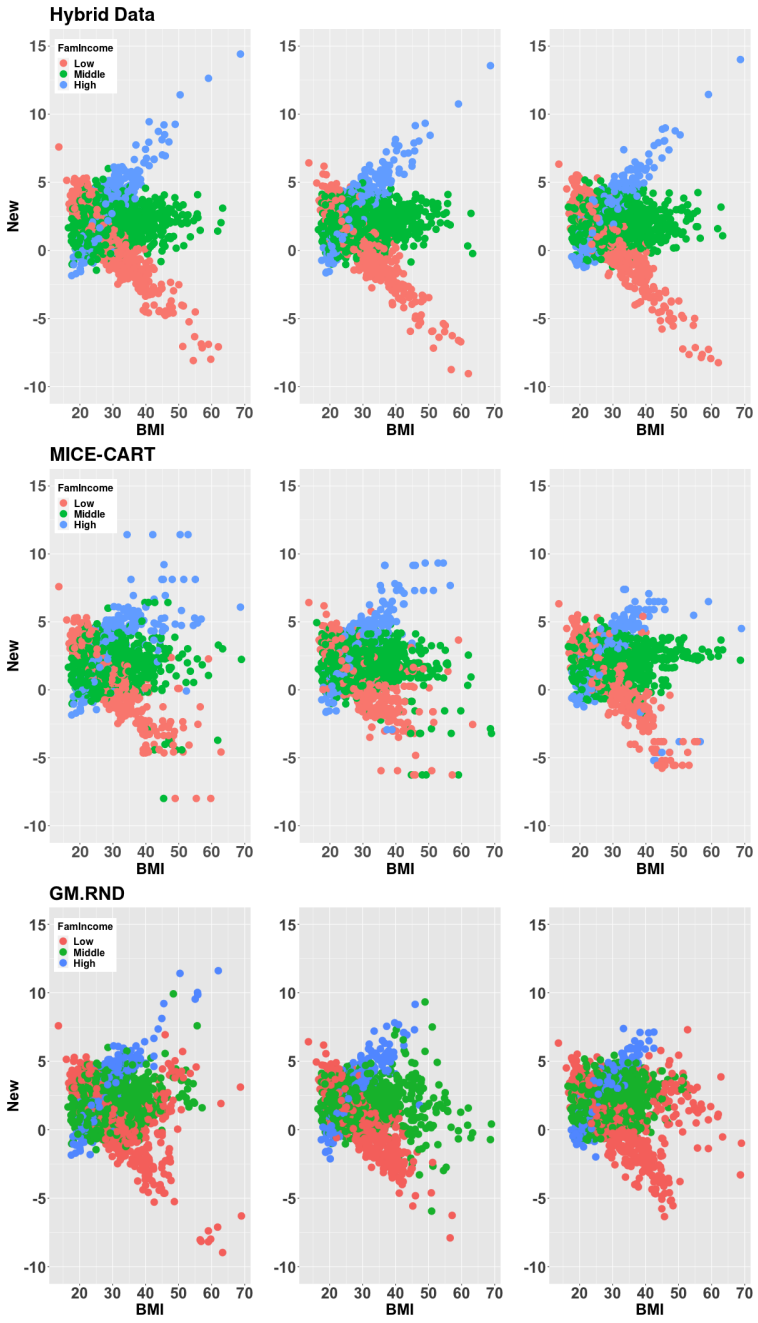

Figure 17: MICE-CART completed data sets compared to ground truth from the 1st (left panel), 50th (middle panel) and 100th (right panel) iterations of the repeated simulation study from Section 6.2. MICE with CART is unable to capture the interactive relationship between `New`, `Age`, and `BMI`, as demonstrated by misclassified `FI` at the tails of `New` and `BMI`. These classifications leverage the regression fit, yielding inaccurate estimates and uncertainty quantification for the regression model of interest. For GM-RND, the continuous relaxation of `FI` clearly leads to many erroneous imputations, which greatly biases the regression inference.

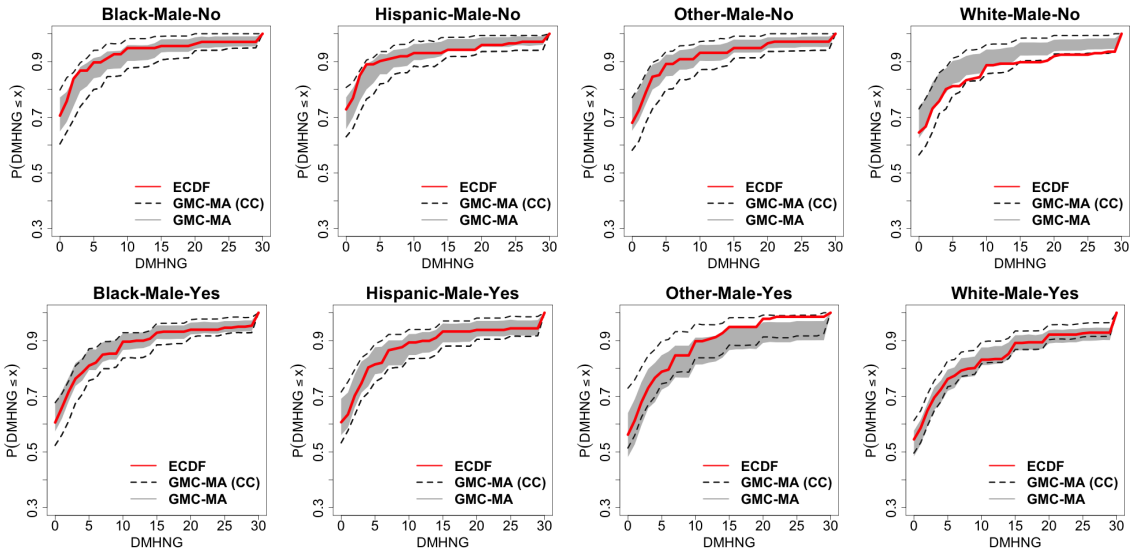

Figure 18: Posterior predictive summaries for models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). Among males, for each male race-gender-marijuana use stratum, we compare the 95% HPD intervals for the posterior predictive ECDFs, and include the ECDF on the CC data 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 MAR—and that CC analysis is unreliable.

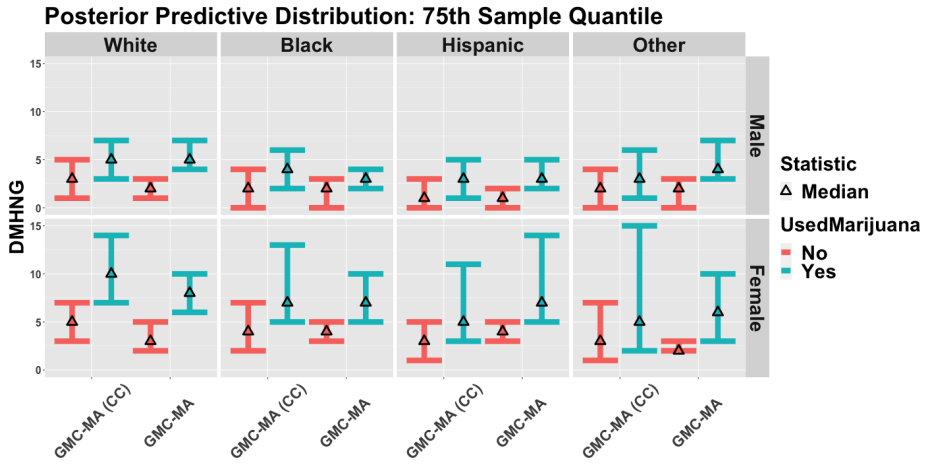

Figure 19: Posterior predictive medians and 95% HPD intervals for the predictive 75th quantiles of `DMHNG` by race-gender-marijuana use and comparing models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). The CC analysis produces wider intervals with more overlap between 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.

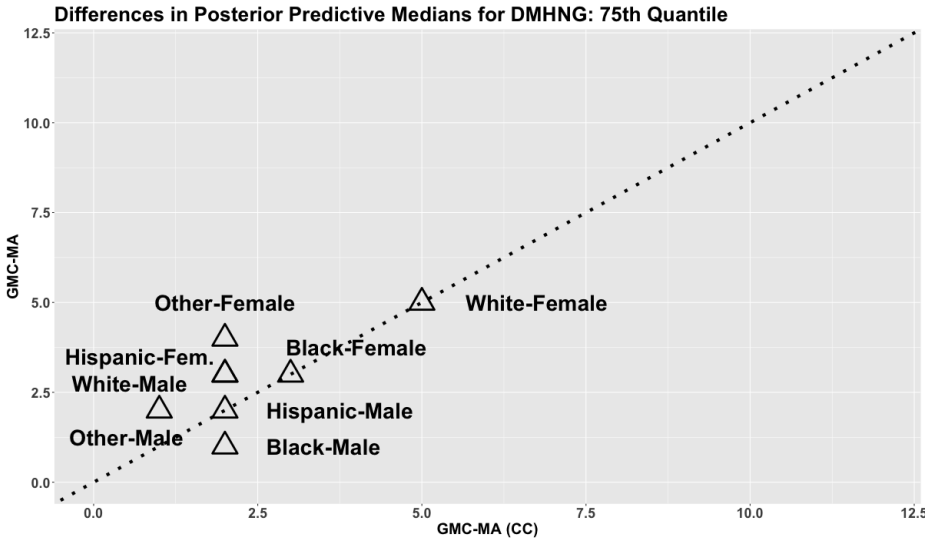

Figure 20: Difference in posterior predictive medians for the predictive 75th quantiles of DMHNG between marijuana users and non-users and comparing models fit to the complete case (CC) dataset (GMC-MA (CC)) and the full dataset (GMC-MA). The CC point estimates attenuate the differences between marijuana 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

Anirban Bhattacharya and David B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, pages 291–306, 2011.

Lane F Burgette and Jerome P Reiter. Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070–1076, 2010.

Noirrit Kiran Chandra, Antonio Canale, and David B Dunson. Escaping the curse of dimensionality in Bayesian model-based clustering. *Journal of Machine Learning Research*, 24(144):1–42, 2023.

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

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

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

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

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

Jiezhun Gu and Subhashis Ghosal. Bayesian ROC curve estimation under binormality using a rank likelihood. *Journal of Statistical Planning and Inference*, 139(6):2076–2083, 2009.

Kristin Gustavson, Tilmann von Soest, Evalill Karevold, and Espen Røysamb. 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, 2012.

Peter D Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283, 2007.

Allan V Horwitz and Teresa L Scheid. *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems*. Cambridge University Press, 1999.

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

Hemant Ishwaran and Lancelot F James. Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):508–532, 2002.

Harry Joe. *Dependence Modeling with Copulas*. CRC press, 2014.

Athanasios Kottas, Peter Müller, and Fernando Quintana. Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14(3):610–625, 2005.

Daniel R Kowal and Bohan Wu. Semiparametric count data regression for self-reported mental health. *Biometrics*, 79(2):1520–1533, 2023.

Roderick J Little. Missing data assumptions. *Annual Review of Statistics and Its Application*, 8:89–107, 2021.

Daniel Manrique-Vallier and Jerome P Reiter. Bayesian multiple imputation for large-scale categorical data with structural zeros. *Survey Methodology*, 40(1):125–135, 2014.

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

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

Jared S Murray and Jerome P Reiter. 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, 2016.

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

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

Ulf Olsson, Fritz Drasgow, and Neil J Dorans. The polyserial correlation coefficient. *Psychometrika*, 47:337–347, 1982.

Michael Pitt, David Chan, and Robert Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

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

Vaibhav Rajan and Sakyajit Bhattacharya. Dependency clustering of mixed data with Gaussian mixture copulas. In *IJCAI*, pages 1967–1973, 2016.

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

Jason Roy, Kirsten J Lum, Bret Zeldow, Jordan D Dworkin, Vincent Lo Re III, and Michael J Daniels. Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics*, 74(4):1193–1202, 2018.

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

Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons, 2004.

Matthew A Taddy and Athanasios Kottas. A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics*, 28(3):357–369, 2010.

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

Lavanya Sita Tekumalla, Vaibhav Rajan, and Chiranjib Bhattacharyya. Vine copulas for mixed data: multi-view clustering for mixed data beyond meta-Gaussian dependencies. *Machine Learning*, 106:1331–1357, 2017.

Ashutosh Tewari, Michael J Giering, and Arvind Raghunathan. Parametric characterization of multimodal distributions with non-Gaussian modes. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 286–292. IEEE, 2011.

Stef Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242, 2007.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011.

Yuxuan Zhao and Madeleine Udell. Matrix completion with quantified uncertainty through low rank Gaussian copula. *Advances in Neural Information Processing Systems*, 33: 20977–20988, 2020a.

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