Bayesian Mixture Modeling for Multivariate Conditional Distributions

Maria DeYoreo · Jerome P. Reiter

Published online: 17 June 2020
© Grace Scientific Publishing 2020

Abstract
We present a Bayesian mixture model for estimating the joint distribution of mixed ordinal, nominal, and continuous data conditional on a set of fixed variables. The modeling strategy is motivated by applied contexts in marketing and the social sciences, in particular data fusion and the analysis of stratified or quota samples. The model uses multivariate normal and categorical mixture kernels for the random variables. It induces dependence between the random and fixed variables through the means of the multivariate normal mixture kernels and via a truncated local Dirichlet process. The latter encourages observations with similar values of the fixed variables to share mixture components. We illustrate use of the model for missing data imputation, in particular data fusion of two surveys, and for the analysis of stratified or quota samples. The data fusion example suggests that the model can estimate underlying relationships in the data and the distributions of the missing values more accurately than several other approaches, including a mixture model applied to the random and fixed variables jointly. We also use the model to analyze consumers’ reading behaviors from a quota sample, i.e., a sample where the empirical distribution of some variables is fixed by design and so should not be modeled as random, conducted by the book publisher HarperCollins.

Keywords Dirichlet process · Fusion · Imputation · Latent · Missing · Mutual information

Jerome P. Reiter: This work was supported by NSF Grant SES 1131897.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42519-020-00109-4) contains supplementary material, which is available to authorized users.

Maria DeYoreo mdeyoreo@rand.org
Jerome P. Reiter jerry@stat.duke.edu

1 RAND Corporation, 1776 Main St., Santa Monica, CA 90401, USA
2 Duke University, Durham, NC 27708, USA
1 Introduction

Bayesian mixture models are flexible and convenient tools for estimating the joint distribution of a set of variables (e.g., [1, 13, 39]). They can capture complex multivariate distributional features for mixed data automatically; they can produce reliable inferences when applied on data with modest sample sizes, e.g., thousands of individuals, and they can enable analysts to quantify uncertainty in inferences, which can be nontrivial in modest sample sizes. For additional discussion of the appealing properties of mixture models, see [40, 42, 44].

Often, however, it is desirable to treat some of the variables as conditioning information rather than random variables. For example, when data are collected using a stratified or quota sampling design, the empirical distribution of the design variables is fixed \textit{a priori}. On principle, it does not make sense to estimate their distribution with uncertainty. In fact, when the sampling is not proportional to population shares, treating the design variables as random can result in badly biased estimates of population-level quantities [16, 32, 51]. As another example, when using mixture models for multiple imputation of missing data [49], it is unnecessary to estimate the marginal distribution of the variables with no missing values. Rather, all we need is the conditional distribution of variables with missingness given those that are fully observed. We note that in such contexts, analysts often seek to characterize entire multivariate conditional distributions, as these provide information on how distributional shapes and relationships change with predictors beyond changes in expected values, which may not suffice for the analysis task at hand (e.g., [14, 21, 55]).

In this article, we present mixture models for estimating the joint distribution of mixed ordinal, nominal, and continuous data variables treated as random conditional on a set of variables treated as fixed. The models use multivariate normal kernels for continuous and ordinal variables (via a probit specification), and independent multinomial kernels for nominal variables. We induce dependence between the random and fixed variables in two ways. First, for continuous and ordinal variables, we let the mean of the multivariate normal distribution within each mixture component be a function of the fixed variables. Second, we encourage observations that have similar values of the fixed variables to share mixture components via a local Dirichlet process [4]. The local Dirichlet process facilitates estimation of the dependence between the nominal random variables and the fixed variables, which is otherwise difficult to capture. It also offers the model additional flexibility to capture relationships between the random continuous/ordinal variables and the fixed variables.

Of course, there are other ways to estimate multivariate conditional distributions using Bayesian mixture models. For example, one alternative is to treat all variables as random, estimate their joint distribution via a mixture model, and derive relevant conditional distributions from the resulting estimates [11, 22, 38, 52]. However, even when the data comprise a (modest-sized) representative sample, a full mixture model can waste fitting power on the joint distribution of the fixed variables, as it seeks to fit the entire joint distribution. This can result
in poor predictive inference for conditional distributions [57]. Another alternative is to use a mixture model in which the mixture weights and/or parameters of the mixture components depend on the fixed variables (e.g., [12, 19, 48]). Models with covariate-dependent weights are more flexible than those with constant weights, but it can be difficult to estimate and obtain inferences from such models [4], which generally do not retain the DP structure marginally. Additionally, it is not clear how to incorporate dependence on multiple continuous and categorical fixed variables under most existing dependent nonparametric models.

The remainder of this article is organized as follows. In Sect. 2, we describe the Bayesian mixture model for multivariate conditional inference with mixed data, which we refer to as CMM-Mix. We also describe a variable selection procedure based on estimated mutual information values [3, 8, 15] that can be used to trim variables from the conditioning set that determines the local weights, which can be useful when the set contains many variables. In Sect. 3, we compare CMM-Mix to a full mixture model in a simulation study of techniques for data fusion [10, 17, 47], which is a type of missing data scenario common in marketing contexts. In Sect. 4, we analyze data from a quota sample from HarperCollins Publishers, in which we seek to understand relationships involving individuals’ reading behaviors and interests; for example, what distinguishes people who own eBooks from those who do not? In Sect. 5, we conclude with future directions for research. This article is accompanied by supplementary material that presents additional results from simulations and data illustrations.

2 Methodology

Suppose that the collected data include \( p \) variables that the analyst treats as random and \( q \) variables that the analyst treats as fixed. For \( i = 1, \ldots, n \), let \( Y_{ij}^{(R)} \in \{1, \ldots, k_j^{(R)}\} \) be the value of ordinal random variable \( j \) for individual \( i \), for \( j = 1, \ldots, p_o \); let \( X_{ij}^{(R)} \in \{1, \ldots, d_j^{(R)}\} \) be the value of nominal random variable \( j \) for individual \( i \), for \( j = 1, \ldots, p_n \); and let \( Z_{ij}^{(R)} \) be the standardized value of continuous random variable \( j \) for individual \( i \), for \( j = 1, \ldots, p_c \). To facilitate modeling, we introduce a latent continuous random variable \( W_{ij}^{(R)} \) for each \( Y_{ij}^{(R)} \). Similarly, for each \( i \), let there be \( q_o \) ordinal fixed variables \( Y_{ij}^{(F)} \in \{1, \ldots, k_j^{(F)}\} \), \( q_n \) nominal fixed variables \( X_{ij}^{(F)} \in \{1, \ldots, d_j^{(F)}\} \), and \( q_c \) standardized continuous fixed variables \( Z_{ij}^{(F)} \). We write each individual’s data as the vector \((Y_{ij}^{(R)}, Z_{ij}^{(R)}, X_{ij}^{(R)}, Y_{ij}^{(F)}, Z_{ij}^{(F)}, X_{ij}^{(F)})\). Writing the variables generically, we seek to construct a mixture model for \( p(Y^{(R)}, Z^{(R)}, X^{(R)} | Y^{(F)}, Z^{(F)}, X^{(F)}) \). To simplify notation, we sometimes refer to the set \((Y^{(R)}, Z^{(R)}, X^{(R)})\) as \( \mathcal{R} \) and the set \((Y^{(F)}, Z^{(F)}, X^{(F)})\) as \( \mathcal{F} \).

2.1 Modeling Strategy: Connecting \( \mathcal{R} \) and \( \mathcal{F} \) via CMM-Mix

Let \( H_i \in \{1, \ldots, N\} \) be a mixture allocation variable representing the component observation \( i \) belongs to, such that \( H_i = l \) if and only if observation \( i \) belongs to
component \( l \). At the level of the data, for any individual \( i \), the model for CMM-Mix is

\[
Y_{ij}^{(R)} = l \text{ iff } \gamma_{j,l-1} < W_{ij}^{(R)} \leq \gamma_{j,l}, \quad l = 1, \ldots, k_j^{(R)}, \quad j = 1, \ldots, p_o \quad (1)
\]

\[
(W_{i}^{(R)}, Z_i^{(R)} \mid \{\beta_h\}, \{\Sigma_h\}, H_i, X_i^{(R)}, \mathcal{F}) \sim N(D(X_i^{(R)}, Y_i^{(F)}, Z_i^{(F)}, X_i^{(F)})\beta_{H_i}, \Sigma_{H_i}) \quad (2)
\]

\[
(X_i^{(R)} \mid \{\psi_h\}, H_i) \sim \prod_{j=1}^{p_n} \text{categ}(\psi_{H_i,1}^{(j)}, \ldots, \psi_{H_i,d_i^{(F)}}^{(j)}). \quad (3)
\]

Here, \( D(\cdot) \) is a design vector of length \( k \) encoding main effects and possibly nonlinear terms identified through exploratory data analysis as helpful for capturing local dependence, and \( \beta_h \) is a \( k \times (p_c + p_o) \) matrix of regression coefficients. The ordinal \( Y^{(R)} \) are determined from \( W^{(R)} \) such that \( Y_{ij}^{(R)} = l \) if and only if \( \gamma_{j,l-1} < W_{ij}^{(R)} \leq \gamma_{j,l} \), for \( l = 1, \ldots, k_j^{(R)} \), and fixed cutoffs \(-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,k_j^{(R)}-1} < \gamma_{j,k_j^{(R)}} = \infty \) (e.g., \([2, 7, 30]\)).

The mixture of regressions helps capture dependence between \( (Y^{(R)}, Z^{(R)}) \) and \( \mathcal{F} \). It also strengthens the connections between \( (Y^{(R)}, Z^{(R)}) \) and \( X^{(R)} \) by allowing for local dependence within components. This regression approach is related to the ANOVA-dependent Dirichlet process (DP) model of De Iorio et al. [6] and also is used by Murray and Reiter [41] in a full mixture model.

With only (1)–(3), the model has to capture dependence between \( X^{(R)} \) and \( \mathcal{F} \) through a convoluted path involving the regressions for \( (Y^{(R)}, Z^{(R)}) \). With modest sample sizes, this path seems unlikely to be up to the task. We therefore strengthen the connections between \( X^{(R)} \) and \( \mathcal{F} \), as well as between \( (Y^{(R)}, Z^{(R)}, X^{(R)}) \) and \( \mathcal{F} \), using a truncated local Dirichlet process (LDP). This provides a prior distribution for a collection of random distributions indexed by fixed variables \( \mathcal{F} \), in which units that have similar values of \( \mathcal{F} \) are assumed to share similar distributions for \( \mathcal{R} \). The dependence between distributions associated with fixed variables \( f \) and \( f' \) decreases as the distance between \( f \) and \( f' \) decreases. As described by Chung and Dunson [4], the LDP possesses many attractive properties, including retaining a marginal DP prior for any \( f \).

In addition to using the LDP, one could consider replacing (3) with a regression of \( X^{(R)} \) on \( \mathcal{F} \), akin to (2). This can add significant computational challenges with MCMC sampling. For example, multinomial logistic regressions for \( X^{(R)} \) within components can introduce a large number of regression parameters for which there are no conjugate priors. We note that when \( \mathcal{F} \) contains only continuous variables, the strategy proposed by Linderman et al. [33] for efficiently sampling dependent multinomial draws could be used.

Each \( H_i \) arises from a subset of \( \{1, \ldots, N\} \), where the particular subset is chosen according to the value of \( \mathcal{F} \) for observation \( i \). The truncated LDP is built from sequences of mutually independent random variables: \( \{V_h : h = 1, 2, \ldots, N\} \) are beta(1, \( \alpha \)) distributed random variables that determine the stick-breaking weights, \( \{\theta_h = (\beta_h, \Sigma_h, \psi_h) : h = 1, 2, \ldots, N\} \) are the atoms, and \( \{\Gamma_h : h = 1, 2, \ldots, N\} \) are
locations in $\mathcal{S}$, the sample space of $\mathcal{F}$. The mixture weights and probabilities associated with an observation having a particular $\mathcal{F} = f$ are determined by the set of $\Gamma_h$ that are in some neighborhood of $f$. Specifically, let $\eta_f = \{ h : d(f, \Gamma_h) \leq d^* \}$ be a set indexing the locations belonging to the $d^*$-neighborhood of $f$. Here $d$ is a distance measure and $d^*$ represents the neighborhood size. Elements $\eta_f$ of $\{ V_h \}$ and $\{ \theta_h \}$ are then used in constructing $G_f$, the random distribution associated with $f$. Chung and Dunson [4] provide a useful graphical representation of the LDP in Fig. 1. The prior on the mixture components and weights is therefore:

$$
H_i \mid V, \Gamma, \mathcal{F}_i = f_i \overset{\text{ind}}{\sim} \sum_{h=1}^{N(f_i)} p_h(f_i) \delta_{\chi_h(f_i)}(\cdot), \quad i = 1, \ldots, n
$$

$$
V_h \mid \alpha \overset{i.i.d.}{\sim} \text{beta}(1, \alpha), \quad h = 1, \ldots, N
$$

$$
\Gamma_h \overset{i.i.d.}{\sim} p(\Gamma_h), \quad h = 1, \ldots, N
$$

where $p_h(f_i) = V_{\chi_h(f_i)} \prod_{j \in \eta_f} (1 - V_{\chi(f_j)})$ for $h = 1, \ldots, N(f_i) - 1$, $\chi_h(f_i)$ is the $h$th-ordered index in $\eta_f$, and $N(f_i) = |\eta_f|$. The last element in each probability vector $p_{N(f_i)}(f_i)$ is determined so that $\sum_{h=1}^{N(f_i)} p_h(f_i) = 1$. Although a global truncation level of $N$ is fixed to a large value to avoid artificially truncating the mixture (see Ishwaran and Zarepour [25] and Ishwaran and James [24] for results related to truncating the DP, and Chung and Dunson [4] for advice specific to the LDP), $N(f_i)$ is still a random quantity. Thus, the number of components in each mixture distribution is data driven. For $p(\Gamma_h)$, we use a product of independent uniforms for each variable in $\mathcal{F}$ to ensure the random locations are well dispersed and cover the predictor space. We discuss the choice of distance function in Sect. 2.3.

The base distributions from which the atoms $\theta_h$ are drawn are given by:

$$
\beta_h \overset{i.i.d.}{\sim} \text{MN}_{k \times (p_o + p_c)}(\beta_0, \text{diag}(\tau_1, \ldots, \tau_k), I_{p_o + p_c}), \quad h = 1, \ldots, N
$$

$$
\Sigma_h \overset{i.i.d.}{\sim} \text{IW}(\nu, S), \quad h = 1, \ldots, N
$$

$$
\psi_h \overset{i.i.d.}{\sim} \text{Dirichlet}(a_{1j}^{(j)}, \ldots, a_{dj}^{(j)}), \quad h = 1, \ldots, N, j = 1, \ldots, p_n
$$

where $\text{MN}_{k \times (p_o + p_c)}$ denotes a matrix-normal distribution of dimension $k$ by $p_o + p_c$. This implies that $\text{vec}(\beta_h) \sim N_{k(p_o + p_c)}(\text{vec}(\beta_0), I_{p_o + p_c} \otimes \text{diag}(\tau_1, \ldots, \tau_k))$, where $\text{vec}(\beta_h)$ denotes the vectorization of $\beta_h$, obtained by stacking its columns.

The model is completed with hyperpriors on the parameters $\beta_0$, $S$, $\tau$, and $\alpha$. Prior specification is discussed in Appendix A.1, including justification for the choice of the base distribution for $\beta_h$.

### 2.2 Model Properties

To describe the properties of CMM-Mix, it is useful first to marginalize over the mixture allocation indicators to obtain the multivariate conditional density for $\mathcal{R}$. We have $f(W^{(R)}, Z^{(R)}, X^{(R)} \mid \mathcal{F} = f) =$

© Springer
\[
\sum_{h=1}^{N(F)} p_h(f) N(W^{(R)}, Z^{(R)}, D(X^{(R)}, f) \beta_{\pi(f)}, \Sigma_{\pi(f)}) \prod_{j=1}^{p_n} \text{categ}(X_j^{(R)}, \psi_{\pi(f)}^{(j)}).
\]

We manipulate this expression to derive statements about the multivariate conditional distributions of the random variables at any \(f\).

Marginalizing (6) over \(W^{(R)}\) and \(Z^{(R)}\), we find that the marginal distribution for \(X^{(R)}\) is a mixture of independent multinomials. In particular, we have \(\text{Pr}(X^{(R)} = \mathbf{x} \mid F = f) = \sum_{l=1}^{N(f)} p_l(f) \prod_{j=1}^{p_n} \psi_{\pi(f), x_j}^{(j)}\). Mixtures of multinomials are quite effective for modeling multivariate categorical data distributions [13]. As desired, the model can capture dependencies between \(X^{(R)}\) and \(F\), as the weights and mixture component parameters depend on \(F\).

The \(f(W^{(R)}, Z^{(R)} \mid X^{(R)}, F)\) is a mixture of multivariate normal linear regressions, with means that are functions of \((X^{(R)}, F)\) and weights that are functions of \(F\). We have

\[
f(W^{(R)}, Z^{(R)} \mid X^{(R)} = \mathbf{x}, F = f) = \sum_{h=1}^{N(f)} p_h(f) N(W^{(R)}, Z^{(R)}, D(x, f) \beta_{\pi(f)}, \Sigma_{\pi(f)}).
\]

From (7), we can integrate out \(Z^{(R)}\) and obtain the probability that \(Y^{(R)}\) takes on a particular combination of ordinal levels \((y_1, \ldots, y_{p_{\phi}})\) conditional on \(F\) and \(X^{(R)}\). This gives \(\text{Pr}(Y^{(R)} = (y_1, \ldots, y_{p_{\phi}}) \mid X^{(R)} = \mathbf{x}, F = f) = \sum_{l=1}^{N(f)} p_l(f) \int_{\gamma_{p_{\phi} \gamma_{p_{\phi}}}} \int_{\gamma_{p_{\phi} \gamma_{p_{\phi}}}} \cdots \int_{\gamma_{p_{\phi} \gamma_{p_{\phi}}}} N(W^{(R)} \mid D(x, f) \beta_{\pi(f)}^{(W)}, \Sigma_{\pi(f)}^{(W)}) dW^{(R)},
\]

where the superscript \((W)\) indicates the portion of \(D(x, f) \beta_{\pi(f)}\) and \(\Sigma_{\pi(f)}\) that corresponds to \(W^{(R)}\).

The weights in (7) and (8) reveal additional flexibility that results from using the LDP. Since the weights are dependent on \(F\), CMM-Mix is able to capture relationships between \((Y^{(R)}, Z^{(R)})\) and \(F\) beyond those implied by \(D(X^{(R)}, F)\). This additional flexibility does not extend to relationships between \((Y^{(R)}, Z^{(R)})\) and \(X^{(R)}\).

### 2.3 Specifying the Distance Function \(d(\cdot, \cdot)\) and the Value of \(d^*\)

We base \(d(\cdot, \cdot)\) on Gower’s generalized coefficient of dissimilarity [18], which is a standard dissimilarity measure for mixed data [28, 36]. The distance between two \(q \times 1\) vectors \(f\) and \(f'\), or \(d(f, f')\), is a weighted sum of the element-wise distances, \(\sum_{j=1}^{q} w_j d_j(f_j, f'_j)\). Each \(d_j\) takes values between 0 and 1, and \(\sum_{j=1}^{q} w_j = 1\) with each \(w_j \geq 0\). For ordinal values, \(d_j(f_j, f'_j) = \max(0, 1 - f_j / f'_j)\). For continuous values, \(d_j(f_j, f'_j) = |f_j - f'_j| / \text{range}(f_j)\). For nominal values, we use the Hamming distance, \(d_j(f_j, f'_j) = 1\) when \(f_j \neq f'_j\), and \(d_j(f_j, f'_j) = 0\) when \(f_j = f'_j\). Hence, values of \(d(f, f')\) near zero indicate similar \((f, f')\), and values near one indicate otherwise.

When \(F\) comprises survey design variables or when \(q\) is modest, we set \(w_j = 1/q\) for \(j = 1, \ldots, q\). This default assignment gives each variable equal weight in
determining similarities. However, one can improve computational efficiency and possibly inferential accuracy by using only a subset of variables in $d(f,f')$, that is, by setting $w_j = 0$ for some $j$. In particular, it may be beneficial to set $w_j = 0$ for variables in $F$ that do not contribute meaningfully to predicting $X^{(R)}$, since the primary function of the LDP is to connect $X^{(R)}$ and $F$. Additionally, one can set $w_j = 0$ for one or more $F$ variables that are highly predictive of one another. In the data fusion example of Sect. 3, updating $\Gamma_j$ for $j = 1, \ldots, N$, takes 27% more time when six variables are used in $d(f,f')$ than when two variables are used. There is also some indication that inference can be made more accurate by using a parsimonious set of variables.

To perform feature selection [20] and determine which $w_j = 0$, we rely on values of mutual information, which can be used to describe dependencies between random variables of any type without assumptions about the nature of their underlying relationships [3]. For generic discrete random variables $A$ and $B$, their mutual information is $I(A,B) = \sum_a \sum_b p(a,b) \log(p(a,b)/p(a)p(b))$; values near zero indicate only weak dependence between $A$ and $B$. In our context, the value of $I(F_i,X_j^{(R)})$ for any $(i,j)$ is, intuitively speaking, the amount of uncertainty in $X_j^{(R)}$ that is explained by $F_i$. We estimate these quantities to derive a single measure of the simultaneous explanatory power of $F_i$ on $X_j^{(R)}$, namely $I_{\max}^{F_i,x_j} = \max\{I(F_i,X_1^{(R)}), \ldots, I(F_i,X_{p-1}^{(R)})\}$. Based on the estimated values of $I_{\max}^{F_i,x_j}$, we use a forward selection procedure to select the set of variables in $F_i$ having $w_j > 0$, i.e., those deemed to have explanatory power beyond a threshold, and let the complementary set have $w_j = 0$. Details of the algorithm are in Appendix A.2.

We still include variables with $w_j = 0$ in the regression for $(W^{(R)}, Z^{(R)})$. If not, we would be forcing these variables to be conditionally independent of the random variables. Even when mutual information values suggest weak dependence with $X_j^{(R)}$, $F_i$ still may be predictive of $(W^{(R)}, Z^{(R)})$. We note that including irrelevant variables in $D(X^{(R)}, F)$ is not problematic computationally (when $q$ is not huge), so that we prefer not to force conditional independence a priori.

We also must specify the neighborhood size $d^*$. As a benchmark, consider what happens when $d^* \rightarrow 1^{-}$. In this case, all $\Gamma_i$ are in the neighborhood of any value of $F_i$, so that $H_i$ in (4) is simply drawn from $\{1, \ldots, N\}$ as in the usual Bayesian mixture model. Hence, to facilitate sharing of components by records with similar values of $F_i$, we seek a $d^*$ away from one. One approach is to determine a $d^*$ so that each observation belongs to the same neighborhood as $r\%$ of the observations on average, for instance $r = 20$ [4]. Alternatively, we can base $d^*$ on interpretations of the distance function. For instance, with all nominal $F_i$, setting $d^* = 0.5$ implies that observations must exactly match on at least 50% of values to be in the same neighborhood and share mixture components. In our data applications, the results were insensitive to different reasonable values of $d^*$.

### 2.4 Posterior Inference and Missing Data Considerations

For posterior inference, we use a Gibbs sampler based on the finite stick-breaking representation of the DP [24]. With missing values, the sampler proceeds via data
augmentation, i.e., given a draw of the parameters, we draw new values of the missing data. We present the posterior full conditionals in Appendix A.3. After MCMC convergence, analysts can use the completed datasets for multiple imputation inferences [23, 49] or directly make posterior inferences from relevant functions of the parameters.

Missing data can cause numerical problems in the algorithm, namely that elements of $\beta_h$, $\Sigma_h$, and $\tau$ can diverge toward very large values. In particular, problems occur when the mixture model samples clusters with all observed values of some ordinal variable equal to the first or last category, e.g., $Y_{ij}^{(R)} = 1$ for all observed $Y_{ij}^{(R)}$ in some cluster. This creates a perfect prediction problem and ensuing estimation difficulties. In our example, the mixture model will strongly favor imputing category 1 for the missing values of $Y_{ij}^{(R)}$ in the cluster. Hence, it will favor making the corresponding latent $W_{ij}$ very negative, which happens when elements of $\beta_h$ are large in magnitude. This in turn can cause some elements of $\tau$ to become extremely large. Related problems arise with other types of perfect predictions within clusters. Perfect prediction is well known to cause problems in maximum likelihood estimation and divergence in regression coefficients in probit regression (e.g., [34]). We note that similar problems can arise in any mixture model (not just CMM-Mix) that uses probit specifications for ordinal variables.

We use a fix that prevents elements of $\beta_h$ from getting too large. Given the standardization of all continuous variables and choice of cutoffs for the latent continuous random variables, practically we need not allow elements of $|\beta_h|$ to exceed 4 or 5, as this still allows the average value of each $Y_j^{(R)}$ to be as small as 1 or as large as $k_j^{(R)}$. We therefore restrict $\tau_j^2 \leq 6$ via a truncated inverse-gamma prior in place of an inverse-gamma prior. This keeps the parameters from diverging to large values without being overly restrictive, since the draws for $\tau_j^2$ are centered far to the left of the truncation value in all of our model implementations. We include plots of posterior samples for $\tau_j^2$ in Sect. 3 of the supplementary material.

### 3 Evaluating CMM-Mix Using a Data Fusion Scenario

In this section, we compare the performance of CMM-Mix and a fully joint mixture model using simulations designed to mimic typical data fusion applications in marketing. Although our main goal is to compare conditional and joint mixture models, we also compare CMM-Mix with three standard predictive modeling strategies that can be used for data fusion, namely nearest neighbor matching, generalized linear models (GLMs) and random forests. Our purpose here is simply to explore whether using CMM-Mix can have advantages over default methods for data fusion. We emphasize that our goal is not to find some best approach for data fusion or other conditional modeling scenarios, as the best model is surely problem-specific. We begin with a brief review of data fusion.
3.1 Review of Data Fusion

Suppose that a common set of demographic variables $A$ is available in two databases, $D_1$ and $D_2$; a set of variables $B_1$ measuring opinions about purchasing products is available only in $D_1$; and a set of variables $B_2$ assessing behaviors relevant for advertising is available only in $D_2$. The marketing company seeks to use $D_1$ and $D_2$ to learn about the joint distribution of $(B_1, B_2)$ conditional on $A$ variables, e.g., to connect purchasing opinions to behaviors. To do so, the company can treat data fusion as a missing data problem, using some predictive model to fill in the intentionally missing values of $B_1$ ($B_2$) for individuals in $D_2$ ($D_1$). Of course, often some individuals in $D_1$ have values of $B_1$ unintentionally missing, and likewise for $D_2$. They also might have values of $A$ missing in both files, although our focus here is on settings where $A$ is complete (or can be completed with a small number of judicious imputations), which is typically required for marketing data fusion to proceed.

This is a challenging prediction problem. Without simultaneous observations of $\{A, B_1, B_2\}$, the company cannot learn the multivariate conditional distribution from the data alone; it must make additional, identifying assumptions about the conditional associations between $B_1$ and $B_2$ given $A$ [29]. The simplest and most common assumption is that $B_1$ and $B_2$ are conditionally independent given $A$. This allows the company to specify flexible models for the conditional distributions $f(B_1 | A)$ and $f(B_2 | A)$ and use these to impute the intentionally missing values.

In marketing, data fusion models often are based on some version of nearest neighbor matching [56, 60]; for example, for every record in $D_1$, find the record in $D_2$ with $A$ values as similar as possible to its own $A$ values according to some distance metric, and use this record’s values of $B_2$ as imputations. Alternatively, one can form the completed datasets by using machine learning algorithms, such as random forests or variants of neural networks, as imputation engines. These methods, however, may not be ideally suited for typical marketing data fusion applications. Many learning algorithms work best with large training sets, which is not the case in our applications of interest. Further, if applied to data fusion, some learning algorithms would generate a single prediction for each missing value. It is well known in the missing data literature [49] that inferences based on a single imputation underestimate uncertainty, which may not matter in large databases but can matter in modest sample sizes. Finally, with some algorithmic approaches, it may not be obvious how to accommodate unintentional item missingness and intentional missingness from the data fusion simultaneously in a principled way.

One way to handle all of these challenges is to concatenate $D_1$ and $D_2$, treat $(A, B_1, B_2)$ as random variables, and use a Bayesian mixture model (e.g., [1, 13, 39, 42]) as the imputation engine (e.g., [16, 27]). One uses the mixture model to estimate the joint distribution, $f(A, B_1, B_2)$, from which one derives $f(B_1 | A)$ and $f(B_2 | A)$ needed to impute missing values in $B_1$ and $B_2$. This is an application of the joint mixture modeling approach to regression, originally proposed by Müller et al. [38] for a continuous response with continuous covariates and later extended to other scenarios (e.g., [22, 52]). However, we care about estimation of $f(B_1, B_2 | A)$, not $f(A, B_1, B_2)$. Thus, as noted in Sect. 1, it may be beneficial to treat the common variables $A$ as conditioning information rather than random variables.
3.2 Constructing the Data Fusion Scenario

We base the simulations on data from a survey of \( n = 3567 \) individuals collected by the book publisher HarperCollins. We treat 11 variables from the survey as \( A \) variables, including six ordinal variables (age, passion for books, reading hours, income, passion for Internet, and opinion on romance in books) and five nominal variables (work status, importance that books that challenge the reader, laptop ownership, eBook reader ownership, use of audiobooks). To construct a data fusion scenario where conditional independence is known to hold, we generate a continuous variable \( Z \), an ordinal variable \( Y \), and a nominal variable \( X \) such that 
\[
f(Z, Y, X | A) = f(Z | A)f(Y | A)f(X | A).
\]
We use normal, probit, and multinomial logistic regressions to generate \( Z, Y \), and \( X \), respectively, using various main effects and interactions involving only \( A \) in the predictor functions. The exact specifications are presented in Sect. 2 of the supplementary material. After binding these simulated variables to \( A \), we blank \((X, Z)\) for the first 1189 rows, \((X, Y)\) for the second 1189 rows, and \((Y, Z)\) for the final 1189 rows, to get three datasets \( D_1, D_2, \) and \( D_3 \). This simulates a data fusion scenario with three databases. This process is repeated 40 times to create 40 unique sets of databases to be fused, which results in Monte Carlo standard errors that are adequately small for our purposes.

3.3 Results

We consider the 11 \( A \) variables as \( \mathcal{F} \), and \((Z, Y, X)\) as \( \mathcal{R} \). In one randomly chosen set of simulated \( \mathcal{R} \), the (normalized) values of \( I_{\text{max}}^{\text{Fl},x} \) to two decimal places in descending order are (.25, .23, .07, .07, .06, .06, .02, .02, .01, .01, .00); other replications yield similar values. Given the number of weak associations, we investigate two possibilities for assigning nonzero weights in \( d(\cdot, \cdot) \). First, we allow only the top two scoring variables—reading hours (ordinal with five categories) and laptop ownership (binary)—to have \( w_j > 0 \); this includes only moderately strong predictors. Second, we allow the top six scoring variables—ordinal variables age, passion for books, and income, and binary variable desire for challenge in reading—to have \( w_j > 0 \); this includes variables that are weakly predictive. The maximum normalized mutual information between any \((\mathcal{F}_i, \mathcal{F}_j)\) pair is 0.27, corresponding to passion for books and reading hours per week, suggesting no serious redundancies among \( \mathcal{F} \).

For each feature selection, we consider three sets of possible \( d^* \) values. For the two feature model, we consider \( d^* \in (0.5, 0.25, 0.125) \). Here, 0.5 is the maximum distance between two individuals’ vectors of reading hours and laptop ownership when only one of the two variables matches. The distance is 0.125 when the laptop ownership matches and reading hours is separated by just one category. This is the most stringent criterion besides requiring that both variables must match exactly, which would set \( d^* = 0 \). The 0.25 is an intermediate value. These \( d^* \) values imply, respectively, that each observation is in the same neighborhood
as 66%, 48% and 34% of the other observations on average. For the six-variable feature selection, we consider \(d^* \in (0.375, 0.30, 0.25)\). These \(d^*\) values imply that each observation is in the same neighborhood as 52%, 34%, and 23% of the other observations on average.

We fit the six versions of CMM-Mix to each concatenated dataset on \{\(Y, Z, X, A\)\} containing missing values, letting \(D(\cdot)\) contain only main effect terms. Thus, although the data-generating mechanism included interactions within \(A\), we do not incorporate this structure into our model, relying on the flexibility provided by the mixture to accommodate such interactions. We refer to the two-feature models as C-2S, C-2M, and C-2L, where the S, M, and L indicate the smallest to largest values of \(d^*\), respectively. We refer to the six feature models as C-6S, C-6M, and C-6L using analogous nomenclature.

For comparisons, we include a fully joint mixture model that puts all 14 variables in \(\mathcal{R}\). In this fully joint mixture model, (1) does not contain \(\mathcal{F}\), and \(H_i\) simply arises from \([1, \ldots, N]\) with probabilities \((p_1, \ldots, p_N)\), with \(p_i = V_i \prod_{r=1}^{i-1}(1 - V_r)\). In other words, the fully joint mixture model is a standard DP mixture model, with multivariate normal and multinomial kernels. For the GLM-based approaches, we fit an ordered probit regression to \(D_1\) to estimate \((Y \mid A)\), a linear regression to \(D_2\) to estimate \((Z \mid A)\), and a multinomial logistic regression to \(D_3\) to estimate \((X \mid A)\). We fit two different sets of GLM specifications: one that contains only main effects for \(A\) and another that contains all main effects as well as two-way interaction terms. We refer to these models as GLM and GLM*, respectively. These can be thought of as reasonable default specifications of basic but commonly used imputation models. We also impute values via random forests (RF), again learning \((Y \mid A)\) from \(D_1\), \((Z \mid A)\) from \(D_2\), and \((X \mid A)\) from \(D_3\), thus assuming conditional independence in the predictive models, an assumption that matches the data generating mechanism. For each method, we create ten completed datasets for use in multiple imputation inferences based on the combining rules of Rubin [49].

We begin with inferences for all 285 cell probabilities associated with the bivariate distributions of \((X, A_j)\) and \((Y, A_j)\) for all \(j\). For each of the 40 simulated datasets, we calculate the proportion of multiple imputation 95% confidence intervals that contain their corresponding ground truths. We also calculate summaries of the absolute errors of the point estimates, including the mean, 25th percentile, and 75th percentile.

**Table 1** Summary of inference for cells corresponding to bivariate discrete distributions with missing data

|                | C-2S | C-2M | C-2L | C-6S | C-6M | C-6L | Joint | GLM   | GLM*  | RF  |
|----------------|------|------|------|------|------|------|-------|-------|-------|-----|
| % CIs covering | 91   | 81   | 80   | 85   | 84   | 79   | 66    | 64    | 67    | 55  |
| Mean Abs. Err. | .74  | .66  | .81  | .66  | .74  | .87  | 1.28  | .92   | .82   | 1.30|
| 25%ile Abs. Err.| .16  | .12  | .11  | .13  | .13  | .13  | .25   | .18   | .18   | .27 |
| 75%ile Abs. Err.| .97  | .86  | .90  | .89  | .91  | 1.00 | 1.61  | 1.18  | 1.01  | 1.67|

Entries include the proportions of 95% multiple imputation confidence intervals (CIs) that contain the truth, and the mean, 25th percentile, and 75th percentile of absolute errors, averaged over all 40 datasets.
As evident in Table 1, the joint model, GLMs and random forests result in poor coverage rates. Coverage is largest under C-2S. The CMM-Mix specifications with the lowest coverage still perform better than all comparison methods. Differences are unlikely due to Monte Carlo error, as the standard errors of reported quantities across the 40 sets of results are relatively small; for example, the estimated standard errors of all coverage rates for the conditional models are less than .006, and the standard error of the coverage rates for the joint model is .016. The average absolute errors are highest under random forests and the joint mixture model, followed by the main effects only regression model (GLM). The GLM with interaction effects (GLM∗) has average errors that are similar to the worst performing CMM-Mix models (C-2L and C-6L).

Among CMM-Mix models, those with small \(d^\ast\) values result in the largest coverage rates. The absolute errors are generally smallest under the small and medium \(d^\ast\) values, but all CMM-Mix models tend to be more accurate than the joint model. The models with small \(d^\ast\) outperform those with large \(d^\ast\), highlighting the benefits of dependent cluster assignments. The two-feature models have larger coverage than the six-feature models with the exception of those with medium \(d^\ast\) values, and the two-feature models have smaller mean absolute errors than the six-feature models, with the exception of the small \(d^\ast\) values.

Turning to \(Z\), inferences for its mean, 25th percentile, and 75th percentile are more accurate under the CMM-Mix models than the joint model. Almost all 95% CIs from the conditional models contain the true mean zero, with all average point estimates less than .027 in absolute value. By comparison, the joint model produces an average point estimate of .093, and more than half of the 95% CIs do not contain zero. The models that use six features are overall less accurate than those that use two features, as shown in Sect. 2 of the supplementary material. The GLM-based models have point estimates that are even worse than the joint mixture model. Random forests perform similarly to CMM-Mix, giving accurate estimates for the mean, 25th, and 75th percentiles of \(Z\).

We estimate \(E(Z \mid A_j)\) for each value of \(A_j\) for all \(j\). Once again, the joint model results in noticeably larger errors than the CMM-Mix models. The average absolute errors of the 46 point estimates from the joint model versus C-2S are shown in Fig. 1; figures for other versions of CMM-Mix are in Sect. 2 of the supplementary material. The mean of these average absolute errors is .183 under the joint model, while the CMM-Mix models produce values ranging from .030 to .048. The means under the alternative methods are as follows: .186 under GLM, .159 under GLM∗, and .053 under random forests.

Taken together, these results suggest that CMM-Mix provides imputations that are closest to the truth. We believe that performance issues for the GLM methods stem from lack of flexibility in the parametric model. We believe that the performance issues for the random forest method stem from the weak predictive power of the method for these modest-sized samples.

While the CMM-Mix models seem to outperform the joint mixture model on these dimensions, we also can use the simulation study to assess which of the two mixture models is most faithful to the conditional independence assumption for data fusion. This investigation is motivated by remarks of Gilula et al. [17], who
questioned whether or not mixture models for data fusion, as proposed by Kamakura and Wedel [27], truly encode conditional independence. In fact, Fosdick et al. [16] found in simulations that a standard mixture model generates stronger estimated associations than implied by conditional independence, but the reason for this was not clear. We focus here on one randomly selected simulation run for simplicity in presentation of results. For each of the ten completed datasets, we fit a regression model of $Z$ conditional on $X$ and $Y$, as well as all variables in $A$ that actually generated the data. We use multiple imputation inference to create point estimates and 95% confidence intervals for all regression coefficients. We also fit the model to the fully observed data without any missing values, referred to as the premissing data, and obtain the true empirical 95% confidence intervals for regression coefficients.

With the full joint model, the 95% confidence intervals for three of the four regression coefficients for $X$ and $Y$ do not contain the true value of zero, whereas with all versions of CMM-Mix and the premissing data all intervals contain zero. All models result in confidence intervals that exclude zero for ten of the 11 nonzero coefficients involving $A$. The average absolute error of the regression coefficient estimates from the joint model is .252, whereas the averages are less than .08 for all CMM-Mix models. Eleven of the 16 regression coefficient confidence intervals from the joint model do not even overlap with their corresponding confidence intervals based on the premissing data; in contrast, all sixteen confidence intervals from CMM-Mix fully contain their corresponding confidence intervals based on the premissing data. We find similar results when fitting regression models that treat $Y$ or $X$ as the response. We also validate that the joint model is less consistent with conditional independence assumptions than the CMM-Mix models using the

![Fig. 1 Average absolute errors of 46 multiple imputation point estimates for the mean of $Z$ conditional on each $A$ variable from the joint model and C-2S](image_url)
strategy of Kunihama and Dunson [31] based on conditional mutual information. Details of all additional results are in Sect. 2 of the supplementary material.

Finally, we investigate the performance of statistical matching [56, 60], a common technique used for data fusion based on a variant of nearest neighbor matching. Here, we create fused datasets via exact matching on the 11 $A$ variables using the StatMatch package in R [9]. For each record $i$ with missing $Y_i$, this method identifies all observations $\{i' \neq i : Y_{i'}$ is observed $\}$ with the smallest Hamming distance from $A_i$, and samples one of their $Y$ values as an imputation for the missing $Y_i$. It operates similarly for observations missing $X$ or $Z$. Using this method, on average, only 65% of confidence intervals associated with bivariate probabilities contain the true values. Thus, the statistical matching approach is not as effective as the conditional mixture model approaches.

In summary, the results suggest that the imputations for the variables with missing data from the joint mixture model are not consistent with conditional independence, whereas those from the CMM-Mix models are. Additionally, the CMM-Mix models estimate the relationships in the data-generating model as or more reliably than the joint model. There is evidence of potential for inferential gains by choosing smaller $d^*$ and a more parsimonious set of features driving the cluster assignments. In Sect. 1 of the supplementary material, we report results from a separate empirical study where the goal is to impute arbitrarily missing items rather than the data fusion setup. The CMM-Mix provides accurate inferences in that study as well.

4 Conditional Inference from a Quota Sample

We now analyze a quota sample conducted by HarperCollins in 2013 to learn about US consumers’ reading behaviors and interests. The sample comprises $n = 3631$ individuals, sampled to achieve fixed quotas in various age, gender, ethnicity, and region (location) groups. We seek inferences for the relationships among income, passion for books, passion for the Internet, reading hours, laptop ownership, eBook ownership, and desire to be challenged in reading. Most variables are complete; only income and reading hours have isolated missingness, which we assume to be at random. Income, passion for books and the Internet, and reading hours are ordinal and comprise $\mathbb{I}(R)$; eBook ownership and laptop ownership are nominal and comprise $\mathbb{I}(R)$; and the measure of desire for challenging books is numerical and comprises $Z(R)$. We condition on all the design variables when fitting CMM-Mix, letting $F$ include age, gender, ethnicity, and region with all $w_j > 0$.

We seek posterior inferences about multivariate relationships among the survey variables, conditional on the variables used in the quota sampling. The underlying multivariate distribution may well be complex, e.g., containing nonstandard distributions and relationships among the continuous and ordinal variables that change with different levels of the categorical variables. Hence, for posterior inference, it is advantageous to use flexible Bayesian mixture models. Since the marginal distribution of the quota variables is fixed by design, we use the CMM-Mix model developed in Sect. 2.
We make full posterior inferences rather than multiple imputation inferences, as some subgroups (i.e., combinations of age, gender, ethnicity, region) have few or no individuals in sample. We fit CMM-Mix with $d^* = 0.25$, which implies that each observation is in the same neighborhood as 14% of the other observations on average, and observations must match on at least of three of the four design variables to be in the same cluster. Posterior predictive model checks do not suggest evidence of lack of model fit; results are in Sect. 3 of the supplementary material.

We focus inferences on individuals aged 25–34, one of the most sought after demographics in marketing [5, 35]. In particular, we contrast the characteristics of individuals in this group who own eBook readers and those who do not own eBook readers. CMM-Mix enables inferences about many other demographic groups; we report only on one here to present a concise and coherent analysis.

Figure 2 displays the distribution of income for eBook reader owners and nonowners for white females aged 25–34 living in the South. Owners of eBook readers tend to have higher incomes than nonowners, with the most striking difference in the percentage of individuals making less than $25000 per year. We obtain these posterior inferences using expression (8), integrating out the nominal random variable laptop, $X_2^{(R)}$, using the estimates for $\Pr(X_2^{(R)} = x) = \sum_{i=1}^{N(F)} p_i(F)p_{(2)}^{(2)}(F)_{\hat{\beta}}(\hat{\eta}(F),x)$. Figure 2 also displays the posterior distribution of weekly reading hours for owners and nonowners. Overall, owners of eBook readers tend to spend more hours reading than nonowners and are noticeably less likely to spend zero hours reading.

These results are in accord with prior research suggesting that higher income individuals are more likely to be frequent readers [26, 46]. Research also suggests that higher income individuals use the Internet more frequently. Building on these findings, we next contrast eBook owners’ and nonowners’ views on the importance of books and the Internet, drilling down by race to provide additional information. Figure 3 displays the probability that individuals aged 25–34 from the South regard books as very important in their lives and the probability that individuals aged 25–34 regard the Internet as very important in their lives, as functions of gender, race, and eBook reader ownership. The posterior inferences suggest that, regardless of income level, white females are most likely to view books as important, and white males are the least likely to view books as important. As might be expected, those who own eBook readers are more likely to view books as very important than those who do not own eBook readers. The posterior inferences also suggest that, regardless of income level, black males are most likely to view the Internet as important, and white females are least likely to view the Internet as important. The results suggest that the importance of the Internet tends to increase as a function of income for both eBook reader owners and nonowners, except for white females. The model suggests little to no interactions between gender and eBook ownership, and race and eBook ownership. Exploratory analysis and posterior predictive checks are consistent with these findings, indicating this is not some artifact of the model specification.

Finally, we examine whether or not eBook reader owners prefer books that challenge them to think more than those who do not own eBook readers. This involves posterior distributions of the continuous variable “challenge,” conditional on eBook
Fig. 2 Left: posterior mean and 90% credible intervals for the probability assigned to each income level for eBook owners (circle symbols) and nonowners (triangle symbols). Income levels 1 through 7 correspond to $<25K$, $25–44K$, $45–74K$, $75–99K$, $100–149K$, $150–199K$, $\geq 200K$. Right: posterior mean and 90% credible intervals for the probability assigned to each level of weekly reading hours for owners versus nonowners. Both displays refer to white females aged 25–34 living in the South.
Fig. 3 Posterior mean estimates for the probability that one views books as very important (top row) and for the probability that one views the Internet as very important (bottom row) as a function of income, gender, and race, for eBook nonowners (left column) and owners (right column). Income levels 1 through 7 correspond to < 25K, 25–44K, 45–74K, 75–99K, 100–149K, 150–199K, ≥ 200K.
reader ownership for each combination of age, gender, ethnicity, and region. Larger values of “challenge” indicate increasing appreciation for books that challenge the reader to think. The distribution
\[
f(Z^{(R)} | \mathcal{F}, X^{(R)}) = \sum_{l=1}^{N(\mathcal{F})} p_l(\mathcal{F}) N(Z^{(R)} | (D(X^{(R)}, \mathcal{F})\beta_{\xi_l(\mathcal{F})})^{(Z)}, \Sigma^{(Z)}),
\]
where the superscript \((Z)\) indicates the portion of \(D(X^{(R)}, \mathcal{F})\beta_{\xi_l(\mathcal{F})}\) and \(\Sigma_{\xi_l(\mathcal{F})}\) that corresponds to \(Z^{(R)}\). To estimate \(f(Z^{(R)} | X_1^{(R)}, \mathcal{F})\), where \(X_1^{(R)}\) refers to the random variable “eBook ownership,” we integrate out \(X_2^{(R)}\) using the estimates for \(\Pr(X_2^{(R)} = x) = \sum_{l=1}^{N(\mathcal{F})} p_l(\mathcal{F}) \psi_{\xi_l(\mathcal{F}), x}\).

Figure 4 displays mean posterior distributions of the challenge variable for four combinations of age, gender, ethnicity, and region. The shapes of the distributions are varied, with some having skewness or bimodality and others being standard unimodal distributions. For almost all combinations of age, gender, ethnicity, and region, the distribution of the challenge variable is centered on larger values for eBook reader owners than for nonowners. This suggests that those who own eBook readers tend to prefer books that are challenging as compared to those who do not own eBook readers.

All of these inferences are conditional on fixed quota variables. For unconditional inferences, for instance national estimates, analysts can integrate out \(\mathcal{F}\) over its distribution in the population. In our context, this can be accomplished by estimating the quantity of interest for each combination of age, gender, ethnicity, and region, and averaging over these estimates according to each group’s population share. For inferences conditional on only age and region, for example, we marginalize over the distribution of ethnicity and gender in the population of US individuals aged 18 and older.

5 Discussion

We conclude with a discussion of future research directions, beginning with the feature selection algorithm for the LDP. The simulations suggest that when the number of variables in \(\mathcal{F}\) is large, one can improve accuracy by selecting only a subset of the variables to include in the distance calculation. We used equal weights for features selected to be in the distance function. It may be beneficial to allow the nonzero weights to differ, for example to be proportional to the values of mutual information with the \(X^{(R)}\) variables.

The CMM-Mix models also may point an alternative path for handling survey weights in Bayesian mixture models, which is an active area of research (e.g., [32, 50, 53]). Since the survey weights are fixed and fully observed, one could include them as an \(\mathcal{F}\) variable. This would make the conditional distribution for the random variables a function of the survey weights, as is done in the Gaussian process model of Si et al. [53]. Here, observations with similar survey weights would be encouraged to share mixture components.

Our work with CMM-Mix reveals problems when using mixture models to impute missing ordinal values that can be explored further. Specifically, the mixture models sometimes favor clusters that are homogeneous in observed values, which
Fig. 4 Mean posterior distribution of the “challenge” variable given that one owns an eBook (solid) or does not own an eBook (dashed). Figures refer to four combinations of age, gender, ethnicity, and region. Left to right: 25–34, female, white, South; 25–34, male, white, Northeast; 35–44, female, Hispanic, West; and 45–54, male, other, West.
encourages imputations also to be at the observed values. This can be undesirable, as it may result in underestimation of imputation uncertainty, particularly when cluster assignments are relatively stable. It may be beneficial somehow to add uncertainty into homogeneous clusters. Related ideas are used by Paiva and Reiter [43], who adjust the weights associated with mixture components in order to generate imputations that follow a nonignorable missing data mechanism.

The data fusion and quota sampling applications that we considered involve modest-sized datasets. If one were to run CMM-Mix on much larger samples, the MCMC algorithm could be challenging to implement efficiently, as is the case for many mixture models. For such cases, one possible approach would be to develop variational approximations, learning the variational parameters of the approximated posterior through optimization (e.g., [58]). Another potential path forward is to split the sample into disjoint sets, implement MCMC independently (in parallel) on each subset, and combine these draws to obtain draws from “subset posteriors” (e.g., [37, 59]). We note that these approaches are approximations and hence may sacrifice accuracy in inferences compared to full estimation of the posterior distribution.

Acknowledgements This research was supported by the National Science Foundation [Grant SES-11-31897].

Compliance with ethical standards

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix

Prior Specification

The particular matrix-normal base distribution for the regression coefficient matrices \( \mathbf{\beta}_k \) implies that elements of \( \mathbf{\beta}_k \) in the same row have the same variance, but the variance differs for elements in different columns. This is a reasonable assumption because the variables in \( \mathbf{Z}^{(R)} \) are standardized, and the cutoffs \( \gamma_{j,1} \) and \( \gamma_{j,k_{(R)}-1} \) for \( j = 1, \ldots, p_o \) can be fixed to values that imply all variables in \( (\mathbf{W}^{(R)}, \mathbf{Z}^{(R)}) \) have the same scale. However, we want to allow the variances of regression coefficients to differ across covariates, hence the use of the \( \mathbf{\tau} = (\tau_1^2, \ldots, \tau_k^2) \) vector.

We assume a conjugate gamma distribution for the precision parameter \( \alpha \). The number of effective global components in the mixture is influenced by \( \alpha \). Thus, the shape and rate parameters can be fixed to small values, such as 0.5, to represent a relatively noninformative prior on the number of global mixture components. We recommend fixing \( a_1 = \cdots = a_{d_j} = 1 \) in the Dirichlet base distribution so that the probability vectors \( \psi_j^{(j)} \) can encompass a wide variety of values.

To specify priors for the remaining parameters, consider the limiting case of the mixture as \( \alpha \to 0^+ \), which results in a single multivariate normal distribution for \( (\mathbf{W}^{(R)}, \mathbf{Z}^{(R)}) \). Our objective is to center and scale the mixture kernel
appropriately. Since the $\mathbf{Z}^{(R)}$ are standardized, fixing $\gamma_{j,1} \approx -3$ and $\gamma_{j,k-1} \approx 3$ places each $\mathbf{Z}_{j}^{(R)}$ and $\mathbf{W}_{j}^{(R)}$ on similar scales. We use a conjugate matrix-normal prior for $\mathbf{\beta}_{0}$, assuming $\mathbf{\beta}_{0} \sim \text{MN}(\mathbf{0}, \mathbf{I}, h\mathbf{I})$. This represents prior information that the regression coefficients are centered at zero. We assume $\tau_{j}^{2} \sim IG(a_{\tau}, b_{\tau})$, possibly truncated to the lie below a finite value (e.g., 6) if numerical problems arise, and $\mathbf{S} \sim \text{W}(\mathbf{a}_{S}, \mathbf{B}_{S})$. Let $v_{j}$ denote an estimate for the variance of $(\mathbf{W}_{j}^{(R)}, \mathbf{Z}_{j}^{(R)})_{j}$, which is given by $(\gamma_{j,k}^{(R)} - \gamma_{j,1})/4)^{2}$, for ordinal variables $j = 1, \ldots, p_{o}$. For continuous variables $j = p_{o} + 1, \ldots, p_{o} + p_{c}$, the quantity $(\text{range}(\mathbf{Z}_{j-p_{o}}^{(R)})/4)^{2}$ can be used as a proxy for variance. Under the standardization and cutoff points described, each $v_{j} \approx 1.5^{2}$. The marginal prior variance for $(\mathbf{W}_{j}^{(R)}, \mathbf{Z}_{j}^{(R)})_{j} | \mathbf{D}$ is $\mathbf{a}_{S}(\mathbf{B}_{S})_{jj}(v_{j} - p_{c} - p_{o} - 1)^{-1} + (b_{\tau}(a_{\tau} - 1)^{-1} + h) \sum_{j=1}^{K} D_{jj}^{2}$. For a default prior, we fix $a_{\tau}, a_{S}$, and $v$ to relatively small values. Because the design vector $\mathbf{D}$ consists of a one for the intercept term followed by only zeroes and ones, the smallest value it can take is 1. To be conservative, we fix $\sum_{j=1}^{K} D_{jj}^{2}$ to 1 and determine values of $\mathbf{B}_{S}, b_{\tau}$, and $h$ such that marginal prior variance is approximately equal to $v$. For instance, we can set $\mathbf{B}_{s} = (v_{j} - p_{c} - p_{o} - 1)(3a_{S})^{-1}\text{diag}(v_{j}, \ldots, v), b_{\tau} = 3^{-1}(a_{\tau} - 1)v$, and $h = 3^{-1}v$. This implies a prior variance for each element of $\mathbf{\beta}_{j}$ that allows a typical value of $Y_{j}^{(R)}$ to be as small as 1 or as large as $k_{j}^{(R)}$. It also centers each diagonal element of $\mathbf{\Sigma}$ at 0.75, which is reasonable given that each $\mathbf{W}_{j}^{(R)}$ and $\mathbf{Z}_{j}^{(R)}$ ranges from around −3 to 3. Although this can be used as a default approach to prior specification, in our experience inferences like density estimates and regression functionals are insensitive to the choice of values for the parameters of these hyperpriors.

**Algorithm for Selecting Variables in $\mathbf{d}^{(r, \cdot)}$**

We select variables in $\mathcal{F}$ in accordance with the mutual information of each $\mathcal{F}_{j}$, where $l = 1, \ldots, q$, with $\mathbf{X}^{(R)}$, as well as the mutual information between pairs of $(\mathcal{F}_{j}, \mathcal{F}_{j})$, $l \neq l'$. Ding and Peng [8] and Peng et al. [45] provide an algorithm referred to as minimal redundancy/maximal relevance (mRMR) which we adapt for our purposes, extending their algorithm to handle multivariate $\mathbf{X}^{(R)}$. The algorithm proceeds as follows.

1. Set $\mathbf{S} = \{\emptyset\}$ and $\mathbf{G} = \{1, \ldots, q\}$. Calculate empirical estimates of $I(\mathcal{F}_{j}, X_{j}^{(R)})$, for $l = 1, \ldots, q$ and $j = 1, \ldots, p_{n}$, and then $I_{j}^{\text{max}}$, for $l = 1, \ldots, q$. Denote these estimates by $\hat{I}(\mathcal{F}_{j}, X_{j}^{(R)})$ and $\hat{I}_{j}^{\text{max}}$.
2. Find $l^{*} = \arg \max_{l \in \mathbf{G}} \{\hat{I}_{j}^{\text{max}}\}$. Set $\mathbf{S} = \mathcal{F}_{l^{*}}$, and $\mathbf{G} = \{1, \ldots, q\} \setminus l^{*}$. Check whether or not to stop adding variables using some stopping criterion.
3. Repeat the following steps until the stopping criterion is satisfied.

   (a) Find $l^{*} = \arg \max_{l \in \mathbf{G}} \{\hat{I}_{j}^{\text{max}} - |\mathbf{S}|^{-1} \sum_{j \in \mathbf{S}} I(\mathcal{F}_{j}, \mathcal{F}_{j})\}$.

   (b) Add $\mathcal{F}_{l^{*}}$ to the set $\mathbf{S}$ and remove $l^{*}$ from $\mathbf{G}$.  

" Springer
The selected subset of variables $\mathcal{F}_S$ is assigned $w_j = 1/|S|$ in the distance function, and all others are assigned $w_j = 0$. To estimate mutual information $I(\mathcal{F}_i, X_j^{(R)})$ for discrete $\mathcal{F}_i$, we use the empirical discrete distribution of $(\mathcal{F}_i, X_j^{(R)})$ in the calculation. When $\mathcal{F}_i$ is continuous, a simple and effective method of estimating the mutual information involves discretizing $\mathcal{F}_i$ [8]. The marginal distribution for $X^{(R)}$ depends on only the variables in $\mathcal{F}$ that are assigned nonzero weight, so one should include all variables that are moderately predictive of $X^{(R)}$.

Related feature selection algorithms [3, 8, 15] stop when a prespecified number of features have been chosen and use cross-validation to determine the optimal number of features. This is impractical and inefficient for our purposes, as the variable selection problem is not the primary inferential or modeling focus. Rather, it is a component of a complex joint model that is introduced to allow for further dependence of $\mathcal{R}$, and especially $X^{(R)}$, on $\mathcal{F}$.

Instead, we make use of measures of redundancy and relevancy. We propose to stop selecting variables when none remain that explain a significant proportion of the uncertainty in any $X^{(R)}$ variable, or when each remaining $\mathcal{F}$ variable to choose from is already well explained by the chosen set $\mathcal{F}_S = \{\mathcal{F}_s : s \in S\}$, where each $s_j \in \{1, \ldots, q\}$, $j = 1, \ldots, |S|$.

For generic random variable $A$, let $H(A)$ be the entropy of $A$. Define a normalized version of mutual information, $I^*(\mathcal{F}_i; X_j^{(R)}) = I(\mathcal{F}_i, X_j^{(R)})/H(X_j^{(R)})$, as the proportion of uncertainty in $X_j^{(R)}$ that is explained by $\mathcal{F}_i$. We use this quantity to determine whether or not $\mathcal{F}_i$ is relevant in explaining $X_j^{(R)}$. Also, let $I^*(\mathcal{F}_i; \mathcal{F}_{l'}) = I(\mathcal{F}_i, \mathcal{F}_{l'})/H(\mathcal{F}_{l'})$ for $l, l' \in \{1, \ldots, q\}$ be the proportion of uncertainty in $\mathcal{F}_{l'}$ that is explained by $\mathcal{F}_i$. As a measure of redundancy between $\mathcal{F}_i$ and $\mathcal{F}_S$, we use $\max_{j \in S} \{I^*(\mathcal{F}_j; \mathcal{F}_i)\}$. If one or more of the variables in $\mathcal{F}_S$ already explains a large amount of the uncertainty in $\mathcal{F}_i$, then $\mathcal{F}_i$ is considered redundant.

We use the relevancy stopping condition $\max_{i \in G} \{\max \{I^*(\mathcal{F}_i, X_1^{(R)}), \ldots, I^*(\mathcal{F}_i, X_q^{(R)})\}\} < t_1$, and the redundancy stopping condition as $\min_{i \in G} \{\max_{j \in S} \{I^*(\mathcal{F}_j; \mathcal{F}_i)\}\} > t_2$. Here, $t_1$ and $t_2$ are user-defined thresholds between 0 and 1. A reasonable value for $t_2$ is around 0.7 to 0.9, and a reasonable value for $t_1$ is closer to 0, such as 0.05 or 0.1. Generally, information values below 0.02 indicate a variable is not predictive, values 0.02 to 0.1 are considered weakly predictive, and values above 0.1 are considered moderately to strongly predictive [54]. The larger $t_1$ is and the smaller $t_2$ is, the moreparsimonious the selected variable set is. Generally, the number of variables one uses will depend on what the variables are (in applications with knowledge about the existence of relationships between the $\mathcal{F}$ and $X^{(R)}$) and the importance of computational efficiency, as the more variables that are selected, the more time it takes the MCMC algorithm to run.
Posterior Inference: Full Conditionals

The Mixing Parameters

The full conditionals for the mixing parameters \( \{ \theta_h = (\beta_h, \Sigma_h, \psi_h) \} \) arise by combining the likelihood terms in (1) and the base distributions in (4), as is standard in DP mixture models. Let \( M_h = \{ H_i = h \} \), or the size of cluster \( h \). The full conditional for \( \beta_h \) is matrix-normal, or \( \text{vec}(\beta_h^T) \) is multivariate normal. Let \( T = \text{diag}(\tau_1^2, \ldots, \tau_k^2) \), and let \( D_h \) be an \( M_h \) by \( k \) matrix obtained by stacking the set of row vectors \( D(\mathbf{X}_i^{(R)}, \mathbf{Y}_i^{(P)}, \mathbf{Z}_i^{(F)}, \mathbf{X}_i^{(F)}) \), such that \( H_i = h \). Let \((W^{(R)}, Z^{(R)})_h \) be similarly defined. The multivariate normal full conditional has covariance matrix \( \Sigma_{\beta_h} = ((T \otimes I_{p_i+p_o})^{-1} + (D_h^T D_h) \otimes \Sigma^{-1}_h)^{-1} \), and mean vector \( \mu_{\beta_h} = (T \otimes I_{p_i+p_o})^{-1} \text{vec}(\beta_0^T) + (D_h^T \otimes \Sigma^{-1}_h) \text{vec}((W^{(R)})_h) \).

Alternatively, we can write the distribution for \((W^{(R)})_h = (W_i^{(R)}, Z_i^{(R)})_h \) such that \( H_i = h \) as \([W_{ij}^{(R)} \mid W_{i-}^{(R)}][(W_{i-}^{(R)})] \) for \( j = 1, \ldots, p_o + p_c \). This depends on the \( j \)th column of \( \beta_h \), \( \beta_h(\cdot, j) \); only through the first univariate normal, which is \( \mathcal{N}(W_{ij}^{(R)}), D_h(\cdot, j) + mu_{h,j}^*, \Sigma_{h,j} \) where \( \mu_{h,j}^* = \Sigma_{h,j}^{-1}(\Sigma_{h,\cdot,j})^{-1}(W_{i-}^{(R)} - D_h(\cdot, j)) \) and the conditional variance is \( \Sigma_{h,j} = \Sigma_{h,\cdot,j}^{-1}(\Sigma_{h,\cdot,j})^{-1} \Sigma_h^{-1} \). Here, \( D_i = D(\mathbf{X}_i^{(R)}, \mathbf{Y}_i^{(P)}, \mathbf{Z}_i^{(F)}, \mathbf{X}_i^{(F)}) \).

We therefore can update each column \( j = 1, \ldots, p_o + p_c \) of the matrix \( \beta_h \) from a multivariate normal with variance \( \Sigma_{h,j} = (T^{-1} + \Sigma_h^{-1}(D_h^T D_h))^{-1} \) and mean \( \mu_{h,j}^* \). The covariance matrices \( \Sigma_{h} \) are updated from the base distribution \( \Sigma_{h} \), and \( \psi_h \).

When cluster \( h \) is empty, i.e., \( M_h = 0 \), we draw from the base distribution, which we can do by simulating \( \text{vec}(\beta_h) \sim \mathcal{N}(\text{vec}(\beta_0), I_{p_i+p_o} \otimes T) \).

The covariance matrices \( \Sigma_{h} \) are updated from the base distribution \( \Sigma_{h} \), and \( \psi_h \).

The mixture allocation variables and local DP parameters

The mixture configuration variables \( H_i \), where \( i = 1, \ldots, n \) are simulated from categorical distributions. We have

\[
p(H_i \mid \ldots) \propto \sum_{j=1}^{N(f)} p_{lj}(f_i) \delta_{\sigma(f_i)}(\cdot),
\]

where \( p_{lj}(f_i) \propto p(f_i) \mathcal{N}(\beta_{\sigma(f_i)} D(\mathbf{X}_i^{(R)}, \mathbf{Y}_i^{(P)}, \mathbf{Z}_i^{(F)}, \mathbf{X}_i^{(F)}), \Sigma_{\sigma(f_i)} \prod_{j=1}^{p_o} \psi^{(j)}_{\sigma(f_i), X_{j}^{(R)}}).

The full conditionals for \( V_h, h = 1, \ldots, N \), are derived as follows:
\[ p(V_h \mid \ldots) \propto \text{Beta}(V_h; 1, \alpha) \prod_{i=1}^{n} \sum_{l=1}^{N(C_i)} \prod_{j=1}^{N(C_i)} p_i(f_i) \delta_{\pi(f_i)}(H_i) \]

\[ \prod_{i : H_i = h} V_h \prod_{i : H_i > h} (1 - V_h) \]

\[\alpha \text{Beta}(V_h; 1, \alpha) \]

\[ H_i \neq \pi_{N(C_i)}(c_i) \quad h \in \eta_f \]

\[ \text{Beta}(V_h; 1 + \sum_{i=1}^{n} 1(H_i = h, H_i \neq \pi_{N(f_i)}(f_i)), \alpha + \sum_{i=1}^{n} 1(H_i > h \text{ and } h \in \eta_f)). \]

The full conditional for \( \Gamma_h \) is

\[ p(\Gamma_h \mid \ldots) \propto p(\Gamma_h) \prod_{|i : H_i = h|} 1(d(f_i, \Gamma_h) < d^*) \]

The full conditional in general depends on the choice of distance function, which in CMM-Mix is \( d(f_i, \Gamma_h) = \sum_{j=1}^{q} w_j d_j^*(f_i, \Gamma_h) \). The term \( \sum_{j=1}^{q} w_j d_j^*(f_i, \Gamma_h) < d^* \) implies \( a_{il} < \Gamma_{hl} < b_{il} \), for any \( l = 1, \ldots, q \). Combined with the product of independent uniform priors for \( \Gamma_h \), the full conditionals for each \( \Gamma_{hl} \) are uniform.

For \( l = 1, \ldots, q \), \( \Gamma_{hl} \) corresponds to ordinal variable \( F_h \), with \( F_h \in \{1, \ldots, k_{l}^{(F)} \} \), and \( d(F_h, \Gamma_h) < d^* \) implies \( a_{il} < \Gamma_{hl} < b_{il} \) where

\[ a_{il} = F_{il} + ((k_{l}^{(F)} - 1)/w_i)(d^* - \sum_{j \neq l} w_j d_j^*(F_{ij}, \Gamma_{hj})) \]

\[ b_{il} = F_{il} - ((k_{l}^{(F)} - 1)/w_i)(d^* - \sum_{j \neq l} w_j d_j^*(F_{ij}, \Gamma_{hj})). \]

Therefore, the full conditional for \( \Gamma_{hl} \) is uniform, with lower bound equal to the smallest value in \( \{1, \ldots, k_{l}^{(F)} \} \) that is greater than \( \max_{(i : H_i = h)} \{a_{il} \} \), and upper bound equal to the largest value in \( \{1, \ldots, k_{l}^{(F)} \} \) that is less than \( \min_{(i : H_i = h)} \{b_{il} \} \). For \( l : w_i = 0 \), \( \Gamma_{hl} \) is simulated from the prior, which is uniform on \( \{1, \ldots, k_{l}^{(F)} \} \).

For \( l = q_o + 1, \ldots, q_o + q_c \), \( \Gamma_{hl} \) corresponds to continuous variable \( F_h \), and under a uniform\( (a_{il}, b_{il}) \) prior for \( \Gamma_{hl} \), \( d(F_h, \Gamma_h) < d^* \) implies \( a_{il} < \Gamma_{hl} < b_{il} \) where

\[ a_{il} = F_{il} + (\max(F_h) - \min(F_h))(d^* - \sum_{j \neq l} w_j d_j^*(F_{ij}, \Gamma_{hj})) / w_i \]

\[ b_{il} = F_{il} - (\max(F_h) - \min(F_h))(d^* - \sum_{j \neq l} w_j d_j^*(F_{ij}, \Gamma_{hj})) / w_i. \]

Therefore, the full conditional for \( \Gamma_{hl} \) is uniform, with lower bound equal to \( \max(\max_{(i : H_i = h)} \{a_{il} \}, a_{il}^{(F)} \) \), and upper bound equal to \( \min(\min_{(i : H_i = h)} \{b_{il} \}, b_{il}^{(F)} \) \). For \( l : w_i = 0 \), \( \Gamma_{hl} \) is simulated from the prior, which is uniform\( (a_{il}, b_{il}) \).

For \( l = q_o + q_c + 1, \ldots, q \), \( \Gamma_{hl} \) corresponds to nominal variable \( F_h \in \{1, \ldots, d_{l}^{(F)} \} \), and \( d(F_h, \Gamma_h) < d^* \) implies

\[ 1(F_{il} \neq \Gamma_{hl}) < b_{il} = \left( d^* - \sum_{j \neq l} w_j d_j^*(F_{ij}, \Gamma_{hj}) \right) / w_i \]

and therefore if \( \min_{(i : H_i = h)} \{b_{il} \} < 1 \), this implies a degenerate full conditional for \( \Gamma_{hk} \) at \( F_{il} \). Otherwise, \( \Gamma_{hl} \) is drawn randomly from \( \{1, \ldots, d_{l}^{(F)} \} \). For \( k : w_k = 0 \), \( \Gamma_{hl} \) is simulated from the prior, which is uniform on \( \{1, \ldots, d_{l}^{(F)} \} \).
The Base Distribution Hyperparameters

The full conditional for $\alpha$ is the same as under the standard DP applied with truncation. Under the prior $\alpha \sim \text{gamma}(a_\alpha, b_\alpha)$, then the posterior for $\alpha$ is proportional to $\text{gamma}(a_\alpha + N, b_\alpha - \sum_{h=1}^{N} \log(1 - V_h))$.

Under the prior $\beta_{0} \sim \text{MN}_{k\times(p_o+p_c)}(0, I, hI)$, the full conditional for each $\beta_{0\text{m},r}$, $m = 1, \ldots, k$, $r = 1, \ldots, p_o + p_c$ is then

$$N((h^{-1} + N\tau_{m}^{-2})^{-1}(\tau_{m}^{-2}\sum_{i=1}^{p_o + p_c} \beta_{(m,r)}), (h^{-1} + N\tau_{m}^{-2})^{-1}).$$

Assuming $\tau_{j}^{2} \sim \text{IG}(a_\tau, b_\tau)$ for $j = 1, \ldots, k$, we obtain a posterior full conditional for $\tau_{j}^{2}$ which is $\text{IG}(a_\tau + 0.5(p_o + p_c), b_\tau + 0.5 \sum_{h=1}^{N} \sum_{j=1}^{p_o + p_c} \beta_{(h,j)} - \beta_{(0,h)}^{2})$.

With a prior $S \sim \text{Wish}(a_S, B_S)$, $S$ is sampled from $\text{Wish}(N\nu + a_S, (B_S^{-1} + \sum_{h=1}^{N} \sum_{i}^{k} \mu_{i}^{-1})^{-1})$.

Latent Continuous Variables and Missing Data Imputation

The latent continuous random variables $W_{ij}^{(R)}$, $i = 1, \ldots, n$, and $j = 1, \ldots, p_o$ have full conditionals $p(W_{ij}^{(R)} \mid \ldots) \propto N(W_{ij}^{(R)}; \tilde{\mu}_{ij}, \tilde{\sigma}_{ij}^{2})$ where $\tilde{\mu}_{ij}$ and $\tilde{\sigma}_{ij}^{2}$ are the conditional normal mean and variance for $W_{ij}^{(R)}$ resulting from the multivariate normal distribution

$$N(W_{i}^{(R)}, Z_{i}^{(R)}, D(X_{i}^{(R)}, Y_{i}^{(F)}, Z_{i}^{(F)}, X_{i}^{(F)}) \beta_{H_{i}}, \Sigma_{H_{i}}).$$

If $Y_{ij}^{(R)}$ is observed, then $W_{ij}^{(R)}$ must lie in the interval $(\gamma_{j, i, y_{ij}^{(R)} - 1}, \gamma_{j, i, y_{ij}^{(R)}})$, i.e., its full conditional is truncated normal. If $Y_{ij}^{(R)}$ is missing, then $W_{ij}^{(R)}$ is sampled from the normal distribution without any truncation.

Similarly, missing $Z_{ij}^{(R)}$ are simulated from the normal distribution $N(Z_{ij}^{(R)}; \tilde{\mu}_{ij}, \tilde{\sigma}_{ij}^{2})$ where $\tilde{\mu}_{ij}$ and $\tilde{\sigma}_{ij}^{2}$ are the conditional normal mean and variance for $Z_{ij}^{(R)}$ resulting from the multivariate normal distribution $N((W_{ij}^{(R)}, Z_{ij}^{(R)}); D(X_{i}^{(R)}, Y_{i}^{(F)}, Z_{i}^{(F)}, X_{i}^{(F)}) \beta_{H_{i}}, \Sigma_{H_{i}})$.

In order to improve mixing, we can block updating the $W_{ij}^{(R)}$ and $Z_{ij}^{(R)}$ using the fact that

$$p(W_{ij}^{(R)} \mid \text{data}, \theta) = p(W_{ij}^{(R)} \mid \text{data}, \theta) \prod_{j=2}^{p_o} p(W_{ij}^{(R)} \mid (W_{ij}^{(R)}, \ldots, W_{ij-1}^{(R)}), \text{data}, \theta),$$

where $\theta$ indicates all model parameters. Letting $D_{ij} = D(X_{i}^{(R)}, Y_{i}^{(F)}, Z_{i}^{(F)}, X_{i}^{(F)})$, we sample $W_{ij}^{(R)} \sim N(W_{ij}^{(R)}; \tilde{D}_{ij}, \beta_{H_{i}, \Sigma_{H_{i}}})$, with truncation to $(\gamma_{1, y_{ij}^{(R)} - 1}, \gamma_{1, y_{ij}^{(R)}})$ if $Y_{ij}^{(R)}$ is observed. For $j = 2, \ldots, p_o$, let $\mu_{ij}^{n}$ and $(\sigma_{ij}^{2})^{n}$ denote the conditional mean and variance for $W_{ij}^{(R)}$, and $\tilde{W}_{ij}^{(R)} \sim N(\tilde{W}_{ij}^{(R)}; \mu_{ij}^{n}, (\sigma_{ij}^{2})^{n})$, with truncation to $(\gamma_{j, i, y_{ij}^{(R)} - 1}, \gamma_{j, i, y_{ij}^{(R)}})$ if $Y_{ij}^{(R)}$ is observed.

Missing $X_{ij}^{(R)}$ are simulated from categorical distributions on $\{1, \ldots, d_{ij}^{(R)}\}$ with probabilities proportional to

$$\left(\psi_{H_{i}, l}^{(R)} N(W_{ij}^{(R)}, Z_{ij}^{(R)}; D_{ij}, \beta_{H_{i}}, \Sigma_{H_{i}}), \ldots, \psi_{H_{i}, d_{ij}^{(R)}}^{(R)} N(W_{ij}^{(R)}, Z_{ij}^{(R)}; D_{ij}, \beta_{H_{i}}, \Sigma_{H_{i}})\right)$$

where $D_{ij}$ indicates the current value of $D(X_{i}^{(R)}, Y_{i}^{(F)}, Z_{i}^{(F)}, X_{i}^{(F)})$ in which $X_{ij}^{(R)} = l$.

References

1. Banerjee A, Murray J, Dunson D (2013) Bayesian learning of joint distributions of objects. In: Proceedings of the 16th international conference on artificial intelligence and statistics
2. Bao J, Hanson TE (2015) Bayesian nonparametric multivariate ordinal regression. Can J Stat 43:337–357
3. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5:537–550
4. Chung Y, Dunson D (2011) The local Dirichlet process. Ann Inst Stat Math 63:59–80
5. Cocotas A (2012) Young women are the most valuable mobile ad demographic. www.businessinsider.com/young-women-are-most-valuable-mobile-ad-demographic-2012-2
6. De Iorio M, Müller P, Rosner G, MacEachern S (2004) An ANOVA model for dependent random measures. J Am Stat Assoc 99:205–215
7. DeYoreo M, Kottas A (2018) Bayesian nonparametric modeling for multivariate ordinal regression. J Comput Gr Stat 27:71–84
8. Ding C, Peng H (2003) Minimum redundancy feature selection from microarray gene expression data. In: Proceedings of the 2003 IEEE, pp 523–528
9. D’Orazio M (2015) StatMatch: Statistical Matching. R Package Version 1(2):3
10. D’Orazio M, Di Zio M, Scanu M (2006) Statistical matching: theory and practice. Wiley, New York
11. Dunson D, Bhattacharya A (2010) Nonparametric Bayes regression and classification through mixtures of product kernels. Bayesian Stat 9:145–164
12. Dunson D, Park J (2008) Kernel stick-breaking processes. Biometrika 95:307–323
13. Dunson D, Xing C (2009) Nonparametric Bayes modeling of multivariate categorical data. J Am Stat Assoc 104:1042–1051
14. Efroymovich S (2007) Conditional density estimation in a regression setting. Ann Stat 35:2504–2535
15. Eztévez P, Tesmer M, Zurada J (2009) Normalized mutual information feature selection. IEEE Trans Neural Netw 20:189–201
16. Fosdick B, DeYoreo M, Reiter J (2016) Categorical data fusion using auxiliary information. Ann Appl Stat 10:1907–1929
17. Gilula Z, McCulloch R, Rossi P (2006) A direct approach to data fusion. J Mark Res 43:73–83
18. Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871
19. Griffin J, Steel M (2006) Order-based dependent Dirichlet processes. J Am Stat Assoc 101:179–194
20. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
21. Hall P, Racine J, Li Q (2004) Cross-validation and the estimation of conditional probability densities. J Am Stat Assoc 99:10151026
22. Hannah L, Blei D, Powell W (2011) Dirichlet process mixtures of generalized linear models. J Mach Learn Res 1:1–33
23. Hu J, Mitra R, Reiter J (2013) Are independent draws necessary for multiple imputation? Am Stat 67:143–149
24. Ishwaran H, James L (2001) Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc 96:161–173
25. Ishwaran H, Zarepour M (2000) Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. Biometrika 87:371–390
26. Jansen J (2010) Use of the internet in higher-income households. http://www.pewinternet.org/2010/11/24/use-of-the-internet-in-higher-income-households
27. Kamakura W, Wedel M (1997) Statistical data fusion for cross tabulation. J Mark Res 34:485–498
28. Kaufman L, Rousseeuw P (1990) Finding groups in data. Wiley, New York
29. Kiesl H, Rässler S (2006) How valid can data fusion be. In: IAB Discussion Paper, 15
30. Kottas A, Müller P, Quintana F (2005) Nonparametric Bayesian modelling for multivariate ordinal data. J Comput Gr Stat 14:610–625
31. Kunihama T, Dunson DB (2016) Nonparametric Bayes inference on conditional independence. Biometrika 103:35–47
32. Kunihama T, Herring A, Halpern C, Dunson D (2016) Nonparametric Bayes modeling with sample survey weights. Stat Probab Lett 113:41–48
33. Linderman S, Johnson M, Adams R (2015) Dependent multinomial models made easy: Stick-breaking with the Pólya-gamma augmentation. In: Neural information processing systems
34. Long S, Freese J (2006) Regression models for categorical dependent variables using Stata. Stata Press, College Station
35. Madrigal A (2012) Sorry, young man, you’re not the most important demographic in tech. http://www.theatlantic.com/technology/archive/2012/06/sorry-young-man-youre-not-the-most-important-demographic-in-tech/258087

36. Maimon O, Rokach L (2010) Data mining and knowledge discovery handbook. Springer, New York
37. Minsker S, Srivastava S, Lin L, Dunson D (2016) Robust and scalable Bayes via a median of subset posterior measures. math.ST. arXiv:1403.2660
38. Müller P, Erkanli A, West M (1996) Bayesian curve fitting using multivariate normal mixtures. Biometrika 83:67–79
39. Müller P, Mitra R (2013) Bayesian nonparametric inference: why and how? Bayesian Anal 8:269–302
40. Müller P, Quintana F (2004) Nonparametric Bayesian data analysis. Stat Sci 19:95–110
41. Murray J, Reiter J (2016) Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. J Am Stat Assoc 111(516):1466–1479
42. Norets A, Pelenis J (2012) Bayesian modeling of joint and conditional distributions. J Econom 168:332–346
43. Paiva T, Reiter JP (2017) Stop or continue data collection: a nonignorable missing data approach for continuous variables. J Off Stat 579–599:33
44. Papageorgiou G, Richardson S, Best N (2015) Bayesian nonparametric models for spatially indexed data of mixed type. J R Stat Soc 77:973–999
45. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependence, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238
46. Rainie L, Zickuhr K, Purcell K, Madden M, Brenner J (2012) The rise of e-reading. http://libraries.pewinternet.org/2012/04/04/part-2-the-general-reading-habits-of-americans
47. Rässler S (2004) Data fusion: identification problems, validity, and multiple imputation. Austrian J Stat 33:153–171
48. Rodriguez A, Dunson D (2011) Nonparametric Bayesian models through probit stick-breaking processes. Bayesian Anal 6:145–177
49. Rubin D (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
50. Savitsky T, Toth D, Sverchkov M (2016) Bayesian estimation under informative sampling. Electron J Stat, To appear
51. Schifeling T, Reiter J (2016) Incorporating marginal prior information in latent class models. Bayesian Anal 11:499–518
52. Shahbaba B, Neal R (2009) Nonlinear modeling using Dirichlet process mixtures. J Mach Learn Res 10:1829–1850
53. Si Y, Pillai N, Gelman A (2015) Bayesian nonparametric weighted sampling inference. Bayesian Anal 10:605–625
54. Siddiqi N (2006) Credit risk scorecards: developing and implementing intelligent credit scoring. Wiley, New York
55. Taddy M, Kottas A (2010) A Bayesian nonparametric approach to inference for quantile regression. J Business Econ Stat 28:357–369
56. van der Putten P, Kok JN, Gupta A (2002) Data fusion through statistical matching. In: Working paper 4342-02, MIT Sloan School of Management
57. Wade S, Dunson D, Perone S, Trippa L (2014) Improving prediction from Dirichlet process mixtures via enrichment. J Mach Learn Res 15:1041–1071
58. Wang C, Paisley J, Blei D (2011) Online variational inference for the hierarchical Dirichlet process. In: International conference on artificial intelligence and statistics, pp 752–760
59. Wang X, Dunson D (2014) Parallelizing MCMC via Weierstrass sample. stat.CO. arXiv:1312.4605
60. Wicken G, Elms S (2009) Demystifying data fusion—the “why?”, the “how?” and the “wow!”. Tech. rep. Advertising Research Foundation Week of Workshops, New York

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