Nonparametric statistical inference and imputation for incomplete categorical data[*]

Chaojie Wang, Linghao Shen, Han Li, and Xiaodan Fan[†]

Missingness in categorical data is a common problem in various real applications. Traditional approaches either utilize only the complete observations or impute the missing data by some ad hoc methods rather than the true conditional distribution of the missing data, thus losing or distorting the rich information in the partial observations. In this paper, we propose the Dirichlet Process Mixture of Collapsed Product-Multinomials (DPMCPM) to model the full data jointly and compute the model efficiently. By fitting an infinite mixture of product-multinomial distributions, DPMCPM is applicable for any categorical data regardless of the true distribution, which may contain complex association among variables. Under the framework of latent class analysis, we show that DPMCPM can model general missing mechanisms by creating an extra category to denote missingness, which implicitly integrates out the missing part with regard to their true conditional distribution. Through simulation studies and a real application, we demonstrate that DPMCPM outperforms existing approaches on statistical inference and imputation for incomplete categorical data of various missing mechanisms. DPMCPM is implemented as the R package `MMDai`, which is available from the Comprehensive R Archive Network at https://cran.r-project.org/web/packages/MMDai/index.html.

Keywords and phrases: Infinite mixture model, Product-multinomial distribution, Missing data, Imputation.

1. INTRODUCTION

Missingness in categorical data is a common problem in many real applications. For examples, in social surveys, data collected by questionnaires are often incomplete because subjects may be unwilling or unable to respond to some items [11]. In biological experiments, data may be incomplete for either biologically-driven or technically-driven reasons [7]. In recommendation system problems [16], analysts often have a dataset as the toy example in Table 1. The

goal is to predict the potential purchase behavior without direct observation, e.g., whether Customer3 and Customer4 would buy Book5.

Table 1. Toy data example with the goal to infer the empty cells. 1: bought; 0: recommended but not bought; empty: missing values.

	Book1	Book2	Book3	Book4	Book5
Customer1	1	1			0
Customer2	1	0			0
Customer3		1	1		
Customer4		1	1		
Customer5			0	1	0
Customer6			0	1	0

When dealing with datasets with missingness, naive approaches, such as the complete-case analysis (CCA) and overall mean imputation, would waste the information in the missing data and may bias the inference [2]. When missingness is high, CCA is hardly applicable due to the lack of complete cases. Advanced methods, such as multiple imputation [18, 20], impose a parametric model on the data and then draw multiple sets of samples to account for the uncertainty of the missing information. For categorical cases, [19] advocated the log-linear model for multiple imputation, which can capture certain types of association among the categorical variables. However, this model works only when the number of variables is small, as the full multi-way cross-tabulation required for the log-linear analysis increases exponentially with the number of variables [27].

There are two basic ideas for imputing multivariate missing data: fully conditional specification (FCS) and joint modeling. FSC [25] specifies a collection of univariate conditional imputation models that condition on all the other variables. A popular application based on FCS is known as Multiple Imputation by Chained Equation (MICE), which specifies a sequence of regression models iteratively [24, 31]. [23] provided the R package `mice` to implement this method efficiently. Although it has been shown to work well for many datasets and become a popular method [9], MICE still has some common drawbacks of FCS. A typical application of MICE is to use multinomial logistic regression for the categorical data, but the relationship among variables may be nonlinear and may involve complex interaction or higher-order effects [15]. Besides, there is no guarantee that the

iterations of sequential regression model will converge to the true posterior distribution of missing values [27]. Other parametric approaches include missMDA based on principal component analysis [10], MIMCA based on correspondence analysis [1], and so on. These parametric methods can be applied in some specific problems but not in general cases. [14] provided a detailed review for current advances of multiple imputation.

From the perspective of data analysts, no matter the data is raw with missingness or has been imputed when arriving, it is important to understand the detailed mechanisms of pre-processing, including imputation. [32] reviewed and discussed the imputation danger caused by the difference of the God's model, imputer's model and analyst's model. They pointed that any attempt of pre-processing to make the data "more usable" implies potential assumptions. In this aspect, joint modeling provides a powerful tool to model the underlying distribution behind the data. For categorical data, [27] proposed a latent class model, i.e., the finite mixture of product-multinomial model. The latent class model can characterize both simple association and complex higher-order interactions among the categorical variables as long as the number of latent class is large enough [13]. [3] proposed the Dirichlet Process Mixture of Products of Multinomial distributions (DPMPM) to model *complete* multivariate categorical datasets, which avoids the arbitrary setting of the number of mixture components. Furthermore, [3] proved that any multivariate categorical data distribution can be approximated by DPMPM for a sufficiently large number of mixture components. [21] generalized the DPMPM framework to analyze incomplete categorical datasets, which works well for low missingness but performs poorly for high missingness since the number of parameters would increase dramatically. Based on the work of [21], [12] introduced a variant of this model for edit-imputation, which accounts for the values that are logically impossible but present due to measurement error. [8] extended this model to nested data structures in the presence of structural zeros. Other related works include the divisive latent class model [26], Bayesian multilevel latent class models [29] and so on. [28] presented a detailed overview of recent researches on multiple imputation using the latent class model.

In this paper, we propose DPMCPM, which extends DPMPM for modelling *incomplete* multivariate categorical data efficiently. DPMCPM inherits some nice properties of DPMPM. It avoids the arbitrary setting of the number of mixture components by using the Dirichlet process. In addition, DPMCPM is applicable for any categorical data regardless of the true distribution and can capture the complex association among variables. Different from the missingness viewed as unknown parameters in [21], DPMCPM creates an extra category to denote the missingness. It shall reduce computation burden and gain better performance

when the missingness is high. Under the framework of the latent class analysis, we show that DPMCPM can model general missing mechanisms. Through simulation studies and a real application, we demonstrate DPMCPM performs better statistical inference and imputation than existing approaches. To our knowledge, this is the first non-parametric tool which can model arbitrary categorical distributions and handle high missingness.

This paper is organized as follows. Section 2 introduces the DPMCPM model and the Gibbs sampler algorithm. Section 3 performs simulation studies on the synthetic data and Section 4 presents a real application in a recommendation system problem. Section 5 concludes the paper.

2. METHOD

2.1 Dirichlet process mixture of product-multinomials

We begin with the finite mixture product-multinomial model for the case of complete dataset $\boldsymbol{x} = \{x_{ij}\}$, where $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Suppose $\boldsymbol{x}$ comprises of n independent samples associated with p categorical variables, and the j-th variable has d_j categories. Let $x_{ij} \in \{1, \cdots, d_j\}$ denote the observed category for the i-th sample in the j-th variable. The finite mixture product-multinomial model assumes that those x_{ij} are generated from a multinomial distribution indexed by a latent variable $z_i \in \{1, \cdots, k\}$. A finite mixture model with k latent components can be expressed as:

$$(1) \qquad x_{ij}|z_i, \boldsymbol{\psi}_{z_i}^{(j)} \sim multinomial(\psi_{z_i 1}^{(j)}, \cdots, \psi_{z_i d_j}^{(j)}),$$

$$(2) \qquad z_i|\Theta \sim multinomial(\theta_1, \cdots, \theta_k),$$

where $\Theta = \{\theta_1, \cdots, \theta_k\}$ and $\boldsymbol{\psi}_{z_i}^{(j)} = \{\psi_{z_i 1}^{(j)}, \cdots, \psi_{z_i d_j}^{(j)}\}$. We further define $\Psi = \{\boldsymbol{\psi}_h^{(j)} : h = 1, \cdots, k; j = 1, \cdots, p\}$ and $\boldsymbol{z} = \{z_i : i = 1, \cdots, n\}$.

[3] proved that any multivariate categorical data distribution can be represented by the mixture distribution in Equation 1 and 2 for a sufficiently large k. However, specifying a good k to avoid over-fitting and over-simplification is non-trivial, and it becomes even harder when the dataset is highly incomplete [3, 21]. This motivates the use of an infinite extension of the finite mixture model, i.e., the Dirichlet process mixture. A Dirichlet process can be represented by various schemes, including the Pólya urn scheme, the Chinese restaurant process and the stick-breaking construction [22]. [3] chose the stick-breaking construction to model the Dirichlet process. However, in practice, the slice Gibbs sampler in their construction may often be trapped in a single component when n is relatively large due to numeric limits and thus fail to identify the correct number of components

[30]. To avoid this drawback, we construct the Dirichlet process by using the Chinese restaurant process in DPMCPM:

$$(3) \quad \begin{aligned} P(z_i = h | \boldsymbol{z}_{-i}) &= \frac{n_{h,-i}}{n + \alpha - 1}, \quad h = 1, \cdots, k; \\ P(z_i = k+1 | \boldsymbol{z}_{-i}) &= \frac{\alpha}{n + \alpha - 1}, \end{aligned}$$

where k denotes the number of previously occupied components and $n_{h,-i}$ denotes the number of samples in the h-th component excluding the i-th sample. Here α is a prior hyper-parameter that acts as the pseudo-count of the number of samples in the new component regardless of the input data.

2.2 Multinomials with incomplete data

[17] introduced a class of missing mechanisms, which are commonly referred as Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). Traditionally, missing mechanisms of a data matrix $\boldsymbol{x}$ are denoted by introducing an indicator matrix $\boldsymbol{r} = (r_{ij})$, where $r_{ij} = 1$ if x_{ij} is observed, and $r_{ij} = 0$ if x_{ij} is missing. The missing rates $p(r_{ij}|\boldsymbol{x}_i)$ under three types of missing mechanisms are defined as follows:

- MCAR: Missingness does not depend on the missing or observed data, i.e.,

$$p(r_{ij}|\boldsymbol{x}_i) = \tau,$$

  where $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$ is the i-th observation and $\tau \in (0,1)$ is a constant which does not depend on $\boldsymbol{x}_i$;

- MAR: Missingness only depends on the observed data, i.e.,

$$p(r_{ij}|\boldsymbol{x}_i) = p(r_{ij}|\boldsymbol{x}_{i,obs}),$$

  where $\boldsymbol{x}_{i,obs}$ denotes the observed values in the i-th observation;

- MNAR: Missingness depends on both the missing and observed data, i.e.,

$$p(r_{ij}|\boldsymbol{x}_i) = p(r_{ij}|\boldsymbol{x}_{i,obs}, \boldsymbol{x}_{i,mis}),$$

  where $\boldsymbol{x}_{i,obs}$ and $\boldsymbol{x}_{i,mis}$ denote the observed and missing values in the i-th observation, respectively.

In DPMCPM, we model the incomplete data by creating an extra category to denote the missingness. Here Equation 4 is used to replace Equation 1 in the DPMPM model:

$$(4) \quad x_{ij}|z_i, \boldsymbol{\psi}_{z_i}^{(j)} \sim multinomial(\psi_{z_i 0}^{(j)}, \psi_{z_i 1}^{(j)}, \cdots, \psi_{z_i d_j}^{(j)}),$$

where the extra 0-th category denotes the case when x_{ij} is missing.

Once we get the parameter estimates using Equation 4, we shall rescale the multinomial probabilities as follows to recover the corresponding parameters in Equation 1:

$$\tilde{\psi}_{hc}^{(j)} = p(x_{ij} = c | x_{ij} \neq 0, z_i = h, \boldsymbol{\psi}_h^{(j)}) = \psi_{hc}^{(j)}/(1 - \psi_{h0}^{(j)}),$$

for $c = 1, \cdots, d_j$. This procedure implicitly integrates out the missing values according to their estimated conditional distributions. As compared to [21], this "collapse" step dramatically shrinks the dimension of the posterior sampling space, thus enables DPMCPM to handle high missingness.

The idea of introducing an extra category comes from [5] for the log-linear model but they did not provide a strict proof. In fact, because any relationship among variables can be approximated by the mixture product-multinomial model for a sufficiently large k, DPMCPM also accommodates any dependencies among the 0-th categories and other parameters, thus it can capture the MCAR, MAR and MNAR missing mechanisms [3, 5]. We have the following theorem:

Theorem 1. *For any general missing rate $p(r_{ij}|\boldsymbol{x}_i)$, there exists at least one set of parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi} = \{\psi_{hc}^{(j)} : h = 1, \cdots, k; j = 1, \cdots, p; c = 0, 1, \cdots, d_j\}$ in DPMCPM, which is equivalent with the corresponding missing rate.*

Proof. See Appendix A.1 for detailed proofs. □

Based on the estimation of $\boldsymbol{\Theta}$ and $\tilde{\boldsymbol{\Psi}} = \{\tilde{\psi}_{hc}^{(j)} : h = 1, \cdots, k; j = 1, \cdots, p; c = 1, \cdots, d_j\}$, we can obtain an approximate of the true distribution up to the Monte Carlo error. With this estimated distribution, any statistical inference and imputation can be performed. For example, if x_{ij} is missing and a single imputation is desired, we can impute $\hat{x}_{ij} = c_{\text{pred}}$ by

$$c_{\text{pred}} = \arg\max_c p(x_{ij} = c | \boldsymbol{x}_{i,obs}, \boldsymbol{\Theta}, \tilde{\boldsymbol{\Psi}}),$$

where $\boldsymbol{x}_{i,obs}$ denotes all of the observed data in the i-th sample and c_{pred} is the predicted value.

2.3 Posterior inference

We use the Bayesian approach to fit the non-parametric model in Equation 2-4 to the data. More specifically, we introduce prior distributions for unknown parameters, and then use a Markov chain Monte Carlo algorithm to sample from their posterior distribution. The converged samples will be used for statistical inference.

For prior distributions, we assume the conjugate prior for $\boldsymbol{\psi}_h^{(j)}$, i.e.,

$$\boldsymbol{\psi}_h^{(j)} \sim Dirichlet(\beta_{j0}, \cdots, \beta_{jd_j}),$$

for $h = 1, \cdots, k$ and $j = 1, \cdots, p$. If the sample size n is small, we suggest using a flat and weak prior by setting $\beta_{j0} = \cdots = \beta_{jd_j} = 1$ for all j. For the Dirichlet prior in Equation 3, [22] suggested that the hyper-parameter α should be a small number, thus we set $\alpha = 0.25$ by default. In a specific application where users have more knowledge about the concentration level, the value of α can be changed according to the prior knowledge.

Since x_{ij}'s are independent conditional on the latent component z_i's and $\boldsymbol{\psi}_{z_i}^{(j)}$'s as in Equation 4, the joint likelihood of $\boldsymbol{x}$ given $\boldsymbol{z}$ and Ψ can be written as

$$p(\boldsymbol{x}|\boldsymbol{z},\Psi) = \prod_{i=1}^{n}\prod_{j=1}^{p} p(x_{ij}|z_i,\boldsymbol{\psi}_{z_i}^{(j)}) = \prod_{i=1}^{n}\prod_{j=1}^{p} \psi_{z_i x_{ij}}^{(j)}.$$

Since $\boldsymbol{z}$ and Ψ are independent, the augmented joint posterior distribution is $p(\boldsymbol{z},\Psi|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{z},\Psi)p(\boldsymbol{z})p(\Psi)$.

To sample from the joint posterior distribution, we use the following Gibbs sampler algorithm 1:

Algorithm 1 Gibbs Sampler Algorithm

1: Initialize parameters $\boldsymbol{z}$ and $\boldsymbol{\Psi}$. For convenience, we can set all samples in different components initially. Set $z_i = i$, $i = 1,\cdots,n$ and $k = n$. Set $\boldsymbol{\psi}_i^{(j)}$ by sampling from $\boldsymbol{\psi}_i^{(j)} \sim Dirichlet(\beta_{j0},\cdots,\beta_{jd_j})$, for $i = 1,\cdots,n$ and $j = 1,\cdots,p$.
2: Update $z_i, i = 1,\cdots,n$, by sampling according to the following weights:

$$P(z_i = h|\boldsymbol{z}_{-i},\boldsymbol{x},\Psi) \propto \frac{n_{h,-i}}{n+\alpha-1}\prod_{j=1}^{p}\psi_{h x_{ij}}^{(j)}, \quad h = 1,\cdots,k;$$

$$P(z_i = k+1|\boldsymbol{z}_{-i},\boldsymbol{x},\Psi)$$
$$\propto \frac{\alpha}{n+\alpha-1}\prod_{j=1}^{p}\frac{\prod_{c=0}^{d_j}\Gamma(I(x_{ij}=c)+\beta_{jc})}{\Gamma(\sum_{c=0}^{d_j}\beta_{jc}+1)}.$$

If $z_i = k+1$, we obtain a new component and increase k by 1. Then we sample a new component parameter $\boldsymbol{\psi}_{k+1}$ from $\boldsymbol{\psi}_{k+1}^{(j)} \sim Dirichlet(I(x_{ij}=0)+\beta_{j0},\cdots,I(x_{ij}=d_j)+\beta_{jd_j})$, for $j = 1,\cdots,p$.
3: Sort the components decreasingly according to the number of samples in each component. Delete empty components and re-calculate k.
4: Update $\boldsymbol{\psi}_h^{(j)}$ by sampling from

$$\boldsymbol{\psi}_h^{(j)} \sim Dirichlet(\sum_{i:z_i=h}I(x_{ij}=0)+\beta_{j0},$$
$$\cdots,\sum_{i:z_i=h}I(x_{ij}=d_j)+\beta_{jd_j}),$$

for $h = 1,\cdots,k$ and $j = 1,\cdots,p$.
5: Repeat Step 2-4 until convergence.
6: Re-scale the multinomial probabilities without the category for missingness:
$$\tilde{\psi}_{hc}^{(j)} = \psi_{hc}^{(j)}/(1-\psi_{h0}^{(j)}),$$
for $h = 1,\cdots,k; j = 1,\cdots,p;$ and $c = 1,\cdots,d_j.$

In Step 2, for $h = 1,\cdots,k$, the conditional posterior distribution is derived as follows:
$$p(z_i = h|\boldsymbol{z}_{-i},\boldsymbol{x},\Psi) = p(z_i = h|\boldsymbol{z}_{-i},\boldsymbol{x}_i,\boldsymbol{\psi}_h)$$
$$\propto p(z_i = h|\boldsymbol{z}_{-i})p(\boldsymbol{x}_i|z_i = h,\boldsymbol{\psi}_h)$$
$$= \frac{n_{h,-i}}{n+\alpha-1}\prod_{j=1}^{p}p(x_{ij}|z_i = h,\boldsymbol{\psi}_h^{(j)})$$
$$= \frac{n_{h,-i}}{n+\alpha-1}\prod_{j=1}^{p}\psi_{h x_{ij}}^{(j)}.$$

For $h = k+1$, we have:

$$p(z_i = k+1|\boldsymbol{z}_{-i},\boldsymbol{x},\Psi) \propto p(z_i = k+1|\boldsymbol{z}_{-i})p(\boldsymbol{x}_i|z_i = k+1)$$
$$= \frac{\alpha}{n+\alpha-1}\prod_{j=1}^{p}p(x_{ij}|z_i = k+1),$$

where

$$p(x_{ij}|z_i = k+1) = \int p(x_{ij}|z_i = k+1,\boldsymbol{\psi}_{k+1}^{(j)})p(\boldsymbol{\psi}_{k+1}^{(j)})d\boldsymbol{\psi}_{k+1}^{(j)}$$
$$= \int \prod_{c=0}^{d_j}(\psi_{k+1,c}^{(j)})^{I(x_{ij}=c)}\cdot\prod_{c=0}^{d_j}(\psi_{k+1,c}^{(j)})^{\beta_{jc}-1}d\boldsymbol{\psi}_{k+1}^{(j)}$$
$$= \frac{\prod_{c=0}^{d_j}\Gamma(I(x_{ij}=c)+\beta_{jc})}{\Gamma(\sum_{c=0}^{d_j}\beta_{jc}+1)}.$$

In Step 4, the conditional posterior distribution comes from:

$$p(\boldsymbol{\psi}_h^{(j)}|\boldsymbol{x},\boldsymbol{z}) \propto p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\psi}_h^{(j)})p(\boldsymbol{\psi}_h^{(j)})$$
$$\propto \prod_{i:z_i=h}\prod_{c=0}^{d_j}(\psi_{hc}^{(j)})^{I(x_{ij}=c)}\cdot\prod_{c=0}^{d_j}(\psi_{hc}^{(j)})^{\beta_{jc}-1}$$
$$= \prod_{c=0}^{d_j}\left[\psi_{hc}^{(j)}\right]^{\sum_{i:z_i=h}I(x_{ij}=c)+\beta_{jc}-1}.$$

The above Gibbs sampler algorithm is implemented in the R package `MMDai`.

2.4 Identifiability

[3] proved that any multivariate categorical distribution can be decomposed to a finite mixture product-multinomial model for some k. It should be noted that this decomposition is not identifiable if no restrictions are placed, where "identifiable" means the decomposition is unique up to label-switching. In fact, [4] proved that the k-component finite mixture of univariate multinomial distribution is identifiable if and only if the number of trials m in multinomial distributions satisfies the condition that $m > 2k - 1$. For the multivariate multinomial distribution, we can also prove that it is identifiable if and only if $m_j > 2k - 1$ for all j. In most real applications, we only have $m_j = 1$ and thus the decomposition is always unidentifiable.

Fortunately, identifiability is not a problem in our analysis. Firstly, we are only interested in the estimation of the joint distribution, which is unique though the decomposition is not. The main issue here is whether the joint distribution can be approximated well enough. [27] discussed the above difference between estimating joint distribution and clustering when the latent class model is adopted. Secondly, with the Dirichlet prior in our Bayesian approach, different decompositions are weighted unequally and the simpler model is preferred. For the multiple non-identifiable decompositions, our algorithm usually converges to the decomposition with the smallest k.

3. SIMULATIONS

In this section, we check the performance of our method and compare it with other methods based on synthetic data. In each simulation experiment, we first synthesize a set of complete data following a given data model, then mask a certain percentage of the observations according to the corresponding missing mechanism to generate the incomplete dataset as our input data, finally use the masked entries to test the performance of statistical inference and imputation. We consider the following missing mechanisms in simulation studies:

- MCAR: The missingness does not depend on observed and missing values,

$$p(x_{ij} = 0) = 0.2,$$

  for all $i = 1, \cdots, n$ and $j = 1, \cdots, p$.
- MAR: The first variable is fully observed and the missingness in other variables depends on the observation in the first variable:

$$p(x_{ij} = 0 | x_{i1} = 1) = 0.1, \quad p(x_{ij} = 0 | x_{i1} = 2) = 0.3,$$

  for $i = 1, \cdots, n$ and $j = 2, \cdots, p$.
- MNAR: The missingness depends on the missing value itself:

$$p(x_{ij} = 0 | x_{ij} = 1) = 0.1, \quad p(x_{ij} = 0 | x_{ij} = 2) = 0.3,$$

  for all $i = 1, \cdots, n$ and $j = 1, \cdots, p$.

Considering the uncertainty of Monte Carlo simulations, we repeat the experiments, including the data synthesis step, for 100 times independently for each data model. To compare the performance with existing methods, we also apply the DPMPM model in [21] and MICE on these datasets. DPMPM and MICE represent two typical ideas for imputing incomplete categorical data respectively: joint modeling by Bayesian latent class model and FCS by iterative regression. See the supplementary materials for the detailed R code, http://intlpress.com/site/pub/files/_supp/sii/2020/0013/0001/SII-2020-0013-0001-s002.zip.

3.1 Case 1: data from mixture model

In this case, we generate binary data ($d_j = 2$ for all j) from a finite mixture of product-multinomial distributions with $n = 50$ and $p = 20$. The true parameters in the mixture model are set as follows: the number of components $k = 3$, Θ and Ψ are sampled from $\Theta \sim Dirichlet(10, \cdots, 10)$ and $\psi_h^{(j)} \sim Dirichlet(0.5, \cdots, 0.5)$ for $h = 1, \cdots, k$ and $j = 1, \cdots, p$, respectively.

First we show that DPMCPM can identify the correct number of components in the mixture model. Figure 1 presents the histogram of the estimated k in 100 replications under different missing mechanisms. In most cases, DPMCPM can capture the true number of components $k = 3$.
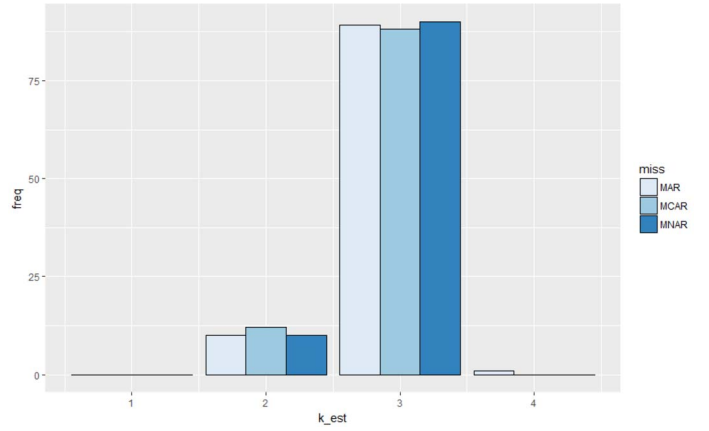

Figure 1. Histograms of the estimated k in 100 replications under different missing mechanisms.

In other cases, DPMCPM may select the simpler model to interpret datasets.

Then we show that DPMCPM can outperform DPMPM and MICE in the imputation of missing values, where the imputation accuracy is defined as follows:

$$\text{imputation accuracy} = \frac{\text{the number of correct imputation}}{\text{the number of masked entries}}.$$

Here the number of correct imputation is calculated based on the true complete data. Table 2 presents the mean and standard error of imputation accuracy of the three methods for 100 replications under different missing mechanisms. It shows that DPMCPM significantly outperformed other methods.

Table 2. Mean and standard error of imputation accuracy of the three methods under different missing mechanisms. Data are generated from the finite mixture product-multinomial model.

	DPMCPM	DPMPM	MICE
MCAR	0.7860(0.044)	0.7031(0.040)	0.5746(0.039)
MAR	0.7744(0.049)	0.6987(0.048)	0.5687(0.052)
MNAR	0.7684(0.050)	0.6811(0.049)	0.5752(0.044)

Besides the imputation accuracy, we can also compare the performance on other statistical inference problems. From DPMCPM, we obtain an estimation of the multivariate multinomial distribution with the estimates of Θ and Ψ. Although the distribution estimation can be a little different from the true distribution due to the existence of missingness and the limited sample size, DPMCPM can perform more or better statistical inference than DPMPM and MICE based on the estimated distribution. Table 3 presents the gap between the estimated correlation matrix and the true

Inference and imputation for incomplete categorical data 21

correlation matrix. The gap is defined as follows:

$$\text{gap} = \sum_i \sum_j (\hat{\sigma}_{ij} - \sigma_{ij})^2,$$

where $\hat{\sigma}_{ij}$ and σ_{ij} are the estimated and true correlation between the i-th and j-th variables, respectively. The mean and standard error of the gap for 100 replications are summarized in Table 3. Because MICE can not provide an estimation of the distribution directly, we use instead the sample distribution from its imputation. The table shows that DPMCPM have better correlation estimation than other methods under all missing mechanisms.

Table 3. Mean and standard error of the gap between the estimated and true correlation matrix under different missing mechanisms.

	DPMCPM	DPMPM	MICE
MCAR	7.5968(2.216)	9.2249(2.656)	11.8448(2.088)
MAR	7.8092(2.366)	8.8777(2.279)	12.1515(2.487)
MNAR	7.1693(2.287)	9.3119(2.549)	11.7781(2.395)

3.2 Case 2: nonlinear association

In addition to the data generated from mixture models, we also perform simulation studies on more general cases. In practice, it is cumbersome to design and simulate general categorical data distribution with large p due to the exponentially increasing size of the contingency table. Here we design a simple nonlinear experiment to demonstrate the power of DPMCPM in this aspect.

As an example, we generate three Bernoulli random variables, V_1, V_2 and V_3, as follows. Assume that $V_1 \sim Bernoulli(0.3)$ and $V_2 \sim Bernoulli(0.5)$ are independent, where $V \sim Bernoulli(p)$ denotes that V follows a Bernoulli distribution with probability $Pr(V = 1) = p$. Let V_3 be the output of the exclusive-or operator on V_1 and V_2 with probability 95% and be an independent Bernoulli random error, i.e., $V_3 \sim Bernoulli(0.5)$, with probability 5%. Essentially, this is a data model with strong nonlinear association. We set $n = 300$, $p = 3$ and the same missing mechanisms as in Section 3.1. The mean and standard error of the imputation accuracy in this study are summarized in Table 4. The results show that MICE does not capture the nonlinear association among variables while DPMCPM does it well. DPMCPM also performs better than DPMPM here.

4. APPLICATION ON RECOMMENDATION SYSTEM PROBLEM

In this section, we present a real application on a recommendation system problem. [6] contributed a dataset about the ratings of movies. The entire ratings table in [6] contains 20000263 ratings on 27278 movies from 138493 users. In real applications of missing data, the true values of missingness

Table 4. Mean and standard error of the imputation accuracy of three methods under different missing mechanisms. Data are generated from the nonlinear association model.

	DPMCPM	DPMPM	MICE
MCAR	0.8527(0.031)	0.7527(0.052)	0.5644(0.038)
MAR	0.8699(0.041)	0.7832(0.058)	0.5054(0.055)
MNAR	0.7935(0.060)	0.6945(0.072)	0.5179(0.041)

are always unknown. For the purpose of cross-validation, we extract a subset of data with low missingness such that we mask a high percent of them for evaluating the imputation accuracy. First, we select the movies which were rated by more than 25% users. Then we remain the users who rated more than 95% of the selected movies. The resulting dataset contains 68861 ratings on 38 movies from 1837 users, where the percentage of missingness is 1.35%. The data matrix *MovieRate* in the MMDai package is the resulted dataset.

Since the original ratings in [6] are ordinal data made on a 5-star scale with half-star increments (0.5 stars–5.0 stars), we consider two plans: (1) transform to binary data according to a cutoff (thus the data is pure categorial); (2) round up to 5-category data (thus the data is ordinal). To evaluate the performance of DPMCPM, we mask 40% of the ratings in *MovieRate* under MCAR, resulting in the incomplete "observed data", where the percentage of missing values is 40.81%. Predicting the unobserved favor of a user on a particular movie is equivalent to impute a missing value in incomplete data.

The analyses in Section 3 are repeated on this real dataset. As an example of recommendation system, we also compare with the classical SVD approximation algorithm for recommendation problems. The data *MovieRate* is masked and imputed for 100 times independently. The mean and standard error of the imputation accuracy are summarized in Table 5. To check the effect of the cutoff used to dichotomize the original ordinal data, we repeat the experiments by varying the cutoff from 4.0 to 3.5 and 3.0. Table 5 shows that DPMCPM has significantly higher imputation accuracy than other approaches on *MovieRate*. The performance improvement is not sensitive to the choice of the cutoff. The improved accuracy can be very helpful for reducing the cost of bad recommendations in real life.

Except for the imputation of missing values, we also explore the clustering structure among movies according to the fitting of DPMCPM. Figure 2 shows the heatmap plots of an input data from one of the 100 experiments (both unordered and ordered versions) and the true data *MovieRate*, where the rows correspond to users and the columns correspond to movies. The order is based on the latent class inferred by DPMCPM. The figure demonstrated that DPMCPM could detect the cluster structure efficiently despite the high missingness in the input data.

To demonstrate the usage of the estimated distribution from DPMCPM, we perform the Fisher exact test to test

Table 5. *Imputation accuracy comparison on the movie data.*

Cutoff	DPMCPM	DPMPM	MICE	SVD
Binary (Cutoff-4.0)	0.7598(0.002)	0.6805(0.003)	0.6859(0.002)	0.6765(0.004)
Binary (Cutoff-3.5)	0.8158(0.002)	0.7424(0.002)	0.7400(0.002)	0.7413(0.003)
Binary (Cutoff-3.0)	0.9075(0.001)	0.8602(0.002)	0.8445(0.002)	0.8599(0.002)
5-Categoty	0.5480(0.003)	0.4321(0.003)	0.4198(0.002)	0.4449(0.005)

Table 6. *The pairs of movies with the least and largest Fisher test p-values.*

| Movie 1	Movie 2	p-value
Star Wars: Episode IV (1977)	Star Wars: Episode V (1980)	0.00004
Star Wars: Episode V (1980)	Star Wars: Episode VI (1983)	0.00023
Star Wars: Episode IV (1977)	Star Wars: Episode VI (1983)	0.00069
The Fugitive (1993)	Jurassic Park (1993)	0.00086
Jurassic Park (1993)	Men in Black (1997)	0.00106
Twelve Monkeys (1995)	Ace Ventura: Pet Detective (1994)	1.00000
Apollo 13 (1995)	Fight Club (1999)	1.00000
Pulp Fiction (1994)	Mission: Impossible (1996)	1.00000
Pulp Fiction (1994)	Independence Day (1996)	1.00000
Speed (1994)	Fight Club (1999)	1.00000
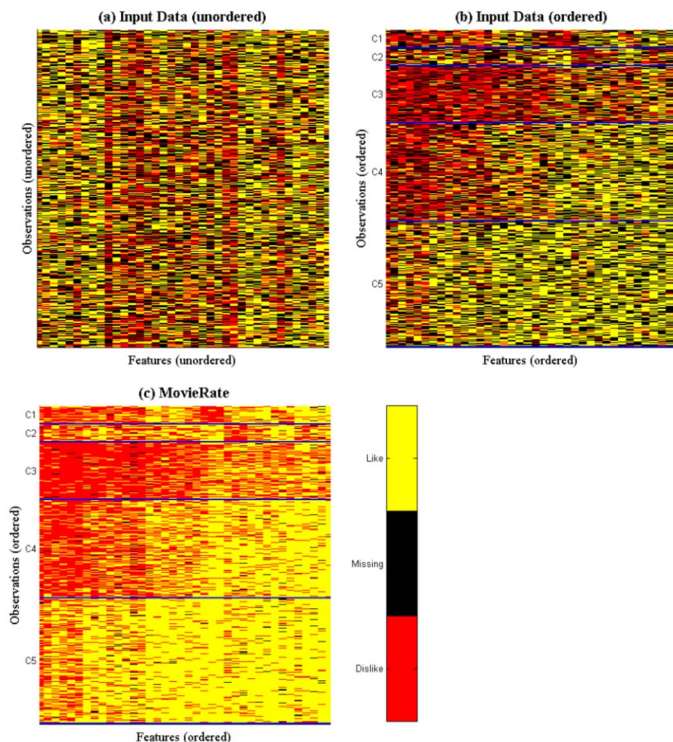
Figure 2. *Heatmap plots of an unordered input Data, the input data ordered according to the model fitting by DPMCPM, and the true data* MovieRate *with the same ordering.*

the independence of each pair of movies according to the estimated joint distribution. The five pairs with the least and largest p-values are reported in Table 6. The association results agree with common sense based on the nature of these movies.

5. CONCLUSION

In this paper, we introduce the DPMCPM method, which implemented multivariate multinomial distribution estimation for categorical data with missing values. DPMCPM inherits some nice properties of DPMPM. It avoids the arbitrary setting of the number of mixture components by using the Dirichlet process. Also, DPMCPM is applicable for any categorical data regardless of the true distribution and can capture the complex association among variables. Unlike DPMPM, DPMCPM can model general missing mechanisms and handle high missingness efficiently, thus DPMCPM can achieve more accurate results in empirical studies. Through simulation studies and a real application, we demonstrate that DPMCPM performs better statistical inference and imputation than other methods.

It shall be noted that approximating a general distribution of high-dimensional categorical data is still a hard task, although DPMCPM could approximate the true general distribution at arbitrary precision theoretically if n is big enough. To estimate the underlying general distribution accurately, the required sample size n still increases exponentially with the number of variables p. Thus our method may be limited when p grows with n. Fortunately, in many applications, there is underlying cluster structures in the dataset. In these cases, the underlying distribution is close to a mixture of product-multinomials, thus the sample size needed for DPMCPM to estimate the underlying distribution accurately is much smaller. For example, in recommendation system problems, it is reasonable to assume that the customers with similar historical activities have similar

tastes. For handling such problems, DPMCPM would have more advantages than traditional approaches. Future work shall try to improve the efficiency for the case that there is no underlying cluster structures in the high-dimensional dataset.

APPENDIX A. PROOF

A.1 Theorem 1

Proof. Here is a constructive proof for Theorem 1.

Assume that there are p variables and the j-th variable has d_j categories. Let $\pi_{c_1,\cdots,c_p} = p(X_1 = c_1, \cdots, X_p = c_p)$ then $\sum_{c_1=1}^{d_1} \cdots \sum_{c_p=p}^{d_p} \pi_{c_1,\cdots,c_p} = 1$.

For all missing mechanisms belong to MCAR, MAR or MNAR family, the missing rate can be written as $p(\boldsymbol{r}|\boldsymbol{x}) = \prod_{i=1}^{n} \prod_{j=1}^{p} p(r_{ij}|\boldsymbol{x}_i)$, which is a function of $\boldsymbol{x}_i$. For each variable, let

$$q_{c_1,\cdots,c_p}^{(j)} = p(r_{ij} = 0|x_{i1} = c_1, \cdots, x_{ip} = c_p),$$

For any joint distribution $\boldsymbol{\Pi}$ with missing mechanism $\mathbf{Q}$ where $\mathbf{Q} = \{q_{c_1,\cdots,c_p}^{(j)}\}$ is a table with $d_1 \times \cdots \times d_p \times p$ cells, we can estimate a new joint probability table by adding an extra category to denote missingness,

$$\tilde{\boldsymbol{\Pi}} = \tilde{\theta}_1 \cdot \tilde{\boldsymbol{\psi}}_1^{(1)} \otimes \cdots \otimes \tilde{\boldsymbol{\psi}}_1^{(p)} + \cdots + \tilde{\theta}_k \cdot \tilde{\boldsymbol{\psi}}_k^{(1)} \otimes \cdots \otimes \tilde{\boldsymbol{\psi}}_k^{(p)}$$

where $\tilde{\boldsymbol{\psi}}_h^{(j)} = (\tilde{\psi}_{h0}^{(j)}, \tilde{\psi}_{h1}^{(j)}, \cdots, \tilde{\psi}_{hd_j}^{(j)})$.

The problem we need to prove is, there exists a set of parameters $\{\tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Psi}}\}$ that can recover the original joint distribution table $\boldsymbol{\Pi}$ and missing mechanism $\mathbf{Q}$ after scaling the extra category in the latent classes. In another word, given $\boldsymbol{\Pi}$ and $\mathbf{Q}$, there exists a set of solution $\{\tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Psi}}\}$ under our estimates.

Note that for any general cases, we always have following decomposition

$$\boldsymbol{\Pi} = \sum_{c_1,\cdots,c_p} \pi_{c_1,\cdots,c_p} \cdot I_{c_1} \otimes \cdots \otimes I_{c_p}.$$

Define a series of latent classes $\mathcal{H} = \{\overline{c_1,\cdots,c_p}\}$ where $\overline{c_1,\cdots,c_p}$ is a notation of latent class. Here we design a set of parameters $\{\tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Psi}}\}$:

$$\tilde{\boldsymbol{\Theta}} = \{\tilde{\theta}_h : h \in \mathcal{H}\}, \quad \text{where} \quad \tilde{\theta}_{\overline{c_1,\cdots,c_p}} = \pi_{c_1,\cdots,c_p},$$

$$\tilde{\boldsymbol{\Psi}} = \{\tilde{\boldsymbol{\psi}}_h^{(j)} : h \in \mathcal{H}\},$$

where $\tilde{\boldsymbol{\psi}}_h^{(j)} = (\tilde{\psi}_{h0}^{(j)}, \tilde{\psi}_{h1}^{(j)} \cdots, \tilde{\psi}_{hd_j}^{(j)})$ is a $d_j + 1$-dimension vector. For the vector $\tilde{\boldsymbol{\psi}}_{\overline{c_1,\cdots,c_p}}^{(j)}$, we have $\tilde{\psi}_{\overline{c_1,\cdots,c_p}0}^{(j)} = q_{c_1,\cdots,c_p}^{(j)}$ and $\tilde{\psi}_{\overline{c_1,\cdots,c_p}c_j}^{(j)} = 1 - q_{c_1,\cdots,c_p}^{(j)}$ with other elements are 0.

Then we re-scale the extra category $\boldsymbol{\psi}_h^{(j)} = (\psi_{h1}^{(j)}, \cdots, \psi_{hd_j}^{(j)})$ where $\psi_{hc}^{(j)} = \tilde{\psi}_{hc}^{(j)}/(1 - \tilde{\psi}_{h0}^{(j)})$. Thus,

we have $\boldsymbol{\psi}_{\overline{c_1,\cdots,c_p}}^{(j)} = I_{c_j}$ where I_{c_j} is a vector with length d_j that the c_j-th element is 1 and other elements are 0.

Under this $\{\tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Psi}}\}$ setting, the joint distribution is

$$\sum_{h \in \mathcal{H}} \tilde{\theta}_h \cdot \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}$$

$$= \sum_{\overline{c_1,\cdots,c_p}} \pi_{c_1,\cdots,c_p} \cdot I_{c_1} \otimes \cdots \otimes I_{c_p} = \boldsymbol{\Pi}.$$

In this decomposition, the posterior probability of latent class

$$p(z_i = \overline{c_1,\cdots,c_p}|x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$= p(z_i = \overline{c_1,\cdots,c_p}, x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$/p(x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$= \pi_{c_1,\cdots,c_p}/\pi_{c_1,\cdots,c_p} = 1.$$

For the missing mechanism, on the level of latent class, we have

$$p(r_{ij} = 1|z_i = \overline{c_1,\cdots,c_p}, x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$= \tilde{\psi}_{\overline{c_1,\cdots,c_p}c_j}^{(j)},$$

then

$$p(r_{ij} = 0|z_i = \overline{c_1,\cdots,c_p}, x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$= 1 - p(r_{ij} = 1|z_i = \overline{c_1,\cdots,c_p}, x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$= 1 - \tilde{\psi}_{\overline{c_1,\cdots,c_p}c_j}^{(j)} = \tilde{\psi}_{\overline{c_1,\cdots,c_p}0}^{(j)}$$

Thus, integrating the level of latent class, we have

$$p(r_{ij} = 0|x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$= \sum_{h \in \mathcal{H}} p(r_{ij} = 0|z_i = h, x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$\cdot p(z_i = h|x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$= p(r_{ij} = 0|z_i = \overline{c_1,\cdots,c_p}, x_{i1} = c_1, \cdots, x_{ip} = c_p)$$
$$= \tilde{\psi}_{\overline{c_1,\cdots,c_p}0}^{(j)} = q_{c_1,\cdots,c_p}^{(j)}.$$

So we show that any joint distribution with general missing mechanism can be captured by our latent class model. Proven.　□

Remark. This proof is a constructive proof. We show that there exists a construction of the latent class model that can capture any general missing mechanisms under any joint distributions. In specific cases, this construction may be not unique since decomposition of joint distribution is multiple, and not optimal in terms of complexity of model. Actually, this proof shows that the estimate can always attain the true joint distribution with right missing mechanism, which is the maximum a posteriori estimate when sample size n is large enough.

REFERENCES

[1] AUDIGIER, V., HUSSON, F. and JOSSE, J. (2017). MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing* **27** 501–518. MR3599686

[2] DONDERS, A. R. T., VAN DER HEIJDEN, G. J., STIJNEN, T. and MOONS, K. G. (2006). Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* **59** 1087–1091.

[3] DUNSON, D. B. and XING, C. (2012). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104** 1042–1051. MR2562004

[4] ELMORE, R. T. and WANG, S. (2003). Identifiability and estimation in finite mixture models with multinomial coefficients. Technical Report, 03-04, Penn State University.

[5] FORMANN, A. K. (2007). Mixture analysis of multivariate categorical data with covariates and missing entries. *Computational Statistics & Data Analysis* **51** 5236–5246. MR2370868

[6] HARPER, F. M. and KONSTAN, J. A. (2016). The movielens datasets: history and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* **5** 19.

[7] HICKS, S. C., TOWNES, F. W., TENG, M. and IRIZARRY, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19** 562-578. MR3867412

[8] HU, J., REITER, J. P., WANG, Q. et al. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis* **13** 183–200. MR3737948

[9] JOLANI, S., DEBRAY, T. P., KOFFIJBERG, H., VAN BUUREN, S. and MOONS, K. G. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine* **34** 1841–1863. MR3334696

[10] JOSSE, J., HUSSON, F. et al. (2016). missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software* **70** 1–31.

[11] KUHA, J., KATSIKATSOU, M. and MOUSTAKI, I. (2018). Latent variable modelling with non-ignorable item non-response: multigroup response propensity models for cross-national analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **181** 1169–1192. MR3876387

[12] MANRIQUE-VALLIER, D. and REITER, J. P. (2017). Bayesian simultaneous edit and imputation for multivariate categorical data. *Journal of the American Statistical Association* **112** 1708–1719. MR3750893

[13] MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models.* John Wiley & Sons, New York. MR1789474

[14] MURRAY, J. S. et al. (2018). Multiple imputation: a review of practical and theoretical findings. *Statistical Science* **33** 142–159. MR3797707

[15] MURRAY, J. S. and REITER, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association* **111** 1466–1479. MR3601702

[16] RICCI, F., ROKACH, L. and SHAPIRA, B. (2015). *Recommender Systems Handbook.* Springer, New York.

[17] RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

[18] RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, New York. MR0899519

[19] SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data.* CRC Press, FL. MR1692799

[20] SCHAFER, J. L. and GRAHAM, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods* **7** 147.

[21] SI, Y. and REITER, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics* **38** 499–521.

[22] TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101** 1566–1581. MR2279480

[23] VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software* **45**.

[24] VAN BUUREN, S. and OUDSHOORN, K. (1999). *Flexible Mutlivariate Imputation by MICE.* TNO, Leiden.

[25] VAN BUUREN, S., BRAND, J. P., GROOTHUIS-OUDSHOORN, C. G. and RUBIN, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76** 1049–1064. MR2307507

[26] VAN DER PALM, D. W., VAN DER ARK, L. A. and VERMUNT, J. K. (2015). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification* 1–21. MR3503203

[27] VERMUNT, J. K., VAN GINKEL, J. R., DER ARK, V., ANDRIES, L. and SIJTSMA, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology* **38** 369–397.

[28] VIDOTTO, D., VERMUNT, J. K. and KAPTEIN, M. C. (2014). Multiple imputation of missing categorical data using latent class models: state of art. *Psychological Test and Assessment Modeling.*

[29] VIDOTTO, D., VERMUNT, J. K. and VAN DEUN, K. (2018). Bayesian multilevel latent class models for the multiple imputation of nested categorical data. *Journal of Educational and Behavioral Statistics* **43** 511–539.

[30] WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation* **36** 45–54. MR2370888

[31] WHITE, I. R., ROYSTON, P. and WOOD, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* **30** 377–399. MR2758870

[32] XIE, X. and MENG, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica* 1485–1545. MR3701490

Chaojie Wang
Faculty of Science
Jiangsu University
Zhenjiang, China
Department of Statistics
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
E-mail address: wang910930@gmail.com

Linghao Shen
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
E-mail address: markshen1991@163.com

Han Li
Department of Risk Management and Insurance
Shenzhen University
Shenzhen, China
E-mail address: hli@szu.edu.cn

Xiaodan Fan
Department of Statistics
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
E-mail address: xfan@cuhk.edu.hk