Bayesian Multiple Imputation for Categorical Data via Latent Class Models

Quanli Wang, Daniel Manrique-Vallier, Jerome P. Reiter, Jingchen Hu

1 Introduction

Many data sets comprise large numbers of exclusively categorical variables subject to item nonresponse. Faced with such item nonresponse, one approach is multiple imputation (Rubin 1987), in which the missing items are filled in by sampling repeatedly from predictive distributions. This creates $M > 1$ completed datasets that can be analyzed or disseminated to others.

This package implements a fully Bayesian, joint modeling approach to multiple imputation for categorical data based on latent class models. The idea is to model the implied contingency table of the categorical variables as a mixture of independent multinomial distributions, estimating the mixture distributions with truncated Dirichlet process prior distributions. Mixtures of multinomials can describe arbitrarily complex dependencies and are computationally expedient, so that they are effective general purpose multiple imputation engines. In contrast to other approaches based on loglinear models or chained equations, the mixture models avoid the need to specify (potentially many) models, which can be a time consuming task with no guarantee of a theoretically coherent set of models. The package also allows for structural zeros, i.e., certain combinations of variables are not possible *a priori*. For example, in the United States it is impossible for children to be married. The package is based on the models described in Si and Reiter (2013) and Manrique-Vallier and Reiter (2014). We note that the package assumes that data are missing at random.

The package includes imputation routines for two settings. The first routine is for data without any structural zeros, i.e., impossible combinations of variables known to have probability zero (Si and Reiter 2013). The second routine is for data with structural zeros (Manrique-Vallier and Reiter 2014). The second routine is computationally more intensive than the first routine. Hence, we advise analysts to use the first routine when possible.

The input is a dataset including only categorical variables with missing values, as well as an optional set of structural zeros. The output includes M multiply-imputed versions of the completed dataset, where the user selects M. Other aspects of the model fitting also are available, as described in the accompanying R help files.

2 The Imputation Model

Suppose that we have a sample of n individuals measured on J categorical variables. Each individual has an associated response vector $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iJ})$, whose components take values from a set of L_j levels. For convenience, we label these levels using consecutive numbers, $x_{ij} \in \{1, ..., L_j\}$, so that $\mathbf{x}_i \in \mathcal{C} = \prod_{j=1}^{J}\{1, ..., L_j\}$. Note that $\mathcal{C}$ includes all combinations of the J variables, including structural zeros, and that each combination $\mathbf{x}$ can be viewed as a cell in the contingency table formed by $\mathcal{C}$. Let $\mathbf{x}_i = (\mathbf{x}_i^{obs}, \mathbf{x}_i^{mis})$, where $\mathbf{x}_i^{obs}$ includes the variables with observed values and $\mathbf{x}_i^{mis}$ includes the variables with missing values. Finally, let $S = \{s_1, \ldots, s_C\}$, where $s_c \in \mathcal{C}$ and $c = 1, \ldots, C < |S|$, be the set of structural zero cells, i.e., $\Pr(\mathbf{x}_i \in S) = 0$.

2.1 Latent Class Model: No Structural Zeros

As an initial step, we describe the Bayesian latent class model without any concerns for structural zeros and without any missing data, i.e., $\mathbf{x}_i = \mathbf{x}_i^{obs}$. This model is a finite mixture of product-multinomial distributions,

$$p(\mathbf{x} \mid \boldsymbol{\lambda}, \boldsymbol{\pi}) = f^{LCM}(\mathbf{x}|\boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_j], \tag{1}$$

where $\boldsymbol{\lambda} = (\lambda_{jk}[l])$, with all $\lambda_{jk}[l] > 0$ and $\sum_{l=1}^{L_j} \lambda_{jk}[l] = 1$. Here, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ with $\sum_{k=1}^{K} \pi_k = 1$. This model corresponds to the generative process,

$$x_{ij} \mid z_i \overset{indep}{\sim} \text{Discrete}_{1:_{L_j}}(\lambda_{jz_i}[1], \ldots, \lambda_{jz_i}[L_j]) \quad \text{for all } i \text{ and } j \tag{2}$$

$$z_i \mid \boldsymbol{\pi} \overset{iid}{\sim} \text{Discrete}_{1:K}(\pi_1, \ldots, \pi_K) \quad \text{for all } i. \tag{3}$$

As notation, let $(\mathcal{X}, \mathcal{Z})$ be a sample of n variates obtained from this process, with $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathcal{Z} = (z_1, \ldots, z_n)$. For K large enough, (1) can represent arbitrary joint distributions for $\mathbf{x}$. And, using the conditional independence representation in (2) and (3), the model can be estimated and simulated from efficiently even for large J.

For prior distributions on $\boldsymbol{\pi}$, we use

$$\lambda_{jk}[\cdot] \overset{indep}{\sim} \text{Dirichlet}(\mathbf{1}_{L_j}) \tag{4}$$

$$\pi_k = V_k \prod_{h<k}(1 - V_h) \tag{5}$$

$$V_k \overset{iid}{\sim} \text{Beta}(1, \alpha) \text{ for } k = 1, \ldots, K-1; V_K = 1 \tag{6}$$

$$\alpha \sim \text{Gamma}(.25, .25) \tag{7}$$

The prior distributions in (4) are equivalent to uniform distributions over the support of the $J \times K$ multinomial conditional probabilities and hence represent vague prior knowledge. The

prior distribution for $\boldsymbol{\pi}$ in (5) – (7) is an example of a finite-dimensional stick-breaking prior distribution. It typically allocates $\mathcal{Z}$ to fewer than K classes, thereby reducing computation and avoiding over-fitting.

Accounting for missing data is straightforward in the MCMC. Conditional on draws of the parameters and latent class indicators, each value in $\mathbf{x}_i^{mis}$ is independent of $\mathbf{x}_i^{obs}$. Thus, one can impute any missing x_{ij} from the corresponding multinomial distribution in (2). Given a completed dataset, one then updates the parameters and latent class indicators from their full conditional distributions. For details, see Si and Reiter (2013).

To obtain M completed datasets for use in multiple imputation, analysts select M of the sampled $\mathbf{x}^{mis}$ after convergence of the Gibbs sampler. These datasets are spaced sufficiently so as to be approximately independent (given $\mathbf{x}^{obs}$). This involves thinning the MCMC samples so that the autocorrelations among parameters are close to zero.

2.2 Truncated Latent Class Model: Some Structural Zeros

The latent class model in (1) does not naturally specify cells with structural zeros *a priori*, because it assumes a positive probability for each cell. Thus, to represent tables with structural zeros, we need to truncate the model so that

$$f^{TLCM}(\mathbf{x} \mid \boldsymbol{\lambda}, \boldsymbol{\pi}, S) \propto 1\{\mathbf{x} \notin S\} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_j]. \tag{8}$$

Obtaining samples from the posterior distribution of parameters $(\boldsymbol{\lambda}, \boldsymbol{\pi})$, conditional on a sample $\mathcal{X}^1 = (\mathbf{x}_1, ..., \mathbf{x}_n)$, can be greatly facilitated by adopting a sample augmentation strategy. We consider $\mathcal{X}^1$ to be the portion of variates that did not fall into the set S from a larger sample, $\mathcal{X}$, generated directly from (1). Let n_0, $\mathcal{X}^0$, and $\mathcal{Z}^0$ be the the (unknown) sample size, response vectors, and latent class labels for the portion of $\mathcal{X}$ that did fall into S. If $p(N) \propto 1/N$, where $N = n_0 + n$, the posterior distribution of $(\boldsymbol{\lambda}, \boldsymbol{\pi})$ under the truncated model can be obtained by integrating the posterior distribution under the augmented sample model over $(n_0, \mathcal{X}^0, \mathcal{Z}^0, \mathcal{Z}^1)$.

This package converts the model into a multiple imputation engine when some items are missing at random. The basic strategy uses a Gibbs sampler. Given a completed dataset $(\mathbf{x}^{obs}, \mathbf{x}^{mis})$, we take a draw of the parameters using the algorithm from Manrique-Vallier (2014). Given a draw of the parameters, we take a draw of $\mathbf{x}^{mis}$, trimming the support of the full conditional distribution of x_{ij} from $\{1, ..., L_j\}$ to only values that avoid $\mathbf{x}_i \in S$, given current values of $\{x_{ij'} : \text{all } j' \neq j\}$.

To obtain M completed datasets for use in multiple imputation, analysts select M of the sampled $\mathbf{x}^{mis}$ after convergence of the Gibbs sampler. These datasets are spaced sufficiently so as to be approximately independent (given $\mathbf{x}^{obs}$). This involves thinning the MCMC samples so that the autocorrelations among parameters are close to zero.

References

1. Manrique-Vallier, D. and Reiter, J. P. (2014), "Bayesian multiple imputation for large-scale categorical data with structural zeros," *Survey Methodology*, 40, 125 - 134.

2. 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.