Bayesian Analysis (2016) **11**, Number 2, pp. 499–518

Incorporating Marginal Prior Information in Latent Class Models

Tracy A. Schifeling* and Jerome P. Reiter†

Abstract. We present an approach to incorporating informative prior beliefs about marginal probabilities into Bayesian latent class models for categorical data. The basic idea is to append synthetic observations to the original data such that (i) the empirical distributions of the desired margins match those of the prior beliefs, and (ii) the values of the remaining variables are left missing. The degree of prior uncertainty is controlled by the number of augmented records. Posterior inferences can be obtained via typical MCMC algorithms for latent class models, tailored to deal efficiently with the missing values in the concatenated data. We illustrate the approach using a variety of simulations based on data from the American Community Survey, including an example of how augmented records can be used to fit latent class models to data from stratified samples.

Keywords: categorical, Dirichlet process, missing, mixture, stratified, survey.

1 Introduction

Mixtures of products of multinomial distributions, also known as latent class models (Goodman, 1974), are used to model multivariate categorical data in many areas of application, including, for example, genomics (Dunson and Xing, 2009), marketing (Kamakura and Wedel, 1997), and political science (Si et al., 2015). They also serve as engines for multiple imputation of missing data (Vermunt et al., 2008; Gebregziabher and DeSantis, 2010; Si and Reiter, 2013; Manrique-Vallier and Reiter, 2014b). The defining feature of latent class models is an assumption of latent conditional independence: within any class the variables follow independent multinomial distributions. This conditional independence makes latent class models particularly useful for contingency tables with large numbers of cells, as the models can capture complex dependence structures automatically. Bayesian versions of latent class models can be efficiently estimated with MCMC algorithms (Ishwaran and James, 2001; Dunson and Xing, 2009; Jain and Neal, 2004).

In many settings amenable to latent class modeling, the analyst may have informative prior beliefs about the distributions of subsets of the variables. For example, the analyst may know with high precision the distributions of demographic variables from external sources, such as censuses or large national surveys. This information could come in the form of joint distributions (e.g., the probabilities of all combinations of gender and race), conditional distributions (e.g., the probabilities of all combinations of race

*Dept. of Statistical Science, Box 90251, Duke University, Durham, NC, 27708, USA, tracy@stat.duke.edu

†Dept. of Statistical Science, Box 90251, Duke University, Durham, NC, 27708, USA, jerry@stat.duke.edu

DOI: 10.1214/15-BA959

given gender), or univariate marginal distributions (e.g., the probabilities for all combinations of race and all combinations of gender, separately). It is not obvious how to incorporate such prior information in Bayesian latent class models because the implied marginal probabilities are tensor products. One approach is the marginally specified prior distribution of Kessler et al. (2015). However, as Kessler et al. (2015) admit, the approximations in this approach can be computationally expensive to implement.

In this article, we propose a simple, yet highly flexible method for incorporating prior information in Bayesian latent class models. The basic idea is to append synthetic observations to the original data such that (i) the empirical distributions of the desired margins match those in the prior beliefs, and (ii) the values for the remaining variables are left completely missing. For example, to add prior information reflecting that 50% of individuals are female, we can append hypothetical records with only gender recorded and all other variables missing, ensuring that half the augmented records have female for gender. The number of added records is a function of the desired level of prior precision: increasing numbers of records implies increasing certainty in the prior marginal probabilities. After adding the hypothetical records, we estimate the latent class model on the concatenated data with MCMC algorithms. For margins with values in the augmented records, the posterior distribution of the corresponding marginal probabilities is pulled toward the empirical distributions in the augmented records. However, adding the augmented data does not distort conditional distributions of the remaining variables (given the variables with augmented data), since by design the augmented data do not offer information about these conditional distributions. Indeed, as we illustrate, because of this feature the augmented records can be leveraged to correct estimates of the joint distribution of all variables for informative sampling.

Our approach to expressing informative prior distributions is related to the approaches suggested in Greenland (2007) and Kunihama and Dunson (2013). Greenland (2007) adds synthetic records to encode a prior distribution for relative risks, and Kunihama and Dunson (2013) represent prior information by generating pseudo-records with values for all variables using pre-specified, generalized linear models. Unlike these methods, by adding partially complete records our approach allows analysts to encode prior information for arbitrary sets of margins.

The remainder of the article is organized as follows. In Section 2, we briefly review the particular latent class model that we use, which is a truncated version of the Dirichlet process mixture of product multinomials model (DPMPM) developed by Dunson and Xing (2009). In Section 3, we present results of simulations illustrating the augmented record approach, including a discussion of how many records to add. In Section 4, we describe how augmented records can be used to account for disproportionate sampling rates in stratified simple random samples. In Section 5, we conclude with a brief discussion of other applications of the augmented record approach.

2 Review of the DPMPM

In describing the DPMPM, we closely follow the presentation in Si and Reiter (2013). Suppose the data comprise n individuals measured on p categorical variables. Let X_{ij}

be the value of variable j for individual i, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $X_i = (X_{i1}, \ldots, X_{ip})$. Without loss of generality, we assume that the possible values of X_{ij} are in $\{1, \ldots, d_j\}$, where $d_j \geq 2$ is the total number of categories for variable j. Let D be the contingency table formed from all levels of all p variables, so that D has $d = d_1 \times d_2 \times \cdots \times d_p$ cells. We denote each cell in D as $(c_1, \ldots, c_p)$, where each $c_j \in \{1, \ldots, d_j\}$. For all cells in D, let $\theta_{c_1,\ldots,c_p} = \Pr(X_{i1} = c_1, \ldots, X_{ip} = c_p)$ be the probability that individual i is in cell $(c_1, \ldots, c_p)$. We require the $\sum_D \theta_{c_1,\ldots,c_p} = 1$. Let $\theta = \{\theta_{c_1,\ldots,c_p} : c_j \in (1, \ldots, d_j), j = 1 \ldots, p\}$ be the collection of all d cell probabilities.

We suppose that each individual i belongs to exactly one of H^* latent classes. For $i = 1, \ldots, n$, let $z_i \in \{1, \ldots, H^*\}$ indicate the class of individual i, and let $\pi_h = \Pr(z_i = h)$. We assume that $\pi = (\pi_1, \ldots, \pi_{H^*})$ is the same for all individuals. Within any class, we suppose that each of the p variables independently follows a class-specific multinomial distribution. This implies that individuals in the same latent class have the same cell probabilities. For any value x, let $\phi_{hjx} = \Pr(X_{ij} = x \mid z_i = h)$ be the probability of $X_{ij} = x$ given that individual i is in class h. Let $\phi = \{\phi_{hjx} : x = 1, \ldots, d_j, j = 1, \ldots, p, h = 1, \ldots, H^*\}$ be the collection of all ϕ_{hjx}. For prior distributions on π and ϕ, we use the truncated stick breaking representation of Sethuraman (1994).

Putting it all together, we have

$$\begin{aligned}
X_{ij}|z_i, \phi &\sim \text{Categorical}(\phi_{z_i j1}, \ldots, \phi_{z_i j d_j}) \quad \text{for all } i, j && (1)\\
z_i|\pi &\sim \text{Categorical}(\pi_1, \ldots, \pi_{H^*}) \quad \text{for all } i && (2)\\
\pi_h &= V_h \prod_{g<h} (1 - V_g) \quad \text{for } h = 1, \ldots, H^* && (3)\\
V_h &\sim \text{Beta}(1, \alpha) \quad \text{for } h = 1, \ldots, H^* - 1, \quad V_{H^*} = 1 && (4)\\
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) && (5)\\
\phi_{hj} = (\phi_{hj1}, \ldots, \phi_{hjd_j}) &\sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}) && (6)
\end{aligned}$$

where the Gamma distribution has mean a_α / b_α.

We set $a_{j1} = \cdots = a_{jd_j} = 1$ for all j to correspond to uniform distributions. Following Dunson and Xing (2009) and Si and Reiter (2013), we set $(a_\alpha = 0.25, b_\alpha = 0.25)$, which represents a small prior sample size and hence vague specification for the Gamma distribution. In practice, we find these specifications allow the data to dominate the prior distribution. The posterior distribution of all parameters can be estimated using a blocked Gibbs sampler (Ishwaran and James, 2001; Si and Reiter, 2013).

We recommend making H^* as large as possible while still offering fast computation. Using an initial proposal for H^*, say $H^* = 30$, analysts can examine the posterior distributions of the sampled number of unique classes across MCMC iterations to diagnose if H^* is large enough. Significant posterior mass at a number of classes equal to H^* suggests that more classes be added. We note that one can use other MCMC algorithms to estimate the posterior distribution that avoid truncation, for example, a slice sampler (Walker, 2007; Dunson and Xing, 2009; Kalli et al., 2009) or an exact blocked sampler (Papaspiliopoulos, 2008).

From (1) and (2), we can see that the probability of any cell $(c_1, \ldots, c_p) \in D$ can be expressed as

$$\theta_{c_1,\ldots,c_p} = \sum_{h=1}^{H^*} \pi_h \prod_{j=1}^{p} \phi_{hjc_j}. \tag{7}$$

Marginal probabilities are computed similarly, taking the product only over the values of j in the margin of interest. This expression reveals the challenge in specifying informative prior distributions for margins in θ: one has to influence both ϕ and π. One possibility is to fix $(a_{j1}, \ldots, a_{jd_j})$ to correspond to the desired prior probabilities with a very large prior sample size that dominates n—this would force the posterior marginal probability to equal the prior marginal probability for variable j. However, this could severely constrain the ability of the model to capture relationships among the other variables since the prior distribution would encourage the latent classes to be comprised of cases with empirical distributions that match the prior distribution.

3 Adding marginal information to the DPMPM model

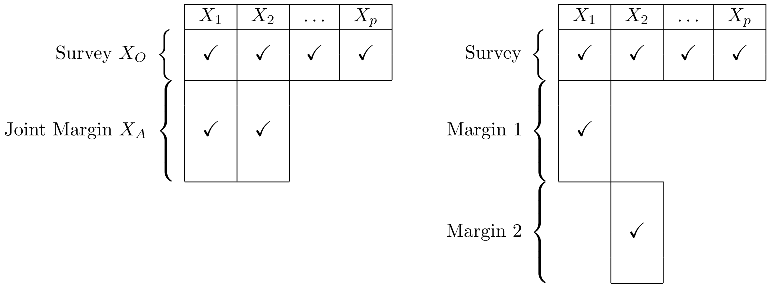
We now turn to the augmented records approach to incorporating prior information about marginal probabilities in the DPMPM model. Let A index the set of variables for which we have informative prior beliefs. Suppose that we create n_A cases to append to the original data. Let X_A include the hypothetically recorded data for the variables in A for the n_A augmented cases; data for all variables not in A are left missing for these cases. Let X_O include all the data for the n cases collected in the sample, and let $X_{obs} = (X_O, X_A)$ be the concatenated data. The exact format of X_{obs} depends on the information in A. When A includes the full joint distribution for $\{X_j : j \in A\}$, the analyst adds X_A as in Figure 1a. When A includes only univariate marginal distributions, the analyst adds augmented data comprising only marginal information for each variable $\{X_j : j \in A\}$, as in Figure 1b. In the latter case, different augmented sample sizes can be used for each margin depending on the levels of prior precision desired by the analyst.

Let $\Theta = \{z_1, \ldots, z_{n+n_A}, \pi, \alpha, \phi\}$. Treating X_A as if it were data, the likelihood function for the augmented data DPMPM is

$$p(X_{obs}|\Theta) = \left(\prod_{j \in A} \prod_{i=1}^{n+n_A} p(X_{ij}|z_i, \phi) \right) \left(\prod_{j \notin A} \prod_{i=1}^{n} p(X_{ij}|z_i, \phi) \right) \tag{8}$$

$$= \left(\prod_{j \in A} \prod_{h=1}^{H^*} \prod_{c_j=1}^{d_j} \phi_{hjc_j}^{\sum_{i=1, z_i=h}^{n+n_A} I(X_{ij}=c_j)} \right) \left(\prod_{j \notin A} \prod_{h=1}^{H^*} \prod_{c_j=1}^{d_j} \phi_{hjc_j}^{\sum_{i=1, z_i=h}^{n} I(X_{ij}=c_j)} \right). \tag{9}$$

Using the default prior distributions in (3)–(6), the posterior distribution of the parameters can be readily estimated with a Gibbs sampler; see Appendix A for the full



(a) Graphical representation of survey plus a joint distribution margin.

(b) Graphical representation of survey plus two disjoint margins.

Figure 1: Graphical representations of augmented surveys.

conditionals. The model allows the n_A additional records to be in any of the latent classes, favoring allocations that best describe X_{obs}.

When n_A is very large, it can be computationally expensive to update each augmented case's z_i one at a time. In many contexts, however, the number of unique combinations in X_A is substantially smaller than n_A; for example, there are two unique combinations when A includes only gender. To update all z_i for the augmented cases, we can compute the conditional probability (given X_A) for each unique combination. We then sample the values of z_i for all augmented records with the same combination at once using a multinomial distribution. When the number of unique combinations in X_A is large, it can be beneficial to update all z_i in parallel. One also can reduce computational burdens by using approximations to the full posterior distribution (e.g., as in Johndrow et al., 2014).

To illustrate the augmented sample approach and the role of n_A, we use three simulation scenarios. In the first scenario, we assume an analyst with very precise (essentially known) estimates of marginal probabilities. Here, we consider prior information comprising a bivariate distribution as in Figure 1a and information comprising two univariate margins as in Figure 1b. In the second scenario, we assume an analyst with imprecise estimates of marginal probabilities. Here, we only show results for prior information comprising a bivariate distribution. We use a small p in these two scenarios to facilitate repeated sampling studies. In the third scenario, we illustrate the approach for a larger p.

For all simulations, and throughout the remainder of the article, we use data from the 2012 American Community Survey (ACS) Public Use Microdata Sample (PUMS) of North Carolina. We include only individuals with age greater than or equal to 18 to avoid structural zeros, i.e., impossible combinations like married five year old. The latent class model from Section 2 does not handle structural zeros correctly without adjustments; see Manrique-Vallier and Reiter (2014a) for an approach that does so. The resulting data comprise $N = 76706$ individuals and the variables in Table 1. In the following simulations, X_O and the information used to generate X_A both come from this ACS PUMS population. In practice, of course, the survey data in X_O and the marginal information for X_A typically come from different sources.

PUMS variable	Categories
Gender	1=male, 2=female
Age	1=18–29, 2=30–44, 3=45–59, 4=60+
Recoded detailed race code	1=White alone, 2=Black or African American alone, 3=American Indian alone, 4=other, 5=two or more races, 6=Asian alone
Educational attainment	1=less than high school diploma, 2=high school diploma or GED or alternative credential, 3=some college, 4=associate's degree or higher
Marital status	1=married, 2=widowed, 3=divorced, 4=separated, 5=never married
Language other than English spoken at home	1=yes speaks another language, 2=no speaks only English
World area of birth	1=US state, 2=PR and US island areas, Oceania and at sea, 3=Latin America, 4=Asia, 5=Europe, 6=Africa, 7=Northern America
Military service	1=yes active duty at some point, 2=no training for Reserves/National Guard only, 3=no never served in the military
When last worked	1=within the past 12 months, 2=1–5 years ago, 3=over 5 years ago or never worked
Disability recode	1=with a disability, 2=without
Health ins. coverage recode	1=with health insurance coverage, 2=no
Mobility status (lived here 1 year ago)	1=yes same house (non movers), 2=no outside US and PR, 3=no different house in US or PR
School enrollment	1=no has not attended in the last 3 months, 2=yes public school or public college, 3=yes private school or college or home school
Recoded detailed Hispanic origin	1=not Spanish/Hispanic/Latino, 2=Spanish/Hispanic/Latino

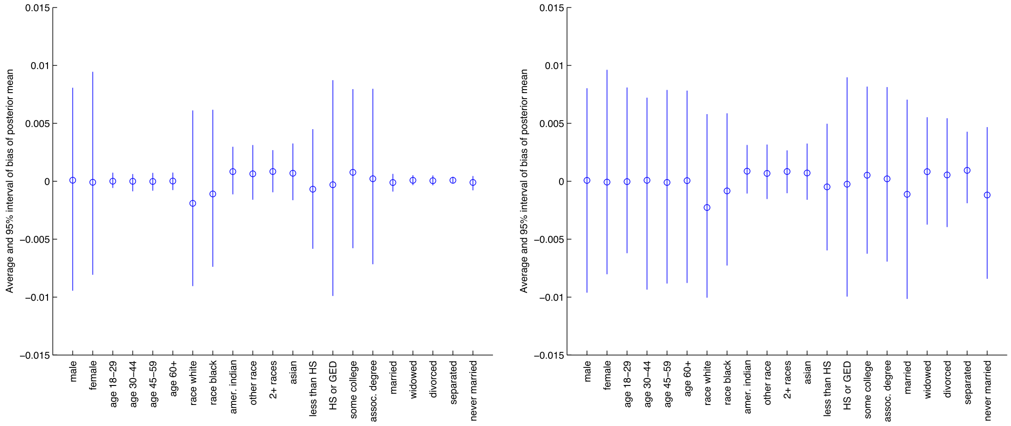
Table 1: Subset of variables from ACS PUMS 2012. Categories for age, race, educational attainment, world area of birth, military service, and Hispanic origin have been collapsed from their original number of levels due to insufficient sample sizes.

3.1 Scenario 1. Adding known margins

When the analyst knows some marginal probabilities precisely, the analyst should augment the sample with enough records so that $n_A \gg n$. As evident from (9), doing so ensures that the information about the marginal probabilities in A comes primarily from X_A. The empirical distributions in X_A are constructed to match the known marginal probabilities.

We illustrate this approach using a repeated sampling simulation, treating the N records in the ACS PUMS data as a population. Each X_O comprises $n = 10000$ randomly sampled individuals from the N records in the ACS PUMS. Each record is measured on $p = 5$ variables including gender, age group, race, educational attainment, and marital status, so that the implied contingency table has $d = 960$ cells. We consider an analyst who knows the joint distribution of age group and marital status in the population, which we take from the N records.

We augment each X_O with $n_A = 100000$ synthetic individuals, setting X_A so that the empirical frequencies of the cross tabulations of age group and marital status match



(a) With $n_A = 100,000$ augmented records.

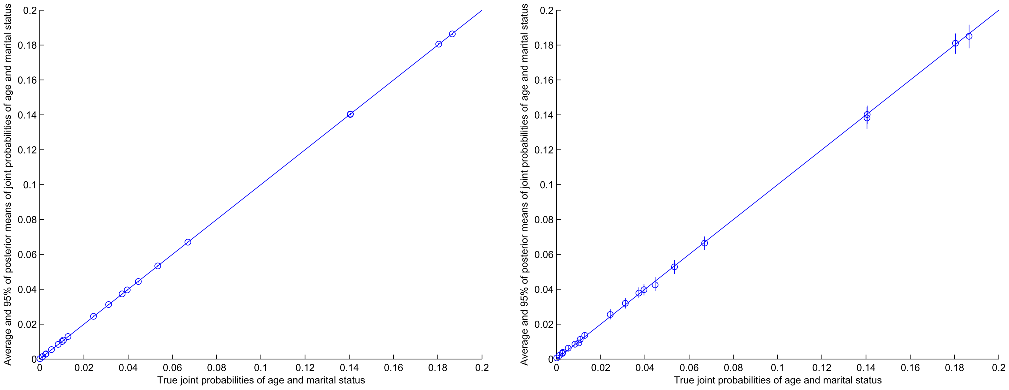
(b) No augmented records.

Figure 2: Distribution across the 100 simulations of differences in posterior means and corresponding population percentages for all marginal probabilities. Left panel displays results with the augmented joint margin of age group and marital status, and right panel displays results based on collected data only.

those from the known joint marginal probabilities. We run the DPMPM model on X_{obs} with $H^* = 30$, running three MCMC chains each for 50,000 iterations and tossing the first 20,000 as burn-in. We identified this number of MCMC iterates as sufficient based on exploratory runs using the diagnostics of Gelman and Rubin (1992) for α and all the univariate marginal probabilities. We repeat the process of generating X_{obs} and fitting the model 100 times. For comparison, we also fit the DPMPM on the 100 sampled X_O without any augmented records.

Figure 2 displays how adding X_A affects the estimates of univariate marginal probabilities. After adding the augmented data, the posterior means of the marginal probabilities for age group and marital status are very close to the frequencies in X_A (which equal the population percentages). In contrast, when the DPMPM is estimated using only X_O, the posterior means for the age group and marital status marginal probabilities are substantially more variable. Figure 3 shows similar patterns for the joint probabilities of age group and marital status. We note that in Figure 2, the posterior means for the marginal probabilities for variables not in A are similar whether or not one adds X_A.

Figure 4 displays the posterior means and corresponding population values for all 960 $\theta_{c_1,\ldots,c_p}$. The posterior means are quite similar whether or not one adds X_A. When not using X_A, the average root mean squared error (RMSE) of the posterior means is 3.8×10^{-4} with 95% of the RMSEs within $(3.2 \times 10^{-4}, 4.6 \times 10^{-4})$. When using X_A, the average RMSE is 3.8×10^{-4}, with 95% of RMSEs within $(3.1 \times 10^{-4}, 4.5 \times 10^{-4})$. These results indicate that using augmented data to represent prior beliefs on marginal probabilities does not distort other aspects of the posterior distribution of θ.



(a) With $n_A = 100,000$ augmented records.

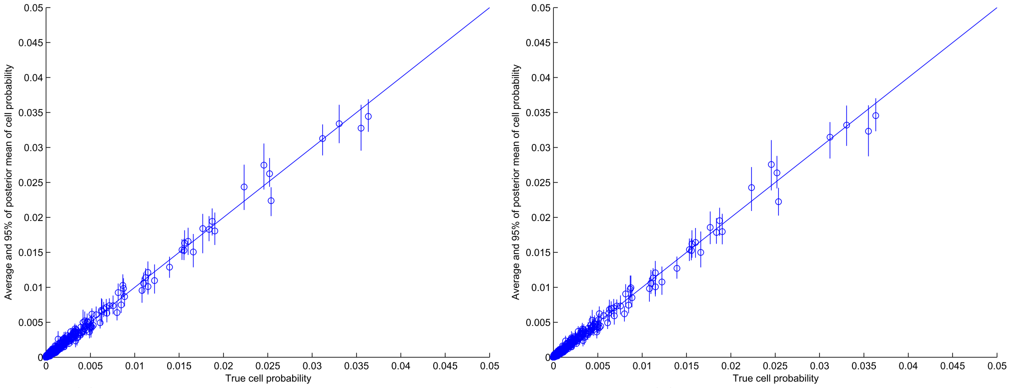
(b) No augmented records.

Figure 3: Distribution across the 100 simulations of posterior means versus corresponding population percentages for joint distribution of age group and marital status. Left panel displays results with the augmented joint margin of age group and marital status, and right panel displays results based on collected data only.

We also run 100 simulations where the analyst precisely knows the distributions of age group and marital status marginally but not jointly. Here, we add $n_A = 200000$ records as in Figure 1b, allocating 100000 to each margin. The results for the 21 univariate marginal probabilities are similar to those in Figure 2a, and the results for the 960 cell probabilities are similar to those in Figure 4a. When using X_A in this scenario, the average RMSE of the posterior means of the 960 probabilities is 3.9×10^{-4} with 95% of RMSEs within $(3.3 \times 10^{-4}, 4.7 \times 10^{-4})$. These RMSEs are not noticeably different from those in the simulation with known joint age–marital status distribution, although they tend to be slightly higher. However, the posterior probabilities in the joint distribution of age group and marital status exhibit variability that is substantially closer to that seen in Figure 3b than in Figure 3a. This is not surprising, as in this scenario X_A does not add information about the conditional distributions for age group and marital status. For brevity, we do not display the figures here.

3.2 Scenario 2. Adding imprecise margins

With imprecise margins, we no longer set $n_A \gg n$; instead, we allow n_A essentially to control the prior precision. Suppose that the analyst's prior beliefs for the probabilities in A are centered at some $\theta_A^{(0)}$. When adding augmented data for joint distributions as in Figure 1a, analysts can think of n_A as the prior sample size in a Dirichlet distribution with shape parameter $\theta_A^{(0)}$. When adding augmented data for marginal distributions only, analysts specify an augmented sample size for each margin separately. In both cases, the analyst can determine n_A by matching the mean and standard deviation in the prior information (e.g., reported estimates of means and standard errors from national surveys) to the first two moments of Dirichlet distributions. For example, Table 2



(a) With $n_A = 100,000$ augmented records.

(b) No augmented records.

Figure 4: Posterior mean estimates of cell probabilities versus corresponding population values for all 960 cells in the table. Left panel displays results with the augmented joint margin of age group and marital status, and right panel displays results based on collected data only.

displays approximate 95% prior intervals on π_j for each possible age group and marital status combination j for various n_A. See Appendix B for further discussion of the reasonableness of interpreting n_A as the prior sample size of a Dirichlet distribution.

To illustrate the incorporation of imprecise marginal information, we modify the simulation from Section 3.1. We add prior information on the joint distribution of age group and marital status, using a prior sample size of $n_A = 10000$. Results are summarized in Figure 5. As intended, the posterior intervals for the age group and marital status marginal and joint probabilities are wider than those estimated with $n_A = 100000$, yet narrower than those estimated with $n_A = 0$. The average RMSE of the 960 posterior means is again similar to the no-margin and precise-margin cases (the average is 3.7×10^{-4} with 95% of RMSEs between 3.2×10^{-4} and 4.4×10^{-4}).

3.3 Scenario 3. Adding information with larger p

We now use a random sample of $n = 10000$ records and the $p = 14$ variables in Table 1, which correspond to a contingency table with more than 8.7 million cells. We add $n_A = 99991$ (not a multiple of 1000 due to rounding considerations) augmented records with recorded multivariate responses to gender, age group, race, educational attainment, marital status, language other than English, and world area of birth. We construct the augmented data as follows. We compute the population percentage of each combination of these seven variables from the N ACS PUMS cases. For example, the population percentage of people who are male, age 18–29, of white race, have less than a high school diploma, are married, who speak another language other than English, and were born in a US state is 0.0039%. The cross-tabulation of these seven variables results in 13440 distinct sub-groups, which we allocate to the n_A cases approximately proportional to their population shares.

Age group, marital status	True percent	$n_A = 100000$	$n_A = 10000$	$n_A = 1000$
Age 18–29, married	3.96	(3.84, 4.08)	(3.61, 4.37)	(2.89, 5.24)
Age 18–29, widowed	0.01	(0.01, 0.02)	(0.00, 0.06)	(0, 0.35)
Age 18–29, divorced	0.29	(0.26, 0.32)	(0.20, 0.41)	(0.11, 0.85)
Age 18–29, separated	0.27	(0.24, 0.31)	(0.19, 0.40)	(0.11, 0.85)
Age 18–29, never married	14.04	(13.82, 14.26)	(13.40, 14.71)	(11.81, 16.04)
Age 30–44, married	14.05	(13.85, 14.26)	(13.37, 14.72)	(11.95, 16.07)
Age 30–44, widowed	0.13	(0.11, 0.16)	(0.08, 0.23)	(0.03, 0.54)
Age 30–44, divorced	2.43	(2.34, 2.54)	(2.13, 2.75)	(1.57, 3.55)
Age 30–44, separated	1.01	(0.95, 1.08)	(0.83, 1.23)	(0.51, 1.81)
Age 30–44, never married	5.34	(5.20, 5.46)	(4.90, 5.81)	(4.03, 6.79)
Age 45–59, married	18.04	(17.81, 18.27)	(17.28, 18.72)	(15.48, 20.39)
Age 45–59, widowed	0.84	(0.79, 0.89)	(0.68, 1.04)	(0.39, 1.49)
Age 45–59, divorced	4.47	(4.34, 4.59)	(4.12, 4.89)	(3.28, 5.88)
Age 45–59, separated	1.08	(1.01, 1.14)	(0.91, 1.30)	(0.63, 1.92)
Age 45–59, never married	3.12	(3.01, 3.23)	(2.79, 3.47)	(2.18, 4.31)
Age 60+, married	18.66	(18.43, 18.90)	(17.84, 19.43)	(16.30, 20.79)
Age 60+, widowed	6.70	(6.56, 6.87)	(6.20, 7.22)	(5.21, 8.21)
Age 60+, divorced	3.73	(3.61, 3.85)	(3.36, 4.09)	(2.68, 5.02)
Age 60+, separated	0.53	(0.49, 0.58)	(0.42, 0.69)	(0.23, 1.17)
Age 60+, never married	1.28	(1.21, 1.35)	(1.07, 1.51)	(0.73, 2.16)

Table 2: 95% prior intervals for margins corresponding to different values of n_A.

To investigate the effects of adding prior information, we examine the Cramér's V statistic for every pair of variables j and j'. This measures strength of bivariate associations. Figure 6a displays the Cramér's V statistic computed from the N observations in the ACS PUMS data. Dunson and Xing (2009) define a model-based version of Cramér's V statistic as

$$\rho^2_{jj'} = \frac{1}{\min\{d_j, d_{j'}\} - 1} \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \frac{(\theta_{c_j,c_{j'}} - \theta_{c_j}\theta_{c_{j'}})^2}{\theta_{c_j}\theta_{c_{j'}}} \tag{10}$$

where $\theta_{c_j,c_{j'}} = \sum_{h=1}^{H^*} \pi_h \phi_{h,j,c_j} \phi_{h,j',c_{j'}}$ and $\theta_{c_j} = \sum_{h=1}^{H^*} \pi_h \phi_{h,j,c_j}$.

We estimate each $\rho^2_{jj'}$ using the models fit to X_{obs} and only to X_0 using a posterior simulation approach, as done by Dunson and Xing (2009). For each analysis, we run three chains of the MCMC algorithm for 80000 iterations after a burn-in of 20000 iterations, and save every 30th draw. Figures 6b and 6c display the posterior means of $\rho_{jj'}$ for all pairs of variables. The posterior means of $\rho_{jj'}$ across the variables are similar to the population Cramér's V statistics whether we use X_{obs} or X_O alone. This would not have been the case if, for example, the augmented data encouraged the model to estimate accurately the distribution of the seven variables in the added margin at the expense of the remaining variables. Put another way, the fit based on X_{obs} is not shrunk toward independence relative to the fit based on X_O.

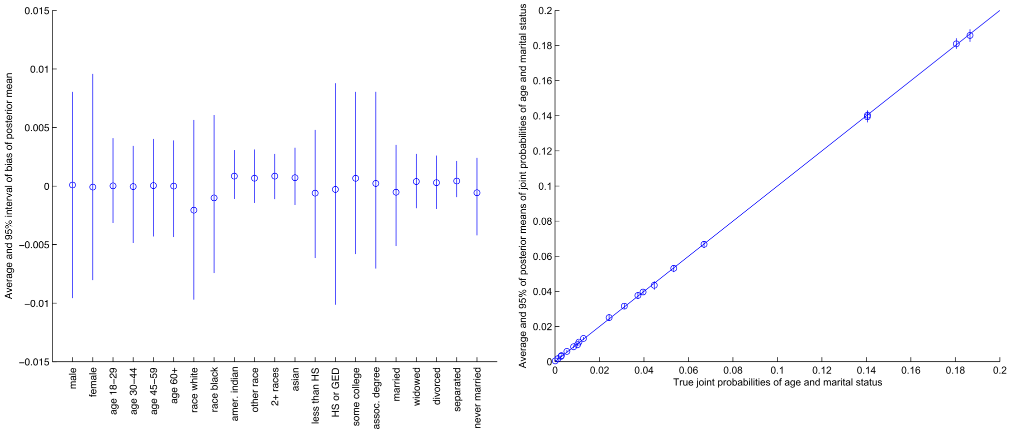


Figure 5: Results of 100 simulation runs when $n_A = 10000$. The left panel displays the distribution of differences in posterior means and corresponding population percentages for all univariate distributions, and the right panel displays the posterior means versus the corresponding population percentages for the joint distribution of age group and marital status.

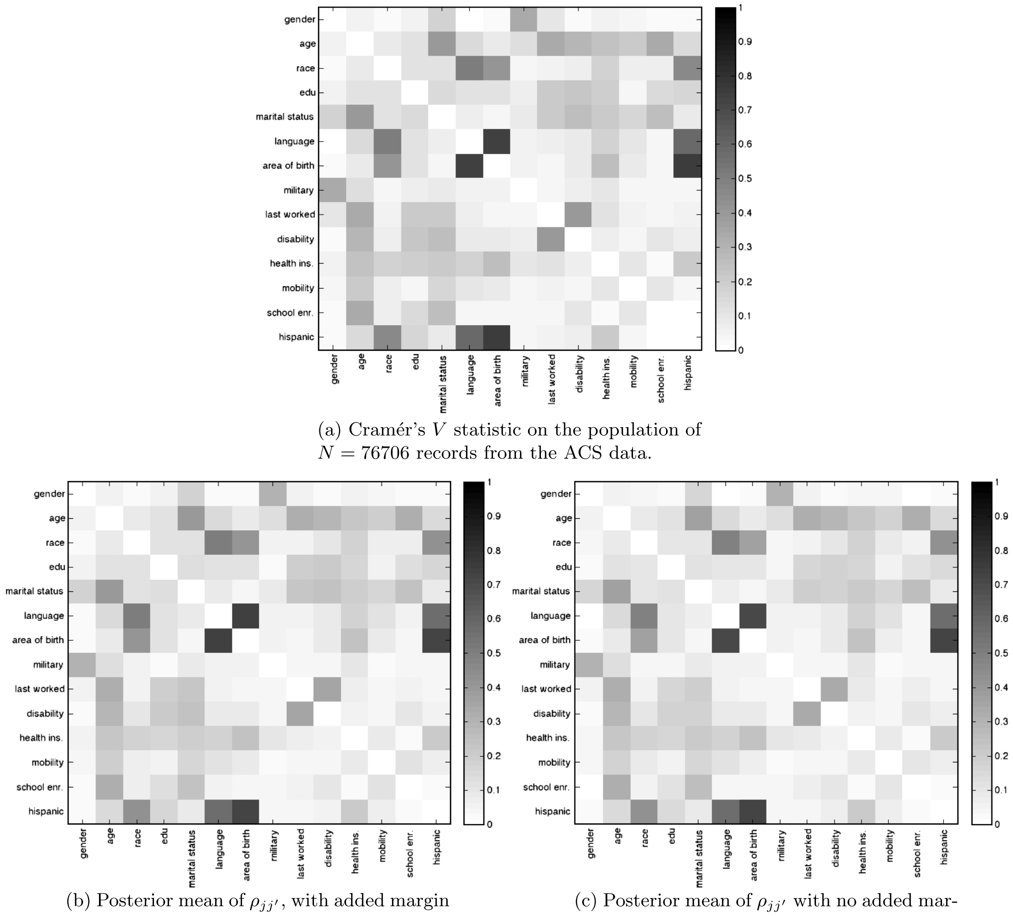
We also consider the joint probabilities in the four-way table involving gender and language spoken other than English (both variables included in A), and school enrollment and Hispanic (not included in A). After adding X_A, the average RMSE of these joint probabilities is 0.0016 with 95% of values between $(0.0009, 0.0025)$. Without adding X_A, the average RMSE of these probabilities is 0.0021 with 95% of values between $(0.0010, 0.0036)$. Thus, even though the school enrollment and Hispanic variables are not included in X_A, using the informative prior distribution improves the estimates of the joint probabilities in this four-way table.

Finally, we note that we ran four additional simulations and got similar results for the model-based Cramér's V statistic and the four-way table.

4 Using augmented records to account for stratified sampling

The DPMPM and other Bayesian latent class models effectively treat X_O as coming from a simple random sample. When this is not the case, these and other joint models can result in unreliable inferences about population parameters. In this section, we illustrate how augmented data can be used to adjust for unequal probabilities of selection resulting from stratified random sampling.

We again treat the ACS PUMS data as the population, and use the same $p = 5$ variables as in the simulation in Section 3.1. We sample $n = 10000$ records comprising simple random samples of 2500 records from each of four strata, namely African Amer-



(a) Cramér's V statistic on the population of $N = 76706$ records from the ACS data.

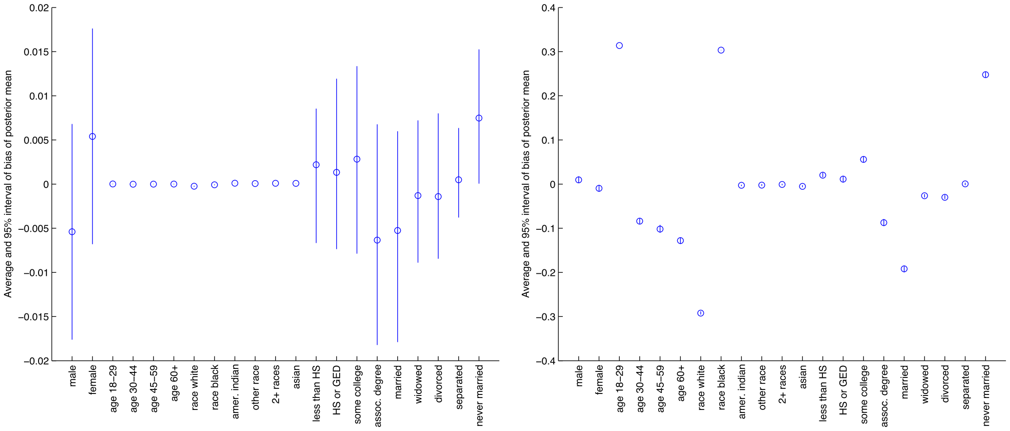
(b) Posterior mean of $\rho_{jj'}$, with added margin on first seven variables.

(c) Posterior mean of $\rho_{jj'}$ with no added margin.

Figure 6: The top figure shows Cramér's V statistic on the population data. The bottom figures show the model-based Cramér's V statistic on the sample of $n = 10000$ records from the ACS data, with and without augmented records.

icans aged 18 to 29, African Americans over age 30, non-African Americans aged 18 to 29, and non-African Americans over age 30. The population shares of the four strata are, in order, 4.2%, 15.3%, 14.4%, and 66.1%. Thus, the stratified sample greatly over-represents younger African Americans and greatly under-represents older non-African Americans. Not surprisingly, when we fit the DPMPM model without correcting for the stratification, the resulting estimates of marginal probabilities are badly biased, as illustrated in Figure 7b.

In many stratified sampling contexts, the population shares of the strata, and hence of the variables defining the stratification, are known and available for analysis. This suggests that we can treat the known shares as precise prior information and use the



(a) With $n_A = 90000$ augmented records.

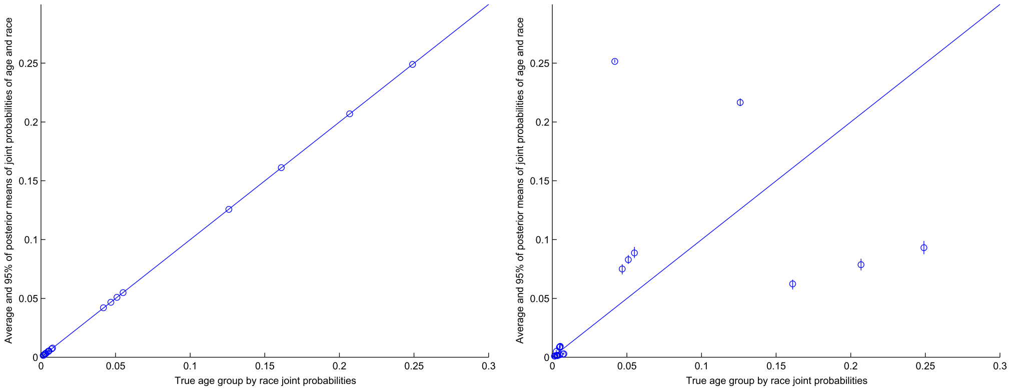
(b) No augmented records.

Figure 7: Difference in posterior means and population quantities for marginal probabilities in the stratified sampling simulation. The left panel fits the model after adding $n_A = 90000$ samples, the right panel fits the DPMPM without any adjustment for stratified sampling. The scales of the vertical axes differ in the two displays to improve interpretation in each display.

techniques of Section 3.1. Specifically, we can create augmented records so that the distributions of the stratification variables in the concatenated data match the known population shares. We set n_A large enough that X_{obs} (including X_O) is centered at the population distribution of the stratification variables with negligible variance. Alternatively, when N is not too large and finite population corrections matter, we can set $n_A = N - n$ and choose X_A so that the distribution of X_{obs} exactly matches that in the population.

We run 100 simulations as follows. For each stratified sample of size 10000, we generate $n_A = 90000$ records so that the distribution of age group and race in the concatenated data closely matches the known population shares, leaving all other variables missing in X_A. We assume the analyst knows the joint distribution of all race and all age group combinations, not just the four probabilities used in the stratification. As shown in Figure 7a, the DPMPM estimated on X_{obs} results in accurate estimates of the marginal probabilities. This is also the case for the joint distribution of age group and race, as shown in Figure 8, and for the 960 cell probabilities, as shown in Figure 9.

We now offer some intuition on how augmented records can adjust estimated joint distributions for stratified sampling. When stratifying on A, by design X_O is not sampled from the population marginal distribution of A. However, because units are collected within strata using simple random samples (this is the standard stratified sampling design), X_O is sampled from the population conditional distribution of $\{X_j : j \notin A\}$ given X_A. Since for large n the DPMPM can accurately estimate the distribution of the generative process for X_O, the DPMPM estimated with only X_O inaccurately estimates the marginal distribution of variables in A, but it should accurately estimate the



(a) With $n_A = 90000$ augmented records.

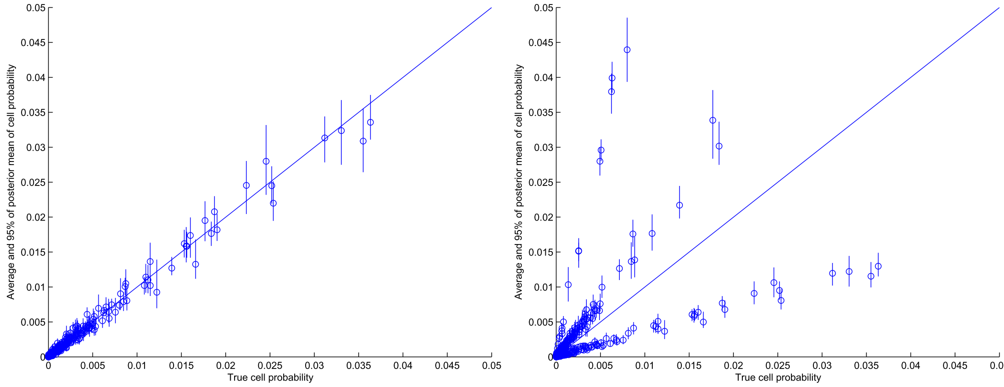
(b) No augmented records.

Figure 8: Posterior mean estimates versus corresponding population values of age group by race joint probabilities. The left panel fits the model with $n_A = 90000$ augmented samples. Across all 100 simulations, all 24 95% credible intervals contain the true joint probability. The right panel fits the model without adjusting for the stratified sampling.

conditional distribution of $\{X_j : j \notin A\}$ given X_A. Since X_A provides information only about the marginal distribution of A, the DPMPM estimated with X_{obs} still should accurately estimate the conditional distributions of $\{X_j : j \notin A\}$ given X_A. However, X_A encourages the DPMPM to estimate the marginal distributions of A accurately. Fusing the accurate estimates of the marginals of A and conditionals given A results in accurate estimates of the joint distribution. We note that the intuition above assumes that X_O includes all variables used in stratification; otherwise, the conditional distributions implied by the DPMPM are likely to be inaccurate.

The augmented records approach can be further understood using the framework put forth by Kunihama et al. (2014). To adjust the DPMPM for stratified sampling, Kunihama et al. (2014) suggest re-weighting the DPMPM mixture components according to their estimated population shares. The estimated shares are derived from sums of the survey weights of the records in X_O. Augmented records serve a similar function: like estimated shares, they increase or decrease the DPMPM mixture weights to reflect the population distribution of A. The DPMPM tends to assign augmented cases to components occupied by observed cases with similar values of A. These augmented records should not change the distributions of A (or the other variables) within components. Rather, they adjust the mixture weights, as the shares of the components reflect the shares of each combination of A in X_{obs}.

Sometimes the available stratum information is coarser than the corresponding variables used in the analysis; for example, the analyst knows the true proportion of African Americans of age 30 and up from metadata about the survey design, but does not know the breakdown of age 30–44 African Americans, age 45–59 African Americans, and age 60 and up African Americans. In this case, the analyst can construct X_A to match the known percentages at the available coarse scale, and allocate the within-stratum records



(a) With $n_A = 90000$ augmented records.

(b) No augmented records.

Figure 9: Posterior mean estimates of cell probabilities versus corresponding population values for all 960 cells in the table. Left panel displays results with $n_A = 90000$ added samples to correct for stratification, and right panel displays results with no added samples.

to match additional prior beliefs about the finer-scale variables. The analyst can create different versions of X_A to reflect different assumptions about the within-stratum allocations, and estimate the DPMPM model on each of the augmented margins as a sensitivity analysis.

5 Concluding remarks

The simulation results presented here suggest that using augmented data is a flexible and convenient way to incorporate prior information about marginal probabilities in latent class models. Augmented categorical cases also could be used to represent prior information on marginal probabilities in other types of mixture models, including models for mixed scale data (e.g., Zhou et al., 2014; Dunson and Bhattacharya, 2011; Wade et al., 2011). The same strategy applies—add n_A cases to reflect prior beliefs about marginal probabilities—with appropriate adjustments to the full conditional distributions. The mixture component indicators for the augmented records can be updated in batch, using only X_A to determine the component probabilities.

The general augmented data approach can be adapted to represent prior information about distributions of continuous variables. For example, to represent the prior belief that the marginal distribution of some continuous variable follows a distribution f, analysts can augment the data with $n_A \gg n$ cases drawn from f. Alternatively, analysts can make n_A small to represent relatively weak prior beliefs about the distribution. Analysts can calibrate n_A by examining the properties of summary quantities, such as moments and quantiles, over repeated draws of n_A values of X_A from the prior distribution. This approach could be used to adjust inferences for probability proportional to size samples. The analyst augments X_O with X_A generated to reflect the known, or at least accurately estimated, size distribution in A. We note that the computations with continuous data generally are more challenging, since typically the number of unique values of X_A will be close to n_A.

The approach could be applied in contexts with non-exchangeable data as well. For example, when data comprise people nested within households, analysts may have prior information from census counts on the number of individuals per household, and the distributions of gender and race within households. Given a sample X_O of households, analysts could append augmented household records reflecting those prior beliefs, and estimate appropriate joint models that account for the nested structure (Hu, 2015).

The augmented data approach potentially could improve inferences in other contexts as well. For example, many surveys suffer from unit nonresponse that is not missing at random. If the analyst has external information about the marginal distributions of some of the missing variables, she can augment the sample in a manner like the stratified sampling application and estimate the model on the concatenated data. In this way, the analyst can adjust inferences for nonignorable nonresponse (assuming the data for the variables not in the augmented margins are missing at random). A similar approach could help correct inferences (again under certain conditions) made with convenience samples. We plan to investigate these applications in future research.

Appendix A: Posterior computation

We use a Gibbs sampler to estimate the posterior distributions of the unknown quantities $(z, V, \pi, \alpha, \phi)$. The full conditionals are similar to those in Si and Reiter (2013), modified to incorporate the augmented data. For the augmented data, we do not fill in the missing values of the variables not in A, preferring to marginalize over the missing data. It would be straightforward to impute these missing values, as each variable is independent within latent classes.

The steps of the Gibbs sampler are as follows:

1. To update z_i for cases in the original data, i.e., where $i = 1, \ldots, n$, sample from a categorical distribution with

$$p(z_i = h | X_{i1}, \ldots X_{ip}, \pi, \phi) = \frac{\pi_h \prod_{j=1}^{p} \phi_{hjX_{ij}}}{\sum_{k=1}^{H^*} \pi_k \prod_{j=1}^{p} \phi_{kjX_{ij}}}. \tag{11}$$

2. To update z_i for cases in the augmented data, i.e., where $i = n+1, \ldots, n+n_A$, sample from a categorical distribution with

$$p(z_i = h | \{X_{ij} : j \in A\}, \pi, \phi) = \frac{\pi_h \prod_{j \in A} \phi_{hjX_{ij}}}{\sum_{k=1}^{H^*} \pi_k \prod_{j \in A} \phi_{kjX_{ij}}}. \tag{12}$$

   This can be done efficiently by sampling values of z for the recorded combinations in X_A. That is, for each recorded combination in X_A, we compute (12) and sample the values of z for all augmented records with that combination using a multinomial distribution.

3. To update V_h for $h = 1, \ldots, H^* - 1$, we sample from

$$p(V_h|\alpha, z) = \text{Beta}\left(n_h + 1, \alpha + \sum_{k=h+1}^{H^*} n_k\right) \tag{13}$$

   where $n_h = \sum_{i=1}^{n+n_A} 1(z_i = h)$. Set $V_{H^*} = 1$. Then, $\pi_h = V_h \prod_{g<h}(1 - V_g)$ for $h = 1, \ldots, H^*$.

4. To update α, sample from

$$p(\alpha|V_1, \ldots, V_{H^*-1}) = \text{Gamma}\left(H^* + a_\alpha - 1, b_\alpha - \log\left(\pi_{H^*}\right)\right). \tag{14}$$

5. To update ϕ_{hj} for variables with no augmented margin, i.e., for $j \notin A$, where $h = 1, \ldots, H^*$, sample from

$$p(\phi_{hj}|X_{obs}, z) = \text{Dirichlet}\left(1 + \sum_{\substack{i=1\\ z_i=h}}^{n} 1(X_{ij} = 1), \ldots, 1 + \sum_{\substack{i=1\\ z_i=h}}^{n} 1(X_{ij} = d_j)\right). \tag{15}$$

6. To update ϕ_{hj} for variables with augmented margin, i.e., for $j \in A$, where $h = 1, \ldots, H^*$, sample from

$$p(\phi_{hj}|X_{obs}, z) = \text{Dirichlet}\left(1 + \sum_{\substack{i=1\\ z_i=h}}^{n+n_A} 1(X_{ij} = 1), \ldots, 1 + \sum_{\substack{i=1\\ z_i=h}}^{n+n_A} 1(X_{ij} = d_j)\right). \tag{16}$$

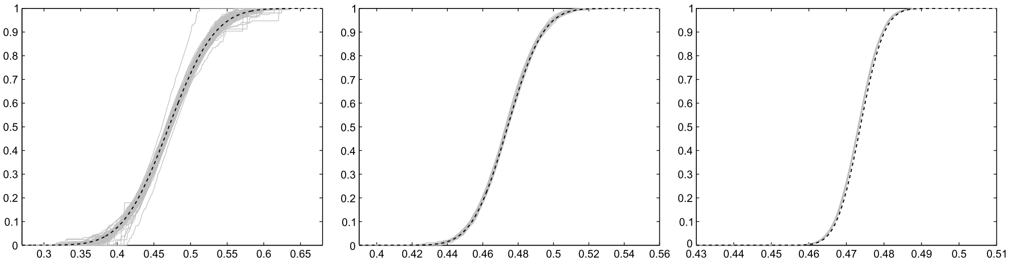
Appendix B: Interpretation of n_A as prior sample size of Dirichlet distribution

In Section 3.2, we suggest thinking of n_A as a prior sample size in a Dirichlet distribution. To illustrate that this is a reasonable interpretation, we now present results of simulation studies in which we approximate the prior distribution on $\phi_{male} = \Pr(\text{gender} = \text{male})$ implied by adding records with only gender recorded.

We take a sample of size $n = 100$ individuals from the PUMS data for whom we observe gender, age group, race, educational attainment, and marital status. We add an augmented sample comprising gender only for $n_A \in \{100, 1000, 10000\}$. We run the DPMPM model on this X_{obs} for $T = 5000$ iterations after the burn-in (also 5000 runs), and save the posterior draws of ϕ_{male}. We repeat the process 100 times.

Rearranging Bayes rule, the implied prior distribution of ϕ_{male} given the collected data X_0 is

$$p(\phi_{male}) = \frac{p(\phi_{male}|X_0)p(X_0)}{p(X_0|\phi_{male})}. \tag{17}$$



(a) $n_A = 100$. (b) $n_A = 1000$. (c) $n_A = 10000$.

Figure 10: Comparison of theoretical Beta CDF to empirical prior CDF under different settings of n_A.

For any simulation run, each of the components on the right hand side of (17) can be readily approximated from the converged MCMC samples. Thus, we can approximate $p(\phi_{male})$ along a grid of values from 0 to 1. Let k_1 be the number of males in the sample of $n = 100$ records. At each multiple of 0.001 between 0 and 1, the approximation is

$$p(\phi_{male} = x) \propto \frac{\frac{1}{T}\sum_{t=1}^{T} I(x - 0.0005 < \phi_{male}^{(t)} \leq x + 0.0005)}{\frac{n!}{k_1!(n-k_1)!}x^{k_1}(1-x)^{n-k_1}}. \tag{18}$$

Figure 10 compares the approximated prior cumulative distributions (in gray) to the theoretical Beta$(k_2 + 1, n_A - k_2 + 1)$ cumulative distributions (in dashed black line), where k_2 is the number of males in the added margin. For all values of n_A, the Beta distribution is a close match, suggesting that it is reasonable to think of n_A as the prior sample size of a Dirichlet distribution.

References

Dunson, D. B. and Bhattacharya, A. (2011). "Nonparametric Bayes Regression and Classification Through Mixtures of Product Kernels." In: Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9, Proceedings of Ninth Valencia International Conference on Bayesian Statistics*. Oxford University Press. MR3204005. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0005. 513

Dunson, D. B. and Xing, C. (2009). "Nonparametric Bayes modeling of multivariate categorical data." *Journal of the American Statistical Association*, 104(487): 1042–1051. MR2562004. doi: http://dx.doi.org/10.1198/jasa.2009.tm08439. 499, 500, 501, 508

Gebregziabher, M. and DeSantis, S. M. (2010). "Latent class based multiple imputation approach for missing categorical data." *Journal of Statistical Planning and Inference*, 140(11): 3252–3262. MR2659852. doi: http://dx.doi.org/10.1016/j.jspi.2010.04.020. 499

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7(4): 457–472. doi: `http://dx.doi.org/10.1214/ss/1177011136`. 505

Goodman, L. A. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models." *Biometrika*, 61(2): 215–231. MR0370936. doi: `http://dx.doi.org/10.1093/biomet/61.2.215`. 499

Greenland, S. (2007). "Prior data for non-normal priors." *Statistics in Medicine*, 26: 3578–3590. MR2393737. doi: `http://dx.doi.org/10.1002/sim.2788`. 500

Hu, J. (2015). "Dirichlet Process Mixture Models for Nested Categorical Data." Ph.D. thesis, Department of Statistical Science, Duke University. MR3347110. 514

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): pp. 161–173. MR1952729. doi: `http://dx.doi.org/10.1198/016214501750332758`. 499, 501

Jain, S. and Neal, R. M. (2004). "A split–merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model." *Journal of Computational and Graphical Statistics*, 13(1): 158–182. MR2044876. doi: `http://dx.doi.org/10.1198/1061860043001`. 499

Johndrow, J., Cron, A., and Dunson, D. B. (2014). "Bayesian tensor factorizations for massive web networks." In: *ISBA World Meeting 2014 in Cancun, Mexico*. 503

Kalli, M., Griffin, J. E., and Walker, S. G. (2009). "Slice sampling mixture models." *Statistics and Computing*, 21: 93–105. MR2746606. doi: `http://dx.doi.org/10.1007/s11222-009-9150-y`. 501

Kamakura, W. A. and Wedel, M. (1997). "Statistical data fusion for cross-tabulation." *Journal of Marketing Research*, 34: 485–498. doi: `http://dx.doi.org/10.2307/3151966`. 499

Kessler, D. C., Hoff, P. D., and Dunson, D. B. (2015). "Marginally specified priors for non-parametric Bayesian estimation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1): 35–58. doi: `http://dx.doi.org/10.1111/rssb.12059`. 500

Kunihama, T. and Dunson, D. B. (2013). "Bayesian modeling of temporal dependence in large sparse contingency tables." *Journal of the American Statistical Association*, 108(504): 1324–1338. MR3174711. doi: `http://dx.doi.org/10.1080/01621459.2013.823866`. 500

Kunihama, T., Herring, A. H., Halpern, C. T., and Dunson, D. B. (2014). "Nonparametric Bayes modeling with sample survey weights." arXiv:1409.5914. 512

Manrique-Vallier, D. and Reiter, J. P. (2014a). "Bayesian estimation of discrete multivariate latent structure models with structural zeros." *Journal of Computational and Graphical Statistics*, 23: 1061–1079. MR3270711. doi: `http://dx.doi.org/10.1080/10618600.2013.844700`. 503

— (2014b). "Bayesian multiple imputation for large-scale categorical data with structural zeros." *Survey Methodology*, 40: 125–134. 499

Papaspiliopoulos, O. (2008). "A note on posterior sampling from Dirichlet mixture models." *Technical Report*, *Centre for Research in Statistical Methodology*. 501

Sethuraman, J. (1994). "A constructive definition of D"irichlet priors. *Statistica Sinica*, 4: 639–650. MR1309433. 501

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(5): 499–521. doi: http://dx.doi.org/10.3102/1076998613480394. 499, 500, 501, 514

Si, Y., Reiter, J. P., and Hillygus, D. S. (2015). "Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples." *Political Analysis*, 23(1): 92–112. doi: http://dx.doi.org/10.1093/pan/mpu009. 499

Vermunt, J. K., Van Ginkel, J. R., Van Der Ark, L. A., and Sijtsma, K. (2008). "Multiple imputation of incomplete categorical data using latent class analysis." *Sociological Methodology*, 38(1): 369–397. 499

Wade, S., Mongelluzzo, S., and Petrone, S. (2011). "An Enriched Conjugate Prior for Bayesian Non-parametric Inference." *Bayesian Analysis*, 6: 359– 385. MR2843536. doi: http://dx.doi.org/10.1214/ba/1339616468. 513

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics – Simulation and Computation*, 36(1): 45–54. MR2370888. doi: http://dx.doi.org/10.1080/03610910601096262. 501

Zhou, J., Bhattacharya, A., Herring, A. H., and Dunson, D. B. (2014). "Bayesian factorizations of big sparse tensors." *Journal of the American Statistical Association*, to appear. doi: http://dx.doi.org/10.1080/01621459.2014.983233. 513

Acknowledgments

This research was supported by grants from the National Science Foundation (NSF-SES-10-61241 and NSF-SES-11-31897) and the SAMSI program Computational Methods in the Social Sciences (DMS-1127914).