Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (ii): New methods for protection of microdata

Non-parametric Bayesian Model for Generating Synthetic Household Data

Prepared by Monika Jingchen Hu and Jerry Reiter, Duke University, U.S.A.

Nonparametric Bayesian Model for Generating Synthetic Household Data

Monika Jingchen Hu and Jerry P. Reiter
Department of Statistical Science, Box 90251, Duke University, Durham,
NC 27708-0251, USA, *jh309@stat.duke.edu*, *jerry@stat.duke.edu*

Abstract. Surveys on household data may comprise only nominal responses (sex, race, marital status etc.) for each household member (such as the Decennial Census Data on Families). In order to protect respondents' privacy, we are developing a fully Bayesian, joint modeling approach for categorical data based on the nested Dirichlet process (Rodriguez et al., 2008). It induces a two-level clustering structure in modeling the household dataset, trying to simultaneously cluster household members within households, and borrow information across households that have similar clusters. The model is applied to a subset of the Current Population Survey March 2011 household dataset synthesis. The results demonstrate its abilities to preserve marginal and multivariate distributions of all nominal variables (dependence structures among variables), and within household relationships, such as difference in race, age and education.

1 Introduction

When releasing microdata to the public, data disseminators typically are required to protect the confidentiality of survey respondents' identities and attribute values. To satisfy these requirements, removing direct identifiers such as names and addresses generally is not efficient to eliminate disclosure risks, so that data must be altered before release to limit the risks of unintended disclosures. Synthetic data is one popular approach, where values of confidential data are replaced with multiple simulations/imputations from statistical models (Reiter, 2011).

Typical models for categorical data imputation and synthesis include sequential regression modeling (Raghunathan et al., 2001) and log-linear models. They have been widely applied successfully, however they have limitations in model selection and estimation, especially in high dimension (Si and Reiter, 2013). For example using sequential regression modeling approach for multiple imputation (MICE) when the number of variables is large, the imputer needs to specify many conditional models. Some users use default settings to include only main effects in the conditional model, which will result in failure to capture complex dependencies that lead to biased inferences. Using log-linear models in high dimension, model selection becomes very challenging.

Some recent work on modeling multivariate unordered categorical data (Dunson and Xing, 2009) approached the problem in the latent class modeling framework (clustering). It developed a nonparametric Bayesian approach which defines a prior (Dirichlet process mixtures of multinomial distributions) with full support on the space of distributions for multiple unordered categorical variables, ensuring that the dependence structure is not restricted a priori. Its application on multiple imputation for incomplete categorical variables in large-scale assessment surveys (Si and Reiter, 2013) demonstrates its abilities of modeling complex dependencies and maintaining computation expediency.

In modeling nominal household data, a particular challenge besides the ones mentioned before is to model relationships among different family members in the same household. This urges us to work on a model structure that is not assuming all individuals to be independently distributed. Along the line of latent class modeling, we propose a two-level clustering for household data synthesis based on the nested Dirichlet process (Rodriguez et al., 2008), which simultaneously clusters household members within households, and borrows information across households that have similar clusters. The posterior computation can be implemented via an efficient Markov Chain Monte Carlo (MCMC) algorithm relying on a blocked Gibss sampler (Ishwaran and James, 2001). The model is applied to a subset of the Current Population Survey March 2011 household data synthesis.

2 The nested dirichlet process model

2.1 Motivation and intuition

In a household survey dataset, there may exist: 3-member families with young parents and a young child; 5-member families with 3 generations under the same roof; 2-member families with aging couple living alone; 4-member families with all members renting and sharing the house together, and many others.

Intuitively, we would like to have households with similar structures to be clustered together, where all members of these households compose a pool of individuals. Inside the pool, individuals with similar personal characteristics (attributes) can be clustered together again for their attributes estimations. The synthetic attribute k of a specific individual A can thus be generated based on the estimate of k in the cluster that A belongs to.

2.2 Model specification

To describe the model, some notations are provided. Let X_{ijk} denote the category that the k-th variable falls into for the j-th member in i-th household, d_k denote the number of categories of variable k. η_i denotes the latent class assignment of household i, and ψ_{ij} denotes the latent class assignment of individual j in household i. $\phi^{(k)}_{fs}$ denotes the multinomial probability vector of variable k in household latent class f and member latent class s. There are total n households and tn individuals in the sample and each has records in p different categorical variables.

The model can be expressed in the following hierarchical form:

$$\begin{aligned}
X_{ijk} &\sim Multinomial(\{1,\ldots,d_k\},\phi^{(k)}_{\eta_i\psi_{ij}1},\ldots,\phi^{(k)}_{\eta_i\psi_{ij}d_k}) \text{ for all i,j,k}\\
\eta_i &\sim Multinomial(\pi_1,\ldots,\pi_\infty)\\
\psi_{ij}|\eta_i &\sim Multinomial(w_{\eta_i 1},\ldots,w_{\eta_i\infty}) \text{ for all i,j}\\
\pi_f &= u_f\prod_{l<f}(1-u_l) \text{ for } f=1,\ldots,\infty\\
w_{fs} &\sim v_{fs}\prod_{l<s}(1-v_{fl}) \text{ for } s=1,\ldots,\infty\\
u_f &\sim Beta(1,\alpha), v_{fs}\sim Beta(1,\beta)\\
\alpha &\sim Gamma(a_\alpha,b_\alpha),\beta\sim Gamma(a_\beta,b_\beta)\\
\phi^{(k)}_{fs} &= (\phi^{(k)}_{fs1},\ldots,\phi^{(k)}_{fsd_k})\sim Dirichlet(a_{k1},\ldots,a_{kd_k})
\end{aligned}$$

This two-level clustering model assumes that: 1. each household is assigned to one underlying latent class (household level clustering); 2. after the household latent class assignment, each member of the household is assigned to one of (another set of) latent classes (member level clustering). Given the latent class, all variables (categorical only) of the individual independently follow their own multinomial distributions. A Dirichlet process prior is induced on the household level latent class mixture probabilities $\pi=(\pi_1,\ldots,\pi_\infty)$ using the stick-breaking representation(Sethuraman, 1994); another Dirichlet process prior is induced on the member level latent class mixture probabilities $w = (w_{\eta_i 1}, \ldots, w_{\eta_i \infty})$ controlled by the concentration parameter β. The conjugate Dirichlet measure of the multinomial likelihood of each categorical variable controlled by the base measure H. Such a prior allows the numbers of latent classes to be infinity.

The algorithm based on the blocked Gibbs sampler (Ishwaran and James, 2001) is applied for posterior computation. It truncates the infinite stick-breaking probabilities at some large number F and S while fast computation is guaranteed. The detailed posterior computation and sampling steps are available upon request.

3 CPS dataset and simulation results

We applied the nested Dirichlet process model to the March 2011 Current Population Survey household dataset, with 6 selected nominal variables (with * re-categorized, ** converted from continuous, and values in parenthesis indicate the number of categories.): *ownership** (3), *sex* (2), *race** (5), *maritalstatus* (6), *education** (5) and *age*** (9). There are 10000 households and 26661 individuals in the sample. We put hyperpriors $a_\alpha = b_\alpha = 0.25$ and $a_{k1} = \cdots = a_{kd_k} = 0.25$. We initialize ϕ's with the maximum likelihood estimates from the original dataset.

The following results are based on the first 10000 households (altogether 26661 individuals), and fixing F=30 and S=15; nrun is 10000, burn-in is 5000 and thin is 10. Marginal and joint distributions of variables in the original dataset and the synthetic dataset (the one generated at the last iteration of the MCMC chain) are compared.

3.1 Marginal distributions

The marginal distributions of variables sex and age are selected to present, and the other variables behave in a similar pattern. They show that the synthetic dataset can preserve the marginal distributions very well.

	1	2
original	12830	13831
synthetic	12663	13998

Table 1: Sex distributions

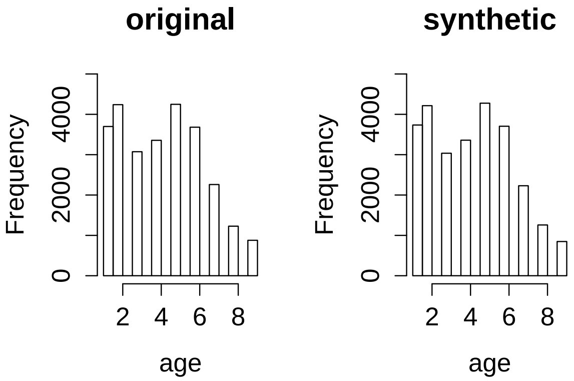

Figure 1: Histograms of age distributions

3.2 Bivariate distributions

The bivariate distributions of variables sex and marital status, education and marital status are selected to present. Pairs of other variables behave in a similar pattern. They show that the synthetic dataset can preserve the dependencies between variables in pair very well.

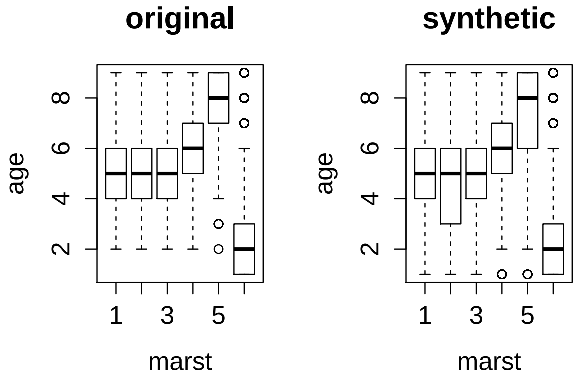

Figure 2: Boxplots of age-marst distributions

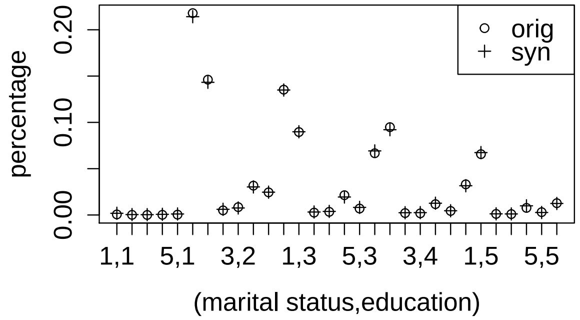

Figure 3: education-marst distributions

3.3 Some measures on within household relationships

We are most interested in how well within household relationships can be preserved using this nested Dirichlet process prior model. We look at the within household race in all households of size 3 in both dataset. In the following table, 0 indicates that all members in the same household have the same race, otherwise 1. Though the original dataset seems to have more households of size 3 having all members in the same race than the synthetic dataset, the performance of the model has done a decent job of capturing this dataset feature.

	0	1
original	1744	97
synthetic	1639	202

Table 2: Within household race for 1841 households of size 3

We also look at the distribution of the largest age difference in households of size 3 and the education difference in households of size 2, comparing them across the original and synthetic datasets. The following histograms show that most of the bars are overlapped well, which means such features are being captured by the model. Other measures follow a similar pattern.

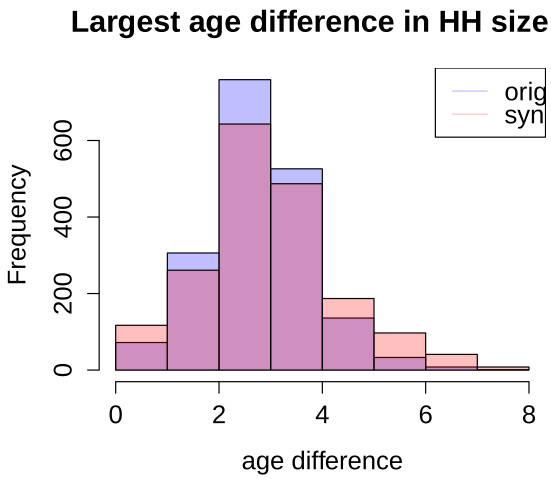

Figure 4: Largest absolute age difference distribution in households of size 3

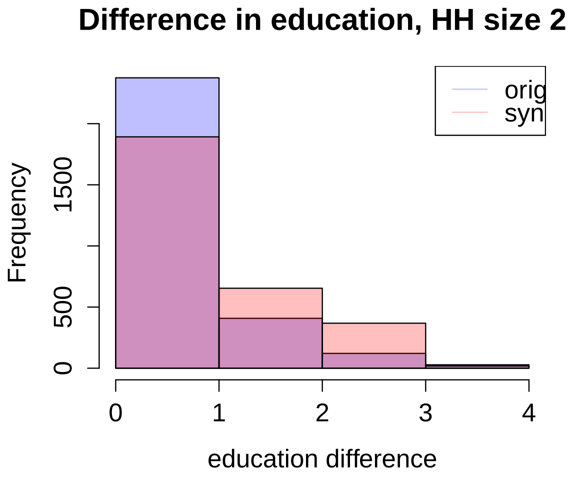

Figure 5: Absolute education difference distribution in households of size 2

4 Discussion

Overall the synthesis results on the Current Population Survey dataset look promising, demonstrating the model's abilities to preserve marginal and multivariate distributions of all nominal variables, and within household relationships. The computation is fairly efficient in $Matlab$ with extensions in $C++$. Traceplots of concentration parameters α and β in the Dirichlet processes seem to show convergence of the MCMC chain. Cluster assignments at different iterations, and parameter estimations in a certain cluster across different iterations seem to be stable, which also suggests the convergence of the chain.

However there are certain features that are not captured fully by our nest Dirichlet prior model. For example in Figure 2, there are no cases of individuals having combinations of age 1 and marital status 1 to 5 in the original dataset, while there exist such cases in the synthetic dataset. Those cases shouldn't be allowed in the model as one wouldn't expect some respondent being 5-year-old with a college degree. These impossible combinations are structural zeros problems which are ongoing work to improve the model performance.

Future work includes developing utility measures and risk measures of this household data synthesis model. Dealing with mixed data types is a natural extension and should be straightforward under this latent class modeling framework.

References

Dunson, D.B., and Xing. C. (2009) "Nonparametric Bayes Modeling of Multivariate Categorical Data", *Journal of the American Statistical Association*, **104(487)**, 1042–1051.

Ishwaran, H., and James, L.F., (2001), "Gibbs sampling for stick-breaking priors", *Journal of the American Statistical Association*, **96**, 161–173.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). "A multivariate technique for multiply imputing missing values using a sequence of regression models technique for multiply imputing missing values using a series of regression models", *Survey Methodology*, **27**, 85–96.

Reiter, J. P. (2011), "Data confidentiality", *Wiley Interdisciplinary Reviews: Computational Statistics*, **3**, 450–456.

Rodriguez, A., Dunson, D.B., and Gelfand, A.E., (2008), "The nested Dirichlet process", *Journal of the American Statistical Association*, **103**, 1131–1154.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors", *Statistica Sinica*, **4**, 639–650.

Si, Y. and Reiter, J. P. (forthcoming), "Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys", *Journal of Educational and Behavioral Statistics*.