The Quasi-Multinomial Synthesizer for Categorical Data

Jingchen Hu* and Nobuaki Hoshino**

Vassar College* and Kanazawa University**

Abstract. We present a new synthesizer for categorical data based on the Quasi-Multinomial distribution. Characteristics of the Quasi-Multinomial distribution provide a tuning parameter, which allows a Quasi-Multinomial synthesizer to control the balance of the utility and the disclosure risks of synthetic data. We develop a Quasi-Multinomial synthesizer based on a popular categorical data synthesizer, the Dirichlet process mixtures of products of multinomial distributions. The general sampling methods and algorithm of the Quasi-Multinomial synthesizer are developed and presented. We illustrate its balance of the utility and the disclosure risks by synthesizing a sample from the American Community Survey.[1]

Keywords: Bayesian, Dirichlet process, microdata, Quasi-Multinomial, synthetic

1 Introduction

The synthetic data approach to data confidentiality has gained attention and momentum in the past two decades. Based on the theory and applications of multiple imputation methodology for missing data problems (Rubin, 1987), statistical models are first estimated from the original confidential data, and then multiply-imputed synthetic data is generated to provide high utility, low risks public microdata. Multiple synthetic datasets should be generated, and appropriate combining rules have been developed to provide accurate point estimates and variance estimates of parameters of interest. Refer to Reiter and Raghunathan (2007); Drechsler (2011) for details of the combining rules.

More recently, nonparametric Bayesian models have been further developed and turned into data synthesizers. Among them, the Dirichlet process mixtures of products of multinomials (DPMPM) synthesizer is worth particular attention. The DPMPM consists of a set of flexible Bayesian latent class models that have been developed to capture complex relationships among multivariate unordered categorical variables (Dunson and Xing, 2009). Hu *et al.* (2014) implemented the DPMPM as a synthesizer on multivariate unordered categorical data and demonstrated its balance between data utility and disclosure risks. Drechsler and Hu (2017+) used the DPMPM synthesizer for generating partially synthetic data with geocoding information. Other work on some version of the DPMPM for synthesis include Manrique-Vallier and Reiter (2014); Hu *et al.* (2018); Manrique-Vallier and Hu (2018). The DPMPM has also been proposed as a multiple imputation engine for missing data problems when all variables are categorical (Si and Reiter, 2013; Akande *et al.*, 2017; Murray, 2018+; Akande *et al.*, 2017+)

[1] This version is a colored version of the published manuscript (with plots in color).

While useful and promising, the characteristics of the utility and disclosure risks tradeoff of the DPMPM synthesizer have not yet been a research focus. The DPMPM synthesizer is based on the multinomial distribution, which has no parameter to control the tradeoff. On the other hand, the Quasi-Multinomial (QM) distribution of Consul and Mittal (1977) is a generalized multinomial distribution with an additional parameter, which can be effectively tuned to deliver a desired balance of utility and disclosure risks in the synthetic data products that statistical agencies would produce, if we construct a synthesizer based on the QM distribution.

In this paper, we focus on developing the QM-DPMPM synthesizer, and comparing it with the DPMPM synthesizer. Section 2 introduces the QM distribution, discusses the sampling methods and proposes an algorithm based on acceptance rejection sampling. The DPMPM and the QM-DPMPM synthesizers are introduced in Section 3. Section 4 presents an illustrative application comparing the two synthesizers, using a sample from the American Community Survey (ACS). Discussions and future work are given in Section 5.

2 The Quasi-Multinomial distribution

2.1 Introducing the Quasi-Multinomial distribution

The Quasi-Multinomial distribution (type 2) is a generalized multinomial distribution proposed by Consul and Mittal (1977). We define the QM distribution by the following probability mass function (pmf):

$$p(y_1, \ldots, y_F) = \frac{n!}{y_1! \ldots y_F!} \frac{1}{(1+n\beta)^{n-1}} \prod_{f=1}^{F} \pi_f (\pi_f + y_f \beta)^{y_f - 1}, \tag{1}$$

where $y_f, f = 1, 2, \ldots, F$, is a nonnegative integer, and $\sum_{f=1}^{F} y_f = n$. Similar to the multinomial distribution, y_f is regarded as the random frequency of the fth cell given a total frequency n. We denote this parameterization of the QM distribution in Equation (1) by QM$(\pi_1, \ldots, \pi_F; n, \beta)$.

We consider the parameter β in Equation (1) as a nonnegative real number. While this pmf is proper when $\beta > -\min_f \pi_f / n$, we disallow negative β to avoid the dependence of the parameter space on parameters themselves. This limitation is necessary for theorems in Section 2.2; it also guarantees that the QM will always increase the data protection.

When $\beta = 0$, Equation (1) reduces to the pmf of the multinomial distribution with cell probabilities $(\pi_1, \ldots, \pi_F)$, which we denote as Multinomial$(\pi_1, \ldots, \pi_F; n)$. As β increases, the variance of any univariate marginal frequency increases (Consul and Mittal, 1975).

Among the similarities of the QM distribution to the multinomial distribution, there are a few points worth mentioning. First, similar to the multinomial distribution, the parameters π_f's of Equation (1) are nonnegative real number, and $\sum_{f=1}^{F} \pi_f = 1$. Hence π_f of the QM distribution is referred to by a cell probability. Second, the expectation of the fth marginal frequency is $n\pi_f$ regardless of β (Hoshino, 2009). In other words, the

fth sample relative frequency is the unbiased estimator of the fth cell probability for all β.

Then we note that since the variance of any univariate marginal frequency increases as β increases, the accuracy of the unbiased estimator of a cell probability can be limited by increasing β. This fact implies that replacing the multinomial distribution with the QM distribution in the generation of synthetic data facilitates to control the balance of the utility and the disclosure risk of synthetic data while the unbiasedness remains to hold. We thus regard β as a tuning parameter of the QM synthesizer determined by a statistical agency.

To sample from the QM distribution, we rely on a special case of it. When $F = 2$, the QM distribution becomes the Quasi-Binomial (QB) distribution (type 2), proposed by Consul and Mittal (1975). We denote the QB distribution by $\mathrm{QB}(\pi; n, \beta)$, with its pmf:

$$p_{QB}(y) = \frac{n!}{y!(n-y)!}\frac{1}{(1+n\beta)^{n-1}}\pi(\pi+y\beta)^{y-1}(1-\pi)(1-\pi+(n-y)\beta)^{n-y-1}, \quad (2)$$

where $y = 0, 1, \ldots, n$, $0 \leq \pi \leq 1, \beta \geq 0$.

2.2 The decomposability of the Quasi-Multinomial distribution

Sampling from the QM distribution can be decomposed into simpler sampling of marginal variables. The characteristic nature of the QM distribution enables two general methods of such decomposition. The first one is the conditional distribution method (Devroye, 1986). The second one is multi-stage sampling.

The first decomposition of sampling from the QM distribution exploits the following general relationship:

$$(Y_1, \ldots, Y_F) \stackrel{d}{=} (Y_1|Y_2, \ldots, Y_F)\ldots(Y_{F-2}|Y_{F-1}, Y_F)(Y_{F-1}|Y_F)Y_F, \quad (3)$$

where "$\stackrel{d}{=}$" denotes equality in distribution.

The right hand side of (3) is the product of the conditional distributions of univariate margins. An explicit formula of these distributions is given below on the QM distribution:

Theorem 1 *If* $(Y_1, \ldots, Y_F) \sim QM(\pi_1, \ldots, \pi_F; n, \beta)$ *then* $(Y_g|Y_{g+1} = y_{g+1}, \ldots, Y_F = y_F) \sim QB(\pi_g/(1 - \sum_{f=g+1}^{F}\pi_f); n - \sum_{f=g+1}^{F} y_f, \beta)$ *for* $g = 1, \ldots, F$.

Theorem 1 reads that $Y_F \sim QB(\pi_F; n, \beta)$. It is widely known that Theorem 1 holds for the case of $\beta = 0$ or the multinomial distribution. Theorem 1 follows from Theorem 2 below.

Combining Equation (3) and Theorem 1, we observe that sampling from the QM distribution is accomplished by sequential sampling from the QB distribution. By symmetry, Theorem 1 holds even after exchanging the indices of variables. Therefore the resulting F dimensional sampling distribution of our procedure does not depend on the order of single margins to sample.

These single margins should be ordered in sampling so that corresponding cell probabilities are decreasing. This sequential sampling from larger cells is known to be efficient on the multinomial distribution (Ho *et al.*, 1979).

The second decomposition exploits another property that the conditional QM distribution given the sum of partial frequencies is again QM:

Theorem 2 *If* $(Y_1, \ldots, Y_F) \sim QM(\pi_1, \ldots, \pi_F; n, \beta)$ *then for* $g = 1, \ldots, F$ *and* $m = 0, \ldots, n$*,*

$$(Y_1, \ldots, Y_g | \sum_{f=1}^{g} Y_f = m) \sim QM(\pi_1/(\sum_{f=1}^{g} \pi_f), \ldots, \pi_g/(\sum_{f=1}^{g} \pi_f); m, \beta). \quad (4)$$

Theorem 2 can be shown by the fact that the QM distribution is closed under the collapse of cells (Hoshino, 2009). It is noteworthy that Theorem 2 holds after exchanging the indices of variables as Theorem 1 does.

Theorem 2 validates two-stage sampling from the QM distribution: The first stage generates the aggregated frequency of $m = Y_1 + \cdots + Y_g \sim \text{QB}(\sum_{f=1}^{g} \pi_f; n, \beta)$; the second stage generates frequencies $Y_1, \ldots, Y_g$ given m, subject to Equation (4). Then the resulting vector $(Y_1, \ldots, Y_F) \sim \text{QM}(\pi_1, \ldots, \pi_F; n, \beta)$. More generally a recursive argument validates multi-stage sampling from the QM distribution.

This type of multi-stage sampling has been used for the multinomial distribution to reduce computing time. For example, Malefaki and Iliopoulos (2007) provide an empirical support to import stages, which might seem redundant. On the QM distribution, we will see that multi-stage sampling can reduce computing time because it lowers the rejection rate of acceptance rejection sampling.

We have shown that sampling from the QM distribution can be expressed as the combination of sequential and multi-stage sampling from the QB distribution. Next, we discuss the method of sampling from the QB distribution.

2.3 Sampling from the Quasi-Binomial distribution

To sample from the QB distribution, we prepare acceptance rejection sampling (also called the rejection method by Devroye (1986)). The key element of acceptance rejection sampling is a "proposal" distribution, which should be close to its "target" distribution and also easy to sample. We use the Beta-Binomial mixture (BB) distribution as our proposal distribution.

If $p \sim \text{Beta}(a_1, a_2)$ and $Y \mid p \sim \text{Binomial}(n, \pi)$ then Y is called $\text{BB}(a_1, a_2; n)$ distributed, with pmf

$$p_{BB}(y) = \frac{n!}{y!(n-y)!} \frac{\Gamma(a.)}{\Gamma(a. + n)} \frac{\Gamma(a_1 + y)}{\Gamma(a_1)} \frac{\Gamma(a_2 + n - y)}{\Gamma(a_2)}, \quad y = 0, \ldots, n. \quad (5)$$

The same as the QB distribution, the BB distribution belongs to the family of the Conditional Compound Poisson distributions (Hoshino, 2009). Equalizing the canonical parameters of this family, the counterpart of $\text{BB}(a_1, a_2; n)$ is $\text{QB}(a_1/(a_1+a_2); n, 1/(a_1+a_2))$, whose pmf is

$$p_{QB}(y) = \frac{n!}{y!(n-y)!} \frac{1}{a.(a. + n)^{n-1}} a_1(a_1 + y)^{y-1} a_2(a_2 + (n-y))^{n-y-1}, \quad (6)$$

where $a. := a_1 + a_2$, and y takes nonnegative integers from 0 to n.

We regard the case of $a_1 = a_2 = 0$ as improper, and this case is excluded from our argument. If $a_1 = 0$ and $a_2 > 0$ then QB$(a_1/(a_1 + a_2); n, 1/(a_1 + a_2))$ degenerates at 0, or it takes 0 with probability one. On the contrary if $a_2 = 0$ and $a_1 > 0$ then QB$(a_1/(a_1 + a_2); n, 1/(a_1 + a_2))$ degenerates at n. Hence these two cases do not need sampling, and they are also excluded from our argument henceforward.

The ratio of Equation (5) to Equation (6) is

$$\frac{p_{BB}(y)}{p_{QB}(y)} = \frac{\Gamma(a. + 1)\Gamma(a_1 + y)\Gamma(a_2 + n - y)(a. + n)^{n-1}}{\Gamma(a. + n)\Gamma(a_1 + 1)\Gamma(a_2 + 1)(a_1 + y)^{y-1}(a_2 + (n - y))^{n-y-1}}, \quad (7)$$

and we are interested in the minimum of Equation (7) with respect to y, which equals the average acceptance rate of our acceptance rejection sampling. The proof of the following Theorem 3 is given in Appendix 1.

Theorem 3 *Suppose that n is a positive integer, and a_1, a_2 are positive real numbers. Denote the value of y that minimizes Equation (7) by y^*. Then $y^* = n$ when $a_1 < a_2$, and $y^* = 0$ when $a_2 < a_1$. When $a_1 = a_2$, (7) is minimized at $y = 0$ and $y = n$.*

By symmetry we only consider the case of $a_1 < a_2$, where the average acceptance rate is assured by Theorem 3 to be

$$\min_y \frac{p_{BB}(y)}{p_{QB}(y)} = \frac{p_{BB}(n)}{p_{QB}(n)} = \frac{\Gamma(a. + 1)\Gamma(a_1 + n)}{\Gamma(a. + n)\Gamma(a_1 + 1)} \left(\frac{a. + n}{a_1 + n}\right)^{n-1} =: r(a_1, a_2, n).$$

This rate coverges to unity in the following sense:

Theorem 4 *Suppose that n is a positive integer, and $0 < \pi < 1$. Then*

$$\lim_{a \to \infty} r(\pi a, (1 - \pi)a, n) = 1.$$

The fact that both BB$(a_1, a_2; n)$ and QB$(a_1/(a_1 + a_2); n, 1/(a_1 + a_2))$ are close to Binomial$(n, a_1/(a_1 + a_2))$ when $(a_1 + a_2)$ is large should suffice to prove Theorem 4. Consequently, our acceptance rejection sampling can be very efficient by taking β very close to 0 for fixed π and n.

On the other hand, one may be interested in the efficiency of our sampling for fixed π and β when n is large. Actually, $r(a_1, a_2, n) = O(n^{-a_2})$ as $n \to \infty$. Therefore large n may cause inefficient sampling, but multi-stage sampling can avoid this situation by repeating the division of samples into two groups of cells, where the sum of cell probabilites should be close to 1/2 since it implies smaller a_2. Refer to Table 2 in Appendix 2 for the summary of average acceptance rates r of our QB sampler.

Next we derive the acceptance rate of our sampling for $a_1 < a_2$ depending on y:

$$\frac{p_{QB}(y)}{p_{BB}(y)} \frac{p_{BB}(n)}{p_{QB}(n)} = \frac{\Gamma(a_1 + n)\Gamma(a_2 + 1)}{\Gamma(a_1 + y)\Gamma(a_2 + n - y)} \frac{(a_1 + y)^{y-1}(a_2 + n - y)^{n-y-1}}{(a_1 + n)^{n-1}} =: \rho(y). \quad (8)$$

Consequently our QB sampler is summarized in the following (note that U$(0, 1)$ is for the standard uniform distribution):

Algorithm 1 (Acceptance rejection sampling from the QB distribution) *The following procedure generates a sample from QB$(a_1/(a_1 + a_2); n, 1/(a_1 + a_2))$ for a positive integer n.*

When $0 < a_1 < a_2$,

1. *Generate* $p \sim$ *Beta*(a_1, a_2)
2. *Generate* $y|p \sim$ *Binomial*(n, p)
3. *Generate* $u \sim U(0, 1)$
4. *If* $u > \rho(y)$ *then goto 1*
5. *Output* y.

When $0 < a_2 < a_1$,

1. *Swap* a_2 *and* a_1.
2. *Generate* $p \sim$ *Beta*(a_1, a_2)
3. *Generate* $y|p \sim$ *Binomial*(n, p)
4. *Generate* $u \sim U(0, 1)$
5. *If* $u > \rho(y)$ *then goto 2*
6. *Output* $n - y$.

3 The DPMPM and QM-DPMPM Synthesizers

3.1 The DPMPM synthesizer

The DPMPM is a Bayesian version of latent class models. Consider a sample $\mathbf{X}$ consists of n records, and each record has p unordered categorical variables. The basic assumption of the DPMPM is that every record $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})$ belongs to one of F underlying unobserved/latent classes. Given the latent class assignment z_i of record i, as in Equation (10), each variable X_{ij} independently follows a multinomial distribution, as in Equation (9). Note that d_j is the number of categories of variable j, and $j = 1, \cdots, p$.

$$X_{ij} \mid z_i, \theta \overset{ind}{\sim} \text{Multinomial}(\theta^{(j)}_{z_i 1}, \ldots, \theta^{(j)}_{z_i d_j}; 1) \quad \text{for all } i, j \tag{9}$$

$$z_i \mid \pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_F; 1) \quad \text{for all } i, \tag{10}$$

The DPMPM effectively clusters records with similar characteristics based on all p variables. Relationships among these p categorical variables are induced by integrating out the latent class assignment z_i. To empower the DPMPM to pick the effective number of occupied latent classes, the truncated stick-breaking representation (Sethuraman, 1994) is used as in Equation (11) through Equation (14),

$$\pi_f = V_f \prod_{l<f} (1 - V_l) \quad \text{for } f = 1, \ldots, F \tag{11}$$

$$V_f \overset{iid}{\sim} \text{Beta}(1, \alpha) \quad \text{for } f = 1, \ldots, F-1, \quad V_F = 1 \tag{12}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \tag{13}$$

$$\theta^{(j)}_f = (\theta^{(j)}_{f1}, \ldots, \theta^{(j)}_{fd_j}) \sim \text{Dirichlet}(a^{(j)}_1, \ldots, a^{(j)}_{d_j}). \tag{14}$$

and a blocked Gibbs sampler is implemented for the Markov chain Monte Carlo sampling procedure (Ishwaran and James, 2001; Si and Reiter, 2013; Hu *et al.*, 2014; Drechsler and Hu, 2017+; Manrique-Vallier and Hu, 2018; Hu *et al.*, 2018).

Let p_0 be the number of variables to be synthesized. To generate one partially synthetic dataset $\mathbf{X}^*_{DPMPM}$ using the DPMPM synthesizer, we first generate sample values of (π, α, θ^s) from the posterior distribution (θ^s contains the sample values of variables to be synthesized). Through a multinomial draw with the samples of π, we can

generate the vector of latent class assignments $\{z_i, i = 1, \cdots, n\}$, as in Equation (10). Then through a multinomial draw with samples of θ^s, we can generate synthetic variable $\{X^*_{ij}, i = 1, \cdots, n, j = 1, \cdots, p_0\}$, as in Equation (9).

The sampling process above can also be described as the process of distributing records over different values of $\mathbf{X}$. The probability of the ith record to take the values of $(x_{i1}, x_{i2}, \cdots, x_{ip_0})$ is expressed as $\prod_{j=1}^{p_0} \theta^{(j)}_{z_i x_{ij}}$. We note that any record in the same latent class f has the same probability of taking the values of $(x_1, x_2, \cdots, x_{p_0})$, which is $p(x_1, x_2, \cdots, x_{p_0}; f) = \prod_{j=1}^{p_0} \theta^{(j)}_{f x_j}$.

Regarding $(x_1, x_2, \cdots, x_{p_0})$ as the address of a cell in a p_0 dimensional contingency table, then $p(x_1, x_2, \cdots, x_{p_0})$ gives the cell probability of the corresponding cell. The total number of the cells is $\prod_{j=1}^{p_0} d_j =: D$, and the generation of one record in the fth latent class is equivalent to one multinomial draw from D cells with probabilities $\{p(x_1, x_2, \cdots, x_{p_0}; f), x_j = 1, 2, \cdots, d_j, j = 1, 2, \ldots, p_0\}$. Abbreviating these probabilities as $q_d, d = 1, 2, \cdots, D$, the DPMPM synthesizer actually distributes individuals over D cells as

$$(n_{f1}, n_{f2}, \ldots, n_{fD}) \sim \text{Multinomial}(q_1, q_2, \cdots, q_D; n_f), \tag{15}$$

where n_{fd} denotes the number of records in the fth latent class taking the values of the dth cell, and n_f is the total number of records in the fth latent class.

It is worthy of note that

$$(n_1, n_2, \cdots, n_F) \sim \text{Multinomial}(\pi_1, \pi_2, \cdots, \pi_F; n). \tag{16}$$

This additional view of the DPMPM leads to our QM-DPMPM synthesizer in the next subsection.

3.2 The QM-DPMPM synthesizer

The QM-DPMPM synthesizer just replaces the multinomial draw in Equation (15) of the DPMPM with

$$(n_{f1}, n_{f2}, \ldots, n_{fD}) \sim \text{QM}(q_1, q_2, \cdots, q_D; n_f, \beta). \tag{17}$$

Hence the QM-DPMPM obviously reduces to the DPMPM when $\beta = 0$. The parameter β is subjectively selected to take the balance of utility and disclosure risk of the synthetic data.

We provide a succinct overview of the QM-DPMPM synthesizer procedure. First we generate sample values of (π, α, θ^s) from the posterior distribution based on the DPMPM model. Through a multinomial draw with the samples of π, we can multinomially distribute n individuals over F latent classes with cell probabilities $\pi = (\pi_1, \ldots, \pi_F)$, as in Equation (16). Finally, we quasi-multinomially distribute n_f individuals over D cells, as in Equation (17). Note that n_{fd} is the number of individuals in cell d of class f.

As we can see, the QM-DPMPM synthesizer generates counts of combinations of synthesized variables. Once the count values of $\{n_{fd}, f = 1, \cdots, F, d = 1, \cdots, D\}$ are drawn, a partially synthetic dataset $\mathbf{X}^*_{QM-DPMPM,\beta}$ is obtained by duplicating the

combinations of variables with $n_{fd} > 1$, and keeping the combinations of variables with $n_{fd} = 1$. Eventually, these synthesized combinations are attached to the un-synthesized variables to produce the partially synthetic dataset $\mathbf{X}^*_{QM-DPMPM,\beta}$.

3.3 Notes on implementation

The `NPBayesImpute` R package is used for the DPMPM implementation. After the Markov chain Monte Carlo (MCMC) is converged, we generate $\mathbf{X}^*_{DPMPM}$ and $\mathbf{X}^*_{QM-DPMPM,\beta}$ within the Gibbs sampler and save these synthetic datasets. We repeat the above processes $m > 1$ times to obtain m synthetic datasets, using approximately independent draws of parameters obtained at MCMC iterations that are far apart.

4 Illustrative Application

We apply the DPMPM and QM-DPMPM synthesizers on a subset of a public available 2012 American Community Survey (ACS) sample. A similar dataset was used in Hu *et al.* (2014). We choose only p = 10 unordered categorical variables from the original p = 14 in Hu *et al.* (2014) because both the DPMPM synthesizer and the QM-DPMPM synthesizer have been developed for unordered categorical variables. To work with ordered categorical variables such as categorized age variables (levels: 1 = 18-29, 2 = 30-44, 3 = 45-59, 4 = 60+), methods such as probit models are needed. The multinomial-based synthesizers cannot incorporate the inherent order in those variables properly.

Our sample has n = 10,000 records and p = 10 unordered categorical variables. We synthesize 5 sensitive variables {SEX, RACE, DIS, HICOV, HISP}, and keep the remaining 5 variables un-synthesized {MAR, LANX, WAOB, MIG, SCH}. See Table 1 for the description and synthesis information of each variable.

We use the methods described in Section 3.1 and Section 3.2 to generate partially synthetic data from the DPMPM and QM-DPMPM synthesizers, respectively. For the QM-DPMPM synthesizer, we consider a sequence of 1000 values of β to assess its effect on the QM-DPMPM synthesizer ($\beta \in \{0.9991, 0.9981, \cdots, 0.0011, 0.0001\}$).

We generate $m = 20$ synthetic datasets from the DPMPM synthesizer, and $m = 20$ synthetic datasets from the QM-DPMPM synthesizer from one of the 1,000 β values ($\beta \in \{0.9991, \cdots, 0.0001\}$). We evaluate the utility and identification disclosure risks of each set of $m = 20$ synthetic datasets. We then evaluate and compare the utility and disclosure risks of $\mathbf{X}^*_{QM-DPMPM,\beta}$ for the range of β to those of $\mathbf{X}^*_{DPMPM}$.

4.1 Utility evaluation

We first compare relative frequencies for various cross tabulations of all 10 variables in the original dataset and in the synthetic datasets. Specifically, we compute the relative frequencies for all one-way tables, two-way tables, and three-way tables. We then compare these relative frequencies in the original data to the synthetic data. Our approach is a modified version of that in Hu *et al.* (2014); Drechsler and Hu (2017+).

Table 1: Variables in the ACS sample, taken from the 2012 ACS public use microdata samples. PR stands for Puerto Rico. The Synthesized column records whether the variable is synthesized (yes) or not (no). The Known column records whether the variable is known to the intruder (yes) or not (no), for identification disclosure risks evaluation.

Variable	Categories	Synthesized	Known
SEX	1 = male, 2 = female	Yes	Yes
RACE	1 = White alone, 2 = Black or African American alone, 3 = American Indian alone, 4 = other, 5 = two or more races, 6 = Asian alone	Yes	Yes
DIS	1 = has a disability, 2 = no disability	Yes	No
HICOV	1 = has health insurance coverage, 2 = no coverage	Yes	No
HISP	1 = not Spanish, Hispanic, or Latino, 2 = Spanish, Hispanic, or Latino	Yes	No
MAR	1 = married, 2 = widowed, 3 = divorced, 4 = separated, 5 = never married	No	Yes
MIG	1 = live in the same house (non movers), 2 = move to outside US and PR, 3 = move to different house in US or PR	No	Yes
LANX	1 = speaks another language, 2 = speaks only English	No	No
WAOB	born in: 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	No	No
SCH	1 = has not attended school in the last 3 months, 2 = in public school or college, 3 = in private school or college or home school	No	No

Figure 1 shows a clear trend of decreasing three-way deviation as β value decreases from 0.9991 to .0001. Plots of one-way and two-way deviation also show clear decreasing trends, and they are omitted for brevity. The shown trend is not surprising. Recall that the parameter β in the QM-DPMPM synthesizer effectively controls its similarity to the DPMPM synthesizer: larger β is associated with the larger variance of any univariate marginal frequency, resulting larger differences from the DPMPM synthesizer. Recall also that the QM-DPMPM synthesizer with $\beta = 0$ is the same as the DPMPM synthesizer. Figure 1 indicates that deviation-based utility of the QM-DPMPM synthesizer is higher when β decreases and approaches 0, and should be the highest (equal to that of the DPMPM synthesizer) when $\beta = 0$. The utility increases much more quickly when β drops under 0.25.

Analysts are often interested in regression analyses using synthetic data. To assess regression-based utility, we run logistic regression of disability status (DIS) on a number of predictors: health insurance coverage (HICOV), migration status (MIG), language use (LANX) and schooling (SCH). All predictors are treated as categorical. To deal

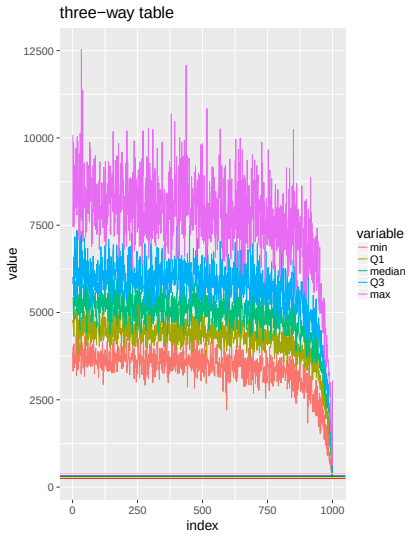

Fig. 1: Three-way table of utility of QM synthesizer with 1000 β's, from 0.9991 (ind. 1) to 0.0001 (ind. 1000). The minimum, Q1, median, Q3, and maximum of the utility of DPMPM synthesizer are 251.38, 284.95, 298.05, 318.66, and 390.78 respectively, and marked on the plot.

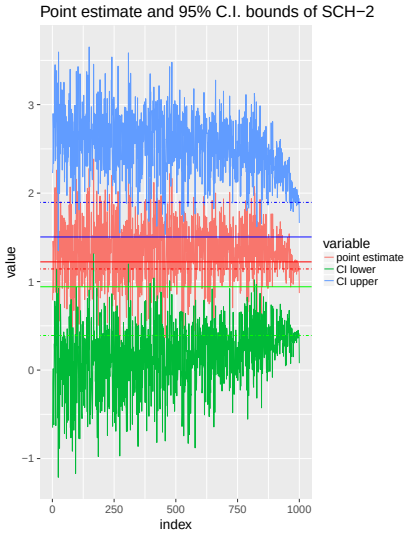

Fig. 2: Utility based on logistic regression coefficient estimates and 95% confidence intervals, for the SCH-2 predictor. Solid horizontal lines are values based on the original data; dashed horizontal lines are values based on the DPMPM synthetic data. β value ranges from 0.9991 (ind. 1) to 0.0001 (ind. 1000).

with the separation problem in logistic regression, we use the `logistf` R package to fit a logistic regression model using Firth's bias reduction method (Firth, 1993) to the original data, the DPMPM synthetic data, and the QM-DPMPM synthetic data.

We obtain the point estimates and 95% confidence intervals from the original and synthetic data and then make comparison to see how close the inferences are (closer to the original means higher utility). For the DPMPM synthetic data and the QM-DPMPM synthetic data, we obtain the point estimates and the 95% confidence intervals by using the combining rules for inference based on partial synthetic data (Reiter, 2003; Drechsler, 2011).

Using the results from the DPMPM synthetic data as a benchmark, we note that while some coefficients are preserved reasonably well (e.g. Intercept, MIG-3, LANX-2 and SCH-2), some differ to some degree (e.g. HICOV-2). However, every 95% confidence interval based on the DPMPM synthetic data includes the point estimate from the original data, indicating a reasonably high level of utility. Table 3 in Appendix 3 contains the detailed results.

Figure 2 plots inferences of the logistic regression coefficient of predictor SCH-2 based on the original data (the horizontal solid line), the DPMPM synthetic data (the

horizontal dashed line), and the QM-DPMPM synthetic data (from 0.9991 (ind. 1) to .0001 (ind. 1000)). Plots on the remaining 6 coefficients show similar pattern, and are omitted for brevity. These results show that overall, the larger the β value, the greater the distance between the inferences based on the original data and those based on the QM-DPMPM synthetic data, indicating less accuracy in the inferences. Moreover, as β decreases, QM-DPMPM inference converges to DPMPM inference, showing the same trend of decreasing one-way, two-way, and three-way deviation as β decreases.

4.2 Identification disclosure risks evaluation

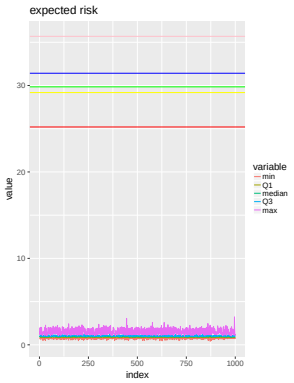

Fig. 3: Expected risk of QM-DPMPM synthesizer with 1000 β's, from 0.9991 (ind. 1) to .0001 (ind. 1000).

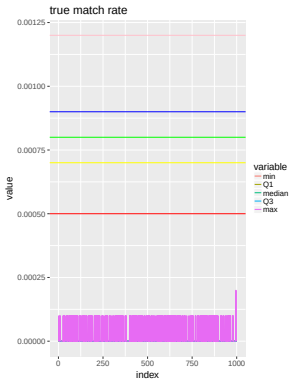

Fig. 4: True match rate of QM-DPMPM synthesizer with 1000 β's, from 0.9991 (ind. 1) to .0001 (ind. 1000).

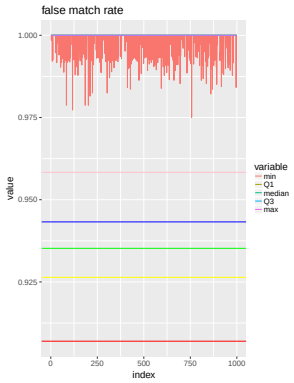

Fig. 5: False match rate of QM-DPMPM synthesizer with 1000 β's, from 0.9991 (ind. 1) to .0001 (ind. 1000).

For partially synthetic data, both identification disclosure and attribute disclosure risks possibly exist (Drechsler, 2011; Hu, 2018+). For illustrative purpose, we consider the identification disclosure risks in our application. That is, we evaluate the probability of identifying a record in the sample by matching with available external information.

Our evaluation approach is a Bayesian probabilistic matching procedure (Duncan and Lambert, 1986, 1989; Lambert, 1993; Fienberg *et al.*, 1997; Reiter, 2005; Drechsler and Reiter, 2008; Reiter and Mitra, 2009; Drechsler and Reiter, 2010; Drechsler and Hu, 2017+). This general evaluation procedure considers the matching probability of a target vector $\mathbf{t}$ available to the intruder. This target record $\mathbf{t}$ contains some un-synthesized variables that are available through external files, denoted as $\mathbf{t}^{A_{us}}$, and some other synthesized and available variables, denoted as $\mathbf{t}^{A_s}$. Therefore $\mathbf{t} = (\mathbf{t}^{A_{us}}, \mathbf{t}^{A_s})$.

The identification disclosure risks evaluation aims at estimating the probability of the intruder being able to identify a record i with the available target vector $\mathbf{t}$, by using the knowledge of un-synthesized variables in $\mathbf{t}^{A_{us}}$ and guessing the synthesized variables in $\mathbf{t}^{A_s}$. In the end, three summaries of identification disclosure probabilities are

produced: i) the expected match risk, an overall summary of all target records being the true match among all records with the highest match probability (similar to the measure proposed in Franconi and Polettini (2004)); ii) the true match rate, the percentage of true unique matches among the target records; and iii) the false match rate, the percentage of false matches among unique matches (Reiter and Mitra, 2009; Drechsler and Reiter, 2010; Drechsler, 2011; Drechsler and Hu, 2017+; Hu, 2018+).

In our application, we assume the intruder knows the sex, race, martial status and migration status of all respondents through external files. Among these variables in $\mathbf{t}$, sex and race are synthesized, while marital status and migration statues are unsynthesized. Therefore, $\mathbf{t}^{A_{us}}$ = {MAR, MIG} and $\mathbf{t}^{A_s}$ = {SEX, RACE}. We treat all n = 10,000 records in the sample as target records. For each of $m = 20$ synthetic datasets we calculate risk measures, and the five number summary of these 20 risk values is plotted for each β in Figure 3 to Figure 5. The five number summary is useful to evaluate the many scenarios of an intruder, who may combine the information of multiple synthetic datasets in various ways.

Figure 3 shows that the expected risk is stable against the change of β, which may seem opposite to our intuition. Nevertheless it reflects the fact that observed true matches are few. Let X be the number of matched sample records to a target record. The expected match risk increases by $1/X$ only when the target record is truly matched among the X records. Hence the expected match risk is close to zero when true matches are few. Then even the expectation of the expected match risk is increasing as $\beta \to 0$, it is hard to observe the increasing trend of the realized expected match risk.

The stable discrepancy of the true match rates in Figure 4 between the QM-DPMPM and the DPMPM may look large, but this discrepancy is not large in a stochastic sense. To see this fact, let us focus on the number of true unique matches; in the case of DPMPM ($\beta = 0$) it ranges from 5 to 12, which are not very far from 0 to 2 of $\beta > 0$.

To confirm the convergence of the true match rate of the QM-DPMPM to that of the DPMPM, we need even smaller β, but we observe the clear convergence of unique match counts of the QM-DPMPM to that of the DPMPM (a plot is omitted for brevity).

5 Concluding Remarks

The properties of the QM distribution and its tuning parameter β have motivated us to develop a QM-DPMPM synthesizer based on the DPMPM synthesizer. We have seen that around $\beta = 0$, the utilities of the QM-DPMPM synthesizer measured in Section 4.1 are very close to those of the DPMPM. On the other hand, risk measures dealt with in Section 4.2 show various speeds of convergence as $\beta \to 0$. This difference implies that with only a slight loss of utility, a statistical agency may be able to generate a much safer data set by employing the QM-DPMPM synthesizer.

We believe the QM-DPMPM synthesizer is a promising method of generating synthetic categorical data that is worth further investigation. It would be interesting to experiment with smaller β values and evaluate the utility-risks tradeoff. Additionally, many other multinomial distribution based categorical data synthesizers can be turned into a QM distribution based synthesizer with desired utility-risks balance.

Bibliography

Akande, O., Li, F., and Reiter, J. P. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician* **71**, 162–170.

Akande, O., Reiter, J. P., and Barrientos, A. F. (2017+). Multiple imputation of missing values in household data with structural zeros. *arXiv:1707.05916* .

Consul, P. C. and Mittal, S. P. (1975). A new urn model with predetermined strategy. *Biometrische Zeitschrift* **17**, 67–75.

Consul, P. C. and Mittal, S. P. (1977). Some discrete multinomial probability models with predetermined strategy. *Biometrische Zeitschrift* **19**, 161–173.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*. Springer: New York.

Drechsler, J. and Hu, J. (2017+). Strategies to facilitate access to detailed geocoding information based on synthetic data. arXiv:1803.05874.

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases*, vol. 5262 of *Lecture Notes in Computer Science*, 227–238. Springer.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach to releasing public use microdata samples of census data. *Journal of the American Statistical Association* **105**, 1347–1357.

Duncan, G. T. and Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association* **10**, 10–28.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7**, 207–217.

Dunson, D. B. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**, 1042–1051.

Fienberg, S. E., Makov, U., and Sanil, A. P. (1997). A bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics* **13**, 75–89.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Franconi, L. and Polettini, S. (2004). Individual risk estimation in $\hat{\mathrm{I}}\frac{1}{4}$-argus: A review. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, vol. 3050 of *Lecture Notes in Computer Science*, 262–272. Springer.

Ho, F. C. M., Gentle, J. E., and Kennedy, W. J. (1979). Generation of random variates from the multinomial distribution. *Proceedings of the American Statistical Association Statistical Computing Section* .

Hoshino, N. (2009). The quasi-multinomial distribution as a tool for disclosure risk assessment. *Journal of Official Statistics* **25**, 269–291.

Hu, J. (2018+). Bayesian estimation of attribute and identification disclosure risks in synthetic data. arXiv:1804.02784.

Hu, J., Reiter, J. P., and Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. In J. Domingo-Ferrer, ed., *Privacy in Statistical Databases*, vol. 8744 of *Lecture Notes in Computer Science*, 185–199. Springer.

Hu, J., Reiter, J. P., and Wang, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis* **13**, 183–200.

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

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics* **9**, 313–331.

Malefaki, S. and Iliopoulos, G. (2007). Simulating from a multinomial distribution with large number of categories. *Computational Statistics and Data Analysis* **51**, 5471–5476.

Manrique-Vallier, D. and Hu, J. (2018). Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society, Series A* to appear.

Manrique-Vallier, D. and Reiter, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics* **23**, 1061–1079.

Murray, J. S. (2018+). Multiple imputation: a review of practical and theoretical findings. *Statistical Science* .

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–188.

Reiter, J. P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association* **100**, 1103–1112.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *The Journal of Privacy and Confidentiality* **1**, 99–110.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

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

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.

Appendix

Appendix 1

Proof of Theorem 3: By symmmetry it suffices to show the case of $a_1 < a_2$. To simplify our argument, write $g(y;a) = \Gamma(a+y)/(\Gamma(a+1)(a+y)^{y-1})$. Then $p_{BB}(y)/p_{QB}(y) = g(y;a_1)g(n-y;a_2)/g(n;a_.)$. Hence $p_{BB}(y)/p_{QB}(y)$ is minimized when $g(y;a_1)g(n-y;a_2)$ is minimized. Also we note that $g(y;a)r(y;a) = g(y+1;a)$, where $r(y;a) = (a+y)^y/(a+y+1)^y$. Since $r(0,a) = 1$ and $r(y,a) < 1$ for $y \geq 1$, $g(y;a)$ is monotonically decreasing when y increases. Therefore to minimize $g(y;a_1)g(n-y;a_2)$, we increase y_1 or y_2 one by one so that $g(y_1;a_1)g(y_2;a_2)$ is more reduced. Denote the ith step values by $(y_1,y_2)_i, i = 1,\ldots,n$. At each step y_1 increases by one if $r(y_1,a) \leq r(y_2,a)$ otherwise y_2 increases by one. Actually $r(0,a_1) = r(0,a_2) = 1$, but if $0 < a_1 < a_2$ then $r(1,a_1) < r(1,a_2)$. Hence we begin with $(y_1,y_2)_1 = (1,0)$. Observing

$$\begin{aligned}\frac{d\log r(y,a)}{dy} &= \log\left(1-\frac{1}{a+y+1}\right) + \frac{y}{(a+y)(a+y+1)} \\ &= -\frac{1}{a+y+1} - \frac{1}{2}(\frac{1}{a+y+1})^2 - \cdots + \frac{y}{(a+y)(a+y+1)} < 0,\end{aligned}$$

we note that $r(y,a) > r(y',a)$ when $y < y'$. Therefore $r(1,a_2) > r(1,a_1) > r(i,a_1)$ for $i \geq 2$, which leads to $(y_1,y_2)_n = (n,0)$. □

Appendix 2

A summary of average acceptance rates r of our QB sampler proposed in Section 2.

Table 2: Average Acceptance Rates: $r(0.1/\beta, 0.9/\beta, n)$

n	β									
	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}
10^1	0.13	0.06	0.03	0.02	0.03	0.06	0.14	0.29	0.50	0.69
10^2	0.00	0.00	0.01	0.07	0.26	0.51	0.71	0.84	0.92	0.96
10^3	0.12	0.34	0.59	0.77	0.87	0.94	0.97	0.98	0.99	1.00
10^4	0.81	0.90	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00
10^5	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10^6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10^7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10^8	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
10^9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Appendix 3

The table contains detailed results of regression-based utility of the DPMPM synthetic data in Section 4.

Table 3: 95% confidence intervals of logistic regression coefficients based on the original data and based on the m = 20 synthetic data generated by the DPMPM synthesizer.

	Original data		DPMPM (m = 20)	
Estimand	Estimate	95% CI	$\bar{q}_{20}$	95% CI
Intercept	2.12	[1.86, 2.39]	2.27	[1.53, 3.01]
HICOV - 2	0.60	[0.43, 0.77]	0.14	[-0.41, 0.69]
MIG - 2	0.39	[-0.61, 1.39]	0.06	[-1.31, 1.43]
MIG - 3	0.04	[-0.12, 0.19]	0.06	[-0.50, 0.63]
LANX - 2	-0.87	[-1.14, -0.60]	-0.96	[-1.71, -0.22]
SCH - 2	1.22	[0.94, 1.50]	1.14	[0.39, 1.90]
SCH - 3	1.63	[1.01, 2.25]	1.07	[0.07, 2.07]

Appendix 4

The table contains the minimum, first quartile, median, third quartile, and maximum of the identification disclosure risks of the DPMPM synthesizer. They are all marked as horizontal lines in Figure 3 to Figure 5.

Table 4: Table of the minimum, first quartile (Q1), median, third quartile (Q3), and maximum of the identification disclosure risks of the DPMPM synthesizer.

Summary	Min	Q1	Median	Q3	Max
Expected risk	25.2011	29.1804	29.8434	31.4175	35.6882
True match rate	0.0005	0.0007	0.0008	0.0009	0.0012
False match rate	0.9070	0.9264	0.9352	0.9433	0.9583