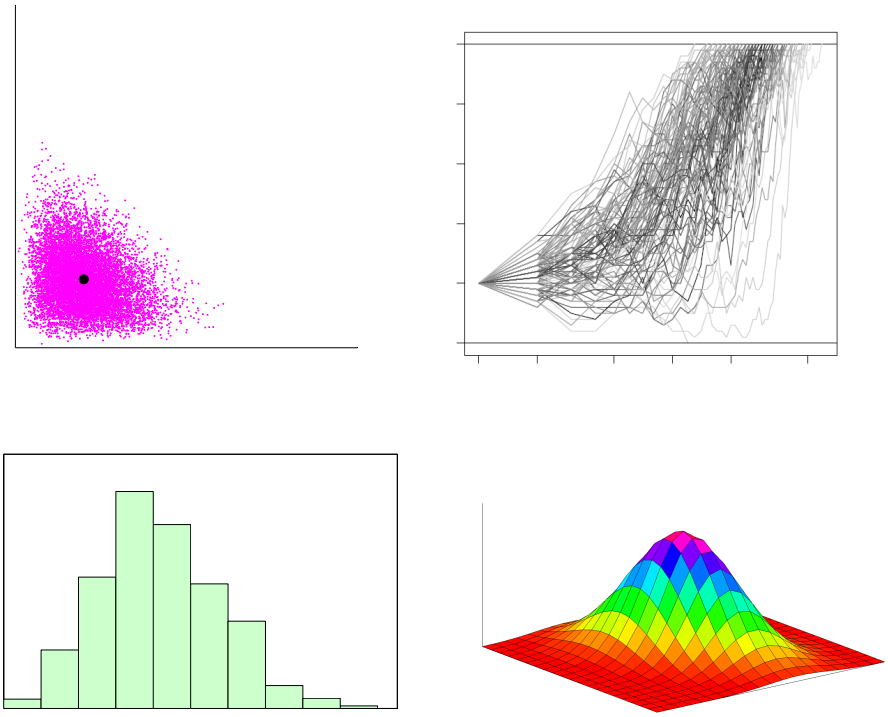
Philipps Universität Marburg

Discussion Papers on Statistics and Quantitative Methods

Using Prior Information in Privacy-Protecting Survey Designs for Categorical Sensitive Variables

Heiko Groenitz

1 / 2013



Download from:
http://www.uni-marburg.de/fb02/statistik/forschung/discpap

Coordination: Prof. Dr. Karlheinz Fleischer • Philipps-University Marburg
Faculty of Business Administration • Department of Statistics
Universitätsstraße 25 • D-35037 Marburg
E-Mail: k.fleischer@wiwi.uni-marburg.de

Using Prior Information in Privacy-Protecting Survey Designs for Categorical Sensitive Variables

Heiko Groenitz[1]

02.01.2013

Abstract

To gather data on sensitive characteristics, such as annual income, tax evasion, insurance fraud or students' cheating behavior, direct questioning is not helpful, because it results in answer refusal or untruthful responses. For this reason, several randomized response (RR) and nonrandomized response (NRR) survey designs, which increase cooperation by protecting the respondents' privacy, have been proposed in the literature. In the first part of this paper, we present a Bayesian extension of a recently published, innovative NRR method for multichotomous sensitive variables. With this extension, the investigator is able to incorporate prior information on the parameter, e.g. based on a previous study, into the estimation and to improve the estimation precision. In particular, we calculate posterior modes with the EM algorithm as well as estimates based on parameter simulation, multiple imputation, and Rao-Blackwellization. The performance of these estimation methods is evaluated in a simulation study. In the second part of this article, we show that for any RR or NRR model, the design matrices of the model play the central role for the Bayes estimation whereas the concrete answer scheme is irrelevant. This observation enables us to widely generalize the calculations from the first part and to establish a common approach for the Bayes inference in RR and NRR designs for categorical sensitive variables. This unified approach covers even multi-stage models and models that require more than one sample.

Zusammenfassung

Zur Datenerhebung bei sensitiven Merkmalen wie Einkommen, Steuerhinterziehung, Versicherungsbetrug oder Prüfungsbetrug ist Direktbefragung problematisch, da sie oft zu Antwortverweigerungen oder Falschantworten führt. Aus diesem Grund wurden in der Literatur verschiedene Randomized-Response- und Nonrandomized-Response-Umfrageverfahren (kurz RR- und NRR-Verfahren), welche die Privatsphäre der Befragten schützen und dadurch deren Kooperationsbereitschaft erhöhen, vorgeschlagen. Im ersten Teil dieses Aufsatzes präsentieren wir eine Bayes-Erweiterung eines kürzlich publizierten NRR-Modells für kategoriale sensitive Merkmale. Durch diese Erweiterung ist es möglich Vorinformation über den Parameter, die zum Beispiel auf einer vorherigen Erhebung basieren könnte, in die Schätzung einzubeziehen und dadurch die Schätzgenauigkeit zu verbessern. Wir ermitteln den Modus der a-posteriori-Verteilung mit dem EM-Algorithmus und berechnen Schätzer basierend auf Parametersimulation, multipler Imputation und Rao-Blackwellisierung. Diese Schätzverfahren werden im Rahmen einer Simulationsstudie verglichen. Im zweiten Teil des Artikels zeigen wir, dass die Designmatrizen des Modells bei jedem RR- / NRR-Modell für kategoriale sensitive Merkmale die zentrale Rolle für die Bayes-Schätzung spielen wohingegen die konkrete Antwortformel irrelevant ist. Diese Beobachtung ermöglicht es uns die Rechnungen aus dem ersten Teil des Aufsatzes weitreichend zu verallgemeinern und einen gemeinsamen Ansatz für die Bayes-Schätzung bei RR- / NRR-Verfahren zu entwickeln. Dieser vereinheitlichte Ansatz deckt sogar mehrstufige Modelle sowie Modelle, welche mehrere Stichproben benötigen, ab.

KEYWORDS: Randomized response; Nonrandomized response; Bayesian estimation; EM algorithm; Data augmentation

[1]Philipps-University Marburg, Department for Statistics (Faculty 02), Universitätsstraße 25, 35032 Marburg, Germany (e-mail: groenitz@staff.uni-marburg.de).

1 Introduction

Let us consider a survey on a sensitive attribute X. For instance, X may represent income classes or the number of times the respondent has evaded taxes. In the case of direct questioning (DQ), many respondents will not reveal the true value of X. Instead, answer refusal and untruthful responses will occur. This leads to a serious bias when estimating the distribution of X based on DQ. For this reason, several randomized response (RR) and nonrandomized response (NRR) techniques have been developed in the literature to obtain trustworthy estimates of the distribution of X. To protect privacy, the respondents are always requested to provide a scrambled answer A instead of the X-value. This practice reduces untruthful answers and answer refusal. The realizations of A and X are observed and missing data, respectively.

A RR technique was first proposed by Warner (1965), whose seminal model has been extended in various dimensions until today. RR models have in common that every respondent is supplied with a randomization device (RD), such as a coin or a deck of cards. The respondents use the RD to conduct a random experiment, whose outcome influences the required scrambled answer. The necessity of running the random experiment is cumbersome. This is why nonrandomized response approaches are coming up in recent years with articles by Tian et al. (2007), Yu et al. (2008), Tan et al. (2009), Tang et al (2009) and Groenitz (2012). NRR models do not need a RD; in such models, the answer depends on an auxiliary variable, and the respondent would give the same answer if he or she was asked again. NRR methods are easy to implement and suitable for face-to-face and e-mail surveys. Compared with RR techniques, NRR methods reduce both survey complexity and study costs.

In privacy-protecting (PP) models (i.e., RR or NRR designs), maximum likelihood (ML) estimates can be derived from the empirical distribution of the scrambled answers. However, for the case in which prior information on the distribution of interest is available, Bayesian methods should be applied to incorporate the prior information. Bayesian estimation means that we collect the prior information in a prior distribution and analyze the observed data posterior distribution. Note that even if there is no prior information, the Bayesian approach with a uniform prior distribution can be recommendable: for this prior, the posterior mode equals the ML estimator (MLE). However, in small samples, the posterior standard deviation and confidence intervals based on posterior quantiles can be expected to be more suitable than the asymptotic standard error of the MLE and confidence intervals based on the asymptotic normality of the MLE.

Bayesian methods (usually based on a Dirichlet prior) have been proposed for some PP designs: Winkler and Franklin (1979) as well as Migon and Tachibana (1997) present Bayesian approaches for Warner's (1965) RR model. O'Hagan (1987) derives Bayes linear estimators for Warner's model and the unrelated question model (UQM) by Horvitz et al. (1967). Unnikrishnan and Kunte (1999) describe a unified model for Warner's model and the UQM as well as a unified model for the common handling of the model by Abul-Ela et al. (1967) and the polychotomous UQM by Greenberg et al. (1969). For both unified models, the Gibbs sampler is used to generate realizations from the posterior distribution. Bayesian inference for Mangat's (1994) RR model can be found in Kim et al. (2006). Tang et al. (2009) suggest a certain NRR model and explain the corresponding Bayesian estimation. Bayesian methods for the NRR methods by Tian et al. (2007) and Yu et al. (2008) can be found in Tian et al. (2009). Barabesi and Marcheselli (2010) propose a Bayesian approach to the joint estimation of the distribution of a binary sensitive variable and the sensitivity level from data collected with a certain two-stage RR scheme. The Bayes estimation for the RR model by Mangat and Singh (1990) is derived in Hussain et al. (2011).

In the first part of this paper, we extend the work by Groenitz (2012), who presents the nonrandomized diagonal model (DM) including ML estimation, in order to have the possibility to incorporate prior information into the estimation and to obtain more precise estimates. In Section 2, we narrate

the diagonal model and derive Bayesian estimates for this model. In particular, we calculate posterior modes via the EM algorithm as well as estimates based on parameter simulation (PS), multiple imputation (MI) and Rao-Blackwellization (RB) for the DM survey design. For PS, MI, RB, the data augmentation algorithm, which generates certain Markov chains, turns out to be beneficial. The quality of PS, MI, RB for a survey according to the diagonal model is investigated in a simulation study.

For the DM, we observe in Section 2 that the design matrix of the model, i.e., a matrix of conditional probabilities, plays the central role for the calculation of posterior modes and estimates based on PS, MI, RB. In the second part of this paper, we show the following generalization of this observation: For any PP survey model dealing with categorical X, the only component of the model that is needed to compute Bayes estimates is the set of design matrices of the model. The concrete answer scheme is irrelevant for Bayes inference. This result enables us to establish a common approach for the Bayes estimation in PP survey designs for categorical sensitive variables in Section 3. This unified approach covers many published and potential PP designs including certain multi-stage designs and designs demanding multiple samples. Here, we derive general formulas that can be applied to a lot of PP models for which Bayesian concepts have not been discussed yet.

2 Bayes estimation for the diagonal model

2.1 Diagonal model

Groenitz (2012) proposed the diagonal model (DM), which can be applied to gather data on a sensitive characteristic $X \in \{1, ..., k\}$. For the DM, a nonsensitive auxiliary variable $W \in \{1, ..., k\}$ (e.g., W may describe the period of birthday) must be specified such that X and W are independent and that the distribution of W is known. The respondent is introduced to give the answer

$$A := [(W - X) \bmod k] + 1. \tag{1}$$

Equation (1) should not be shown to the respondents; instead, every interviewee receives a table that illustrates (1). E.g., for $k = 4$, we have

X/W	$W = 1$	$W = 2$	$W = 3$	$W = 4$
$X = 1$	**1**	**2**	**3**	**4**
$X = 2$	**4**	**1**	**2**	**3**
$X = 3$	**3**	**4**	**1**	**2**
$X = 4$	**2**	**3**	**4**	**1**

The number in the interior of the table is the required answer A. Notice, the answers A do not restrict the possible X-values. Hence, we assume that the interviewees cooperate and reveal their values of A. We remark that the DM is applicable even if all the values of X are sensitive (e.g., if the values of X correspond to income classes).

Throughout this article, let π_i, c_i, λ_i be the proportion of units in the population having attribute $X = i$, $W = i$, $A = i$, respectively. Moreover, define $C(i, j)$ to be the proportion of individuals having $A = i$ among the persons with $X = j$. We then have $(\lambda_1, ..., \lambda_k)^T = C \cdot (\pi_1, ..., \pi_k)^T$ with the $k \times k$ matrix $C = [C(i, j)]_{ij}$, where every row of C is a left-cyclic shift of the row above and the first row of C is equal to $(c_1, ..., c_k)$. C is called the "design matrix" and plays an important role for the Bayes estimation in the DM.

2.2 Basic principles and definitions for Bayes estimation

We assume a simple random sample with replacement (SRSWR) of n units has been drawn. These n persons are introduced to answer according to the DM answer formula (1). Let X_i and A_i be the i-th respondent's value of X and A, respectively. Consequently, $\mathbf{A} = (A_1, ..., A_n)$ and $\mathbf{X} = (X_1, ..., X_n)$ represent the observed data and the missing data, respectively. Thus, a DM survey generates a data structure that corresponds to a special missing data problem. For this reason, we can apply known missing data methods, e.g., EM algorithm or data augmentation, to incorporate prior information into the estimation for the DM.

In the subsequent subsections, we derive Bayes estimates for the unknown $\pi = (\pi_1, ..., \pi_{k-1})^T \in \mathbb{R}^{k-1}$. In a Bayesian view, π is treated as a realization of a random variable Π. The prior information about π is collected in a prior distribution defined by a density f_Π, which is specified by the investigator. In this article, we focus on Dirichlet prior distributions. In Subsection 2.3, we explain a possibility to convert prior information into a concrete Dirichlet distribution. In addition to f_Π, the conditional distribution of the complete data $(\mathbf{X}, \mathbf{A})$ given Π must be defined. We denote the corresponding density by $f_{\mathbf{X},\mathbf{A}\,|\,\Pi}(\cdot, \cdot\,|\,\pi)$, and set for $x_j, a_j \in \{1, ..., k\}$

$$f_{\mathbf{X},\mathbf{A}\,|\,\Pi}(\mathbf{x}, \mathbf{a}\,|\,\pi) = \prod_{j=1}^{n} C(a_j, x_j) \cdot \pi_{x_j}, \tag{2}$$

where $\mathbf{x} = (x_j)_j$, $\mathbf{a} = (a_j)_j$. That is, we have conditional independence of the n vectors (X_j, A_j) given Π. It follows that

$$f_{\mathbf{X}\,|\,\mathbf{A},\Pi}(\mathbf{x}\,|\,\mathbf{a}, \pi) = \prod_{j=1}^{n} \frac{C(a_j, x_j) \cdot \pi_{x_j}}{f_{A_j\,|\,\Pi}(a_j\,|\,\pi)}, \tag{3}$$

where $f_{A_j\,|\,\Pi}(\alpha\,|\,\pi)$ is the entry number $\alpha \in \{1, ..., k\}$ of vector $C \cdot (\pi_1, ..., \pi_k)^T$.

Assume a value $\mathbf{a}$ of $\mathbf{A}$ has been observed in the survey. The basic idea is to evaluate the posterior distribution of Π given $\mathbf{a}$ and the distribution of $\mathbf{X}$ given $\mathbf{a}$. In Subsection 2.4, we compute posterior modes with the EM algorithm, and in 2.5, we describe ways based on the data augmentation algorithm (in particular, parameter simulation and multiple imputation) to estimate the true proportion π. Estimators derived by the idea of Rao-Blackwell's theorem are considered in 2.6.

2.3 Dirichlet prior distributions

The random vector $\Pi = (\Pi_1, ..., \Pi_{k-1})$ is Dirichlet distributed if it has Lebesgue density

$$f_\Pi(\pi) = f_\Pi(\pi_1, ..., \pi_{k-1}) = K \cdot \pi_1^{\delta_1 - 1} \cdots \pi_{k-1}^{\delta_{k-1} - 1} \cdot (1 - \sum_{i=1}^{k-1} \pi_i)^{\delta_k - 1} \cdot 1_{E_{k-1}}(\pi), \tag{4}$$

where $E_{k-1} = \{(x_1, ..., x_{k-1}) \in [0,1]^{k-1} \,:\, x_1 + ... + x_{k-1} \leq 1\}$, $\delta = (\delta_1, ..., \delta_k)$ is a vector of parameters with $\delta_i > 0$ and K is a constant depending on δ. We will usually write $\Pi \sim Di(\delta)$ in the sequel. Let us assume that $(\hat{\pi}_1^{(p)}, ..., \hat{\pi}_k^{(p)})^T$ is the investigator's guess for the unknown proportions. This guess may be based on a previous study. One option to convert this guess into a Dirichlet distribution is as follows. Choose a proportionality factor d, and define δ_i to be proportional to $\hat{\pi}_i^{(p)}$, i.e, $\delta_i = \hat{\pi}_i^{(p)} \cdot d$. Let $(D_1, ..., D_{k-1})$ be Dirichlet distributed with these δ_i. Then, we have $\mathbb{E}(D_i) = \hat{\pi}_i^{(p)}$ and $Var(D_i) = \hat{\pi}_i^{(p)}(1 - \hat{\pi}_i^{(p)})/(d+1)$. Obviously, small and large d result in a large and small variance, respectively. If the investigator feels certain that his or her guess is close to the true vector of proportions for the current study, a relatively large d should be chosen. If the investigator is unsure, a relatively small d will reflect this uncertainty.

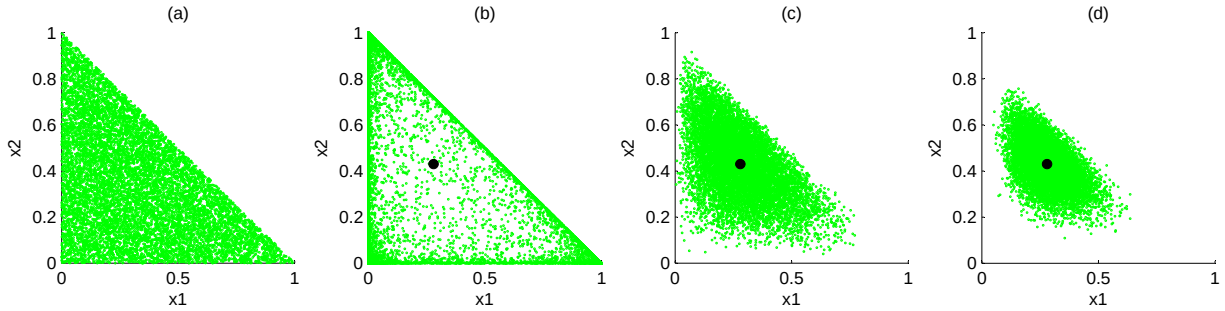


Figure 1: *Scatter plots of each 10000 random numbers from several Dirichlet distributions. In (a), we have* $\delta = (1, 1, 1)$*, for (b)-(c) we use* δ_i *as described in Subsection 2.3 where* $d = 0.5$ *in (b),* $d = 10$ *in (c) and* $d = 25$ *in (d). The black point equals* $(0.28, 0.43)$*, which is the investigator's guess for the unknown* π_1 *and* π_2*.*

The scatter plots of each 10000 draws from several Dirichlet distributions for $k = 3$ can be found in Figure 1. Realizations of the Dirichlet distribution can be obtained from Gamma distributed random variables, see Gentle (1998), p. 111. For $\delta = (1, 1, 1)$, the points (x_1, x_2) are uniformly scattered on E_2. This corresponds to a situation without prior information. For the figures (b) - (d), we define $(0.28, 0.43, 0.29)$ to be the investigator's guess. In (b), we use $d = 0.5$ and δ_i as described above. It seems that there are more realizations close to the boundaries $x_1 = 0$, $x_2 = 0$, and $x_1 + x_2 = 1$ than realizations close to $(0.28, 0.43)$. Thus, $d = 0.5$ seems inappropriate. In (c), we have $d = 10$, and the draws form a point cloud around $(0.28, 0.43)$. The extent of this point cloud is larger than the extent of the point cloud in (d) where $d = 25$. That is, situation (d) corresponds to a larger certainty concerning the guess for the unknown true proportions.

2.4 Posterior modes for the diagonal model

As described in Dempster, Laird, Rubin (1977) for general missing data situations, the EM algorithm can be applied to generate a sequence $\pi^{(t)}$ that converges to the posterior mode, i.e, the mode of the observed data posterior density $f_{\Pi \mid \mathbf{A}}(\cdot \mid \mathbf{a})$. In particular, we have

$$\log f_{\Pi \mid \mathbf{X},\mathbf{A}}(\pi \mid \mathbf{x}, \mathbf{a}) = \log f_{\mathbf{A} \mid \Pi}(\mathbf{a} \mid \pi) + \log f_{\mathbf{X} \mid \mathbf{A},\Pi}(\mathbf{x} \mid \mathbf{a}, \pi) + \log f_{\Pi}(\pi) + \textit{constant}. \tag{5}$$

Let $\pi^{(t)}$ be available from iteration t. Computing the expectation with respect to the distribution given by $f_{\mathbf{X} \mid \mathbf{A},\Pi}(\cdot \mid \mathbf{a}, \pi^{(t)})$ yields

$$Q(\pi \mid \pi^{(t)}) + \log f_{\Pi}(\pi) = \log f_{\Pi \mid \mathbf{A}}(\pi \mid \mathbf{a}) + H(\pi \mid \pi^{(t)}) + \textit{constant},$$

where

$$\begin{aligned} Q(\pi \mid \pi^{(t)}) &= \int \log f_{\mathbf{X},\mathbf{A} \mid \Pi}(\mathbf{x}, \mathbf{a} \mid \pi) \cdot f_{\mathbf{X} \mid \mathbf{A},\Pi}(\mathbf{x} \mid \mathbf{a}, \pi^{(t)}) \, \partial \mathbf{x} \\ H(\pi \mid \pi^{(t)}) &= \int \log f_{\mathbf{X} \mid \mathbf{A},\Pi}(\mathbf{x} \mid \mathbf{a}, \pi) \cdot f_{\mathbf{X} \mid \mathbf{A},\Pi}(\mathbf{x} \mid \mathbf{a}, \pi^{(t)}) \, \partial \mathbf{x}. \end{aligned}$$

Notice that $Q(\pi \mid \pi^{(t)})$ equals the conditional expectation of the complete data log-likelihood given the observed data and $\pi^{(t)}$. In the E step of iteration $t + 1$, the function $Q^*(\cdot \mid \pi^{(t)})$ with $Q^*(\pi \mid \pi^{(t)}) = Q(\pi \mid \pi^{(t)}) + \log f_{\Pi}(\pi)$ is calculated. In the subsequent M step, we find $\pi^{(t+1)}$, which is the maximum of $Q^*(\cdot \mid \pi^{(t)})$. This $\pi^{(t+1)}$ increases the value of the observed data posterior density, i.e., it fulfills $f_{\Pi \mid \mathbf{A}}(\pi^{(t+1)} \mid \mathbf{a}) \geq f_{\Pi \mid \mathbf{A}}(\pi^{(t)} \mid \mathbf{a})$. A possible starting value is $(1/k, ..., 1/k)^T$. A detailed description of

this general scheme can be also found in Schafer (2000), Chapter 3.2.

Adopting this general scheme to a survey according to the diagonal model, we have for $\pi = (\pi_1, ..., \pi_{k-1})$, $\pi_k = 1 - \pi_1 - ... - \pi_{k-1}$ (apart from a constant)

$$Q(\pi \,|\, \pi^{(t)}) = \sum_{i=1}^{k} \hat{m}_i^{(t)} \cdot \log \pi_i \text{ and } Q^*(\pi \,|\, \pi^{(t)}) = \sum_{i=1}^{k} \left(\delta_i - 1 + \hat{m}_i^{(t)}\right) \cdot \log \pi_i \tag{6}$$

with $\hat{m}_i^{(t)} = \sum_{j=1}^{k} n_j \cdot \pi_i^{(t)} \cdot C(j,i) / f_{A_1 \,|\, \Pi}(j \,|\, \pi^{(t)})$, where n_j is the number of respondents in the sample giving answer j. We remark that $\hat{m}_i^{(t)}$ is equal to the sum of the i-th column of the $k \times k$ matrix

$$C \,.^* \left[\left[\tilde{n}^T ./\, \lambda(\pi^{(t)})\right] \cdot (\pi_1^{(t)}, ..., \pi_k^{(t)})\right].$$

Here, the signs $.^*$ and $./$ stand for componentwise multiplication and division, respectively, and

$$\tilde{n} = (n_1, ..., n_k) \text{ and } \lambda(\pi^{(t)}) = (f_{A_1 \,|\, \Pi}(1 \,|\, \pi^{(t)}), ..., f_{A_1 \,|\, \Pi}(k \,|\, \pi^{(t)}))^T$$

hold. The maximum of the function $Q^*(\cdot \,|\, \pi^{(t)})$ is given by $\pi_i^{(t+1)} = (\delta_i - 1 + \hat{m}_i^{(t)}) / (n - k + \delta_1 + ... + \delta_k)$.

2.5 Parameter simulation and multiple imputation for the diagonal model

Beyond finding the posterior mode, we can draw realizations from $f_{\Pi \,|\, \mathbf{A}}(\cdot \,|\, \mathbf{a})$ and $f_{\mathbf{X} \,|\, \mathbf{A}}(\cdot \,|\, \mathbf{a})$. To draw from these distributions, the data augmentation (DA) algorithm by Tanner and Wong (1987) is most convenient. The DA algorithm generates realizations $(\mathbf{x}^{(t)}, \pi^{(t)})$ of a Markov chain, short MC, $(\mathbf{X}^{(t)}, \Pi^{(t)})$ for $t \in \mathbb{N}$. This Markov chain converges in distribution to $f_{\mathbf{X}, \Pi \,|\, \mathbf{A}}(\cdot, \cdot \,|\, \mathbf{a})$. Thus, by integration, the sequence $(\Pi^{(t)})$ has the asymptotic distribution $f_{\Pi \,|\, \mathbf{A}}(\cdot \,|\, \mathbf{a})$.

Let us consider the diagonal model survey design and a prior distribution given by $f_\Pi \sim Di(\delta)$ with fixed and known parameter δ. The DA algorithm proceeds as follows. Let $\pi^{(t-1)} = (\pi_1^{(t-1)}, ..., \pi_{k-1}^{(t-1)})^T$ and $\pi_k^{(t-1)} = 1 - \sum_{i=1}^{k-1} \pi_i^{(t-1)}$ be available from the preceding iteration $t-1$. The next iteration t consists of the imputation step (I step) and the posterior step (P step):

I step: Drawing from $f_{\mathbf{X} \,|\, \mathbf{A}, \Pi}(\cdot \,|\, \mathbf{a}, \pi^{(t-1)})$ can be done by generating independent realizations x_j $(j = 1, ..., n)$, where x_j must be drawn according to the density $f_{X_j \,|\, A_j, \Pi}(\cdot \,|\, a_j, \pi^{(t-1)})$. However, we only need the frequency of value i $(i = 1, ..., k)$ among the values x_j for the subsequent P step. For this reason, let $m^{(t)}(i,j)$ describe the in iteration t simulated number of persons who have X-value j among the persons in the sample who give answer i. We draw

$$(m^{(t)}(i,1), ..., m^{(t)}(i,k)) \sim Multinomial(n_i, \gamma_i^{(t)}).$$

The vector $\gamma_i^{(t)}$ contains the cell probabilities and is defined to be the i-th row of the $k \times k$ matrix

$$C \,.^* \left[\left[(1, \cdots, 1)^T ./\, \lambda(\pi^{(t-1)})\right] \cdot \left(\pi_1^{(t-1)}, ..., \pi_k^{(t-1)}\right)\right],$$

where

$$\lambda(\pi^{(t-1)}) = (f_{A_1 \,|\, \Pi}(1 \,|\, \pi^{(t-1)}), ..., f_{A_1 \,|\, \Pi}(k \,|\, \pi^{(t-1)}))^T.$$

Set $m_j^{(t)} = \sum_{i=1}^{k} m^{(t)}(i,j)$, which is the simulated number of persons having $X = j$ in iteration t.

P step: We simulate realizations $(\pi_1^{(t)}, ..., \pi_{k-1}^{(t)})^T$ from $f_{\Pi \,|\, \mathbf{X}, \mathbf{A}}(\cdot \,|\, \mathbf{x}^{(t)}, \mathbf{a})$, which is the density corresponding to the $Di(m_1^{(t)} + \delta_1, ..., m_k^{(t)} + \delta_k)$ distribution.

To determine a starting value $\pi^{(0)}$, one option is to draw an outcome from the prior density. Alternatively, $\pi_i^{(0)} = 1/k$ can be used.

If t is large, then $\pi^{(t)}$ can be treated as realization from $f_{\Pi \mid \mathbf{A}}(\cdot \mid \mathbf{a})$. Assume we have generated one Markov chain of length $L_2 \in \mathbb{N}$. We delete $m^{(t)} = (m_1^{(t)}, ..., m_k^{(t)})$ and $\pi^{(t)}$ from the burn-in period $t = 1, ..., L_3 - 1$ and save them for $t = L_3, ..., L_2$. Thus, there remains a sequence $(m^{(t)}, \pi^{(t)})$ of length $L_2 - L_3 + 1$. We have two ways to extract information from this sequence. The first way is referred to as parameter simulation (see e.g., Schafer (2000), p. 89) and considers the $\pi^{(t)}$. The mean and the empirical standard deviation of the $\pi_i^{(t)}$ can be used as an estimate for the true proportion π_i and as a measure for the estimation precision, respectively. The empirical $\alpha/2$ and $1 - \alpha/2$ quantiles can be used as lower and upper bounds of a $1 - \alpha$ confidence interval (CI) for π_i. A slightly different strategy is to view the $m^{(t)} = (m_1^{(t)}, ..., m_k^{(t)})$, $t = L_3, ..., L_2$ as multiple imputations for the unobserved variables $(\sum_{j=1}^n 1_{\{X_j=1\}}, ..., \sum_{j=1}^n 1_{\{X_j=k\}})$. Each imputation $m^{(t)}$ results in an estimate $m^{(t)}/n$ for the unknown vector $(\pi_1, ..., \pi_k)$. That is, we obtain $L_2 - L_3 + 1$ estimates for π_i, which can be combined to a single estimate by using the mean. The empirical standard deviation and the $\alpha/2$ and $1 - \alpha/2$ quantiles of the $L_2 - L_3 + 1$ estimates for π_i are suitable to measure the estimation precision and to construct a $1 - \alpha$ CI for π_i, respectively.

In the last paragraph, we analyzed realizations of a single Markov chain, that is, we have considered a dependent sample. Of course, an alternative approach is given by simulating $L_1 \in \mathbb{N}$ independent Markov chains and saving only the values from the last iteration of each chain. It follows that we have L_1 independent draws from $f_{\Pi \mid \mathbf{A}}(\cdot \mid \mathbf{a})$ and L_1 independent multiple imputations, which can be evaluated analogously to the dependent quantities of the last paragraph.

2.6 Diagonal model estimates motivated by the Rao-Blackwell Theorem

Parameter simulation with a single Markov chain results in an estimate $s = (L_2 - L_3 + 1)^{-1} \sum_{t=L_3}^{L_2} \pi^{(t)}$ for the observed data posterior mean $\mathbb{E}(\Pi \mid \mathbf{A} = \mathbf{a})$. This s is used to estimate the true proportions π_i. In the context of a general missing data situation, Schafer (2000), section 4.2.3, discusses an estimate based on the idea of the Rao-Blackwell theorem. Applied to our situation of diagonal model interviews, this estimate is given by

$$\tilde{s} = (L_2 - L_3 + 1)^{-1} \sum_{t=L_3}^{L_2} \mathbb{E}(\Pi \mid \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a}). \tag{7}$$

The distribution of Π given $\mathbf{a}$ and $\mathbf{x}^{(t)}$ appears in the P step of DA. Thus, we have

$$\mathbb{E}(\Pi \mid \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a}) = \frac{(m_1^{(t)} + \delta_1, ..., m_{k-1}^{(t)} + \delta_{k-1})^T}{(n + \delta_1 + ... + \delta_k)},$$

where $m_j^{(t)}$ is again the simulated count of persons having $X = j$ in iteration t. The components of $\tilde{s}$ provide estimates for the unknown π_i. Analogously to Section 2.5, the empirical standard deviation and quantiles of $\mathbb{E}(\Pi_i \mid \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a})$, $t = L_3, ..., L_2$ can be used to measure precision and to construct confidence intervals for π_i, respectively. Obviously, instead of analyzing a single dependent Markov chain, it is also possible to generate $L_2 - L_3 + 1$ independent Markov chains of length L_3, where only the last iteration of each chain is saved for the estimation.

2.7 Simulation study

The simulations in this section are conducted to assess the benefit and the quality of the estimation procedures given in Sections 2.4-2.6. We run all simulations with MATLAB. We choose the true

parameter $\pi = (0.3, 0.4, 0.3)$, which may represent the proportions of persons in certain income classes, and $(\mathbb{P}(W = 1), ..., \mathbb{P}(W = 3)) = (2/3, 1/6, 1/6)$, where W represents a nonsensitive auxiliary characteristic. Groenitz (2012) presents ways to construct a W for a given distribution and shows that the above distribution of W provides a medium degree of privacy protection. The design matrix is then given by

$$C = \begin{pmatrix} c_1 & c_2 & c_3 \\ c_2 & c_3 & c_1 \\ c_3 & c_1 & c_2 \end{pmatrix} = \begin{pmatrix} 2/3 & 1/6 & 1/6 \\ 1/6 & 1/6 & 2/3 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}.$$

We consider sample sizes $n \in \{100, 300\}$, the confidence level $1 - \alpha = 0.95$, and three Dirichlet(δ) prior distributions whose scatter plots appear in Figure 1. In particular, we study $\delta^{(1)} = (1, 1, 1)$, $\delta^{(2)} = (2.8, 4.3, 2.9)$, and $\delta^{(3)} = (7, 10.75, 7.25)$. The first is the noninformative prior, the second and third are informative priors. Both informative priors correspond to an investigator's guess $(\hat{\pi}_1^{(p)}, \hat{\pi}_2^{(p)}, \hat{\pi}_3^{(p)}) = (0.28, 0.43, 0.29)$ with $d^{(2)} = 10$ and $d^{(3)} = 25$, i.e, prior three indicates a larger certainty about the guess than prior two. In other words, prior three is more informative than prior two.

The simulation procedure is as follows. We draw 1000 samples of size n. In each sample, we calculate the posterior mode and apply parameter simulation (PS), multiple imputation (MI), and Rao-Blackwellization (RB) according to Sections 2.4-2.6 to calculate estimates and confidence intervals for the true π_i. The estimation quality is evaluated by the average estimate for π_i, the empirical MSE of the estimates for π_i, the empirical width, and the empirical coverage probability (CP) of the confidence intervals for π_i. The simulation results for PS, MI, and RB based on a single dependent Markov chain of length 1000 with burn-in period $t = 1, ..., 500$ are reported in Table 1 in the appendix.
For each of the methods PS, MI, and RB and for both considered sample sizes, we recognize that the average estimates are always close to the true proportions. The simulated MSEs and the widths of the CIs decrease as the prior becomes more informative. Additionally, we observe the tendency that the more informative the prior, the higher the coverage probabilities.

Reduced MSEs and shorter CIs are the effects caused by increasing the sample size.

Comparing the MSEs of the estimates for π_i, we find that RB and PS have nearly identical values, whereas MI shows the largest MSEs. The confidence widths of RB are smaller than the widths of MI, and PS delivers the widest CIs. However, RB has the lowest and PS has clearly the highest CPs. Due to the MSE results and the highest CPs, we evaluate PS to be the best method.

For comparison, we calculate the maximum likelihood estimates (MLEs) for each 1000 samples of size $n = 300$ and $n = 100$ and compute Bootstrap CIs (without normality assumption) for the π_i for each sample from $B = 2000$ Bootstrap replications, see Groenitz (2012), Section 3.2 and 3.3. The average ML estimates (see Table 3 in the appendix) are close to the true proportions. Consider $n = 300$ first. For the uniform prior ($\delta^{(1)}$), the CI widths and CPs for PS are slightly smaller than for ML. The MSEs of PS and ML are close to each other. The reason is that the posterior variance is a consistent estimate for the large sample variance of the ML estimator (see e.g., Little and Rubin (2002), Section 9.2.4). Parameter simulation with the informative prior with $\delta^{(2)}$ reduces the MSEs provided by ML by up to approximately 20%, and the more informative prior with $\delta^{(3)}$ leads to a reduction by approximately 40%.
We next examine $n = 100$. We notice that PS with the noninformative prior has smaller MSEs than ML. Moreover, we point out that PS with $\delta^{(2)}$ and $\delta^{(3)}$ decreases the MSEs of ML by approximately 40% and 75%, respectively. The widths of the CIs for π_i decrease by approximately 15% for $\delta^{(2)}$ and 30% for $\delta^{(3)}$ by using PS instead of ML.
For both informative priors and both sample sizes, there is a tendency that the CPs of PS are larger than the CPs of ML and overachieve the 95% level.
The estimates generated by PS are posterior means. On average, these posterior means are close to

the posterior modes (see appendix, Table 4). The MSEs of the posterior means and modes are quite similar for $n = 300$. In the case $n = 100$, the posterior modes provide a bit higher MSEs. We remark that the posterior mode for the uniform prior equals the MLE, if both are calculated from the same sample. This explains that the average MLEs and posterior means as well as the corresponding MSEs in Tables 3 and 4 are close to each other.

We also have conducted simulations in which the Bayes estimates were computed with the help of independent Markov chains. In particular, for each of 1000 simulated samples, we have calculated the PS, MI, and RB estimates from 500 independent chains of length 501, where only the last iteration of each chain is saved for the estimation. The simulation results are provided in Table 2. We discover that the above statements regarding estimates based on a single MC remain valid for the estimation with independent chains.

In sum, we emphasize that the estimation accuracy can be significantly improved by using Bayesian methods when prior information is available.

3 Common approach for Bayes estimation in privacy-protecting survey designs

Studying the calculations to obtain posterior modes and estimates based on parameter simulation, multiple imputation, and Rao-Blackwellization in Section 2, we observe that the design matrix C is the only component of the diagonal model that influences these calculations. Let us now consider an arbitrary PP design for $X \in \{1, ..., k\}$ with k_A possible scrambled answers and S required samples (in the DM, k_A equals k and $S = 1$). For each sample, we then have one design matrix. In the sequel, we restrict to PP designs whose design matrices do not contain nuisance terms, i.e., unknown parameters. For such a design, the only model component that is needed to compute Bayes estimates is the set of design matrices. That is, all relevant model information is stored in the design matrices - it does not matter whether we consider a RR or NRR method, moreover, the concrete answer scheme is irrelevant. Hence, most PP models for categorical X can be handled by a common approach. This fact has not been addressed in existing papers about Bayesian inference in PP models.
In Subsection 3.1, we give the design matrices for some PP models. Subsequently, in Subsection 3.2, we develop a general framework for Bayes estimation in PP designs for categorical X. Here, we generalize the calculations from Section 2 in order to cover many PP designs including certain multi-stage and multi-sample techniques.

3.1 Other privacy-protecting designs for categorical sensitive variables

We consider PP designs (i.e., RR or NRR models) for categorical sensitive variables $X \in \{1, ..., k\}$ with k_A possible answers (coded with $1, ..., k_A$) and S required samples. The complete data, i.e., the union of missing and observed data, are given by the vectors $(X_{sj}, A_{sj})_{sj}$ where X_{sj} and A_{sj} denote the X-value and the scrambled answer of respondent j in sample s, respectively ($s = 1, ..., S; j = 1, ..., n_s$). We demand the following conditions:

(M1) The $n = n_1 + ... + n_S$ vectors (X_{sj}, A_{sj}) are independent. Further, for $s = 1, ..., S$, the n_s vectors $(X_{s1}, A_{s1}), ..., (X_{s,n_s}, A_{s,n_s})$ are identically distributed, and $X_{sj} \sim X$ for all indices s, j.

(M2) The $k_A \times k$ matrices of conditional probabilities $C_s = [C_s(i, j)]_{ij} = [\mathbb{P}(A_{s1} = i \,|\, X_{s1} = j)]_{ij}$ have known entries ($s = 1, ..., S$).

Assumption (M1) means that the design needs S independent simple random samples with replacement (SRSWR) where the distribution of the scrambled answer is allowed to alter in different samples. We call the matrices C_s "design matrices". We next provide some examples of PP survey techniques, for which (M1)-(M2) are satisfied. All PP designs considered in the sequel are assumed to be applied to a SRSWR (for $S = 1$) respectively $S \geq 1$ independent SRSWR.

The RR model by Warner (1965) considers $X \in \{1, 2\}$ and needs one SRSWR. Each respondent draws and answers one of the questions "Do you have $X = 1$?" and "Do you have $X = 2$?". The first question is drawn with known probability c. The possible answers are "yes" and "no" (coded with 1 and 2). Then, the rows of $C = C_1$ are known and given by $(c, 1 - c)$ and $(1 - c, c)$.

The RR design by Abul-Ela, Greenberg, Horvitz (1967) is applicable to $X \in \{1, ..., k\}$, $k \geq 2$, and needs $S = k - 1$ independent samples (each sample is a SRSWR). The interviewees select and answer one of the k questions "Do you have $X = j$?" $(j = 1, ..., k)$. The probability c_{sj} $(s = 1, ..., k - 1; j = 1, ..., k)$ that question j is selected in sample s is determined by the RD and is known. Coding "yes" and "no" by 1 and 2 results in the $2 \times k$ matrices C_s having the j-th column equal to $(c_{sj},\ 1-c_{sj})^T$ $(s = 1, ..., k-1)$.

The unrelated question model (UQM) - see Horvitz et al. (1967) and Greenberg et al. (1969) - is constructed for a sensitive $X \in \{1, 2\}$. According to the result of a random experiment, each interviewee answers either "Do you have $X = 1$?" or "Do you have $Y = 1$?" where $Y \in \{1, 2\}$ is an unrelated nonsensitive variable. Let c be the known probability that the first question is selected, and assume $\phi = \mathbb{P}(Y = 1)$ to be known. Then, the UQM requires a single SRSWR, and we have $C = C_1$ with rows $(c + (1 - c)\phi,\ (1 - c)\phi)$ and $((1 - c)(1 - \phi),\ (1 - c)(1 - \phi) + c)$. If the distribution of Y is unknown, the UQM needs two independent SRSWR. In this case, we can define the new variable

$$\tilde{X} \in \{1, ..., 4\} \tag{8}$$

that attains the values 1, 2, 3, 4 if (X, Y) attains $(1, 1)$, $(1, 2)$, $(2, 1)$, $(2, 2)$, respectively. This $\tilde{X}$ plays the role of X from (M1) and (M2). Let c_{s1} be the known probability that question 1 is selected in sample s. It follows that C_s has the rows $(1,\ c_{s1},\ 1 - c_{s1},\ 0)$ and $(0,\ 1 - c_{s1},\ c_{s1},\ 1)$.

Omitting details, we also can fulfill (M1)-(M2) for the RR methods for $X \in \{1, ..., k\}$ $(k \geq 2)$ suggested by Eriksson (1973), and Liu et al. (1975).
The two-stage RR design by Mangat and Singh (1990) considers $X \in \{1, 2\}$. In the first stage, each respondent conducts a random experiment that decides whether the question "Do you have $X = 1$?" must be answered or whether the respondent has to go to stage two. In stage two, another random experiment must be accomplished by the interviewee. According to its outcome, either the question "Do you have $X = 1$?" or "Do you have $X = 2$?" must be answered. This model needs one SRSWR, and $C = C_1$ has the known rows $(T + (1 - T)c, (1 - T)(1 - c))$ and $((1 - T)(1 - c), T + (1 - T)c)$, where T is the probability that the experiment in stage one decides that the question must be answered and c is the probability of drawing the first question in stage two.
Omitting certain details again, for the RR model by Mangat (1994), (M1)-(M2) are fulfilled, where $k_A = 2$, $S = 1$, and $C = C_1$ with rows $(1, 1 - c)$ and $(0, c)$ for a $c \in (0, 1)$.

Quatember (2009) presents a standardized RR model for $X \in \{1, 2\}$ and explains that 16 survey designs are special cases of his model. In this standardized design, each interviewee draws randomly one of the five instructions:

1:	Answer "Do you have $X = 1$?"	2:	Answer "Do you have $X = 2$?"		
3:	Answer "Do you have $Y = 1$?"	4:	Say "yes"	5:	Say "no"

Here, $Y \in \{1, 2\}$ is a nonsensitive characteristic. Let us consider a single SRSWR, set $\phi = \mathbb{P}(Y = 1)$, and define c_i to be the probability that instruction i is drawn. Coding answers "yes" and

"no" with 1 and 2 yields the 2×2 design matrix with rows $(c_1 + c_3\phi + c_4,\ c_2 + c_3\phi + c_4)$ and $(c_2 + c_3(1-\phi) + c_5,\ c_1 + c_3(1-\phi) + c_5)$ and (M1)-(M2) are fulfilled.

The properties (M1)-(M2) are also satisfied for the following NRR models: the hidden sensitivity model by Tian et al. (2007), the crosswise and triangular model by Yu et al. (2008), and the multicategory model by Tang et al. (2009). For instance, Tang et al. (2009) consider $X \in \{1, ..., k\}$, $k \geq 2$. The respondent's answer depends on the value of X and on the value of a nonsensitive auxiliary variable $W \in \{1, ..., k\}$, which is independent of X and possesses a known distribution (e.g., W may describe the period of the birthday). If $X = 1$, an answer equal to the value of W is required. For $X = i$, the response i ($i = 2, ..., k$) must be given. The design needs a single SRSWR. The first column of the $k \times k$ matrix $C = C_1$ equals $(P(W = 1), ..., \mathbb{P}(W = k))^T$, and column i ($i = 2, ..., k$) is a vector having entry i equal to 1 and all other entries equal to 0.

We finish this section with a model that violates (M2): the two-trial UQM by Horvitz et al. (1967) is for $X \in \{1, 2\}$ and needs $S = 2$ independent SRSWR. Each respondent selects one of the questions "Do you have $X = 1$?" or "Do you have $Y = 1$?" with the help of a random experiment (Y is again an unrelated variable). Subsequently, the selection is repeated. The possible answers are 1=("yes", "yes"), 2=("yes", "no"), 3=("no", "yes"), 4=("no", "no"). The distribution of Y is unknown, and independence between X and Y is assumed. Then, we have

$$C_s = \begin{pmatrix} c_{s1}^2 + 2c_{s1}c_{s2}\phi + c_{s2}^2\phi & c_{s2}^2\phi \\ c_{s1}c_{s2}(1-\phi) & c_{s1}c_{s2}\phi \\ c_{s1}c_{s2}(1-\phi) & c_{s1}c_{s2}\phi \\ c_{s2}^2(1-\phi) & c_{s1}^2 + 2c_{s1}c_{s2}(1-\phi) + c_{s2}^2(1-\phi) \end{pmatrix}$$

with $s \in \{1, 2\}$, where $\phi = \mathbb{P}(Y = 1)$, c_{s1} is the known probability that question 1 is selected in sample s, and $c_{s2} = 1 - c_{s1}$. Since ϕ is unknown, (M2) does not hold. A possible remedy is to abandon the independence assumption for X and Y and to consider $\tilde{X}$ from (8) again. $\tilde{X}$ plays the role of X in (M1)-(M2) with

$$C_s = \begin{pmatrix} 1 & c_{s1}^2 & c_{s2}^2 & 0 \\ 0 & c_{s1}c_{s2} & c_{s1}c_{s2} & 0 \\ 0 & c_{s1}c_{s2} & c_{s1}c_{s2} & 0 \\ 0 & c_{s2}^2 & c_{s1}^2 & 1 \end{pmatrix},$$

where $s \in \{1, 2\}$. This version of the two-trial UQM, which can be found in Bourke and Moran (1988), Section 2, satisfies (M1)-(M2).

3.2 Bayes estimation in PP models

The calculations from Section 2 can be generalized to arbitrary randomized response and nonrandomized response survey techniques with (M1)-(M2). For such a model, the missing data $\mathbf{X}$ and observed data $\mathbf{A}$ are given by $(X_{sj})_{sj}$ and $(A_{sj})_{sj}$, respectively ($s = 1, ..., S$; $j = 1, ..., n_s$). Set for $x_{sj} \in \{1, ..., k\}$ and $a_{sj} = \{1, ..., k_A\}$

$$f_{\mathbf{X},\mathbf{A}\,|\,\Pi}(\mathbf{x}, \mathbf{a}\,|\,\pi) = \prod_{s=1}^{S} \prod_{j=1}^{n_s} C_s(a_{sj}, x_{sj}) \cdot \pi_{x_{sj}},$$

where the C_s are the design matrices of the PP model and $\mathbf{x} = (x_{sj})_{sj}$, $\mathbf{a} = (a_{sj})_{sj}$. Accordingly, we have

$$f_{\mathbf{X}\,|\,\mathbf{A},\Pi}(\mathbf{x}\,|\,\mathbf{a}, \pi) = \prod_{s=1}^{S} \prod_{j=1}^{n_s} \frac{C_s(a_{sj}, x_{sj}) \cdot \pi_{x_{sj}}}{f_{A_{sj}\,|\,\Pi}(a_{sj}\,|\,\pi)},$$

where $f_{A_{sj}\,|\,\Pi}(\alpha\,|\,\pi)$ is the entry number $\alpha \in \{1, ..., k_A\}$ of vector $C_s \cdot (\pi_1, ..., \pi_k)^T$. As in Section 2, we focus on Dirichlet prior distributions.

To calculate the posterior mode in a PP design with (M1)-(M2), (6) becomes

$$Q(\pi \,|\, \pi^{(t)}) = \sum_{s=1}^{S} \sum_{i=1}^{k} \hat{m}_{si}^{(t)} \cdot \log \pi_i \text{ and } Q^*(\pi \,|\, \pi^{(t)}) = \sum_{i=1}^{k} \left(\delta_i - 1 + \sum_{s=1}^{S} \hat{m}_{si}^{(t)} \right) \cdot \log \pi_i$$

with $\hat{m}_{si}^{(t)} = \sum_{j=1}^{k_A} n_{sj} \cdot \pi_i^{(t)} \cdot C_s(j,i) / f_{A_{s1} \,|\, \Pi}(j \,|\, \pi^{(t)})$, where n_{sj} is the number of respondents in sample s giving answer j. The term $\hat{m}_{si}^{(t)}$ is equal to the sum of the i-th column of the $k_A \times k$ matrix

$$C_s \,.^* \left[\left[\tilde{n}_s^T ./ \, \lambda_s(\pi^{(t)}) \right] \cdot (\pi_1^{(t)}, ..., \pi_k^{(t)}) \right] \text{ with}$$

$$\tilde{n}_s = (n_{s1}, ..., n_{sk_A}) \text{ and } \lambda_s(\pi^{(t)}) = (f_{A_{s1} \,|\, \Pi}(1 \,|\, \pi^{(t)}), ..., f_{A_{s1} \,|\, \Pi}(k_A \,|\, \pi^{(t)}))^T.$$

Maximization of $Q^*(\cdot \,|\, \pi^{(t)})$ results in $\pi_i^{(t+1)} = (\delta_i - 1 + \sum_{s=1}^{S} \hat{m}_{si}^{(t)}) / (n - k + \delta_1 + ... + \delta_k)$.

To conduct parameter simulation and to obtain multiple imputations, data augmentation for a general privacy-protecting survey design proceeds as follows:
I step: It suffices to simulate the number of sample units with $X = j$. Let $m_s^{(t)}(i,j)$ be the in iteration t simulated number of persons who have X-value j among the persons who give answer i in sample s. Draw

$$(m_s^{(t)}(i,1), ..., m_s^{(t)}(i,k)) \sim Multinomial(n_{si}, \gamma_{s,i}^{(t)}).$$

The vector $\gamma_{s,i}^{(t)}$ contains the cell probabilities and is defined to be the i-th row of the $k_A \times k$ matrix

$$C_s \,.^* \left[\left[(1, \cdots, 1)^T ./ \, \lambda_s(\pi^{(t-1)}) \right] \cdot \left(\pi_1^{(t-1)}, ..., \pi_k^{(t-1)} \right) \right],$$

where

$$\lambda_s(\pi^{(t-1)}) = (f_{A_{s1} \,|\, \Pi}(1 \,|\, \pi^{(t-1)}), ..., f_{A_{s1} \,|\, \Pi}(k_A \,|\, \pi^{(t-1)}))^T.$$

Obviously, the cell probabilities depend (apart from the parameters of the preceding iteration) only on the design matrices. The desired number of persons having $X = j$ in iteration t is then $m_j^{(t)} = \sum_{s=1}^{S} \sum_{i=1}^{k_A} m_s^{(t)}(i,j)$.

P step: Draw a new parameter $(\pi_1^{(t)}, ..., \pi_{k-1}^{(t)})^T$ from $f_{\Pi \,|\, \mathbf{X}, \mathbf{A}}(\cdot \,|\, \mathbf{x}^{(t)}, \mathbf{a})$, a density corresponding to the $Di(m_1^{(t)} + \delta_1, ..., m_k^{(t)} + \delta_k)$ distribution.

Rao-Blackwellized estimates for a general PP design can be obtained analogously to Subsection 2.6 by averaging conditional expectations. In particular, the estimate is given by

$$\tilde{s} = (L_2 - L_3 + 1)^{-1} \sum_{t=L_3}^{L_2} \mathbb{E}(\Pi \,|\, \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a}).$$

with (compare P step of data augmentation above)

$$\mathbb{E}(\Pi \,|\, \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a}) = \frac{(m_1^{(t)} + \delta_1, ..., m_{k-1}^{(t)} + \delta_{k-1})^T}{(n + \delta_1 + ... + \delta_k)},$$

where $m_j^{(t)}$ is again the simulated count of persons having $X = j$ in iteration t.

4 Summary

Survey concepts that protect the respondents' privacy are important to obtain reliable data on sensitive characteristics. To exploit prior information on the distribution of the sensitive variable, the application of Bayesian methods is appealing. In this paper, we have developed a Bayesian extension of the privacy-protecting, nonrandomized diagonal model survey technique by Groenitz (2012). We illustrated in simulations that precision can be significantly improved by incorporating available prior information into the estimation. In the second part of this paper, we found that for any privacy-protecting survey design dealing with categorical sensitive characteristics, all relevant model information is stored in the design matrices. For this reason, we were able to present the Bayes inference for privacy-protecting models in a general framework that covers a lot of randomized and nonrandomized response methods.

References

[1] Abul-Ela, A.A., Greenberg, B.G., Horvitz, D.G.: A Multi-Proportions Randomized Response Model. Journal of the American Statistical Association 62, 990-1008 (1967)

[2] Barabesi, L., Marcheselli, M.: Bayesian estimation of proportion and sensitivity level in randomized response procedures. Metrika 72, 75-88 (2010)

[3] Bourke, P.D., Moran, M.A.: Estimating Proportions From Randomized Response Data Using the EM Algorithm. Journal of the American Statistical Association 83, 964-968 (1988)

[4] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39, 1-38 (1977)

[5] Eriksson, S.A.: A New Model for Randomized Response. International Statistical Review 41, 101-113 (1973)

[6] Gentle, J.E.: Random Number Generation and Monte Carlo Methods. Springer (1998)

[7] Greenberg, B.G., Abul-Ela, A.A., Simmons, W.R., Horvitz, D.G.: The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association 64, 520-539 (1969)

[8] Groenitz, H.: A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables. Metrika, DOI: 10.1007/s00184-012-0406-8 (2012).

[9] Horvitz, D.G., Shah, B.V., Simmons, W.R.: The Unrelated Question Randomized Response Model. Proceedings of the Social Statistics Section, American Statistical Association, 65-72 (1967)

[10] Hussain, Z., Cheema, S.A., Zafar, S.: Extension of Mangat Randomized Response Model. International Journal of Business and Social Science 2, 261-266 (2011)

[11] Kim, J.M., Tebbs, J.M., An, S.W.: Extensions of Mangat's randomized-response model. Journal of Statistical Planning and Inference 136, 1554-1567 (2006)

[12] Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley (2002)

[13] Liu, P.T., Chow, L.P., Mosley, W.H.: Use of the Randomized Response Technique With a New Randomizing Device. Journal of the American Statistical Association 70, 329-332 (1975)

[14] Mangat, N.S.: An Improved Randomized Response Strategy. Journal of the Royal Statistical Society B 56, 93-95 (1994)

[15] Mangat, N.S., Singh, R.: An Alternative Randomized Response Procedure. Biometrika 77, 439-442 (1990)

[16] Migon, H.S., Tachibana, V.M.: Bayesian approximations in randomized response model. Computational Statistics & Data Analysis 24, 401-409 (1997)

[17] O'Hagan, A.: Bayes Linear Estimators for Randomized Response Models. Journal of the American Statistical Association 82, 207-214 (1987)

[18] Quatember, A.: A standardization of randomized response strategies. Statistics Canada, Survey Methodology 35, 143-152 (2009)

[19] Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC (2000)

[20] Tan, M.T., Tian, G.L., Tang, M.L.: Sample Surveys with Sensitive Questions: A Nonrandomized Response Approach. The American Statistician 63, 9-16 (2009)

[21] Tanner, M.A., Wong, W.H.: The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association 82, 528-540 (1987)

[22] Tang, M.L., Tian G.L., Tang, N.S., Liu, Z.: A new non-randomized multi-category response model for surveys with a single sensitive question: Design and analysis. Journal of the Korean Statistical Society 38, 339-349 (2009)

[23] Tian, G.L., Yu, J.W., Tang, M.L., Geng, Z.: A new non-randomized model for analysing sensitive questions with binary outcomes. Statistics in Medicine 26, 4238-4252 (2007)

[24] Tian, G.L., Yuen, K.C., Tang, M.L., Tan, M.T.: Bayesian non-randomized response models for surveys with sensitive questions. Statistics and its interface 2, 13-25 (2009)

[25] Unnikrishnan, N.K., Kunte, S.: Bayesian analysis for randomized response models. The Indian Journal of Statistics 61, Series B, 422-432 (1999)

[26] Warner, S.L.: Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association 60, 63-69 (1965)

[27] Winkler, R.L., Franklin, L.A.: Warner's Randomized Response Model: A Bayesian Approach. Journal of the American Statistical Association 74, 207-214 (1979)

[28] Yu, J.W., Tian, G.L., Tang, M.L.: Two new models for survey sampling with sensitive characteristic: design and analysis. Metrika 67, 251-263 (2008)

A Appendix: Simulation Outputs

This appendix contains the simulation results described in Section 2.7.

		$n = 300$ - estimation based on a single Markov chain											
		Parameter simulation				Multiple imputation				Rao-Blackwellization			
		av.est.	MSE	width	CP	av.est.	MSE	width	CP	av.est.	MSE	width	CP
$\delta^{(1)}$	π_1	0.2986	0.0027	0.2071	0.9540	0.2982	0.0028	0.1827	0.9300	0.2986	0.0027	0.1809	0.9260
	π_2	0.3972	0.0029	0.2140	0.9410	0.3979	0.0030	0.1873	0.9140	0.3972	0.0029	0.1854	0.9070
	π_3	0.3043	0.0028	0.2075	0.9470	0.3039	0.0028	0.1830	0.9180	0.3042	0.0028	0.1812	0.9140
$\delta^{(2)}$	π_1	0.2969	0.0022	0.1970	0.9610	0.2974	0.0023	0.1760	0.9250	0.2969	0.0022	0.1704	0.9190
	π_2	0.4070	0.0025	0.2047	0.9610	0.4063	0.0027	0.1812	0.9240	0.4070	0.0025	0.1753	0.9180
	π_3	0.2961	0.0027	0.1971	0.9330	0.2963	0.0028	0.1758	0.9130	0.2961	0.0026	0.1701	0.9030
$\delta^{(3)}$	π_1	0.2942	0.0017	0.1799	0.9720	0.2954	0.0019	0.1645	0.9470	0.2942	0.0016	0.1518	0.9380
	π_2	0.4077	0.0018	0.1886	0.9740	0.4058	0.0021	0.1700	0.9450	0.4076	0.0018	0.1569	0.9420
	π_3	0.2981	0.0015	0.1803	0.9740	0.2988	0.0018	0.1644	0.9490	0.2981	0.0015	0.1518	0.9450
		$n = 100$ - estimation based on a single Markov chain											
		Parameter simulation				Multiple imputation				Rao-Blackwellization			
		av.est.	MSE	width	CP	av.est.	MSE	width	CP	av.est.	MSE	width	CP
$\delta^{(1)}$	π_1	0.2956	0.0078	0.3460	0.9470	0.2945	0.0083	0.3142	0.9140	0.2957	0.0078	0.3050	0.9030
	π_2	0.3985	0.0082	0.3625	0.9450	0.4004	0.0087	0.3249	0.9170	0.3985	0.0082	0.3154	0.9060
	π_3	0.3059	0.0078	0.3477	0.9480	0.3050	0.0082	0.3154	0.9220	0.3058	0.0077	0.3063	0.9100
$\delta^{(2)}$	π_1	0.2974	0.0046	0.3047	0.9670	0.2991	0.0056	0.2836	0.9340	0.2974	0.0046	0.2578	0.9290
	π_2	0.4090	0.0053	0.3189	0.9720	0.4070	0.0064	0.2923	0.9400	0.4091	0.0053	0.2657	0.9300
	π_3	0.2936	0.0046	0.3027	0.9700	0.2939	0.0056	0.2815	0.9450	0.2936	0.0046	0.2559	0.9350
$\delta^{(3)}$	π_1	0.2898	0.0023	0.2514	0.9900	0.2922	0.0035	0.2476	0.9680	0.2897	0.0023	0.1981	0.9570
	π_2	0.4151	0.0026	0.2673	0.9880	0.4115	0.0039	0.2595	0.9660	0.4152	0.0026	0.2076	0.9510
	π_3	0.2951	0.0021	0.2514	0.9960	0.2963	0.0033	0.2470	0.9740	0.2950	0.0021	0.1976	0.9580

Table 1: *Simulation results for PS, MI, RB based on a single Markov chain. The performance of the estimation strategies is assessed in terms of the average estimate for π_i, the simulated MSE of the estimates for π_i, the empirical width and coverage probability of the confidence intervals for π_i ($\alpha = 5\%$). The true proportions are given by* (0.3, 0.4, 0.3).

$n = 300$ - estimation based on independent Markov chains												
		Parameter simulation				Multiple imputation				Rao-Blackwellization			
		av.est.	MSE	width	CP	av.est.	MSE	width	CP	av.est.	MSE	width	CP
$\delta^{(1)}$	π_1	0.2971	0.0027	0.2080	0.9550	0.2968	0.0028	0.1837	0.9200	0.2971	0.0027	0.1819	0.9110
	π_2	0.4004	0.0032	0.2155	0.9490	0.4010	0.0032	0.1883	0.9140	0.4004	0.0032	0.1864	0.9110
	π_3	0.3024	0.0029	0.2083	0.9440	0.3022	0.0030	0.1838	0.9080	0.3025	0.0029	0.1819	0.9030
$\delta^{(2)}$	π_1	0.2963	0.0024	0.1983	0.9490	0.2969	0.0025	0.1767	0.9180	0.2963	0.0024	0.1710	0.9120
	π_2	0.4074	0.0026	0.2058	0.9510	0.4066	0.0028	0.1818	0.9140	0.4074	0.0026	0.1760	0.9090
	π_3	0.2963	0.0022	0.1982	0.9570	0.2965	0.0024	0.1770	0.9210	0.2963	0.0022	0.1713	0.9150
$\delta^{(3)}$	π_1	0.2944	0.0017	0.1814	0.9690	0.2955	0.0019	0.1653	0.9360	0.2943	0.0017	0.1526	0.9310
	π_2	0.4091	0.0018	0.1899	0.9740	0.4074	0.0021	0.1712	0.9370	0.4091	0.0018	0.1580	0.9280
	π_3	0.2965	0.0017	0.1811	0.9650	0.2971	0.0020	0.1653	0.9310	0.2965	0.0017	0.1526	0.9290
$n = 100$ - estimation based on independent Markov chains													
		Parameter simulation				Multiple imputation				Rao-Blackwellization			
		av.est.	MSE	width	CP	av.est.	MSE	width	CP	av.est.	MSE	width	CP
$\delta^{(1)}$	π_1	0.3000	0.0071	0.3504	0.9590	0.2991	0.0076	0.3186	0.9350	0.3001	0.0071	0.3094	0.9280
	π_2	0.3956	0.0082	0.3645	0.9520	0.3975	0.0087	0.3276	0.9300	0.3957	0.0083	0.3180	0.9140
	π_3	0.3043	0.0085	0.3499	0.9420	0.3034	0.0089	0.3171	0.9080	0.3043	0.0084	0.3078	0.8990
$\delta^{(2)}$	π_1	0.2911	0.0047	0.3040	0.9710	0.2921	0.0057	0.2823	0.9360	0.2910	0.0047	0.2566	0.9240
	π_2	0.4080	0.0049	0.3212	0.9780	0.4059	0.0059	0.2942	0.9520	0.4081	0.0049	0.2675	0.9430
	π_3	0.3009	0.0045	0.3058	0.9820	0.3021	0.0054	0.2841	0.9510	0.3010	0.0045	0.2583	0.9380
$\delta^{(3)}$	π_1	0.2880	0.0022	0.2513	0.9980	0.2900	0.0032	0.2478	0.9800	0.2880	0.0022	0.1982	0.9680
	π_2	0.4166	0.0028	0.2683	0.9910	0.4133	0.0041	0.2602	0.9700	0.4166	0.0028	0.2081	0.9600
	π_3	0.2954	0.0022	0.2528	0.9930	0.2968	0.0034	0.2486	0.9680	0.2954	0.0022	0.1988	0.9560

Table 2: *Simulation results for PS, MI, RB based on independent Markov chains. The performance of the estimation strategies is assessed in terms of the average estimate for π_i, the simulated MSE of the estimates for π_i, the empirical width and coverage probability of the confidence intervals for π_i ($\alpha = 5\%$). The true proportions are given by* (0.3, 0.4, 0.3).

ML estimation for $n = 300$				
	av.est.	MSE	width	coverage
π_1	0.2996	0.0028	0.2097	0.9580
π_2	0.4008	0.0030	0.2174	0.9510
π_3	0.2996	0.0028	0.2102	0.9470
ML estimation for $n = 100$				
π_1	0.3024	0.0084	0.3587	0.9580
π_2	0.4008	0.0094	0.3735	0.9510
π_3	0.2968	0.0083	0.3584	0.9500

Table 3: *This table contains the simulation results for the ML estimation based on 1000 samples. Average ML estimates for π_i, empirical MSEs for the ML estimates as well as empirical widths and coverage probabilities for Bootstrap CIs ($\alpha = 5\%$) reported. The true proportions are given by* (0.3, 0.4, 0.3).

		Posterior modes			
		$n = 300$		$n = 100$	
		av. est.	MSE	av. est.	MSE
$\delta^{(1)}$	π_1	0.2979	0.0027	0.2942	0.0086
	π_2	0.3982	0.0030	0.4013	0.0089
	π_3	0.3040	0.0028	0.3045	0.0084
$\delta^{(2)}$	π_1	0.2964	0.0022	0.2960	0.0052
	π_2	0.4080	0.0026	0.4126	0.0060
	π_3	0.2956	0.0027	0.2914	0.0052
$\delta^{(3)}$	π_1	0.2940	0.0017	0.2880	0.0026
	π_2	0.4085	0.0019	0.4186	0.0030
	π_3	0.2976	0.0016	0.2934	0.0024

Table 4: *Simulation results for the observed data posterior mode. The table reports the average posterior mode and the corresponding empirical MSE. The true proportions are given by* (0.3, 0.4, 0.3).