

# A new privacy-protecting survey design for multichotomous sensitive variables

Heiko Groenitz

Received: 13 April 2012 / Published online: 9 August 2012  
© Springer-Verlag 2012

**Abstract** In this paper, we propose the diagonal model (DM), a survey technique for multicategorical sensitive variables. The DM is a nonrandomized response method; that is, the DM avoids the use of any randomization device. Thus, both survey complexity and study costs are reduced. The DM does not require that at least one outcome of the sensitive variable is nonsensitive. Thus, the model can even be applied to characteristics like income which are sensitive as a whole. We describe the maximum likelihood estimation for the distribution of the sensitive variable and show that the EM algorithm is beneficial to calculate the estimates. Subsequently, we present asymptotic as well as bootstrap confidence intervals. Applying properties of circulant matrices, we show the connection between efficiency loss and the degree of privacy protection (DPP). Here, we prove that the efficiency loss has a lower bound that depends on the DPP. Moreover, for any desired DPP, we derive model parameters that ensure the largest possible efficiency.

**Keywords** Nonrandomized response method · Randomized response model · EM algorithm · Untruthful answers · Circulant matrix

## 1 Introduction

Sensitive variables often appear in surveys. For instance, an interviewer could ask: “How much do you earn?” or “Have you ever evaded taxes?”. However, when such sensitive questions are asked, some interviewees will refuse to respond or will give an untruthful answer. To estimate the distribution of sensitive variables, countless randomized response (RR) models have been developed since the paper by Warner (1965).

---

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

What the RR models have in common is that every respondent is supplied with a randomization device (RD). A RD is an instrument (e.g., a coin or a deck of cards) used by the interviewee to conduct a random experiment, where the experiment has— for a fixed respondent—at least two possible results. The outcome of the experiment influences the answer. It follows that RR techniques have a lack of “reproducibility” (the same respondent may give different answers if the survey is conducted again; see [Tan et al. \(2009\)](#), p. 15). A different approach was discussed by [Tian et al. \(2007\)](#), [Yu et al. \(2008\)](#), [Tan et al. \(2009\)](#), and [Tang et al. \(2009\)](#), see Sect. 2. These authors proposed some nonrandomized response (NRR) models. That is, their models do not require any RD, thus reducing both survey complexity and study costs.

The NRR models of the previously mentioned authors are not applicable to multicategorical variables like income when all categories are sensitive. To overcome this drawback, we have developed the nonrandomized diagonal model (DM), which is presented in Sect. 3.1. In Sect. 3.2, we derive the maximum likelihood estimator for the distribution of the sensitive variable. The computation of standard errors and confidence intervals is discussed in Sect. 3.3. Subsequently, in Sect. 3.4, we explain how the design matrices in the diagonal model are related to circulant matrices. In Sect. 3.5, we investigate the degree of privacy protection and the efficiency of the estimator depending on the model parameters. Here, under the assumption of a large sample, we will obtain two key results: first, there are optimal and nonoptimal choices for the model parameters, and second, a decrease in the degree of privacy protection (DPP) is the “price” for increasing efficiency.

In our simulation study (Sect. 4), we examine whether the asymptotic results from Sect. 3.5 hold for fixed, not “too large” sample sizes.

## 2 Nonrandomized response models

[Tian et al. \(2007\)](#) proposed the hidden sensitivity (HS) model, which does not involve a RD. This model is used to study the association of two sensitive characteristics with binary outcomes. As an example, the authors consider the variables  $X_1, X_2 \in \{0, 1\}$ , where  $X_1 = 1$  if the respondent takes drugs and  $X_2 = 1$  if the respondent has AIDS. The crosswise model, which can be found in [Yu et al. \(2008\)](#) and [Tan et al. \(2009\)](#), is a nonrandomized version of Warner’s model and is suitable for two-valued sensitive variables  $X \in \{0, 1\}$ . The model requires choosing a nonsensitive auxiliary variable  $W \in \{0, 1\}$  with a known distribution such that  $X$  and  $W$  are independent. [Tan et al. \(2009\)](#) suggested using the respondent’s birthday to construct  $W$ . For instance,  $W$  may indicate if the respondent is born between January and May. If  $X = 0$  and  $W = 0$ , the respondent answers  $A = 1$ . The answer  $A = 1$  is also required if  $X = 1$  and  $W = 1$ . In any other case, the answer  $A = 0$  must be given. Another NRR model is the multi-category (MC) model proposed by [Tang et al. \(2009\)](#), which is applicable to sensitive variables  $X \in \{1, \dots, k\}$ ,  $k \geq 2$ . For this model, an important assumption is made:

$$\text{at least one value of } X, \text{ say } X = 1, \text{ is nonsensitive.} \quad (1)$$

The MC model demands the choice of a nonsensitive auxiliary variable  $W \in \{1, \dots, k\}$  with a known distribution so that  $X$  and  $W$  are assumed to be independent. The period

of the birthday was suggested as a concrete  $W$ . If  $X = 1$ , the respondent gives an answer  $A$  equal to his or her value of  $W$ . If  $X = j$  ( $j = 2, \dots, k$ ), the answer  $A = j$  is given. The distribution of  $X$  can be estimated from the empirical distribution of answers  $A$ . When  $k = 2$ , the MC model equals the triangular model proposed by Yu et al. (2008). The authors of the MC model argue that truthful answers  $A$  can be expected, because the value  $X = 1$  is nonsensitive. However, we doubt that completely truthful answers will be obtained. To show the reason, consider a respondent with a sensitive attribute  $X \neq 1$ . This person may be enticed to respond untruthfully  $A = 1$  with the intention of pretending that he or she possesses the nonsensitive characteristic  $X = 1$ . Furthermore, assumption (1) constrains the applicability of the MC model to a subclass of sensitive variables. For instance, consider  $X \in \{1, \dots, k\}$  describing income classes. Then, all values of  $X$  are usually sensitive and private. For such  $X$ , we cannot find any outcome  $x$  of  $X$  with the property: each respondent who possesses the attribute  $x$  is willing to reveal it. Hence, the MC model may not perform well with such variables. To avoid these disadvantages of the MC model, we have developed a new nonrandomized response model, which is presented in the following section.

### 3 Diagonal model

#### 3.1 The answer formula

Let  $X \in \{1, \dots, k\}$  be a sensitive variable. As in the crosswise model, we choose a nonsensitive auxiliary variable  $W$ , but now with values  $1, \dots, k$ . We assume that the distribution of  $W$  is known and that  $X$  and  $W$  are independent. The respondent gives the answer

$$A := [(W - X) \bmod k] + 1. \tag{2}$$

Formula (2) should not be presented to the respondents, because some of them may be not familiar with the modular arithmetic. Instead, every respondent receives a table that gives a simple illustration of (2). For example, for  $k = 4$ , such a table would look as follows:

The number in the table provides the required answer  $A$  depending on  $X$  and  $W$ . If we artificially expand this table, we obtain the following form (not presented to the interviewees).

| $X/W$   | $W = 1$  | $W = 2$  | $W = 3$  | $W = 4$  |
|---------|----------|----------|----------|----------|
| $X = 1$ | <b>1</b> | <b>2</b> | <b>3</b> | <b>4</b> |
| $X = 2$ | <b>4</b> | <b>1</b> | <b>2</b> | <b>3</b> |
| $X = 3$ | <b>3</b> | <b>4</b> | <b>1</b> | <b>2</b> |
| $X = 4$ | <b>2</b> | <b>3</b> | <b>4</b> | <b>1</b> |

This table has four (boldfaced) diagonals of length four, where the answer  $A$  describes the diagonal the respondent belongs to. This explains the name

| $X/W$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
|-------|---|---|---|---|---|---|---|
| 1     | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| 2     | 4 | 1 | 2 | 3 | 4 | 1 | 2 |
| 3     | 3 | 4 | 1 | 2 | 3 | 4 | 1 |
| 4     | 2 | 3 | 4 | 1 | 2 | 3 | 4 |

“diagonal model”. Note that it is not possible to identify the  $X$  value with the help of the answer. Unlike the MC model, the answers  $A$  do not restrict the possible  $X$  values. Thus, there is no temptation to give an untruthful answer. Hence, we assume no nonresponses and truthful answers. Moreover, the diagonal model is applicable even if all the outcomes of  $X$  are sensitive (e.g., if  $X$  represents income classes).

We complete the section with an example for  $W$  where  $k = 4$ : let  $W$  depend on the birthday of the respondent’s mother, where  $W = 1$  if the mother was born between January 1st and August 16th,  $W = 2$  if the mother was born between August 17th and October 1st, and  $W = 3$  if the mother was born between October 2nd and November 16th. Otherwise, define  $W = 4$ . Ignoring leap years and assuming a uniform distribution of births over 365 days of the year, we have  $(\mathbb{P}(W = 1), \dots, \mathbb{P}(W = 4)) = (\frac{228}{365}, \frac{46}{365}, \frac{46}{365}, \frac{45}{365})$ . If census data are available that provide deviating probabilities, one should work with them. In principle, the respondent’s birthday can also be used to construct  $W$ . However, this variable is dangerous if the respondent thinks that the survey agency knows such data. Another example of an auxiliary characteristic is “last digits of the mother’s phone number”.

### 3.2 The maximum likelihood estimator (MLE) for $\pi$

Define  $\pi_i := \mathbb{P}(X = i)$ ,  $\pi := (\pi_1, \dots, \pi_k)^T$ ,  $c_i := \mathbb{P}(W = i)$  and  $c := (c_1, \dots, c_k)$  where  $\pi$  is unknown and  $c$  is known. The answer pattern implies that  $\lambda := (\lambda_1, \dots, \lambda_k)^T := (\mathbb{P}(A = 1), \dots, \mathbb{P}(A = k))^T = C_0 \cdot \pi$ . Here,  $C_0$  is a  $k \times k$  matrix where every row is a left-cyclic shift of the row above. The first row of  $C_0$  equals  $c$ . We call the probabilities  $c_1, \dots, c_k$  the model parameters and  $C_0$  the design matrix induced by  $c$ . To estimate  $\pi$ , assume  $n$  persons are interviewed. Let  $X_i$ ,  $W_i$  and  $A_i$  denote the  $i$ -th respondent’s value of the sensitive characteristic, the value of the auxiliary characteristic and the given answer, respectively. We assume:

- (D1) The  $n$  vectors  $(X_i, W_i)$ ,  $i = 1, \dots, n$ , are i.i.d. with  $(X_i, W_i) \sim (X, W)$ .
- (D2)  $X_i$  and  $W_i$  are independent for every  $i$ .
- (D3)  $c_1, \dots, c_k \neq 0$ . (Otherwise, if a  $c_i$  equaled zero, every answer  $A$  would restrict the possible  $X$  values.)
- (D4) The matrix  $C_0$  is invertible.

Assumption (D1) can be fulfilled by selecting the respondents according to simple random sampling with replacement. Further, let  $n_i$  be the absolute frequency of answer  $A = i$  and define  $h_i := n_i/n$  as well as  $h := (h_1, \dots, h_k)^T$ . The likelihood function for  $\pi$  is

$$L(\pi) = \prod_{i=1}^k \mathbb{P}(A = i)^{n_i}, \quad \pi \in \{(x_1, \dots, x_k)^T : x_i \in [0, 1], x_1 + \dots + x_k = 1\} =: D. \tag{3}$$

Define  $E := \{C_0 \cdot x : x \in D\}$ . It can be shown by standard calculus that  $\hat{\pi} = C_0^{-1} \cdot h$  is the unique MLE for  $\pi$  if  $h \in E$ . If  $\pi_1, \dots, \pi_k > 0$ , we have  $\mathbb{P}(h \in E) \rightarrow 1$  as  $n \rightarrow \infty$  due to the weak law of large numbers. However,  $h \notin E$  (that is  $C_0^{-1}h \notin D$ ) may occur. In this case, the estimation of  $\pi$  is most convenient with the expectation maximization (EM) algorithm. The complete data matrix is  $Y = (A_i, X_i)_{i=1, \dots, n}$ . With realizations  $(a_i, x_i)$  the complete data log-likelihood for  $\pi \in D$  is equal to

$$\begin{aligned} l_{com}(\pi; a_1, \dots, a_n, x_1, \dots, x_n) &= \log \left( \prod_{i=1}^n \mathbb{P}(A_i = a_i, X_i = x_i) \right) \\ &= C + \sum_{i=1}^k m_i \cdot \log \pi_i, \end{aligned} \tag{4}$$

where  $C$  is a negligible constant and  $m_i = \sum_{j=1}^n 1_{\{x_j=i\}}$ , that is  $m_i$  is the number of persons in the sample whose value of the sensitive variable equals  $i$ . The observed data log-likelihood for  $\pi \in D$  equals  $\log L$  where  $L$  is from (3). For the E step, suppose we have an estimate  $\pi^{(t)}$  from the preceding iteration  $t$  and let  $M_i$  be the random variables corresponding to the values  $m_i$ . Then

$$\hat{m}_i := \mathbb{E}_{\pi^{(t)}} (M_i \mid A_1 = a_1, \dots, A_n = a_n) = \sum_{j=1}^k n_j \cdot \frac{\pi_i^{(t)} \cdot C_0(i, j)}{\mathbb{P}_{\pi^{(t)}}(A = j)}, \tag{5}$$

where  $C_0(i, j)$  denotes entry  $(i, j)$  of the design matrix  $C_0$ . Thus, the log-likelihood (4) is estimated by  $\widehat{l}_{com}(\pi) = C + \sum_{i=1}^k \hat{m}_i \cdot \log \pi_i$ . At the succeeding M step, the maximum of  $\widehat{l}_{com}$ , which is given by  $\pi_i^{(t+1)} = \hat{m}_i / \sum_{j=1}^k \hat{m}_j$ , is calculated. As initial parameter  $\pi_i^{(0)} = 1/k$  can be used.

### 3.3 Standard errors and confidence intervals

To express the dependence of  $h$  and the MLE  $\hat{\pi}$  on  $n$ , we will occasionally write  $h_n$  and  $\hat{\pi}_n$  in the sequel. The multivariate central limit theorem implies  $\sqrt{n}(h_n - \lambda) \xrightarrow{\mathcal{L}} N(0, \Sigma)$  with  $\Sigma = \text{diag}(\lambda) - \lambda\lambda^T$ . Note,  $N(0, \Sigma)$  is a singular normal distribution. Consider the estimator  $\tilde{\pi}_n := C_0^{-1}h_n$ . If  $h_n \in E$ , then  $\hat{\pi}_n = \tilde{\pi}_n$  holds. The estimator  $\tilde{\pi}_n$  is unbiased and we have  $\text{Var}(\tilde{\pi}_n) = \frac{1}{n} \cdot C_0^{-1} \cdot \Sigma \cdot C_0^{-1} = \frac{1}{n} \cdot (\Gamma + \Delta)$  with

$$\Gamma := C_0^{-1} \cdot \text{diag}(\lambda) \cdot C_0^{-1} - \text{diag}(\pi) \quad \text{and} \quad \Delta := \text{diag}(\pi) - \pi \cdot \pi^T \tag{6}$$

By the continuous mapping theorem  $\sqrt{n}(\tilde{\pi}_n - \pi) \xrightarrow{\mathcal{L}} N(0, \Gamma + \Delta)$ . We claim  $\sqrt{n}(\hat{\pi}_n - \pi) \xrightarrow{\mathcal{L}} N(0, \Gamma + \Delta)$ . To prove the claim, it suffices to show the stochastic convergence of  $\delta_n := \sqrt{n}(\hat{\pi}_n - \tilde{\pi}_n)$  to 0: For  $\varepsilon > 0$  we have  $\mathbb{P}(|\delta_n| > \varepsilon) \leq \mathbb{P}(h_n \notin E) \rightarrow 0 (n \rightarrow \infty)$ , because  $h_n$  converges almost surely to  $\lambda$ .

Now consider a function  $\psi = f(\pi) \in \mathbb{R}$  of the parameter  $\pi$ . The MLE for  $\psi$  is given by  $\hat{\psi} = f(\hat{\pi})$ . Let  $\hat{\Gamma}$  and  $\hat{\Delta}$  arise by using the MLE  $\hat{\pi}$  instead of  $\pi$  as well as  $C_0\hat{\pi}$  instead of  $\lambda$  in (6) and define  $S_{\hat{\psi}}^2 = \frac{1}{n} \cdot \nabla^T f(\hat{\pi}) \cdot (\hat{\Gamma} + \hat{\Delta}) \cdot \nabla f(\hat{\pi})$  where  $\nabla$  denotes the gradient. Applying the delta method (cf. Van der Vaart 2007, p. 25) and Slutsky’s theorem  $(\hat{\psi} - \psi)/S_{\hat{\psi}}$  converges in distribution to the standard normal distribution. Thus,  $AS := [\hat{\psi} - z_{1-\alpha/2} \cdot S_{\hat{\psi}}, \hat{\psi} + z_{1-\alpha/2} \cdot S_{\hat{\psi}}]$  is an asymptotic  $(1 - \alpha)$  confidence interval (CI) for  $\psi$ . Here,  $z_{1-\alpha/2}$  is the  $(1 - \alpha/2)$  quantile of the standard normal distribution.

Bootstrap standard errors and bootstrap confidence intervals can be derived according to Efron and Tibshirani (1993), chapters 6 and 12–13. Assume  $\hat{\pi}$  and  $\hat{\psi}$  are calculated. A bootstrap replication  $\hat{\psi}^{(i)}, i = 1, \dots, B$ , can be generated by drawing new answer frequencies  $n^{(i)} := (n_1^{(i)}, \dots, n_k^{(i)}) \sim \text{Multinomial}(n, C_0\hat{\pi})$ , computing the MLE  $\hat{\pi}^{(i)}$  based on  $n^{(i)}$  and defining  $\hat{\psi}^{(i)} := f(\hat{\pi}^{(i)})$ . Then, the bootstrap estimate  $\hat{SE}(\hat{\psi})$  for the standard error of  $\hat{\psi}$  is given by the empirical standard deviation of the replications  $\hat{\psi}^{(1)}, \dots, \hat{\psi}^{(B)}$ . Further, if  $(\hat{\psi} - \psi)/\hat{SE}(\hat{\psi})$  is approximately standard normal,  $BT1 := [\hat{\psi} - z_{1-\alpha/2} \cdot \hat{SE}(\hat{\psi}), \hat{\psi} + z_{1-\alpha/2} \cdot \hat{SE}(\hat{\psi})]$  is a  $(1 - \alpha)$  bootstrap CI for  $\psi$ . Alternatively, a  $(1 - \alpha)$  bootstrap CI for  $\psi$  without the normality assumption is given by  $BT2 := [\psi_{\alpha/2}, \psi_{1-\alpha/2}]$  where the bounds are the  $\alpha/2$  and  $(1 - \alpha/2)$  quantiles of the replications  $\hat{\psi}^{(1)}, \dots, \hat{\psi}^{(B)}$ .

In simulations, which are not included in the paper, we have considered  $k = 3, n \in \{100, 200\}$ , several values of  $\pi$  and  $c$ , as well as linear functions  $f$  (e.g., the projection  $f(x_1, x_2, x_3) = x_2$ ). We have seen that  $AS$  and  $BT2$  usually provide empirical coverage probabilities (CP) between 93 and 96% ( $\alpha = 5\%$ ). The CP of  $BT1$  turned out to be lower than the CP of  $AS$  and  $BT2$ . Furthermore, the average width of  $AS$  was mostly larger than the average width of  $BT2$ . For these reasons, we recommend the use of the  $BT2$  confidence intervals.

(A MATLAB program that computes the MLE for  $\pi$  and confidence intervals for  $\psi$  is provided as online supplemental material.)

### 3.4 The special shape of the design matrix

In this section, we show the connection between  $C_0$  and circulant matrices. Further, we mention some properties of circulant matrices that can be found in Gray (2006), chapter 3, and will be applied in the next section.

If we permute the rows of  $C_0$ , we obtain a matrix  $C$  where every row of  $C$  is a right-cyclic shift of the row above. In particular, define  $C = S \cdot C_0$  where the entries of  $S$  are given by  $S(i, j) = 1_{\{C_0(i, j)=c_1\}}$ .  $C$  is a circulant matrix (circulant matrices are a subset of Toeplitz matrices). Occasionally, we will write  $C = \text{circulant}(c_1, \dots, c_k)$ . Define  $\psi_1, \dots, \psi_k$  by

$$\psi_l = \sum_{m=1}^k \exp\left(-2\pi i \cdot \frac{(l-1) \cdot (m-1)}{k}\right) \cdot c_m. \tag{7}$$

Then,  $\psi := (\psi_1, \dots, \psi_k)$  is a vector of eigenvalues of  $C$  and  $\psi$  can be computed by multiplying the  $k \times k$  Fourier matrix  $F = [\exp(-2\pi i \cdot (l-1) \cdot (m-1)/k)]_{l,m}$  with the vector  $c^T = (c_1, \dots, c_k)^T$ , i.e.,  $\psi$  is the discrete Fourier transform of  $c^T$ . Moreover,  $C^{-1}$  is circulant if  $C$  is circulant and invertible.

### 3.5 Efficiency loss and degree of privacy protection

To study the efficiency of the MLE  $\hat{\pi} = \hat{\pi}_n$ , the mean squared error  $MSE(\hat{\pi}_n) = \mathbb{E}((\hat{\pi}_n - \pi)(\hat{\pi}_n - \pi)^T)$  is suitable. The whole matrix  $MSE(\hat{\pi}_n)$  is cumbersome, so we use  $trace(MSE(\hat{\pi}_n)) = MSE(\hat{\pi}_{n,1}) + \dots + MSE(\hat{\pi}_{n,k})$ , i.e., the sum of the mean squared errors of the components  $\hat{\pi}_{n,j}$ , as a measure of the estimation inaccuracy. Define  $\tilde{\pi}_n$  as in Sect. 3.3. We claim

$$\hat{T} / \tilde{T} := trace(MSE(\hat{\pi}_n)) / trace(MSE(\tilde{\pi}_n)) \xrightarrow{n} 1. \tag{8}$$

To see (8), note that  $n \cdot MSE(\tilde{\pi}_n) = \Gamma + \Delta$ , as described in Sect. 3.3. Because  $\hat{\pi}_n = \tilde{\pi}_n$  for large  $n$ , we have  $\hat{T} / \tilde{T} = (n \cdot \hat{T}) / (n \cdot \tilde{T}) \xrightarrow{n} trace(\Gamma + \Delta) / trace(\Gamma + \Delta) = 1$ . That is,  $\hat{\pi}_n$  and  $\tilde{\pi}_n$  are equally efficient as  $n \rightarrow \infty$ . Hence, for large  $n$ ,

$$\begin{aligned} trace(MSE(\hat{\pi}_n)) &\approx trace(MSE(\tilde{\pi}_n)) = trace(Var(\tilde{\pi}_n)) \\ &= trace(\Gamma/n) + trace(\Delta/n). \end{aligned} \tag{9}$$

Imagine the respondents are interviewed directly, i.e., they are requested to tell their values of  $X$ , and assume that truthful answers are given. If  $n$  people are asked, the MLE for  $\pi$  is given by  $\hat{\pi}^{dir} = (\hat{\pi}_1^{dir}, \dots, \hat{\pi}_k^{dir})^T$  with  $\hat{\pi}_i^{dir} = \frac{1}{n} \sum_{j=1}^n 1_{\{X_j=i\}}$ .

The variance matrix of  $\hat{\pi}^{dir}$  is then given by  $\Delta/n$ . Since  $\frac{1}{n} \cdot trace(\Gamma) \approx trace(MSE(\hat{\pi}_n)) - trace(MSE(\hat{\pi}^{dir}))$ , the quantity  $\frac{1}{n} \cdot trace(\Gamma)$  describes the efficiency loss caused by questioning according to the diagonal model, instead of direct questioning.

The efficiency loss  $trace(\Gamma)/n$  converges to zero as  $n$  increases and we have:

**Theorem 1** *Let  $\pi$  and  $c = (c_1, \dots, c_k)$  describe the distribution of the sensitive variable  $X$  and the auxiliary variable  $W$ , respectively. Further let  $\Gamma$  be defined by (6). Then:*

- (a)  $trace(\Gamma)$  depends on  $c$ , but not on  $\pi$ .
- (b)  $trace(\Gamma) \geq 0$ .
- (c)  $trace(\Gamma)$  attains its minimum value zero iff  $W$  has a degenerate distribution (i.e.,  $W$  is a constant variable).

*Proof* Define  $C_0$  as design matrix induced by  $c$  and  $C = circulant(c_1, \dots, c_k)$ . Let  $S$  be given by  $S(i, j) = 1_{\{C_0(i,j)=c_1\}}$ . Then,  $S$  is a self-inverse permutation matrix.

The equation  $C_0 = S \cdot C$  implies  $D_0 := C_0^{-1} = C^{-1}S$ , i.e.  $D_0$  arises by column permutation of a circulant matrix. Thus, every row of  $D_0$  is a left-cyclic shift of the row above. Let  $(d_1, d_2, \dots, d_k)$  denote the first row of  $D_0$ . The vector of the diagonal elements of  $\Gamma$  fulfills

$$\tilde{\Gamma} := (\Gamma(1, 1), \dots, \Gamma(k, k))^T = (D_0^{(2)} \cdot C_0 - I) \cdot \pi \tag{10}$$

where  $D_0^{(2)}$  is the componentwise square of  $D_0$ . For (10), the definitions of  $\Gamma$  and  $D_0$  as well as the identity  $\pi = D_0 \cdot \lambda$  yield  $\tilde{\Gamma} = (D_0^{(2)} - D_0) \cdot \lambda = (D_0^{(2)} - D_0) \cdot C_0 \cdot \pi$ . This verifies (10).

Using the shape of  $C_0$  and  $D_0^{(2)}$  it is easy to verify that  $E := (D_0^{(2)} \cdot C_0 - I)$  is circulant. Since  $trace(\Gamma)$  equals the sum of the first column of  $E$ , we have

$$trace(\Gamma) = d_1^2 + \dots + d_k^2 - 1, \tag{11}$$

which shows (a). Consider an arbitrary eigenvalue  $\psi$  of  $C_0$  with eigenvector  $v = (v_1, \dots, v_k)^T$  where the 1-norm of  $v$  equals 1. Then, by the triangle inequality:

$$|\psi| = \|C_0 v\|_1 \leq (|v_1| + \dots + |v_k|) (c_1 + \dots + c_k) = 1. \tag{12}$$

Further, for  $i = 1, \dots, k$  let  $\psi_i$  and  $\psi_i(M)$  denote the (not necessarily distinct) eigenvalues of  $C_0$  and an arbitrary  $k \times k$ -matrix  $M$ , respectively. By a well-known property of the Frobenius norm  $\|\cdot\|_F$  (see Gentle 2007, p. 132) and (12) we have

$$d_1^2 + \dots + d_k^2 = \frac{1}{k} \|D_0\|_F^2 = \frac{1}{k} \sum_{i=1}^k \psi_i(D_0 \cdot D_0) = \frac{1}{k} \left( \frac{1}{\psi_1^2} + \dots + \frac{1}{\psi_k^2} \right) \geq 1, \tag{13}$$

that is,  $trace(\Gamma) \geq 0$  as claimed in (b). Further, we have

$$trace(\Gamma) = 0 \iff \psi_1^2 = \psi_2^2 = \dots = \psi_k^2 = 1 \iff C_0 \cdot C_0 = I.$$

Consider  $c := (c_1, \dots, c_k)$  and assume  $c_j = 1$  and  $c_i = 0$  for  $i \neq j$ . Then, clearly  $C_0^2 = I$  holds. Conversely, consider a  $c$  with  $C_0^2 = I$ . It follows that  $c_1^2 + \dots + c_k^2 = 1$ . Thus  $c_1(1 - c_1) + \dots + c_k(1 - c_k) = 0$ , which is only satisfied if  $c_j = 1$  for some  $j \in \{1, \dots, k\}$  and  $c_i = 0$  for any  $i \neq j$ . Altogether we have  $trace(\Gamma) = 0$  iff  $c$  corresponds to a degenerate distribution.  $\square$

Theorem 1 (a) says that the efficiency loss  $trace(\Gamma)$  is independent of the unknown  $\pi$ . Occasionally, we will write  $trace(\Gamma(c))$  to symbolize the dependence of  $trace(\Gamma)$  on  $c$ . Furthermore, Theorem 1(c) states that the diagonal model would be most efficient for a constant auxiliary variable  $W$ . However, such a  $W$  is not appropriate, because the interviewer can conclude the  $X$  value from answer (2). This would be contradictory to the idea of the diagonal model and would result in nonresponses and untruthful answers.



In the remainder of this section, we will have a closer look at the connection between the DPP and efficiency loss.

As explained above, the privacy is not protected if  $W$  has a degenerate distribution. In contrast, the privacy is protected as much as possible in the case of an exactly uniformly distributed  $W$ ; in this case, any answer  $A$  according to the diagonal model does not change the probability that the variable  $X$  attains a value  $x$ , i.e.,  $A$  and  $X$  are independent (notice the constant likelihood function in this case). To measure the DPP, it seems reasonable to consider the closeness of the distribution of  $W$  to a degenerate and a uniform distribution. This motivates us to quantify the DPP by the empirical standard deviation of the vector  $c$

$$\sigma(c) := std(c) = \sqrt{(c_1^2 + \dots + c_k^2)/(k - 1) - [(k - 1)k]^{-1}} \in [0, \sqrt{1/k}]. \tag{14}$$

A large  $std(c)$  means an approach to a degenerate distribution and consequently a small DPP, while a small  $std(c)$  indicates that the distribution of  $W$  is close to a uniform distribution and a high DPP is available.

The efficiency loss  $trace(\Gamma)$  has a lower bound depending on  $\sigma$ :

**Theorem 2** (a) *Consider a vector  $c = (c_1, \dots, c_k)$  describing the distribution of the auxiliary variable  $W$ . Let  $\sigma > 0$  and  $trace(\Gamma)$  denote the corresponding DPP and efficiency loss due to indirect questioning, respectively. Then*

$$trace(\Gamma) \geq \frac{(k - 1) \left(\frac{1}{k} - \sigma^2\right)}{k\sigma^2}. \tag{15}$$

(b) *Let  $\sigma \in \left(0, \sqrt{\frac{1}{k}}\right]$  be given. Define  $c = (c_1, \dots, c_k)$  with*

$$c_1 = \frac{1}{k} + \frac{k - 1}{k} \sqrt{k\sigma^2} \quad \text{and} \quad c_2 = \dots = c_k = \frac{1}{k} - \frac{1}{k} \sqrt{k\sigma^2}. \tag{16}$$

*Then, we have  $std(c) = \sigma$  and  $trace(\Gamma(c)) = \frac{(k-1)(\frac{1}{k}-\sigma^2)}{k\sigma^2}$ .*

*Proof* (a) We first minimize the function

$$f(x) = f(x_1, \dots, x_{k-1}) := \frac{1}{k} \left( \frac{1}{x_1^2} + \dots + \frac{1}{x_{k-1}^2} \right) - 1 + \frac{1}{k}, \quad x_i \neq 0$$

under the restriction  $g(x) = x_1^2 + \dots + x_{k-1}^2 - k(k - 1)\sigma^2 = 0$ . Since  $f \rightarrow \infty$  if  $x_i \rightarrow 0$ ,  $f$  possesses a global minimum  $m^* = (m_1^*, \dots, m_{k-1}^*)$ . By Lagrange multipliers, one can show  $(m_1^*)^2 = \dots = (m_{k-1}^*)^2 = k\sigma^2$ . Then

$$f(m^*) = \frac{(k-1)\left(\frac{1}{k} - \sigma^2\right)}{k\sigma^2} \leq f(x), \text{ for all } x \text{ with } x_i \neq 0, g(x) = 0. \tag{17}$$

Now define  $C_0$  as the design matrix formed with the vector  $c$ . For  $i = 1, \dots, k$  let  $\psi_i$  and  $\psi_i(C_0^2)$  denote the eigenvalues of  $C_0$  and  $C_0^2$ , respectively. By a property of the Frobenius norm  $\|\cdot\|_F$  (Gentle 2007, p. 132)

$$c_1^2 + \dots + c_k^2 = \frac{1}{k} \|C_0\|_F^2 = \frac{1}{k} \sum_{i=1}^k \psi_i(C_0^2) = \frac{1}{k} (\psi_1^2 + \dots + \psi_k^2).$$

holds. Thus, with (14) we obtain

$$\sigma^2 = \frac{1}{(k-1)k} (\psi_1^2 + \dots + \psi_k^2 - 1) = \frac{1}{(k-1)k} (\psi_2^2 + \dots + \psi_k^2) \tag{18}$$

where the latter equality is true since  $C_0$  has an eigenvalue equal to 1 (with eigenvector  $(1, \dots, 1)^T$ ). According to (11) and (13) we have

$$\text{trace}(\Gamma) = \frac{1}{k} \left( \frac{1}{\psi_1^2} + \dots + \frac{1}{\psi_k^2} \right) - 1 = \frac{1}{k} \left( \frac{1}{\psi_2^2} + \dots + \frac{1}{\psi_k^2} \right) - 1 + \frac{1}{k}. \tag{19}$$

Then, the claim follows by application of (17).

(b) Consider  $c$  as in the claim. Let  $C_0$  be the corresponding design matrix with eigenvalues  $\psi_1, \dots, \psi_k$  and  $x = (x_1, \dots, x_k)$  the first row of  $C_0^2$ . It follows straightforwardly that

$$x_1 = \frac{1}{k}(1 + (k-1)k\sigma^2) \text{ and } x_2 = \dots = x_k = \frac{1}{k}(1 - k\sigma^2). \tag{20}$$

Since  $C_0^2$  is a circulant matrix, the eigenvalues can be obtained by a discrete Fourier transform of  $x$ , cf. (7). Then,  $\psi_l^2 = \sum_{i=1}^k x_i = 1$  and with the formula for the sum of a geometric sequence we have for  $l = 2, \dots, k$

$$\begin{aligned} \psi_l^2 &= \sum_{m=1}^k \exp\left(-2\pi i \cdot \frac{(l-1) \cdot (m-1)}{k}\right) \cdot x_m \\ &= x_1 + x_2 \cdot \sum_{m=1}^{k-1} \exp\left(-2\pi i \cdot \frac{(l-1) \cdot m}{k}\right) = x_1 - x_2 = k\sigma^2. \end{aligned}$$

Hence, by (18) and (19) we can compute that  $std(c)$  equals the given  $\sigma$  and that  $\text{trace}(\Gamma(c))$  attains the lower bound  $\frac{(k-1)\left(\frac{1}{k} - \sigma^2\right)}{k\sigma^2}$ . □

We immediately attach two remarks concerning this theorem:

(1) If  $k = 2$  or  $k = 3$ , then for every  $c = (c_1, \dots, c_k)$ , the point  $(std(c), trace(\Gamma(c)))$  is located on the curve  $\left\{ \left( \sigma, \frac{(k-1)(\frac{1}{k}-\sigma^2)}{k\sigma^2} \right) : 0 \leq \sigma \leq \sqrt{\frac{1}{k}} \right\}$ .

For  $k = 2$ : One eigenvalue of  $C_0$ , say  $\psi_1$ , is always equal to 1 and  $\psi_2^2$  equals  $2\sigma^2$  due to (18). Then, (19) implies  $trace(\Gamma(c)) = (\frac{1}{2}-\sigma^2) \cdot (2\sigma^2)^{-1}$ . Similar in case  $k = 3$ : We obtain  $\psi_1 = 1$  and  $\psi_2^2 + \psi_3^2 = 6\sigma^2$ . Additionally,  $1 = c_1 + c_2 + c_3 = trace(C_0) = \psi_1 + \psi_2 + \psi_3$  implies  $\psi_2 = -\psi_3$ . Thus,  $\psi_2^2 = \psi_3^2 = 3\sigma^2$  and the claim follows with (19).

(2) Fix a  $\sigma > 0$  and consider  $c$  given in Theorem 2 (b). This  $c$  is not the only vector with a provided DPP equal to  $\sigma$  and a caused efficiency loss  $trace(\Gamma)$  that attains the lower bound  $\frac{(k-1)(\frac{1}{k}-\sigma^2)}{k\sigma^2}$ .

Assume  $c' \neq c$  has design matrix  $C'_0$  and  $(C'_0)^2 = circulant(x_1, \dots, x_k)$  with the  $x_i$  from (20). Then, according to the proof of theorem 2(b),  $std(c') = \sigma$  and  $trace(\Gamma(c')) = (k-1)(\frac{1}{k}-\sigma^2) \cdot (k\sigma^2)^{-1}$  hold, that is,  $c'$  would be an alternative to  $c$ . Now, such a vector  $c'$  can be found by permuting  $c$ . However, other choices are also possible: For instance, for  $k = 4$  and  $\sigma = 0.4$  we can use  $(0.45 + \sqrt{0.1591}, 0.02, 0.45 - \sqrt{0.1591}, 0.08)$ , too. □

Theorem 2 contains two core statements: first, there are optimal and nonoptimal distributions of  $W$ . A distribution of  $W$  is not optimal if the corresponding vector  $c$  has a standard deviation  $\sigma$ , but  $trace(\Gamma(c)) \neq \frac{(k-1)(\frac{1}{k}-\sigma^2)}{k\sigma^2}$ . In this case,  $c$  would lead to an efficiency loss larger than necessary for a DPP in the amount of  $\sigma$ . A vector  $c$  is optimal if and only if  $(\sigma(c), trace(\Gamma(c)))$  is an element of the “optimality curve”  $\mathcal{O} := \{(\sigma, \gamma^*(\sigma)) : \sigma \in (0, \sqrt{1/k}], \gamma^*(\sigma) = (k-1)(k^{-1}-\sigma^2)/(k\sigma^2)\}$ . Of course, it is reasonable to use only optimal vectors  $c$ . For these vectors, the efficiency loss is a function of the DPP. Since  $\gamma^*$  is a decreasing function, the efficiency loss decreases if the standard deviation increases. This implies the second key result: Increasing efficiency of the diagonal model corresponds to decreasing protection of the respondents’ privacy.

### 4 Simulation study

Remember that the results in 3.5 were obtained under the assumption of a large sample size  $n$ . In the subsequent simulation study, we are going to examine the validity of the results for a concrete example with different sample sizes. We run all simulations using MATLAB, Version 7.11.

#### 4.1 Simulated MSE sum

For our simulations, we consider a concrete variable  $X$  with a known distribution. Let  $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_k)^T$  denote the DM estimator for a known  $\pi$ . For a fixed and possibly “small” sample size  $n$ , we cannot calculate  $trace(MSE(\hat{\pi}))$  with the help of Sect. 3.5, because (9) holds only for “large”  $n$ . Therefore, assume  $l$  independent realizations of  $\hat{\pi}$ —denoted with  $\hat{\pi}^{(i)} = (\hat{\pi}_1^{(i)}, \dots, \hat{\pi}_k^{(i)})^T, i = 1, \dots, l$ —are available.

Define  $MSE(\hat{\pi}_j) := \frac{1}{l} \sum_{i=1}^l (\hat{\pi}_j^{(i)} - \pi_j)^2$  for  $j = 1, \dots, k$ . The simulated MSE sum  $trace(MSE(\hat{\pi})) := \sum_{j=1}^k MSE(\hat{\pi}_j)$  is then a suitable approximation for the efficiency measure  $trace(MSE(\hat{\pi}))$ .

#### 4.2 DPP and efficiency in a simulated example

We assume that the attribute “income” is the subject of a survey. For instance, a company could be interested in the income of customers, because products and prices must be attuned to it. We use Germany’s income distribution published by the [Federal Statistical Office, Germany \(2009\)](#). Suppose we focus on four income classes (in Euro per year): below 10,000 ( $X = 1$ ), 10,000–24,999 ( $X = 2$ ), 25,000–49,999 ( $X = 3$ ), and more than 50,000 ( $X = 4$ ). According to the data of the Federal Statistical Office, the distribution  $\mathbb{P}_X$  is given by

$$\mathbb{P}_X \sim \pi = (\pi_1 \ \pi_2 \ \pi_3 \ \pi_4)^T = (0.2371 \ 0.3203 \ 0.3343 \ 0.1083)^T. \tag{21}$$

To express the dependence of the DM estimator on the sample size  $n$  and on the distribution of  $W$ , i.e., on the vector  $c = (c_1, \dots, c_4)$ , we write  $\hat{\pi}_n(c)$ . First, we investigate the efficiency of  $\hat{\pi}_n(c)$  for randomly drawn vectors  $c$ . In particular, 500 vectors are drawn so that they are preferably uniformly scattered over the set of valid points  $D_k = \{(x_1, \dots, x_k) \in [0, 1]^k : x_1 + \dots + x_k = 1\}$ . To obtain one such vector  $c$ , we firstly generate  $\tilde{c} := (c_1, \dots, c_{k-1})$  from a Dirichlet distribution with parameter  $(1, \dots, 1)$ , see [Gentle \(1998\)](#), p. 111. Subsequently, we define  $c_k = 1 - (c_1 + \dots + c_{k-1})$ .

For each drawn vector  $c$ :

1. Calculate  $\sigma(c)$ .
2. Generate  $l = 1,000$  independent simple random samples with replacement of size  $n \in \{50, 100, 250\}$  from the distribution  $\mathbb{P}_X$  in (21). For each sample, we compute the DM estimator, so we have  $l$  independent realizations of  $\hat{\pi}_n(c)$ .
3. Calculate the simulated MSE sum  $trace(MSE(\hat{\pi}_n(c)))$ .

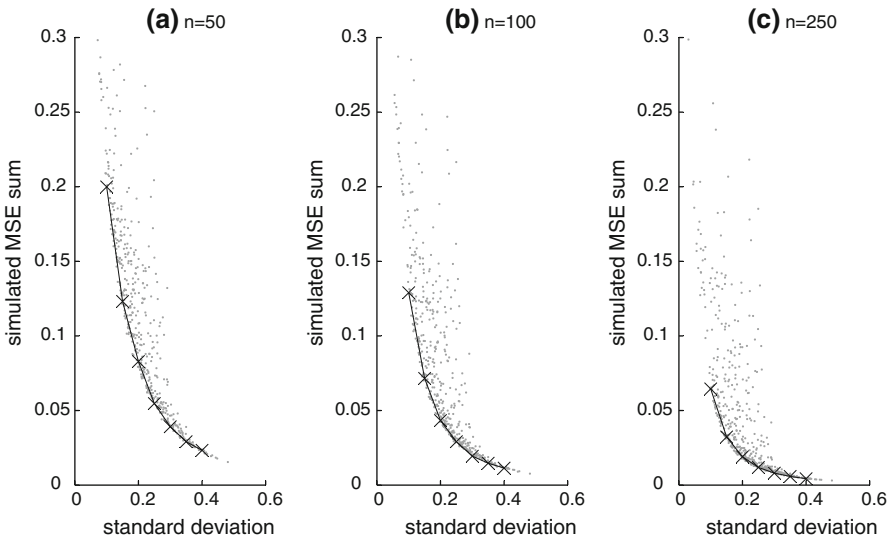
Subsequently, we choose different vectors  $c^*$  according to Theorem 2 (b)—see Table 1—and run steps 2 and 3 for these vectors.

For all considered vectors, we plot the simulated MSE sum against the standard deviation in Fig. 1. Here, we see that the key results described after Theorem 2, which were obtained under the assumption of a “large”  $n$ , remain valid for  $n \in \{50, 100, 250\}$ . In particular, the point clouds for the randomly drawn vectors  $c$  have lower bounds. Thus, there are optimal and nonoptimal vectors  $c$ . For all considered  $n$ , the points  $(\sigma(c^{*(i)}), trace(MSE(\hat{\pi}_n(c^{*(i)})))$ , which are marked with  $\times$  in the figure, are located close to this bound. In other words,  $c^{*(i)}$  ( $i = 1, \dots, 7$ ) are good choices for the distribution of  $W$  for the corresponding degrees of privacy protection.

Furthermore, if we connect the points marked with  $\times$ , we obtain a decreasing function. Thus, an increasing standard deviation implies a decreasing estimation

**Table 1** The vectors  $c^*$  that we used from Theorem 2(b)

| $i$ | $c^{*(i)} = (c_1^{*(i)}, \dots, c_4^{*(i)})$ |       |       |       | $\sigma(c^{*(i)})$ |
|-----|--|-------|-------|-------|--------------------|
| 1   | 0.400  | 0.200 | 0.200 | 0.200 | 0.10               |
| 2   | 0.475  | 0.175 | 0.175 | 0.175 | 0.15               |
| 3   | 0.550  | 0.150 | 0.150 | 0.150 | 0.20               |
| 4   | 0.625  | 0.125 | 0.125 | 0.125 | 0.25               |
| 5   | 0.700  | 0.100 | 0.100 | 0.100 | 0.30               |
| 6   | 0.775  | 0.075 | 0.075 | 0.075 | 0.35               |
| 7   | 0.850  | 0.050 | 0.050 | 0.050 | 0.40               |



**Fig. 1** Plots of simulated MSE sum against standard deviation for different sample sizes. Each point (·) corresponds to a vector  $c$  drawn randomly from  $D_4$ , and each black cross corresponds to a vector  $c^{*(i)}$

inaccuracy. Hence, a decreasing DPP corresponds to increasing efficiency. Finally, for any  $c^{*(i)}$ , the simulated MSE sum decreases with increasing sample size.

In practice, the investigator should first fix a DPP  $\sigma$  around the middle of the range (0, 0.5]. For this  $\sigma$ , he or she should then choose a vector  $c$  using Theorem 2 (b). Subsequently, an auxiliary attribute  $W$  such as the “birthday of the respondent’s mother” or the “last digits of the mother’s phone number” should be adapted to the chosen  $c$ . For example, define  $W$  as in the concrete example of Sect. 3.1, where  $c = (c_1, \dots, c_4) = \left(\frac{228}{365}, \frac{46}{365}, \frac{46}{365}, \frac{45}{365}\right) \approx (0.625, 0.126, 0.126, 0.123)$ . This  $c$  is then a good approximation to  $c^{*(4)}$  from Table 1, which was obtained from Theorem 2 (b).

**5 Summary**

In this paper, we have presented the nonrandomized diagonal model, which is a privacy-protecting survey design for multichotomous sensitive variables. The proposed method

can be conducted even if all values of the considered variable are sensitive. We have shown that the formula for the answer can be easily explained to the respondents. After deriving the MLE and confidence intervals, we discussed the DPP and efficiency. Here, a mathematical function for the dependence of the efficiency on the model parameters was derived in Theorem 2. This enables the interviewer to choose optimal model parameters for a desired DPP.

**Acknowledgments** The author would like to thank a referee for valuable comments and suggestions.

## References

- Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, London
- Federal Statistical Office, Germany (2009) Lohn- und Einkommensteuer - Fachserie 14 Reihe 7.1 - 2004. [online], document number 2140710049004. Available at [www.destatis.com](http://www.destatis.com) (only in German language)
- Gentle JE (1998) Random number generation and Monte Carlo methods. Springer, Berlin
- Gentle JE (2007) Matrix algebra: theory, computations, and applications in statistics. Springer, Berlin
- Gray RM (2006) Toeplitz and circulant matrices: a review. Now
- Tan MT, Tian GL, Tang ML (2009) Sample surveys with sensitive questions: a nonrandomized response approach. *Am Stat* 63:9–16
- Tang ML, Tian GL, Tang NS, Liu Z (2009) A new non-randomized multi-category response model for surveys with a single sensitive question: design and analysis. *J Korean Stat Soc* 38:339–349
- Tian GL, Yu JW, Tang ML, Geng Z (2007) A new non-randomized model for analysing sensitive questions with binary outcomes. *Stat Med* 26:4238–4252
- Van der Vaart AW (2007) Asymptotic statistics. Cambridge University Press, Cambridge
- Warner SL (1965) Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *J Am Stat Assoc* 60:63–69
- Yu JW, Tian GL, Tang ML (2008) Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika* 67:251–263