ISSN 1066-5307, Mathematical Methods of Statistics, 2016, Vol. 25, No. 2, pp. 133–144. © Allerton Press, Inc., 2016.

Risk-Optimal Estimators for Survey Procedures with Certain Indirect Questions

H. Groenitz[1*]

[1]*Philipps-Univ. Marburg, School of Business and Economics, Marburg, Germany*

Received November 13, 2015; in final form, June 2, 2016

Abstract—Surveys usually consist of a list of direct questions. However respondents reluctantly provide direct information on sensitive topics such as socially undesired behavior (e.g., social fraud, discrimination, tax evasion), income or political preferences. For this reason, the diagonal technique (DT), an indirect questioning procedure has been proposed in the literature. In this paper, we consider multiple categorical target variables where all or some of the variables are gathered by the DT. The maximum likelihood (ML) estimator for the joint distribution depends on the setup of the survey procedure, i.e., on certain parameters to adjust. We conduct a decision-theoretic analysis and derive risk-optimal ML estimators. The special point of our investigation is the incorporation of the degree of privacy protection (DPP). In particular, in the class of ML estimators corresponding to a given DPP, we detect an estimator with the lowest risk, i.e., with the highest quality.

Keywords: optimal decision theory, loss function, Kronecker product, circulant matrix, Frobenius norm, sensitive characteristic.

2000 Mathematics Subject Classification: 62Cxx, 62D05, 62F12.

DOI: 10.3103/S1066530716020046

1. INTRODUCTION

Statistical surveys are usually based on direct questions, that is, the respondents are instructed to tell their values for the variables involved in the survey. However, it may occur that some variables are difficult to gather by direct questioning. For instance, for the characteristic income, which is frequently relevant in social surveys on living conditions and market research surveys, answer refusal often occurs. In addition to answer refusal, untruthful answer can be expected. Possible reasons for these problems are that people with large income may be afraid of envy and that people with small income may be embarrassed. As another example, assume that a study on social fraud, especially the economic loss through moonlighting, is intended to be conducted. Then, a direct question on, for example, a person's average monthly revenues from undeclared work is very critical and will cause answer refusal and socially desired answers. The same problems with the data quality exist also for other sensitive topics (e.g., tax evasion, insurance fraud, political preferences, cheating in exams, discrimination) if direct questioning is applied.

These troubles motivate the application of indirect questioning procedures, which safeguard the respondents' privacy, but deliver data that allow inference for the variables of interest (e.g., Fox and Tracy (1986), Chaudhuri and Christofides (2013) as well as Tian and Tang (2014) give overviews). One such method that has recently been published is the diagonal technique (DT), see Groenitz (2014a, 2014b). We review the DT in Section 2. Groenitz (2014a, 2014b) considers only one attribute of interest. In this paper, we now address multiple variables $X_1, \ldots, X_v$. In the first part of the article (Section 3), we assume that each variable is gathered by the DT. In the second part (Section 4), we address surveys that involve direct questions and questions according to the DT. In both parts, we consider the maximum likelihood (ML) estimator for the joint distribution of the v variables and its asymptotic variance. The ML estimator depends on the setup or configuration of the survey procedure. A setup means certain parameters that must be determined. We conduct a decision-theoretic analysis and derive risk-optimal

*E-mail: groenitz@staff.uni-marburg.de

ML estimators, or equivalently, optimal setups for the survey procedure. The special feature of our decision-theoretic investigation is to involve the degree of privacy protection (DPP) offered by a setup of the survey procedure. In detail, for the class of ML estimators corresponding to a given DPP, we search the one with the lowest risk, that is, with the best accuracy.

2. INDIRECT INFORMATION BY DIAGONAL TECHNIQUE

Groenitz (2014a) proposes a questioning technique for surveys, namely the diagonal technique, to obtain indirect information on a categorical variable X_1. The DT is beneficial when X_1 is a sensitive variable, which is difficult to gather by a direct question. For example, X_1 may describe the preferred political party (i.e., political preferences), income, the number how often a person has conducted insurance fraud or the earnings from undeclared work (in each case, recorded in classes). The idea of the DT is that the interviewees do not tell the X_1 value, but give an indirect answer A_1, which does not imply the X_1 value and, thus, protects the respondent's privacy. The following definition fixes more details on the DT.

Definition 1. Let X_1 and W_1 be characteristics with possible values $1, \ldots, k_1$. Set

$$A_1 = [(W_1 - X_1) \bmod k_1] + 1, \tag{1}$$

define the vector $c_1 = (c_{11}, \ldots, c_{1,k_1}) = \big(\mathbb{P}(W_1 = 1), \ldots, \mathbb{P}(W_1 = k_1)\big)$ and the matrix $C_1 = [\mathbb{P}(A_1 = p \mid X_1 = q)]_{p,q=1,\ldots,k_1}$. A questioning procedure for a survey in which each respondent is instructed to give the response A_1 and not to reveal the values of X_1 and W_1 is called "diagonal technique with setup c_1" or "diagonal technique with coupling matrix C_1" if:

- X_1 and W_1 are independent.
- W_1 possesses a known distribution.
- The matrix C_1 is invertible.

For the DT with coupling matrix C_1, each row of the matrix C_1 is a left-cyclic shift of the row above and the first row of C_1 equals c_1. The matrix C_1 couples the distributions of A_1 and X_1, in particular, we have $\big(\mathbb{P}(A_1 = 1), \ldots, \mathbb{P}(A_1 = k_1)\big)^\top = C \cdot \big(\mathbb{P}(X_1 = 1), \ldots, \mathbb{P}(X_1 = k_1)\big)^\top$. This explains the name coupling matrix.

We give an example for the DT: Let X_1 describe the monthly earnings from undeclared work, where $X_1 = 1$, $X_1 = 2$, $X_1 = 3$, and $X_1 = 4$ represent earnings in the amount of 0, 1–50, 51–200, and more than 200 Euro, respectively. Furthermore, let W_1 be based on the number formed by the last three digits of the interviewee's telephone number. If this number is ≤ 624, 625–749, 750–874, 875–999, we set $W_1 = 1$, $W_1 = 2$, $W_1 = 3$, $W_1 = 4$, respectively. For instance, telephone number 9478648 results in number 648 and $W_1 = 2$. It is reasonable to assume that X_1 and W_1 are independent and that the distribution of W_1 is given by $\big(\mathbb{P}(W_1 = 1), \ldots, \mathbb{P}(W_1 = 4)\big) = (0.625,\ 0.125,\ 0.125,\ 0.125)$. Moreover, let us assume that each respondent is instructed to provide an indirect answer A_1 according to Table 1.

Table 1. Table of the required indirect answer A_1

$X_1 \backslash W_1$	$W_1 = 1$	$W_1 = 2$	$W_1 = 3$	$W_1 = 4$
$X_1 = 1$	1	2	3	4
$X_1 = 2$	4	1	2	3
$X_1 = 3$	3	4	1	2
$X_1 = 4$	2	3	4	1

Then, we have

$$C_1 = \begin{pmatrix} 0.625 & 0.125 & 0.125 & 0.125 \\ 0.125 & 0.125 & 0.125 & 0.625 \\ 0.125 & 0.125 & 0.625 & 0.125 \\ 0.125 & 0.625 & 0.125 & 0.125 \end{pmatrix}$$

and the questioning procedure is a DT with coupling matrix C_1. Table 1 illustrates that for every scrambled response A_1, all X_1 values are still possible, that is, a participant does not reveal the outcome of X_1 and his or her privacy is protected. Thus, for sensitive X_1, we can expect a higher cooperation of the interviewees compared to direct questions on X_1. In this context, higher cooperation means less answer refusal and less untruthful answers. Based on the indirect answers A_1 of many persons in a sample, we can estimate the distribution of X_1 as shown in Groenitz (2014a, 2014b).

3. THE ITERATIVE DIAGONAL TECHNIQUE

3.1. Definition of Iterative Diagonal Technique

We consider v categorical attributes $X_1, \ldots, X_v$ and assume that indirect data on each attribute are gathered by the DT from Section 2. That is, the DT is applied iteratively. This motivates the term iterative diagonal technique as specified in the following definition.

Definition 2. Let $X_1, \ldots, X_v$ be characteristics, where X_i has the possible values $1, \ldots, k_i$ $(i = 1, \ldots, v)$. Let $W_1, \ldots, W_v$ be further characteristics, where W_i also has the possible values $1, \ldots, k_i$, set $A_i = [(W_i - X_i) \bmod k_i] + 1$, define $c_i = (c_{i1}, \ldots, c_{i,k_i}) = \big(\mathbb{P}(W_i = 1), \ldots, \mathbb{P}(W_i = k_i)\big)$ and $C_i = [\mathbb{P}(A_i = p \mid X_i = q)]_{p,q=1,\ldots,k_i}$ $(i = 1, \ldots, v)$. A questioning procedure for a survey in which each respondent is instructed to give one of the responses $A_1, \ldots, A_v$ and not to reveal the values of $X_1, \ldots, X_v$ and $W_1, \ldots, W_v$ is called “iterative diagonal technique with setup $c_1, \ldots, c_v$” or “iterative diagonal technique with coupling matrices $C_1, \ldots, C_v$” if:

- The two vectors $(W_1, \ldots, W_v)$ and $(X_1, \ldots, X_v)$ are independent.
- The v characteristics $W_1, \ldots, W_v$ are independent.
- W_i possesses a known distribution $(i = 1, \ldots, v)$.
- The matrix C_i is invertible $(i = 1, \ldots, v)$.

3.2. Maximum Likelihood Estimators

In this subsection, we address the ML estimation for the joint distribution of v attributes $X_1, \ldots, X_v$ gathered by the iterative DT. We define for $i_1 = 1, \ldots, k_1, \ldots, i_v = 1, \ldots, k_v$

$$\pi_{i_1,\ldots,i_v} = \mathbb{P}(X_1 = i_1, \ldots, X_v = i_v) \qquad \text{and} \qquad \lambda_{i_1,\ldots,i_v} = \mathbb{P}(A_1 = i_1, \ldots, A_v = i_v).$$

Let π be the vector of length $k = \prod_{j=1}^{v} k_j$ whose entries are the $\pi_{i_1,\ldots,i_v}$ $(i_1 = 1, \ldots, k_1, \ldots, i_v = 1, \ldots, k_v)$, where the entries are sorted first by index i_1, then by index i_2 and so on. Similarly to π, we define λ, which contains the entries $\lambda_{i_1,\ldots,i_v}$. For instance, for $k_1 = k_2 = 2$, $k_3 = 3$, we have

$$\pi = \Big(\pi_{1,1,1},\ \pi_{1,1,2},\ \pi_{1,1,3},\ \pi_{1,2,1},\ \pi_{1,2,2},\ \pi_{1,2,3},\ \pi_{2,1,1},\ \pi_{2,1,2},\ \pi_{2,1,3},\ \pi_{2,2,1},\ \pi_{2,2,2},\ \pi_{2,2,3}\Big)^\top .$$

In the sequel, we need the Kronecker matrix product, denoted by the symbol $\otimes$. The Kronecker product of two matrices $R \in \mathbb{R}^{r_1 \times r_2}$ and $S \in \mathbb{R}^{s_1 \times s_2}$ is given by

$$R \otimes S = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1,r_2} \\ R_{21} & R_{22} & \cdots & R_{2,r_2} \\ \vdots & \vdots & & \vdots \\ R_{r_1,1} & R_{r_1,2} & \cdots & R_{r_1,r_2} \end{pmatrix} \otimes S = \left(\begin{array}{c|c|c|c} R_{11} \cdot S & R_{12} \cdot S & \cdots & R_{1,r_2} \cdot S \\ \hline R_{21} \cdot S & R_{22} \cdot S & \cdots & R_{2,r_2} \cdot S \\ \hline \vdots & \vdots & & \vdots \\ \hline R_{r_1,1} \cdot S & R_{r_1,2} \cdot S & \cdots & R_{r_1,r_2} \cdot S \end{array}\right),$$

that is, $R \otimes S$ is a matrix of size $r_1 s_1 \times r_2 s_2$. With this, we have the following coupling of the distributions of $(A_1, \ldots, A_v)$ and $(X_1, \ldots, X_v)$ for an iterative DT.

Theorem 1. *For an iterative DT with coupling matrices $C_1, \ldots, C_v$, we introduce the matrix $C = C_1 \otimes C_2 \otimes \ldots \otimes C_v$ and have*

$$\lambda = C \cdot \pi. \tag{2}$$

Proof. For $l = 1, \ldots, v$, the characteristic A_l is a function of X_l and W_l. In particular, $A_l = f_l(X_l, W_l)$ with $f_l(x, w) = [(w - x) \bmod k_l] + 1$ for $x, w \in \{1, \ldots, k_l\}$ holds. Then, we have

$$\begin{aligned}
\mathbb{P}(A_1 = a_1, \ldots, A_v = a_v) &= \sum_{i_1,\ldots,i_v} \mathbb{P}\Big(\bigcap_{l=1}^{v} \{A_l = a_l\} \mid \bigcap_{m=1}^{v} \{X_m = i_m\}\Big) \cdot \pi_{i_1,\ldots,i_v} \\
&= \sum_{i_1,\ldots,i_v} \mathbb{P}\Big(\bigcap_{l=1}^{v} \{f_l(X_l, W_l) = a_l\} \mid \bigcap_{m=1}^{v} \{X_m = i_m\}\Big) \cdot \pi_{i_1,\ldots,i_v} \\
&= \sum_{i_1,\ldots,i_v} \mathbb{P}\Big(\bigcap_{l=1}^{v} \{f_l(i_l, W_l) = a_l\} \mid \bigcap_{m=1}^{v} \{X_m = i_m\}\Big) \cdot \pi_{i_1,\ldots,i_v} \\
&= \sum_{i_1,\ldots,i_v} \mathbb{P}\Big(\bigcap_{l=1}^{v} \{f_l(i_l, W_l) = a_l\}\Big) \cdot \pi_{i_1,\ldots,i_v} \\
&= \sum_{i_1,\ldots,i_v} \prod_{l=1}^{v} \mathbb{P}\big(f_l(i_l, W_l) = a_l\big) \cdot \pi_{i_1,\ldots,i_v} \\
&= \sum_{i_1,\ldots,i_v} \prod_{l=1}^{v} \mathbb{P}\big(f_l(X_l, W_l) = a_l \mid X_l = i_l\big) \cdot \pi_{i_1,\ldots,i_v} \\
&= \sum_{i_1,\ldots,i_v} \prod_{l=1}^{v} \mathbb{P}\big(A_l = a_l \mid X_l = i_l\big) \cdot \pi_{i_1,\ldots,i_v} \\
&= \sum_{i_1,\ldots,i_v} \prod_{l=1}^{v} C_l(a_l, i_l) \cdot \pi_{i_1,\ldots,i_v} \\
&= \sum_{i_1,\ldots,i_v} C_1(a_1, i_1) \cdots C_v(a_v, i_v) \cdot \pi_{i_1,\ldots,i_v},
\end{aligned}$$

where $C_l(p, q)$ is the entry (p, q) of the matrix $C_l = [\mathbb{P}(A_l = p \mid X_l = q)]_{p,q=1,\ldots,k_l}$. The last equation implies (2). □

Theorem 1 is the starting point for estimates for the joint distribution given by π. Let us consider a sample of n respondents drawn by simple random sampling with replacement (SRSWR), let $n_{i_1,\ldots,i_v}$ be the number of persons in the sample giving answers $A_1 = i_1, \ldots, A_v = i_v$, and collect these numbers in a column vector $\tilde{n}$ where the entries are sorted first by i_1, second by i_2 and so on. Set

$\hat{\lambda}_{i_1,\ldots,i_v} = n_{i_1,\ldots,i_v}/n$, that is, $\hat{\lambda}_{i_1,\ldots,i_v}$ is the relative frequency of persons in the sample responding $A_1 = i_1, \ldots, A_v = i_v$. Further, set $\hat{\lambda} = \tilde{n}/n$, i.e., $\hat{\lambda}$ is arranged similarly to λ.

Theorem 2. (a) *The estimator*

$$\tilde{\pi} = C^{-1} \cdot \hat{\lambda} = (C_1 \otimes C_2 \otimes \ldots \otimes C_v)^{-1} \cdot \hat{\lambda} = (C_1^{-1} \otimes C_2^{-1} \otimes \ldots \otimes C_v^{-1}) \cdot \hat{\lambda}$$

is unbiased for π.

(b) *If* $\tilde{\pi}$ *has all components in* $[0,1]$*, it is equal to an ML estimator.*

Proof. (a) We clearly have $\mathbb{E}(\tilde{\pi}) = C^{-1} \cdot n^{-1} \cdot \mathbb{E}(\tilde{n}) = C^{-1} \cdot n^{-1} \cdot n \cdot \lambda = \pi$.

(b) Define the set

$$D_k = \{(x_1, \ldots, x_k)^\top : x_i \in [0,1] \text{ and } x_1 + \cdots + x_k = 1\} \tag{3}$$

and the functions $u(x) = C \cdot x$ ($x \in D_k$) and $v(y) = \tilde{n}^\top \cdot \log(y)$ ($y \in D_k$). Here, log is applied componentwise. Notice that the function v possesses the maximum at $\tilde{n}/n = \hat{\lambda}$. The log-likelihood function for π is given by $l(\pi) = v(u(\pi))$ for $\pi \in D_k$. Let us assume that $\tilde{\pi} \in D_k$ holds, i.e., $\tilde{\pi}$ has all components in $[0,1]$. For an arbitrary $\pi' \in D_k$, we then have

$$l(\pi') = v(u(\pi')) \leq v(\hat{\lambda}) = v(C \cdot C^{-1} \cdot \hat{\lambda}) = v(u(C^{-1} \cdot \hat{\lambda})) = l(\tilde{\pi}).$$

□

Remark 1. It may occur that $\tilde{\pi}$ possesses components outside [0,1]. Then, the maximization of the log-likelihood for π is typically conducted by iterative methods.

The next theorem addresses the asymptotic distribution of the ML estimator for our target quantity π, especially its asymptotic variance.

Theorem 3. *Let each component of* π *be in the open interval* $(0,1)$*. For the ML estimator* $\hat{\pi}$ *for* π*, we then have the asymptotic normality*

$$\sqrt{n} \cdot (\hat{\pi} - \pi) \quad \xrightarrow{d} \quad N(0,\ C^{-1} \cdot (\mathrm{diag}(\lambda) - \lambda\lambda^\top) \cdot C^{-1}) \qquad \textit{for} \quad n \to \infty.$$

Proof. Due to (2), λ and π are coupled by a matrix $C = C_1 \otimes \ldots \otimes C_v$. In this proof, it is convenient to reindex the components of π, λ and $\tilde{n}$ by counting from $1, \ldots, k$ with $k = \prod_{i=1}^{v} k_i$. That is, e.g., $\pi = (\pi_1, \ldots, \pi_k)^\top$ and $\tilde{n} = (n_1, \ldots, n_k)^\top$ hold. For $\tilde{\pi}$ from Theorem 2 as well as D_k from (3), we first show

$$\mathbb{P}(\tilde{\pi} \notin D_k) \longrightarrow 0 \quad (n \to \infty). \tag{4}$$

The weak law of large numbers implies the stochastic convergence $\hat{\lambda} = \tilde{n}/n \xrightarrow{\mathbb{P}} \lambda$. It follows that $\tilde{\pi} = C^{-1}\hat{\lambda} \xrightarrow{\mathbb{P}} C^{-1}\lambda = \pi$. For $\varepsilon > 0$, we set

$$U_\varepsilon\big((\pi_1, \ldots, \pi_{k-1})^\top\big) = \{(x_1, \ldots, x_{k-1})^\top \in \mathbb{R}^{k-1} \colon \|(x_1, \ldots, x_{k-1})^\top - (\pi_1, \ldots, \pi_{k-1})^\top\| < \varepsilon\}.$$

We now choose some $\varepsilon > 0$ with

$$U_\varepsilon\big((\pi_1, \ldots, \pi_{k-1})^\top\big) \subset \{(x_1, \ldots, x_{k-1})^\top \in \mathbb{R}^{k-1} \colon x_i \in (0,1), x_1 + \cdots + x_{k-1} < 1\}.$$

Finding such an ε is possible, because we have assumed that each component of π is in $(0,1)$. Then,

$$\begin{aligned}
\mathbb{P}(\tilde{\pi} \notin D_k) &= \mathbb{P}\Big((\tilde{\pi}_1, \ldots, \tilde{\pi}_{k-1})^\top \notin \{(x_1, \ldots, x_{k-1})^\top \in \mathbb{R}^{k-1} \colon x_i \in [0,1], x_1 + \cdots + x_{k-1} \leq 1\}\Big) \\
&\leq \mathbb{P}\Big((\tilde{\pi}_1, \ldots, \tilde{\pi}_{k-1})^\top \notin U_\varepsilon((\pi_1, \ldots, \pi_{k-1})^\top)\Big) \\
&= \mathbb{P}\Big(\|(\tilde{\pi}_1, \ldots, \tilde{\pi}_{k-1})^\top - (\pi_1, \ldots, \pi_{k-1})^\top\| \geq \varepsilon\Big) \longrightarrow 0 \quad (n \to \infty)
\end{aligned}$$

holds due to the convergence of $\tilde{\pi}$ to π in probability. That is, we have (4). By a multivariate central limit theorem (see, for instance, Borovkov (2013, p. 214, Corollary 8.6.1)), we obtain $\sqrt{n}\cdot(\hat{\lambda}-\lambda)\xrightarrow{d} N\big(0,\operatorname{diag}(\lambda)-\lambda\lambda^\top\big)$. By a continuous mapping theorem,

$$C^{-1}\cdot\sqrt{n}\cdot(\hat{\lambda}-\lambda)=\sqrt{n}\cdot(\tilde{\pi}-\pi)$$
$$\xrightarrow{d}\quad C^{-1}\cdot N(0,\operatorname{diag}(\lambda)-\lambda\lambda^\top)=N(0,C^{-1}\cdot(\operatorname{diag}(\lambda)-\lambda\lambda^\top)\cdot C^{-1})$$

follows. We now demonstrate that $\sqrt{n}\cdot(\hat{\pi}-\pi)-\sqrt{n}\cdot(\tilde{\pi}-\pi)$ converges in probability to 0. For every $\varepsilon>0$, it is true that

$$\mathbb{P}\Big(\|\sqrt{n}\cdot(\hat{\pi}-\pi)-\sqrt{n}\cdot(\tilde{\pi}-\pi)\|\geq\varepsilon\Big)$$
$$=\mathbb{P}\Big(\|\sqrt{n}\cdot(\hat{\pi}-\tilde{\pi})\|\geq\varepsilon\Big)\leq\mathbb{P}(\tilde{\pi}\notin D_k)\longrightarrow 0\quad(n\to\infty)$$

according to Theorems 2(b) and (4). Since $\sqrt{n}\cdot(\tilde{\pi}-\pi)\xrightarrow{d} N\big(0,C^{-1}\cdot(\operatorname{diag}(\lambda)-\lambda\lambda^\top)\cdot C^{-1}\big)$ and $\sqrt{n}\cdot(\hat{\pi}-\pi)-\sqrt{n}(\tilde{\pi}-\pi)\xrightarrow{\mathbb{P}}0$, we also have

$$\sqrt{n}\cdot(\hat{\pi}-\pi)\xrightarrow{d} N(0,C^{-1}\cdot(\operatorname{diag}(\lambda)-\lambda\lambda^\top)\cdot C^{-1}).$$

□

3.3. Risk-Optimal Estimators for Iterative Diagonal Technique

In this subsection, we derive optimal ML estimators, or equivalently optimal setups, for the iterative DT by a decision-theoretic analysis. As already stated in Definition 2, a setup is given by the v vectors $c_1,\ldots,c_v$ with $c_i=(c_{i1},\ldots,c_{i,k_i})=(\mathbb{P}(W_i=1),\ldots,\mathbb{P}(W_i=k_i))$, i.e., by the distributions of the auxiliary characteristics $W_1,\ldots,W_v$. The special feature of our decision-theoretic analysis will be the incorporation of the degree of privacy protection (DPP) provided by a setup of the iterative DT. In the class of ML estimators corresponding to a fixed DPP, we search an estimator with the lowest risk, i.e., with the highest quality. We measure the DPP as follows.

Definition 3. Let us consider the iterative DT with setup $c_1,\ldots,c_v$. The DPP for the ith question ($i=1,\ldots,v$) is quantified by the empirical standard deviation of the vector $c_i=(c_{i1},\ldots,c_{i,k_i})$ given by

$$\sigma_i=\sigma(c_i)=\operatorname{std}(c_i)=\sqrt{(c_{i1}^2+\cdots+c_{i,k_i}^2)/(k_i-1)-[(k_i-1)\cdot k_i]^{-1}}\in[0,\sqrt{1/k_i}].$$

The overall DPP is then measured by the vector $(\sigma_1,\ldots,\sigma_v)$.

A small σ_i indicates that the distribution of the auxiliary attribute W_i is close to a uniform distribution. In this case, the privacy is protected much, because for a uniformly distributed W_i, X_i and A_i are independent. A large value of σ_i indicates that the distribution of W_i is close to a degenerate distribution, i.e., W_i is nearly constant. In this case, the privacy is protected sparsely, because for a constant W_i, the answer A_i implies the X_i value.

The ML estimator $\hat{\pi}$ for π indeed depends on the setup of the iterative DT, thus we write $\hat{\pi}(c_1,\ldots,c_v)$ instead of $\hat{\pi}$ on occasion. With the loss function

$$(\pi,\hat{\pi})\mapsto\operatorname{tr}\big((\hat{\pi}-\pi)(\hat{\pi}-\pi)^\top\big),$$

the risk of the ML estimator would be

$$\mathbb{E}\big(\operatorname{tr}((\hat{\pi}(c_1,\ldots,c_v)-\pi)\cdot(\hat{\pi}(c_1,\ldots,c_v)-\pi)^\top)\big)=\operatorname{tr}\big(\operatorname{MSE}(\hat{\pi}(c_1,\ldots,c_v))\big).$$

Since we have not derived a general explicit form of the ML estimator (cf. Theorem 2 and Remark 1), it is more convenient to work with the asymptotic risk motivated by Theorem 3:

$$R_{\hat{\pi}(c_1,\ldots,c_v)}(\pi):=\operatorname{tr}\big(n^{-1}\cdot C^{-1}\cdot(\operatorname{diag}(\lambda)-\lambda\lambda^\top)\cdot C^{-1}\big).\tag{5}$$

In the class $\{\hat{\pi}(c_1,\ldots,c_v)\colon\ \operatorname{std}(c_1)=\sigma_1,\ldots,\operatorname{std}(c_v)=\sigma_v\}$, where $\sigma_1,\ldots,\sigma_v$ are given, that is, in the class of ML estimators corresponding to a certain DPP, we now search an estimator with the smallest asymptotic risk (5). According to the following theorem, the asymptotic risk has a lower bound that depends on the DPP $(\sigma_1,\ldots,\sigma_v)$.

Theorem 4. *We have the inequality*

$$n \cdot R_{\hat{\pi}(c_1,\ldots,c_v)}(\pi) \geq \prod_{i=1}^{v} \left(\frac{(k_i - 1)(k_i^{-1} - \sigma_i^2)}{k_i \sigma_i^2} + 1 \right) - \operatorname{tr}(\pi\pi^\top). \tag{6}$$

Proof. Define $D_i = C_i^{-1}$ $(i = 1, \ldots, v)$ and $D = C^{-1}$. We have $D = D_1 \otimes \cdots \otimes D_v$. Since each row of C_i is a left-cyclic shift of the row above, each row of D_i is also a left-cyclic shift of the row above. Let $(d_{i1}, \ldots, d_{i,k_i})$ be the first row of D_i and let $(d_1, \ldots, d_k)$ be the first row of D. It is true that

$$C^{-1} \cdot \big(\operatorname{diag}(\lambda) - \lambda\lambda^\top \big) \cdot C^{-1} = D \cdot \operatorname{diag}(\lambda) \cdot D - \pi\pi^\top.$$

Each row of D contains the same entries (the order of the entries differs from row to row). The entries are the products $d_{1,j_1} \cdots d_{v,j_v}$ $(j_1 = 1, \ldots, k_1; \ldots; j_v = 1, \ldots, k_v)$. Consequently,

$$\operatorname{tr}\big(D \cdot \operatorname{diag}(\lambda) \cdot D\big) = d_1^2 + \cdots + d_k^2$$

follows. Now, define $\alpha_{i1}, \ldots, \alpha_{i,k_i}$ to be the eigenvalues of C_i and $\psi_1, \ldots, \psi_k$ the eigenvalues of C. Furthermore, let $\|\cdot\|_F$ be the Frobenius norm of a matrix. It is true that

$$d_1^2 + \cdots + d_k^2 = k^{-1} \cdot \|D\|_F^2 = k^{-1} \cdot (\psi_1^{-2} + \cdots + \psi_k^{-2}) = k^{-1} \cdot \prod_{i=1}^{v} (\alpha_{i1}^{-2} + \cdots + \alpha_{i,k_i}^{-2}),$$

where the last equality holds since the eigenvalues of the Kronecker product $C_1 \otimes \cdots \otimes C_v$ are given by the products $\alpha_{1,i_1} \cdots \alpha_{v,i_v}$ $(i_1 = 1, \ldots, k_1, \ldots, i_v = 1, \ldots, k_v)$. It follows that

$$n \cdot R_{\hat{\pi}(c_1,\ldots,c_v)}(\pi) = k^{-1} \cdot \prod_{i=1}^{v} (\alpha_{i1}^{-2} + \cdots + \alpha_{i,k_i}^{-2}) - \operatorname{tr}(\pi\pi^\top). \tag{7}$$

We further have for $i = 1, \ldots, v$ the identities

$$\begin{aligned} \sigma_i^2 &= (k_i - 1)^{-1} \cdot k_i^{-1} \cdot [k_i \cdot (c_{i1}^2 + \cdots + c_{i,k_i}^2) - 1] \\ &= (k_i - 1)^{-1} \cdot k_i^{-1} \cdot [\|C_i\|_F^2 - 1] = (k_i - 1)^{-1} \cdot k_i^{-1} \cdot [\alpha_{i1}^2 + \cdots + \alpha_{i,k_i}^2 - 1]. \end{aligned} \tag{8}$$

Note, that 1 is an eigenvalue of C_i (with eigenvector $(1, \ldots, 1)^\top$), say $\alpha_{i,k_i} = 1$. We set $q_i = k_i - 1$. To prove (6), we have to minimize the function

$$\begin{aligned} f(\alpha_{11}, \ldots, \alpha_{1,q_1}, \ldots, \alpha_{v1}, \ldots, \alpha_{v,q_v}) &= \prod_{i=1}^{v} k_i^{-1} (\alpha_{i,1}^{-2} + \cdots + \alpha_{i,q_i}^{-2} + 1) \\ &=: \prod_{i=1}^{v} k_i^{-1} \cdot g_i(\alpha_{i1}, \ldots, \alpha_{i,q_i}) \end{aligned}$$

under the v restrictions

$$r_i(\alpha_{i1}, \ldots, \alpha_{i,q_i}) = (k_i - 1)^{-1} \cdot k_i^{-1} \cdot [\alpha_{i1}^2 + \cdots + \alpha_{i,q_i}^2] - \sigma_1^2 = 0 \quad (i = 1, \ldots, v).$$

This is a separated minimization problem, that is, it suffices to minimize each g_i under

$$r_i(\alpha_{i1}, \ldots, \alpha_{i,q_i}) = 0.$$

Each single minimization can be conducted with Lagrange multipliers. We obtain

$$f(\alpha_{11}, \ldots, \alpha_{1,q_1}, \ldots, \alpha_{v1}, \ldots, \alpha_{v,q_v}) \geq \prod_{i=1}^{v} \left(\frac{(k_i - 1)(k_i^{-1} - \sigma_i^2)}{k_i \sigma_i^2} + 1 \right).$$

Consequently, (6) holds. □

The next theorem shows how to find a setup of iterative DT that leads to a given DPP and to the lower bound of Theorem 4.

Theorem 5. *Let* $(\sigma_1, \dots, \sigma_v)$ *be a chosen DPP and set for* $i = 1, \dots, v$

$$c_{i1} = k_i^{-1} + (k_i - 1)/k_i \cdot \sqrt{k_i \cdot \sigma_i^2}, \qquad c_{i2} = \dots = c_{i,k_i} = k_i^{-1} - k_i^{-1} \cdot \sqrt{k_i \cdot \sigma_i^2}. \tag{9}$$

Then, $\mathrm{std}(c_i) = \sigma_i$ *holds for every* $i = 1, \dots, v$ *and* $n \cdot R_{\hat{\pi}(c_1,\dots,c_v)}(\pi)$ *equals the right side of* (6).

Proof. The rows of C_i^2 arise by right-cyclic shift of the row above, i.e., C_i^2 is a circulant matrix. Say, the first row of C_i^2 is $(x_{i1}, \dots, x_{i,k_i})$. By some calculation, we can show $x_{i1} = k_i^{-1}(1 + (k_i - 1)k_i\sigma_i^2)$ and $x_{i2} = \dots = x_{i,k_i} = k_i^{-1}(1 - k_i\sigma_i^2)$. Since C_i^2 is a circulant matrix, its eigenvalues can be computed by discrete Fourier transform of $(x_{i1}, \dots, x_{i,k_i})$, see Gray (2006, Chapter 3). That is, to obtain the eigenvalues of C_i^2, we can calculate $\sum_{m=1}^{k_i} \exp(-2\pi \mathrm{i} \cdot \frac{(l-1)(m-1)}{k_i}) \cdot x_{im}$ for $l = 1, \dots, k_i$. We obtain eigenvalue 1 and eigenvalue $k_i\sigma_i^2$ (with multiplicity $k_i - 1$). The application of (7) and (8) completes the proof. □

Due to the Theorems 4 and 5, the ML estimator corresponding to the setup (9) possesses the lowest asymptotic risk among the estimators in the class $\{\hat{\pi}(c_1, \dots, c_v)\colon \mathrm{std}(c_1) = \sigma_1, \dots, \mathrm{std}(c_v) = \sigma_v\}$. Thus, for a given DPP $(\sigma_1, \dots, \sigma_v)$, the setup (9) is an optimal setup of the iterative DT. In other words, for the given DPP $(\sigma_1, \dots, \sigma_v)$, the choice of the configuration (9) is the best decision.

Figure 1 illustrates the findings of this subsection graphically. For this figure, we set $k_1 = 3$, $k_2 = 4$ and fix a certain π. For the left part of this figure, we then randomly select a large number of vectors c_1 and c_2 such that these vectors are preferably uniformly scattered. For each pair c_1 and c_2, we compute the corresponding DPP (σ_1, σ_2) and $n \cdot R_{\hat{\pi}(c_1,\dots,c_v)}(\pi)$. This results in the shown point cloud. This point cloud has a lower bound depending on (σ_1, σ_2). The lower bound is depicted in the right part of Fig. 1. Moreover, each bold point in the right part of Fig.1 belongs to an optimal setup c_1, c_2 according to (9) for a given DPP (σ_1, σ_2). These bold points are located exactly on the lower bound of the point cloud of the randomly drawn vectors.

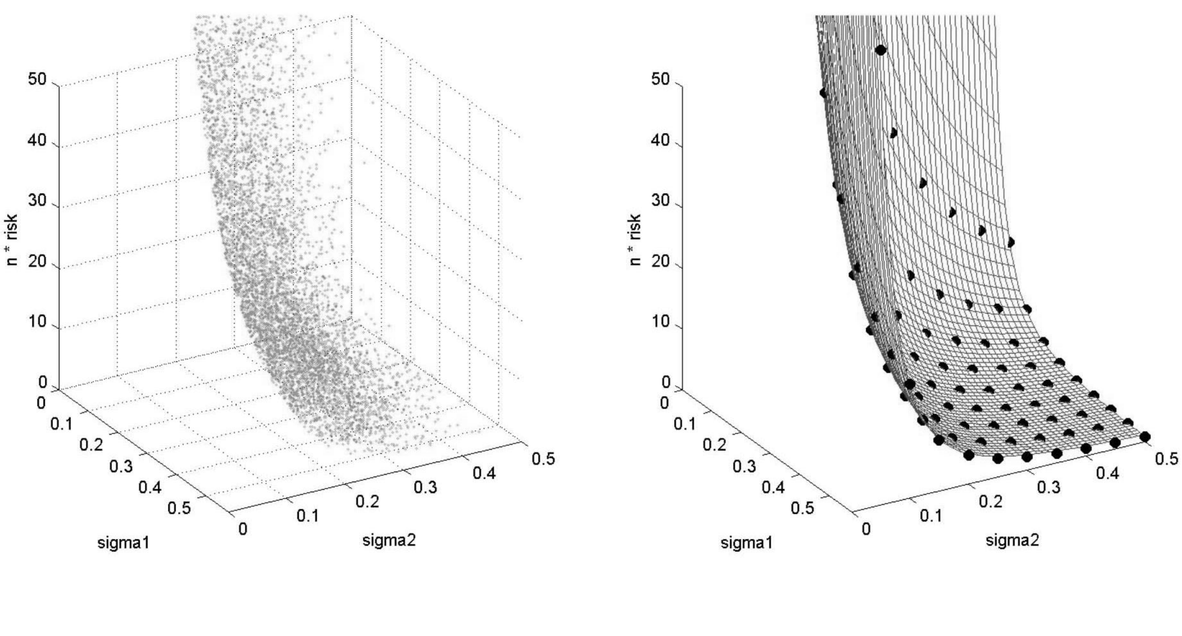

Fig. 1. Illustration of optimal and nonoptimal setups for the iterative DT

4. COMBINATIONS OF DIRECT QUESTIONS AND DT

We consider v categorical characteristics $X_1, \dots, X_v$, where X_i should have the possible values $1, \dots, k_i$. That is, there are overall $k = k_1 \cdots k_v$ categories. Among the v characteristics, the first v' attributes $X_1, \dots, X_{v'}$ are assumed to be gathered by ordinary direct questioning, while the remaining attributes $X_{v'+1}, \dots, X_v$ are assumed to be gathered by the DT. In practice, $X_1, \dots, X_{v'}$ will be non-sensitive characteristics (e.g., age, gender) and the $X_{v'+1}, \dots, X_v$ will concern sensitive stigmatizing characteristics (e.g., tax evasion, social fraud, income). Let A_i be the ith response of an interviewee ($i = 1, \dots, v$). Then, for $i = 1, \dots, v'$, $A_i = X_i$ holds. For $i = v' + 1, \dots, v$, the answer A_i is a function of X_i and an auxiliary attribute W_i, namely $A_i = [(W_i - X_i) \bmod k_i] + 1$.

4.1. Assumptions and Maximum Likelihood Estimation

In this section, we make the following assumptions:

– The two vectors $(W_{v'+1}, \dots, W_v)$ and $(X_1, \dots, X_v)$ are independent. (10)

– The $v - v'$ attributes $W_{v'+1}, \dots, W_v$ are independent. (11)

– The characteristic W_i has a known distribution ($i = v' + 1, \dots, v$). (12)

– The matrix $C_i = [\mathbb{P}(A_i = p \mid X_i = q)]_{p,q=1,\dots,k_i}$ is invertible ($i = v' + 1, \dots, v$). (13)

We define π and λ similarly to Section 3 and denote the $l \times l$ identity matrix by I_l. For our combination of direct questions and the DT, the coupling between the vectors λ and π is presented in the following theorem.

Theorem 6. *Under the assumptions* (10) *and* (11)*, we have*

$$\lambda = \left(I_{k_1} \otimes \dots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \dots \otimes C_v\right) \cdot \pi. \tag{14}$$

Proof. The claim can be proved by steps similar to the proof of Theorem 1. □

For the estimation, assume again that n persons are selected by SRSWR. We define $n_{i_1,\dots,i_v}$ and $\hat{\lambda}_{i_1,\dots,i_v}$ as well as the vectors $\tilde{n}$ and $\hat{\lambda}$ similarly to Section 3. For example, $n_{i_1,\dots,i_v}$ is the number of sample units giving answers $A_1 = i_1, \dots, A_v = i_v$. The next theorem shows an important representation of the log-likelihood corresponding to our current situation.

Theorem 7. *For our combination of v' times direct questioning and $v - v'$ times questioning according to the DT, the log-likelihood is given by*

$$\begin{aligned} l(\pi) &= \tilde{n}^\top \cdot \log\left[\left(I_{k_1} \otimes \dots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \dots \otimes C_v\right) \cdot \pi\right] \\ &= {n^*}^\top \cdot \log\left[S_{k_1} \otimes \dots \otimes S_{k_{v'}} \otimes C_{v'+1} \otimes \dots \otimes C_v) \cdot \pi\right], \end{aligned} \tag{15}$$

where $n^* = (S_{k_1} \otimes \dots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \dots \otimes I_{k_v}) \cdot \tilde{n}$ and S_l is an $l \times l$ matrix whose entry (p, q) is equal to 1 *if $p = q = 1$ or $p + q = l + 2$ and equal to* 0 *otherwise.*

Proof. The first equality follows directly from the definition of the log-likelihood and (14). For the second equality, notice that multiplying a column vector from the left with the matrix $S_{k_1} \otimes \dots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \dots \otimes I_{k_v}$ causes a permutation of the column vector's elements. With this and a computation rule for Kronecker products, we can write

$$\begin{aligned} l(\pi) &= [S_{k_1} \otimes \dots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \dots \otimes I_{k_v} \cdot \tilde{n}]^\top \\ &\quad \times \log\left[(S_{k_1} \otimes \dots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \dots \otimes I_{k_v}) \cdot (I_{k_1} \otimes \dots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \dots \otimes C_v) \cdot \pi\right] \\ &= {n^*}^\top \cdot \log\left[(S_{k_1} \otimes \dots \otimes S_{k_{v'}} \otimes C_{v'+1} \otimes \dots \otimes C_v) \cdot \pi\right]. \end{aligned}$$

□

Remark 2. The function (15) from Theorem 7 equals the log-likelihood function corresponding to the iterative DT with coupling matrices $S_{k_1}, \dots, S_{k_{v'}}, C_{v'+1}, \dots, C_v$ according to Definition 2 and observed absolute answer frequencies n^*.

According to this Remark 2, the ML estimation for the combination of direct questions and the DT can be traced back to the ML estimation for the iterative DT from Section 3. We now come to the asymptotic normality of the ML estimator for the combination of direct questions and the DT.

Theorem 8. *Consider our combination of v' times direct questioning and $v - v'$ times questioning according to the DT and let each component of π be located in the open interval* $(0, 1)$. *For the ML estimator $\hat{\pi}$ for π, we then have for $n \to \infty$ the asymptotic normality*

$$\sqrt{n} \cdot (\hat{\pi} - \pi) \quad \xrightarrow{d} \quad N(0, \nu) \quad \textit{with}$$

$$\begin{aligned}\nu &= (I_{k_1} \otimes \cdots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1} \cdot (\operatorname{diag}(\lambda) - \lambda\lambda^\top) \\ &\qquad \times (I_{k_1} \otimes \cdots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1} \qquad (16) \\ &= (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1} \cdot (\operatorname{diag}(\lambda^*) - \lambda^* {\lambda^*}^\top) \\ &\qquad \times \left(S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v\right)^{-1} \qquad (17)\end{aligned}$$

with matrices S_l as in Theorem 7 and $\lambda^* = \left(S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}\right) \cdot \lambda$.

Proof. The asymptotic normality with the variance representation (16) can be established with computations similarly to the proof of Theorem 3. For formula (17), note that the product $(S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}) \cdot (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v})$ equals the identity matrix. Further, note that multiplying a diagonal matrix first from the left with $S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}$ and then from the right with $S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}$ yields a permutation of the diagonal elements of the diagonal matrix. Then,

$$\begin{aligned}\nu &= (I_{k_1} \otimes \cdots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1} \\ &\quad \times (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}) \cdot (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}) \\ &\quad \times (\operatorname{diag}(\lambda) - \lambda\lambda^\top) \\ &\quad \times (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}) \cdot (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}) \\ &\quad \times (I_{k_1} \otimes \cdots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1} \\ &= (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1} \\ &\quad \times [(S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}) \cdot \operatorname{diag}(\lambda) \\ &\quad \times (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes I_{k_{v'+1}} \otimes \cdots \otimes I_{k_v}) - \lambda^* {\lambda^*}^\top] \\ &\quad \times (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1} \\ &= (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1} \cdot [\operatorname{diag}(\lambda^*) - \lambda^* {\lambda^*}^\top] \\ &\quad \times (S_{k_1} \otimes \cdots \otimes S_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v)^{-1}\end{aligned}$$

holds. This is exactly the claim. □

Remark 3. The asymptotic variance of the ML estimator for the combination of v' times direct questioning and $v - v'$ questioning according to the DT equals the asymptotic variance of the ML estimator for the iterative DT with coupling matrices $S_{k_1}, \dots, S_{k_{v'}}, C_{v'+1}, \dots, C_v$.

4.2. Risk-Optimal Estimators for Combinations of Direct Questioning and Diagonal Technique

We now establish optimal ML estimators or optimal setups for the combination of direct questions and the DT by decision-theoretic considerations similar to Subsection 3.3. The specifications to adjust are the distributions of $W_{v'+1}, \dots, W_v$, i.e., the probability masses $c_i = (c_{i1}, \dots, c_{i,k_i}) = (\mathbb{P}(W_i = 1), \dots, \mathbb{P}(W_i = k_i))$ for $i = v' + 1, \dots, v$. Therefore, a setup is now given by the vectors $c_{v'+1}, \dots, c_v$. We measure the DPP for gathering X_i $(i = v' + 1, \dots, v)$ again by the empirical standard deviation $\sigma_i = \mathrm{std}(c_i)$. The overall DPP is now the vector $(\sigma_{v'+1}, \dots, \sigma_v)$. As (asymptotic) risk of an ML estimator, we use the trace of its asymptotic variance, that is, according to Theorem 8, we use

$$R_{\hat{\pi}(c_{v'+1},\dots,c_v)}(\pi) = \mathrm{tr}\Big[n^{-1} \cdot \big(I_{k_1} \otimes \cdots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v\big)^{-1} \times \big(\mathrm{diag}(\lambda) - \lambda\lambda^\top\big) \cdot \big(I_{k_1} \otimes \cdots \otimes I_{k_{v'}} \otimes C_{v'+1} \otimes \cdots \otimes C_v\big)^{-1}\Big]. \qquad (18)$$

In the class of ML estimators corresponding to a prefixed DPP $(\sigma_{v'+1}, \dots, \sigma_v)$, i.e., in $\{\hat{\pi}(c_{v'+1}, \dots, c_v)$: $\mathrm{std}(c_{v'+1}) = \sigma_{v'+1}, \dots, \mathrm{std}(c_v) = \sigma_v\}$, we would like to find the one with the lowest risk (18). For this, we present Theorems 9 and 10.

Theorem 9. *For our combination of v' times direct questioning and $v - v'$ times DT questioning, we have*

$$n \cdot R_{\hat{\pi}(c_{v'+1},\dots,c_v)}(\pi) \geq \prod_{i=v'+1}^{v} \left(\frac{(k_i - 1)(k_i^{-1} - \sigma_i^2)}{k_i \sigma_i^2} + 1 \right) - \mathrm{tr}(\pi\pi^\top). \qquad (19)$$

Proof. The asymptotic variance of the ML estimator for our combination of direct questioning and DT questioning equals the asymptotic variance of the ML estimator for the iterative DT with coupling matrices $S_{k_1}, \dots, S_{k_{v'}}, C_{v'+1}, \dots, C_v$ due to Remark 3. For this iterative DT, we obtain a DPP for question i $(i = 1, \dots, v')$ equal to $\sqrt{1/k_i}$. Due to Theorem 4,

$$n \cdot R_{\hat{\pi}(c_{v'+1},\dots,c_v)}(\pi) \geq \prod_{i=1}^{v'} (0 + 1) \cdot \prod_{i=v'+1}^{v} \left(\frac{(k_i - 1)(k_i^{-1} - \sigma_i^2)}{k_i \sigma_i^2} + 1 \right) - \mathrm{tr}(\pi\pi^\top).$$

□

Theorem 10. *Consider a combination of v' times direct questioning and $v - v'$ times DT questioning. Let $(\sigma_{v'+1}, \dots, \sigma_v)$ be chosen. Define for $i = v' + 1, \dots, v$*

$$c_{i1} = k_i^{-1} + (k_i - 1)/k_i \cdot \sqrt{k_i \cdot \sigma_i^2}, \qquad c_{i2} = \cdots = c_{i,k_i} = k_i^{-1} - k_i^{-1} \cdot \sqrt{k_i \cdot \sigma_i^2}. \qquad (20)$$

It follows that $\mathrm{std}(c_i) = \sigma_i$ *for every $i = v' + 1, \dots, v$ and that*

$$n \cdot R_{\hat{\pi}(c_{v'+1},\dots,c_v)}(\pi) = \prod_{i=v'+1}^{v} \left(\frac{(k_i - 1)(k_i^{-1} - \sigma_i^2)}{k_i \sigma_i^2} + 1 \right) - \mathrm{tr}(\pi\pi^\top). \qquad (21)$$

Proof. The asymptotic variance of the ML estimator for the survey procedure of Theorem 10 equals the asymptotic variance of the ML estimator of the iterative DT with coupling matrices $S_{k_1}, \dots, S_{k_{v'}}, C_{v'+1}, \dots, C_v$, where C_i $(i = v' + 1, \dots, v)$ has the first row $c_i = (c_{i1}, \dots, c_{i,k_i})$ from (20). For $i = 1, \dots, v'$, S_{k_i} has the first row $(1, 0, 0, \dots, 0)$. Theorem 5 implies then that $\mathrm{std}(c_i) = \sigma_i$ $(i = v' + 1, \dots, v)$ and that (21) holds. □

Theorems 9 and 10 imply that the ML estimator for the setup (20) is risk-optimal among the ML estimators corresponding to a given DPP $(\sigma_{v'+1}, \dots, \sigma_v)$. In other words, the survey setup given through (20) is the best setup for this DPP.

5. SUMMARY AND CONCLUDING REMARKS

In this article, we have addressed the estimation of the joint distribution of categorical variables based on survey data when the diagonal technique is involved in the survey. First, we have considered the case that data on each characteristic are collected by the DT. Second, we treated combinations of ordinary direct questions and the DT. The second case is probably of more practical meaning than asking only DT questions throughout. The reason is that surveys will typically comprise nonsensitive and sensitive characteristics rather than unexceptional sensitive attributes. For both cases, we have derived optimal ML estimators by decision-theoretic arguments involving the DPP and the asymptotic risk of the estimator. For this point, which is clearly advantageous in comparison with the existing literature on indirect survey procedures, properties of, e.g., circulant matrices, the Frobenius norm, the Kronecker product, and permutation matrices were used. Our optimality results motivate the following strategy for a survey agency: First, select a DPP. Second, select an optimal ML estimator for this DPP via Theorem 5 or Theorem 10. This is equivalent to choosing an optimal setup for the survey procedure. Third, adapt concrete auxiliary characteristics on the optimal setup, e.g., based on birthday periods, telephone numbers, or house numbers.

ACKNOWLEDGMENTS

The author would like to thank the Editor Professor Levit and a referee for their helpful comments and suggestions.

REFERENCES

1. A. A. Borovkov, *Probability Theory* (Springer, 2013).
2. A. Chaudhuri and T. C. Christofides, *Indirect Questioning in Sample Surveys* (Springer, 2013).
3. J. A. Fox and P. E. Tracy, *Randomized Response — A Method for Sensitive Surveys* (Sage, Beverly Hills 1986).
4. R. M. Gray, *Toeplitz and Circulant Matrices: A Review* (Now Publ., 2006).
5. H. Groenitz, "A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables", Metrika **77**, 211–224 (2014a).
6. H. Groenitz, "Applying the Nonrandomized Diagonal Model to Estimate a Sensitive Distribution in Complex Sample Surveys", J. Statist. Theory and Practice **8**, 319–342 (2014b).
7. G. L. Tian and M. L. Tang, *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys* (CRC Press, 2014).