Noname manuscript No.
(will be inserted by the editor)

An Information Theoretic approach to Post Randomization Methods under Differential Privacy

Fadhel Ayed · Marco Battiston · Federico Camerlenghi

Received: date / Accepted: date

Abstract Post Randomization Methods (PRAM) are among the most popular disclosure limitation techniques for both categorical and continuous data. In the categorical case, given a stochastic matrix M and a specified variable, an individual belonging to category i is changed to category j with probability $M_{i,j}$. Every approach to choose the randomization matrix M has to balance between two desiderata: 1) preserving as much statistical information from the raw data as possible; 2) guaranteeing the privacy of individuals in the dataset. This trade-off has generally been shown to be very challenging to solve. In this work, we use recent tools from the computer science literature and propose to choose M as the solution of a constrained maximization problems. Specifically, M is chosen as the solution of a constrained maximization problem, where we maximize the Mutual Information between raw and transformed data, given the constraint that the transformation satisfies the notion of Differential Privacy. For the general Categorical model, it is shown how this maximization problem reduces to a convex linear programming and can be therefore solved with known optimization algorithms.

F. Ayed
Oxford University,
24-29 St Giles', Oxford OX1 3LB, UK.
E-mail: fadhel.ayed@gmail.com

M. Battiston
Lancaster University,
Fylde College, Bailrigg, Lancaster, LA1 4YF, UK.
E-mail: m.battiston@lancaster.ac.uk

F. Camerlenghi
University of Milano - Bicocca,
Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy.
E-mail: federico.camerlenghi@unimib.it
Also affiliated to Collegio Carlo Alberto (Turin, Italy) and BIDSA at Bocconi University (Milan, Italy).

Keywords Post Randomization Methods · Disclosure risk · Mutual Information · Differential Privacy · Categorical Variables.

1 Introduction

Data from census or survey studies are among the most useful sources of information for social and political studies. However, when statistical and governmental agencies release microdata to the public, they often encounter ethical and moral issues concerning the possible privacy leak for individuals present in the dataset. Anonymization techniques, like encrypting or removing personally identifiable information, have been widely used with the hope of ensuring privacy protection. However, recent studies by Gymrek et al. (2013), Homer et al. (2008), Narayanan and Shmatikov (2008), Sweeney (1997) have shown that, even after removing directly identifying variables, like names or national insurance numbers, the potential for breaches of confidentiality is still present. Specifically, an intruder might still be able to identify individuals by cross-classifying categorical variables in the dataset and matching them with some external database. This kind of privacy problems have been widely considered in the statistical literature and different measures of *disclosure risk* have been proposed to assess the riskiness of specific dataset.

Different disclosure limitation techniques have been proposed, like rounding, suppression of extreme values or entire variables, sampling or perturbation techniques. *Post Randomization Methods* (PRAM) are among the most used techniques for disclosure risk limitation. See De Wolf et al. (1997), Gouweleeuw et al. (1997), Kooiman et al. (1997). With these techniques, before releasing the dataset, the data curator randomly changes the values of some categorical identifying variables, like gender, job or age, of some individuals in the dataset. In a recent paper, Shlomo and Skinner (2010) consider PRAM and random data swapping of a geographical variable and propose a way of computing measures of disclosure risk to assess whether these techniques have been effective in "privatizing" the dataset. The choice of the geographical variable is motivated by the fact that, by swapping or changing it, it is usually less likely to generate unreasonable combinations of categorical variables, like for instance a pregnant man or a 10 year old lawyer. In order to implement PRAM, they consider a stochastic matrix M, where the (i, j) entry of this matrix gives the probability that an individual from location i has his geographical variable swapped to location j. Given this known matrix M, Shlomo and Skinner (2010) suggest some measures of risk and related estimation methods. However, an open problem is to decide how the data curator should actually choose the matrix M in order to guarantee an effective level of privacy.

Over the last ten years, a new approach to data protection, called *Differential Privacy* (see Dwork et al. (2006)), has become more and more popular in the computer science literature and has been implemented in their security protocols by IT companies (Eland (2015), Erlingsson et al. (2014),Machanavajjhala et al. (2008)). This new framework finds its roots in the cryptography literature and prescribes to transform the original data, containing sensitive information, using a channel or *mechanism* Q into a sanitized dataset. The mechanism Q should be chosen carefully, in such a way that, by only looking at the released dataset, an intruder will have very low probability of guessing correctly the presence or absence of a specific individual in the raw data and, therefore, the privacy of the latter will be preserved. Differential Privacy formalizes mathematically this intuitive idea. We will provide a short review on it in Subsection 2.3.

In this work, we bring together ideas from the Disclosure Risk and Differential Privacy literature to propose a formal way of choosing the stochastic matrix M used in PRAM. Specifically, when choosing M, we need to balance two conflicting goals: 1) on the one hand, we want the application of M to make the dataset somehow private; 2) on the other hand, we also want that the released dataset preserves as much statistical information as possible from the raw data. In order to balance this trade-off, we propose to choose M as the solution of a constrained maximization problem. We maximize the Mutual Information between the released and the raw dataset, hence guaranteeing preservation of statistical information and achieving goal 2). Mutual Information is a common measure of dependence between random variables used in probability and information theory. In order to guarantee also goal 1), we introduce a constraint in the maximization problem by imposing that the application M satisfies differential privacy, therefore the resulting mechanism based on M can formally be considered private. We show how this optimization problem results in a convex maximization problem under linear contraints and can therefore being solved efficiently by known optimization algorithms.

The rest of this work is organized as follows. In Section 2, first we will briefly review the disclosure risk problem in Subsection 2.1 and then the tools needed for our approach. Specifically, we review Mutual Information in Subsection 2.2 and Differential Privacy in Subsection 2.3. In Section 3, we formalize the proposed constrained maximization problem to choose the stochastic matrix M in PRAM and show that this choice is made by solving a convex optimization problem under linear constraints. Section 4 contains a simulation study showing first the effect of the Diffential Privacy constraint on simulated data and then the effect of different choices of M using a real dataset of a survey of New York residents. Finally, a concluding remarks section closes the work. Proofs of the statements are deferred to the Appendix.

2 Literature Review

2.1 Disclosure Risk Limitation with categorical variables

In disclosure risk problems, we usually have microdata of n individuals, where for each individual we can observe two distinct sets of variables: 1) some variables, usually called *sensitive variables*, containing private information, e.g. health status or salary; 2) some identifying categorical variables, usually called *key variables*, e.g. gender, age or job. Disclosure problems arise because an intruder may be able to identify individuals in the dataset by cross-classifying their corresponding key variables and matching them to some external source of information. If the matching is correct, the intruder will be able to disclose the information contained in the sensitive variables.

Formally, let us assume we have J categorical key variables in the dataset, observed for a sample of n individuals, collected from a population of size N. Each variable has n_j possible categories labelled, without loss of generality, from 1 up to n_j. The observation for individual i, $X_i = (X_{i1}, \ldots, X_{iJ})$, therefore takes values in the state space $\mathcal{C} := \prod_{j=1}^{J}\{1 \ldots, n_j\}$. This set has $K := |\mathcal{C}| = \prod_{j=1}^{J} n_j$ values, corresponding to all possible cross-classification of the J key variables. The information about the sample is usually given through the sample frequency vector $(f_1, \ldots, f_K)$, where f_i counts how many individuals of the sample have been observed with the particular combination of cross-classified key variables corresponding to cell i. $(F_1, \ldots, F_K)$ denotes the corresponding vector frequencies when considering the whole population of N individuals.

The earliest papers to consider disclosure risk problems include Bethlehem et al. (1990), Duncan and Lambert (1986), Duncan and Lambert (1989), Lambert (1993). These works propose different measures of disclosure risk and possible ways to estimates them under different model choice. Skinner and Elliot (2002), Skinner et al. (1994) review the most popular among measures of disclosure risk. These measures depend on the sample frequencies $(f_1, \ldots, f_K)$ and usually focus on small frequencies, especially those having frequency 1, called *sample uniques*. The individuals belonging to these cells are those with the highest risk of their sensitive information being disclosed. Specifically, suppose that an individual is the only one both in the sample and in population to have a specific combination of key variables. Then, his key variables can be matched to an external database, and therefore this match will be perfect, i.e. correct with probability one, and his sensitive information will be therefore disclosed.

We usually distinguish between two groups of *measures of disclosure risk*:

1. **Record Level** (or per-record) measures: they assign a measure of risk for each data point. Among the most popular, there are

$$\begin{aligned} r_{1k} &= \mathbb{P}(F_k = 1 | f_k = 1), \\ r_{2k} &= \mathbb{E}(1/F_k | f_k = 1). \end{aligned} \tag{1}$$

   $k \in \{1, \ldots, K\}$. The first measure provides the probability that a sample unique is also population unique. The second tells the probability that if we select a sample unique and guess uniformly about his identity, we pick him correctly. The first measure is less conservative and is always smaller than the second.
2. **File level** measures: they provide an overall measure of risk for a dataset and are usually defined by aggregating the record level. Popular examples are

$$\tau_1 = \sum_{k: f_k = 1} r_{1k}, \quad \tau_2 = \sum_{k: f_k = 1} r_{2k}. \tag{2}$$

These measures of disclosure risk are estimated using the data $(f_1, \ldots, f_K)$ under different modelling choices. For example, Skinner and Shlomo (2008), Shlomo and Skinner (2010) consider the estimation of these measures under log-linear models for the population and sample frequencies. Under this model choice, the indexes (1) and (2), can be derived in closed form and estimated using plug-in MLE estimators. A different modelling approach, proposed in Manrique-Vallier and Reiter (2012),Manrique-Vallier and Reiter (2014), is to apply grade of membership models, which provide very accurate estimates for (2). For a quite recent review on disclosure risk problems, the reader is referred to Matthews and Harel (2011).

If the estimated values of (1) and (2) are too high, then the data curator should apply a disclosure limitation technique to the dataset before releasing it to the public. Some possibilities are for example rounding, suppression of extreme values or entire variables, subsampling or perturbation techniques. See Willenborg and de Waal (2001) for a review of different disclosure limitation techniques.

2.2 Mutual Information

Let X be a discrete random variable taking values on a finite set $\mathcal{X}$ and having probability mass function $p_X(x)$. The *(Shannon) entropy* of X is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x) = -\mathbb{E}(\log(p_X(x)))$$

and it is a measure of uncertainty about the distribution of X. $H(X)$ is always non-negative, takes value 0 when p_X is a point mass in one of the support points and it is maximized when p_X is uniform, $p_X(x) = \frac{1}{|\mathcal{X}|}$ $\forall x \in \mathcal{X}$, in which case $H(X) = \log |\mathcal{X}|$.
Similarly, given two discrete random variables X and Z, their *joint entropy* is defined as

$$H(X, Z) = -\sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} p_{(X,Z)}(x, z) \log p_{(X,Z)}(x, z),$$

where $p_{(X,Z)}$ denotes the joint mass function on $\mathcal{X} \times \mathcal{Z}$. $H(X, Z)$ measures the joint uncertainty of X and Z taken together.
Besides the *conditional entropy* of Z given X is defined as

$$\begin{aligned} H(Z|X) &= -\sum_{x,z} p_{X,Z}(x, z) \log(P_{Z|X=x}(z)) \\ &= H(X, Z) - H(X) \end{aligned} \tag{3}$$

and quantifies the amount of information needed to describe the outcome of Z given that the value of X is known. If Z and X are independent, the conditional entropy $H(Z|X)$ coincides with $H(Z)$.

The *mutual information* between X and Z is defined as

$$I(X,Z) = \sum_{z\in\mathcal{Z}}\sum_{x\in\mathcal{X}} P_{(X,Z)}(x,z)\log\left(\frac{p_{(X,Z)}(x,z)}{p_X(x)p_Z(z)}\right)$$

where $p_X, p_Z, p_{(X,Z)}$ are respectively the marginal and joint distributions of X and Z. From the definition of $I(X,Z)$ it follows that

$$I(X,Z) = D_{KL}(p_{(X,Z)}||p_X p_Z) \tag{4}$$

where D_{KL} denotes the Kullback-Leibler divergence. Therefore, $I(X,Z)$ measures the divergence between the joint distribution of X and Z and the product of their marginals. From (4), it also follows that $I(X,Z) \geq 0$, and $I(X,Z) = 0$ if and only if X and Z are independent.

An important equality connecting the mutual information $I(X,Z)$ with the marginal and joint entropies is

$$I(X,Z) = H(X) + H(Z) - H(X,Z). \tag{5}$$

This formula is the base of the so-called $3H$ principle to estimate $I(X,Z)$, in which the three H entropy terms on the right hand side are estimated from the data and plugged into (5) to obtain an estimate $\widehat{I(X,Z)}$.

For a review on entropy, mutual information and their properties, see for example Gibbs and Su (2002); Gray (2011) and references therein.

2.3 Differential Privacy

Differential Privacy is a notion recently proposed in the computer science literature by Dwork et al. (2006); Dwork and Roth (2014) mathematically formalize the idea that the presence or absence of an individual in the raw data should have a limited impact on the transformed data, in order for the latter to be considered privatized. Formally, let $X_{1:n} = (X_1,\dots,X_n)$ be a set of observations, taking values in a state space $\mathcal{X}^n \subseteq \mathbb{R}^n$, containing sensitive information. A *mechanism* is simply a conditional distribution Q that, given the raw dataset $X_{1:n}$, returns a transformed dataset $Z_{1:k_n} = (Z_1,\dots,Z_{k_n})$, with $\mathcal{Z}^{k_n} \subseteq \mathbb{R}^{k_n}$, to be released to the public, where the sample sizes of $X_{1:n}$ and $Z_{1:n}$ are allowed to be different. Differential Privacy is a property of Q that guarantees that it should be very difficult for an intruder to recover the sensitive information of $X_{1:n}$ by having access only to $Z_{1:k_n}$ and is defined as follows.

Definition 1 (α-Differential Privacy, Dwork et al. (2006)) The mechanism Q satisfies α-Differential Privacy if

$$\sup_{S\in\sigma(\mathcal{Z}^n)} \frac{Q(Z_{1:n}\in S|X_{1:n})}{Q(Z_{1:n}\in S|X'_{1:n})} \leq \exp(\alpha) \tag{6}$$

for all $X_{1:n}, X'_{1:n} \in \mathcal{X}^n$ s.t. $d_H(X_{1:n}, X'_{1:n}) = 1$, where d_H denotes the Hamming distance, $d_H(X_{1:n}, X'_{1:n}) = \sum_{i=1}^n \mathbb{I}(X_i \neq X'_i)$ and $\mathbb{I}$ is the indicator function of the event inside brackets.

For small values of α the right hand side of (6) is approximately 1. Therefore, if Q satisfies Differential Privacy, (6) guarantees that the output database $Z_{1:n}$ has basically the same probability of having been generated from either one of two *neighboring databases* $X_{1:n}$, $X'_{1:n}$, i.e. databases differing in only one entry. See Rinott at al. (2018) for a statistical viewpoint of differential privacy.

Differential Privacy has been studied in a wide range of problems, differing among them in the way data is collected and/or released to the end user. The two most important classifications are between Global vs Local privacy, and Interactive vs Non-Interactive models. In the *Global (or Centralized) model* of privacy, each individual sends his data to the data curator who privatizes the entire data set centrally. Alternatively, in the *Local (or Decentralized) model*, each user privatizes his own data before sending it to the data curator. In this latter model, data also remains secret to the possibly untrusted curator. In the *Non-Interactive (or Off-line) model*, the transformed data set $Z_{1:n}$ is released in one spot and each end user has access to it to perform his statistical analysis. In the *Interactive (or On-line) model* however, no data set is directly released to the public, but each end user can ask queries f about $X_{1:n}$ to the data holder who will reply with a noisy version of the true answer $f(X_{1:n})$.

There have been many extensions and generalizations of the notion (6) of Differential Privacy proposed over the last ten years, in order to accommodate for different areas of applications and state spaces of the input and output data. Among them, we mention (α,δ)-Differential Privacy (Dwork and Roth (2014)), vertex and edge Differential Privacy for network models (Borgs et al. (2015)), zero-mean Concentrated Differential Privacy (Bun and Steine (2016)), randomised differential privacy (Happ et al. (2011)) or ρ Differential Privacy (Chatzikokolakis et al. (2013), Dimitrakakis et al. (2017)), where the Hamming distance d_H is (6) is replaced by possibly any distance ρ, and many others. However, since it is not possible to review all the many extensions of Differential Privacy here, we refer to Dwork and Roth (2014) for a quite updated review on different applications and extensions of Differential Privacy. To conclude this brief review, we recall one of the most important properties of any Differential Private mechanism: *post processing*, see Dwork and Roth (2014). This property guarantees that if the output $Z_{1:n}$ of any α-Differential Private mechanism is further processed and gone through another mechanism (depending only on $Z_{1:n}$, and not on $X_{1:n}$), then the resulting output will also be α-Differential Private. Therefore, there will be no chance of any leak of privacy simply by post-processing the released data $Z_{1:n}$.

3 An information-theoretic approach to PRAM using Differential Privacy

Post Randomization Method is a popular perturbation method for disclosure risk limitation. It is connected to randomized response techniques described by Warner (1965). In the former approach, the raw data are perturbed by the data holder after having being collected, while in the latter, the perturbation is directly applied by the respondents during the interviewing process. We remind that PRAM was introduced by Kooiman et al. (1997) and further explored by Gouweleeuw et al. (1997) and De Wolf et al. (1997). Given raw microdata, PRAM produces a new dataset where some of entries are randomly changed according to a prescribed probability mechanism. The randomness introduced by the mechanism implies that matching a record in the perturbed dataset may actually be a mismatch instead of a true match, hence making usual disclosure matching attempts less reliable.

Shlomo and Skinner (2010) consider the problem of disclosure risk estimation when the microdata has gone through either a PRAM or data swapping process. They perturb the geographical key variable using a stochastic matrix M, i.e. every row of M sums to one, where M_{ij} provides the probability that an individual from location i is changed to location j. Shlomo and Skinner (2010) then proceed to discuss the problem of how to estimate the measures of risk presented in Subsection 2.1, but without providing any tangible rule on how to choose M, which is not the main goal of that paper.

In this work, we propose a novel approach to choose the randomization matrix M in PRAM. Specifically, we propose to choose it as the solution of a constrained maximization problem, in which we maximize the mutual information between raw data $X_{1:n}$ and released data $Z_{1:n}$, under the constraint that the perturbation mechanism satisfies the Differential Privacy condition (6). Other optimization approaches for PRAM were already considered by Willnborg (1999) and Willenborg (2000), using different target functions and constraints. See also Section 5.5 of Willenborg and de Waal (2001). However, these choices usually result in a difficult maximization problems and often rely on approximation methods.

We argue that the choice of Mutual Information and Differential Privacy have several advantages. First, Mutual Information and Differential Privacy are very natural notions and popular measures of information similarity and privacy guaranty that have been widely considered in Information Theory and Machine Learning. Second, as it will be shown shortly, the resulting maximization problem reduces to a convex maximization problem under a set of linear constraints, hence it can be solved efficiently by well known optimization tools, like the Simplex method which is implemented in most of the commonly used computational softwares, like Matlab or R. Finally, the level of privacy guaranteed by the proposed methodology is tuned by a single tuning parameter α, which can be chosen by the data curator to achieve the desired level of privacy in a very simple manner. In subsection 4.1, we will show empirically how the choice of this parameter affects the estimation of the parameters, hence providing some evidence and guidance on how to choose it.

3.1 Model of PRAM

We propose to choose M as the solution of the following constrained maximization program

$$\max_{M \text{ satisfies (6)}} I(X_{1:n}, Z_{1:n}). \tag{7}$$

We will consider the case of randomly changing the values of a key variable with S possible outcomes, e.g. the geographical location. $X_i \in \{1, \dots, S\}$ is the corresponding categorical random variable, having probabilities $p = (p_1, \dots, p_S)$, and therefore $\mathbb{P}(X_i = j) = p_j$. We consider the class of all randomizing matrices of the following form

$$M = \begin{bmatrix} q_1 & \frac{1-q_1}{S-1} & \frac{1-q_1}{S-1} & \cdots & \frac{1-q_1}{S-1} \\ \frac{1-q_2}{S-1} & q_2 & \frac{1-q_2}{S-1} & \cdots & \frac{1-q_2}{S-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1-q_S}{S-1} & \frac{1-q_S}{S-1} & \frac{1-q_S}{S-1} & \cdots & q_S \end{bmatrix} \tag{8}$$

for an unknown parameter vector $q = (q_1, \dots, q_S)$. This corresponds to the case in which, given that X_i belongs to category j, then its transformed value Z_i will either remain unchanged with probability q_j, or will be changed to one of the other $S-1$ categories, chosen uniformly at random, with probability $1 - q_j$. Therefore, the conditional distribution of Z_i given X_i is

$$Q(Z_i|X_i) = q_{X_i}^{\mathbb{I}(Z_i=X_i)} \left(\frac{1-q_{X_i}}{S-1}\right)^{\mathbb{I}(Z_i \neq X_i)}.$$

To underline the dependency on the vector q, we will sometimes write Q_q. It is easy to check that the marginal of Z_i is given by

$$\mathbb{P}(Z_i = j) = p_j q_j + \sum_{k \neq j} p_k \frac{1-q_k}{S-1} =: m_j \tag{9}$$

for $j \in \{1,..,S\}$. We remark that the vector $m = (m_1, \dots, m_S)$ can be computed in linear time in the dimension S by first computing the quantity $\sum_{k=1}^S p_k \frac{1-q_k}{S-1}$.

In the non interactive setting that we are considering, i.e. when Z_i only depends on X_i, the conditional distribution of $Z_{1:n}$ factorizes and can be written as

$$Q(Z_{1:n}|X_{1:n}) = \prod_{i=1}^n Q(Z_i|X_i).$$

Plugging it into (6), the Differential Privacy condition simplifies into

$$\sup_{Z_i, X_i \neq X_i'} \frac{Q(Z_i|X_i)}{Q(Z_i|X_i')} \leq e^\alpha.$$

Depending on the value of Z_i, the quotient $Q(Z_i|X_i)/Q(Z_i|X_i')$ can take one of three values. If $Z_i = X_i$, then it is equal to $(S-1)q_{X_i}/(1-q_{X_i'})$. If $Z_i = X_i'$, then it is equal to $(1-q_{X_i})/(S-1)q_{X_i'}$. Finally, if Z_i is different from both X_i and X_i', then the quotient is equal to $(1-q_{X_i})/(1-q_{X_i'})$. Therefore, the privacy condition specializes into the following set of constraints

$$\max\left(\frac{(S-1)q_k}{1-q_{k'}}, \frac{1-q_k}{(S-1)q_{k'}}, \frac{1-q_k}{1-q_{k'}}\mathbb{I}(S \geq 3)\right) \leq e^\alpha \tag{10}$$

for any couple $k \neq k' \in \{1, \dots, S\}$. We notice that this set of conditions can be expressed as a linear constraint. Specifically,

Fact 1: There exists a matrix C and a vector b_α (depending on α) such that the set of differential privacy constraints (10) can be rewritten as the following linear constraint

$$Cq^T \leq b_\alpha, \tag{11}$$

where q is the vector $q = (q_1, .., q_S)$ and $\leq$ denotes entry-wise inequality. C and b_α are given in Appendix.

In general, computing $I(X, Z)$ takes of an order of $|\mathcal{X}||\mathcal{Z}|$ operations, meaning that here it should be quadratic in S. However, due to the particular form of the matrix M considered here, this computation can be achieved linearly in S. Let us recall from Subsection 2.2, that $H(Z)$ denotes the Shannon entropy of the random variable Z and $H(Z|X)$ the conditional entropy of Z given X. To underline the dependency on q, we denote $f(q) := I(X, Z)$. We use the following known identity, which can immediately be derived from (5) and (3),

$$f(q) = I(X, Z) = H(Z) - H(Z|X),$$

which leads to the simpler form

$$f(q) = \sum_{x=1}^S p_x \left(q_x \log q_x + (1-q_x) \log \frac{1-q_x}{S-1}\right) - \sum_{z=1}^S m_z \log m_z$$

where we recall that $m = (m_1, \dots, m_J)$ denotes the marginal distribution of Z given in (9). Let us start by noticing that f is minimal, equal to 0, for $q_1 = .. = q_S = \frac{1}{S}$. In the Appendix, we show that f is convex in q, which, together with differential privacy constraint (11), implies that the problem (7) is a linearly constrained convex program, i.e. we are maximizing a convex function under a set of linear constraints. As a consequence, the following proposition follows,

Proposition 1 *Any optimal q solution of the program* (7)*, lays within the vertices of the convex polytope formed by all the feasible points.*

It follows from the previous proposition that finding the optimal matrix M of general form (8) requires finding the vertices of the feasible set. In Section 3.3 we will give some properties of this feasible set, which might make the search faster. In the following paragraph, we will provide the optimal M for several sub-cases of (8).

3.2 Examples

In this section we show how we can use Proposition 1 to give the explicit solutions of the program (7) for several particular examples of interest.

3.2.1 Binary key variable with symmetric M

We start from the simplest case of a categorical variable with only two possible categories denoted $\mathcal{X} = \{0, 1\}$ and symmetric M with $q_1 = q_2 = q$. We will abuse our notations by writing q both the scalar value in $[0, 1]$ and the corresponding two-dimensional vector $(q, q)^T$ having both coordinates equal to this value. We are considering binary symmetric matrices of the following form,

$$M = \begin{bmatrix} q & 1-q \\ 1-q & q \end{bmatrix}.$$

In this setting, the Differential Privacy condition (10) specializes into

$$\max\left(\frac{q}{1-q}, \frac{1-q}{q}\right) \leq e^{\alpha},$$

which simplifies to $q \in [\frac{1}{1+e^{\alpha}}, \frac{e^{\alpha}}{1+e^{\alpha}}]$. In such a situation, the constrained maximization problem can actually be solved analytically by derivation of the target function. However, from Proposition 1, it is already known that the optimal q is among the boundaries of the feasible region. Let $\psi : \{0,1\} \rightarrow \{0,1\}$ be defined as $\psi(x) = 1-x$. Since ψ is one-to-one, it follows that $I(X,Z) = I(X,\psi(Z))$. Moreover, by noticing that $\psi(Z)$ has conditional distribution Q_{1-q}, we can deduce that $I(X,\psi(Z)) = f(1-q)$, and therefore $f(q) = f(1-q)$. Hence, the optimal q are both boundaries points, $\frac{1}{1+e^{\alpha}}$ and $\frac{e^{\alpha}}{1+e^{\alpha}}$.

There are two interesting properties appearing in this simple example. First, we understand that there are two solutions of the program. Second, these solutions are independent of p, the marginal of X.

3.2.2 Binary key variable with any M

The previous argument can be easily extended to the non-symmetric case,

$$M = \begin{bmatrix} q_1 & 1-q_1 \\ 1-q_2 & q_2 \end{bmatrix}.$$

In this setting, the convex polytope generated by the linear constraints has four vertices, specifically (q_1, q_2) belongs to the following set

$$\left\{(1,0), (0,1), \left(\frac{1}{1+e^{\alpha}}, \frac{1}{1+e^{\alpha}}\right), \left(\frac{e^{\alpha}}{1+e^{\alpha}}, \frac{e^{\alpha}}{1+e^{\alpha}}\right)\right\}$$

If either (q_1, q_2) is equal to $(1,0)$ or $(0,1)$, then the Mutual Information $I(X_{1:n}, Z_{1:n})$ is null, since Z_i will be constant and independent of X_i. Therefore, the only optimal solutions are the two symmetric matrices derived in the symmetric case.

3.2.3 Symmetric M

Let us now consider the case of a categorical variable with S categories and symmetric M. Specifically, we consider $\mathcal{X} = \{1, \ldots, S\}$ and M of the form (8) with $q_1 = q_2 = \cdots = q_S$. We again abuse our notation by denoting with q both the scalar in $[0,1]$ and the corresponding S dimensional vector with all entries equal to this value. The differential privacy condition (10) specializes into $\max\left(\frac{(S-1)q}{1-q}, \frac{1-q}{(S-1)q}\right) \leq e^{\alpha}$, which leads to $q \in [\frac{e^{-\alpha}}{S-1+e^{-\alpha}}, \frac{e^{\alpha}}{S-1+e^{\alpha}}]$. As before, following from Proposition 1, the optimal q are the boundary values.

3.3 Feasible set

In our experiments, we have experienced that routine optimization functions implemented in standard software, e.g. Matlab, can solve the optimization problem (7) extremely quickly. However, when the number of possible categories S becomes very large, the optimization might become time consuming. For this reason, in the following Proposition, we provide a description of all possible vectors $q = (q_1, \ldots, q_S)$ that can arise as vertices of the convex polytope generated by the Differential Privacy constraints (10) when S is large enough. This result should help to speed up the search for the optimal vertices among all feasible points given by (10).

Proposition 2 *For $S \geq 4$, if $\alpha \leq \log(S + \sqrt{S(S-4)} - 2) - \log 2$, then, up to permutations, the vertices of the convex polytope formed by all feasible points are:*

1. $q_k = v_{\alpha}$, $\forall k \in \{1, \ldots, S\}$;
2. $q_k = v_{-\alpha}$, $\forall k \in \{1, \ldots, S\}$;
3. $q_k = v_{i_k \alpha}$, *with* $i_k = \pm 1$ *and* $2 \leq \#\{k\ s.t. i_k = 1\} \leq S-2$, $\forall k \in \{1, \ldots, S\}$;
4. $q_1 = v_{\min}$, $q_k = v_{\alpha}\cdot$, $\forall k \in \{2, \ldots, S\}$;
5. $q_1 = v_{\max}$, $q_k = v_{-\alpha}$, $\forall k \in \{2, \ldots, S\}$;

where $v_x = \frac{e^x}{e^x+S-1}$, $v_{\min} = \frac{e^{-\alpha}}{e^{\alpha}+S-1}$ *and* $v_{\max} = \frac{e^{\alpha}}{e^{-\alpha}+S-1}$.

Common values of α are generally within the range $[0,2]$, which means that the conditions of previous Proposition are satisfied when $S \geq 10$. Contrary to the symmetric case, the optimal M will depend on p. In the following section, we will show some simulations illustrating for some values of p which of the vertices in Proposition 2 are optimal.

4 Simulations

4.1 Simulation Study

We consider different simulates scenarios, where the observations $X_{1:n}$ are generated from a categorical distribution with S possible outcomes, having known probabilities $p = (p_1, \ldots, p_S)$. In the first scenario, we set $S = 10$ and consider the following vector of probabilities

$$p = (0.3, 0.1, 0.2, 0.08, 0.02, 0.04, 0.06, 0.1, 0.01, 0.09).$$

We consider the following values of $\alpha = 0.5, 1, 1.5, 2$, and we determine the corresponding optimal vectors of $q = (q_1, \dots, q_S)$ that solve (7). We select a sample size $n = 10^4$. As explained in Section 3, the Differential Privacy condition can be expressed as a set of linear constraints as in (11). The optimal q is then determined numerically by solving the constrained maximization problem via the optimization function in Matlab. Besides we have also generated the corresponding privatized dataset $Z_{1:n}$ using the determined values of $(q_1, \dots, q_S)$ for the different choices of α. The determined values of q are reported in Table 1. From Proposition 2, we know that, up to permutations, there are only 5 possible different scenarios and the q_k's may assume only 4 different values, corresponding to $v_\alpha, v_{-\alpha}, v_{\min}$ and $v_{\max}$. Hence in Table 1 we have reported the number of times the q_k's assume these values for the different choices of α. In order to investigate the effect of differential privacy, for the four values of α considered here, we have reported the MLE of the vector of probabilities p obtained using the observed sample $Z_{1:n}$. The results are represented in Figure 1, all the simulations are averaged over 100 iterations. For each value of the categorical variable $k \in \{1, \dots, 10\}$, we have reported the estimated p_k's, and each blue star corresponds to the MLE of p_k in one of the 100 experiments. The solid red line links the averaged estimates of the p_k's over the 100 runs, while the true values of the probabilities p_k are represented in yellow. It is apparent that as α increases, the estimates improve and the variability of the estimates decreases, hence the higher α, the weaker the privacy mechanism.

α	$\# v_\alpha$	$\# v_{-\alpha}$	$\# v_{\min}$	$\# v_{\max}$
0.5	4	6	0	0
1	5	5	0	0
1.5	2	8	0	0
2	0	9	0	1

Table 1 Scenario I: the number of times the q_k's assume the four possible values $v_\alpha, v_{-\alpha}, v_{\min}$ and $v_{\max}$, under different choices of α.

Figure 1 about here.

In the second scenario we have generated the data using the vector of probabilities

$$p = (0.0336, 0.1059, 0.1697, 0.0962, 0.0180,\\ 0.0062, 0.1097, 0.0005, 0.1233, 0.3369).$$

As before we report the values of the q_k's for different choices of α in Table 2, besides the estimated probabilities p_k's are reported in Figure 2. The simulations are averaged over 100 iterations.

α	$\# v_\alpha$	$\# v_{-\alpha}$	$\# v_{\min}$	$\# v_{\max}$
0.5	7	3	0	0
1	6	4	0	0
1.5	6	4	0	0
2	0	9	0	1

Table 2 Scenario II: the number of times the q_k's assume the four possible values $v_\alpha, v_{-\alpha}, v_{\min}$ and $v_{\max}$, under different choices of α.

Figure 2 about here.

We consider now a third scenario, in which $S = 30$ and we have generated the data using the vector of probabilities p obtained as a normalization of 30 independent gamma random variables with parameters $(1, 5)$, more precisely we have generated $G_k \sim \text{Gamma}(1, 5)$ for $k = 1, \dots, 30$ and we have put $p_k := G_k / \sum_{s=1}^S G_s$. As before we report the vectors of q for different values of α in Table 3 and the estimated probabilities averaged over 100 iterations in Figure 3, where again $n = 10^4$ is the sample size.

α	$\# v_\alpha$	$\# v_{-\alpha}$	$\# v_{\min}$	$\# v_{\max}$
0.5	29	0	1	0
1	29	0	1	0
1.5	30	0	0	0
2	29	0	1	0

Table 3 Scenario III: the number of times the q_k's assume the four possible values $v_\alpha, v_{-\alpha}, v_{\min}$ and $v_{\max}$, under different choices of α.

Figure 3 about here.

In the last scenario IV, we assume again that $S = 30$ and we have generated the data using the vector of probabilities p defined by

$$p_1 = 0.05, \quad p_k = 0.95/29 \text{ for } k \geq 2.$$

We report the vectors of q for different values of α in Table 4 and the estimated probabilities averaged over 100 iterations in Figure 4, where the sample size equals $n = 10^4$.

α	$\# v_\alpha$	$\# v_{-\alpha}$	$\# v_{\min}$	$\# v_{\max}$
0.5	0	29	0	1
1	30	0	0	0
1.5	30	0	0	0
2	30	0	0	0

Table 4 Scenario IV: the number of times the q_k's assume the four possible values $v_\alpha, v_{-\alpha}, v_{\min}$ and $v_{\max}$, under different choices of α.

Figure 4 about here.

4.2 Real Data

We finally test the performance of our strategy on some benchmark datasets from the public use microdata sample of the U.S. 2000 census for the state of New York, Ruggles et al. (2010). The data contains the values of ten categorical variables of 953076 individuals: ownership of dwelling (3 levels), mortgage status (4 levels), age (9 levels), sex (2 levels), marital status (6 levels), single race identification (5 levels), educational attainment (11 levels), employment status (4 levels), work disability status (3 levels), and veteran status (3 levels).

For ease of illustration we consider the sex variable, which has two possible categories (female or male), therefore $S = 2$ and $q = q_1 = q_2$. We have already seen that the optimal q lies on the boundaries of the interval $J_\alpha := [1/(e^\alpha + 1), e^\alpha/(1 + e^\alpha)]$. We have estimated the probabilities of the two possible categories using the sample mean, thus obtaining $p_1 = 0.48$ and $p_2 = 0.52$. In our numerical experiments we have considered $\alpha = 0.05$, and for different values of $q \in J_\alpha$ we have generated the privatized dataset $Z_{1:n}$ estimating p_1 and p_2. More precisely, in Figure 5 we have considered six values of $q \in [0.4875, 0.5125]$, and we reported the estimates of p_1 and p_2 averaged over 100 iterations. Each panel corresponds to a different q, each blue star corresponds to the estimated value in one of the 100 experiments based on the privatized sample $Z_{1:n}$. The solid blue line links the averaged estimates of the p_k's over the 100 runs, while the true values of the probabilities are represented in yellow. From the top left to the bottom right, we have chosen $q = 0.4875, 0.4925, 0.4975, 0.5025, 0.5075, 0.5125$: from the theory developed in the paper it is not surprising to realize that the values on the boundary lead to more reliable estimates, indeed they maximize the mutual information between X and Z. In Figure 6 we reported the estimated mutual information between X and Z for different values of $q \in J_{0.05}$, in order to do that we have estimated $P_Z(k)$ and $P_X(k)$ using the corresponding sample means for each $k = 1, 2$.

Figures 5–6 about here.

5 Conclusions and future work

In this work, we have proposed a novel approach to choose the randomizing matrix M in PRAM. This approach applies popular tools from computer science to derive M as the solution of a constrained optimization problem, in which the Mutual Information between raw and transformed data is maximized, under the constraint that the transformation satisfies Differential Privacy. The proposed approach has the advantage to be quick and easy to implement. Also, the desired level of privacy can be tuned by a single parameter α.

There are different ways in which the present work could be extended. A first possible direction of research is to understand how to tune the Differential Privacy parameter α, which regulates the desired level of privacy, using the classical measures of risk (1) and (2). Specifically, given the choice of some model, α can be chosen in such a way that the estimate of the disclosure risk index computed on the transformed dataset matches or falls below a particular threshold value. A second direction of research is to generalize the proposed procedure to the case in which a few categorical variables are jointly perturbed. The proposed methodology can be extended to this case following similar lines. In particular, an individual will be randomly moved from one frequency cell to another using a $K \times K$ stochastic matrix, where K is the number of cells after cross-classifying the variables we want to jointly perturb. In this setting, it will be important to study what further structure the $K \times K$ matrix should have in order to avoid structural zeros combinations. Further, another direction of research is to examine the problem using other formulations of Differential Privacy. Specifically, the definition of Differential Privacy as in (6) is known to provide a very strong privacy guarantee. Therefore, generalizing the proposed methodology to other formulations and relaxations of Differential Privacy, as those mentioned in Subsection 1, might be an interesting topic.

Some other important research directions have also been suggested by the reviewers. Specifically, the proposed approach is focused on perturbing categorical variables, which are usually the most sensitive in terms of disclosure risk. However, a direction of research can be to study the problem for other datatypes, possibly including also some continuous data. Another line of research is to study theoretical guarantees in terms of preservation of utility for different classes of queries. In computer science, with a query it is usually meant a statistics of the observations, or function of some sufficient statistics. A crucial problem consists to quantify and analyse the expected distance (risk) of some classes of queries computed on raw and realised dataset. Similar contributions in this direction are Smith (2011) and Duchi et al. (2018). A useful extension of the proposed methodology would focus on different structures for the matrix M, rather than with uniform off-diagonal rows as in (8). An interesting example of application suggested by one reviewer, in which imposing non-uniform off-diagonal rows would be important, is in spatial modelling, when perturbing a geographical variable. In this context, a more suitable structure for M would allow for the geographical category to have a higher probability to be swapped with a spatially neighboring category rather than to one very far from the true observed value. Within this context, the optimal choice of M will have to balance between the higher randomization to achieve the same level of Differential Privacy and the benefit in statistical utility that follows from geographically localised perturbation for any later spatial analysis. Finally, another extension could be to include all variables in the mutual information in the maximization (7), applying privacy perturbation and the Differential Privacy constraint only to a subset of them. If the included and excluded variables are modelled as independent, the solution of the maximization problem M should be unaltered. Instead, in the dependent case, the optimal solution M might depend also on the non-perturbed variables and the maximization problem could become analytically much more challenging.

Aknowlegdment

The authors thank the Associate Editor and two anonymous referees, whose constructive comments and suggestions have been appreciated and helped to improve the paper. Federico Camerlenghi received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Federico Camerlenghi gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

Appendix

5.1 Proof of Proposition 1

To underline the dependency on q, we will sometimes use the notation Q_q. We need to show that f is convex. Let $q' = (q'_1, .., q'_S)^T$ and $\theta \in [0, 1]$. Let $k, l \in \{1, .., S\}$ such that $k \neq l$.

$$\begin{aligned} Q_{\theta q+(1-\theta)q'}(Z = k|X = k) &= \theta q_k + (1-\theta)q'_k \\ &= \theta Q_q(Z = k|X = k) + (1-\theta)Q_{q'}(Z = k|X = k). \end{aligned}$$

Besides,

$$\begin{aligned} &Q_{\theta q+(1-\theta)q'}(Z = l|X = k) \\ &\quad = \frac{1 - (\theta q_k + (1-\theta)q'_k)}{S-1} \\ &\quad = \frac{\theta(1-q_k) + (1-\theta)(1-q'_k)}{S-1} \\ &\quad = \theta Q_q(Z = l|X = k) + (1-\theta)Q_{q'}(Z = l|X = k). \end{aligned}$$

Therefore, $Q_{\theta q+(1-\theta)q'} = \theta Q_q + (1-\theta)Q_{q'}$. It is known that for a fixed marginal distribution of one of the variables, the mutual information is convex in the conditional distribution of the second, see for example Theorem 2.7.4 of Cover and Thomas (2012). Therefore, $f(\theta q + (1-\theta)q') \leq \theta f(q) + (1-\theta)f(q')$, and hence f is convex.

5.2 Fact 1: Set of feasible parameters q

We start by writing explicitly the linear constraints (11) on q. Let $\mathcal{T}_\alpha$ be the convex polytope of all q satisfying α-differential privacy. Let $\mathcal{S}_\alpha$ be the planar polygon defined by the set of equations

$$(S-1)x + e^\alpha y \leq e^\alpha \tag{12}$$
$$(S-1)y + e^\alpha x \leq e^\alpha \tag{13}$$
$$-(S-1)e^\alpha y - x \leq -1 \tag{14}$$
$$-(S-1)e^\alpha x + y \leq -1 \tag{15}$$
$$e^\alpha y - x \leq e^\alpha - 1 \tag{16}$$
$$e^\alpha x - y \leq e^\alpha - 1 \tag{17}$$

The set of feasible points is then characterized by

$$(q_1, \dots, q_S) \in \mathcal{T}_\alpha \iff \forall(k,l),\ (q_k, q_l) \in \mathcal{S}_\alpha.$$

This set characterized by the $3S(S-1)$ linear constraints given by equations (12) to (17) can thus be defined as the set of solutions of the equation $Cq^T \leq b_\alpha$, where C has dimension $3S(S-1) \times S$ and b_α is a $3S(S-1)$-dimensional vector.

5.3 Proof of Proposition 2

Equations (12) to (15) define a quadrilateral whose vertices are

$$\begin{aligned} u_\alpha &= \left(\frac{(S-1)e^\alpha - 1}{S(S-2)}, \frac{(S-1)e^{-\alpha} - 1}{S(S-2)}\right), \\ u_{-\alpha} &= \left(\frac{(S-1)e^{-\alpha} - 1}{S(S-2)}, \frac{(S-1)e^\alpha - 1}{S(S-2)}\right), \\ v_\alpha &= \left(\frac{e^\alpha}{S-1+e^\alpha}, \frac{e^\alpha}{S-1+e^\alpha}\right), \\ v_{-\alpha} &= \left(\frac{e^{-\alpha}}{S-1+e^{-\alpha}}, \frac{e^{-\alpha}}{S-1+e^{-\alpha}}\right). \end{aligned}$$

The points v_α and $v_{-\alpha}$ always satisfy (16) and (17). Besides, for $S \geq 4$, if $\alpha \leq \log(S + \sqrt{S(S-4)} - 2) - \log 2$, then u_α and $u_{-\alpha}$ also satisfy (16) and (17). In such a setting, equations (16) and (17) are redundant and hence can be omitted when defining $\mathcal{T}_\alpha$. Common values of α are generally within the range $[0, 2]$, therefore equations (16) and (17) are omitted when $S \geq 10$. In the following, we will suppose that $S \geq 4$ and $\alpha \leq \log(S + \sqrt{S(S-4)} - 2) - \log 2$. Let $(q_1, \ldots, q_S) \in \mathcal{T}_\alpha$, since (q_2, q_3) satisfy (15), we have that $q_3 \leq 1 - (S-1)e^\alpha q_2$. Therefore, using the fact that (q_1, q_3) satisfy (12), and the symmetry of the constraints, we can deduce that any (q_k, q_l) satisfy

$$y - e^{2\alpha} x \geq 1 \tag{18}$$

$$x - e^{2\alpha} y \geq 1 \tag{19}$$

Equations (13) and (18) give that $q_k \leq \frac{e^\alpha}{e^{-\alpha}+S-1} = v_{\max}$ and equations (15) and (18) give $q_k \geq \frac{e^{-\alpha}}{e^{\alpha}+S-1} = v_{\min}$. Let $\mathcal{V}$ be the set of points defined up to permutations by

1. $\forall k,\ q_k = \frac{e^\alpha}{e^\alpha+S-1} = v_\alpha$
2. $\forall k,\ q_k = \frac{e^{-\alpha}}{e^{-\alpha}+S-1} = v_{-\alpha}$
3. $\forall k,\ q_k = v_{i_k\alpha}$ with $i_k = \pm 1$ and $2 \leq \#\{k \text{ s.t } i_k = 1\} \leq S-2$
4. $q_1 = v_{\min},\ q_{k\geq 2} = v_\alpha$
5. $q_1 = v_{\max},\ q_{k\geq 2} = v_{-\alpha}$.

Under the assumption that $S \geq 4$ and $\alpha \leq \log(S + \sqrt{S(S-4)} - 2) - \log 2$, it is straightforward to verify that $\mathcal{V} \subset \mathcal{T}_\alpha$. In the following, we show that any element of $\mathcal{T}_\alpha$ is a convex combination of points of $\mathcal{V}$. In order to do so, we will use the following Lemma.

Lemma 1 *For $S \geq 2$, if q satisfies differential privacy, then at most one of its coordinates is larger than v_α and at most one is smaller than $v_{-\alpha}$*

Proof This trivially follow from the constraint (10). Indeed, suppose that $q_i > \frac{e^\alpha}{S-1+e^\alpha}$. Then for any other q_j, using formula (19),

$$(S-1)\frac{e^\alpha}{S-1+e^\alpha} + e^\alpha q_j < (S-1)q_i + e^\alpha q_j \leq e^\alpha$$

Therefore,

$$(S-1)\frac{1}{S-1+e^\alpha} + q_j < 1$$

$$q_j < 1 - (S-1)\frac{1}{S-1+e^\alpha} = \frac{e^\alpha}{S-1+e^\alpha}$$

Similarly suppose both $q_i < \frac{e^{-\alpha}}{S-1+e^{-\alpha}}$. For any other q_j, from formula (21),

$$(S-1)e^\alpha \frac{e^{-\alpha}}{S-1+e^{-\alpha}} + q_j > (S-1)e^\alpha q_i + q_j \geq 1$$

Therefore

$$\frac{(S-1)}{S-1+e^{-\alpha}} + q_j > 1$$

$$q_j > \frac{e^{-\alpha}}{S-1+e^{-\alpha}}$$

□

Let $(q_1, \ldots, q_S) \in \mathcal{T}_\alpha$, using previous Lemma, we know that up to permutations, one of 4 settings is possible:

1. For all k, $v_{-\alpha} \leq q_k \leq v_\alpha$.
2. $v_\alpha < q_1 \leq \frac{e^\alpha}{e^{-\alpha}+S-1} = v_{\max}$, and for $k \geq 2$, $v_{-\alpha} \leq q_k \leq v_\alpha$
3. $v_{\min} = \frac{e^{-\alpha}}{e^{\alpha}+S-1} \leq q_1 < v_{-\alpha}$, and for $k \geq 2$, $v_{-\alpha} \leq q_k \leq v_\alpha$
4. $v_\alpha < q_1 \leq v_{\max}$, $v_{\min} \leq q_2 < v_{-\alpha}$, and for $k \geq 3$, $v_{-\alpha} \leq q_k \leq v_\alpha$

The first setting is the most straightforward, indeed since $v_{\min} < v_{-\alpha}$ and $v_{\max} > v_\alpha$, we find that all the points $(v_{i_k\alpha})_{1\leq k\leq S}$ for any sequence $(i_k)_{1\leq k\leq S} \in \{-1, 1\}^S$, are within the convex hull of $\mathcal{V}$ and so does the whole hypercube generated by those 2^S points.

The second and third settings have similar proof, that we will explicit for the second setting. As said in the previous remark, the point $(v_\alpha, v_{-\alpha}, \ldots, v_{-\alpha})$ belongs to the convex hull of $\mathcal{V}$. Let $k \geq 2$, we know that $q_k \geq v_{-\alpha}$. Besides, since $(q_k, q_1) \in \mathcal{S}_\alpha$, (12) gives that (q_k, q_1) is below the line passing through $(v_{-\alpha}, v_{\max})$ and (v_α, v_α). Hence, denoting $\theta = \frac{q_1 - v_\alpha}{v_{\max} - v_\alpha}$, we find that

$$v_{-\alpha} \leq q_k \leq \theta v_{-\alpha} + (1-\theta)v_\alpha.$$

Therefore, we only need to show that any point $(q_1, x_2, \ldots, x_S)$ is in the convex hull of $\mathcal{V}$ for any sequence $(x_k)_{k\geq 2} \in \{v_{-\alpha},\ \theta v_{-\alpha} + (1-\theta)v_\alpha\}^{S-1}$. Let $(q_1, x_1, \ldots, x_{S-1})$ be such a point. Let $(i_k)_{2\leq k\leq S}$ such that $i_k = -1$ if $x_k = v_{-\alpha}$, and $i_k = 1$ otherwise. Now, from previous setting we know that $(v_\alpha, v_{i_2\alpha}, \ldots, v_{i_S\alpha})$ is in the convex hull of $\mathcal{V}$, and so does $(v_{\max}, v_{-\alpha}, \cdots, v_{-\alpha})$. We conclude as we notice that

$$\theta(v_{\max}, v_{-\alpha}, \cdots, v_{-\alpha}) + (1-\theta)(v_\alpha, v_{i_2\alpha}, \ldots, v_{i_S\alpha}) \\ = (q_1, x_2, \ldots, x_S)$$

The proof of the last setting is similar to the previous one. Equation (19) together with $x \geq v_\alpha$ and $y \leq v_{-\alpha}$ define a triangle within which (q_1, q_2) lays. The points $(v_{\max}, v_{-\alpha}), (v_\alpha, v_{\min})$ and $(v_\alpha, v_{-\alpha})$ are the three vertices of the triangle. Therefore, denoting $\theta_1 = \frac{q_1 - v_\alpha}{v_{\max} - v_\alpha}$ and $\theta_2 = \frac{v_{-\alpha} - q_2}{v_{-\alpha} - v_{\min}}$, we have that $0 \leq \theta_1, \theta_2 \leq 1$, $\theta_1 + \theta_2 \leq 1$ and (q_1, q_2) equals

$$\theta_1(v_{\max}, v_{-\alpha}) + \theta_2(v_\alpha, v_{\min}) + (1 - \theta_1 - \theta_2)(v_\alpha, v_{-\alpha}).$$

Let $k \geq 3$, since $(q_k, q_1) \in \mathcal{S}_\alpha$, (12) implies that (q_k, q_1) is below the line passing through $(v_{-\alpha}, v_{\max})$ and (v_α, v_α). Similarly, since $(q_2, q_k) \in \mathcal{S}_\alpha$, (15) implies that (q_2, q_k) is above the line passing through $(v_{\min}, v_\alpha)$ and $(v_{-\alpha}, v_{-\alpha})$. Therefore, q_k satisfies

$$\theta_2 v_\alpha + (1-\theta_2) v_{-\alpha} \leq q_k \leq \theta_1 v_{-\alpha} + (1-\theta_1) v_\alpha.$$

As previously, we only need to show that any point $(q_1, q_2, x_3, \dots, x_S)$ is in the convex hull of $\mathcal{V}$ for any $(x_k)_{k\geq 3} \in \{\theta_2 v_\alpha + (1-\theta_2) v_{-\alpha},\ \theta_1 v_{-\alpha} + (1-\theta_1) v_\alpha\}^{S-2}$. Let $(x_k)_{k\geq 3}$ be such a sequence. Let $(i_k)_{k\geq 3}$ defined by $i_k = -1$ if $x_k = \theta_2 v_\alpha + (1-\theta_2) v_{-\alpha}$ and $i_k = 1$ otherwise. We conclude by noticing that

$$\begin{aligned}&(q_1, x_2, \dots, x_S)\\ &\quad = \theta_1(v_{\max}, v_{-\alpha}, \cdots, v_{-\alpha}) + \theta_2(v_\alpha, v_{\min}, v_\alpha, \cdots, v_\alpha)\\ &\qquad + (1-\theta_1-\theta_2)(v_\alpha, v_{-\alpha}, v_{i_3\alpha}, \dots, v_{i_s\alpha}).\end{aligned}$$

References

Bethlehem, J.G., Keller, W.J., Pannekoek, J.: Disclosure control of microdata. J. Amer. Statist. Assoc. **85**, 38–45 (1990)

Borgs, C., Chayes, J., Smith, A.: Private Graphon Estimation for Sparse Graphs. ArXiv:1506.06162 (2015)

Bun, M., Steine, T.: Concentrated Differential Privacy: simplifications, extensions, and lower bounds. In: Theory and Cryptography - 14th International Conference, pp. 635–658 (2016)

Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C.: Broadening the scope of Differential Privacy using metrics. In: Privacy Enhancing Technologies, pp. 82–102. Springer Berlin Heidelberg (2013)

De Wolf, P.P., Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J.: Reflection on PRAM. Report, Department of Statistical Methods, Statistics Netherlands. Voorburg (1997)

Dimitrakakis, C., Nelson, B., Zhang, Z., Mitrokotsa, A., & Rubistein, B.I.P.: Differential Privacy for Bayesian Inference through Posterior Sampling. J. Mach. Learn. Res. **18**, 1–39 (2017)

Duchi, J.C., Jordan, M.I. & Wainwright, M.J.: Minimax optimal procedures for locally private estimation. JASA **113**(521), 182–201 (2018)

Duncan, G.T., Lambert, D.: Disclosure-Limited Data Dissemination. J. Amer. Statist. Assoc. **81**, 10–28 (1986)

Duncan, G.T., Lambert, D.: The Risk of Disclosure for Microdata. J. Bus. Econ. Stat. **7**, 207-217 (1989)

Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In Proc. of the Third Theory of Cryptography Conference, pp. 265–284 (2006)

Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**, 211–407 (2014)

Eland, A.: Tackling Urban Mobility with Technology by Andrew Eland. Google Policy Europe Blog, Nov 18, 2015.

Erlingsson, U., Pihur, V., Korolova, A.: RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In: Proceedings of the 21st ACM Conference on Computer and Communications Security (2014)

Gibbs, A.L., Su, F.E.: On Choosing and Bounding Probability Metrics. Int. Stat. Rev. **70**, 419–435 (2002)

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., De Wolf, P.P.: Post Randomization for Statistical Disclosure Control: Theory and Implementation. Report, Department of Statistical Methods, Statistics Netherlands. Voorburg (1997)

Gray, R.M.: Entropy and Information Theory. 2nd Edition. Springer, (2011).

Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. Science **339**, 321–324 (2013)

Hall, R., Rinaldo, A., Wasserman, L.: Random differential privacy. ArXiv:1112.2680 (2011)

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. **4**, e100016 (2008)

Kooiman, P., Willenborg, L., Gouweleeuw, J.: PRAM: A Method for Disclosure Limitation of Microdata. Report, Department of Statistical Methods, Statistics Netherlands, Voorburg (1997)

Lambert, D.: Measures of Disclosure Risk and Harm. J. Off. Stat. **9**, 313–331 (1993)

Machanavajjhala, A., Kifer, D., Abowd, J.M., Gehrke, J., Vilhuber, L.: Privacy: Theory meets Practice on the Map. In: Proceedings of the 24th International Conference on Data Engineering (2008)

Manrique-Vallier, D., Reiter, J.P.: Bayesian estimation of discrete multivariate latent structure models with structural zeros. J. Comput. Graph. Statist. **23**, 1061–1079 (2014)

Manrique-Vallier, D., Reiter, J.P.: Estimating identification disclosure risk using mixed membership models. J. Amer. Statist. Assoc. **107**, 1385–1394 (2012)

Matthews, G.J., Harel, O.: Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. Statist. Surv. **5**, 1-29 (2011)

Narayanan, A., Shmatikov, V.: Robust de-anonymization of large datasets. In *Proc. IEEE Security & Privacy Conference*, pp. 111–125 (2008)

Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B. and Sobek, M.: Integrated public use microdata series: Version 5.0 [Machine-readable database]. University of Minnesota, Minneapolis. Available at https://usa.ipums.org/usa/ (2010)

Rinott, Y., O'Keefe, C.M., Shlomo, N., Skinner, C.: Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. Statist. Sci. **33**, 358–385 (2018)

Shlomo, N., Skinner, C.J.: Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. Ann. of App. Stat. **4**(3), 1291–1310 (2010)

Skinner, C.J, Elliot, M.J: A Measure of Disclosure Risk for Microdata. J. Roy. Statist. Soc. B **64**, 855–867 (2002)

Skinner, C.J. & Shlomo, N.. Assessing identification risk in survey microdata using log-linear models. J. Amer. Statist. Assoc. **103**, 989–1001 (2008)

Skinner, C., Marsh, C., Openshaw, S.,Wymer, C.: Disclosure control for census microdata. J. Off. Stat. **10**, 31–51 (1994)

Smith, A.: Privacy-preserving statistical estimation with optimal convergence rates. In Proc. of the Forty-Third Annula ACM Symposium on the Theory of Computing, (2011)

Sweeney, L.: Waeving technology and policy together to maintain confidentiality. J. Law Med. Ethics **25**, 98-110 (1997)

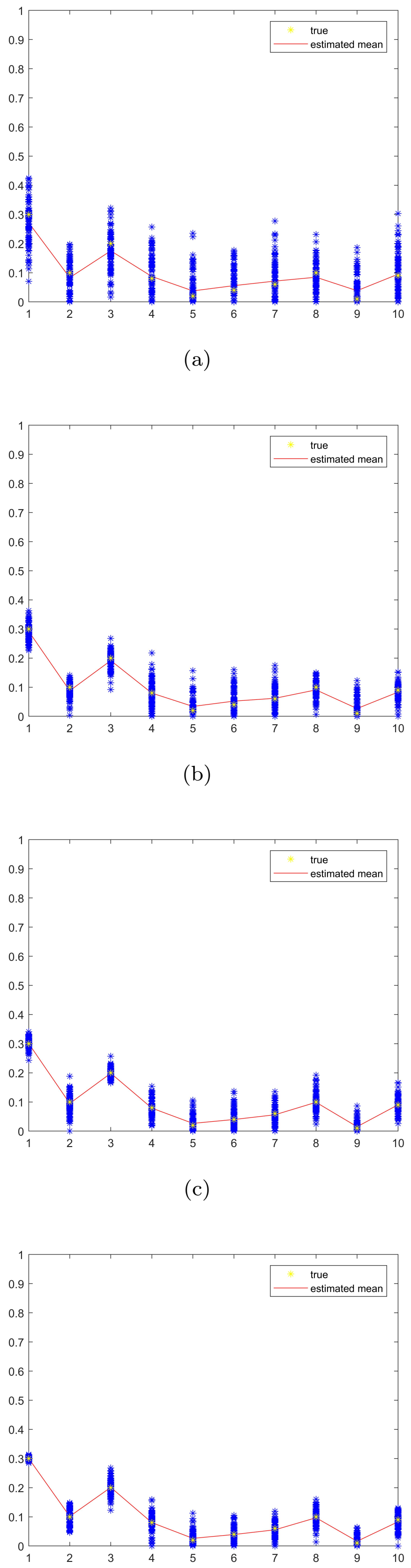
Warner, S.: Randomized response: a survey technique for eliminating evasive answer bias. J. Amer. Statist. Assoc. **60**(309), 63–69 (1965)

Willnborg, L.: Optimization Models for PRAM Matrices. Report, Department of Statistical Methods, Statistics Netherlands (1999)

Willenborg, L.: Optimality Models for PRAM. Proceedings Compstat 2000, Utrecht, 21-25 August, Physica-Verlag, Heidelberg (2000)

Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control in Practice. *Lecture Notes in Statistics*, **155**. Springer, New York (2001)

Cover, T. M., Thomas, J. A.: Elements of information theory. John Wiley & Sons (2012)



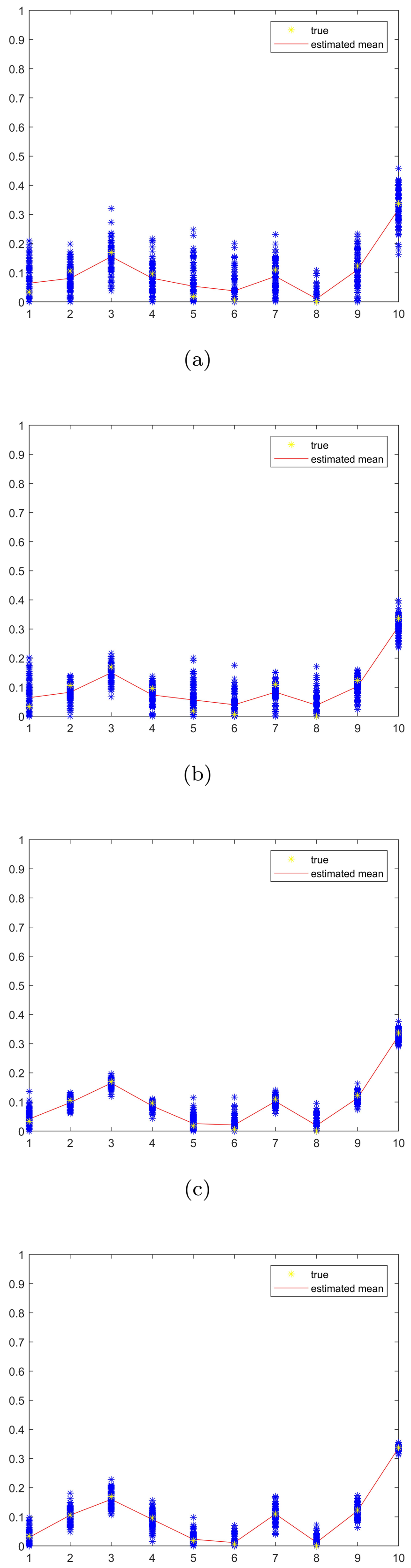
(a)

(b)

(c)

(d)

Fig. 1 Scenario I: estimates of the true probabilities generating the data. The x-axis encodes the $S = 10$ possible categories, for each one the yellow point represents the true probability p_k, while the solid red line connects the estimated probabilities averaged over 100 iterations.



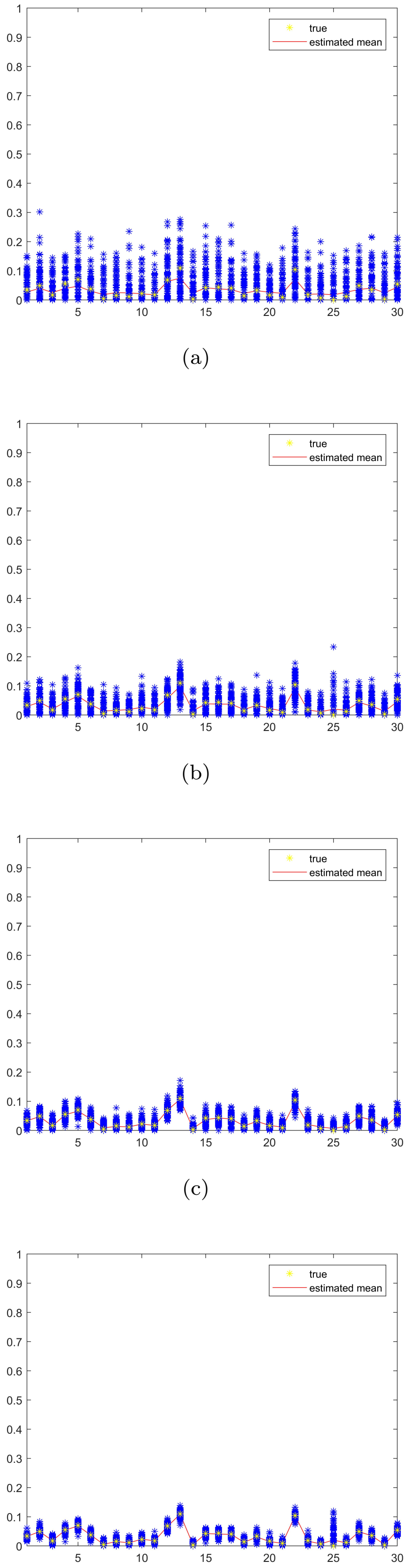
(a)

(b)

(c)

(d)

Fig. 2 Scenario II: estimates of the true probabilities generating the data. The x-axis encodes the $S = 10$ possible categories, for each one the yellow point represents the true probability p_k, while the solid red line connects the estimated probabilities averaged over 100 iterations.



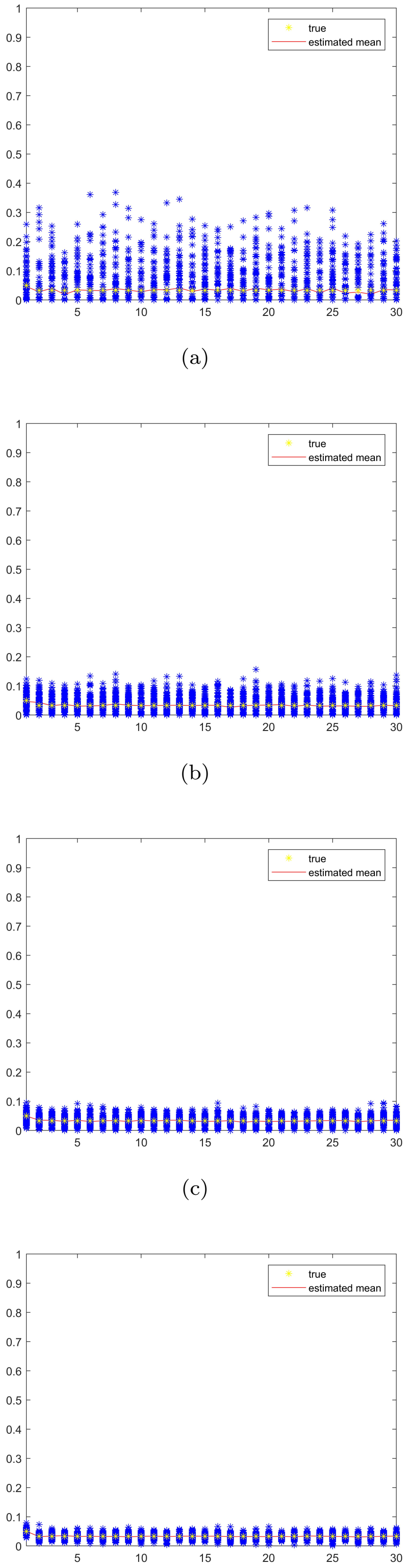
(a)

(b)

(c)

(d)

Fig. 3 Scenario III: estimates of the true probabilities generating the data. The x-axis encodes the $S = 30$ possible categories, for each one the yellow point represents the true probability p_k, while the solid red line connects the estimated probabilities averaged over 100 iterations.



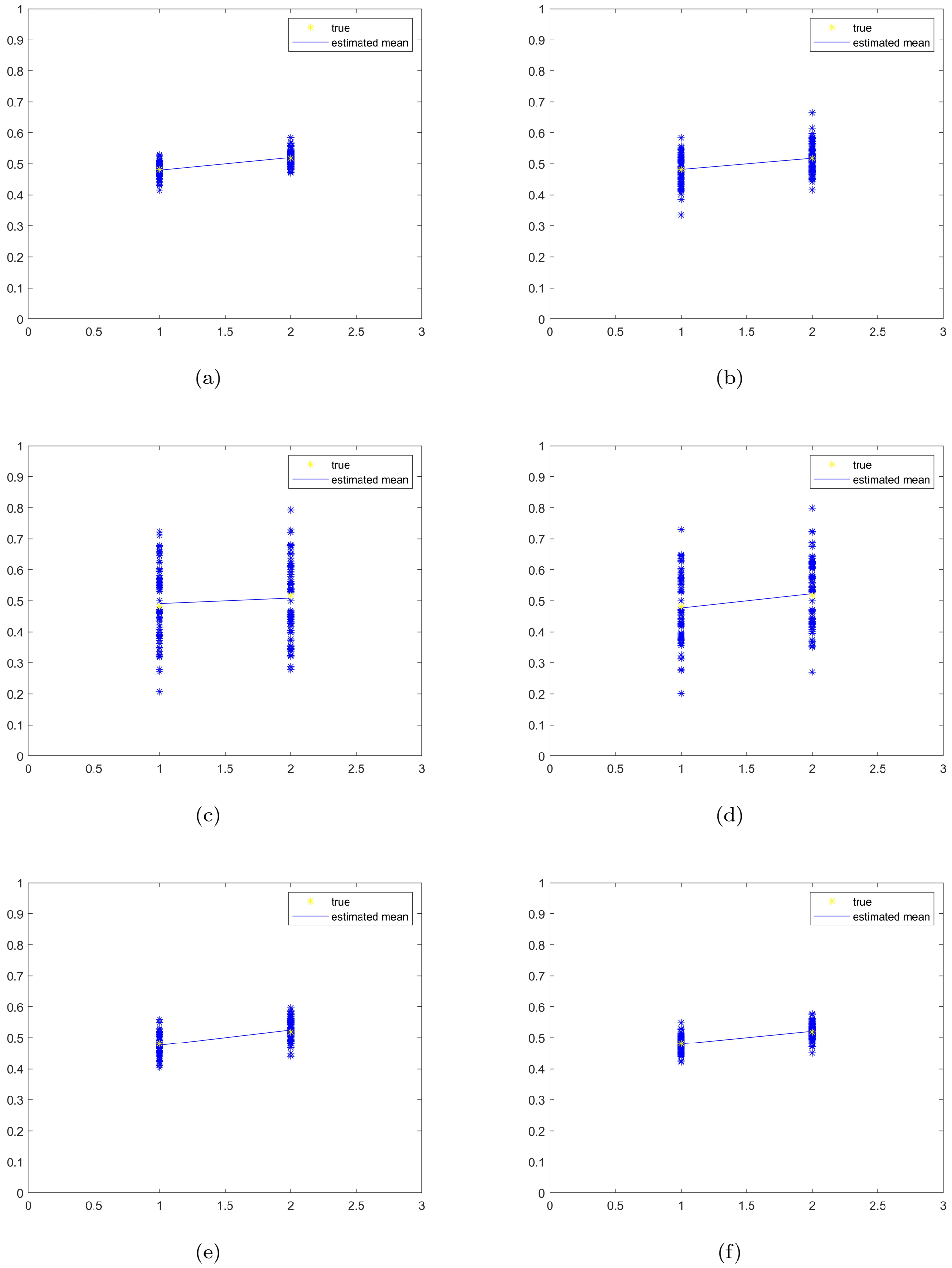
(a)

(b)

(c)

(d)

Fig. 4 Scenario IV: estimates of the true probabilities generating the data. The x-axis encodes the $S = 30$ possible categories, for each one the yellow point represents the true probability p_k, while the solid red line connects the estimated probabilities averaged over 100 iterations.



(a) (b)

(c) (d)

(e) (f)

Fig. 5 NY dataset: estimates of the true probabilities generating the data. The x-axis encodes the $S = 2$ possible categories (female or male), for each one the yellow point represents the true probability p_k, while the solid blue line connects the estimated probabilities averaged over 100 iterations.

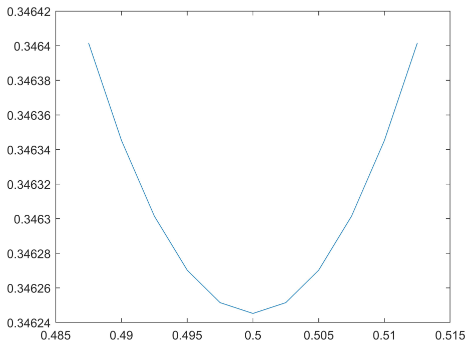


Fig. 6 NY dataset: mutual information as a function of q.