Estimating Joint Probability of Independently Randomized Multi-dimensional Data

Hiroaki Kikuchi[1,2,a)] Josep Domingo-Ferrer[2]

Abstract: Randomization of multi-dimensional data has several issues. Applying randomization to each attribute independently incurs an exponential grownup of possible values of composed domains that follows a huge computational time and correlations among attributes may be lost. In this paper, we show that the accurate estimation of joint probability distributions between attributes are feasible from the independently randomized multidimensional data. We show a simple but general scheme that computers composed randomization matrix and estimates the joint probabilities based on the inverse of randomization matrix. The estimation accuracy is evaluated in a model that takes a number of users, a number of values of attribute, a privacy budget and a correlation measure.

Keywords: local differential privacy, randomized response

1. Introduction

With widely spread of IoT devices, our daily activities are scanned and monitored in our digital society. Highly advanced smartphones keep scanning multi-dimensional vital data and help to suggest a healthy lifestyle. For example, Shen et al. [2] proposed method to aggregate high dimensional data to make better demand response for smart grid. Saint-Maurice et al. [1] found that a greater number of steps per day was significantly associated with a lower risk of all-cause mortality in adults in the US. These data are useful but often highly private data.

Local differential privacy (LDP) is a challenge for private data analysis, where data is locally anonymized before sending to data controller who estimates the frequency of responses accurately. Erlingsson et al. at Google proposed a LDP algorithm, Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR)[12]. Google Chrome extension uses RAPPOR to collect Windows process names and Chrome Homepages in privacy guaranteed manner.

Unfortunately, the curse of dimensionality does not allow the LDP approach to apply to multi-dimensional data. If we naïvely apply the randomization to every attribute independently, it leads to low-quality estimation because of the following reasons:

- (*Exponential growing up domain*) The number of values (categories) of the Cartesian product of multiple domains grows exponentially. The analysis of aggregated randomization matrix needs a high computational costs.
- (*Association loss*) Independent randomization of attributes violates the associations among attributes. The randomized values tend to be distributed uniformly and the highly associated pair of data may be hidden over the aggregated domains.
- (*Sparse domain*) The possible combination of values increase exponentially, while the number of individuals is constant. Therefore, the dense of records becomes lower with the increment of dimensions. The sparse randomized data incurs estimation accuracy loss.

To address the dimensionality issues, we propose a simple but general method to estimate joint probability accurately. Our proposed method, called RR-Ind-Joint, randomizes all attributes of multi-dimensional data *independently*, aggregates some randomization methods of subset of attributes *jointly* into a single randomization matrix for which the joint probability distributions are estimated. We show a simple way to compute the aggregated randomization matrix, which conveys conditional probabilities from the domain of original data to the randomized domain. The inverse of the aggregated matrix allows to estimate the original joint probability from the randomized data.

Our proposed method has some advantages against the existing schemes in terms of dimensionality issues. RR-Ind-Joint is efficient in terms of computation cost. The combined domains grows exponentially to the number of attributes m to aggregate, while the complexity of matrix inversion is linear to m. We will show it in Section 3.2.2. Hence, it can apply to high-dimensional data as far as a capacity allows.

RR-Ind-Joint is accurate in estimation of joint probabilities. The randomizations of attributes are performed independently, and hence the joint conditional probability can be obtained by the product of probabilities. Hence, the ag-

[1] Meiji University
[2] Universitat Rovira i Virgili
[a)] kikn@meiji.ac.jp

gregated multi-dimensional data *retains the association of attributes slightly*, which can be recovered by the inverse of aggregated randomization matrix. When we assume observing empirical probability in sufficient precision, the estimation error can be ignored. It estimates joint probability distributions of the original multi-dimensional data regardless of the the strength of associations among attributes.

Observed frequencies of values may slightly vary depending on randomization probability, the number of values in domains and the number of individuals. Hence, this causes a small estimation error. The error decreases as the size of survey increases and sufficient number of individuals participate. In order to guarantee the utility loss in privacy preservation, we propose a mathematical model with the above parameters to identify the magnitude of estimation error. We also conduct an experiment to verify our estimation of error using a synthesized data.

RR-Ind-Joint estimates joint probability based on the inverse of aggregated randomization matrix. We prove that an arbitrary number of dimensions is possible to be aggregated in Corollary 1 and guarantee that the aggregated randomization matrix is always non-singular in Section 3.2.2. Hence, it apples high-dimensional data even if the combined domain is sparse.

RR-Ind-Joint guarantees the differential privacy for specified privacy budget. Since attribute randomizations are performed independently, the whole privacy budgets can be simply given the sum of all budgets due to the sequential compositional theorem of LDP [11].

Our contributions of this work are as follows:

- We propose a new LDP scheme RR-Ind-Joint that deals with multi-dimensional data and estimates the joint probability distribution from the observed frequencies of the randomized data.
- We prove some useful properties of the proposed schemes that can be applied to arbitrary number of dimension, the estimation error does not depends on the strength of the association of the original multi-dimensional data.
- We show the experimental results using synthesized data and the open data to evaluate the estimation loss in terms of several parameters, the correlation among attributes, the size of domain, and the number of individuals.

The rest of our paper is organized as follows. In Section 2, we give some fundamental definitions and review some existing works related to the multi-dimensional anonymization. Section 3 proposes our scheme, provides an algorithm for estimation. We also discuss the privacy and and the primary factor of error of the estimation error. After dementing an example of toy data, we report the experimental results using some synthesized and open data in Section 4 to verify that our model for estimation error as claimed. In Section 5, we conclude our study based on the proved theorem and the experimental results.

Algorithm 1 Randomization RR(X)

```
$x_i \leftarrow$ input of party $i$ for attribute $X$.
$P \leftarrow$ a randomization matrix for attribute $X$.
for all Party $i = 1, \ldots, n$ do
    $y_i \leftarrow \begin{cases} x_i & \text{w.p. } = p_{uu},\ x_i \text{ is } u\text{-th element} \\ v & \text{w.p. } = p_{uv} = q \end{cases}$
end for
return the randomized response $y_1, \ldots, y_n$.
```

2. Fundamental Definitions

2.1 Randomized Response

Definition 1 Let X be a set of d elements. A $d \times d$ matrix of probabilities

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1d} \\ \vdots & \ddots & \vdots \\ p_{d1} & \cdots & p_{dd} \end{pmatrix},$$

is a *randomization matrix* of X if and only if $p_{11}+\cdots+p_{dd} = 1$ and p_{uv} is a conditional probability of a randomized element to be v given u of X i.e., $p_{uv} = Pr(Y = v|X = u)$ for all $u, v \in \{1, \ldots, d\}$.

A randomized response (RR) is a randomized mechanism that input X of d possible values $a_1, \ldots, a_d$ is randomized to the response Y according to P. By $Y = \mathsf{RR}_P(X)$, we denote an algorithm defined in Algorithm 1. The goal of RR is to estimate the frequency of a of X, which can be given the most likelihood estimation as $\frac{f_{Y=a}/n-q}{p-q}$, where n is the number of responders.

More correctly, letting $\pi_1, \ldots, \pi_d$ be proportions of respondents whose true values fall in each of d values of X and λ_a be the empirical probabilities of the observed Y being v, we can write $(\lambda_1, \ldots, \lambda_d)^T = P^T(\pi_1, \ldots, \pi_d)^T$. According to [9], an unbiased estimator π can be computed as

$$\hat{\pi} = (P^T)^{-1}\hat{\lambda},$$

where $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_d)^T$ is the vector of observed empirical probabilities of Y.

2.2 Multi-Dimensional RR

RR works well for a single attribute. However, when multiple attributes $X^1, \ldots, X^m$ are given, the estimation is not so trivial because of the curse of dimensionality. Applying RR simultaneously to all attributes suffers the exponentially grown number of possible values in the Cartesian product of attributes. Alternatively, applying RR independently to each attribute may loss the association between the attributes. To overcome the dimensionality issue, [3] proposed two basic protocols RR-Independent and RR-Joint as follows.

2.2.1 RR-Independent

Each party apply RR independently for m attributes $X^1, \ldots, X^m$ in a dataset simply as $Y = (Y^1, \ldots, Y^m)$, where $Y^j = \mathsf{RR}_{P^j}(X^j)$. After estimating the marginal probabilities for j-th attribute as $\hat{\pi}^j = ({P^j}^T)^{-1}\hat{\lambda}^j$, the joint probability distribution for $X^1, \ldots, X^m$ is estimated by the

Algorithm 2 Estimation from RR-Independent $\mathsf{RR\text{-}Ind}(Y)$

```
1: $\hat{\lambda}^j \leftarrow$ observed empirical probability for attribute $A^j$.
2: for all $j = 1, \ldots, m$ do
3:     estimate the marginal probability $\hat{\pi}^j \leftarrow ((P^j)^T)^{-1}\hat{\lambda}^j$
4: end for
5: return the joint probability of $A^1 \times \cdots \times A^m$ estimated as $\hat{\Sigma}^{(1,\ldots,m)}_{\mathsf{RR-ind}} \leftarrow \hat{\pi}^1(x^1)\ldots\hat{\pi}^m(x^m)$
```

product of the each marginal distributions as

$$\Sigma^{(X_1,\ldots,X_m)}_{\mathsf{RR-Ind}}(a_1,\ldots,a_m) = \hat{\pi}^1(a_1)\cdots\hat{\pi}^m(a_m).$$

Algorithm 2 show the steps.

2.2.2 RR-Joint

Regarding values of m attributes as a single tuple $(a_1, \ldots, a_m)$, parties apply a single RR to the tuple using a $d_1 \times \cdots \times d_m$ randomization matrix for attributes $X^1, \ldots, X^m$, respectively. The joint probability $\hat{\Sigma}^{(X_1,\ldots,X_m)}_{\mathsf{RR-Joint}}$ can be estimated as $(P^T)^{-1}\hat{\lambda}^{X_1,\ldots,X_m}$.

According to [3], RR-Joint has some drawbacks; (1) the number of values (categories) of the Cartesian product grows exponentially. (2) the inverse of the randomization matrix needs a high computational costs. (3) the number of users n that is less than the product of of m domains of attributes such that $n < |X^1|\cdots|X^m|$ incurs estimation accuracy loss.

To overcome the issues, several multi-dimensional RR schemes were proposed in [3], including RR-Clusters splitting attributes into clusters so that the independence between clusters can be assumed with reasonable computational cost.

2.3 Related Works

Ren et al. studied a LDP scheme called LoPub, estimating multi-dimensional joint probability distributions in [4]. They perturbs a multi-dimensional data encoded binary vectors using Bloom filter (similar to RAPPOR [12]) and combine a Lasso regression with an Expectation Maximization to estimate the joint probabilities accurately. They also show a synthesized data that preserves the utility of the original data in the sense that classification accuracies for some machine learning algorithms are preserved as original.

Ye et al. proposed the LDP protocol designed for a key-value data in [5]. Their scheme encodes a numerical data into discrete values according to a probability which is proportional to the value and then perturbs the key jointly with the encoded value using a variation of randomized response. The associations between key and value are preserved from the randomized pairs with the privacy of input is guaranteed in a specified privacy budget. Their scheme is classified as a two-dimensional RR with nominal (key) and numerical (value) attributes.

3. Proposed Method

3.1 Idea

Dimensional issues in RR exist when we consider all m joint probability distributions. However, most use-cases do not need the association among *all attributes* and two or three associations are useful enough for many cases. For example, a key-value pair (2 dimension) has been used in variety of applications with SQL database. With a limited number of attributes, it is not so hard to perform inverse of randomization matrix in terms of computational time or memory consumption. We demonstrate that the 3-way joint probabilities of "Adult" dataset can be estimated accurately in this work.

Our method guarantees the differential privacy as same as RR-Independent, where the privacy budget is given the total sum of m privacy budgets, $m\epsilon$, due to the sequential compositional theorem of LDP [11]. We perturb data as same as RR-Independent, but estimate similar to RR-Joint. Hence, we call the method as RR-Ind-Joint.

One of the concern of multi-dimensional RR is the estimation error. The estimation accuracy may be reduced with a complex conditions, e.g., when the number of users n is not sufficient for the product of domains, when the privacy budget ϵ is too small to preserve the properties, when the number of of values d in attributes is too large, and so on. To quantify the estimation error, we prove some useful properties to bound the error in some simple model with assumptions.

3.2 RR-Ind-Joint

3.2.1 Randomization

Suppose that all m attributes are independently randomized for m randomization matrices $P^1, \ldots, P^m$. To estimate joint probability distributions, we need to aggregate all independent randomizations into a unified matrix in some way. The following theorem shows the composition of two random matrices.

Theorem 1 Let P^i and P^j be $d^i \times d^i$ and $d^j \times d^j$ randomization matrices for attributes A^i and A^j, respectively. The $(d^i d^j) \times (d^i d^j)$ matrix of defined as Kronecker product $P^i \otimes P^j$ is a randomization matrix for Cartesian product of domains A^i and A^j, that is, p_{uv} of $P^i \otimes P^j$ is a conditional probability $p_{uv} = Pr[(Y^i, Y^j) = u|(X^i, X^j) = v]$.

Proof: Let u and v be tuples of attributes A^i and A^j such that $u = (y^i, y^j)$ and $v = (x^i, x^j)$. From the premises, for attributes A^i and A^j, $p^i_{x^i y^i} = Pr[Y^i = y^i|X^i = x^i]$ and $p^j_{x^j y^j} = Pr[Y^j = y^j|X^j = x^j]$ hold. According to the definition of Kronecker product, we have the $(d^i d^j) \times (d^i d^j)$ matrix as

$$P^i \otimes P^j = \begin{pmatrix} p_{11}P^j & \cdots & p_{1d^i}P^j \\ \vdots & \ddots & \vdots \\ p_{d^i 1}P^j & \cdots & p_{d^i d^i}P^j \end{pmatrix},$$

where element p_{uv} is $p^i_{y^i x^i} \cdot p^j_{y^j x^j}$, which is equal to the joint probability of u and v because two randomizations are independently performed as

$$Pr[(Y^i, Y^j) = u|(X^i, X^j) = v]$$
$$= Pr[Y^i = y^i|X^i = x^i]Pr[Y^j = y^j|X^j = x^j].$$

□

Algorithm 3 Estimation in RR-Ind-Joint

```
$Y^j \leftarrow \textsf{RR-ind}_{P^j}(X^j)$ for $j = 1, \dots, m$.
$\hat{\lambda}^{Y^1,\dots,Y^m} \leftarrow$ observed empirical probability for attributes $(Y^1, \dots, Y^m)$.
return the joint probability estimated as $\hat{\Sigma}^{(1,\cdots,m)} = ((P^{i_1} \otimes \cdots \otimes P^{i_k})^T)^{-1}\hat{\lambda}^{Y^1,\dots,Y^m}$.
```

Due to the associativity of Kronecker products, i.e., $(A \otimes B) \otimes C = A \otimes (B \otimes C)$, we can easily extend the randomization matrix to multi-dimensional attribute.

Corollary 1 Let $P^{i_1}, \dots, P^{i_k}$ be randomization matrices for k attributes $A^{i_1}, \dots, A^{i_k}$. A matrix defined by $P^{i_1} \otimes \cdots \otimes P^{i_k}$ is a randomization matrix for $|A^{i_1}| \times \cdots \times |A^{i_k}|$ where $|A^i|$ is a domain of attribute A^i.

Proof: It is straightforward by recursively applying Theorem 1 to every two attributes in turn. □

3.2.2 Estimation

We consider to estimate the joint probability of attributes A^i and A^j from the independently randomized response $Y^1, \dots, Y^m$.

Algorithm 3 shows the procedure for estimate the joint probabilities of k attributes from independently randomized responses $Y^1 \dots Y^k$. Note that the Cartesian product of k attributes is in the order consistent with that of $(P^{i_1} \otimes \cdots \otimes P^{i_k})$.

The inverse of Kronecker product can be given as

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1},$$

which stats that the composed randomization matrix $P^{i_1} \otimes \dots \otimes P^{i_k}$ is non-singular if and only if $P^{i_1}, \dots, P^{i_k}$ are non-singular. It also means that the cost for computing matrix inversion is linear to the number of attributes to be aggregated because and the Kronecker product of all separate inverses gives the inverse of the aggregated matrix. Computing the inverse of randomization matrix does not need any record and hence it can be preprocessed. Each randomization matrix is determined only by the size of domain d and the privacy budget ϵ. Hence, some matrices can be identical and help to save computational cost. Therefore, our proposed method is efficient in terms of computational cost.

3.3 Example

Consider a dataset X on $n = 10$ parties with two attributes A and B having domain $|A| = \{a_1, a_2\}$ and $|B| = \{b_1, b_2\}$. The empirical (true) joint probability distribution of X is as follows:

$$\begin{aligned}
\Sigma_{AB}(a_1, b_1) &= 4/10, \\
\Sigma_{AB}(a_2, b_1) &= 2/10, \\
\Sigma_{AB}(a_1, b_2) &= 0, \\
\Sigma_{AB}(a_2, b_2) &= 4/10.
\end{aligned}$$

This yields marginal distributions $\pi_A = (0.4, 0.6)$ and $\pi_B = (0.6, 0.4)$. We denote frequencies of X by 2×2 matrix as

$$f^X = \begin{pmatrix} 4 & 0 \\ 2 & 4 \end{pmatrix},$$

which indicates frequencies of $(a_1, b_1), (a_2, b_1), (a_1, b_2), (a_2, b_2)$ of X, respectively.

With $p_A = p_B = 3/4$, we have randomization matrices for A and B as

$$P^A = \begin{pmatrix} p_A & q_A \\ q_A & p_A \end{pmatrix} = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix} = P^B,$$

for which n parties randomize their two responses x_i^A, x_i^B independently. Suppose that the randomized $Y^A = \textsf{RR}_{P^A}(X^A)$ and $Y^B = \textsf{RR}_{P^B}(X^B)$ are observed as

$$f^Y = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix},$$

for which the empirical probabilities of Y are $\lambda^A = (0.5, 0.5)$ and $\lambda^B = (0.5, 0.5)$. Note that V statistics of A and B is $V_{AB}(Y) = 0.2$, which is reduced from that of the original dataset $V_{AB}(X) = 0.67$. Clearly, the correlation between A and B is partially reduced by independent randomizations.

RR-Ind estimates the joint probabilities as the product of estimated marginal distributions $\hat{\pi}^A$ and $\hat{\pi}^B$ as

$$\hat{\Sigma}^{AB}_{\textsf{RR-Ind}} = \begin{pmatrix} 0.24 & 0.36 \\ 0.16 & 0.24 \end{pmatrix},$$

which preserves the marginal distributions $\hat{\pi}^A_{\textsf{RR-ind}} = (0.4, 0.6) = \pi^A$ and $\hat{\pi}^B_{\textsf{RR-ind}} = (0.6, 0.4) = \pi^B$, but fails estimating the joint probabilities correctly. It has MAE$= 0.16$ and $V_{AB}(\hat{\Sigma}^{AB}_{\textsf{RR-Ind}}) = 1 \times 10^{-19} (= 0)$.

RR-Ind-Joint regards independent two randomizations as joint processing of

$$\lambda^{AB} = (P^A \otimes P^B)\pi^{AB},$$

where $P^A \otimes P^B$ is the aggregated randomization matrices

$$\begin{pmatrix} p_a p_b & p_a q_b & q_a p_b & q_a q_b \\ p_a q_b & p_a p_b & q_a q_b & q_a p_b \\ q_a p_b & q_a q_b & p_a p_b & p_a q_b \\ q_a q_b & q_a p_b & p_a q_b & p_a p_b \end{pmatrix} = \frac{1}{16} \begin{pmatrix} 9 & 3 & 3 & 1 \\ 3 & 9 & 1 & 3 \\ 3 & 1 & 9 & 3 \\ 1 & 3 & 3 & 9 \end{pmatrix}.$$

Given the observed the empirical distributions of Y as $\lambda^{AB}(a_1, b_1) = 3/10$, $\lambda^{AB}(a_2, b_1) = 2/10$, $\lambda^{AB}(a_1, b_2) = 2/10$, $\lambda^{AB}(a_2, b_2) = 3/10$, we estimate the joint probability distribution of two randomized attributes as

$$\hat{\Sigma}^{AB}_{\textsf{RR-Ind-Joint}} = (P^A \otimes P^B)^{-1}\lambda^{AB} = (0.375, 0.075, 0.275, 0.375)$$

The estimated joint probabilities have smaller error than that of RR-Ind, i.e., MAE(RR-Ind-Joint) $= 0.05$ and and the correlation between attributes is $V^{AB}(\hat{\Sigma}^{AB}_{\textsf{RR-Ind-Joint}}) = 0.8$, which is close to the original $V_{AB}(X) = 0.67$. The estimation error is caused by rounding frequencies of RR when $n = 10$. It is improved as MAE $= 0.01$ when $n = 100$.

3.4 Privacy

The privacy of the schema RR-ind-joint is that of RR-Ind. Suppose a simple RR that responses x with probability of $p = \frac{e^{epsilon}}{e^{\epsilon}+d-1}$ and with probability $q = 1 - p = \frac{d-1}{e^{\epsilon}+d-1}$ responses a randomly chosen value in $|A|$.

Theorem 2 RR-Ind-Joint satisfies $(\epsilon, 0)$-LDP for attribute A. With m attributes $A^1, \ldots, A^m$, RR-Ind-Joint satisfies $(m\epsilon, 0)$-LDP.

Proof: For any $x, x' \in |A|$ such that $x \neq x'$, and any $y \in |A|$

$$\frac{Pr[RR(x) = y]}{Pr[RR(x') = y]} = \frac{p}{q} = e^{\epsilon}$$

Because m attributes are perturbed independently, the sequential decomposition theorem stats that RR-Ind-Joint satisfies $(m\epsilon, 0)$-LDP. □

3.5 Estimation Error

We evaluate the accuracy loss by means of MAE (Mean Absolute Error) of the estimated joint probability distributions, defined as $MAE = 1/d^2 \sum_{x \in |A| \times |B|} |\Sigma^{AB}(x) - \hat{\Sigma}^{AB}(x)|$ for several environments.

First, we show the bound of MAE of RR-Independent.

Theorem 3 Let A and B be two attributes with the Cramer's V statistics V and the same number of values in both domains, $d = |A| = |B|$. MSE of RR-Ind is less than V^2/d.

Proof: The definition of V statistics is $V = \sqrt{\chi^2/n(d-1)}$. Squaring and dividing both sides by d, we have

$$\begin{aligned} V^2/d &= \frac{\chi^2/n}{d(d-1)} \leq \frac{1}{nd^2} \sum_{i=1}^{d^2} \frac{(o_i - e_i)^2}{e_i} \\ &= \frac{1}{d^2} \sum_{a \in |A|, b \in |B|} \frac{(o_{(a,b)}/n - \hat{\lambda}_a \hat{\lambda}_b)^2}{\hat{\lambda}_a \hat{\lambda}_b} \\ &\leq \frac{1}{d^2} \sum_{a \in |A|, b \in |B|} (\Sigma^{AB}(a,b) - \hat{\lambda}_a \hat{\lambda}_b)^2 \\ &= MSE(\Sigma^{AB}). \end{aligned}$$

Note that expected value e_i is the mean of binomial distribution of $p = \Sigma^{AB}_{RRInd}$ and n trials, i.e., $np = n\hat{\lambda}^A(a)\hat{\lambda}^B(b)$. The last inequality holds when $\hat{\lambda}_a \hat{\lambda}_b \leq 1.0$. □

Taking squared root of the both sides, we estimate that MAE of RR-Independent is proportional to $V/\sqrt{d}$

MAE of RR-Ind-Joint does not depend on V because it estimates the joint probability of attributes by the inverse of randomization matrix. RR-Ind-Joint has no estimation error as far as all randomization matrices for attributes are non-singular. Differential private randomization matrix with $p = \frac{e^{\epsilon}}{e^{\epsilon}+d-1}$, $q = \frac{1}{e^{\epsilon}+d-1}$ becomes singular only when $\epsilon = 0$ and $p = 1/d$. It is not hard to avoid the trivial case $\epsilon = 0$, we confirm that all randomization matrices are non-singular.

However, the estimation of RR-Ind-Joint suffers a rounding error of empirical probability distribution $\lambda^{AB}(Y)$. The observed probability of (a, b) of Y is the fraction of users who send (a, b) of n users. Hence, the precision of of empirical probability $\lambda^{AB}(Y)$ is $1/n$. Let us remind the estimation error in the example in Section 3.3, where MAE= 0.05 when $n = 10$. We consider a model of empirical probability as the form,

$$\hat{\lambda} = \lambda + \Delta\lambda,$$

where $\Delta\lambda$ is the rounding error. The example in Section 3.3 has the instance,

$$\hat{\lambda}^{AB} = \begin{pmatrix} 3/10 \\ 2/10 \\ 2/10 \\ 3/10 \end{pmatrix} = \begin{pmatrix} 0.2875 \\ 0.1625 \\ 0.2625 \\ 0.2875 \end{pmatrix} + \begin{pmatrix} -0.0125 \\ -0.0375 \\ +0.0625 \\ -0.0125 \end{pmatrix}$$

where the last vector is the rounding error $\Delta\lambda$ for $n = 10$.

We regard a rounding error as a uniform distribution over $[-1/n, 1/n]$, which holds $E[\Delta\lambda] = 0$ and $E[|\Delta\lambda|] = 1/2n$. With the model, the estimation of joint probability is

$$\hat{\Sigma} = P^{-1}\hat{\lambda} = P^{-1}(\lambda - \Delta\lambda) = \Sigma - P^{-1}\Delta\lambda.$$

The last term of the above formula is the source of MAE of RR-Ind-Joint. It is the linear combination of d^2 uniform distribution and can be approximated as normal distribution with the mean increasing with $1/n$. It is not trivial to estimate $P^{-1}\Delta\lambda$ because the inverse of aggregated random matrix has elements distributed widely than $[0, 1]$. We say that it increases as n decreases, as d increases, and as p decreases.

4. Evaluation

4.1 Data

To quantify utility loss in processing RR and estimating, we synthesize a dataset with two attributes A and B having marginal probability $\lambda^A = \lambda^B$ distributed in $Pr(A = a) = c/a$ for $a = 2, \ldots, d$ and a constant $c = 1/(\sum_{a=2}^{d} 1/a)$. The domain of attribute A is denoted by $|A| = \{c/2, \ldots, c/d\}$, where d is the number of unique values in attribute A. The correlation between attributes is controlled for Cramer's V statistics $v = V_{AB} \in [0, 1]$.

Figures 1, 2, 3 and 4 show the joint probability distributions A and B with $n = 1000, d = 10, v = 0.5$, for the synthesized data $\Sigma^{AB}(X)$, the perturbed data $Y = \mathsf{RR}(X)$ with $\epsilon = 1$ $\lambda^{AB}(Y)$, the estimated probability by RR-Ind $\hat{\Sigma}^{AB}_{\mathsf{RRInd}}(X)$ and the estimated probability by RR-Ind-Joint $\hat{\Sigma}^{AB}_{\mathsf{RRIndJoint}}(X)$, respectively. We observe that the joint probability of the given data X with Cramer's V of $v = 0.5$ has strong correlation at the diagonal elements in the Cartesian product $|A| \times |B|$, which is distributed widely in the perturbed data Y. The RR-Ind fails to estimate the strong correlation between two attributes in $\hat{\Sigma}^{AB}_{\mathsf{RRInd}}(X)$. While, the RR-Ind-Joint estimates the joint probabilities more accurately in Figure 4. Estimated probabilities are not exactly same to that of the original X because the precision of the empirical distribution λ^{AB} depends on environmental parameters, e.g., the number of individuals n, the size of

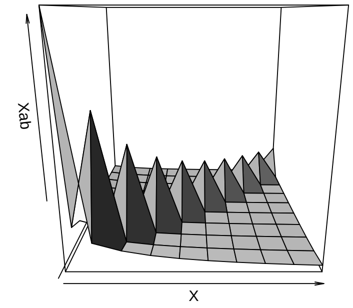


Fig. 1 Synthesized data $\Sigma^{AB}(X)$

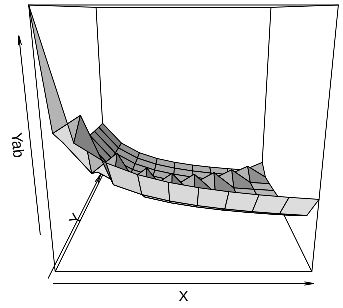


Fig. 2 Perturbed data $\lambda^{AB}(Y)$

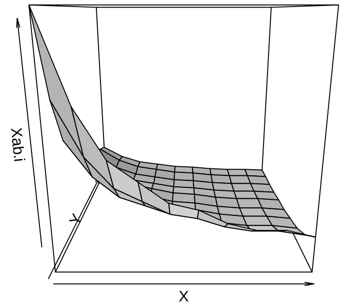


Fig. 3 Estimated data RR-Ind $\hat{\Sigma}^{AB}_{\mathsf{RRInd}}(X)$

attribute domain (the number of unique values) d, privacy budget ϵ and the correlation of two attributes. Hence, we evaluate the accuracy loss in terms of these parameters.

4.2 Results (synthesized data)

Figures 5, 6, 5 and 8 show the MAE with regards to Cramer's V statistics $v \in [0, 1]$, the MAE with regards to privacy budget $\epsilon = 0.1, \ldots, 2$, the MAE with regards to the number of individuals $n = 10, 100, 1000, 10000$, the MAE

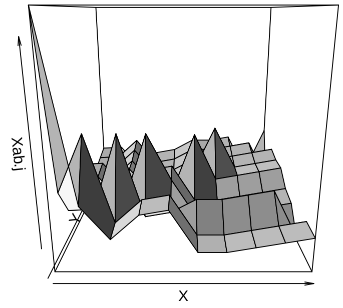


Fig. 4 Estimated data RR-Ind $\hat{\Sigma}^{AB}_{\mathsf{RRIndJoint}}(X)$

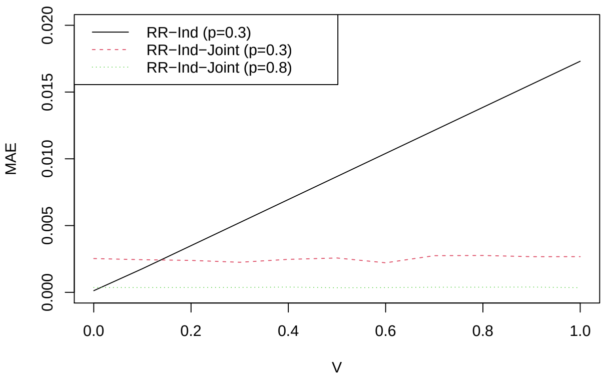


Fig. 5 MAE with regards to correlation v

with regards to the size of domains (the number of unique values in attribute) $d(= |A| = |B|) = 2, \ldots, 20$, respectively.

Figure 5 shows that the estimation error of RR-Ind depends on the correlation between attributes. MAE is proportional to V with two extreme cases; 0 when A and B are independent ($V = 0$) and the highest when A completely depends on B ($V = 1$). RR-Independent estimates the joint probability with the product of two marginal probabilities as $\hat{\Sigma}^{AB}(a, b) = \hat{\sigma}^{A}(a)\hat{\sigma}^{B}(b)$ under an assumption of independent attributes, for which $V = 0$. Hence, the estimation error is linear to V that is considered as a fraction of independent pair of values (a, b) over $d \times d$ pairs. While, MAE of RR-Ind-Joint does not depend on V. It estimates joint probabilities accurately whatever attributes are correlated.

Figure 6 shows that MAE of RR-Ind-Joint decreases as privacy budget ϵ increases, which follows in turn the increases of the probabilities of retaining. It also indicates that MAE of RR-Independent is constant because the primary part of estimation error is caused by the strength of correlation between attributes and the effect of privacy budget is hide of the error.

MAE of RR-Ind-Joint depends on the number of users n and the size of domains $d = |A|$. We observe the reduction of MAE with n in Figure 7. MAE of RR-Ind-Joint decreases in the order of $1/n$ when $n > 1000$. MAE also tends to

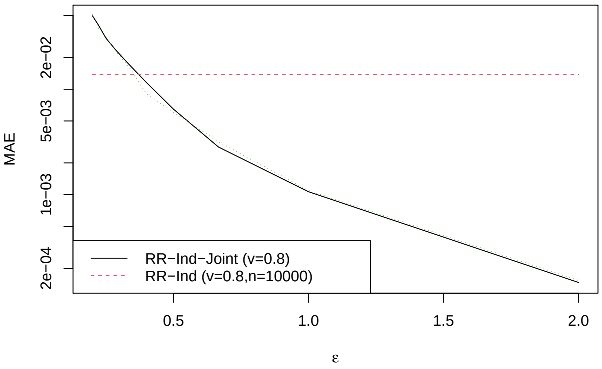


Fig. 6 MAE with regards to privacy budget ϵ

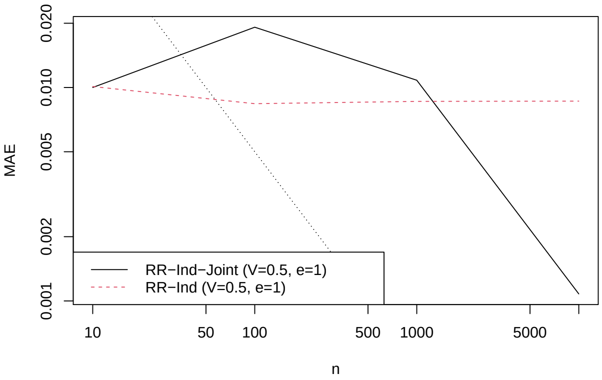


Fig. 7 MAE with regards to number of individuals n

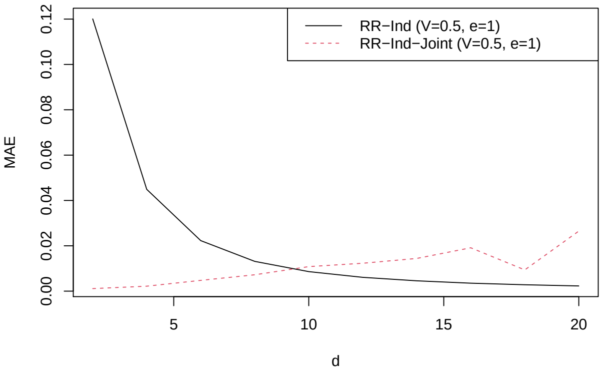


Fig. 8 MAE with regards to the size of domain $d(= |A|)$

increase with d in Figure 8. Therefore, we claim that RR-Ind-Joint estimation needs sufficient number of users and has a limitation of number of attributes to aggregate.

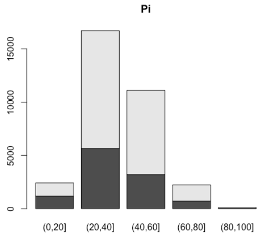
The reduction of MAE with increasing d is consistent with Theorem 3 that stats MAE is linear to $1/\sqrt{d}$.

4.3 Result (Adults)

Table 1 shows the MSE in some two attributes in UCI Adult dataset, where $n = 32561$ and probability to retain data $p = 0.5$.

Table 1 Mean Squared Error in Adult data

	sex	income	sex	race	education	occupation
d	2	2	2	5	16	15
V	0.2159802		0.1181155		0.1873341	
RR Ind	0.00188		0.00011		2.14×10^{-5}	
RR Ind Joint	2.93×10^{-33}		3.61×10^{-33}		5.48×10^{-33}	



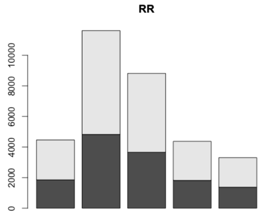
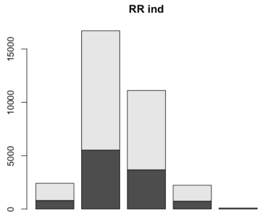


Fig. 9 Histogram of “Age” and “sex”

Fig. 10 Randomized Y



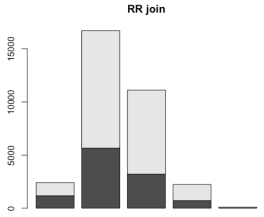


Fig. 11 Estimated $\hat{\Sigma}_{\mathsf{RRInd}}$

Fig. 12 Estimated $\hat{\Sigma}_{\mathsf{RRIndJoint}}$

Fig 9 shows the frequency distributions of male (light) and female (dark) and age categorized in 20-years-old bins. The continuous data are quantified into five categories in this experiment. Figs. 10, 11, and 12 show the joint frequency distributions performed by RR(X), RR-Independent and RR-Ind-Joint, respectively. The estimation of RR-Ind-Joint is close to the original distribution Σ.

5. Conclusion

In this paper, we have studied the randomization of a multi-dimensional data. Due ot the curse of dimensionality, naïvely independent randomization of attributes suffers several issues, e.g., the combination of domain grows up exponentially, it violates the association among attributes, and there are too small records to cover the combined domain, i.e., sparse domain issue.

Our proposed method RR-Ind-Joint addresses the dimensionality issues, by aggregating independently randomized matrix into a single randomization matrix and estimating the joint probability distribution of the original attribute by inverse of the aggregated randomization matrix. We show some useful theorems that guarantees that the proposed scheme can be applied to arbtrary number of dimensions regardless of the strength of association among attributes. The estimation is quite accurate in comparison with the one of the existing scheme. Our experiments using synthesized and open data verified that the estimation error of RR-Ind-Joint does not depend on the association between attributes.

We plan to compare the estimation accuracy with some of the state-of-arts schemes in multi-dimensional LDP schemes as one of future works.

Acknowledgment

This work was supported by JST, CREST Grant Number JPMJCR21M1, Japan.

References

[1] Pedro F. Saint-Maurice, et al. “Association of Daily Step Count and Step Intensity With Mortality Among US Adults,” JAMA, 323(12) pp.1151-1160, 2020.

[2] H. Shen, M. Zhang, and J. Shen, “Efficient privacy-preserving cubedata aggregation scheme for smart grids,” IEEE Trans. Inf. Forensics and Security, vol. 12, no. 6, pp. 1369-1381, 2017.

[3] J. Domingo-Ferrer and J. Soria-Comas, “Multi-Dimensional Randomized Response,” in IEEE Transactions on Knowledge and Data Engineering, 2022, `doi:10.1109/TKDE.2020.3045759`.

[4] X. Ren et al., “LoPub : High-Dimensional Crowdsourced Data Publication With Local Differential Privacy,” in IEEE Transactions on Information Forensics and Security, vol. 13, no. 9, pp. 2151-2166, Sept. 2018, `doi:10.1109/TIFS.2018.2812146`.

[5] Q. Ye, H. Hu, X. Meng, H. Zheng, “PrivKV : Key-Value Data Collection with Local Differential Privacy”, IEEE S&P, pp. 294-308, 2019.

[6] C. Dwark, F. McSherry, K. Nissim, A. Smith, “Calibrating noise to sensitivity in private data analysis,” TCC, Vol. 3876, p. 265-284, 2006.

[7] J. C. Duchi, M. I. Jordan, M. J. Wainwright, “Local privacy and statistical minimax rates,” FOCS, pp. 429-438, 2013.

[8] P. Kairouz, S. Oh, and P. Viswanat, “Extremal mechanisms for local differential privacy”, NIPS, pp. 2879-2887, 2014.

[9] S. L. Warner, “Randomized response: A survey technique for eliminating evasive answer bias”, Journal of the American Statistical Association, pp. 63-69, 1965.

[10] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, J. Shin, “Collecting and analyzing data from smart device users with local differential privacy”, `arXiv:1606.05053`, 2016.

[11] F. McSherry, “Privacy integrated queries: An extensible platform for privacy-preserving data analysis”, SIGMOD, pp. 19-30, 2009.

[12] Úlfar Erlingsson, Vasyl Pihur, Aleksandra Korolova, “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”, ACM Conference on Computer and Communications Security, pp.1054-1067, 2014.

[13] “Learning with Privacy at Scale” `https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html` (accessed on 2019).

[14] Kairouz, P., Oh, S., Viswanath, P., “The Composition Theorem for Differential Privacy” Proceedings of the 32nd International Conference on Machine Learning, 37, pp. 1376-1385, 2015.

[15] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy”, Found. Trends Theor. Comput. Sci. 9, 3-4, 211-407, 2014.