Econometrics and Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Differentially Private Goodness-of-Fit Tests for Continuous Variables

Seung Woo Kwak[a], Jeongyoun Ahn[b], Jaewoo Lee[c], Cheolwoo Park[d,*]

[a] Department of Statistics, Seoul National University, Seoul, 08826, Republic of Korea
[b] Department of Industrial & Systems Engineering, KAIST, Daejeon, 34141, Republic of Korea
[c] Department of Computer Science, University of Georgia, Athens, GA 30602, USA
[d] Department of Mathematical Sciences, KAIST, Daejeon, 34141, Republic of Korea

ARTICLE INFO

Article history:
Received 29 January 2021
Revised 28 September 2021
Accepted 30 September 2021
Available online xxx

Keywords:
Continuous random variables
Differential privacy
Discretization
Goodness-of fit-test

ABSTRACT

Data privacy is a growing concern in modern data analyses as more and more types of information about individuals are collected and shared. Statistical analysis in consideration of privacy is thus becoming an exciting area of research. Differential privacy can provide a means by which one can measure the stochastic risk of violating the privacy of individuals that can result from conducting an analysis, such as a simple query from a database and a hypothesis test. The main interest of the work is a goodness-of-fit test that compares the sampled data to a known distribution. Many differentially private goodness-of-fit tests have been proposed for discrete random variables, but little work has been done for continuous variables. The objective is to review some existing tests that guarantee differential privacy for discrete random variables, and to propose an extension to continuous cases via a discretization process. The proposed test procedures are demonstrated through simulated examples and applied to the Household Financial Welfare Survey of South Korea in 2018.

1. Introduction

Each day large quantities of data are collected, analyzed, and released by many institutions, organizations, or companies. Researchers or analysts use these data to find patterns, construct models to understand underlying phenomena, and predict the future. Accumulation of data at a large scale helps to build better models and make more accurate predictions, but the risk of leaking private sensitive information is a major concern.

For instance, large corporations and tech companies regularly collect user data to aid in improving users' experience and providing personalized services. However, such data often contain sensitive information, creating a situation where privacy violations can occur. One might think that sanitization, a process of removing sensitive private information from data, is sufficient to protect privacy. Nevertheless, it has been known that the anonymization may not be sufficient to keep private information from adversaries. Narayanan and Shmatikov (2008) demonstrated that it is possible to identify Netflix subscribers based on their reviews and information that can be found from the Internet Movie Database (IMDb). This linkage attack is possible because those two websites provide similar information about the users. Another example of privacy breach is given in Sweeney (2013), which used three datasets to identify individuals: the hospitalization database of Washington state in 2011, a collection of news stories in the state in 2011, and an online public records service for basic

* Corresponding author.:
E-mail address: parkcw2021@kaist.ac.kr (C. Park).

https://doi.org/10.1016/j.ecosta.2021.09.007

demographics on Americans. They made guesses on 81 individuals, 43% of which were correct. These two examples show that the anonymization of data before release is not enough to maintain the privacy of individuals when auxiliary data that contain similar information are available.

Reiter (2004) pointed out that statistical agencies and organizations should disseminate data that are safe from attacks yet still informative and ready for statistical analyses. Providing raw data would keep the utility of the data, but it has a high risk of leaking private information. Therefore, it has become imperative to develop methods for simultaneously protecting privacy and preserving the utility of data in many fields.

Masking is a natural way of achieving privacy by modifying the content of the original data. It can protect sensitive information and reduce the risk of disclosure. However, it can also cause information loss due to the changes in the original data. Another problem of masking is that the disclosure risk still exists even after masking because there is much auxiliary information to identify individuals' sensitive information as can be seen in the two examples above. To resolve this problem, differential privacy was introduced by Dwork et al. (2006b).

Differential privacy is a framework to quantify preserved individual privacy while releasing useful information from data. Differential privacy requires near-indistinguishability of whether an individual belongs to a particular dataset or not, based on the released information. The essential part of a differentially private algorithm is data perturbation or randomization, which makes speculation on a specific individual difficult. A conventional way to perturb data is adding noise to the raw data. The amount of perturbation is determined so that it is large enough to hide whether a specific individual is included or not, but small enough to retain important information in the data. A key concept in differential privacy is the sensitivity of a function, or a statistic. The sensitivity is defined as the maximum change in the value of the function caused by modifying one observation in the dataset (Dwork et al., 2006b).

Since its introduction, differential privacy has impacted data-releasing designs and noise-adding mechanisms. If there exists a possibility of deducing sensitive private information from the results of statistical analysis, differential privacy can be an effective solution. For example, from a publicly available machine learning model, a malicious attacker can learn private information in the training dataset. Hence, some machine learning methods have been equipped with differentially private algorithms(Abadi et al., 2016; McMahan et al., 2018; Beimel et al., 2019; Bun et al., 2020; Kaplan et al., 2020). Differential privacy has also been incorporated into regression analysis (Zhang et al., 2012; Wang, 2018) and parameter estimation (Amin et al., 2019; Kamath et al., 2019; Liu and Oh, 2019; Biswas et al., 2020; Brunel and Avella-Medina, 2020; Kamath et al., 2020; Tzamos et al., 2020).

Our main interest lies in differentially private goodness-of-fit (GOF) tests. The GOF tests determine whether the assumed distribution for the population stands true or not based on the collected sample. However, the sample used in testing may contain highly sensitive information about subjects, and thus the privacy of individuals can be compromised when the results of the test are released. Most of the existing differentially private GOF tests have been developed for discrete random variables. For example, Rogers and Kifer (2017) and Gaboardi et al. (2016) developed tests based on asymptotic or approximate distributions of the differentially private χ^2 GOF test statistics with perturbation mechanisms. Cai et al. (2017) constructed a GOF test that consists of two differentially private steps. The compositional property of differential privacy guarantees that the results of the process satisfy differential privacy. We note that they focus on level α testing even if their targeted α values are different from a conventional level. Since additional noise has an impact on a test statistic, the variance of the statistic tends to be larger. As a result, it makes sense that most of differentially private hypothesis tests require a higher α compared to non-private counterparts. Berrett and Butucea (2020) proposed differentially private GOF tests based on local differential privacy. Canonne et al. (2019) proposed algorithms for differentially private hypothesis testing on high-dimensional distributions focusing on the product distribution over $\{\pm 1\}^d$ or the multivariate normal distribution with known covariance matrix.

A main objective of this paper is to give a brief survey on differentially private goodness-of-fit tests for discrete random variables and develop differentially private tests for continuous random variables. There are GOF tests for continuous cases, such as maximum mean discrepancy (MMD, (Gretton et al., 2012)), kernelized stein discrepancy (KSD, (Liu et al., 2016)), or Kolmogorov-Smirnov test (Stephens, 1970). However, designing a differentially private GOF test for a continuous variable is challenging because of the high sensitivity of the test statistic. Many continuous distributions have a range over $(-\infty, \infty)$ or $(0, \infty)$, which leads to infinite sensitivity for certain statistics. For example, the sensitivity of the sample mean for a normal distribution is infinite since a change made by a single observation could be unbounded. This would cause large noises to be added for perturbation. Subsequently, the information or the signal contained in the original data would be perturbed too much, which in turn would make the test too conservative or even unreliable. In such a case, we need more observations so as not to be dominated by perturbation, or develop a process that can appropriately bound the sensitivity of the variable or the test statistic. If a test such as Kolmogorov-Smirnov depends on the differences between two empirical distribution functions, randomizing the differences can be considered. But in this case, the distances could be negative during a randomization process.

Our strategy to achieve differential privacy of a GOF test for a continuous variable is via discretization of the continuous distribution by dividing the domain into non-overlapping intervals with the same length. Dividing the domain into bins is inspired by the work of Balakrishnan and Wasserman (2019). They introduce a recursive partitioning scheme for Lipschitz densities. The scheme divides a continuous distribution into a multinomial distribution. Their partitioning scheme consists of two stages: partitioning, and pruning. In the partitioning procedure, the algorithm divides the domain of a continuous distribution based on the probability of each bin. Their algorithm finds the upper and lower bounds of each bin so that the

volume or probability of each bin is large enough to distinguish distributions. If the probability of a bin is too large, the bin is divided into two bins. Then, the pruning procedure merges the bins with very small probabilities based on the pruning level. Hence, the bounds of pruning and dividing are important for their partitioning method since the number of categories is determined by the bounds, and the number of bins can be determined automatically. However, the method requires many parameters to determine the boundaries, which prevents practitioners from implementing it easily. Also, we note that they do not consider differentially private GOF tests.

The proposed binning procedure with equal probabilities is simple to implement, and its sensitivity is limited by one owing to discretization. As mentioned above, the maximum possible change of one observation can be infinite in continuous cases, but in a histogram it is just one count difference in a bin. After binning, we apply existing differentially private GOF tests for discrete random variables to the discretized continuous variable. To our best knowledge, there has been no attempts so far to develop differentially private GOF tests for continuous variables.

The rest of the paper is organized as follows. Section 2 reviews background on differential privacy and some existing differentially private GOF test for discrete random variables. In Section 3, the proposed differentially private GOF tests for continuous random variables through discetization is presented. The performance of the proposed test procedures are demonstrated via simulated examples in Section 4, and applied to the Household Financial Welfare Survey of South Korea in 2018 in Section 5. We conclude with some discussion in Section 6.

2. Review of Differentially Private GOF Tests for Discrete Variables

In this section, we briefly review the mathematical framework of differential privacy in Section 2.1 and differentially private GOF tests for discrete random variables in Section 2.2.

2.1. Differential Privacy

Differential privacy requires that a randomized algorithm returns similar outputs when executed on similar input databases, while protecting sensitive information (Dwork et al., 2006b). Hence, differential privacy restricts possible changes in the output of a randomized algorithm that can occur by a small change in its input.

Assume that a database D is a collection of records from $\mathcal{X}$, a set of discrete values that each data value is from. Suppose each value in the database D can be categorized into one of $i \in \{1, 2, \ldots, d\}$ where d denotes the number of categories in $\mathcal{X}$. Then $\mathbf{x} \in \mathbb{N}^d$, in which each entry x_i of $\mathbf{x}$ represents the number of elements in category i. Here, $\mathbb{N}$ is the set of natural numbers including zero.

Next, we define a distance that represents the difference between two databases. A natural distance between two databases D and D' is the Hamming distance (Wasserman and Zhou, 2010), which counts how many elements are different by comparing the observations in the two databases. Given two databases $D = \{D_1, \ldots, D_n\}$ and $D' = \{D'_1, \ldots, D'_n\}$, let $\kappa(D, D')$ denote the Hamming distance between D and D', i.e., $\kappa(D, D') = \left|\{i : D_i \neq D'_i\}\right|$ where $|\cdot|$ denotes the cardinality of a set. We say D and D' are neighbors if they are different only by a single observation. We define differential privacy as follows.

Definition 2.1 (Differential Privacy, Dwork et al. (2006b,a))**.** A randomized algorithm $\mathcal{M}$ with domain on the real line $\mathbb{R}$ is (ϵ, δ)-differentially private if for all $S \subseteq \text{Range}(\mathcal{M})$ and for all D, D' such that $\kappa(D, D') \leq 1$:

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr\left[\mathcal{M}(D') \in S\right] + \delta.$$

If $\delta = 0$, $\mathcal{M}$ is ϵ-differentially private.

The ranges of ϵ and δ are usually given as $\epsilon > 0$ and $0 \leq \delta \ll 1/|D|$ where $|D|$ is the sample size of the given database D (Dwork and Roth, 2014). By the definition of ϵ-differential privacy, if ϵ is very small, the algorithm will return the same result with a high probability for two neighboring databases. It means that the randomness of the algorithm disguises the impact of differences between two datasets on the output of the algorithm. Thus, we can say that a small ϵ can strongly protect privacy. Since a change in the output is probabilistically bounded, it is hard to distinguish whether the change comes from the original data or noise. Hence, an outcome of a differentially private analysis is essentially indistinguishable whether an individual is included in the dataset, or not. However, there is a trade-off between utility and privacy because the useful information in the dataset can also be disguised by the additional randomness to protect privacy.

In what follows, we introduce the ℓ_1 and ℓ_2 sensitivities, and the Laplace and Gaussian mechanisms to perturb the outputs of a function. Denote the ℓ_p norm of a vector $\mathbf{x}$ with d elements by $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{\frac{1}{p}}$, and the ℓ_p distance of two vectors $\mathbf{x}$ and $\mathbf{y}$ having d elements by $\|\mathbf{x} - \mathbf{y}\|_p = \left(\sum_{i=1}^d |x_i - y_i|^p\right)^{\frac{1}{p}}$.

Then, the ℓ_p sensitivity is defined as follows.

Definition 2.2. (ℓ_p-Sensitivity, Dwork et al. (2006b)) The ℓ_p-sensitivity of a function $f : \mathbb{N}^d \to \mathbb{R}^k$ is:

$$\Delta_p f = \max_{\substack{\mathbf{x},\mathbf{y} \in \mathbb{N}^d \\ \kappa(\mathbf{x},\mathbf{y})=2}} \|f(\mathbf{x}) - f(\mathbf{y})\|_p.$$

By Definition 2.2, the sensitivity quantifies the impact of a single observation on the function f in the worst case. This determines the uncertainty we need to introduce to the output in order to hide the participation of a particular individual. In other words, the sensitivity determines the variability of noise for perturbation on its output to preserve privacy. Based on the neighboring databases we use, ℓ_1-sensitivity of histogram is 2. Using the sensitivity of a function f, the randomization mechanism can be defined as follows.

Theorem 2.1. *(Randomization Mechanism, Dwork et al. (2006b, 2006a)) Given any function $f : \mathbb{N}^d \to \mathbb{R}^k$, the randomization mechanism is defined as:*

$$\mathcal{M}_{\mathcal{D}}(\mathbf{x}, f(\cdot), \epsilon, \delta) = f(\mathbf{x}) + (Y_1, \ldots, Y_k).$$

If Y_i are i.i.d. random variables from $\mathcal{D} = Lap(\Delta_1 f/\epsilon)$ and $\delta = 0$. Then, $\mathcal{M}_{\mathcal{D}}$ is the Laplacian mechanism satisfying ϵ-differential privacy. If Y_i are i.i.d. random variables from $\mathcal{D} = N(0, \sigma^2)$ with $\sigma = \frac{\Delta_2 f\sqrt{2\ln(2/\delta)}}{\epsilon}$. Then $\mathcal{M}_{\mathcal{D}}$ is the Gaussian mechanism satisfying (ϵ, δ)-differential privacy.

The Laplace distribution with scale b, $Lap(b)$, is $p(x|b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$, $x \in (-\infty, \infty)$. The Laplace mechanism uses Laplacian noise to perturb statistics. Another type of noise is the Gaussian noise. To satisfy $(\epsilon, \delta)-$differential privacy using the Gaussian noise, the ℓ_2 sensitivity of a function f, $\Delta_2 f$, is required.

As with the Laplace mechanism, the Gaussian mechanism adds the noise generated from a normal distribution with mean zero and standard deviation determined by the ℓ_2 sensitivity of the function and differential privacy parameters ϵ and δ.

When we compose several differentially private algorithms, the total differential privacy can be obtained by Theorem 2.2. It means that an algorithm can consist of several differentially private ones which satisfy the targeted privacy parameter values.

Theorem 2.2. *(Composition of Differentially Private Algorithms, Dwork et al. (2006a)) Let $\mathcal{M}_i : \mathbb{N}^d \to \mathcal{R}_i$ be an (ϵ_i, δ_i)-differentially private algorithm for $i \in \{1, \ldots, m\}$. If $\mathcal{M}_{[m]} : \mathbb{N}^d \to (\mathcal{R}_1, \ldots, \mathcal{R}_m)$ is defined to be*

$$\mathcal{M}_{[m]}(\mathbf{x}) = (\mathcal{M}_1(\mathbf{x}), \ldots, \mathcal{M}_m(\mathbf{x})),$$

then $\mathcal{M}_{[m]}$ is $(\sum_{i=1}^m \epsilon_i, \sum_{i=1}^m \delta_i)$-differentially private.

Note that the privacy bound in Theorem 2.2 becomes tight when δ_i are all zero. That is, the approximate DP with $\delta > 0$ does not exactly quantify the privacy guarantee in compositions, unlike the original pure DP that only ϵ is involved. When a differentially private algorithm is composed with a non-private algorithm, it is immune to such post-processing. In other words, once an algorithm protects an individual's privacy, a user cannot increase the privacy loss by applying a non-private process to the private algorithm. The following proposition states that composition of a data-independent mapping g with an (ϵ, δ)-differentially private algorithm $\mathcal{M}$ is also (ϵ, δ)-differentially private.

Proposition 2.1. *(Post-Processing, Dwork and Roth (2014)). Let $\mathcal{M} : \mathbb{N}^d \to \mathcal{R}$ be a randomized algorithm that is (ϵ, δ)-differentially private. Let $g : \mathcal{R} \to \mathcal{R}'$ be an arbitrary (randomized) mapping. Then $g \circ \mathcal{M} : \mathbb{N}^d \to \mathcal{R}'$ is (ϵ, δ)-differentially private.*

Differential privacy is a mathematically rigorous definition of privacy tailored to the analysis of large datasets and equipped with a formal measure of privacy loss. Because it considers the biggest possible difference, it adds relatively large noise and perturbs much, compared to non-private results. To balance this issue, there are some methods to provide tighter bounds compared to the pure differential privacy (ϵ-differential privacy) or approximate differential privacy ((ϵ, δ)-differential privacy).

Zero-Concentrated Differential Privacy (zCDP, Bun and Steinke (2016)) provides a tighter bound for composition of differentially private algorithms. Since it is comparable to (ϵ, δ)-differential privacy, the zCDP can be used to develop a differentially private GOF test. The zCDP is defined through the Rényi divergence (van Erven and Harremoës, 2014), which is a divergence measure between two probability distributions. Let P and Q be two arbitrary distributions on the same measure space, and p and q be the corresponding probability density functions, respectively.

Definition 2.3. (Rényi divergence, van Erven and Harremoës (2014)). For two probability distributions P and Q defined over $\mathbb{R}$, the Rényi divergence of order $\alpha > 1$ is

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1}\ln E_{X\sim Q}\left(\frac{p(X)}{q(X)}\right)^\alpha,$$

where $X \sim Q$ means that X follows the distribution Q.

$D_1(P||Q)$ is set to be $\lim_{\alpha\to 1} D_\alpha(P||Q)$, and is equal to the Kullback-Leibler divergence (van Erven and Harremoës, 2014). The zCDP is defined via the Rényi divergence between two distributions of a randomized algorithm based on two neighboring databases.

Definition 2.4. (zCDP, Bun and Steinke (2016)). A randomized mechanism $\mathcal{M}: \mathbb{N}^d \to \mathbb{R}^k$ is (ξ, ρ)-zero-concentrated differentially private (henceforth (ξ, ρ)-zCDP) if, for all D, D' differing on a single entry and all $\alpha \in (1, \infty)$,

$$D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \xi + \rho\alpha.$$

Define ρ-zCDP to be $(0, \rho)$-zCDP.

By taking $\alpha \to \infty$, we have $D_\infty(M(D)|M(D')) \leq \epsilon$, which results in classic differential privacy (Mironov, 2017). The zCDP is related to (ϵ, δ)-differential privacy and ϵ-differential privacy, as illustrated in Theorem 2.3.

Theorem 2.3. *(Bun and Steinke (2016)) If $\mathcal{M}$ is ϵ-differentially private, $\mathcal{M}$ is $\frac{\epsilon^2}{2}$-zCDP. Further, if $\mathcal{M}$ is ρ-zCDP, then $\mathcal{M}$ is $(\rho + 2\sqrt{\rho \ln(1/\delta)}, \delta)$-differentially private for every $\delta > 0$.*

It is known that ρ-zCDP with $\rho \approx \frac{\epsilon^2}{4\log(1/\delta)}$ guarantees to achieve (ϵ, δ)-DP (Bun and Steinke, 2016). Therefore, we can compare an (ϵ, δ)-differentially private algorithm and a ρ-zCDP algorithm applied on a dataset with the same level of privacy. Theorem 2.4 illustrates that the Gaussian mechanism satisfies ρ-zCDP.

Theorem 2.4. *(Bun and Steinke (2016)) For a function $f: \mathbb{N}^d \to \mathbb{R}^k$, the Gaussian mechanism $\mathcal{M}_G$ for $\mathbf{x} \in \mathbb{N}^d$ is:*

$$\mathcal{M}_G(\mathbf{x}, f(\cdot), \epsilon, \rho) = f(\mathbf{x}) + (Y_1, \ldots, Y_k),$$

where $\sigma = \frac{\Delta_2 f}{\sqrt{2\rho}}$.

The following proposition states that post-processing to the results from zCDP maintains the privacy under post-processing.

Proposition 2.2. *(Post-Processing, Bun and Steinke (2016)). Let $\mathcal{M}: \mathbb{N}^d \to \mathcal{R}$ and $g: \mathcal{R} \to \mathcal{R}'$ be (randomized) algorithms. If $\mathcal{M}$ is ρ-zCDP, $g \circ \mathcal{M}: \mathbb{N}^d \to \mathcal{R}'$ is ρ-zCDP.*

Thus, Propositions 2.1 and 2.2 together guarantee that a GOF test satisfies differential privacy or zCDP when the input data satisfy differential privacy or zCDP, respectively. In our method, we develop differentially private GOF tests for continuous variables via transforming a continuous random variable into a discrete random variable, and applying differentially private GOF tests which are based on Propositions 2.1 and 2.2, and Theorem 2.2.

2.2. Differentially Private Goodness-of-Fit Tests for Discrete Variables

In this section, we review some existing differentially private GOF tests for discrete variables. Suppose there are d bins (categories) and we observe the count for each bin. The Private and Sample Efficient Identity Testing (Priv'IT) method for discrete variables is introduced by Cai et al. (2017). The hypotheses of interest are

$$H_0: \mathbf{p} = \mathbf{p^0} \quad \text{vs} \quad H_1: \mathbf{p} \neq \mathbf{p^0}, \tag{1}$$

where $\mathbf{p}$ is the population distribution of a given sample, $\mathbf{p^0}$ is the hypothesized distribution. This hypothesis test determines whether the two distributions are farther apart than a certain threshold, d_α.

The Priv'IT algorithm consists of two steps. The first step is a filtering step that rejects the null hypothesis when the sample distribution is far enough from $\mathbf{p^0}$. In the first step, we reject the null if the deviation between the noise-contaminated histogram counts and the ones expected under $\mathbf{p^0}$ is too large. The second step is a statistical step that uses the actually observed counts. The χ^2-style statistic, Z, is defined as:

$$Z = \frac{2}{n{d_\alpha}^2} \sum_{i \in A} \frac{(N_i - np_i^0)^2 - N_i}{np_i^0},$$

where n is the total number of observations, N_i is the observed count in the ith bin, and p_i^0 is the proportion of observations in the ith category that fall within the set defined by the hypothesized distribution $\mathbf{p^0}$, $A = \{i : p_i^0 \geq \frac{c_1 d_\alpha}{d}$, where $i = 1, \ldots, d\}$. We set $c_1 = 1/4$. Then we reject the null with the following rule.

$$P(\text{Reject } H_0) = \begin{cases} 0 & \text{if } Z \leq 0; \\ z & \text{if } 0 < Z < 1; \\ 1 & \text{if } Z \geq 1. \end{cases}$$

An important aspect of this test is that the filtering step guarantees $(0, c_2\epsilon/4)$-differential privacy and the statistical step guarantees $(0, c_2\epsilon/4)$-differential privacy, where Cai et al. (2017) set $c_2 = 3/40$. By the composition of differential privacy in Theorem 2.2, the entire algorithm satisfies $(0, c_2\epsilon/2)$-differential privacy, which implies ϵ-differential privacy for the test. If the first step does not reject the null, the statistic Z is ϵ-Lipschitz with respect to the counts. Thus, if two neighboring input datasets differ by only one observation, then the resulting statistics Z_1 and Z_2 can differ by at most ϵ. Hence, by Theorem 2.2, the test result is ϵ-differentially private.

Gaboardi et al. (2016) proposed two versions of differentially private χ^2 (DP χ^2) GOF tests. One uses Monte Carlo simulation to find rejection criteria, and the other uses an asymptotic distribution of the test statistic. Monte Carlo simulation can be applied when the noise is generated from a Laplace or a normal distribution, while the asymptotic distribution can be used for the Gaussian noise case only. The test statistic is

$$Q_{\mathcal{D}}^2 = \sum_{i=1}^{d} \frac{(N_i + Y_i - np_i^0)^2}{np_i^0},$$

where Y_i are i.i.d random variables from $\mathcal{D}$. If $\mathcal{D}$ is a Laplace distribution, the test is ϵ-differentially private. If $\mathcal{D}$ is a normal distribution, the test is (ϵ, δ)-differentially private by Theorem 2.1. The test statistic $Q_{\mathcal{D}}^2$ is similar to the Pearson's χ^2 test statistic except that the counts are perturbed.

The algorithm of the Monte Carlo simulation test (MC-DP χ^2 GOF) generates noise and computes $Q_{\mathcal{D}}^2$ K times. Then, we have $q_{(1)}, \ldots, q_{(K)}$ which represent a distribution of $Q_{\mathcal{D}}^2$ under the null hypothesis where $q_{(t)}$ is the tth order statistic. Thus, $Pr(Q_{\mathcal{D}}^2 > q_{(t)}) < \alpha$ where $t = \lceil (K+1)(1-\alpha) \rceil$, and it achieves a level α test. When compared with the non-private χ^2 test, however, the MC-DP χ^2 GOF test requires more observations to have the same power and significance level because the noise is included in the statistic. Since all elements to generate the null distribution Q_D are known, one can easily obtain the distribution and conduct hypothesis test.

To illustrate the asymptotic approach (Asymptotic-DP χ^2 GOF) in detail, define the random vector $\mathbf{U} = (U_1, \ldots, U_d)^T$ where $U_i = \frac{N_i - np_i^0}{\sqrt{np_i^0}}$, $i \in \{1, \ldots, d\}$. Also, define the standardized Gaussian noise random vector, $\mathbf{V} = (Y_1/\sigma(\epsilon, \delta_n), \ldots, Y_d/\sigma(\epsilon, \delta_n))^T \sim N(0, I_d)$. Let $\mathbf{W} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}$. Note that $\mathbf{W}$ converges in distribution to $N(0, \Sigma')$ where Σ' is the $2d \times 2d$ block diagonal matrix

$$\Sigma' = \begin{bmatrix} \Sigma & 0 \\ 0 & I_d \end{bmatrix}, \quad \text{where} \quad \Sigma = I_d - \sqrt{\mathbf{p}^0}\sqrt{\mathbf{p}^0}^T.$$

Note that Σ' is idempotent. We can express the differentially private χ^2 statistic as a quadratic form $Q^2 = \mathbf{W}^T A \mathbf{W}$, where $Q^2 = Q_{\mathcal{D}}^2$ when $\mathcal{D}$ is the normal distribution with mean zero, standard deviation σ, and the positive semi-definite $2d \times 2d$ matrix A:

$$A = \begin{bmatrix} I_d & \Lambda \\ \Lambda & \Lambda^2 \end{bmatrix}, \quad \text{where} \quad \Lambda = \text{Diag}\left(\frac{\sigma(\epsilon, \delta_n)}{\sqrt{n\mathbf{p}^0}} \right).$$

Theorem 2.5. *(Gaboardi et al. (2016)) Let $\mathbf{W} \sim N(0, \Sigma')$ where Σ' has rank $r \le 2d$. Then $\mathbf{W}^T A \mathbf{W}$ is distributed as $\sum_{i=1}^{r} \lambda_i \chi_1^{2,i}$ where $\{\lambda_i\}_{i=1}^r$ are the eigenvalues of $B^T A B$ for $B \in \mathbb{R}^{2d \times r}$ such that $BB^T = \Sigma'$ and $B^T B = I_r$, and $\{\chi_1^{2,i}\}_{i=1}^r$ is a set of r independent χ^2 random variables with 1 degree of freedom.*

The result of Theorem 2.5 is used to find a critical value τ_α:

$$\Pr\left[\sum_{i=1}^{r} \lambda_i \chi_1^{2,i} \ge \tau_\alpha \right] = \alpha,$$

for a given significance level α. Note that τ_α is a function of $n, \epsilon, \delta, \alpha$ and $\mathbf{p}^0$, but not the data. If the critical value includes the information of the data, it should also be perturbed for privacy purposes. As with the MC-DP χ^2 GOF test, the Asymptotic-DP χ^2 GOF test needs a bigger sample size to achieve a level α test compared to non-private χ^2 GOF tests because of noise. In R, τ_α can be obtained from the library `CompQuadForm`.

The zCDP GOF test (Rogers and Kifer, 2017) utilizes the Gaussian mechanism in Theorem 2.4 and has two types of tests: projected test and unprojected test. Define the zCDP statistic $\mathbf{Z}_\rho = (Z_{\rho,1}, \ldots, Z_{\rho,d})^T$ where

$$Z_{\rho,i} = \frac{N_i + Y_i - np_i^0}{\sqrt{n}}, \tag{2}$$

and Y_i are i.i.d. random variables drawn from $N(0, 1/\rho)$. The random vector $Z_{\rho_n,i}$ in (2) has the following asymptotic distribution under the null hypothesis: if $n\rho_n \to \infty$, $\mathbf{Z}_{\rho_n} \xrightarrow{D} N(0, \Sigma)$ where $\Sigma = Diag\{\mathbf{p}^0\} - \mathbf{p}^0(\mathbf{p}^0)^T$; if $n\rho_n \to \rho > 0$, $\mathbf{Z}_{\rho_n} \xrightarrow{D} N(0, \Sigma_\rho)$ where $\Sigma_\rho = \Sigma + 1/\rho \cdot I_d$.

The unprojected zCDP GOF test is defined as:

$$Q_{\rho_n} = (\mathbf{Z}_{\rho_n})^T \Sigma_{n\rho_n}^{-1} \mathbf{Z}_{\rho_n},$$

which converges in distribution to χ_d^2 if $n\rho_n \to \rho$. In contrast, if $n\rho_n \to \infty$, the unprojected private test statistic does not approach the χ^2 test for fixed ρ, and $\Sigma_{n\rho_n}$ converges to a singular matrix. We note that when $n\rho_n \to \rho$, the degrees of freedom for the χ^2 test, d, is greater than that of a non-private test statistic, and subsequently its critical value is greater

than the non-private counterparts. Thus, rejecting the null becomes more difficult when the null is false, and the test becomes less powerful. To resolve this issue, a projected private test statistic is proposed.

Note that Σ_ρ has the eigenvalue of $1/\rho$ corresponding to the eigenvector $\mathbf{1} = (1, \cdots, 1)^T$, and that this direction can be thought as pure noise. Hence, the projection $\mathbf{Z}_\rho$ onto the subspace orthogonal to $\mathbf{1}$ can eliminate the impact of pure noise. This enforces the constraint that the entries in $\mathbf{Z}_\rho$ add up to 0. Then, the projected zCDP GOF test statistic $\mathcal{Q}_\rho$ is defined as:

$$\mathcal{Q}_\rho = (\mathbf{Z}_\rho)^T \mathbf{P} \Sigma_{n\rho}^{-1} \mathbf{P} \mathbf{Z}_\rho,$$

where $\mathbf{P} = I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^T$. The advantage of using the test statistic $\mathcal{Q}_\rho$ is that it converges in distribution to χ^2_{d-1} instead of χ^2_d regardless of whether $n\rho_n \to \rho > 0$ or $n\rho_n \to \infty$.

In comparison, the Priv'IT test has better power compared to the DP χ^2 and the zCDP GOF tests when the sample size is small because of the filtering step. Unlike the DP χ^2 and the zCDP methods, the rejection of the Priv'IT test is determined randomly. Sometimes the decision just depends on the probability of getting large noise. However, the test procedure produces a quite high false positive rate.

3. Proposed Differentially Private GOF Tests for Continuous Variables

In this section, we propose differentially private GOF tests for one-dimensional continuous variables. Recall that we need to perturb observations or to randomize the result to satisfy Definition 2.1. If we perturb a test statistic constructed directly from a continuous variable, the sensitivity of the test statistic can be infinite. Therefore, our strategy is to convert a continuous variable to a discrete one by partitioning. Then, we apply a differentially private GOF test for discrete random variables to the discretized data.

In partitioning, there are two possible approaches: equal length and equal probability. In the equal length approach, the domain is divided into equal length intervals. This approach has an advantage that the shape of a discretized distribution is similar to the shape of the original distribution. But, for a tail bin (the first or the last bin of the discretized distribution), its probability may be greater than the neighboring bins. Moreover, finding a start or an end bin of a distribution might not be possible if the domain is unbounded. If we employ the equal probability approach, a set of bin boundaries of the discretized distribution with d bins can be expressed as:

$$H = \left\{ u_i : F_X(u_{i+1}) - F_X(u_i) = \frac{1}{d}, \text{ for } i = 0, 1, \ldots, d-1 \right\}, \tag{3}$$

where F_X is the cumulative function of X, and u_0 is the minimum and u_d is the maximum of a random variable X, which could be $-\infty$ and ∞, respectively. We propose to take the equal probability approach.

In conducting a GOF test with the equal probability approach, the hypothesized distribution $\mathbf{p^0}$ is defined as $\mathbf{p^0} = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right)^T$. Based on H in (3), we can obtain a set of counts N corresponding to each bin from a dataset D with sample size n:

$$N = \{N_i : N_i = |\{z_l : u_{i-1} \le z_l < u_i, z_l \in D, l = 1, 2, \ldots, n\}|, u_i \in H, i = 1, \ldots, d\},$$

where $|D|$ means the cardinality of a set D and N_i is the count in the ith bin. The discretization procedure is concisely described in Algorithm 1.

Algorithm 1 Discretization

1: **input**: Hypothesized distribution $\mathbf{p^0}$; dataset $\mathbf{x}$; the number of bins d
2: Let $F_{\mathbf{p^0}}$ be the cumulative distribution function corresponding to $\mathbf{p^0}$ and find $u_0 < u_1 < \ldots < u_d$ satisfying

$$F_{\mathbf{p^0}}(u_i) - F_{\mathbf{p^0}}(u_{i-1}) = \frac{1}{d},$$

where u_0 and u_d are end points of the domain of the distribution $\mathbf{p^0}$.
3: Let N_i be the number of observations in the ith category and $x_k \in \mathbf{x}$.
4: **for** $i \in \{1, \ldots, d\}$ **do**
5: **for** $k \in \{1, \ldots, n\}$ **do**
6: **if** $u_{i-1} < x_k < u_i$ **then**
7: **return** $N_i \leftarrow N_i + 1$
8: **end if**
9: **end for**
10: **end for**

Figure 1 shows an example of discretized distributions with $d = 100$. The hypothesized distribution is the standard normal distribution as displayed in the left panel. Using H in (3) computed from the hypothesized distribution, we can transform other continuous distributions into discrete ones. The right panel shows the χ^2 distributions with degrees of freedom

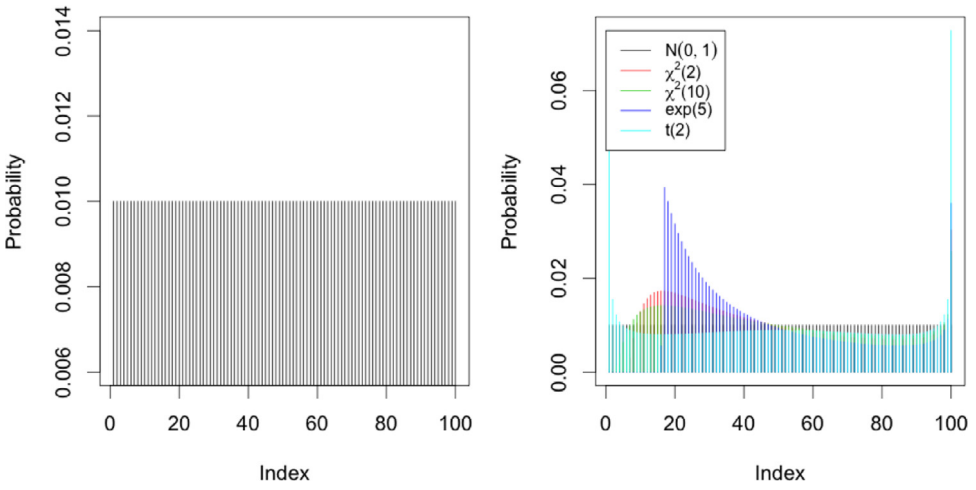

Fig. 1. The left panel displays the discretization of the standard normal distribution with equal probability and $d = 100$. The right panel shows a comparison of χ^2 distributions with degrees of freedom 2 and 10, the exponential distribution with rate 5, and the student-t distribution with degrees of freedom 2. They are discretized by the the boundaries obtained from the standard normal distribution in the left panel.

2 and 10, the exponential distribution with rate 5, and the student-t distribution with degrees of freedom 2, based on the boundaries computed from the hypothesized distribution in the left panel. In order to compare the discretized distributions, the distributions are scaled by their theoretical mean and standard deviation. In the plot, the bars higher than those of the hypothesized distribution implies higher probability mass in the corresponding regions. It can be seen that the χ^2 distributions and the exponential distribution have higher probability around the center of the left side. Also, the student-t distribution has thicker tails than the normal distribution. Using these differences, we can conduct a GOF test. We apply the Priv'IT, DP χ^2, and zCDP GOF tests introduced in Section 2.2 on the discretized variables to guarantee differential privacy.

4. Simulation

We demonstrate the performance of the proposed differentially private GOF tests through simulated examples. We discretize continuous data as explained in Section 3, and apply the differentially private GOF tests discussed in Section 2.2: Priv'IT, MC-DP and Asymptotic-DP χ^2, and unprojected and projected zCDP GOF tests.

4.1. Settings

We consider various privacy parameters for ϵ- and (ϵ, δ)-differential privacy algorithms. The parameter ϵ is set as 0.1, 0.2, 0.4, 0.6, or 1.0 and δ is set as 10^{-6} or 10^{-5}. In the Priv'IT test, $d_\alpha = 0.1$ in (1). The privacy parameter ρ in the zCDP GOF test is determined by Theorem 2.3. For the MC-DP χ^2 GOF test, the null distribution of the test statistic is obtained with 100 iterations. The sensitivities are $\ell_1 = 2$ and $\ell_2 = \sqrt{2}$ when one observation moves from a bin to another. The hypothesized distribution $\mathbf{p^0}$ of the simulation is the standard normal and the alternative distributions are the χ^2 distribution with degrees of freedom 2 and 10, and the exponential distribution with rate 5. The number of bins is determined as in Freedman and Diaconis (1981); $d = Cn^{1/3}$, where n is sample size and C is a positive constant. We consider six sample sizes, n =1500, 3000, 5000, 10000, 15000, and 20000 with five different C values, 3, 3.5, 4, 4.5, and 5. The noise generated from the Laplace distribution is used for the Priv'IT and MC-DP χ^2 GOF tests. The normal distribution is used to generate the noise for the MC-DP χ^2 and the asymptotic-DP χ^2, and zCDP GOF tests. We note that the Priv'IT test is a level $1/3$ test, while the DP χ^2 and zCDP GOF tests are level 0.05 tests. Each algorithm is repeated 150 times to compute p-values.

4.2. Results

Figure 2 shows the power and type I errors of the Priv'IT GOF test with different numbers of bins for the simulated data with various sample sizes. In each plot, the hypothesized distribution, standard normal distribution, is displayed with the black solid line, and three alternative distributions, $\chi^2(2)$, $\chi^2(10)$, and exp(5), are displayed with the red dashed, blue dotted, and orange dot-dashed lines, respectively. The gray dotted line in each plot indicates the significance level guaranteed by the test, which is 1/3 in this case. We can see that the Priv'IT test shows high power even for small sample size and the small number of bins. Also, high power is achieved regardless of the alternative distributions. However, when the sample size is small, it can be seen that the false positive rate (black solid line) is somewhat high even though it is still under the given significance level. The false positive rate becomes lower as the sample size increases. Overall, the Priv'IT GOF test shows consistent performance over different numbers of bins and sample sizes.

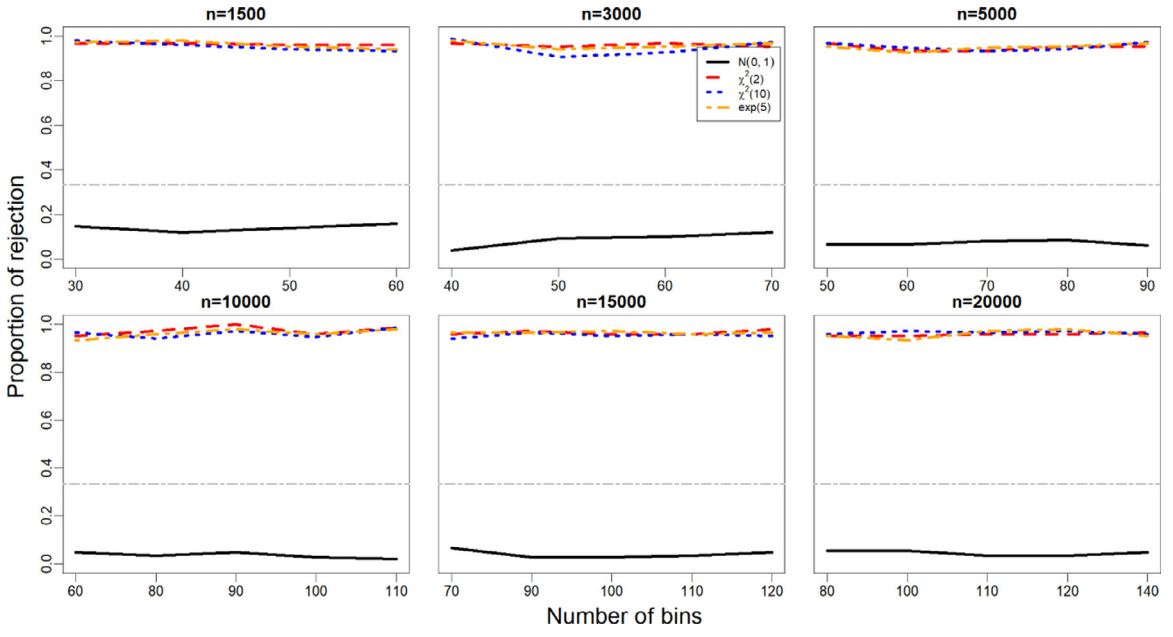

Fig. 2. The plots of power and type I error of the Priv'IT test with different numbers of bins on the simulated data when n =1500, 3000, 5000, 10000, 15000, and 20000. The hypothesized distribution is the standard normal distribution (black solid) and three alternative distributions, $\chi^2(2)$ (red dashed), $\chi^2(10)$ (blue dotted) and exp(5) (orange dot-dashed), are considered. The black solid line indicates the false positive rate and the other lines denote the true positive rates of the test obtained from 150 repetitions. The gray dotted line in each plot indicates the significance level of the test, 1/3.

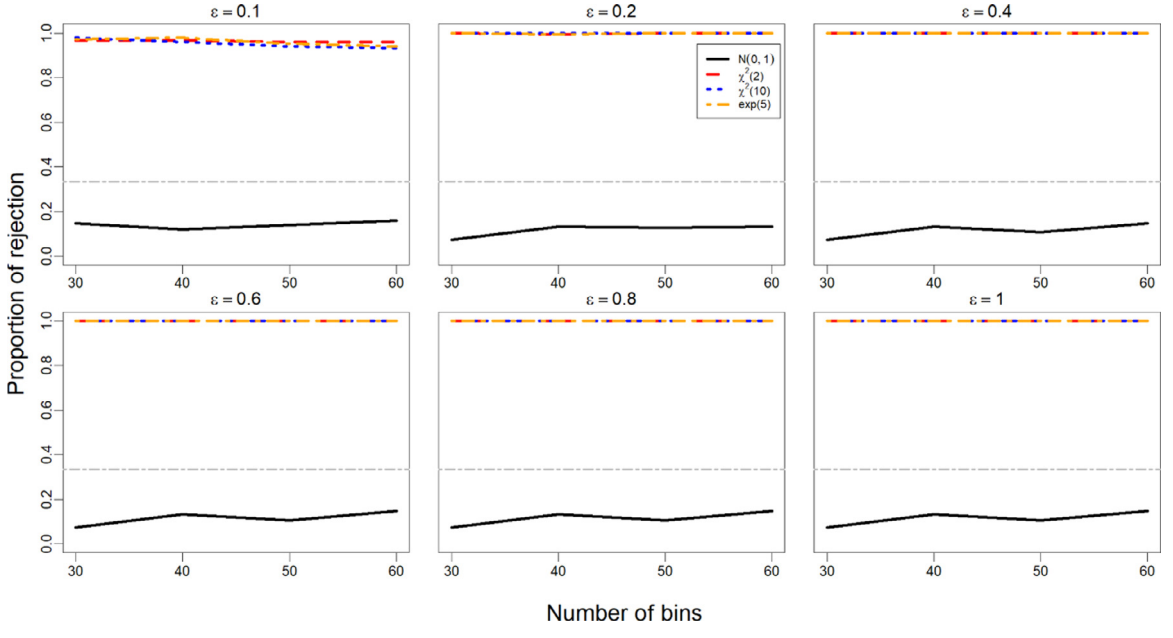

Fig. 3. The plots of power and type I error of the Priv'IT test with ϵ=0.1, 0.2, 0.4, 0.6, 0.8, 1.0 when n=1500.

Figure 3 shows the proportion of rejection for different ϵ values for the Priv'IT test. We use the same 150 samples with sample size 1500, and the results stay almost identical regardless of ϵ. As ϵ gets larger, the variance of the noise becomes smaller and the Priv'IT mechanism depends more on the second step mentioned in Section 2.2.

For the MC-DP χ^2 GOF test, two types of noise, the Laplace and Gaussian, can be used. Figure 4 shows the results of applying the MC-DP χ^2 test when adding the Laplace noise. We can see that type I errors (black solid) are close to the given significance level, 0.05 (gray dotted). However, when the alternative distribution is $\chi^2(10)$ (blue dotted), the power of the test is low for $n = 1500$ and $n = 3000$ and decreases over the number of bins even when $n = 5000$. When the sample size is 1500, the power for $\chi^2(2)$ (red dashed) and exp(5) (orange dot-dashed) shows a downward trend as the number of

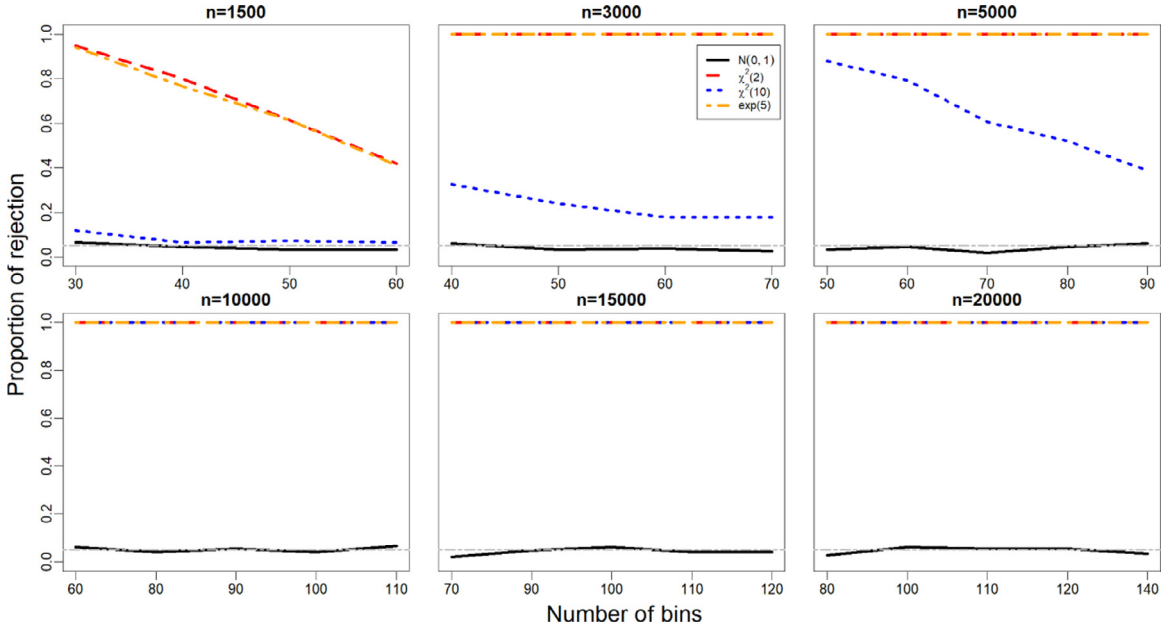

Fig. 4. The power and type I error of the MC-DP χ^2 GOF test with different numbers of bins based on the Laplace noise, when sample sizes are 1500, 3000, 5000, 10000, 15000, and 20000. The gray dotted line in each plot indicates the significance level of the test, 0.05.

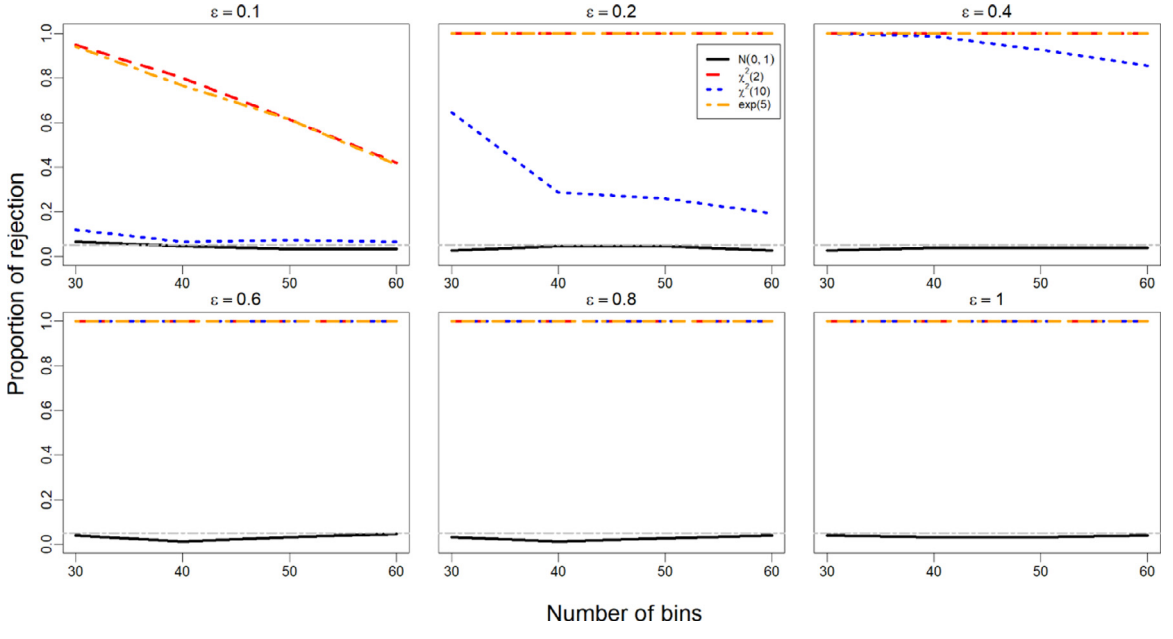

Fig. 5. The power and type I error of the Laplacian MC-DP χ^2 GOF test with ϵ=0.1, 0.2, 0.4, 0.6, 0.8, 1.0 when n=1500. The gray dotted line in each plot indicates the significance level of the test, 0.05.

bins increases. This implies that the sample size and the number of bins are important factors to determine the power of the MC-DP χ^2 GOF test particularly when the sample size is below 10000. The type I errors of the MC-DP χ^2 GOF test with the Laplacian mechanism in Figure 5 stays around the nominal significance level regardless of ϵ when the samples were generated from the standard normal distribution. For the other distributions, the power improves as ϵ increases, and stays around 1 when $\epsilon \geq 0.6$.

Figure 6 shows the results of the MC-DP χ^2 test with the Gaussian noise added. The type I errors stay stable around the given significance level. However, we can observe lower power and more downward trends with the number of bins, compared to the Laplace noise case in Figure 4. Since the Gaussian noise has greater variance than the Laplace noise, the test statistic might be more vulnerable to the Gaussian noise and yield lower power especially when the sample size is

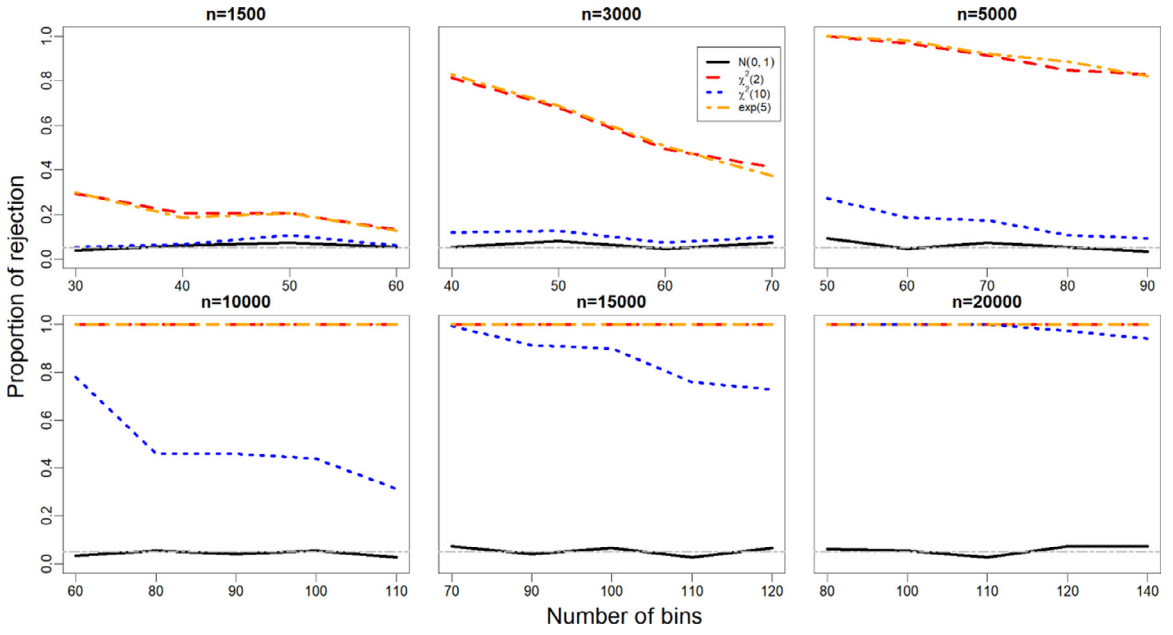

Fig. 6. The power and type I error of the MC-DP χ^2 GOF test with different numbers of bins based on the Gaussian noise, when sample sizes are 1500, 3000, 5000, 10000, 15000, and 20000. The gray dotted line in each plot indicates the significance level of the test, 0.05.

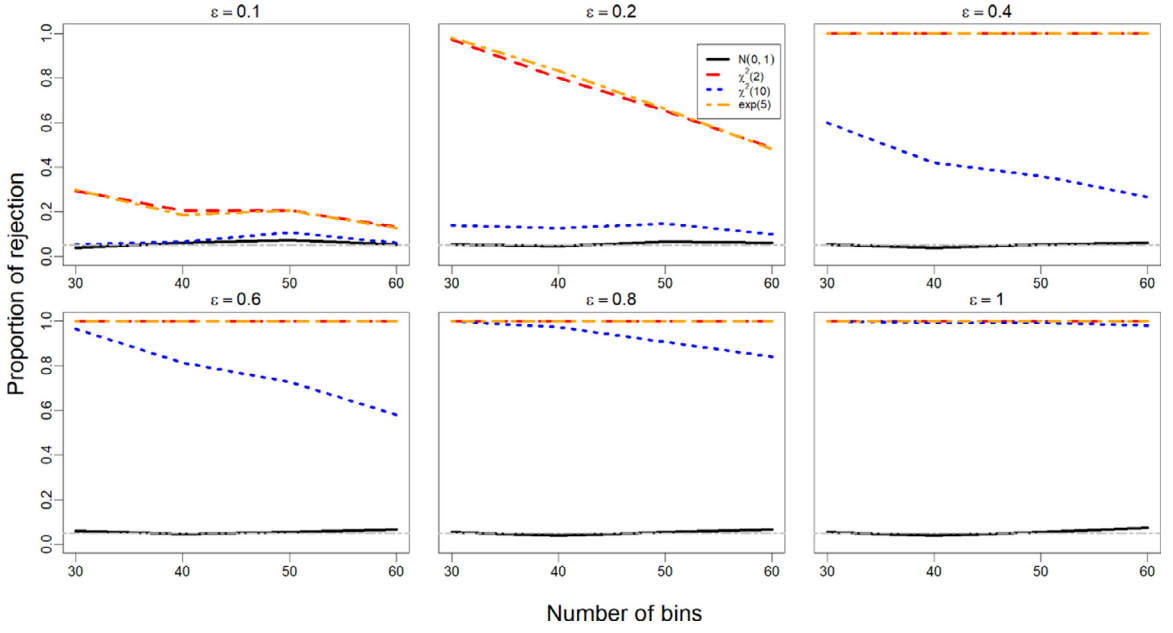

Fig. 7. The power and type I error of the Gaussian MC-DP χ^2 GOF test with ϵ=0.1, 0.2, 0.4, 0.6, 0.8, 1.0, and $\delta = 10^{-5}$ when n =1500. The gray dotted line in each plot indicates the significance level of the test, 0.05.

relatively small. Therefore, the Gaussian mechanism needs a larger sample size than the Laplace mechanism to reduce the impact of the noise added to the bin counts. This is supported by the bottom right plot when $n = 20000$, which shows stable power over the number of bins. Figure 7 shows the proportion of rejection with ϵ for the Gaussian MC-DP χ^2 GOF test when $n = 1500$ and $\delta = 10^{-5}$. The type I error stays around the significance level when the population distribution is normal. However, when ϵ is small and the number of bins is large, the Gaussian MC-DP χ^2 tests yield a low power under non-normality. It achieves higher power as ϵ gets larger.

For the Asymptotic-DP χ^2 test, the results in Figures 8 and 9 can be similarly interpreted as those of the Gaussian MC-DP χ^2 GOF test in Figures 6 and 7. Because the Monte Carlo eventually converges to the asymptotic distribution of the statistic as n increases, the similarity in the results between the two tests is not surprising.

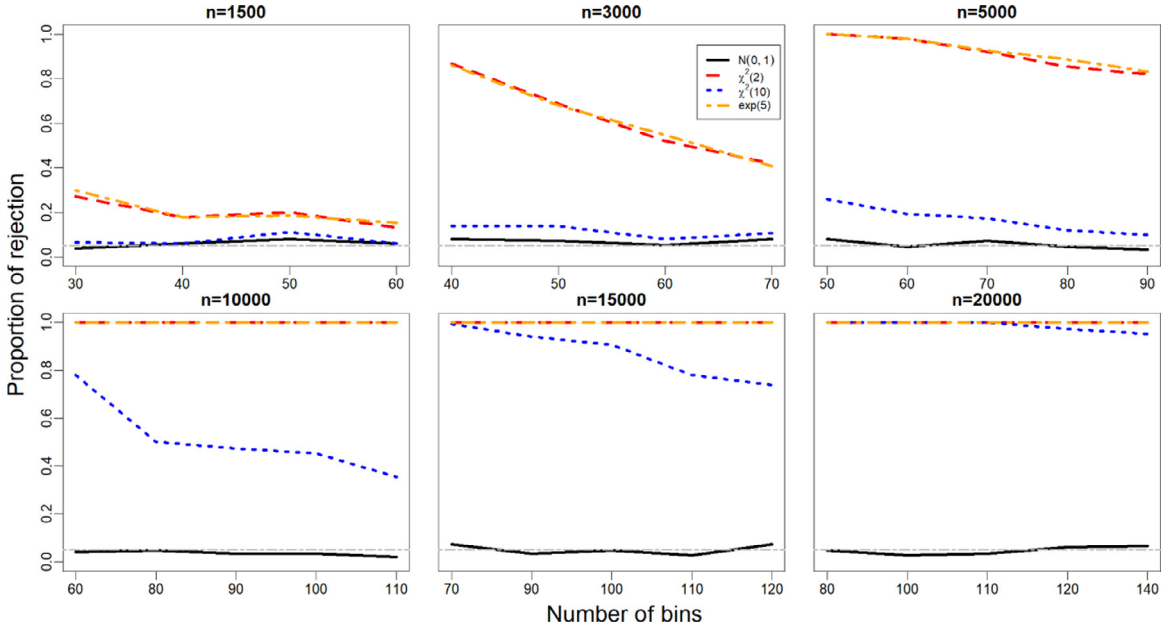

Fig. 8. The power and type I error of the Asymptotic-DP χ^2 GOF test with different numbers of bins based on the Gaussian noise, when sample sizes are 1500, 3000, 5000, 10000, 15000, and 20000. The gray dotted line in each plot indicates the significance level of the test, 0.05.

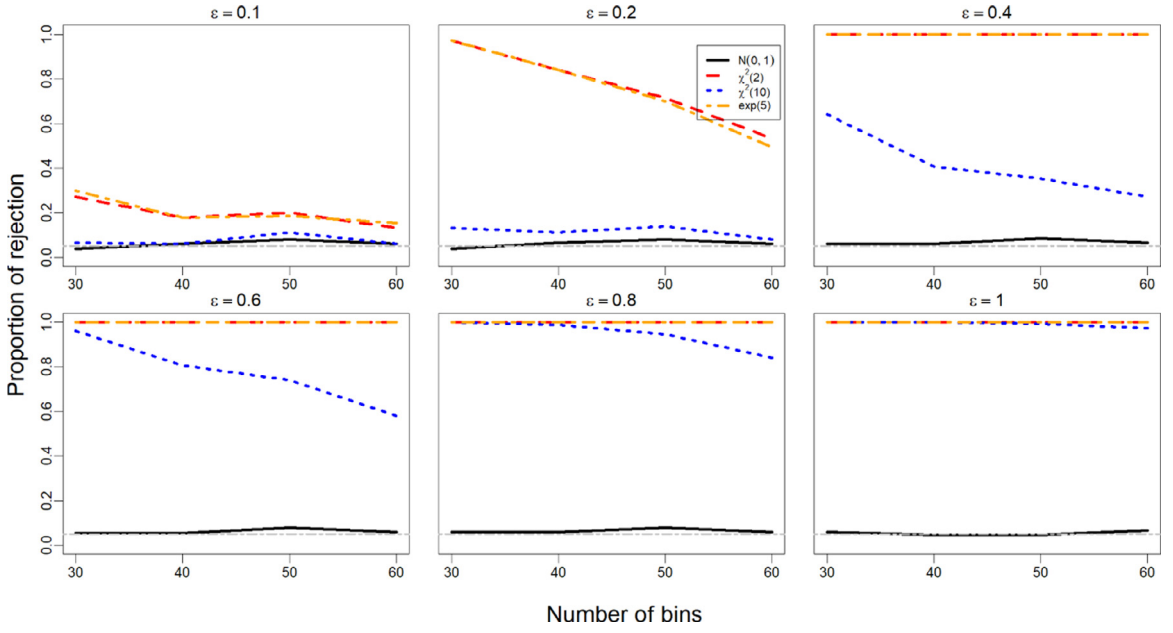

Fig. 9. The power and type I error of the Asymptotic-DP χ^2 GOF test with ϵ=0.1, 0.2, 0.4, 0.6, 0.8, 1.0, and $\delta = 10^{-5}$ when n =1500. The gray dotted line in each plot indicates the significance level of the test, 0.05.

Figure 10 displays the power and type I error results of the unprojected zCDP GOF test. The results are similar to those of DP χ^2 GOF tests with the Gaussian noise (Figures 6 and 8); although, the power of the unprojected zCDP test is slightly higher when the sample size is large. Figure 11 shows better performance of the unprojected zCDP GOF test compared to the Gaussian and Asymptotic MC-DP GOF χ^2 tests in terms of power.

The projected zCDP GOF test in Figure 12 shows quite similar results with those of the unprojected zCDP GOF test in Figure 10. Recall the difference between the two zCDP test statistics, $\mathcal{Q}_\rho$ and Q_ρ, is $Q_\rho - \mathcal{Q}_\rho = \frac{\rho}{d}\left(\sum_{k=1}^{d} Y_k\right)^2$. Because we set small $\rho \approx 0.0002$ to satisfy (ϵ, δ)-differential privacy with $\epsilon = 0.1$ and $\delta = 10^{-5}$, the difference between the two test

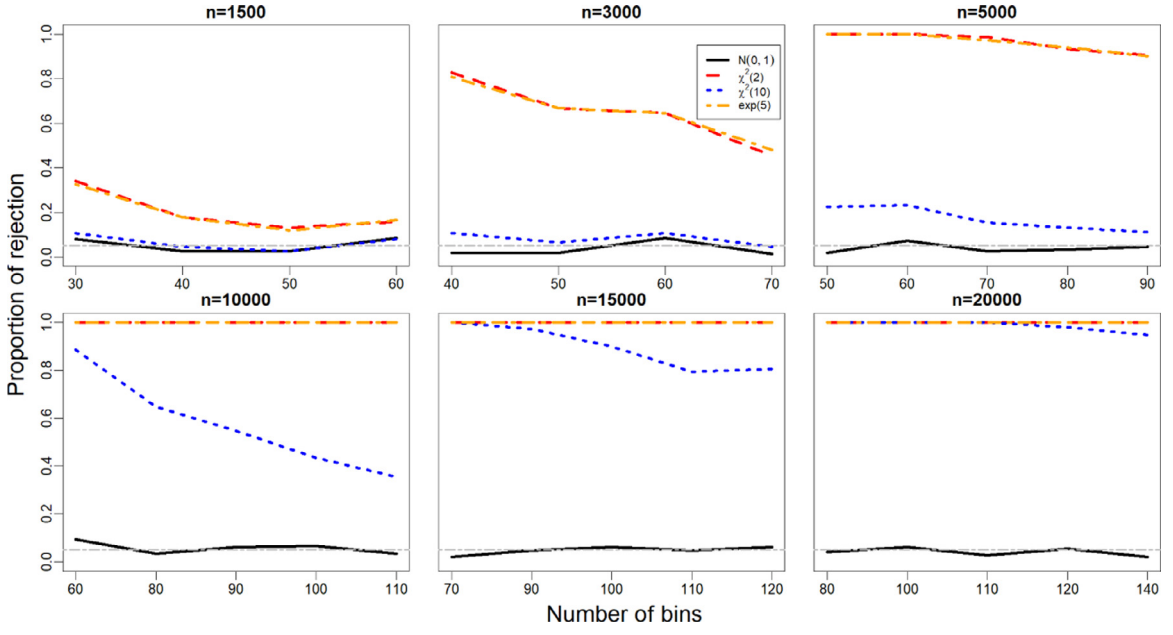

Fig. 10. The power and type I error of the unprojected zCDP GOF test with different numbers of bins based on the Gaussian noise, when sample sizes are 1500, 3000, 5000, 10000, 15000, and 20000. The gray dotted line in each plot indicates the significance level of the test, 0.05.

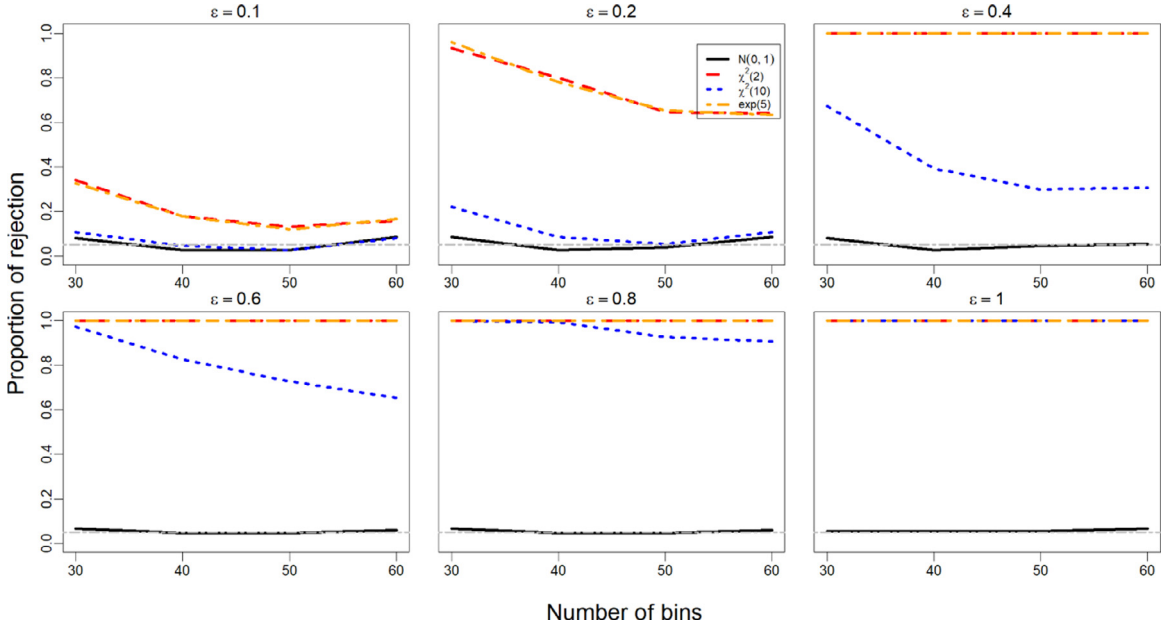

Fig. 11. The power and type I error of the unprojected zCDP GOF test with ϵ=0.1, 0.2, 0.4, 0.6, 0.8, 1.0, and $\delta = 10^{-5}$ when n =1500. The gray dotted line in each plot indicates the significance level of the test, 0.05.

statistics stays small. Furthermore, since $Y_k \sim N(0, \sigma^2)$, as d increase, $\sum_{k=1}^{d} Y_k/d$ will converge to zero. We also observe that Figure 13 and Figure 11 look similar. As we allow more privacy budget (larger ϵ), the test achieves higher power.

In Figure 14, we compare the private tests with the non-private Pearson's χ^2 test with discretization for various population distributions and sample sizes. Each row corresponds to the population distribution of samples, $N(0, 1)$, $\chi^2(2)$, $\chi^2(10)$, and exp(5), respectively. The non-private test almost always rejects the null hypothesis (normal) when the population distribution is not a normal distribution, and fails to reject the null when it is a normal distribution. These results show that the non-private test with discretization performs well on our simulation settings. As sample size gets larger, the private tests show similar results to those of the non-private test. We note when the population distribution is $\chi^2(10)$, the gaussian

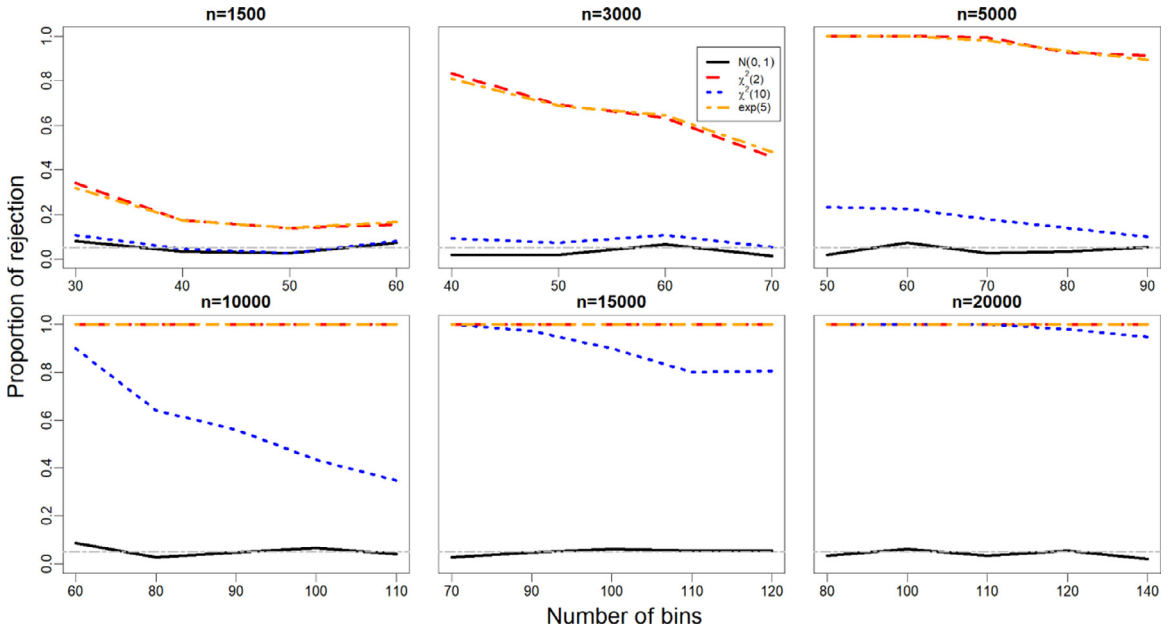

Fig. 12. The power and type I error of the projected zCDP GOF test with different numbers of bins based on the Gaussian noise, when sample sizes are 1500, 3000, 5000, 10000, 15000, and 20000. The gray dotted line in each plot indicates the significance level of the test, 0.05.

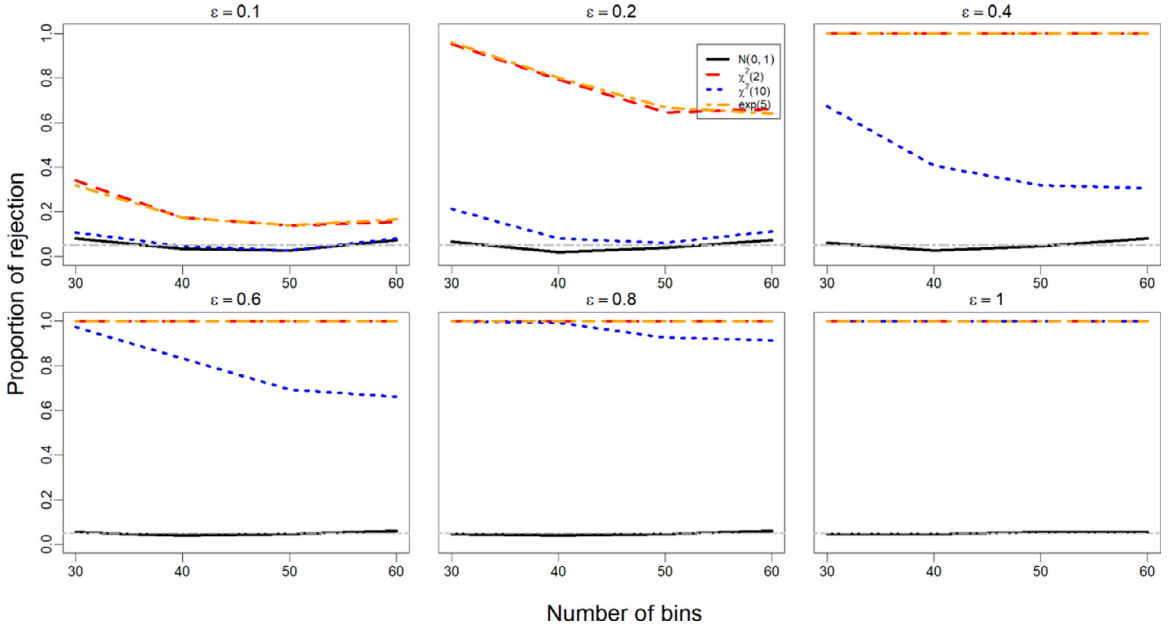

Fig. 13. The power and type I error of the projected zCDP GOF test with ϵ=0.1, 0.2, 0.4, 0.6, 0.8, 1.0, and $\delta = 10^{-5}$ when n =1500. The gray dotted line in each plot indicates the significance level of the test, 0.05.

mechanism based tests, MC-DP χ^2 (Gaussian and asymptotic) tests and zCDP (both projected and unprojected) tests, do not work well, even when the sample size is as large as 8000. On the contrary, the Laplacian mechanism-based ones produce the results close to those of the non-private test as sample size increases.

Overall, the results show that the proposed approach offers differentially private GOF tests for one-dimensional continuous random variables. The simulation confirms that all three types of tests achieve the given significance level, and the DP χ^2 and zCDP GOF tests have lower type I errors than the Priv'IT test by design. In terms of power, Priv'IT produces high power regardless of sample size and number of bins compared to the other tests. The MC-DP χ^2 GOF test based on the Laplace noise has greater power (Figure 4) compared to the Gaussian noise (Figure 6). We can also see from Figures 4, 6, and 8 that the Laplacian mechanism shows better performance under the same ϵ and sample size. This can be viewed as a

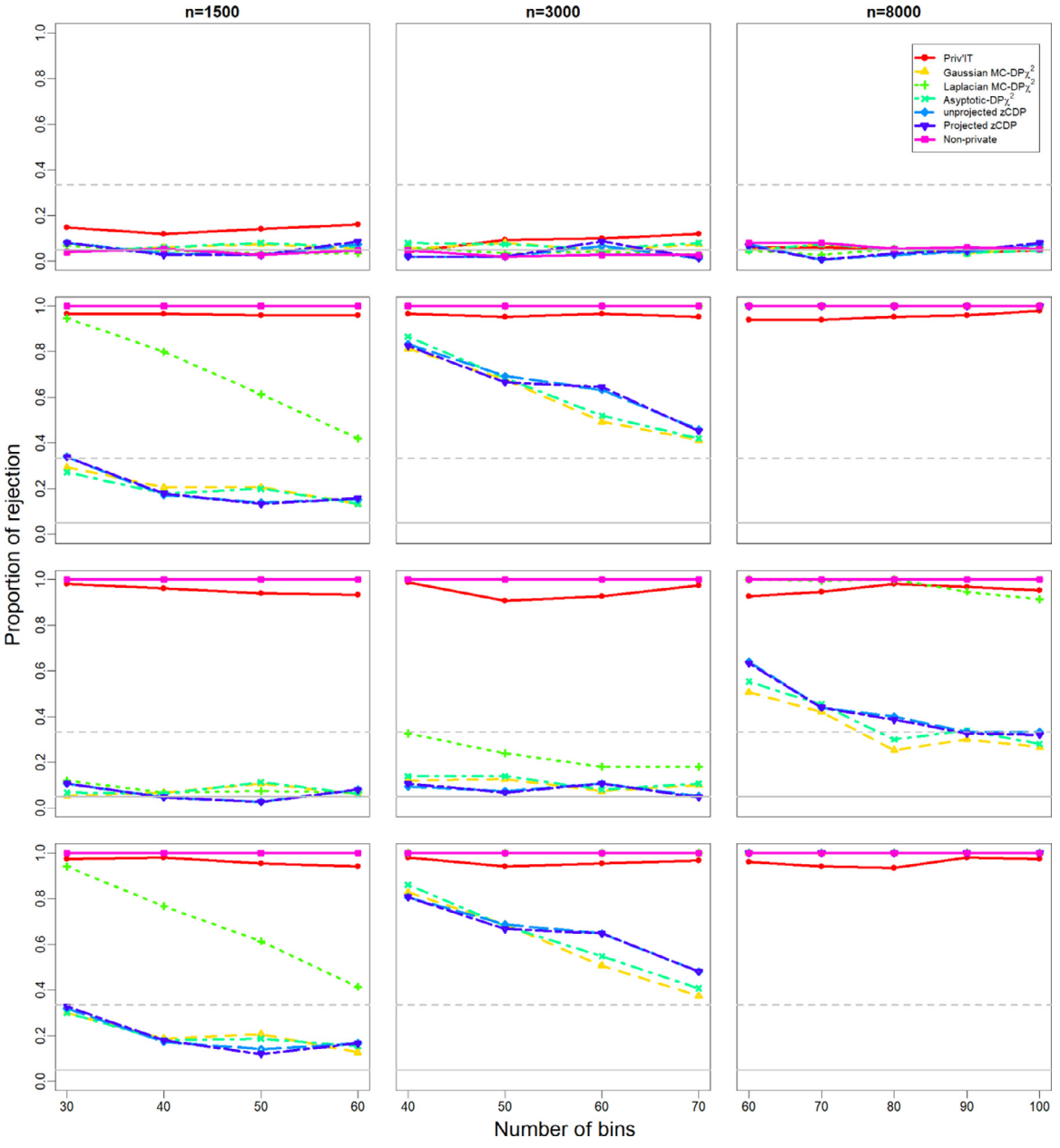

Fig. 14. Discretized GOF test results of Priv'IT, Gaussian MC-DP χ^2, Laplacian MC-DP χ^2, Asymptotic-DP χ^2, Unprojected zCDP, Projected zCDP, and non-private tests. Here, ϵ=0.1 and $\delta = 10^{-6}$ are used. Each row corresponds to the population distribution of samples, $N(0,1)$, $\chi^2(2)$, $\chi^2(10)$, and exp(5), respectively. The gray lines in each plot indicate the significance level of the test. The gray dotted lines indicate the significance level of 1/3 for Priv'IT and gray lines indicate 0.05 for the other tests.

comparison of private tests between $\delta = 0$ and $\delta > 0$ because the Laplacian mechanism guarantees $(\epsilon, 0)$-differential privacy while the Gaussian mechanism guarantees (ϵ, δ)-differential privacy. This may occur due to the large variance noise of the Gaussian mechanism. As δ gets close to 0, the standard deviation $\sigma^2(\epsilon, \delta)$ in the Gaussian mechanism becomes large since the standard deviation is proportional to $\sqrt{\log(1/\delta)}$. The (ϵ, δ)-differential privacy GOF tests and zCDP GOF tests, under the Gaussian mechanism, exhibit similar patterns in our simulation (Figures 6–13). It is also shown that the unprojected and projected zCDP GOF tests behave similarly in our settings.

The simulation also shows a possible drawback of the proposed discretization approach when an alternative distribution becomes similar to the hypothesized distribution during the randomization process. For example, when the alternative distribution is $\chi^2(10)$, the power of the tests tends to be low even when the sample size is 8000. This is because the difference between observed and expected counts of the bin gets relatively smaller than perturbation on the bin. In addition, when

Table 1
Non-private GOF test results: Kolmogorov-Smirnov and Anderson-Darling tests, and χ^2 GOF test after discretization. The letter R in the table represents the rejection of the null hypothesis.

Tests	K-S test	A-D test	χ^2 GOF test				
			80	90	110	120	130
Total Income	R	R	R	R	R	R	R
Wage and Salary	R	R	R	R	R	R	R
Business Income	R	R	R	R	R	R	R
Property Income	R	R	R	R	R	R	R

the sample size is 8000 and the number of bins is 100, the power tends to become a little less than those from smaller numbers of bins. It indicates that an appropriate number of bins is important in discretization.

We have conducted an additional simulation study with the Cauchy distribution, mixture of two or three distributions, and t distribution with small and large degrees of freedom and non central parameters. The main lessons from the results based on these distributions are consistent with those presented in this section. Hence, we have included the additional simulation results in Supplementary Material.

5. Real Data Analysis

In this section, we apply the proposed differentially private approach to real data. The dataset is from the Household Financial Welfare Survey of South Korea in 2018, which was collected by Statistics Korea (KOSTAT), Financial Supervisory Service (FSS), and Bank of Korea. The dataset can be freely downloaded from Microdata Integrated Service (https://mdis.kostat.go.kr/eng/index.do). The dataset includes assets and income of 18,640 South Koreans, and we focus on the income variable.

Income is the household income, which is the sum of the following types: wage and salary income, business income, property income, and transfer income. Wage and salary income is a house earning from working. Business income is earned from any business activity including house rent and equipment rental. Property income is income received by: virtue of owning properties such as land rent, owning financial assets such as interest, and ownership of capital equipment such as profit. Transfer income can be obtained from public or non-public sources. The public source is mainly government. Non-public sources include relatives, patrons, or sponsors. To conduct the private GOF tests, we consider not only the total income, but also wage and salary income, business income, and property income separately, which consist of about 80% of the total income. Figures 15 (a)–(d) display the histograms of four log-transformed income variables. The zero values of wage and salary income, business income, and property income are removed from analysis. We note that all four histogram show a skewed shape.

In conducting the differentially private GOF tests, we consider the following hypotheses,

$$H_0 : \mathbf{p} = \mathbf{p^0} \ \ vs \ \ H_1 : \mathbf{p} \neq \mathbf{p^0},$$

where $\mathbf{p^0}$ is a log-normal distribution. While a log-normal distribution is commonly used to describe the income distribution, there are other possible options(Campano and Salvatore, 2006).

We first conduct non-private GOF tests using the Kolmogorov-Smirnov test (Stephens, 1970), Anderson-Darling test (Anderson and Darling, 1954), and χ^2 GOF test, and then compare their results with those of the proposed differentially private GOF test. The Kolmogorov-Smirnov test and Anderson-Darling test can be directly applied to the continuous income variables; and, the non-private χ^2 GOF test is applied after the same equal probability discretization process described in Section 3 with the number of bins $d = 80, 90, \ldots, 130$. Table 1 reports the test results for the total income, wage and salary income, business income, and property income variables. All non-private tests conclude that all the variables considered in the tests do not follow a log-normal distribution.

Next, we conduct the proposed differentially private tests using the Priv'IT, MC-DP (Laplace and Gaussian) and Asymptotic-DP χ^2, and unprojected and projected zCDP GOF tests. Because a randomization process is involved in differentially private tests, all tests are repeated 150 times to compute the proportion of rejection.

The proportion of rejection for the total income, wage and salary income, business income, and property income variables are drawn in Figure 16 for the six tests. In Figure 16(a), the proportion of rejection for the total income variable suggest that the total income variable does not follow a log-normal distribution. It can be seen that the proportion of rejection stays around 1 or close to 1 for the MC-DP χ^2 and Priv'IT tests. However, it decreases with the number of bins for the zCDP and DP χ^2 GOF tests using the Gaussian mechanism, varying from 0.92 to 0.487. As discussed in Section 4, the Gaussian mechanism adds noise with larger variance compared to the Laplace mechanism.

The wage and salary income variable explains a large proportion of the total income variable since it has a large correlation coefficient ($r = 0.749$) and accounts for 52% of the total income. As can be seen in Figure 16 (b), the proportion of rejection stays close to 1 (from 0.947 to 1) for all tests and for the given range of number of bins. The second largest proportion of the total income comes from business income, which accounts for about 20%. From Figure 16(c), the proportion of rejection also stays close to 1 (from 0.867 to 1) for all tests and for the number of bins. The results for the property

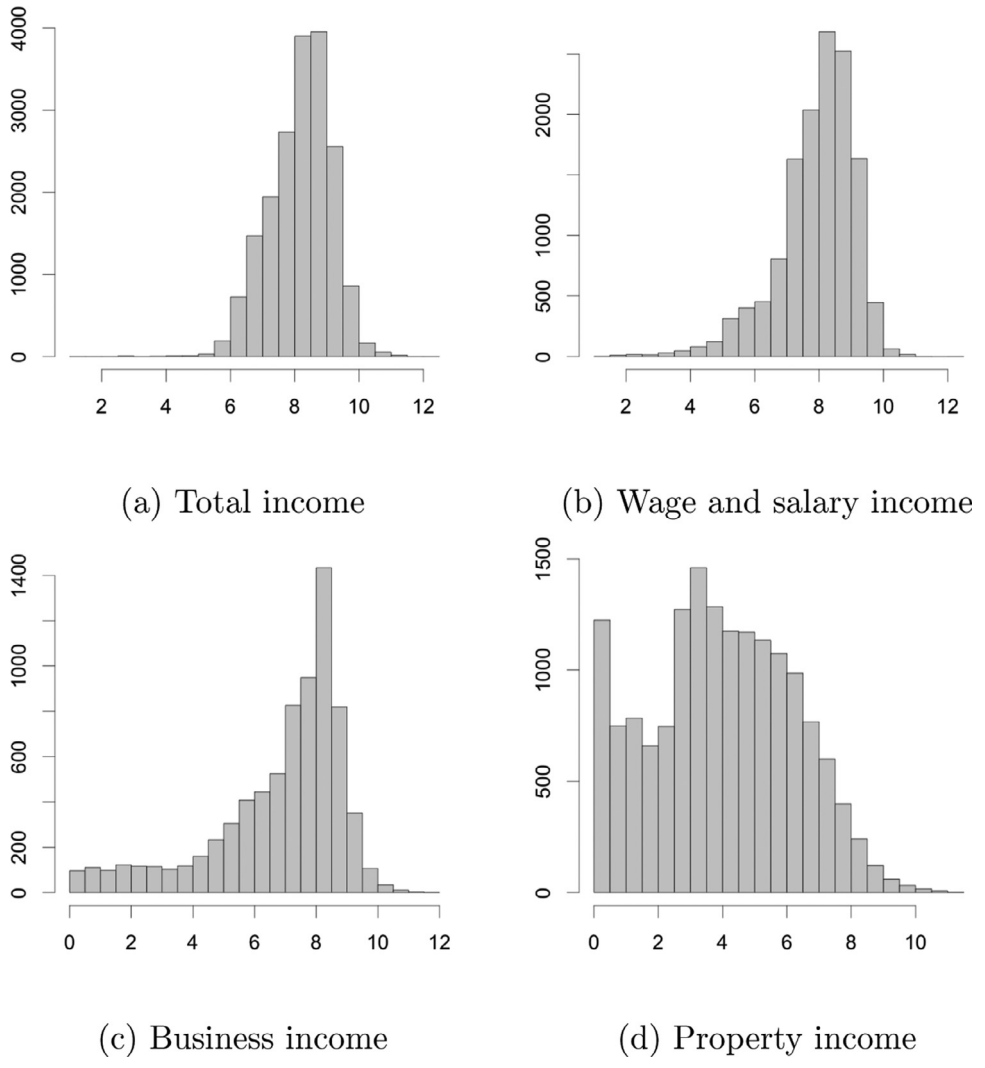

Fig. 15. Four log-transformed income variables.

income variable in Figure 16(d) also show that all tests yield high proportions of rejection, varying from 0.94 to 1 over the number of bins. The results for the three sub-variables coincide with the non-private tests that the variables do not follow a log-normal distribution, regardless of the tests and the number of bins.

6. Discussion

We develop differentially private GOF tests for continuous random variables by combining the equal probability discretization and differentially private GOF tests for discrete random variables such as Priv'IT, DP χ^2, and zCDP GOF tests. The discretization of a continuous random variable is essential to control the sensitivity of continuous variables. By discretizing a continuous distribution, the sensitivity becomes small and we can apply a differentially private GOF test for a discrete random variable to the discretized variable. The simulation results demonstrate that the discretization is an effective approach to achieve differential privacy of GOF tests for a one-dimensional continuous random variable. For the real data analysis, the conclusions of the differentially private GOF tests are consistent with those of the non-private GOF tests.

From the simulation and real data analysis, it can be seen that the number of bins affects the performance of some of the tests we consider. The proposed approach roughly determines the number of bins based on the sample size, while the recursive partitioning scheme of Balakrishnan and Wasserman (2019) computes it in a process of diving the domain. If we provide an appropriate number of bins for the discretization, the differentially private GOF tests would guarantee reasonable power of test while satisfying the level of test. Moreover, the proposed approach cannot be easily extended to high-dimensional continuous variables due to the curse of dimensionality. We propose differential privacy of GOF tests for high-dimensional continuous random variables as our future work.

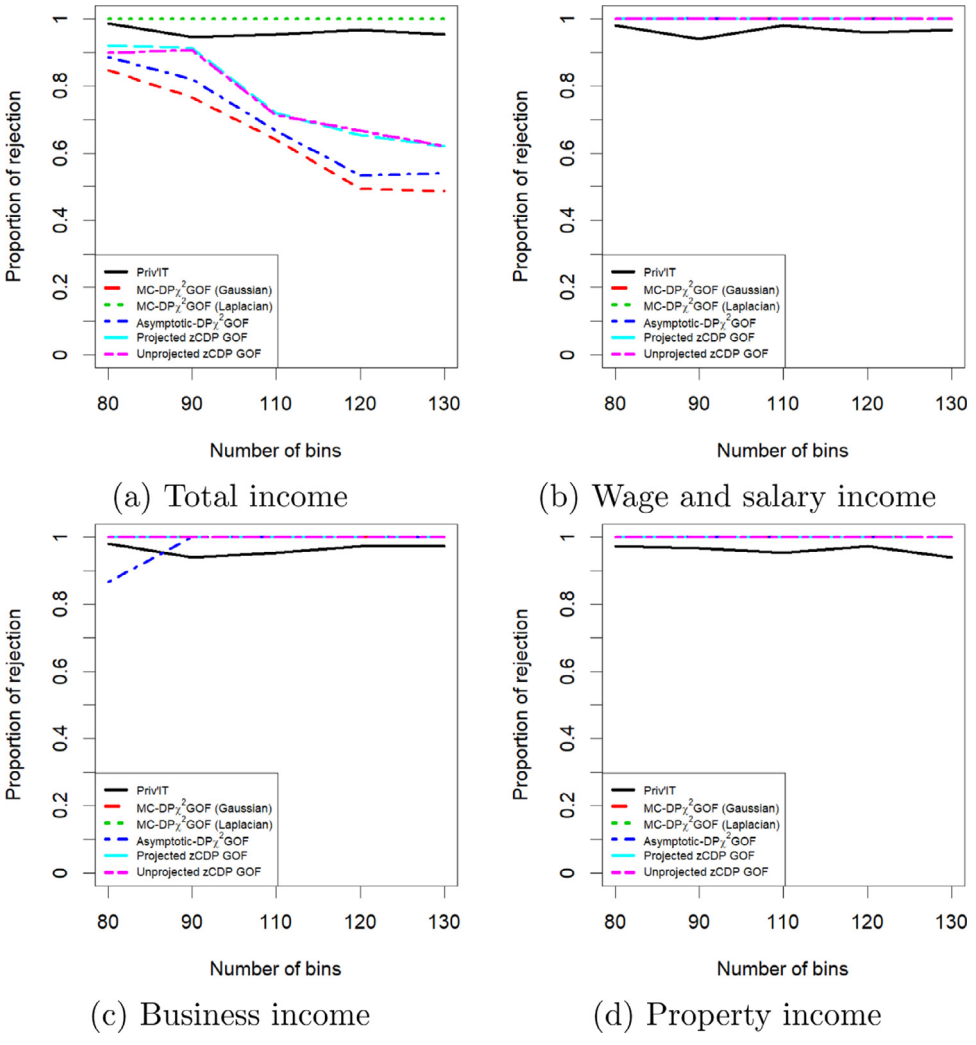

Fig. 16. Results from differentially private tests for four income variables.

In this work, we assume that observations are standardized thus do not need to estimate the location and scale parameters in the distributions. For sun-standardized data, one could apply a DP parameter estimation method (Amin et al., 2019; Kamath et al., 2019; Liu and Oh, 2019; Biswas et al., 2020; Brunel and Avella-Medina, 2020; Kamath et al., 2020; Tzamos et al., 2020). If DP GOF test results are released along with DP estimates, a separate privacy budget needs to be allocated for the entire procedure to satisfy (ϵ, δ)-differential privacy, based on the composition theorem.

Acknowledgement

We appreciate the comments from two reviewers and Associate Editor. Jeongyoun Ahn's work was supported in part by Basic Science Research Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A2C1093526). Jaewoo Lee's work was supported by the National Science Foundation under Grant No. CNS-1943046.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L., 2016. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. In: CCS '16. ACM, New York, NY, USA, pp. 308–318. doi:10.1145/2976749.2978318.

Amin, K., Dick, T., Kulesza, A., Munoz, A., Vassilvitskii, S., 2019. Differentially private covariance estimation. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d., Fox, E., Garnett, R. (Eds.), Advances in neural information processing systems, vol. 32. Curran Associates, Inc., pp. 14213–14222. https://proceedings.neurips.cc/paper/2019/file/4158f6d19559955bae372bb00f6204e4-Paper.pdf

Anderson, T.W., Darling, D.A., 1954. A test of goodness of fit. Journal of the American Statistical Association 49 (268), 765–769. http://www.jstor.org/stable/2281537

Balakrishnan, S., Wasserman, L., 2019. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. Ann. Statist. 47 (4), 1893–1927. doi:10.1214/18-AOS1729.

Beimel, A., Moran, S., Nissim, K., Stemmer, U., 2019. Private center points and learning of halfspaces. In: Beygelzimer, A., Hsu, D. (Eds.), Proceedings of the thirty-second conference on learning theory. In: Proceedings of Machine Learning Research, vol. 99. PMLR, Phoenix, USA, pp. 269–282. http://proceedings.mlr.press/v99/beimel19a.html

Berrett, T.B., Butucea, C., 2020. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. arXiv: Statistics Theory.

Biswas, S., Dong, Y., Kamath, G., & Ullman, J. (2020). CoinPress: Practical Private Mean and Covariance Estimation. arXiv e-prints, (p.arXiv:2006.06618).

Brunel, V.-E., & Avella-Medina, M. (2020). Propose, test, release: Differentially private estimation with high probability.

Bun, M., Carmosino, M.L., Sorrell, J., 2020. Efficient, noise-tolerant, and private learning via boosting. In: Abernethy, J., Agarwal, S. (Eds.), Proceedings of thirty third conference on learning theory. In: Proceedings of Machine Learning Research, vol. 125. PMLR, pp. 1031–1077. http://proceedings.mlr.press/v125/bun20a.html

Bun, M., Steinke, T., 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In: Hirt, M., Smith, A. (Eds.), Theory of cryptography. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 635–658.

Cai, B., Daskalakis, C., Kamath, G., 2017. Priv'IT: Private and sample efficient identity testing. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th international conference on machine learning. In: Proceedings of Machine Learning Research, vol. 70. PMLR, International Convention Centre, Sydney, Australia, pp. 635–644. http://proceedings.mlr.press/v70/cai17a.html

Campano, F., Salvatore, D., 2006. Income Distribution. Oxford University Press, Oxford.

Canonne, C. L., Kamath, G., McMillan, A., Ullman, J., & Zakynthinou, L. (2019). Private Identity Testing for High-Dimensional Distributions. arXiv e-prints, (p.arXiv:1905.11947).

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M., 2006a. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In: Advances in cryptology - EUROCRYPT 2006, 25th annual international conference on the theory and applications of cryptographic techniques. In: Lecture Notes in Computer Science, vol. 4004. Springer, pp. 486–503. doi:10.1007/11761679_29. https://iacr.org/archive/eurocrypt2006/40040493/40040493.pdf

Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006b. Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi, S., Rabin, T. (Eds.), Theory of cryptography. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 265–284.

Dwork, C., Roth, A., 2014. The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. 9 (3-4), 211–407. doi:10.1561/0400000042.

van Erven, T., Harremoës, P., 2014. Rényi divergence and kullback-leibler divergence. IEEE Transactions on Information Theory 60, 3797–3820.

Freedman, D.A., Diaconis, P., 1981. On the histogram as a density estimator:l2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 57 (4), 453–476. doi:10.1007/BF01025868.

Gaboardi, M., Lim, H., Rogers, R., Vadhan, S., 2016. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In: Balcan, M.F., Weinberger, K.Q. (Eds.), Proceedings of the 33rd international conference on machine learning. In: Proceedings of Machine Learning Research, vol. 48. PMLR, New York, New York, USA, pp. 2111–2120. http://proceedings.mlr.press/v48/rogers16.html

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. J. Mach. Learn. Res. 13, 723–773. http://dl.acm.org/citation.cfm?id=2188385.2188410

Kamath, G., Li, J., Singhal, V., Ullman, J., 2019. Privately learning high-dimensional distributions. In: Beygelzimer, A., Hsu, D. (Eds.), Proceedings of the thirty-second conference on learning theory. In: Proceedings of Machine Learning Research, vol. 99. PMLR, Phoenix, USA, pp. 1853–1902. http://proceedings.mlr.press/v99/kamath19a.html

Kamath, G., Singhal, V., Ullman, J., 2020. Private mean estimation of heavy-tailed distributions. In: Abernethy, J., Agarwal, S. (Eds.), Proceedings of thirty third conference on learning theory. In: Proceedings of Machine Learning Research, vol. 125. PMLR, pp. 2204–2235. http://proceedings.mlr.press/v125/kamath20a.html

Kaplan, H., Ligett, K., Mansour, Y., Naor, M., Stemmer, U., 2020. Privately learning thresholds: Closing the exponential gap. In: Abernethy, J.D., Agarwal, S. (Eds.), Proceedings of thirty third conference on learning theory. In: Proceedings of Machine Learning Research, vol. 125. PMLR, pp. 2263–2285. http://proceedings.mlr.press/v125/kaplan20a.html

Liu, Q., Lee, J.D., Jordan, M., 2016. A kernelized stein discrepancy for goodness-of-fit tests. In: Proceedings of the 33rd international conference on international conference on machine learning - volume 48. In: ICML'16. JMLR.org, pp. 276–284. http://dl.acm.org/citation.cfm?id=3045390.3045421

Liu, X., Oh, S., 2019. Minimax optimal estimation of approximate differential privacy on neighboring databases. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d., Fox, E., Garnett, R. (Eds.), Advances in neural information processing systems, vol. 32. Curran Associates, Inc., pp. 2417–2428. https://proceedings.neurips.cc/paper/2019/file/7a674153c63cff1ad7f0e261c369ab2c-Paper.pdf

McMahan, H.B., Ramage, D., Talwar, K., Zhang, L., 2018. Learning differentially private recurrent language models. In: International conference on learning representations. https://openreview.net/forum?id=BJ0hF1Z0b

Mironov, I., 2017. Rényi differential privacy. In: 30th IEEE computer security foundations symposium, CSF 2017, santa barbara, ca, usa, august 21-25, 2017. IEEE Computer Society, pp. 263–275. doi:10.1109/CSF.2017.11.

Narayanan, A., Shmatikov, V., 2008. Robust de-anonymization of large sparse datasets. In: Proceedings of the 2008 IEEE symposium on security and privacy. In: SP '08. IEEE Computer Society, Washington, DC, USA, pp. 111–125. doi:10.1109/SP.2008.33.

Reiter, J.P., 2004. New approaches to data dissemination: A glimpse into the future (?). CHANCE 17 (3), 11–15. doi:10.1080/09332480.2004.10554907.

Rogers, R., Kifer, D., 2017. A New Class of Private Chi-Square Hypothesis Tests. In: Singh, A., Zhu, J. (Eds.), Proceedings of the 20th international conference on artificial intelligence and statistics. In: Proceedings of Machine Learning Research, vol. 54. PMLR, Fort Lauderdale, FL, USA, pp. 991–1000. http://proceedings.mlr.press/v54/rogers17a.html

Stephens, M.A., 1970. Use of the kolmogorovsmirnov, cramérvon mises and related statistics without extensive tables. Journal of the Royal Statistical Society: Series B (Methodological) 32 (1), 115–122. doi:10.1111/j.2517-6161.1970.tb00821.x.

Sweeney, L. (2013). Matching known patients to health records in washington state data. Data Privacy Lab, IQSS, Harvard Universityhttp://thedatamap.org/risks.html.

Tzamos, C., Vlatakis-Gkaragkounis, E.-V., & Zadik, I. (2020). Optimal Private Median Estimation under Minimal Distributional Assumptions. arXiv e-prints, (p. arXiv:2011.06202).

Wang, Y.-X., 2018. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In: Globerson, A., Silva, R. (Eds.), Proceedings of the thirty-fourth conference on uncertainty in artificial intelligence, UAI 2018, monterey, california, usa, august 6-10, 2018. AUAI Press, pp. 93–103. http://auai.org/uai2018/proceedings/papers/40.pdf

Wasserman, L., Zhou, S., 2010. A statistical framework for differential privacy. Journal of the American Statistical Association 105 (489), 375–389. doi:10.1198/jasa.2009.tm08651.

Zhang, J., Zhang, Z., Xiao, X., Yang, Y., Winslett, M., 2012. Functional mechanism: Regression analysis under differential privacy. Proc. VLDB Endow. 5 (11), 13641375. doi:10.14778/2350229.2350253.