Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing

Marco Gaboardi GABOARDI@BUFFALO.EDU
University at Buffalo, SUNY

Hyun Woo Lim LIMHYUN@G.UCLA.EDU
University of California, Los Angeles

Ryan Rogers RYROGERS@SAS.UPENN.EDU
University of Pennsylvania

Salil P. Vadhan SALIL@SEAS.HARVARD.EDU
Harvard University

Abstract

Hypothesis testing is a useful statistical tool in determining whether a given model should be rejected based on a sample from the population. Sample data may contain sensitive information about individuals, such as medical information. Thus it is important to design statistical tests that guarantee the privacy of subjects in the data. In this work, we study hypothesis testing subject to differential privacy, specifically chi-squared tests for goodness of fit for multinomial data and independence between two categorical variables.

1. Introduction

Hypothesis testing provides a systematic way to test given models based on a sample, so that with high confidence a data analyst may conclude that the model is incorrect or not. However, these data samples may contain highly sensitive information about the subjects and so the privacy of individuals can be compromised when the results of a data analysis are released. For example, in the area of *genome-wide associaton studies* (GWAS) Homer et al. (2008) have shown that it is possible to identify subjects in a data set based on publicly available aggregate statistics.

A way to address this concern is by developing new techniques to support privacy-preserving data analysis. An approach that is gaining more and more attention by the statistics and data analysis community is *differential privacy* (Dwork et al., 2006b), which originated in theoretical computer science. In this work, we seek to develop hypothesis tests that are differentially private and that give conclusions similar to standard, non-private hypothesis tests.

We focus here on two classical tests for data drawn from a multinomial distribution: *goodness of fit test*, which determines whether the data was in fact drawn from a multinomial distribution with probability vector $\mathbf{p}^0$; and *independence test*, which tests whether two categorical random variables are independent of each other. Both tests depend on the *chi-squared* statistic, which is used to determine whether the data is likely or not under the given model.

To guarantee differential privacy, we consider adding Laplace and Gaussian noise to the counts of categorical data. Using the noisy data we can form a *private* chi-squared statistic. It turns out that the classical hypothesis tests perform poorly when used with this modified statistic because they ignore the fact that noise was added. To improve this situation, we develop four new tests that account for the additional noise due to privacy. In particular, we give two tests based on a *Monte Carlo* approach to testing the null hypothesis and two tests based on an *asymptotic* distribution of the private chi-squared distribution factoring in the noise distribution.

Our four differentially private tests achieve a target level $1-\alpha$ *significance*, i.e. they reject with probability at most α when the null hypothesis holds (in some cases, we provide a rigorous proof of this fact and in others, it is experimentally verified). This guarantees limited Type I errors. However, all of our tests do lose *power*; that is when the null hypothesis is false, they correctly reject with lower proba-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

bility than the classical hypothesis tests. This corresponds to an increase of Type II errors. We empirically show that we can recover a level of power similar to the one achieved by the classical versions by adding more samples.

2. Related Work

There has been a myriad of work dealing with the application of differential privacy in statistical inference. One of the first works that put differential privacy in the language of statistics is Wasserman and Zhou (2010), which studies rates of convergence of distributions based on differentially private data released from the *exponential mechanism* (McSherry and Talwar, 2007). In a result of great generality, Smith (2011) shows that for a wide class of statistics T, there is a differentially private statistic that converges in distribution to the same asymptotic distribution as T. However, having the correct asymptotic distribution does not ensure that only statistically significant conclusions are drawn at finite sample sizes, and indeed we observe that this fails dramatically for the most natural differentially private algorithms. Thus, we study how to ensure significance and optimize power at small sample sizes by focusing on two basic statistical tests.

A tempting first approach to developing a hypothesis test for categorical data that is also differentially private is to either add noise directly to the chi-squared statistic that will ensure differential privacy or to add noise to each cell count (as we do in this work) and use a classical test with the private counts. For the former method, the amount of noise that must be added to ensure privacy can be unbounded in the worst case. However, motivated by applications to genome-wide association studies (GWAS), Uhler et al. (2013) and Yu et al. (2014) place restrictions on the form of the data or what is known to the data analyst to reduce the scale of the noise that needs to be added. The work of Johnson and Shmatikov (2013) adds noise to each cell of a contingency table, but then uses classical statistical tests on the private version of the data, which we show can have very poor significance (see Figures 1 and 2). Additionally, Uhler et al. (2013) look at 3×2 contingency tables that are evenly split between the two columns, and study releasing differentially private χ^2-statistics of the most relevant SNPs for certain diseases by perturbing the table of counts, the χ^2-statistic itself, and the p-values for the underlying test. The only one of these works that explicitly examine significance and power in hypothesis testing (as we do here) is (Uhler et al., 2013), which shows that perturbing the p-values in independence testing does not perform much better than a random test, independent of a selected threshold, e.g. α. In fact, (Uhler et al., 2013) goes as far as to say that basing inference on perturbed p-values "seems impossible." An interesting direction for future work would be to apply the *distance-score mechanism* introduced by (Johnson and Shmatikov, 2013) and later improved by (Yu et al., 2014; Yu and Ji, 2014; Simmons and Berger, 2016), to achieve a target level of significance and high power in hypothesis testing for GWAS data.

If we assume that there is some prior estimates for the contingency table cell probabilities, Vu and Slavkovic (2009) determine the sample size adjustment for the Pearson chi-squared independence test that uses the private counts to achieve the same power as the test with the original counts. Several other works have shown negative experimental results for using classical inference on statistics that have been altered for differential privacy (Fienberg et al., 2010; Karwa and Slavković, 2012; Karwa and Slavković, 2016).

Another problem that arises when noise is added to the cells in a contingency table is that the entries may neither be positive nor sum to a known value n. Several works have focused on this problem, where they seek to release a contingency table in a differentially private way that also satisfies some known *consistency* properties of the underlying data (Barak et al., 2007; Li et al., 2010; Hardt et al., 2012; Li and Miklau, 2012; Gaboardi et al., 2014). For independence testing, we use techniques from Lee et al. (2015) to find the most likely contingency table given the noisy version of it so that we can then estimate the cell probabilities that generated the table. This two step procedure to estimate parameters given a differentially private statistic is inspired by the work of Karwa and Slavković (2016) for estimating parameters in the β-model for random graphs.

Independent of our work, Wang et al. (2015) also look at hypothesis testing with categorical data subject to differential privacy. They mainly consider adding Laplace noise to the data but point out that their method also generalizes to arbitrary noise distributions. However, in order to compute critical values, they resort to Monte Carlo methods to sample from the asymptotic distribution. Our Monte Carlo approach samples from the *exact* distribution from the underlying null hypothesis, which, unlike sampling from the asymptotic distribution, guarantees significance at least $1 - \alpha$ in goodness of fit tests at finite sample sizes. We only focus on Gaussian noise in our asymptotic analysis due to there being existing methods for finding tail probabilities (and hence critical values) for the resulting distributions, but our approaches can be generalized for arbitrary noise distributions. Further, we also consider the power of each of our differentially private tests.

3. Differential Privacy Preliminaries

We start with a brief overview of differential privacy. In order to define differential privacy, we first define *neighboring* databases $\mathbf{d}, \mathbf{d}'$ from some class of databases D^n where they differ in an individual's data but are equal among the rest of the data, e.g. $\mathbf{d} = (d_1, \cdots, d_i, \cdots d_n)$ and $\mathbf{d}' = (d_1, \cdots, d_i', \cdots, d_n)$ where $d_i \neq d_i'$. We will consider n to be known and public.

Definition 3.1 (Differential Privacy (Dwork et al., 2006b)). Let $M : D^n \to O$ be some randomized mechanism For $\epsilon, \delta > 0$ we say that M is (ϵ, δ)-differentially private if for any neighboring databases $\mathbf{d}, \mathbf{d}' \in D^n$ and any subset of outcomes $S \subseteq O$ we have

$$\Pr[M(\mathbf{d}) \in S] \leq e^{\epsilon} \Pr[M(\mathbf{d}') \in S] + \delta.$$

If $\delta = 0$, then we simply say M is ϵ-differentially private. As is common in the privacy literature, we will think of ϵ as a small constant, e.g. 0.1, and $\delta \ll 1/n$ as cryptographically small, where we sometimes write δ_n to explicitly show its dependence on n.

A typical differentially private mechanism is to add carefully calibrated noise to some quantity that a data analyst is interested in. We can release a differentially private answer to a function $\phi : D^n \to \mathbb{R}^d$ by adding independent noise to each component of ϕ. The scale of the noise we add depends on the impact any individual can have on the outcome. We use the *global sensitivity* of ϕ to quantify this impact, which we define for $i = 1, 2$ as:

$$GS_i(\phi) = \max_{\mathbf{d}, \mathbf{d}' \text{neighboring in } D^n} \{||\phi(\mathbf{d}) - \phi(\mathbf{d}')||_i\}.$$

Lemma 3.2 (Dwork et al. (2006a;b)). *Let $\phi : D^n \to \mathbb{R}^d$ have global sensitivity $GS_i(\phi)$ for $i = 1, 2$. Then the mechanism $M_{\mathcal{D}} : D^n \to \mathbb{R}^d$ where $M_{\mathcal{D}}(\mathbf{d}) = \phi(\mathbf{d}) + (Z_1, \cdots, Z_d)^T \quad \{Z_i\} \overset{i.i.d.}{\sim} \mathcal{D}$ is ϵ-differentially private if $\mathcal{D} = Laplace\left(\frac{GS_1(\phi)}{\epsilon}\right)$ or (ϵ, δ)-differentially private if $\mathcal{D} = N(0, \sigma^2)$ with $\sigma = \frac{GS_2(\phi)\sqrt{2\ln(2/\delta)}}{\epsilon}$.*

There are many useful properties of differentially private mechanisms. The one we will use in this paper is referred to as *post-processing*, which ensures privacy no matter what we do with the outcome of M.

Lemma 3.3. *(Dwork et al., 2006b)Let $M : D^n \to O$ be (ϵ, δ)-differentially private and $\psi : O \to O'$ be some arbitrary mapping from O to O'. Then $\psi \circ M : D^n \to O'$ remains (ϵ, δ)-differentially private.*

The tests that we present will be differentially private, assuming n is known and public, because we will add Laplace or Gaussian noise as in Lemma 3.2 to the vector of counts in goodness of fit testing

4. Hypothesis Testing Preliminaries

Given sampled data from a population, we wish to test whether the data came from a specific model, which is given as a *null hypothesis* H_0. We will denote our test as an algorithm $\mathcal{A}$ that takes a dataset $\mathbf{X}$, significance level $1 - \alpha$ and null hypothesis H_0 and returns a decision of whether to reject H_0 or not. We would like to design our test so that we achieve Type I error at most α, that is $\Pr[\mathcal{A}(\mathbf{X}; \alpha, H_0) = \text{Reject}|H_0] \leq \alpha$ while also achieving a small Type II error $\beta = \Pr[\mathcal{A}(\mathbf{X}; \alpha, H_0) = \text{Reject}|H_1]$ when the model is actually some *alternate* $H_1 \neq H_0$. Note that the probability is taken over the randomness from the data generation and the possible randomness from the algorithm $\mathcal{A}$ itself. It is common to refer to $1 - \alpha$ as the *significance* of test $\mathcal{A}$ and $1 - \beta$ as the *power* of $\mathcal{A}$. We think of bounding Type I error as a *hard constraint* in our tests and then hope to minimize Type II error.

5. Goodness of Fit Test

We consider $\mathbf{X} = (X_1, \cdots, X_d)^T \sim \text{Multinomial}(n, \mathbf{p})$ where $\mathbf{p} = (p_1, \cdots, p_d)$ and $\sum_{i=1}^d p_i = 1$. For a *goodness of fit test*, we want to test the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}^0$. A common way to test this is based on the *chi-squared* statistic Q^2 where $Q^2 = \sum_{i=1}^d \frac{\left(X_i - np_i^0\right)^2}{np_i^0}$. We present the classical chi-squared *goodness of fit test* in Algorithm 1, which compares the chi-squared statistic Q^2 to a threshold $\chi^2_{d-1,1-\alpha}$ that depends on a desired level of significance $1 - \alpha$ as well as the dimension of the data. The threshold $\chi^2_{d-1,1-\alpha}$ satisfies the following relationship: $\Pr\left[\chi^2_{d-1} \geq \chi^2_{d-1,1-\alpha}\right] = \alpha$. where χ^2_{d-1} is a chi-squared random variable with $d - 1$ degrees of freedom.

Algorithm 1 Goodness of Fit Test for Multinomial Data

procedure $\texttt{GOF}(\mathbf{x}, \alpha, H_0 : \mathbf{p} = \mathbf{p}^0)$
  Compute Q^2.
  if $Q^2 > \chi^2_{d-1,1-\alpha}$ **then** Decision $\leftarrow$ Reject
  else Decision $\leftarrow$ Fail to Reject
  return Decision.

The reason why we compare Q^2 with the chi-squared distribution is because of the following classical result.

Theorem 5.1. *(Bishop et al., 1975) Assuming $H_0 : \mathbf{p} = \mathbf{p}^0$ holds, the statistic Q^2 converges in distribution to a chi-squared with $d - 1$ degrees of freedom, i.e. $Q^2 \xrightarrow{D} \chi^2_{d-1}$.*

Note that this does not guarantee that $\Pr\left[Q^2 > \chi^2_{d-1,1-\alpha}\right] \leq \alpha$ for finite samples, nevertheless the test works well and is widely used in practice.

5.1. Differentially Private Chi-Squared Statistic

To ensure differential privacy, we add independent noise to each component of $\mathbf{X}$, which we will either use Laplace or Gaussian noise. The function g that outputs the counts in the d cells has global sensitivity $GS_1(g) = 2$ and $GS_2(g) = \sqrt{2}$ because one individual may move from one cell count (decreasing the count by 1) to another (increasing the cell count by 1). We then form the *private chi-squared* statistic $Q^2_{\mathcal{D}}$ based on the noisy counts,

$$Q^2_{\mathcal{D}} = \sum_{i=1}^{d} \frac{\left(X_i + Z_i - np_i^0\right)^2}{np_i^0}, \qquad \{Z_i\} \overset{i.i.d.}{\sim} \mathcal{D} \quad (1)$$

where the distributions for the noise that we consider include $\mathcal{D} = Laplace(2/\epsilon)$ and $\mathcal{D} = N(0, \sigma^2)$ where $\sigma \equiv \sigma(\epsilon, \delta_n) = \frac{2\sqrt{\ln(2/\delta_n)}}{\epsilon}$. We will denote the ϵ-differentially private statistic as Q^2_{Lap} and the (ϵ,δ)-differentially private statistic as Q^2_{Gauss} based on whether we use Laplace or Gaussian noise, respectively. Recall that in the original goodness of fit test without privacy in Algorithm 1 we compare the distribution of Q^2 with that of a chi-squared random variable with $d-1$ degrees of freedom. We show in the supplementary materials that $Q^2_{\mathcal{D}}$ still has the same asymptotic distribution, with certain conditions on δ_n.

It then seems natural to use GOF on the private chi-squared statistic as if we had the actual chi-squared statistic that did not introduce noise to each count since both private and nonprivate statistics have the same asymptotic distribution. We will show in our results in Section 7 that if we were to simply compare the private statistic to the critical value $\chi^2_{d-1,1-\alpha}$, we will typically not get a good significance level even for relatively large n. In the following lemma we show that for every realization of data, the statistic $Q^2_{\mathcal{D}}$ is expected to be larger than the actual chi-squared statistic Q^2. See the supplementary materials for the proof.

Lemma 5.2. *For each realization $\mathbf{X} = \mathbf{x}$, we have $\mathbb{E}_{\mathcal{D}}\left[Q^2_{\mathcal{D}}|\mathbf{x}\right] \geq Q^2$, where $\mathcal{D}$ has mean zero.*

This result suggests that the significance threshold for the private version of the chi-squared statistic $Q^2_{\mathcal{D}}$ should be higher than the standard one. Otherwise, we would reject H_0 too easily using the classical test, which we show in our experimental results. This motivates the need to develop new tests that account for the distribution of the noise.

5.2. Monte Carlo Test: MCGOF$_{\mathcal{D}}$

Given some null hypothesis $\mathbf{p}^0$ and statistic $Q^2_{\mathcal{D}}$, we want to determine a threshold τ^α such that $Q^2_{\mathcal{D}} > \tau^\alpha$ at most an α fraction of the time when the null hypothesis is true. As a first approach, we determine threshold τ^α using a Monte Carlo (MC) approach by sampling from the distribution of $Q^2_{\mathcal{D}}$, where $\mathbf{X} \sim \text{Multinomial}(n, \mathbf{p}^0)$ and $\{Z_i\} \overset{i.i.d.}{\sim} \mathcal{D}$ for both Laplace and Gaussian noise. We give our MC based test MCGOF in Algorithm 2. We show in the supplementary materials that MCGOF achieves at least our target significance $1-\alpha$ when we choose $k > 1/\alpha$ many samples.

Algorithm 2 MC Goodness of Fit

procedure MCGOF$_{\mathcal{D}}(\mathbf{x}, (\epsilon,\delta), \alpha, H_0 : \mathbf{p} = \mathbf{p}^0)$
  Compute $q = Q^2_{\mathcal{D}}$ (1).
  Select $k > 1/\alpha$.
  Sample $q_1, \cdots, q_k$ i.i.d. from the distribution of $Q^2_{\mathcal{D}}$.
  Sort the samples $q_{(1)} \leq \cdots \leq q_{(k)}$.
  Compute threshold $q_{(t)}$ where $t = \lceil (k+1)(1-\alpha) \rceil$.
  if $q > q_{(t)}$ **then** Decision $\leftarrow$ Reject
  else Decision $\leftarrow$ Fail to Reject
  return Decision

Theorem 5.3. *The test MCGOF$_{\mathcal{D}}(\cdot, (\epsilon,\delta), \alpha, \mathbf{p}^0)$ has significance at least $1-\alpha$, i.e. $\Pr\left[\text{MCGOF}_{\mathcal{D}}(\mathbf{X}, (\epsilon,\delta), \alpha, \mathbf{p}^0) = \text{Reject} \,|H_0\right] \leq \alpha$.*

In Section 7, we present the empirical power results for MCGOF$_{\mathcal{D}}$ (along with all our other tests) when we fix an alternative hypothesis.

5.3. Asymptotic Approach: PrivGOF

In this section we attempt to determine a more analytical approximation to the distribution of Q^2_{Gauss}. We focus on Gaussian noise because it is more compatible with the asymptotic analysis of GOF, as opposed to Laplace noise. Consider the random vector $\mathbf{U} = (U_1, \cdots, U_d)$ where $U_i = \frac{X_i - np_i^0}{\sqrt{np_i^0}}$ for any $i \in [d]$. We then introduce the Gaussian noise random vector as $\mathbf{V} = (Z_1/\sigma(\epsilon,\delta_n), \cdots, Z_d/\sigma(\epsilon,\delta_n))^T \sim N(\mathbf{0}, I_d)$. Let $\mathbf{W} \in \mathbb{R}^{2d}$ be the concatenated vector defined as $\mathbf{W} = \binom{\mathbf{U}}{\mathbf{V}}$.

Note that $\mathbf{W} \overset{D}{\to} N(\mathbf{0}, \Sigma')$ where the covariance matrix is the $2d$ by $2d$ block matrix

$$\Sigma' = \begin{bmatrix} \Sigma & 0 \\ 0 & I_d \end{bmatrix}, \quad \text{where} \quad \Sigma = I_d - \sqrt{\mathbf{p}^0}\sqrt{\mathbf{p}^0}^T \quad (2)$$

Since Σ is idempotent, so is Σ'. We next define the $2d$ x $2d$ positive semi-definite matrix A (composed of four d by d block matrices) as

$$A = \begin{bmatrix} I_d & \Lambda \\ \Lambda & \Lambda^2 \end{bmatrix} \quad \text{where} \quad \Lambda = \text{Diag}\left(\frac{\sigma(\epsilon,\delta_n)}{\sqrt{n\mathbf{p}^0}}\right) \quad (3)$$

We can then rewrite our private chi-squared statistic as a quadratic form $Q^2_{Gauss} = \mathbf{W}^T A \mathbf{W}$.

Remark 5.4. If we have $\sigma(\epsilon,\delta_n)/\sqrt{n\mathbf{p}^0} \to$ constant then the asymptotic distribution of Q^2_{Gauss} would be a quadratic form of multivariate normals.

Similar to the classical goodness of fit test we consider the limiting case that the random vector $\mathbf{U}$ is actually a multivariate normal, which will result in $\mathbf{W}$ being multivariate normal as well. We next want to be able to calculate the distribution of the quadratic form of normals $\mathbf{W}^T A \mathbf{W}$. Note that we will write $\{\chi_1^{2,i}\}_{i=1}^r$ as a set of r independent chi-squared random variables with one degree of freedom.

Theorem 5.5. *Let $\mathbf{W} \sim N(\mathbf{0}, \Sigma')$ where Σ' is idempotent and has rank $r \leq 2d$. Then the distribution of $\mathbf{W}^T A \mathbf{W}$ where A is positive semi-definite is $\sum_{i=1}^r \lambda_i \chi_1^{2,i}$ where $\{\lambda_i\}_{i=1}^r$ are the eigenvalues of $B^T A B$ where $B \in \mathbb{R}^{2d \times r}$ such that $BB^T = \Sigma'$ and $B^T B = I_r$.*

Note that in the non-private case, the coefficients $\{\lambda_i\}$ in Theorem 5.5 become the eigenvalues for the rank $d-1$ idempotent matrix Σ, thus resulting in a χ^2_{d-1} distribution. We use the result of Theorem 5.5 in order to find a threshold that will achieve the desired significance level $1-\alpha$, as in the classical chi-squared goodness of fit test. We then set the threshold τ^α to satisfy the following:

$$\Pr\left[\sum_{i=1}^r \lambda_i \chi_1^{2,i} \geq \tau^\alpha\right] = \alpha \tag{4}$$

for $\{\lambda_i\}$ found in Theorem 5.5. Note, the threshold τ^α is a function of $n, \epsilon, \delta, \alpha$ and $\mathbf{p}^0$, but not the data.

We present our modified goodness of fit test when we are dealing with differentially private counts in Algorithm 3.

Algorithm 3 Private Chi-Squared Goodness of Fit Test

procedure `PrivGOF`$(\mathbf{x}, (\epsilon, \delta), \alpha, H_0 : \mathbf{p} = \mathbf{p}^0)$
  Set $\sigma = 2\sqrt{\log(2/\delta)}/\epsilon$.
  Compute Q^2_{Gauss} from (1) and τ^α that satisfies (4).
  if $Q^2_{Gauss} > \tau^\alpha$ **then** Decision $\leftarrow$ Reject
  else Decision $\leftarrow$ Fail to Reject
  return Decision

6. Independence Testing

We now consider the problem of testing whether two random variables $\mathbf{Y}^{(1)} \sim \mathrm{Multinomial}(1, \pi^{(1)})$ and $\mathbf{Y}^{(2)} \sim \mathrm{Multinomial}(1, \pi^{(2)})$ are independent of each other. We then form the null hypothesis $H_0 : \mathbf{Y}^{(1)} \perp \mathbf{Y}^{(2)}$, i.e. they are independent. One approach to testing H_0 is to sample n joint outcomes of $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ and count the number of observed outcomes, $X_{i,j}$ which is the number of times $Y_i^{(1)} = 1$ and $Y_j^{(2)} = 1$ in the n trials, so that we can summarize all joint outcomes as a contingency table $\mathbf{X} = (X_{i,j}) \sim \mathrm{Multinomial}(n, \mathbf{p})$, where $p_{i,j}$ is the probability that $Y_i^{(1)} = 1$ and $Y_j^{(2)} = 1$. We will write the full contingency table of counts $\mathbf{X} = (X_{i,j})$ as a vector with the ordering convention that we start from the top row and move from left to right across the contingency table.

We want to calculate the chi-squared statistic as in the goodness of fit section (where now the summation is over all joint outcomes i and j), but now we do not know the true proportion $\mathbf{p} = (p_{i,j})$ which depends on $\pi^{(1)}$ and $\pi^{(2)}$. However, we can use the *maximum likelihood estimator* (MLE) $\widehat{\mathbf{p}}$ for the probability vector $\mathbf{p}$ subject to H_0 to form the statistic $\widehat{Q}^2 = \sum_{i,j} \frac{(X_{i,j} - n\widehat{p}_{i,j})^2}{n\widehat{p}_{i,j}}$. Note that under the null hypothesis we can write $\mathbf{p}$ as a function of $\pi^{(1)}$ and $\pi^{(2)}$,

$$\mathbf{p} = \mathbf{f}(\pi^{(1)}, \pi^{(2)}) \text{ where } f_{i,j}(\pi^{(1)}, \pi^{(2)}) = \pi_i^{(1)} \cdot \pi_j^{(2)}. \tag{5}$$

Further, we can write the MLE $\widehat{\mathbf{p}}$ as described below.

Lemma 6.1 ((Bishop et al., 1975)). *Given $\mathbf{X}$, which is n samples of joint outcomes of $\mathbf{Y}^{(1)} \sim \mathrm{Multinomial}(1, \pi^{(1)})$ and $\mathbf{Y}^{(2)} \sim \mathrm{Multinomial}(1, \pi^{(2)})$, if $\mathbf{Y}^{(1)} \perp \mathbf{Y}^{(2)}$, then the MLE for $\mathbf{p} = \mathbf{f}(\pi^{(1)}, \pi^{(2)})$ for $\mathbf{f}$ given in (5) is the following: $\widehat{\mathbf{p}} = \mathbf{f}(\widehat{\pi}^{(1)}, \widehat{\pi}^{(2)})$ where*

$$\widehat{\pi}_i^{(1)} = X_{i,\cdot}/n, \; \widehat{\pi}_j^{(2)} = X_{\cdot,j}/n \quad \textit{for } i \in [r], j \in [c] \tag{6}$$

where $X_{i,\cdot} = \sum_{j=1}^r X_{i,j}$ and $X_{\cdot,j} = \sum_{i=1}^c X_{i,j}$.

We then state another classical result that gives the asymptotic distribution of $\widehat{Q}^2$ given H_0.

Theorem 6.2. *(Bishop et al., 1975) Given the assumptions in Lemma 6.1, the statistic $\widehat{Q}^2 \xrightarrow{D} \chi^2_\nu$ for $\nu = (r-1)(c-1)$.*

Algorithm 4 Pearson Chi-Squared Independence Test

procedure `Indep`$(\mathbf{x}, \alpha)$
  $\widehat{\mathbf{p}} \leftarrow$ MLE calculation in (6)
  Compute $\widehat{Q}^2$ and set $\nu = (r-1)(c-1)$.
  if $\widehat{Q}^2 > \chi^2_{\nu,1-\alpha}$ and all entries of $\mathbf{x}$ are at least 5 **then**
  Decision $\leftarrow$ Reject
  else Decision $\leftarrow$ Fail to Reject
  return Decision.

The chi-squared independence test is then to compare the statistic $\widehat{Q}^2$, with the value $\chi^2_{\nu,1-\alpha}$ for a $1-\alpha$ significance test. We formally give the Pearson Chi-Squared test in Algorithm 4. An often used "rule of thumb" (Triola, 2014) with this test is that it can only be used if all the cell counts are at least 5, otherwise the test Fails to Reject H_0. We will follow this rule of thumb in our tests.

Similar to our prior analysis for goodness of fit, we aim to understand the asymptotic distribution from Theorem 6.2. First, we can define $\widehat{\mathbf{U}}$ in terms of the MLE $\widehat{\mathbf{p}}$ given in (6):

$$\widehat{U}_{i,j} = (X_{i,j} - n\widehat{p}_{i,j})/\sqrt{n\widehat{p}_{i,j}}. \tag{7}$$

The following classical result gives the asymptotic distribution of $\widehat{\mathbf{U}}$ under H_0, which also proves Theorem 6.2.

Lemma 6.3. *(Bishop et al., 1975) With the same hypotheses as Lemma 6.1, the random vector $\widehat{\mathbf{U}} \xrightarrow{D} N(0, \Sigma_{ind})$ where $\Sigma_{ind} = I_{rc} - \sqrt{\mathbf{p}} \cdot \sqrt{\mathbf{p}}^T - \Gamma(\Gamma^T\Gamma)^{-1}\Gamma^T$ with $\mathbf{f}$ given in (5), and $\Gamma = \mathrm{Diag}(\sqrt{\mathbf{p}})^{-1} \cdot \nabla \mathbf{f}(\pi^{(1)}, \pi^{(2)})$.*

In order to do a test that is similar to `Indep` given in Algorithm 4, we need to determine an estimate for $\pi^{(1)}$ and $\pi^{(2)}$ where we are only given access to the noisy cell counts.

6.1. Estimating Parameters with Private Counts

We now assume that we do not have access to the counts $X_{i,j}$ in a contingency table but instead we have $W_{i,j} = X_{i,j} + Z_{i,j}$ where $Z_{i,j} \sim \mathcal{D}$ for Laplace or Gaussian noise given in (1) and we want to perform a test for independence. We consider the full likelihood of the noisy $r \times c$ contingency table $\Pr\left[\mathbf{X} + \mathbf{Z} = \mathbf{w} | H_0, \pi^{(1)}, \pi^{(2)}\right]$ to find the best estimates for $\{\pi^{(i)}\}$ given the noisy counts.

Algorithm 5 Two Step MLE Calculation

procedure $\texttt{2MLE}(\mathbf{X}+\mathbf{Z}=\mathbf{w})$
- $\widetilde{\mathbf{x}} \leftarrow$ Solution to (8).
- **if** $\mathcal{D} =$ Gauss **then** set $\gamma = 1$
- **if** $\mathcal{D} =$ Lap **then** set $0 < \gamma \ll 1$.
- **if** Any cell of $\widetilde{\mathbf{x}}$ is less than 5 **then** $\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)} \leftarrow$ NULL
- **else** $\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)} \leftarrow$ MLE with $\widetilde{\mathbf{x}}$ given in (6).
- **return** $\widetilde{\pi}^{(1)}$ and $\widetilde{\pi}^{(2)}$.

We will follow a two step procedure similar to the work of (Karwa and Slavković, 2016). We will first find the most likely contingency table given the noisy data $\mathbf{w}$ and then find the most likely probability vectors under the null hypothesis that could have generated that denoised contingency table (this is not equivalent to maximizing the full likelihood, but it seems to work well as our experiments later show). For the latter step, we use Equation (6) to get the MLE for $\pi^{(1)}$ and $\pi^{(2)}$ given a vector of counts $\mathbf{x}$. For the first step, we need to minimize $||\mathbf{w} - \mathbf{x}||$ subject to $\sum_{i,j} x_{i,j} = n$ and $x_{i,j} \geq 0$ where the norm in the objective is either ℓ_1 for Laplace noise or ℓ_2 for Gaussian noise.

Note that for Laplace noise, the above optimization problem does not give a unique solution and it is not clear which contingency table $\mathbf{x}$ to use. One solution to overcome this is to add a *regularizer* to the objective value. We will follow the work of (Lee et al., 2015) to overcome this problem by using an *elastic net* regularizer (Zou and Hastie, 2005):

$$\begin{aligned} \underset{\mathbf{x}}{\text{argmin}} \quad & (1-\gamma) \cdot ||\mathbf{w} - \mathbf{x}||_1 + \gamma \cdot ||\mathbf{w} - \mathbf{x}||_2^2 \\ s.t. \quad & \sum_{i,j} x_{i,j} = n, \qquad x_{i,j} \geq 0. \end{aligned} \tag{8}$$

where if we use Gaussian noise, we set $\gamma = 1$ and if we use Laplace noise then we pick a small $\gamma > 0$ and then solve the resulting program. Our two step procedure for finding an approximate MLE for $\pi^{(1)}$ and $\pi^{(2)}$ based on our noisy vector of counts $\mathbf{w}$ is given in Algorithm 5, where we take into account the rule of thumb from `Indep` and return NULL if any computed table has counts less than 5.

We will denote $\widetilde{\mathbf{p}}$ to be the probability vector of $\mathbf{f}$ from (5) applied to the result of $\texttt{2MLE}(\mathbf{X}+\mathbf{Z})$. We now write down the private chi-squared statistic when we use the estimate $\widetilde{\mathbf{p}}$ in place of the actual (unknown) probability vector $\mathbf{p}$:

$$\widetilde{Q}^2_{\mathcal{D}} = \sum_{i,j} \frac{\left(X_{i,j} + Z_{i,j} - n\widetilde{p}_{i,j}\right)^2}{n\widetilde{p}_{i,j}} \quad \{Z_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{D}. \tag{9}$$

Algorithm 6 MC Independence Testing

procedure $\texttt{MCIndep}_{\mathcal{D}}(\mathbf{x}, (\epsilon, \delta), \alpha)$
- $\mathbf{w} \leftarrow \mathbf{x} + \mathbf{Z}$, where $\{Z_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{D}$ and $\mathcal{D}$ given in (1).
- $(\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)}) \leftarrow \texttt{2MLE}(\mathbf{w})$ and $\widetilde{\mathbf{p}} \leftarrow \mathbf{f}(\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)})$.
- **if** $(\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)}) ==$ NULL **then**
  - **return** Fail to Reject
- **else**
  - $\widetilde{q} \leftarrow \widetilde{Q}^2_{\mathcal{D}}$ using $\mathbf{w}$ and $\widetilde{\mathbf{p}}$.
  - Set $k > 1/\alpha$ and $q \leftarrow$ NULL.
  - **for** $t \in [k]$ **do**
    - Generate contingency table $\widetilde{\mathbf{x}}$ using $\left(\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)}\right)$.
    - $\widetilde{\mathbf{w}} \leftarrow \widetilde{\mathbf{x}} + \mathbf{Z}$, where $\{Z_{i,j}\} \overset{i.i.d.}{\sim} \mathcal{D}$.
    - $(\widetilde{\widetilde{\pi}}^{1}, \pi^{(2)}) \leftarrow \texttt{2MLE}(\widetilde{\mathbf{w}})$.
    - **if** $(\widetilde{\widetilde{\pi}}^{1}, \widetilde{\widetilde{\pi}}^{2}) ==$ NULL **then**
    - **return** Fail to Reject.
    - **else** Compute $\widetilde{Q}^2_{\mathcal{D}}$ from (9), add it to array q.
  - $\widetilde{\tau}^{\alpha} \leftarrow$ the $\lceil (k+1)(1-\alpha) \rceil$ ranked statistic in q.
  - **if** $\widetilde{q} > \widetilde{\tau}^{\alpha}$ **then** **return** Reject H_0.
  - **else** **return** Fail to Reject H_0.

6.2. Monte Carlo Test: $\texttt{MCIndep}_{\mathcal{D}}$

We first follow a similar procedure as in Section 5.2 but using the parameter estimates from `2MLE` instead of the actual (unknown) probabilities. Our procedure $\texttt{MCIndep}_{\mathcal{D}}$ (given in Algorithm 6) works as follows: given a dataset $\mathbf{x}$, we will add the appropriately scaled Laplace or Gaussian noise to ensure differential privacy to get the noisy table $\mathbf{w}$. Then we use `2MLE` on the private data to get approximates to the parameters $\pi^{(i)}$, which we denote as $\widetilde{\pi}^{(i)}$ for $i = 1, 2$. Using these probability estimates, we sample $k > 1/\alpha$ many contingency tables and noise terms to get k different values for $\widetilde{Q}^2_{\mathcal{D}}$ and choose the $\lceil (k+1)(1-\alpha) \rceil$ ranked statistic as our threshold $\widetilde{\tau}^{\alpha}$. If at any stage `2MLE` returns NULL, then the test Fails to Reject H_0. We formally give our test $\texttt{MCIndep}_{\mathcal{D}}$ in Algorithm 6.

6.3. Asymptotic Approach: `PrivIndep`

We will now focus on the analytical form of our private statistic when Guassian noise is added. We can then write $\widetilde{Q}^2_{Gauss} = \widetilde{\mathbf{W}}^T \widetilde{A} \widetilde{\mathbf{W}}$ in its quadratic form, which is similar

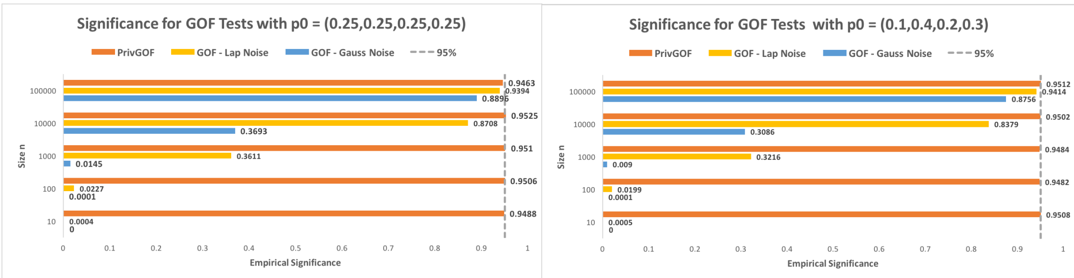

Figure 1. Significance of test `PrivGOF` in $10,000$ trials with $(\epsilon, \delta) = (0.1, 10^{-6})$.

Algorithm 7 Private Independence Test for $r \times c$ tables

procedure $\texttt{PrivIndep}(\mathbf{x}, (\epsilon, \delta), 1-\alpha)$
  $\mathbf{w} \leftarrow \mathbf{x} + \mathbf{Z}$ where $\mathbf{Z} \sim N(0, \sigma^2 I_{r \cdot c})$.
  $\left(\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)}\right) \leftarrow \texttt{2MLE}(\mathbf{w})$.
  if $\left(\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)}\right) ==$ NULL **then**
    return Fail to Reject
  else
    $\widetilde{\mathbf{p}} \leftarrow \mathbf{f}\left(\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)}\right)$ for $\mathbf{f}$ given in (5).
    Set $\widetilde{Q}^2_{Gauss}$ from (9) with $\mathbf{w}$ and $\widetilde{\tau}^\alpha$ from (11).
    if $\widetilde{Q}^2_{Gauss} > \widetilde{\tau}^\alpha$ **then** **return** Reject
    else **return** Fail to Reject

to the form of Q^2_{Gauss} where $\widetilde{\mathbf{W}} = \binom{\widetilde{\mathbf{U}}}{\mathbf{V}}$ with $\widetilde{\mathbf{U}}$ set in (7) except with $\widetilde{\mathbf{p}}$ used instead of the given $\mathbf{p}^0$ in the goodness of fit testing. Further, we denote $\widetilde{A}$ as A in (3) but with $\widetilde{\mathbf{p}}$ instead of $\mathbf{p}^0$. We will use the $2rc$ by $2rc$ block matrix $\widetilde{\Sigma}'_{ind}$ to estimate the covariance of $\widetilde{\mathbf{W}}$, where

$$\widetilde{\Sigma}'_{ind} = \begin{bmatrix} \widetilde{\Sigma}_{ind} & 0 \\ 0 & I_{rc} \end{bmatrix} \tag{10}$$

and $\widetilde{\Sigma}_{ind}$ is the matrix Σ_{ind} in Lemma 6.3, except we use our estimates $\widetilde{\pi}^{(1)}, \widetilde{\pi}^{(2)}$, or $\widetilde{\mathbf{p}}$ whenever we need to use the actual (unknown) parameters.

Thus, if we are given a differentially private version of a contingency table where each cell has added independent Gaussian noise with variance σ^2, we calculate $\widetilde{Q}^2_{Gauss}$ and compare it to the threshold $\widetilde{\tau}^\alpha$ where

$$\Pr\left[\sum_{i=1}^{rc} \widetilde{\lambda}_i \chi^{2,i}_1 \geq \widetilde{\tau}^\alpha\right] = \alpha \tag{11}$$

with $\{\widetilde{\lambda}_i\}$ being the eigenvalues of $\widetilde{B}^T \widetilde{A} \widetilde{B}$ with rank $\nu = rc + (r-1)(c-1)$ matrix $\widetilde{B} \in \mathbb{R}^{2rc, \nu}$ where $\widetilde{B}\widetilde{B}^T = \widetilde{\Sigma}'_{ind}$. Our new independence test `PrivIndep` is given in Algorithm 7, where `2MLE` estimates $\pi^{(i)}$ for $i = 1, 2$ and `PrivIndep` Fails to Reject if `2MLE` returns NULL.

7. Significance Results

We now show how each of our tests perform on simulated data when H_0 holds in goodness of fit and independence testing. We fix our desired significance $1 - \alpha = 0.95$ and privacy level $(\epsilon, \delta) = (0.1, 10^{-6})$ in all of our tests.

By Theorem 5.3, we know that $\texttt{MCGOF}_{\mathcal{D}}$ will have significance at least $1 - \alpha$. We then turn to our test `PrivGOF` to compute the proportion of trials that failed to reject $H_0 : \mathbf{p} = \mathbf{p}^0$ when it holds. In Figure 1 we give several different null hypotheses $\mathbf{p}^0$ and sample sizes n to show that `PrivGOF` achieves near 0.95 significance in all our tested cases. We also compare our results with how the original test `GOF` would perform if used on the private counts with either Laplace and Gaussian noise.

We then turn to independence testing for 2×2 contingency tables using both $\texttt{MCIndep}_{\mathcal{D}}$ and `PrivIndep`. Note that our methods do apply to arbitrary $k \times \ell$ tables and run in time $\text{poly}(k, \ell, \log(n))$ plus the time for the iterative Imhof method to find the critical values. In Figure 2 we compute the empirical significance of both of our tests and compare it to how `Indep` performs on the nonprivate data. For $\texttt{MCIndep}_{\mathcal{D}}$ and `PrivIndep` we sample 1,000 trials for various parameters $\pi^{(1)}, \pi^{(2)}$, and n that could have generated the contingency tables. We set the number of samples $k = 50$ in $\texttt{MCIndep}_{\mathcal{D}}$ regardless of the noise we added and when we use Laplace noise, we set $\gamma = 0.01$ as the parameter in `2MLE`.

We also plot the critical values of our various tests in Figure 3. For both `PrivGOF` and `PrivIndep` we used the package in R "CompQuadForm" that has various methods for finding estimates to the tail probabilities for quadratic forms of normals, of which we used the "imhof" method (Imhof, 1961) to approximate the threshold for each test. Note that in $\texttt{MCIndep}_{\mathcal{D}}$ and `PrivIndep` each trial has a different threshold, so we give the average over all trials.

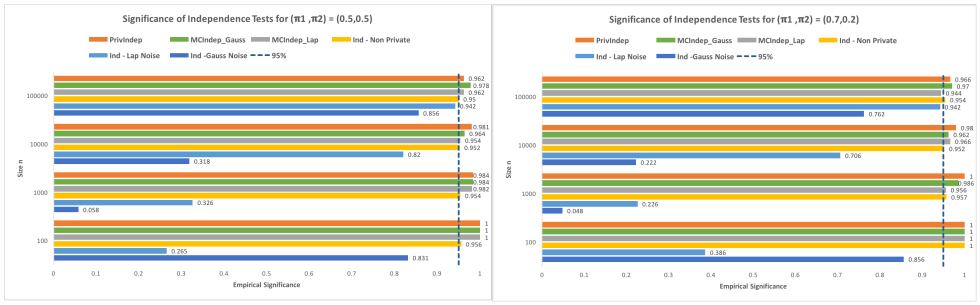

Figure 2. Significance of `Indep` when used on a contingency table with added Laplace or Gaussian noise compared to $\texttt{MCIndep}_{\mathcal{D}}$ for both Laplace and Gaussian noise and `PrivIndep` in 1,000 trials with $(\epsilon, \delta) = (0.1, 10^{-6})$ and $\alpha = 0.05$.

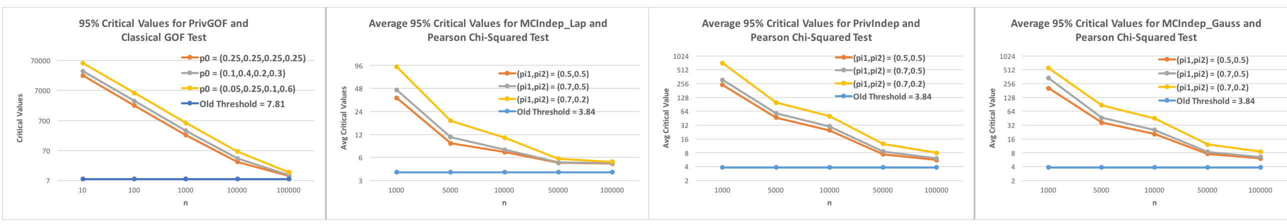

Figure 3. Comparison of the (average) critical values for all of our tests with $\alpha = 0.05$ and $(\epsilon, \delta) = (0.1, 10^{-6})$.

8. Power Results

We now want to show that our tests correctly reject H_0 when it is false, fixing parameters $\alpha = 0.05$ and $(\epsilon, \delta) = (0.1, 10^{-6})$. For our two goodness of fit tests, $\texttt{MCGOF}_{\mathcal{D}}$ (with $k = 100$) and `PrivGOF` we test whether the multinomial data came from $\mathbf{p}^0 = (1/4, 1/4, 1/4, 1/4)$ when it was actually sampled from $\mathbf{p}^1 = \mathbf{p}^0 + 0.01 \cdot (1, -1, 1, -1)$. We compare each of our tests with the classical `Indep` test that uses the unaltered data in Figure 4. We then find the proportion of 1,000 trials that each of our tests rejected $H_0 : \mathbf{p} = \mathbf{p}^0$ for various n. Note that `Indep` has difficulty distinguishing $\mathbf{p}^0$ and $\mathbf{p}^1$ for reasonable sample sizes.

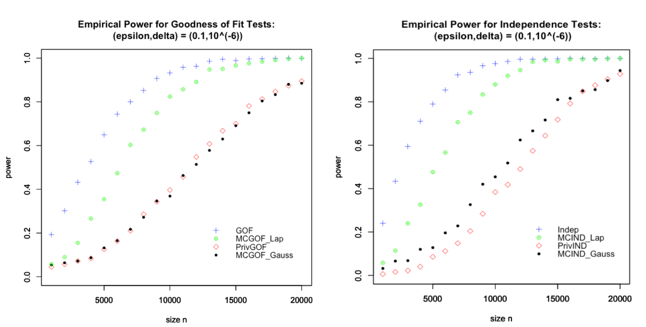

Figure 4. Power Plots of $\texttt{MCGOF}_{\mathcal{D}}$ and `PrivGOF`, as well as our independence tests $\texttt{MCIndep}_{\mathcal{D}}$ and `PrivIndep` compared with the classical tests, with $(\epsilon, \delta) = (0.1, 10^{-6})$.

We then turn to independence testing for 2×2 tables with our two differentially private tests $\texttt{MCIndep}_{\mathcal{D}}$ and `PrivIndep`. We fix the alternate H_1 : $\text{Cov}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}) = \Delta > 0$ so that $\mathbf{Y}^{(1)} \sim \text{Bern}(\pi^{(1)} = 1/2)$ and $\mathbf{Y}^{(2)} \sim \text{Bern}(\pi^{(2)} = 1/2)$ are not independent. We then sample contingency tables from a multinomial distribution with probability $\mathbf{p}^1 = (1/4, 1/4, 1/4, 1/4) + \Delta(1, -1, 1, -1)$ and various sizes n. We compute the proportion of 1,000 trials that $\texttt{MCIndep}_{\mathcal{D}}$ and `PrivIndep` rejected $H_0 : \mathbf{Y}^{(1)} \perp \mathbf{Y}^{(2)}$ and $\Delta = 0.01$ in Figure 4. For $\texttt{MCIndep}_{\mathcal{D}}$ we set the number of samples $k = 50$ and when we use Laplace noise, we set $\gamma = 0.01$ in `2MLE`.

9. Conclusion

We proposed new hypothesis tests based on a private version of the chi-squared statistic for goodness of fit and independence tests. For each test, we showed analytically or experimentally that we can achieve significance close to the target $1 - \alpha$ level similar to the nonprivate tests. We also showed that all the tests have a loss in power with respect to the non-private classical tests, with methods using Laplace noise outperforming those with Gaussian noise, due to the fact that the Gaussian noise has higher variance (to achieve the same level of privacy). Experimentally we show for 2×2 tables that with less than 3000 additional samples the tests with Laplace noise achieve the same power as the classical tests. Typically, one would expect differential privacy to require the sample size to blow up by a multiplicative $1/\epsilon$ factor. However, we see a better performance because the noise is dominated by the sampling error.

Acknowledgements

This work is part of the "Privacy Tools for Sharing Research Data" project based at Harvard, supported by NSF grant CNS-1237235 as well as a grant from the Sloan Foundation. This work has also been partially supported by the "PrivInfer - Programming Languages for Differential Privacy: Conditioning and Inference" EPSRC project EP/M022358/1 and by the University of Dundee, UK. Salil Vadhan's work was also supported by a Simons Investigator grant and was done in part while visiting the Department of Applied Mathematics and the Shing-Tung Yau Center at National Chiao-Tung University in Taiwan. We would like to thank the following people for helpful discussions: Dan Kifer, Aaron Roth, Aleksandra B. Slavković, Or Sheffet, Adam Smith, and a number of others involved with the Privacy Tools for Sharing Research Data project. A special thanks to Vishesh Karwa for helping us with statistics background and the suggestion to use a two step MLE procedure.

References

Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China*, pages 273–282, 2007.

Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland. Discrete multivariate analysis: Theory and practice, 1975.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on The Theory and Applications of Cryptographic Techniques*, EUROCRYPT'06, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06*, pages 265–284, 2006b.

Stephen E. Fienberg, Alessandro Rinaldo, and Xiaolin Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Proceedings of the 2010 International Conference on Privacy in Statistical Databases*, PSD'10, pages 187–199, Berlin, Heidelberg, 2010. Springer-Verlag.

Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Zhiwei Steven Wu. Dual query: Practical private query release for high dimensional data. In *International Conference on Machine Learning*, 2014.

Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. In *Conference on Neural Information Processing Systems (NIPS)*, pages 2348–2356, 2012.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8), 08 2008.

J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3-4):419–426, 1961.

Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1079–1087, New York, NY, USA, 2013. ACM.

Vishesh Karwa and Aleksandra Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Ann. Statist.*, 44(P1):87–112, 02 2016.

Vishesh Karwa and Aleksandra Slavković. Differentially private graphical degree sequences and synthetic graphs. In Josep Domingo-Ferrer and Ilenia Tinnirello, editors, *Privacy in Statistical Databases*, volume 7556 of *Lecture Notes in Computer Science*, pages 273–285. Springer Berlin Heidelberg, 2012.

Jaewoo Lee, Yue Wang, and Daniel Kifer. Maximum likelihood postprocessing for differential privacy under consistency constraints. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 635–644, New York, NY, USA, 2015. ACM.

Chao Li and Gerome Miklau. An adaptive mechanism for accurate query answering under differential privacy. *Proc. VLDB Endow.*, 5(6):514–525, February 2012.

Chao Li, Michael Hay, Vibhor Rastogi, Gerome Miklau, and Andrew McGregor. Optimizing linear counting queries under differential privacy. In *ACM SIGACT–SIGMOD–SIGART Symposium on Principles of Database Systems (PODS)*, pages 123–134, 2010.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Washington, DC, USA, 2007. IEEE Computer Society.

Sean Simmons and Bonnie Berger. Realizing privacy preserving genome-wide association studies. *Bioinformatics*, 32(9):1293–1300, 2016. doi: 10.1093/bioinformatics/btw009. URL `http://bioinformatics.oxfordjournals.org/content/32/9/1293.abstract`.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

M.F. Triola. *Essentials of Statistics*. Pearson Education, 2014. ISBN 9780321924636. URL `https://books.google.com/books?id=QZN-AgAAQBAJ`.

Caroline Uhler, Aleksandra Slavkovic, and Stephen E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013.

Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 138–143, Washington, DC, USA, 2009. IEEE Computer Society.

Y. Wang, J. Lee, and D. Kifer. Differentially Private Hypothesis Testing, Revisited. *ArXiv e-prints*, November 2015.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Fei Yu and Zhanglong Ji. Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to idash healthcare privacy protection challenge. *BMC Medical Informatics and Decision Making*, 14(1):1–8, 2014. ISSN 1472-6947. doi: 10.1186/1472-6947-14-S1-S3. URL `http://dx.doi.org/10.1186/1472-6947-14-S1-S3`.

Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavkovic, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.