A Differentially Private Kernel Two-Sample Test

Anant Raj*[1][0000−0002−0320−4733], Ho Chung Leon Law*[2][0000−0003−4101−5041] (✉), Dino Sejdinovic[2][0000−0001−5547−9213], and Mijung Park[1][0000−0003−1771−6104]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
{anant.raj, mijung.park}@tuebingen.mpg.de
[2] Department of Statistics, University of Oxford, United Kingdom
{ho.law, dino.sejdinovic}@stats.ox.ac.uk

Abstract. Kernel two-sample testing is a useful statistical tool in determining whether data samples arise from different distributions without imposing any parametric assumptions on those distributions. However, raw data samples can expose sensitive information about individuals who participate in scientific studies, which makes the current tests vulnerable to privacy breaches. Hence, we design a new framework for kernel two-sample testing conforming to differential privacy constraints, in order to guarantee the privacy of subjects in the data. Unlike existing differentially private parametric tests that simply add noise to data, kernel-based testing imposes a challenge due to a complex dependence of test statistics on the raw data, as these statistics correspond to estimators of distances between representations of probability measures in Hilbert spaces. Our approach considers finite dimensional approximations to those representations. As a result, a simple chi-squared test is obtained, where a test statistic depends on a mean and covariance of empirical differences between the samples, which we perturb for a privacy guarantee. We investigate the utility of our framework in two realistic settings and conclude that our method requires only a relatively modest increase in sample size to achieve a similar level of power to the non-private tests in both settings.

Keywords: Differential privacy · kernel two-sample test

1 Introduction

Several recent works suggest that it is possible to identify subjects that have participated in scientific studies based on publicly available aggregate statistics (cf. [21, 24] among many others). The *differential privacy* formalism [8] provides a way to quantify the amount of information on whether or not a single individual's data is included (or modified) in the data and also provides rigorous privacy guarantees in the presence of *arbitrary side information.*

An important tool in statistical inference is *two-sample testing*, in which samples from two probability distributions are compared in order to test the

[*] denote authors with equal contribution.

null hypothesis that the two underlying distributions are identical against the general alternative that they are different. In this paper, we focus on the non-parametric, *kernel-based* two-sample testing approach and investigate the utility of this framework in a differentially private setting. The kernel-based two-sample testing was introduced by Gretton et al [15, 16] who considers an estimator of maximum mean discrepancy (MMD) [3], the distance between embeddings of probability measures in a reproducing kernel Hilbert space (RKHS) (See [27] for a recent review), as a test statistic for the nonparametric two-sample problem.

Many existing differentially private testing methods are based on categorical data, i.e. counts [12, 13, 29], in which case a natural way to achieve privacy is simply adding noise to these counts. However, when we consider a more general input space $\mathcal{X}$ for testing, the amount of noise needed to privatise the data essentially becomes the order of diameter of the input space (explained in Appendix **??**). For spaces such as $\mathbb{R}^d$, the level of noise that needs to be added can destroy the utility of the data as well as that of the test.

Here we take an alternative approach and privatise only the quantities that are required for the test. In particular, for the two-sample testing, we only require the empirical kernel embedding $\frac{1}{N}\sum_i k(\mathbf{x}_i, \cdot)$ corresponding to a dataset, where $\mathbf{x}_i \in \mathcal{X}$ and k is some positive definite kernel. Now, as the kernel embedding lives in $\mathcal{H}_k$, a space of functions, a natural way to protect them is to add Gaussian process noise as suggested in [20] (discussed in Appendix **??**). Although sufficient for situations where the functions themselves are of interest, embeddings impaired by a Gaussian process does not lie in the same RKHS [31], and hence one cannot estimate the RKHS distances between such noisy embeddings. Alternatively, one could consider adding noise to an estimator of MMD [16]. However, asymptotic null distributions of these estimators are data dependent and the test thresholds are typically computed by permutation testing or by eigendecomposing centred kernel matrices of the data [17]. In this case neither of these approaches is available in a differentially private setting as they both require further access to data.

Contribution In this paper, we build a differentially private two-sample testing framework, by considering *analytic representations* of probability measures [6, 23], aimed at large scale testing scenarios. Through this formulation, we are able to obtain a test statistic that is based on means and covariances of feature vectors of the data. This suggests that privatisation of summary statistics of the data is sufficient to make the testing differential private, implying a reduction of level of noise needed versus adding to the data directly (as summary statistics are less sensitive to individual changes). Further, we show that while the asymptotic distribution under the null hypothesis of the test statistic does not depend on the data, unlike the non-private case, using the asymptotic null distribution to compute p-values can lead to grossly miscalibrated Type I control. Hence, we propose a remedy for this problem, and give approximations of the finite-sample null distributions, yielding good Type I error control and power-privacy tradeoffs experimentally in Sec. 6.

Related work To the best of our knowledge, this paper is the *first* to propose a two sample test in a differential private setting. Although, there are various differentially private hypothesis test in the literature, most of these revolve around categorical data [12, 13, 29] on chi-squared tests. This is very different to our work, which considers the problem of identifying whether two distributions are equal to each other. Further, while there are several works that connect kernel methods with differential privacy, including [2, 20, 22], none of these attempts to make the kernel-based two sample testing procedure private. It is also important to emphasise that in a hypothesis testing, it is not sufficient to make the test statistic differentially private, as one has to carefully construct the distribution under the null hypothesis in a differential private manner, taking into account the level of noise added.

Motivation and setting We now present the two privacy scenarios that we consider and motivate their usage. In the first scenario, we assume there is a trusted curator and also an untrusted tester, in which we want to protect data from. In this setting, the trusted curator has access to the two datasets and computes the mean and covariance of the empirical differences between the feature vectors. The curator can protect the data in two different ways: (1) perturb mean and covariance separately and release them; or (2) compute the statistic without perturbations and add noise to it directly. The tester can now take these perturbed quantities and performs the test at a desired significance level. Here, we separate the entities of tester and curator, as sometimes a decision whether to reject or not is of interest, for example one can imagine that the tester may require the test-statistic/p-values for multiple hypothesis testing corrections. In the second scenario, we assume that there are two data-owners, each having one dataset each, and a tester. In this case, as no party trust each other, each data-owner has to perturb their own mean and covariance of the feature vectors and release them to the tester. Under these two settings, we will exploit various differentially private mechanisms and empirically study the utility of the proposed framework. We start by providing a brief background on kernels, differential privacy and the two privacy settings we consider in this paper in Sec. 2. We derive essential tools for the proposed test in Sec. 3 and Sec. 4, and describe approximations to finite-sample null distributions in Sec. 5. Finally, we illustrate the effectiveness of our algorithm in Sec. 6.

2 Background

2.1 Mean embedding and smooth characteristic function tests

First introduced by [6] and then extended and further analyzed by [23], these two tests are the state-of-the-art kernel based testing approaches applicable to large datasets. Here, we will focus on the approach by [23], and in particular on the mean embedding (ME) and on a characterisation approach based on the smooth characteristic function (SCF). Assume that we observe samples $\{\mathbf{x}_i\}_{i=1}^n \sim P$

and $\{\mathbf{y}_i\}_{i=1}^n \sim Q$, where P and Q are some probability measures on $\mathbb{R}^D$. Now our goal is to test the null hypothesis $\mathbf{H}_0 : P = Q$ against all alternatives. Both ME and SCF tests consider finite-dimensional feature representations of the empirical measures P_n and Q_n corresponding to the samples $\{\mathbf{x}_i\}_{i=1}^n \sim P$ and $\{\mathbf{y}_i\}_{i=1}^n \sim Q$ respectively. The ME test considers feature representation given by $\boldsymbol{\phi}_{P_n} = \frac{1}{n}\sum_{i=1}^n [k(\mathbf{x}_i, T_1), \cdots, k(\mathbf{x}_i, T_J)] \in \mathbb{R}^J$, for a given set of test locations $\{T_j\}_{j=1}^J$, i.e. it evaluates the kernel mean embedding $\frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \cdot)$ of P_n at those locations. We write $\mathbf{w}_n = \boldsymbol{\phi}_{P_n} - \boldsymbol{\phi}_{Q_n}$ to be the difference of the feature vectors of the empirical measures P_n and Q_n. If we write

$$\mathbf{z}_i = \Big[k(\mathbf{x}_i, T_1) - k(\mathbf{y}_i, T_1), \cdots, k(\mathbf{x}_i, T_J) - k(\mathbf{y}_i, T_J)\Big],$$

then $\mathbf{w}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$. We also define the empirical covariance matrix $\boldsymbol{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{z}_i - \mathbf{w}_n)(\mathbf{z}_i - \mathbf{w}_n)^\top$. The final statistic is given by

$$s_n = n\ \mathbf{w}_n^\top (\varSigma_n + \gamma_n I)^{-1} \mathbf{w}_n, \tag{1}$$

where, as [23] suggest, a regularization term $\gamma_n I$ is added onto the empirical covariance matrix for numerical stability. This regularization parameter will also play an important role in analyzing sensitivity of this statistic in a differentially private setting. Following [23, Theorem 2], one should take $\gamma_n \to 0$ as $n \to \infty$, and in particular, γ_n should decrease at a rate of $\mathcal{O}(n^{-1/4})$. The SCF setting uses the statistic of the same form, but considers features based on empirical characteristic functions [28]. Thus, it suffices to set $\mathbf{z}_i \in \mathbb{R}^J$ to

$$\begin{aligned}\mathbf{z}_i = \Big[&g(\mathbf{x}_i)\cos(\mathbf{x}_i^\top T_j) - g(\mathbf{y}_i)\cos(\mathbf{y}_i^\top T_j),\\ &g(\mathbf{x}_i)\sin(\mathbf{x}_i^\top T_j) - g(\mathbf{y}_i)\sin(\mathbf{y}_i^\top T_j)\Big]_{j=1}^J,\end{aligned}$$

where $\{T_j\}_{j=1}^{J/2}$ is a given set of frequencies, and g is a given function which has an effect of smoothing the characteristic function estimates (cf. [6] for derivation). The test then proceeds in the same way as the ME version. For both cases, the distribution of the test statistic (1) under the null hypothesis $\mathbf{H}_0 : P = Q$ converges to a chi-squared distribution with J degrees of freedom. This follows from a central limit theorem argument whereby $\sqrt{n}\mathbf{w}_n$ converges in law to a zero-mean multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{z}\mathbf{z}^\top]$, while $\boldsymbol{\Sigma}_n + \gamma_n I \to \boldsymbol{\Sigma}$ in probability.

While [6] uses random distribution features (i.e. test locations/frequencies $\{T_j\}_j$ are sampled randomly from a predefined distribution), [23] selects test locations/frequencies $\{T_j\}_j$ which maximize the test power, yielding interpretable differences between the distributions under consideration. Throughout the paper, we assume that we use bounded kernels in the ME test (e.g. Gaussian and Laplace Kernel), in particular $k(\mathbf{x}, \mathbf{y}) \leq \kappa/2, \quad \forall \mathbf{x}, \mathbf{y}$, and that the weighting function in the SCF test is also bounded: $h(\mathbf{x}) \leq \kappa/2$ Hence, $||\mathbf{z}_i||_2 \leq \kappa\sqrt{J}$ in both cases, for any $i \in [1, n]$.

2.2 Differential privacy

Given an algorithm $\mathcal{M}$ and neighbouring datasets $\mathcal{D}$, $\mathcal{D}'$ differing by a single entry, the *privacy loss* of an outcome o is

$$L^{(o)} = \log \frac{Pr(\mathcal{M}_{(\mathcal{D})} = o)}{Pr(\mathcal{M}_{(\mathcal{D}')} = o)}. \tag{2}$$

The mechanism $\mathcal{M}$ is called ϵ-DP if and only if $|L^{(o)}| \leq \epsilon, \forall o, \mathcal{D}, \mathcal{D}'$. A weaker version of the above is (ϵ, δ)-DP, if and only if $|L^{(o)}| \leq \epsilon$, with probability at least $1 - \delta$. The definition states that a single individual's participation in the data do not change the output probabilities by much, hence this limits the amount of information that the algorithm reveals about any one individual.

A differentially private algorithm is designed by adding noise to the algorithms' outputs. Suppose a deterministic function $h : \mathcal{D} \mapsto \mathbb{R}^p$ computed on sensitive data $\mathcal{D}$ outputs a p-dimensional vector quantity. In order to make h private, we can add noise in function h, where the level of noise is calibrated to the *global sensitivity* GS_h [7], defined by the maximum difference in terms of L_2-norm $||h(\mathcal{D}) - h(\mathcal{D}')||_2$, for neighboring $\mathcal{D}$ and $\mathcal{D}'$ (i.e. differ by one data sample). In the case of Gaussian mechanism (Theorem 3.22 in [9]), the output is perturbed by

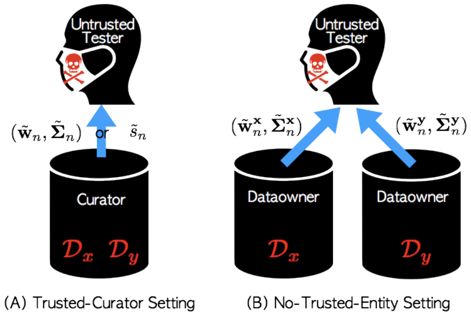
$$\tilde{h}(\mathcal{D}) = h(\mathcal{D}) + \mathcal{N}(0, GS_h^2 \sigma^2 \mathbf{I}_p)$$

The perturbed function $\tilde{h}(\mathcal{D})$ is then (ϵ, δ)-DP, where $\sigma \geq \sqrt{2 \log(1.25/\delta)}/\epsilon$, for $\epsilon \in (0, 1)$ (See the proof of Theorem 3.22 in [9] why σ has such a form). When constructing our tests, we will use two important properties of differential privacy. The composability theorem [7] tells us that the strength of privacy guarantee degrades with repeated use of DP-algorithms. In particular, when two differentially private subroutines are combined, where each one guarantees (ϵ_1, δ_1)-DP and (ϵ_2, δ_2)-DP respectively by adding independent noise, the parameters are simply composed by $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$. Furthermore, post-processing invariance [7] tells us that the composition of any arbitrary data-independent mapping with an (ϵ, δ)-DP algorithm is also (ϵ, δ)-DP. Here below in the next section, we discuss the two privacy settings which we are considering for our study in this paper.

2.3 Privacy settings

We consider the two different privacy settings as shown in Fig. 1:

(A) Trusted-curator (TC) setting There is a trusted entity called curator that handles datasets and outputs the private test statistic, either in terms of perturbed $\tilde{\mathbf{w}}_n$ and $\tilde{\mathbf{\Sigma}}_n$, or in terms of perturbed test statistic $\tilde{s}_n$. An untrusted tester performs a chi-square test given these quantities.


Untrusted Tester
$(\tilde{\mathbf{w}}_n, \tilde{\mathbf{\Sigma}}_n)$ or $\tilde{s}_n$
Curator
$\mathcal{D}_x$ $\mathcal{D}_y$
Untrusted Tester
$(\tilde{\mathbf{w}}_n^{\mathbf{x}}, \tilde{\mathbf{\Sigma}}_n^{\mathbf{x}})$
$(\tilde{\mathbf{w}}_n^{\mathbf{y}}, \tilde{\mathbf{\Sigma}}_n^{\mathbf{y}})$
Dataowner
$\mathcal{D}_x$
Dataowner
$\mathcal{D}_y$

(A) Trusted-Curator Setting (B) No-Trusted-Entity Setting

Fig. 1: Two privacy settings. **(A)** A trusted curator releases a private test statistic or private mean and covariance of empirical differences between the features. **(B)** Data owners release private feature means and covariances calculated from their samples. In both cases, an untrusted tester performs a test using the private quantities.

(B) No-trusted-entity (NTE) setting Each data owner outputs private mean and covariance of the feature vectors computed on their own dataset, meaning that the owner of dataset $\mathcal{D}_x$ outputs $\tilde{\mathbf{w}}_n^{\mathbf{x}}$ and $\tilde{\mathbf{\Sigma}}_n^{\mathbf{x}}$ and the owner of dataset $\mathcal{D}_y$ outputs $\tilde{\mathbf{w}}_n^{\mathbf{y}}$ and $\tilde{\mathbf{\Sigma}}_n^{\mathbf{y}}$. An untrusted tester performs a chi-squared test given these quantities.

It is worth noting that the NTE setting is different from the typical two-party model considered in the differential privacy literature. In the two-party model, it is typically assumed that Alice owns a dataset $\mathcal{D}_x$ and Bob owns a dataset $\mathcal{D}_y$, and they wish to compute some functionality $f(\mathcal{D}_x, \mathcal{D}_y)$ in a differentially private manner. In this case, the interest is to obtain a two-sided ϵ-differentially private protocol for f, i.e., each party's view of the protocol should be a differentially private function of the other party's input. For instance, the probability of Alice's views conditioned on $\mathcal{D}_y$ and $\mathcal{D}_{y'}$ should be e^ϵ multiplicatively close to each other, where $\mathcal{D}_y$ and $\mathcal{D}_{y'}$ are adjacent datasets [14, 26]. On the other hand, in our NTE setting, we are not considering a joint function that takes two datasets. Rather, we consider a function (statistics) which each data-owner computes given their own dataset independent of the dataset that the other party has. We would like to analyze how the performance of the test run by an untrusted third party using those separately released DP statistics from each party degrades with the level of DP guarantees.

3 Trusted-curator setting

In this setting, a trusted curator releases either a private test statistic or private mean and covariance which a tester can use to perform a chi-square test. Given a total privacy budget (ϵ, δ), when we perturb mean and covariance separately, we spend (ϵ_1, δ_1) for mean perturbation and (ϵ_2, δ_2) for covariance perturbation, such that $\epsilon = \epsilon_1 + \epsilon_2$ and $\delta = \delta_1 + \delta_2$.

3.1 Perturbing mean and covariance

Mean perturbation We obtain a private mean by adding Gaussian noise based on the analytic Gaussian mechanism recently proposed in [1]. The main reason for using this Gaussian mechanism over the original [9] is that it provides a DP guarantee with smaller noise.

For $\mathbf{w}_n : \mathcal{D} \to \mathbb{R}^J$ that has the global L2-sensitivity $GS_2(\mathbf{w}_n)$, the analytic Gaussian mechanism produces $\tilde{\mathbf{w}}_n(\mathcal{D}) = \mathbf{w}_n(\mathcal{D}) + \mathbf{n}, \quad$ where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_J, \sigma_{\mathbf{n}}^2 \mathbf{I}_{J \times J})$. Then $\tilde{\mathbf{w}}_n(\mathcal{D})$ is (ϵ_1, δ_1)-differentially private mean vector if $\sigma_{\mathbf{n}}$ follows the regime in Theorem 9 of [9], here implicitly $\sigma_{\mathbf{n}}$ depends on $GS_2(\mathbf{w}_n), \epsilon_1$ and δ_1. Assuming an entry difference between two parts of datasets $\mathcal{D} = (\mathcal{D}_x, \mathcal{D}_y)$ and $\mathcal{D}' = (\mathcal{D}'_x, \mathcal{D}'_y)$ the global sensitivity is simply

$$\begin{aligned} GS_2(\mathbf{w}_n) &= \max_{\mathcal{D}, \mathcal{D}'} \|\mathbf{w}_n(\mathcal{D}) - \mathbf{w}_n(\mathcal{D}')\|_2 \\ &= \max_{\mathbf{z}_i, \mathbf{z}'_i} \tfrac{1}{n} \|\mathbf{z}_i - \mathbf{z}'_i\|_2 \leq \tfrac{\kappa \sqrt{J}}{n}. \end{aligned} \quad (3)$$

where $\mathbf{z}_i$ is as the corresponding feature maps defined in Sec. 2.

Covariance perturbation To obtain a private covariance, we consider [10] which utilises Gaussian noise. Here since the covariance matrix is given by $\mathbf{\Sigma}_n = \mathbf{\Lambda} - \frac{n}{n-1} \mathbf{w}_n \mathbf{w}_n^\top$, where $\mathbf{\Lambda} = \frac{1}{n-1} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top$, we can simply privatize the covariance by simply perturbing the 2nd-moment matrix $\mathbf{\Lambda}$ and using the private mean $\tilde{\mathbf{w}}_n$, i.e., $\tilde{\mathbf{\Sigma}}_n = \tilde{\mathbf{\Lambda}} - \frac{n}{n-1} \tilde{\mathbf{w}}_n \tilde{\mathbf{w}}_n^\top$. To construct the 2nd-moment matrix $\tilde{\mathbf{\Lambda}}$ that is (ϵ_2, δ_2)-differentially private, we use $\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda} + \mathbf{\Psi}$, where $\mathbf{\Psi}$ is obtained as follows:

1. Sample from $\boldsymbol{\eta} \sim \mathcal{N}(0, \beta^2 \mathbf{I}_{J(J+1)/2})$, where β is a function of global sensitivity $GS(\mathbf{\Lambda}), \epsilon_2, \delta_2$, outlined in Theorem **??** in the appendix.
2. Construct an upper triangular matrix (including diagonal) with entries from $\boldsymbol{\eta}$.
3. Copy the upper part to the lower part so that resulting matrix $\mathbf{\Psi}$ becomes symmetric.

Now using the composability theorem [7] gives us that $\tilde{\mathbf{\Sigma}}_n$ is (ϵ, δ)-differentially private.

3.2 Perturbing test statistic

The trusted-curator can also release a differentially private statistic, to do this we use the analytic Gaussian mechanism as before, perturbing the statistic by adding Gaussian noise. To use the mechanism, we need to calculate the global sensitivity needed of the test statistic $s_n = \mathbf{w}_n^\top (\mathbf{\Sigma}_n + \gamma_n I)^{-1} \mathbf{w}_n$, which we provide in this Theorem (proof in Appendix **??**):

Theorem 1. *Given the definitions of* $\mathbf{w}_n$ *and* $\mathbf{\Lambda}_n$*, and the L2-norm bound on* $\mathbf{z}_i$*'s, the global sensitivity* $GS_2(s_n)$ *of the test statistic* s_n *is* $\frac{4\kappa^2 J\sqrt{J}}{n\gamma_n}\left(1+\frac{\kappa^2 J}{n-1}\right)$*, where* γ_n *is a regularization parameter, which we set to be smaller than the smallest eigenvalue of* $\mathbf{\Lambda}$*.*

4 No-trusted-entity setting

In this setting, the two samples $\{\mathbf{x}_i\}_{i=1}^{n_\mathbf{x}} \sim P$ and $\{\mathbf{y}_j\}_{j=1}^{n_\mathbf{y}} \sim Q$ reside with different data owners each of which wish to protect their samples in a differentially private manner. Note that in this context we allow the size of each sample to be different. The data owners first need to agree on the given kernel k as well as on the test locations $\{T_j\}_{j=1}^{J}$. We denote now $\mathbf{z}_i^\mathbf{x} = \Big[k(\mathbf{x}_i, T_1), \cdots, k(\mathbf{x}_i, T_J)\Big]^\top$ in the case of the ME test or $\mathbf{z}_i^\mathbf{x} = \Big[h(\mathbf{x}_i)\cos(\mathbf{x}_i^\top T_j), h(\mathbf{x}_i)\sin(\mathbf{x}_i^\top T_j)\Big]_{j=1}^{J}$ in the case of the SCF test. Also, we denote

$$\mathbf{w}_{n_\mathbf{x}}^\mathbf{x} = \frac{1}{n_\mathbf{x}}\sum_{i=1}^{n}\mathbf{z}_i^\mathbf{x} \qquad \mathbf{\Sigma}_{n_\mathbf{x}}^\mathbf{x} = \frac{1}{n_\mathbf{x}-1}\sum_{i=1}^{n_\mathbf{x}}(\mathbf{z}_i^\mathbf{x} - \mathbf{w}_{n_\mathbf{x}}^\mathbf{x})(\mathbf{z}_i^\mathbf{x} - \mathbf{w}_{n_\mathbf{x}}^\mathbf{x})^\top$$

and similarly for the sample $\{\mathbf{y}_j\}_{j=1}^{n_\mathbf{y}} \sim Q$. The respective means and covariances $\mathbf{w}_{n_\mathbf{x}}^\mathbf{x}$, $\mathbf{\Sigma}_{n_\mathbf{x}}^\mathbf{x}$ and $\mathbf{w}_{n_\mathbf{y}}^\mathbf{y}$, $\mathbf{\Sigma}_{n_\mathbf{y}}^\mathbf{y}$ are computed by their data owners, which then impair them independently with noise according to the sensitivity analysis described in Section 3.1. As a result we obtain differentially private means and covariances $\tilde{\mathbf{w}}_{n_\mathbf{x}}^\mathbf{x}$, $\tilde{\mathbf{\Sigma}}_{n_\mathbf{x}}^\mathbf{x}$ and $\tilde{\mathbf{w}}_{n_\mathbf{y}}^\mathbf{y}$, $\tilde{\mathbf{\Sigma}}_{n_\mathbf{y}}^\mathbf{y}$ at their respective users. All these quantities are then released to the tester whose role is to compute the test statistic and the corresponding p-value. In particular, the tester uses the statistic given by

$$\tilde{s}_{n_\mathbf{x},n_\mathbf{y}} = \frac{n_\mathbf{x} n_\mathbf{y}}{n_\mathbf{x}+n_\mathbf{y}}(\tilde{\mathbf{w}}_{n_\mathbf{x}}^\mathbf{x} - \tilde{\mathbf{w}}_{n_\mathbf{y}}^\mathbf{y})^\top(\tilde{\mathbf{\Sigma}}_{n_\mathbf{x},n_\mathbf{y}} + \gamma_n I)^{-1}(\tilde{\mathbf{w}}_{n_\mathbf{x}}^\mathbf{x} - \tilde{\mathbf{w}}_{n_\mathbf{y}}^\mathbf{y}),$$

where $\tilde{\mathbf{\Sigma}}_{n_\mathbf{x},n_\mathbf{y}} = \frac{(n_\mathbf{x}-1)\tilde{\mathbf{\Sigma}}_{n_\mathbf{x}}^\mathbf{x} + (n_\mathbf{y}-1)\tilde{\mathbf{\Sigma}}_{n_\mathbf{y}}^\mathbf{y}}{n_\mathbf{x}+n_\mathbf{y}-2}$ is the pooled covariance estimate.

5 Analysis of null distributions

In the previous sections, we discussed necessary tools to make the kernel two sample tests private in two different settings by considering sensitivity analysis of quantities of interest.[4]. In this section, we consider the distributions of the test statistics under the null hypothesis $P = Q$ for each of the two settings.

5.1 Trusted-curator setting: perturbed mean and covariance

In this scheme, noise is added both to the mean vector $\mathbf{w}_n$ and to the covariance matrix $\mathbf{\Sigma}_n$ (by dividing the privacy budget between these two quantities). Let

[4] See Appendix **??** and **??** for other possible approaches.

us denote the perturbed mean by $\tilde{\mathbf{w}}_n$ and perturbed covariance with $\tilde{\mathbf{\Sigma}}_n$. The noisy version of the test statistic $\tilde{s}_n$ is then given by

$$\tilde{s}_n = n\tilde{\mathbf{w}}_n^\top \, (\tilde{\mathbf{\Sigma}}_n + \gamma_n I)^{-1}\tilde{\mathbf{w}}_n \tag{4}$$

where γ_n is a regularization parameter just like in the non-private statistic (1). We show below that the asymptotic null distribution (as sample size $n \to \infty$) of this private test statistic is in fact identical to that of the non-private test statistic. Intuitively, this is to be expected: as the number of samples increases, the contribution to the aggregate statistics of any individual observation diminishes, and the variance of the added noise goes to zero.

Theorem 2. *Assuming the Gaussian noise for $\tilde{\mathbf{w}}_n$ with the sensitivity bound in* (3) *and the perturbation mechanism introduced in Section 3.1 for $\tilde{\mathbf{\Sigma}}_n$, $\tilde{s}_n$ and s_n converge to the same limit in distribution, as $n \to \infty$. Also, under the alternate, $\tilde{s}_n = s_n(1+\epsilon)$ and ϵ goes down as $\mathcal{O}(n^{-1+\gamma})$.*

Proof is provided in Appendix **??**. We here assume that under the alternate the relation $\mathbf{w}_n^\top \mathbf{\Sigma}^{-1}\mathbf{w}_n \geq \mathcal{O}(n^{-\gamma})$ for $\gamma < 1$ holds. Based on the Theorem, it is tempting to ignore the additive noise and rely on the asymptotic null distribution. However, as demonstrated in Sec. 6, such tests have a *grossly mis-calibrated Type I error*, hence we propose a non-asymptotic regime in order to improve approximations of the null distribution when computing the test threshold.

In particular, let's start by recalling that we previously relied on $\sqrt{n}\mathbf{w}_n$ converging to a zero-mean multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma})$, with $\mathbf{\Sigma} = \mathbb{E}[\mathbf{z}\mathbf{z}^\top]$ [6]. In the private setting, we will also approximate the distribution of $\sqrt{n}\tilde{\mathbf{w}}_n$ with a multivariate normal, but consider explicit non-asymptotic covariances which appear in the test statistic. Namely, the covariance of $\sqrt{n}\tilde{\mathbf{w}}_n$ is $\mathbf{\Sigma} + n\sigma_{\mathbf{n}}^2 I$ and its mean is 0, so we will approximate its distribution by $\mathcal{N}(0, \mathbf{\Sigma} + n\sigma_{\mathbf{n}}^2 I)$. The test statistic can be understood as a squared norm of the vector $\sqrt{n}\left(\tilde{\mathbf{\Sigma}}_n + \gamma_n I\right)^{-1/2}\tilde{\mathbf{w}}_n$. Under the normal approximation to $\sqrt{n}\tilde{\mathbf{w}}_n$ and by treating $\tilde{\mathbf{\Sigma}}_n$ as fixed (note that this is a quantity released to the tester), $\sqrt{n}\left(\tilde{\mathbf{\Sigma}}_n + \gamma_n I\right)^{-1/2}\tilde{\mathbf{w}}_n$ is another multivariate normal, i.e. $\mathcal{N}(0, \mathbf{C})$, where

$$\mathbf{C} = (\tilde{\mathbf{\Sigma}}_n + \gamma_n I)^{-1/2}(\mathbf{\Sigma} + n\sigma_{\mathbf{n}}^2 I)(\tilde{\mathbf{\Sigma}}_n + \gamma_n I)^{-1/2}.$$

The overall statistic thus follows a distribution given by a weighted sum $\sum_{j=1}^{J}\lambda_j\chi_j^2$ of independent chi-squared distributed random variables, with the weights λ_j given by the eigenvalues of $\mathbf{C}$. Note that this approximation to the null distribution depends on a *non-private* true covariance $\mathbf{\Sigma}$. While that is clearly not available to the tester, we propose to simply replace this quantity with the privatized empirical covariance, i.e. $\tilde{\mathbf{\Sigma}}_n$, so that the tester approximates the null distribution with $\sum_{j=1}^{J}\tilde{\lambda}_j\chi_j^2$, where $\tilde{\lambda}_j$ are the eigenvalues of

$$\tilde{\mathbf{C}} = (\tilde{\mathbf{\Sigma}}_n + \gamma_n I)^{-1}(\tilde{\mathbf{\Sigma}}_n + n\sigma_{\mathbf{n}}^2 I),$$

i.e. $\tilde{\lambda}_j = \frac{\tau_j + n\sigma_{\mathbf{n}}^2}{\tau_j + \gamma_n}$, where $\{\tau_j\}$ are the eigenvalues of $\tilde{\mathbf{\Sigma}}_n$ (note that $\tilde{\lambda}_j \to 1$ as $n \to \infty$ recovering back the asymptotic null). This approach, while a heuristic, gives a correct Type I control, good power performance and is differentially private. This is unlike the approach which relies on the asymptotic null distribution and ignores the presence of privatizing noise. We demonstrate this empirically in Sec. 6.

5.2 Trusted-curator setting: perturbed test statistic

In this section, we will consider how directly perturbing the test statistic impacts the null distribution. To achieve private test statistics, we showed that we can simply add Gaussian noise [5] using the Gaussian mechanism, described in Section 3.2. Similarly to Theorem 2, we have a similar theorem below, which says that the perturbed statistic then has the same asymptotic null distribution as the original statistic.

Theorem 3. *Using the noise variance $\sigma_\eta^2(\epsilon, \delta, n)$ defined by the upper bound in Theorem 1, $\tilde{s}_n$ and s_n converge to the same limit in distribution, as $n \to \infty$. More specifically, the error between s_n and $\tilde{s}_n$ goes down approximately at the rate of $\mathcal{O}(n^{-1/2})$.*

The proof follows immediately from $\sigma_\eta(\epsilon, \delta, n) \to 0$, as $n \to \infty$. The specific order of convergence directly comes after applying the Chebysev inequality since the variance σ^2 is of the order of $\mathcal{O}(n^{-1})$. As in the case of perturbed mean and covariance, we consider approximating the null distribution with the sum of the chi-squared with J degrees of freedom and a normal $\mathcal{N}(0, \sigma_\eta^2(\epsilon, \delta, n))$, i.e., the distribution of the true statistic is approximated with its asymptotic version, whereas we use exact non-asymptotic distribution of the added noise. The test threshold can then easily be computed by a Monte Carlo test which repeatedly simulates the sum of these two random variables. It is important to note that since $\sigma_\eta^2(\epsilon, \delta, n)$ is *independent of the data* (Appendix **??**), an untrusted tester can simulate the approximate null distribution without compromising privacy.

5.3 No-trusted-entity setting

Similarly as in section 5.1, as $n_{\mathbf{x}}, n_{\mathbf{y}} \to \infty$ such that $n_{\mathbf{x}}/n_{\mathbf{y}} \to \rho \in (0, 1)$, asymptotic null distribution of this test statistic remains unchanged as in the non-private setting, i.e. it is the chi-squared distribution with J degrees of freedom. However, by again considering the non-asymptotic case and applying a chi-squared approximation, we get improved power and type I control. In particular, the test statistic is close to a weighted sum $\sum_{j=1}^{J} \lambda_j \chi_j^2$ of independent

[5] While this may produce negative privatized test statistics, the test threshold is appropriately adjusted for this. See Appendix **??** and **??** for alternative approaches for privatizing the test statistic.

chi-square distributed random variables, with the weights λ_j given by the eigenvalues of

$$\mathbf{C} = \frac{n_{\mathbf{x}} n_{\mathbf{y}}}{n_{\mathbf{x}} + n_{\mathbf{y}}} (\tilde{\mathbf{\Sigma}}_{n_{\mathbf{x}},n_{\mathbf{y}}} + \gamma_n I)^{-1/2} (\mathbf{\Sigma}^{\mathbf{x}}/n_{\mathbf{x}} + \mathbf{\Sigma}^{\mathbf{y}}/n_{\mathbf{y}} + (\sigma^2_{n_{\mathbf{x}}} + \sigma^2_{n_{\mathbf{y}}})I)(\tilde{\mathbf{\Sigma}}_{n_{\mathbf{x}},n_{\mathbf{y}}} + \gamma_n I)^{-1/2}$$

where $\mathbf{\Sigma}^{\mathbf{x}}$ and $\mathbf{\Sigma}^{\mathbf{y}}$ are the true covariances within each of the samples, $\sigma^2_{n_{\mathbf{x}}}$ and $\sigma^2_{n_{\mathbf{y}}}$ are the variances of the noise added to the mean vectors $\mathbf{w}_{n_{\mathbf{x}}}$ and $\mathbf{w}_{n_{\mathbf{y}}}$, respectively. While $\mathbf{\Sigma}^{\mathbf{x}}$ and $\mathbf{\Sigma}^{\mathbf{y}}$ are clearly not available to the tester, the tester can replace them with their privatized empirical versions $\tilde{\mathbf{\Sigma}}^{\mathbf{x}}_{n_{\mathbf{x}}}$ and $\tilde{\mathbf{\Sigma}}^{\mathbf{y}}_{n_{\mathbf{y}}}$ and compute eigenvalues $\tilde{\lambda}_j$ of

$$\tilde{\mathbf{C}} = \frac{n_{\mathbf{x}} n_{\mathbf{y}}}{n_{\mathbf{x}} + n_{\mathbf{y}}} (\tilde{\mathbf{\Sigma}}_{n_{\mathbf{x}},n_{\mathbf{y}}} + \gamma_n I)^{-1/2} (\tilde{\mathbf{\Sigma}}^{\mathbf{x}}_{n_{\mathbf{x}}}/n_{\mathbf{x}} + \tilde{\mathbf{\Sigma}}^{\mathbf{y}}_{n_{\mathbf{y}}}/n_{\mathbf{y}} + (\sigma^2_{n_{\mathbf{x}}} + \sigma^2_{n_{\mathbf{y}}})I)(\tilde{\mathbf{\Sigma}}_{n_{\mathbf{x}},n_{\mathbf{y}}} + \gamma_n I)^{-1/2}$$

Note that this is a differentially private quantity. Similarly as in the trusted-curator setting, we demonstrate that this corrected approximation to the null distribution leads to significant improvements in power and Type I control.

6 Experiments

Here we demonstrate the effectiveness of our private kernel two-sample test[6] on both synthetic and real problems, for testing $\mathbf{H_0} : P = Q$. The total sample size is denoted by N and the number of test set samples by n. We set the significance level to $\alpha = 0.01$. Unless specified otherwise use the isotropic Gaussian kernel with a bandwidth θ and fix the number of test locations to $J = 5$. Under the trusted-curator (TC) setting, we use 20% of the samples N as an independent training set to optimize the test locations and θ using gradient descent as in [23]. Under the no-trusted-entity (NTE) setting, we randomly sample J locations and calculate the median heuristic bandwidth [18].

For all our experiments, we average them over 500 runs, where each run repeats the simulation or randomly samples without replacement from the data set. We then report the empirical estimate of $\mathcal{P}(\tilde{s}_n > T_\alpha)$, computed by proportion of times the statistic $\tilde{s}_n$ is greater than the T_α, where T_α is the test threshold provided by the corresponding approximation to the null distribution. Regularization parameter $\gamma = \gamma_n$ is fixed to 0.001 for TC under perturbed test statistics (TCS), with the choice of this investigated in Figure **??**. In the trusted-curator mean covariance perturbation (TCMC) and NTE, given the privacy budget of (ϵ, δ), we use $(0.5\epsilon, 0.5\delta)$ to perturb the mean and covariance seperately. We compare these to its non-private counterpart ME and SCF, as there are no available appropriate baseline to compare against. We will also demonstrate the importance of using an approximated finite-null distribution versus the asymptotic null distribution. More details and experiments can be found in Appendix **??**.

[6] Code is available at https://github.com/hcllaw/private_tst

6.1 Synthetic data

We demonstrate our tests on 4 separate synthetic problems, namely, Same Gaussian (SG), Gaussian mean difference (GMD), Gaussian variance difference (GVD) and Blobs, with the specifications of P and Q summarized in Table. 2. The same experimental setup was used in [23]. For the Blobs dataset, we use the SCF approach as the baseline, and also the basis for our algorithms, since [6, 23] showed that SCF outperforms the ME test here.

Data	**P**	**Q**
SG	$\mathcal{N}(0, I_{50})$	$\mathcal{N}(0, I_{50})$
GMD	$\mathcal{N}(0, I_{100})$	$\mathcal{N}((1, 0, \ldots, 0)^\top, I_{100})$
GVD	$\mathcal{N}(0, I_{50})$	$\mathcal{N}(0, diag(2, 1, \ldots, 1))$

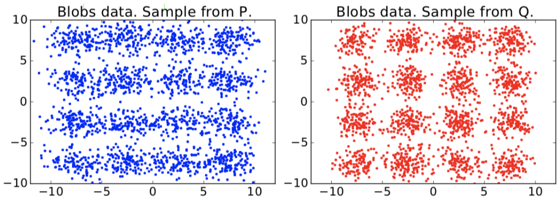


Fig. 2: Blobs data sampled from **P** on the left and from **Q** in the right.

Varying privacy level ϵ We now fix the test sample size n to be 10 000, and vary ϵ between 0 and 5 with a fixed $\delta = 1e - 5$. The results are shown in the top row of Figure 4. For SG dataset, where $\mathbf{H_0} : P = Q$ is true, we can see that if one simply applies the asymptotic null distribution of a χ^2 on top, we will obtain a massively inflated type I error. This is however not the case for TCMC, TCS and NTE, where the type I error is approximately controlled at the right level, this is shown more clearly in Figure **??** in the Appendix. In GMD, GVD and Blobs dataset, the null hypothesis does not hold, and we see that our algorithms indeed discover this difference. As expected we observe a trade-off between privacy level and power, for increasing privacy (decreasing ϵ), we have less power. These experiments also reveals the order of performance of these algorithms, i.e. TCS > TCMC > NTE. This is not surprising, as for TCMC and NTE, we are pertubing the mean and covariance separately, rather than the statistic directly which is the direct quantity we want to protect. The power analysis for the SVD and Blobs dataset also reveal the interesting nature of sampling versus optimisation in our two settings. In the SVD dataset, we observe that NTE performs better than TCS and TCMC, however if we use the same test locations and bandwidth of NTE for TCS and TCMC, the order of performance is as we expect, better for sampling over optimization. However, in the Blobs dataset, we observe that NTE has little or no power, because this dataset is sensitive to the choice of test frequency locations, highlighting the importance of optimisation in this case.

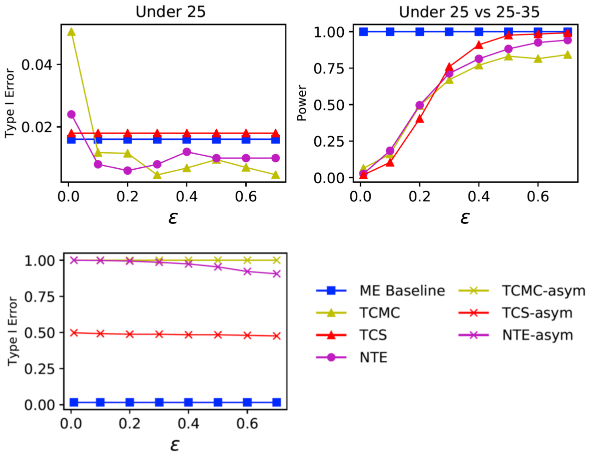


Fig. 3: Type I error for the under25 only test, Power for the under25 vs 25to35 test over 500 runs, with $n = 2500, \delta = 1e^{-5}$. *-asym represents using the asymptotic χ^2 null distribution.

Varying test sample size n We now fix $\epsilon = 2.5$, $\delta = 1.0^{-5}$ and vary n from 1000 to 15 000. The results are shown in the bottom row of Figure 4. The results for the SG dataset further reinforce the importance of not simply using the asymptotic null distribution, as even at very large sample size, the type I error is still inflated when naively computing the test threshold form a chi-squared distribution. This is not the case for TCMC, TCS and NTE, where the type I error is approximately controlled at the correct level for all sample sizes, as shown in Figure **??** in the Appendix.

6.2 Real data: Celebrity age data

We now demonstrate our tests on a real life celebrity age dataset [30], containing 397 949 images of 19 545 celebrities and their corresponding age labels. Here, we will follow the preprocessing of [25], where images from the same celebrity are placed into the same bag, and the bag label is calculated as the mean age of that celebrity's images and use this to construct two datasets, under25 and 25to35. Here the under25 dataset is the images where the corresponding celebrity's bag label is < 25, and the 25to35 dataset is the images corresponding to the celebrity's bag label that is between 25 and 35. The dataset under25 contains 58095 images, and the dataset 25to35 contains 126415 images. For this experiment, we will focus on using the ME version of the test and consider the kernel

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{||\varphi(\mathbf{x}) - \varphi(\mathbf{y})||^2}{2\theta^2}\right)$$

where $\varphi(x) : \mathbb{R}^{256\times256} \to \mathbb{R}^{4096}$ is the feature map learnt by the CNN in [30], mapping the image in the original pixel space to the last layer. For our experiment, we take $N = 3125$, and use 20% of the data for sampling test locations, and calculation of the median heuristic bandwidth. Note here we do not perform optimization, due to the large dimension of the feature map φ. We now perform two tests, for one test we compare samples from under25 only (i.e. $H_0 : P = Q$ holds), and the other we compares samples from under25 to samples from 25to35 (i.e. $\mathbf{H_0} : P = Q$ does not hold). The results are shown in Figure 3 for ϵ from 0.1 to 0.7. We observe that in the under25 only test, the TCMC, TCS and NTE all achieve the correct Type I error rate, this is unlike their counterpart that uses the χ^2 asymptotic null distribution. In the under25 vs 25to35 two sample test, we see that our algorithms can achieve a high power (with little samples) at a high level of privacy, protecting the original images from malicious intent.

Table 1: Synthetic problems (Null hypothesis $\mathbf{H_0}$ holds only for SG). Gaussian Mixtures in $\mathbb{R}^2$, also studied in [6, 19, 23].

7 Conclusion

While kernel-based hypothesis testing provides flexible statistical tools for data analysis, its utility in differentially private settings is not well understood. We investigated differentially private kernel-based two-sample testing procedures, by making use of the sensitivity bounds on the quantities used in the test statistics. While asymptotic null distributions for the modified procedures remain unchanged, ignoring additive noise can lead to an inflated number of false positives. Thus, we propose new approximations of the null distributions under the private regime which give improved Type I control and good power-privacy tradeoffs, as demonstrated in extensive numerical evaluations.

References

[1] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. 2018.

[2] Matej Balog, Ilya Tolstikhin, and Bernhard Schölkopf. Differentially Private Database Release via Kernel Mean Embeddings. October 2017. arXiv: 1710.01641.

[3] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, July 2006.

[4] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *JMLR*, 12(Mar):1069–1109, 2011.

[5] Xinjia Chen. A new generalization of chebyshev inequality for random vectors. *arXiv preprint arXiv:0707.0805*, 2007.

[6] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, pages 1981–1989, 2015.

[7] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Eurocrypt*, volume 4004, pages 486–503. Springer, 2006.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876, pages 265–284. Springer, 2006.

[9] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, August 2014.

[10] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: optimal bounds for privacy-preserving principal component analysis. In *Symposium on Theory of Computing, STOC 2014*, pages 11–20, 2014.

[11] S. Flaxman, D. Sejdinovic, J.P. Cunningham, and S. Filippi. Bayesian Learning of Kernel Embeddings. In *UAI*, pages 182–191, 2016.

[12] Marco Gaboardi, Hyun Woo Lim, Ryan Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *ICML - Volume 48*, ICML'16, pages 2111–2120, 2016.

[13] Marco Gaboardi and Ryan M. Rogers. Local private hypothesis testing: Chi-square tests. *CoRR*, abs/1709.07155, 2017.

[14] Vipul Goyal, Dakshita Khurana, Ilya Mironov, Omkant Pandey, and Amit Sahai. Do distributed differentially-private protocols require oblivious transfer?. In *ICALP*, pages 29:1–29:15, 2016.

[15] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*, pages 513–520. MIT Press, 2007.

[16] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(1):723–773, March 2012.

[17] Arthur Gretton, Kenji Fukumizu, Zaïd Harchaoui, and Bharath K. Sriperumbudur. A fast, consistent kernel two-sample test. In *NIPS*, pages 673–681. 2009.

[18] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *NIPS*, 2012.

[19] Arthur Gretton, Bharath K. Sriperumbudur, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, and Kenji Fukumizu. Optimal kernel choice for large-scale two-sample tests. In *NIPS*. 2012.

[20] Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *JMLR*, 14(Feb):703–727, 2013.

[21] 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 Genetics*, 4(8):1–9, 08 2008.

[22] Prateek Jain and Abhradeep Thakurta. Differentially private learning with kernels. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 118–126, July 2013.

[23] Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *NIPS*, 2016.

[24] Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *ACM SIGKDD 2013*, 2013.

[25] Ho Chung Leon Law, Dougal J Sutherland, Dino Sejdinovic, and Seth Flaxman. Bayesian approaches to distribution regression. In *UAI*, 2017.

[26] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan. The limits of two-party differential privacy. In *IEEE*, Oct 2010.

[27] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[28] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[29] Ryan Rogers and Daniel Kifer. A new class of private chi-square hypothesis tests. In *Artificial Intelligence and Statistics*, pages 991–1000, 2017.

[30] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2018.

[31] G. Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, 1990.

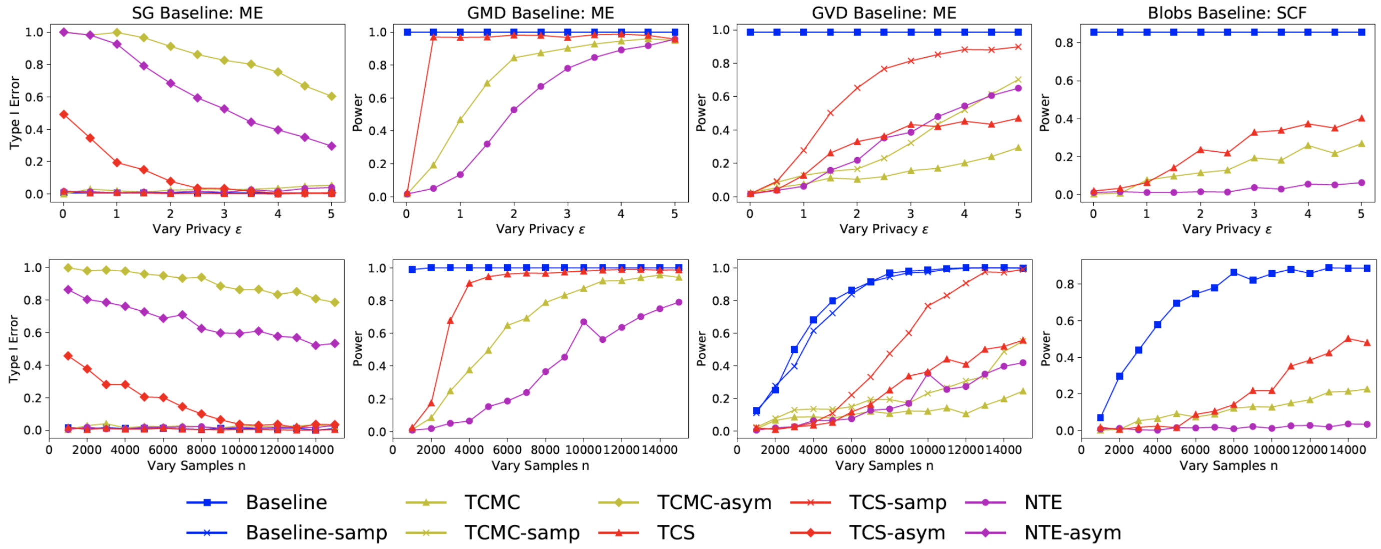


Fig. 4: Type I error for the SG dataset , Power for the GMD, GVD, Blobs dataset over 500 runs, with $\delta = 1e^{-}5$. **Top**: Varying ϵ with $n = 10000$. **Bottom**: Varying n with $\epsilon = 2.5$. Here *-asym represents using the asymptotic χ^2 null distribution, while *-samp represents sampling locations and using the median heuristic bandwidth.