Multivariate Mean Comparison Under Differential Privacy

Martin Dunsche(✉), Tim Kutta, and Holger Dette

Ruhr-University, Bochum, Germany
{martin.dunsche,tim.kutta,holger.dette}@ruhr-uni-bochum.de

Abstract. The comparison of multivariate population means is a central task of statistical inference . While statistical theory provides a variety of analysis tools, they usually do not protect individuals' privacy. This knowledge can create incentives for participants in a study to conceal their true data (especially for outliers), which might result in a distorted analysis. In this paper, we address this problem by developing a hypothesis test for multivariate mean comparisons that guarantees differential privacy to users. The test statistic is based on the popular Hotelling's t^2-statistic, which has a natural interpretation in terms of the Mahalanobis distance. In order to control the type-1-error, we present a bootstrap algorithm under differential privacy that provably yields a reliable test decision. In an empirical study, we demonstrate the applicability of this approach.

Keywords: Differential privacy · Private testing · Private bootstrap

1 Introduction

Over the last decades, the availability of large databases has transformed statistical practice. While data mining flourishes, users are concerned about increasing transparency vis-à-vis third parties. To address this problem, new analysis tools have been devised that balance precise inference with solid privacy guarantees.

In this context, statistical tests that operate under *differential privacy* (DP) are of interest: Statistical tests are the standard tool to validate hypotheses regarding data samples and to this day form the spine of most empirical sciences. Performing tests under DP means determining general trends in the data, while masking individual contribution. This makes it hard for adversaries to retrieve unpublished, personal information from the published analysis.

Related Works: In recent years, hypothesis testing under DP has gained increasing attention. In a seminal work [20] introduces a privatization method, for a broad class of test statistics, that guarantees DP without impairing asymptotic performance. Other theoretical aspects such as optimal tests under DP are considered in [3]. Besides such theoretical investigations, a number of privatized tests have been devised to replace classical inference, where sensitive data is at

J. Domingo-Ferrer and M. Laurent (Eds.): PSD 2022, LNCS 13463, pp. 31–45, 2022.
https://doi.org/10.1007/978-3-031-13945-1_3

stake. For example [11] and [17] consider privatizations of classical goodness of fit tests for categorical data, tailored to applications in genetic research, where privacy of study participants is paramount. In a closely related work, [22] use privatized likelihood-ratio statistics to validate various assumptions for tabular data. Besides, [19] propose a method for privatizations in small sample regimes.

A cornerstone of statistical analysis is the study of population means and accordingly this subject has attracted particular attention. For example, [6] develop a private t-test to compare population means under local differential privacy, while [16] consider the multivariate case in the global setting. [12] and [7] construct private confidence intervals for the mean (which is equivalent to the one-sample t-test) under global DP and [21] suggests a differentially private ANOVA. Moreover, [4] present privatizations for a number of non-parametric tests (such as Wilcoxon signed-rank tests) and [10] devise general confidence intervals for exponential families.

A key problem of statistical inference under DP consists in the fact that privatization inflates the variance of the test statistics. If this is not taken into account properly, it can destabilize subsequent analysis and lead to the "discovery" of spurious effects. To address these problems, recent works (such as [11] and [10]) have employed resampling procedures that explicitly incorporate the effects of privatization and are therefore more reliable than tests based on standard, asymptotic theory.

Our Contributions: In this work, we present a test for multivariate mean comparisons under pure-DP, based on the popular Hotelling's t^2-statistic. We retrieve the effect that asymptotic test decisions work under DP, as long as privatizations are weak, whereas for strong privatizations, they yield distorted results (see Sect. 4 for details). As a remedy, we consider a parametric bootstrap that cuts false rejections and is provably consistent for increasing sample size. This method can be extended to other testing problems, is easy to implement (even for non-expert users) and can be efficiently automatized as part of larger data disseminating structures. We demonstrate the efficacy of our approach, even for higher dimensions and strong privatizations, in a simulation study. The proofs of all mathematical results are deferred to the Appendix. The work most closely related to our paper is [16], who consider Hotelling's t^2-statistic for approximate DP and propose a test based on a (heuristic) resampling strategy. In contrast to this paper, we focus on pure-DP, employ a different privatization mechanism and a parametric bootstrap test, for which we provide a rigorous proof of its validity (see Sect. 3.2).

2 Mathematical Background

In this section, we provide the mathematical context for private mean comparisons, beginning with a general introduction into two sample tests. Subsequently, we discuss Hotelling's t^2-test, which is a standard tool to assess mean deviations. Finally, we define the notion of differential privacy and consider key properties, such as stability under post-processing. Readers familiar with any of these topics can skip the respective section.

2.1 Statistical Tests for Two Samples

In this work, we are interested in testing statistical hypotheses regarding the distribution of two data samples (of random vectors) $X_1, ..., X_{n_1}$ and $Y_1, ..., Y_{n_2}$.

Statistical tests are decision rules that select one out of two rivaling hypotheses H_0 and H_1, where H_0 is referred to as the "null hypothesis" (default belief) and H_1 as the "alternative". To make this decision, a statistical test creates a summary statistic $S := S(X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2})$ from the data and based on S determines whether to keep H_0, or to switch to H_1. Typically, the decision to reject H_0 in favor of H_1 is made, if S surpasses a certain threshold q, above which, the value of S seems at odds with H_0. In this situation, the threshold q may or may not depend on the data samples.

Given the randomness in statistical data, there is always a risk of making the wrong decision. Hypothesis-alternative-pairs (H_0, H_1) are usually formulated such that mistakenly keeping H_0 inflicts only minor costs on the user, while wrongly switching to H_1 produces major ones. In this spirit, tests are constructed to keep the risk of false rejection below a predetermined level α, i.e. $\mathbb{P}_{H_0}(S > q) \leq \alpha$, which is referred to as the *nominal level* (or *type-1-error*). Commonly, the nominal level is chosen as $\alpha \in \{0.1, 0.05, 0.01\}$. Notice that α can be regarded as an input parameter of the threshold $q = q(\alpha)$. Even though sometimes an exact nominal level can be guaranteed, in practice most tests only satisfy an asymptotic nominal level, i.e. $\limsup_{n_1,n_2\to\infty} \mathbb{P}_{H_0}(S > q(\alpha)) \leq \alpha$. Besides controlling the type-1-error, a reasonable test has to be *consistent*, i.e. it has to reject H_0 if H_1 holds and sufficient data is available. In terms of the summary statistic S, this means that S increases for larger data samples and transgresses $q(\alpha)$ with growing probability $\lim_{n_1,n_2\to\infty} \mathbb{P}_{H_1}(S > q(\alpha)) = 1$.

2.2 Hotelling's t^2-Test

We now consider a specific test for the comparison of multivariate means: Suppose that two independent samples of random vectors $X_1, ..., X_{n_1}$ and $Y_1, ..., Y_{n_2}$ are given, both stemming from the d-dimensional cube $[-m, m]^d$, where $m > 0$ and $d \in \mathbb{N}$. Furthermore, we assume that both samples consist of independent identically distributed (i.i.d) observations. Conceptually, each vector corresponds to the data of one individual and we want to use these to test the "hypothesis-alternative"-pair

$$H_0 : \mu_X = \mu_Y \ , \qquad H_1 : \mu_X \neq \mu_Y \ , \tag{2.1}$$

where $\mu_X := \mathbb{E}[X_1] \in \mathbb{R}^d, \mu_Y := \mathbb{E}[Y_1] \in \mathbb{R}^d$ denote the respective expectations. A standard way to test (2.1) is provided by *Hotelling's t^2-test*, which is based on the test statistic

$$t^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{X} - \bar{Y})^T \hat{\Sigma}^{-1} (\bar{X} - \bar{Y}) \ , \tag{2.2}$$

where $\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ and $\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ denote the respective sample means and the pooled sample covariance is given by

$$\hat{\Sigma} = \frac{(n_1 - 1)\hat{\Sigma}_X + (n_2 - 1)\hat{\Sigma}_Y}{n_1 + n_2 - 2}.$$

Here, $\hat{\Sigma}_X = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \mu_X)(X_i - \mu_X)^\top$ and $\hat{\Sigma}_Y = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \mu_Y)(Y_i - \mu_Y)^\top$ denote the sample covariance matrices of X_1 and Y_1, respectively. Assuming that $\Sigma_X = \Sigma_Y$ (a standard condition for Hotelling's t^2-test) $\hat{\Sigma}$ is a consistent estimator for the common covariance.

We briefly formulate a few observations regarding the t^2-statistic:

i) In the simple case of $d = 1$, the t^2-statistic collapses to the (squared) statistic of the better-known two sample t-test.
ii) We can rewrite
$$t^2 = \frac{n_1 n_2}{n_1 + n_2}\left\|\hat{\Sigma}^{-1/2}(\bar{X} - \bar{Y})\right\|_2^2.$$
As a consequence, the t^2-statistic is non-negative and assumes high values if $\bar{X} - \bar{Y} \approx \mu_X - \mu_Y$ is large in the norm.
iii) The t^2-statistic is closely related to the Mahalanobis distance, which is a standard measure for multivariate mean comparisons (see [5]).

In order to formulate a statistical test based on the t^2-statistic, we consider its large sample behavior. Under the hypothesis $\sqrt{n_1 n_2/(n_1 + n_2)}\hat{\Sigma}^{-1/2}(\bar{X} - \bar{Y})$ follows (approximately) a d-dimensional, standard normal distribution, such that its squared norm (that is the t^2-statistic) is approximately χ^2_d distributed (chi-squared with d degrees of freedom). Now if $q_{1-\alpha}$ denotes the upper α-quantile of the χ^2_d distribution, the test decision "reject H_0 if $t^2 > q_{1-\alpha}$", yields a consistent, asymptotic level α-test for any $\alpha \in (0, 1)$. For details on Hotelling's t^2-test we refer to [15].

2.3 Differential Privacy

Differential privacy (DP) has over the last decade become the de facto gold standard in privacy assessment of data disseminating procedures (see e.g. [9,14] or [18]). Intuitively, DP describes the difficulty of inferring individual inputs from the releases of a randomized algorithm. This notion is well suited to a statistical framework, where a trusted institution, like a hospital, publishes results of a study (algorithmic releases), but candidates would prefer to conceal participation (individual inputs). To make this notion mathematically rigorous, we consider databases $\mathbf{x}, \mathbf{x}' \in \mathcal{D}^n$, where $\mathcal{D}$ is some set, and call them *adjacent* or *neighboring*, if they differ in only one entry.

Definition 2.3.1. A randomized algorithm $A : \mathcal{D}^n \to \mathbb{R}$ is called *ε-differentially private* for some $\varepsilon > 0$, if for any measurable event $E \subset \mathbb{R}$ and any adjacent $\mathbf{x}, \mathbf{x}'$

$$\mathbb{P}(A(\mathbf{x}) \in E) \leq e^{\varepsilon}\, \mathbb{P}(A(\mathbf{x}') \in E) \tag{2.3}$$

holds.

Condition (2.3) requires that the distribution of $A(\mathbf{x})$ does not change too much, if one entry of $\mathbf{x}$ is exchanged (where small ε correspond to less change and thus stronger privacy guarantees). In statistical applications, private algorithms are usually assembled modularly: They take as building blocks some well-known private algorithms (e.g., the Laplace or Exponential Mechanism), use them to privatize key variables (empirical mean, variance etc.) and aggregate the privatized statistic. This approach is justified by two stability properties of DP: Firstly, privacy preservation under post-processing, which ensures that if A satisfies ε-DP, so does any measurable transformation $h(A)$. Secondly, the composition theorem that maintains at least $\sum_{i=1}^{k} \varepsilon_i$-DP of a vector $(A_1, ..., A_k)$ of algorithms, where A_i are independent ε_i-differentially private algorithms. In the next section, we employ such a modular privatization of the Hotelling's t^2-statistic for private mean comparison. We conclude our discussion on privacy with a small remark on the role of the "trusted curator".

Remark 2.3.1. Discussions of (global) DP usually rely on the existence of some "trusted curator" who aggregates and privatizes data before publication. In reality this role could be filled by an automatized, cryptographic protocol (secure multi-party computation), which calculates and privatizes the statistic before publication without any party having access to the full data set (for details see [2,13]). This process has the positive side effect that it prevents a curator from re-privatizing if an output seems too outlandish (overturning privacy in the process).

3 Privatized Mean Comparison

In this section, we introduce a privatized version t^{DP} of Hotelling's t^2-statistic. Analogous to the traditional t^2-statistic, the rejection rule "$t^{DP} > q_{1-\alpha}$" yields in principle a consistent, asymptotic level-α test for H_0 (see Theorem 3.1.2). However, empirical rejection rates often exceed the prescribed nominal level α for a combination of low sample sizes and high privatization (see Example B). As a consequence, we devise a parametric bootstrap for a data-driven rejection rule. We validate this approach theoretically (Theorem 3.2.1) and demonstrate empirically a good approximation of the nominal level in Sect. 4.

3.1 Privatization of the t^2-Statistic

We begin this section by formulating the Assumptions of the following, theoretical results:

Assumption 3.1.1. (1) The samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent, each consisting of i.i.d. observations and are both supported on the cube $[-m, m]^d$, for some known $m > 0$.

(2) The covariance matrices

$$\Sigma_X := \mathbb{E}[(X_1 - \mu_X)(X_1 - \mu_X)^T]; \quad \Sigma_Y := \mathbb{E}[(Y_1 - \mu_Y)(Y_1 - \mu_Y)^T] \ .$$

are identical and invertible.

(3) The sample sizes n_1, n_2 are of the same order. That is with $n := n_1 + n_2$ we have

$$\lim_{n\to\infty} \frac{n_i}{n} = \xi_i \in (0,1) \qquad i = 1, 2.$$

We briefly comment on the Assumptions made.

Remark 3.1.1. (1): The assumption of independent observations is common in the literature on machine learning and justified in many instances. Boundedness of the data -with some known bound- is an important precondition for standard methods of privatization (such as the below discussed Laplace Mechanism or the ED algorithm). Generalization are usually possible (see e.g., [20]) but lie beyond the scope of this paper.

(2): Invertibility of the covariance matrices is necessary to define the Mahalanobis distance. If this assumption is violated, either using another distance measure (defining a different test) or a prior reduction of dimensions is advisable.

Equality of the matrices $\Sigma_X = \Sigma_Y$ is assumed for ease of presentation, but can be dropped, if the pooled estimate $\hat{\Sigma}$ is replaced by the re-weighted version

$$\hat{\Sigma}^{\neq} := \frac{n_2\hat{\Sigma}_X + n_1\hat{\Sigma}_Y}{n_1 + n_2}.$$

(3): We assume that asymptotically the size of each group is non-negligible. This assumption is standard in the analysis of two sample tests and implies that the noise in the estimates of both groups is of equal magnitude. If this was not the case and e.g. $\xi_1 = 0$ (in practice $n_1 << n_2$) it is more appropriate to model the situation as a one-sample test (as μ_Y is basically known).

Recall the definition of Hotelling's t^2-statistic in (2.2). By construction, we can express the t^2-statistic as a deterministic function of four data dependent entities: The sample means $\bar{X}, \bar{Y}$ and the sample covariance matrices $\hat{\Sigma}_X, \hat{\Sigma}_Y$. According to the *composition-* and *post-processing theorem* of DP (see Sect. 2.3) we can privatize the t^2-statistic by privatizing each of these inputs.

For the privatization of the sample means, we use the popular *Laplace Mechanism* (see [8], p.32): It is well-known that $\bar{X}^{DP} := \bar{X} + Z$ and $\bar{Y}^{DP} := \bar{Y} + Z'$ fulfill $\varepsilon/4$-DP, if $Z = (Z_1, \ldots, Z_d)^T$ and $Z' = (Z'_1, \ldots, Z'_d)^T$ consist of independent random variables $Z_k \sim Lap(0, \frac{2md}{n_1(\varepsilon/4)})$ and $Z'_k \sim Lap(0, \frac{2md}{n_2(\varepsilon/4)})$ for $k = 1, ..., d$.

For the privatization of the covariance matrices $\hat{\Sigma}_X, \hat{\Sigma}_Y$ we employ the ED Mechanism, specified in the Appendix (which is a simple adaption of the Algorithm proposed in [1]). We can thus define differentially private estimates $\hat{\Sigma}_X^{DP} := ED(\hat{\Sigma}_X, \varepsilon/4)$ and $\hat{\Sigma}_Y^{DP} := ED(\hat{\Sigma}_Y, \varepsilon/4)$, both satisfying $\varepsilon/4$-DP. We point out that the outputs of ED are always covariance matrices (positive semi-definite and symmetric). Therewith, we can define a privatized pooled sample covariance matrix as

$$\hat{\Sigma}^{DP} := \frac{(n_1 - 1)\hat{\Sigma}_X^{DP} + (n_2 - 1)\hat{\Sigma}_Y^{DP}}{n_1 + n_2 - 2} + diag(c_1 + c_2)\ ,$$

Algorithm 1. Privatized statistics (PS)

Input: means: $\bar{X}$, $\bar{Y}$, covariance matrices: $\hat{\Sigma}_X$, $\hat{\Sigma}_Y$, privacy level: ε

Output: $\bar{X}^{DP}$, $\bar{Y}^{DP}$, $\hat{\Sigma}_X^{DP}$, $\hat{\Sigma}_Y^{DP}$

function PS($\bar{X}$, $\bar{Y}$, $\hat{\Sigma}_X$, $\hat{\Sigma}_Y$, ε)
  for $i = 1, \ldots, d$ **do**
    Generate $Z_i \sim Lap(0, \frac{2md}{n_1 \varepsilon/4})$
    Generate $Z_i' \sim Lap(0, \frac{2md}{n_2 \varepsilon/4})$
  end for
  Set $\bar{X}^{DP} := \bar{X} + (Z_1, ..., Z_d)$, $\bar{Y}^{DP} := \bar{Y} + (Z_1', ..., Z_d')$
  Set $\hat{\Sigma}_X^{DP} = ED(\hat{\Sigma}_X, \varepsilon/4)$, $\hat{\Sigma}_Y^{DP} = ED(\hat{\Sigma}_Y, \varepsilon/4)$
  return $\bar{X}^{DP}, \bar{Y}^{DP}, \hat{\Sigma}_X^{DP}, \hat{\Sigma}_Y^{DP}$
end function

where $c_1 := 2(\frac{2md}{n_1(\varepsilon/4)})^2, c_2 := 2(\frac{2md}{n_2(\varepsilon/4)})^2$ are corrections accounting for variance increase, due to the mean privatizations. Finally, we can formulate a privatized version of the Hotelling's t^2-statistic as follows:

$$\begin{aligned} t^{DP} &= \frac{n_1 n_2}{n_1 + n_2} (\bar{X}^{DP} - \bar{Y}^{DP})^T [\hat{\Sigma}^{DP}]^{-1} (\bar{X}^{DP} - \bar{Y}^{DP}) \\ &= \frac{n_1 n_2}{n_1 + n_2} \left\| [\hat{\Sigma}^{DP}]^{-1/2} (\bar{X}^{DP} - \bar{Y}^{DP}) \right\|_2^2 \end{aligned} \tag{3.1}$$

Theorem 3.1.1. The privatized t^2-statistic t^{DP} is ε-*differentially private.*

In the one dimensional case, the covariance privatization by ED boils down to an application of the Laplace Mechanism and t^{DP} has a simple closed form.

Example 3.1.1. (Privatization in $d = 1$) Assume that $d = 1$. Then the data $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ originates from the interval $[-m, m]$ and we can write the privatized test statistic as

$$t^{DP} = \frac{n_1 n_2}{n_1 + n_2} \frac{(\bar{X}^{DP} - \bar{Y}^{DP})^2}{(\sigma^{DP})^2} ,$$

where

$$\begin{aligned} (\sigma^{DP})^2 := &\frac{(n_1 - 1)(|\hat{\sigma}_X + L_1|) + (n_2 - 1)(|\hat{\sigma}_Y + L_2|)}{n_1 + n_2 - 2} \\ &+ 2\Big(\frac{2m}{n_1(\varepsilon/4)}\Big)^2 + 2\Big(\frac{2m}{n_2(\varepsilon/4)}\Big)^2 . \end{aligned}$$

Here, L_1 and L_2 follow a centered Laplace distribution, with variance specified in the Appendix. Note that the privatization of $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ is conforming with the privatization of Algorithm ED (see Appendix), since the first (and only) eigenvalue is the sample variance itself, while privatization of eigenvectors is a non-issue for $d = 1$.

As for the non-privatized t^2-statistic, we can prove under H_0 that t^{DP} approximates a χ^2_d-distribution as $n_1, n_2 \to \infty$. This means that (at least for large sample sizes) the perturbations introduced by the Laplace noise and the ED-algorithm are negligible.

Algorithm 2. Privatized Hotelling's t^2-test (PHT)

Input: means: $\bar{X}^{DP}$, $\bar{Y}^{DP}$, covariance matrices: $\hat{\Sigma}_X^{DP}$, $\hat{\Sigma}_Y^{DP}$, quantile: q

Output: *choice* $\in \{0, 1\}$ coding for acceptation (0) or rejection (1) of H_0

function PHT($\bar{X}^{DP}$, $\bar{Y}^{DP}$, $\hat{\Sigma}_X^{DP}$, $\hat{\Sigma}_Y^{DP}$, q)
    Compute t^{DP} (defined in 3.1)
    Define *choice* $= 0$
    if $t^{DP} > q$ **then**
        Set *choice* $= 1$
    end if
    return *choice*
end function

Theorem 3.1.2. The decision rule "reject if

$$t^{DP} > q_{1-\alpha} \tag{3.2}$$

(Algorithm 2)" where $q = q_{1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2_d distribution, yields a consistent, asymptotic level-α test for the hypotheses (2.1).

Theorem 3.1.2 underpins the assertion that "asymptotically, privatizations do not matter". Yet in practice, privatizations can have a dramatic impact on the (finite sample) performance of tests.

3.2 Bootstrap

In this section, we consider a modified rejection rule for H_0, based on t^{DP}, that circumvents the problem of inflated type-1-error (see Example B). Privatizations increase variance and therefore t^{DP} is less strongly concentrated than t^2, leading to excessive transgressions of the threshold $q_{1-\alpha}$. Consequently, to guarantee an accurate approximation of the nominal level, a different threshold is necessary.

Hypothetically, if we knew the true distribution of t^{DP} under H_0, we could analytically calculate the exact α-quantile $q_{1-\alpha}^{exact}$ and use the rejection rule "$t^{DP} > q_{1-\alpha}^{exact}$". Of course, in practice, these quantiles are not available, but we can use a *parametric bootstrap* to approximate $q_{1-\alpha}^{exact}$ by an empirical version $q_{1-\alpha}^{*}$ calculated from the data. In Algorithm 3 we describe the systematic derivation of $q_{1-\alpha}^{*}$.

Algorithm 3. Quantile Bootstrap (QB)

Input: Covariance matrices: $\hat{\Sigma}_X^{DP}$, $\hat{\Sigma}_Y^{DP}$, sample sizes: n_1,n_2, bootstrap iterations: B

Output: Empirical $1-\alpha$ quantile of t^{DP}: $q_{1-\alpha}^*$.

function QB($\hat{\Sigma}_X^{DP}$, $\hat{\Sigma}_Y^{DP}$, n_1,n_2, B)
  for $i = 1, \ldots, B$ **do**
    Sample $\bar{X}^* \sim \mathcal{N}(0, \frac{\hat{\Sigma}_X^{DP}}{n_1})$ and $\bar{Y}^* \sim \mathcal{N}(0, \frac{\hat{\Sigma}_Y^{DP}}{n_2})$
    for $k = 1, \ldots, d$ **do**
      Generate $Z_k \sim Lap(0, \frac{2md}{n_1(\varepsilon/4)})$
      Generate $Z'_k \sim Lap(0, \frac{2md}{n_2(\varepsilon/4)})$
    end for
    Define $\bar{X}^{DP*} := \bar{X}^* + (Z_1, ..., Z_d)$
    Define $\bar{Y}^{DP*} := \bar{Y}^* + (Z'_1, ..., Z'_d)$
    Define $t_i^{DP^*} := \frac{n_1 n_2}{n_1+n_2} \left\| [\hat{\Sigma}^{DP}]^{-1/2} (\bar{X}^{DP*} - \bar{Y}^{DP*}) \right\|_2^2$
  end for
  Sort statistics in ascending order: $(t_{(1)}^{DP^*}, ..., t_{(B)}^{DP^*}) = sort((t_1^{DP^*}, ..., t_B^{DP^*}))$
  Define $q_{1-\alpha}^* := t^{DP^*}{}_{(\lfloor (1-\alpha)B \rfloor)}$
  return $q_{1-\alpha}^*$
end function

Algorithm 3 creates B *bootstrap versions* $t_1^{DP^*}, ..., t_B^{DP^*}$, that mimic the behavior of t^{DP}. So, e.g., $\bar{X}^{DP*}$ (in $t_i^{DP^*}$) has a distribution close to that of $\bar{X}^{DP}$ (in t^{DP}), which, if centered, is approximately normal with covariance matrix Σ_X/n_1. As a consequence of this parallel construction, the *empirical* $1-\alpha$*-quantile* $q_{1-\alpha}^*$ is close to the true $(1-\alpha)$-quantile of the distribution of t^{DP}, at least if the number B of bootstrap replications is sufficiently large. In practice, the choice of B depends on α (where small α require larger B), but our simulations suggest that for a few hundred iterations the results are already reasonable even for nominal levels as small as 1%.

Theorem 3.2.1. The decision rule "reject if

$$t^{DP} > q_{1-\alpha}^* \tag{3.3}$$

(Algorithm 2)", where $q_{1-\alpha}^*$ is chosen by Algorithm 3, yields a consistent, asymptotic level-α test in the sense that

$$\lim_{B\to\infty} \lim_{n_1,n_2\to\infty} \mathbb{P}_{H_0}(t^{DP} > q_{1-\alpha}^*) = \alpha,$$

(level α) and

$$\lim_{n_1,n_2\to\infty} \mathbb{P}_{H_1}(t^{DP} > q_{1-\alpha}^*) = 1$$

(consistency).

4 Simulation

In this section we investigate the empirical properties of our methodology by means of a small simulation study.

Data Generation: In the following, the first sample $X_1, ..., X_{n_1}$ is drawn from the uniform distribution on the d-dimensional cube $[-\sqrt{3}, \sqrt{3}]^d$, whereas the second sample $Y_1, ..., Y_{n_2}$ is uniformly drawn from the shifted cube $[-\sqrt{3} + a/\sqrt{d}, \sqrt{3} + a/\sqrt{d}]^d$. Here, $a \geq 0$ determines the mean difference of the two samples. In particular $a = 0$ corresponds to the hypothesis $\mu_X = \mu_Y = (0, ..., 0)^T$, whereas for $a > 0$, $\|\mu_X - \mu_Y\|_2 = a$. We also point out that both samples have the same covariance matrix $\Sigma_X = \Sigma_Y = Id_{d\times d}$. As a consequence, deviations in each component of $\mu_X - \mu_Y$ have equal influence on the rejection probability.

Parameter Settings: In the following we discuss various settings: We consider different group sizes n, between 10^2 and 10^5, privacy levels $\varepsilon = 1/10, 1/2, 1, 5$ and dimensions $d = 1, 10, 30$. The nominal level α is fixed at 5% and the number of bootstrap samples is consistently $B = 200$. All below results are based on 1000 simulation runs.

Empirical Type-1-Error: We begin by studying the behavior of our test decisions under the null hypothesis ($a = 0$). In Table 1 we report the empirical rejection probabilities for the bootstrap test (3.3) (top) and the asymptotic test (3.2) (bottom). The empirical findings confirm our theoretical results from the previous Section.

On the one hand, we observe that the bootstrap test approximates the nominal-level reasonably well (compare Theorem 3.2.1), even in scenarios with small sample size and high dimensions. In contrast, the validity of the asymptotic test (3.2) depends on the negligibility of privatization effects (see discussion of Theorem 3.1.2). Consequently, it works best for large ε and large sample sizes. However, for higher dimensions d, the asymptotic approach breaks down quickly, in the face of more noise by privatizations and thus stronger digressions from the limiting distribution.

Table 1. Empirical type-1-error

	ε	d =1				d =10				d =30			
		$n_1 = n_2$											
		10^2	10^3	10^4	10^5	10^2	10^3	10^4	10^5	10^2	10^3	10^4	10^5
test (3.3)	0.1	0.052	0.046	0.051	0.048	0.058	0.05	0.068	0.063	0.04	0.057	0.056	0.062
	0.5	0.054	0.05	0.059	0.05	0.039	0.06	0.057	0.052	0.054	0.054	0.06	0.056
	1	0.053	0.05	0.054	0.053	0.048	0.061	0.038	0.069	0.048	0.063	0.056	0.054
	5	0.041	0.053	0.043	0.053	0.055	0.053	0.056	0.051	0.044	0.05	0.062	0.052
test (3.2)	0.1	0.738	0.676	0.328	0.093	1	1	1	1	1	1	1	1
	0.5	0.4	0.154	0.055	0.058	1	1	1	0.891	1	1	1	1
	1	0.24	0.063	0.057	0.044	1	1	0.993	0.428	1	1	1	1
	5	0.054	0.047	0.045	0.039	0.990	0.933	0.181	0.062	1	1	0.999	0.301

Empirical Power: Next we consider the power of our test. Given the poor performance of the asymptotic test (3.2) in higher dimensions (the key interest

of this paper) we restrict our analysis to the bootstrap test (3.3) for the sake of brevity. In the following, we consider the alternative for $a = 1$. Recall that $\|\mu_X - \mu_Y\|_2 = a$ is independent of the dimension. However, we expect more power in low dimensions due to weaker privatization. In Fig. 1, we display a panel of empirical power curves, each graphic reflecting a different choice of the privacy parameter ($\varepsilon = 1/10, 1/2, 1, 5$) and each curve corresponding to a different dimension ($d = 1, 10, 30$). The group size is reported in logarithmic scale on the x-axis and the rejection probability on the y-axis. As might be expected, low dimensions and weak privatizations (i.e., large ε) are directly associated with a sharper increase of the power curves and smaller sample sizes to attain high power. For instance, moving from $\varepsilon = 1/2$ (high privatization) to the less demanding $\varepsilon = 5$ (low privatization) means that a power of 90% is attained with group sizes that are about an order of magnitude smaller. Similarly, increasing dimension translates into lower power: To attain for $\varepsilon = 0.1$ and $d = 30$, high power requires samples of a few ten thousand observations (see Fig. 1(a)). Even though such numbers are not in excess of those used in related studies (see e.g. [6]) nor of those raised by large tech cooperations, this trend indicates that comparing means of even higher dimensional populations might require (private) pre-processing to reduce dimensions.

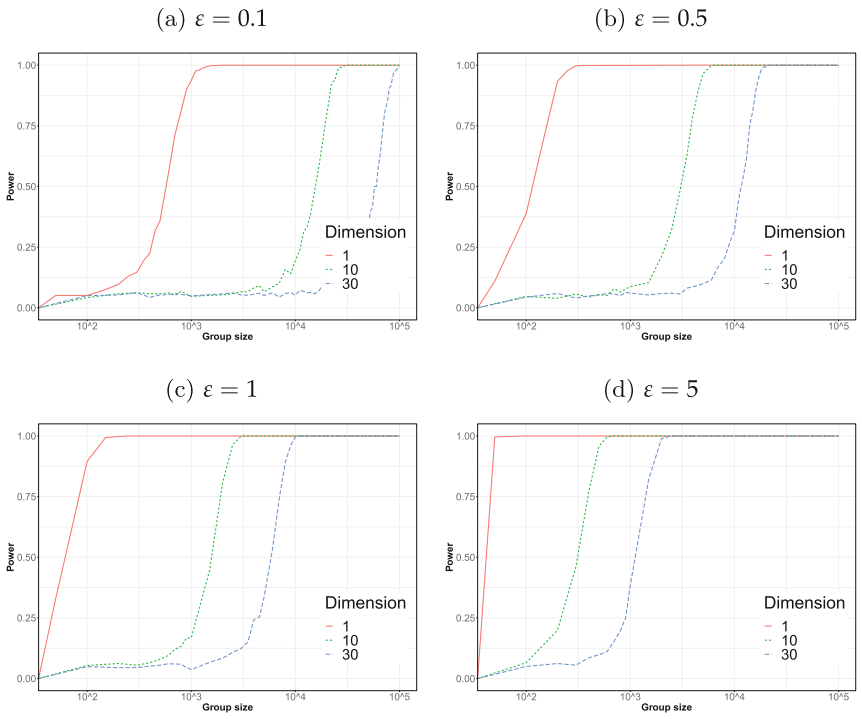

Fig. 1. Simulated power of the bootstrap test (3.3) under a uniform alternative for $\varepsilon = 0.1, 0.5, 1, 5$ and different group sizes.

5 Conclusion

In this paper, we have considered a new way to test multidimensional mean differences under the constraint of differential privacy. Our test employs a privatized version of the popular Hotelling's t^2-statistic, together with a bootstraped rejection rule. While strong privacy requirements always go hand in hand with a loss in power, the test presented in this paper respects the nominal level α with high precision, even for moderate sample sizes, high dimensions and strong privatizations. The empirical advantages are underpinned by theoretical guarantees for large samples. Given the easy implementation and reliable performance, the test can be used as an automatized part of larger analytical structures.

Acknowledgement. This work was partially funded by the DFG under Germany's Excellence Strategy - EXC 2092 CASA - 390781972.

A Proofs

For the proofs, we refer to the arxiv version (see https://arxiv.org/abs/2110.07996).

B Effects of Privatization - Example

In most instances, privatizing a test statistic has no influence on its asymptotic behavior, s.t. rejection rules based on asymptotic quantiles remain theoretically valid. However, empirical studies demonstrate, that in practice even moderate privacy levels can lead to inflated type-1-errors – in our case because the quantiles of the χ^2_d-distribution do not provide good approximations for those of t^{DP}.

To illustrate this effect we consider the case $d = 1$, discussed in Example 3.1.1 for samples of sizes $n_1 = n_2 = 500$, both of which drawn according to the same density, $f(t) \propto \exp(-2t^2)$ on the interval $[-1, 1]$. We simulate the quantile functions (inverse of the distribution function) of χ^2_1 and t^{DP} respectively for privacy levels $\varepsilon = 1, 4$. Figure 2 indicates that for moderate privacy guarantees ($\varepsilon = 4$) the distribution of t^{DP} is close to that of the χ^2_1, s.t. for instance $\mathbb{P}_{H_0}(t^{DP} > q_{0.95}) \approx 6.8\%$ (where again $q_{1-\alpha}$ is the α quantile of the χ^2_1-distribution). This approximation seems reasonable, but it deteriorates quickly for smaller ε. Indeed, for $\varepsilon = 1$ we observe that $\mathbb{P}_{H_0}(t^{DP} > q_{0.95}) \approx 18.9\%$, which is a dramatic error. This effect is still more pronounced in higher dimensions and much larger sample sizes are needed to mitigate it (for details see Table 1).

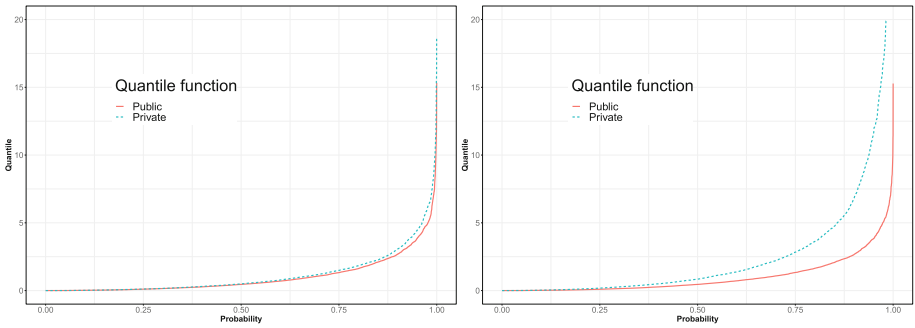

Fig. 2. Simulated quantile functions for χ^2_1 (red) and t^{DP} (blue) for privacy levels $\varepsilon = 4$ (left) and $\varepsilon = 1$ (right)

Summarizing this discussion, we recommend to use Hotelling's t^2-test (3.2) based on the privatized statistic t^{DP} with the standard (asymptotic) quantiles only in situations where sample sizes are large, the dimension is small and privatizations are weak. In all other cases, specifically for larger dimension and stronger privatization, the quantiles have to be adapted to avoid inflated rejection errors under the null hypothesis.

C Algorithms

In the following, we will state two algorithms which describe the covariance privatization. Here, Algorithm 5 **ED** is used for the privatization, while Algorithm 4 describes the eigenvector sampling process. In Algorithm 5 **ED** the privatization budget is not supposed to be separated (for eigenvalues and eigenvectors) in the case $d = 1$ (as eigenvector privatization is unnecessary for $d = 1$). For more details, see [1].

Algorithm 4. Eigenvector sampling

Input: $\tilde{C} \in \mathbb{R}^{q\times q}$, privacy parameter ε

Output: Eigenvector u.

function SAMPLE($\tilde{C}$,ε)
  Define $A := -\frac{\varepsilon}{4}\tilde{C} + \frac{\varepsilon}{4}\hat{\lambda}_1 I_q$, where $\hat{\lambda}_1$ denotes the largest eigenvalue of C.
  Define $\Omega = I_q + 2A/b$, where b satisfies $\sum_{i=1}^{q} \frac{1}{b+2\lambda_i(A)} = 1$.
  Define $M := \exp(-(q-b)/2)(q/b)^{q/2}$.
  Set $ANS = 0$
  while $ANS = 0$ **do**
    Sample $X \sim \mathcal{N}_q\left(0, \Omega^{-1}\right)$ and set $u := z/\left\|z\right\|_2$.
    With probability $\frac{\exp(-u^T Au)}{M(u^T\Omega u)^{q/2}}$ $ANS = 1$
    return u.
  end while
end function

Algorithm 5. Covariance estimation with algorithm **ED**

Input: $\hat{C} \in \mathbb{R}^{d\times d}$, privacy parameter ε, sample size n

Output: Privatized covariance matrix $\hat{\Sigma}^{DP}$

Separate the privacy budget uniformly in $d+1$ parts, i.e. each step $\frac{\varepsilon}{d+1}$
function ED($\hat{C}$,ε,n)
    Initialize $C_1 := \frac{n\hat{C}}{dm^2}$, $P_1 := I_d$.
    Privatize the eigenvalue vector by $(\bar{\lambda}_1, \ldots, \bar{\lambda}_d)^T = \left|(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)^T + \left(Lap\left(\frac{2}{(\varepsilon/(d+1))}\right), \ldots, Lap\left(\frac{2}{(\varepsilon/(d+1))}\right)\right)^T\right|$.
    for $i = 1, \ldots, d-1$ **do**
        Sample $\bar{u}_i \in S^{d-i}$ with $\bar{u}_i := Sample(\hat{C}, \frac{\varepsilon}{d+1})$ and let $\bar{v}_i := P_i^T \bar{u}_i$.
        Find an orthonormal basis $P_{i+1} \in \mathbb{R}^{(d-i)\times d}$ orthogonal to $\bar{v}_1, \ldots, \bar{v}_i$.
        Let $C_{i+1} := P_{i+1}\hat{C}P_{i+1}^T \in \mathbb{R}^{(d-i)\times(d-i)}$.
    end for
    Sample $\bar{u}_d \in S^0$ proportional to $f_{C_d}(u) = \exp\left((\frac{\varepsilon_i}{4})u^T C_d u\right)$ and let $\bar{v}_d := P_d^T \bar{u}_d$.
    $C^{ED} := \sum_{i=1}^{d} \bar{\lambda}_i \bar{v}_i \bar{v}_i^T$.
    return $\hat{\Sigma}^{DP} = \frac{1}{n} C^{ED}$
end function

References

1. Amin, K., Dick, T., Kulesza, A., Munoz, A., Vassilvitskii, S.: Differentially private covariance estimation. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019). proceedings.neurips.cc/paper/2019/file/4158f6d19559955bae372bb00f6204e4-Paper.pdf
2. Bogetoft, P., et al.: Secure Multiparty Computation Goes Live. In: Dingledine, R., Golle, P. (eds.) FC 2009. LNCS, vol. 5628, pp. 325–343. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03549-4_20
3. Canonne, C.L., Kamath, G., McMillan, A., Smith, A., Ullman, J.: The structure of optimal private tests for simple hypotheses. In: Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, pp. 310–321 (2019)
4. Couch, S., Kazan, Z., Shi, K., Bray, A., Groce, A.: Differentially private nonparametric hypothesis testing. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 737–751 (2019)
5. De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.: The mahalanobis distance. Chemomet. Intell. Lab. Syst.**50**(1), 1–18 (2000). https://doi.org/10.1016/S0169-7439(99)00047-7. www.sciencedirect.com/science/article/pii/S0169743999000477
6. Ding, B., Nori, H., Li, P., Allen, J.: Comparing population means under local differential privacy: with significance and power. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
7. Du, W., Foot, C., Moniot, M., Bray, A., Groce, A.: Differentially private confidence intervals. arXiv preprint arXiv:2001.02285 (2020)
8. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014). https://doi.org/10.1561/0400000042

9. Erlingsson, U., Pihur, V., Korolova, A.: Rappor: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 1054–1067. CCS 2014, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2660267.2660348
10. Ferrando, C., Wang, S., Sheldon, D.: General-purpose differentially-private confidence intervals. arXiv preprint arXiv:2006.07749 (2020)
11. Gaboardi, M., Lim, H., Rogers, R., Vadhan, S.: Differentially private chi-squared hypothesis testing: goodness of fit and independence testing. In: International Conference on Machine Learning, pp. 2111–2120. PMLR (2016)
12. Karwa, V., Vadhan, S.: Finite sample differentially private confidence intervals. arXiv preprint arXiv:1711.03908 (2017)
13. Lindell, Y.: Secure multiparty computation for privacy preserving data mining. In: Encyclopedia of Data Warehousing and Mining, pp. 1005–1009. IGI Global (2005)
14. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map, pp. 277–286, April 2008. https://doi.org/10.1109/ICDE.2008.4497436
15. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map, pp. 277–286, April 2008. https://doi.org/10.1109/ICDE.2008.4497436
16. Raj, A., Law, H.C.L., Sejdinovic, D., Park, M.: A differentially private kernel two-sample test. In: Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C. (eds.) ECML PKDD 2019. LNCS (LNAI), vol. 11906, pp. 697–724. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46150-8_41
17. Rogers, R., Kifer, D.: A new class of private chi-square hypothesis tests. In: Singh, A., Zhu, J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 991–1000. PMLR, Fort Lauderdale, FL, USA 20–22 April 2017. http://proceedings.mlr.press/v54/rogers17a.html
18. Rogers, R., et al.: Linkedin's audience engagements api: A privacy preserving data analytics system at scale. arXiv preprint arXiv:2002.05839 (2020)
19. Sei, Y., Ohsuga, A.: Privacy-preserving chi-squared test of independence for small samples. BioData Mining **14**(1), 1–25 (2021)
20. Smith, A.: Privacy-preserving statistical estimation with optimal convergence rates. In: Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing, pp. 813–822 (2011)
21. Swanberg, M., Globus-Harris, I., Griffith, I., Ritz, A., Groce, A., Bray, A.: Improved differentially private analysis of variance. arXiv preprint arXiv:1903.00534 (2019)
22. Wang, Y., Lee, J., Kifer, D.: Revisiting differentially private hypothesis tests for categorical data. arXiv, Cryptography and Security (2015)