Robust Kernel Hypothesis Testing under Data Corruption

Antonin Schrab[⋆]
Centre for Artificial Intelligence
Gatsby Computational Neuroscience Unit
University College London & Inria London

Ilmun Kim[⋆]
Department of Statistics and Data Science
Department of Applied Statistics
Yonsei University

Abstract

We propose a general method for constructing robust permutation tests under data corruption. The proposed tests effectively control the non-asymptotic type I error under data corruption, and we prove their consistency in power under minimal conditions. This contributes to the practical deployment of hypothesis tests for real-world applications with potential adversarial attacks. For the two-sample and independence settings, we show that our kernel robust tests are minimax optimal, in the sense that they are guaranteed to be non-asymptotically powerful against alternatives uniformly separated from the null in the kernel MMD and HSIC metrics at some optimal rate (tight with matching lower bound). We point out that existing differentially private tests can be adapted to be robust to data corruption, and we demonstrate in experiments that our proposed tests achieve much higher power than these private tests. Finally, we provide publicly available implementations and empirically illustrate the practicality of our robust tests.

1 INTRODUCTION

One of the fundamental goals of statistical machine learning is to quantify departures from non-parametric distributional null hypotheses, with various applications in causal discovery, model calibration, clinical trials, *etc.* However, after collecting real-world data, practitioners rarely observe that the null holds. This can be explained by the inherently intricate process that data collection represents. In fact, *real data is messy*, often consisting of outliers and corrupted or wrong values. This illustrates that these nulls are often too strong to be observed in real-world applications, hence ultimately preventing practitioners from using such tests. To overcome these practical challenges, we consider a framework with relaxed null hypotheses for which distributional conditions only need to hold on a large proportion of the data (as opposed to the entire dataset), hence facilitating the deployment of robust tests in practice. Specifically, we focus on scenarios where some data points are arbitrarily corrupted by an adversary, and aim to develop methods for constructing robust tests that fulfill the following desiderata: the considered tests are (i) straightforward to implement, (ii) applicable to an important set of nonparametric testing problems, (iii) maintain non-asymptotic validity under data corruption, and (iv) attain minimax optimal power in certain settings. With this goal in mind, we first review some relevant work and highlight our key contributions.

Robust testing. There is an extensive body of work on robust testing developed under various contamination models. One popular framework is Huber's ϵ-contamination model (Huber, 1964) where the observed data are assumed to be drawn from a mixture distribution $(1-\epsilon)P+\epsilon G$ with P as the target distribution and G as the contamination distribution. Our work is concerned with a stronger adversarial contamination model where an adversary selects r observations out of n data points and replace them with arbitrary values. Under different contamination models, there has been a flurry of recent work on robust testing such as the two-point testing problem (Chen et al., 2016; Li et al., 2023), covariance testing (Diakonikolas and Kane, 2021), mean testing (Diakonikolas et al., 2017; George and Canonne, 2022; Canonne et al., 2023), identity testing (Acharya et al., 2021b). See Diakonikolas and Kane (2019) for a recent survey on robust statistics. We aim to advance the field by studying robust testing for exchangeability (see Sec-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s). [⋆]Equal contributions.

tion 2.1 for a formal setup), and establishing minimax testing rates in terms of the kernel metrics under data corruption. A distinct line of work explores a different notion of robustness, where the goal is to determine whether a given dataset lies within an uncertainty set centered around an empirical distribution with uncertainty defined using various metrics (Gao et al., 2018; Wang and Xie, 2022; Sun and Zou, 2021, 2022, 2023). This approach contrasts with our focus on worst-case contamination as it seeks to account for distributional ambiguity rather than adversarial data corruption.

Permutation tests. As mentioned earlier, our main interest is in designing robust tests for assessing the exchangeability of data under the null hypothesis. A gold standard method for testing exchangeability is the permutation test, which leverages the permutation-invariant properties of the data distribution under the null hypothesis. This ensures rigorous control of the type I error rate as demonstrated in Romano and Wolf (*e.g.*, 2005); Hemerik and Goeman (*e.g.*, 2018). When the data are arbitrarily corrupted, however, the exchangeability assumption no longer holds, and naively applying the permutation test using the corrupted data can potentially inflate the type I error rate. To address this miscalibration issue, it is essential to either modify the test statistic or adjust the permutation critical value under data corruption. This process inevitably compromises the power of the test, and the key challenge is then to balance a trade-off between robustness and power. We approach this problem under the minimax testing framework as in Albert et al. (2022); Schrab et al. (2023); Acharya et al. (2021b); Kim and Schrab (2023) and explain the fundamental limit of testing under data corruption.

Kernel MMD and HSIC. The Maximum Mean Discrepancy (MMD, Gretton et al., 2012a) and the Hilbert–Schmidt Independence Criterion (HSIC, Gretton et al., 2005a) are two prominent kernel-based measures for assessing homogeneity and dependence between two random quantities. Since their introduction, there have been significant developments in related topics, including time-efficient kernel tests (Gretton et al., 2012b; Zaremba et al., 2013; Zhao and Meng, 2015; Yamada et al., 2019; Schrab et al., 2022b; Domingo-Enrich et al., 2023), adaptive kernel selections (Schrab et al., 2023, 2022b,a; Hagrass et al., 2024; Domingo-Enrich et al., 2023; Biggs et al., 2024; Chatterjee and Bhattacharya, 2023) and minimax optimality of kernel testing (Albert et al., 2022; Li and Yuan, 2024; Schrab et al., 2023; Shekhar et al., 2022, 2023; Hagrass et al., 2024). However, prior work has predominantly focused on standard i.i.d. data without any perturbation, potentially limiting their applicability to real-world scenarios with anomalies or adversarial attacks. While there exist some studies on the robustness of MMD statistics in terms of estimation (Briol et al., 2019, Section 3.3 and Chérief-Abdellatif and Alquier, 2022, Section 3.2.3) and in terms of Bayesian inference (Chérief-Abdellatif and Alquier, 2020; Dellaporta et al., 2022; Dellaporta and Damoulas, 2023; Legramanti et al., 2025; Fazeli-Asl et al., 2024), there has been limited understanding of their performance in robust testing. To fill this gap, we consider the two-sample and independence testing problems as running examples of our general methods, and propose robust tests based on the MMD and HSIC statistics. We then show that the proposed tests guarantee rigorous control of testing errors even in the presence of data corruption, and achieve minimax optimality across all levels of data corruption.

Outline and summary. The remainder of this paper is organised as follows. In Section 2, we formally introduce our *testing under data corruption* framework, and propose our DC procedure (Algorithm 1) for constructing robust permutation tests from any statistic with finite sensitivity. We prove that the validity of this procedure (non-asymptotic control of the type I error even under data corruption) and present conditions for consistency (asymptotic control of the type II error). In Sections 3 and 4, we construct robust kernel tests for the two-sample and independence testing frameworks, respectively. For each robust test, we prove minimax optimality under data corruption in terms of the respective MMD and HSIC kernel metrics. In Section 5, we discuss related work on differential privacy which can be leveraged to construct robust tests. We empirically verify the validity of all robust tests in Section 6, and highlight, on corrupted synthetic and real-world data, the significantly higher power achieved by our DC tests compared to alternative private tests of Kim and Schrab (2023). Section 7 closes the paper with further discussions. We defer the technical proofs of the main results to the appendix.

Notation. Consider two ordered sets $\mathcal{X}_n := (X_1, \ldots, X_n)$ and $\mathcal{Y}_n := (Y_1, \ldots, Y_n)$. The Hamming distance $d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) := \sum_{i=1}^n \mathbb{1}(X_i \neq Y_i)$ is the number of indices for which the entries of the two ordered sets are different. We let $[n]$ denote the set $\{1, \ldots, n\}$ and $[n]_0 := \{0, \ldots, n\}$. We define $\boldsymbol{\Pi}_n$ to be the set of all permutations of $[n]$. Given a permutation $\boldsymbol{\pi} \in \boldsymbol{\Pi}_n$, we denote by $\mathcal{X}_n^{\boldsymbol{\pi}}$ the permuted ordered set $(X_{\boldsymbol{\pi}(1)}, \ldots, X_{\boldsymbol{\pi}(n)})$. For two sequences of real numbers a_n and b_n, we write $a_n \lesssim b_n$ if $a_n \leq C b_n$ for some constant $C > 0$, and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The global sensitivity Δ_T of a statistic T is defined as the maximum difference in the statistic's output when evaluated on permuted datasets whose entries differ by at most one, that is

$$\Delta_T \coloneqq \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n,\mathcal{Y}_n\,:\,d_{\mathrm{ham}}(\mathcal{X}_n,\mathcal{Y}_n)\le 1} \left|T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}})\right|.$$

Given values $M_0,\ldots,M_B$, the $(1-\alpha)$-quantile of the M_i's is defined as $q_{1-\alpha} \coloneqq \inf\big\{t \in \mathbb{R} : \frac{1}{B+1}\sum_{i=0}^{B} \mathbb{1}(M_i \le t) \ge 1-\alpha\big\}$. We let $\mathsf{Laplace}(0,1)$ denote a Laplace distribution with location and scale parameters $(0,1)$. We also let $\mathsf{Gaussian}(\mu,\sigma,d)$ denote a d-dimensional distribution with each dimension being independent and following a Gaussian distribution with mean μ and standard deviation σ. Similarly, $\mathsf{Geometric}(p,d)$ denotes a d-dimensional distribution with each dimension being independent and following a geometric distribution taking values in $\{0,1,2,\ldots\}$ with parameter $p\in(0,1)$.

2 CONSTRUCTION OF HYPOTHESIS TESTS ROBUST TO DATA CORRUPTION

In Section 2.1, we set the stage and formalise our framework for robust testing in the present of data corruption. In Section 2.2, we introduce a novel procedure (DC) that enables the construction of robust tests under data corruption using any test statistic with finite sensitivity.

2.1 Robust testing framework under data corruption

Standard testing framework. Consider a set of distributions $\mathcal{P}$ partitioned into disjoint subsets $\mathcal{P}_0$ and $\mathcal{P}_1$. Given a set $\widetilde{\mathcal{X}}_n$ consisting of n random samples drawn i.i.d. from $P\in\mathcal{P}$, the aim of hypothesis testing is to test whether the null $\mathcal{H}_0\colon P\in\mathcal{P}_0$, or the alternative $\mathcal{H}_1\colon P\in\mathcal{P}_1$, holds.

Testing under data corruption framework. In this setting, we do not have access to $\widetilde{\mathcal{X}}_n$ but only to a corrupted version of it, denoted simply by $\mathcal{X}_n$, where up to r samples of $\widetilde{\mathcal{X}}_n$ might have been corrupted (possibly in an adversarial manner). We receive only the set $\mathcal{X}_n$ of size n with no knowledge of which samples, if any, have been corrupted. The aim is still to test whether $\mathcal{H}_0\colon P\in\mathcal{P}_0$ or $\mathcal{H}_1\colon P\in\mathcal{P}_1$, but we can only assume $n-r$ samples of $\mathcal{X}_n$ are actually drawn from $P\in\mathcal{P}$ (as r samples might have been manipulated).

Robustness. Hypothesis tests designed for this setting are *robust* to data corruption: under the null, manipulating up to r samples will not make the test deviate from the null. This can also be thought of as enlarging the null hypothesis as it only needs to hold for at least $n-r$ samples of the data rather than for all n samples. The number r of maximum sample manipulations to be robust to is specified by the user depending on the application. If $r=0$, we recover the standard testing framework. As r increases, the test becomes more robust but less powerful (*i.e.*, there is a trade-off between robustness and power). If $r=n$, the null would never be rejected.

Exchangeability. We restrict our attention to testing frameworks under which the exchangeability assumption holds (Lehmann and Romano, 2005, Chapter 15.2) in the non-corrupted setting, which means that, under the null, $\widetilde{\mathcal{X}}_n$ is exchangeable: for any permutation $\boldsymbol{\pi}\in\boldsymbol{\Pi}_n$, the joint distributions of $\widetilde{\mathcal{X}}_n$, and of $\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}$, are the same. For example, the two-sample and independence testing frameworks, serving as primary applications of our proposal, satisfy this assumption.

2.2 DC procedure: Robust test construction against data corruption

We assume the maximum number of corrupted samples r to be fixed apriori. The aim is then to construct a robust test that is *valid* in the sense that it controls the type I error at the desired level α even when up to r samples are arbitrarily corrupted.

We now propose a procedure to construct tests which are robust to data corruption. First, recall that, given a statistic T, the classical permutation test is defined as rejecting the null when $T_0 > q_{1-\alpha}(T_0\ldots,T_B)$ where $T_0 = T(\mathcal{X}_n)$ and $T_i = T(\mathcal{X}_n^{\boldsymbol{\pi}_i})$, $i\in[B]$ for B permutations $\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_B\in\boldsymbol{\Pi}_n$. Such a test in the usual setting (no data corruption) controls the type I error at level α non-asymptotically (*e.g.*, Hemerik and Goeman, 2018). For our setting with at most r corrupted samples, we instead define our DC test as rejecting the null when $T_0 > q_{1-\alpha}(T_0\ldots,T_B) + 2r\Delta_T$ (see details in Algorithm 1), where Δ_T is the global sensitivity of T. We prove that this results in a well-calibrated non-asymptotic test under r data corruption, and provide sufficient conditions to guarantee its consistency.

Algorithm 1 Robust DC procedure

```
Inputs: Data $\mathcal{X}_n$, robustness $r$, level $\alpha$, statistic $T$,
permutation number $B$.

Generate i.i.d. permutations $\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_B$ of $[n]$.
Set $\boldsymbol{\pi}_0 = \mathrm{Id}$ and compute global sensitivity $\Delta_T$.
Compute $T_i = T(\mathcal{X}_n^{\boldsymbol{\pi}_i})$, $i\in[B]_0$.
Compute $(1-\alpha)$-quantile $q$ of $T_0,\ldots,T_B$.

Output: Reject $\mathcal{H}_0$ if $T_0 > q + 2r\Delta_T$.
```

Lemma 1 (Validity & consistency of DC). *(i) The DC test of Algorithm 1 has non-asymptotic level control under r data corruption. (ii) Let $P_1 \in \mathcal{P}_1$ be an alternative distribution. Assume $\alpha \in (0,1)$ fixed. For any sequence of B_n of permutation numbers satisfying $\min_{n\in\mathbb{N}} B_n > \alpha^{-1} - 1$, the DC test is consistent in the sense that $\lim_{n\to\infty} \mathbb{P}_{P_1}\big(\text{DC rejects } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1$ if*

$$\lim_{n\to\infty} \mathbb{P}_{P_1}\big(\, T(\widetilde{\mathcal{X}}_n) \;>\; T(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}) + 4r\Delta_T \big) = 1,$$

where the probability is taken with respect to the (uniformly) random permutation $\boldsymbol{\pi}$ of $[n]$, and to $\widetilde{\mathcal{X}}_n$ i.i.d. drawn from P_1.

We highlight that Lemma 1 is applicable to any test statistic with finite global sensitivity Δ_T. Moreover, the validity under data corruption holds in any testing framework for which exchangeability holds under the (uncorrupted) null hypothesis. Finally, we stress that knowledge of the maximum number of potential corruptions r is sufficient for all theoretical guarantees in this paper, including Lemma 1, to hold; knowing the exact number of corrupted samples is not required.

We next turn to two-sample and independence kernel testing as specific applications of these general methods, and present more refined results.

3 TWO-SAMPLE KERNEL TESTING ROBUST TO DATA CORRUPTION

Robust two-sample testing. We now consider the two-sample testing framework which, in the non-corrupted setting, satisfies the exchangeability assumption under the null (*e.g.*, Schrab et al., 2023, Proposition 1). In our robust setting, we are given samples $Y_1, \dots, Y_n$ and $Z_1, \dots, Z_m$, where at most r of the $m+n$ samples have been corrupted (we do not know which ones, if any). Of the remaining samples (at least $m+n-r$ of them), the Y_i's are drawn from P while the Z_i's are drawn from Q. We are interested in testing whether $P = Q$. The data corruption could potentially be applied in an adversarial way; for example, some of the Y_i's samples could be replaced by new samples from Q. As before, the maximum number of corrupted samples r is specified by the user and is considered to be known. We restrict ourselves to the case $r \le n = \min(m,n)$, as the setting $r > n$ suffers from the same issue as when $r = n$: all the samples from P could be corrupted, in which case all information about P would be lost and we would not be able to test whether $P = Q$.

MMD statistic. As a divergence for this two-sample problem, we use the kernel MMD introduced by Gretton et al. (2012a) which, for a kernel k, is defined as

$$\begin{aligned}&\mathrm{MMD}_k(P,Q)\\ :=\;&\sqrt{\mathbb{E}_P[k(Y,Y')] - 2\mathbb{E}_{P,Q}[k(Y,Z)] + \mathbb{E}_Q[k(Z,Z')]}\end{aligned}$$

for probability distributions P and Q. The MMD is well-suited for the two-sample problem as it can distinguish between any two distributions in the sense that $\mathrm{MMD}_k(P,Q) = 0$ if and only if $P = Q$, given that the kernel k is characteristic (Fukumizu et al., 2008). For a given sample $\mathcal{X}_{n+m} := (Y_1, \dots, Y_n, Z_1, \dots, Z_m)$ generated as described above, the quadratic-time empirical MMD statistic is

$$\begin{aligned}&\widehat{\mathrm{MMD}}(\mathcal{X}_{n+m})\\ :=\;&\Bigg(\frac{1}{n^2}\sum_{i,j=1}^{n} k(Y_i,Y_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(Z_i,Z_j)\\ &\qquad - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(Y_i,Z_j)\Bigg)^{1/2}\end{aligned}$$

which is a plug-in estimator, and its square can be expressed as a two-sample V-statistic. For a kernel bounded everywhere by K, Kim and Schrab (2023, Lemma 5) provide an upper bound on the global sensitivity of this MMD statistic $\Delta_{\widehat{\mathrm{MMD}}} \le \sqrt{2K}/\min(n,m)$, which is guaranteed to be tight when the kernel k is translation invariant.

dcMMD. We construct our robust two-sample dcMMD test by applying the procedure of Algorithm 1 with the MMD statistic $T = \widehat{\mathrm{MMD}}$ and global sensitivity $\Delta_T = \sqrt{2K}/\min(n,m)$. It is immediately clear from Lemma 1 that the robust dcMMD test controls the type I error at the desired level α non-asymptotically, even under r data corruption. We next prove its consistency in the data corruption framework, which ensures that any fixed distributions P and Q with $P \ne Q$ can be distinguished with probability one by the dcMMD test for large enough sample sizes. This test scales quadratically with the sample sizes.

Lemma 2 (Consistency of dcMMD). *Suppose that the kernel k is characteristic, non-negative, and bounded everywhere by K. Assume that $n \le m$ and $r/n \to 0$ as $n \to \infty$, and that a sequence of permutation numbers B_n satisfies $\min_{n\in\mathbb{N}} B_n > \alpha^{-1} - 1$. Then, for any fixed P and Q with $P \ne Q$, dcMMD is consistent in the sense that*

$$\lim_{n\to\infty} \mathbb{P}_{P,Q}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1.$$

The proof of Lemma 2 relies on proving that the sufficient condition of Lemma 1 is satisfied. We show that this condition is met for the proposed dcMMD test with characteristic and bounded kernel. Having obtained guarantees for the asymptotic power of dcMMD against fixed alternatives, we now consider the more challenging task of providing power guarantees that hold uniformly over classes of alternatives shrinking towards the null as the sample sizes increase. We choose these classes to be separated from the null in the MMD metric, and prove that dcMMD achieves the optimal uniform separation rate (with matching upper and lower bounds). This minimax result holds with respect to the smallest sample size, to the data corruption level, and to the quantities controlling the type I and II errors; hence, demonstrating the optimality of dcMMD in the data corruption setting.

Theorem 1 (Minimax optimal uniform separation of dcMMD)**.** *Suppose that the kernel k is characteristic, non-negative, and bounded everywhere by K. For $\alpha, \beta \in (0,1)$, assume that the number of permutations B is greater than $3\alpha^{-2}\{\log(8/\beta)+\alpha(1-\alpha)\}$, and that $n \asymp m$ with $n \leq m$.*

(i) (Uniform separation) The dcMMD test is guaranteed to have high power, i.e., $\mathbb{P}_{P,Q}(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}) \geq 1-\beta$ for any distributions P and Q separated as

$$\mathrm{MMD}_k(P,Q) \geq C_K \max\left\{\sqrt{\frac{\max\{\log(e/\alpha),\log(e/\beta)\}}{n}},\ \frac{r}{n}\right\}$$

for some positive constant C_K depending on K.

(ii) (Minimax optimality) Further assuming that the kernel is translation invariant and non-constant as in Assumption 1 in the appendix, this separation rate is optimal in terms of the smallest sample size n, of the data corruption level r, and of the testing errors α, β with $\alpha \asymp \beta$ and $\alpha+\beta < 0.4$.

Theorem 1 demonstrates that the dcMMD test achieves minimax rate optimality in terms of the MMD metric. Further remarks regarding Theorem 1 are in order, which we present below.

Low corruption regime. We first note that when $r \lesssim \sqrt{n\max\{\log(e/\alpha),\log(e/\beta)\}}$, the dominating term in the condition of Theorem 1(i) is the first one, which is the minimax rate in the usual no-corruption setting (Kim and Schrab, 2023, Theorem 8). In this low corruption regime, the two robust dcMMD tests achieve the same minimax optimal rate as the standard MMD test in the non-corrupted setting. This demonstrates that we can obtain robustness against corruption of up to r samples, where $r \lesssim \sqrt{n\max\{\log(e/\alpha),\log(e/\beta)\}}$, without compromising power in terms of uniform separation rate.

High corruption regime. When we have $\sqrt{n\max\{\log(e/\alpha),\log(e/\beta)\}} \lesssim r < n$, our robust dcMMD test achieves high power when MMD_k exceeds r/n, which is the optimal rate in this high corruption regime as we have proved in Theorem 1(ii). It is worth pointing out that the separation rate r/n in the high corruption regime is independent of α and β, unlike the one obtained in the low corruption regime. However, we emphasise that both α and β cannot be taken to be arbitrarily small as this would break the condition $\sqrt{n\max\{\log(e/\alpha),\log(e/\beta)\}} \lesssim r$.

Total corruption regime. When $r = n$ (or $r \geq n$), the separation rate becomes an arbitrary constant. This makes the condition for MMD_k in Theorem 1(i) vacuous since MMD_k is bounded above by $\sqrt{2K}$ for any distributions P and Q. In fact, in this regime, all samples from P could be adversarially replaced by samples from Q, so there is no hope of distinguishing between the two distributions.

4 INDEPENDENCE KERNEL TESTING ROBUST TO DATA CORRUPTION

Robust independence testing. Another testing framework satisfying the exchangeability assumption under the null is the independence testing one (*e.g.*, Albert et al., 2022, Proposition 1). The robust independence testing problem is defined as follows. Given paired samples $\mathcal{X}_n = \big((Y_i,Z_i)\big)_{i=1}^n$ where at most $r \in [n]$ of them have been corrupted (no knowledge of which ones are corrupted, if any), and the remaining paired samples (at least $n-r$ of them) are drawn from some joint distribution P_{YZ}. We test whether that joint distribution P_{YZ} is equal to the product of its marginals $P_Y \times P_Z$, that is, we test for independence among the non-corrupted paired samples. Again, the maximum number of corrupted samples r is considered known, usually specified by the user.

HSIC statistic. As a measure of dependence, we use the HSIC (Gretton et al., 2005b) which, for kernels k and ℓ, is defined as

$$\mathrm{HSIC}_{k,\ell}(P_{YZ}) := \Big(\mathbb{E}_{P_{YZ}}\big[k_{Y,Y'}\ell_{Z,Z'}\big] - 2\mathbb{E}_{P_{YZ}}\big[\mathbb{E}_{P_Y}[k_{Y,Y'}]\mathbb{E}_{P_Z}[\ell_{Z,Z'}]\big] + \mathbb{E}_{P_Y}\big[k_{Y,Y'}\big]\mathbb{E}_{P_Z}\big[\ell_{Z,Z'}\big]\Big)^{1/2}$$

with the condensed notation $k_{Y,Y'}$ and $\ell_{Z,Z'}$ for $k(Y,Y')$ and $\ell(Z,Z')$. Provided that both kernels are characteristic (Gretton, 2015), the HSIC characterises dependence in the sense that $\mathrm{HSIC}_{k,\ell}(P_{YZ}) = 0$ if and only if $P_{YZ} = P_Y \times P_Z$; hence, it is suitable as a building block for independence testing. Given paired samples $\mathcal{X}_n = \big((Y_i, Z_i)\big)_{i=1}^n$ generated by the process described above, the empirical HSIC statistic is

$$\widehat{\mathrm{HSIC}}(\mathcal{X}_n) \;\coloneqq\; \Bigg(\frac{1}{n^2}\sum_{i,j=1}^{n} k_{i,j}\ell_{i,j} - \frac{2}{n^3}\sum_{i,j_1,j_2=1}^{n} k_{i,j_1}\ell_{i,j_2} + \frac{1}{n^4}\sum_{i_1,i_2,j_1,j_2=1}^{n} k_{i_1,j_1}\ell_{i_2,j_2}\Bigg)^{1/2}$$

where $k_{i,j}$ and $\ell_{i,j}$ denote $k(Y_i, Y_j)$ and $\ell(Z_i, Z_j)$ for $i, j \in [n]$. This quadratic-time plug-in estimator can be written as a one-sample fourth-order V-statistic. For kernels k and ℓ bounded everywhere by K and L, respectively, the global sensitivity of the HSIC statistic is bounded as $\Delta_{\widehat{\mathrm{HSIC}}} \leq 4\sqrt{KL}(n-1)/n^2$, which is asymptotically tight as shown by Kim and Schrab (2023, Lemma 6).

dcHSIC. Under the data corruption setting, the robust dcHSIC is constructed by applying the DC procedure of Algorithm 1 with the HSIC statistic $T = \widehat{\mathrm{HSIC}}$ and global sensitivity Δ_T as $4\sqrt{KL}(n-1)/n^2$. The validity of dcHSIC is guaranteed by Lemma 1 with non-asymptotic type I error control at the desired level α under r data corruption. The next lemma proves that the dcHSIC test is consistent against any fixed alternative P_{YZ} with $P_{YZ} \neq P_Y \times P_Z$. In other words, this test can detect any fixed dependence between Y and Z with probability one for large enough sample size. The dcHSIC test scales quadratically with n.

Lemma 3 (Consistency of dcHSIC)**.** *Suppose that the kernels k and ℓ are characteristic, non-negative, and bounded everywhere by K and L, respectively. Assume that $r/n \to 0$ as $n \to \infty$, and that a sequence of permutation numbers B_n satisfies $\min_{n\in\mathbb{N}} B_n > \alpha^{-1} - 1$. Then, for any fixed joint distribution P_{YZ} with $P_{YZ} \neq P_Y \times P_Z$, dcHSIC is consistent in the sense that*

$$\lim_{n\to\infty} \mathbb{P}_{P_{YZ}}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1.$$

Similarly to the robust two-sample testing setting, we can also obtain power guarantees for dcHSIC in terms of uniform separation rates. More precisely, the type II errors of dcHSIC can be controlled non-asymptotically whenever $\mathrm{HSIC}_{k,\ell}(P_{YZ})$ is larger than the minimax rate (which tends to zero as the sample size grows), highlighting the optimality of dcHSIC under the data corruption framework.

Theorem 2 (Minimax optimal uniform separation of dcHSIC)**.** *Suppose that the kernels k and ℓ are characteristic, non-negative, and bounded everywhere by K and L, respectively. For $\alpha, \beta \in (0,1)$, assume that the number of permutations B is greater than $\alpha^{-2}\big\{\log(8/\beta) + \alpha(1-\alpha)\big\}$.*

(i) (Uniform separation) The dcHSIC test is guaranteed to have high power, i.e., $\mathbb{P}_{P_{YZ}}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) \geq 1 - \beta$ for any joint distribution P_{YZ} separated as

$$\mathrm{HSIC}_{k,\ell}(P_{YZ}) \geq C_{K,L} \max\Bigg\{\sqrt{\frac{\max\{\log(e/\alpha), \log(e/\beta)\}}{n}},\ \frac{r}{n}\Bigg\}$$

for some positive constant $C_{K,L}$ depending on K and L.

(ii) (Minimax optimality) Further assuming that the kernels k and ℓ are translation invariant and non-constant as in Assumption 2 in the appendix, this separation rate is optimal in terms of the sample size n, of the data corruption level r, and of the testing errors α, β with $\alpha \asymp \beta$ and $\alpha + \beta < 0.4$.

The interpretation of the independence uniform separation rate of Theorem 2 mirrors the one following Theorem 1 for the two-sample case. In the low corruption regime with $r \lesssim \sqrt{n \max\{\log(e/\alpha), \log(e/\beta)\}}$, robustness comes for free as the minimax rate is the same as in the non-corrupted setting. In the high corruption regime with $\sqrt{n \max\{\log(e/\alpha), \log(e/\beta)\}} \lesssim r < n$, only the alternatives for which $\mathrm{HSIC}_{k,\ell}(P_{YZ})$ exceeds r/n can be detected with high power, and this rate is optimal with respect to both n and r by Theorem 2(ii). In the total corruption regime where $r = n$, the uniform separation condition is never satisfied and any test is non-informative as all information is lost. Finally, we highlight again that our dcHSIC test is minimax optimal across all regimes of data corruption.

5 RELATED WORK: ROBUSTNESS VIA DIFFERENTIAL PRIVACY

Robust tests can also be constructed from existing differentially private tests. In Section 5.1, we present background information on differential privacy. In Section 5.2, we construct robust tests by leveraging the existing differential privatisation (DP) method of Kim and Schrab (2023) for permutation tests. We prove their validity and defer its power analysis to Appendix D.

5.1 Differential privacy background

Differential privacy (Dwork et al., 2006) is a popular framework for private data analysis that aims to return an output insensitive to any single individual's data point. As discussed in Dwork and Lei (2009); Avella-Medina (2020), there is a close connection between robust statistics and differential privacy. Both frameworks require that algorithms remain stable under data perturbation. Recent research has delved into extending non-private tests to privacy settings and studying their asymptotic properties (Fienberg et al., 2011; Couch et al., 2019; Campbell et al., 2018; Rogers and Kifer, 2017; Raj et al., 2020) and non-asymptotic properties (Cai et al., 2017; Aliakbarpour et al., 2018, 2019; Acharya et al., 2019; Kim and Schrab, 2023).

5.2 DP procedure: Robust test construction via differentially privacy

Differential privacy (DP) is closely related to statistical robustness (Dwork and Lei, 2009; Avella-Medina, 2020). This connection naturally motivates us to leverage the recent advances in differential privatisation of hypothesis tests (Peña and Barrientos, 2022; Kazan et al., 2023; Kim and Schrab, 2023) for our data corruption setting. Consider any valid ε-DP hypothesis test with adjusted lower level $\alpha e^{-r\varepsilon}$ and pure DP parameter $\varepsilon > 0$. Under the null, the DP group privacy property (Dwork et al., 2014, Theorem 2.2) ensures that

$$\begin{aligned} & \sup_{P_0 \in \mathcal{P}_0} \mathbb{P}_{P_0}\big(\text{reject } \mathcal{H}_0 \mid r \text{ corrupted data}\big) \\ \leq \;& e^{r\varepsilon} \sup_{P_0 \in \mathcal{P}_0} \mathbb{P}_{P_0}\big(\text{reject } \mathcal{H}_0 \mid \text{uncorrupted data}\big) \quad (1) \\ \leq \;& \alpha. \end{aligned}$$

This guarantees that any valid ε-DP test with adjusted level $\alpha e^{-r\varepsilon}$ is well-calibrated under data corruption in the sense that its type I error is controlled by α.

We choose to focus on the differential privatisation approach of Kim and Schrab (2023) for permutation tests, as it benefits from stronger theoretical guarantees. In particular, their resulting DP tests are guaranteed to be valid at every sample size (Kim and Schrab, 2023, Theorem 1), so type I error control under data corruption of Equation (1) holds non-asymptotically. This is actually crucial as the corruption and privacy parameters r and ε usually depend on the sample size. Consequently, the adjusted level $\alpha e^{-r\varepsilon}$ also depends on the sample size, which can modify the asymptotic behaviour of type I error control. We simply refer to this test construction based on Kim and Schrab (2023) with the adjusted level as the DP procedure (see details in Algorithm 2).

The consistency of the DP test against data corruption can be guaranteed (*i.e.*, for a fixed alternative distribution, the test power converges to one asymptotically) under minimal conditions. Such a result is presented in Appendix D, alongside with a power separation analysis of the two-sample dpMMD and independence dpHSIC tests, constructed from Algorithm 2.

Algorithm 2 Robust DP procedure
Adapted from Kim and Schrab, 2023

```
Inputs: Data $\mathcal{X}_n$, robustness $r$, level $\alpha$, statistic
    $T$, permutation number $B$, privacy $\varepsilon$.
Generate i.i.d. $\zeta_0, \dots, \zeta_B \sim \mathsf{Laplace}(0,1)$.
Generate i.i.d. permutations $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_B$ of $[n]$.
Set $\boldsymbol{\pi}_0 = \mathrm{Id}$ and compute global sensitivity $\Delta_T$.
Compute $M_i = T(\mathcal{X}_n^{\boldsymbol{\pi}_i}) + 2\zeta_i \Delta_T \varepsilon^{-1}$, $i \in [B]_0$.
Compute $(1-\alpha e^{-r\varepsilon})$-quantile $q$ of $M_0, \dots, M_B$.
Output: Reject $\mathcal{H}_0$ if $M_0 > q$.
```

6 EXPERIMENTS

In this section, we provide simulation results for two-sample and independence testing on both synthetic and real-world data, focusing on the robustness properties of the designed tests. Our initial focus is to demonstrate that the robust tests control the type I error whenever r or fewer entries have been corrupted, a task which the non-robust tests fail to accomplish. Assuming this control, the aim is then to have test power, *e.g.*, to detect when strictly more than r samples have been corrupted.

We stress that the DC and DP methods (Algorithms 1 and 2) are calibrated to control the type I error even under the most severe corruption scenarios. Under the strongest c-corruption scenario where c denotes the number of corrupted data points, we expect the rejection rate of the robust tests with robustness parameter r to be exactly at α when $c = r$ (as the bounds in Equations (1) and (2) are tight in this setting). We also expect test power to start increasing as soon as $c > r$. In practice, and in our experiments, however, we typically encounter milder corruption scenarios under which the robust tests are conservative for $c \leq r$, resulting in a type I error substantially smaller than α. This implies a delay in the observed increase of test power, which sometimes happens only for values of c much larger than r.

In the two-sample setting, we compare the rejection rates achieved by our proposed robust dcMMD test against the dpMMD test adapted from Kim and Schrab (2023), as well as by the non-robust permutation-based MMD test. Similarly, we compare the performance of all three dcHSIC, dpHSIC and HSIC tests for the independence testing problem. All tests are run using Gaussian kernels and 500 permutations. For both frameworks, we consider synthetic Gaussian simulations (mean shift and mixture), and experiments based on the real-world IMDb 'Large Movie Review Dataset' of Maas et al. (2011) (see captions of Figures 1 and 2 for experimental details, rejec-

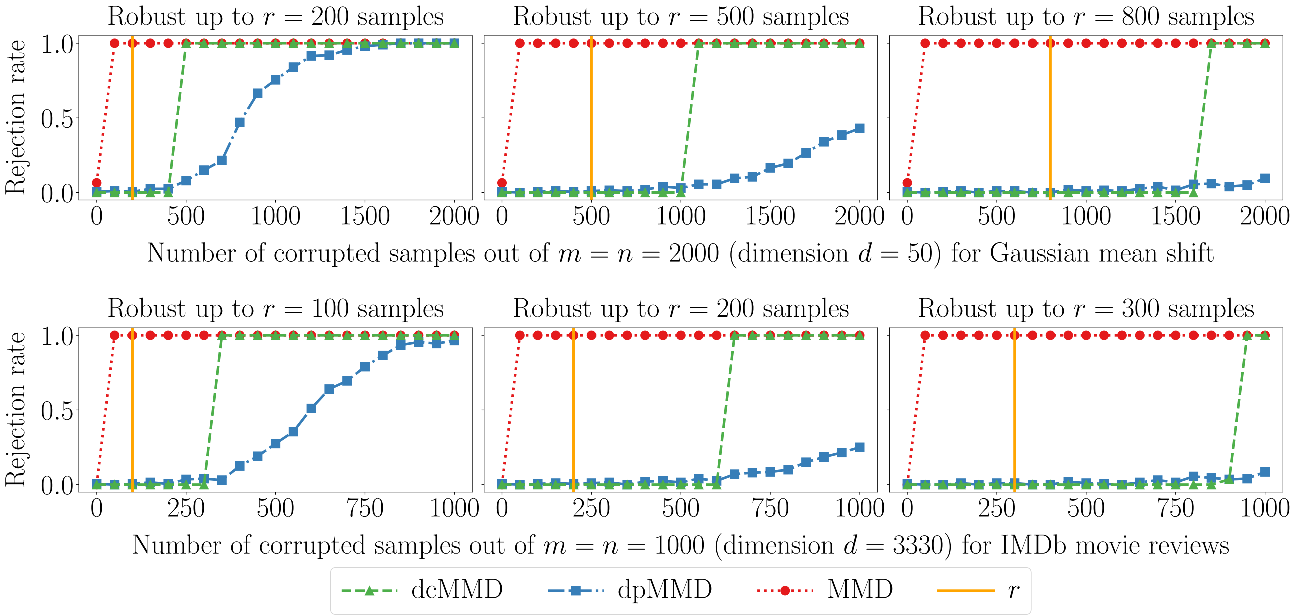

Figure 1: Two-sample experiments robust up to r corrupted samples. To have valid level, a robust test needs to control the rejection rate by $\alpha = 0.05$ when fewer than r samples are corrupted. To be powerful, the robust test needs to have a high rejection rate when more than r samples are corrupted. *(Top row: Gaussian mean shift)* Both samples are originally i.i.d. drawn from $\mathsf{Gaussian}(0, 1/10, 50)$, entries of one sample are corrupted being replaced by samples from $\mathsf{Gaussian}(1000, 1/10, 50)$. *(Bottom row: IMDb movie reviews)* Both samples originally consist of movie reviews (using a bag of 3330 words representation). Corrupted entries for one sample are replaced by samples from $\mathsf{Geometric}(0.05, 3330)$.

tion rates plotted are averaged over 200 repetitions).

The same trends are observed across all experiments of Figures 1 and 2. First, we note that all tests control the type I error at $\alpha = 0.05$ in the uncorrupted setting. As soon as some samples are corrupted, the MMD and HSIC tests achieve rejection rate 1, *i.e.*, these tests are not robust to data corruption. In contrast, we observe that the DC and DP tests control the rejection rate to be lower than α when at most r samples are corrupted, as theoretically guaranteed by Lemmas 1 and 4.

In the alternative regime with more than r corrupted samples, the DC and DP tests behave differently, with our proposed DC tests clearly always achieving higher power than the DP tests adapted from Kim and Schrab (2023). The DC rejection rate rapidly increases to 1 after some threshold is reached in the number of corrupted samples, which is a desired property: the rejection rate is controlled to be lower than α when robustness is needed, and the power becomes 1 when robustness is no longer required. We observe that the DP rejection rate starts increasing around the same threshold as for DC, but, unlike DC, its growth is slow and gradual. Overall, while both procedures control the type I error under data corruption, the DC tests are much more powerful than the DP tests.

As aforementioned, the fact that this threshold can sometimes be larger than r can be explained by the design of the tests to be robust to any r corruption processes, including stronger ones than those used in these experiments. Nonetheless, we highlight the strong performance of the DC tests. For example, dcMMD is able to detect differences in distributions using only 2000 samples while being robust to the corruption of up to 800 of these samples (see Figure 1), which is remarkable.

We note that the experiments in Figures 1 and 2 are presented as the null hypothesis holding and being corrupted towards some alternative hypothesis (*e.g.*, reading the plots from left to right). They can equivalently be interpreted as some alternative hypothesis holding and being corrupted towards the null hypothesis (*e.g.*, reading the plots from right to left).

7 CONCLUSION

In this paper, we introduced a general approach for constructing robust permutation tests under data corruption. This method, called DC procedure and presented in Algorithm 1, avoids the use of random noise injection (required for the DP procedure of Kim and Schrab (2023) presented in Algorithm 2), making it more suitable for settings where reproducibility is critical. By employing the permutation principle, under the exchangeability assumption, both approaches ensure non-asymptotic validity under r data corruption.

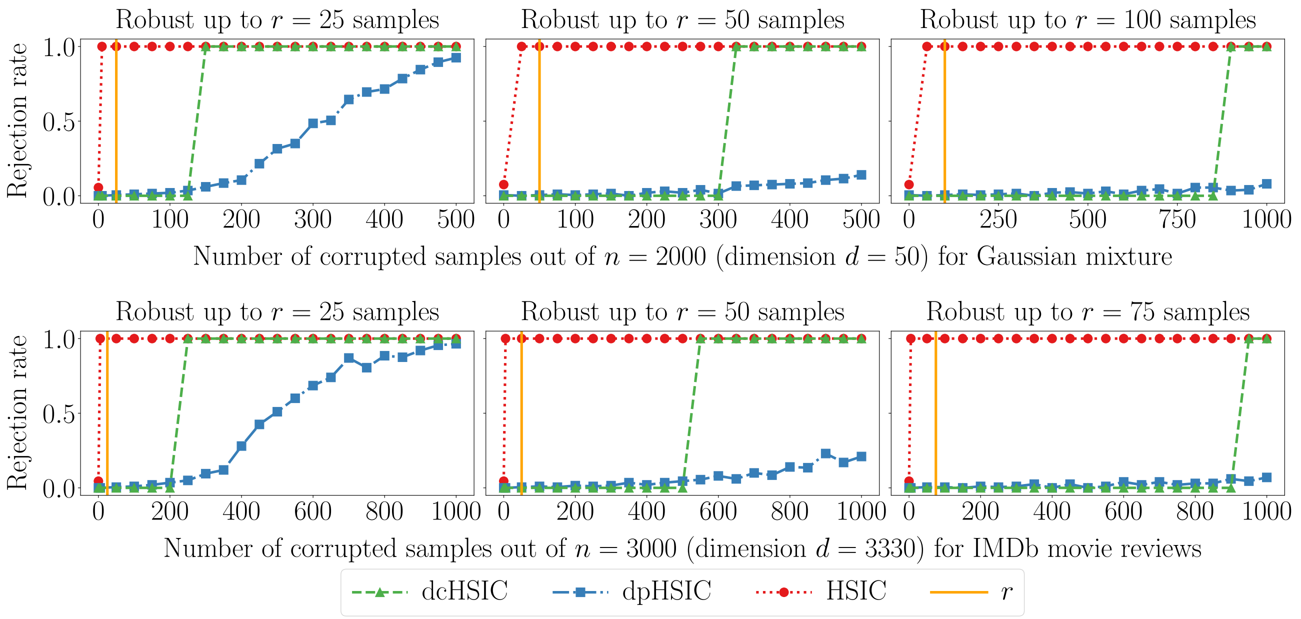

Figure 2: Independence experiments robust up to r corrupted samples. To have valid level, a robust test needs to control the rejection rate by $\alpha = 0.05$ when fewer than r samples are corrupted. To be powerful, the robust test needs to have a high rejection rate when more than r samples are corrupted. *(Top row: Gaussian mixture)* Paired samples (X, Y) are originally i.i.d. drawn from two $\mathsf{Gaussian}(0, {}^1\!/\!_{10}, 50)$. Corrupted samples are replaced by $(X, X + \epsilon)$ where $\epsilon \sim \mathsf{Gaussian}(0, {}^1\!/\!_{10}, 50)$ and where $X \sim \mathsf{Gaussian}(s1000, {}^1\!/\!_{10}, 50)$ with $s = 1$ for half of the corrupted samples and $s = -1$ for the other half. *(Bottom row: IMDb movie reviews)* Paired samples (X, Y) originally consist of two independent reviews (represented using a bag of 3330 words). Corrupted samples are replaced by $(X + s, X + s + \epsilon)$ where $X \sim \mathsf{Geometric}(0.05, 3330)$, $\epsilon \sim \mathsf{Gaussian}(0, {}^1\!/\!_{10}, 3330)$ and with $s = 0$ for half of the corrupted samples and $s = 5$ for the other half.

In terms of power guarantees, we established the minimal conditions under which the DC test maintains consistency under data corruption against any fixed alternatives. We illustrated our general frameworks in the context of kernel two-sample and independence testing, and showed that our kernel robust DC tests achieve minimax optimal power in terms of the MMD and HSIC metrics. Additionally, we demonstrated that in low corruption regimes, the robustness property can be attained without compromising the minimum separation rate in terms of power. Empirical results were presented to illustrate the finite-sample performance of our robust DC tests which significantly outperform the DP tests in terms of test power.

Limitations and future work. Our work opens up several fruitful directions for future research. As illustrated in our empirical studies, the proposed methods might be conservative in certain settings, as they are designed to safeguard testing errors against worst-case attack scenarios. It would be interesting to consider milder and potentially more structured attack scenarios (*e.g.*, Huber's ϵ-contamination model), and to refine our DC procedure to mitigate this conservative nature. Furthermore, expanding our methods to other types of test statistics and different testing problems is worth exploring. One can also attempt to develop computationally efficient robust tests by exploiting recent advancements in time-efficient kernel tests (*e.g.*, Domingo-Enrich et al., 2023; Schrab et al., 2022b). Lastly, studying minimax rates for robust testing in terms of other metrics (*e.g.*, Wasserstein and L_p metrics) presents an interesting avenue for future work.

Acknowledgments

Antonin Schrab acknowledges support from the U.K. Research and Innovation (grant EP/S021566/1). Ilmun Kim acknowledges support from the Basic Science Research Program through the National Research Foundation of Korea (2022R1A4A1033384), and the Korea government (MSIT) (RS-2023-00211073).

References

Acharya, J., Canonne, C., Freitag, C., and Tyagi, H. (2019). Test without trust: Optimal locally private distribution testing. In *The 22nd International Conference on Artificial Intelligence and Statistics*.

Acharya, J., Sun, Z., and Zhang, H. (2021a). Differentially Private Assouad, Fano, and Le Cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR.

Acharya, J., Sun, Z., and Zhang, H. (2021b). Robust Testing and Estimation under Manipulation

Attacks. In *International Conference on Machine Learning*, pages 43–53. PMLR.

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022). Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858–879.

Aliakbarpour, M., Diakonikolas, I., Kane, D., and Rubinfeld, R. (2019). Private Testing of Distributions via Sample Permutations. *Advances in Neural Information Processing Systems*, 32.

Aliakbarpour, M., Diakonikolas, I., and Rubinfeld, R. (2018). Differentially Private Identity and Closeness Testing of Discrete Distributions. *In International Conference on Machine Learning*, pages 169–178.

Avella-Medina, M. (2020). The Role of Robust Statistics in Private Data Analysis. *CHANCE*, 33(4):37–42.

Biggs, F., Schrab, A., and Gretton, A. (2024). MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting. *Advances in Neural Information Processing Systems*, 36.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019). Statistical Inference for Generative Models with Maximum Mean Discrepancy. *arXiv preprint arXiv:1906.05944*.

Cai, B., Daskalakis, C., and Kamath, G. (2017). Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning*, pages 635–644. PMLR.

Campbell, Z., Bray, A., Ritz, A., and Groce, A. (2018). Differentially private ANOVA testing. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 281–285. IEEE.

Canonne, C., Hopkins, S. B., Li, J., Liu, A., and Narayanan, S. (2023). The Full Landscape of Robust Mean Testing: Sharp Separations between Oblivious and Adaptive Contamination. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE.

Chatterjee, A. and Bhattacharya, B. B. (2023). Boosting the power of kernel two-sample tests. *arXiv preprint arXiv:2302.10687*.

Chen, M., Gao, C., and Ren, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electronic Journal of Statistics*, 10(2):3752 – 3774.

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR.

Chérief-Abdellatif, B.-E. and Alquier, P. (2022). Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1):181–213.

Couch, S., Kazan, Z., Shi, K., Bray, A., and Groce, A. (2019). Differentially private nonparametric hypothesis testing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 737–751.

Dellaporta, C. and Damoulas, T. (2023). Robust Bayesian Inference for Berkson and Classical Measurement Error Models. *arXiv preprint arXiv:2306.01468*.

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. (2022). Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 943–970. PMLR.

Diakonikolas, I. and Kane, D. M. (2019). Recent Advances in Algorithmic High-Dimensional Robust Statistics. *arXiv preprint arXiv:1911.05911*.

Diakonikolas, I. and Kane, D. M. (2021). The sample complexity of robust covariance testing. In *Conference on Learning Theory*, pages 1511–1521. PMLR.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2017). Statistical Query Lower Bounds for Robust Estimation of High-Dimensional Gaussians and Gaussian Mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE.

Domingo-Enrich, C., Dwivedi, R., and Mackey, L. (2023). Compress then test: Powerful kernel testing in near-linear time. In *International Conference on Artificial Intelligence and Statistics*.

Dwork, C. and Lei, J. (2009). Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Dwork, C., Roth, A., et al. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Fazeli-Asl, F., Zhang, M. M., and Lin, L. (2024). A semi-bayesian nonparametric estimator of the maximum mean discrepancy measure: Applications in goodness-of-fit testing and generative adversarial networks. *Transactions on Machine Learning Research.*

Fienberg, S. E., Slavkovic, A., and Uhler, C. (2011). Privacy preserving GWAS data sharing. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 628–635. IEEE.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 1, pages 489–496.

Gao, R., Xie, L., Xie, Y., and Xu, H. (2018). Robust Hypothesis Testing Using Wasserstein Uncertainty Sets. In *Advances in Neural Information Processing Systems*, volume 31.

George, A. J. and Canonne, C. L. (2022). Robust testing in high-dimensional sparse models. *Advances in Neural Information Processing Systems*, 35:16469–16480.

Gretton, A. (2015). A simpler condition for consistency of a kernel independence test. *arXiv preprint arXiv:1501.06103*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a). Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems*, 25.

Hagrass, O., Sriperumbudur, B. K., and Li, B. (2024). Spectral Regularized Kernel Two-Sample Tests. *The Annals of Statistics*, 52(3):1076–1101.

Hemerik, J. and Goeman, J. (2018). Exact testing with random permutations. *Test*, 27(4):811–825.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Kazan, Z., Shi, K., Groce, A., and Bray, A. P. (2023). The test of tests: A framework for differentially private hypothesis testing. In *International Conference on Machine Learning*, pages 16131–16151. PMLR.

Kim, I. (2021). Comparing a large number of multivariate distributions. *Bernoulli*, 27(1):419–441.

Kim, I. and Schrab, A. (2023). Differentially Private Permutation Tests: Applications to Kernel Methods. *arXiv preprint arXiv:2310.19043*.

Le Cam, L. (1973). Convergence of Estimates Under Dimensionality Restrictions. *The Annals of Statistics*, 1(1):38–53.

Le Cam, L. (2012). *Asymptotic Methods in Statistical Decision Theory.* Springer Science & Business Media.

Legramanti, S., Durante, D., and Alquier, P. (2025). Concentration of discrepancy-based approximate Bayesian computation via Rademacher complexity. *The Annals of Statistics*, 53(1):37–60.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, volume 3. Springer.

Li, M., Berrett, T. B., and Yu, Y. (2023). On robustness and local differential privacy. *The Annals of Statistics*, 51(2):717–737.

Li, T. and Yuan, M. (2024). On the Optimality of Gaussian Kernel Based Nonparametric Tests against Smooth Alternatives. *Journal of Machine Learning Research*, 25(334):1–62.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA.

Peña, V. and Barrientos, A. F. (2022). Differentially Private Hypothesis Testing with the Subsampled and Aggregated Randomized Response Mechanism. *arXiv preprint arXiv:2208.06803*.

Raj, A., Law, H. C. L., Sejdinovic, D., and Park, M. (2020). A differentially private kernel two-sample test. *Lecture Notes in Computer Science*, 11906.

Rogers, R. and Kifer, D. (2017). A new class of private chi-square hypothesis tests. In *Artificial Intelligence and Statistics*, pages 991–1000. PMLR.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Schrab, A., Guedj, B., and Gretton, A. (2022a). KSD Aggregated Goodness-of-fit Test. *Advances in Neural Information Processing Systems*, 35:32624–32638.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD Aggregated Two-Sample Test. *Journal of Machine Learning Research*, 24(194):1–81.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022b). Efficient Aggregated Kernel Tests using Incomplete U-statistics. *Advances in Neural Information Processing Systems*, 35:18793–18807.

Shekhar, S., Kim, I., and Ramdas, A. (2022). A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*.

Shekhar, S., Kim, I., and Ramdas, A. (2023). A permutation-free kernel independence test. *Journal of Machine Learning Research*, 24(369):1–68.

Sun, Z. and Zou, S. (2021). A Data-Driven Approach to Robust Hypothesis Testing Using Kernel MMD Uncertainty Sets. In *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE.

Sun, Z. and Zou, S. (2022). Robust hypothesis testing with kernel uncertainty sets. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 3309–3314. IEEE.

Sun, Z. and Zou, S. (2023). Kernel robust hypothesis testing. *IEEE Transactions on Information Theory*, 69(10):6619–6638.

Wang, J. and Xie, Y. (2022). A Data-Driven Approach to Robust Hypothesis Testing Using Sinkhorn Uncertainty Sets. In *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE.

Yamada, M., Wu, D., Tsai, Y.-H. H., Takeuchi, I., Salakhutdinov, R., and Fukumizu, K. (2019). Post selection inference with incomplete maximum mean discrepancy estimator. *International Conference on Learning Representations*.

Zaremba, W., Gretton, A., and Blaschko, M. (2013). B-test: A non-parametric, low variance kernel two-sample test. *Advances in Neural Information Processing Systems*, 26.

Zhao, J. and Meng, D. (2015). FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation*, 27(6):1345–1372.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable (outputs are binary so error bars are deterministic given the number of repetitions)
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Yes
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes
   (d) Information about consent from data providers/curators. Yes
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Supplementary Materials
Robust Kernel Hypothesis Testing under Data Corruption

In this supplementary material, we discuss computational resources (Appendix A), present assumptions (Appendix B), derive proofs of the theoretical results presented in the main text (Appendix C), and offer a power analysis of the DP tests (Appendix D).

A EXPERIMENTAL RESOURCES

Reproducible code in JAX (Bradbury et al., 2018) is available at `https://github.com/antoninschrab/dckernel-paper` under the MIT License. Experiments were run on a 24Gb RTX A5000 GPU, total compute time is of the order of ten hours. For the DP tests, we adapt the implementation of Kim and Schrab (2023) published under the MIT License. We use the publicly available IMDb dataset of Maas et al. (2011).

B ASSUMPTIONS

This section collects several technical assumptions used in the main text to streamline our presentation. We first make the following assumptions on kernels.

Assumption 1 (Conditions on kernel for Theorem 1)**.** *Let $k : \mathbb{S} \times \mathbb{S} \mapsto \mathbb{R}$ be a reproducing kernel defined on $\mathbb{S}$.*

(i) Assume that the kernel k is characteristic, non-negative, and bounded everywhere by a constant K, i.e., $0 \leq k(x, y) \leq K$ for all $x, y \in \mathbb{S}$.

(ii) Setting $\mathbb{S} = \mathbb{R}^d$, assume that the kernel k is translation invariant on $\mathbb{S}$, i.e., there exists a symmetric positive definite function κ such that $k(x, y) = \kappa(x - y)$ for all $x, y \in \mathbb{R}^d$. Moreover, assume that the kernel is non-constant in the sense that there exists a positive constant η such that $\kappa(0) - \kappa(z) \geq \eta$ for some $z \in \mathbb{R}^d$.

Assumption 2 (Conditions on kernels for Theorem 2)**.** *Let $k : \mathbb{S}_Y \times \mathbb{S}_Y \mapsto \mathbb{R}$ and $\ell : \mathbb{S}_Z \times \mathbb{S}_Z \mapsto \mathbb{R}$ be reproducing kernels defined on $\mathbb{S}_Y$ and $\mathbb{S}_Z$, respectively.*

(i) Assume that the kernels k and ℓ are characteristic, non-negative, and bounded everywhere by constants K and L, respectively, i.e., $0 \leq k(y, y') \leq K$ for all $y, y' \in \mathbb{S}_Y$ and $0 \leq \ell(z, z') \leq L$ for all $z, z' \in \mathbb{S}_Z$.

(ii) Setting $\mathbb{S}_Y = \mathbb{R}^{d_Y}$ and $\mathbb{S}_Z = \mathbb{R}^{d_Z}$, assume that the kernels k and ℓ are translation invariant on $\mathbb{S}_Y$ and $\mathbb{S}_Z$, i.e., there exist symmetric positive definite functions κ_Y and κ_Z such that $k(y, y') = \kappa_Y(y - y')$ for all $y, y' \in \mathbb{R}^{d_Y}$ and $\ell(z, z') = \kappa_Z(z - z')$ for all $z, z' \in \mathbb{R}^{d_Z}$. Moreover, assume that the kernels are non-constant in the sense that there exist positive constants η_Y, η_Z such that $\kappa_Y(0) - \kappa_Y(y_0) \geq \eta_Y$ for some $y_0 \in \mathbb{R}^{d_Y}$ and $\kappa_Z(0) - \kappa_Z(z_0) \geq \eta_Z$ for some $z_0 \in \mathbb{R}^{d_Z}$.

As discussed in Kim and Schrab (2023), commonly used kernels, such as the Gaussian kernel and the Laplace kernel, satisfy the above assumptions. We next describe the assumption for B, the number of permutations, used in Theorems 1, 2, 3 and 4.

Assumption 3. *We make the following assumptions on the number of permutations B for Theorems 1, 2, 3 and 4:*

- *Permutation numbers for dpMMD/dpHSIC: $B \geq 6\alpha_{\mathtt{dp}}^{-1} \log(2/\beta_{\mathtt{dp}})$ where $\alpha_{\mathtt{dp}} = e^{-r\varepsilon}\alpha$ and $\beta_{\mathtt{dp}} = e^{-r\varepsilon}\beta$.*
- *Permutation numbers for dcMMD/dcHSIC: $B \geq 3\alpha^{-2}\big\{\log\big(8/\beta\big) + \alpha(1 - \alpha)\big\}$.*

The reason for the difference in the required number of permutations between the DP test and the DC test stems from their reliance on different techniques. Specifically, we can guarantee that $M_0, M_1, \ldots, M_B$ in Algorithm 2 are all distinct with probability one, due to the injection of continuous noise to the test statistics. This property allows us to employ the multiplicative Chernoff inequality as in Kim and Schrab (2023, Lemma 21). On the other hand, there is no guarantee that $T_0, T_1, \ldots, T_B$ are distinct in Algorithm 1, for which we apply the Dvoretzky–Kiefer–Wolfowitz inequality to analyse the random permutation distribution, as in Schrab et al. (2023, Proposition 4). If we randomly break ties in $T_0, T_1, \ldots, T_B$, the condition for dcMMD/dcHSIC can be improved to $6\alpha^{-1}\log(2/\beta)$.

C PROOFS OF DC PROCEDURE MINIMAX OPTIMALITY

This section collects the proofs of the technical results in the main text.

Additional notation. Let $X \sim P$ and $Y \sim Q$. We use $d_{\mathrm{TV}}(P, Q)$ to denote the total variation (TV) distance between P and Q. With an abuse of notation, we also use $d_{\mathrm{TV}}(X, Y)$ to denote $d_{\mathrm{TV}}(P, Q)$. As in the main text, the subscript P in $\mathbb{P}_P$ is used to emphasise that the underlying data are generated from the distribution P. We often omit the dependence on P in $\mathbb{P}_P$ whenever it is implicitly clear from the context. We stress that the probabilities are also taken with respect to $\boldsymbol{\pi}$, that is, with respect to the uniformly random draw of permutations.

C.1 Proof of Lemma 1

Lemma 1 (Validity & consistency of DC)**.** *(i) The DC test of Algorithm 1 has non-asymptotic level control under r data corruption. (ii) Let $P_1 \in \mathcal{P}_1$ be an alternative distribution. Assume $\alpha \in (0,1)$ fixed. For any sequence of B_n of permutation numbers satisfying $\min_{n\in\mathbb{N}} B_n > \alpha^{-1} - 1$, the DC test is consistent in the sense that $\lim_{n\to\infty} \mathbb{P}_{P_1}\big(\text{DC rejects } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1$ if*

$$\lim_{n\to\infty} \mathbb{P}_{P_1}\big(T(\widetilde{\mathcal{X}}_n) \;>\; T(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}) + 4r\Delta_T\big) = 1,$$

where the probability is taken with respect to the (uniformly) random permutation $\boldsymbol{\pi}$ of $[n]$, and to $\widetilde{\mathcal{X}}_n$ i.i.d. drawn from P_1.

Proof. (i) To simplify the notation, define $U_0 = T(\widetilde{\mathcal{X}}_n)$, $U_i = T(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}_i})$ for $i \in [B]$, (*i.e.*, the test statistics based on the uncorrupted dataset $\widetilde{\mathcal{X}}_n$ and its permuted counterparts $\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}_i}$) and denote the $(1-\alpha)$-quantile of $U_0, U_1, \ldots, U_B$ as $\widetilde{q}$. Equipped with this notation, observe that $|U_i - T_i| \le r\Delta_T$ for each $i \in [B]_0$, which follows by the repeated use of the triangle inequality and by using the definition of global sensitivity, *i.e.*, $\Delta_T = \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}_n} \sup_{\mathcal{X}_n, \mathcal{Y}_n : d_{\mathrm{ham}}(\mathcal{X}_n, \mathcal{Y}_n) \le 1} |T(\mathcal{X}_n^{\boldsymbol{\pi}}) - T(\mathcal{Y}_n^{\boldsymbol{\pi}})|$. This inequality yields

$$\mathbb{1}(T_0 > q + 2r\Delta_T) \le \mathbb{1}(U_0 + r\Delta_T > \widetilde{q} - r\Delta_T + 2r\Delta_T) = \mathbb{1}(U_0 > \widetilde{q}). \tag{2}$$

Since $U_0, U_1, \ldots, U_B$ are exchangeable under the null, we have $\mathbb{P}_{P_0}(T_0 > q + 2r\Delta_T) \le \mathbb{P}_P(U_0 > \widetilde{q}) \le \alpha$ for any $P_0 \in \mathcal{P}_0$ (Romano and Wolf, 2005, Lemma 1), which proves the validity result.

(ii) Using the same notation, we first note that the type II error of the DC test is bounded above as follows

$$\mathbb{P}_{P_1}(T_0 \le q + 2r\Delta_T) \le \mathbb{P}_{P_1}(U_0 - r\Delta_T \le \widetilde{q} + r\Delta_T + 2r\Delta_T) = \mathbb{P}_{P_1}(U_0 \le \widetilde{q} + 4r\Delta_T), \tag{3}$$

where we use the fact that $|U_i - T_i| \le r\Delta_T$ for each $i \in [B]_0$. Therefore, in order to prove that the DC test is consistent, it suffices to show that the above upper bound for the type II error converges to zero as $n \to \infty$ under the given conditions. Now, denoting $\overline{U}_i = U_i + 4r\Delta_T$ for $i \in [B]_0$, by definitions of quantiles, we obtain

$$\begin{aligned}
\mathbb{1}\left(\frac{1}{B+1}\left\{\sum_{i=1}^{B} \mathbb{1}(U_0 \le \overline{U}_i) + 1\right\} \le \alpha\right) &= \mathbb{1}\left(\frac{1}{B+1}\sum_{i=0}^{B} \mathbb{1}(U_0 \le \overline{U}_i) \le \alpha\right) \\
&= \mathbb{1}(U_0 > \widetilde{q} + 4r\Delta_T).
\end{aligned}$$

Using this alternative representation of the test in conjunction with Kim and Schrab (2023, Lemma 8), it follows that

$$\lim_{n\to\infty} \mathbb{P}_{P_1}(U_0 > \widetilde{q} + 4r\Delta_T) = 1 \quad \text{equivalently} \quad \lim_{n\to\infty} \mathbb{P}_{P_1}(U_0 \leq \widetilde{q} + 4r\Delta_T) = 0,$$

if $\lim_{n\to\infty} \mathbb{P}_{P_1}(U_0 \leq \overline{U}_1) = \lim_{n\to\infty} \mathbb{P}_{P_1}(T(\widetilde{\mathcal{X}}_n) \leq T(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}) + 4r\Delta_T) = 0$ and $\min_{n\in\mathbb{N}} B_n > \alpha^{-1} - 1$. This completes the proof of Lemma 1. □

C.2 Proof of Lemma 2

Lemma 2 (Consistency of dcMMD)**.** *Suppose that the kernel k is characteristic, non-negative, and bounded everywhere by K. Assume that $n \leq m$ and $r/n \to 0$ as $n \to \infty$, and that a sequence of permutation numbers B_n satisfies $\min_{n\in\mathbb{N}} B_n > \alpha^{-1} - 1$. Then, for any fixed P and Q with $P \neq Q$, dcMMD is consistent in the sense that*

$$\lim_{n\to\infty} \mathbb{P}_{P,Q}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1.$$

Proof. In order to prove the consistency of dcMMD, it is enough to verify the condition in Lemma 1:

$$\lim_{n\to\infty} \mathbb{P}_{P_1}\bigg(\widehat{\mathrm{MMD}}(\widetilde{\mathcal{X}}_{n+m}) \ > \ \widehat{\mathrm{MMD}}(\widetilde{\mathcal{X}}^{\boldsymbol{\pi}}_{n+m}) + 4r\frac{\sqrt{2K}}{n} \bigg) = 1.$$

As $r/n \to 0$, we need to verify that

$$\lim_{n\to\infty} \mathbb{P}_{P,Q}\Big(\widehat{\mathrm{MMD}}(\widetilde{\mathcal{X}}_{n+m}) > \ \widehat{\mathrm{MMD}}(\widetilde{\mathcal{X}}^{\boldsymbol{\pi}}_{n+m}) \Big) = 1,$$

which holds by the result of Kim and Schrab (2023, Theorem 5). This proves that dcMMD is consistent. □

C.3 Proof of Theorem 1

Theorem 1 (Minimax optimal uniform separation of dcMMD)**.** *Suppose that the kernel k is characteristic, non-negative, and bounded everywhere by K. For $\alpha, \beta \in (0,1)$, assume that the number of permutations B is greater than $3\alpha^{-2}\big\{\log\big(8/\beta\big) + \alpha(1-\alpha)\big\}$, and that $n \asymp m$ with $n \leq m$.*

(i) (Uniform separation) The dcMMD test is guaranteed to have high power, i.e., $\mathbb{P}_{P,Q}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) \geq 1 - \beta$ for any distributions P and Q separated as

$$\begin{aligned} &\mathrm{MMD}_k(P,Q) \\ \geq \ & C_K \max\bigg\{ \sqrt{\frac{\max\{\log(e/\alpha), \log(e/\beta)\}}{n}}, \ \frac{r}{n} \bigg\} \end{aligned}$$

for some positive constant C_K depending on K.

(ii) (Minimax optimality) Further assuming that the kernel is translation invariant and non-constant as in Assumption 1 in the appendix, this separation rate is optimal in terms of the smallest sample size n, of the data corruption level r, and of the testing errors α, β with $\alpha \asymp \beta$ and $\alpha + \beta < 0.4$.

Proof. We emphasise that from Lemma 1, dcMMD controls the non-asymptotic type I error rate at level α. Hence, we focus on the type II error guarantees.

(i) As shown in Equation (3), the type II error of the DC test based on corrupted data $\mathcal{X}_{n+m}$ is bounded above by the type II error of the modified DC test based on uncorrupted data $\widetilde{\mathcal{X}}_{n+m}$. This modified DC test uses the cutoff value $\widetilde{q} + 4r\Delta_T$ instead of $q + 2r\Delta_T$. This slightly inflated cutoff value only affects the constant factor in the minimum uniform separation. We therefore assume that the data are *not corrupted* throughout the proof and derive the minimum uniform separation for the dcMMD test.

Denoting the $(1-\alpha/2)$-quantile of the full permutation distribution as

$$q_{\infty,1-\alpha/2} = \inf\bigg\{ u \in \mathbb{R} : 1 - \frac{\alpha}{2} \leq \frac{1}{(n+m)!} \sum_{\boldsymbol{\pi}\in\boldsymbol{\Pi}_{n+m}} \mathbb{1}\big(\widehat{\mathrm{MMD}}(\widetilde{\mathcal{X}}^{\boldsymbol{\pi}}_{n+m}) \leq u\big) \bigg\},$$

define the event that $E_1 := \{\widetilde{q} \leq q_{\infty,1-\alpha/2}\}$. With $B \geq 3\alpha^{-2}\big\{\log\big(8/\beta\big) + \alpha(1-\alpha)\big\}$, following the proof of Schrab et al. (2023, Proposition 4) shows that $\mathbb{P}_{P,Q}(E_1) \geq 1 - \beta/2$. Moreover, Kim (2021, Theorem 5.1) with the condition $n \asymp m$ ensures that

$$q_{\infty,1-\alpha/2} \leq C_{1,K}\sqrt{\frac{\log(2/\alpha)}{n}},$$

where $C_{1,K}, C_{2,K}, \ldots$ are constants only depending on K. Define another event

$$E_2 := \left\{\left|\mathrm{MMD}_k(P,Q) - \widehat{\mathrm{MMD}}(\widetilde{\mathcal{X}}_{n+m})\right| \leq C_{2,K}\sqrt{\frac{\log(4/\beta)}{n}}\right\},$$

which satisfies $\mathbb{P}_{P,Q}(E_2) \geq 1 - \beta/2$ by Gretton et al. (2012a, Theorem 7). With the two events E_1 and E_2, holding with high probability, the type II error of the dcMMD test is bounded above as

$$\begin{aligned}
&\mathbb{P}_{P,Q}\big(\text{dcMMD fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\big) \\
=\ &\mathbb{P}_{P,Q}\left(\widehat{\mathrm{MMD}}(\widetilde{\mathcal{X}}_{n+m}) \leq \widetilde{q} + \frac{4r\sqrt{2K}}{n}\right) \\
\leq\ &\mathbb{P}_{P,Q}\left(\widehat{\mathrm{MMD}}(\widetilde{\mathcal{X}}_{n+m}) \leq C_{1,K}\sqrt{\frac{\log(2/\alpha)}{n}} + \frac{4r\sqrt{2K}}{n}\right) + \mathbb{P}_{P,Q}(E_1^c) \\
\leq\ &\mathbb{P}_{P,Q}\left(\mathrm{MMD}_k(P,Q) - C_{2,K}\sqrt{\frac{\log(4/\beta)}{n}} \leq C_{1,K}\sqrt{\frac{\log(2/\alpha)}{n}} + \frac{4r\sqrt{2K}}{n}\right) \\
&\qquad + \mathbb{P}_{P,Q}(E_1^c) + \mathbb{P}_{P,Q}(E_2^c) \\
\leq\ &\mathbb{P}_{P,Q}\left(\mathrm{MMD}_k(P,Q) \leq C_{3,K}\max\left\{\sqrt{\frac{\max\{\log(e/\alpha), \log(e/\beta)\}}{n}}, \frac{r}{n}\right\}\right) + \beta = \beta,
\end{aligned}$$

where the last equality holds by taking $C_K > C_{3,K}$ in the theorem statement.

(ii) We aim to prove that if the separation parameter ρ in Equation (7) is smaller than the following threshold:

$$\rho \leq C_\eta \max\left\{\min\left(\sqrt{\frac{\log(e/(\alpha+\beta))}{n}}, 1\right), \frac{r}{n}\right\},$$

no test can have power greater than $1-\beta$ uniformly over $\mathcal{P}_{\mathrm{MMD}_k}(\rho)$ where $C_\eta > 0$ is a small constant depending on η in Assumption 1. The first part of the lower bound, involving $n^{-1/2}$, was obtained in Kim and Schrab (2023, Appendix E.10.1), except that we have $\log(e/(\alpha+\beta))$ instead of $\log(1/(\alpha+\beta))$. Under the condition $\alpha + \beta < 0.4$, it can be seen that both are the same rate by adjusting the constant factor C_η. Hence, it is enough to prove that the second part of the threshold, involving r/n.

As remarked in Kim and Schrab (2023, Appendix E.10.1), the minimax separation for two-sample testing is not smaller than that for one-sample testing. Therefore, it suffices to derive the lower bound result for one-sample testing. Specifically, given i.i.d. observations $Y_1, \ldots, Y_n$ drawn from P, we are interested in distinguishing $\mathcal{H}_0 : P = Q_0$ against $\mathcal{H}_1 : \mathrm{MMD}_k(P, Q_0) \geq \rho$ where Q_0 is a hypothesised (known) distribution, which will be specified later on.

Recall that a test is simply a function ϕ taking as input data and returning a binary value for whether the test rejects the null. We denote a collection of level α tests for one-sample testing under r data corruption as

$$\Phi_{\alpha,r,Q_0} = \left\{\phi : \sup_{M_r \in \mathcal{M}_r} \mathbb{E}_{Y^n \sim Q_0}[\phi(M_r(Y^n))] \leq \alpha\right\},$$

where $\mathcal{M}_r$ is the collection of r-manipulation attack functions, *i.e.*, each function $M_r : \mathbb{R}^{n\times d} \mapsto \mathbb{R}^{n\times d}$ in $\mathcal{M}_r$ first chooses r components of $Y^n = (Y_1, \ldots, Y_n)$ and changes them to arbitrary values on the same support. We define the class of alternatives for the one-sample problem as

$$\mathcal{P}_1(\rho) \coloneqq \big\{P : \mathrm{MMD}_k(P, Q_0) \geq \rho\big\}.$$

Choose some specific distribution $P_0 \in \mathcal{P}_1(\rho)$ and $M_r^* \in \mathcal{M}_r$ specified later on. By Le Cam's two point method (Le Cam, 1973, 2012), the minimax type II error satisfies

$$\begin{aligned}
&\inf_{\phi \in \Phi_{\alpha,r,Q_0}} \sup_{M_r \in \mathcal{M}_r} \sup_{P \in \mathcal{P}_1(\rho)} \mathbb{E}_{Y^n \sim P}[1 - \phi(M_r(Y^n))] \\
&\geq \inf_{\phi \in \Phi_{\alpha,r,Q_0}} \mathbb{E}_{Y^n \sim P_0}[1 - \phi(M_r^*(Y^n))] \\
&= 1 - \sup_{\phi \in \Phi_{\alpha,r,Q_0}} \mathbb{E}_{Y^n \sim P_0}[\phi(M_r^*(Y^n))] \\
&= 1 - \sup_{\phi \in \Phi_{\alpha,r,Q_0}} \Big\{ \mathbb{E}_{Y^n \sim P_0}[\phi(M_r^*(Y^n))] - \mathbb{E}_{Y^n \sim Q_0}[\phi(Y^n)] + \mathbb{E}_{Y^n \sim Q_0}[\phi(Y^n)] \Big\} \\
&\geq 1 - \alpha - \sup_{\phi \in \Phi_{\alpha,r,Q_0}} \Big\{ \mathbb{E}_{Y^n \sim P_0}[\phi(M_r^*(Y^n))] - \mathbb{E}_{Y^n \sim Q_0}[\phi(Y^n)] \Big\} \\
&\geq 1 - \alpha - d_{\mathrm{TV}}\big(M_r^*(Y_{P_0}^n), Y_{Q_0}^n\big) \\
&= 1 - \alpha - d_{\mathrm{TV}}\big(M_r^*(Y_{P_0}^n), F(Y_{P_0}^n)\big),
\end{aligned}$$

where F is an optimal transport such that $Y_{Q_0}^n \overset{d}{=} F(Y_{P_0}^n)$, and

$$\inf_{R \in \pi(P_0,Q_0)} \mathbb{E}_{(Y_{P_0}^n, Y_{Q_0}^n) \sim R}\big[d_{\mathrm{Ham}}(Y_{P_0}^n, Y_{Q_0}^n))\big] = \mathbb{E}\big[d_{\mathrm{Ham}}(Y_{P_0}^n, F(Y_{P_0}^n))\big], \tag{4}$$

which is guaranteed to exist (minimiser of optimal transport). Here $\pi(P_0, Q_0)$ denotes the set of all couplings between P_0 and Q_0. Now, following the proof of Acharya et al. (2021b, Theorem 1), take M_r^* as

$$M_r^*(Y_{P_0}^n) = \begin{cases} F(Y_{P_0}^n), & \text{if } d_{\mathrm{Ham}}(Y_{P_0}^n, F(Y_{P_0}^n)) \leq r, \\ Y_{P_0}^n, & \text{if } d_{\mathrm{Ham}}(Y_{P_0}^n, F(Y_{P_0}^n)) > r, \end{cases}$$

which corrupts at most r samples by construction, and hence belongs to $\mathcal{M}_r$. Hence, we get

$$\begin{aligned}
1 - \alpha - d_{\mathrm{TV}}\big(M_r^*(Y_{P_0}^n), F(Y_{P_0}^n)\big) &\overset{\text{(i)}}{\geq} 1 - \alpha - \mathbb{P}\big(d_{\mathrm{Ham}}(Y_{P_0}^n, F(Y_{P_0}^n)) > r\big) \\
&\overset{\text{(ii)}}{\geq} 1 - \alpha - r^{-1}\mathbb{E}\big[d_{\mathrm{Ham}}(Y_{P_0}^n, F(Y_{P_0}^n))\big] \\
&\overset{\text{(iii)}}{\geq} 1 - \alpha - r^{-1} n d_{\mathrm{TV}}(P_0, Q_0),
\end{aligned}$$

where step (i) follows by the coupling lemma of the TV distance, *i.e.*, $d_{\mathrm{TV}}(X, Y) \leq \mathbb{P}(X \neq Y)$, step (ii) uses Markov's inequality, and step (iii) follows by Acharya et al. (2021a, Lemma 20) along with the definition of F in (4). Therefore, if

$$d_{\mathrm{TV}}(P_0, Q_0) \leq \frac{r}{n}(1 - \alpha - \beta), \tag{5}$$

then the minimax type II error is bounded below by β. Finally, as in Kim and Schrab (2023, Appendix E.10.1), we choose $P_0 = p_0\delta_x + (1 - p_0)\delta_v$ and $Q_0 = q_0\delta_x + (1 - q_0)\delta_v$, where $x, v \in \mathbb{R}^d$, $0 < p_0, q_0 < 1$ and δ_x is a Dirac measure at x, which yields

$$\mathrm{MMD}_k(P_0, Q_0) = \sqrt{2\big(\kappa(0) - \kappa(x - v)\big)} \underbrace{|p_0 - q_0|}_{= d_{\mathrm{TV}}(P_0, Q_0)} .$$

Choose x and v such that $\kappa(0) - \kappa(x - v) \geq \eta$. Moreover let $q_0 = 1/2$ and $p_0 = 1/2 + r/(2n)$. Then since $\alpha + \beta < 0.4$, the condition (5) holds as

$$d_{\mathrm{TV}}(P_0, Q_0) = \frac{r}{2n} \leq \frac{r}{n}(1 - \alpha - \beta)$$

and the corresponding MMD is upper bounded as

$$\mathrm{MMD}_k(P_0, Q_0) \geq \frac{\sqrt{2\eta}}{2} \frac{r}{n}.$$

Hence, the second part of the lower bound holds. This together with the condition $\alpha \asymp \beta$ completes the proof of Theorem 1(ii). □

C.4 Proof of Lemma 3

Lemma 3 (Consistency of dcHSIC)**.** *Suppose that the kernels k and ℓ are characteristic, non-negative, and bounded everywhere by K and L, respectively. Assume that $r/n \to 0$ as $n \to \infty$, and that a sequence of permutation numbers B_n satisfies $\min_{n\in\mathbb{N}} B_n > \alpha^{-1} - 1$. Then, for any fixed joint distribution P_{YZ} with $P_{YZ} \neq P_Y \times P_Z$, dcHSIC is consistent in the sense that*

$$\lim_{n\to\infty} \mathbb{P}_{P_{YZ}}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1.$$

Proof. The proof of Lemma 3 mirrors the one of Lemma 2. In order to prove the consistency of dcHSIC, it is enough to verify the condition in Lemma 1:

$$\lim_{n\to\infty} \mathbb{P}_{P_{YZ}}\Bigg(\widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n) \;>\; \widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}) + \frac{16r(n-1)\sqrt{KL}}{n^2}\Bigg) = 1.$$

Since $r/n \to 0$, it is enough to verify that

$$\lim_{n\to\infty} \mathbb{P}_{P_{YZ}}\Big(\widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n) > \widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}})\Big) = 1,$$

which was shown in the proof of Kim and Schrab (2023, Theorem 6). Therefore, dcHSIC is consistent. □

C.5 Proof of Theorem 2

Theorem 2 (Minimax optimal uniform separation of dcHSIC)**.** *Suppose that the kernels k and ℓ are characteristic, non-negative, and bounded everywhere by K and L, respectively. For $\alpha, \beta \in (0,1)$, assume that the number of permutations B is greater than $\alpha^{-2}\{\log(8/\beta) + \alpha(1-\alpha)\}$.*

(i) (Uniform separation) The dcHSIC test is guaranteed to have high power, i.e., $\mathbb{P}_{P_{YZ}}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) \geq 1 - \beta$ for any joint distribution P_{YZ} separated as

$$\begin{aligned}&\mathrm{HSIC}_{k,\ell}(P_{YZ})\\ \geq\;& C_{K,L} \max\Bigg\{\sqrt{\frac{\max\{\log(e/\alpha), \log(e/\beta)\}}{n}},\; \frac{r}{n}\Bigg\}\end{aligned}$$

for some positive constant $C_{K,L}$ depending on K and L.

(ii) (Minimax optimality) Further assuming that the kernels k and ℓ are translation invariant and non-constant as in Assumption 2 in the appendix, this separation rate is optimal in terms of the sample size n, of the data corruption level r, and of the testing errors α, β with $\alpha \asymp \beta$ and $\alpha + \beta < 0.4$.

Proof. We stress that from Lemma 1, dpHSIC controls the non-asymptotic type I error rate at level α, so we focus on the type II error guarantees.

(i) The proof for the uniform separation of dcHSIC closely resembles that of dcMMD in Theorem 1. The only difference is that we need to use concentration inequalities for the empirical HSIC rather than the empirical MMD. To present details, we assume that the data are *not corrupted* throughout the proof and derive the minimum uniform separation for dcHSIC. As explained in the proof of Theorem 1, this assumption does not affect the separation rate due to the inequality given in Equation (3).

Denote the $(1-\alpha/2)$-quantile of the full permutation distribution as

$$q_{\infty,1-\alpha/2} = \inf\Bigg\{u \in \mathbb{R} : 1 - \frac{\alpha}{2} \leq \frac{1}{n!}\sum_{\boldsymbol{\pi}\in\boldsymbol{\Pi}_n} \mathbb{1}\big(\widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}) \leq u\big)\Bigg\},$$

and define the event that $E_1 := \{\widetilde{q} \leq q_{\infty,1-\alpha/2}\}$. With $B \geq 3\alpha^{-2}\{\log(8/\beta) + \alpha(1-\alpha)\}$, we have $\mathbb{P}_{P_{YZ}}(E_1) \geq 1 - \beta/2$, which is shown in the proof of Schrab et al. (2023, Proposition 4). Moreover, Kim and Schrab (2023, Lemma 12) ensures that

$$q_{\infty,1-\alpha/2} \leq C_{1,K,L}\sqrt{\frac{1}{n}\max\Bigg\{\log\Big(\frac{2}{\alpha}\Big), \sqrt{\log\Big(\frac{2}{\alpha}\Big)}, 1\Bigg\}} \leq C_{2,K,L}\sqrt{\frac{\log(e/\alpha)}{n}}$$

where $C_{1,K,L}, C_{2,K,L}, \ldots$ are constants depending on K and L. Define another event

$$E_2 := \left\{ \left| \mathrm{HSIC}_{k,\ell}(P_{YZ}) - \widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n) \right| \leq C_{3,K,L} \sqrt{\frac{\log(e/\beta)}{n}} \right\}$$

which holds with $\mathbb{P}_{P_{YZ}}(E_2) \geq 1 - \beta/2$ by Kim and Schrab (2023, Lemma 14). With the two events E_1 and E_2, holding with high probability, the type II error of dcHSIC is bounded above as

$$\begin{aligned}
& \mathbb{P}_{P_{YZ}}\big(\text{dcHSIC fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\big) \\
= \; & \mathbb{P}_{P_{YZ}}\left(\widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n) \leq \widetilde{q} + \frac{16r(n-1)\sqrt{KL}}{n^2} \right) \\
\leq \; & \mathbb{P}_{P_{YZ}}\left(\widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n) \leq C_{2,K,L} \sqrt{\frac{\log(e/\alpha)}{n}} + \frac{16r\sqrt{KL}}{n} \right) + \mathbb{P}_{P_{YZ}}(E_1^c) \\
\leq \; & \mathbb{P}_{P_{YZ}}\left(\mathrm{HSIC}_{k,\ell}(P_{YZ}) - C_{3,K,L} \sqrt{\frac{\log(e/\beta)}{n}} \leq C_{2,K,L} \sqrt{\frac{\log(e/\alpha)}{n}} + \frac{16r\sqrt{KL}}{n} \right) \\
& \qquad + \mathbb{P}_{P_{YZ}}(E_1^c) + \mathbb{P}_{P_{YZ}}(E_2^c) \\
\leq \; & \mathbb{P}_{P_{YZ}}\left(\mathrm{HSIC}_{k,\ell}(P_{YZ}) \leq C_{4,K,L} \max\left\{ \sqrt{\frac{\max\{\log(e/\alpha), \log(e/\beta)\}}{n}}, \, \frac{r}{n} \right\} \right) + \beta = \beta,
\end{aligned}$$

where the last equality holds by taking $C_{K,L} > C_{4,K,L}$ in the theorem statement. This proves the uniform separation of dcHSIC.

(ii) We would like to show that if the separation parameter ρ in Equation (8) is smaller than the following threshold:

$$\rho \leq C_{\eta_Y,\eta_Z} \max\left\{ \min\left(\sqrt{\frac{\log(e/(\alpha+\beta))}{n}}, 1 \right), \, \frac{r}{n} \right\},$$

no test can have power greater than $1-\beta$ uniformly over $\mathcal{P}_{\mathrm{HSIC}_{k,\ell}}(\rho)$ where $C_{\eta_Y,\eta_Z} > 0$ is a small constant depending on η_Y, η_Z in Assumption 2. The first part of the lower bound, involving $n^{-1/2}$, was obtained in Kim and Schrab (2023, Appendix F.4.1) with a slight modification to a constant factor C_{η_Y,η_Z} to replace $\log(1/(\alpha+\beta))$ with $\log(e/(\alpha+\beta))$ where $\alpha+\beta < 0.4$. Hence, it is enough to prove the second part of the threshold, involving r/n. The proof of this claim is essentially the same as the proof of Theorem 1(ii). To prove the desired result, it hence suffices to find a distribution P_{YZ} such that

$$\begin{aligned}
& d_{\mathrm{TV}}(P_{YZ}, P_Y \times P_Z) \leq \frac{r}{n}(1-\alpha-\beta) \quad \text{and} \\
& \mathrm{HSIC}_{k,\ell}(P_{YZ}) \geq C_{\eta_Y,\eta_Z} \frac{r}{n}.
\end{aligned} \tag{6}$$

To this end, as in Kim and Schrab (2023, Appendix F.4.1), construct the distribution of (Y,Z) as

$$\begin{aligned}
& \mathbb{P}(Y=y_1, Z=z_1) = \mathbb{P}(Y=y_2, Z=z_2) = 1/4 + \mu \quad \text{and} \\
& \mathbb{P}(Y=y_1, Z=z_2) = \mathbb{P}(Y=y_2, Z=z_1) = 1/4 - \mu,
\end{aligned}$$

for some distinct $y_1, y_2 \in \mathbb{R}^{d_Y}$ and $z_1, z_2 \in \mathbb{R}^{d_Z}$ such that $y_1 - y_2 = y_0$ and $z_1 - z_2 = z_0$, and $\mu \in (0, 1/4]$. Following the calculations in Kim and Schrab (2023, Appendix F.4.1), the HSIC of such P_{YZ} and the TV distance between P_{YZ} and $P_Y \times P_Z$ can be computed as

$$\mathrm{HSIC}_{k,\ell}(P_{YZ}) \geq 2\mu\sqrt{\eta_Y \eta_Z} \quad \text{and} \quad d_{\mathrm{TV}}(P_{YZ}, P_Y \times P_Z) = 2\mu.$$

Therefore, given that $\alpha + \beta < 0.4$, the condition (6) is fulfilled by setting $\mu = r/(4n)$. This, together with the condition $\alpha \asymp \beta$, completes the proof of Theorem 2(ii). □

D ROBUSTNESS PROPERTIES OF THE RELATED DP PROCEDURE

In this section, we derive results on the robustness properties of the DP tests adapted from Kim and Schrab (2023) and presented in Section 5 and Algorithm 2.

First, we guarantee the validity of DP tests under robustness, and provide a condition guaranteeing their consistency in the data corruption framework.

Lemma 4 (Validity & consistency of DP). *(i) The DP test of Algorithm 2 has non-asymptotic level control under r data corruption. (ii) Let $P_1 \in \mathcal{P}_1$ be an alternative distribution. Assume $\alpha \in (0,1)$ fixed and $r\varepsilon \leq \nu$ for all $n \in \mathbb{N}$ where ν is some positive constant. For any sequence of B_n of permutation numbers satisfying $\min_{n\in\mathbb{N}} B_n > e^{\nu}\alpha^{-1} - 1$, the DP test is consistent in the sense that $\lim_{n\to\infty} \mathbb{P}_{P_1}\big(\text{DP rejects } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1$ if*

$$\lim_{n\to\infty} \mathbb{P}_{P_1}\left(T(\widetilde{\mathcal{X}}_n) + \frac{2\zeta\Delta_T}{\varepsilon} > \; T(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}) + \frac{2\zeta'\Delta_T}{\varepsilon}\right) = 1$$

where the probability is taken with respect to the (uniformly) random permutation $\boldsymbol{\pi}$ of $[n]$, to $\widetilde{\mathcal{X}}_n$ i.i.d. drawn from P_1, and to ζ and ζ' i.i.d. $\mathsf{Laplace}(0,1)$ *noise.*

Proof. (i) The non-asympotic validity of the adjusted DP test in the data corruption framework follows from Equation (1) since the test using the uncorrupted data is non-asymptotically valid (in the usual framework) and differentially private (Kim and Schrab, 2023, Theorems 1 and 2).

(ii) By the DP group privacy property (Dwork et al., 2014, Theorem 2.2), we get the type II error bounded above as

$$\begin{aligned}\mathbb{P}_{P_1}\big(\text{DP fails to reject } \mathcal{H}_0 \mid \text{corrupted data}\big) &\leq \; e^{r\varepsilon}\mathbb{P}_{P_1}\big(\text{DP fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\big) \\ &\leq \; e^{\nu}\mathbb{P}_{P_1}\big(\text{DP fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\big)\end{aligned}$$

where the second inequality uses the condition $r\varepsilon \leq \nu$. Moreover, since the $(1-\alpha e^{-r\varepsilon})$-quantile of $M_0,\ldots,M_B$ is smaller than or equal to the $(1-\alpha e^{-\nu})$-quantile of $M_0,\ldots,M_B$, say $q_{1-\alpha e^{-\nu}}$, we have

$$\begin{aligned}\mathbb{P}_{P_1}\big(\text{DP fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\big) &= \mathbb{P}_{P_1}\big(T_0 \leq q \mid \text{uncorrupted data}\big) \\ &\leq \mathbb{P}_{P_1}\big(T_0 \leq q_{1-\alpha e^{-\nu}} \mid \text{uncorrupted data}\big).\end{aligned}$$

By Kim and Schrab (2023, Theorem 3), the above bound converges to zero as $n \to \infty$ under the conditions of Lemma 4. This proves that the DP test is consistent in power. □

Using the above result, we can now show the consistency of the dpMMD and dpHSIC tests under data corruption.

Lemma 5 (Consistency of dpMMD). *Suppose that the kernel k is characteristic, non-negative, and bounded everywhere by K. Assume that $n \leq m$, $r\varepsilon \leq \nu$ for some positive constant ν and $1/(\varepsilon n) \to 0$ as $n \to \infty$, and that a sequence of permutation numbers B_n satisfies $\min_{n\in\mathbb{N}} B_n > e^{\nu}\alpha^{-1} - 1$. Then, for any fixed P and Q with $P \neq Q$, dpMMD is consistent in the sense that*

$$\lim_{n\to\infty} \mathbb{P}_{P,Q}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1.$$

Proof. We need to verify the condition in Lemma 4 that

$$\lim_{n\to\infty} \mathbb{P}_{P,Q}\left(\widehat{\text{MMD}}(\widetilde{\mathcal{X}}_{n+m}) + 2\zeta\frac{\sqrt{2K}}{n}\varepsilon^{-1} > \; \widehat{\text{MMD}}(\widetilde{\mathcal{X}}_{n+m}^{\boldsymbol{\pi}}) + 2\zeta'\frac{\sqrt{2K}}{n}\varepsilon^{-1}\right) = 1.$$

Indeed, since $1/(\varepsilon n) \to 0$, it suffices to verify that

$$\lim_{n\to\infty} \mathbb{P}_{P,Q}\left(\widehat{\text{MMD}}(\widetilde{\mathcal{X}}_{n+m}) > \; \widehat{\text{MMD}}(\widetilde{\mathcal{X}}_{n+m}^{\boldsymbol{\pi}})\right) = 1,$$

which was shown in the proof of Kim and Schrab (2023, Theorem 5). Therefore, dpMMD is consistent. □

Lemma 6 (Consistency of dpHSIC). *Suppose that the kernels k and ℓ are characteristic, non-negative, and bounded everywhere by K and L, respectively. Assume that $r\varepsilon \leq \nu$ for some positive constant ν and $1/(\varepsilon n) \to 0$ as $n \to \infty$, and that a sequence of permutation numbers B_n satisfies $\min_{n\in\mathbb{N}} B_n > e^{\nu}\alpha^{-1} - 1$. Then, for any fixed joint distribution P_{YZ} with $P_{YZ} \neq P_Y \times P_Z$, dpHSIC is consistent in the sense that*

$$\lim_{n\to\infty} \mathbb{P}_{P_{YZ}}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) = 1.$$

Proof. The proof of Lemma 6 follows a similar approach to that of Lemma 5. By Lemma 4, the consistency of dpHSIC holds provided that

$$\lim_{n\to\infty} \mathbb{P}_{P_{YZ}}\left(\widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n) + 2\zeta\frac{4(n-1)\sqrt{KL}}{n^2}\varepsilon^{-1} > \widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}) + 2\zeta'\frac{4(n-1)\sqrt{KL}}{n^2}\varepsilon^{-1} \right) = 1.$$

Since $1/(\varepsilon n) \to 0$, it suffices to verify that

$$\lim_{n\to\infty} \mathbb{P}_{P_{YZ}}\left(\widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n) > \widehat{\mathrm{HSIC}}(\widetilde{\mathcal{X}}_n^{\boldsymbol{\pi}}) \right) = 1,$$

which was shown in the proof of Kim and Schrab (2023, Theorem 6). Therefore, dpHSIC is consistent. □

Finally, we can derive uniform separation rates for the robust DP tests in terms of MMD and HSIC separation.

Theorem 3 (Uniform separation of dpMMD)**.** *Suppose that the kernel k is characteristic, non-negative, and bounded everywhere by K. For $\alpha, \beta \in (0,1)$, assume that the number of permutations is greater than $6\alpha_{\mathsf{dp}}^{-1}\log(2/\beta_{\mathsf{dp}})$ where $\alpha_{\mathsf{dp}} = e^{-r\varepsilon}\alpha$ and $\beta_{\mathsf{dp}} = e^{-r\varepsilon}\beta$, setting $\varepsilon = r^{-1}\max\{\log(e/\alpha), \log(e/\beta)\}$, and $n \asymp m$. The dpMMD test is guaranteed to have high power, i.e., $\mathbb{P}_{P,Q}(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}) \geq 1-\beta$ for any distributions P and Q separated as*

$$\mathrm{MMD}_k(P,Q) \;\geq\; C_K \max\left\{ \sqrt{\frac{\max\{\log(e/\alpha), \log(e/\beta)\}}{n}},\; \frac{r}{n} \right\}$$

for some positive constant C_K depending on K.

Proof. We stress that from Lemma 4, the dpMMD test has non-asymptotic level α. Therefore, we focus on controlling the type II error. Recall that the dpMMD test with level α is defined with privacy parameters $\varepsilon = r^{-1}\max\bigl\{\log(e/\alpha), \log(e/\beta)\bigr\}$ and $\delta = 0$, with adjusted level parameter $\alpha_{\mathsf{dp}} := e^{-r\varepsilon}\alpha$. As in the proof of Lemma 4, we may use the DP group property to bound the type II error of dpMMD as

$$\begin{aligned}
&\mathbb{P}_{P,Q}\bigl(\text{dpMMD fails to reject } \mathcal{H}_0 \mid \text{corrupted data}\bigr)\\
\leq\;& e^{r\varepsilon}\mathbb{P}_{P,Q}\bigl(\text{dpMMD fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\bigr)\\
\leq\;& e^{r\varepsilon}\beta_{\mathsf{dp}} := \beta.
\end{aligned}$$

Let us define the class of pairs of distributions, which are ρ-separated in terms of the MMD metric, as

$$\mathcal{P}_{\mathrm{MMD}_k}(\rho) = \bigl\{(P,Q) : \mathrm{MMD}_k(P,Q) \geq \rho\bigr\}. \tag{7}$$

Now, by leveraging Kim and Schrab (2023, Theorem 7) with $n \asymp m$ and $B \geq 6\alpha_{\mathsf{dp}}^{-1}\log(2\beta_{\mathsf{dp}}^{-1})$, we see that the minimum value of ρ that controls the type II error as

$$\sup_{(P,Q)\in\mathcal{P}_{\mathrm{MMD}_k}(\rho)} \mathbb{P}_{P,Q}\bigl(\text{dpMMD fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\bigr) \;\leq\; \beta_{\mathsf{dp}},$$

satisfies

$$\begin{aligned}
\rho \;&\leq\; C_K' \max\left\{ \sqrt{\frac{\max\bigl\{\log(e/\alpha_{\mathsf{dp}}), \log(e/\beta_{\mathsf{dp}})\bigr\}}{n}},\; \frac{\max\bigl\{\log(e/\alpha_{\mathsf{dp}}), \log(e/\beta_{\mathsf{dp}})\bigr\}}{n\varepsilon} \right\}\\
&\leq\; C_K' \max\left\{ \sqrt{\frac{r\varepsilon + \max\bigl\{\log(e/\alpha), \log(e/\beta)\bigr\}}{n}},\; \frac{r\varepsilon + \max\bigl\{\log(e/\alpha), \log(e/\beta)\bigr\}}{n\varepsilon} \right\}\\
&\leq\; C_K \max\left\{ \sqrt{\frac{r\varepsilon}{n}},\; \frac{r}{n},\; \sqrt{\frac{\max\bigl\{\log(e/\alpha), \log(e/\beta)\bigr\}}{n}},\; \frac{\max\bigl\{\log(e/\alpha), \log(e/\beta)\bigr\}}{n\varepsilon} \right\}\\
&=\; C_K \max\left\{ \sqrt{\frac{\max\bigl\{\log(e/\alpha), \log(e/\beta)\bigr\}}{n}},\; \frac{r}{n} \right\},
\end{aligned}$$

where C_K', C_K are constants depending on K. The last equality holds since we set $\varepsilon = r^{-1}\max\bigl\{\log(e/\alpha), \log(e/\beta)\bigr\}$. This proves the uniform separation of the dpMMD test. □

Theorem 4 (Uniform separation of dpHSIC)**.** *Suppose that the kernels k and ℓ are characteristic, non-negative, and bounded everywhere by K and L, respectively. For $\alpha, \beta \in (0,1)$, assume that the number of permutations is greater than $6\alpha_{\mathtt{dp}}^{-1}\log(2/\beta_{\mathtt{dp}})$ where $\alpha_{\mathtt{dp}} = e^{-r\varepsilon}\alpha$ and $\beta_{\mathtt{dp}} = e^{-r\varepsilon}\beta$, setting $\varepsilon = r^{-1}\max\{\log(e/\alpha), \log(e/\beta)\}$. The dpHSIC test is guaranteed to have high power, i.e., $\mathbb{P}_{P_{YZ}}\big(\text{reject } \mathcal{H}_0 \,|\, r \text{ corrupted data}\big) \geq 1-\beta$ for any joint distribution P_{YZ} separated as*

$$\mathrm{HSIC}_{k,\ell}(P_{YZ}) \;\geq\; C_{K,L}\max\bigg\{\sqrt{\frac{\max\{\log(e/\alpha), \log(e/\beta)\}}{n}},\; \frac{r}{n}\bigg\}$$

for some positive constant $C_{K,L}$ depending on K and L.

Proof. We emphasize that from Lemma 4, the dpHSIC test has non-asymptotic level α. Therefore, we focus on controlling the type II error. Recall that the dpHSIC test with level α is defined with privacy parameters $\varepsilon = r^{-1}\max\big\{\log(e/\alpha), \log(e/\beta)\big\}$ and $\delta = 0$, with adjusted level parameter $\alpha_{\mathtt{dp}} \coloneqq e^{-r\varepsilon}\alpha$. As in the proof of Lemma 4, we may use the DP group property to bound the type II error of dpHSIC as

$$\begin{aligned}
&\mathbb{P}_{P_{YZ}}\big(\text{dpHSIC fails to reject } \mathcal{H}_0 \mid \text{corrupted data}\big)\\
\leq\;& e^{r\varepsilon}\mathbb{P}_{P_{YZ}}\big(\text{dpHSIC fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\big)\\
\leq\;& e^{r\varepsilon}\beta_{\mathtt{dp}} \coloneqq \beta.
\end{aligned}$$

Let us define the class of distributions, which are ρ-separated in terms of the HSIC metric, as

$$\mathcal{P}_{\mathrm{HSIC}_{k,\ell}}(\rho) = \big\{P_{YZ} : \mathrm{HSIC}_{k,\ell}(P_{YZ}) \geq \rho\big\}. \tag{8}$$

Now, by leveraging Kim and Schrab (2023, Theorem 12) and $B \geq 6\alpha_{\mathtt{dp}}^{-1}\log(2\beta_{\mathtt{dp}}^{-1})$, we see that the minimum value of ρ that controls the type II error as

$$\sup_{P_{YZ}\in\mathcal{P}_{\mathrm{HSIC}_{k,\ell}}(\rho)} \mathbb{P}_{P_{YZ}}\big(\text{dpHSIC fails to reject } \mathcal{H}_0 \mid \text{uncorrupted data}\big) \leq \beta_{\mathtt{dp}},$$

satisfies

$$\begin{aligned}
\rho \;&\leq\; C'_{K,L}\max\Bigg\{\sqrt{\frac{\max\big\{\log(e/\alpha_{\mathtt{dp}}),\, \log(e/\beta_{\mathtt{dp}})\big\}}{n}},\; \frac{\max\big\{\log(e/\alpha_{\mathtt{dp}}),\, \log(e/\beta_{\mathtt{dp}})\big\}}{n\varepsilon}\Bigg\}\\
&\leq\; C'_{K,L}\max\Bigg\{\sqrt{\frac{r\varepsilon + \max\big\{\log(e/\alpha),\, \log(e/\beta)\big\}}{n}},\; \frac{r\varepsilon + \max\big\{\log(e/\alpha),\, \log(e/\beta)\big\}}{n\varepsilon}\Bigg\}\\
&\leq\; C_{K,L}\max\Bigg\{\sqrt{\frac{r\varepsilon}{n}},\, \frac{r}{n},\, \sqrt{\frac{\max\big\{\log(e/\alpha),\, \log(e/\beta)\big\}}{n}},\; \frac{\max\big\{\log(e/\alpha),\, \log(e/\beta)\big\}}{n\varepsilon}\Bigg\}\\
&=\; C_{K,L}\max\Bigg\{\sqrt{\frac{\max\big\{\log(e/\alpha),\, \log(e/\beta)\big\}}{n}},\, \frac{r}{n}\Bigg\},
\end{aligned}$$

where $C'_{K,L}, C_{K,L}$ are constants depending on K and L, and where the last equality holds since we set $\varepsilon = r^{-1}\max\big\{\log(e/\alpha), \log(e/\beta)\big\}$. This proves the uniform separation of the dpHSIC test. □