Learning Kernel Tests Without Data Splitting

Jonas M. Kübler **Wittawat Jitkrittum*** **Bernhard Schölkopf** **Krikamol Muandet**
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{jmkuebler, bs, krikamol}@tue.mpg.de, wittawatj@gmail.com

Abstract

Modern large-scale kernel-based tests such as maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) optimize kernel hyperparameters on a held-out sample via data splitting to obtain the most powerful test statistics. While data splitting results in a tractable null distribution, it suffers from a reduction in test power due to smaller test sample size. Inspired by the selective inference framework, we propose an approach that enables learning the hyperparameters and testing on the full sample without data splitting. Our approach can correctly calibrate the test in the presence of such dependency, and yield a test threshold in closed form. At the same significance level, our approach's test power is empirically larger than that of the data-splitting approach, regardless of its split proportion.

1 Introduction

Statistical hypothesis testing is a ubiquitous problem in numerous fields ranging from astronomy and high-energy physics to medicine and psychology [1]. Given a hypothesis about a natural phenomenon, it prescribes a systematic way to test the hypothesis empirically [2]. Two-sample testing, for instance, addresses whether two samples originate from the same process, which is instrumental in experimental science such as psychology, medicine, and economics. This procedure of rejecting false hypotheses while retaining the correct ones governs most advances in science.

Traditionally, test statistics are usually fixed prior to the testing phase. In modern-day hypothesis testing, however, practitioners often face a large family of test statistics from which the best one must be selected before performing the test. For instance, the popular kernel-based two-sample tests [3, 4] and goodness-of-fit tests [5, 6] require the specification of a kernel function and its parameter values. Abundant evidence suggests that finding good parameter values for these tests improves their performance in the testing phase [4, 7–9]. As a result, several approaches have recently been proposed to learn optimal tests directly from data using different techniques such as optimized kernels [4, 9–13], classifier two-sample tests [14, 15], and deep neural networks [16, 17], to name a few. In other words, the modern-day hypothesis testing has become a two-stage "learn-then-test" problem.

Special care must be taken in the subsequent testing when optimal tests are learned from data. If the same data is used for both learning and testing, it becomes harder to derive the asymptotic null distribution because the selected test and the data are now dependent. In this case, conducting the tests as if the test statistics are independent from the data leads to an uncontrollable false positive rate, see, e.g., our experimental results. While permutation testing can be applied [18], it is too computationally prohibitive for real-world applications. Up to now, the most prevalent solution is *data splitting*: the data is randomly split into two parts, of which the former is used for learning the test while the latter is used for testing. Although data splitting is simple and in principle leads to the correct false positive rate, its downside is a potential loss of power.

In this paper, we investigate the two-stage "learn-then-test" problem in the context of modern kernel-based tests [3–6] where the choice of kernel function and its parameters play an important role. The

*Now with Google Research

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

key question is *whether it is possible to employ the full sample for both learning and testing phase without data splitting, while correctly calibrating the test in the presence of such dependency*. We provide an affirmative answer if we learn the test from a vector of jointly normal base test statistics, e.g., the linear-time MMD estimates of multiple kernels. The empirical results suggest that, at the same significance level, the test power of our approach is larger than that of the data-splitting approach, regardless of the split proportion (cf. Section 5). The code for the experiments is available at https://github.com/MPI-IS/tests-wo-splitting.

2 Preliminaries

We start with some background material on conventional hypothesis testing and review linear-time kernel two-sample tests. In what follows, we will use $[d] := \{1, \ldots, d\}$ to denote the set of natural numbers up to $d \in \mathbb{N}$, $v \geq 0$ to denote that all entries of $v \in \mathbb{R}^d$ are non-negative, e_i to denote the i-th Cartesian unit vector, and $k \cdot k := k \circ k$.

Statistical hypothesis testing. Let Z be a random variable taking values in $Z \subseteq \mathbb{R}^p$ distributed according to a distribution P. The goal of statistical hypothesis testing is to decide whether some *null hypothesis* H_0 about P can be rejected in favor of an *alternative hypothesis* H_A based on empirical data [2, 19]. Let h be a real-valued function such that $0 < \mathbb{E}[h^2(Z)] < \infty$. In this work, we consider testing the null hypothesis $H_0 : \mathbb{E}[h(Z)] = 0$ against the one-sided alternative hypothesis $H_A : \mathbb{E}[h(Z)] > 0$ for reasons which will become clear later. To do so, we define the *test statistic* $\hat{\tau}(Z_n) = \frac{1}{n}\sum_{i=1}^{n} h(z_i)$ as the empirical mean of h based on a sample $Z_n := \{z_1, \ldots, z_n\}$ drawn i.i.d. from P^n. We reject H_0 if the observed test statistic $\hat{\tau}(Z_n)$ is *significantly* larger than what we would expect if H_0 was true, i.e., if $P(\tau(Z_n) < \hat{\tau}(Z_n) \mid H_0) > 1 - \alpha$. Here α is a *significance level* and controls the probability of incorrectly rejecting H_0 (Type-I error). For sufficiently large n we can work with the asymptotic distribution of $\tau(Z_n)$, which is characterized by the Central Limit Theorem [20].

Lemma 1. *Let* $\mu := \mathbb{E}[h(Z)]$ *and* $\sigma^2 := \mathrm{Var}[h(Z)]$. *Then, the test statistic converges in distribution to a Gaussian distribution, i.e.,* $\sqrt{n}(\tau(Z_n) - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$:

Let Φ be the CDF of the standard normal and Φ^{-1} its inverse. We define the test threshold $t_\alpha = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha) + \mu$ as the $(1-\alpha)$-quantile of the null distribution so that $P(\tau(Z_n) < t_\alpha \mid H_0) = 1 - \alpha$ and we reject H_0 simply if $\hat{\tau}(Z_n) > t_\alpha$. Besides correctly controlling the Type-I error, the test should also reject H_0 as often as possible when P actually satisfies the alternative H_A. The probability of making a Type-II error is defined as $P(\tau(Z_n) < t_\alpha \mid H_A)$, i.e., the probability of failing to reject H_0 when it is false. A powerful test has a small Type-II error while keeping the Type-I error at α. Since Lemma 1 holds for any P, and thus both under null and alternative hypotheses, the asymptotic probability of a Type-II error is [4]

$$P(\tau(Z_n) < t_\alpha \mid H_A) \to \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sqrt{n}\mu}{\sigma}\right). \tag{1}$$

Since Φ is monotonic, this probability decreases with $\frac{\sqrt{n}\mu}{\sigma}$, which we interpret as a signal-to-noise ratio (SNR). It is therefore desirable to find test statistics with high SNR.

Kernel two-sample testing. As an example that can be expressed in the above form we present kernel two-sample tests. Given two samples X_n and Y_n drawn from distributions P and Q, the two-sample test aims to decide whether P and Q are different, i.e., $H_0 : P = Q$ and $H_A : P \neq Q$. A popular test statistic for this problem is the maximum mean discrepancy (MMD) of Gretton et al. [3], which is defined based on a positive definite kernel function k [21]: $\mathrm{MMD}^2[P, Q] = \mathbb{E}[k(x, x') + k(y, y') - k(x, y') - k(x', y)] = \mathbb{E}[h(x, x', y, y')]$, where x, x' are independent draws from P, y, y' are independent draws from Q, and $h(x, x', y, y') := k(x, x') + k(y, y') - k(x, y') - k(x', y)$. A minimum-variance unbiased estimator of MMD^2 is given by a second-order U-statistic [20]. However, this estimator scales quadratically with the sample size, and the distribution under H_0 is not available in closed form. Thus it has to be simulated either via a bootstrapping approach or via a permutation of the samples. For large sample size, the computational requirements become prohibitive [3]. In this work, we assume we are in this regime. To circumvent these computational burdens, Gretton et al. [3] suggest a "linear-time" MMD estimate that scales linearly with sample size and is asymptotically normally distributed under both null and alternative hypotheses. Specifically,

let $X_{2n} = \{x_1, \ldots, x_{2n}\}$ and $Y_{2n} = \{y_1, \ldots, y_{2n}\}$, i.e., the samples are of the same (even) size. We can define $z_i := (x_i, x_{n+i}, y_i, y_{n+i})$ and $(Z_n) := \frac{1}{n}\sum_{i=1}^{n} h(z_i)$ as the test statistic, which by Lemma 1 is asymptotically normally distributed. Furthermore, if the kernel is characteristic [12], it is guaranteed that $\mathrm{MMD}^2(P, Q) = 0$ if $P = Q$ and $\mathrm{MMD}^2(P, Q) > 0$ otherwise. Therefore, a one-sided test is sufficient.

Other well-known examples are goodness-of-fit tests based on the kernelized Stein discrepancy (KSD), which also has a linear time estimate [5, 6]. In our experiments, we focus on the kernel two-sample test, but point out that our theoretical treatment in Section 3 is more general and can be applied to other problems, e.g., KSD goodness-of-fit tests, but also beyond kernel methods.

3 Selective hypothesis tests

Statistical lore tells us not to use the same data for learning and testing. We now discuss whether it is indeed possible to use the same data for selecting a test statistic from a candidate set and conducting the selected test [2, 3]. The key to controllable Type-I errors is that we need to adjust the test threshold to account for the selection event. As before, let Z_n denote the data we collected. Let $\mathcal{I} = \{\ _i\}_{i \in \mathcal{I}}$ be a countable set of candidate test statistics that we evaluate on the data Z_n, and $\{t^i\}_{i \in \mathcal{I}}$ the respective test thresholds. Assume that $\{A_i\}_{i \in \mathcal{I}}$ are disjoint selection events depending on Z_n and that their outcomes determine which test statistic out of $\mathcal{I}$ we apply. Thus, all the tests and events are generally dependent via Z_n. To define a well-calibrated test, we need to control the overall Type-I error, i.e., $P(\text{reject} \mid H_0)$. Using the law of total probability, we can rewrite this in terms of the selected tests

$$P(\text{reject} \mid H_0) = \sum_{i \in \mathcal{I}} P(\ _i > t^i \mid A_i, H_0) P(A_i \mid H_0). \tag{2}$$

To control the Type-I error $P(\text{reject} \mid H_0)$, it thus suffices to control $P(\ _i > t^i \mid A_i, H_0)$ for each $i \in \mathcal{I}$, i.e., the test thresholds need to take into account the conditioning on the selection event A_i. A naive approach would wrongly calibrate the test such that $P(\ _i > t^i \mid H_0)$, not accounting for the selection A_i and thus would result in an uncontrollable Type-I error. On the other hand, this reasoning directly tells us why data splitting works. There, A_i is evaluated on a split of Z_n that is independent of the split used to compute $\ _i$ and hence $P(\ _i > t^i \mid A_i, H_0) = P(\ _i > t^i \mid H_0)$.

Selecting tests with high power. Our objective in selecting the test statistic is to maximize the power of the selected test. To this end, we start from $d \geq 2$ different base functions $h_1, \ldots, h_d$. Based on observed data $Z_n = \{z_1, \ldots, z_n\} \sim P^n$, we can compute the base test statistics $\ _u := \ _u(Z_n) = \frac{1}{n}\sum_{i=1}^{n} h_u(z_i)$ for $u \in [d]$. Let $\quad := (\ _1, \ldots, \ _d)^\top$ and $\quad := E[h(Z)]$, where $h(Z) = (h_1(Z), \ldots, h_d(Z))^\top$. Asymptotically, we have $\sqrt{n}(\quad) \to_d N(0, \ )$, with the variance of the asymptotic distribution given by $= \mathrm{Cov}[h(Z)]$.[2] Now, for any $\in \mathbb{R}^d \setminus \{0\}$ that is independent of , the normalized test statistic $:= \frac{\ ^\top}{(\ ^\top \ )^{\frac{1}{2}}}$ is asymptotically normal, i.e., $\sqrt{n}\,\frac{\ ^\top}{(\ ^\top \ )^{\frac{1}{2}}} \to_d N(0, 1)$. Following our considerations of Section 2, the test with the highest power is defined by

$$^1 := \operatorname*{argmax}_{k\ k=1} \frac{\ ^\top}{(\ ^\top \ )^{\frac{1}{2}}} = \frac{^{-1}}{k\ ^{-1}k}; \tag{3}$$

where the constraint $k\ k = 1$ is to ensure that the solution is unique, since the objective of the maximization is a homogeneous function of order 0 in . The explicit form of ^{1} is proven in Appendix C.2. Obviously, in practice, is not known, so we use an estimate of to select . The standard strategy to do so is to split the sample Z_n into two independent sets and estimate $_{te}$ and $_{te}$, i.e., two independent training and test realizations [4, 8, 9, 13]. One can then choose a suitable by using $_{tr}$ as a proxy for . Then one tests with this and $_{te}$. However, to our knowledge, there exists no principled way to decide in which proportion to split the data, which will generally influence the power, as shown in our experimental results in Section 5.

[2] In practice, we work with an estimate $\hat{}$ of the covariance obtained from Z_n, which is justified since $\sqrt{n}\,\hat{}^{-\frac{1}{2}}(\quad) \to_d N(0, I_d)$ for consistent estimates of the covariance.

Our approach to maximizing the utility of the observed dataset is to use it for both learning and testing. To do so, we have to derive an adjustment to the distribution of the statistic under the null, in the spirit of the selective hypothesis testing described above. We will consider three different candidate sets $\mathcal{T}$ of test statistics, which are all constructed from the base test statistics T_{base}. To do so, we will work with the asymptotic distribution of $\sqrt{n}\,\bar{\beta}$ under the null. To keep the notation concise, we include the $\sqrt{n}$ dependence into Σ. Thus, we will assume $\beta \sim \mathcal{N}(0, \Sigma)$, where Σ is known and strictly positive. We provide the generalization to singular covariance in Appendix E.

To select the test statistics, we maximize the SNR $\eta = \eta^\top \beta / (\eta^\top \Sigma \eta)^{\frac{1}{2}}$ and thus the test power over three different sets of candidate test statistics: $\mathcal{T}_{\text{base}} = \{ \eta^\top \beta \mid \eta \in \{e_1, \dots, e_d, -e_1, \dots, -e_d\} \}$, i.e., we directly select from the base test statistics; $\mathcal{T}_{\text{Wald}} = \{ \eta^\top \beta \mid \eta^\top \Sigma \eta = 1 \}$, where we allow for arbitrary linear combinations, 3; $\mathcal{T}_{\text{OST}} = \{ \eta^\top \beta \mid \eta \geq 0, \eta^\top \Sigma \eta = 1 \}$, where we constrain the allowed values to increase the power (see below). The rule for selecting the test statistic from these sets is simply to select the one with the highest value. To design selective hypothesis tests, we need to derive suitable selection events and the distribution of the maximum test statistic conditioned on its selection.

3.1 Selection from a finite candidate set

We start with $\mathcal{T}_{\text{base}} = \{ \eta^\top \beta \mid \eta \in \{e_1, \dots, e_d, -e_1, \dots, -e_d\} \}$ and use the test statistic $T_{\text{base}} = \max_{\eta \in \mathcal{T}_{\text{base}}} \eta^\top \beta$. Since the selection is from a countable set and the selected statistic is a projection, we can use the polyhedral lemma of Lee et al. [24] to derive the conditional distributions. Therefore, we denote $u = \operatorname{argmax}_{u \in [d]} \frac{u^\top \beta}{\sigma_u}$, with $\sigma_u := (\Sigma_{uu})^{\frac{1}{2}}$, and obtain $T_{\text{base}} = \frac{u^\top \beta}{\sigma_u}$. The following corollary characterizes the conditional distribution. The proof is given in Appendix C.1.

Corollary 1. Let $\beta \sim \mathcal{N}(\mu, \Sigma)$, $z := \frac{e_u^\top \beta}{\sigma_u^2}$, $V^-(z) = \max_{j \in [d], j \neq u} \frac{\Sigma_{uj}}{\sigma_u^2} z - \frac{\Sigma_{uj} \mu_j - e_u^\top \mu}{\sigma_u^2}$, and $\mathcal{TN}(\mu, \sigma^2, a, b)$ denote a normal distribution with mean μ and variance σ^2 truncated at a and b. Then the following statement holds:

$$\left. \frac{u^\top \beta}{\sigma_u} \;\middle|\; u = \operatorname*{argmax}_{u \in [d]} \frac{u^\top \beta}{\sigma_u}, z = z^* \right. \overset{d}{=} \mathcal{TN}\left(\frac{u^\top \mu}{\sigma_u}, 1, V^-(z^*), V^+ = \infty \right), \tag{4}$$

This scenario arises, for example, in kernel-based tests when the kernel parameters are chosen from a grid of predefined values [3, 4]. Corollary 1 allows us to test using the same set of data that was used to select the test statistic, by providing the corrected asymptotic distribution (4). The only downside is its dependence on the parameter grid. To overcome this limitation, several works have proposed to optimize for the parameters directly [4, 9–12]. Unfortunately, we cannot apply Corollary 1 directly to this scenario.

3.2 Learning from an uncountable candidate set

To allow for more flexible tests, in the following we consider the candidate sets $\mathcal{T}_{\text{Wald}}$ and $\mathcal{T}_{\text{OST}}$ that contain uncountably many tests. For these sets, we cannot directly use (4) to derive conditional tests, since the probability of selecting some given tests is 0. However, we show that it is possible in both cases to rewrite the test statistic such that we can build conditional tests based on (4). First, for $\mathcal{T}_{\text{Wald}}$, we rewrite the entire test statistic including the maximization in closed form. Second, for $\mathcal{T}_{\text{OST}}$ we derive suitable measurable selection events that allow us to rewrite the conditional test statistic in closed form and derive their distributions in Theorem 1.

Wald Test. We first allow for arbitrary linear combinations of the base test statistics. Therefore, define $\mathcal{T}_{\text{Wald}} = \{ \eta^\top \beta \mid \eta^\top \Sigma \eta = 1 \}$ and $T_{\text{Wald}} := \max_{\eta \in \mathcal{T}_{\text{Wald}}} \eta^\top \beta$. We denote the optimal η for this set as $\eta_{\text{Wald}} := \operatorname{argmax}_{\eta \; \eta^\top \Sigma \eta = 1} \frac{\eta^\top \beta}{(\eta^\top \Sigma \eta)^{\frac{1}{2}}}$. This optimization problem is the same as (3), hence $\eta_{\text{Wald}} = \frac{\Sigma^{-1} \nu}{\nu^\top \Sigma^{-1} \nu}$; and we can rewrite the "Wald" test statistic as $T_{\text{Wald}} = \frac{\eta_{\text{Wald}}^\top \beta}{(\eta_{\text{Wald}}^\top \Sigma \eta_{\text{Wald}})^{\frac{1}{2}}} = (\beta^\top \Sigma^{-1} \beta)^{\frac{1}{2}} = \|\beta\|_\Sigma$. Note that T_{Wald} contains uncountably many tests. However, instead of deriving individual conditional distributions, we can directly derive the distribution of the maximized test statistic, since η_{Wald} can be written in closed form. In fact, under the null, we have $\Sigma^{-\frac{1}{2}} \beta \sim \mathcal{N}(0, I_d)$ and T_{Wald} follows a chi distribution with d degrees of freedom. Surprisingly, the presented approach results in the classic Wald test statistic [25], which originally was defined directly in closed form.

One-sided test (OST). The original Wald test was de ned to optimally test $H_0 : = 0$ against the alternative $H_A : \neq 0$ [25]. Thus, it ignores the fact that we only test against the "one-sided" alternative 0, which suf ces since we consider linear-time estimates of the squared MMD as test statistics and their population values are non-negative. Multiplying with (3) yields $^1 = \frac{}{k \ ^1 \ k}$: Using 0, we nd $^1 \quad 0$. Thus, we have prior knowledge over the asymptotically optimal combination ^{1} . To incorporate this, we a priori constrain the considered values of by the condition 0. Thus we de ne $T_{OST} = f \quad j \quad 0; k \quad k = 1 g$, where the norm constraint $k \quad k = 1$ is added to make the maximum unique. We suggest using the test statistic $_{OST} := \max_{2 T_{OST}}$. Before we derive suitable conditional distributions for this test statistic, we rewrite it in a canonical form:

Remark 1. De ne $:= \quad , := \quad ^1$, and $^0 := \quad ^1 \quad ^1 = \quad ^1$: This implies $N (0; \ ^0)$ and $_{OST} := \max_k \ _{k=1 ; \quad 0} \frac{^>}{(\ ^> \ )^{\frac{1}{2}}} = \max_k \ _{k=1 ; \quad 0} \frac{^>}{(\ ^> \ ^0 \ )^{\frac{1}{2}}}$:

Thus in the following, we focus on the canonical form, where the constraints are simply positivity constraints. For ease of notation, we stick with and instead of and ^{0}. We will thus analyze the distribution of

$$\max_k \ _{k=1 ; \quad 0} \frac{^>}{(\ ^> \ )^{\frac{1}{2}}} = \frac{^>}{(\ ^> \ )^{\frac{1}{2}}}; \tag{5}$$

where $(\ ) := \operatorname{argmax}_k \ _{k=1 ; \quad 0} \frac{^>}{(\ ^> \ )^{\frac{1}{2}}}$. We emphasize that $(\ )$ is a random variable that is determined by . For conciseness, however, we will use and keep the dependency implicit. We nd the solution of (5) by solving an equivalent convex optimization problem, which we provide in Appendix B. We need to characterize the distribution of (5) under the null hypothesis, i.e., $N (0; \ )$. Since we are not able to give an analytic form for, it is hard to directly compute the distribution of $_{OST}$ as we did for the Wald test. In Section 3.1 we were able to work around this by deriving the distribution conditioned on the selection of In the present case, however, there are uncountably many values that can take, so for some the probability is zero. Hence, the reasoning of (2) does not apply and we cannot use the PSI framework of Lee et al. [24].

Our approach to solving this is the following. Instead of directly conditioning on the explicit value of , we condition on the active set For a given , we de ne the active set as $U := f u j \ _u \neq 0 g$ [d]. Note that the active set is a function of, de ned via (5). In Theorem 1 we show that given the active set, we can derive a closed-form expression for, and we can characterize the distribution of the test statistic conditioned on the active set. Figure 1 depicts the intuition behind Theorem 1 and Appendix A contains the full proof. In the following, let denote a chi distribution with degrees of freedom and $TN (0; 1; a; 1)$ denote the distribution of a standard normal RV truncated from below at a, i.e., with CDF $F^a(x) = \frac{(x) \ (a)}{1 \ (a)}$:

Theorem 1. Let $N (0;)$ be a normal RV in R^d with positive de nite covariance matrix . Let be de ned as in (5), $U := f u j \ _u \neq 0 g$, $l := jUj$, $z := \ l_d \ \frac{^>}{^> \ }$, and V as in Corollary 1. Then, the following statements hold.

1.) If $l = 1$: $\max_k \ _{k=1 ; \quad 0} \frac{^>}{(\ ^> \ )^{\frac{1}{2}}} U; z = z \overset{d}{=} TN(0; 1; V \ (z); 1)$:

2.) If $l \quad 2$: $\max_k \ _{k=1 ; \quad 0} \frac{^>}{(\ ^> \ )^{\frac{1}{2}}} U \overset{d}{=} \ _l$:

With Theorem 1 and Remark 1, we are able to de ne conditional hypothesis tests with the test statistic $_{OST}$. First, we transform our observation according to Remark 1 to obtain it in canonical form, i.e., $\hat{} \ ! \quad ^1 \hat{}$ and $\ ! \quad ^1$. Then we solve the optimization problem (5) to nd . Next, we de ne the active set U, by checking which entries of are non-zero. Theorem 1 characterizes the distribution $_{OST}$ conditioned on the selection. We can then de ne a test threshold that accounts for the selection of U, i.e.,

$$t = \begin{cases} ^1((1 \quad)(1 \quad (V)) + (V)) & \text{if } jUj = 1 ; \\ \ _l^1(1 \quad) & \text{if } jUj = l \quad 2; \end{cases} \tag{6}$$

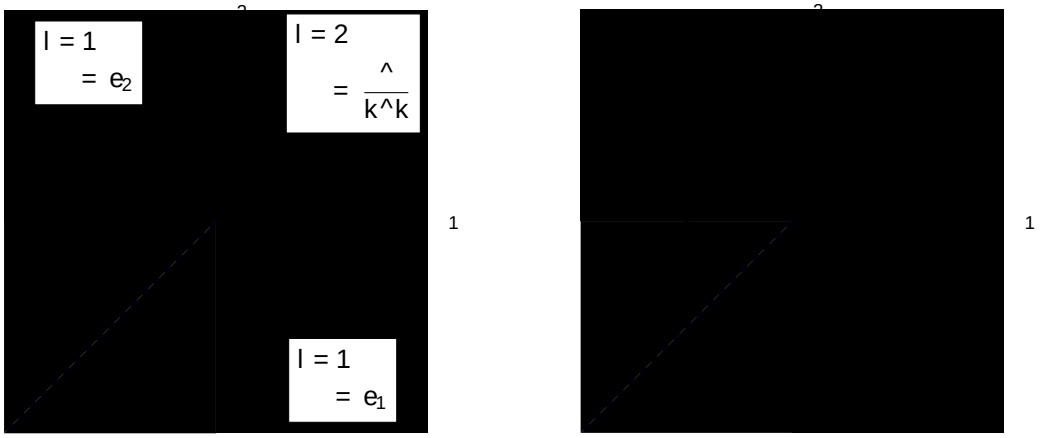

Figure 1: Geometric interpretation of Theorem 1 for = 2 and unit covariance = I (denoted by the black dotted unit-circle). Left: If ^ is in the positive quadrant (green), the constraints of the optimization are not active and the optimal direction is the same as for the Wald test, hence the distribution of the test statistic follows 2. When ^ is in the orange or purple zone, one of the constraints is active and is a canonical unit-vector. Right: If l = 1, for example when only the rst direction is active, we additionally condition on z = z, which is independent of the value of > since z is orthogonal to . For the observed value z, we only select = e1 if > V . If this was not the case, then would lie in the orange/vertically lined region and we would select = e2. This explains the truncated behavior and is in analogy to the results of Lee et al. [24].

with 1 being the inverse CDF of a chi distribution with degrees of freedom, which we can evaluate using standard libraries, e.g., Jones [26]. We can then reject the null, if the observed value of the optimized test statistic exceeds this threshold, i.e., > t . We summarize the entire approach in Algorithm 1.

4 Related work

Our work is best positioned in the context of modern statistical tests with tunable hyperparameters. Gretton et al. [4] were the rst to propose a kernel two-sample test that optimizes the kernel hyperparameters by maximizing the test power. This in uential work has led to further development of optimized kernel-based tests [7–12]. Since any universally consistent binary classi er can be used to construct a valid two-sample test, [27, 28], Kim et al. [14], Lopez-Paz and Oquab [15] used classi cation accuracy as a proxy to train machine learning models for two-sample tests. Kirchler et al. [17], Cai et al. [29] studied this further, and Cheng and Cloninger [16] proposed using the difference of a trained deep network's expected logit values as the test statistic for two-sample tests.

All the aforementioned “learn-then-test” approaches optimize hyperparameters (e.g., kernels, weights in a network) on a training set which is split from the full dataset. While the null distribution becomes tractable due to the independence between the optimized hyperparameters and the test set, there is a potential reduction of test power because of a smaller test set. This observation is the main motivation for our consideration of selective hypothesis tests, which allow the full dataset to be used for both training and testing by correcting for the dependency, as we discuss in Section 3.

More broadly, properly assessing the strength of potential associations that have been previously learned from the data falls under an emerging sub eld of statistics known as selective inference [30]. A seminal work of Lee et al. [24] proposed a post-selection inference (PSI) framework to characterize the valid distribution of a post-selection estimator where model selection is performed by the Lasso [31]. The PSI framework has been applied to kernel tests, albeit in different context, for selecting the most informative features for supervised learning [32, 33], selecting a subset of features that best discriminates two samples [34], as well as selecting a model with the best t from a list of candidate models [35]. All these applications of the PSI framework consider a nite candidate set. Our Theorem 1 can be seen as an extension of the previously known results of Lee et al. [24].

uncountable candidate sets. To our knowledge, our work is the first to explicitly maximize test power by using the same data for selecting and testing.

Unfortunately, we cannot directly use our results to optimize tests based on complete U-statistics estimates of the MMD, which would be desirable since those estimates have lower variance than the linear version we use. The difficulty arises since our method requires asymptotic normality under the null, which is not the case for complete U-statistics [33]. To circumvent this problem, Yamada et al. [34] considered incomplete U-statistics [36] and Zaremba et al. [37] used a Block estimate of the MMD. Under the null, these approaches either have approximately asymptotic normal distribution [34] or require a higher sample size to reach the asymptotic normality [37]. In principle thus our approach is applicable with these methods if one is willed to assume asymptotic normality and to neglect the induced errors. Besides that, since the linear-time estimate has lowest computational cost, it should generally be used in the large-data, constraint-computation regime [4]. On the other hand one should consider the other approaches when the computational efforts are not the limiting factor.

Moreover, under the assumption that $N(\cdot;\cdot)$, similar scenarios have previously been investigated in the traditional statistical literature, but the idea of data splitting is not considered there. In particular, our construction of t_{Wald} turned out to coincide with the test statistic suggested in Wald [25]. The one-sided version t_{OST} also has a twin named "chi-bar-square" test previously considered in Kudo [38]. While their test statistic is constructed to be always non-negative, t_{OST} can be negative. Furthermore, they derived the distribution of the test statistic by decomposing the distribution into 2^d selection events, which, however, may represent a quite difficult problem [39, p. 54]. Our work circumvents this difficulty by defining a conditional test, which does not require calculating any probability of the selection events. Another difference is that our approach only defines $2^d - 1$ different active sets, by enforcing $\delta \neq 0$. It is instructive to note that there exist other more complicate settings of "learn-then-test" scenarios in which the normality assumption may not hold [15–17, 29]. Extending our work towards these scenarios remains an open, yet promising problem to consider.

5 Experiments

We demonstrate the advantages of OST over data-splitting approaches and the Wald test with kernel two-sample testing problems as described in Section 2. For an extensive description of the experiments we refer to Appendix D. We consider three different datasets with different input dimensions p. 1. DIFF VAR ($p = 1$): $P = N(0;1)$ and $Q = N(0;1.5)$. 2. MNIST ($p = 49$): We consider downsampled 7x7 images of the MNIST dataset [40], where P contains all the digits and Q only uneven digits. 3. Blobs ($p = 2$): A mixture of anisotropic Gaussians where the covariance matrix of the Gaussians have different orientations for P and Q. We denote by k_{lin} the linear kernel, and $k_\sim$ the Gaussian kernel with bandwidth $\sim$. For each dataset we consider three different base sets of kernels K and choose $\sim$ with the median heuristic: (a) $d = 1 : K = [k_\sim]$, (b) $d = 2 : K = [k_\sim; k_{\text{lin}}]$, (c) $d = 6 : K = [k_{0.25\sim}; k_{0.5\sim}; k_\sim; k_{2\sim}; k_{4\sim}; k_{\text{lin}}]$. From the base set of kernels we estimate the base set of test statistics using the linear-time MMD estimates. We compare four different approaches: i) OST, ii) WALD, iii) SPLIT: Data splitting similar to the approach in Gretton et al. [4], but with the same constraints as in OST. SPLIT 0.1 denotes that 10% of the data are used for learning δ and 90% are used for testing, iv) NAIVE: Similar to splitting but all the data is used for learning and testing without correcting for the dependency. The NAIVE approach is not a well-calibrated test. For all the setups we estimate the Type-II error for various sample sizes at a level $\alpha = 0.05$. Error rates are estimated over 5000 independent trials and the results are shown in Figure 2. In Appendix D.1, we also investigate the Type-I error and show that all methods except for NAIVE correctly control the Type-I error at a rate α. Note that all of the methods scale with $\mathcal{O}(n)$ and the difference in computational cost are negligible.

The experimental results in Figure 2 support the main claims of this paper. First, comparing OST with SPLIT, we conclude that using all the data in an integrated approach is always better (or equally good) than any data splitting approach. Second, comparing OST to WALD, we conclude that adding a priori information ($\delta \neq 0$) to reduce the class of considered tests in a sensible way leads to higher (or equally high) test power. Another interesting observation is in the results of the data-splitting approach. Looking at the DIFF VAR experiment, in the leftmost plot, we can see that the errors are monotonically increasing with the portion of data used to select the test. Since there is only one test, the more data we use to select the test, the higher the error (less data remains for testing). In the

Figure 2: Type-II errors obtained from different experiments. The rows (columns) correspond to different datasets (sets of base kernels). For all considered cases, C2ST outperforms all the (well-calibrated) competing methods, i.e. FSST and WALD.

middle plot, selection becomes important. Hence, we can see that the gap in performance between all data-splitting approach reduces. However, the order is still consistent with the previous plot. Interestingly, in the rightmost plot, learning becomes even more important. Now, the order changes. If we use too little data for learning the test (SPLIT 0.1), the error is high. However, if we use too much data for learning the test (SPLIT 0.8), the error will be high as well. That is, there is a trade-off in how much data one should use for selecting the test, and for conducting the test. The optimal proportion depends on the problem and can thus in general not be determined a priori.

In the Appendix D.3 we also compare k_{base} to a selection of a base test via the data-splitting approach. Here, SPLIT 0.1 consistently performs better than the other split approaches, which is plausible, since the class of considered tests $\mathcal{T}_{k_{\text{base}}}$ is quite small. SPLIT 0.1 can even be better than k_{base}, see discussion in Appendix D.3.

In Figure 3, we additionally consider a constructed 50D dataset where the distributions share the first three moments and all uneven moments vanish (Figure 7 in the appendix). We compare the results for different sets of $d \in [5]$ base kernels $\mathcal{K} = [k_{\text{pol}}^1, \dots, k_{\text{pol}}^d]$, where $k_{\text{pol}}^u(x, y) = (x^\top y)^u$ denotes the homogeneous polynomial kernel of order u. By construction, k_{pol}^u does not contain any information about the difference of P and Q, for $u \in \{1, 3\}$. Thus, for $d = 3$ the well-calibrated methods have a Type-II error of $1 - \alpha$. Only the NAIVE approach already overfits to the noise. Adding the fourth order polynomial adds helpful information and all the methods improve performance. However, adding the fifth order, which again only contains noise, leads to an increased error rate. We interpret this as bias-variance tradeoff that should be considered in the construction of the base set $\mathcal{K}$.

Algorithm 1 One-Sided Test (OST)

input , $\hat{} = \sqrt{n}\mathrm{MMD}^2(P;Q)$,
$\hat{}^{\top}\hat{} = {}^{-1}\hat{}$ {Apply Remark 1}
$^{\top} = {}^{-1}$ {Apply Remark 1}
$= \operatorname{argmax}_{k\ k=1;\ 0} \frac{{}^{\top}\hat{}}{({}^{\top}{}^{\top})^{\frac{1}{2}}}$
$U = \{u \mid u \in [d];\ {}_u > 0\}$
$\hat{z} = \hat{} \frac{}{{}^{\top}\hat{}}$
$l = |U|$
if $l \geq 2$ then
$t = {}_{l}^{-1}(1 \quad)$
if $l = 1$ then
$V = \max_{u \in U} \frac{\hat{z}_u ({}^{\top}{}^{\top})^{\frac{1}{2}}}{{}_{uu}^{\frac{1}{2}} ({}^{\top}{}^{\top})^{\frac{1}{2}} ({})_u}$
$t = {}^{-1}((1 \quad)(1 \quad (V)) + (V))$
if $t < \frac{{}^{\top}\hat{}}{({}^{\top}{}^{\top})^{\frac{1}{2}}}$ then
Reject H_0

Figure 3: Type-II errors when the rst polynomial kernels are used for a two-sample test with symmetric distributions with the equal covariance (Figure 7 in the appendix). OST outperforms all the (well-calibrated) competitors.

In Appendix D.2 we compare how the constraints 0, as suggested in Gretton et al. [4], work in comparison to the OST approach. We nd that while the constraints 0 lead to consistently higher power than the Wald test, the simple positivity constraints can lead to both, better or worse power depending on the problem. We thus recommend using the OST.

6 Conclusion

Previous work used data splitting to exclude dependencies when optimizing a hypothesis test. This work is the rst step towards using all the data for learning and testing. Our approach uses asymptotic joint normality of a prede ned set of test statistics to derive the conditional null distributions in closed form. We investigated the example of kernel two-sample tests, where we use linear-time MMD estimates of multiple kernels as a base set of test statistics. We experimentally veri ed that an integrated approach outperforms the existing data-splitting approach of Gretton et al. [4]. Thus data splitting, although theoretically easy to justify, does not ef ciently use the data. Further, we experimentally showed that a one-sided test (OST), using prior information about the alternative hypothesis, leads to an increase in test power compared to the more general Wald test. Since the estimates of the base test statistics are linear in the sample size and the null distributions are derived analytically, the whole procedure is computationally cheap. However, it is an open question whether and how this work can be generalized to problems where the class of candidate tests is not directly constructed from a base set of jointly normal test statistics.

Broader impact

Hypothesis testing and valid inference after model selection are fundamental problems in statistics, which have recently attracted increasing attention also in machine learning. Kernel tests such as MMD are not only used for statistical testing, but also to design algorithms for deep learning and GANs [41, 42]. The question of how to select the test statistic naturally arises in kernel-based tests because of the kernel choice problem. Our work shows that it is possible to overcome the need of (wasteful and often heuristic) data splitting when designing hypothesis tests with feasible null distribution. Since this comes without relevant increase in computational resources we expect the proposed method to replace the data splitting approach in applications that t the framework considered in this work. Theorem 1 is also applicable beyond hypothesis testing and extends the previously known PSI framework proposed by Lee et al. [24].

Acknowledgments and Disclosure of Funding

The authors thank Arthur Gretton, Will Fithian, and Kenji Fukumizu for helpful discussion. JMK thanks Simon Buchholz for helpful discussions and pointing out a simpli cation of Lemma 2.

References

[1] J. Neyman and E. S. Pearson. On the problem of the most ef cient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Charact , 231:289–337, 1933.

[2] E. L. Lehmann and Joseph P. Roman . Testing statistical hypotheses . Springer Texts in Statistics. Springer, third edition, 2005.

[3] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test . Journal of Machine Learning Resear , 13:723–773, 2012.

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

[5] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of- t tests. In ICML, 2016.

[6] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of t. In ICML, 2016.

[7] Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In ICLR, 2017.

[8] Meyer Scetbon and Gael Varoquaux. Comparing distributions: L1 geometry improves kernel two-sample testing. In NeurIPS, 2019.

[9] Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In NeurIPS, 2016.

[10] Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton. Informative features for model comparison. In NeurIPS, 2018.

[11] Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of- t test. In NeurIPS, 2017.

[12] Wittawat Jitkrittum, Zoltán Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In ICML, 2017.

[13] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests. arXiv:2002.09116, 2020.

[14] Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classi cation accuracy as a proxy for two sample testing. arXiv:1602.02210, accepted to Annals of Stat., 2016.

[15] David Lopez-Paz and Maxime Oquab. Revisiting classi er two-sample tests. In ICLR, 2017.

[16] Xiuyuan Cheng and Alexander Cloninger. Classi cation logit two-sample testing by neural networks. arXiv:1909.11298, 2019.

[17] Matthias Kirchler, Shahryar Khorasani, Marius Kloft, and Christoph Lippert. Two-sample testing using deep learning. arXiv:1910.06239, 2019.

[18] R.A. Fisher. The design of experiments . Oliver and Boyd, 1935.

[19] T.W. Anderson. An Introduction to Multivariate Statistical Analysis . Wiley Series in Probability and Statistics. Wiley, 2003.

[20] Robert J Ser ing. Approximation theorems of mathematical statistics . John Wiley & Sons, 1980.

[21] Bernhard Schölkopf and Alexander Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond . MIT Press, Cambridge, MA, USA, 2002.

[22] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. Journal of Machine Learning Research, 11:1517–1561, 2010.

[23] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. arXiv:1410.2597v4, 2017.

[24] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. Ann. Statist., 44(3):907–927, 06 2016.

[25] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 54(3):426–482, 1943.

[26] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scienti c tools for Python, 2001.

[27] Jerome H. Friedman. On multivariate goodness of t and two sample testing. stConf C030908: THPD002, 2003.

[28] Kenji Fukumizu, Arthur Gretton, Gert R. Lanckriet, Bernhard Schölkopf, and Bharath K. Sriperumbudur. Kernel choice and classi ability for rkhs embeddings of probability distributions. In NeurIPS, 2009.

[29] Haiyan Cai, Bryan Goggin, and Qingtang Jiang. Two-sample test based on classi cation probability. Statistical Analysis and Data Mining: The ASA Data Science Journal, 13(1):5–13, 2020.

[30] Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. Proceedings of the National Academy of Sciences, 112(25):7629–7634, 2015.

[31] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B), 58:267–288, 1996.

[32] Makoto Yamada, Yuta Umezu, Kenji Fukumizu, and Ichiro Takeuchi. Post selection inference with kernels. In AISTATS, 2018.

[33] Lot Slim, Clément Chatelain, Chloe-Agathe Azencott, and Jean-Philippe Vert. kernelPSI: a post-selection inference framework for nonlinear variable selection. ICML, 2019.

[34] Makoto Yamada, Denny Wu, Yao-Hung Hubert Tsai, Hirofumi Ohta, Ruslan Salakhutdinov, Ichiro Takeuchi, and Kenji Fukumizu. Post selection inference with incomplete maximum mean discrepancy estimator. In ICLR, 2019.

[35] Jen Ning Lim, Makoto Yamada, Bernhard Schölkopf, and Wittawat Jitkrittum. Kernel Stein tests for multiple model comparison. In NeurIPS, 2019.

[36] Svante Janson. The asymptotic distributions of incomplete u-statistics. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 66(4):495–505, Sep 1984.

[37] Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In NeurIPS, pages 755–763, 2013.

[38] Akio Kudo. A multivariate analogue of the one-sided test. Biometrika, 50(3/4):403–418, 1963.

[39] A. Shapiro. Towards a uni ed theory of inequality constrained testing in multivariate analysis. International Statistical Review / Revue Internationale de Statistique, 56(1):49–62, 1988.

[40] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2, 2010.

[41] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. ICML, 2015.

[42] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In NeurIPS, 2017.

[43] Lieven Vandenberghe. The cvxopt linear and quadratic cone program solvers. 2010.

A Proof of Theorem 1

In this section we prove the main theorem. The outline of the proof is as follows: We first characterize the "selection event", i.e., we characterize under which conditions each active set is selected. This is done with Lemmas 2 and 3. For the case = 1 we then show that the PSI framework of Lee et al. [24] can be applied and we recover the result of Corollary 1. It is not surprising, that for the = 1 case the PSI framework works, since corresponds to a single fixed and the probability of selecting it is greater than 0. For the case = 2, we show, that the considered test statistic essentially takes the same form as the Wald test but only on the active dimensions. Thus it follows a distribution. This distribution does not change even if we explicitly condition on the selection of . This is because the randomness that determines which active set is selected is independent of the value of the selected test statistic. Before we start with the proof we collect some notation we introduce for the proof.
Notation:

The objective of the optimization is $f(\eta) := \frac{\eta^\top \Sigma \eta}{(\eta^\top \eta)^{\frac{1}{2}}}$.

Projector onto the active subspace (leaving the dependency on implicit):

$$:= \sum_{u \in U} e_u e_u^\top ;$$

where e_u denotes the u-th Cartesian unit vector in $\mathbb{R}^d$.

$z := \left(I_d - \frac{\eta \eta^\top}{\eta^\top \eta}\right) = \frac{\eta^\top}{\eta^\top \eta}$.

denotes the pseudoinverse of .

As a first step, we need to characterize which values of correspond to which active set . This is done with Lemma 2, which we prove separately in A.1.

Lemma 2. Let $U := \{u \mid \eta_u \neq 0\}$. Then,

$$= \operatorname*{argmax}_{k \; k=1,\ldots,|U|} \frac{\eta^\top}{(\eta^\top \eta)^{\frac{1}{2}}}$$

if and only if all of the following conditions hold:

1. $\frac{\partial}{\partial \eta_u} \frac{\eta^\top}{(\eta^\top \eta)^{\frac{1}{2}}} = \begin{cases} 0 & \text{if } u \in U \quad \text{(a)}; \\ = 0 & \text{if } u \in U \quad \text{(b)}; \end{cases}$

2. $\frac{\eta^\top}{(\eta^\top \eta)^{\frac{1}{2}}} \quad p \frac{\eta_u}{\eta_u \eta_u}; \qquad \forall u \in U,$

3. $\begin{cases} \eta_u = 0 & \forall u \in U \quad \text{(a)}; \\ \eta_u > 0 & \forall u \in U \quad \text{(b)}, \\ k \quad k = 1 & \text{(c)}. \end{cases}$

Intuitively, Condition1(b) ensures that is a local maximum of the objective function for the active dimensions. Condition 1(a) ensures that $\forall u \notin U$, increasing η_u does not improve the SNR. Condition 2 is harder to interpret, but is needed in cases where all entries of are negative. Condition 3 enforces that lies in the feasible set of (5).

Note that $\eta^\top$ is essentially a one-dimensional RV. We define another random variable

$$z := \left(I_d - \frac{\eta \eta^\top}{\eta^\top \eta}\right) = \frac{\eta^\top}{\eta^\top \eta}: \tag{7}$$

In Appendix A.2, we show that is closely related to the partial derivatives of the objective function and we have

$$\frac{\partial}{\partial \eta_u} \frac{\eta^\top}{(\eta^\top \eta)^{\frac{1}{2}}} = \frac{z}{(\eta^\top \eta)^{\frac{1}{2}}}: \tag{8}$$

We can then rewrite the conditions of Lemma 2 as follows.

Lemma 3. The conditions of Lemma 2 are equivalent to

1. $z_u \geq 0 \quad \forall u \in U$ (a);
   $z_u = 0 \quad \forall u \notin U$ (b);
2. $\frac{\Sigma^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}}} \in V(z)$, with
   $V(z) := \max_{u \in U} \frac{z_u (\Sigma^{>}\Sigma)^{\frac{1}{2}}}{\frac{1}{2}_{uu} (\Sigma^{>}\Sigma)^{\frac{1}{2}} (\Sigma)_u}$;
3. $_u = 0 \quad \forall u \in U$ (a);
   $_u > 0 \quad \forall u \notin U$ (b),
   $k \quad k = 1$ (c).

Proof of Lemma 3. Condition 1 directly follows from (8). The second condition follows by inserting the de nition of z

$$
\begin{aligned}
& \frac{\Sigma^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}}} \; p\frac{}{uu} \\
, \; & \frac{\Sigma^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}}} \; p\frac{z_u}{uu} + e_u^{>} \; \frac{\Sigma^{>}}{\Sigma^{>}} \; p\frac{}{uu} \\
, \; & \frac{\Sigma^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}}} \; 1 \; \frac{e_u^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}} p\frac{}{uu}} \; p\frac{z_u}{uu} \\
, \; & \frac{\Sigma^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}}} \; (\Sigma^{>}\Sigma)^{\frac{1}{2}} p\frac{}{uu} \; e_u^{>} \; z_u (\Sigma^{>}\Sigma)^{\frac{1}{2}} \\
, \; & \frac{\Sigma^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}}} \; \frac{z_u (\Sigma^{>}\Sigma)^{\frac{1}{2}}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}} p\frac{}{uu} \; e_u^{>}};
\end{aligned}
$$

where we used $\frac{1}{2}_{uu} (\Sigma^{>}\Sigma)^{\frac{1}{2}} (\Sigma)_u > 0$, which holds since Σ is positive and we only consider u such that $e_u \neq$. □

Note that $V(z)$ is always non-positive by Condition 1 and the positivity of With the above two lemmas we are able to prove Theorem 1.

Proof of Theorem 1. We prove the two cases $l = 1$ and $l \geq 2$ separately.

1.): Let $u \in [d]$ such that $U = \{u\}$. In this case, by Condition 3, $= e_u$. We shall now see how Lemma 3 constrains the distribution of . For Condition 1(b), we have $z_u = 0$ by the de nition of z. So there only remain the constraints 1(a) and 2. Using the de nition of z, we can rewrite 1(a) as

$$
\left(I_d \quad e_u \frac{e_u^{>}}{u \; u} \right) \quad 0 \quad \forall u \in U \quad \Leftrightarrow \quad A^{[1(b)]} \quad 0;
$$

where $A^{[1(b)]}$ is the matrix $I_d \quad e_u \frac{e_u^{>}}{u \; u}$ and we used that its u-th row contains only zeros. Note that Condition 2 is the same as used in Section 3.1. Thus we can de ne the matrix $A^{[2]}$ as we do in the proof of Corollary 1. We have now all the remaining constraints as linear inequalities of and thus we can nd the conditional distribution by applying Theorem 2. De ning $= \frac{e_u}{(\Sigma \; T \; \Sigma)^{\frac{1}{2}}}$ and $c := \quad ^{>} \quad ^{1}$, we get $A^{[1(b)]} c = 0$. Note that whenever $(Ac)_j = 0$, the constraint does not change anything in Theorem 2. Thus the result follows by using $A = A^{[2]}$ and application of Theorem 2.

An alternative proof can be done by noting that z is independent of $\frac{\Sigma^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}}}$ if we consider $= e_u$ as xed. Thus, the ful llment of Condition 1b) is independent of $\frac{\Sigma^{>}}{(\Sigma^{>}\Sigma)^{\frac{1}{2}}}$. Since the

unconditional distribution of $\frac{\eta^\top}{(\eta^\top \Sigma \eta)^{\frac{1}{2}}}$ follows a standard normal, adding Condition 2 results in a truncated normal.

2.) Next, we consider the case $|U| \geq 2$. Again we will be considering the conditions as stated in Lemma 3. As we state in (15), we have $\nu^\top \eta \geq 0$ and thus Condition 2 is fulfilled, since v^0 is always non-positive. Thus, we can neglect Condition 2. Our first step will be to find a closed form function h_U such that $\eta = h_U(\eta)$ (this function will only hold true if U is actually the active set). Defining the projector onto the active subspace $P = \sum_{u \in U} e_u e_u^\top$, by Condition 3(a) we have $P\eta = \eta$. Using (7), we can rewrite Condition 1(b) as

$$z = 0 \overset{(7)}{\Leftrightarrow} P\eta = \frac{\Sigma\eta}{\eta^\top \Sigma \eta} \overset{3(a)}{\Leftrightarrow} P\eta = \frac{P\Sigma\eta}{\eta^\top \Sigma \eta}. \tag{9}$$

This defines a system of non-trivial equations and by Condition 3, has $|U|$ free parameters. We define A as the pseudoinverse of $P\Sigma$.[3] For the pseudoinverse it is easy to show that $AP\Sigma = P$. Since $P\Sigma$ has full rank, a possible solution of (9) necessarily has to be of the form $\eta = cA\eta$ for some $c \in \mathbb{R}$. Plugging this into (9), we get $c = \frac{\eta^\top \Sigma \eta}{\eta^\top A \eta}$. Using (15) we get $0 < \frac{\eta^\top \Sigma \eta}{(\eta^\top A \eta)^{\frac{1}{2}}} = \frac{1}{c}$. Hence, $c > 0$. Using $\eta^\top A \eta = 1$ we get $c = \frac{1}{\eta^\top \Sigma \eta}$. Thus, given that the active set is U, we found a closed-form solution for η as a function of Σ, i.e.,

$$\eta = h_U(\Sigma) := \frac{A\eta}{\eta^\top \Sigma \eta}. \tag{10}$$

Note that so far we did not use Condition 3(b), so this formula itself does not ensure the positivity of η.

Replacing η in the definition (7) of z with its closed form, the constant c cancels, and we get

$$z = \frac{A\eta}{\eta^\top \Sigma \eta}.$$

Note that $P\Sigma = \Sigma P$ and $(A)_{uu'} = A_{uu'}$ if $u, u' \in U$. This implies that $z_u = 0$ if $u \in U$ and thus also $z^\top \eta = 0$.

Let us now define $\tilde{X} := (\Sigma)^{\frac{1}{2}}\eta$, resulting in $\tilde{X}_u = 0$ for all $u \in U$. Since $\tilde{X}$ and z are both linear transformations of η they are jointly normally distributed. In Appendix A.3 we show that $\tilde{X}$ and z are uncorrelated. This, together with the joint normality, implies that they are independent, i.e.,

$$\tilde{X} \perp z. \tag{11}$$

Further the non-zero coordinates of $\tilde{X}$ are jointly distributed according to a $|I|$-dimensional standard normal distribution. Hence, its euclidean norm follows a chi-distribution

$$\|\tilde{X}\|_2 \sim \chi_{|I|}. \tag{12}$$

Let us summarize how we used all the conditions of Lemma 3 and finish the proof. We used 1(b), 3(a), and 3(c) to show (10). We thus still need to condition on 1(a), and 3(b). Conditioning on 1(a) can be done using the independence of z and $\tilde{X}$. To condition on 3(b), we rewrite it in terms of $\tilde{X}$, i.e., for all $u \in U$ we have

$$\nu_u > 0, \quad \eta_u, \quad (\Sigma)^{\frac{1}{2}}\tilde{X}_u > 0, \quad (\Sigma)^{\frac{1}{2}}\frac{\tilde{X}}{\|\tilde{X}\|_2}_u > 0.$$

Thus it only depends on the direction of $\tilde{X}$. Since the non-trivial entries of $\tilde{X}$ follow a standard normal, the direction of $\tilde{X}$ is independent of its norm, i.e.,

$$\|\tilde{X}\|_2 \perp \frac{\tilde{X}}{\|\tilde{X}\|_2}. \tag{13}$$

[3] For intuition, assume WLOG that $U = \{1, \dots, |I|\}$. The pseudoinverse is then simply the inverse of the blockmatrix padded with zeros.

Figure 4: Numerical veri cation of Theorem 1. For the histogram, we generate a random covariance matrix $\in \mathbb{R}^{4\times 4}$ and sample $\sim \mathcal{N}(0,)$. We solve (5) and only accept the samples for which the active set is $U = \{1, 2\}$. The orange line is the theoretical distribution according to Theorem 1, which is given by a chi distribution with two degrees of freedom. For the speci c example the acceptance rate is $\mathbb{P}(U = \{1, 2\}) \approx 4\%$.

In the end we get

$$\frac{}{(\quad)^{\frac{1}{2}}} \text{ Conditions 1, 2, 3} \overset{(10)}{\iff} \frac{^{>}}{(\quad)^{\frac{1}{2}}} \text{ Conditions 1(a), 3(b)}$$

$$\iff \begin{pmatrix} z_u \quad 0 \quad \forall u \in U, \\ (\quad)^{\frac{1}{2}} \frac{X^{-}}{kX^{-}k} \quad u > 0 \quad \forall u \in U \end{pmatrix} \overset{(11)}{\iff} \begin{bmatrix} kX^{-}k_2 \\ \end{bmatrix} \iff \quad .$$

□

A.1 Proof of Lemma 2

Proof of Lemma 2. Since the objective is a homogeneous function of order zero, we can make the proof by considering the optimization without the constraint $k = 1$.

The necessity of the conditions is trivial to show. We thus only show the suf ciency. The fourth condition ensures that is in the feasible set. For the other conditions, assume there exists $\in \mathbb{R}^d$ such that $u \quad 0$ for all $u \in [d]$ and $\frac{^{>}}{(\ ^{>}\ )^{\frac{1}{2}}} > \frac{^{>}}{(\ ^{>}\quad)^{\frac{1}{2}}}$. In the following we show that this implies that at least one of the conditions above is violated, and hence the conditions are suf cient. We separate two cases: i), where $^{>} \quad 0$, and ii) $^{>} < 0$.

i) Assume $\beta^{>} \neq 0$. We have

$$\begin{aligned}
\beta^{>} r \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} &= \\
&= \sum_{u \in [d]} u \frac{\partial}{\partial u} \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} = \\
&= \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} - \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{3}{2}}} \beta^{>} \\
&= \frac{\beta^{>}\beta^{\frac{1}{2}}}{(\beta^{>}\beta)^{\frac{1}{2}}} - \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} - \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} \frac{\beta^{>}\beta}{(\beta^{>}\beta)^{\frac{1}{2}}(\beta^{>}\beta)^{\frac{1}{2}}} \\
&> \frac{\beta^{>}\beta^{\frac{1}{2}}}{(\beta^{>}\beta)^{\frac{1}{2}}} - \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} - \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} \frac{\beta^{>}\beta}{(\beta^{>}\beta)^{\frac{1}{2}}(\beta^{>}\beta)^{\frac{1}{2}}} \\
&= \frac{\beta^{>}\beta^{\frac{1}{2}}}{(\beta^{>}\beta)^{\frac{1}{2}}} \left(1 - \frac{\beta^{>}\beta}{(\beta^{>}\beta)^{\frac{1}{2}}(\beta^{>}\beta)^{\frac{1}{2}}}\right) \\
&> 0,
\end{aligned}$$

where we used the assumption $\frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} > \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}}$ for the first inequality and $\beta^{>}\beta \neq 0$ and the Cauchy-Schwarz inequality to arrive at the last line. Since, by assumption, $\beta_u \neq 0$ for all u, this implies $\frac{\partial}{\partial u} \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} \neq 0$ for some u and thus is a contradiction to Condition 1.

ii) Assume $\beta^{>} < 0$. We define $u = \arg\max_{u \in [d]} \frac{u}{(e_u^{>} e_u)^{\frac{1}{2}}}$. By the third condition and the assumption $\beta^{>} < 0$, we have $0 > \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} \frac{u}{(e_u^{>} e_u)^{\frac{1}{2}}}$. This implies $\beta_u < 0$. We then get

$$\begin{aligned}
\frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} &= \sum_{u \in [d]} u \frac{u}{(\beta^{>}\beta)^{\frac{1}{2}}} = \sum_{u \in [d]} u \frac{u\, e_u^{>} e_u^{\frac{1}{2}}}{(\beta^{>}\beta)^{\frac{1}{2}}(e_u^{>} e_u)^{\frac{1}{2}}} \\
&\sum_{u \in [d]} u \frac{u\, e_u^{>} e_u^{\frac{1}{2}}}{e_u^{>} e_u^{\frac{1}{2}}(\beta^{>}\beta)^{\frac{1}{2}}} \\
&= \frac{u}{e_u^{>} e_u^{\frac{1}{2}}} \frac{\sum_{u \in [d]} u\, e_u^{>} e_u^{\frac{1}{2}}}{(\beta^{>}\beta)^{\frac{1}{2}}} \\
&\frac{u}{e_u^{>} e_u^{\frac{1}{2}}} \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}},
\end{aligned}$$

where to arrive at the last line we used $\beta_u < 0$ and the triangle inequality $\sum_{u \in [d]} u\, e_u^{>} e_u^{\frac{1}{2}} = \sum_{u \in [d]} u k^{\frac{1}{2}} e_u k k \sum_{u \in [d]} u^{\frac{1}{2}} e_u k = k^{\frac{1}{2}} k = \beta^{>}\beta^{\frac{1}{2}}$. Thus this violates the assumption $\frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} > \frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}}$.

Note that the above inequalities also hold for $\frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}}$. Thus we get that $\frac{\beta^{>}}{(\beta^{>}\beta)^{\frac{1}{2}}} = \frac{u}{(e_u^{>} e_u)^{\frac{1}{2}}}$. This implies that $l = |U| = 1$. Thus the following statements hold true:

$$\text{i)} \quad \beta^{>} < 0 \quad \Rightarrow \quad l = 1; \tag{14}$$

$$\text{ii)} \quad l \geq 2 \quad \Rightarrow \quad \beta^{>} \geq 0. \tag{15}$$

□

A.2 Gradient of objective

We overload the notation and de ne $z := \frac{\Sigma\beta}{\sqrt{\beta^\top \Sigma \beta}}$ similar as in (7) but for any β. Then

$$\begin{aligned}
\nabla_\beta f(\beta) &= \nabla_\beta \frac{\Sigma\beta}{(\beta^\top \Sigma \beta)^{\frac{1}{2}}} \\
&= \frac{(\beta^\top \Sigma \beta)^{\frac{1}{2}} \nabla_\beta (\Sigma\beta) - \Sigma\beta \nabla_\beta ((\beta^\top \Sigma \beta)^{\frac{1}{2}})}{\beta^\top \Sigma \beta} \\
&= \frac{(\beta^\top \Sigma \beta)^{\frac{1}{2}} \Sigma - \frac{1}{2} \Sigma\beta ((\beta^\top \Sigma \beta)^{-\frac{1}{2}}) 2\Sigma\beta}{\beta^\top \Sigma \beta} \\
&= \frac{1}{(\beta^\top \Sigma \beta)^{\frac{1}{2}}} \Sigma - \frac{\Sigma\beta}{(\beta^\top \Sigma \beta)} z \qquad (16) \\
&= \frac{1}{(\beta^\top \Sigma \beta)^{\frac{1}{2}}} z.
\end{aligned}$$

A.3 Proof of Equation (11)

In the proof of Theorem 1 we used that $\tilde{X}_t$ and z are independent. Which we prove here. Since $\tilde{X}_t$ and z are jointly normal, we only need to show that they are uncorrelated. To do so recall that we are only interested in the distribution under the null and hence $0 = E[\tilde{X}_t] = E[\tilde{X}_t z] = E[z]$. Since $\tilde{X}_u = 0$ for all $u \in U$ and $z_u^0 = 0$ for all $u^0 \in U$, it suf ces to show that $\tilde{X}_j$ is uncorrelated with z_i for all $i \in U, j \in U$.

$$\begin{aligned}
\mathrm{Cov}[z_i, \tilde{X}_j] &= E[z_i \tilde{X}_j] = E\Big[{}_i (\quad)_i ((\ )^{\frac{1}{2}})_j \Big] \\
&= \sum_{u \in U} ((\ )^{\frac{1}{2}})_{ju} E[{}_i, {}_u] - \sum_{s,t,u \in U} ((\ )^{\frac{1}{2}})_{ju} {}_{is} {}_{st} E[{}_t {}_u] \\
&= \sum_{u \in U} ((\ )^{\frac{1}{2}})_{ju} {}_{iu} - \sum_{s,t,u \in U} ((\ )^{\frac{1}{2}})_{ju} {}_{is} {}_{st} {}_{tu} \\
&= (\ )^{\frac{1}{2}}_{ji} - (\ )^{\frac{1}{2}}_{ij} \\
&= (\ )^{\frac{1}{2}}_{ji} - (\ )^{\frac{1}{2}}_{ij} = 0.
\end{aligned}$$

Thus $\tilde{X}$ and z are uncorrelated and independent.

B Solution of the continuous optimization problem

The presented solution is similarly described in Gretton et al. [14, Sec. 4]. There an L_1 norm constraint was used, which, however does not change anything. For completeness we include it here. We de ne

$$f(\beta) := \frac{\Sigma\beta}{(\beta^\top \Sigma \beta)^{\frac{1}{2}}},$$

and we want to nd

$$\beta^* = \operatorname*{argmax}_{\beta \neq 0, k=1} \frac{\Sigma\beta}{(\beta^\top \Sigma \beta)^{\frac{1}{2}}}.$$

Since f is a homogeneous function of order 0 in β we have $f(c\beta) = f(\beta)$ for any $c > 0$. We can thus solve the relaxed problem (we implicitly exclude $\beta = 0$)

$$\beta^* = \operatorname*{argmax}_{\beta} f(\beta).$$

The solution of the original problem is then simply given as a rescaled version of the relaxed problem $= \frac{^0}{k\ ^0 k}$. We shall solve the relaxed problem for two different cases.

i) $9u \in [d] : {}_u \quad 0$.
In this case, we know that $\max_{\ 0} f(\ ) \quad 0$ and hence $^0 = \operatorname{argmax} f(\ ),\ ^0 = \operatorname{argmax} f^2(\ )$. The set $S := \{ \in R^d \mid 0; f(\ ) \quad 0\}$ is convex and the functions $g_1(\ ) := (\ ^{\top}\ )^2$ and $g_2(\ ) := {}^{\top}\ $ are convex (recall that is a positive matrix). Thus our problem becomes

$$^0 = \operatorname*{argmax}_{\in S} \frac{g_1(\ )}{g_2(\ )};$$

which is a concave fractional program. In our implementation we solve it by xing $^{\top}\ = a$ for some $a > 0$ and then minimizing the denominator. Thus we are solving the quadratic optimization problem

$$\begin{aligned} &\text{minimize} && {}^{\top} \\ &\text{subject to:} && 0 \\ & && {}^{\top}\ = a: \end{aligned}$$

We solve this problem with the CVXOPT python package [43].

ii) ${}_u < 0 \forall u \in [d]$.
In this case we have ${}^{\top}\ < 0$. By (15) we have $l = 1$. Thus we simply $= e_u$, where $u = \operatorname*{argmax}_{u \in [d]} \frac{u}{u;u}$.

Note that in the case $= 0$, is not well de ned and we could randomly select any. However, the probability of this happening is 0.

C Other proofs

C.1 Proof of Corollary 1

As we pointed out in the main paper, when selecting a test from a countable number of test that can be written as projections of the base tests we can use the results of Lee et al. [24]. For completeness we explicitly include the relevant theorem.

Theorem 2 (Polyhedral Lemma 2 [24], Theorem 5.2) Let $N(\ ;\ ),\ ; \in R^d,\ \in R^{d\ d}$ positive de nite, and $A \in R^{s\ d}$, $b \in R^s$ for some $s \in N$. De ne $c := {}^{\top}\ ^1$ and $z := I_d \quad c\ ^{\top}$. Then we have

$$^{\top} \mid A \quad b; z = z \overset{d}{=} TN\left(^{\top}\ ;\ ^{\top}\ ; V^-(z); V^+(z)\right);$$

where $TN(\ ;\ ^2; a; b)$ denotes a Gaussian distribution with mean and variance ^{2} that is truncated at a and b. Here

$$V^-(z) := \max_{j:(Ac)_j < 0} \frac{b_j \quad (Az)_j}{(Ac)_j}; \quad V^+(z) := \min_{j:(Ac)_j > 0} \frac{b_j \quad (Az)_j}{(Ac)_j}:$$

Note that c is simply a xed vector, z is a random variable that can be shown to be independent of $^{\top}$. The result enables us to draw a realization from the random variable (RV) and select if $A\hat{} \quad b$. Since the truncation points of the Gaussian only depend on z, and z is independent of $^{\top}$, we can compute a reliable value of $^{\top}\hat{}$ by using Theorem (2).

Proof of Corollary 1. We need the distribution of $\frac{u}{u;}$ after conditioning on the selection of . To obtain this distribution we rst need to characterize the event that leads to the selection of the

selection event simply is $u = \operatorname{argmax}_{u \in [d]} \frac{u}{u}, \quad \frac{u}{u} \quad \frac{u}{u}$ for all $u \in [d]$. Therefore, define the matrix $A := \operatorname{diag}(\frac{1}{1}; \ldots; \frac{1}{d}) \quad \frac{1}{u} A(u)$, where $\operatorname{diag}(\cdot)$ defines a $d \quad d$ matrix with the arguments on its diagonal and zeros everywhere else and $A(u)$ is a $d \quad d$ matrix with ones in the column given by its argument and zeros everywhere else. It follows that $(AA^{\top})_j = \frac{j}{j} \quad \frac{u}{u}$; and $u = \operatorname{argmax}_{u \in [d]} \frac{u}{u}$ is equivalent to $A \quad 0 =: b$. Apart from this we define $:= \frac{e_u}{u}$, so that $^{\top} \quad = \frac{u}{u}$. Then we can define $c := \quad ^{\top} \quad ^{1}$ and $z := I_d \quad c \, ^{\top}$ as in Theorem 2, and denote by z the value of the random variable z that we observed (note that this coincides with the definition we used for the Corollary). By our definitions we have $(Ac)_j = \frac{ju = j \quad u}{u} = \frac{1}{u \quad j} (\quad u \, j \quad \quad u \, j)$. Since is positive definite, $(Ac)_j < 0$ if $j \in u$ and $(Ac)_u = 0$. Thus according to Theorem 2, Z^{+} is an optimization over an empty set and we can set it to Further $(Az)_j = \frac{1}{u \quad j} \quad j \quad u \quad \frac{u \, j}{u} \quad u \quad .$

Combining the previous two expressions we obtain $\frac{(Az)_j}{(Ac)_j} = \frac{j \quad u \quad \frac{u \, j}{u} \quad u}{u \, j \quad u \, j} = \frac{u \quad z_j}{u \, j \quad u \, j}$. We can then directly apply Theorem 2 and the result follows. □

C.2 Proof of Equation (3)

In the main paper we omitted the proof of the closed form solution of ^{1} We thus need to show

$$\operatorname*{argmax}_{k \quad k=1} \frac{^{\top}}{(\quad ^{\top} \quad)^{\frac{1}{2}}} = \frac{^{1}}{k \quad ^{1} \quad k}:$$

Proof. We are only interested in ^{1} if the alternative hypothesis is true and thus at least one entry of is positive. We further assume that the covariance has full rank. Hence there exists a > 0 such that $^{\top} \quad > b$ for all with $k \quad k = 1$, i.e., the denominator $(\quad ^{\top} \quad)^{\frac{1}{2}}$ is strictly positive and has a lower bound. Since $\in 0$, this implies that $\max_{k \quad k=1} \frac{^{\top}}{(\quad ^{\top} \quad)^{\frac{1}{2}}} > 0$. Also the nominator has an upper bound which is given by $^{\top} \quad ^{\top} = k \quad k$ if $k \quad k = 1$. Hence the whole maximization is upper bounded. Since the unit sphere in $\mathbb{R}^n$ is a compact set, we can conclude that the maximum of the objective is attained. Thus it suffices to show that for all ^{1} the objective is not maximized. In the following, we use that the objective of the maximization is a homogeneous function of order 0 in and hence we can relax the constraint $k \quad k = 1$ to $\in 0$ (note that this not affect the existence of the maximum). As we showed in Appendix A.2, the gradient of the objective function is given by

$$r \quad \frac{^{\top}}{(\quad ^{\top} \quad)^{\frac{1}{2}}} = \frac{1}{(\quad ^{\top} \quad)^{\frac{1}{2}}} \quad \quad \frac{^{\top}}{(\quad ^{\top} \quad)} \quad :$$

Setting the gradient to zero we obtain

$$r \quad \frac{^{\top}}{(\quad ^{\top} \quad)^{\frac{1}{2}}} = 0 \, , \quad = c \quad ^{1} \quad \text{for some } c \in \mathbb{R}:$$

If $c < 0$ the objective attains a negative value, since ^{1} is a strictly positive matrix, and thus does not correspond to the global maximum, which we already know to be positive. Thus, the maximum has to be attained for some $c > 0$. Using the constraint $k \quad k = 1$ it follows that the global optimum is attained at ^{1} . □

D Experimental details and further experiments

We first give some details on the experiments we showed in the main paper. For all the experiments we start with a set of base kernels $K = [k_1; \ldots; k_d]$ that are chosen independently of the observed data samples $X = \{x_1; \ldots; x_{2n}\} \quad P^{2n}$ and $Y = \{y_1; \ldots; y_{2n}\} \quad Q^{2n}$. First, we define $z_i := (x_i, x_{n+i}, y_i, y_{n+i})$ and compile X and Y into $Z = \{z_1; \ldots; z_n\}$. For each kernel we define $h_i(z) := h_i(x; x'; y; y') := k_i(x; x') + k_i(y; y') \quad k_i(x; y') \quad k_i(y; x')$. For all the methods we estimate the covariance matrix on the whole dataset as

$$\hat{}_{ij} = \frac{1}{n} \sum_{k=1}^{n} h_i(z_k) h_j(z_k) \quad \frac{1}{n} \sum_{k=1}^{n} h_i(z_k) \frac{1}{n} \sum_{k'=1}^{n} h_j(z_{k'}):$$

We then further assume that $= \hat{}$ which is justified since the CLT also works with a consistent estimate of the covariance. For all the methods that do not split the data (OST, WALD, and NAIVE) we estimate the entries of as

$$\hat{}_i = \sqrt{n}\,\mathrm{MMD}^2_{\mathrm{lin}}(P,Q) = \sqrt{n}\,\frac{1}{n}\sum_{k=1}^{n} h_i(z_k),$$

i.e., we directly absorb the $\sqrt{n}$ dependence of the asymptotic distribution into . For data splitting we estimate $\hat{}_{\mathrm{tr}}$ on a split of the data and $\hat{}_{\mathrm{te}}$ on the other split. For example, SPLIT 0.3 means that 30% of the data are used to estimate $\hat{}_{\mathrm{tr}}$ and 70% used to estimate $\hat{}_{\mathrm{te}}$. We assume that the number of samples in the respective subsets are even and otherwise neglect some samples.

Methods We compare four different methods:

i) OST: The test we recommend to use, as described in Algorithm 1.

ii) WALD: The Wald test, which does not take into account the prior information 0.

iii) SPLIT: Data splitting similar to the approach in Gretton et al. [4]. SPLIT 0.3 denotes that 30% of the data are used for learning and 70% are used for testing. Here we first, learn on the training sample, i.e., $= \arg\max_{k\ k=1;\ 0} \frac{(\hat{}^{>}_{\mathrm{tr}})^2}{(\hat{}^{>}_{\mathrm{tr}}\hat{}_{\mathrm{tr}})^{\frac{1}{2}}}$. We then use the test statistic $\frac{\hat{}^{>}_{\mathrm{te}}}{(\hat{}^{>}\hat{})^{\frac{1}{2}}}$, which follows a standard normal under the null. This differs from the approach in Gretton et al. [4], since we optimize with the constraints 0, whereas Gretton et al. [4] suggested a simple positivity constraint 0. We discuss this in Section D.2.

iv) NAIVE: Two stage procedure where all the data is used for learning and testing without correcting for the dependency, i.e., without splitting the data. Thus the test statistic is the same as for OST, but we work with the wrong null distribution, i.e., the one that is only valid for data splitting. This approach is not a well-calibrated test, see Fig. 8 and hence is useless.

Datasets The DIFF VAR dataset is a simple one-dimensional toy dataset, where $P = N(0;1)$ and $Q = N(0;1.5)$.

The Blobs dataset was constructed using a mixture of 2D Gaussians on a 3 3 grid. The centers of the Gaussians are set to $_1;\ldots;\ _9 = (0;0);(0;1);(0;2);(1;0);(1;1);(1;2);(2;0);(2;1);(2;2)$ and the covariances are $_P = \mathrm{diag}(0.1;0.3)$ and $_Q = \mathrm{diag}(0.3;0.1)$. Samples from P and Q are shown in Figure 5. The Blobs dataset is constructed such that the main variance in the data does not reflect the difference between P and Q, which happens on a smaller length scale. This is inspired by Gretton et al. [4], where similar data has been considered to showcase that such problems benefit from careful kernel choice. We can reproduce this behavior with our results, which show that for this dataset the performance is bad if one only considers the median heuristic Gaussian kernel together with a linear kernel.

The MNIST dataset was constructed by first downsampling all the images to 7 7 pixels (originally 28 28), by simply averaging over fields of 4 4 pixels. We define P to contain all the digits, while Q only contains uneven digits. For our experiments we draw with replacement from the images in the database. Some samples from both distributions are shown in Figure 6.

Experiments for Figure 3 For Figure 3 we constructed a 1D data set such that both P and Q are symmetric (thus all uneven moments vanish) and have the same variance, see Figure 7.

D.1 Type-I errors

To verify which methods are theoretically justified, i.e., control the Type-I error at a level $= 0.05$, we run the following experiments, similar to the experiments in the main paper, where Q.

1. DIFF VAR ($p = 1$): $P = N(0;1)$ and $Q = N(0;1)$.
2. MNIST ($p = 49$): We consider downsampled 7x7 images of the MNIST dataset [43], where P contains all the digits and $Q = P$.

Figure 5: Samples from the BLOBS dataset.

Figure 6: Samples from downsampled MNIST dataset. P (left) contains all digits, while Q (right) only contains uneven digits.

3. BLOBS ($p = 2$): A mixture of anisotropic Gaussians and $P = Q$.

The results are in Figure 8. All the methods except LIN correctly control the Type-I error at a rate $= 0.05$ even for relatively small sample sizes. Note that all the described approaches rely on the asymptotic distribution. The critical sample size, at which it is safe to use, generally depends on the distributions P and Q and also the kernel functions. A good approach to simulating Type-I errors in in two-sample testing problems is to merge the samples and then randomly split them again. If the estimated Type-I error is significantly larger than, working with the asymptotic distribution is not reliable.

D.2 Comparison of the constraints

In Section 3.2 we motivate to constrain the set of considered to obey 0, thus incorporating the knowledge 0. All our experiments suggest that this constraint indeed improves test power as compared to the general Wald test. In Gretton et al. [4] a different constraint was chosen. There is constrained to be positive, i.e., 0. The motivation for their constraint is that the sum of positive definite (pd) kernel functions is again a pd kernel function [21]. Thus, by constraining 0 one ensures that $k = \sum_{u=1}^{d} {}_u k_u$ is also a pd kernel. While this is sensible from a kernel perspective, it is unclear whether this is smart from a hypothesis testing viewpoint. From the latter perspective we do not necessarily care whether or not de nes a pd kernel. Our approach instead was purely motivated to increase test power over the Wald test. In Figure 9 we thus compare the two different

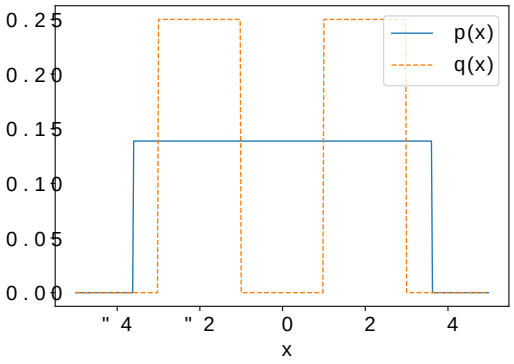

Figure 7: Probability density functions used for the experiment in Figure 3 of the main paper. Both distributions are symmetric and are constructed to have the same variance.

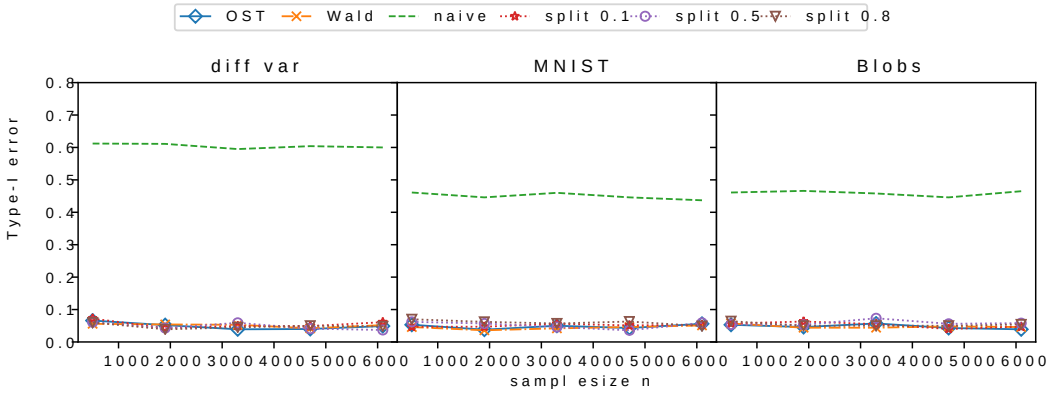

Figure 8: Type-I errors for similar distributions as the one considered in the main paper. To simulate type-I errors we choose distributions $P = Q$ that are similar to the ones considered for the Type-II errors. We see that all well-calibrated methods reliably control the Type-I error at a rate = 0:05, and conclude that working with the asymptotic distributions is well justified for the considered examples. The NAIVE approach fails to control the error, as it overfits in the training phase without a correction in the testing phase.

constraints to the Wald test on the examples that were also investigated in the main paper with $d = 6$ kernels (again five Gaussian kernels and a linear kernel).

From Figure 9 we observe that the positivity constraint of Gretton et al. [4] does not allow for general conclusions. Depending on the problem, the positivity constraint can both lead to higher or lower test power than the Wald test or tests with the constraint 0. It will thus generally depend on the problem at hand which constraint is better. However, at least the approach we recommend (0) seems to guarantee a test power at least as high as the Wald test, whereas the positivity constraint can also be worse. As long as one has not a clear indication that the positivity constraint leads to better performance, we thus recommend the constraint 0.

D.3 Discrete selection from T_{base}

In this experiment, we use the same datasets and base kernels as for the experiment in the main paper. Instead of considering T_{Wald} and T_{OST}, we consider T_{base}. We thus only compare to a data-splitting approach where also one of the base test statistics is selected. For completeness, we also include the NAIVE approach, which again overfits for $d > 1$. Note that the thresholds for $_{\text{base}}$ can be computed with Corollary 1 and do not rely on Theorem 1. The results are shown in Figure 10, again averaged over 5000 independent trials. In most of the cases, we observe that $_{\text{base}}$ outperforms the data-splitting

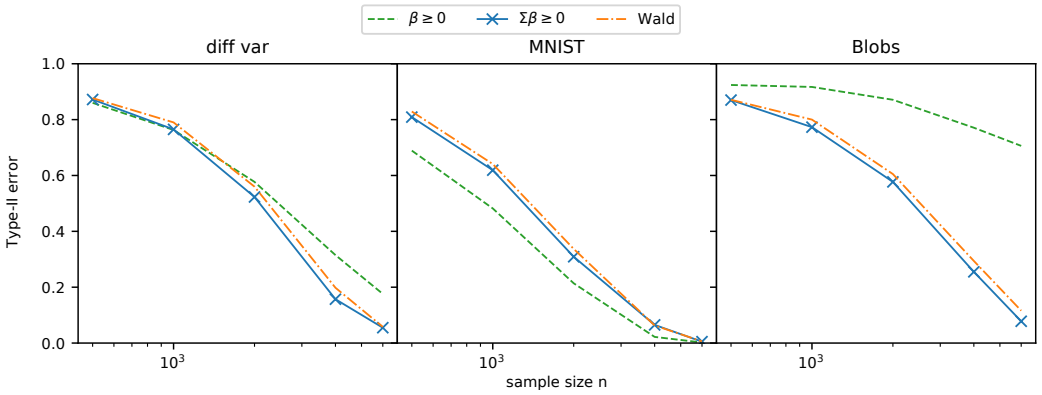

Figure 9: Comparison of the different constraints: In the main paper we argue that OST is a principled approach to constraint the class of considered tests, when 0 is guaranteed. Gretton et al. [4] suggested a different constraint 0. With Theorem 1, we can also work with these constraints without data-splitting. The results suggest that indeed OST is a meaningful way to constrain the class of tests, as it consistently outperforms the Wald test. On the other hand the constraint suggested by Gretton et al. [4], can only be seen as a heuristic. For some cases it performs better than the Wald test and the OST, but it can also perform worse.

approaches. However, for the MNIST dataset and $d = 2$, the splitting approach that uses 10% for learning and 90% for testing does perform slightly better. Our attempt to explain this behavior lies in the truncation V^- of the conditional distribution. While for OST, we can show that V^- 0 (see proof of Theorem 1), for Corollary 1, V^- cannot be bounded. If V^- is very large, the selected test is very conservative. We acknowledge that this is not a sufficient analysis of this phenomenon, but leave a more theoretical treatment for future work.

E Singular covariance matrices

In the main paper we assumed that is strictly positive, i.e., non-singular. However, in practice, some eigenvalues of the covariance matrix can be sufficiently close to zero to cause numerical problems. In the case of the kernel two-sample test, this can happen if we consider kernels that are too similar and thus cause redundancy in our observations. In practice, this happens for example if we consider Gaussian kernels with too similar bandwidths on an easy problem.

Note on regularization: One strategy to recover the numerical stability of the algorithm is to regularize the covariance matrix ! + I. Doing this indeed increases the numerical stability, since it leads to a well-behaved condition number. However, it also makes the whole approach more conservative, since the (artificially) increased variance decreases the value of the test statistic compared to the threshold. This leads to an increase of Type-II error and thus a loss of power. To evade this, we suggest the more elaborate strategy below.

Since is symmetric, there exists an orthonormal basis $fv_i g_{i\in[d]}$ and non-negative numbers $f\ _i g_{i\in[d]}$ such that

$$= \sum_{i=1} {}_i v_i v_i^\top :$$

If is singular, we can assume WLOG that there exists d_0 2 $[d]$ such that $_i = 0$ if i d_0 and hence

$$= \sum_{i=d_0+1} {}_i v_i v_i^\top :$$

Now if $v_i^\top$ 6 0 for some i 2 $[d_0]$, we immediately know that 6 0 and could reject. In other words the signal-to-noise ratio along this direction is infinite. Thus, in the following we assume

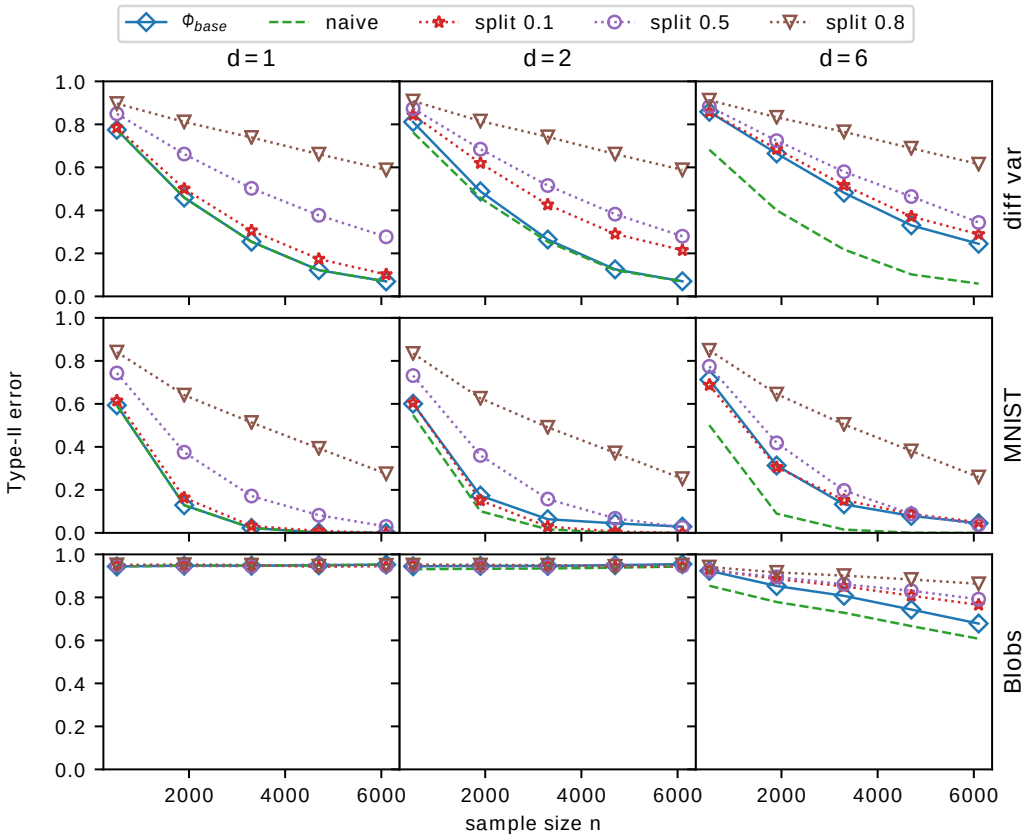

Figure 10: Type-II errors for discrete selection, i.e., the class of considered tests is T_{base}. The rows (columns) correspond to different datasets (sets of base kernels). Similar as in Figure 2, our approach $_{\text{base}}$ outperforms the splitting approaches in most cases. However, for the MNIST dataset and $d = 2$ we see that the splitting approach with 10% training and 90% testing data (SPLIT0.1) performs better.

$v_i^\top = 0$ for all $i \in [d_0]$, and hence, $\sum_{i=d_0+1}^{d} v_i v_i^\top = $. We can then rewrite the objective as follows

$$\max_{\geq \mathbf{0}} \frac{\top}{(\top)^{\frac{1}{2}}} = \sum_{i=d_0+1}^{d} \max_{{}_i v_i v_i^\top \geq \mathbf{0}} \frac{\top \sum_{i=d_0+1}^{d} v_i v_i^\top}{(\top \sum_{i=d_0+1}^{d} {}_i v_i v_i^\top)^{\frac{1}{2}}}.$$

Now define $:= \sum_{i=d_0+1}^{d} {}_i v_i v_i^\top$. Since is symmetric its pseudoinverse is given as $^+ = \sum_{i=d_0+1}^{d} \frac{1}{{}_i} v_i v_i^\top$ and we get

$$\sum_{i=d_0+1}^{d} \max_{{}_i v_i v_i^\top \geq \mathbf{0}} \frac{\top \sum_{i=d_0+1}^{d} v_i v_i^\top}{(\top \sum_{i=d_0+1}^{d} {}_i v_i v_i^\top)^{\frac{1}{2}}} = \max_{\geq \mathbf{0}} \frac{\top \ ^+}{(\top \ ^+)^{\frac{1}{2}}}.$$

Similar as in Remark 1 we can define $:= ^+$ and $' = ^+$. However, in Theorem 1 we assumed that the covariance is not singular. Therefore in Theorem 1 we used $I = jUj$, which corresponded to the rank of (see Appendix A). However, in the present case the rank of $^+$ does not equal the number of non-zero entries of . Therefore we use $I = \text{rank}(^+)$. With this we can apply Theorem 1 and get the conditional distribution under the null.

In practice, we have to treat the covariance matrix as singular if its condition number is below some threshold, as otherwise the numerical precision does not suffice to invert matrices faithfully.