Biometrika (2025), **112**, 1, asae048 https://doi.org/10.1093/biomet/asae048
Advance Access publication 28 September 2024

Boosting the power of kernel two-sample tests

By A. CHATTERJEE and B. B. BHATTACHARYA

Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, 265 37th Street, Philadelphia, Pennsylvania 19104, U.S.A.
anirbanc@wharton.upenn.edu bhaswar@wharton.upenn.edu

Summary

The kernel two-sample test based on the maximum mean discrepancy is one of the most popular methods for detecting differences between two distributions over general metric spaces. In this paper we propose a method to boost the power of the kernel test by combining maximum mean discrepancy estimates over multiple kernels using their Mahalanobis distance. We derive the asymptotic null distribution of the proposed test statistic and use a multiplier bootstrap approach to efficiently compute the rejection region. The resulting test is universally consistent and, since it is obtained by aggregating over a collection of kernels/bandwidths, is more powerful in detecting a wide range of alternatives in finite samples. We also derive the distribution of the test statistic for both fixed and local contiguous alternatives. The latter, in particular, implies that the proposed test is statistically efficient, that is, it has nontrivial asymptotic (Pitman) efficiency. The consistency properties of the Mahalanobis and other natural aggregation methods are also explored when the number of kernels is allowed to grow with the sample size. Extensive numerical experiments are performed on both synthetic and real-world datasets to illustrate the efficacy of the proposed method over single-kernel tests. The computational complexity of the proposed method is also studied, both theoretically and in simulations. Our asymptotic results rely on deriving the joint distribution of the maximum mean discrepancy estimates using the framework of multiple stochastic integrals, which is more broadly useful, specifically, in understanding the efficiency properties of recently proposed adaptive maximum mean discrepancy tests based on kernel aggregation and also in developing more computationally efficient, linear-time tests that combine multiple kernels. We conclude with an application of the Mahalanobis aggregation method for kernels with diverging scaling parameters.

Some key words: Kernel method; Nonparametric two-sample testing; Pitman efficiency; U-statistic.

1. Introduction

Given two probability distributions P and Q on a separable metric space $\mathcal{X}$, the two-sample problem is to test the hypothesis

$$H_0 : P = Q \quad \text{versus} \quad H_1 : P \neq Q, \tag{1}$$

based on independent and identically distributed samples $\mathcal{X}_m := \{X_1, X_2, \ldots, X_m\}$ and $\mathcal{Y}_n := \{Y_1, Y_2, \ldots, Y_n\}$ from distributions P and Q, respectively. This is a classical

problem that has been extensively studied, especially in the parametric regime, where the data are assumed to have certain low-dimensional functional forms. However, parametric methods often perform poorly for misspecified models, especially when the number of nuisance parameters is large, and for non-Euclidean data. This necessitates the development of nonparametric methods, which make minimal distributional assumptions on the data, but remain powerful for a wide class of alternatives.

For univariate data, there are several well-known nonparametric tests such as the two-sample Kolmogorov–Smirnoff maximum deviation test (Smirnov, 1948), the Wald–Wolfowitz runs test (Wald & Wolfowitz, 1940), the rank-sum test (Mann & Whitney, 1947; Wilcoxon, 1947) and the Cramér–von Mises test (Anderson, 1962). Efforts to generalize these univariate methods to higher dimensions date back to Weiss (1960) and Bickel (1969). Thereafter, several nonparametric methods for multivariate two-sample testing have been proposed over the years. These include tests based on geometric graphs (Friedman & Rafsky, 1979; Henze, 1984; Schilling, 1986; Hall & Tajvidi, 2002; Rosenbaum, 2005; Biswas et al., 2014; Chen & Friedman, 2017; Bhattacharya, 2019), tests based on data-depth (Liu & Singh, 1993), the energy distance test (see the 2003 Bowling Green State University technical report by G. J. Székely, Baringhaus & Franz, 2004; Székely & Rizzo, 2004; Aslan & Zech, 2005; Székely & Rizzo, 2013), kernel maximum mean discrepancy tests (Gretton et al., 2009, 2012a,b; Sejdinovic et al., 2013; Chwialkowski et al., 2015; Ramdas et al., 2015, 2017; Shekhar et al., 2022; Song & Chen, 2023; Zhang et al., 2024), ball divergence (Pan et al., 2018; Banerjee & Ghosh, 2022), projection-averaging (Kim et al., 2020) and classifier-based tests (Lopez-Paz & Oquab, 2017; Kim et al., 2021), among others. Recently, a distribution-free version of the energy distance test has been proposed by Deb & Sen (2021) using the emerging theory of multivariate ranks based on optimal transport.

Among the aforementioned methods kernel-based tests have emerged as a powerful technique for detecting distributional differences on general domains. The basic idea is to quantify the discrepancy between the two distributions P and Q in terms of the largest difference in expectation between $f(X)$ and $f(Y)$, for $X \sim P$ and $Y \sim Q$, over functions f in the unit ball of a reproducing kernel Hilbert space (RKHS) defined on $\mathcal{X}$. This is called the maximum mean discrepancy (MMD) between distributions P and Q (see (2) for the precise definition), which can be conveniently estimated from the data in terms of the pairwise kernel dissimilarities; see § 2.1 for details. For characteristic kernels (see Assumption 1), a useful property of the MMD is that it takes value zero if and only if distributions P and Q are the same. Consequently, the test that rejects H_0 for large values of the estimated MMD is universally consistent. The power of the test converges to 1 as the sample size increases for hypothesis (1); see Gretton et al. (2012a) for further details.

Although the kernel two-sample test is widely used and has found numerous applications, it often performs poorly for high-dimensional problems (Ramdas et al., 2015) and its empirical performance depends heavily on the choice of the kernel. Kernels are usually parameterized by their bandwidths, and the most popular strategy for choosing the kernel bandwidth is the median heuristic, where the bandwidth is chosen to be the median of the pairwise distances of the pooled sample (Gretton et al., 2012a). Despite its popularity, there is limited understanding of the median heuristic and empirical results demonstrate that the median heuristic performs poorly when differences between the two distributions occur at a scale that differs significantly from the median of the interpoint distances. Another approach is to split the data and estimate the kernel by maximizing an approximate empirical power on the held-out data (Gretton et al., 2012b; Liu et al., 2020). This, however, can lead to loss in power for smaller sample sizes.

In this paper we propose a strategy for augmenting the power of the classical single-kernel two-sample test by borrowing strengths from multiple kernels. Specifically, we propose a new test statistic that combines MMD estimates from $r \geqslant 1$ kernels using their sample Mahalanobis distance. The advantage of aggregating across a collection of kernels/bandwidths is that the test can simultaneously deal with cases that require both small and large bandwidths, and, hence, detect both global and local differences more effectively. We illustrate the effectiveness of our method through a wide range of results, including a holistic study of its asymptotic properties, finite-sample and real-data performance, computational complexity, and comparison with other aggregation methods.

To begin with, we derive the joint distribution of the vector of MMD estimates under H_0, which can be described using bivariate stochastic integrals, and, as a consequence, derive the asymptotic distribution of the Mahalanobis aggregated MMD (Mahalanobis MMD) statistic under H_0. Moreover, using the kernel Gram matrix representation, we develop a multiplier bootstrap approach that allows us to efficiently compute the rejection threshold for the Mahalanobis MMD statistic and show that the resulting test is universally consistent. Next, we derive the distribution of the proposed test against local alternatives in the well-known contamination model. In the Supplementary Material we derive the joint distribution of MMD estimates and, consequently, that of the Mahalanobis MMD statistic, under the alternative.

To compliment the theoretical results, we perform extensive simulations to compare our Mahalanobis MMD-based test with various single-kernel MMD tests, with bandwidths chosen based on the median heuristic. The experiments show that the Mahalanobis MMD method outperforms the single-kernel tests and also the graph-based Friedman–Rafsky test (Friedman & Rafsky, 1979) across a range of alternatives and dimensions, showcasing the efficacy of our aggregation method. To further reinforce the benefits of our aggregation scheme, we also compare the Mahalanobis MMD test with bandwidth-optimized single-kernel tests, as in Gretton et al. (2012b) and Liu et al. (2020), and with p-value combination methods.

To understand the computational complexity, we analyse the running time of the Mahalanobis MMD tests and also report the trade-off between power and computation time of the Mahalanobis MMD and the single-kernel MMD tests in simulations. We also implement our Mahalanobis aggregation strategy for the linear-time statistic (Gretton et al., 2012a, § 6), derive the corresponding asymptotic theory and report its finite-sample performance. The multiplier bootstrap also emerges as the more computationally efficient option than the permutation test for calibrating the Mahalanobis MMD statistic.

Next, we apply the proposed method to compare images of digits in the noisy MNIST dataset. The Mahalanobis MMD effectively distinguishes different digits for significantly more noisy images compared to its single-kernel counterparts, again illustrating the advantage of using multiple kernels.

We also investigate the behaviour of the Mahalanobis and other aggregation strategies when the number of kernels is allowed to grow with the sample size. Specifically, we derive consistent tests based on the Mahalanobis method, as well as maximum and L_2-type aggregations, in the growing r regime.

Our results on the joint distribution for multiple kernels are also more broadly useful in understanding the asymptotic properties of general aggregation strategies. To demonstrate this, we present two applications. We propose an asymptotic implementation of the adaptive MMD test recently proposed by Schrab et al. (2023), and derive its asymptotic local

power. Numerical results comparing the Mahalanobis MMD method and the aforementioned adaptive test are also reported in the Supplementary Material. We also derive the asymptotic distribution of the Mahalanobis MMD statistic for kernels with bandwidths depending on the sample size. Specifically, we show that, when the scaling parameters are chosen proportional to the optimal bandwidth, as in Li & Yuan (2019) and Schrab et al. (2023), then the vector of MMD estimates has a multivariate normal distribution under the null. Using this, we construct a test that aggregates multiple kernels with a chi-squared distribution under H_0.

The codes for all the experiments are available at https://github.com/anirbanc96/MMMD-boost-kernel-two-sample.

2. Kernel maximum mean discrepancy and Mahalanobis aggregation

2.1. *Kernel maximum mean discrepancy*

Suppose that $\mathcal{X}$ is a separable metric space and that $\mathscr{B}(\mathcal{X})$ is the sigma-algebra generated by the open sets of $\mathcal{X}$. Denote by $\mathcal{P}(\mathcal{X})$ the collection of all probability distributions on $\{\mathcal{X}, \mathscr{B}(\mathcal{X})\}$. Suppose that $P, Q \in \mathcal{P}(X)$ and that $X \sim P$ and $Y \sim Q$ are random variables distributed as P and Q, respectively. Throughout, we assume that P and Q are nonatomic. The maximum mean discrepancy between P and Q is defined as

$$\mathrm{MMD}[\mathcal{F}, P, Q] = \sup_{f \in \mathcal{F}} \{\mathrm{E}_{X\sim P}[f(X)] - \mathrm{E}_{Y\sim Q}[f(Y)]\}, \tag{2}$$

where $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space $\mathcal{H}$ defined on $\mathcal{X}$ (Aronszajn, 1950). Since $\mathcal{H}$ is an RKHS, for every $x \in \mathcal{X}$, the evaluation map operator $\eta_x \colon \mathcal{H} \to \mathbb{R}$ given $\eta_x(f) = f(x)$ is continuous. Thus, by the Riesz representation theorem (Reed & Simon, 1980, Theorem II.4), for each $x \in \mathcal{X}$, there is a feature mapping $\psi_x \in \mathcal{H}$ such that $f(x) = \langle f, \psi_x \rangle_{\mathcal{H}}$ for every $f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. The feature mapping takes the canonical form $\psi_x(\cdot) = \mathsf{K}(x, \cdot)$, where $\mathsf{K} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel. This, in particular, implies that $\mathsf{K}(x, y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$. Extending the notion of a feature map, an element $\mu_P \in \mathcal{H}$ is defined to be the mean embedding of $P \in \mathcal{P}(\mathcal{X})$ if

$$\langle f, \mu_P \rangle_{\mathcal{H}} = \mathrm{E}_{X\sim P}[f(X)] \tag{3}$$

for all $f \in \mathcal{H}$. By the canonical form of the feature map, it follows that

$$\mu_P(t) := \int_{\mathcal{X}} \psi_t(x)\, \mathrm{d}P(x) = \mathrm{E}_{X\sim P}[\psi_t(X)] = \mathrm{E}_{X\sim P}[\mathsf{K}(t, X)]. \tag{4}$$

Throughout, we make the following assumption.

Assumption 1. The kernel $\mathsf{K} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies the following conditions:

(i) $\mathrm{E}_{X\sim P}[\mathsf{K}(X, X)^{1/2}] < \infty$ and $\mathrm{E}_{Y\sim Q}[\mathsf{K}(Y, Y)^{1/2}] < \infty$,
(ii) K is characteristic, that is, the mean embedding $\mu \colon \mathcal{P}(\mathcal{X}) \to \mathcal{H}$ is a one-to-one (injective) function.

Assumption 1 ensures that $\mu_P, \mu_Q \in \mathcal{H}$ (see Lemma 3 of Gretton et al., 2012a and Lemma 2.1 of Park & Muandet, 2020) and that the MMD defines a metric on $\mathcal{P}(\mathcal{X})$. Then

the MMD can be expressed as the distance between mean embeddings in $\mathcal{H}$ (see Lemma 4 of Gretton et al., 2012a):

$$\mathrm{MMD}^2[\mathcal{F}, P, Q] = \|\mu_P - \mu_Q\|^2_{\mathcal{H}} \tag{5}$$

with $\|\cdot\|_{\mathcal{H}}$ the norm corresponding to the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. This implies that $\mathrm{MMD}^2[\mathcal{F}, P, Q] = 0$ if and only if $P = Q$. Expanding the square in (5) and using representation (4), it follows that

$$\mathrm{MMD}^2[\mathcal{F}, P, Q] = \mathrm{E}_{X,X'\sim P}[\mathsf{K}(X, X')] + \mathrm{E}_{Y,Y'\sim Q}[\mathsf{K}(Y, Y')] - 2\mathrm{E}_{X\sim P,\, Y\sim Q}[\mathsf{K}(X, Y)];$$

see Gretton et al. (2012a, Lemma 6) for details. Therefore, based on independent and identically distributed observations $\mathscr{X}_m := \{X_1, X_2, \ldots, X_m\}$ and $\mathscr{Y}_n := \{Y_1, Y_2, \ldots, Y_n\}$, a natural unbiased estimate of $\mathrm{MMD}^2[\mathcal{F}, P, Q]$ is given by

$$\mathrm{MMD}^2[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n] = \mathcal{W}_{\mathscr{X}_m} + \mathcal{W}_{\mathscr{Y}_n} - 2\mathcal{B}_{\mathscr{X}_m,\mathscr{Y}_n}, \tag{6}$$

where

$$\mathcal{W}_{\mathscr{X}_m} := \frac{1}{m(m-1)} \sum_{1\leqslant i\neq j\leqslant m} \mathsf{K}(X_i, X_j) \quad \text{and} \quad \mathcal{W}_{\mathscr{Y}_n} := \frac{1}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} \mathsf{K}(Y_i, Y_j) \tag{7}$$

are the averages of the kernel dissimilarities within the samples in $\mathscr{X}_m$ and $\mathscr{Y}_n$, respectively, and

$$\mathcal{B}_{\mathscr{X}_m,\mathscr{Y}_n} := \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathsf{K}(X_i, Y_j) \tag{8}$$

is the average of the kernel dissimilarities between the samples in $\mathscr{X}_m$ and $\mathscr{Y}_n$. Throughout, we assume that $N := m + n \to \infty$ such that

$$\frac{m}{m+n} \to \rho \in (0, 1). \tag{9}$$

Then $\mathrm{MMD}^2[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n]$ is a consistent estimate of $\mathrm{MMD}^2[\mathcal{F}, P, Q]$ (see Theorem 7 of Gretton et al., 2012a), that is,

$$\mathrm{MMD}^2[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{\mathrm{P}} \mathrm{MMD}^2[\mathcal{F}, P, Q]. \tag{10}$$

Hence, the test that rejects H_0 in (1) for large values of $\mathrm{MMD}^2[\mathsf{K}, \mathscr{X}_m, \mathscr{Y}_n]$ is universally consistent. In fact, for the consistency result, it suffices to assume that $\min\{m, n\} \to \infty$. The existence of the limit in (9) will be required for deriving the asymptotic distribution of the test statistic.

2.2. *Aggregating multiple kernels*

Fix $r \geqslant 1$, and suppose that $\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r$ are r distinct kernels each of which satisfy Assumption 1. Denote the vector of MMD estimates as

$$\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] = (\mathrm{MMD}^2[\mathsf{K}_1, \mathscr{X}_m, \mathscr{Y}_n], \ldots, \mathrm{MMD}^2[\mathsf{K}_r, \mathscr{X}_m, \mathscr{Y}_n])^{\mathrm{T}}, \tag{11}$$

where $\mathcal{K} := \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$. In this paper we propose a new test statistic that combines the contributions of the different kernels using the Mahalanobis distance of the vector $\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$ as

$$(\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n])^{\mathrm{T}} \underline{S}^{-1} (\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]), \tag{12}$$

where $\underline{S}$ is a consistent estimate of the limiting covariance matrix of $\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$ under H_0, which we denote by $\underline{\Sigma}_{H_0} = \{(\sigma_{ab})\}_{1 \leqslant a,b \leqslant r}$. Adjusting by the covariance matrix $\underline{S}$ places the contributions of the individual MMD estimates on the same scale and by selecting a range of kernels/bandwidths in $\mathcal{K}$ one can detect more fine-grained deviations from H_0, leading to significant power improvements, as will be seen in § 6 below. In the Supplementary Material we present general conditions under which $\underline{\Sigma}_{H_0}$ is invertible, which, in particular, hold for any collection of Gaussian or Laplace kernels.

In Corollary 1 below we compute

$$\begin{aligned} \sigma_{ab} &:= \lim_{N \to \infty} (m+n)^2 (\mathrm{cov}_{H_0}\{\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]\})_{ab} \\ &= \frac{2}{\rho^2(1-\rho)^2} \mathrm{E}_{X,X' \sim P}[\mathsf{K}_a^\circ(X, X') \mathsf{K}_b^\circ(X, X')], \end{aligned} \tag{13}$$

where

$$\mathsf{K}_a^\circ(x, y) = \mathsf{K}_a(x, y) - \mathrm{E}_{X \sim P} \mathsf{K}_a(X, y) - \mathrm{E}_{X' \sim P} \mathsf{K}_a(x, X') + \mathrm{E}_{X,X' \sim P} \mathsf{K}_a(X, X') \tag{14}$$

is the centred version of kernel K_a for $1 \leqslant a \leqslant r$. Therefore, a natural empirical estimate of $\underline{\Sigma}_{H_0}$ is given by $\underline{\hat{\Sigma}} = \{(\hat{\sigma}_{ab})\}_{1 \leqslant a,b \leqslant r}$, where

$$\hat{\sigma}_{ab} = \frac{2}{\hat{\rho}^2(1-\hat{\rho})^2} \frac{1}{m^2} \sum_{1 \leqslant i,j \leqslant m} \hat{\mathsf{K}}_a^\circ(X_i, X_j) \hat{\mathsf{K}}_b^\circ(X_i, X_j) \tag{15}$$

with

$$\hat{\mathsf{K}}_a^\circ(x, y) = \mathsf{K}_a(x, y) - \frac{1}{m} \sum_{u=1}^{m} \mathsf{K}_a(X_u, y) - \frac{1}{m} \sum_{v=1}^{m} \mathsf{K}_a(x, X_v) + \frac{1}{m^2} \sum_{u,v=1}^{m} \mathsf{K}_a(X_u, X_v) \tag{16}$$

the empirical analogue of K_a° and $\hat{\rho} = m/(m+n)$. Therefore, choosing $\underline{S} = \underline{\hat{\Sigma}}$ in (12), we define the Mahalanobis MMD statistic as

$$T_{m,n} := (\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n])^{\mathrm{T}} \underline{\hat{\Sigma}}^{-1} (\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]). \tag{17}$$

In Corollary 1 below we show that $\underline{\hat{\Sigma}} \xrightarrow{\mathrm{P}} \underline{\Sigma}_{H_0}$; hence, (10) implies that

$$T_{m,n} \xrightarrow{\mathrm{P}} (\mathrm{MMD}^2[\underline{\mathcal{F}}, P, Q])^{\mathrm{T}} \underline{\Sigma}_{H_0}^{-1} (\mathrm{MMD}^2[\underline{\mathcal{F}}, P, Q]) := T_{\mathcal{K}}, \tag{18}$$

where $\underline{\mathcal{F}} = \{\mathcal{F}_1, \ldots, \mathcal{F}_r\}$, with $\mathcal{F}_a$ the unit ball in the RKHS of K_a for all $1 \leqslant a \leqslant r$, and

$$\mathrm{MMD}^2[\underline{\mathcal{F}}, P, Q] = (\mathrm{MMD}^2[\mathcal{F}_1, P, Q], \ldots, \mathrm{MMD}^2[\mathcal{F}_r, P, Q])^{\mathrm{T}}. \tag{19}$$

Note that $T_{\mathcal{K}} = 0$ under H_0 and $T_{\mathcal{K}} > 0$ whenever $P \neq Q$. Hence, a test rejecting H_0 for large values of $T_{m,n}$ will be universally consistent. However, to construct a test based on $T_{m,n}$, we need to chose a cut-off (rejection region) based on the data. The first step towards this is to derive the limiting null distribution of $\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$.

3. Asymptotic null distribution

In this section we derive the limiting distribution of the vector of MMD estimates (11) under H_0 and, consequently, that of the proposed statistic $T_{m,n}$, using the framework of multiple Weiner–Itô stochastic integrals. We recall the definition and basic properties of multiple Weiner–Itô stochastic integrals in the Supplementary Material.

Theorem 1. *Suppose that $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ is a collection of r distinct kernels such that K_a satisfies Assumption 1 and $\mathsf{K}_a \in L^2(\mathcal{X}^2, P^2)$ for $1 \leqslant a \leqslant r$. Then, under H_0, in the asymptotic regime (9),*

$$(m+n)\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{\mathrm{D}} G_{\mathcal{K}} := \frac{1}{\rho(1-\rho)}\{I_2(\mathsf{K}_1^\circ), I_2(\mathsf{K}_2^\circ), \ldots, I_2(\mathsf{K}_r^\circ)\}^{\mathrm{T}}, \tag{20}$$

where $I_2(\cdot)$ is the bivariate multiple Weiner–Itô stochastic integral, as defined in the Supplementary Material, and K_a° is defined in (14) for $1 \leqslant a \leqslant r$. Moreover, the characteristic function of $G_{\mathcal{K}}$ at $\underline{\eta} = (\eta_1, \eta_2, \ldots, \eta_r)^T \in \mathbb{R}^r$ is given by

$$\Phi(\underline{\eta}) := \mathrm{E}[e^{\iota \underline{\eta}^T G_{\mathcal{K}}}] = \prod_{\lambda \in \Lambda(\underline{\eta})} \frac{\exp\{-\iota\lambda/\rho(1-\rho)\}}{\{1 - 2\iota\lambda/\rho(1-\rho)\}^{1/2}}, \tag{21}$$

where $\Lambda(\underline{\eta})$ is the set of eigenvalues with repetitions of the Hilbert–Schmidt operator $\mathcal{H}_{\mathcal{K},\underline{\eta}}\colon L^2(\mathcal{X}, P) \to L^2(\mathcal{X}, P)$ defined as

$$\mathcal{H}_{\mathcal{K},\underline{\eta}}[f(x)] = \int_{\mathcal{X}} \left(\sum_{a=1}^{r} \eta_a \mathsf{K}_a^\circ(x, y) \right) f(y)\, \mathrm{d}P(y). \tag{22}$$

The proof of Theorem 1 is given in the Supplementary Material. For an alternate representation of the limiting distribution in (20), see Remark 2 in the Supplementary Material.

Theorem 1 allows us to obtain the limiting distribution of any smooth function of finitely many MMD estimates under H_0. In particular, for the Mahalanobis MMD statistic $T_{m,n}$ in (17), we have the following result. The proof is given in the Supplementary Material.

Corollary 1. *Suppose that $\underline{\Sigma}_{H_0} := \{(\sigma_{ab})\}_{1 \leqslant a,b \leqslant r}$ and $\hat{\underline{\Sigma}} := \{(\hat{\sigma}_{ab})\}_{1 \leqslant a,b \leqslant r}$ are as in (13) and (15), respectively. Then*

$$\sigma_{ab} = \frac{2}{\rho^2(1-\rho)^2} \mathrm{E}_{X,X' \sim P}[\mathsf{K}_a^\circ(X, X')\mathsf{K}_b^\circ(X, X')], \tag{23}$$

where K_a° *for* $1 \leqslant a \leqslant r$ *is as defined in* (14)*. Moreover, in the asymptotic regime* (9)*,*

$$\hat{\sigma}_{ab} \to \sigma_{ab} \tag{24}$$

almost surely for $1 \leqslant a, b \leqslant r$*. Furthermore, under* H_0*, for* $G_{\mathcal{K}}$ *as in* (20)*,*

$$(m+n)^2 T_{m,n} \xrightarrow{\mathrm{D}} G_{\mathcal{K}}^{\mathrm{T}} \underline{\Sigma}_{H_0}^{-1} G_{\mathcal{K}}. \tag{25}$$

4. Calibration using the Gaussian multiplier bootstrap

In order to apply Corollary 1 to obtain a valid level α test based on $T_{m,n}$, we need to estimate the quantiles of the limiting distribution in (25), which depends on the unknown distribution P. Although the distribution in (25) does not have a tractable closed form, we can efficiently estimate its quantiles based on the samples $\mathscr{X}_m = \{X_1, X_2, \ldots, X_m\}$, using the kernel Gram matrix representation of the MMD estimate and the Gaussian multiplier bootstrap. To this end, for each kernel K_a, define its Gram matrix based on $\mathscr{X}_m$ as

$$\hat{\underline{\mathsf{K}}}_a = \{\mathsf{K}_a(X_i, X_j)\}_{1 \leqslant i,j \leqslant m},$$

and their centred versions as

$$\hat{\underline{\mathsf{K}}}_a^\circ = \underline{C}_m \hat{\underline{\mathsf{K}}}_a \underline{C}_m / m = \left(\frac{\hat{\mathsf{K}}_a^\circ(X_i, X_j)}{m} \right)_{1 \leqslant i,j \leqslant m}, \quad \text{where} \quad \underline{C}_m = I - \frac{1}{m} \underline{1} \cdot \underline{1}^{\mathrm{T}}, \tag{26}$$

with $\hat{\mathsf{K}}_a^\circ$ defined as in (16) for $1 \leqslant a \leqslant r$. For $\hat{\rho} := m/(m+n)$, define

$$\mathcal{E}(\mathcal{K}, \mathscr{X}_m) := \begin{pmatrix} \underline{Z}_m^{\mathrm{T}} \hat{\underline{\mathsf{K}}}_1^\circ \underline{Z}_m - \mathrm{tr}[\hat{\underline{\mathsf{K}}}_1^\circ]/\hat{\rho}(1-\hat{\rho}) \\ \underline{Z}_m^{\mathrm{T}} \hat{\underline{\mathsf{K}}}_2^\circ \underline{Z}_m - \mathrm{tr}[\hat{\underline{\mathsf{K}}}_2^\circ]/\hat{\rho}(1-\hat{\rho}) \\ \vdots \\ \underline{Z}_m^{\mathrm{T}} \hat{\underline{\mathsf{K}}}_m^\circ \underline{Z}_m - \mathrm{tr}[\hat{\underline{\mathsf{K}}}_m^\circ]/\hat{\rho}(1-\hat{\rho}) \end{pmatrix}, \tag{27}$$

where $\underline{Z}_m \sim \mathcal{N}_m\{0, I/\hat{\rho}(1-\hat{\rho})\}$ independent of $\mathscr{X}_m$. In the following theorem we show that the distribution of $\mathcal{E}(\mathcal{K}, \mathscr{X}_m)$ conditional on $\mathscr{X}_m$ converges to $G_{\mathcal{K}}$, as in (20).

THEOREM 2. *Suppose that* $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ *is a collection of* $r \geqslant 1$ *distinct kernels such that* K_a *satisfies Assumption* 1 *and* K_a *is bounded for all* $1 \leqslant a \leqslant r$*. Then, under* H_0*, in the asymptotic regime* (9)*,* $\mathcal{E}(\mathcal{K}, \mathscr{X}_m) \mid \mathscr{X}_m \xrightarrow{D} G_{\mathcal{K}}$ *almost surely, where* $G_{\mathcal{K}}$ *is as defined in* (20)*.*

The proof of Theorem 2 is given in the Supplementary Material. It shows that the asymptotic distribution of $\mathcal{E}(\mathcal{K}, \mathscr{X}_m) \mid \mathscr{X}_m$ is the same as that of $(m+n)\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n]$. Since $\mathcal{E}(\mathcal{K}, \mathscr{X}_m) \mid \mathscr{X}_m$ is completely determined by the data $\mathscr{X}_m$, we can use it to approximate the quantiles of any nice functions $G_{\mathcal{K}}$. To this end, define

$$\hat{T}_m := \mathcal{E}(\mathcal{K}, \mathscr{X}_m)^{\mathrm{T}} \hat{\underline{\Sigma}}^{-1} \mathcal{E}(\mathcal{K}, \mathscr{X}_m). \tag{28}$$

Now, by a direct computation,

$$\operatorname{var}[\mathcal{E}(\mathcal{K}, \mathscr{X}_m) \mid \mathscr{X}_m] = \frac{2}{\hat{\rho}^2(1-\hat{\rho})^2}\{(\operatorname{tr}[\underline{\hat{\mathbf{K}}}_a^{\circ}\underline{\hat{\mathbf{K}}}_b^{\circ}])\}_{1\leqslant a,b\leqslant m}.$$

Hence, from the proof of Corollary 1, specifically (24), it follows that $\operatorname{var}[\mathcal{E}(\mathcal{K}, \mathscr{X}_m) \mid \mathscr{X}_m] = \underline{\hat{\Sigma}} \xrightarrow{\text{a.s.}} \underline{\Sigma}_{H_0}$. This combined with Theorem 2 implies that, under H_0,

$$\hat{T}_m \mid \mathscr{X}_m \xrightarrow{\text{D}} G_{\mathcal{K}}^{\mathrm{T}}\underline{\Sigma}_{H_0}^{-1}G_{\mathcal{K}}$$

almost surely. This shows that $\hat{T}_m$ has the same limiting distribution as $(m+n)^2T_{m,n}$ under H_0 (recall (25)); hence, we can use the quantiles of $\hat{T}_m$ to calibrate the statistic $T_{m,n}$. Specifically, for $\alpha \in (0,1)$, denote by $\hat{q}_{\alpha,m}$ the αth quantile of distribution $\hat{T}_m \mid \mathscr{X}_m$ and consider the test function

$$\phi_{m,n} = \mathbb{1}\{(m+n)^2T_{m,n} > \hat{q}_{1-\alpha,m}\}. \tag{29}$$

Corollary 1, (18) and (28) now imply the following result.

COROLLARY 2 (CONSISTENCY). *Suppose that the assumptions of Theorem 2 hold and that $\phi_{m,n}$ is defined as above. Then $\lim_{m,n\to\infty} \mathrm{E}_{H_0}[\phi_{m,n}] = \alpha$. Moreover, for any $P \neq Q$, $\lim_{m,n\to\infty} \mathrm{E}_{H_1}[\phi_{m,n}] = 1$, that is, $\phi_{m,n}$ is universally consistent.*

The result above shows that the Mahalanobis MMD statistic with cut-off chosen using the multiplier bootstrap method attains the exact asymptotic level and is universally consistent. In practice, to compute $\hat{q}_{1-\alpha,m}$, we generate B replicates $\{\hat{T}_m^{(1)}, \hat{T}_m^{(2)}, \ldots, \hat{T}_m^{(B)}\}$ of $\hat{T}_m$, based on B independent copies of $\underline{Z}_m$, and choose $\hat{q}_{1-\alpha,m}$ to be the sample αth quantile of $\{\hat{T}_m^{(1)}, \hat{T}_m^{(2)}, \ldots, \hat{T}_m^{(B)}\}$.

Remark 1. While implementing the test we often replace $\underline{\hat{\Sigma}}^{-1}$ in (17) and (28) by $(\underline{\hat{\Sigma}} + \lambda\underline{I}_m)^{-1}$ for some suitably chosen regularization parameter $\lambda > 0$. Although the limiting covariance matrix $\underline{\Sigma}_{H_0}$ is invertible (see Corollary L.1 in the Supplementary Material), and hence $\underline{\hat{\Sigma}}$ is also invertible for large sample sizes with probability 1, adding a small regularization provides numerical stability in finite samples. In fact, the conclusions in Corollary 2 remain valid for any choice of $\lambda = \lambda(\mathscr{X}_m)$ converging almost surely to a deterministic constant $\lambda_0 > 0$; see § 6 below for more details on the choice of λ in our experiments.

5. LOCAL ASYMPTOTIC POWER

Throughout this section, we assume that $\mathcal{X} = \mathbb{R}^d$ and that distributions P and Q have densities f_P and f_Q with respect to the Lebesgue measure in $\mathbb{R}^d$. To quantify the notion of local alternatives, we adopt the commonly used contamination model:

$$f_Q(\cdot) = (1-\delta)f_P(\cdot) + \delta g(\cdot). \tag{30}$$

Here $\delta \in [0,1)$ and $g \neq f_P$ is a probability density function with respect to the Lebesgue measure in $\mathbb{R}^d$ such that the following assumption holds.

Assumption 2. The support of g is contained in that of $f_P(\cdot)$ and $0 < \mathrm{var}_{X\sim P}[g(X)/f_P(X)] < \infty$.

Under this assumption, contiguous local alternatives are obtained by considering local perturbations of the mixing proportion δ as (see Ch. 12 of Lehmann & Romano, 2005)

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta = h/\sqrt{N} \tag{31}$$

for some $h \neq 0$ and $N = m + n$. The following theorem derives the distribution of the Mahalanobis MMD statistic $T_{m,n}$ under H_1 as above.

THEOREM 3. *Suppose that $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ is a collection of r distinct kernels such that K_a satisfies Assumption 1 and $\mathsf{K}_a \in L^2(\mathcal{X}^2, P^2)$ for $1 \leqslant a \leqslant r$. Then, under H_1 as in (31), in the asymptotic regime (9),*

$$(m+n)\mathrm{MMD}^2[\mathcal{K}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{D} G_{\mathcal{K},h} := \begin{pmatrix} \gamma I_2(\mathsf{K}_1^\circ) + 2h\gamma^{1/2} I_1(\mathsf{K}_1^\circ[g/f_P]) + h^2\mu_1 \\ \gamma I_2(\mathsf{K}_2^\circ) + 2h\gamma^{1/2} I_1(\mathsf{K}_2^\circ[g/f_P]) + h^2\mu_2 \\ \vdots \\ \gamma I_2(\mathsf{K}_r^\circ) + 2h\gamma^{1/2} I_1(\mathsf{K}_r^\circ[g/f_P]) + h^2\mu_r \end{pmatrix}, \tag{32}$$

where $\gamma = 1/\rho(1-\rho)$, $\mathsf{K}_a^\circ[g/f_P](x) := \int_{\mathcal{X}} \mathsf{K}_a^\circ(x,y) g(y)\,\mathrm{d}y$,

$$\mu_a := \mathrm{E}\left[\mathsf{K}_a^\circ(X, X') \frac{g(X)g(X')}{f_P(X)f_P(X')}\right] \tag{33}$$

and K_a° is defined in (14) for $1 \leqslant a \leqslant r$.

The proof of the theorem is given in the Supplementary Material. The following result is an immediate consequence of the above result together with the continuous mapping theorem and Corollary 1.

COROLLARY 3. *Under H_1 as in (31), $(m+n)^2 T_{m,n} \xrightarrow{\mathrm{D}} G_{\mathcal{K},h}^{\mathrm{T}} \underline{\Sigma}_{H_0}^{-1} G_{\mathcal{K},h}$.*

Using Corollary 3, we can derive the limiting local power of test $\phi_{m,n}$ in (29). Specifically, suppose that $F_{\mathcal{K},h}$ denotes the cumulative distribution function of $G_{\mathcal{K},h}^{\mathrm{T}} \underline{\Sigma}_{H_0}^{-1} G_{\mathcal{K},h}$ and that $q_{1-\alpha}$ is the $(1-\alpha)$th quantile of distribution $G_{\mathcal{K}}^{\mathrm{T}} \underline{\Sigma}_{H_0}^{-1} G_{\mathcal{K}}$. Note that $G_{\mathcal{K},0} = G_{\mathcal{K}}$. Since $\hat{q}_{1-\alpha,m} \mid \mathscr{X}_m \to q_{1-\alpha}$ almost surely, Corollary 3 implies that the asymptotic power of $\phi_{m,n}$ under H_1 as in (31) is given by $\lim_{m,n\to\infty} \mathrm{E}_{H_1}[\phi_{m,n}] = 1 - F_{\mathcal{K},h}(q_{1-\alpha})$. This implies that $\phi_{m,n}$ has nontrivial asymptotic (Pitman) efficiency and is rate optimal, in the sense that $\lim_{|h|\to\infty} \lim_{m,n\to\infty} \mathrm{E}_{H_1}[\phi_{m,n}] = 1$.

6. NUMERICAL EXPERIMENTS

6.1. *Choice of kernels and experimental parameters*

In this section, we study the finite-sample performance of the proposed Mahalanobis MMD (abbreviated as MMMD in the figures) test across a range of simulation settings. Specifically, we compare the Mahalanobis MMD test to the single-kernel MMD

test (Gretton et al., 2009) and the graph-based Friedman–Rafsky (FR) test (Friedman & Rafsky, 1979). Additional simulations are given in the Supplementary Material. Throughout, we set the significance level $\alpha = 0.05$.

For single-kernel tests, we use the Gaussian and Laplace kernels

$$\mathsf{K}_{\mathrm{GAUSS}}(x,y) = e^{-\|x-y\|^2/\sigma^2} \quad \text{and} \quad \mathsf{K}_{\mathrm{LAP}}(x,y) = e^{-\|x-y\|/\sigma} \tag{34}$$

with the bandwidth σ chosen using the median heuristic $\sigma^2 := \lambda^2_{\mathrm{med}} = \mathrm{median}\{\|Z_i - Z_j\|^2 : 1 \leqslant i < j \leqslant n\}$, where $\mathscr{X}_m \cup \mathscr{Y}_n = \{Z_1, Z_2, \ldots, Z_N\}$ is the pooled sample and $\|\cdot\|$ denotes the Euclidean norm. We refer to these tests as `Gauss MMD` and `LAP MMD`, respectively.

For the Mahalanobis MMD statistic, we use multiple Gaussian kernels, multiple Laplace kernels or a combination of Gaussian and Laplace kernels, with different bandwidths chosen as follows.

(i) `Gauss Mahalanobis MMD`: this is the Mahalanobis MMD statistic with five Gaussian kernels with bandwidths

$$\underline{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = \left(\frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}, 2\right)\lambda_{\mathrm{med}}. \tag{35}$$

(ii) `LAP Mahalanobis MMD`: this is the Mahalanobis MMD statistic with five Laplace kernels with bandwidths

$$\underline{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = \left(\frac{1}{2}, \frac{1}{\sqrt{2}}, 1, \sqrt{2}, 2\right)\lambda_{\mathrm{med}}. \tag{36}$$

(iii) `Mixed Mahalanobis MMD`: this is the Mahalanobis MMD statistic with three Gaussian kernels and three Laplace kernels with the same set of bandwidths

$$\underline{\sigma} = (\sigma_1, \sigma_2, \sigma_3) = \left(\frac{1}{\sqrt{2}}, 1, \sqrt{2}\right)\lambda_{\mathrm{med}}. \tag{37}$$

In our implementation we choose the regularity parameter λ (recall Remark 1) as $\lambda = 10^{-5} \times \min_{1 \leqslant a \leqslant r} \hat{\sigma}_{aa}$ for $\hat{\sigma}_{aa} > 0$, as in (15). Since λ converges to $10^{-5} \times \min_{1 \leqslant a \leqslant r} \sigma_{aa}$ almost surely (recall Corollary 1), the results in Corollary 2 remain valid. The cut-offs of the tests are chosen based on the multiplier bootstrap as in (29) using $B = 500$ resamples.

Finally, for the Friedman–Rafsky test, we use the implementation in the `R` package `gTests` (R Development Core Team, 2024), with the 5-MST (minimum spanning tree), which is the recommended practical choice in Chen & Friedman (2017).

6.2. *Dependence on dimension*

In this section we study the performance of the different tests as the dimension varies in the following settings. We fix sample sizes $m = n = 100$, vary the dimension over $d \in \{5, 10, 25, 50, 75, 100, 150\}$ and compute the empirical power by averaging over 500 iterations.

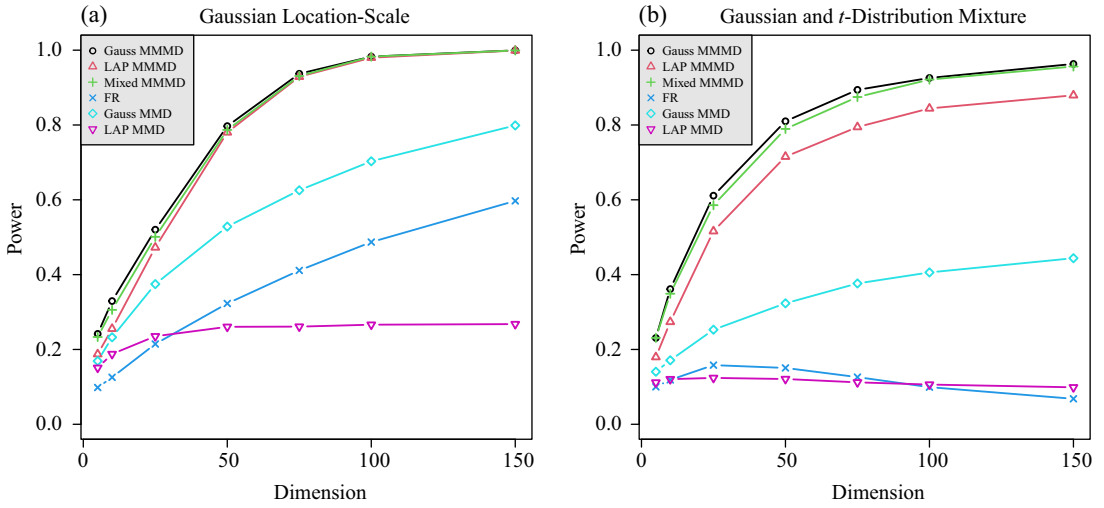

Fig. 1. Empirical powers in (a) Setting 1 and (b) Setting 2.

Setting 1 (Gaussian location-scale). Here, we consider $P = \mathcal{N}_d(0, \underline{\Sigma}_0)$ and $Q = \mathcal{N}_d(0.1\underline{1}, 1.15\underline{\Sigma}_0)$, where $\underline{\Sigma}_0 = \{(0.5^{|i-j|})\}_{1\leqslant i,j\leqslant d}$; see Fig. 1(a).

Setting 2 (Gaussian and t-distribution mixture). Here, we consider $P = \frac{1}{2}\mathcal{N}_d(0, \underline{\Sigma}_0) + \frac{1}{2}t_{10}(0, \underline{\Sigma}_0)$ and $Q = \frac{1}{2}\mathcal{N}_d(0, 1.22\underline{\Sigma}_0) + \frac{1}{2}t_{10}(0, 1.22\underline{\Sigma}_0)$, where $\underline{\Sigma}_0$ is as above; see Fig. 1(b).

The plots show that the multiple kernel Mahalanobis MMD tests have significantly more power than the single-kernel MMD tests and the FR test in both settings. Overall, the `Gauss Mahalanobis MMD` and the `Mixed Mahalanobis MMD` tests perform the best, closely followed by the `Lap Mahalanobis MMD`. This also shows the advantage of aggregating kernels across a range of dimensions, from low dimensions to dimensions that are comparable and even larger than the sample size. Additional simulations are provided in the Supplementary Material.

6.3. *Mixture alternatives*

In this section we evaluate the performance of the tests for mixture alternatives by varying the mixing proportion. To this end, suppose that $\underline{\Sigma}_0 = \{(0.5^{|i-j|})\}_{1\leqslant i,j\leqslant d}$ and consider

$$P = \varepsilon\mathcal{N}_d(0, \underline{\Sigma}_0) + (1-\varepsilon)t_{10}(0, \underline{\Sigma}_0) \quad \text{and} \quad Q = \varepsilon\mathcal{N}_d(0, 1.25\underline{\Sigma}_0) + (1-\varepsilon)t_{10}(0, 1.25\underline{\Sigma}_0).$$

Figure 2 shows the empirical power, averaged over 500 iterations, of the different tests as ε varies over $[0, 1]$, with sample sizes $m = n = 100$ and dimensions $d = 30$ (see Fig. 2(a)) and $d = 150$ (see Fig. 2 (b)). In both cases, the Mahalanobis MMD tests outperform the single-kernel tests and the Friedman–Rafsky test, again illustrating the versatility of the aggregated tests.

6.4. *Computational complexity of the Mahalanobis MMD test*

In the Supplementary Material we analyse the computational complexity of the Mahalanobis MMD test, when the rejection region is chosen based on B replications of statistic $\hat{T}_m$ from (28). In particular, we show that the computational cost of the Mahalanobis MMD

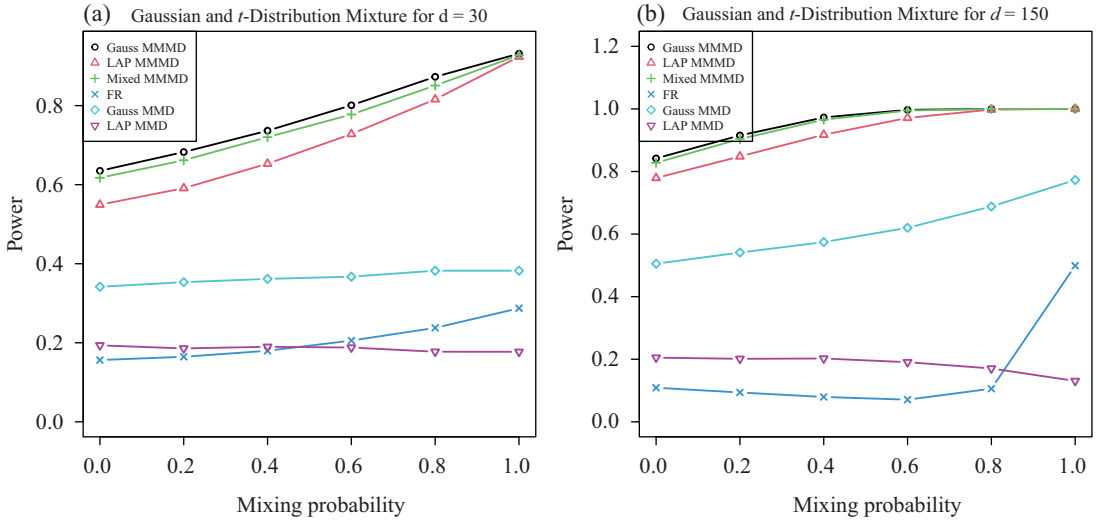

Fig. 2. Empirical powers as a function of the mixing proportion for (a) $d = 30$ and (b) $d = 150$.

test is $O(r^2N^2 + BrN^2 + B\log B)$, assuming that $r < N$. In practice, the number of resamples B is usually chosen to be much larger than the number of kernels r, in which case the time complexity simplifies to $O(BrN^2 + B\log B)$. In fact, realistically, one only aggregates over a bounded number of kernels, that is, $r = O(1)$, in which case the computational costs of the Mahalanobis MMD test and the MMD test differ only by a constant factor. In the Supplementary Material we also compare the running times of the MMD and the Mahalanobis MMD tests in simulations (see Table 1). Our experiments show that the Mahalanobis MMD tests provide significant power enhancement over the MMD test, with only a small increase in the computation time.

Remark 2. One way to reduce the computation cost of the MMD test from $O(N^2)$ to $O(N)$ is the linear-time MMD statistic (Gretton et al., 2012a, § 6). In the Supplementary Material we apply the Mahalanobis aggregation strategy to combine linear-time statistics over multiple kernels and develop the associated theory. We also compare the power of the aggregated linear-time MMD tests with their single-kernel counterparts and also with the quadratic time Mahalanobis MMD tests in simulations.

6.5. *Comparison with the permutation test*

Another alternative to choosing the rejection threshold for $T_{m,n}$ is the permutation method. In fact, the permutation principle can be applied to calibrate any two-sample test statistic based on the sample quantiles of the test statistic computed on B permuted versions of the pooled data $\mathscr{X}_m \cup \mathscr{Y}_n$. The resulting test is guaranteed to control the Type-I error in finite samples. In this paper we adopt the multiplier bootstrap over the permutation method for the following two reasons.

Firstly, the independence of the Gaussian multipliers makes the asymptotic theory of the multiplier bootstrap method more tractable. Consequently, we are able to provide a holistic asymptotic theory for the multiplier bootstrap–based Mahalanobis MMD test, including limiting distributions, under both the null and the alternative, consistency and local power analysis.

Secondly, the multiplier bootstrap is computationally more efficient than the permutation method, both in terms of their asymptotic running times as well as power versus computation time trade-off in finite samples. To obtain the permutation p-value, we have to compute the Mahalanobis MMD statistic $T_{m,n}$ (recall (17)) for each of the B random permutations of N samples, where $N = m + n$ is the total number of samples. Since with r kernels it takes $O(r^2N^2)$ time to compute $T_{m,n}$, the time complexity for the permutation test is $O(Br^2N^2)$, where B is the number of permutations. On the other hand, we know from § 6.4 that the time complexity of the multiplier bootstrap-based Mahalanobis MMD test (29) is $O(r^2N^2 + BrN^2 + B\log B)$, which has a better dependence on r than the permutation test. Even for fixed r one can see significant gains in computation time in finite samples. In particular, our simulations show that the Type-I error and power of the multiplier bootstrap and the permutation methods are comparable, but the computation time of the multiplier bootstrap method is much faster.

6.6. *Comparisons with bandwidth-optimized MMD tests and p-value combination methods*

Recall that in the previous sections while implementing the MMD test we chose the bandwidths for the Gaussian and Laplace kernels based on the median heuristic. Although this is the common choice in practice (Gretton et al., 2012a; Ramdas et al., 2015), it remains a heuristic because there is no theoretical understanding of its validity. To address this issue, there have been studies that aim to find the best single-kernel test by optimizing the bandwidth in such a way that the asymptotic power is maximized. This approach was first proposed by Gretton et al. (2012b) for the linear-time MMD test, which was subsequently extended to the quadratic time MMD test by Sutherland et al. (2021). The method involves splitting the data into two parts and using the first part to select the bandwidth by maximizing asymptotic power, or, equivalently, by maximizing the ratio (see Liu et al., 2020) $\mathrm{MMD}^2[\mathsf{K}_\lambda, \mathscr{X}_m, \mathscr{Y}_n]/\hat{\sigma}^2_\lambda$, where $\hat{\sigma}^2_\lambda$ is a regularized estimator of the asymptotic variance of $\mathrm{MMD}^2[\mathsf{K}_\lambda, \mathscr{X}_m, \mathscr{Y}_n]$ under H_1. In the Supplementary Material we provide empirical comparisons of our test based on multiple kernels with the bandwidth-optimized single-kernel test in different simulations. To mitigate the effect of data splitting, we also implement the single-kernel tests with twice the amount of data as in Schrab et al. (2023, § 5.3). This emulates an oracle choice of bandwidth and represents the best single-kernel MMD test for the given data. In all the settings considered, the Mahalanobis MMD tests have improved power more than the bandwidth-optimized single-kernel tests. Also, Mahalanobis MMD tests with Gaussian/Laplace kernels have better power than the Gaussian/Laplace oracle MMD test (where the bandwidth is optimized with double the sample size), respectively. The bandwidths for the kernels in the Mahalanobis MMD tests are chosen as in (35), (36) and (37), respectively, which requires no optimization or data splitting. Even so, the multiple kernel Mahalanobis MMD test is able to outperform the 'best' single kernel, demonstrating the effectiveness of our aggregation scheme.

Another possible aggregation strategy is to consider tests that combine p-values for multiple single-kernel MMD tests. To illustrate how our aggregation strategy compares with p-value combination methods, we consider the following experimental set-up. We implement the `Gauss Mahalanobis MMD` test based on five different Gaussian kernels with respective bandwidths $\underline{\sigma} = (\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5) = (1/2, 1/\sqrt{2}, 1, \sqrt{2}, 2)\lambda_{\mathrm{med}}$, where λ_{med} is defined after (34). The `Gauss Mahalanobis MMD` test is calibrated using the multiplier bootstrap with $B = 500$ resamples. For comparison, we consider the following p-value combinations.

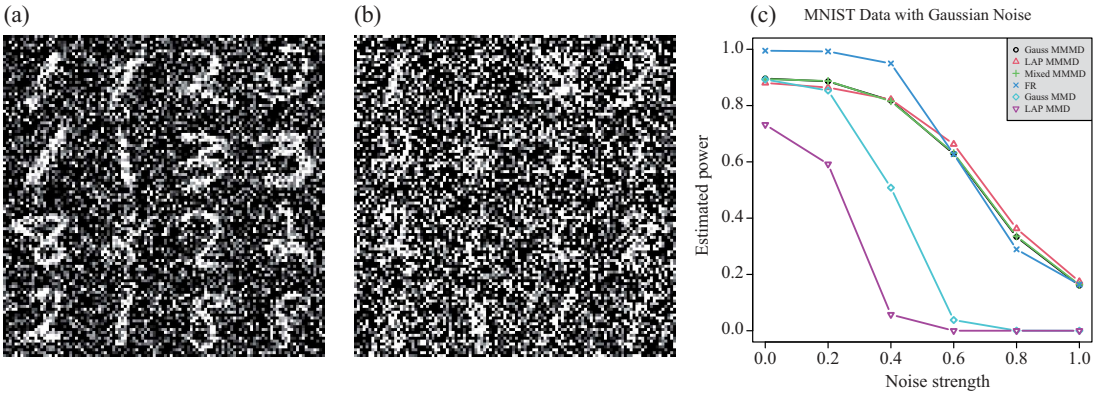

Fig. 3. MNIST data with additive Gaussian noise with (a) $\sigma = 0.6$, (b) $\sigma = 1$ and (c) estimated powers of the tests with increasing noise strength.

(i) Bonferonni setting: reject H_0 if $5 \min_{1 \leqslant i \leqslant 5} p_i \leqslant \alpha$.
(ii) Harmonic mean: reject H_0 if $2.214\,7295 \log(5) / \sum_{i=1}^{5} p_i^{-1} \leqslant \alpha$.
(iii) Bonferonni and geometric mean: reject H_0 if $2 \min\{5 \min_{i=1}^{5} p_i, e \prod_{i=1}^{5} p_i^{1/5}\} \leqslant \alpha$.

Here, p_i denotes the p-value of the MMD test for a Gaussian kernel with bandwidth σ_i for all $1 \leqslant i \leqslant 5$, and the significance level $\alpha = 0.05$. The validity of the above p-value combinations follows from Vovk & Wang (2020).

The results of our experiments are given in the Supplementary Material. In all the simulation settings considered, the `Gauss Mahalanobis MMD` test emerges as the clear winner. This suggests that it is more advantageous to adopt our aggregation strategy over p-value combination methods for boosting the performance of kernel two-sample tests.

7. Real data applications

7.1. *MNIST with additive Gaussian noise*

In this subsection we illustrate the performance of the proposed test in detecting different set of digits from the Modified National Institute of Standards and Technology (MNIST) database when independent and identically distributed Gaussian noise with standard deviation σ is added to each pixel. Figure 3 shows such noisy data for (a) $\sigma = 0.6$ and (b) $\sigma = 1$.

To evaluate the proposed method, we consider the sets of digits $P = \{1, 2, 3\}$ and $Q = \{1, 2, 8\}$, and vary the standard error $\sigma \in (0, 0.2, 0.4, 0.6, 0.8, 1)$. For each σ, we draw 100 samples with replacements from the two sets and check if the tests successfully reject H_0 at level $\alpha = 0.5$. We repeat this experiment 500 times to estimate the power. Figure 3(c) shows performance of the above-mentioned tests, where we plot the power over the index of pairs of sets of digits. This shows that, for the clean data and small noise levels, the single-kernel `Gauss MMD` test performs comparably to the Mahalanobis MMD tests. However, for larger noise levels, the Mahalanobis MMD tests perform much better than the single-kernel tests. The Friedman–Rafsky test also performs well in this case across the range of noise levels.

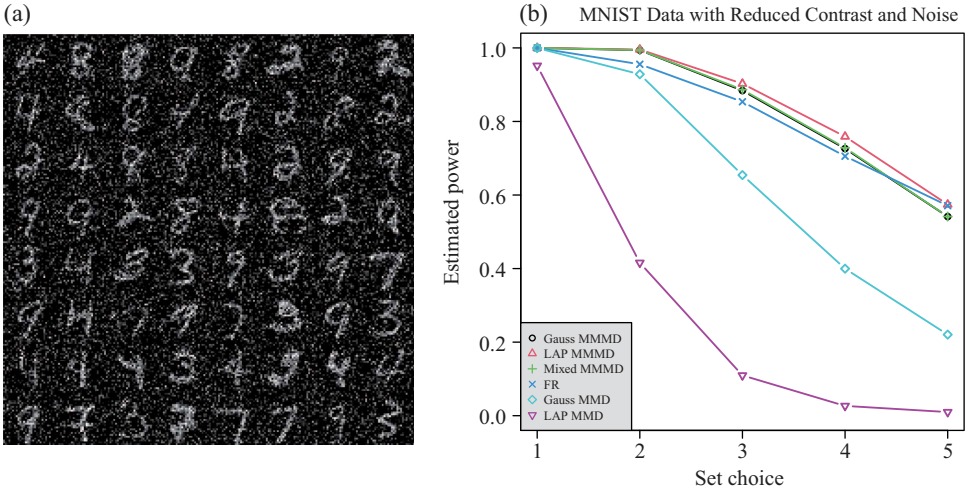

Fig. 4. (a) MNIST dataset with reduced contrast and additive noise and (b) estimated power.

7.2. *MNIST with reduced contrast and additive Gaussian noise*

In this subsection we illustrate the performance of the different tests on the noisy version of the MNIST dataset considered in Basu et al. (2017), which is publicly available at https://csc.lsu.edu/saikat/n-mnist/. Here, in addition to additive Gaussian noise the contrast of the images is also reduced. Specifically, the contrast range is scaled down to half and an additive Gaussian noise is introduced with a signal-to-noise ratio of 12. This emulates background clutter along with significant change in lighting conditions; see Fig. 4 for an example of such a noisy image.

We evaluate the performance of the different test for the following five pairs of sets of digits: (i) $P = \{2, 4, 8, 9\}$ and $Q = \{3, 4, 7, 9\}$; (ii) $P = \{1, 2, 4, 8, 9\}$ and $Q = \{1, 3, 4, 7, 9\}$; (iii) $P = \{0, 1, 2, 4, 8, 9\}$ and $Q = \{0, 1, 3, 4, 7, 9\}$; (iv) $P = \{0, 1, 2, 4, 5, 8, 9\}$ and $Q = \{0, 1, 3, 4, 5, 7, 9\}$; and (v) $P = \{0, 1, 2, 4, 5, 6, 8, 9\}$ and $Q = \{0, 1, 3, 4, 5, 6, 7, 9\}$. For each of the five cases, we draw 150 samples with replacements from the two sets and check if the tests successfully reject H_0 at level $\alpha = 0.5$. We repeat this experiment 500 times to estimate the power. Figure 4 shows the power of the different methods for the above five sets. In this case, the multiple kernel tests and the Friedman–Rafsky test overall has the highest power across the five sets, followed by the `Gauss MMD` and the `Lap MMD`.

8. AGGREGATION WITH AN INCREASING NUMBER OF KERNELS

8.1. *Maximum and L_2 aggregations*

There are many ways in which one can combine multiple kernels into a test statistic. For instance, we could consider maximum or L_2-based aggregations as (assuming that $m = n$ for simplicity)

$$T_m^{\max} := \max_{a=1}^{r_m} \mathrm{MMD}^2[\mathsf{K}_a, \mathscr{X}_m, \mathscr{Y}_m] \quad \text{and} \quad T_m^{L_2} := \|\mathrm{MMD}^2[\mathcal{K}_{r_m}, \mathscr{X}_m, \mathscr{Y}_m]\|_2,$$

where $r = r_m$ depends on m and $\mathcal{K}_{r_m} := \{\mathsf{K}_a : 1 \leqslant a \leqslant r_m\}$. The consistency and asymptotic distribution of these statistics when r is fixed follow from the results in § 2.2 and Theorem 1,

respectively. In the following proposition, using uniform convergence bounds for the MMD estimate, we construct tests based on $T_m^{\max}$ and $T_m^{L_2}$ that are consistent in the growing r regime. The proof is given in the Supplementary Material.

PROPOSITION 1. *Suppose that* $\mathcal{K}_{r_m} = \{\mathsf{K}_a : 1 \leqslant a \leqslant r_m\}$ *is a collection of* r_m *distinct characteristic kernels such that* $0 \leqslant \mathsf{K}_a \leqslant K$ *for all* $1 \leqslant a \leqslant r_m$. *Fix* $\alpha \in (0,1)$, *and consider the test functions*

$$\phi_m^{\max} := \mathbb{1}\left\{|T_m^{\max}| > C\left(\frac{1}{m}\log\frac{6r_m}{\alpha}\right)^{1/2}\right\},$$
$$\phi_m^{L_2} := \mathbb{1}\left\{|T_m^{L_2}| > C\left(\frac{r_m}{m}\log\frac{6r_m}{\alpha}\right)^{1/2}\right\},$$

where $C := 8K$. *Then both* $\phi_m^{\max}$ *and* $\phi_m^{L_2}$ *have level* α *in finite samples. Moreover,* $\phi_m^{\max}$ *and* $\phi_m^{L_2}$ *are asymptotically consistent for* (1) *if* $\log r_m = o(m)$ *and* $r_m \log r_m = o(m)$, *respectively.*

8.2. *Mahalanobis aggregation*

In the growing r regime the Mahalanobis MMD statistic takes the form

$$T_m^{\mathrm{MA}} := (\mathrm{MMD}^2[\mathcal{K}_{r_m}, \mathscr{X}_m, \mathscr{Y}_m])^{\mathrm{T}}\,\hat{\underline{\Sigma}}_{r_m}^{-1}(\mathrm{MMD}^2[\mathcal{K}_{r_m}, \mathscr{X}_m, \mathscr{Y}_m]), \tag{38}$$

where $\hat{\underline{\Sigma}}_{r_m}$ has entries defined in (15).

THEOREM 4. *Suppose that the assumptions of Proposition 1 hold. Then the test*

$$\phi_m^{MA} := \mathbb{1}\left\{|T_m^{MA}| > \frac{64K^2}{\sqrt{m}}\right\}$$

is asymptotically consistent whenever $\lim_{m\to\infty}\inf_{1\leqslant a\leqslant r_m}\{\mathrm{MMD}^2[\mathcal{F}_a, P, Q]\} > 0$ *and* $r_m \log r_m = o(\lambda_m\sqrt{m})$, *where* λ_m *is the smallest eigenvalue of* $\underline{\Sigma}_{r_m}$.

The proof of Theorem 4 is given in the Supplementary Material. Essentially, the result shows that T_m^{MA} leads to a consistent test for $r_m = o(\lambda_m\sqrt{m})$, ignoring logarithmic factors. In comparison, for the maximum aggregation, one can have r_m grow subexponentially in m and the L_2 aggregation allows any sublinear growth for r_m (recall Proposition 1). One of the challenges in dealing with the Mahalanobis MMD statistic in the growing r_m regime is that, in addition to the concentration of the vector of the MMD statistic, one has to ensure the concentration of $\underline{\Sigma}_{r_m}$, which necessitates a more stringent requirement on r_m, in comparison to the L_2 aggregation, to guarantee consistency. Towards this, it is expected that the lowest eigenvalue of the population covariance matrix $\underline{\Sigma}_{r_m}$ plays a role in how large r_m can be.

Improving the dependence on r in the above results and investigating the behaviour of the various aggregation strategies when r is comparable or even larger than m are important future directions. However, from a practical standpoint one needs to exercise caution while selecting r. Although the aggregated tests remain consistent under appropriate growth conditions on r, in finite samples the power of the tests saturate and the tests also become conservative when r is large. This latter issue, which is already apparent for the single-kernel test when the cut-off is chosen based on concentration inequalities (see § 4 of Gretton et al.,

2012a), becomes more significant when r grows with m. Moreover, the computation of $\hat{\underline{\Sigma}}^{-1}$ becomes less stable when r becomes too large. In practice, as we see in the simulations, there is already significant improvement in power over single-kernel tests just by aggregating over a few, up to five, kernels. Further exploring the interplay between the choice of r, the Type-I error and power is an interesting future direction.

9. Broader scope I: local power of adaptive MMD tests

The idea of using multiple kernels/bandwidths has recently emerged as a popular alternative to selecting a single bandwidth for developing adaptive kernel two-sample tests that do not require data splitting. In this direction, Kübler et al. (2020) proposed a method that does not require data splitting using the framework of postselection inference. However, this method requires asymptotic normality of the test statistic under H_0; hence, it is restricted to the linear-time MMD estimate (Gretton et al., 2012a, § 6), which leads to loss in power when compared to the more commonly used quadratic time estimate (7). Fromont et al. (2012, 2013) and, more recently, Schrab et al. (2023) introduced another non-asymptotic aggregated test, hereafter referred to as MMDAgg, that is adaptive minimax up to an iterated logarithmic term over Sobolev balls.

Our aggregation strategy leads to a test that can be efficiently implemented, enjoys improved empirical power over single-kernel tests for a range of alternatives and scales well in high dimensions. Moreover, our theoretical results apply to general aggregation schemes, using which we can obtain the asymptotic local power of the aforementioned MMDAgg test. To demonstrate this, in this section we propose an asymptotic implementation of the MMDAgg test and sketch a heuristic argument that derives its limiting local power in the contamination model (31). The argument can be made rigorous by using tools from empirical process theory; however, since the purpose of this section is more illustrative than technical, we have not pursued this direction.

To describe the asymptotic version of the MMDAgg test, suppose that $\mathcal{K} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_r\}$ is a finite collection of kernels and that $\mathcal{W} := \{w_1, w_2, \ldots, w_r\}$ is an associated collection of positive weights such that $\sum_{s=1}^{r} w_s \leqslant 1$. Moreover, for $\alpha \in (0, 1)$ and $1 \leqslant s \leqslant r$, let $\hat{q}_{1-\alpha,s,m}$ be the αth quantile of the distribution

$$\mathscr{E}(\mathsf{K}_s, \mathscr{X}_m) := \underline{Z}_m^{\mathrm{T}} \hat{\underline{\mathsf{K}}}_s^{\circ} \underline{Z}_m - \frac{1}{\hat{\rho}(1-\hat{\rho})} \operatorname{tr}[\hat{\underline{\mathsf{K}}}_s^{\circ}],$$

where $\hat{\underline{\mathsf{K}}}_s^{\circ}$ is as defined in (26) for $1 \leqslant a \leqslant r$ and $\underline{Z}_m \sim \mathcal{N}_m\{0, I/\hat{\rho}(1-\hat{\rho})\}$ is independent of $\mathscr{X}_m$. The idea of the MMDAgg test is to reject H_0 if any one of the individual (single-kernel) tests based on the kernels in $\mathcal{K}$ rejects H_0 for a specially chosen cut-off; see Schrab et al. (2023, § 3.5) for details. Here, we consider an alternative implementation of the MMDAgg test based on the Gaussian multiplier bootstrap discussed in § 4. To this end, define

$$u^*_{\alpha,m} := \arg\max \left\{ u \in (0, L) \colon \operatorname{pr}\left(\max_{1 \leqslant s \leqslant r} \{\mathscr{E}(\mathsf{K}_s, \mathscr{X}_m) - \hat{q}_{1-uw_s,s,m}\} > 0 \;\middle|\; \mathscr{X}_m \right) \leqslant \alpha \right\},$$

where $L := \min_{1 \leqslant s \leqslant r} w_s^{-1}$. The probability on the right-hand side above is over the randomness of $\underline{Z}_m$, conditional on $\mathscr{X}_m$; hence, $u^*_{\alpha,m}$ can be computed from the data by a grid search

over $u \in (0, L)$. The MMDAgg test would then reject H_0 if

$$\phi^{\mathrm{MMDAgg}}_{m,n,\alpha} := \mathbb{1}\left\{ \max_{1 \leqslant s \leqslant r} \{\mathrm{MMD}^2[\mathsf{K}_s, \mathscr{X}_m, \mathscr{Y}_n] - \hat{q}_{1-w_s u^*_\alpha, s, m}\} > 0 \right\}. \tag{39}$$

To describe the asymptotic properties of this test, let $q_{\alpha,s}$ be the αth quantile of the distribution $I_2(\mathsf{K}^\circ_s)/\rho(1-\rho)$ for $1 \leqslant s \leqslant r$. Then, for each fixed $u \in (0, L)$, by Theorem 1, Slutsky's theorem and the continuous mapping theorem, as $m \to \infty$,

$$\max_{1 \leqslant s \leqslant r} \{\mathscr{E}(\mathsf{K}_s, \mathscr{X}_m) - \hat{q}_{1-uw_s,s,m}\} \xrightarrow{\mathrm{D}} \max_{1 \leqslant s \leqslant r} \left\{ \frac{1}{\rho(1-\rho)} I_2(\mathsf{K}^\circ_s) - q_{1-uw_s,s} \right\},$$

since $\hat{q}_{1-uw_s,s,m} | \mathscr{X}_m \xrightarrow{\mathrm{a.s.}} q_{1-uw_s,s}$. Therefore, for each fixed $u \in (0, L)$, as $m \to \infty$,

$$\begin{aligned} &\mathrm{pr}\left(\max_{1 \leqslant s \leqslant r} \{\mathscr{E}(\mathsf{K}_s, \mathscr{X}_m) - \hat{q}_{1-uw_s,s,m}\} > 0 \,\middle|\, \mathscr{X}_m \right) \\ &\quad \to \mathrm{pr}\left(\max_{1 \leqslant s \leqslant r} \left\{ \frac{1}{\rho(1-\rho)} I_2(\mathsf{K}^\circ_s) - q_{1-uw_s,s} \right\} > 0 \right). \end{aligned}$$

Now, since the convergence of the quantiles is uniform, we expect the following to hold as $m \to \infty$: $u^*_{\alpha,m} \xrightarrow{\mathrm{a.s.}} u^*_\alpha$ and $\hat{q}_{1-w_s u^*_{\alpha,m}, s, m} \mid \mathscr{X}_m \xrightarrow{\mathrm{a.s.}} q_{1-w_s u^*_\alpha, s}$, where

$$u^*_\alpha := \arg\max \left[u \in (0, L) \colon \mathrm{pr}\left(\max_{1 \leqslant s \leqslant r} \left\{ \frac{1}{\rho(1-\rho)} I_2(\mathsf{K}^\circ_s) - q_{1-uw_s,s} \right\} > 0 \right) \leqslant \alpha \right].$$

Hence, under H_1 as in (31), by Theorem 3, Slutsky's theorem and the continuous mapping theorem,

$$\max_{1 \leqslant s \leqslant r} \{\mathrm{MMD}^2[\mathsf{K}_s, \mathscr{X}_m, \mathscr{Y}_n] - \hat{q}_{1-w_s u^*_\alpha, s, m}\} \xrightarrow{\mathrm{D}} \max_{1 \leqslant s \leqslant r} \{G_{\mathsf{K}_s,h} - q_{1-w_s u^*_\alpha, s}\},$$

where $G_{\mathsf{K}_s,h} := \gamma I_2(\mathsf{K}^\circ_s) + 2h\gamma^{1/2} I_1(\mathsf{K}^\circ_s[g/f_P]) + h^2 \mu_s$, and μ_s is as defined in (33). Therefore, the limiting power of test (39) is given by

$$\begin{aligned} \lim_{m,n \to \infty} \mathrm{E}_{H_1}[\phi^{\mathrm{MMDAgg}}_{m,n,\alpha}] &= \mathrm{pr}\left(\max_{1 \leqslant s \leqslant r} \{G_{\mathsf{K}_s,h} - q_{1-w_s u^*_\alpha, s}\} > 0 \right) \\ &= 1 - \underline{F}_{\mathcal{K},h}(q_{1-w_1 u^*_\alpha, 1}, \ldots, q_{1-w_r u^*_\alpha, r}), \end{aligned}$$

where $\underline{F}_{\mathcal{K},h}$ is the cumulative distribution function of vector $(G_{\mathsf{K}_1,h}, G_{\mathsf{K}_2,h}, \ldots, G_{\mathsf{K}_r,h})^{\mathrm{T}}$.

Numerical results comparing the empirical power of the Mahalanobis MMD test with the MMDAgg test are reported in the Supplementary Material. The experiments show that the Mahalanobis MMD test has better power than the MMDAgg test for a range of alternatives, which include perturbed uniform distributions in the Sobolev class, as well as mixture and local alternatives. This showcases both the practical relevance of the Mahalanobis aggregation strategy and the broader scope of our asymptotic results.

10. BROADER SCOPE II: AGGREGATION WITH DIVERGING BANDWIDTHS

In the previous sections we established the universal consistency and derived the asymptotic null distribution of the Mahalanobis MMD test for kernels with fixed bandwidths. However, in practice, bandwidths are often chosen in a data-driven manner that depends on the sample size N. For instance, to obtain tests that are optimal, in detecting smooth departures from the null hypothesis, the scaling parameter $\lambda := 1/\sigma^2$ has to diverge with the sample size; see Li & Yuan (2019) and Schrab et al. (2023). For such choices of the scaling parameter, the test statistic has an asymptotically normal distribution under H_0; hence, the rejection threshold can be readily obtained without any permutation/bootstrap resampling. Combining this idea with our aggregation strategy, in this section we construct a new test that combines multiple Gaussian kernels, with appropriately chosen diverging scaling parameters, that has a multivariate normal distribution under H_0.

For $\lambda > 0$, let $\mathsf{K}_\lambda(x, y) := e^{-\lambda\|x-y\|^2}$ be the Gaussian kernel with scaling parameter λ. For $r \geqslant 1$, consider the collection of kernels $\mathcal{K}_{\underline{\nu}} := \{\mathsf{K}_{\nu_a} : 1 \leqslant a \leqslant r\}$, where $\underline{\nu} = (\nu_1, \nu_2, \ldots, \nu_r)$ is a set of scaling parameters that can possibly depend on N. Throughout this section, we make the following assumptions.

Assumption 3. There exists $\{\eta_s > 0 : 1 \leqslant s \leqslant r\}$ such that $\nu_s = \eta_s \lambda_N$ for all $1 \leqslant s \leqslant r$, where $\lambda_N = o(N^{4/d})$ such that $\lambda_N \to \infty$, in the asymptotic regime (9).

Assumption 4. Suppose that $\mathcal{X} = \mathbb{R}^d$ for $d \geqslant 1$ and that distribution P has a density $f_P \in L_2(\mathbb{R}^d)$ with respect to the Lebesgue measure on $\mathbb{R}^d$.

Under these assumptions, we have the following theorem.

THEOREM 5. *Suppose that the collection of kernels $\mathcal{K}_{\underline{\nu}}$ satisfies Assumption 3 and that Assumption 4 holds. Then, under H_0, in the asymptotic regime (9),*

$$\frac{mn}{2^{1/2}(m+n)} \lambda_N^{d/4} \mathrm{MMD}^2[\mathcal{K}_{\underline{\nu}}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{\mathrm{D}} \mathcal{N}_r(0, \Gamma),$$

where $\Gamma = (\gamma_{ab})_{1 \leqslant a,b \leqslant r}$ is an $r \times r$ matrix with entries $\gamma_{ab} = \pi^{d/2}\|f_P\|_2^2/(\eta_a + \eta_b)^{d/2}$ for $1 \leqslant a, b \leqslant r$.

In Lemma K.6 in the Supplementary Material we provide a consistent estimate $\|\hat{f}_P\|_2^2$ of $\|f_P\|_2^2$, as in Theorem 4 of Li & Yuan (2019). Combining Theorem 5 and Lemma K.6 gives the following result.

COROLLARY 4. *Suppose that the conditions of Theorem 5 hold. Then, under H_0, in the asymptotic regime (9),*

$$\underline{V}_{m,n} := \frac{mn}{2^{1/2}(m+n)} \lambda_N^{d/4} \|\hat{f}_P\|_2^{-1} \mathrm{MMD}^2[\mathcal{K}_{\underline{\nu}}, \mathscr{X}_m, \mathscr{Y}_n] \xrightarrow{\mathrm{D}} \mathcal{N}_r(0, \tilde{\underline{\Gamma}}),$$

where $\|\hat{f}_P\|_2$ is defined in Lemma K.6 in the Supplementary Material and $\tilde{\underline{\Gamma}} = (\tilde{\gamma}_{ab})_{1 \leqslant a,b \leqslant r}$ is an $r \times r$ matrix with entries $\tilde{\gamma}_{ab} = \{\pi/(\eta_a + \eta_b)\}^{d/2}$ for $1 \leqslant a, b \leqslant r$. Consequently, $\{\underline{V}_{m,n}^{\mathrm{T}} \tilde{\underline{\Gamma}}^{-1} \underline{V}_{m,n} > \chi^2_{r,1-\alpha}\}$ is an asymptotically level-α test.

The test above has a tractable chi-squared distribution under H_0; hence, its rejection region can be readily obtained without any bootstrap resampling (unlike the general test

with fixed bandwidths discussed in § 4). Furthermore, the test in Corollary 4 will be optimal in detecting certain smooth alternatives for an appropriately chosen bandwidth, depending on the smoothness parameter, similar to the single-kernel test; see § 3 of Li & Yuan (2019). Moreover, we expect the test in Corollary 4 to have better power than its single-sample counterpart in finite samples for specific types of smooth alternatives.

Acknowledgement

The authors are grateful to the editor, associate editor and the referees for their insightful comments that led to several new results and greatly improved the quality and presentation of the paper. Bhattacharya was supported by the National Science Foundation and a Sloan Research Fellowship.

Supplementary material

The Supplementary Material includes proofs of the theorems, additional simulations, the definition and properties of multiple Weiner–Itô stochastic integrals, and a derivation of the joint distribution of the MMD estimates under the alternative.

References

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *Ann. Math. Statist.* **33**, 1148–59.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.* **68**, 337–404.

Aslan, B. & Zech, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *J. Statist. Comp. Simul.* **75**, 109–19.

Banerjee, B. & Ghosh, A. K. (2022). On high dimensional behaviour of some two-sample tests based on ball divergence. *arXiv:* 2212.08566v1.

Baringhaus, L. & Franz, C. (2004). On a new multivariate two-sample test. *J. Mult. Anal.* **88**, 190–206.

Basu, S., Karki, M., Ganguly, S., DiBiano, R., Mukhopadhyay, S., Gayaka, S., Kannan, R. & Nemani, R. (2017). Learning sparse feature representations using probabilistic quadtrees and deep belief nets. *Neural Proces. Lett.* **45**, 855–67.

Bhattacharya, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *J. R. Statist. Soc.* B **81**, 575–602.

Bickel, P. J. (1969). A distribution free version of the Smirnov two sample test in the *p*-variate case. *Ann. Math. Statist.* **40**, 1–23.

Biswas, M., Mukhopadhyay, M. & Ghosh, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika* **101**, 913–26.

Chen, H. & Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Am. Statist. Assoc.* **112**, 397–409.

Chwialkowski, K. P., Ramdas, A., Sejdinovic, D. & Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. In *Proc. 28th Int. Conf. Neural Info. Proces. Syst.*, vol. 2, pp. 1981–9. Cambridge, MA: MIT Press.

Deb, N. & Sen, B. (2021). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *J. Am. Statist. Assoc.* **118**, 192–207.

Friedman, J. H. & Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7**, 697–717.

Fromont, M., Laurent, B. & Reynaud-Bouret, P. (2013). The two-sample problem for poisson processes: adaptive tests with a nonasymptotic wild bootstrap approach. *Ann. Statist.* **41**, 1431–61.

Fromont, M., Lerasle, M. & Reynaud-Bouret, P. (2012). Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *Proc. 25th Ann. Conf. Learn. Theory*, pp. 23.1–23. PMLR.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012a). A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–73.

Gretton, A., Fukumizu, K., Harchaoui, Z. & Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. In *Proc. 22nd Int. Conf. Neural Info. Proces. Syst.*, pp. 673–81. Red Hook, NY: Curran Associates.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K. & Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Proc. 25th Int. Conf. Neural Info. Proces. Syst.*, vol. 1, pp. 1205–13. Red Hook, NY: Curran Associates.

Hall, P. & Tajvidi, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89**, 359–74.

Henze, N. (1984). On the number of random points with nearest neighbour of the same type and a multivariate two-sample test. *Metrika* **31**, 259–73.

Kim, I., Balakrishnan, S. & Wasserman, L. (2020). Robust multivariate nonparametric tests via projection averaging. *Ann. Statist.* **48**, 3417–41.

Kim, I., Ramdas, A., Singh, A. & Wasserman, L. (2021). Classification accuracy as a proxy for two-sample testing. *Ann. Statist.* **49**, 411–34.

Kübler, J., Jitkrittum, W., Schölkopf, B. & Muandet, K. (2020). Learning kernel tests without data splitting. In *Proc. 34th Int. Conf. Neural Info. Proces. Syst.*, vol. 1, pp. 6245–55. Red Hook, NY: Curran Associates.

Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. New York: Springer.

Li, T. & Yuan, M. (2019). On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives. *arXiv:* 1909.03302v1.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A. & Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *Proc. 37th Int. Conf. Mach. Learn.*, pp. 6316–26. PMLR.

Liu, R. Y. & Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Am. Statist. Assoc.* **88**, 252–60.

Lopez-Paz, D. & Oquab, M. (2017). Revisiting classifier two-sample tests. In *Int. Conf. Learn. Representations*, pp. 1895–909. Red Hook, NY: Curran Associates.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18**, 50–60.

Pan, W., Tian, Y., Wang, X. & Zhang, H. (2018). Ball divergence: nonparametric two sample test. *Ann. Statist.* **46**, 1109–37.

Park, J. & Muandet, K. (2020). A measure-theoretic approach to kernel conditional mean embeddings. In *Proc. 34th Int. Conf. Neural Info. Proces. Syst.*, pp. 21247–59. Red Hook, NY: Curran Associates.

Ramdas, A., García Trillos, N. & Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19**, 47.

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A. & Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proc. 29th AAAI Conf. Artif. Intel.*, pp. 3571–7. New York: Washington, DC: AAAI Press.

R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Reed, M. & Simon, B. (1980). *Methods of Modern Mathematical Physics. I: Functional Analysis*. New York: Academic Press.

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Statist. Soc.* B **67**, 515–30.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Am. Statist. Assoc.* **81**, 799–806.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B. & Gretton, A. (2023). MMD aggregated two-sample test. *J. Mach. Learn. Res.* **24**, 1–81.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41**, 2263–91.

Shekhar, S., Kim, I. & Ramdas, A. (2022). A permutation-free kernel two-sample test. *arXiv:* 2211.14908v1.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.* **19**, 279–81.

Song, H. & Chen, H. (2023). Generalized kernel two-sample tests. *arXiv:* 2011.06127v4.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A. & Gretton, A. (2021). Generative models and model criticism via optimized maximum mean discrepancy. *arXiv:* 1611.04488v6.

Székely, G. J. & Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat* **5**, 1249–72.

Székely, G. J. & Rizzo, M. L. (2013). Energy statistics: a class of statistics based on distances. *J. Statist. Plan. Infer.* **143**, 1249–72.

Vovk, V. & Wang, R. (2020). Combining *p*-values via averaging. *Biometrika* **107**, 791–808.

Wald, A. & Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.* **11**, 147–62.

Weiss, L. (1960). Two-sample tests for multivariate distributions. *Ann. Math. Statist.* **31**, 159–64.

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics* **3**, 119–22.

Zhang, J.-T., Guo, J. & Zhou, B. (2024). Testing equality of several distributions in separable metric spaces: a maximum mean discrepancy based approach. *J. Economet.* **239**, 105286.

[*Received on* 1 *March* 2023. *Editorial decision on* 4 *September* 2024]