Summary and Discussion of: "A Kernel Two-Sample Test"

36-825: Statistics Journal Club

Sam Adhikari and Anthony Platanios

October 14, 2014

1 Introduction

The paper reviewed in this discussion addresses the problem of comparing samples from two probability distributions, by proposing statistical tests of the null hypothesis that these distributions are equal against the alternative hypothesis that these distributions are different (we call this the two-sample problem). Such tests have application in a variety of areas, as discussed in Gretton et al. (2012). In bioinformatics, for example, we want to compare microarray data from identical tissue types as measured by different laboratories, to detect whether the data may be analyzed jointly, or whether differences in experimental procedures have caused systematic differences in the data distributions. As a second application example, in database attribute matching it is desirable to merge databases containing multiple fields where it is not known in advance which fields correspond to each other. In that case the fields are matched by maximizing the similarity in the distributions of their entries.

The ideas presented in the paper can be divided into three main categories mentioned below:

1. Defining a distance metric for probability distributions based on distances between their Hilbert space mean embeddings. We call this metric the maximum mean discrepancy (MMD).
2. Computing empirical estimates of the MMD.
3. Performing hypothesis tests using empirical MMD estimates as the test statistic, in order to determine whether two samples were drawn from the same distribution or not.

Problem statement: Let x and y be random variables defined on a topological space $\mathcal{X}$, with distributions p and q. Define $X := \{x_1, \dots, x_m\}$ and $Y := \{y_1, \dots, y_n\}$ as independently and identically distributed samples from p and q respectively. The goal of this paper is to formulate a test statistic and do hypothesis testing to decide whether $p \neq q$.

Notation: Throughout this document we use the notation conventions used in the paper. Define $\mathbf{E}_x[f(x)] := \mathbf{E}_{x\sim p}[f(x)]$ and $\mathbf{E}_y[f(y)] := \mathbf{E}_{y\sim q}[f(y)]$, to denote expectations with respect to p and q respectively.

2 The Maximum Mean Discrepancy (MMD)

We want to test whether distributions p and q are different on the basis of samples drawn from each of them, by finding a well behaved (e.g., smooth) function which is large on the points drawn from p, and small (as negative as possible) on the points drawn from q. We use as our test statistic the difference between the mean function values on the two samples; when this is large, the samples are likely from different distributions. We call this test statistic the Maximum Mean Discrepancy (MMD).

The quality of the MMD as a statistic depends on the class $\mathcal{F}$ of smooth functions that define it. On one hand, $\mathcal{F}$ must be "rich enough" so that the population MMD vanishes if and only if $p = q$. On the other hand, for the test to be consistent in power, $\mathcal{F}$ needs to be "restrictive" enough for the empirical estimate of the MMD to converge quickly to its expectation as the sample size increases.

Lets start by reviewing a condition that lets us differentiate between p and q.

Lemma 1. *Let $(\mathcal{X},d)$ be a metric space, and let p, q be two Borel probability measures defined on $\mathcal{X}$. Then $p = q$ if and only if $\mathbf{E}_x(f(x)) = \mathbf{E}_y(f(y))\ \forall f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$.*

Note that while Lemma 1 defines a class of functions to uniquely identify whether $p = q$, $C(\mathcal{X})$ is a very rich function class and it is computationally impractical to work with such rich class in finite sample settings. This paper aims to present a way to use a more restrictive functional class $\mathcal{F}$ to measure differences between p and q in finite sample settings, in realizable time.

Definition 1. *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ and let p, q, x, y, X, Y be defined as above. Then, the maximum mean discrepancy (MMD) is defined as:*

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \left(\mathbf{E}_x[f(y)] - \mathbf{E}_y[f(y)] \right). \tag{1}$$

We want a smooth function $f \in \mathcal{F}$ that maximizes the difference in the expectations of the function, where the first expectation is over p and the second expectation is taken over q. A biased empirical estimate of the MMD can be obtained by replacing the population expectations with empirical expectations computed on the samples X and Y:

$$\mathrm{MMD}_b[\mathcal{F}, X, Y] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right). \tag{2}$$

Note that, $\mathbf{E}[\mathrm{MMD}_b[\mathcal{F}, X, Y]] \geq \mathrm{MMD}[\mathcal{F}, p, q]$ by Jensen's inequality.

The goal is then to identify a function class that is rich enough to uniquely identify whether $p = q$, yet restrictive enough to provide useful finite sample estimates.

3 The MMD in Reproducing Kernel Hilbert Spaces

A unit ball in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is used as the MMD function class $\mathcal{F}$. This class of functions has the following properties:

- The operator of evaluation δ_x mapping $f \in \mathcal{H}$ to $f(x) \in \mathbb{R}$ is continuous.
- By the Riesz representer theorem, we know that there exists a feature mapping $\phi(x) : \mathcal{X} \to \mathbb{R}$ such that $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$. $\phi(x)$ takes the canonical form $k(x, \cdot)$, where $k(x_1, x_2) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite, and the notation $k(x, \cdot)$ indicates the kernel has one argument fixed at x, and the second free.
- We have that $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = k(x, y)$.

Definition 2. *The mean embedding of p is defined as an element $\mu_p \in \mathcal{H}$ such that $\mathbf{E}_x[f] = \langle f, \mu_p \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.*

We now define a condition under which the mean embedding $\mu_p \in \mathcal{H}$ exists and can be used to later define the MMD in RKHS $\mathcal{H}$.

Lemma 2. *If $k(\cdot, \cdot)$ is measurable and $\mathbf{E}_x \sqrt{k(x, x)} < \infty$, then $\mu_p \in \mathcal{H}$.*

The proof for this lemma can be found in the main paper.

Lemma 3. *Assuming the condition in Lemma 2 for the existence of mean embeddings μ_p and μ_q is satisfied, we have that:*

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \|\mu_p - \mu_q\|^2_{\mathcal{H}}, \tag{3}$$

where $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$.

Proof. We have that:

$$\begin{aligned} \mathrm{MMD}^2[\mathcal{F}, p, q] &= \left[\sup_{f : \|f\|_{\mathcal{H}} \leq 1} (\mathbf{E}_x[f(y)] - \mathbf{E}_y[f(y)]) \right]^2, \\ &= \left[\sup_{f : \|f\|_{\mathcal{H}} \leq 1} \langle \mu_p - \mu_q, f \rangle_{\mathcal{H}} \right]^2, \\ &= \|\mu_p - \mu_q\|^2_{\mathcal{H}}, \end{aligned} \tag{4}$$

where the last step follows from the dual norm definition. □

We now establish a condition on the RKHS $\mathcal{H}$ under which the mean embedding μ_p is injective, which indicates that $\mathrm{MMD}[\mathcal{F}, p, q] = 0$ is a metric on the Borel probability measures on $\mathcal{X}$. This condition only holds for *universal RKHSs*.

Theorem 1. *Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, defined on the compact metric space $\mathcal{X}$, with associated continuous kernel $k(\cdot, \cdot)$. Then, MMD$[\mathcal{F}, p, q] = 0$ if and only if $p = q$.*

Proof. If $p = q$ then by Lemma 1, $\mathbf{E}_x f(x) = \mathbf{E}_y f(y) \Rightarrow \text{MMD}[\mathcal{F}, p, q] = 0$. We now prove the converse. By the universality of $\mathcal{H}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$ there exists a $g \in \mathcal{H}$, such that $\|f - g\|_\infty \leq \epsilon$. Therefore, we have that:

$$\begin{aligned}\|\mathbf{E}_x f(x) - \mathbf{E}_y f(y)\| &\leq |\mathbf{E}_x f(x) = \mathbf{E}_x g(x)| + |\mathbf{E}_y g(y) - \mathbf{E}_y f(y)| + |\mathbf{E}_x g(x) - \mathbf{E}_y g(y)|, \\ &\leq \mathbf{E}_x |f(x) - g(x)| + \mathbf{E}_y |g(y) - f(y)| + \langle g, \mu_p - \mu_q \rangle_{\mathcal{H}}, \\ &\leq 2\epsilon,\end{aligned} \tag{5}$$

for all $f \in C(\mathcal{X})$ and $\epsilon > 0$, which implies that $p = q$. In this derivation we made use of Jensen's inequality and of the fact that $\text{MMD}[\mathcal{F}, p, q] = 0 \Rightarrow \mu_p = \mu_q$. □

While this result establishes the mapping μ_p is injective for universal kernels on compact domains, this result can be shown to hold in more general cases, provided that we are using a *characteristic kernel*. It has been shown that both the Gaussian and the Laplace kernels are characteristic on $\mathbb{R}^d$ and thus the associated MMD is a metric on distributions for that domain.

Now we show how we can write the MMD in terms of the kernel function associated with the RKHS we are dealing with, and how we can estimate it using a finite set of samples from the two distributions of interest.

Lemma 4. *Given x and x' independent random variables with distribution p, and y and y' independent random variables with distribution q, the squared population MMD can be written as:*

$$\text{MMD}^2[\mathcal{F}, p, q] = \mathbf{E}_{x,x'}[k(x, x')] - 2\mathbf{E}_{x,y}[k(x, y)] + \mathbf{E}_{y,y'}[k(y, y')], \tag{6}$$

where x' is an independent copy of x with the same distribution, and y' is an independent copy of y. An unbiased empirical estimate is a sum of two U-statistics and a sample average:

$$\begin{aligned}\text{MMD}_u^2[\mathcal{F}, X, Y] = &\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(y_i, y_j) \\ &- \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j).\end{aligned} \tag{7}$$

Proof. We have that:

$$\begin{aligned}\text{MMD}_u^2[\mathcal{F}, X, Y] &= \|\mu_p - \mu_q\|_{\mathcal{H}}^2, \\ &= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q \rangle_{\mathcal{H}}, \\ &= \mathbf{E}_{x,x'} \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} + \mathbf{E}_{y,y'} \langle \phi(y), \phi(y') \rangle_{\mathcal{H}} - 2\mathbf{E}_{x,y} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}, \\ &= \mathbf{E}_{x,x'}[k(x, x')] - 2\mathbf{E}_{x,y}[k(x, y)] + \mathbf{E}_{y,y'}[k(y, y')].\end{aligned} \tag{8}$$

□

The biased MMD estimate of equation 2 can also be written in a similar way, by replacing the U-statistics with V-statistics:

$$\text{MMD}_b[\mathcal{F}, X, Y] = \left[\frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(y_i, y_j) \right]^{\frac{1}{2}}. \tag{9}$$

3.1 Witness Function of the MMD for RKHSs

Definition 3. *We define f^* as the RKHS function that attains the supremum in the MMD definition, and $\hat{f}^*$ to be the function that attains the supremum in the empirical MMD definition. Then from Lemma 3:*

$$f^*(t) \propto \langle \phi(t), \mu_p - \mu_q \rangle_{\mathcal{H}} \quad = \mathbf{E}_x[k(x,t)] - \mathbf{E}_y[k(y,t)], \tag{10}$$

$$\hat{f}^*(t) \propto \langle \phi(t), \mu_X - \mu_Y \rangle_{\mathcal{H}} = \frac{1}{m}\sum_{i=1}^{m} k(x_i,t) - \frac{1}{n}\sum_{i=1}^{n} k(y_i,t), \tag{11}$$

where $\mu_X = m^{-1}\sum_{i=1}^{m}\phi(x_i)$, and $\mu_Y = n^{-1}\sum_{i=1}^{n}\phi(y_i)$ similarly. This result follows from the fact that the vector ν maximizing $\langle \nu, x\rangle_{\mathcal{H}}$ in the unit ball we are considering is $\nu = x/\|x\|_{\mathcal{H}}$. We call the function that achieves the supremum, the witness function because it is the function that witnesses the difference in the two distributions.

Some plots of witness functions for the MMD that help understand the intuition behind it are shown in figures 1 and 2.

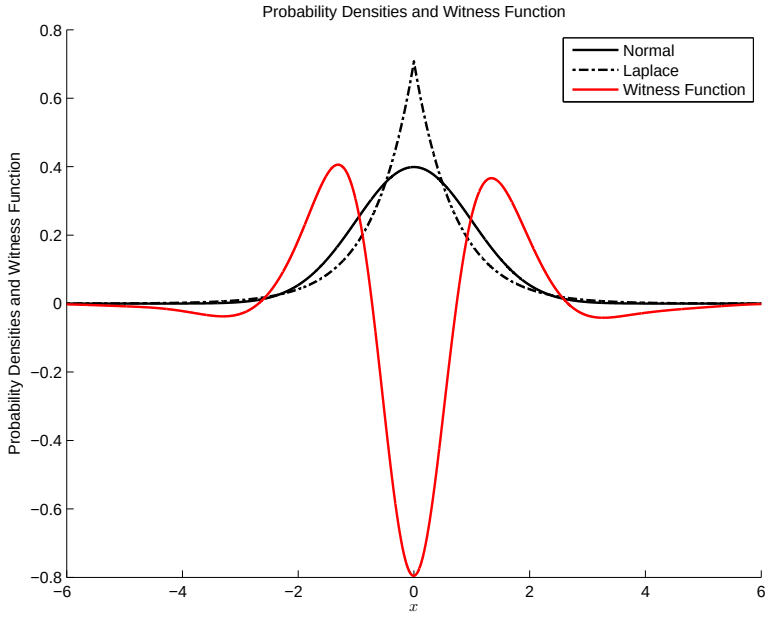

Figure 1: Our two distributions are a $\mathcal{N}(0,1)$ and a Laplace$(0,1)$. We drew $20,000$ samples from each one and we computed the MMD witness function using a Gaussian kernel with width $\sigma = 0.5$. We see that the witness function is positive whenever the Normal probability is greater than the Laplace probability and negative when the Laplace probability is greater. The magnitude of the witness function is a direct reflection of the amount by which one density exceeds the other, insofar as the smoothness constraint permits it. This result is intuitive and one can easily see why such a function would maximize the mean discrepancy in the associated RKHS.

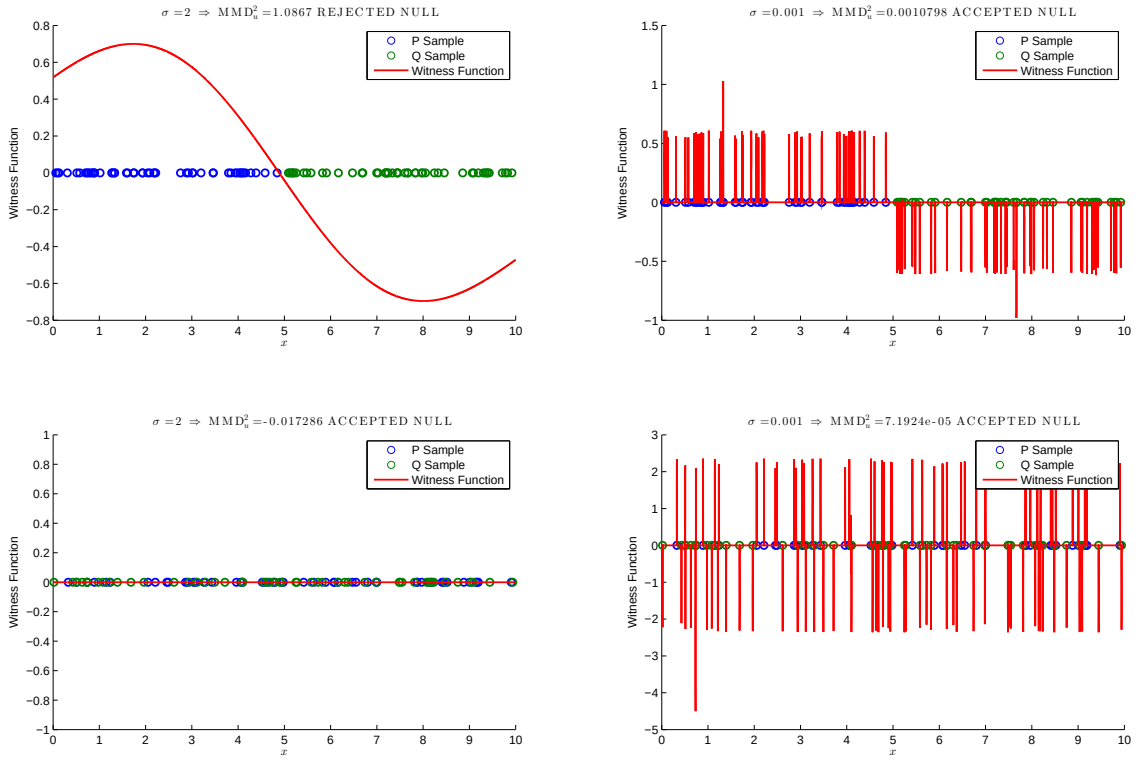

Figure 2: Here we show one-dimensional samples from distribution p as blue circles and one-dimensional samples from q as green circles and plot the corresponding MMD witness function for each case in red color. We also mention what kernel width σ was used for the MMD and whether the null hypothesis was accepted or rejected. We used the Gaussian kernel to create these simulations. In the top left it is easy to understand why the witness function looks that way and everything is fine since the null hypothesis is rejected which seems like the right thing based on the samples distributions. On the top right we have the same samples, but a very small kernel width was used. In this case the witness function becomes too "wiggly" and we see that the null hypothesis was accepted. This shows how important the selection of the kernel width value is. A reasonable choice for the kernel width is to make it equal to the median distance among all pairwise distances of points in our two data samples. The bottom figures show a corresponding example for when the two samples are drawn from the same distribution. Note that hypothesis testing with the MMD is discussed in section 4.

4 Hypothesis Testing

Given independent and identically distributed samples $X \sim p$, $Y \sim q$, our goal is to define a statistical test that distinguishes between the null hypothesis $H_0 : p = q$ and the alternative hypothesis $H_A : p \neq q$. This is performed by comparing test statistic (T) with a particular threshold (t) for a given level (α) test such that, if the test statistic exceeds the threshold the null hypothesis is rejected at the level. For an α level test, t is chosen such that $P(T > t) \leq \alpha$, where P is the distribution of T when null is true. The acceptance region

of the test is defined as the set of real numbers below the threshold, i.e. the region where null is more likely. In this section we first present some definitions related to hypothesis testing and then present some statistical tests that are performed using the MMD as the test statistic.

Definition 4. *A type I error is when H_0 is rejected while actually being true.*

Definition 5. *A type II error is when H_0 is accepted while actually being false. The power of a statistical test is defined as 1 minus the type II error.*

Definition 6. *For an α-level test, the probability of a type I error is upper bounded by α.*

Definition 7. *An α-level test is consistent if it achieves a type II error of zero in the limit of infinite data samples.*

Our goal is to create an α-level test, where we bound the type I error and make the type II error as low as possible.

4.1 Tests Based on Uniform Convergence Bounds

In this section, we first establish guarantees on convergence of empirical estimate of MMD (both biased and unbiased) to population MMD in probability. Then we proceed by finding threshold for level α test under the null.

4.1.1 Bound on the Biased Statistic and Test

We first provide a uniform convergence bound for the biased MMD estimate. We then use that bound to derive a hypothesis test that uses the MMD as its test statistic.

Theorem 2. *Let p, q, X, Y be defined as in the previous section, and assume $0 \leq k(x, y) \leq K$. Then, we have that:*

$$P_{X,Y}\left\{|\mathrm{MMD}_b[\mathcal{F}, X, Y] - \mathrm{MMD}[\mathcal{F}, p, q]| > 2\left(\sqrt{\frac{K}{m}} + \sqrt{\frac{K}{n}}\right) + \epsilon\right\} \leq 2e^{\frac{-\epsilon^2 mn}{2K(m+n)}}, \quad (12)$$

where, $P_{X,Y}$ denotes the probability over the m-sample X and the n-sample Y.

We now use Theorem 2 to compute a bound on probability for when $p = q$ and $m = n$, which we will use to derive our test threshold for accepting or rejecting the null hypothesis.

Theorem 3. *Under the conditions of Theorem 2 where additionally $p = q$ and $m = n$, we have that:*

$$\mathrm{MMD}_b[\mathcal{F}, X, Y] \leq m^{-1/2}\sqrt{2\mathbf{E}_{x,x'}[k(x, x) - k(x, x')]} + \epsilon \leq \sqrt{\frac{2K}{m}} + \epsilon, \quad (13)$$

with probability at least $1 - e^{-\frac{\epsilon^2}{4K}}$.

Corollary 4. *A hypothesis test of level α for the null hypothesis $p = q$, that is, for $\mathrm{MMD}[\mathcal{F}, p, q] = 0$, has the acceptance region $\mathrm{MMD}_b[\mathcal{F}, X, Y] < \sqrt{2K/m}(1 + \sqrt{2\log(\alpha^{-1})})$. This result was obtained by observing that $\alpha = e^{-\epsilon^2 m/4K} \Rightarrow \epsilon = \sqrt{4K\log(\alpha^{-1})/m}$.*

Note that that this test is *distribution-free* (i.e., the test threshold does not depend on the particular distribution that generated the sample). Theorem 2 guarantees the consistency of the test against fixed alternatives, and that the Type II error probability decreases to zero at rate $O(m^{1/2})$, assuming that $m = n$. Proofs for the results presented in this section are provided in the main paper's appendix.

4.1.2 Bound on the Unbiased Statistic and Test

By using the large deviation bound on U-statistics of Hoeffding (1963, p. 25) we derive the following result.

Theorem 5. *Assume $0 \leq k(x_i, x_j) \leq K$. Then, we have that:*

$$P_{X,Y}\left\{\mathrm{MMD}_u^2[\mathcal{F}, X, Y] - \mathrm{MMD}^2[\mathcal{F}, p, q] > t\right\} \leq e^{\frac{-t^2 m_2}{8K^2}}, \tag{14}$$

where $m_2 := \lfloor m/2 \rfloor$ (the same bound applies for deviations of $-t$ and below).

Then, in a similar way as in the previous section, we obtain the following hypothesis test.

Corollary 6. *A hypothesis test of level α for the null hypothesis $p = q$ has the acceptance region $\mathrm{MMD}_u^2 < (4K/\sqrt{m})\sqrt{\log(\alpha^{-1})}$.*

This is also a *distribution-free* test. Note that for a sufficiently large m, the threshold in Corollary 6 will be lower than that in Corollary 4 (it decreases at the rate of $O(m^{-1})$ instead of $O(m^{-1/2})$), and the type II error of the corresponding test will be lower, for a given type I error bound.

Note that both bounds we obtained based on uniform convergence bounds are rather conservative for a number of reasons; they do not take the actual distributions into account, and they are finite sample size, distribution-free bounds that hold even in the worst case scenario.

4.2 Tests Based on the Asymptotic Distribution of the Unbiased Statistic

The tests presented in this section use the asymptotic distribution of MMD_u^2 under H_0, which can be defined as follows.

Theorem 7. *Let $\tilde{k}(x_i, x_j)$ be the kernel between feature space mappings from which the mean embedding of p has been subtracted:*

$$\begin{aligned}\tilde{k}(x_i, x_j) &:= \langle \phi(x_i) - \mu_p, \phi(x_j) - \mu_p \rangle_{\mathcal{H}}, \\ &= k(x_i, x_j) - \mathbf{E}_x k(x_i, x) - \mathbf{E}_x k(x, x_j) + \mathbf{E}_{x,x'} k(x, x'),\end{aligned} \tag{15}$$

where x' is an independent copy of x drawn from p. Assume $\tilde{k} \in L_2(\mathcal{X} \times \mathcal{X}, p \times p)$ (i.e., the centred kernel is square integrable, which is true for all p when the kernel is bounded), and that for $t = m + n$, $\lim_{n,m\to\infty} m/t = \rho_x$, and $\lim_{n,m\to\infty} n/t = \rho_y := 1 - \rho_x$, for fixed $0 < \rho_x < 1$. Then, under H_0, MMD_u^2 converges in distribution to:

$$t\mathrm{MMD}_u^2[\mathcal{F}, X, Y] \rightsquigarrow \sum_{l=1}^{\infty} \lambda_l \left[\left(\frac{a_l}{\sqrt{\rho_x}} - \frac{b_l}{\sqrt{\rho_y}}\right)^2 - \frac{1}{\rho_x \rho_y}\right], \tag{16}$$

where $a_l \sim \mathcal{N}(0,1)$ and $b_l \sim \mathcal{N}(0,1)$ are infinite sequences of independent random variables, and the λ_l are the eigenvalues of:

$$\int_{\mathcal{X}} \tilde{k}(x,x')\psi_l(x)dp(x) = \lambda_l \psi_l(x') \tag{17}$$

Our goal is to determine whether the empirical test statistic MMD_u^2 is so large as to be outside the $1-\alpha$ quantile of the null distribution defined in this theorem, which gives an α-level test. In figure 3 we illustrate the MMD density under both the null and alternative hypotheses by approximating it empirically for $p = q$ and $p \neq q$.

4.2.1 Bootstrap Test (Permutation Test)

Under the null hypothesis we know that $p = q$ and so we know that if we shuffle the two data samples that we have together and draw two new data samples out of them, the MMD value should not change. Using this reasoning we can estimate the CDF of the MMD under the null hypothesis by re-shuffling the data samples many times and re-computing the MMD value. Then, if the original MMD value is higher than the $1-\alpha$ quantile of the empirical CDF we reject the null hypothesis, otherwise we accept it. The more data samples we have and the more bootstrapping iterations we do, the more powerful our test should become. Note that this is also a distribution-free hypothesis test.

4.2.2 Pearson Approximation Test

Another approach is to approximate the null distribution by fitting Pearson curves to its first four moments. Taking advantage of the degeneracy of the U-statistic, for $m = n$ we have that:

$$\begin{aligned}
\mathbf{E}\left[\left(\text{MMD}_u^2\right)^2\right] &= \frac{2}{m(m-1)}\mathbf{E}_{z,z'}[h^2(z,z')] \text{ and} \\
\mathbf{E}\left[\left(\text{MMD}_u^2\right)^3\right] &= \frac{8(m-2)}{m^2(m-1)^2}\mathbf{E}_{z,z'}\left[h(z,z')\mathbf{E}_{z''}[h(z,z'')h(z',z'')]\right] + O(m^{-4}),
\end{aligned} \tag{18}$$

where $h(z,z') = k(x,x') + k(y,y') - k(x,y') - k(x',y)$, $z = (x,y) \sim p \times q$ where x and y are independent, and z' and z'' are independent copies of z. The fourth moment, $\mathbf{E}\left[\left(\text{MMD}_u^2\right)^4\right]$ is not computed because it is both very small (i.e., $O(m^{-4})$) and very expensive to compute (i.e., $O(m^4)$). Instead we set $\text{kurt}(\text{MMD}_u^2) = 2\left(\text{skew}\left(\text{MMD}_u^2\right)^2 + 1\right)$, which is an approximation based on the bound $\text{kurt}(\text{MMD}_u^2) \geq \text{skew}\left(\text{MMD}_u^2\right)^2 + 1$. In figure 4 below we illustrate the empirical CDFs along with the Pearson approximation CDFs for the same examples as in figure 3. Using the obtained CDF approximation we can compute an appropriate testing threshold. This hypothesis testing method is faster than the bootstrap method, described in the previous section, but it is an approximation and is thus expected to have less power.

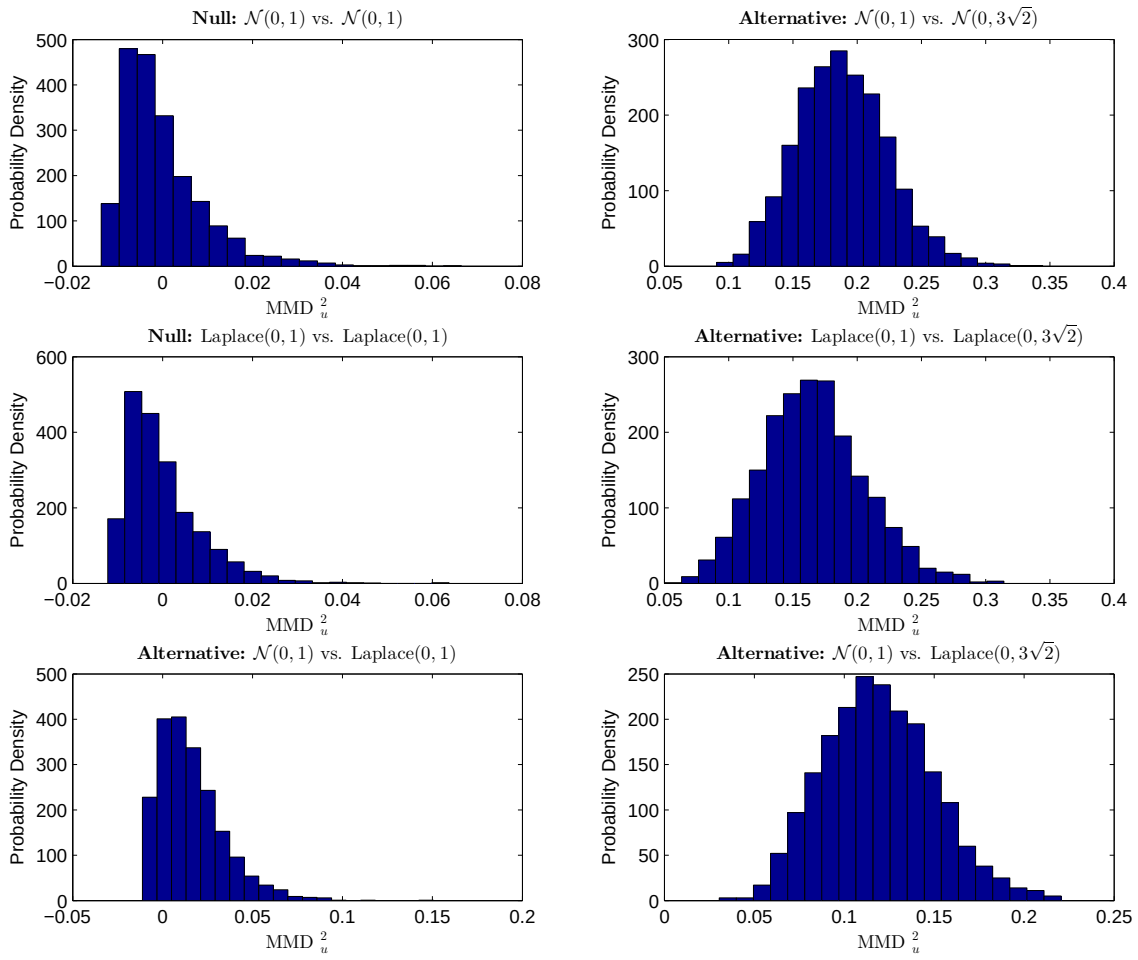

Figure 3: The empirical MMD_u^2 distribution under both the null and the alternative hypothesis, for several different cases. In order to obtain the empirical distributions we sampled 100 samples from each distribution and computed the MMD value, $2,000$ times independently, in order to obtain $2,000$ MMD values and construct the histograms. We did that for each case separately. It can be seen that in most cases the distribution under the null hypothesis is skewed left towards 0 and is very different than the distribution under the alternative hypothesis. However, in the bottom left plot we show an example where the distribution under the alternative hypothesis is skewed left towards 0 and is harder to distinguish whether this is a distribution of the MMD under the null hypothesis or not.

4.2.3 Gamma Approximation Test

An alternative, even more computationally efficient approach for testing than the Pearson approximation method, is using a two-parameter Gamma approximation for the CDF of the biased MMD estimate. We have that:

$$m\text{MMD}_b \sim \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, \tag{19}$$

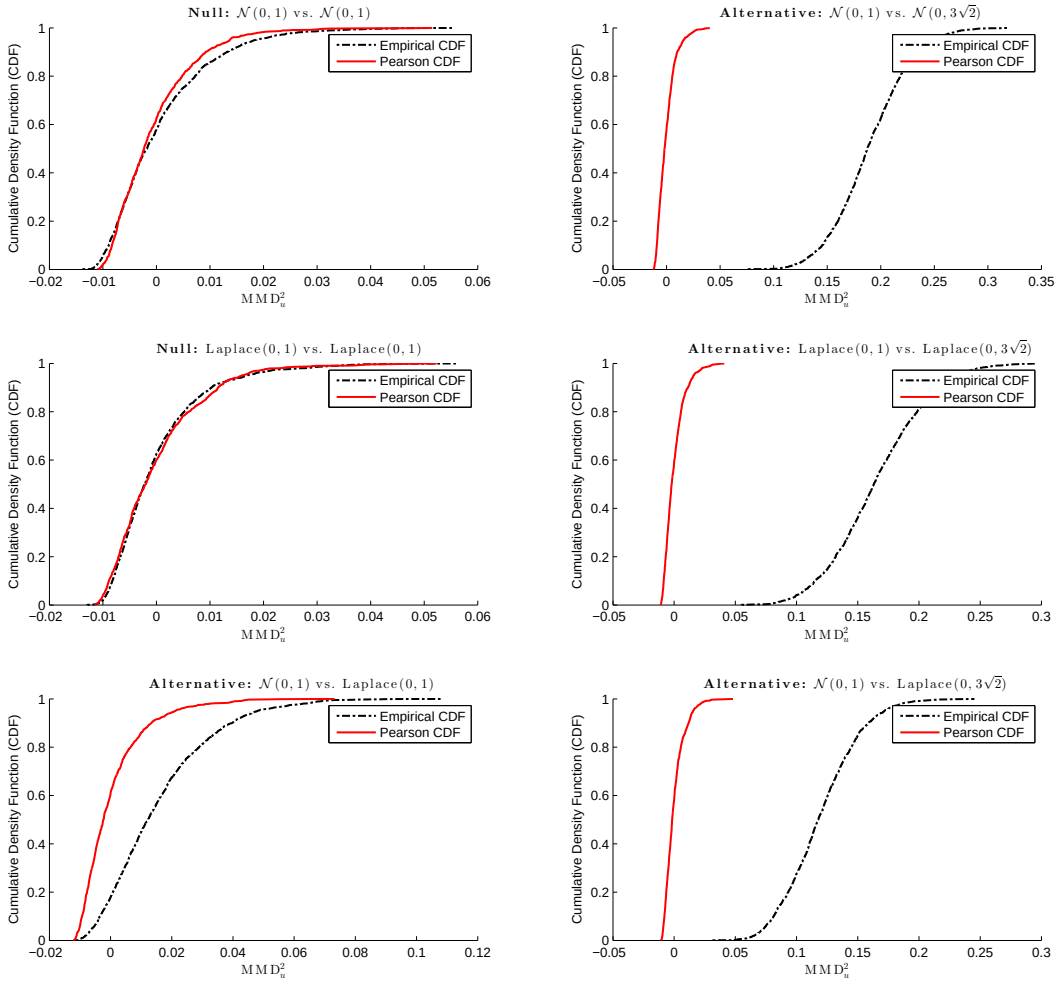

Figure 4: The empirical MMD_u^2 distribution under both the null and the alternative hypothesis and the corresponding CDF obtained using the Pearson approximation. For all these cases it is clear that the Pearson approximation works well for approximating the CDF of the unbiased empirical MMD estimate when the null hypothesis is true. The plots for the alternative hypothesis are not that meaningful, because in obtaining that approximation, it has been assumed that the null hypothesis is true. They have been included here for completeness, in order to parallel the plots of figure 3. The experiments that we run to obtain these plots are the same as the experiments we run to obtain the results of figure 3.

where $\alpha = \frac{(\mathbf{E}[\mathrm{MMD}_b])^2}{\mathrm{Var}[\mathrm{MMD}_b]}$ and $\beta = \frac{m\mathrm{Var}[\mathrm{MMD}_b]}{\mathbf{E}[\mathrm{MMD}_b]}$. This method is less accurate than the Pearson approximation method, but its computational complexity is $O(m^2)$, whereas the computational complexity of the Pearson approximation method is $O(m^3)$.

4.2.4 Spectrum Approximation Test

In Gretton et al. (2009), it was established that for large sample sizes, the null distribution of the MMD approaches an infinite weighted sum of independent χ_1^2 random variables. In that paper the authors describe a method of approximating that distribution by using empirical eigenvalues estimates. We are not going to describe this method here because it is fairly involved and not described in the main paper, but we are going to use it in our simulations. Interested readers can refer to Gretton et al. (2009) for further details on the method.

5 Comparison of the Tests

We performed two experiments to compare the tests described in the previous section as we vary the dimensionality of the problem at hand. We also compare against Hotelling's T-Squared test, which is expected to perform well when our two samples come from Normal distributions with different mean but same covariance matrices (or simply variance for the one-dimensional case). We run two separate simulations, which are described in the following two sections.

5.1 Normal Distributions with Different Mean and Same Covariance

We perform the following experiment for several different problem dimensionalities. For each problem dimensionality d, we repeat the following steps 100 times:

1. We go through steps 2-4 for 15 different values of μ logarithmically spaced from 0.05 to 50.
2. We sample 250 data samples from a $\mathcal{N}(0, \mathrm{I})$, where $\mathrm{I} \in \mathbb{R}^d$.
3. We sample 250 data samples from a $\mathcal{N}(\mu\mathbf{1}/\sqrt{d}, \mathrm{I})$, where $\mathbf{1}$ is a d dimensional vector of ones. Note that the division of each element of the mean vector by $\sqrt{d}$ retains the same Euclidean distance between the means for all dimensionalities (i.e., the hypothesis testing problem remains equivalently difficult to perform – we talk more about this particular topic in section 6).
4. Perform hypothesis testing using all of the tests described in the previous section.
5. Store the number of times the null hypothesis is accepted (i.e., the number of times we get a type II error).

In figure 5 we present the results of this experiment. It is clear that power drops with increasing problem dimensionality, which is expected. Another result that is expected is that the tests that are based on the uniform convergence bounds of the MMD are very weak compared to the other tests. That is because the assumptions made when deriving the testing threshold are too general and conservative. Finally, Hotelling's T-Squared test (labeled as "t-test" in the plot in order to follow the main paper naming conventions) performs better than all tests in low problem dimensions, but its performance degrades quickly when we consider higher dimensional problems. That is due to the sample covariance matrix of the data becoming very badly conditioned. This problem could probably be mitigated by using a good shrinkage method for that matrix. It is important to note here

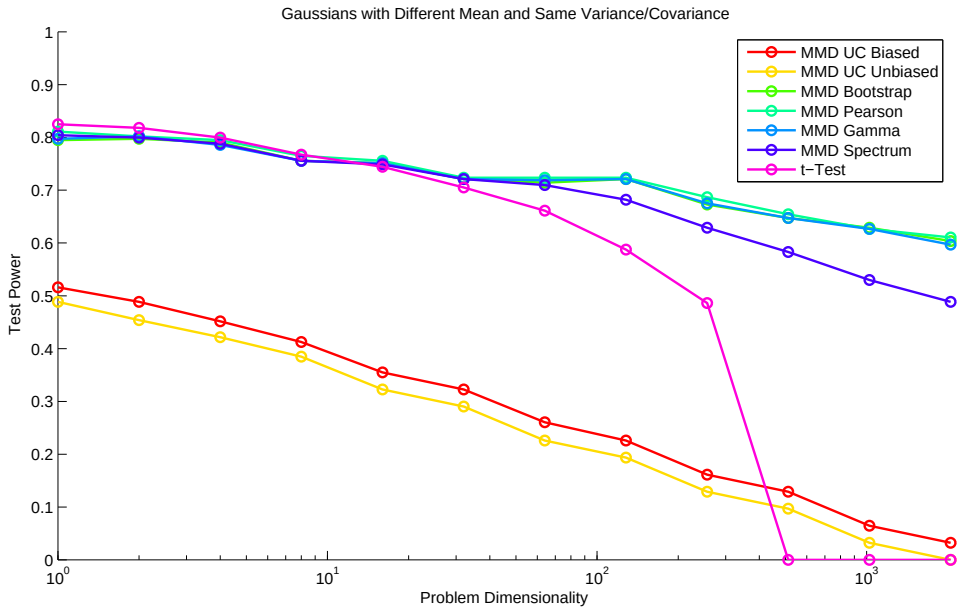

Figure 5: Our results for our two-sample test experiment when we consider Normal distributions with different mean and same covariance.

that we expect Hotelling's T-Squared test to perform that well since it is designed to test whether the means of two samples with the same covariance are the same or not, and it is not as general as the MMD-based two-sample test.

5.2 Normal Distributions with Same Mean and Different Covariance

We perform the following experiment for several different problem dimensionalities. For each problem dimensionality d, we repeat the following steps 100 times:

1. We go through steps 2-4 for 10 different values of σ^2 logarithmically spaced from 1.05 to 100.
2. We sample 250 data samples from a $\mathcal{N}(0, \mathrm{I})$, where $\mathrm{I} \in \mathbb{R}^d$.
3. We sample 250 data samples from a $\mathcal{N}(0, \sigma^2 \mathrm{I})$. Note that in this case, as opposed to the previous experiment, the hypothesis testing problem becomes easier as we increase the dimensionality of the problem, because the variance spreads the same amount over more dimensions and so the power of the test might even increase as we increase the problem dimensionality (we talk more about this particular topic in section 6).
4. Perform hypothesis testing using all of the tests described in the previous section.
5. Store the number of times the null hypothesis is accepted (i.e., the number of times we get a type II error).

In figure 6 we present the results of this experiment. First of all, we see that the power of the MMD tests does not really decrease with increasing dimensionality which could be explained by the point we made in item 3 of the above list. We discuss this further in the following section. We also observe that Hotelling's T-Squared test fails completely in this

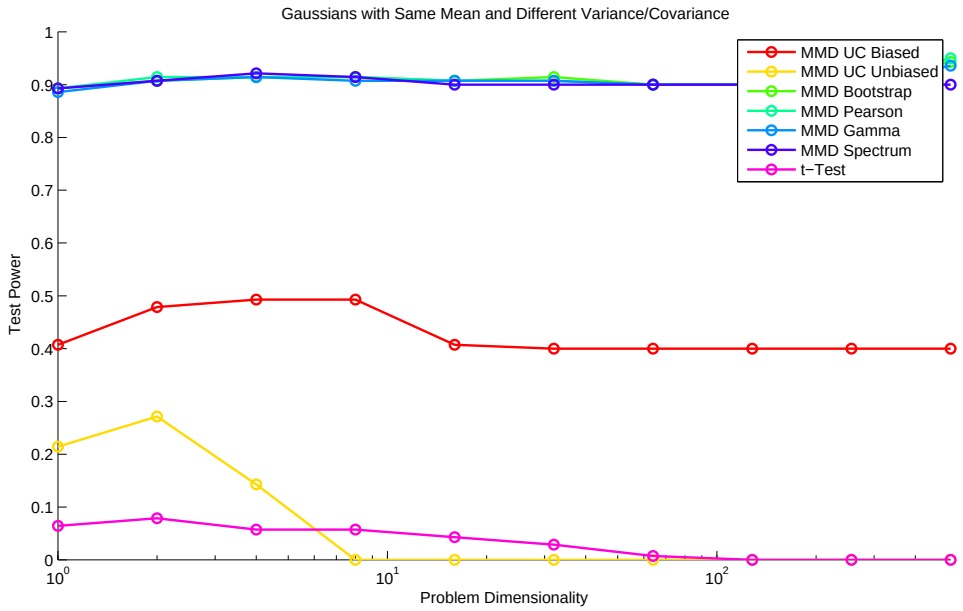

Figure 6: Our results for our two-sample test experiment when we consider Normal distributions with same mean and different covariance.

case, which is expected because it is not designed to deal with this sort of problem. The tests that are based on the uniform convergence bounds of the MMD are once again very weak compared to the other tests, which is also expected. Finally, we see that all the other MMD-based tests perform very well. It might have been more fair to compare them to the chi-squared test in this case, which is designed to deal with this problem, but unfortunately there was not enough time for us to implement and perform experiments with that test as well.

6 Discussion

One of the main points made in the previous section is that we have to be very careful on how we setup our problem as we increase its dimensionality in order to keep its difficulty constant (i.e., in order to not make it easier or harder to deal with), so that different methods become comparable. We showed a simple way in which we can do that for our first experiment. We could do something similar for the second experiment, by scaling only one of the diagonal entries of the covariance matrix by σ^2, instead of scaling all of them. We did not do that because we wanted to show how this scaling is important and how our results can become unintuitive and unreasonable (e.g., test power increasing with increasing problem dimensionality) if we are not careful enough. We are not going to go into any more detail about this problem because it deserves a long discussion. However, there is an upcoming paper at a major machine learning conference this year (2014), where the authors have proven that in order for the test to be fair, the Kullback-Leibler (KL) divergence of the two distributions from which our two samples are drawn from should remain constant as we

increase the problem dimensionality (the paper is currently under review and therefore the authors need to remain anonymous for now). Our proposed scalings for the mean and the covariance matrix have that effect for the Normal distribution, but it is more difficult to deal with other, more complicated, distributions. The authors of that paper manage to show that *the power of the MMD-based test actually drops exponentially as the dimensionality of the problem increases.*

An important point made in the discussion during the class was about the MMD defined using other functional classes besides the unit ball in an RKHS. A student, Aaditya Ramdas, pointed out that for different functional classes MMD will still be a distance metric with some norm defined in that class, but might not necessarily be computationally feasible or have close form solution. Two examples which are also provided in the paper but not discussed in this manuscript are: (i) the Kolmogorov-Smirnov test where the MMD is defined on a Banach space with supremum norm, and (ii) a test based on a metric where the function class is defined to be the space of bounded Lipschitz functions.

Finally, an interesting comment was made about the MMD uniform convergence bounds. In section 4 we saw that the uniform convergence bounds of both the biased and the unbiased MMD estimates depend on the assumption that the kernel function is bounded above. We are not entirely sure how the results change if this assumption is not satisfied.

References

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research (JMLR)*, vol. 13, pp. 723–773, Mar. 2012.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur, "A fast, consistent kernel two-sample test," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., 2009, pp. 673–681.