Kernel MMD, the Median Heuristic and Distance Correlation in High Dimensions

Sashank J. Reddi*
Machine Learning Department
Carnegie Mellon University
sjakkamr@cs.cmu.edu

Aaditya Ramdas*
Machine Learning Department
Carnegie Mellon University
aramdas@cs.cmu.edu

Barnabás Póczos
Machine Learning Department
Carnegie Mellon University
bapoczos@cs.cmu.edu

Aarti Singh
Machine Learning Department
Carnegie Mellon University
aarti@cs.cmu.edu

Larry Wasserman
Department of Statistics
Carnegie Mellon University
larry@stat.cmu.edu

June 10, 2014

Abstract

This paper is about two related methods for two sample testing and independence testing which have emerged over the last decade: Maximum Mean Discrepancy (MMD) for the former problem and Distance Correlation (dCor) for the latter. Both these methods have been suggested for high-dimensional problems, and sometimes claimed to be unaffected by increasing dimensionality of the samples. We will show theoretically and practically that the power of both methods (for different reasons) does actually decrease polynomially with dimension. We also analyze the median heuristic, which is a method for choosing tuning parameters of translation invariant kernels. We show that different bandwidth choices could result in the MMD decaying polynomially or even exponentially in dimension.

1 Introduction

Nonparametric two-sample testing and independence testing are two related problems of paramount importance in statistics. In the former, we have two sets of samples and we would like to determine if these were drawn from the same or different distributions. In the latter, we have one set of samples from a multivariate distribution, and we would like to determine if the joint is the product of marginals or not. The two problems are related because an algorithm for testing the former can be used to test the latter.

Kernel maximum mean discrepancy is a quantity introduced in Gretton et al. [2012a] to tackle the first problem using Reproducing Kernel Hilbert Spaces (RKHSs). In brief, one can embed the two probability distributions as functions in the RKHS, and calculate the squared RKHS-norm of their difference (called the MMD). For characteristic kernels like the Gaussian and Laplace kernels we will later consider, the MMD is zero iff the distributions are equal (see Gretton et al. [2012a]). The corresponding test uses empirical distributions for plug-in estimators (described later) and is consistent (for fixed dimension, power tends to one as number of samples becomes infinite) against any single fixed alternative.

Distance correlation is a quantity introduced in Székely et al. [2007] to tackle the second problem using distances between pairs of points. The population quantity is a weighted norm of difference between characteristic functions of the joint and product-of-marginal distributions, which is zero if and only if the random variables are independent. Empirically, one can calculate the matrix dot-product between the two pairwise centered distance matrices (one for each random variable) giving a consistent test against any dependent alternative.

We will explore the behavior of these related methods when the number of dimensions could be as large as, or larger than the number of samples. We will challenge existing folklore that the "performance" of both tests is unaffected by the underlying

*Both student authors had equal contribution.

dimensionality by explaining the source of both misconceptions. We demonstrate theoretically and experimentally that the *power* of both these methods actually goes down as d increases relative to n. We will also see explicit examples where the median heuristic for bandwidth selection leads to good power and when it is suboptimal.

2 Summary of Contributions

Kernel Maximum Mean Discrepancy (MMD) Gretton et al. [2012a] showed that the estimated MMD converges to the true MMD at rate $O(n^{-1/2})$ independently of dimension d. This gives the impression that the two sample test works well for large d. The result is correct but possibly misleading. We will see that the *true value* of the population MMD can be polynomially or even exponentially small in d (we were notified that a special case of Corollary 1 was earlier independently noted by Balakrishnan [2013], but do not know other examples). Also, while it is known that MMD^2 is smaller than the KL-divergence, for the first time we give several examples where it can be polynomially or exponentially smaller in d than KL. This does indicate (not imply) that the test might have low power, and we indeed experimentally demonstrate that the power against *fair* alternatives (discussed later) degrades polynomially in d.

Median Heuristic A crucial issue when using the Laplace kernel ($\exp(-\|x-x'\|_1/\gamma)$) or the Gaussian kernel ($\exp(-\|x-x'\|^2/2\gamma^2$) for MMD is the choice of the associated bandwidth γ. One of the most common heuristic choices for γ in the literature, is to choose it as the median distance between all pairs of points. This is called the median heuristic Gretton et al. [2012a]. We will show an example (separated Gaussian distributions with Gaussian kernel) where if the median heuristic is used, the population MMD will (theoretically) drop to 0 polynomially in d and if the bandwidth is of smaller order than the median distance, then it could drop exponentially. A similar conclusion holds for a second example of same mean Gaussians with different variances. In a third example, separated Laplace distributions with Laplace kernel, if the median heuristic is used, the population MMD will (theoretically) drop to 0 exponentially in d; with a larger bandwidth choice, however, the MMD drops to zero only polynomially in d. All our theoretical predictions are also validated by simulations. In the above examples, choosing the bandwidth optimally to maximize the MMD did also experimentally maximize the power against *fair* alternatives (it has been noted in Gretton et al. [2012b] that choosing the bandwidth to maximize MMD is sometimes better than the median heuristic). However, in all simulations, power goes to zero as $d \to \infty$ for all settings of the bandwidth.

Distance Correlation (dCor) Székely and Rizzo [2013] studied distance correlation in high dimensions where they considered the following example. (X, Y) are drawn from a standard normal, and even though they are independent, as dimension is increased (keeping number of samples fixed), the sample test statistic approaches one even though the true $dCor$ is zero. So even though the test statistic is consistent with n increasing and d fixed, in high dimensions its value approaches 1. They thus motivate an unbiased $dCor$ statistic ("$udCor$"), and show that for the above example, it is well behaved (centered at zero) as d increases. However, this only tells half the story - several other facts also matter for the complete picture. Specifically, the quantity that matters is the power, and the behavior of null and alternate distributions of biased and unbiased test statistics determine their power. We will empirically show that there is no difference in the polynomial decay of power of $dCor$ and $udCor$ against *fair* alternatives. We also experimentally demonstrate the different reasons that they have low power. To the best of our knowledge, there has been no prior attempt to study the power of distance correlation in the literature, in either low or high dimensions.

Due to limited space, we only provide a brief introduction to MMD and dCor, and we refer the reader to the aforementioned papers for detailed treatment. We will first get into the details of our results about MMD and the median heuristic (Sec. 3.5), returning to $dCor$ in Sec. 4.

3 The Power of MMD in High Dimensions

Let $\mathcal{P}$ be a class of continuous distributions on topological space $\mathcal{X}$. Our goal is to test

$$H_0 : p = q \quad \text{against} \quad H_1 : p \neq q$$

where $p, q \in \mathcal{P}$ We construct a test for the hypothesis from samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ from distributions p and q, respectively. To do so, one defines a divergence measure $\rho(p, q)$ such that: (a) $\rho(p, q) \geq 0$ for all $p, q \in \mathcal{F}$ and (b) $\rho(p, q) = 0$ if and only if $p = q$. We are interested in the high-dimensional regime i.e $\mathcal{X} \subseteq \mathbb{R}^d$ for large d, possibly larger than n. Most existing non-parametric methods for this problem, like KL divergence, suffer from the curse of dimensionality - if the smoothness of densities p and q doesn't grow with d, then the estimators generally require exponentially many samples in dimension to obtain a good estimate of the measure (see Birge and Massart [1995], Laurent [1996], Kerkyacharian and Picard

[1996], Bickel and Ritov [1988], Hero and Michel [1999]). However, there is some folklore that MMD does not suffer from such a curse.

Let us first introduce some known results before delving into our detailed analysis that will show that this folklore is false and that MMD's power decays with d against *fair* alternatives.

Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. The MMD is defined as:

$$\text{MMD}(\mathcal{F}, p, q) := \sup_{f \in \mathcal{F}} \mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{y\sim q}[f(y)].$$

We restrict our attention to the case where function class $\mathcal{F}$ is a unit ball in a RKHS $(\mathcal{H}, k)$ where we assume that k is a bounded, continuous and positive definite kernel function. In this case, it can be shown that $\text{MMD}(p, q) := \|\mu_p - \mu_q\|_{\mathcal{H}}$ where $\mu_p = \mathbb{E}_{x\sim p}[k(x, .)]$ for any distribution $p \in \mathcal{P}$. Furthermore, it is also well-known that $\text{MMD}(p, q) = 0$ iff $p = q$ when we use characteristic kernels Gretton et al. [2012a]. The following is a biased estimator for MMD^2 from Gretton et al. [2012a]:

$$\text{MMD}_b^2(p, q) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i, x_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y_i, y_j) - 2\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j).$$

A similar unbiased estimator exists without the $k(x_i, x_i), k(y_j, y_j)$ terms Gretton et al. [2012a]. The convergence of the above estimator to their respective measures is shown by the following result.

Theorem 1. *Gretton et al. [2012a] Suppose $0 \le k(x, x) \le K$, then with probability at least $1 - \delta$, we have*

$$|\text{MMD}_b^2(p, q) - \text{MMD}^2(p, q)| \le 2\left(\left(\frac{K}{n}\right)^{1/2} + \left(\frac{K}{m}\right)^{1/2}\right)\left(1 + \log\left(\frac{2}{\delta}\right)\right).$$

The unbiased estimator has a very similar convergence rate. Since this rate is independent of dimension, it is sometimes claimed not to suffer from any curse of dimensionality. We will show that this claim is misleading, i.e. hypothesis testing using such quantities can still suffer from the curse.

3.1 The Difficulty of Analytically Characterizing Power

The power of a test depends on the distribution of the statistic under H_0 and H_1. If the distributions are nearly Gaussian, the mean statistic and its standard deviation (s.d.) under both H_0 (μ_0, σ_0) and H_1 (μ_1, σ_1) play a role in determining power (the probability mass of the alternate distribution beyond a predetermined α in the right tail of the null distribution). Characterizing the *asymptotic* (as d is fixed and n goes to infinity) behavior of the test statistic under the null and alternative is usually hard. For example, the above MMD_b^2 estimator has a null distribution which is an infinite weighted sum of chi-squared variables Gretton et al. [2012a]. A different linear-time statistic called MMD_l^2 is unbiased and has an asymptotic normal distribution Gretton et al. [2012a]. However, the associated quantities $\mu_0, \sigma_0, \mu_1, \sigma_1$ actually vary with d and n. Further, in the high-dimensional setting, classical large sample theory does not apply as d can be comparable to or larger than n, and calculations assuming the "asymptotically" normal distribution can be misleading.

Specifically, consider $Q := \frac{\sqrt{n}(\text{MMD}_l^2 - \text{MMD}^2)}{\sigma_1} \rightsquigarrow N(0, 1) \stackrel{d}{=}: Z$ as shown in Gretton et al. [2012a]. Thus, an asymptotic level α test rejects when $\text{MMD}_l^2 > \sigma_0 z_\alpha/\sqrt{n}$. Under the alternate, the power is

$$\mathbb{P}\left(\text{MMD}_l^2 > \frac{\sigma_0 z_\alpha}{\sqrt{n}}\right) = \mathbb{P}\left(Q > \frac{\sigma_0 z_\alpha}{\sigma_1} - \frac{\sqrt{n}\text{MMD}^2}{\sigma_1}\right) \approx \mathbb{P}\left(Z > \frac{\sigma_0 z_\alpha}{\sigma_1} - \nu_n\right),$$

where $\nu_{n,d} = \sqrt{n}\text{MMD}^2/\sigma_1$ is the non-centrality parameter. The power will tend to one and the test will be consistent only if $\nu_{n,d} \to \infty$. Hence, one might be tempted to use $\nu_{n,d}$ to measure the effectiveness of the test, and indeed choosing the kernel (or bandwidth) to maximize $\nu_{n,d}$ was studied by Gretton et al. [2012b]. However, in the high dimensional setting the normal approximation used in the last step can be extremely poor, as we have experimentally verified. Development of high dimensional theory, like Barry-Esseen bounds to explicitly characterize the closeness of Q and Z, is needed.

Hence on the issue of power, we will only demonstrate carefully designed experiments showing that MMD does suffer from the curse of dimensionality against reasonable alternatives. We will give two examples where explicit calculations for the *population value* of MMD^2 are possible (not necessarily implying anything about power) and demonstrate that MMD^2 can be much smaller than the KL-divergence. These examples will also yield insights into the crucial bandwidth choice.

3.2 Relating MMD^2, TV, KL

To simplify our analysis, let us restrict ourselves to translation invariant kernels i.e. for all δ, we have $k(x+\delta, x'+\delta) = k(x, x')$. For these kernels, it is relatively easy to characterize MMD^2.

Lemma 1. *For translation invariant kernels, there exists a pdf s such that*

$$\mathrm{MMD}^2(p,q) = \int s(w)|\Phi_p(w) - \Phi_q(w)|^2 dw,$$

where Φ_p, Φ_q denote the characteristic functions of p, q respectively.

The above lemma can be proved using Bochner's theorem (Appendix). Note that

$$|\Phi_p(w) - \Phi_q(w)| \;=\; \left|\int_x \exp(iw^\top x)(p(x)-q(x))dx\right| \;\leq\; \int_x |p(x)-q(x)|dx = \mathrm{TV}(p,q).$$

From the fact that $|\Phi_p(w) - \Phi_q(w)| \leq \mathrm{TV}(p,q)$ and Pinsker's inequality, we can conclude

Lemma 2. *For translation invariant kernels,* $\mathrm{MMD}^2(p,q) \leq TV^2(p,q) \leq 2KL(p,q)$.

A more general version of the above lemma for all kernels (with a different constant than 1) is presented in Sriperumbudur et al. [2012] (Proposition 5.1). The aforementioned result gives an intuitive justification that, in general, MMD^2 is smaller than the other well known non-parametric divergence measures, whose estimators suffer from the curse of dimensionality. We will see that MMD^2 can be polynomially, and sometimes exponentially smaller than KL, and while that does not immediately imply lower power, it is an important determining factor. The proofs of the following examples are in the Appendix.

3.3 Example: Gaussian Kernel for Different Mean, Same Covariance Normal Distributions

Theorem 2. *Let $\mu_1, \mu_2 \in \mathbb{R}^d$. Suppose $p : \mathcal{N}(\mu_1, \Sigma)$ and $q : \mathcal{N}(\mu_2, \Sigma)$. Then MMD^2 between p and q using a Gaussian kernel with bandwidth γ is,*

$$\mathrm{MMD}^2(p,q) \;=\; 2\left(\frac{\gamma^2}{2}\right)^{d/2} \frac{1-\exp(-(\mu_1-\mu_2)^\top(\Sigma+\gamma^2 I/2)^{-1}(\mu_1-\mu_2)/4)}{|\Sigma+\gamma^2 I/2|^{1/2}}.$$

Suppose $\Sigma = \sigma^2 I$. Using Taylor's theorem for $1-e^{-x} \approx x$ and ignoring $-\frac{x^2}{2}$ and other smaller remainder terms for clarity, Then the above expression simplifies to

$$\mathrm{MMD}^2(p,q) \;\approx\; \frac{\|\mu_1-\mu_2\|^2}{\gamma^2(1+2\sigma^2/\gamma^2)^{d/2+1}}.$$

Keep in mind that the KL-divergence in the case of $\Sigma = \sigma^2 I$ is given by

$$KL(p,q) \;=\; \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) \;=\; \frac{\|\mu_1-\mu_2\|^2}{2\sigma^2}.$$

3.4 Example: Gaussian Kernel for Same Mean, Different Covariance Normal Distribution

The next example is for the Gaussian kernel with product Gaussian distributions having the same mean and different variances. Example 3 in Sec. 4.2 of Sriperumbudur et al. [2012] has related calculations, with a different aim that we discuss in Section 3.7.

Theorem 3. *Suppose $p : \otimes_{i=1}^{d-1}\mathcal{N}(0,\sigma^2)\otimes\mathcal{N}(0,\tau^2)$ and $q : \otimes_{i=1}^{d}\mathcal{N}(0,\sigma^2)$. Then MMD^2 between q and p using a Gaussian kernel with bandwidth γ is*

$$\mathrm{MMD}^2(p,q) \approx \frac{(\tau^2-\sigma^2)^2}{\gamma^4(1+4\sigma^2/\gamma^2)^{d/2-1/2}}.$$

By Taylor's theorem for $\log x$, the KL divergence in this case is approximately given by

$$\begin{aligned} KL(p,q) \;&=\; \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) - d - \log(\det\Sigma_0/\det\Sigma_1)) \\ &=\; \frac{1}{2}(\tau^2/\sigma^2 - 1 - \log(\tau^2/\sigma^2)) \;\approx\; \frac{(\tau^2-\sigma^2)^2}{4\sigma^4}. \end{aligned}$$

3.5 Bandwidth Choice and the Median Heuristic

We investigate how bandwidth choice affects the population MMD^2 for the example in Theorem 2 (corollaries for Theorem 3 are similar). In what follows, scaling bandwidth choices by a constant does not change the qualitative behavior, so we leave out constants for simplicity. For clarity in the following corollaries, we also ignore the Taylor residuals, and use $(1+1/d)^d \approx e$ for large d.

Underestimated bandwidth

Corollary 1. *Suppose $\Sigma = \sigma^2 I$. If we choose $\gamma = \sigma d^{1/2-\epsilon}$ for $0 < \epsilon \leq 1/2$, then*

$$\mathrm{MMD}^2(p,q) \approx \frac{\|\mu_1 - \mu_2\|^2}{\sigma^2(d^{1-2\epsilon}+2)\exp(d^{2\epsilon}/2)}.$$

Hence, the population MMD^2 goes to zero exponentially fast in d (verified by experiments that follow). The special case of a constant bandwidth with $\epsilon = 1/2$ has already been noted by Balakrishnan [2013].

The median heuristic. We approximate the choice of the median heuristic by $\gamma^2 = \mathbb{E}\|x_i - x_j\|^2$. Note that when $\Sigma = \sigma^2 I$, we have $\mathbb{E}\|x_i - x_j\|^2 \approx 2\sigma^2 d + \|\mu_1 - \mu_2\|^2$ (the first term dominates, see Sec.3.7 for explanation). Also, the experimental median (Sec.3.7) is exactly of this order.

Corollary 2. *Suppose $\Sigma = \sigma^2 I$. If we choose $\gamma = \sigma\sqrt{d}$, then*

$$\mathrm{MMD}^2(p,q) \approx \frac{\|\mu_1 - \mu_2\|^2}{\sigma^2(d+2)e}.$$

Note the population MMD^2 goes to zero polynomially as $1/d$. This is the largest MMD value one can hope for, but it is still smaller than the KL divergence by a factor of $1/d$.

Overestimating the bandwidth

Corollary 3. *Suppose $\Sigma = \sigma^2 I$. If $\gamma = \sigma d^{1/2+\epsilon}$ for $\epsilon > 0$, then*

$$\mathrm{MMD}^2(p,q) \approx \frac{\|\mu_1 - \mu_2\|^2}{\sigma^2(d^{1+2\epsilon}+2)\exp(1/2d^{2\epsilon})}.$$

Hence, the population MMD^2 goes to zero polynomially as $1/d^{1+2\epsilon}$, since $\exp(1/2d^{2\epsilon}) \approx 1$ for large d. So one pays very little for overestimating the bandwidth, compared to underestimating it.

3.6 Example : Laplace Kernel for Different Mean, Same Variance Laplace Distributions

Theorem 4 (MMD^2 Approximation)**.** *Let $\mu_1, \mu_2 \in \mathbb{R}^d$. Suppose $p : \otimes_i \mathrm{Laplace}(\mu_{1,i}, \sigma)$ and $q : \otimes_i \mathrm{Laplace}(\mu_{2,i}, \sigma)$. Using a Laplace kernel with bandwidth γ, we have*

$$\mathrm{MMD}^2(p,q) \;\approx\; \frac{\|\mu_1 - \mu_2\|^2}{2\sigma\gamma\,(1+\sigma/\gamma)^d}.$$

Note that by applying Taylor's theorem for $e^{-x} \approx 1 - x + x^2/2$, we have

$$KL(p,q) = e^{-\frac{\|\mu_1-\mu_2\|}{\sigma}} - 1 + \tfrac{\|\mu_1-\mu_2\|}{\sigma} \;\approx\; \frac{\|\mu_1 - \mu_2\|^2}{2\sigma^2}.$$

Once again, $\mathbb{E}\|x_i - x_j\|^2 \approx 2\sigma^2 d$ and indeed experimentally the median heuristic chooses $\gamma \approx \sigma\sqrt{d}$. This time, the median heuristic is suboptimal and MMD^2 drops exponentially in d. A larger bandwidth of $\gamma = \sigma d$ is optimal, making the denominator $\approx \sigma^2 de$. An overestimated bandwidth again leads to only a slow polynomial drop in MMD. In summary:

Corollary 4 (Underestimated bandwidth, median heuristic)**.** *If we choose $\gamma = \sigma d^{1-\epsilon}$ for $0 < \epsilon < 1$,*

$$\mathrm{MMD}^2(p,q) \approx \frac{\|\mu_1 - \mu_2\|^2}{2\sigma^2 d^{1-\epsilon}\exp(d^\epsilon)}.$$

Corollary 5 (Correct or Overestimated bandwidth)**.** *If we choose $\gamma = \sigma d^{1+\epsilon}$, for $\epsilon \geq 0$*

$$\mathrm{MMD}^2(p,q) \approx \frac{\|\mu_1 - \mu_2\|^2}{2\sigma^2 d^{1+\epsilon}\exp(1/d^\epsilon)}.$$

3.7 MMD^2 Power for the mean-separated Gaussian and Laplace examples

The null hypothesis is either chosen as $p = q = \mathcal{N}(0, \sigma^2 I)$ or $p = q = \otimes_{i=1}^{d} \mathrm{Laplace}(0, \sigma)$. For the alternative, p is the same and we choose $q = N(\mu, \sigma^2 I)$ or $q = \otimes_{i=1}^{d} \mathrm{Laplace}(\mu_i, \sigma)$. The choice of μ is subtle - any effect on power should not arise from the unfair choice of alternative. We choose to keep $\|\mu\|^2/\sigma^2$ constant, for example by setting $\mu = (1, 0, ...,)$ for all d. This can be justified by :

- The KL divergence between the two distributions equals (or scales like) $\|\mu\|^2/2\sigma^2$ in both cases, and hence by keeping the KL constant with d, we are not making it information theoretically harder or easier to distinguish the hypotheses as d grows.
- This quantity represents the Mahalanobis distance $\mu^T \Sigma^{-1} \mu$ which is considered as signal-to-noise-ratio, and stays constant with d.

Fig. 1 does confirm that power drops with dimension in both settings. The experiments in the Appendix of Gretton et al. [2012b] do use (alas, with no justification) our fair choice of alternatives, and they also observe decaying power with d (represented in terms of type 2 error). We contrast this with the choice of Fig. 3 in Sriperumbudur et al. [2012], which was not a power study but an empirical study of convergence rates for estimation of MMD^2, where the authors choose to let the mean separation be $(1, 1, 1, ..., 1)$ which makes the problem easier with dimension. Also, they use it to argue that the mean squared error (as summarised by Thm 1) with increasing n is indeed independent of d. Another relevant comparison is with Fig. 5A in Gretton et al. [2012a] where they show an extremely slow decrease in power with dimension for the same example of mean-shifted Gaussians with Gaussian kernel that we consider. Since the details are not in the paper, it was verified by personal communication that the bandwidth was chosen to maximize MMD, but that the means we chosen such that $\|\mu_1 - \mu_2\|$ equaled d.

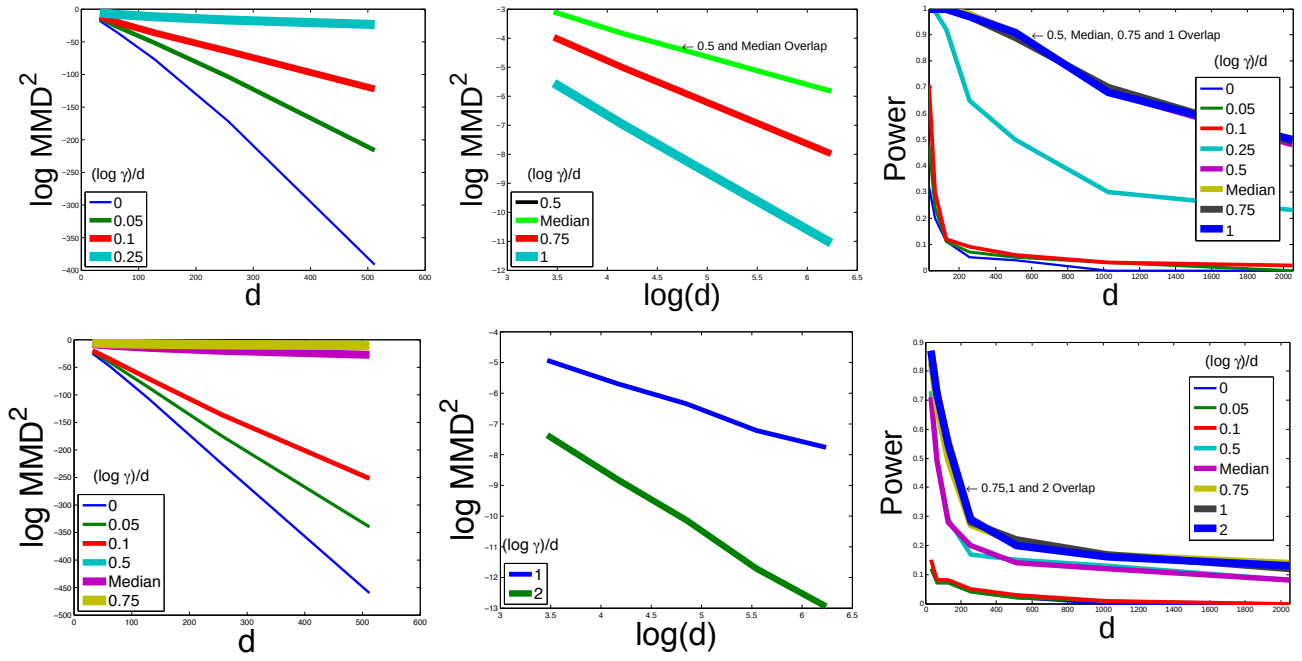

Figure 1: In all panels thicker lines represent larger bandwidth choices. The top row is for the Gaussian kernel, when the data is drawn from two Gaussians with $\sigma^2 = 1$ and constant mean separation $\|\mu_1 - \mu_2\|^2 = 1$. The bottom row is for the Laplace kernel when the data is drawn from a Laplace product distribution with $\sigma^2 = 1$ and constant mean separation $\|\mu_1 - \mu_2\|^2 = 1$. The left panels plot the population $\log \mathrm{MMD}^2$ against d (estimated from 1000 samples) for underestimated bandwidths of $1, d^{0.05}, d^{0.1}, d^{0.2}$ for Gaussian, and $d^{0.05}, d^{0.1}$,median, $d^{0.5}, d^{0.75}$ for Laplace, and show MMD^2 exponentially decreasing with d as predicted for both. The middle panels plot $\log \mathrm{MMD}^2$ against $\log d$ for optimal and overestimated bandwidths of median,$d^{0.5}, d^{0.75}, d$ for Gaussian, and d, d^2 for Laplace, and show MMD^2 decreasing polynomially in d as predicted for both. The rightmost panels plot Power vs d for all aforementioned bandwidths - the power always decreases polynomially with d (as can be verified by the logPower vs $\log d$ plot in Fig.5), but the slowest drop in power is with the near-optimal or overestimated bandwidth.

We also ran a simulation to verify that our derived expressions and approximations for MMD^2 are accurate. Fig 3 in the Appendix shows the results.

4 The Power of Distance Correlation (dCor) in high dimensions

Now we discuss nonparametric independence testing. Given n samples $(x_i, y_i) \in \mathbb{R}^{d_x+d_y}$ ($d_x \neq d_y$ is allowed) drawn from a joint distribution P_{XY} with marginals P_X, P_Y, we would like to test

$$H_0 : P_{XY} = P_X P_Y \quad \text{against} \quad H_1 : P_{XY} \neq P_X P_Y.$$

The authors of Székely et al. [2007] introduce a test statistic called (squared) distance covariance which is defined as

$$dCov_n^2(X,Y) \;=\; \frac{1}{n^2}\mathrm{tr}(\tilde{A}\tilde{B}) \;=\; \frac{1}{n^2}\sum_{i,j=1}^{n}\tilde{A}_{ij}\tilde{B}_{ij}. \tag{1}$$

Here, $\tilde{A} = HAH, \tilde{B} = HBH$ where $H = I - 11^T/n$ is a centering matrix, and A, B are distance matrices for X, Y respectively, i.e. $A_{ij} = \|x_i - x_j\|, B_{ij} = \|y_i - y_j\|$. One can use other negative definite metrics instead of Euclidean norms to generalize the definition to metric spaces Lyons [2013]. The above expression is different from the presentation in the original papers (but mathematically equivalent). They then define (squared) distance correlation as the normalized version of $dCov^2$:

$$dCor_n^2(X,Y) \;=\; \frac{dCov_n^2(X,Y)}{\sqrt{dCov_n^2(X,X)dCov_n^2(Y,Y)}}.$$

$dCor_n^2$ is always between $[0,1]$, and unlike correlation, the population $dCor^2 = 0$ iff X, Y are independent, and Székely et al. [2007] proves it is consistent against any fixed alternatives (with finite second moments).

There is an interesting connection with MMD^2 which justifies its appearance in this paper. The MMD^2 between $\mu_{P_{XY}}$ and $\mu_{P_X \times P_Y}$ is called HSIC (see Gretton et al. [2005]), which has the sample expression:

$$HSIC_n \;=\; \frac{1}{n^2}\mathrm{tr}(\tilde{K}\tilde{L}) \;=\; \frac{1}{n^2}\sum_{i,j=1}^{n}\tilde{K}_{ij}\tilde{L}_{ij}, \tag{2}$$

where $\tilde{K} = HKH, \tilde{L} = HLH$, H is defined as before and K, L are kernel matrices i.e. $K_{ij} = k(x_i, x_j), L_{ij} = l(y_i, y_j)$. The striking similarity between Eqs.(1) and (2) is not coincidental - Sejdinovic et al. [2013] recently showed for every negative definite metric, there exists a positive definite kernel, and for every positive definite kernel, there exists a negative definite metric, such that $HSIC$ equals $dCov^2$. Hence, $dCor^2$ and MMD are very strongly related quantities, for very related problems.

In Székely and Rizzo [2013], the authors advocate the use of a modified unbiased $dCor$ (called $udCor$), claiming that it works well for large d. We will describe experiments that demonstrate that $dCor^2$ and $udCor^2$ as test statistics both suffer in high d, but for a slightly different reason than for MMD^2.

When (X, Y) are drawn from a standard Gaussian, the authors of Székely and Rizzo [2013] show that the biased $dCor_n \to 1$, if n is kept fixed and d_x, d_y are increased. Then, they show that their unbiased $udCor_n$ hovers around 0 in the same situation (even when $d_x, d_y \gg n$), and conclude that it performs well in high-dimensions. The bias is indeed zero, but we argue that the variance of $udCor_n$ remains the same order as $dCor_n$. Székely and Rizzo [2013] show that when the null is true, $udCor_n$ is well behaved. However, the right followup question to ask is - when the alternate hypothesis is true, how does the statistic behave in high dimensions? Below we demonstrate that in this case it fails to detect such dependence in high dimensions, i.e. its power goes to zero.

4.1 Experimental Verification

Here we carefully design a simple simulation experiment to demonstrate this decrease in power with dimension, with some subtleties in choice of the alternative hypothesis that are quite crucial. For the null hypothesis of independent variables, we let (x, y) be sampled from a d-dimensional standard normal, like in Székely and Rizzo [2013]. For the alternative hypothesis, we need to make a choice about how to change the covariance matrix, such that as d increases, the problem neither gets easier nor harder. We choose to make a *constant* number of off-diagonal elements non-zero, i.e. we don't change the number or value of non-zero off-diagonal elements as d increases. The marginals are still standard Gaussians; the non-zero elements are only in the cross-diagonal blocks indicating dependence between X,Y.

One can argue that a constant number of non-zeros (not growing with d) is the fairest choice, which does not increase/decrease (with d) the amount of information provided to the statistician:

1. All the information to decide between null and alternate is captured in the covariance matrix. From classical information theory, the Gaussian entropy $\log\det\Sigma$ is the amount of information encoded in Σ, which (with our choice) remains constant as d increases.

2. Another information theoretic quantity of relevance is the mutual information (MI) between X, Y. Since the MI between Gaussians is given by $\log \frac{\det \Sigma}{\det \Sigma_X \det \Sigma_Y}$, one can easily check that the mutual information between X, Y stays constant as dimension increases.

3. All one is trying to do is differentiate Σ from I, so $\|\Sigma - I\|_F^2$ is also a measure of difficulty of the problem. Even if the statistician detects dependence caused by a single off-diagonal element, he will reject the null. With our choice, $\|\Sigma - I\|_F^2$ does stay constant with d.

This is similar in spirit to the MMD^2 case, where we justified our alternative by verifying that the signal to noise ratio and the KL-divergence weren't changing with d. Figure 2 provides a detailed analysis of the power of $dCor, udCor$. P_0, P_1 represent the distributions of the corresponding test statistic (possibly not normally distributed) under H_0 and H_1 and $\mu_0, \sigma_0, \mu_1, \sigma_1$ are their mean and standard deviation. The explanations in the figure subtext bring out the complicated scenario of $\mu_0, \mu_1, \sigma_0, \sigma_1$ all changing with d - i.e. P_0, P_1 change with d.

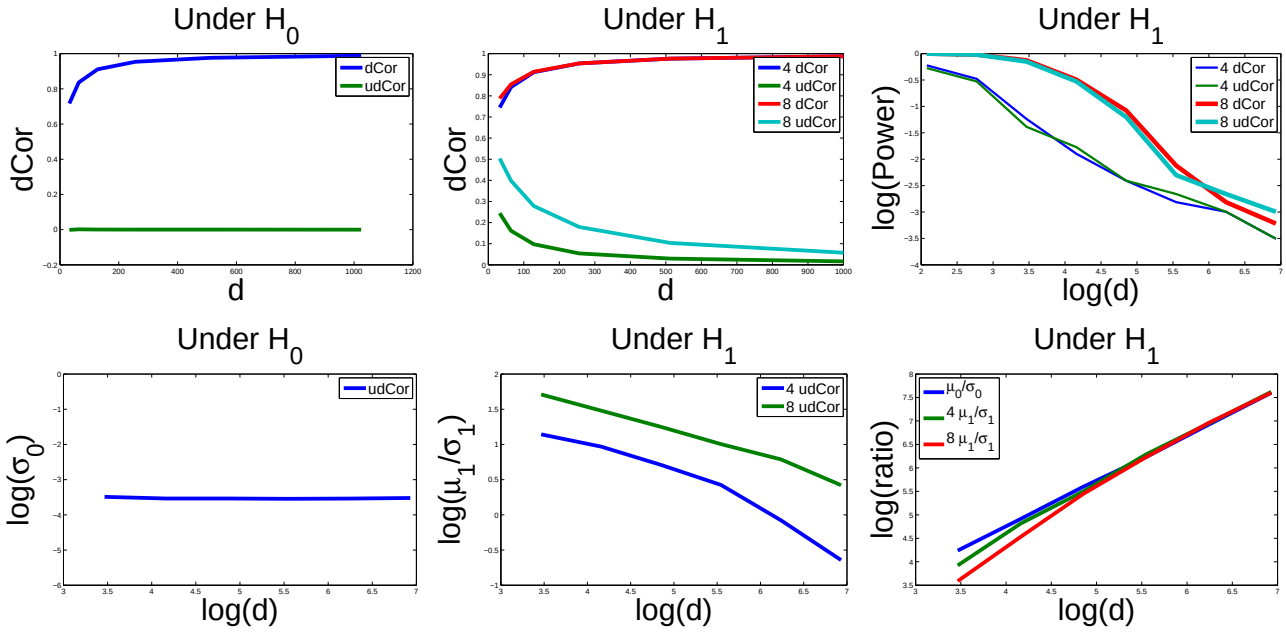

Figure 2: All plots were produced by averaging quantities over 1000 repetitions. The choice of H_1 is as described in the text, with 4 and 8 nonzeros. TOP ROW. The left panel shows that $dCor_n \to 1, udCor_n \approx 0$, as predicted by Székely and Rizzo [2013]. The middle panel shows that $dCor_n \to 1, udCor_n \to 0$, similar to the null. The right panel demonstrates that the power of $dCor, udCor$ decreases polynomially in d for the same alternatives and the unbiasedness of $udCor$ seen in the left panel doesn't improve its power. BOTTOM ROW. The behavior of $udCor$ (first two panels) and $dCor$ (last panel) are different, these panels investigate the *reasons*. The left panel shows that the σ_0 does not change with d, and along with the panel above it shows that $udCor$'s P_0 is unchanging with d and centered at 0. The second panel however *suggests* that $udCor$'s P_1 approaches the P_0 rapidly as a polynomial in d, i.e. even though σ_1 is shrinking with d (Appendix Fig.6), μ_1 is shrinking even faster than σ_1, leading to larger overlap between P_0, P_1 and the observed decrease in power. The last panel shows a different story for $dCor$ (ratio is either μ_0/σ_0 or μ_1/σ_1); Since the means μ_0, μ_1 slowly converge to 1 as d increases, the σ_0, σ_1 must decrease with d (Appendix Fig.6), so that their ratio continues to increase. However, since the ratios are converging to each other as d increases, P_0, P_1 approach each other faster than their standard deviations drop explaining the observed decrease in power.

5 Conclusion

In summary, we believe that we have made a strong case for the first time that the power of the closely related kernel and distance based tests both suffer from the curse of dimensionality against *fair* alternatives. In the process, we also undertook a detailed study of bandwidth choices and explicitly demonstrated cases when and why the median heuristic works and fails, and made a case for overestimating the bandwidth. The reasons for the observed power decay can be complicated, and a better theory is necessary to understand the null and alternate distributions in high dimensions.

References

S. Balakrishnan. *Finding and Leveraging Structure in Learning Problems*. PhD thesis, Carnegie Mellon University, 2013.

P. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

L. Birge and P. Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, pages 11–29, 1995.

D.H. Fremlin. *Measure Theory*. Number v. 2 in Measure theory. Torres Fremlin, 2000. ISBN 9780953812905.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

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

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

A. Hero and O. Michel. Estimation of Rényi information divergence via pruned minimal spanning trees. In *Higher-Order Statistics, 1999. Proceedings of the IEEE Signal Processing Workshop on*, pages 264–268. IEEE, 1999.

G. Kerkyacharian and D. Picard. Estimating nonquadratic functionals of a density using haar wavelets. *The Annals of Statistics*, 24(2):485–507, 1996.

B. Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

R. Lyons. Distance covariance in metric spaces. *Annals of Probability*, 41(5):3284–3305, 2013.

W. Rudin. *Fourier analysis on groups*. Interscience Publishers, New York, 1962.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, et al. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

B. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schoelkopf, and G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

G.J. Székely and M.L. Rizzo. The distance correlation t-test of independence in high dimension. *J. Multivariate Analysis*, 117:193–213, 2013.

G.J. Székely, M.L. Rizzo, and Bakirov N.K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Appendix

A Proof of Lemma 1

Proof. From definition of MMD^2, we have

$$\mathrm{MMD}^2(p,q) = \int_{x,x'} k(x,x')p(x)p(x')dxdx' + \int_{x,x'} k(x,x')q(x)q(x')dxdx' - 2\int_{x,x'} k(x,x')p(x)q(x')dxdx'.$$

From Bochner's theorem (see [Rudin, 1962]) for translation invariant kernels, we know $k(x,x') = \int_w s(w)e^{iw^\top x}e^{-iw^\top x'}dw$ where s is the fourier transform of the kernel. Substituting the above equality in the definition of MMD^2, we have the required result. □

B Proof of Theorem 2

Proof. Since Gaussian kernel is a translation invariant kernel, we can use Lemma 1 to derive the MMD^2 in this case. It is well-known that the Fourier transform $s(w)$ of Gaussian kernel is Gaussian distribution. Substituting the characteristic function of normal distribution in Lemma 1, we have

$$\begin{aligned}
\mathrm{MMD}^2(p,q) &= \int_w \left(\gamma^2/2\pi\right)^{d/2} \exp\left(-\gamma^2\|w\|^2/2\right) \left|\exp(i\mu_1^\top w - w^\top\Sigma w/2) - \exp(i\mu_1^\top w - w^\top\Sigma w/2)\right|^2 dw \\
&= \left(\gamma^2/2\pi\right)^{d/2} \int_w \exp\left(-w^\top\Sigma w\right) \exp\left(-\gamma^2\|w\|^2/2\right) \left|\exp(i\mu_1^\top w) - \exp(i\mu_2^\top w)\right|^2 dw \\
&= \left(\gamma^2/2\pi\right)^{d/2} \int_w \exp\left(-w^\top(\Sigma+\gamma^2 I/2)w\right) \left(2 - \exp\left(-i(\mu_1-\mu_2)^\top w\right) - \exp\left(-i(\mu_2-\mu_1)^\top w\right)\right) dw \\
&= 2\left(\gamma^2/2\pi\right)^{d/2} \int_w \exp\left(-w^\top(\Sigma+\gamma^2 I/2)w\right) \left(1 - \exp\left(-i(\mu_1-\mu_2)^\top w\right)\right) dw \qquad (3)
\end{aligned}$$

The third step follows from definition of complex conjugate. In what follows, we do the following change of variable $u = (\Sigma+\gamma^2 I/2)^{1/2}w$. Consider the following term:

$$\begin{aligned}
&\int_w \exp\left(-w^\top(\Sigma+\gamma^2 I/2)w\right) \exp\left(-i(\mu_1-\mu_2)^\top w\right) dw \\
&\quad= \int_u \exp - \left(u^\top u + i(\mu_1-\mu_2)^\top(\Sigma+\gamma^2 I/2)^{-1/2}u\right) |\Sigma+\gamma^2 I/2|^{-1/2} du \\
&\quad= |\Sigma+\gamma^2 I/2|^{-1/2} \exp(-(\mu_1-\mu_2)^\top(\Sigma+\gamma^2 I/2)^{-1}(\mu_1-\mu_2)/4)\times \\
&\qquad\qquad \int_u \exp - \left(\|u - i(\Sigma+\gamma^2 I/2)^{-1/2}(\mu_1-\mu_2)/2\|^2\right) du \\
&\quad= \pi^{d/2}|\Sigma+\gamma^2 I/2|^{-1/2} \exp(-(\mu_1-\mu_2)^\top(\Sigma+\gamma^2 I/2)^{-1}(\mu_1-\mu_2)/4)
\end{aligned}$$

The second step follows from well-known theory of change of variables (see Theorem 263D of Fremlin [2000]). By substituting the above equality in Equation 3, we get the required result. □

C Proof of Proposition 1

Proposition 1. *Suppose $\lambda \neq \sigma$, then we have,*

$$\int_{-\infty}^{\infty} \exp\left(-\frac{|x-\lambda|}{\gamma}\right)\exp\left(-\frac{|x|}{\sigma}\right)dx = \frac{e^{-|\lambda|/\sigma}}{1/\gamma+1/\sigma}+\frac{e^{-|\lambda|/\gamma}}{1/\sigma-1/\gamma}-\frac{e^{-|\lambda|/\sigma}}{1/\sigma-1/\gamma}+\frac{e^{-|\lambda|/\gamma}}{1/\gamma+1/\sigma}$$

and when $\lambda = \sigma$, we have,

$$\int_{-\infty}^{\infty} \exp\left(-\frac{|x-\lambda|}{\sigma}\right)\exp\left(-\frac{|x|}{\sigma}\right)dx = \frac{e^{-|\lambda|/\sigma}}{1/\gamma+1/\sigma}+|\lambda|e^{-|\lambda|/\sigma}+\frac{e^{-|\lambda|/\gamma}}{1/\gamma+1/\sigma}$$

Proof. We show this when $\lambda \leq 0$ as an example proof:

$$\begin{aligned}\int_{-\infty}^{\infty} \exp\left(-\frac{|x-\lambda|}{\gamma}\right)\exp\left(-\frac{|x|}{\sigma}\right)dx &= \int_{-\infty}^{\lambda} \exp\left(\frac{x-\lambda}{\gamma}\right)\exp\left(\frac{x}{\sigma}\right)dx+\int_{\lambda}^{0} \exp\left(\frac{\lambda-x}{\gamma}\right)\exp\left(\frac{x}{\sigma}\right)dx \\ &\quad +\int_{0}^{\infty} \exp\left(\frac{\lambda-x}{\gamma}\right)\exp\left(-\frac{x}{\sigma}\right)dx \\ &= \frac{e^{-\lambda/\gamma}e^{\lambda/\sigma+\lambda/\gamma}}{1/\gamma+1/\sigma}+\frac{e^{-\lambda/\gamma}(1-e^{-\lambda/\gamma+\lambda/\sigma})}{1/\sigma-1/\gamma}+\frac{e^{\lambda/\gamma}}{1/\gamma+1/\sigma}\end{aligned}$$

Also, when $\gamma = \sigma$, we obtain the same expression for the first and last terms. However, the middle term has the following constant integrand, thereby, leading to the required expression.

$$\int_{\lambda}^{0} \exp\left(\frac{\lambda-x}{\gamma}\right)\exp\left(\frac{x}{\sigma}\right)dx = |\lambda|e^{-|\lambda|/\sigma}.$$

□

D Proof of Proposition 2

Proposition 2. *Let $\psi = \sigma/\gamma$. Then we have,*

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{|x-x'|}{\gamma}\right)\frac{1}{4\sigma^2}\exp\left(-\frac{|x-\mu|}{\sigma}\right)\exp\left(-\frac{|x'|}{\sigma}\right)dxdx' \\ =\ & -\frac{1}{2}e^{-|\mu|/\sigma}\left(\frac{\psi+|\mu|/\gamma}{1-\psi^2}\right)+\frac{1}{1-\psi^2}\left(-\frac{\psi e^{-|\mu|/\sigma}}{1-\psi^2}+\frac{e^{-|\mu|/\gamma}}{1-\psi^2}\right) \\ =\ & -\frac{\mu^2}{4\sigma\gamma(1+\psi)^2}+\frac{2+\psi}{2(1+\psi)^2}+O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right)\end{aligned}$$

Proof. We first integrate with respect to x' using the Proposition 1 to get

$$\frac{1}{4\sigma^2}\int_{-\infty}^{\infty}\left(\frac{e^{-|x|/\sigma}}{1/\gamma+1/\sigma}+\frac{e^{-|x|/\gamma}}{1/\sigma-1/\gamma}-\frac{e^{-|x|/\sigma}}{1/\sigma-1/\gamma}+\frac{e^{-|x|/\gamma}}{1/\gamma+1/\sigma}\right)\exp\left(-\frac{|x-\mu|}{\sigma}\right)dx$$

We then integrate these terms once again using both parts of Proposition 1 to get the first equality. We simplify the second

equation in the following manner:

$$
\begin{aligned}
& -\frac{1}{2}e^{-|\mu|/\sigma}\left(\frac{\psi+|\mu|/\gamma}{1-\psi^2}\right)+\frac{1}{1-\psi^2}\left(-\frac{\psi e^{-|\mu|/\sigma}}{1-\psi^2}+\frac{e^{-|\mu|/\gamma}}{1-\psi^2}\right) \\
= & -\frac{1}{2}\left(1-\frac{|\mu|}{\sigma}+\frac{|\mu|^2}{2\sigma^2}\right)\left(\frac{\psi+|\mu|/\gamma}{1-\psi^2}\right)+\frac{1}{1-\psi^2}\left(-\frac{(\sigma/\gamma-|\mu|/\gamma+\mu^2/2\sigma\gamma)}{1-\psi^2}+\frac{1-|\mu|/\gamma+\mu^2/2\gamma^2}{1-\psi^2}\right) \\
& \quad +O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right) \\
= & -\frac{1}{2(1-\psi^2)}\left(\psi-\frac{\mu^2}{2\sigma\gamma}+\frac{|\mu|^3}{2\sigma^2\gamma}\right)+\frac{1}{(1-\psi^2)^2}\left(1-\psi-\frac{\mu^2}{2\sigma\gamma}+\frac{\mu^2}{2\gamma^2}\right) \\
& \quad +O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right) \\
= & -\frac{1}{2(1-\psi^2)}\left(\psi-\frac{\mu^2}{2\sigma\gamma}\right)+\frac{(1-\mu^2/2\sigma\gamma)(1-\psi)}{(1-\psi^2)^2}+O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right) \\
= & \frac{1}{1-\psi^2}\left(-\frac{\psi}{2}+\frac{1}{2}\frac{\mu^2}{2\sigma\gamma}\right)+\frac{1}{1-\psi^2}\left(\frac{1}{1+\psi}-\frac{\mu^2}{(1+\psi)2\sigma\gamma}\right)+O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right) \\
= & -\frac{\mu^2}{4\sigma\gamma(1+\psi)^2}+\frac{2+\psi}{2(1+\psi)^2}+O\left(\frac{|\mu|^3}{\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu|^3}{\gamma^3(1-\psi^2)^2}\right)
\end{aligned}
$$

□

E Proof of Theorem 4

Proof. Recall that we use Laplace kernel, i.e., $k(x,x')=\exp(-\|x-x'\|_1/\gamma)$. By using the definition of MMD^2, we have

$$\mathrm{MMD}^2=\int_{x,x'}(p(x)p(x')+q(x)q(x')-2p(x)q(x'))k(x,x')dxdx'. \tag{4}$$

Consider the term $\int_{x,x'}p(x)q(x')k(x,x')dxdx'$. The other terms can be calculated in a similar manner. Let $\psi=\sigma/\gamma$ and $\beta=(1+\psi/2)/(1+\psi)^2$. We have,

$$
\begin{aligned}
\int_{x,x'}p(x)q(x')k(x,x')dxdx' &= \prod_{i=1}^{d}\int_{x_i,x_i'}\exp\left(-\frac{|x-x'|}{\gamma}\right)\frac{1}{4\sigma^2}\exp\left(-\frac{|x-\mu|}{\sigma}\right)\exp\left(-\frac{|x'|}{\sigma}\right)dx_idx)i' \\
&= \prod_{i=1}^{d}\beta\left(1-\frac{\mu_i^2}{4\beta\sigma\gamma(1+\psi)^2}+O\left(\frac{|\mu_i|^3}{\beta\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu_i|^3}{\beta\gamma^3(1-\psi^2)^2}\right)\right) \\
&= \beta^d\left(1-\frac{\|\mu\|^2}{4\beta\sigma\gamma(1+\psi)}+O\left(\frac{|\mu_i|^3}{\beta\sigma^2\gamma(1-\psi^2)^2}\right)-O\left(\frac{|\mu_i|^3}{\beta\gamma^3(1-\psi^2)^2}\right)\right)
\end{aligned}
$$

The first step follows from the fact that both Laplace kernel and Laplace distribution decompose over the coordinates. The second step follows from Proposition 2. Substituting the above expression in Equation 4, we get,

$$\mathrm{MMD}^2=\frac{\beta^{d-1}\|\mu\|^2}{2\sigma\gamma(1+\psi)}-O\left(\frac{\beta^{d-1}\|\mu\|_3^3}{\sigma^2\gamma(1-\psi^2)^2}\right)+O\left(\frac{\beta^{d-1}\|\mu\|_3^3}{\gamma^3(1-\psi^2)^2}\right).$$

□

F Verifying MMD approximations

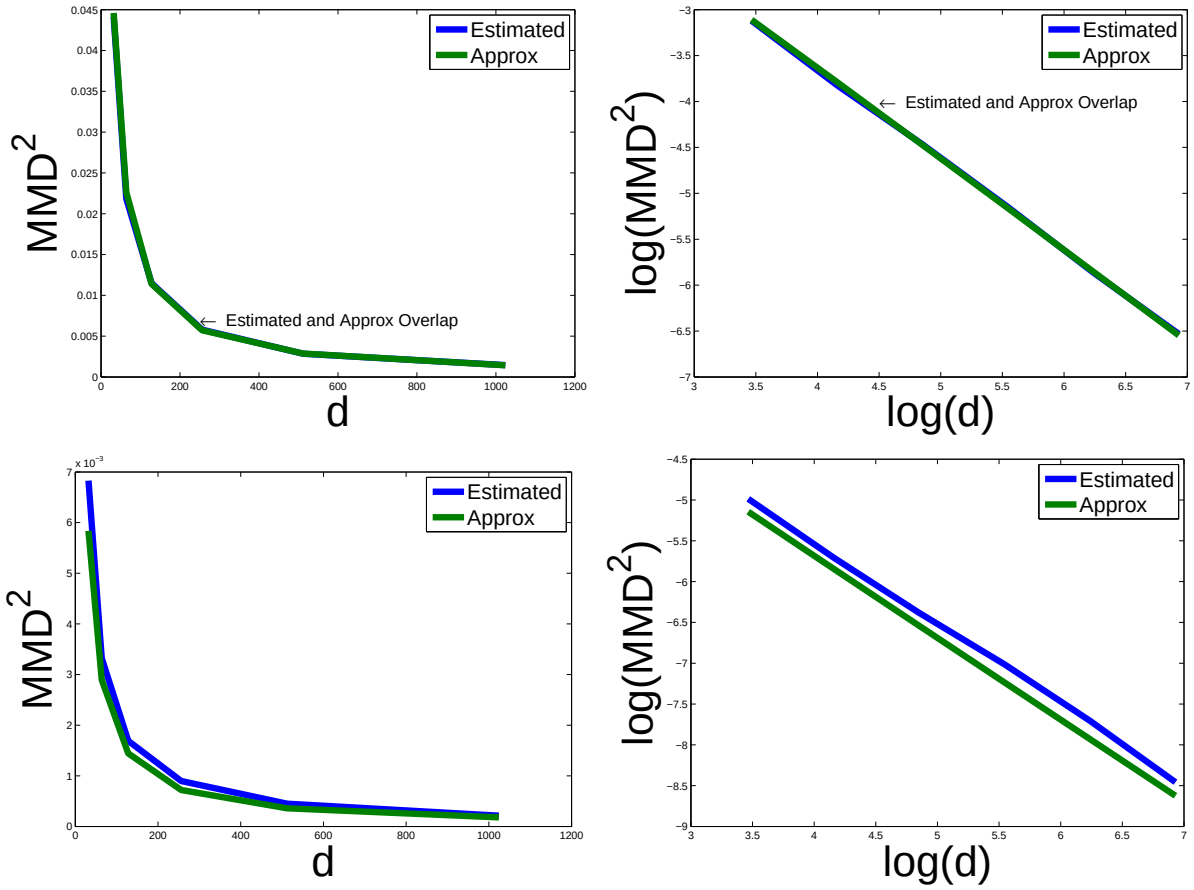

Figure 3: Top left: MMD vs d, for Gaussian kernel with optimal $\sigma\sqrt{d}$ bandwidth, as estimated from data and approximated by formula. Top right: same but for Log(MMD). Bottom left: MMD vs d, for Laplace kernel with optimal σd bandwidth, estimated from data and approximated by formula. Bottom right: same but for Log(MMD). The Log Plots also show the right scaling that MMD decays as $1/d$ with the right choice of bandwidth, with a constant mean separation of $\mu = (1, 0, 0, ..., 0)$.

G Biased MMD for Gaussian Distribution

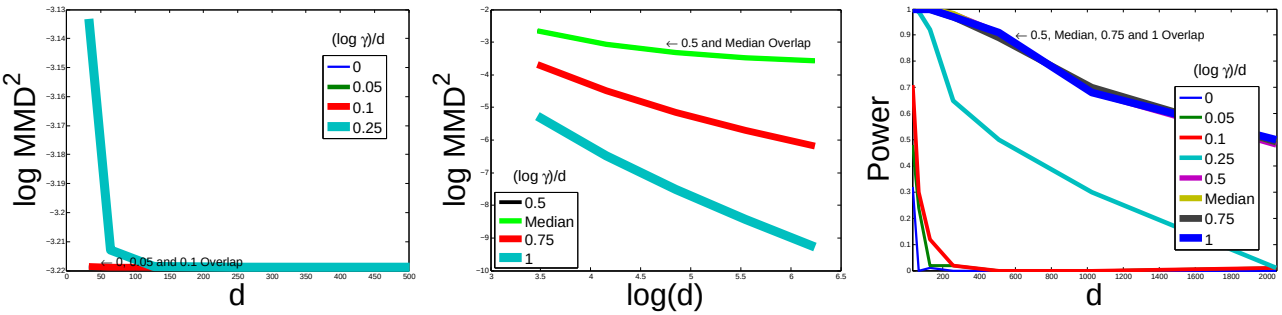

Figure 4: Plots for Biased MMD with Gaussian kernel, when the data is drawn from two Gaussians with $\sigma^2 = 1$ and constant mean separation $\|\mu_1 - \mu_2\|^2 = 1$. With respect to the selection of bandwidth γ, the power of Biased MMD has similar behavior as Unbiased MMD.

H Verifying Power Plots decay polynomially

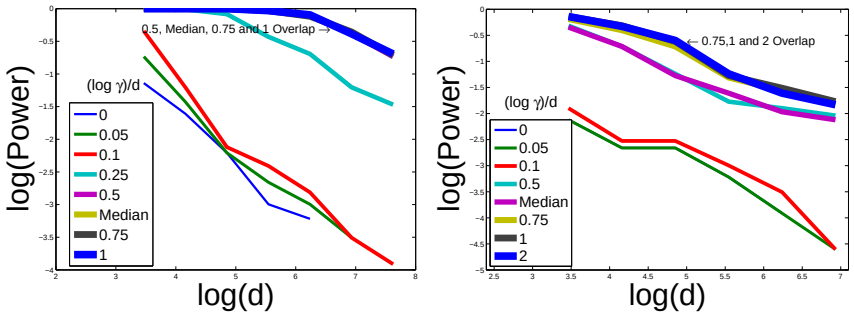

Figure 5: The aim of this figure is to show that the seemingly exponential drop-off in Fig 1 is actually a polynomial decay of power with d.

I Standard deviation of $udCor, dCor$

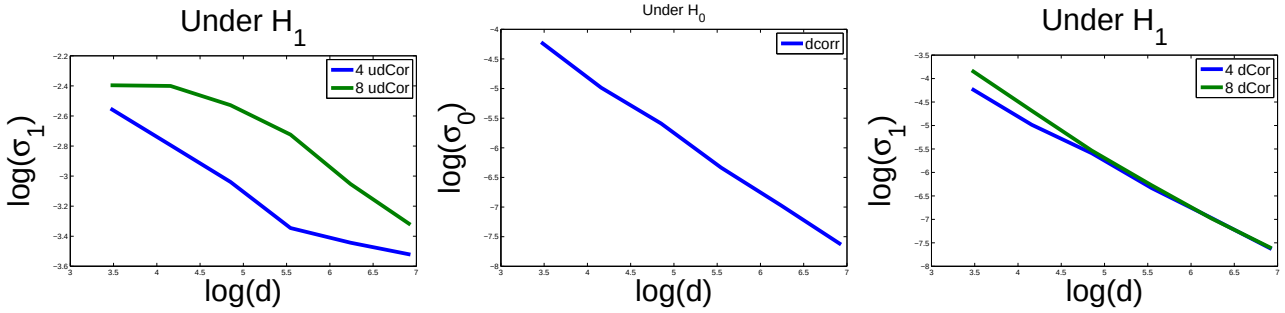

Figure 6: First plot: The standard deviation of $udCor$ under alternative H_1 decreases with dimension. Next two plots: The standard deviation of $dCor$ under H_0 and H_1 decreases with dimension.

J MMD between Gaussians with same mean, different variances

Suppose $P = \otimes_{i=1}^{d} N(0, \sigma^2) \otimes N(0, a^2)$ and $Q = \otimes_{i=1}^{d} N(0, \sigma^2) \otimes N(0, b^2)$. If a, b are of the same order as σ then the median heuristic will still pick $\gamma \approx \sigma\sqrt{d}$ for bandwidth γ of the Gaussian kernel. First we note that for distributions with the same mean, by Taylor's theorem,

$$\begin{aligned} KL(P, Q) &= \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0 - d - \log(\det \Sigma_0)/\det \Sigma_1)) = \frac{1}{2}(a^2/b^2 - 1 - \log(a^2/b^2)) \\ &\approx \frac{(a^2/b^2 - 1)^2}{4} \end{aligned}$$

The MMD^2 can be derived (approximated using $(1+x)^n \approx 1 + nx$ for small x) as

$$\begin{aligned} & \frac{1}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} \left(\frac{1}{\sqrt{1 + 4a^2/\gamma^2}} + \frac{1}{\sqrt{1 + 4b^2/\gamma^2}} - \frac{2}{\sqrt{1 + 2(a^2 + b^2)/\gamma^2}} \right) \\ \approx & \frac{1}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} \left(\frac{1}{1 + 2a^2/\gamma^2} + \frac{1}{1 + 2b^2/\gamma^2} - \frac{2}{1 + (a^2 + b^2)/\gamma^2} \right) \\ \approx & \frac{1}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} \left(\frac{1}{\sqrt{1 + 2a^2/\gamma^2}} - \frac{1}{\sqrt{1 + 2b^2/\gamma^2}} \right)^2 \\ \approx & \frac{1}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} \left((1 - a^2/\gamma^2) - (1 - b^2/\gamma^2) \right)^2 \\ = & \frac{b^4/\gamma^4}{(1 + 4\sigma^2/\gamma^2)^{d/2-1/2}} (a^2/b^2 - 1)^2 \end{aligned}$$

If γ is chosen by the median heuristic (optimal in this case), we see that this is smaller than KL by $\sigma^4 d^2 e / b^4$. If it is chosen as constant, it can be exponentially smaller than KL.

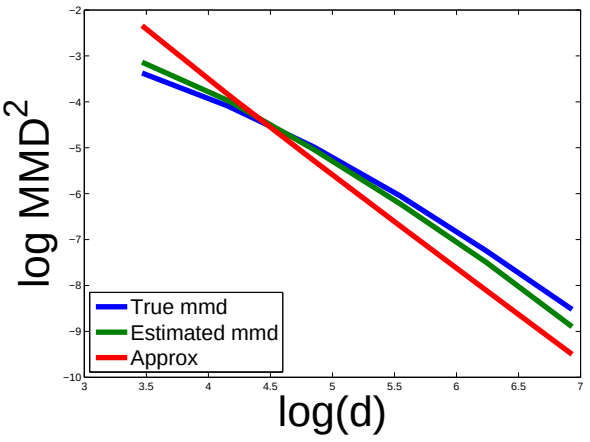

Figure 7: This is to verify the above approximation. The straight line is our final approximation in the theorem. The other two are the true MMD by formula, and the MMD from data. This shows that they are all very similar, grow at the same rate, and the approximation is useful.