Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

On the Decreasing Power of Kernel and Distance Based Nonparametric Hypothesis Tests in High Dimensions

Aaditya Ramdas,[1,2,*] **Sashank J. Reddi,**[1,*] **Barnabás Póczos,**[1] **Aarti Singh,**[1] **Larry Wasserman**[2]
Machine Learning Department[1] and Department of Statistics[2]
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA - 15213, USA

Abstract

This paper is about two related decision theoretic problems, nonparametric two-sample testing and independence testing. There is a belief that two recently proposed solutions, based on kernels and distances between pairs of points, behave well in high-dimensional settings. We identify different sources of misconception that give rise to the above belief. Specifically, we differentiate the hardness of estimation of test statistics from the hardness of testing whether these statistics are zero or not, and explicitly discuss a notion of "fair" alternative hypotheses for these problems as dimension increases. We then demonstrate that the power of these tests actually drops polynomially with increasing dimension against fair alternatives. We end with some theoretical insights and shed light on the *median heuristic* for kernel bandwidth selection. Our work advances the current understanding of the power of modern nonparametric hypothesis tests in high dimensions.

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 distribution 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.

More formally, the problem of two-sample or homogeneity testing can be described as follows. Given m samples $x_1, ..., x_m$ drawn from a distribution P supported on $\mathcal{X} \subseteq \mathbb{R}^d$ and n samples $y_1, ..., y_n$ drawn from a distribution Q supported on $\mathcal{Y} \subseteq \mathbb{R}^d$, we would like to tell which of the following hypotheses is true:

$$H_0 : P = Q \text{ vs. } H_1 : P \neq Q$$

Similarly, the problem of independence testing can be described as follows. Given n samples (x_i, y_i) for $i \in \{1, ..., n\}$ where $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^q$, that are drawn from a joint distribution P_{XY} supported on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{p+q}$, we would to tell which of the following hypotheses is true:

$$H_0 : P_{XY} = P_X \times P_Y \text{ vs. } H_1 : P_{XY} \neq P_X \times P_Y$$

where P_X, P_Y are the marginals of P_{XY} w.r.t. X, Y.

In both cases, H_0 is called the *null hypothesis* and H_1 is called the *alternate hypothesis*. Both problems are considered in the *nonparametric* setting, in the sense that no parametric assumptions are made about any of the aforementioned distributions.

A recent class of popular approaches for this problem (and a related two-sample testing problem) involve the use of test statistics based on quantities defined in reproducing kernel Hilbert spaces (RKHSs) (Gretton et al. 2012a; Eric, Bach, and Harchaoui 2008; Gretton et al. 2005; Fukumizu et al. 2008) that are computed using kernels evaluated on pairs of points. A related set of approaches were developed in parallel based on pairwise distances between points, as exemplified for independence testing by distance correlation, introduced in (Székely, Rizzo, and Bakirov 2007) and further discussed or extended in (Székely and Rizzo 2009; Lyons 2013; Székely and Rizzo 2013; Sejdinovic et al. 2013). We summarize these in the next subsection.

This paper is about existing folklore that these methods "work well" in high-dimensions. We will identify and address the different sources of misconception which lead to this faulty belief. One of the main misconceptions is that while it is true for the normal means problem, estimating the mean of Gaussian is harder than deciding whether the mean is non-zero or not, this is not true in general. Indeed, the test statistics that we will deal with have the opposite behavior - they have low estimation error that is independent of dimension, but the decision problem of whether they are nonzero or not gets harder in higher dimensions, causing the tests to have low power. Indeed, we will demonstrate that against a class of "fair" alternatives, the power of both sets of approaches degrades with dimension for both types of problems (two-sample or independence testing).

The takeaway message of this paper is - kernel and distance based hypothesis tests *do* suffer from decaying power in high dimensions (even though the current literature is often misinterpreted to claim the opposite). We provide some mathematical reasoning accompanied by solid intuitions as to why this should be the case. However, settling the issue completely and formally is important future work.

[*] Both authors had equal contribution.

Two-Sample Testing using kernels

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive-definite kernel corresponding to RKHS H_k with inner-product $\langle ., . \rangle_k$ - see (Schölkopf and Smola 2002) for an introduction. Let k correspond to feature maps at x denoted by $\phi_x \in H_k$ respectively satisfying $\phi_x(x') = \langle \phi_x, \phi'_x \rangle_k = k(x, x')$. The mean embedding of P is defined as $\mu_P := \mathbb{E}_{x \sim P} \phi_x$ whose empirical estimate is $\hat{\mu}_P := \frac{1}{n} \sum_{i=1}^{n} \phi_{x_i}$. Then, the Maximum Mean Discrepancy (MMD) is defined as

$$\text{MMD}^2(P, Q) := \|\mu_P - \mu_Q\|_k^2$$

where $\|.\|_k$ is the norm induced by $\langle ,., \rangle_k$, i.e. $\|f\|_k^2 = \langle f, f, \rangle_k$ for every $f \in H_k$. The corresponding empirical test statistic is defined as

$$\text{MMD}_b^2(P, Q) := \|\hat{\mu}_P - \hat{\mu}_Q\|_k^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j). \quad (1)$$

The subscript b indicates that it is a biased estimator of MMD^2. The unbiased estimator is calculated by excluding the $k(x_i, x_i), k(y_i, y_i)$ terms from the above sample expression, let us call that MMD_u^2. It is important to note that every statement/experiment in this paper about the power of MMD_b^2 qualitatively holds true for MMD_u^2 also.

Independence Testing using distances

The authors of (Székely, Rizzo, and Bakirov 2007) introduce an empirical test statistic called (squared) distance covariance which is defined as

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

where, $\tilde{A} = HAH, \tilde{B} = HBH$ where $H = I - 11^T/n$ is a centering matrix, and A, B are distance matrices for X, Y respectively, i.e. $A_{ij} = \|x_i - x_j\|, B_{ij} = \|y_i - y_j\|$. The subscript n suggests that it is an empirical quantity based on n samples. The corresponding population quantity turns out to be a weighted norm of the difference between characteristic functions of the joint and product-of-marginal distributions, see (Székely, Rizzo, and Bakirov 2007).

The expression in Equation 2 is different from the presentation in the original papers (but mathematically equivalent). They then define (squared) distance correlation $dCor_n^2$ as the normalized version of $dCov_n^2$:

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

One can use other distance metrics instead of Euclidean norms to generalize the definition to metric spaces, see (Lyons 2013). As before, the above expressions don't yield unbiased estimates of the population quantities, and (Székely and Rizzo 2013) discusses how to debias them. However, as for MMD, it is important to note that every statement/experiment in this paper about the power of $dCor_n^2$ qualitatively holds true for $dCov_n^2$, and both their unbiased versions also.

The relationship between kernels and distances

As mentioned earlier, the two problems of two-sample and independence testing are related because any algorithm for the former yields an algorithm for the latter. Indeed, corresponding to MMD, there exists a test statistic using kernels called HSIC, see (Gretton et al. 2005), for the independence testing problem. The sample expression for HSIC looks a lot like Eq.(2), except where A and B represent the pairwise kernel matrices instead of distance matrices. Similarly, corresponding to $dCov$, there exists a test statistic using distances for the two-sample testing problem, whose empirical statistic matches that of Eq.(1), except using distances instead of kernels. This is not a coincidence. Informally, for every positive-definite kernel, there exists a negative-definite metric, and vice-versa, such that these quantities are equal; see (Sejdinovic et al. 2013) for more formal statements.

When a *characteristic* kernel, see (Gretton et al. 2012a) for a definition, or its corresponding distance metric is used, the population quantities corresponding to all the test statistics equals zero iff the null hypothesis is true. In other words $MMD = 0$ iff $P = Q$, $dCor = dCov = 0$ iff X, Y are independent. It suffices to note that this paper will only be dealing with distances or kernels satisfying this property.

Permutation testing and power simulations

A permutation-based test for any of the above test statistics T proceeds in the following manner :

1. Calculate the test statistic T on the given sample.
2. (Independence) Keeping the order of $x_1, ..., x_n$ fixed, randomly permute $y_1, ..., y_n$, and recompute the permuted statistic T. This destroys dependence between xs, ys and behaves like one draw from the null distribution of T.

2'. (Two-sample) Randomly permute the $m+n$ observations, call the first m of them your xs and the remaining your ys, and now recompute the permuted statistic T. This behaves like one draw from the null distribution of the test statistic.

3. Repeat step 2 a large number of times to get an accurate estimate of the null distribution of T. For a prespecified type-1 error α, calculate threshold t_α in the right tail of the null distribution.
4. Reject H_0 if $T > t_\alpha$.

This test is proved to be *consistent* against any fixed alternative, in the sense that as $n \to \infty$ for a fixed type-1 error, the type-2 error goes to 0, or the power goes to 1. Empirically, the power can be calculated using simulations as:

1. Choose a distribution P_{XY} (or P, Q) such that H_1 is true. Fix a sample size n (or m, n).
2. (Independence) Draw n samples, run the independence test. (Two-sample) Draw m samples from P and n from Q, run the two-sample test. A rejection of H_0 is a success. This is one trial.
3. Repeat step 2 a large number of times (conduct many independent trials).
4. The power is the fraction of successes (rejections of H_0) to the total number of trials.

Note that the power depends on the alternative P_{XY} or P, Q.

Paper Organization

In Section 2, we discuss the misconceptions that exist regarding the supposedly good behavior of these tests in high dimensions. In Section 3, we demonstrate that against *fair* alternatives, the power of kernel and distance based hypothesis tests degrades with dimension. In Section 4, we provide some initial insights as to why this might be the case, and the role of the bandwidth choice (when relevant) in test power.

2 Misconceptions about power in high-dimensions

Hypothesis tests are typically judged along one metric - test power. To elaborate, for a fixed type-1 error, we look at how small type-2 error is, or equivalently how large the power is. Further, one may also study the rate at which the power improves to approach one or degrades to zero (with increasing number of points, or even increasing dimension). So when a hypothesis test is said to "work well" or "perform well", it is understood to mean that it has high power with a controlled type-1 error.

We believe that there are a variety of reasons why people believe that the power of the aforementioned hypothesis tests does not degrade with the underlying dimension of the data. We first outline and address these, since they will improve our general understanding of these tests and guide us in our experiment design in Section 3.

Claims of good performance

A proponent of distance-based tests claims, on Page 17 of the tutorial presentation (Székely 2014), that *"The power of dCor test for independence is very good especially for high dimensions p,q"*. In other words, not only does he claim that it does not get worse, but it gets better in high dimensions. Unfortunately, this is not backed up with evidence, and in Section 3, we will provide evidence to the contrary.

Given the strong relationship between kernel-based and distance-based methods described in the introduction, one might be led to conclude that kernel-based tests also get better, or at least not worse, in high dimensions. Again, this is not true, as we will see in Section 3.

Estimation of MMD^2 is independent of dimension

It is proved in (Gretton et al. 2012a) that the rate of convergence of the estimators of MMD^2 to the population quantity is $O(1/\sqrt{n})$, independent of the underlying dimension of the data. Formally, suppose $0 \leq k(x,x) \leq K$, then with probability at least $1-\delta$, we have

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

A similar statement is also true for the unbiased estimator. This error is indeed independent of dimension, in the sense that in every dimension (large or small), the convergence rate is the same, and the rate does not degrade in higher dimensions. This was also demonstrated empirically in Fig. 3 of (Sriperumbudur et al. 2012a).

However, one must not mix up estimation error with test power. While it is true that estimation does not degrade with dimension, it is possible that test power does (as we will demonstrate in Section 3). This leads us to our next point.

Estimation vs Testing

In the normal means problem, one has samples from a Gaussian distribution, and we have one of two objectives - either estimate the mean of the Gaussian, or test whether the mean of the Gaussian is zero or not. In this setting, it is well known and easily checked that *estimation of the mean is harder than testing if the mean is zero or not*.

Using this same intuition, one might be tempted to assume that hypothesis testing is generally easier than estimation, or specifically like that Gaussian mean case that *estimation of the MMD is harder than testing if the MMD is zero or not*.

However, this is an incorrect assumption, and the intuition attained from the Gaussian setting can be misleading.

On a similar note, (Székely and Rizzo 2013) note that even when P, Q are independent, if n is fixed and $d \to \infty$ then the biased $dCor \to 1$. Then, they show how to form an unbiased $dCor$ (called $udCor$) so that $udCor \to 0$ as one might desire, even in high dimensions. However, they seem to be satisfied with good estimation of the population $udCor$ value (0 in this case), which does not imply good test power. As we shall see in our experiments, in terms of power, unbiased $udCor$ does no better than biased $dCor$.

No discussion about alternatives

One of the most crucial points for examining test power with increasing dimension is the choice of alternative hypothesis. Most experiments in (Gretton et al. 2012a; Székely, Rizzo, and Bakirov 2007; Gretton et al. 2005) are conducted without an explicit discussion or justification for the sequences of chosen alternatives. For example, consider the case of two-sample testing below. As the underlying dimension increases, if the two distributions "approach" each other in some sense, then the simulations might suggest that test power degrades; conversely if the distributions "diverge" in some sense, then the simulations might suggest that test power does not actually degrade much.

Let us illustrate the lack of discussion/emphasis on the choice of alternatives in the current literature. Assume P, Q are spherical Gaussians with the same variance, but different means. For simplicity, say that in every dimension, the mean is always at the origin for P. When P and Q are one-dimensional, say that the mean of Q is at the point 1 - when dimension varies, we need to decide (for the purposes of simulation) how to change the mean of Q. Two possible suggestions are $(1, 0, 0, 0, ..., 0)$ and $(1, 1, 1, 1, ..., 1)$, and it is possibly unclear which is a *fairer* choice. In Fig. 5A of (Gretton et al. 2012a), the authors choose the latter (verified by personal communication) and find that the power is only very slowly affected by dimension. In experiments in the appendix of (Gretton et al. 2012b), the authors choose the former and find that the power decreases fast with dimension. Fig. 3 in (Sriperumbudur et al. 2012a) also makes the latter choice, though only for verifying estimation error decay rate. In all cases, there is no justification of these choices.

Our point is the following - when n is fixed and d increasing, or both are increasing, it is clearly possible to empirically demonstrate any desired behavior of power (i.e. increasing, fairly constant, decreasing) in simulations, by appropriately changing the choice of alternatives. This raises the question - what is a good or *fair* choice of alternatives by which we will not be misled? We now discuss our proposal for this problem.

Fair Alternatives

We propose the following notion of fair alternatives - for two-sample testing as dimension increases, the Kullback Leibler (KL) divergence between the pairs of distributions should remain constant, and for independence testing as dimension increases, the mutual information (MI) between X, Y should remain constant.

Our proposal is guided by the fact that KL-divergence (and MI) is a fundamental information-theoretic quantity that is well-known to determine the hardness of hypothesis testing problems, for example via lower bounds using variants of Fano's inequality, see (Tsybakov 2010). By keeping the KL (or MI) constant, we are not making the problem artificially harder or easier (in the information-theoretic sense) as dimension increases.

Let us make one point clear - we are *not* promoting the use of KL or MI as test statistics, or saying that one should estimate these quantities from data. We are also not comparing the performance to MMD/HSIC to the performance of KL/MI. We are only suggesting that one way of calibrating our simulations, so that our simulations are fair representations of true underlying behavior, is to make parameter choices so that KL/MI between the distributions stay constant as the dimension increases.

For the aforementioned example of the Gaussians, the choice of $(1, 0, 0, 0, ..., 0)$ turns out to be a fair alternative, while $(1, 1, 1, ..., 1)$ increases the KL and makes the problem artificially easier. If we fix n, a method would work well in high-dimensions if its power remained the same irrespective of dimension, against fair alternatives. In the next section, we will demonstrate using variety of examples, that the power of kernel and distance based tests decays with increasing dimension against fair alternatives.

3 Simple Demonstrations of Decaying Power

As we mentioned in the introduction, we will be working with characteristic kernel. Two such kernels we consider here are also translation invariant - Gaussian $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{\gamma^2}\right)$ and Laplace $k(x, y) = \exp\left(-\frac{\|x-y\|}{\gamma}\right)$, both of which have a bandwidth parameter γ. One of the most common ways in the literature to choose this bandwidth is using the *median heuristic*, see (Schölkopf and Smola 2002), according to which γ is chosen to be the median of all pairwise distances. It is a heuristic because there is no theoretical understanding of when it is a good choice.

In our experiments, we will consider a range of bandwidth choices - from much smaller to much larger than what the median heuristic would choose - and plot the power for each of these. The y-axis will always represent power, and the x-axis will always represent increasing dimension. There was no perceivable difference between using biased and unbiased MMD^2, so all plots apply for both estimators.

(A) Mean-separated Gaussians, Gaussian kernel

Here P, Q are chosen as Gaussians with covariance matrix I. P is centered at the origin, while Q is centered at $(1, 0, ..., 0)$ so that $KL(P, Q)$ is kept constant. A simple calculation shows that the median heuristic chooses $\gamma \approx \sqrt{d}$ - we run the experiment for $\gamma = d^\alpha$ for $\alpha \in [0, 1]$. As seen in Figure 1, the power decays with d for all bandwidth choices. Interestingly, the median heuristic maximizes the power.

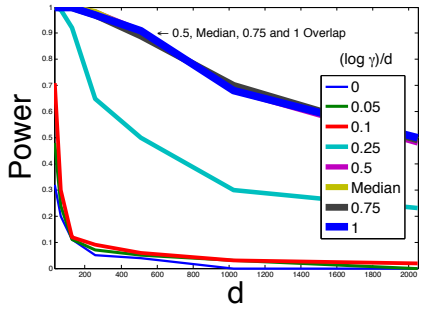

Figure 1: MMD Power vs d of for mean-separated Gaussians using Gaussian kernel with bandwidths $d^\alpha, \alpha \in [0, 1]$.

(B) Mean-separated Laplaces, Laplace kernel

Here P, Q are both the product of d independent univariate Laplace distributions with the same variance. As before, P is centered at the origin, while Q is centered at $(1, 0, 0, ..., 0)$ - Section 4 shows that this choice keeps $KL(P, Q)$ constant. Here too, the median heuristic chooses γ on the order of $\sqrt{d}$, and again we run the experiment for $\gamma = d^\alpha$ for $\alpha \in [0, 2]$. Once again, note that the power decays with d for all the bandwidth choices. However, this is an example where the median heuristic does not maximize the power - larger choices like $\gamma = d, d^2$ work better (see Figure 2).

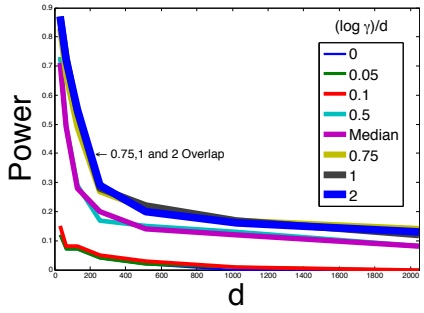

Figure 2: MMD Power vs d for mean-separated Laplaces using Laplace kernel with bandwidths $d^\alpha, \alpha \in [0, 2]$.

(C) Non-diagonal covariance matrix Gaussians

Let us consider the case of independence testing and $dCor$ to show that (as expected) this behavior is not restricted to two-sample testing or MMD^2. Here, P, Q will both be origin-centered d dimensional gaussians. If they were independent, their joint covariance matrix Σ would be I. Instead, we ensure that a constant number (say 4 or 8) of off-diagonal entries in the covariance matrix are non-zero. We keep the number of non-zeros constant as dimension increases. One can verify that this keeps the mutual information constant as dimension increases (as well as other quantities like $\log\det\Sigma$, which is the amount of information encoded in Σ, and $\|\Sigma - I\|_F^2$ which is relevant since we are really trying to detect any deviation of Σ from I). Figure 3 shows that the power of $dCor, udCor$ both drop with dimension - hence debiasing the test statistic does make the *value* of the test statistic more accurate but it does *not* improve the corresponding power.

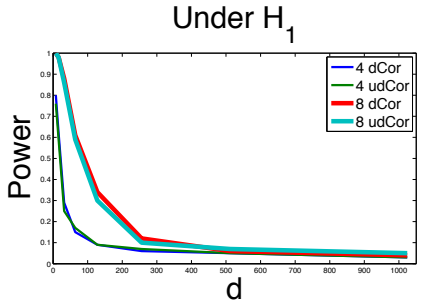

Figure 3: Power vs d of dCor and unbiased dCor (udCor) for the dependent Gaussians example, with 4 or 8 off-diagonal non-zeros in the joint covariance matrix.

(D) Differing-variance Gaussians, Gaussian kernel

We take $P = \otimes_{i=1}^{d-1}\mathcal{N}(0,1) \otimes \mathcal{N}(0,4)$ and $Q = \otimes_{i=1}^{d}\mathcal{N}(0,1)$ (both are origin centered). As we shall see in the next section, this choice keeps KL constant.

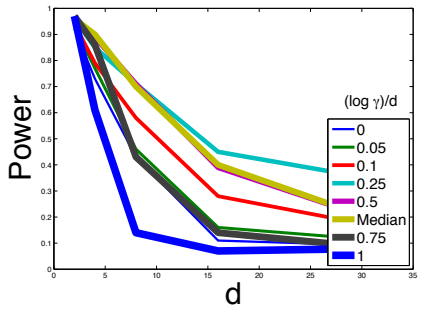

Figure 4: MMD Power vs d for Gaussians differing in variance using Gaussian kernel with bandwidths $d^\alpha, \alpha \in [0,1]$.

It is easy to see in Fig.4 that the power of MMD decays with dimension, for all choices of the bandwidth parameter.

4 MMD^2 vs KL

Here, we shed light on why the power of MMD^2 might degrade with dimension, against alternatives where KL is kept constant. We actually calculate the MMD^2 for the aforementioned examples (A), (B) and (D), and compare it to KL.

It is known that $\text{MMD}^2(p,q) \leq KL(p,q)$ (Sriperumbudur et al. 2012b). We show that it can be smaller than the KL by polynomial or even exponential factors in d - in all our previous examples, while KL was kept constant, MMD was actually shrinking to zero polynomially or exponentially fast. This discussion will also bring out the role of the bandwidth choice, especially the median heuristic.

(A) Mean-separated Gaussians, Gaussian kernel

Some special cases of the following calculations appear in (Balakrishnan 2013) and (Sriperumbudur et al. 2012a). Our results are more general, and unlike them we clearly analyze the role of the bandwidth choice. We also simplify the calculations to make direct comparisons to KL divergence possible, unlike earlier work which had different aims.

Proposition 1. *Suppose $p = \mathcal{N}(\mu_1, \Sigma)$ and $q = \mathcal{N}(\mu_2, \Sigma)$. Using a Gaussian kernel with bandwidth γ, MMD^2 =*

$$2\left(\frac{\gamma^2}{2}\right)^{d/2} \frac{1 - \exp(-\Delta^\top(\Sigma + \gamma^2 I/2)^{-1}\Delta/4)}{|\Sigma + \gamma^2 I/2|^{1/2}}.$$

where $\Delta = \mu_1 - \mu_2 \in \mathbb{R}^d$.

The above proposition (proved in Appendix A of the full version[1]) looks rather daunting. Let us derive a revealing corollary, which involves a simple approximation by Taylor's theorem.

Corollary 1. *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*

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

Recall that when $\Sigma = \sigma^2 I$, the KL is given by

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

Let us now see how the bandwidth choice affects the MMD. 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 assume d is large so that $(1 + 1/d)^d \approx e$.

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

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

[1]Full version of the paper can be found at http://arxiv.org/abs/1406.2083.

Hence, the population MMD^2 goes to zero exponentially fast in d as $\exp(d^{2\epsilon}/2)$, verified in Fig. 5, and is exponentially smaller than $KL(p,q)$.

Observation 2 (median heuristic). *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 that when $\Sigma = \sigma^2 I$, we have $\mathbb{E}\|x_i - x_j\|^2 \approx 2\sigma^2 d + \|\mu_1 - \mu_2\|^2$ which is dominated by the first term as d increases. This indicates that the median heuristic chooses $\gamma \approx \sigma\sqrt{d}$, verified in Fig.5. Here 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 $1/d$.

Observation 3 (overestimated bandwidth). *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. Here too, the MMD is a factor $1/d$ smaller than the KL.

We demonstrate in Fig.5 that our approximations are actually accurate, by calculating the population MMD as a function of d for each bandwidth choice. The population MMD is approximated by calculating the empirical MMD after drawing a very large number of samples so that the approximation error is small.

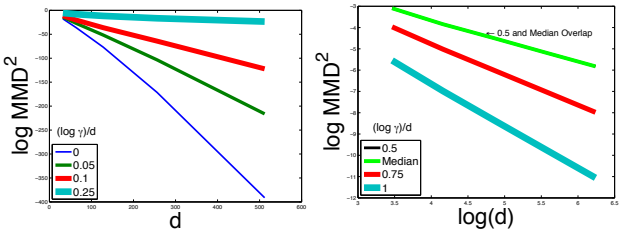

Figure 5: MMD^2 vs d for mean-separated Gaussians using Gaussian kernel. The left panel shows behavior predicted by Observations 1,2 and the right by Observation 3.

(B) Mean-separated Laplaces, Laplace kernel

In the previous example, the median heuristic maximized the MMD. However, this is not always the case and now we present one such example where the median heuristic results an exponentially small MMD. We use Taylor approximations to yield expressions that are insightful.

Proposition 2. *Let* $\mu_1, \mu_2 \in \mathbb{R}^d$. *If* $p = \otimes_i \mathrm{Laplace}(\mu_{1,i}, \sigma)$ *and* $q = \otimes_i \mathrm{Laplace}(\mu_{2,i}, \sigma)$, *using a Laplace kernel with bandwidth* γ, *we have*

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

It is proved in Appendix A of the full version and the accuracy of approximation is verified in Appendix B of the full version . It can be checked that

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

using Taylor's theorem, $e^{-x} \approx 1 - x + x^2/2 + o(x^2)$.

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

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

It is easily derived that $\mathbb{E}\|x_i - x_j\|^2 \approx 2\sigma^2 d$ so the median heuristic chooses $\gamma \approx \sigma\sqrt{d}$, experimentally verified in Fig.6. This time, *the median heuristic is suboptimal* and MMD^2 drops to zero exponentially in d, also making it exponentially smaller than KL.

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

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

A bandwidth of $\gamma = \sigma d$ is optimal, making the denominator $\approx \sigma^2 de$, which is still a factor $1/d$ smaller than KL. An overestimated bandwidth again leads to a slow polynomial drop in MMD. This behavior is verified in Fig. 6.

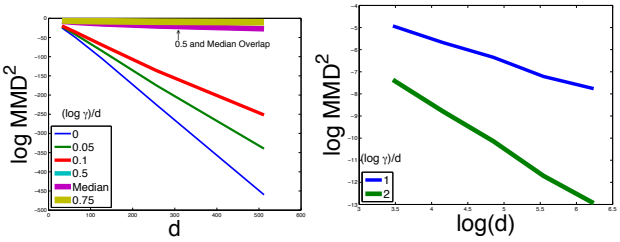

Figure 6: MMD^2 vs d for mean-separated Laplaces using Laplace kernel. The left panel shows behavior predicted by Observation 4 and right panel by Observation 5.

(D) Differing-variance Gaussians, Gaussian kernel

Example 3 in Sec. 4.2 of (Sriperumbudur et al. 2012a) has related calculations, again with a different aim. We again use Taylor approximations to yield insightful expressions.

Proposition 3. *Suppose* $p = \otimes_{i=1}^{d-1}\mathcal{N}(0,\sigma^2)\otimes\mathcal{N}(0,\tau^2)$ *and* $q = \otimes_{i=1}^{d}\mathcal{N}(0,\sigma^2)$. *For a Gaussian kernel of bandwidth* γ,

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

It is proved in Appendix A of the full version and the accuracy of approximation is verified in Appendix B of the full version. It is easy to verify that

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

where we used Taylor's theorem for $\log x$. These calculations for MMD, KL suggest that the observations made for the earlier example of mean-separated Gaussians carry forward qualitatively here as well, verified by Fig.7.

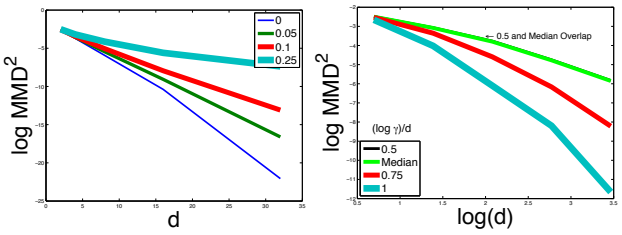

Figure 7: MMD^2 vs d for Gaussian distributions with differing variance using Gaussian kernel. The behavior in boths panels is very similar to Fig.5 as predicted by Proposition 3.

5 Conclusion

This paper addressed an important issue in our understanding of the power of recent nonparametric hypothesis tests. We identified the various reasons why misconceptions exist about the power of these tests. Using our proposal of fair alternatives, we clearly demonstrate that the power of biased/unbiased kernel/distance based two-sample/independence tests all degrade with dimension.

We also provided an understanding of how a popular kernel-based test statistic, the Maximum Mean Discrepancy (MMD), behaves with dimension and bandwidth choice - its value drops to zero polynomially (at best) with dimension even when the KL-divergence is kept constant - shedding some light on why the power degrades with dimension (differentiating the empirical quantity from zero becomes harder as the population value approaches zero).

This paper provides an important advancement in our current understanding of the power of modern nonparametric hypothesis tests in high dimensions. While it does *not* completely settle the question of how these tests behave in high dimensions, it is a crucial first step.

Acknowledgements

This work is supported in part by NSF grants IIS-1247658 and IIS-1250350.

References

Balakrishnan, S. 2013. *Finding and Leveraging Structure in Learning Problems*. Ph.D. Dissertation, Carnegie Mellon University.

Eric, M.; Bach, F. R.; and Harchaoui, Z. 2008. Testing for homogeneity with kernel fisher discriminant analysis. In Platt, J.; Koller, D.; Singer, Y.; and Roweis, S., eds., *Advances in Neural Information Processing Systems 20*, 609–616. Curran Associates, Inc.

Fukumizu, K.; Gretton, A.; Sun, X.; and Schölkopf, B. 2008. Kernel measures of conditional dependence. In *NIPS 20*, 489–496. Cambridge, MA: MIT Press.

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

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

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

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

Schölkopf, B., and Smola, A. J. 2002. *Learning with Kernels*. Cambridge, MA: MIT Press.

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

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

Sriperumbudur, B. K.; Fukumizu, K.; Gretton, A.; Schölkopf, B.; and Lanckriet, G. R. G. 2012b. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics* 6(0):1550–1599.

Székely, G. J., and Rizzo, M. L. 2009. Brownian distance covariance. *The Annals of Applied Statistics* 3(4):1236–1265.

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

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

Székely, G. 2014. Distance correlation. *Workshop on Nonparametric Methods of Dependence, Columbia University, http://dependence2013.wikischolars.columbia.edu/file/view/Szekely Columbia Workshop.ppt*.

Tsybakov, A. 2010. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York.