Summary of
"On the High-dimensional Power of Linear-time Kernel Two-Sample Testing under Mean-shift Alternatives"

Wittawat Jitkrittum (wittawatj@gmail.com)

November 25, 2015

Created on 22 Nov 2015. This is a summary of Ramdas et al. [2014]. The paper also appeared in AISTATS 2015.

Summary

Given two samples $\{x_i\}_{i=1}^n \sim P$ and $\{y_i\}_{i=1}^n \sim Q$ with unknown P, Q defined on $\mathbb{R}^d$, the goal of the two sample test is to test the hypotheses $H_0 : P = Q$ v.s. $H_1 : P \neq Q$. A nonparameteric kernel-based test which considers such general alternatives was recently proposed by Gretton et al. [2012]. Although the power of the test was studied under the setting that $n \to \infty$ with fixed d, it is unclear how the power is affected when $(n, d) \to \infty$. The main contribution of the paper is to characterize the power of the linear MMD (with a Gaussian kernel) test under a mean-shift alternative (i.e., $H_0 : \mathbb{E}_{x\sim P}[x] = \mathbb{E}_{y\sim Q}[y]$ and H_1 :$\mathbb{E}_{x\sim P}[x] \neq \mathbb{E}_{y\sim Q}[y]$) under the $(n, d) \to \infty$ setting.

1 Hypothesis testing

Let $X^{(n)} := \{x_i\}_{i=1}^n$ and $Y^{(n)} := \{y_i\}_{i=1}^n$. A test is a function to a specific hypotheses H_0 and H_1, that takes $X^{(n)}$ and $Y^{(n)}$ and outputs either 0 or 1, where 1 indicates the rejection of H_0, and 0 means failure to reject H_0 due to insufficient evidence. The **type-1 error** α or false positive rate is defined as

$$\alpha = p(\text{reject } H_0 \mid H_0 \text{ is true}).$$

The **type-2 error** β or false negative rate is defined as

$$\beta = p(\text{not reject } H_0 \mid H_1 \text{ is true}).$$

Generally decreasing one will increase the other. We refer to $1 - \beta$ as the **power** of the test i.e., the probability of correctly rejecting H_0. Many tests compute a **test statistic** $T := T(X^{(n)}, Y^{(n)})$ and reject H_0 if $T > c_\alpha$ where the rejection threshold c_α depends on the distribution of T under H_0, and a prechosen **significance level** α.

1.1 Two-sample test with MMD

One of the most popular tests for nonparameteric two-sample testing is the kernel two-sample test proposed by Gretton et al. [2012]. The test uses maximum mean discrepancy (MMD) as the test statistic T. Given a symmetric positive definite kernel function $k(x, y)$, MMD is defined as

$$\mathrm{MMD}^2(P, Q) := \mathbb{E}_{x\sim P}\mathbb{E}_{x'\sim P}k(x, x') + \mathbb{E}_{y\sim Q}\mathbb{E}_{y'\sim Q}k(y, y') - 2\mathbb{E}_{x\sim P}\mathbb{E}_{y\sim Q}k(x, y).$$

An unbiased estimator is given by

$$\mathrm{MMD}_u^2 = \frac{1}{n(n-1)}\sum_{i\neq j}^n k(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i\neq j}^n k(y_i, y_j) - \frac{2}{n^2}\sum_{i,j=1}^n k(x_i, y_j),$$

which can be computed in $O(n^2)$. A linear unbiased statistic is given by

$$\mathrm{MMD}_l^2 = \frac{1}{n/2}\sum_{i=1}^{n/2} \left[k(x_{2i-1}, x_{2i}) + k(y_{2i-1}, y_{2i}) - k(x_{2i-1}, y_{2i}) - k(y_{2i-1}, x_{2i})\right].$$

2 Power of MMD_l

Let $z_i := (x_i, y_i)$. Define $h(z_i, z_j)$ as

$$h(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i).$$

Then, we have $\mathrm{MMD}_l^2 = \frac{1}{n/2}\sum_{i=1}^{n/2} h(z_{2i-1}, z_{2i})$. Gretton et al. [2012] showed that under both H_0 and H_1 with fixed d and $n \to \infty$, we have

$$F := \frac{\sqrt{n}(\mathrm{MMD}_l^2 - \mathrm{MMD}^2)}{\sqrt{V}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $V = 2\mathbb{V}_{z,z'}h(z,z')$. Equivalently, $\mathrm{MMD}_l^2 \sim \mathcal{N}(\mathrm{MMD}^2, V/n)$. Note that under $H_0 : P = Q$, we have $\mathrm{MMD}^2 = 0$. The test rejects H_0 if $\mathrm{MMD}_l^2 > z_\alpha\sqrt{\frac{\hat{V}}{n}}$ where z_α is the $(1-\alpha)$ quantile of $\mathcal{N}(0,1)$, and $\hat{V}$ is the empirical estimator of V.

The test power β is x

$$\begin{aligned}
\beta = P_{H_1}\left(\mathrm{MMD}_l^2 > z_\alpha\sqrt{\frac{\hat{V}}{n}}\right) &= P_{H_1}\left(\frac{\sqrt{n}(\mathrm{MMD}_l^2 - \mathrm{MMD}^2)}{\sqrt{V}} > \sqrt{\frac{n}{V}}\left[z_\alpha\sqrt{\frac{\hat{V}}{n}} - \mathrm{MMD}^2\right]\right) \\
&= P_{H_1}\left(F > \sqrt{\frac{\hat{V}}{V}}z_\alpha - \sqrt{\frac{n}{V}}\mathrm{MMD}^2\right) && (1) \\
(\text{as } n\to\infty, d \text{ fixed}) &\to P_{H_1}\left(Z > z_\alpha - \sqrt{\frac{n}{V}}\mathrm{MMD}^2\right) && (2) \\
&= 1 - \Phi\left(z_\alpha - \sqrt{\frac{n}{V}}\mathrm{MMD}^2\right) \\
&= \Phi\left(\sqrt{\frac{n}{V}}\mathrm{MMD}^2 - z_\alpha\right), && (3)
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ and Φ is the CDF of Z. Observe that the power behaves like $\Phi(\Theta(\sqrt{n}))$. The expression for power holds for general alternatives i.e., $H_0 : P = Q$.

2.1 Challenges in high dimensions

The goal of this paper is to characterize the power as $(n,d) \to \infty$, not just $n \to \infty$, under the mean-shift alternatives i.e., $H_0 : \mu_P = \mu_Q$ and $H_1 : \mu_P \neq \mu_Q$ where $\mu_P = \mathbb{E}_{x\sim P}[x]$. There are four challenges

1. MMD does not depend on n. But, it depends on d.
2. The variance V in the last line depends on d.
3. We relied on the fact that $F \to Z$ as $n \to \infty$. Does it still hold if $(n,d) \to \infty$?
4. We need $\hat{V} \to V$ in Eq. 2. Is it still true if $(n,d) \to \infty$?

We will see that the power can still be characterized under the assumption that $(n,d) \to \infty$.

3 Assumptions and contributions

Assumptions

1. **(A1)** $x_i = Us_i + \mu_P \in \mathbb{R}^d$. Similarly, $y_i = Ut_i + \mu_Q$. The d coordinates are i.i.d. zero-mean, and U is orthogonal i.e., $UU^\top = I$.
   - Note that U is the same for both x and y. Also, it is assumed that the moments of s and t are the same. This should not be a problem as it only makes H_0 more difficult to reject. The derived power should still holds even when the moments of s and t are different. The authors did not comment on this point.

- A1 implies that $\mathbb{V}[x] = \mathbb{V}[y] = \sigma^2 I$ because $\mathbb{V}[x] = \mathbb{E}_s[Uss^\top U^\top] = U(\sigma^2 I)U^\top$. This means x and y have spherical covariances, but not necessarily follow a Gaussian. According to the paper, the results still hold even with a diagonal covariance.

2. **(A2)** $\mathbb{E}[|s_{(k)}|^6] < \infty$ and $\mathbb{E}[|t_{(k)}|^6] < \infty$ where $\cdot_{(k)}$ means the k^{th} coordinate. This implies that all moments up to 6 exist.
   - The existence of 3^{rd} and 4^{th} moments is needed for the Berry-Esseen lemma to guarantee that $F \to Z$ and $\hat{V} \to V$ when $(n, d) \to \infty$.
   - The existence of 6^{th} moments is to bound the Taylor expansion of $\exp(-x)$ for the Gaussian kernel.

3. **(A3)** Assume $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{\gamma^2}\right)$, a Gaussian kernel with width γ.

3.1 Main result

Define $\delta := \mu_P - \mu_Q$. Let $\sigma^2 := \mathbb{E}[s^2_{(k)}] = \mathbb{E}[t^2_{(k)}] = \mathbb{V}[x_{(k)}]$.

Theorem 1. *Consider the mean-shift alternatives. Assume $\gamma = \Omega(\sqrt{d})$. Under A1-A3, with $(n, d) \to \infty$, the asymptotic test power β of MMD^2_l is*

$$\beta = \Phi\left(\frac{\sqrt{n}\|\delta\|^2}{\sqrt{8d\sigma^4 + 8\sigma^2\|\delta\|^2}} - z_\alpha\right).$$

The notation $\gamma = \Omega(\sqrt{d})$ means $\limsup_{d\to\infty} \gamma/\sqrt{d} > 0$. That is, γ grows at least as fast as $\sqrt{d}$.

Remarks about the theorem

1. The power is independent of the Gaussian bandwidth γ as long as $\gamma = \Omega(\sqrt{d})$. In particular, this growth rate applies to the popular median heuristic. The assumption that $\gamma = \Omega(\sqrt{d})$ is to control the residual term in the Taylor expansion of the Gaussian kernel.
2. Define the signal to noise ratio (SNR) as $\Psi := \|\delta\|/\sigma$. The power behaves like $\Phi\left(\frac{\sqrt{n}\Psi^2}{\sqrt{8d+8\Psi^2}} - z_\alpha\right)$. A natural question to ask: when is the power independent of d? This is characterized by the following corollaries.

Corollary 1. *If $\Psi = o(\sqrt{d})$ i.e., $\lim_{d\to\infty} \Psi/\sqrt{d} = 0$ (SNR is small), then $\beta \to 1$ at the rate $\Phi\left(\sqrt{n}\Psi^2/\sqrt{d}\right)$.*

Corollary 2. *If $\Psi = \omega(\sqrt{d})$ i.e., $\lim_{d\to\infty} \sqrt{d}/\Psi = 0$ (SNR is large), then $\beta \to 1$ at the rate $\Phi(\sqrt{n}\Psi)$.*

The switch occurs at Ψ being on the order of $\sqrt{d}$.

4 Proof of the theorem

We need to address the four challenges in Sec. 2.1. We first consider four lemmas.

Lemma 1. *Under A1-A3 and $\gamma = \Omega(\sqrt{d})$,*

$$\mathrm{MMD}^2 = \frac{2\|\delta\|^2}{\gamma^2}(1 + o(1)).$$

The idea of the proof relies on the use of Taylor's expansion for the Gaussian kernel. The first term of the MMD^2 is

$$\mathbb{E}_x\mathbb{E}_{x'} k(x, x') = \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-y\|^2}{\gamma^2}\right) p(x)p(y)\,\mathrm{d}x\mathrm{d}y = \left(1 - \frac{2\sigma^2}{\gamma^2}\right)^d,$$

relying on Taylor expansion of $\exp(-x)$ around 0 i.e., $\exp\left(-\frac{\|x-y\|^2}{\gamma^2}\right) = 1 - \frac{\|x-y\|^2}{\gamma^2} + \frac{\exp\left(-\lambda\frac{\|x-y\|^2}{\gamma^2}\right)\|x-y\|^2}{2\gamma^4}$ for some $\lambda \in [0, 1]$. This gives the expression for MMD^2 to be used in Eq. 3.

Lemma 1 is wrong? Kacper and Arthur found that the expression of MMD^2 does not match the one given, when applied to the case where P, Q are Gaussians.

Lemma 2. *Under A1-A3 and* $\gamma = \Omega(\sqrt{d})$, $V = \mathbb{V}_{z,z'} h(z, z')$ *(from here on the definition of* V *changed a bit) is given by*

$$V = \frac{16 d\sigma^4 + 16\sigma^2 \|\delta\|^2}{\gamma^4}(1 + o(1)).$$

The proof involves a long algebra as a result of expanding the definition $V = \mathbb{E}_{z,z'} h^2(z, z') - \mathrm{MMD}^4$. This will be used in Eq. 3.

Lemma 3 (Berry-Esseen bound)**.** *Under A1-A3 and* $\gamma = \Omega(\sqrt{d})$,

$$\sup_t \left| p\left(\frac{\sqrt{n}\left(\mathrm{MMD}_l^2 - \mathrm{MMD}^2\right)}{\sqrt{2V}} \le t \right) - \Phi(t) \right| \le 20/\sqrt{n}.$$

The proof is more or less directly given by the Berry-Esseen theorem which gives a bound on the difference of CDF of $\frac{\sqrt{n}}{\sigma} \times$empirical average (which follows a Gaussian by the central limit theorem) and $\Phi(t)$. Berry-Esseen theorem requires the existence of the third moment. We need this to conclude that $F \to Z$ in Eq. 2.

Lemma 4. *Under A1-A3 and* $\gamma = \Omega(\sqrt{d})$, $\sqrt{\hat{V}/V} = 1 + O_P(n^{-1/4})$.

The proof relies on a general theorem characterizing the bias and variance of U-statistics given in Serfling [1980, Theorem A, Sec. 2.2.3].

Proof of the theorem

- By Lemma 4, $\hat{V} \to V$ in Eq. 1.
- By Lemma 3, $F \to Z$ in Eq. 2 when $(n, d) \to \infty$.
- We get the theorem by using Lemma 1 to rewrite MMD^2 in Eq. 3, and Lemma 2 to rewrite V in Eq. 3.

References

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, March 2012. URL http://jmlr.csail.mit.edu/papers/v13/gretton12a.html. 1, 2

Aaditya Ramdas, Sashank J. Reddi, Barnabas Poczos, Aarti Singh, and Larry Wasserman. On the High-dimensional Power of Linear-time Kernel Two-Sample Testing under Mean-difference Alternatives. *arXiv:1411.6314 [cs, math, stat]*, November 2014. URL http://arxiv.org/abs/1411.6314. arXiv: 1411.6314. 1

R. J. Serfling. *Approximation theorems of mathematical statistics.* Wiley series in probability and mathematical statistics. Wiley, New York, 1980. ISBN 0-471-02403-1. URL http://opac.inria.fr/record=b1086006. Includes indexes. 4