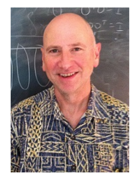
Two Sample Testing – Free Lunches & Comp-Stat Tradeoffs

Aaditya Ramdas*, Sashank Reddi*, Barnabas Poczos, Aarti Singh, Larry Wasserman


Carnegie Mellon University





Introduction

Thought Experiment: Red vs Blue Pill

4-week Experiment: Each day, give Neo either R or B pill, measure real-valued thyroxine concentration (X or Y respectively).
Data: R: $X_1, ..., X_{14} \sim P$ (iid) and B: $Y_1, ..., Y_{14} \sim Q$ (iid).
Mean-difference alternatives:

$$H_0 : \mathbb{E}_P[X] = \mathbb{E}_Q[Y] \quad \text{vs.} \quad H_1 : \mathbb{E}_P[X] \neq \mathbb{E}_Q[Y]$$

General alternatives:

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

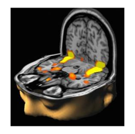
Two-sample test: inputs data, outputs 0 or 1. "Stochastic proof by contradiction". **Nonparametric**: no assumptions about P, Q.

Real Experiment: Faces vs Houses

Question: Does brain region R differentiate faces and non-faces?

- Show someone a face:

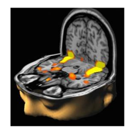
- Measure brain activity (in, say, 500 voxels)



- Repeat 200 times

- Show someone a house:

- Measure brain activity (in, say, 500 voxels)



- Repeat 200 times

Data: $X_1, ..., X_{200} \sim P \in \mathbb{R}^{500}$ and $Y_1, ..., Y_{200} \sim Q \in \mathbb{R}^{500}$.
Test: Mean-difference alternative or general alternative.

Errors, and power

Two ways that a two-sample test could be wrong:

1. **False Positive**: When $P = Q$ (H_0), but the test returns 1. The type-1 error α is the probability of a false positive.

   High α - false discoveries - dangerous! Control at (say) 0.05.

2. **False Negative**: When $P \neq Q$ (H_1), but the test returns 0. The type-2 error β is the probability of a false negative.

   High β implies low power $\phi := 1 - \beta$ - very weak test incapable of detecting real differences that do exist.

A test is (classically) **consistent** if, while controlling α at any level (say 0.05), the power goes to 1 as the number of samples $n \to \infty$.

A test is (high-dim) **consistent** if, while controlling α at any level (say 0.05), the power goes to 1 as $(n, d) \to \infty$.

Open Questions

Tests and Tradeoffs

1. **Parametric + Mean-Difference Alternative**
   Eg: Threshold Hotelling's Statistic $(\bar{X} - \bar{Y})^T S^{-1} (\bar{X} - \bar{Y})$
   Eg: Random Projections variant : Lopes-Jacob-Wainwright'12
2. **Nonparametric + Mean-Difference Alternative**
   Eg: Diagonalize/drop S (SD, BS, CQ: Chen+Qin).
3. **Nonparametric + General Alternative**
   Eg: Threshold the empirical Gaussian MMD (G-MMD^2)

$$\frac{1}{\binom{n}{2}} \sum_{i \neq j} e^{-\frac{\|x_i - x_j\|^2}{\gamma^2}} + \frac{1}{\binom{n}{2}} \sum_{i \neq j} e^{-\frac{\|y_i - y_j\|^2}{\gamma^2}} - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} e^{-\frac{\|x_i - y_j\|^2}{\gamma^2}}$$

   Eg: Euclidean Energy Dist. (E-ED): $e^{-\frac{\|x_i - x_j\|^2}{\gamma^2}} \to \|x_i - x_j\|$
   Fact: Population G-$\mathrm{MMD}^2 = 0$ iff $P = Q$ (also E-ED)

Q1: How do the latter tests perform in the former setting?

Q2: What is the role of bandwidth γ on test power?

The Quadratic-time and Linear-time Statistics

Define (k is the Gaussian kernel or negative Euclidean distance)

$$h := h(x, x', y, y') = k(x, x') + k(y, y') - k(x, y') - k(x', y)$$

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

$$\mathrm{MMD}_l^2 = \frac{1}{n/2} \sum_{i=1}^{n/2} h(x_{2i-1}, x_{2i}, y_{2i-1}, y_{2i})$$

It is easy to define a sub-quadratic time block-based variant that looks at n^{2x} entries for $0.5 \leq x \leq 1$.

Q3: What is the tradeoff of computation vs power for these tests?

How did we land here? Misconceptions!

Q4: How do G-MMD^2 and E-ED perform in high-dimensions?

1. Szekely[1]: "The power of dCor test for independence is very good especially for high dimensions p,q". No proof!
2. Gretton et al[2]: "Estimation of MMD^2 is independent of d". True, but...Power $\neq$ estimation error! MMD^2 is $1/poly(d)$!
3. Misleading Experiments[2,] - P and Q are mean-separated Gaussians with means $(0, 0, ..., 0)$ and $(1, 1, ..., 1)$.
4. Misleading Experiments[3] - Table 1[3] shows that unbiased dCov *estimates* dCov well (dCov hovers near 0 under independence).

[1] W'shop on Nonparam. Measures of Dependence (Columbia Univ, May'14)
[2] A Kernel Two Sample Test, Gretton et al. (JMLR'12)
[3] The dist. corr. t-test of indep. in high dim., Szekely et al (JMA'14)

Some Details...

Test based on MMD_l^2

Define $V = 2\mathrm{Var}[h]$. By CLT

$$\frac{\sqrt{n}(\mathrm{MMD}_l^2 - \mathrm{MMD}^2)}{\sqrt{V}} \rightsquigarrow N(0, 1)$$

Let v be the empirical counterpart of V. Define the test

$$\text{Reject when} \quad \sqrt{n} \frac{\mathrm{MMD}_l^2}{\sqrt{v}} > z_\alpha$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$ for standard Gaussian cdf Φ, i.e. $P(Z > z_\alpha) = \alpha$ for standard Gaussian random variable Z.

The type-1 error and power of this test changes with

1. Number of points n
2. Underlying dimensionality d of x, y
3. The signal-to-noise ratio $\Psi = \|\mathbb{E}_P[X] - \mathbb{E}_Q[X]\|^2 / \sigma^2$
4. The bandwidth of the Gaussian kernel γ

The "Classical" Power of MMD_l^2

If Pr denotes the probability under H_1, and Φ is the standard normal cdf, the power is

$$\begin{aligned}
& \Pr\left(\frac{\sqrt{n}\mathrm{MMD}_l^2}{\sqrt{v}} > z_\alpha\right) \\
= \; & \Pr\left(\frac{\sqrt{n}(\mathrm{MMD}_l^2 - \mathrm{MMD}^2)}{\sqrt{V}} > \sqrt{\frac{v}{V}} z_\alpha - \frac{\sqrt{n}\mathrm{MMD}^2}{\sqrt{V}}\right) \\
\xrightarrow{n \to \infty} \; & \Pr\left(Z > z_\alpha - \frac{\sqrt{n}\mathrm{MMD}^2}{\sqrt{V}}\right) \\
= \; & 1 - \Phi\left(z_\alpha - \frac{\sqrt{n}\mathrm{MMD}^2}{\sqrt{V}}\right) \\
= \; & \Phi\left(\frac{\sqrt{n}\mathrm{MMD}^2}{\sqrt{V}} - z_\alpha\right)
\end{aligned}$$

This behaves like $\Phi(\sqrt{n})$ since the population MMD^2 and V are constants that are both independent of n.

Challenges in the high dimensional setting

1. MMD^2 depends on dimension
2. V depends on dimension
3. $(n, d) \to \infty$ at any rate
4. Does $v/V \to 1$ even if $(n, d) \to \infty$?

We will address these by

1. Non-asymptotic, finite-sample Berry-Esseen theorem
2. Calculating MMD^2, V explicitly by Taylor expansions
3. Concentration bounds in terms of fourth moments

Answers!

Summary of Some Results

1. **Explicit characterization of power** as a function of $n, d, \Psi \approx KL(P, Q)$ in the high-dimensional setting as $(n, d) \to \infty$, for nonparametric P, Q differing in their means.
2. **A clear and smooth computation-statistics tradeoff** if computation scales as n^{2x} for $0.5 \leq x \leq 1$, then the power in the low SNR (low Ψ) regime is $\boxed{\approx \Phi(n^x \Psi^2 / \sqrt{d})}$.
3. **The power is independent of Gaussian kernel bandwidth**, as long as it is chosen large enough as $\Omega(\sqrt{d})$, which happens to be the choice made by the popular "median heuristic".
4. **Energy Distance & Gaussian MMD have the same power** in this setting with a mean-difference between distributions.
5. **Free Lunch!** ED and GMMD have the same power as specialized tests that have been designed in the literature to test for mean differences (like Chen+Qin, Srivastava+Du).

A Simulation

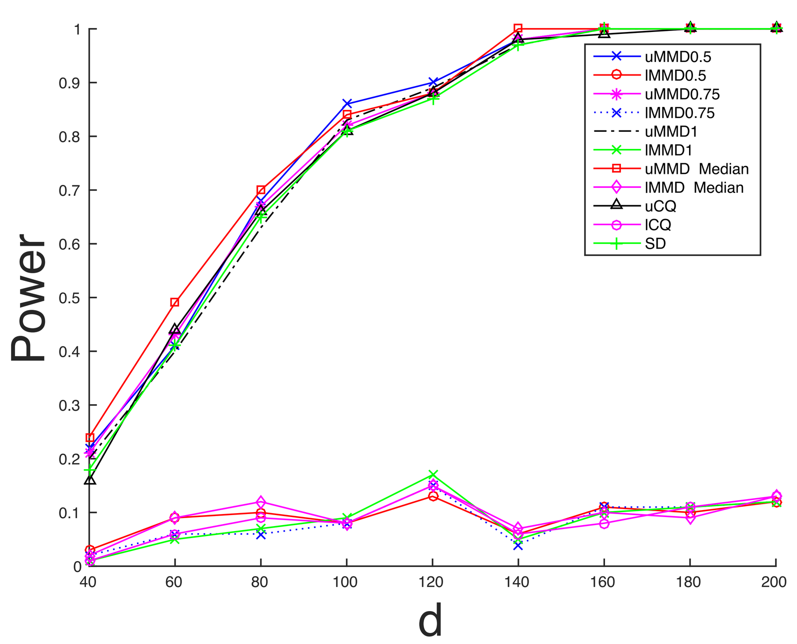


Figure : **Parameters:** P,Q Gaussians, $d = 40, 60, ..., 200$, $n = d$, $\Psi = 1$.
Top set: U-statistics (G-MMD for many bandwidths, E-ED, CQ, SD).
Bottom set: Linear-time statistics.

Summary of Some Techniques

1. G-$\mathrm{MMD}^2 \approx \|\delta\|^2 / \gamma^2$. Recall $\gamma = \Omega(\sqrt{d})$.
   Implication: This is why estimation error alone is misleading.
2. Variance V of test statistic also decays with d, but slower.
   Implication: This is why power decays with d.
3. Ratio of $v/V \to 1$ as $n \to \infty$ independent of how d grows.
   Implication: Studentization works fine.
4. (Linear) The right-hand side of the Berry Esseen lemma $10 \frac{\xi_3}{V^{3/2}\sqrt{n}}$ is actually $\leq 20/\sqrt{n}$, independent of d!
   Implication: Null, alternate distributions are always Gaussian.
5. (U-statistic) Martingale central limit theorem works out.
   Implication: Null, alternate distributions are Gaussian, *not* infinite sums of weighted chi-squared distributions (fixed d).