Introduction

Thought Experiment: Red vs Blue Pill

4-week Experiment: Each day, give Neo either R or B pill, measure real-valued thymine concentration [X vs Y, respectively].

Data: R: X1, . . . , Xn1 ∼ P (data) and B: Y1, . . . , Yn2 ∼ Q (data).

Mean-difference alternatives:


General alternatives:

H0: P = Q vs. H1: P ̸= 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 distinguish faces from 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: X1, . . . , Xn2 ∼ P in E300 and Y1, . . . , Yn2 ∼ Q in E300.

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 (H0), but the test returns 1. The type-1 error α is the probability of a false positive.

2. False Negative: When P ̸= Q (H1), but the test returns 0. The type-2 error β is the probability of a false negative.

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

High β implies low power: 1 − β – 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 → ∞.

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

Open Questions

1. Parametric + Mean-Difference Alternative
   Eg: Threshold Hotelling’s Statistic (X−Y)TΣ−1(X−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

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

1. Number of points n
2. Underlying dimensionality d1, d2
3. The signal-to-noise ratio Φ = E[|E[X]|]−E[|X|]/σ2
4. The bandwidth of the Gaussian kernel γ

The “Classical” Power of MMD2

If Pγ denotes the probability under H0 and Φ is the standard normal cdf, the power is

\[
P_1 \left( \frac{\sqrt{\text{MMD}}}{\sqrt{\text{MMD}}} \right) = 1 - \Phi \left( \frac{\sqrt{\text{MMD}}}\sqrt{\text{MMD}} \right)
\]

This behaves like Φ(\sqrt{\bar{\lambda}d}) since the population MMD2 and Φ are constants that are both independent of n.

Challenges in the high dimensional setting

1. MMD2 depends on dimension
2. V depends on dimension
3. (n, d) → ∞ at any rate
4. Does v/|V| → 1 even if (n, d) → ∞?

We will address these by

1. Non-asymptotic, finite-sample Berry-Esseen theorem
2. Calculating MMD2, V explicitly by Taylor expansions
3. Concentration bounds in terms of fourth moments

Some Details...

Test based on MMD2

Define V = 2Var(δ). By CLT

\[
\frac{\sqrt{n}MMD}{\sqrt{\text{MMD}}} \overset{d}{\to} N(0, 1)
\]

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

Reject when \(\sqrt{n}v_{\text{emp}} \geq z_α\)

where \(z_α = \Phi^{-1}(1-α)\) for standard Gaussian cdf, \(Φ, i.e. P(v > z_α) = α\) 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 d1, d2
3. The signal-to-noise ratio Φ = E[|E[X]|]−E[|X|]/σ2
4. The bandwidth of the Gaussian kernel γ

How did we land here? Misconceptions!

4. How do G-MMD2 and E-E perform in high-dimensions?

1. Sashank Reddi*, Aarti Singh, Barnabas Poczos, Larry Wasserman

Answers!

Summary of Some Results

1. Explicit characterization of power as a function of n, d, Φ = N(0, 1) in the high-dimensional setting as (n, d) → ∞, for nonparametric P, Q differing in their means.

2. A clear and smooth computability-statistics tradeoff if computation scales as nγ for 0.5 ≤ γ ≤ 1, then the power in the low SNR (low Φ) regime is 0.

3. The power is independent of Gaussian kernel bandwidth, as long as it is chosen large enough as Ω(\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 G-MMD have the same power as specialized tests that have been designed in the literature to test for mean differences (like Chen+Qin, Srivastava+Dua).

Summary of Some Techniques

1. G-MMD2 \(\approx \|\bar{X} - \bar{Y}\|^2\) Recall: \(\bar{X} \sim \Omega(\sqrt{d})\).
   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| → 1 as n → ∞ independent of how d grows. Implication: Studentization works fine.

4. \(\|\cdot\|\) The right-hand side of the Berry Esseen lemma \(10/|\text{MMD}|\) is actually ≤ 20/\sqrt{d}, independent of d! Implication: Null, alternate distributions are always Gaussian.


A Simulation

Figure: Parameters: P: Q Gaussian, d = 40, 60, . . . , 200, n = d, Φ = 1. Top set: U-statistics (G-MMD for many bandwidths, E-E, E-D, CQ, SD). Bottom set: Linear-time statistics.