Nonparametric Inference and Geometric Probability

(The Curious Case of Dimension 8)

Bhaswar B. Bhattacharya

University of Pennsylvania

November 14, 2022

Outline

The Goodness-of-Fit and the Two-Sample Problems

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a densiy f in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test

$$H_0 : f = f_0 \quad \text{versus} \quad H_1 : f \neq f_0,$$

where f_0 is some specified density in $\mathbb{R}^d$.

The Goodness-of-Fit and the Two-Sample Problems

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a densiy f in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test

$$H_0 : f = f_0 \quad \text{versus} \quad H_1 : f \neq f_0,$$

  where f_0 is some specified density in $\mathbb{R}^d$.

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_m\}$ and $\mathscr{Y}_n = \{Y_1, Y_2, \ldots, Y_n\}$ be i.i.d. samples from densities f and g in $\mathbb{R}^d$, respectively. The *two-sample problem* is to test

$$H_0 : f = g \quad \text{versus} \quad H_1 : f \neq g.$$

The Goodness-of-Fit and the Two-Sample Problems

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a densiy f in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test

$$H_0 : f = f_0 \quad \text{versus} \quad H_1 : f \neq f_0,$$

  where f_0 is some specified density in $\mathbb{R}^d$.
- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_m\}$ and $\mathscr{Y}_n = \{Y_1, Y_2, \ldots, Y_n\}$ be i.i.d. samples from densities f and g in $\mathbb{R}^d$, respectively. The *two-sample problem* is to test

$$H_0 : f = g \quad \text{versus} \quad H_1 : f \neq g.$$

- *Parametric Analogues*: Suppose $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ is a parametric family of distributions in $\mathbb{R}^d$, where $\Theta \subseteq \mathbb{R}^p$ is the parameter space.
  - *Goodness-of-fit problem*: For a specified value $\theta_0 \in \Theta$ consider

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

The Goodness-of-Fit and the Two-Sample Problems

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a densiy f in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test
$$H_0 : f = f_0 \quad \text{versus} \quad H_1 : f \neq f_0,$$
where f_0 is some specified density in $\mathbb{R}^d$.
- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_m\}$ and $\mathscr{Y}_n = \{Y_1, Y_2, \ldots, Y_n\}$ be i.i.d. samples from densities f and g in $\mathbb{R}^d$, respectively. The *two-sample problem* is to test
$$H_0 : f = g \quad \text{versus} \quad H_1 : f \neq g.$$
- *Parametric Analogues*: Suppose $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ is a parametric family of distributions in $\mathbb{R}^d$, where $\Theta \subseteq \mathbb{R}^p$ is the parameter space.
  - *Goodness-of-fit problem*: For a specified value $\theta_0 \in \Theta$ consider
  $$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$
  - *Two-sample problem*:
  $$H_0 : \theta_1 = \theta_2 \quad \text{versus} \quad H_1 : \theta_1 \neq \theta_2.$$

The Goodness-of-Fit and the Two-Sample Problems

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a densiy f in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test

$$H_0 : f = f_0 \quad \text{versus} \quad H_1 : f \neq f_0,$$

  where f_0 is some specified density in $\mathbb{R}^d$.

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_m\}$ and $\mathscr{Y}_n = \{Y_1, Y_2, \ldots, Y_n\}$ be i.i.d. samples from densities f and g in $\mathbb{R}^d$, respectively. The *two-sample problem* is to test

$$H_0 : f = g \quad \text{versus} \quad H_1 : f \neq g.$$

- *Parametric Analogues*: Suppose $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ is a parametric family of distributions in $\mathbb{R}^d$, where $\Theta \subseteq \mathbb{R}^p$ is the parameter space.
  - *Goodness-of-fit problem*: For a specified value $\theta_0 \in \Theta$ consider

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

  - *Two-sample problem*:

$$H_0 : \theta_1 = \theta_2 \quad \text{versus} \quad H_1 : \theta_1 \neq \theta_2.$$

- Throughout we will consider the asymptotic regime where $m, n \to \infty$, such that $\frac{m}{m+n} \to p \in (0, 1)$, and the dimension is fixed.

(Asymptotically) Distribution-Free Tests

- What are (asymptotically) distribution-free tests?

(Asymptotically) Distribution-Free Tests

- What are (asymptotically) distribution-free tests?
  - *Goodness-of-fit problem*: The (asymptotic) null distribution of the test statistic does not depend on null distribution f_0.
    - *Classical univariate tests*: Kolmogorov-Smirnov test.
    - We will discuss multivariate analogues: *Bickel-Brieman spacings test*.

(Asymptotically) Distribution-Free Tests

- What are (asymptotically) distribution-free tests?
  - *Goodness-of-fit problem*: The (asymptotic) null distribution of the test statistic does not depend on null distribution f_0.
    - *Classical univariate tests*: Kolmogorov-Smirnov test.
    - We will discuss multivariate analogues: *Bickel-Brieman spacings test*.
  - *Two-sample problem*: The (asymptotic) null distribution of the test statistic does not depend on the unknown null distribution $f = g$.
    - *Classical univariate tests*: Wald-Wolfowitz runs test, Mann-Whitney test.

(Asymptotically) Distribution-Free Tests

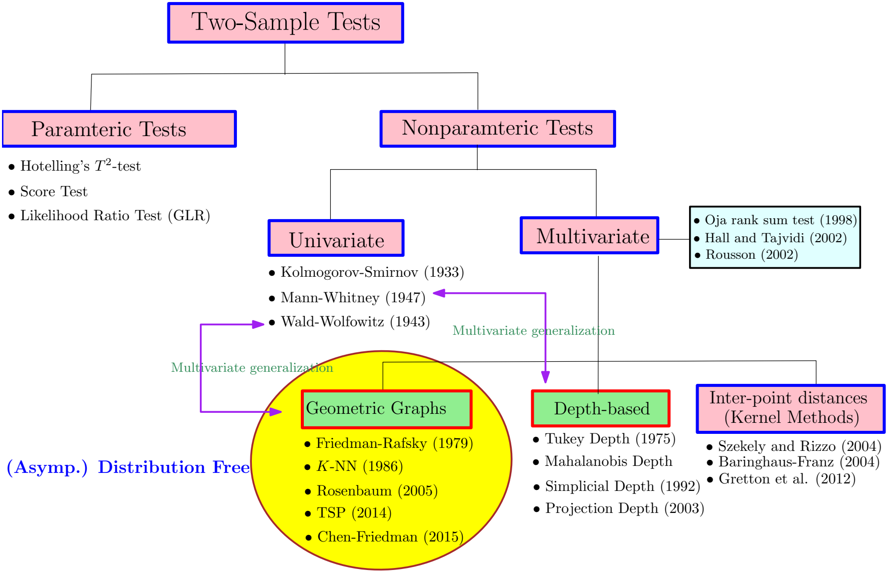
- What are (asymptotically) distribution-free tests?
  - *Goodness-of-fit problem*: The (asymptotic) null distribution of the test statistic does not depend on null distribution f_0.
    - *Classical univariate tests*: Kolmogorov-Smirnov test.
    - We will discuss multivariate analogues: *Bickel-Brieman spacings test*.
  - *Two-sample problem*: The (asymptotic) null distribution of the test statistic does not depend on the unknown null distribution $f = g$.
    - *Classical univariate tests*: Wald-Wolfowitz runs test, Mann-Whitney test.
    - We will discuss multivariate analogues: *Friedman-Rafsky test*, *nearest-neighbor based tests*, *cross-match test*, among others.

(Asymptotically) Distribution-Free Tests

- What are (asymptotically) distribution-free tests?
  - *Goodness-of-fit problem*: The (asymptotic) null distribution of the test statistic does not depend on null distribution f_0.
    - *Classical univariate tests*: Kolmogorov-Smirnov test.
    - We will discuss multivariate analogues: *Bickel-Brieman spacings test*.
  - *Two-sample problem*: The (asymptotic) null distribution of the test statistic does not depend on the unknown null distribution $f = g$.
    - *Classical univariate tests*: Wald-Wolfowitz runs test, Mann-Whitney test.
    - We will discuss multivariate analogues: *Friedman-Rafsky test*, *nearest-neighbor based tests*, *cross-match test*, among others.
- Most (if not all) distribution-free goodness-of-fit/two-sample tests are based on geometric graphs, like *nearest-neighbor graphs*, *minimum spanning trees*, *matchings*, etc.

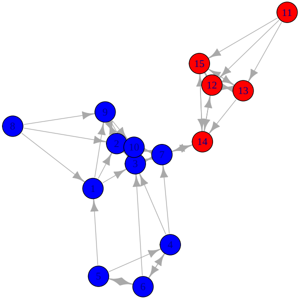
Outline

Two-Sample Tests: An Overview



Test Based on Nearest Neighbors Graphs

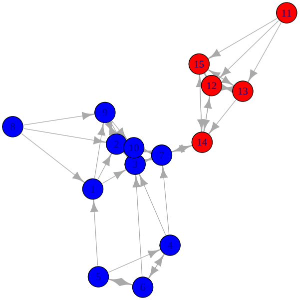
Bivariate normal data. Location shift. 3-NN graph.

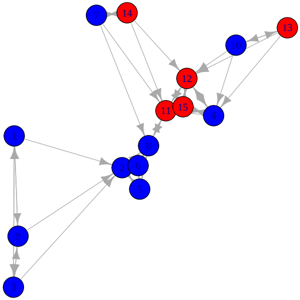


$$\Delta = 2, \quad T(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_m)) = 1.$$

Test Based on Nearest Neighbors Graphs

Bivariate normal data. Location shift. 3-NN graph.





$\Delta = 2, \quad T(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_m)) = 1.$ $\qquad \Delta = 0.05, \quad T(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_m)) = 7.$

Graph Based Two-Sample Tests

- Let $\mathscr{G}$ be a *graph functional* in $\mathbb{R}^d$.

Graph Based Two-Sample Tests

- Let $\mathscr{G}$ be a *graph functional* in $\mathbb{R}^d$. For any finite $S \subset \mathbb{R}^d$, $\mathscr{G}(S)$ is a graph with vertex set S.

Graph Based Two-Sample Tests

- Let $\mathscr{G}$ be a *graph functional* in $\mathbb{R}^d$. For any finite $S \subset \mathbb{R}^d$, $\mathscr{G}(S)$ is a graph with vertex set S.
- The *2-sample test statistic based on the graph functional $\mathscr{G}$* is defined as

$$\begin{aligned}\underbrace{T(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))}_{T(\mathscr{G})} &:= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{1}\{(X_i, Y_j) \in E(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))\} \\ &= \#\text{ edges across the two samples.}\end{aligned}$$

Graph Based Two-Sample Tests

- Let $\mathscr{G}$ be a *graph functional* in $\mathbb{R}^d$. For any finite $S \subset \mathbb{R}^d$, $\mathscr{G}(S)$ is a graph with vertex set S.
- The *2-sample test statistic based on the graph functional $\mathscr{G}$* is defined as

$$\underbrace{T(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))}_{T(\mathscr{G})} := \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{1}\{(X_i, Y_j) \in E(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))\}$$

$$= \# \text{ edges across the two samples.}$$

- Reject when $T(\mathscr{G})$ is small. Calibrate using asymptotic distribution.

Graph Based Two-Sample Tests

- Let $\mathscr{G}$ be a *graph functional* in $\mathbb{R}^d$. For any finite $S \subset \mathbb{R}^d$, $\mathscr{G}(S)$ is a graph with vertex set S.
- The *2-sample test statistic based on the graph functional* $\mathscr{G}$ is defined as

$$\underbrace{T(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))}_{T(\mathscr{G})} := \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{1}\{(X_i, Y_j) \in E(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))\}$$

$$= \# \text{ edges across the two samples.}$$

- Reject when $T(\mathscr{G})$ is small. Calibrate using asymptotic distribution. Reject when $\{T(\mathscr{G}) < C_{m,n}\}$, where $C_{m,n}$ is such that

$$\lim_{m,n\to\infty} \mathbb{P}_{H_0}(T(\mathscr{G}) < C_{m,n}) = \alpha.$$

Graph Based Two-Sample Tests

- Let $\mathscr{G}$ be a *graph functional* in $\mathbb{R}^d$. For any finite $S \subset \mathbb{R}^d$, $\mathscr{G}(S)$ is a graph with vertex set S.
- The *2-sample test statistic based on the graph functional $\mathscr{G}$* is defined as

$$\underbrace{T(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))}_{T(\mathscr{G})} := \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{1}\{(X_i, Y_j) \in E(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))\}$$

$$= \# \text{ edges across the two samples.}$$

- Reject when $T(\mathscr{G})$ is small. Calibrate using asymptotic distribution. Reject when $\{T(\mathscr{G}) < C_{m,n}\}$, where $C_{m,n}$ is such that

$$\lim_{m,n\to\infty} \mathbb{P}_{H_0}(T(\mathscr{G}) < C_{m,n}) = \alpha.$$

- In dimension 1: The *Wald-Wolfowitz runs test* (1940) counts the number of runs.

Graph Based Two-Sample Tests

- Let $\mathscr{G}$ be a *graph functional* in $\mathbb{R}^d$. For any finite $S \subset \mathbb{R}^d$, $\mathscr{G}(S)$ is a graph with vertex set S.
- The *2-sample test statistic based on the graph functional $\mathscr{G}$* is defined as

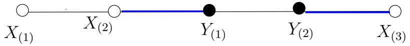
$$\underbrace{T(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))}_{T(\mathscr{G})} := \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{1}\{(X_i, Y_j) \in E(\mathscr{G}(\mathscr{X}_m \cup \mathscr{Y}_n))\}$$

$$= \# \text{ edges across the two samples}.$$

- Reject when $T(\mathscr{G})$ is small. Calibrate using asymptotic distribution. Reject when $\{T(\mathscr{G}) < C_{m,n}\}$, where $C_{m,n}$ is such that

$$\lim_{m,n\to\infty} \mathbb{P}_{H_0}(T(\mathscr{G}) < C_{m,n}) = \alpha.$$

- In dimension 1: The *Wald-Wolfowitz runs test* (1940) counts the number of runs. *This is a graph based test where $\mathscr{G} = P_N$ is the path.*



Friedman-Rafsky Test (1979)

Definition (Minimal Spanning Tree (MST))

- Given a finite set $S \subset \mathbb{R}^d$, a *spanning tree* of S is a connected graph with vertex-set S and no cycles.
- A *minimal spanning tree* (MST) of S, denoted by $\mathcal{T}(S)$, is a spanning tree with the smallest length, sum of Euclidean lengths of the edges.

Friedman-Rafsky Test (1979)

Definition (Minimal Spanning Tree (MST))

- Given a finite set $S \subset \mathbb{R}^d$, a *spanning tree* of S is a connected graph with vertex-set S and no cycles.
- A *minimal spanning tree* (MST) of S, denoted by $\mathcal{T}(S)$, is a spanning tree with the smallest length, sum of Euclidean lengths of the edges.

- The FR-test rejects H_0 for *small* values of

$$\frac{\sum_{i=1}^m \sum_{j=1}^n \mathbf{1}\{(X_i, Y_j) \in E(\mathcal{T}(\mathscr{X}_m \cup \mathscr{Y}_n))\}}{N-1}.$$

- *When two distributions are different, the number edges across samples 1 and 2 should be small.*

Friedman-Rafsky Test (1979)

Definition (Minimal Spanning Tree (MST))

- Given a finite set $S \subset \mathbb{R}^d$, a *spanning tree* of S is a connected graph with vertex-set S and no cycles.
- A *minimal spanning tree* (MST) of S, denoted by $\mathcal{T}(S)$, is a spanning tree with the smallest length, sum of Euclidean lengths of the edges.

- The FR-test rejects H_0 for *small* values of

$$\frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}\{(X_i, Y_j) \in E(\mathcal{T}(\mathscr{X}_m \cup \mathscr{Y}_n))\}}{N-1}.$$

- *When two distributions are different, the number edges across samples 1 and 2 should be small.* This is precisely the Wald-Wolfowitz runs test in $d = 1$.

Friedman-Rafsky Test (1979)

Definition (Minimal Spanning Tree (MST))

- Given a finite set $S \subset \mathbb{R}^d$, a *spanning tree* of S is a connected graph with vertex-set S and no cycles.
- A *minimal spanning tree* (MST) of S, denoted by $\mathcal{T}(S)$, is a spanning tree with the smallest length, sum of Euclidean lengths of the edges.

- The FR-test rejects H_0 for *small* values of

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{1}\{(X_i, Y_j) \in E(\mathcal{T}(\mathscr{X}_m \cup \mathscr{Y}_n))\}}{N-1}.$$

- *When two distributions are different, the number edges across samples 1 and 2 should be small.* This is precisely the Wald-Wolfowitz runs test in $d = 1$.
- Other geometric graphs are often used:
  - *K-NN Test*: $\mathscr{G}$ is the K-nearest neighbor graph (Henze (1988), Schilling (1989)).
  - *Cross Match Test*: $\mathscr{G}$ is the minimum non-bipartite matching (Rosenbaum (2005)).

Properties of Tests on Geometric Graphs

- Asymptotic normality under the null of the centered statistic

$$\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) := \sqrt{N}\left(\frac{T(\mathscr{G}(\mathcal{Z}_N))}{|E(\mathscr{G}(\mathcal{Z}_N))|} - \frac{mn}{N(N-1)}\right) \xrightarrow{D} N(0, \sigma^2_{\mathscr{G}}).$$

as $N := m + n \to \infty$ such that $\frac{m}{N} \to p \in (0, 1)$.

Properties of Tests on Geometric Graphs

- Asymptotic normality under the null of the centered statistic

$$\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) := \sqrt{N}\left(\frac{T(\mathscr{G}(\mathcal{Z}_N))}{|E(\mathscr{G}(\mathcal{Z}_N))|} - \frac{mn}{N(N-1)}\right) \xrightarrow{D} N(0, \sigma^2_{\mathscr{G}}).$$

  as $N := m + n \to \infty$ such that $\frac{m}{N} \to p \in (0, 1)$.
- *Asymptotically distribution free*: $\sigma^2_{\mathscr{G}}$ does not depend on f.

Properties of Tests on Geometric Graphs

- Asymptotic normality under the null of the centered statistic

$$\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) := \sqrt{N}\left(\frac{T(\mathscr{G}(\mathcal{Z}_N))}{|E(\mathscr{G}(\mathcal{Z}_N))|} - \frac{mn}{N(N-1)}\right) \xrightarrow{D} N(0, \sigma^2_{\mathscr{G}}).$$

  as $N := m + n \to \infty$ such that $\frac{m}{N} \to p \in (0,1)$.
- *Asymptotically distribution free*: $\sigma^2_{\mathscr{G}}$ does not depend on f.
- The level α test based on $\mathscr{G}$ has rejection region

$$\{\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) < -z_\alpha \sigma_{\mathscr{G}}\},$$

  where z_α is the $(1-\alpha)$-th quantile of the standard normal.

Properties of Tests on Geometric Graphs

- Asymptotic normality under the null of the centered statistic

$$\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) := \sqrt{N}\left(\frac{T(\mathscr{G}(\mathcal{Z}_N))}{|E(\mathscr{G}(\mathcal{Z}_N))|} - \frac{mn}{N(N-1)}\right) \xrightarrow{D} N(0, \sigma^2_{\mathscr{G}}).$$

  as $N := m + n \to \infty$ such that $\frac{m}{N} \to p \in (0,1)$.
- *Asymptotically distribution free*: $\sigma^2_{\mathscr{G}}$ does not depend on f.
- The level α test based on $\mathscr{G}$ has rejection region

$$\{\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) < -z_\alpha \sigma_{\mathscr{G}}\},$$

  where z_α is the $(1-\alpha)$-th quantile of the standard normal.
- *Consistent against fixed alternatives*. Power goes to 1 whenever the two distributions differ on a set of positive measure (for parametric models, when $\theta_1 - \theta_2 = \Delta$, where $\Delta \neq 0$ is fixed).

Properties of Tests on Geometric Graphs

- Asymptotic normality under the null of the centered statistic

$$\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) := \sqrt{N}\left(\frac{T(\mathscr{G}(\mathcal{Z}_N))}{|E(\mathscr{G}(\mathcal{Z}_N))|} - \frac{mn}{N(N-1)}\right) \xrightarrow{D} N(0, \sigma_{\mathscr{G}}^2).$$

  as $N := m + n \to \infty$ such that $\frac{m}{N} \to p \in (0,1)$.
- *Asymptotically distribution free*: $\sigma_{\mathscr{G}}^2$ does not depend on f.
- The level α test based on $\mathscr{G}$ has rejection region

$$\{\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) < -z_\alpha \sigma_{\mathscr{G}}\},$$

  where z_α is the $(1-\alpha)$-th quantile of the standard normal.
- *Consistent against fixed alternatives*. Power goes to 1 whenever the two distributions differ on a set of positive measure (for parametric models, when $\theta_1 - \theta_2 = \Delta$, where $\Delta \neq 0$ is fixed).
- *How can we compare these tests*? Power against local alternatives. Asymptotic (Pitman) efficiency.
  - $H_0 : \theta_2 - \theta_1 = 0, \quad$ versus $\quad H_1 : \theta_2 - \theta_1 = \frac{h}{\sqrt{N}}$, for $h \in \mathbb{R}^p$.

Properties of Tests on Geometric Graphs

- Asymptotic normality under the null of the centered statistic

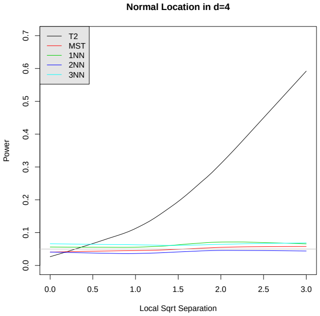
$$\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) := \sqrt{N}\left(\frac{T(\mathscr{G}(\mathcal{Z}_N))}{|E(\mathscr{G}(\mathcal{Z}_N))|} - \frac{mn}{N(N-1)}\right) \xrightarrow{D} N(0, \sigma^2_{\mathscr{G}}).$$

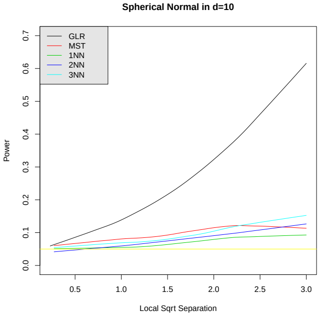
  as $N := m + n \to \infty$ such that $\frac{m}{N} \to p \in (0, 1)$.
- *Asymptotically distribution free*: $\sigma^2_{\mathscr{G}}$ does not depend on f.
- The level α test based on $\mathscr{G}$ has rejection region

$$\{\mathcal{R}(\mathscr{G}(\mathcal{Z}_N)) < -z_\alpha \sigma_{\mathscr{G}}\},$$

  where z_α is the $(1-\alpha)$-th quantile of the standard normal.
- *Consistent against fixed alternatives*. Power goes to 1 whenever the two distributions differ on a set of positive measure (for parametric models, when $\theta_1 - \theta_2 = \Delta$, where $\Delta \neq 0$ is fixed).
- *How can we compare these tests*? Power against local alternatives. Asymptotic (Pitman) efficiency.
  - $H_0 : \theta_2 - \theta_1 = 0, \quad$ versus $\quad H_1 : \theta_2 - \theta_1 = \frac{h}{\sqrt{N}}$, for $h \in \mathbb{R}^p$.
  - The performances of the different tests can be compared using these limiting power functions.

Asymptotic Efficiency of Graph-Based Tests





$$\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$$
$$H_0 : \theta_1 - \theta_2 = 0 \quad \text{vs} \quad H_1 : \theta_1 - \theta_2 = \frac{\delta \mathbf{1}}{\sqrt{N}}$$

$$\mathbb{P}_\sigma \sim N(0, \sigma^2 \mathrm{I})$$
$$H_0 : \sigma_1 - \sigma_2 = 0 \quad \text{vs} \quad H_1 : \sigma_1 - \sigma_2 = \frac{\delta}{\sqrt{N}}.$$

Asymptotic Efficiency of Graph-Based Tests

(Informal) Theorem (B. (2019))

The asymptotic efficiency of the two-sample test based on an undirected graph functional $\mathscr{G}$ is

$$\mathrm{AE}(\mathscr{G}) = \frac{|C(r) \int \langle h, \nabla f(z|\theta_0)\rangle \lambda(z) \mathrm{d}z|}{\sqrt{\{\gamma_0(1-r) + (\gamma_1 - 2)(1-2r)\}}},$$

Asymptotic Efficiency of Graph-Based Tests

(Informal) Theorem (B. (2019))

The asymptotic efficiency of the two-sample test based on an undirected graph functional $\mathscr{G}$ is

$$\mathrm{AE}(\mathscr{G}) = \frac{|C(r)\int \langle h, \nabla f(z|\theta_0)\rangle \lambda(z) \mathrm{d}z|}{\sqrt{\{\gamma_0(1-r)+(\gamma_1-2)(1-2r)\}}},$$

where

- $C(r)$ *is a constant that only depends on* $r := 2p(1-p)$,
- *for* $\mathcal{V}_N := \{V_1, V_2, \ldots, V_N\}$ *i.i.d. with density* $f(\cdot|\theta_0)$,

$$\frac{N}{|E(\mathscr{G}(\mathcal{V}_N))|} \xrightarrow{P} \gamma_0, \ \ \textit{and} \ \ \frac{N\overbrace{|T_2(\mathscr{G}(\mathcal{V}_N))|}^{2-stars}}{|E(\mathscr{G}(\mathcal{V}_N))|^2} \xrightarrow{P} \gamma_1.$$

Asymptotic Efficiency of Graph-Based Tests

(Informal) Theorem (B. (2019))

The asymptotic efficiency of the two-sample test based on an undirected graph functional $\mathscr{G}$ is

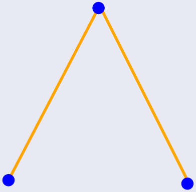
$$\mathrm{AE}(\mathscr{G}) = \frac{|C(r)\int\langle h, \nabla f(z|\theta_0)\rangle \lambda(z)\mathrm{d}z|}{\sqrt{\{\gamma_0(1-r)+(\gamma_1-2)(1-2r)\}}},$$

where

- *$C(r)$ is a constant that only depends on $r := 2p(1-p)$,*
- *for $\mathcal{V}_N := \{V_1, V_2, \ldots, V_N\}$ i.i.d. with density $f(\cdot|\theta_0)$,*

$$\frac{N}{|E(\mathscr{G}(\mathcal{V}_N))|} \xrightarrow{P} \gamma_0, \text{ and } \frac{N\overbrace{|T_2(\mathscr{G}(\mathcal{V}_N))|}^{2-stars}}{|E(\mathscr{G}(\mathcal{V}_N))|^2} \xrightarrow{P} \gamma_1.$$

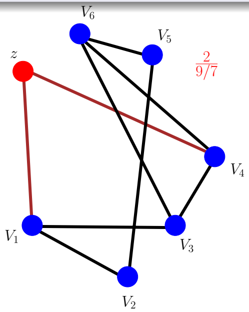
$K_{1,2}$



Asymptotic Efficiency of Graph-Based Tests

The function $\lambda(\cdot)$

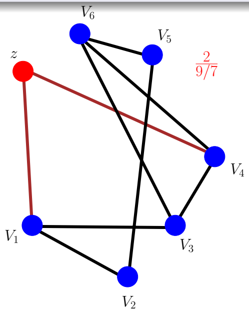
$$\lambda(z) = \lim_{N\to\infty} \mathbb{E}\frac{d(z, \mathscr{G}(\mathcal{V}_N^z))}{|E(\mathscr{G}(\mathcal{V}_N^z))|/N}$$
$$= \lim_{N\to\infty} \mathbb{E}\frac{\text{degree of vertex } z \text{ in } \mathscr{G}(\mathcal{V}_N \cup \{z\})}{\text{average degree of the graph}}.$$



Asymptotic Efficiency of Graph-Based Tests

The function $\lambda(\cdot)$

$$\lambda(z) = \lim_{N\to\infty} \mathbb{E}\frac{d(z, \mathscr{G}(\mathcal{V}_N^z))}{|E(\mathscr{G}(\mathcal{V}_N^z))|/N}$$
$$= \lim_{N\to\infty} \mathbb{E}\frac{\text{degree of vertex } z \text{ in } \mathscr{G}(\mathcal{V}_N \cup \{z\})}{\text{average degree of the graph}}.$$

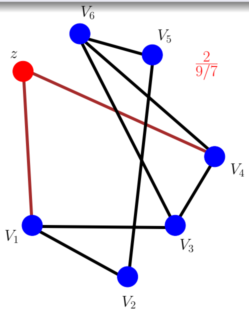


- *The function λ is like a 'centrality' measure. Small values of λ correspond to extreme points.*

Asymptotic Efficiency of Graph-Based Tests

The function $\lambda(\cdot)$

$$\begin{aligned}\lambda(z) &= \lim_{N\to\infty} \mathbb{E}\frac{d(z, \mathscr{G}(\mathcal{V}_N^z))}{|E(\mathscr{G}(\mathcal{V}_N^z))|/N} \\ &= \lim_{N\to\infty} \mathbb{E}\frac{\text{degree of vertex } z \text{ in } \mathscr{G}(\mathcal{V}_N \cup \{z\})}{\text{average degree of the graph}}.\end{aligned}$$



- *The function λ is like a 'centrality' measure. Small values of λ correspond to extreme points.*
- If $\mathscr{G} = MST$,

$$\frac{d(z, \mathscr{G}(\mathcal{V}_N^z))}{|E(\mathscr{G}(\mathcal{V}_N^z))|/N} \asymp d(z, \mathscr{G}(\mathcal{V}_N^z)).$$

Example: Friedman-Rafsy Test (MST)

- In this case, $\gamma_0 = 1$

Example: Friedman-Rafsy Test (MST)

- In this case, $\gamma_0 = 1$ and

$$\gamma_1 = \lim_{N\to\infty} \frac{T_2(\mathscr{G}(\mathcal{V}_N))}{N} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \binom{d(V_i, \mathscr{G}(\mathcal{V}_N))}{2} \xrightarrow{P} \frac{1}{2}\,\mathrm{Var}(D_d)+1.$$

Example: Friedman-Rafsy Test (MST)

- In this case, $\gamma_0 = 1$ and

$$\gamma_1 = \lim_{N\to\infty} \frac{T_2(\mathscr{G}(\mathcal{V}_N))}{N} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \binom{d(V_i, \mathscr{G}(\mathcal{V}_N))}{2} \xrightarrow{P} \frac{1}{2}\,\mathrm{Var}(D_d)+1.$$

- What is D_d?
  - Aldous and Steele (1992) defined the MSF for *infinite* point sets which are locally finite, using the Prim's algorithm.
  - Look at the MSF on a *Poisson process of rate* 1 *with point* 0 *added to it*. *D_d is the degree of the vertex 0 in this graph.*

Example: Friedman-Rafsy Test (MST)

- In this case, $\gamma_0 = 1$ and

$$\gamma_1 = \lim_{N\to\infty} \frac{T_2(\mathscr{G}(\mathcal{V}_N))}{N} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \binom{d(V_i, \mathscr{G}(\mathcal{V}_N))}{2} \xrightarrow{P} \frac{1}{2} \mathrm{Var}(D_d) + 1.$$

- Aldous and Steele (1992) showed that

$$\lambda(z) = \lim_{N\to\infty} \mathbb{E}(d(z, \mathscr{G}(\mathcal{Z}_N))) = 2,$$

is independent of z.

Example: Friedman-Rafsy Test (MST)

- In this case, $\gamma_0 = 1$ and

$$\gamma_1 = \lim_{N\to\infty} \frac{T_2(\mathscr{G}(\mathcal{V}_N))}{N} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \binom{d(V_i, \mathscr{G}(\mathcal{V}_N))}{2} \xrightarrow{P} \frac{1}{2}\,\mathrm{Var}(D_d)+1.$$

- Aldous and Steele (1992) showed that

$$\lambda(z) = \lim_{N\to\infty} \mathbb{E}(d(z, \mathscr{G}(\mathcal{Z}_N))) = 2,$$

is independent of z. Therefore, the numerator is

$$\int \langle h, \nabla f(z|\theta_0)\rangle \lambda(z)\mathrm{d}z = 0.$$

Example: Friedman-Rafsy Test (MST)

- In this case, $\gamma_0 = 1$ and

$$\gamma_1 = \lim_{N\to\infty} \frac{T_2(\mathscr{G}(\mathcal{V}_N))}{N} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \binom{d(V_i,\mathscr{G}(\mathcal{V}_N))}{2} \xrightarrow{P} \frac{1}{2}\,\mathrm{Var}(D_d)+1.$$

- Aldous and Steele (1992) showed that

$$\lambda(z) = \lim_{N\to\infty} \mathbb{E}(d(z,\mathscr{G}(\mathcal{Z}_N))) = 2,$$

is independent of z. Therefore, the numerator is

$$\int \langle h, \nabla f(z|\theta_0)\rangle \lambda(z)\mathrm{d}z = 0.$$

Theorem

The asymptotic (Pitman) efficiency of the test based on the MST is zero.

Stabilizing Graphs

- Convergence to the limiting Poisson graph.
- Local dependence.

Stabilizing Graphs

- Convergence to the limiting Poisson graph.
- Local dependence.

Definition (Penrose and Yukich (2003))

A translation and scale invariant graph functional $\mathscr{G}$ *stabilizes* on $\mathscr{P}_\lambda$ if there exists a random but almost surely finite variable R such that

$$E(0, \mathscr{G}(\mathscr{P}_{\lambda,0})) = E(0, \mathscr{G}(\mathscr{P}_{\lambda,0} \cap B(0,R) \cup \mathscr{A})),$$

for all finite $\mathscr{A} \subset \mathbb{R}^d \backslash B(0,R)$.

Stabilizing Graphs

- Convergence to the limiting Poisson graph.
- Local dependence.

Definition (Penrose and Yukich (2003))

A translation and scale invariant graph functional $\mathscr{G}$ *stabilizes* on $\mathscr{P}_\lambda$ if there exists a random but almost surely finite variable R such that

$$E(0, \mathscr{G}(\mathscr{P}_{\lambda,0})) = E(0, \mathscr{G}(\mathscr{P}_{\lambda,0} \cap B(0,R) \cup \mathscr{A})),$$

for all finite $\mathscr{A} \subset \mathbb{R}^d \backslash B(0,R)$.

- Includes MST, K-NN, Delaunay graphs, etc.

Efficiency of Tests Based on Stabilizing Graphs

Theorem (B. (2019))

Let $\mathscr{G}$ be any translation and scale invariant graph functional which stabilizing $\mathscr{P}_1$, such that

$$\underbrace{\frac{\overbrace{\Delta(\mathscr{G}(\mathcal{Z}_N))}^{\text{max degree}}}{\underbrace{|E(\mathscr{G}(\mathcal{Z}_N))|/N}_{\text{average degree}}} = O_P(1)}_{\text{normality condition}}, \quad \text{and} \quad \underbrace{\sup_{N \in \mathbb{N}} \mathbb{E}\left(d(Z_1, \mathscr{G}(\mathcal{Z}_N))^s\right) < \infty}_{\text{moment condition}},$$

for some $s > 4$. Then the asymptotic efficiency of the two-sample test based on $\mathscr{G}$ is zero.

Efficiency of Tests Based on Stabilizing Graphs

Theorem (B. (2019))

Let $\mathscr{G}$ be any translation and scale invariant graph functional which stabilizing $\mathscr{P}_1$, such that

$$\underbrace{\frac{\overbrace{\Delta(\mathscr{G}(\mathcal{Z}_N))}^{\text{max degree}}}{\underbrace{|E(\mathscr{G}(\mathcal{Z}_N))|/N}_{\text{average degree}}} = O_P(1)}_{\text{normality condition}}, \quad \text{and} \quad \underbrace{\sup_{N\in\mathbb{N}} \mathbb{E}\left(d(Z_1, \mathscr{G}(\mathcal{Z}_N))^s\right) < \infty}_{\text{moment condition}},$$

for some $s > 4$. Then the asymptotic efficiency of the two-sample test based on $\mathscr{G}$ is zero.

Corollary

The asymptotic efficiencies of the tests based on the MST or the K-NN graphs are zero.

What Next?

What Next?

- *How can we compare these tests*? For what sequence $\{\varepsilon_N\}_{N\geq 1}$ going to zero, can graph-based two-sample tests *detect* the hypothesis:

$$H_0 : \theta_2 - \theta_1 = 0, \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N.$$

What Next?

- *How can we compare these tests*? For what sequence $\{\varepsilon_N\}_{N\geq 1}$ going to zero, can graph-based two-sample tests *detect* the hypothesis:

$$H_0 : \theta_2 - \theta_1 = 0, \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N.$$

  - The above result shows $\varepsilon_N = \frac{h}{\sqrt{N}}$ is too hard: zero Pitman efficiency.

What Next?

- *How can we compare these tests*? For what sequence $\{\varepsilon_N\}_{N\geq 1}$ going to zero, can graph-based two-sample tests *detect* the hypothesis:

$$H_0 : \theta_2 - \theta_1 = 0, \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N.$$

  - The above result shows $\varepsilon_N = \frac{h}{\sqrt{N}}$ is too hard: zero Pitman efficiency.
  - *What is the detection threshold?* A sequence $a_N \to 0$, such that when

$$\begin{cases} ||\varepsilon_N|| \ll a_N & \text{the limiting power of the test is less than } \alpha, \\ & \end{cases}$$

What Next?

- *How can we compare these tests*? For what sequence $\{\varepsilon_N\}_{N\geq 1}$ going to zero, can graph-based two-sample tests *detect* the hypothesis:

$$H_0 : \theta_2 - \theta_1 = 0, \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N.$$

  - The above result shows $\varepsilon_N = \frac{h}{\sqrt{N}}$ is too hard: zero Pitman efficiency.
  - *What is the detection threshold?* A sequence $a_N \to 0$, such that when

$$\begin{cases} ||\varepsilon_N|| \ll a_N & \text{the limiting power of the test is less than } \alpha, \\ ||\varepsilon_N|| \gg a_N & \text{the limiting power of the test is 1.} \end{cases}$$

A Heuristic Calculation

- Consider the hypothesis

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N,$$

such that $||\varepsilon_N|| \to 0$.

A Heuristic Calculation

- Consider the hypothesis

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N,$$

  such that $||\varepsilon_N|| \to 0$.

- *Guessing the detection threshold*:

$$\begin{aligned} & N^{-\frac{1}{2}} \left\{ T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N))) \right\} \\ =& N^{-\frac{1}{2}} \left\{ T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N))) \right\} + N^{-\frac{1}{2}} \left\{ \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N))) \right. \\ =& T_1 + T_2. \end{aligned}$$

A Heuristic Calculation

- Consider the hypothesis

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N,$$

  such that $||\varepsilon_N|| \to 0$.
- *Guessing the detection threshold*:

$$\begin{aligned} & N^{-\frac{1}{2}} \left\{ T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N))) \right\} \\ =& N^{-\frac{1}{2}} \left\{ T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N))) \right\} + N^{-\frac{1}{2}} \left\{ \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N))) \right. \\ =& T_1 + T_2. \end{aligned}$$

(*CLT under alternative*) Under H_1, $T_1 \xrightarrow{D} N(0, \sigma^2(\theta_1, \theta_2, p))$?

A Heuristic Calculation

- Consider the hypothesis

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N,$$

such that $||\varepsilon_N|| \to 0$.

- *Guessing the detection threshold*:

$$\begin{aligned} & N^{-\frac{1}{2}} \left\{ T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N))) \right\} \\ =& N^{-\frac{1}{2}} \left\{ T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N))) \right\} + N^{-\frac{1}{2}} \left\{ \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N))) \right. \\ =& T_1 + T_2. \end{aligned}$$

(*CLT under alternative*) Under H_1, $T_1 \xrightarrow{D} N(0, \sigma^2(\theta_1, \theta_2, p))$?
(*Mean difference*) Derive the limit of T_2, when $\theta_2 - \theta_1 = \varepsilon_N \to 0$.

A Heuristic Calculation

- Consider the hypothesis

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N,$$

  such that $||\varepsilon_N|| \to 0$.
- *Guessing the detection threshold*:

$$\begin{aligned}
&N^{-\frac{1}{2}}\left\{T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N)))\right\}\\
=&N^{-\frac{1}{2}}\left\{T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N)))\right\} + N^{-\frac{1}{2}}\left\{\mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N)))\right.\\
=&T_1 + T_2.
\end{aligned}$$

(*CLT under alternative*) Under H_1, $T_1 \xrightarrow{D} N(0, \sigma^2(\theta_1, \theta_2, p))$?
(*Mean difference*) Derive the limit of T_2, when $\theta_2 - \theta_1 = \varepsilon_N \to 0$. If $\theta_2 - \theta_1 = \frac{h}{\sqrt{N}}$,

$$\begin{aligned}
T_2 = N^{-\frac{1}{2}}(\underbrace{\delta_N(\theta_1, \theta_2, p)}_{\mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N))} - \underbrace{\delta_N(\theta_1, \theta_1, p)}_{\mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N))}) &\approx N^{-\frac{1}{2}}(\langle\theta_2 - \theta_1, \nabla\delta_N(\theta_1, \theta_1, p)\rangle)\\
&\approx \frac{1}{N}(\langle h, \nabla\delta_N(\theta_1, \theta_1, p)\rangle) \to 0.
\end{aligned}$$

A Heuristic Calculation

- Consider the hypothesis

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N,$$

  such that $||\varepsilon_N|| \to 0$.
- *Guessing the detection threshold*:

$$\begin{aligned} & N^{-\frac{1}{2}} \left\{T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N)))\right\} \\ =& N^{-\frac{1}{2}} \left\{T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N)))\right\} + N^{-\frac{1}{2}} \left\{\mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N)) - \mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N)))\right. \\ =& T_1 + T_2. \end{aligned}$$

(*CLT under alternative*) Under H_1, $T_1 \xrightarrow{D} N(0, \sigma^2(\theta_1, \theta_2, p))$?
(*Mean difference*) Derive the limit of T_2, when $\theta_2 - \theta_1 = \varepsilon_N \to 0$.
If $\theta_2 - \theta_1 = \frac{h}{N^{1/4}}$,

$$N^{-\frac{1}{2}} \Big(\underbrace{\delta_N(\theta_1, \theta_2, p)}_{\mathbb{E}_{H_1}(T(\mathscr{G}(\mathcal{Z}_N)))} - \underbrace{\delta_N(\theta_1, \theta_1, p)}_{\mathbb{E}_{H_0}(T(\mathscr{G}(\mathcal{Z}_N)))} \Big) \approx N^{-\frac{1}{2}} \left(\langle(\theta_2 - \theta_1), \mathrm{H}\delta_N(\theta_1, \theta_1, p)(\theta_2 - \theta_1)\rangle\right)$$

$$= \frac{1}{N} \left(\langle h, \mathrm{H}\delta_N(\theta_1, \theta_1, p) h\rangle\right).$$

CLT Under Alternative: The K-NN Graph

(Informal) Theorem (B. (2020))

For the two-sample test based on the directed K-NN graph functional $\mathcal{N}_K$, in the Poissonized setting,

$$N^{-\frac{1}{2}}\left\{T(\mathcal{N}_K(\mathcal{Z}'_N)) - \mathbb{E}_{H_1}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\right\} \xrightarrow{D} N(0, \sigma_K^2(f, g, p)).$$

- Proved for the Wald's run test ($d = 1$) by Lehmann (1953).

CLT Under Alternative: The K-NN Graph

(Informal) Theorem (B. (2020))

For the two-sample test based on the directed K-NN graph functional $\mathcal{N}_K$, in the Poissonized setting,

$$N^{-\frac{1}{2}}\left\{T(\mathcal{N}_K(\mathcal{Z}'_N)) - \mathbb{E}_{H_1}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\right\} \xrightarrow{D} N(0, \sigma_K^2(f, g, p)).$$

- Proved for the Wald's run test ($d = 1$) by Lehmann (1953).
- It is also known that (Henze and Penrose (1999))

$$\frac{1}{N}\mathbb{E}_{H_1}T(\mathscr{G}(\mathcal{Z}'_N)) \to \underbrace{\int \frac{pqf(x)g(x)}{(pf(x)+qg(x))}\mathrm{d}x}$$

CLT Under Alternative: The K-NN Graph

(Informal) Theorem (B. (2020))

For the two-sample test based on the directed K-NN graph functional $\mathcal{N}_K$, in the Poissonized setting,

$$N^{-\frac{1}{2}}\left\{T(\mathcal{N}_K(\mathcal{Z}'_N)) - \mathbb{E}_{H_1}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\right\} \xrightarrow{D} N(0, \sigma_K^2(f, g, p)).$$

- Proved for the Wald's run test ($d = 1$) by Lehmann (1953).
- It is also known that (Henze and Penrose (1999))

$$\frac{1}{N}\mathbb{E}_{H_1} T(\mathscr{G}(\mathcal{Z}'_N)) \to \underbrace{\int \frac{pqf(x)g(x)}{(pf(x)+qg(x))}\mathrm{d}x}_{\delta(f,g,p)}.$$

CLT Under Alternative: The K-NN Graph

(Informal) Theorem (B. (2020))

For the two-sample test based on the directed K-NN graph functional $\mathcal{N}_K$, in the Poissonized setting,

$$N^{-\frac{1}{2}}\left\{T(\mathcal{N}_K(\mathcal{Z}_N')) - \mathbb{E}_{H_1}(T(\mathcal{N}_K(\mathcal{Z}_N')))\right\} \xrightarrow{D} N(0, \sigma_K^2(f,g,p)).$$

- Proved for the Wald's run test ($d=1$) by Lehmann (1953).
- It is also known that (Henze and Penrose (1999))

$$\frac{1}{N}\mathbb{E}_{H_1}T(\mathscr{G}(\mathcal{Z}_N')) \to \underbrace{\int \frac{pqf(x)g(x)}{(pf(x)+qg(x))}\mathrm{d}x}_{\delta(f,g,p)}.$$

- Can we say

$$N^{\frac{1}{2}}\left\{\frac{1}{N}T(\mathcal{N}_K(\mathcal{Z}_N')) - \delta(f,g,p)\right\} \xrightarrow{D} N(0, \sigma_K^2(f,g,p))?$$

  - Need to show $\sqrt{N}\left(\frac{1}{N}\mathbb{E}T(\mathcal{N}_K(\mathcal{Z}_N')) - \delta(f,g,p)\right) \to 0$?

CLT Under Alternative: The K-NN Graph

(Informal) Theorem (B. (2020))

For the two-sample test based on the directed K-NN graph functional $\mathcal{N}_K$, in the Poissonized setting,

$$N^{-\frac{1}{2}}\left\{T(\mathcal{N}_K(\mathcal{Z}'_N)) - \mathbb{E}_{H_1}(T(\mathcal{N}_K(\mathcal{Z}'_N)))\right\} \xrightarrow{D} N(0, \sigma_K^2(f, g, p)).$$

- Proved for the Wald's run test ($d = 1$) by Lehmann (1953).
- It is also known that (Henze and Penrose (1999))

$$\frac{1}{N}\mathbb{E}_{H_1} T(\mathscr{G}(\mathcal{Z}'_N)) \to \underbrace{\int \frac{pqf(x)g(x)}{(pf(x)+qg(x))}\mathrm{d}x}_{\delta(f,g,p)}.$$

- Can we say

$$N^{\frac{1}{2}}\left\{\frac{1}{N}T(\mathcal{N}_K(\mathcal{Z}'_N)) - \delta(f,g,p)\right\} \xrightarrow{D} N(0, \sigma_K^2(f, g, p))?$$

  - Need to show $\sqrt{N}\left(\frac{1}{N}\mathbb{E}T(\mathcal{N}_K(\mathcal{Z}'_N)) - \delta(f,g,p)\right) \to 0$?
  - In dimension 1, *R. Savage pointed out an issue in Lehmann's original proof.*

The Mean Difference

this manner. Since then it has been pointed out to me by R. Savage that when the limit result for

$$\left(\frac{W}{m} - E\left(\frac{W}{m}\right)\right) \Big/ \sigma\left(\frac{W}{m}\right)$$

we replace

$$E(W/m)$$

by

$$2\int_0^1 g'(x)/(\gamma + g'(x))\, dx,$$

the error is of the order

$$\sqrt{m}\left[E(W/m) - 2\int_0^1 g'(x)/(\gamma + g'(x))\, dx\right],$$

as is seen from (5.4). Thus (5.3) is not enough to guarantee the validity of this substitution. However, the numerical results obtained seemed sufficiently interesting to leave them in, in the hope that a proof of their validity will soon be forthcoming.

The Mean Difference

this manner. Since then it has been pointed out to me by R. Savage that when the limit result for

$$\left(\frac{W}{m} - E\left(\frac{W}{m}\right)\right) \Big/ \sigma\left(\frac{W}{m}\right)$$

we replace

$$E(W/m)$$

by

$$2\int_0^1 g'(x)/(\gamma + g'(x))\,dx,$$

the error is of the order

$$\sqrt{m}\left[E(W/m) - 2\int_0^1 g'(x)/(\gamma + g'(x))\,dx\right],$$

as is seen from (5.4). Thus (5.3) is not enough to guarantee the validity of this substitution. However, the numerical results obtained seemed sufficiently interesting to leave them in, in the hope that a proof of their validity will soon be forthcoming.

- For dimension 1, $\frac{1}{N}\mathbb{E}T(\mathcal{N}_K(\mathcal{Z}'_N)) - \delta(f,g,p) = o(1/\sqrt{N})$, *and the Lehmann claim can be easily validated.*

The Mean Difference

this manner. Since then it has been pointed out to me by R. Savage that when the limit result for

$$\left(\frac{W}{m} - E\left(\frac{W}{m}\right)\right) \Big/ \sigma\left(\frac{W}{m}\right)$$

we replace

$$E(W/m)$$

by

$$2\int_0^1 g'(x)/(\gamma + g'(x))\, dx,$$

the error is of the order

$$\sqrt{m}\left[E(W/m) - 2\int_0^1 g'(x)/(\gamma + g'(x))\, dx\right],$$

as is seen from (5.4). Thus (5.3) is not enough to guarantee the validity of this substitution. However, the numerical results obtained seemed sufficiently interesting to leave them in, in the hope that a proof of their validity will soon be forthcoming.

- For dimension 1, $\frac{1}{N}\mathbb{E}T(\mathcal{N}_K(\mathcal{Z}_N')) - \delta(f,g,p) = o(1/\sqrt{N})$, *and the Lehmann claim can be easily validated.*
- *Is this true for dimension d?*

The Mean Difference

this manner. Since then it has been pointed out to me by R. Savage that when the limit result for

$$\left(\frac{W}{m} - E\left(\frac{W}{m}\right)\right) \Big/ \sigma\left(\frac{W}{m}\right)$$

we replace

$$E(W/m)$$

by

$$2\int_0^1 g'(x)/(\gamma + g'(x))\, dx,$$

the error is of the order

$$\sqrt{m}\left[E(W/m) - 2\int_0^1 g'(x)/(\gamma + g'(x))\, dx\right],$$

as is seen from (5.4). Thus (5.3) is not enough to guarantee the validity of this substitution. However, the numerical results obtained seemed sufficiently interesting to leave them in, in the hope that a proof of their validity will soon be forthcoming.

- For dimension 1, $\frac{1}{N}\mathbb{E}T(\mathcal{N}_K(\mathcal{Z}'_N)) - \delta(f,g,p) = o(1/\sqrt{N})$, *and the Lehmann claim can be easily validated.*
- *Is this true for dimension d?* If yes, then the test will have power against $O(N^{-\frac{1}{4}})$ alternatives, and the heuristic would be correct.

The Mean Difference

this manner. Since then it has been pointed out to me by R. Savage that when the limit result for

$$\left(\frac{W}{m} - E\left(\frac{W}{m}\right)\right) \Big/ \sigma\left(\frac{W}{m}\right)$$

we replace

$$E(W/m)$$

by

$$2\int_0^1 g'(x)/(\gamma + g'(x))\ dx,$$

the error is of the order

$$\sqrt{m}\left[E(W/m) - 2\int_0^1 g'(x)/(\gamma + g'(x))\ dx\right],$$

as is seen from (5.4). Thus (5.3) is not enough to guarantee the validity of this substitution. However, the numerical results obtained seemed sufficiently interesting to leave them in, in the hope that a proof of their validity will soon be forthcoming.

- For dimension 1, $\frac{1}{N}\mathbb{E}T(\mathcal{N}_K(\mathcal{Z}'_N)) - \delta(f,g,p) = o(1/\sqrt{N})$, *and the Lehmann claim can be easily validated.*
- *Is this true for dimension d?* If yes, then the test will have power against $O(N^{-\frac{1}{4}})$ alternatives, and the heuristic would be correct. Otherwise, the rate of convergence competes with the Hessian term to determine the scaling for local power.

Case 1: Dimension Less or Equals 8

Theorem (B. (2020))

Suppose dimension $d \leq 8$. Then the limiting power of the directed K-NN test is given by

$$\begin{cases} \alpha & if \quad ||N^{\frac{1}{4}}\varepsilon_N|| \to 0, \\ \Phi\left(z_\alpha + c_{K,\theta_1}(h)\right) & if \quad N^{\frac{1}{4}}\varepsilon_N \to h, \\ 1 & if \quad ||N^{\frac{1}{4}}\varepsilon_N|| \to \infty. \end{cases}$$

Case 1: Dimension Less or Equals 8

Theorem (B. (2020))

Suppose dimension $d \leq 8$. Then the limiting power of the directed K-NN test is given by

$$\begin{cases} \alpha & if \quad ||N^{\frac{1}{4}}\varepsilon_N|| \to 0, \\ \Phi\left(z_\alpha + c_{K,\theta_1}(h)\right) & if \quad N^{\frac{1}{4}}\varepsilon_N \to h, \\ 1 & if \quad ||N^{\frac{1}{4}}\varepsilon_N|| \to \infty. \end{cases}$$

- The heuristic is correct: The detection threshold is at $O(N^{-\frac{1}{4}})$ and is driven by the Hessian term (*second-order efficiency*).

Case 1: Dimension Less or Equals 8

Theorem (B. (2020))

Suppose dimension $d \leq 8$. Then the limiting power of the directed K-NN test is given by

$$\begin{cases} \alpha & if \quad ||N^{\frac{1}{4}}\varepsilon_N|| \to 0, \\ \Phi\left(z_\alpha + c_{K,\theta_1}(h)\right) & if \quad N^{\frac{1}{4}}\varepsilon_N \to h, \\ 1 & if \quad ||N^{\frac{1}{4}}\varepsilon_N|| \to \infty. \end{cases}$$

- The heuristic is correct: The detection threshold is at $O(N^{-\frac{1}{4}})$ and is driven by the Hessian term (*second-order efficiency*).
- *What is $c_{K,\theta_1}(h)$?*

$$c_{K,\theta_1}(h) = \begin{cases} \frac{r^2 K}{2\sigma_K}\mathbb{E}\left[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)}\right]^2 & if \quad d \leq 7, \\ \frac{r^2 K}{2\sigma_K}\mathbb{E}\left[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)}\right]^2 + \underbrace{b_{K,\theta_1}(h)}_{\text{correction term}} & if \quad d = 8. \end{cases}$$

Case 1: Simulations

- For a *fixed direction $h \in \mathbb{R}^p$*, consider the hypothesis

$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b},$$

as b varies from $(0, 1)$.

Case 1: Simulations

- For a *fixed direction* $h \in \mathbb{R}^p$, consider the hypothesis

$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b},$$

as b varies from $(0, 1)$.

  - $b = 0$: Corresponds to fixed alternatives.

Case 1: Simulations

- For a *fixed direction $h \in \mathbb{R}^p$*, consider the hypothesis

$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b},$$

as b varies from $(0, 1)$.

  - $b = 0$: Corresponds to fixed alternatives.
  - $b = 0.5$: Parametric detection rate.
  - $b = 0.25$: Predicted rate of detection for the K-NN test.

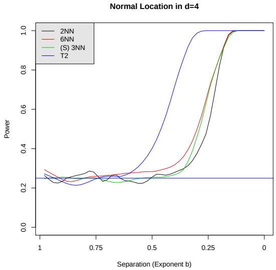
Case 1: Simulations

- For a *fixed direction* $h \in \mathbb{R}^p$, consider the hypothesis

$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b},$$

as b varies from $(0, 1)$.

  - $b = 0$: Corresponds to fixed alternatives.
  - $b = 0.5$: Parametric detection rate.
  - $b = 0.25$: Predicted rate of detection for the K-NN test.



$$\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$$

$$H_0 : \theta_1 - \theta_2 = 0 \quad \text{vs} \quad H_1 : \theta_1 - \theta_2 = \frac{\mathbf{1}}{N^b}$$

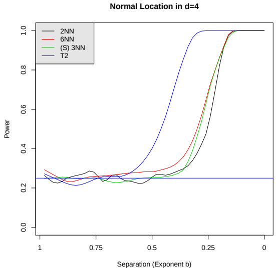
Case 1: Simulations

- For a *fixed direction* $h \in \mathbb{R}^p$, consider the hypothesis

$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b},$$

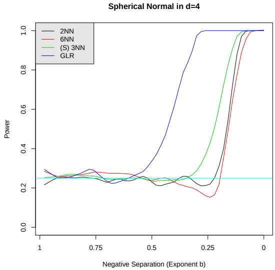
as b varies from $(0, 1)$.

  - $b = 0$: Corresponds to fixed alternatives.
  - $b = 0.5$: Parametric detection rate.
  - $b = 0.25$: Predicted rate of detection for the K-NN test.



$\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$

$H_0 : \theta_1 - \theta_2 = 0 \quad \text{vs} \quad H_1 : \theta_1 - \theta_2 = \frac{\mathbf{1}}{N^b}$



$\mathbb{P}_\sigma \sim N(0, \sigma^2 \mathrm{I})$

$H_0 : \sigma_1 - \sigma_2 = 0 \quad \text{vs} \quad H_1 : \sigma_1 - \sigma_2 = \frac{2}{N^b}.$

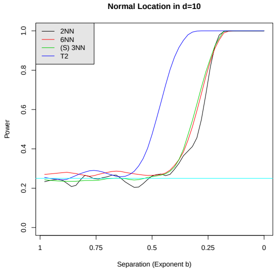
Case 2: Dimension Greater Than 8

- Consider dimension $d = 10$. For a fixed direction $h \in \mathbb{R}^p$, consider the hypothesis, as b varies from $(0, 1)$,
$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b}.$$

Case 2: Dimension Greater Than 8

- Consider dimension $d = 10$. For a fixed direction $h \in \mathbb{R}^p$, consider the hypothesis, as b varies from $(0, 1)$,

$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b}.$$



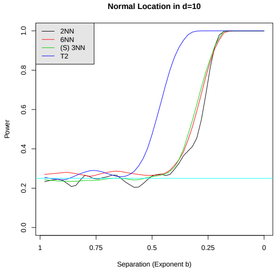
$$\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$$
$$H_0 : \theta_1 - \theta_2 = 0 \quad \text{vs} \quad H_1 : \theta_1 - \theta_2 = \frac{\mathbf{1}}{N^b}$$

Threshold still around at $b = 0.25$.

Case 2: Dimension Greater Than 8

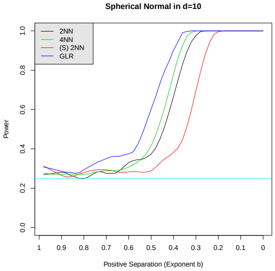
- Consider dimension $d = 10$. For a fixed direction $h \in \mathbb{R}^p$, consider the hypothesis, as b varies from $(0, 1)$,

$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b}.$$



$$\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$$
$$H_0 : \theta_1 - \theta_2 = 0 \quad \text{vs} \quad H_1 : \theta_1 - \theta_2 = \frac{1}{N^b}$$

Threshold still around at $b = 0.25$.



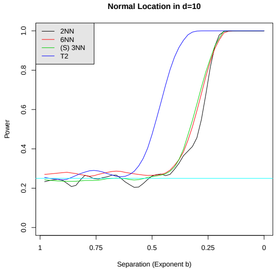
$$\mathbb{P}_\sigma \sim N(0, \sigma^2 \mathrm{I})$$
$$H_0 : \sigma_1 - \sigma_2 = 0 \quad \text{vs} \quad H_1 : \sigma_1 - \sigma_2 = \frac{2}{N^b}.$$

Threshold moves closer to $b = 0.5$.

Case 2: Dimension Greater Than 8

- Consider dimension $d = 10$. For a fixed direction $h \in \mathbb{R}^p$, consider the hypothesis, as b varies from $(0, 1)$,

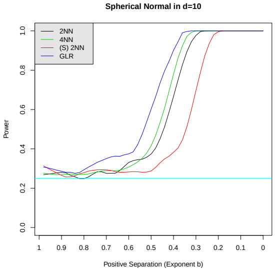
$$H_0 : \theta_2 = \theta_1 \quad \text{versus} \quad H_1 : \theta_2 = \theta_1 + \frac{h}{N^b}.$$



$\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$

$H_0 : \theta_1 - \theta_2 = 0 \quad \text{vs} \quad H_1 : \theta_1 - \theta_2 = \frac{\mathbf{1}}{N^b}$

Threshold still around at $b = 0.25$.



$\mathbb{P}_\sigma \sim N(0, \sigma^2 \mathrm{I})$

$H_0 : \sigma_1 - \sigma_2 = 0 \quad \text{vs} \quad H_1 : \sigma_1 - \sigma_2 = \frac{2}{N^b}$.

Threshold moves closer to $b = 0.5$.

- The detection threshold might not be universal: Depends on the distribution of the data and the sign of the alternative.

Case 2: Dimension Greater Than 8

Theorem (Continued) (B. (2020))

Suppose dimension $d \geq 9$. Then the limiting power of the K-NN test is given by

$$\begin{cases} \alpha & if \quad ||N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N|| \to 0, \\ \Phi\left(z_\alpha + b_{K,\theta_1}(h)\right) & if \quad N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N \to h, \\ \end{cases}$$

Case 2: Dimension Greater Than 8

Theorem (Continued) (B. (2020))

Suppose dimension $d \geq 9$. Then the limiting power of the K-NN test is given by

$$\begin{cases} \alpha & if & ||N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N|| \to 0, \\ \Phi\left(z_\alpha + b_{K,\theta_1}(h)\right) & if & N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N \to h, \\ 1/0 & if & ||N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N|| \to \infty \text{ and } ||N^{\frac{2}{d}}\varepsilon_N|| \to 0, \\ 1 & if & ||N^{\frac{2}{d}}\varepsilon_N|| \to \infty, \end{cases}$$

Case 2: Dimension Greater Than 8

Theorem (Continued) (B. (2020))

Suppose dimension $d \geq 9$. Then the limiting power of the K-NN test is given by

$$\begin{cases} \alpha & if \quad \|N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N\| \to 0, \\ \Phi\left(z_\alpha + b_{K,\theta_1}(h)\right) & if \quad N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N \to h, \\ 1/0 & if \quad \|N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N\| \to \infty \text{ and } \|N^{\frac{2}{d}}\varepsilon_N\| \to 0, \\ 1 & if \quad \|N^{\frac{2}{d}}\varepsilon_N\| \to \infty, \end{cases}$$

where $b_{K,\theta_1}(h) := -\lambda(p,K)\int h^\top \nabla_{\theta_1}\left(\frac{\mathrm{tr}(\mathrm{H}_x f(x|\theta_1))}{f(x|\theta_1)}\right) f^{\frac{d-2}{d}}(x|\theta_1)\mathrm{d}x.$

Case 2: Dimension Greater Than 8

Theorem (Continued) (B. (2020))

Suppose dimension $d \geq 9$. Then the limiting power of the K-NN test is given by

$$\begin{cases} \alpha & if \quad ||N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N|| \to 0, \\ \Phi\left(z_\alpha + b_{K,\theta_1}(h)\right) & if \quad N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N \to h, \\ 1/0 & if \quad ||N^{\frac{1}{2}-\frac{2}{d}}\varepsilon_N|| \to \infty \text{ and } ||N^{\frac{2}{d}}\varepsilon_N|| \to 0, \\ 1 & if \quad ||N^{\frac{2}{d}}\varepsilon_N|| \to \infty, \end{cases}$$

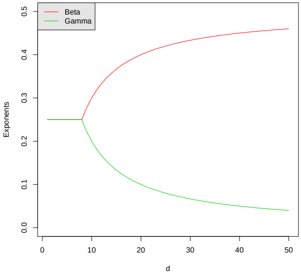
where $b_{K,\theta_1}(h) := -\lambda(p,K)\int h^\top \nabla_{\theta_1}\left(\frac{\mathrm{tr}(\mathrm{H}_x f(x|\theta_1))}{f(x|\theta_1)}\right) f^{\frac{d-2}{d}}(x|\theta_1)\mathrm{d}x.$

- *The heuristic is incorrect for dimensions greater than 8*: The detection threshold is driven by the rate of convergence of the gradient term.
- *The detection threshold might not be universal*: Depends on the distribution of the data and the sign of the alternative.

Summarizing the Result: Critical Exponents

In general, there are two *critical exponents*,

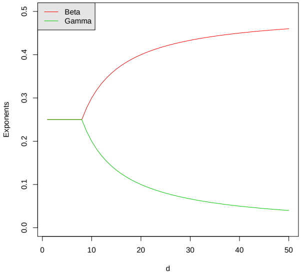
$$\beta_d = \begin{cases} \frac{1}{4} \\ \frac{1}{2} - \frac{2}{d}, \end{cases} \quad \gamma_d = \begin{cases} \frac{1}{4} & \text{if} \quad d \leq 8 \\ \frac{2}{d} & \text{if} \quad d \geq 9. \end{cases}$$



Summarizing the Result: Critical Exponents

In general, there are two *critical exponents*,

$$\beta_d = \begin{cases} \frac{1}{4} \\ \frac{1}{2} - \frac{2}{d}, \end{cases} \quad \gamma_d = \begin{cases} \frac{1}{4} & \text{if} \quad d \leq 8 \\ \frac{2}{d} & \text{if} \quad d \geq 9. \end{cases}$$



Theorem (Restated) (B. (2020))

Consider testing $H_0 : \theta_2 - \theta_1 = 0$ versus $H_1 : \theta_2 - \theta_1 = \varepsilon_N$, based on the directed K-NN graph functional. Then

- *If $||N^{\beta_d} \varepsilon_N|| \to 0$, the limiting power of the test is α.*
- *If $||N^{\gamma_d} \varepsilon_N|| \to \infty$, the limiting power of the test is 1.*

Outline

1. Preliminaries

2. Two-Sample Tests Based on Geometric Graphs
   - Definitions and Properties
   - Asymptotic Efficiency of Graph-Based Tests
   - Detection Thresholds

3. More Examples
   - Goodness-of-Fit Tests Based on Geometric Graphs
   - Independence Tests Based on Geometric Graphs

Goodness-of-Fit Tests

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test

$$H_0 : F = F_0 \quad \text{versus} \quad H_1 : F \neq F_0.$$

Goodness-of-Fit Tests

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test

$$H_0 : F = F_0 \quad \text{versus} \quad H_1 : F \neq F_0.$$

- Well-known asymptotically distribution-free univariate tests:
  - *Chi-squared test*: Fix cells and compare the observed frequencies in each cell with the expected frequency under H_0.

Goodness-of-Fit Tests

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test

$$H_0 : F = F_0 \quad \text{versus} \quad H_1 : F \neq F_0.$$

- Well-known asymptotically distribution-free univariate tests:
  - *Chi-squared test*: Fix cells and compare the observed frequencies in each cell with the expected frequency under H_0.
  - *Spacings Test*: For each X_i, define its 1-*step spacing* as

$$D_i = F_0(X_{(i+1)}) - F_0(X_{(i)}).$$

  Consider tests based on $\{nD_i\}_{1 \leq i \leq n}$.

Goodness-of-Fit Tests

- Let $\mathscr{X}_m = \{X_1, X_2, \ldots, X_n\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^d$. The *goodness-of-fit problem* is to test

$$H_0 : F = F_0 \quad \text{versus} \quad H_1 : F \neq F_0.$$

- Well-known asymptotically distribution-free univariate tests:
  - *Chi-squared test*: Fix cells and compare the observed frequencies in each cell with the expected frequency under H_0.
  - *Spacings Test*: For each X_i, define its 1-*step spacing* as

$$D_i = F_0(X_{(i+1)}) - F_0(X_{(i)}).$$

  Consider tests based on $\{nD_i\}_{1 \leq i \leq n}$. For example, for a (known) function $u : [0, \infty) \to \mathbb{R}$, reject H_0 for large values of

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ u(nD_i) - \mathbb{E}_0[u(nD_i)] \right\} \right|.$$

  Common choices of functions are $u(x) = e^{-x}$ or $u(x) = \log x$. (Pyke (1965), Hall (1986))

The Multivariate Spacings Test

- *How to define multivariate spacings?*

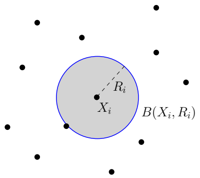
The Multivariate Spacings Test

- *How to define multivariate spacings?* Use "nearest-neighbor" balls (Bickel and Breiman (1983)).

- For each point X_i, define its *multivariate spacing* as

$$\mu_{F_0}(X_i) := F_0(B(X_i, R_i)) = \int_{B(X_i, R_i)} f_0(z)\mathrm{d}z,$$

where $B(x, r)$ is the ball of radius r around x, and R_i is the the nearest-neighbor distance from X_i.



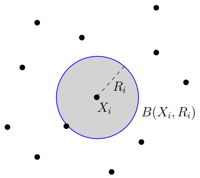
The Multivariate Spacings Test

- *How to define multivariate spacings?* Use "nearest-neighbor" balls (Bickel and Breiman (1983)).

- For each point X_i, define its *multivariate spacing* as

$$\mu_{F_0}(X_i) := F_0(B(X_i, R_i)) = \int_{B(X_i, R_i)} f_0(z)\mathrm{d}z,$$

  where $B(x, r)$ is the ball of radius r around x, and R_i is the the nearest-neighbor distance from X_i.



- *Multivariate Spacings Test*: For a (known) function $u : [0, \infty) \to \mathbb{R}$, reject H_0 for large values of

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))] \right\} \right|.$$

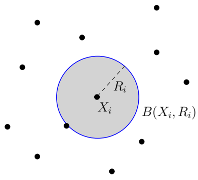
The Multivariate Spacings Test

- *How to define multivariate spacings?* Use "nearest-neighbor" balls (Bickel and Breiman (1983)).

- For each point X_i, define its *multivariate spacing* as

$$\mu_{F_0}(X_i) := F_0(B(X_i, R_i)) = \int_{B(X_i,R_i)} f_0(z)\mathrm{d}z,$$

where $B(x, r)$ is the ball of radius r around x, and R_i is the the nearest-neighbor distance from X_i.



- *Multivariate Spacings Test*: For a (known) function $u : [0,\infty) \to \mathbb{R}$, reject H_0 for large values of

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \right|.$$

- *The Bickel-Breiman Approximation*: Replace $\mu_{F_0}(X_i)$ by

$$D_i = \mathrm{Vol}(B(0,1))f_0(X_i)R_i^d.$$

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

where $\sigma^2(u)$ *does not depend on* F_0.

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

  where $\sigma^2(u)$ *does not depend on* F_0.

- The level α test has rejection region $\{|T_n(u)| > z_{\frac{\alpha}{2}}\sigma(u)\}$.

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

  where $\sigma^2(u)$ *does not depend on* F_0.
- The level α test has rejection region $\{|T_n(u)| > z_{\frac{\alpha}{2}}\sigma(u)\}$.
- *Consistent against fixed alternatives*. Power goes to 1 whenever $\theta \neq \theta_0$.

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

  where $\sigma^2(u)$ *does not depend on* F_0.
- The level α test has rejection region $\{|T_n(u)| > z_{\frac{\alpha}{2}}\sigma(u)\}$.
- *Consistent against fixed alternatives*. Power goes to 1 whenever $\theta \neq \theta_0$.
- *How can we compare these tests*? What sequence $\{\varepsilon_n\}_{n\geq 1}$ going to zero is detectable:

$$H_0 : \theta = \theta_0, \quad \text{versus} \quad H_1 : \theta = \theta_0 + \varepsilon_n.$$

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

  where $\sigma^2(u)$ *does not depend on* F_0.

- The level α test has rejection region $\{|T_n(u)| > z_{\frac{\alpha}{2}}\sigma(u)\}$.
- *Consistent against fixed alternatives*. Power goes to 1 whenever $\theta \neq \theta_0$.
- *How can we compare these tests*? What sequence $\{\varepsilon_n\}_{n\geq 1}$ going to zero is detectable:

$$H_0 : \theta = \theta_0, \quad \text{versus} \quad H_1 : \theta = \theta_0 + \varepsilon_n.$$

- As before, can be shown that $\varepsilon_n = \frac{h}{\sqrt{n}}$ is too hard: zero Pitman efficiency.

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

  where $\sigma^2(u)$ *does not depend on* F_0.
- The level α test has rejection region $\{|T_n(u)| > z_{\frac{\alpha}{2}} \sigma(u)\}$.
- *Consistent against fixed alternatives*. Power goes to 1 whenever $\theta \neq \theta_0$.
- *How can we compare these tests*? What sequence $\{\varepsilon_n\}_{n \geq 1}$ going to zero is detectable:

$$H_0 : \theta = \theta_0, \quad \text{versus} \quad H_1 : \theta = \theta_0 + \varepsilon_n.$$

- As before, can be shown that $\varepsilon_n = \frac{h}{\sqrt{n}}$ is too hard: zero Pitman efficiency. *What is the detection threshold?*

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

  where $\sigma^2(u)$ *does not depend on* F_0.
- The level α test has rejection region $\{|T_n(u)| > z_{\frac{\alpha}{2}}\sigma(u)\}$.
- *Consistent against fixed alternatives*. Power goes to 1 whenever $\theta \neq \theta_0$.
- *How can we compare these tests*? What sequence $\{\varepsilon_n\}_{n\geq 1}$ going to zero is detectable:

$$H_0 : \theta = \theta_0, \quad \text{versus} \quad H_1 : \theta = \theta_0 + \varepsilon_n.$$

- As before, can be shown that $\varepsilon_n = \frac{h}{\sqrt{n}}$ is too hard: zero Pitman efficiency. *What is the detection threshold?*
  - *Dimension* 1: $O(N^{-\frac{1}{4}})$. (Holst and Rao (1981))

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

  where $\sigma^2(u)$ *does not depend on* F_0.
- The level α test has rejection region $\{|T_n(u)| > z_{\frac{\alpha}{2}}\sigma(u)\}$.
- *Consistent against fixed alternatives*. Power goes to 1 whenever $\theta \neq \theta_0$.
- *How can we compare these tests*? What sequence $\{\varepsilon_n\}_{n\geq 1}$ going to zero is detectable:

$$H_0 : \theta = \theta_0, \quad \text{versus} \quad H_1 : \theta = \theta_0 + \varepsilon_n.$$

- As before, can be shown that $\varepsilon_n = \frac{h}{\sqrt{n}}$ is too hard: zero Pitman efficiency. *What is the detection threshold?*
  - *Dimension* 1: $O(N^{-\frac{1}{4}})$. (Holst and Rao (1981))
  - *Higher dimensions*: $O(N^{-\frac{1}{4}})$, *if* $d < 8$. (Zhou and Rao (1993))

Properties of the Spacings Test

- *Asymptotically distribution free*: Under H_0,

$$T_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u(n\mu_{F_0}(X_i)) - \mathbb{E}_0[u(n\mu_{F_0}(X_i))]\} \xrightarrow{D} N(0, \sigma^2(u)),$$

  where $\sigma^2(u)$ *does not depend on* F_0.
- The level α test has rejection region $\{|T_n(u)| > z_{\frac{\alpha}{2}} \sigma(u)\}$.
- *Consistent against fixed alternatives*. Power goes to 1 whenever $\theta \neq \theta_0$.
- *How can we compare these tests*? What sequence $\{\varepsilon_n\}_{n \geq 1}$ going to zero is detectable:

$$H_0 : \theta = \theta_0, \quad \text{versus} \quad H_1 : \theta = \theta_0 + \varepsilon_n.$$

- As before, can be shown that $\varepsilon_n = \frac{h}{\sqrt{n}}$ is too hard: zero Pitman efficiency. *What is the detection threshold?*
  - *Dimension* 1: $O(N^{-\frac{1}{4}})$. (Holst and Rao (1981))
  - *Higher dimensions*: $O(N^{-\frac{1}{4}})$, *if* $d < 8$. (Zhou and Rao (1993))
  - *Dimension 8 or higher*: Threshold changes. *Curious case of dimension 8 appears again.* (B. (2022+))

Graph Based Independence Tests

- Suppose $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^{d_1+d_2}$.

Graph Based Independence Tests

- Suppose $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^{d_1+d_2}$. *Denote the marginal distributions of X_1 and Y_1 by F_1 and F_2, respectively.*

Graph Based Independence Tests

- Suppose $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^{d_1+d_2}$. *Denote the marginal distributions of X_1 and Y_1 by F_1 and F_2, respectively.*
- The *independence testing problem* is the following hypotheses:

$$H_0 : F = F_1 \otimes F_2 \quad \text{versus} \quad H_1 : F \neq F_1 \otimes F_2.$$

Graph Based Independence Tests

- Suppose $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^{d_1+d_2}$. *Denote the marginal distributions of X_1 and Y_1 by F_1 and F_2, respectively.*
- The *independence testing problem* is the following hypotheses:

$$H_0 : F = F_1 \otimes F_2 \quad \text{versus} \quad H_1 : F \neq F_1 \otimes F_2.$$

- *Chatterjee's Correlation Coefficient*: A measure of correlation between 2 variables in dimension 1. (Chatterjee (2020))
    - Based on the relative ranks of data points. Has several attractive properties.

Graph Based Independence Tests

- Suppose $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^{d_1+d_2}$. *Denote the marginal distributions of X_1 and Y_1 by F_1 and F_2, respectively.*
- The *independence testing problem* is the following hypotheses:

$$H_0 : F = F_1 \otimes F_2 \quad \text{versus} \quad H_1 : F \neq F_1 \otimes F_2.$$

- *Chatterjee's Correlation Coefficient*: A measure of correlation between 2 variables in dimension 1. (Chatterjee (2020))
  - Based on the relative ranks of data points. Has several attractive properties.
  - This has been generalized to higher-dimensions using geometric graphs (Deb, Ghosal, and Sen (2020)).

Graph Based Independence Tests

- Suppose $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^{d_1+d_2}$. *Denote the marginal distributions of X_1 and Y_1 by F_1 and F_2, respectively.*
- The *independence testing problem* is the following hypotheses:

$$H_0 : F = F_1 \otimes F_2 \quad \text{versus} \quad H_1 : F \neq F_1 \otimes F_2.$$

- *Chatterjee's Correlation Coefficient*: A measure of correlation between 2 variables in dimension 1. (Chatterjee (2020))
    - Based on the relative ranks of data points. Has several attractive properties.
    - This has been generalized to higher-dimensions using geometric graphs (Deb, Ghosal, and Sen (2020)).
    - Has zero Pitman efficiency in dimension 1. (Shi, Drton, and Han (2022))

Graph Based Independence Tests

- Suppose $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^{d_1+d_2}$. *Denote the marginal distributions of X_1 and Y_1 by F_1 and F_2, respectively.*
- The *independence testing problem* is the following hypotheses:

$$H_0 : F = F_1 \otimes F_2 \quad \text{versus} \quad H_1 : F \neq F_1 \otimes F_2.$$

- *Chatterjee's Correlation Coefficient*: A measure of correlation between 2 variables in dimension 1. (Chatterjee (2020))
  - Based on the relative ranks of data points. Has several attractive properties.
  - This has been generalized to higher-dimensions using geometric graphs (Deb, Ghosal, and Sen (2020)).
  - Has zero Pitman efficiency in dimension 1. (Shi, Drton, and Han (2022))
- *What is the detection threshold?*
  - *Dimension* 1: $O(N^{-\frac{1}{4}})$. (Auddy, Deb, and Nandy (2021))

Graph Based Independence Tests

- Suppose $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be i.i.d. samples from a distribution F in $\mathbb{R}^{d_1+d_2}$. *Denote the marginal distributions of X_1 and Y_1 by F_1 and F_2, respectively.*
- The *independence testing problem* is the following hypotheses:

$$H_0 : F = F_1 \otimes F_2 \quad \text{versus} \quad H_1 : F \neq F_1 \otimes F_2.$$

- *Chatterjee's Correlation Coefficient*: A measure of correlation between 2 variables in dimension 1. (Chatterjee (2020))
  - Based on the relative ranks of data points. Has several attractive properties.
  - This has been generalized to higher-dimensions using geometric graphs (Deb, Ghosal, and Sen (2020)).
  - Has zero Pitman efficiency in dimension 1. (Shi, Drton, and Han (2022))
- *What is the detection threshold?*
  - *Dimension* 1: $O(N^{-\frac{1}{4}})$. (Auddy, Deb, and Nandy (2021))
  - *Curious case of dimension 8 is expected to appear in higher dimensions.*

Summary

- Random geometric graphs are important tools for constructing non-parametric tests.

Summary

- Random geometric graphs are important tools for constructing non-parametric tests.
- These tests are usually *distribution-free* but are often *Pitman inefficient*.

Summary

- Random geometric graphs are important tools for constructing non-parametric tests.
- These tests are usually *distribution-free* but are often *Pitman inefficient*.
- Rate of detection at $O(N^{-\frac{1}{4}})$ for dimension up to 8 (*second-order efficiency*).

Summary

- Random geometric graphs are important tools for constructing non-parametric tests.
- These tests are usually *distribution-free* but are often *Pitman inefficient*.
- Rate of detection at $O(N^{-\frac{1}{4}})$ for dimension up to 8 (*second-order efficiency*).
- *The rate of detection has a curious phase transition in dimension 8.*

Summary

- Random geometric graphs are important tools for constructing non-parametric tests.
- These tests are usually *distribution-free* but are often *Pitman inefficient*.
- Rate of detection at $O(N^{-\frac{1}{4}})$ for dimension up to 8 (*second-order efficiency*).
- *The rate of detection has a curious phase transition in dimension 8.*
- The uniform spanning forest changes its geometry at dimension 8 (Pemantle (1991), Benjamini, Kesten, Peres, and Schramm (2003)).

Summary

- Random geometric graphs are important tools for constructing non-parametric tests.
- These tests are usually *distribution-free* but are often *Pitman inefficient*.
- Rate of detection at $O(N^{-\frac{1}{4}})$ for dimension up to 8 (*second-order efficiency*).
- *The rate of detection has a curious phase transition in dimension 8.*
- The uniform spanning forest changes its geometry at dimension 8 (Pemantle (1991), Benjamini, Kesten, Peres, and Schramm (2003)).
- *Connections?*

Thank

You

Parsing the Theorem: The Good and the Bad

(Fix an alternative direction $h \in \mathbb{R}^p$ and suppose $\varepsilon_N = \delta_N h$, such that $\delta_N \to 0$.)

-

Parsing the Theorem: The Good and the Bad

(Fix an alternative direction $h \in \mathbb{R}^p$ and suppose $\varepsilon_N = \delta_N h$, such that $\delta_N \to 0$.)

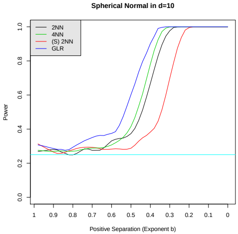
- $b_{K,\theta_1}(h) > 0$:

Parsing the Theorem: The Good and the Bad

(Fix an alternative direction $h \in \mathbb{R}^p$ and suppose $\varepsilon_N = \delta_N h$, such that $\delta_N \to 0$.)

- $b_{K,\theta_1}(h) > 0$: *These are the 'good' directions.* Detection threshold improves with dimension. *Blessing of dimensionality.*

$$\begin{cases} \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\left(z_\alpha + \lambda b_{K,\theta_1}(h)\right) > \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 1 & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty. \end{cases}$$



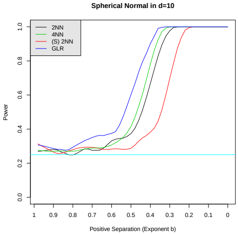
$$H_1 : \sigma_1 - \sigma_2 = \frac{2}{N^b}.$$

Parsing the Theorem: The Good and the Bad

(Fix an alternative direction $h \in \mathbb{R}^p$ and suppose $\varepsilon_N = \delta_N h$, such that $\delta_N \to 0$.)

- $b_{K,\theta_1}(h) > 0$: *These are the 'good' directions.* Detection threshold improves with dimension. *Blessing of dimensionality.*

$$\begin{cases} \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\left(z_\alpha + \lambda b_{K,\theta_1}(h)\right) > \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 1 & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty. \end{cases}$$



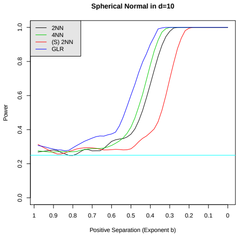
$$H_1 : \sigma_1 - \sigma_2 = \frac{2}{N^b}.$$

Parsing the Theorem: The Good and the Bad

(Fix an alternative direction $h \in \mathbb{R}^p$ and suppose $\varepsilon_N = \delta_N h$, such that $\delta_N \to 0$.)

- $b_{K,\theta_1}(h) > 0$: *These are the 'good' directions.* Detection threshold improves with dimension. *Blessing of dimensionality.*

$$\begin{cases} \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\left(z_\alpha + \lambda b_{K,\theta_1}(h)\right) > \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 1 & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty. \end{cases}$$



$H_1 : \sigma_1 - \sigma_2 = \frac{2}{N^b}$.

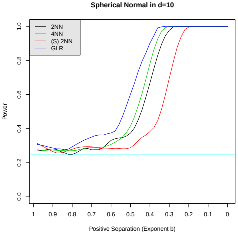
- $b_{K,\theta_1}(h) < 0$:

Parsing the Theorem: The Good and the Bad

(Fix an alternative direction $h \in \mathbb{R}^p$ and suppose $\varepsilon_N = \delta_N h$, such that $\delta_N \to 0$.)

- $b_{K,\theta_1}(h) > 0$: *These are the 'good' directions.* Detection threshold improves with dimension. *Blessing of dimensionality.*

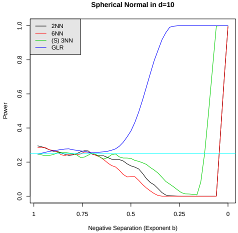
$$\begin{cases} \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\left(z_\alpha + \lambda b_{K,\theta_1}(h)\right) > \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 1 & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty. \end{cases}$$



$H_1 : \sigma_1 - \sigma_2 = \frac{2}{N^b}$.

- $b_{K,\theta_1}(h) < 0$: *These are the 'bad' directions.* Detection threshold worsens with dimension.

$$\begin{cases} \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\left(z_\alpha + \lambda b_{K,\theta_1}(h)\right) < \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 0 & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty \text{ and } N^{\frac{2}{d}}\delta_N \to 0, \\ 1 & N^{\frac{2}{d}}\delta_N \to \infty. \end{cases}$$



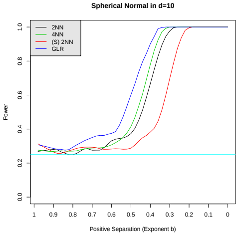
$H_1 : \sigma_1 - \sigma_2 = -\frac{2}{N^b}$.

Parsing the Theorem: The Good and the Bad

(Fix an alternative direction $h \in \mathbb{R}^p$ and suppose $\varepsilon_N = \delta_N h$, such that $\delta_N \to 0$.)

- $b_{K,\theta_1}(h) > 0$: *These are the 'good' directions.* Detection threshold improves with dimension. *Blessing of dimensionality.*

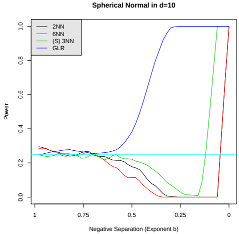
$$\begin{cases} \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\left(z_\alpha + \lambda b_{K,\theta_1}(h)\right) > \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 1 & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty. \end{cases}$$



$$H_1 : \sigma_1 - \sigma_2 = \frac{2}{N^b}.$$

- $b_{K,\theta_1}(h) < 0$: *These are the 'bad' directions.* Detection threshold worsens with dimension. *Non-monotonicity of power.*

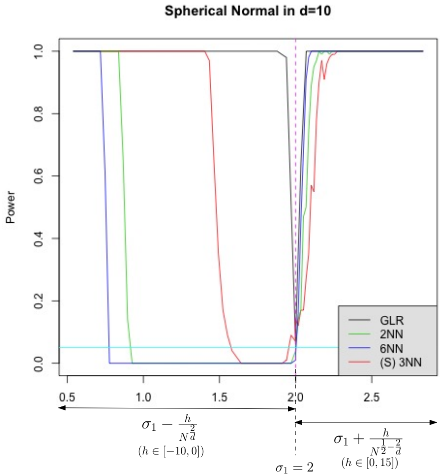
$$\begin{cases} \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to 0, \\ \Phi\left(z_\alpha + \lambda b_{K,\theta_1}(h)\right) < \alpha & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \lambda > 0, \\ 0 & N^{\frac{1}{2}-\frac{2}{d}}\delta_N \to \infty \text{ and } N^{\frac{2}{d}}\delta_N \to 0, \\ 1 & N^{\frac{2}{d}}\delta_N \to \infty. \end{cases}$$



$$H_1 : \sigma_1 - \sigma_2 = -\frac{2}{N^b}.$$

Zooming in at Thresholds: Spherical Normal

- We can zoom in at the two thresholds $O(N^{-\frac{1}{2}+\frac{2}{d}})$ and $O(N^{-\frac{2}{d}})$, and observe the phase transitions of the power function.



Degenerate Directions: Normal Location

- What about $b_{K,\theta}(h) = 0$? These are the “degenerate directions”.

Degenerate Directions: Normal Location

- What about $b_{K,\theta}(h) = 0$? These are the "degenerate directions".
- Consider the normal location family $\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$, and suppose

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N.$$

- Here, $b_{K,\theta}(h) = 0$, that is, all directions are degenerate.

Degenerate Directions: Normal Location

- What about $b_{K,\theta}(h) = 0$? These are the "degenerate directions".
- Consider the normal location family $\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$, and suppose

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N.$$

- Here, $b_{K,\theta}(h) = 0$, that is, all directions are degenerate.
- Direct calculations show that the two-sample test based on the directed K-NN graph satisfies
  - If $||N^{\frac{1}{4}}\varepsilon_N|| \to 0$, the limiting power of the test is α.
  - If $||N^{\frac{1}{4}}\varepsilon_N|| \to \infty$, the limiting power of the test is 1.

Degenerate Directions: Normal Location

- What about $b_{K,\theta}(h) = 0$? These are the "degenerate directions".
- Consider the normal location family $\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$, and suppose

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N.$$

- Here, $b_{K,\theta}(h) = 0$, that is, all directions are degenerate.
- Direct calculations show that the two-sample test based on the directed K-NN graph satisfies
  - If $||N^{\frac{1}{4}}\varepsilon_N|| \to 0$, the limiting power of the test is α.
  - If $||N^{\frac{1}{4}}\varepsilon_N|| \to \infty$, the limiting power of the test is 1.
  - If $\varepsilon_N = hN^{-\frac{1}{4}}$, for some $h \in \mathbb{R}^p \backslash \{\mathbf{0}\}$, the limiting power of the test is

$$\Phi\left(z_\alpha + \frac{r^2 K}{2\sigma_K}\mathbb{E}_{\mu_1}(h^\top X)^2\right),$$

where $V_d = |B(0,1)|$ is the volume of the unit ball in $\mathbb{R}^d$, and σ_K is the null variance.

Degenerate Directions: Normal Location

- What about $b_{K,\theta}(h) = 0$? These are the "degenerate directions".
- Consider the normal location family $\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$, and suppose

$$H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = \varepsilon_N .$$

- Here, $b_{K,\theta}(h) = 0$, that is, all directions are degenerate.
- Direct calculations show that the two-sample test based on the directed K-NN graph satisfies
  - If $||N^{\frac{1}{4}} \varepsilon_N|| \to 0$, the limiting power of the test is α.
  - If $||N^{\frac{1}{4}} \varepsilon_N|| \to \infty$, the limiting power of the test is 1.
  - If $\varepsilon_N = hN^{-\frac{1}{4}}$, for some $h \in \mathbb{R}^p \backslash \{\mathbf{0}\}$, the limiting power of the test is

$$\Phi\left(z_\alpha + \frac{r^2 K}{2\sigma_K} \mathbb{E}_{\mu_1} (h^\top X)^2 \right),$$

  where $V_d = |B(0,1)|$ is the volume of the unit ball in $\mathbb{R}^d$, and σ_K is the null variance.
- The rate is same across all dimensions (*second-order efficiency*).

Implications

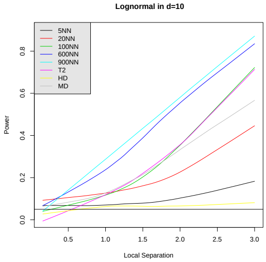
- *Which test should we use?*

Implications

- *Which test should we use? Can we get distribution-free tests with non-zero Pitman efficiency?*

Implications

- *Which test should we use? Can we get distribution-free tests with non-zero Pitman efficiency?*
  - Consider K-NN graphs with $K = K_N \to \infty$.
  - In this case, $O(N^{-\frac{1}{2}})$ detection thresholds (Pitman efficiency) can be attained, when K grows with N sufficiently fast.
  - *How fast is fast enough?* Trade-off with computation time.



$$\mathbb{P}_\theta \sim \exp(N(\theta, \mathrm{I}))$$
$$H_1 : \theta_1 - \theta_2 = \frac{\delta \cdot \mathbf{1}}{\sqrt{N}}.$$

Case 1: Dimension Less or Equals 8

Theorem (Zhou and Rao (1993), B. (2019+))

Suppose dimension $d \leq 8$. Then the limiting power of the multivariate spacings test, for a fixed function $u : [0, \infty) \to \mathbb{R}$, is given by

$$\begin{cases} \alpha & if \quad ||n^{\frac{1}{4}} \varepsilon_n|| \to 0, \\ \Phi\left(-z_{\frac{\alpha}{2}} + c_{\theta_0}(u,h)\right) + \Phi\left(-z_{\frac{\alpha}{2}} - c_{\theta_0}(u,h)\right) & if \quad ||n^{\frac{1}{4}} \varepsilon_n|| \to h, \\ 1 & if \quad ||n^{\frac{1}{4}} \varepsilon_n|| \to \infty. \end{cases}$$

Case 1: Dimension Less or Equals 8

Theorem (Zhou and Rao (1993), B. (2019+))

Suppose dimension $d \leq 8$. Then the limiting power of the multivariate spacings test, for a fixed function $u : [0, \infty) \to \mathbb{R}$, is given by

$$\begin{cases} \alpha & if \quad ||n^{\frac{1}{4}}\varepsilon_n|| \to 0, \\ \Phi\left(-z_{\frac{\alpha}{2}} + c_{\theta_0}(u,h)\right) + \Phi\left(-z_{\frac{\alpha}{2}} - c_{\theta_0}(u,h)\right) & if \quad ||n^{\frac{1}{4}}\varepsilon_n|| \to h, \\ 1 & if \quad ||n^{\frac{1}{4}}\varepsilon_n|| \to \infty. \end{cases}$$

- *What is $c_{\theta_0}(u, h)$?*

$$c_{\theta_0}(u,h) = \begin{cases} \frac{1}{\sigma(u)}\mathbb{E}\left[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)}\right]^2 \int_0^\infty e^{-t}(\frac{t^2}{2} - t)u'(t)\mathrm{d}t & if \quad d \leq 7, \\ \frac{1}{\sigma(u)}\mathbb{E}\left[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)}\right]^2 \int_0^\infty e^{-t}(\frac{t^2}{2} - t)u'(t)\mathrm{d}t + \underbrace{b_{\theta_0}(u,h)}_{\text{correction term}} & if \quad d = 8. \end{cases}$$

Case 1: Dimension Less or Equals 8

Theorem (Zhou and Rao (1993), B. (2019+))

Suppose dimension $d \leq 8$. Then the limiting power of the multivariate spacings test, for a fixed function $u : [0, \infty) \to \mathbb{R}$, is given by

$$\begin{cases} \alpha & if \quad ||n^{\frac{1}{4}} \varepsilon_n|| \to 0, \\ \Phi\left(-z_{\frac{\alpha}{2}} + c_{\theta_0}(u,h)\right) + \Phi\left(-z_{\frac{\alpha}{2}} - c_{\theta_0}(u,h)\right) & if \quad ||n^{\frac{1}{4}} \varepsilon_n|| \to h, \\ 1 & if \quad ||n^{\frac{1}{4}} \varepsilon_n|| \to \infty. \end{cases}$$

- *What is $c_{\theta_0}(u,h)$?*

$$c_{\theta_0}(u,h) = \begin{cases} \frac{1}{\sigma(u)} \mathbb{E}\left[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)}\right]^2 \int_0^\infty e^{-t}(\frac{t^2}{2} - t)u'(t)\mathrm{d}t & if \quad d \leq 7, \\ \frac{1}{\sigma(u)} \mathbb{E}\left[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)}\right]^2 \int_0^\infty e^{-t}(\frac{t^2}{2} - t)u'(t)\mathrm{d}t + \underbrace{b_{\theta_0}(u,h)}_{\textit{correction term}} & if \quad d = 8. \end{cases}$$

  - Can be optimized over u to obtain the "optimal" test among the class of tests $T_n(u)$:

Case 1: Dimension Less or Equals 8

Theorem (Zhou and Rao (1993), B. (2019+))

Suppose dimension $d \leq 8$. Then the limiting power of the multivariate spacings test, for a fixed function $u : [0, \infty) \to \mathbb{R}$, is given by

$$\begin{cases} \alpha & if \quad ||n^{\frac{1}{4}}\varepsilon_n|| \to 0, \\ \Phi\left(-z_{\frac{\alpha}{2}} + c_{\theta_0}(u,h)\right) + \Phi\left(-z_{\frac{\alpha}{2}} - c_{\theta_0}(u,h)\right) & if \quad ||n^{\frac{1}{4}}\varepsilon_n|| \to h, \\ 1 & if \quad ||n^{\frac{1}{4}}\varepsilon_n|| \to \infty. \end{cases}$$

- *What is $c_{\theta_0}(u, h)$?*

$$c_{\theta_0}(u,h) = \begin{cases} \frac{1}{\sigma(u)}\mathbb{E}\left[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)}\right]^2 \int_0^\infty e^{-t}(\frac{t^2}{2} - t)u'(t)\mathrm{d}t & if \quad d \leq 7, \\ \frac{1}{\sigma(u)}\mathbb{E}\left[\frac{h^\top \nabla_{\theta_1} f(X|\theta_1)}{f(X|\theta_1)}\right]^2 \int_0^\infty e^{-t}(\frac{t^2}{2} - t)u'(t)\mathrm{d}t + \underbrace{b_{\theta_0}(u,h)}_{\text{correction term}} & if \quad d = 8. \end{cases}$$

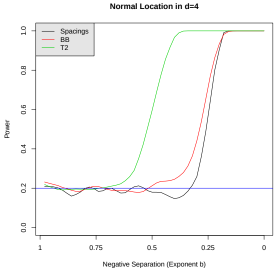
- Can be optimized over u to obtain the "optimal" test among the class of tests $T_n(u)$: *Global test for uniformity, irrespective of the alternative.*

Case 1: Simulations

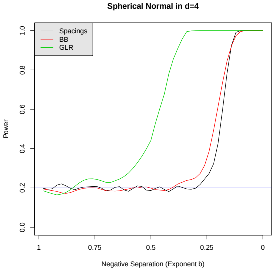
- For a fixed direction $h \in \mathbb{R}^p$, consider the hypothesis

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_0 + \frac{h}{N^b},$$

as b varies from $(0, 1)$.



$\mathbb{P}_\theta \sim N(\theta, \mathrm{I})$
$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta = -\frac{\mathbf{1}}{N^b}$



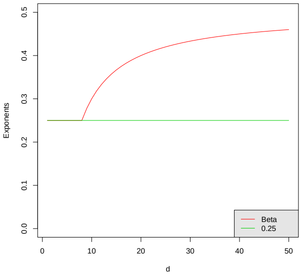
$\mathbb{P}_\sigma \sim N(0, \sigma^2 \mathrm{I})$
$H_0 : \sigma = 4 \quad \text{vs} \quad H_1 : \sigma = 4 - \frac{3}{N^b}.$

Case 2: Dimension Greater Than 8

Again, there are two *critical exponents*,

$$\beta_d = \begin{cases} \frac{1}{4} & \text{if} \quad d \leq 8 \\ \frac{1}{2} - \frac{2}{d} & \text{if} \quad d \geq 9, \end{cases}$$

and the *constant* $\frac{1}{4}$.

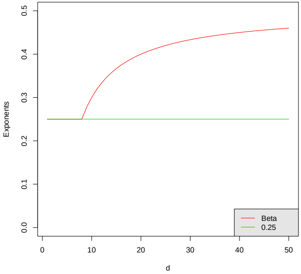


Case 2: Dimension Greater Than 8

Again, there are two *critical exponents*,

$$\beta_d = \begin{cases} \frac{1}{4} & \text{if} \quad d \leq 8 \\ \frac{1}{2} - \frac{2}{d} & \text{if} \quad d \geq 9, \end{cases}$$

and the *constant* $\frac{1}{4}$.



Theorem (B. (2019+))

The limiting power of the multivariate spacings test, for a fixed function $u : [0, \infty) \to \mathbb{R}$*, satisfies:*

- *If* $||N^{\beta_d} \varepsilon_N|| \to 0$*, the limiting power of the test is* α*.*
- *If* $||N^{\frac{1}{4}} \varepsilon_N|| \to \infty$*, the limiting power of the test is* 1*.*