Selection Detection and Two-Sample-Testing: Generalized Greenwood Statistics and their Applications

Đan Daniel Erdmann-Pham, Jonathan Terhorst & Yun S. Song

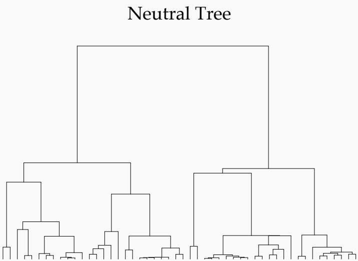
University of California, Berkeley

July 9, 2019

SPA 2019

Two Problems

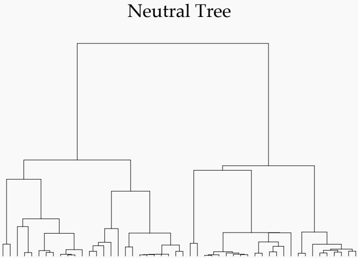
Population Genetics: Detecting Selective Pressure



- At each depth, leaf set sizes are approximately equidistributed
- Leaf set sizes are highly unbalanced close to the root

- Given a tree, how can we tell whether it was generated under selection or not?
- Data allows computation of sum of squares of leaf set sizes

Population Genetics: Detecting Selective Pressure

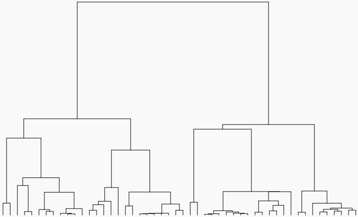


- At each depth, leaf set sizes are approximately equidistributed
- Leaf set sizes are highly unbalanced close to the root

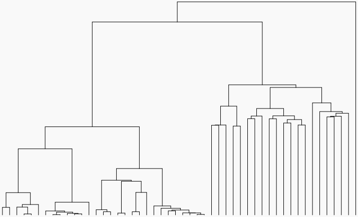
- Given a tree, how can we tell whether it was generated under selection or not?
- Data allows computation of sum of squares of leaf set sizes

Population Genetics: Detecting Selective Pressure

Neutral Tree



Tree with Selection

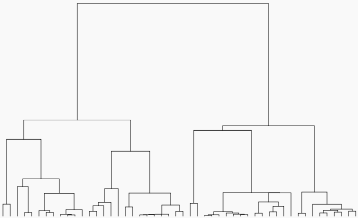


- At each depth, leaf set sizes are approximately equidistributed
- Leaf set sizes are highly unbalanced close to the root

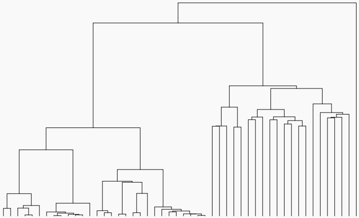
- Given a tree, how can we tell whether it was generated under selection or not?
- Data allows computation of sum of squares of leaf set sizes

Population Genetics: Detecting Selective Pressure

Neutral Tree



Tree with Selection



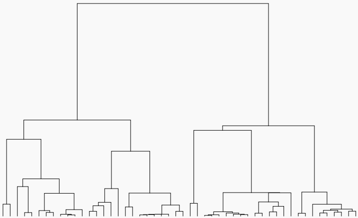
- At each depth, leaf set sizes are approximately equidistributed

- Leaf set sizes are highly unbalanced close to the root

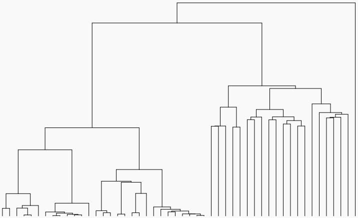
- Given a tree, how can we tell whether it was generated under selection or not?
- Data allows computation of sum of squares of leaf set sizes

Population Genetics: Detecting Selective Pressure

Neutral Tree



Tree with Selection

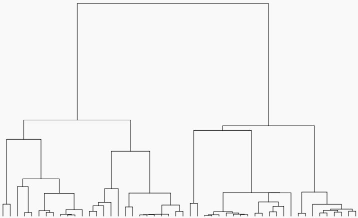


- At each depth, leaf set sizes are approximately equidistributed
- Leaf set sizes are highly unbalanced close to the root

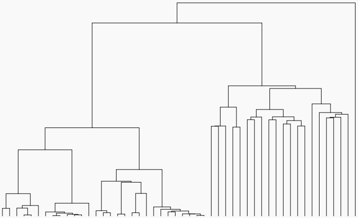
- Given a tree, how can we tell whether it was generated under selection or not?
- Data allows computation of sum of squares of leaf set sizes

Population Genetics: Detecting Selective Pressure

Neutral Tree



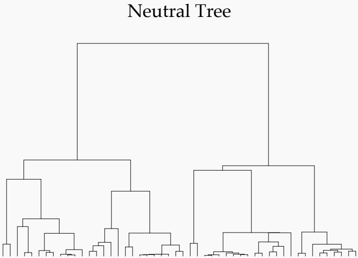
Tree with Selection

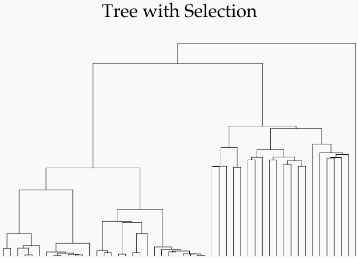


- At each depth, leaf set sizes are approximately equidistributed
- Leaf set sizes are highly unbalanced close to the root

- Given a tree, how can we tell whether it was generated under selection or not?
- Data allows computation of sum of squares of leaf set sizes

Population Genetics: Detecting Selective Pressure





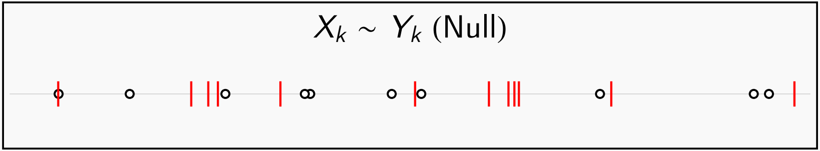
- At each depth, leaf set sizes are approximately equidistributed
- Leaf set sizes are highly unbalanced close to the root

- Given a tree, how can we tell whether it was generated under selection or not?
- Data allows computation of sum of squares of leaf set sizes

Two-Sample Tests: Comparing $\{X_k\}_{k\in[n]}$ and $\{Y_k\}_{k\in[m]}$

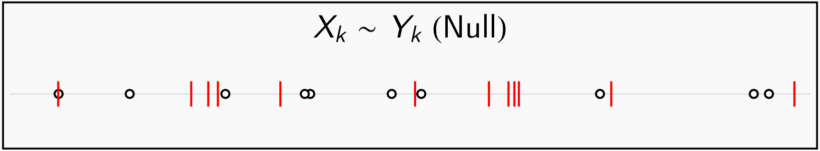
How to test the hypothesis whether $\{X_k\}$ and $\{Y_k\}$ are identically distributed?

Two-Sample Tests: Comparing $\{X_k\}_{k\in[n]}$ and $\{Y_k\}_{k\in[m]}$



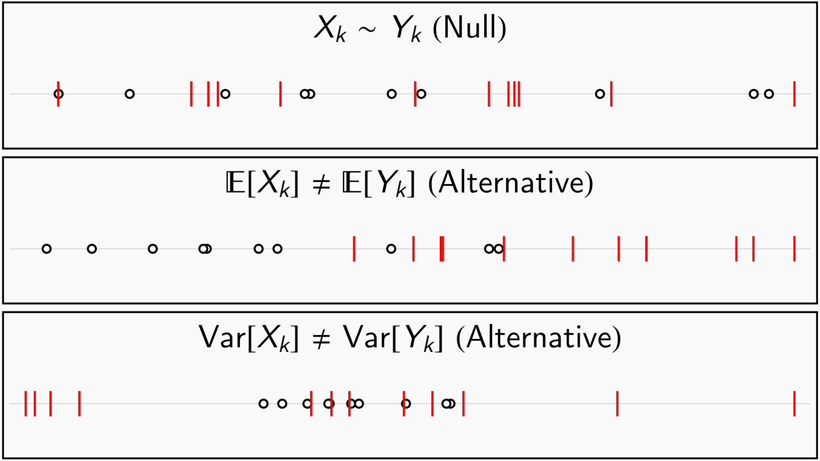
How to test the hypothesis whether $\{X_k\}$ and $\{Y_k\}$ are identically distributed?

Two-Sample Tests: Comparing $\{X_k\}_{k\in[n]}$ and $\{Y_k\}_{k\in[m]}$



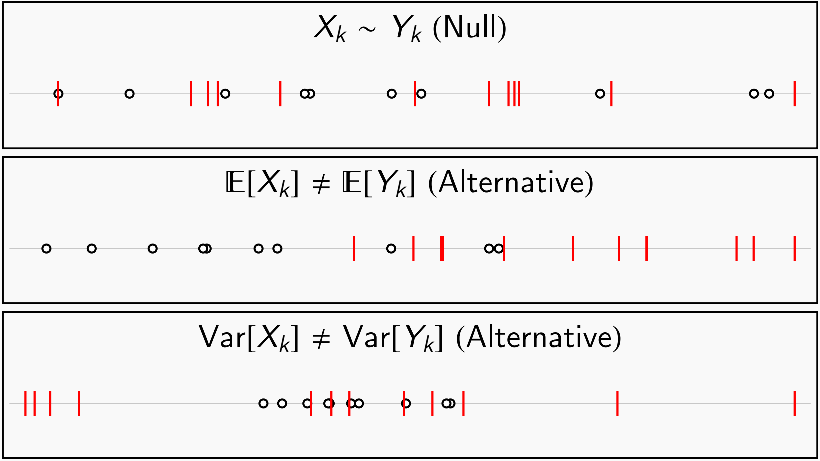
How to test the hypothesis whether $\{X_k\}$ and $\{Y_k\}$ are identically distributed?

Two-Sample Tests: Comparing $\{X_k\}_{k \in [n]}$ and $\{Y_k\}_{k \in [m]}$



How to test the hypothesis whether $\{X_k\}$ and $\{Y_k\}$ are identically distributed?

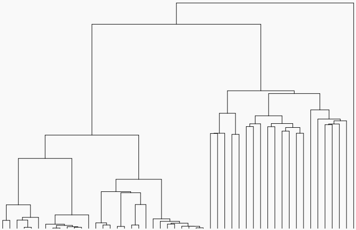
Two-Sample Tests: Comparing $\{X_k\}_{k\in[n]}$ and $\{Y_k\}_{k\in[m]}$



How to test the hypothesis whether $\{X_k\}$ and $\{Y_k\}$ are identically distributed?

Sampling uniformly from the k-dimensional simplex Δ^{k-1}

Balls and bins

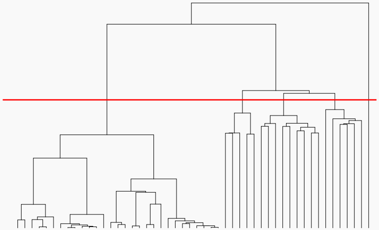


- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins

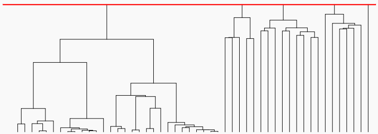


- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '90)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

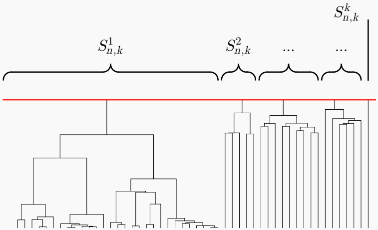
Balls and bins

- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins

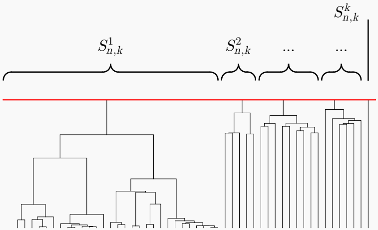


- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins

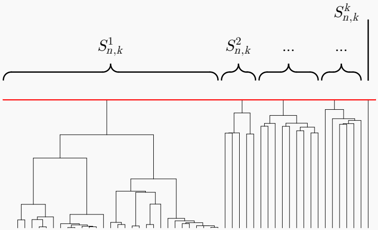


- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|_2^2$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '90)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



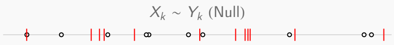
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|_2^2$?

What is the distribution of $\|S_{n,k}\|_2^2$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



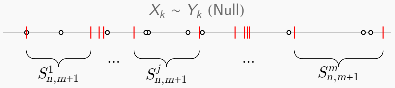
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|_2^2$?

What is the distribution of $\|S_{n,k}\|_2^2$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



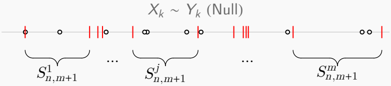
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|_2^2$?

What is the distribution of $\|S_{n,k}\|_2^2$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



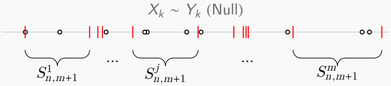
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|^p_{p,w}$

What is the distribution of $\|S_{n,k}\|^p_{p,w}$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



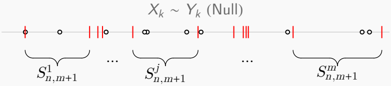
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|^p_{p,w}$

What is the distribution of $\|S_{n,k}\|^p_{p,w}$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



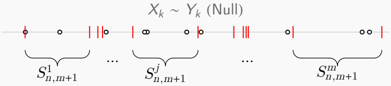
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|^p_{p,w}$

What is the distribution of $\|S_{n,k}\|^p_{p,w}$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



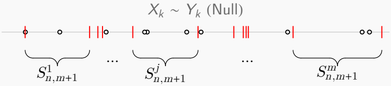
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|^p_{p,w}$

What is the distribution of $\|S_{n,k}\|^p_{p,w}$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



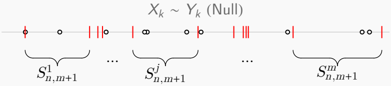
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|^p_{p,w}$

What is the distribution of $\|S_{n,k}\|^p_{p,w}$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



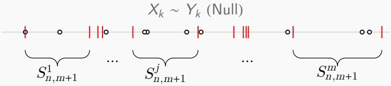
- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|^p_{p,w}$

What is the distribution of $\|S_{n,k}\|^p_{p,w}$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Balls and bins



- $S_{n,k} \sim \mathrm{U}\left(n \cdot \Delta^{k-1} \cap \mathbb{Z}_+\right)$ (Bose-Einstein-Distribution)
- Can we perform hypothesis testing based on $\|S_{n,k}\|_{p,w}^p$

What is the distribution of $\|S_{n,k}\|_{p,w}^p$?

Limit as $n \to \infty$ for fixed k

- Greenwood Statistic (Greenwood '46)
- Some moments, CLT, statistical efficiency (Moran '47, '51, '53)
- Geometry: intersection of L^1 and L^2 balls
  - Up to $k = 3$ (Gardner '52)
  - Large deviations (Schechtner, Zinn '00)
- Tabulation of z-scores up to $k = 20$ (Burrows '79, Currie '81, Stephens '81)

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} \left[x^n\right] \left(S_m(x)\right)^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon}\log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon}\log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}\left([0,1]\right)$

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} [x^n] (S_m(x))^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon}\log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon}\log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}([0,1])$

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} [x^n] (S_m(x))^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon} \log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon} \log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}([0,1])$

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} [x^n] \left(S_m(x)\right)^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon}\log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon}\log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}([0,1])$

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} \left[x^n\right] \left(S_m(x)\right)^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon}\log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon}\log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}\left([0,1]\right)$

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} [x^n] \left(S_m(x)\right)^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon}\log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon}\log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}([0,1])$

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} [x^n] (S_m(x))^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon}\log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon}\log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}([0,1])$

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} [x^n] (S_m(x))^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon} \log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon} \log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}([0,1])$

Results

Observation (Recursion)

Let $G(x) = \sum_{m=0}^{\infty} \mathrm{Li}_{-2m}(x)/m!$, then

$$\mathbb{E}\|S_{n,k}\|_2^{2m} = \frac{m!}{\binom{n-1}{k-1}} [x^n] \left(S_m(x)\right)^{\star(k)} .$$

Corollaries (Discrete)

1. ε-approximation in $O\left(\frac{n}{\varepsilon} \log\left(\frac{n}{\varepsilon}\right) + \frac{n}{\varepsilon} \log k\right)$ time
2. Conservative hypothesis tests
3. Alternative Scaling limits: CLT, LLN, large deviations

Corollaries (Continuous)

1. Continuum approximation: $\|F_{n,k} - F_k\|_\infty \in O\left(n^{-1}\right)$
2. Monotonicity: $F_{n,k} - F_k \geq 0$
3. Regularity: $F_k \in C^{k-3}\left([0,1]\right)$

Application to Two-Sample Testing

Comparing Non-Parametric Two-Sample Tests

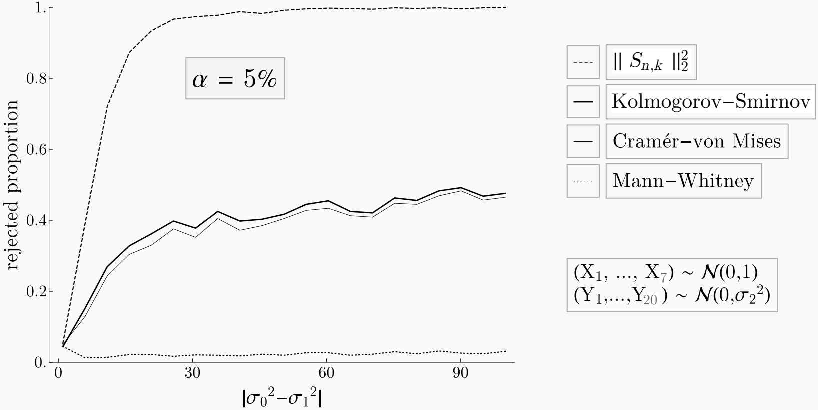


Figure: Hypothesis testing based on $\|S_{n,k}\|_2^2$ is more sensitive to variance changes than common other two-sample tests.

Comparing Non-Parametric Two-Sample Tests

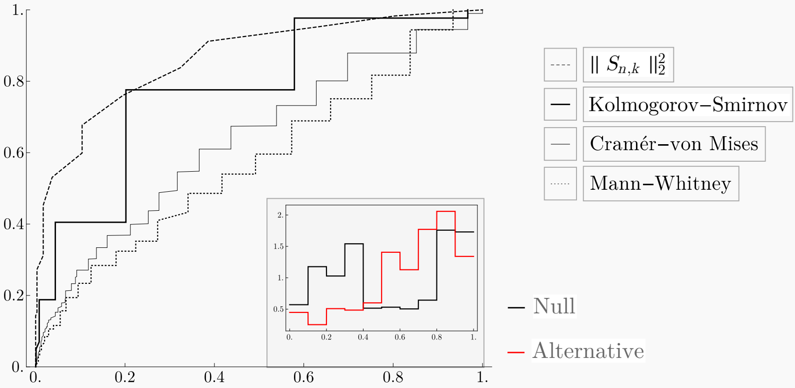


Figure: Hypothesis testing based on $\|S_{n,k}\|_2^2$ is more sensitive to compound mean and variance changes than common other two-sample tests, for randomly generated null and alternative of common support.

Comparing Non-Parametric Two-Sample Tests

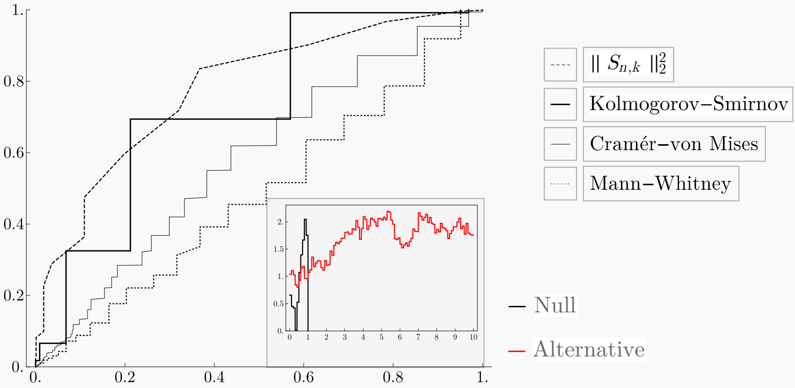


Figure: Hypothesis testing based on $\|S_{n,k}\|_2^2$ is more sensitive to compound mean and variance changes than common other two-sample tests, for randomly generated null and alternative of distinct support.

New Perspectives on old Questions

What happened?

1. Discretized continuous Greenwood Statistic
2. Understood discretized problem through generating functions of moments
3. CDF reconstruction from moments, CLT, transfer to continuous problem
4. Application to two-sample testing

What happens now?

1. Apply hypothesis test to real data
2. Quantify more precisely the power against given classes of alternatives

New Perspectives on old Questions

What happened?

1. Discretized continuous Greenwood Statistic
2. Understood discretized problem through generating functions of moments
3. CDF reconstruction from moments, CLT, transfer to continuous problem
4. Application to two-sample testing

What happens now?

1. Apply hypothesis test to real data
2. Quantify more precisely the power against given classes of alternatives

New Perspectives on old Questions

What happened?

1. Discretized continuous Greenwood Statistic
2. Understood discretized problem through generating functions of moments
3. CDF reconstruction from moments, CLT, transfer to continuous problem
4. Application to two-sample testing

What happens now?

1. Apply hypothesis test to real data
2. Quantify more precisely the power against given classes of alternatives

New Perspectives on old Questions

What happened?

1. Discretized continuous Greenwood Statistic
2. Understood discretized problem through generating functions of moments
3. CDF reconstruction from moments, CLT, transfer to continuous problem
4. Application to two-sample testing

What happens now?

1. Apply hypothesis test to real data
2. Quantify more precisely the power against given classes of alternatives

New Perspectives on old Questions

What happened?

1. Discretized continuous Greenwood Statistic
2. Understood discretized problem through generating functions of moments
3. CDF reconstruction from moments, CLT, transfer to continuous problem
4. Application to two-sample testing

What happens now?

1. Apply hypothesis test to real data
2. Quantify more precisely the power against given classes of alternatives

New Perspectives on old Questions

What happened?

1. Discretized continuous Greenwood Statistic
2. Understood discretized problem through generating functions of moments
3. CDF reconstruction from moments, CLT, transfer to continuous problem
4. Application to two-sample testing

What happens now?

1. Apply hypothesis test to real data
2. Quantify more precisely the power against given classes of alternatives

New Perspectives on old Questions

What happened?

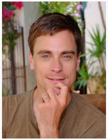
1. Discretized continuous Greenwood Statistic
2. Understood discretized problem through generating functions of moments
3. CDF reconstruction from moments, CLT, transfer to continuous problem
4. Application to two-sample testing

What happens now?

1. Apply hypothesis test to real data
2. Quantify more precisely the power against given classes of alternatives

Acknowledgements

Jonathan Terhorst Yun Song



Jonathan Fischer

THE UNIVERSITY OF CALIFORNIA · BERKELEY · LET THERE BE LIGHT · 1868

Funding:

NIH National Institutes of Health

the David & Lucile Packard FOUNDATION

SIMONS FOUNDATION

German Academic Scholarship Foundation

Studienstiftung des deutschen Volkes

CHAN ZUCKERBERG BIOHUB