Chapter 3.8: Two-Sample Problem

Amine Hadji

University of Leiden

November 9, 2020

What if?

Imagine having two random samples $(X_i)_{i=1}^m$ and $(Y_i)_{i=1}^n$ and wondering

What if?

Imagine having two random samples $(X_i)_{i=1}^{m}$ and $(Y_i)_{i=1}^{n}$ and wondering

- Is the first sample coming from a certain distribution?
- Are the two samples coming from the same dristribution?

Overview

1. Kolmogorov-Smirnov tests
2. Two-Sample Permutation EP/Bootstrap

Kolmogorov distribution

Definition

Kolmogorov distribution is the distribution of the random variable

$$\sup_{t\in[0,1]} |B(t)|$$

where $B(t)$ is the Brownian bridge.

Kolmogorov distribution

Definition

Kolmogorov distribution is the distribution of the random variable

$$\sup_{t \in [0,1]} |B(t)|$$

where $B(t)$ is the Brownian bridge. The cumulative distribution function of $\sup_{t \in [0,1]} |B(t)|$ is given by

$$P(\sup_{t \in [0,1]} |B(t)| \leq x) = 1 - 2 \sum_{k=1}^{+\infty} (-1)^{k-1} e^{-2k^2x^2}.$$

Kolmogorov-Smirnov test

Let $(X_i)_{i=1}^m \in \mathbb{R}$. We want to test the null-hypothesis H_0: $(X_i)_{i=1}^m \sim F$ (a cumulative distribution function)

Definition

The Kolmogorov-Smirnov statistic is defined as

$$D_m = \sqrt{m} \sup_x |\mathbb{F}_m - F|$$

where $\mathbb{F}_m$ is the empirical distribution function of the $(X_i)_{i=1}^m$.

Kolmogorov-Smirnov test

Let $(X_i)_{i=1}^m \in \mathbb{R}$. We want to test the null-hypothesis H_0: $(X_i)_{i=1}^m \sim F$ (a cumulative distribution function)

Definition

The Kolmogorov-Smirnov statistic is defined as

$$D_m = \sqrt{m} \sup_x |\mathbb{F}_m - F|$$

where $\mathbb{F}_m$ is the empirical distribution function of the $(X_i)_{i=1}^m$.

Proposition

Under the null-hypothesis

$$D_m/\sqrt{m} \xrightarrow{a.s} 0, \quad \text{and} \quad D_m \to \sup_{t \in [0,1]} |B(t)|.$$

Kolmogorov-Smirnov test

Definition

Since

$$P(D_m > c_\alpha) \to \alpha = 2\sum_{k=1}^{+\infty}(-1)^{k-1}e^{-2k^2c_\alpha^2}$$

the K-S test with significance α rejects H_0 if $D_m > c_\alpha$

Kolmogorov-Smirnov test two-sample

Let $(X_i)_{i=1}^m$ and $(Y_i)_{i=1}^n$ in $\mathbb{R}$. We want to test the null-hypothesis H_0: the two samples come from the same distribution

Definition

The two-sample Kolmogorov-Smirnov statistic is defined as

$$D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_x |\mathbb{F}_m - \mathbb{F}'_n|$$

Kolmogorov-Smirnov test two-sample

Let $(X_i)_{i=1}^m$ and $(Y_i)_{i=1}^n$ in $\mathbb{R}$. We want to test the null-hypothesis H_0: the two samples come from the same distribution

Definition

The two-sample Kolmogorov-Smirnov statistic is defined as

$$D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_x |\mathbb{F}_m - \mathbb{F}'_n|$$

Here again, the K-S test with significance α rejects H_0 if $D_{n,m} > c_\alpha$.

Two-Sample Permutation test

The two-sample K-S test is quite useful. However, it would be nice to generalize it somehow

Two-Sample Permutation test

The two-sample K-S test is quite useful. However, it would be nice to generalize it somehow
Let $(X_i)_{i=1}^m$ and $(Y_i)_{i=1}^n$ iid sample from distributions P and Q.
We want to test the null-hypothesis H_0: $P = Q$.

Two-Sample Permutation test

The two-sample K-S test is quite useful. However, it would be nice to generalize it somehow

Let $(X_i)_{i=1}^m$ and $(Y_i)_{i=1}^n$ iid sample from distributions P and Q. We want to test the null-hypothesis H_0: $P = Q$.

Definition

The Kolmogorov-Smirnov statistic is defined as

$$D_{m,n} = \sqrt{\frac{mn}{m+n}} \|\mathbb{P}_m - \mathbb{Q}_n\|_{\mathcal{F}}$$

where $\mathcal{F}$ is a class of measurable function, and $\mathbb{P}_m$ and $\mathbb{Q}_m$ are empirical measures.

Donsker class

If $\mathcal{F}$ is Donsker, then $\mathbb{G}_m := \sqrt{m}(\mathbb{P}_m - P) \to \mathbb{G}_P$ and $\mathbb{G}'_n := \sqrt{n}(\mathbb{Q}_n - Q) \to \mathbb{G}_Q$ (two independent Brownian bridges)

Donsker class

If $\mathcal{F}$ is Donsker, then $\mathbb{G}_m := \sqrt{m}(\mathbb{P}_m - P) \to \mathbb{G}_P$ and $\mathbb{G}'_n := \sqrt{n}(\mathbb{Q}_n - Q) \to \mathbb{G}_Q$ (two independent Brownian bridges)

$$D_{m,n} = \left\| \sqrt{\frac{n}{N}}\mathbb{G}_m - \sqrt{\frac{m}{N}}\mathbb{G}'_n + \sqrt{\frac{mn}{N}}(P - Q) \right\|_{\mathcal{F}}$$

with $N = n + m$

Donsker class

If $\mathcal{F}$ is Donsker, then $\mathbb{G}_m := \sqrt{m}(\mathbb{P}_m - P) \to \mathbb{G}_P$ and $\mathbb{G}'_n := \sqrt{n}(\mathbb{Q}_n - Q) \to \mathbb{G}_Q$ (two independent Brownian bridges)

$$D_{m,n} = \left\| \sqrt{\frac{n}{N}}\mathbb{G}_m - \sqrt{\frac{m}{N}}\mathbb{G}'_n + \sqrt{\frac{mn}{N}}(P - Q) \right\|_{\mathcal{F}}$$

with $N = n + m$

- If $\|P - Q\|_{\mathcal{F}} > 0$, then $D_{m,n} \to +\infty$
- If $P = Q$, then $D_{m,n} \to \|\sqrt{1-\lambda}\mathbb{G}_P - \sqrt{\lambda}\mathbb{G}_Q\|_{\mathcal{F}}$ which has the same distribution as $\|\mathbb{G}_P\|_{\mathcal{P}}$

(with $m/N \to \lambda$)

Donsker class

If $\mathcal{F}$ is Donsker, then $\mathbb{G}_m := \sqrt{m}(\mathbb{P}_m - P) \to \mathbb{G}_P$ and $\mathbb{G}'_n := \sqrt{n}(\mathbb{Q}_n - Q) \to \mathbb{G}_Q$ (two independent Brownian bridges)

Donsker class

If $\mathcal{F}$ is Donsker, then $\mathbb{G}_m := \sqrt{m}(\mathbb{P}_m - P) \to \mathbb{G}_P$ and $\mathbb{G}'_n := \sqrt{n}(\mathbb{Q}_n - Q) \to \mathbb{G}_Q$ (two independent Brownian bridges)
The test with significance α rejects H_0 if $D_{m,n} > c_{m,n}$ where

$$c_{m,n} \to c_P := \inf\{t : P(\|\mathbb{G}_P\|_{\mathcal{F}} > t) \leq \alpha\}.$$

Donsker class

If $\mathcal{F}$ is Donsker, then $\mathbb{G}_m := \sqrt{m}(\mathbb{P}_m - P) \to \mathbb{G}_P$ and $\mathbb{G}'_n := \sqrt{n}(\mathbb{Q}_n - Q) \to \mathbb{G}_Q$ (two independent Brownian bridges)
The test with significance α rejects H_0 if $D_{m,n} > c_{m,n}$ where

$$c_{m,n} \to c_P := \inf\{t : P(\|\mathbb{G}_P\|_{\mathcal{F}} > t) \leq \alpha\}.$$

However, c_P depends the underlying measure, P. It is impossible to find $c_{m,n}$ with the given convergence property for every P in H_0

Pooled empirical measure

What can we do now?

Pooled empirical measure

What can we do now?
Answer: Use a pooled empirical measure

$$\mathbb{H}_N := \frac{1}{N}\sum_{i=1}^{N}\delta_{Z_{Ni}} = \lambda_N \mathbb{P}_m + (1-\lambda_N)\mathbb{Q}_n$$

where $(Z_{Ni})_{i=1}^N$ represents the pooled data $(X_i)_{i=1}^m$ and $(Y_i)_{i=1}^n$ and $\lambda_N = m/N$.

Pooled empirical measure

What can we do now?
Answer: Use a pooled empirical measure

$$\mathbb{H}_N := \frac{1}{N}\sum_{i=1}^{N}\delta_{Z_{Ni}} = \lambda_N \mathbb{P}_m + (1-\lambda_N)\mathbb{Q}_n$$

where $(Z_{Ni})_{i=1}^{N}$ represents the pooled data $(X_i)_{i=1}^{m}$ and $(Y_i)_{i=1}^{n}$ and $\lambda_N = m/N$.

$$\mathbb{P}_m - \mathbb{H}_N = (1-\lambda_N)(\mathbb{P}_m - \mathbb{Q}_n)$$

Permutation Empirical Process

If we have been attentive, we might have noticed something

Permutation Empirical Process

If we have been attentive, we might have noticed something
Let a random vector $R = (R_i)_{i=1}^n$ from the uniform distribution on the set of permutations, we can define

$$\tilde{\mathbb{P}}_{m,N} = \frac{1}{m}\sum_{i=1}^{m}\delta_{Z_{NR_i}}, \quad \tilde{\mathbb{Q}}_{n,N} = \frac{1}{n}\sum_{i=m+1}^{N}\delta_{Z_{NR_i}}.$$

Permutation Empirical Process

If we have been attentive, we might have noticed something
Let a random vector $R = (R_i)_{i=1}^n$ from the uniform distribution on the set of permutations, we can define

$$\tilde{\mathbb{P}}_{m,N} = \frac{1}{m}\sum_{i=1}^{m} \delta_{Z_{NR_i}}, \quad \tilde{\mathbb{Q}}_{n,N} = \frac{1}{n}\sum_{i=m+1}^{N} \delta_{Z_{NR_i}}.$$

then

$$\tilde{D}_{m,n} = \sqrt{\frac{mn}{m+n}}\|\tilde{\mathbb{P}}_{m,N} - \tilde{\mathbb{Q}}_{n,N}\|_{\mathcal{F}}$$

The test with significance α rejects H_0 if $\tilde{D}_{m,n} > \tilde{c}_{m,n}$ where

$$\tilde{c}_{m,n} \to c_H := \inf\{t : P(\|\mathbb{G}_H\|_{\mathcal{F}} > t) \leq \alpha\}.$$

Permutation Theorem

Theorem

Let $\mathcal{F}$ be a class of measurable functions that is Donsker under both P and Q and satisfies both $\|P\|_{\mathcal{F}} < \infty$ and $\|Q\|_{\mathcal{F}} < \infty$. If $m, n \to +\infty$ such that $m/N \to \lambda \in (0,1)$, then

$$\sqrt{m}(\tilde{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_H$$

given $X_1, X_2, ..., Y_1, Y_2, ...$ in probability. Here $\mathbb{G}_H$ is a tight Brownian bridge process corresponding to the measure $H = \lambda P + (1-\lambda)Q$.

Permutation Theorem

Theorem

Let $\mathcal{F}$ be a class of measurable functions that is Donsker under both P and Q and satisfies both $\|P\|_{\mathcal{F}} < \infty$ and $\|Q\|_{\mathcal{F}} < \infty$. If $m, n \to +\infty$ such that $m/N \to \lambda \in (0,1)$, then

$$\sqrt{m}(\tilde{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_H$$

given $X_1, X_2, ..., Y_1, Y_2, ...$ in probability. Here $\mathbb{G}_H$ is a tight Brownian bridge process corresponding to the measure $H = \lambda P + (1-\lambda)Q$.
If in addition $\mathcal{F}$ possesses an envelope function F with both $P^*F^2 < \infty$ and $Q^*F^2 < \infty$, then also

$$\sqrt{m}(\tilde{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_H$$

given almost every sequence $X_1, X_2, ..., Y_1, Y_2, ...$

Permutation Theorem

Main idea of the proof:
(wlog $Pf = 0$)
Step 1:

Permutation Theorem

Main idea of the proof:
(wlog $Pf = 0$)
Step 1:

- Use the CLT for two-sample rank statistics, to show for all f with $Hf = 0$ and $Hf^2 < \infty$

Permutation Theorem

Main idea of the proof:
(wlog $Pf = 0$)
Step 1:

- Use the CLT for two-sample rank statistics, to show for all f with $Hf = 0$ and $Hf^2 < \infty$
- $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} (f(Z_{NR_i}) - \mathbb{H}_N f) \rightsquigarrow N(0, (1-\lambda)Hf^2)$ given almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$

CLT for two-sample rank statistics

Let $n > 0$, let $(a_i)_{i=1}^n$ and $(b_i)_{i=1}^n$ be real numbers, and let $R = (R_i)_{i=1}^n$ from the uniform distribution on the set of permutations. Consider the rank statistic

$$S_n = \sum_{i=1}^n b_i a_{R_i}.$$

The mean and variance of S_n are equal to $ES_n = n\bar{a}\bar{b}$ and $V(S_n) = A_n^2 B_n^2/(n-1)$, where A_n^2 and B_n^2 are the sums of squared deviations from the respective means of the a's and b's

CLT for two-sample rank statistics

Proposition (Rank central limit theorem)

Suppose that

$$\max_{1\leq n} \frac{|a_i - \bar{a}|}{A_n} \to 0, \quad \max_{1\leq n} \frac{|b_i - \bar{b}|}{B_n} \to 0$$

Then the sequence $(S_n - ES_n)/\sigma(S_n)$ *converges in distribution to a standard normal distribution iff*

$$\sum_{\sqrt{n}|a_i - \bar{a}||b_j - \bar{b}| > \epsilon A_n B_n} \frac{|a_i - \bar{a}|^2 |b_i - \bar{b}|^2}{A_n^2 B_n^2} \to 0$$

for every $\epsilon > 0$

Permutation Theorem

Main idea of the proof:
(wlog $Pf = 0$)
Step 1:

- Use the CLT for two-sample rank statistics, to show for all f with $Hf = 0$ and $Hf^2 < \infty$
- $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} (f(Z_{NR_i}) - \mathbb{H}_N f) \rightsquigarrow N(0, (1-\lambda)Hf^2)$ given almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$

Hoeffding's inequality

Proposition

Let $\{c_i\}_{i=1}^N$ *be elements of an arbitrary vector space* V*, and let* $(U_i) = i = 1^n$ *and* $(V_i) = i = 1^n$ *denote samples of size* $n \leq N$ *drawn without and with replacement, respectively, from* $\{c_i\}_{i=1}^N$*. Then, for every convex function* ϕ *from* V *to* $\mathbb{R}$

$$E\phi\Big(\sum_{i=1}^{n} U_i\Big) \leq E\phi\Big(\sum_{i=1}^{n} V_i\Big)$$

Permutation Theorem

Step 2:

- Let $(\hat{Z}_{Ni})_{i=1}^{N}$ an iid sample from $\mathbb{H}_N$

Permutation Theorem

Step 2:

- Let $(\hat{Z}_{Ni})_{i=1}^{N}$ an iid sample from $\mathbb{H}_N$
- Set $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}; H(f-g)^2 < \delta^2\}$

Permutation Theorem

Step 2:

- Let $(\hat{Z}_{Ni})_{i=1}^{N}$ an iid sample from $\mathbb{H}_N$
- Set $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}; H(f-g)^2 < \delta^2\}$
- Use Hoeffding's inequality

Permutation Theorem

Step 2:

- Let $(\hat{Z}_{Ni})_{i=1}^{N}$ an iid sample from $\mathbb{H}_N$
- Set $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}; H(f-g)^2 < \delta^2\}$
- Use Hoeffding's inequality

$$\begin{aligned} E_R\|\sqrt{m}(\tilde{\mathbb{P}}_{m,N} - \mathbb{H}_N)\|_{\mathcal{F}_\delta} &= E_R\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(\delta_{Z_{R_i}} - \mathbb{H}_N)\right\|_{\mathcal{F}_\delta} \\ &\leq E_{\hat{Z}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(\delta_{\hat{Z}_{Ni}} - \mathbb{H}_N)\right\|_{\mathcal{F}_\delta} \end{aligned}$$

Permutation Theorem

Step 3: Sequence of inequalities

Step 3: Sequence of inequalities
Poissonization inequality with $(\tilde{N}_i)_{i=1}^N$ iid symetrized $P(m/N)$

$$
\begin{aligned}
E_{\hat{Z}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(\delta_{\hat{Z}_{Ni}} - \mathbb{H}_N)\right\|_{\mathcal{F}_\delta} &\leq 4E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{N}\tilde{N}_i\delta_{Z_{R_i}}\right\|_{\mathcal{F}_\delta} \\
&\lesssim E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\tilde{N}_i\delta_{X_i}\right\|_{\mathcal{F}_\delta} \\
&+ E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\tilde{N}_i\delta_{Y_i}\right\|_{\mathcal{F}_\delta}
\end{aligned}
$$

Permutation Theorem

Step 3: Sequence of inequalities
By Theorem 2.9.2 [Multiplier Central Limit Theorems], the $m^{1/2}\sum_{i=1}^{m}\tilde{N}_i\delta_{X_i}\to Z_P$ and $n^{1/2}\sum_{i=1}^{n}\tilde{N}_i\delta_{Y_i}\to Z_P$, two tight Gaussian processes.

$$
\begin{aligned}
E_{\hat{Z}}\Big\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(\delta_{\hat{Z}_{Ni}}-\mathbb{H}_N)\Big\|_{\mathcal{F}_\delta} &\leq 4E_{\tilde{N}}\Big\|\frac{1}{\sqrt{m}}\sum_{i=1}^{N}\tilde{N}_i\delta_{Z_{R_i}}\Big\|_{\mathcal{F}_\delta}\\
&\lesssim E_{\tilde{N}}\Big\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\tilde{N}_i\delta_{X_i}\Big\|_{\mathcal{F}_\delta}\\
&+E_{\tilde{N}}\Big\|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\tilde{N}_i\delta_{Y_i}\Big\|_{\mathcal{F}_\delta}
\end{aligned}
$$

Permutation Theorem

Step 3: Sequence of inequalities
We finally manage to have this bound

$$E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\tilde{N}_i\delta_{X_i}\right\|_{\mathcal{F}_\delta}+E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\tilde{N}_i\delta_{Y_i}\right\|_{\mathcal{F}_\delta}$$

Step 3: Sequence of inequalities
We finally manage to have this bound

$$E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\tilde{N}_i\delta_{X_i}\right\|_{\mathcal{F}_\delta}+E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\tilde{N}_i\delta_{Y_i}\right\|_{\mathcal{F}_\delta}$$

By Theorem 2.9.2 [Multiplier Central Limit Theorems], the $m^{1/2}\sum_{i=1}^{m}\tilde{N}_i\delta_{X_i}\to Z_P$ and $n^{1/2}\sum_{i=1}^{n}\tilde{N}_i\delta_{Y_i}\to Z_P$, two tight Gaussian processes.

Permutation Theorem

Step 3: Sequence of inequalities
We finally manage to have this bound

$$E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\tilde{N}_i\delta_{X_i}\right\|_{\mathcal{F}_\delta}+E_{\tilde{N}}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\tilde{N}_i\delta_{Y_i}\right\|_{\mathcal{F}_\delta}$$

By Theorem 2.9.2 [Multiplier Central Limit Theorems], the $m^{1/2}\sum_{i=1}^{m}\tilde{N}_i\delta_{X_i}\to Z_P$ and $n^{1/2}\sum_{i=1}^{n}\tilde{N}_i\delta_{Y_i}\to Z_P$, two tight Gaussian processes.

By Lemma 2.3.11 [Symmetrization and Measurability], the outer expectation is asymptotically bounded by a multiple $E\|Z_P\|_{\mathcal{F}_\delta}+\sqrt{(1-\lambda)/\lambda}E\|Z_Q\|_{\mathcal{F}_\delta}\to 0$ as $\delta\to 0$.

Two-Sample Bootstrap

Can we use bootstrap to this method?

Two-Sample Bootstrap

Can we use bootstrap to this method?

Of course!

Two-Sample Bootstrap

Can we use bootstrap to this method?
Of course! Let $\hat{Z} = (\hat{Z}_i)_{i=1}^N$ iid sample from $\mathbb{H}_N$. The two-sample bootstrap empirical measures

$$\hat{\mathbb{P}}_{m,N} = \frac{1}{m}\sum_{i=1}^{m}\delta_{\hat{Z}_i}, \quad \hat{\mathbb{Q}}_{n,N} = \frac{1}{n}\sum_{i=m+1}^{N}\delta_{\hat{Z}_{m+i}}.$$

Two-Sample Bootstrap

Can we use bootstrap to this method?
Of course! Let $\hat{Z} = (\hat{Z}_i)_{i=1}^N$ iid sample from $\mathbb{H}_N$. The two-sample bootstrap empirical measures

$$\hat{\mathbb{P}}_{m,N} = \frac{1}{m}\sum_{i=1}^{m} \delta_{\hat{Z}_i}, \quad \hat{\mathbb{Q}}_{n,N} = \frac{1}{n}\sum_{i=m+1}^{N} \delta_{\hat{Z}_{m+i}}.$$

then

$$\hat{D}_{m,n} = \sqrt{\frac{mn}{m+n}} \|\hat{\mathbb{P}}_{m,N} - \hat{\mathbb{Q}}_{n,N}\|_{\mathcal{F}}$$

The test with significance α rejects H_0 if $\hat{D}_{m,n} > \hat{c}_{m,n}$ where

$$\hat{c}_{m,n} \to c_H := \inf\{t : P(\|\mathbb{G}_H\|_{\mathcal{F}} > t) \leq \alpha\}.$$

Two-Sample Bootstrap Theorem

Theorem

Let $\mathcal{F}$ be a class of measurable functions that is Donsker under both P and Q and satisfies both $\|P\|_{\mathcal{F}} < \infty$ and $\|Q\|_{\mathcal{F}} < \infty$. If $m, n \to +\infty$ such that $m/N \to \lambda \in (0,1)$, then

$$\sqrt{m}(\hat{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \sqrt{1-\lambda}\,\mathbb{G}_H$$

given $X_1, X_2, ..., Y_1, Y_2, ...$ in probability. Here $\mathbb{G}_H$ is a tight Brownian bridge process corresponding to the measure $H = \lambda P + (1-\lambda)Q$.

Two-Sample Bootstrap Theorem

Theorem

Let $\mathcal{F}$ be a class of measurable functions that is Donsker under both P and Q and satisfies both $\|P\|_{\mathcal{F}} < \infty$ and $\|Q\|_{\mathcal{F}} < \infty$. If $m, n \to +\infty$ such that $m/N \to \lambda \in (0,1)$, then

$$\sqrt{m}(\hat{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_H$$

given $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ in probability. Here $\mathbb{G}_H$ is a tight Brownian bridge process corresponding to the measure $H = \lambda P + (1-\lambda)Q$.
If in addition $\mathcal{F}$ possesses an envelope function F with both $P^* F^2 < \infty$ and $Q^* F^2 < \infty$, then also

$$\sqrt{m}(\hat{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_H$$

given almost every sequence $X_1, X_2, \ldots, Y_1, Y_2, \ldots$

Two-Sample Bootstrap Theorem

Main idea of the proof:

Two-Sample Bootstrap Theorem

Main idea of the proof:
Exactly the same as the previous one.

Two-Sample Bootstrap Theorem

Main idea of the proof:
Exactly the same as the previous one. Actually, even simpler because no need for Hoeffding's inequality.

Trivia

- K-S test is a special case of the test in this Chapter
- The previous results can easily be generalized for a k-sample problem
- They can also be used to test independence in iid pairs of variable

Thank you

Any questions?