2020

On a Projective Ensemble Approach to Two Sample Test for Equality of Distributions

Zhimei Li　Yaowu Zhang　　Shanghai University of Finance and Economics

CONTENTS

1 Introduction

1.1 Research Question

$$\mathrm{x} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p \sim F$$

$$\mathrm{y} = (Y_1, \ldots, Y_p)^T \in \mathbb{R}^p \sim G$$

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

1.2 Value of Research

Testing whether two samples come from the same population is one of the most fundamental problems in statistics and has applications in a wide range of areas. For example, we can check the consistency of the distribution of training samples and test samples

1.3 Research Method

- We apply the idea of projections and develop a new **projective ensemble approach** for testing equality of distributions.
- This method has the following advantages:

1. Simple **closed-form**, **no tuning parameters**,
2. Be computed in **quadratic time**,
3. **Be insensitive to the dimension**, **consistent** against all fixed alternatives,
4. No moment assumption, **robust to the outliers**.

1 Introduction

Motivation

- Some existing methods can be implemented in quadratic time but have been reported to be sensitive to heavy-tailed data,
- Robust counterparts are computationally challenging with a cubic time complexity.

So we want to improve the approach proposed by Kim et al. (2020), and propose a robust test, meanwhile reduce the computational cost.

1.4 Related literature

Normality assumption:

Mean vector; covariance matrices

examples

The Student's t test;
Hotelling's T^2 test;
Bai & Saranadasa (1996);
Li & Chen (2012);
Cai et al. (2014);
Cai & Liu (2016)

disadvantages

- The first two moments are not sufficient to characterize the distribution
- May be inconsistent when the normality assumption violates

nonparametric approaches:

Use a measure of difference between F_m and G_n as the test statistic

examples

Kolmogorov-Smirnov test statistic (Smirnov, 1939):

$$\sqrt{nm/(n+m)}\,\sup_{t\in\mathbb{R}}\left|F_m(t)-G_n(t)\right|$$

Cramér-von Mises (CvM) test statistic (Anderson, 1962) and Anderson-Darling statistic (Darling, 1957) :

$$\frac{mn}{m+n}\int_{-\infty}^{\infty}\left\{F_m(t)-G_n(t)\right\}^2\omega\left\{H_{m+n}(t)\right\}dH_{m,n}(t),$$

Introduction

Advantages

When p = 1,

- Consistent against any fixed alternatives, distribution free under the null,
- No moment conditions are required,
- Free of tuning parameters,

Dis-advantages

- Difficult to generalize to multivariate cases (Kim et al., 2020).
- Suffer from significant power loss when p increases.

Introduction

graph-based tests

$p \geq 2$

reproducing kernel Hilbert space (RKHS)

- k minimum spanning tree graphs;
- k nearest neighbor graphs.

disadvantages

- Inconsistent
- Rely on selecting tuning parameters

- Maximum mean discrepancy (MMD) test statistic based on RKHS;
- Energy statistic (be a special case of the MMD).

1 Introduction

Kim et al. (2020)

$$\iint \{F_{\boldsymbol{\beta}}(t) - G_{\boldsymbol{\beta}}(t)\}^2 dH_{\boldsymbol{\beta}}(t) d\lambda(\boldsymbol{\beta}), \qquad (1)$$

Where:

$$F_\beta(t) = \mathrm{P}(\beta^\mathrm{T}\mathrm{x} \le t), G_\beta(t) = \mathrm{P}(\beta^\mathrm{T}\mathrm{y} \le t),$$

$$H_\beta(t) = \tau F_\beta(t) + (1-\tau)G_\beta(t),$$

$$\lim_{\min(m,n)\to\infty} \tau = m/(m+n)$$

$\lambda(\beta)$ is the uniform probability measure on the p-dimensional unit sphere

$$\mathcal{S}^{p-1} \stackrel{\text{def}}{=} \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\| = 1\}$$

$H_{\boldsymbol{\beta}}(t) = t$

energy statistic
(Baringhaus& Franz, 2004)

Table: Comparison of Projection-averaging approach and energy statistic

	Projection-averaging approach	Energy statistic
Advantages	• nonnegative and equal to zero if and only if F = G • have a simple closed-form expression • free of tuning parameters	
	robust to heavy-tailed distributions or outliers	quadratic computations
Disadvantages	cubic computations	energy distance is only well-defined under the moment condition (finite first moment)

$$\iint \{F_{\boldsymbol{\beta}}(t) - G_{\boldsymbol{\beta}}(t)\}^2 dH(\boldsymbol{\beta}, t) = 0. \qquad (2)$$

Projection-averaging approach focused on the case that β^Tx and β^Ty have continuous distribution functions for all $\beta \in S^{p-1}$, whereas we are targeting on a more general case and we do not need such continuous distribution assumption.

These observations motivate us to carefully choose other weight functions such that

1. The integration in (2) equals zero if and only if x and y are equally distributed;
2. The choice of $H(\beta, t)$ does not depend on unknown functions which are difficult to estimate;
3. The integration in (2) has a closed-form expression, and is finite without any moment conditions.

We apply the idea of projections and develop a new projective ensemble approach for testing equality of distributions.

2 Projective Ensemble Test

2.1 Motivation

The integration in Eq.(2) can be rewritten as

$$\begin{aligned}&\iint F_{\boldsymbol{\beta}}^2(t)dH(\boldsymbol{\beta},t) - 2\iint F_{\boldsymbol{\beta}}(t)G_{\boldsymbol{\beta}}(t)dH(\boldsymbol{\beta},t) + \iint G_{\boldsymbol{\beta}}^2(t)dH(\boldsymbol{\beta},t)\\ = &\boxed{\iint E\left\{I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_2 \le t)\right\} dH(\boldsymbol{\beta},t)} - 2\iint E\left\{I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}_2 \le t)\right\} dH(\boldsymbol{\beta},t)\\ &+ \iint E\left\{I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}_2 \le t)\right\} dH(\boldsymbol{\beta},t),\end{aligned}$$

In order to obtain a closed-form expression, we need to evaluate the three integrations in the above display. We take the first integration for example. By adopting Fubini's theorem, it suffices to find H(β,t) such that the following integration

$$\iint I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \le t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_2 \le t)dH(\boldsymbol{\beta},t)$$

has a closed form for given $x_1$and x_2

2 Projective Ensemble Test

Lemma 1. *(Gupta, 1963) Let* $(Z_1, Z_2)^T$ *be bivariate normal with mean* $\mathbf{0}$*. The correlation between* Z_1 *and* Z_2 *is* ρ*, then*

$$P(Z_1 \geq 0, Z_2 \geq 0) = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho).$$

By treating x_1 and x_2 as constants, $(\beta, t)^T$ as a $p+1$ dimensional **multivariate joint normal random vector** with cumulative distribution function $H(\beta, t)$, the integration can be expressed as

$$\begin{aligned} & \iint I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \leq t, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_2 \leq t) dH(\boldsymbol{\beta}, t) \\ = \;& \Pr\left(t - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1 \geq 0, t - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_2 \geq 0 \,\middle|\, \mathbf{x}_1, \mathbf{x}_2\right) \\ = \;& \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(\frac{1 + \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_2}{\sqrt{1 + \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1}\sqrt{1 + \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_2}}\right). \end{aligned}$$

Consequently, the integration in (2) can be expressed in a closed form, which is shown in the following Theorem.

Theorem 1. *Suppose* $\mathbf{x} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ *and* $\mathbf{y} = (Y_1, \ldots, Y_p)^T \in \mathbb{R}^p$ *are two* p*-dimensional random vectors whose distribution functions are* F *and* G*, respectively.* $\mathbf{x}_1, \mathbf{x}_2$ *and* $\mathbf{y}_1, \mathbf{y}_2$ *are two independent copies of* $\mathbf{x}$ *and* $\mathbf{y}$*, respectively. Let* $H(\boldsymbol{\beta}, t)$ *be the cumulative distribution function of a* $p+1$ *dimensional multivariate joint normal random vector with mean* $\mathbf{0}$ *and covariance* $\mathbf{I}_{p+1}$*. Then*

$$\begin{aligned} T &= 2\pi \iint \{F_{\boldsymbol{\beta}}(t) - G_{\boldsymbol{\beta}}(t)\}^2 dH(\boldsymbol{\beta}, t) \\ &= T_1 - 2T_2 + T_3, \end{aligned} \tag{3}$$

where T_1*,* T_2 *and* T_3 *are defined as*

$$T_1 \stackrel{def}{=} E \arcsin\left(\frac{1 + \mathbf{x}_1^T\mathbf{x}_2}{\sqrt{1 + \mathbf{x}_1^T\mathbf{x}_1}\sqrt{1 + \mathbf{x}_2^T\mathbf{x}_2}}\right),$$

$$T_2 \stackrel{def}{=} E \arcsin\left(\frac{1 + \mathbf{x}_1^T\mathbf{y}_2}{\sqrt{1 + \mathbf{x}_1^T\mathbf{x}_1}\sqrt{1 + \mathbf{y}_2^T\mathbf{y}_2}}\right),$$

$$T_3 \stackrel{def}{=} E \arcsin\left(\frac{1 + \mathbf{y}_1^T\mathbf{y}_2}{\sqrt{1 + \mathbf{y}_1^T\mathbf{y}_1}\sqrt{1 + \mathbf{y}_2^T\mathbf{y}_2}}\right).$$

In addition, T *is nonnegative and equals zero if and only if* $F = G$*.*

2 Projective Ensemble Test

2.2 Asymptotic properties

At the sample level, we estimate T_1, T_2, and T_3 by

$$\widehat{T}_1 \overset{\text{def}}{=} \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\arcsin\left(\frac{1+\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j}{\sqrt{1+\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i}\sqrt{1+\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j}}\right),$$

$$\widehat{T}_2 \overset{\text{def}}{=} \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\arcsin\left(\frac{1+\mathbf{x}_i^{\mathrm{T}}\mathbf{y}_j}{\sqrt{1+\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i}\sqrt{1+\mathbf{y}_j^{\mathrm{T}}\mathbf{y}_j}}\right),$$

$$\widehat{T}_3 \overset{\text{def}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\arcsin\left(\frac{1+\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_j}{\sqrt{1+\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i}\sqrt{1+\mathbf{y}_j^{\mathrm{T}}\mathbf{y}_j}}\right).$$

V-statistic

Complexity: $O\{(m+n)^2\}$

2 Projective Ensemble Test

asymptotic properties of the test statistic under the null hypothesis

Theorem 2. *Under the null hypothesis, that is, $F = G$, as $\min(m, n) \to \infty$, $\frac{mn}{m+n}\widehat{T}$ converges in distribution to*

$$2\pi \iint \{\zeta(\boldsymbol{\beta}, t)\}^2 dH(\boldsymbol{\beta}, t),$$

where $\zeta(\boldsymbol{\beta}, t)$ is a gaussian random process with mean zero and covariance function, $\operatorname{cov}\{\zeta(\boldsymbol{\beta}, t), \zeta(\boldsymbol{\alpha}, s)\}$, is given by

$$P\left(\boldsymbol{\beta}^T\mathbf{x} \le t, \boldsymbol{\alpha}^T\mathbf{x} \le s\right) - P\left(\boldsymbol{\beta}^T\mathbf{x} \le t\right) P\left(\boldsymbol{\alpha}^T\mathbf{x} \le s\right). \quad (4)$$

No moment condition

No continuity assumption

$mn/(m+n)^2 \to \tau(1-\tau)$ as $\min(m, n) \to \infty$,
τ is the limit value of $m/(m+n)$.

→ $(m+n)$ consistent

Projective Ensemble Test

Under the global alternative, F ≠ G and the difference between the two distribution functions does not vary with the sample size.

Theorem 3. *Under the global alternative hypothesis, as* $\min(m,n) \to \infty$, $(m+n)^{1/2}(\widehat{T} - T)$ *converges in distribution to*

$$\mathcal{N}\left\{0, \frac{4(1-\tau)\mathrm{var}(Z_1) + 4\tau\mathrm{var}(Z_2)}{\tau(1-\tau)}\right\}$$

where $\tau \in (0,1)$ *is the limit value of* $m/(m+n)$, Z_1 *and* Z_2 *are defined in* (S.3.1) *and* (S.3.2) *in the Supplementary Material, respectively.*

$$E\left\{\arcsin\left(\frac{1+\widetilde{\mathbf{x}}^{\mathrm{T}}\mathbf{x}}{\sqrt{1+\widetilde{\mathbf{x}}^{\mathrm{T}}\widetilde{\mathbf{x}}}\sqrt{1+\mathbf{x}^{\mathrm{T}}\mathbf{x}}}\right) - \arcsin\left(\frac{1+\mathbf{x}^{\mathrm{T}}\widetilde{\mathbf{y}}}{\sqrt{1+\mathbf{x}^{\mathrm{T}}\mathbf{x}}\sqrt{1+\widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}}}\right) \middle| \mathbf{x}\right\} \quad \text{(S.3.1)}$$

$$E\left\{\arcsin\left(\frac{1+\widetilde{\mathbf{x}}^{\mathrm{T}}\mathbf{y}}{\sqrt{1+\widetilde{\mathbf{x}}^{\mathrm{T}}\widetilde{\mathbf{x}}}\sqrt{1+\mathbf{y}^{\mathrm{T}}\mathbf{y}}}\right) - \arcsin\left(\frac{1+\widetilde{\mathbf{y}}^{\mathrm{T}}\mathbf{y}}{\sqrt{1+\widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}}\sqrt{1+\mathbf{y}^{\mathrm{T}}\mathbf{y}}}\right) \middle| \mathbf{y}\right\} \quad \text{(S.3.2)}$$

2 Projective Ensemble Test

Under the local alternative, $F \neq G$ but the difference between the two distribution functions diminishes as the sample size increases. We consider a sequence of local alternatives as follows:

$$H_{1l} : \mathrm{P}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \leq t) = \mathrm{P}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y} \leq t) + (m+n)^{-1/2}\ell(\boldsymbol{\beta}, t),$$

where $\ell(\boldsymbol{\beta}, t)$ is a function depending only on $\boldsymbol{\beta}$ and t such that $\iint \ell^2(\boldsymbol{\beta}, t) dH(\boldsymbol{\beta}, t)$ exists.

Theorem 4. *Under the local alternative hypothesis, as* $\min(m, n) \to \infty$, $\frac{mn}{m+n}\widehat{T}$ *converges in distribution to*

$$2\pi \iint \{\zeta(\boldsymbol{\beta}, t) + \tau^{1/2}(1-\tau)^{1/2}\ell(\boldsymbol{\beta}, t)\}^2 dH(\boldsymbol{\beta}, t),$$

where τ *is the limit value of* $m/(m+n)$, $\zeta(\boldsymbol{\beta}, t)$ *is a gaussian random process defined in Theorem 2.*

That is, as long as the difference is larger than $O\left\{(m+n)^{-1/2}\right\}$, it can be consistently detected by our proposed test with probability tending to one.

2 Projective Ensemble Test

1. Let $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{m+n}\} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{y}_1, \ldots, \mathbf{y}_n\}$ denote the pooled samples. Randomly permute the pooled samples to obtain $\{\mathbf{z}_1^*, \mathbf{z}_2^*, \ldots, \mathbf{z}_{m+n}^*\}$.
2. Select the first m observations from the pooled samples as $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_m^*\}$, and the rest observations as $\{\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*\}$.
3. Based on the two randomly permuted samples $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_m^*\}$ and $\{\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*\}$, calculate the test statistic to obtain $\widehat{T}^*$.
4. Repeat steps 1 to 3 for B times to obtain $\widehat{T}_b^*, b = 1, 2, \ldots, B$. The associated p-value is given by

$$B^{-1} \sum_{b=1}^{B} I\left(\widehat{T}_b^* \geq \widehat{T}\right),$$

where $I(\cdot)$ is an indicator function. Reject the null hypothesis if the p-value is smaller than the given significance level.

Theorem 5. *As* $\min(m, n) \to \infty$,

$$\sup_{t \geq 0} \left| P\left\{ mn/(m+n)\widehat{T}^* \leq t \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_n \right\} - P(T_\infty^* \leq t) \right|$$

converges in probability to 0, where T_∞^* *is defined as*

$$T_\infty^* = 2\pi \iint \{\zeta^*(\boldsymbol{\beta}, t)\}^2 dH(\boldsymbol{\beta}, t)$$

and $\zeta^*(\boldsymbol{\beta}, t)$ *is a gaussian random process with mean zero and covariance function,* $\operatorname{cov}\{\zeta^*(\boldsymbol{\beta}, t), \zeta^*(\boldsymbol{\alpha}, s)\}$*, given by*

$$(1-\tau)\left\{P\left(\boldsymbol{\beta}^T\mathbf{y} \leq t, \boldsymbol{\alpha}^T\mathbf{y} \leq s\right) - P(\boldsymbol{\beta}^T\mathbf{y} \leq t)P(\boldsymbol{\alpha}^T\mathbf{y} \leq s)\right\}$$
$$+\tau\left\{P\left(\boldsymbol{\beta}^T\mathbf{x} \leq t, \boldsymbol{\alpha}^T\mathbf{x} \leq s\right) - P(\boldsymbol{\beta}^T\mathbf{x} \leq t)P(\boldsymbol{\alpha}^T\mathbf{x} \leq s)\right\}. \quad (5)$$

3 Numerical Studies

We generate the samples $\{\mathbf{x}_i, \mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n_x\}$, $\{\mathbf{y}_i, \mathbf{y}_i \in \mathbb{R}^p, i = 1, \ldots, n_y\}$, and $\{\mathbf{z}_i, \mathbf{z}_i \in \mathbb{R}^p, i = 1, \ldots, n_z\}$ independently from $t_d(\mu_x \mathbf{1}_p, \sigma_x^2 \mathbf{I}_p)$, $t_d(\mu_y \mathbf{1}_p, \sigma_y^2 \mathbf{I}_p)$, and $t_d(\mu_z \mathbf{1}_p, \sigma_z^2 \mathbf{I}_p)$, respectively.

$$\mu_x = 0,\ \sigma_x = 1,$$
$$\mu_y = 1,\ \sigma_y = 1,$$
$$\mu_z = 1,\ \sigma_z = 2.$$

Compare x and y to inspect location shift
Compare y and z to inspect scale difference
Compare x and z to inspect both location shift and scale difference

Throughout the experiment, we set the significance level as 0.05. We repeat each experiment 1000 times and determine the critical values with 1000 permutations.

1. Normal distributions, $n_x = n_y = n_z = 20, p = 10$;
2. Cauchy distributions, $n_x = n_y = n_z = 20, p = 10$;
3. Cauchy distributions, $n_x = 20$, $n_y = 20, n_z = 40, p = 100$;
4. Normal distributions, $n_x = n_y = \{20, 50, 100\}, p = 10$.

3 Numerical Studies

We compare the performance of the projection ensemble based test (“PE”) with other competing nonparametric tests.

1. the projection-averaging based Cramér-von Mises test (Kim et al., 2020, “CvM”),
2. the k nearest neighbor test (Henze, 1988, “NN”),
3. the modified k nearest neighbor test (Mondal et al., 2015, “MGB”),
4. the energy statistic based test (Székely & Rizzo, 2004, “Energy”),
5. the inter-point distance test (Biswas & Ghosh, 2014, “BG”),
6. the cross-match test (Rosenbaum, 2005, “CM”),
7. ball divergence test (Pan et al., 2018, “Ball”).

3 Numerical Studies

Case 1: Normal distributions, $n_x = n_y = n_z = 20, p = 10$;

Table 1: The empirical powers for different methods when all the samples are generated from multivariate normal distributions at significance level $\alpha = 0.05$.

	PE	CvM	NN	MGB	Energy	BG	CM	Ball
Location	1.000	1.000	0.999	0.996	1.000	1.000	0.994	1.000
Scale	0.713	0.880	0.723	1.000	0.989	1.000	0.169	1.000
Location-Scale	0.966	1.000	0.997	1.000	1.000	1.000	0.765	1.000

The cross-match test is not efficient in detecting the scale difference may be mainly because it relies on some tuning parameters.

3 Numerical Studies

Case 2: Cauchy distributions, $n_x = n_y = n_z = 20, p = 10$;
Case 3: Cauchy distributions, $n_x = 20$, $n_y = 20, n_z = 40, p = 100$;

Table 2: The empirical powers for different methods when all the samples are generated from Cauchy distributions at significance level $\alpha = 0.05$.

	PE	CvM	NN	MGB	Energy	BG	CM	Ball
Location	1.000	1.000	0.139	0.056	0.043	0.048	1.000	0.036
Scale	0.550	0.555	0.311	0.601	0.302	0.217	0.108	0.577
Location-Scale	0.998	1.000	0.335	0.582	0.289	0.215	0.872	0.581

Table 3: The empirical powers for different methods when all the samples are generated from Cauchy distributions and the sample sizes are imbalanced at significance level $\alpha = 0.05$.

	PE	CvM	NN	MGB	Energy	BG	CM	Ball
Location	0.999	0.999	0.084	0.040	0.055	0.052	0.996	0.037
Scale	0.397	0.477	0.000	0.737	0.259	0.026	0.045	0.520
Location-Scale	1.000	1.000	0.000	0.857	0.372	0.089	0.943	0.728

3 Numerical Studies

Case 4: Normal distributions, $n_x = n_y = \{20,50,100\}, p = 10$.

Table 4: The average running time (in milliseconds) for different methods.

n	PE	CvM	NN	MGB	Energy	BG	CM	Ball
20	0.16	6.11	1.00	0.98	0.11	0.11	3.69	4.49
50	0.27	37.19	3.17	2.72	0.29	0.30	6.28	20.22
100	0.63	213.69	7.46	7.36	1.17	1.06	16.20	81.56

heavy computations

Summary

- our method is comparable with the projection-averaging based Cramér-von Mises test in terms of power performance,
- be superior to the other tests across almost all the cases, especially in the presence of the heavy-tailed distributions.
- more computationally efficient than the projection-averaging based Cramér-von Mises test .

Application

Dataset

UCI machine learning repository: Daily Demand Forecasting Orders Data Set

Question

inspect whether the demand on Friday is significantly different from other weekdays.

Features

Non urgent order (X_1),
Urgent order (X_2),
Three order types (X_3, X_4, X_5),
Fiscal sector orders (X_6),
Orders from the traffic controller sector(X_7),
Three kinds of banking orders (X_8, X_9, X_{10}),
Total orders (X_{11}).

Table 5. The empirical p-values for different methods for the daily demand forecasting orders data set.

	PE	CvM	NN	MGB
P-VALUE	0.008	0.004	0.249	0.180

	ENERGY	BG	CM	BALL
P-VALUE	0.011	0.210	0.554	0.112

Permutation 1000 times

Cauchy combination test statistic:

$$T = \frac{1}{8}\sum_{i=1}^{8} \tan\left\{(0.5 - p_i)\,\pi\right\}$$

$$p = 1/2 - \pi^{-1}\arctan(T).$$

- The corresponding p-value is 0.0164
- the demand on Friday is significantly different from other weekdays

$\alpha = 0.05$

4 Conclusion and Discussion

Conclusion

- We apply the idea of **projections** and propose a robust test for the multivariate two-sample problem.
- It is demonstrated that with **a suitable choice of the ensemble approach**, we can obtain a test, which is superior to most existing tests, especially in the presence of the heavy-tailed distributions.
- Moreover, it is **comparable** with the projection-averaging based Cramér-von Mises test in terms of power performance, but much **more efficient** in terms of **computation**.

Discussion

It's necessary to continue reducing the computational cost:

- In univariate cases, we can adopt AVL tree-type implementation to develop an efficient algorithm with complexity $O\{(m+n)\log(m+n)\}$
- In multivariate cases, we can approximate the test statistic with random projections, whose computational cost can be reduced to $O\{(m+n)K\log(m+n)\}$ and memory cost $O\{\max(m+n,K)\}$, where K is the number of random projections.

THANK YOU