Josua's Discussion

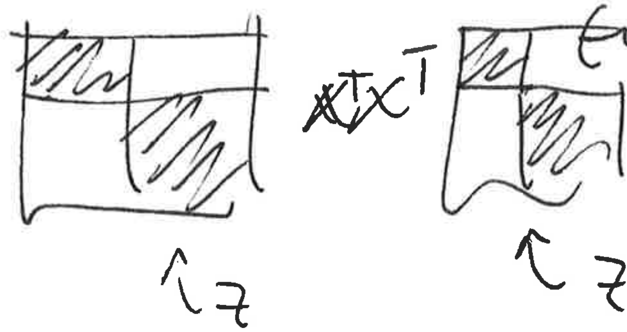
$A = \{x_1, x_2\}$ $\quad$ $T = -|x_1 - x_2|^2 - |x_3 - x_4|^2 + 2|x_1 - x_3|^2 + 2|x_2 - x_4|^2$

$B = \{x_3, x_4\}$

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

we want $T(X, Z) \overset{?}{=} \mathrm{Tr}(Z X X^T Z)$

quadratic function



$$T^* = \max_Z \mathrm{Tr}(Z X X^T Z)$$

$\boxed{\text{something like this.}}$

Jan 22, 2017

Write T in a quadratic form in terms of a matrix Z that partions the data points into the right classes.

Goal: $T^* = \max_{Z} Tr(Z^T D Z)$. ?

Thus test statistic ~ clustering.

$\boxed{\text{Paper: Energy statistics, Székely, Rizzo 2013}}$

E-stats: function of distances between statistical observations.

Typically more general and powerfull than classical alternatives such as correlation, F-stats, ...

Ex.: $$\begin{cases} V_n = \frac{1}{n^2} \sum_{i,j=1}^{n} h(X_i, X_j) \\ h(x_i, x_j) = h(|x_i - x_j|_d) \end{cases}$$ func. of Euclidean d-dimentional distance.

Székely proposed | Rotational invariant

$$2\int_{-\infty}^{\infty} (F(x) - G(x))^2 dx = 2\mathbb{E}|X - Y| - \mathbb{E}|X - X'| - \mathbb{E}|Y - Y'|$$

where $X \sim F$
$Y \sim G$

X' is a copy of X, $X' \overset{iid}{\sim} F$
Y' " Y, $Y' \overset{iid}{\sim} G$

It can be shown that this quantity is ≥ 0 and $= 0$ iff $X, Y \sim F$, identically distributed.

Energy distance: $\mathcal{E}(X,Y) = 2\mathbb{E}|X-Y|_d - \mathbb{E}|X-X'|_d - \mathbb{E}|Y-Y'|_d$

Prop: $$\mathcal{E}(X,Y) = \frac{1}{c_d}\int_{\mathbb{R}^d} \frac{|\hat{f}(t) - \hat{g}(t)|^2 dt}{|t|_d^{d+1}}$$

$$c_d = \frac{\pi^{(d+1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}$$

$\Rightarrow$ $\mathcal{E}(X,Y) \geq 0$ with equality iff X, Y are identically distributed.

Proof. Use characteristic function, and Parseval-Plancherel.

If X, Y live in a space with metric $\delta(X,Y)$ then,

$$\mathcal{E}(X,Y) \equiv 2\mathbb{E}\,\delta(X,Y) - \mathbb{E}\,\delta(X,X') - \mathbb{E}\,\delta(Y-Y')$$

carefull with the previous Prop. which may not hold.

Why special?
L_2 distance (weighted) is $\mathcal{E}$ under assumption of Rotation and scaling invariance.

- Look in the literature for: Maximum Mean Discrepancy (Kernel Methods).
- Klebanov (2005) N-distances and their Applications.

Testing equal distributions

H_0: X and Y have the same distr.
H_1: They don't.

$\{x_1, \dots, x_n\} \to$ sample from X
$\{y_1, \dots, y_m\} \to$ sample ~~from~~ Y

$$A = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \| x_i - y_j \|$$

$$B = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \| x_i - x_j \|$$

$$C = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \| y_i - y_j \|$$

E-statistic energy

$$\boxed{\mathcal{E}(X, Y) \equiv 2A - B - C}$$

$\to$ T-statistic test.

$$\boxed{T = \frac{n \cdot m}{n+m} \mathcal{E}(X, Y)}$$

Under H_0, $T \xrightarrow{d}$ quadratic form independent of Normal r.v. ?

Under H_1, $T \xrightarrow{d} \infty$

$$H = \frac{\mathcal{E}^2(X, Y)}{2\mathbb{E}\|X - Y\|} = \frac{2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|}{2\mathbb{E}\|X - Y\|}$$

$H = 0$ when $X \sim Y$.

This is normality test

From the paper: $H_0: F = F_0$, $H_1: F \neq F_0$ goodness-of-fit

$$\mathcal{E}(X, F_0) = \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}|x_i - X| - \mathbb{E}|X - X'| - \frac{1}{n^2} \sum_{l=1}^{n} \sum_{m=1}^{n} |x_l - x_m|$$

$n\mathcal{E} \to \infty$ for H_1
$n\mathcal{E} \to$ some distr for H_0.

$\alpha \in (0,2)$ <u>generalization</u> (4)

(1) $\mathcal{E}^{(\alpha)}(X,Y) = 2\mathbb{E}|X-Y|^\alpha - \mathbb{E}|X-X'|^\alpha - \mathbb{E}|Y-Y'|^\alpha$

(2) $\mathcal{E}^{(2)}(X,Y) = 2|\mathbb{E}X - \mathbb{E}Y|^2$

$\mathcal{E}^{(\alpha)}(X,Y) \geq 0$ with equality $\Leftrightarrow X,Y \sim F$.
This does not hold for (2), since it gives 0 if $\mathbb{E}X = \mathbb{E}Y$.

Conditionally negative function:
$z_{ij} \equiv x_i - y_j$. $\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} c_i \bar{c}_j \phi(z_{ij}) \leq 0$

$\forall\, c_i, c_j \in \mathbb{C}$, whenever $\sum_i c_i = 0$.

In this case we can replace $|X-Y| \to \phi(X,Y)$, its <u>better</u> if ϕ is <u>strictly</u> negative.

<u>Testing for equal distributions</u>

$X = \{x_1, x_2, \ldots, x_{n_1}\}$ iid random samples.
$Y = \{y_1, \ldots, y_{n_2}\}$

$$\mathcal{E}_{n_1,n_2}(X,Y) = \underbrace{\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}|x_i - y_m|}_{2A} - \underbrace{\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|x_i - x_j|}_{B} - \underbrace{\frac{1}{n_2^2}\sum_{l=1}^{n_2}\sum_{m=1}^{n_2}|y_l - y_m|}_{C} = 2A - B - C$$

$$T_{n_1,n_2} = \frac{n_1 n_2}{n_1+n_2} \mathcal{E}_{n_1,n_2}$$

H_0 : $X, Y \sim F$, small T
H_1 : X, Y diff. distn, large T

Distance Components (DISCO)

$H_0 : F_1 = F_2 = \ldots = F_k \quad k \geq 2.$

two samples $A = \{a_1, \ldots, a_{n_1}\}$
$B = \{b_1, \ldots, b_{n_2}\}$

let $g_\alpha(A,B) \equiv \frac{1}{n_1 n_2} \sum_{l=1}^{n_1} \sum_{m=1}^{n_2} |a_l - b_m|^\alpha \quad 0 < \alpha \leq 2.$

Let $A_1, A_2, \ldots, A_K$ be samples of sizes $n_1, n_2, \ldots, n_K$ and $\sum_{j=1}^{k} n_j = N$.

total dispersion

$$T_\alpha = T_\alpha(A_1, \ldots, A_K) = \frac{N}{2} g_\alpha(A, A)$$

A is the pooled sample of size N. I guess this means puting all the data from every sample into A.

The within-sample dispersion is

$$W_\alpha = W_\alpha(A_1, \ldots, A_K) = \sum_{j=1}^{k} \frac{n_j}{2} g_\alpha(A_j, A_j)$$

Between-sample energy statistics: ⑥

$$S_{n,\alpha} = \sum_{1 \le j < k \le K} \left(\frac{n_j + n_k}{2N}\right)\left[\frac{n_j n_k}{n_j + n_k} \mathcal{E}^{(\alpha)}_{n_j, n_k}(A_j, A_k)\right]$$

$$\sum_{j=1}^{K} \sum_{k=1}^{K}, \quad j<k$$

$$S_{n,\alpha} = \sum_{\substack{j=1 \\ j<k}}^{k} \sum_{k=1}^{k} \frac{n_j n_k}{2N}\Big(2 g_\alpha(A_j, A_k) - g_\alpha(A_j, A_j) - g_\alpha(A_k, A_k)\Big)$$

For $0 < \alpha \le 2$ we have $T_\alpha = S_\alpha + W_\alpha$

If $0 < \alpha < 2$ we have a statistically consistent test of equality of distr.

If $\alpha = 2$, ~~then~~ $\mathcal{E}$ can be zero if the means of the distributions are identical.

Paper: Equivalence of Distance-Based And RKHS-Based Statistics in Hypothesis Testing, Ann. Stat (2013)

energy distance
dist. covariance ⟷ maximum mean discrepancy (MMD)

(statistics)

this is a distance between embeddings of distr. to reproducing kernel Hilbert spaces (RKHS). (Mach. Learning)

To any positive definite kernel, MMD is "kinda" equivalent to Energy distance.

⑦

(*) They show that E-dist. most commonly used in stats. is just one member of a parametric family of kernels, and other choices can yield more powerful tests.

E-stats (statistics)

two-sample testing in Euclidean space
Szekely, Rizzo 2004, 2005
Baringhaus, Franz 2004
Dependence measure: dist covariance

link ?

RKHS (Mach. Learning)

Embeddings of prob. distr. into reproducing kernel Hilbert spaces.
Test stat: ~~dist. between~~ difference between embeddings.
Maximum Discrepancy (MMD).
Gretton et al. (2007, 2012)

Seems to be more general and recover E-stats in some limit?

two-sample testing: E-dist is a maximum mean discrepancy.

E-dist arise from a particular choice of kernel!

Semi-Metric of Negative Type:

Z is a set. $\rho: Z \times Z \to [0, \infty)$ s.t. $\rho(z, z') = 0 \iff z = z'$, and $\rho(z, z') = \rho(z', z)$. (Z, ρ) is a semi-metric space.

Negative type: $\sum_{i=1}^{n} \alpha_i = 0$, $\sum_{i,j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$

Prop. If ρ is of Negative type, so is ρ^q for $q \in (0,1)$. ρ is a semimetric of Nega. type $\Leftrightarrow$ $\exists$ a Hilbert space $\mathcal{H}$ and $\varphi : \mathcal{Z} \to \mathcal{H}$ (injective) s.t. $\rho(z, z') = \|\varphi(z) - \varphi(z')\|_{\mathcal{H}}^2$

So $\rho^{1/2}$ is a metric, even though ρ is a semimetric.

Energy Distance: measure of stat. distance between two prob. measures P and Q on $\mathbb{R}^d$:

$$D_E(P, Q) = 2\mathbb{E}\|Z - W\|_2 - \mathbb{E}\|Z - Z'\|_2 - \mathbb{E}\|W - W'\|_2$$

$Z, Z' \overset{iid}{\sim} P$ (i) $D_E(P,Q) \geq 0$

$W, W' \overset{iid}{\sim} Q$ (ii) $D_E(P,Q) > 0$ if $P \neq Q$.

This can be generalized to

$$D_{E,\rho}(P, Q) = 2\mathbb{E}\rho(Z, W) - \mathbb{E}\rho(Z, Z') - \mathbb{E}\rho(W, W')$$

ρ must satisfy certain conditions so $D_{E,\rho}$ is a metric.

Kernel-Based Approach:

RKHS. $\mathcal{H}$ is a Hilbert space of real valued functions defined on a set $\mathcal{Z}$. $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if:

1. $k(\cdot, z) \in \mathcal{H}, \forall z \in \mathcal{Z}$.
2. $\langle f, k(\cdot, z)\rangle_{\mathcal{H}} = f(z), \forall z \in \mathcal{Z}, \forall f \in \mathcal{H}$.

If $\mathcal{H}$ has such a k then $\mathcal{H}$ is a reproducing kernel Hilbert space.

<u>Theo.</u> (Moore–Aronszajn) To every symmetric, positive definite function $k: Z \times Z \to \mathbb{R}$, there is an RKHS $\mathcal{H}_k$ of real valued functions.
$\varphi: Z \to \mathcal{H}_k$, $\varphi: z \mapsto k(\cdot, z)$ is the canonical feature map.

<u>Def.:</u> Let k be a kernel on Z, and ν a prob measure on Z. The kernel embedding of ν into RKHS is $\mu_k(\nu) \in \mathcal{H}_k$ s.t.
$$\int f(z)\, d\nu(z) = \langle f, \mu_k(\nu) \rangle_{\mathcal{H}_k} \quad \forall f \in \mathcal{H}_k.$$

<u>Def.:</u> (MMD) Let k be a kernel on Z. The MMD γ_k between P, Q prob. measures on Z. The MMD γ_k between P and Q is $\gamma_k(P, Q) \equiv \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}$

usefull formula:
$$\gamma_k^2(P, Q) = \mathbb{E}\, k(Z, Z') + \mathbb{E}\, k(W, W') - 2\mathbb{E}\, k(Z, W)$$

$Z, Z' \overset{iid}{\sim} P$
$W, W' \overset{iid}{\sim} Q$.

if the restriction of μ_k to some prob space is well defined and injective γ_k is a metric.

<u>Lemma</u> Let $\rho: Z \times Z \to \mathbb{R}$ be a semimetric. Let $z_0 \in Z$, and denote
$$k(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z')$$

Then k is positive definit iff ρ is conditionally negative, or a semimetric of negative type.

Distance-induced kernel:

$$k(z,z') = \frac{1}{2}\{\rho(z,z_0) + \rho(z',z_0) - \rho(z,z')\}$$

also just called distance kernel.

Prop. 1. $\rho(z,z') = k(z,z) + k(z',z') - 2k(z,z')$

$$= \| k(\cdot,z) - k(\cdot,z') \|^2_{\mathcal{H}_k}$$

2. $z \mapsto k(\cdot,z)$ is injective

Ex.: $\mathcal{Z} \subseteq \mathbb{R}^d$, $\rho_q(z,z') = \|z-z'\|^q \quad 0 < q \leq 2.$

$$k_q(z,z') = \frac{1}{2}\{\|z-z_0\|^q + \|z'-z_0\|^q - \|z-z'\|^q\}$$

if $z_0 = 0 \Rightarrow k_q(z,z') = \frac{1}{2}\{\|z\|^q + \|z'\|^q - \|z-z'\|^q\}$

Let k be a nondegenerate kernel on $\mathcal{Z}$. The

$$\rho(z,z') = k(z,z') + k(z',z) - 2k(z,z')$$

defines a valid semimetric ρ of negative type.

k generates ρ.

if k and k' generates the same ρ, k and k' are equivalent.

Prop. k and $\tilde{k}$ are equivalent iff

$$\tilde{k}(z,z') = k(z,z') + f(z) + f(z')$$

for some shift function $f: \mathcal{Z} \to \mathbb{R}$.

Equivalence of MMD and Energy Distance

For every ρ, $D_{E,\rho}$ is related to the MMD associated to a kernel k that generates ρ.

Theo. (Z,ρ) is a semimetric space of negative type and k is a kernel that generates ρ. Then

$$D_{E,\rho}(P,Q) = 2\gamma_k^2(P,Q)$$

Recall $\gamma_k^2(P,Q) = \mathbb{E}k(z,z') + \mathbb{E}k(w,w') - 2\mathbb{E}k(z,w)$

$Z = \{z_i\}_{i=1}^m \overset{iid}{\sim} P$
$W = \{w_i\}_{i=1}^n \overset{iid}{\sim} Q$ • The empirical estimate is

$$\hat{\gamma}_k(Z,W) = \gamma_k^2\left(\frac{1}{m}\sum_{i=1}^m \delta_{z_i}, \frac{1}{n}\sum_{j=1}^n \delta_{w_j}\right)$$

$$= \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(z_i,z_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(w_i,w_j)$$

$$- \frac{2}{m.n}\sum_{i=1}^m\sum_{j=1}^n k(z_i,w_j)$$

Theo. $Z = \{z_i\}_{i=1}^m$, $W = \{W_i\}_{i=1}^m$ are two iid samples from P. Assume $S_{\tilde{k}_P}$ is trace class

$$\frac{m}{2}\hat{\gamma}_k^2(Z,W) \rightsquigarrow \sum_{i=1}^\infty \lambda_i N_i^2$$

where $N_i \overset{iid}{\sim} N(0,1)$ and λ_i are the eigenvalues of $S_{\tilde{k}_P}$.

Two-sample Experiment

Two Multivariate Gaussians, where the
(1) means differ in 1D only and all variances are equal.

(2) Two Gaussians but with identical means, but variance that differ in a single dimension.

(3) One distr. is a Gaussian in 1D, and the other is a Gaussian in 1D with a sinusoidal perturbation of increasing frequency.

Exponent $q < 1$ seems to give good results in (3).

Conclusion: Energy distance is a particular case of a larger class of discrepancy measures.

Kernel based methods can be applied to data that do not lie in E^n.

A New test for Multivariate Normality
Székely, Rizzo (2005)

New class of rotation invariant and consistent goodness-of-fit tests for multivariate distributions based on Euclidean distance between sample elements.

The most widely applied tests of normality are based on Mardia's multivariate generalization of skewness and kurtosis.

Let $X_1, \dots, X_n$ be a d-dimensional sample.
T_n is affine invariant if $T_n(A(X_1), \dots, A(X_n)) = T_n(X_1, \dots, X_n)$

for every affine transformation $A: \mathbb{R}^d \to \mathbb{R}^d$.
A goodness-of-fit test of $H_0: F \in \mathcal{F}$ versus $H_1: F \notin \mathcal{F}$ is consistent against all fixed alternatives if the probability of rejecting the null hypothesis $\to 1$ when $n \to \infty$ and the actual distribution of the sampled population is not in $\mathcal{F}$.

V-statistic with kernel:

$$h(x, y) = \mathbb{E}\|x - Y\| + \mathbb{E}\|y - Y\| - \mathbb{E}\|Y - Y'\| - \|x - y\|$$

$\|x\| = (x^T x)^{1/2}$ Euclidean

$X, X' \overset{iid}{\sim} F$, Y is a d-dimensional random vector.

$$\mathcal{E}_n = \frac{1}{n} \sum_{j,k=1}^{n} h(X_j, X_k).$$

$\frac{\mathcal{E}_n}{n}$ is a degenerate V-statistic.

If X, X', Y, Y' are independent then

$$\mathbb{E}\, h(X, Y) = \underbrace{2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|}_{\geq 0}$$

with equality iff $X, Y \sim F$.

S is a non-empty set. $\gamma : S \times S \to \mathbb{R}$ is negotive definite if (14)

$$\sum_{j,k=1}^{n} \gamma(x_j, x_k) r_j r_k \leq 0$$

where $\sum_{j=1}^{n} r_j = 0$. Continuous analogue is

$$\iint_{S\,S} dQ(x)\, dQ(y)\, \gamma(x,y)\, r(x) r(y) \leq 0.$$

when $\int dQ(x)\, r x = 0$ for some measure Q.

Prop. If $\gamma(x,y) = \|x-y\|$ is the Euclidean dist. then it γ is strictly negotive.

Thm. $2\mathbb{E}\gamma(X,Y) - \mathbb{E}\gamma(X,X') - \mathbb{E}\gamma(Y,Y') \geqslant 0$ iff γ is negative definite. if γ is strictly negotive the equality holds ~~iff~~ $X \overset{d}{=} Y$.

Proof. Let $X \sim \mu$, $Y \sim \nu$. Let Q be a arbitrary prob measure. ~~Define~~ $r(x) = \frac{d\mu(x)}{dQ} - \frac{d\nu(x)}{dQ}$

We have

$$\int \gamma(x,y)\, d\mu(x) d\nu(y) + \int \gamma(x,y) d\mu(x) d\nu(y) - \int \gamma(x,y) d\mu(x) d\mu(x) - \int \gamma(x,y) d\nu(x) d\nu(y) = -\int \gamma(x,y) \cdot r(x) r(y)\, dQ(x) dQ(y) \geqslant 0!$$

$$\cancel{\frac{d\mu(x) d\nu(y)}{dQ\, dQ} - \frac{d\nu(x) d\nu(y)}{dQ\, dQ} = r(x) \cdot r(y)}$$

$$r(x) \cdot r(y) = \left(\frac{d\mu(x)}{dQ} - \frac{d\nu(x)}{dQ}\right)\left(\frac{d\mu(y)}{dQ} - \frac{d\nu(y)}{dQ}\right)$$
$$= \frac{d\mu(x) d\mu(y)}{dQ\, dQ} - \frac{d\mu(x) d\nu(y)}{dQ\, dQ} - \frac{d\nu(x) d\mu(y)}{dQ\, dQ} + \frac{d\nu(x) d\nu(y)}{dQ\, dQ}$$

We can use $\mathcal{J} = \|\cdot\|_d$ in $\mathbb{R}^d$ the equality implies $X \stackrel{d}{=} Y$.

Suppose $X_1, \dots, X_n$ is a random sample from F, and $x_1, \dots, x_n$ are the observed values of the random sample. $H_0: F = F_0$ vs. $H_1: F \neq F_0$.

$$\mathcal{E}_n = n\left(\frac{2}{n}\sum_{j=1}^{n} \mathbb{E}\|x_j - X\| - \mathbb{E}\|X - X'\| - \frac{1}{n^2}\sum_{j,k=1}^{n} \|x_j - x_k\|\right)$$

$X, X' \stackrel{iid}{\sim} F_0$.

If $F_0 = N_d(\mu, \Sigma)$ then $y_j \equiv \Sigma^{-1/2}(x_j - \mu)$ and use

$$\mathcal{E}_n = n\left(\frac{2}{n}\sum_{j=1}^{n} \mathbb{E}\|y_j - Z\| - \mathbb{E}\|Z - Z'\| - \frac{1}{n^2}\sum_{j,k=1}^{n}\|y_j - y_k\|\right)$$

where $Z, Z' \stackrel{iid}{\sim} N_d(0, I)$

It's possible to find formulas for $\mathbb{E}[\dots]$.

Then one estimates

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^T$$

$$Y_j = \hat{\Sigma}^{-1/2}(X_j - \hat{\mu})$$

The null hypothesis is rejected if $\mathcal{E}_n$ is large.

Projection Pursuit (PP): $X \in \mathbb{R}^d \sim N_d(\mu, \Sigma)$
iff $a^T X \sim N_1(a^T\mu, a^T \Sigma a)$ for all $a \in \mathbb{R}^d$.
The PP tests the "worst" 1D projection with a goodness-of-fit method.

$$C_n^* = \sup_{\substack{a \in \mathbb{R}^d \\ \|a\|=1}} C_n(a^T x_1, \ldots, a^T x_n)$$

reject the H_0 if C_n^* is large.

C_n^* can be approximated by finding a finite set of projections determined by $\{a_1, a_2, \ldots, a_m\}$ uniformly scattered in $\mathbb{R}^d$.

Testing for equal Distributions in High Dimensions

Székely, Rizzo 2004

Nonparametric test for equality of two or more multivariate distr. based on Euclidean distance. Distr. are not specified. Test is consistent. Prob. more sensitive than nearest neighbors and performs well in high dim. Computational complexity indep. of dimension and number of data points!

Kolmogorov-Smirnov
Crámer-von Mises → do not generalize to D>1 well in a "distribution free" manner.
...
1D

Suppose $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ are independ. random samples of $\mathbb{R}^d$, according to F_1 and F_2.

two-sample prob: $H_0: F_1 = F_2$
$H_1: F_1 \neq F_2$

n-samples prob: $H_0: F_1 = \dots = F_n$
$H_1: F_i \neq F_j$ for some $i \neq j$

{Empirically: $\mathcal{E}$-test way superior than nearest neighbor in high dimensions.}

$A = \{a_1, \dots, a_{n_1}\}$ disjoint non-empty sets of $\mathbb{R}^d$
$B = \{b_1, \dots, b_{n_2}\}$

$$\mathcal{E}(A,B) \equiv \frac{n_1 n_2}{n_1+n_2}\left\{ \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|a_i - b_j\| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|a_i - a_j\| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|b_i - b_j\| \right\} \geq 0$$

- See Székely, Rizzo (2003), Hierarchical clustering via joint between-within distances.

Two-sample prob.

$A = \{X_1, \dots, X_{n_1}\}$ $\quad \mu_{AB} = \mathbb{E}\|X - Y\|$
$B = \{Y_1, \dots, Y_{n_2}\}$ $\quad \mu_A = \mathbb{E}\|X_1 - X_2\|$
$\mu_B = \mathbb{E}\|Y_1 - Y_2\|$

$$\mathbb{E}[\mathcal{E}(A,B)] = \frac{n_1 n_2}{n_1+n_2}\left\{ \frac{2}{\cancel{n_1 n_2}}\cancel{n_1 n_2}\,\mu_{AB} - \frac{1}{n_1^2}\cancel{n_1^2}\,n_1(n_1-1)\,\mu_A - \frac{1}{n_2^2}\cancel{n_2^2}\,n_2(n_2-1)\,\mu_B \right\}$$

$$\mathbb{E}[\mathcal{E}(A,B)] = \frac{n_1 n_2}{n_1+n_2}\left\{2\mu_{AB} - \frac{n_1-1}{n_1}\mu_A - \frac{n_2-1}{n_2}\mu_B\right\}$$

$$= \frac{n_1 n_2}{n_1+n_2}\left\{2\mu_{AB} - \mu_A - \mu_B\right\} + \frac{n_2}{n_1+n_2}\mu_A + \frac{n_1 \mu_B}{n_1+n_2}$$

if $X \overset{D}{=} Y \Rightarrow \mu_{AB} = \mu_A = \mu_B$, $\quad = 0 \Rightarrow \mathbb{E}[\mathcal{E}(A,B)] = \mu_{AB}$

if $X \overset{D}{\neq} Y \Rightarrow \quad \neq 0 \Rightarrow \cancel{\mathbb{E}[\mathcal{E}(A,B)]}$

$$\mathbb{E}[\mathcal{E}(A,B)] = \frac{n_1 n_2}{n} C + \underbrace{\frac{n_2}{n}}_{c_2}\mu_A + \underbrace{\frac{n_1}{n}}_{c_1}\mu_B$$

Fixed ratio

$$= c_1 c_2\, n\, C$$

so $\mathbb{E}[\mathcal{E}(A,B)]$ is $\sim n$ (assymptotically)

as $n \to \infty$, under H_0, $\mathbb{E}[\mathcal{E}(A,B)] \to$ cte

under H_1, $\mathbb{E}[\mathcal{E}(A,B)] \to \infty$

Not only $\mathbb{E}[\mathcal{E}]$ but $\mathcal{E}$ itself either converges or diverges. (in distr.)

conditional on data

$$\mathcal{E}(A,B) = \frac{n_1 n_2}{n} C$$

$$= \frac{n_1}{n} \frac{n_2}{n} \cdot n \cdot C$$

fixed ratios (assymptotically).

$$C = 2\mu_{AB} - \mu_A - \mu_B = 0 \iff X \overset{D}{=} Y$$

Two-sample test stats.

$$\mathcal{E}_{n_1,n_2} = \frac{n_1 n_2}{n_1+n_2}\left(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}\|X_i - Y_m\| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|X_i - X_j\| - \frac{1}{n_2^2}\sum_{l=1}^{n_2}\sum_{m=1}^{n_2}\|Y_l - Y_m\|\right)$$

under H_0 or H_1, a random permutation

$$W_1^{(\pi)}, \dots, W_n^{(\pi)} \qquad n_1+n_2=n$$

of $X_1,\dots,X_{n_1}, Y_1,\dots Y_{n_2}$, is equal in distribution to a random sample of size n from the mixture W where W samples from X with prob. $\frac{n_1}{n}$ and from Y with prob $\frac{n_2}{n}$.

$$\lim_{n\to\infty} P(\mathcal{E}_n > c_\alpha) = \alpha \in (0,1)$$

Reject H_0 if $\mathcal{E}_n > c_\alpha$

Implementation

- Permutation test Approach, see Efron 1993, chap 15 An introduction to the bootstrap.

$$\begin{matrix} A_1, & A_2, & \dots, & A_k \\ \sim & \sim & & \sim \\ F_1, & F_2, & \dots, & F_k \end{matrix} \xrightarrow[\text{sample}]{\text{pooled}} \{W_1,\dots,W_n\} = A_1 \cup \dots \cup A_k$$

$$n = \sum n_i$$

under H_0, $W_1,\dots,W_n$ are iid, with distr. F.
If the desired significance level is α,

resample (without replacement) from $W_1, \ldots, W_n$

B samples of size n so that $(B+1)\alpha$ is an integer.

$$m_j = \sum_{i=1}^{j} n_i \qquad m_0 = 0.$$

$\{W_1^{(b)}, \ldots, W_n^{(b)}\} \to$ one bootstrap sample.
$b = 1 \ldots B$

Compute $\mathcal{E}_n^{(b)}$ from $A_i^{(b)} = \{W_{m_{i-1}+1}^{(b)}, \ldots W_{m_i}^{(b)}\}$
$i = 1 \ldots K$

The bootstrap estimate of $P_n(\mathcal{E}_n \leq t)$ is

$$\frac{1}{B} \sum_{b=1}^{B} I(\mathcal{E}_n^{(b)} \leq t)$$

Reject H_0 if the observed $\mathcal{E}_n$ exceeds $100(1-\alpha)\%$ of the replicates $\mathcal{E}_n^{(b)}$