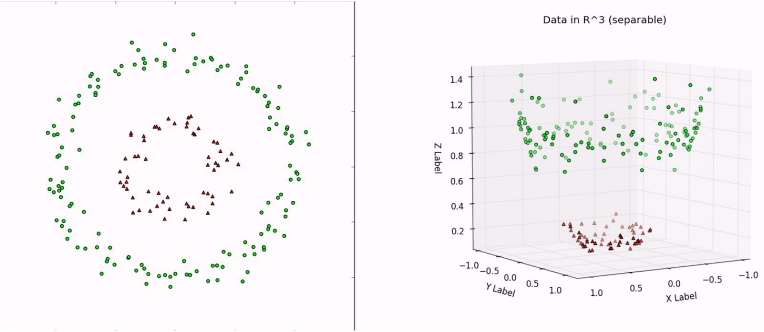
A New Framework for Distance-based Metrics in High Dimension

Xianyang Zhang

Going Beyond Linearity

In many statistical and machine learning problems, it is important to extract nonlinear features.

- Spline, trees, neural networks, **kernel tricks**...



Embedding

Measuring distances between probability distributions through kernel embedding:

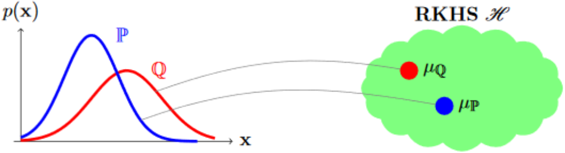


Figure 1.1: Embedding of marginal distributions: Each distribution is mapped into an RKHS via an expectation operation.

$\mathbb{P}, \mathbb{Q} \quad \overset{\text{embedding}}{\Longrightarrow} \quad \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}$ in some RKHS

$$\mathbb{P} = \mathbb{Q} \quad \Longleftrightarrow \quad \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2 = 0$$

$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2$ typically has a nice and simple expression.

Distance and Kernel-based Methods

- Testing for similarities in two datasets
  - **energy distance**, maximum mean discrepancy
- Determine strength of dependence ($\mathbb{P} = \mathbb{P}_{XY}$ and $\mathbb{Q} = \mathbb{P}_X\mathbb{P}_Y$)
  - **distance covariance**, Hilbert-Schmidt independence criterion
- Measuring strength of
  - *conditional dependence*
  - *mutual dependence*
  - *interaction dependence*
  - *conditional mean/quantile dependence*

Distance and Kernel-based Metrics

- Distance and kernel-based measures are getting popular.

- Can be applied to a wide range of statistical problems.

Applications

- Signal detection
- High-dimensional feature screening
- Undirected/directed graph modeling
- Change-point detection
- Independent component analysis
- Dimension reduction
- Time series analysis
- Training Generative Adversarial Network
- ......

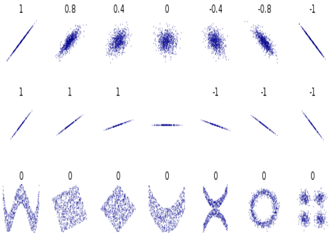
This Talk

For two high dimensional random vectors **X** and **Y**, we are interested in studying the behaviors of

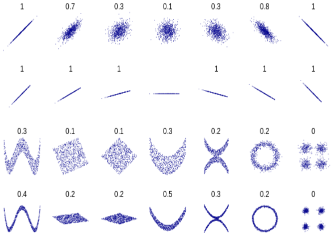
- Dependence metrics: $\|\mu_{\mathbb{P}_{\mathbf{XY}}} - \mu_{\mathbb{P}_{\mathbf{X}}\mathbb{P}_{\mathbf{Y}}}\|^2$
- Homogeneity metrics: $\|\mu_{\mathbb{P}_{\mathbf{X}}} - \mu_{\mathbb{P}_{\mathbf{Y}}}\|^2$

Distance Covariance

Distance Covariance (Székely, Rizzo and Bakirov, 2007) is a powerful measure of dependence between two random vectors of *arbitrary but fixed dimensions*.



(a) Correlation



(b) Distance Correlation

Distance Covariance and Correlation

Consider two random vectors $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$ with $\mathbb{E}\|\mathbf{X}\|_p < \infty$ and $\mathbb{E}\|\mathbf{Y}\|_q < \infty$, where $\|\cdot\|_p$ denotes the Euclidean norm of $\mathbb{R}^p$.

(Squared) Distance Covariance

$$dCov^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{\mathbf{X},\mathbf{Y}}(t, s) - f_{\mathbf{X}}(t) f_{\mathbf{Y}}(s)|^2}{\|t\|_p^{p+1} \|s\|_q^{q+1}} dtds,$$

where $f_{\mathbf{X}}$, $f_{\mathbf{Y}}$ and $f_{\mathbf{X},\mathbf{Y}}$ are the individual and joint characteristic functions of $\mathbf{X}$ and $\mathbf{Y}$. $dCov(\mathbf{X}, \mathbf{Y}) = 0$ **if and only if** $\mathbf{X}$ and $\mathbf{Y}$ are independent.

(Squared) Distance Correlation

$$dCor^2(\mathbf{X}, \mathbf{Y}) = \frac{dCov^2(\mathbf{X}, \mathbf{Y})}{\sqrt{dCov^2(\mathbf{X}, \mathbf{X}) dCov^2(\mathbf{Y}, \mathbf{Y})}}.$$

An Integration Formula and Alternative expression

- **Integration formula**:

$$\int_{\mathbb{R}^p} \frac{1-\cos(t^\top x)}{c_p \|t\|_p^{p+1}} dt = \|x\|_p.$$

- **Alternative expression**:

$$\begin{aligned} dCov^2(\mathbf{X},\mathbf{Y}) =& \mathbb{E}\|\mathbf{X}-\mathbf{X}'\|_p\|\mathbf{Y}-\mathbf{Y}'\|_q - 2\mathbb{E}\|\mathbf{X}-\mathbf{X}'\|_p\|\mathbf{Y}-\mathbf{Y}''\|_q \\ &+ \mathbb{E}\|\mathbf{X}-\mathbf{X}'\|_p\mathbb{E}\|\mathbf{Y}-\mathbf{Y}'\|_q \\ =& \mathbb{E}U_{\mathbf{X}}(\mathbf{X},\mathbf{X}')U_{\mathbf{Y}}(\mathbf{Y},\mathbf{Y}'), \end{aligned}$$

where $(\mathbf{X}',\mathbf{Y}')$ and $(\mathbf{X}'',\mathbf{Y}'')$ are independent copies of $(\mathbf{X},\mathbf{Y})$, and

$$U_{\mathbf{X}}(\mathbf{x},\mathbf{x}') = \mathbb{E}\|\mathbf{x}-\mathbf{X}'\|_p + \mathbb{E}\|\mathbf{X}-\mathbf{x}'\|_p - \|\mathbf{x}-\mathbf{x}'\|_p - \mathbb{E}\|\mathbf{X}-\mathbf{X}'\|_p,$$

for $\mathbf{x},\mathbf{x}' \in \mathbb{R}^p$.

Estimation

Unbiased Estimator

Given $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})$ and $\mathbf{Y}_i = (Y_{i,1}, \ldots, Y_{i,q})$ for $1 \leq i \leq n$, define

$$\widehat{dCov^2}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n(n-3)} \sum_{k \neq l} \widehat{U}_{\mathbf{X}}(\mathbf{X}_k, \mathbf{X}_l) \widehat{U}_{\mathbf{Y}}(\mathbf{Y}_k, \mathbf{Y}_l),$$

where $\widehat{U}_{\mathbf{X}}(\mathbf{X}_k, \mathbf{X}_l)$ is the sample version of $U_{\mathbf{X}}(\mathbf{X}_k, \mathbf{X}_l)$.

When p and q are fixed and $n \to +\infty$:

- $\widehat{dCov^2}$ is an **unbiased** and **strongly consistent** estimator for $dCov^2$.

Question: What happens if p and q are large?

Intuition: Euclidean Distance in High Dimension

- As $p \to +\infty$,

$$\frac{\|\mathbf{X}-\mathbf{X}'\|_p}{\tau_{\mathbf{X}}} = \sqrt{1+\frac{\|\mathbf{X}-\mathbf{X}'\|_p^2-\tau_{\mathbf{X}}^2}{\tau_{\mathbf{X}}^2}}$$
$$\approx 1+\frac{\|\mathbf{X}-\mathbf{X}'\|_p^2-\tau_{\mathbf{X}}^2}{2\tau_{\mathbf{X}}^2} + \text{Remainder term},$$

  where $\tau_{\mathbf{X}}^2 = \mathbb{E}\|\mathbf{X}-\mathbf{X}'\|_p^2$.

- Euclidean distance behaves as **the squared Euclidean distance.**

- Squared Euclidean distance fails to capture **nonlinearity** as

$$U_{\mathbf{X}}(\mathbf{x},\mathbf{x}') \approx \frac{(\mathbf{x}-\mathbb{E}[\mathbf{X}])^\top(\mathbf{x}'-\mathbb{E}[\mathbf{X}])}{\tau_{\mathbf{X}}},$$
$$dCov^2(\mathbf{X},\mathbf{Y}) = \mathbb{E}U_{\mathbf{X}}(\mathbf{X},\mathbf{X}')U_{\mathbf{Y}}(\mathbf{Y},\mathbf{Y}') \approx \frac{1}{\tau_{\mathbf{X}}\tau_{\mathbf{Y}}}\|\mathrm{cov}(\mathbf{X},\mathbf{Y})\|_F^2.$$

Asymptotic Expansions

Population dCov

Under some assumptions and as $p, q \to +\infty$,

$$dCov^2(\mathbf{X}, \mathbf{Y}) = \underbrace{\frac{1}{\tau} \sum_{i=1}^{p} \sum_{j=1}^{q} cov^2(X_i, Y_j)}_{\text{leading term}} + \text{Remainder term},$$

where *cov* denotes the covariance, and $\tau = \tau_{\mathbf{X}} \tau_{\mathbf{Y}} > 0$.

Ratio of $dCov^2(\mathbf{X}, \mathbf{Y})$ and the leading term:

$p = 20$	$p = 40$	$p = 60$	$p = 80$	$p = 100$
0.980	0.993	0.994	0.989	0.997

Asymptotic Expansions

Sample dCov

Under mild assumptions, as $p, q \to +\infty$ and n is fixed or growing slowly,

$$\widehat{dCov^2}(\mathbf{X}, \mathbf{Y}) = \frac{1}{\tau} \sum_{i=1}^{p} \sum_{j=1}^{q} \widehat{cov^2}(X_i, Y_j) + \text{Remainder term},$$

where $\widehat{cov^2}$ denotes the unbiased estimator of the squared covariance.

Fixed p, Growing q

If $\mathbf{X}$ is low-dimensional and $\mathbf{Y}$ is high-dimensional (fixed p, growing q),

$$
\begin{aligned}
dCov^2(\mathbf{X},\mathbf{Y}) &\approx \frac{1}{\tau_{\mathbf{Y}}}\mathbb{E}[U_{\mathbf{X}}(\mathbf{X},\mathbf{X}')(\mathbf{Y}-\mathbb{E}[\mathbf{Y}])^\top(\mathbf{Y}'-\mathbb{E}[\mathbf{Y}])] \\
&= \frac{1}{\tau_{\mathbf{Y}}}\sum_{j=1}^{q}\underbrace{\mathbb{E}[U_{\mathbf{X}}(\mathbf{X},\mathbf{X}')(Y_j-\mathbb{E}[Y_j])(Y_j'-\mathbb{E}[Y_j])]}_{MDD^2(Y_j|\mathbf{X})}.
\end{aligned}
$$

Martingale difference divergence (Shao and Zhang, 2014)

$MDD^2(Y_j|\mathbf{X})$ denotes the (squared) martingale difference divergence which characterizes the *conditional mean dependence* of Y_j given $\mathbf{X}$ in the sense that

$$E[Y_j|\mathbf{X}] = E[Y_j] \text{ almost surely} \quad \text{iff} \quad MDD^2(Y_j|\mathbf{X}) = 0.$$

Curse of Dimensionality

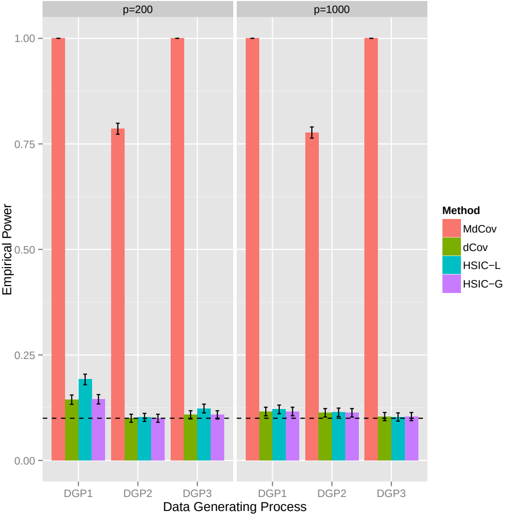
Arbitrary dependence *Conditional mean dependence*

$$
\begin{tikzcd}[column sep=huge, row sep=huge]
dCov^2(\mathbf{X}, \mathbf{Y}) \arrow[r, "q \to +\infty"] \arrow[d, "p \to +\infty"'] \arrow[dr, "{p, q \to +\infty}"] & \sum_{j=1}^{q} MDD^2(Y_j|\mathbf{X}) \arrow[d, "p \to +\infty"] \\
\sum_{i=1}^{p} MDD^2(X_i|\mathbf{Y}) \arrow[r, "q \to +\infty"'] & \sum_{i=1}^{p} \sum_{j=1}^{q} \mathrm{cov}^2(X_i, Y_j)
\end{tikzcd}
$$

Conditional mean dependence *Linear dependence*

Numerical Evidence

Two high-dimensional vectors **X** and **Y** are nonlinearly dependently.



New Metrics for Euclidean Space

- A metric space $(\mathcal{X}, \rho)$ is said to have negative type if for all $n \geq 1$, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{R}$ with $\sum_{i=1}^n a_i = 0$, we have

$$\sum_{i,j=1}^{n} a_i a_j \rho(\mathbf{x}_i, \mathbf{x}_j) \leq 0.$$

  Suppose $P, Q \in \mathcal{M}_1(\mathcal{X})$ with finite first moments. When $(\mathcal{X}, \rho)$ has negative type,

$$\int_{\mathcal{X}} \rho(x_1, x_2)\, d(P - Q)^2(x_1, x_2) \leq 0. \tag{1}$$

  We say that $(\mathcal{X}, \rho)$ has strong negative type if it has negative type and the equality in (1) holds only when $P = Q$.

- $(\mathbb{R}^p, \|\cdot - \cdot\|_p)$ is of strong negative type.

- Replacing $\|\cdot - \cdot\|_p$ by any $\rho(\cdot, \cdot)$ that is of strong negative type preserves the property: $dCov_\rho(\mathbf{X}, \mathbf{Y}) = 0$ **if and only if X** and **Y** are independent.

New Metrics for Euclidean Space

A new class of distances

For $\mathbf{x} \in \mathbb{R}^p$, we partition $\mathbf{x}$ into R subvectors:

$$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_R), \quad \mathbf{x}_i \in \mathbb{R}^{p_i}.$$

Let ρ_i be a metric of strong negative type on $\mathbb{R}^{p_i}$. Define

$$K_{\mathbf{p}}(\mathbf{x}, \mathbf{x}') = \sqrt{\rho_1(\mathbf{x}_1, \mathbf{x}_1') + \cdots + \rho_R(\mathbf{x}_R, \mathbf{x}_R')},$$

where $\mathbf{p} = (p_1, \ldots, p_R)$.

- $K_{\mathbf{p}}$ is a metric of strong negative type on $\mathbb{R}^p$.
- If $p_i = 1$ and $\rho_i(x_i, x_i') = |x_i - x_i'|$, then $K_{\mathbf{p}}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_1^{1/2}$.
- If $\rho_i(\mathbf{x}_i, \mathbf{x}_i') = \|\mathbf{x}_i - \mathbf{x}_i'\|^2$, then K becomes the usual Euclidean distance. However, ρ_i is no longer a metric (but a semi-metric).
- Euclidean distance does not belong to the above class.

Properties of Generalized dCov

Generalized dCov

Given the metric K introduced above, define

$$GdCov^2(\mathbf{X},\mathbf{Y}) = \mathbb{E}K_{\mathbf{p}}(\mathbf{X},\mathbf{X}')K_{\mathbf{q}}(\mathbf{Y},\mathbf{Y}') - 2\mathbb{E}K_{\mathbf{p}}(\mathbf{X},\mathbf{X}')K_{\mathbf{q}}(\mathbf{Y},\mathbf{Y}'') \\ + \mathbb{E}K_{\mathbf{p}}(\mathbf{X},\mathbf{X}')\mathbb{E}K_{\mathbf{q}}(\mathbf{Y},\mathbf{Y}').$$

for two random vectors $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$

- $GdCov^2(\mathbf{X},\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.
- An estimator for $GdCov^2$, denoted by $\widehat{GdCov}^2$, can be constructed by replacing the usual Euclidean distance by the new distance K in $\widehat{dCov}^2$.
- When p, q are fixed and $n \to +\infty$, $\widehat{GdCov}^2$ inherits all the nice properties of the usual dCov.

Asymptotic Expansions

Population GdCov

When $\rho_i(\mathbf{x}_i, \mathbf{x}'_i) = \|\mathbf{x}_i - \mathbf{x}'_i\|$ and $R, S \to +\infty$,

$$GdCov^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{4\tau} \sum_{i=1}^{R} \sum_{j=1}^{S} dcov^2(\mathbf{X}_i, \mathbf{Y}_j) + \text{Remainder term.}$$

Sample GdCov

When $\rho_i(\mathbf{x}_i, \mathbf{x}'_i) = \|\mathbf{x}_i - \mathbf{x}'_i\|$, $R, S \to +\infty$ and n is fixed or growing slowly,

$$\widehat{GdCov^2}(\mathbf{X}, \mathbf{Y}) = \frac{1}{4\tau} \sum_{i=1}^{R} \sum_{j=1}^{S} \widehat{dcov^2}(\mathbf{X}_i, \mathbf{Y}_j) + \text{Remainder term,}$$

where $\widehat{dcov^2}$ denotes the unbiased estimator of the squared dCov.

High dimensional t-test

Define

$$T_n = \sqrt{v_n - 1}\frac{\widehat{GdCor}^2(\mathbf{X}, \mathbf{Y})}{\sqrt{1 - (\widehat{GdCor}^2(\mathbf{X}, \mathbf{Y}))^2}},$$

where $v_n = n(n-3)/2$ and

$$\widehat{GdCor}^2(\mathbf{X}, \mathbf{Y}) = \frac{\widehat{GdCov}^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\widehat{GdCov}^2(\mathbf{X}, \mathbf{X})\widehat{GdCov}^2(\mathbf{Y}, \mathbf{Y})}}.$$

High-dimensional t-test

Fixed n

As $R, S \to +\infty$,

$$T_n \to^d \begin{cases} t_{v_n - 1}, & \text{if } \mathbf{X}, \mathbf{Y} \text{ are independent,} \\ t_{v_n - 1, W}, & \text{if } \mathbf{X}, \mathbf{Y} \text{ are dependent,} \end{cases}$$

with $W^2 \sim c\chi^2_{v_n}$. If $\rho_i(\mathbf{x}_i, \mathbf{x}_i') = \|\mathbf{x}_i - \mathbf{x}_i'\|$, c is proportional to

$$\lim_{R,S \to +\infty} \frac{1}{RS} \sum_{i=1}^{R} \sum_{j=1}^{S} dcov^2(\mathbf{X}_i, \mathbf{Y}_j).$$

Growing n

As $R, S \to +\infty$,

$$T_n \to^d N(0, 1), \quad \text{if } \mathbf{X}, \mathbf{Y} \text{ are independent.}$$

Data Example

- Monthly stock returns of $p = 127$ companies under the finance sector and $q = 125$ companies under the healthcare sector.
- The dependence among financial asset returns is usually nonlinear.

Table: p-values corresponding to the different tests for cross-sector independence of stock returns.

GdCov-E	GdCov-L	GdCov-G	dCov	HSIC-L	HSIC-G
5.70×10^{-13}	2.36×10^{-10}	7.99×10^{-11}	0.120	0.093	0.040

A Quick Summary

- *GdCov* completely characterizes dependence in low dimension.
- *GdCov* detects groupwise nonlinear dependence in high dimension.
- The computational complexity of the t-test only grows linearly with the dimensions p, q.
- Grouping allows us to go from *pairwise dependence* to *groupwise dependence*.
  - Motivated by applications: gene set, neighboring regions.
  - Random grouping.

Homogeneity Metrics

- A classical problem in statistics is to test if

$$P = Q \quad \text{for two probability measures } P, Q.$$

- Kolmogorov-Smirnov test, Wald-Wolfowitz runs test, k-nearest neighbor (k-NN) graphs......

- Energy distance:

$$\begin{aligned} ED(\mathbf{X}, \mathbf{Y}) =& 2\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|_p - \mathbb{E}\|\mathbf{X} - \mathbf{X}'\|_p - \mathbb{E}\|\mathbf{Y} - \mathbf{Y}'\|_p \\ =& \|\Pi(P) - \Pi(Q)\|_{\mathcal{H}}^2, \end{aligned}$$

for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^p$ and $\mathbf{X} \sim P, \mathbf{Y} \sim Q$. An important property:

$$ED(\mathbf{X}, \mathbf{Y}) = 0 \text{ iff } P = Q.$$

Unbiased Estimator

Given $\{\mathbf{X}_i\}_{i=1}^n$ and $\{\mathbf{Y}_i\}_{i=1}^m$, define

$$\widehat{ED}(\mathbf{X}, \mathbf{Y}) = \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \|\mathbf{X}_i - \mathbf{Y}_j\|_p - \frac{1}{n(n-1)} \sum_{i \neq j} \|\mathbf{X}_i - \mathbf{X}_j\|_p - \frac{1}{m(m-1)} \sum_{i \neq j} \|\mathbf{Y}_i - \mathbf{Y}_j\|_p.$$

When p is fixed and $n \to +\infty$:

- $\widehat{ED}(\mathbf{X}, \mathbf{Y})$ is unbiased and strongly consistent.

Generalized Energy Distance

Generalized ED

For two random vectors $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^p$, define

$$\text{GED}(\mathbf{X}, \mathbf{Y}) = 2\mathbb{E}K_{\mathbf{p}}(\mathbf{X}, \mathbf{Y}) - \mathbb{E}K_{\mathbf{p}}(\mathbf{X}, \mathbf{X}') - \mathbb{E}K_{\mathbf{p}}(\mathbf{Y}, \mathbf{Y}').$$

- The generalized energy distance (GED) and its sample version ($\widehat{GED}$) are defined by replacing the Euclidean distance by K.
- All nice properties of energy distance are preserved for fixed p.

Curse of dimensionality

As $p \to +\infty$ and n is fixed or growing slowly, both $\widehat{ED}$ and $\widehat{GED}$ have the expansions

$$C + \text{Asymptotic normal term} + \text{Remainder term.}$$

For ED

$$C = 0 \quad \text{iff} \quad \mu_{\mathbf{X}} = \mu_{\mathbf{Y}} \text{ and } \mathrm{tr}(\Sigma_{\mathbf{X}}) = \mathrm{tr}(\Sigma_{\mathbf{Y}}).$$

For GED

$$C = 0 \quad \text{iff} \quad \mathbf{X}_i \text{ and } \mathbf{Y}_i \text{ have the same distribution.}$$

High Dimensional Two-sample t-test

Define

$$T_{n,m} = \frac{\widehat{GED}}{\sqrt{\left(\frac{1}{nm} + \frac{1}{2n(n-1)} + \frac{1}{2m(m-1)}\right) S_{n,m}}},$$

where

$$S_{n,m} = \frac{4v_{n,m}\widehat{CGdCov}^2(\mathbf{X},\mathbf{Y}) + 4v_n\widehat{GdCov}^2(\mathbf{X},\mathbf{X}) + 4v_m\widehat{GdCov}^2(\mathbf{Y},\mathbf{Y})}{(n-1)(m-1)+v_n+v_m}$$

is the *pool sample variance estimator*.

High Dimensional Two-sample t-test

Asymptotics

When $P = Q$ and $p \to +\infty$,

$$T_{n,m} \to^d \begin{cases} t_{(n-1)(m-1)+v_n+v_m}, & \text{if } n, m \text{ are fixed,} \\ N(0,1), & \text{if } n, m \to +\infty. \end{cases}$$

When $P \neq Q$ and $p \to +\infty$,

$$\frac{\widehat{GED} - C}{\sqrt{\left(\frac{1}{nm} + \frac{1}{2n(n-1)} + \frac{1}{2m(m-1)}\right) S_{n,m}}} \to^d \text{Nondegenerate limit.}$$

Numerical Results

Table: Power comparison among GED, ED and MMD, where $P \neq Q$ but have the same first two moments.

				GED		ED		MMD	
	n	m	p	10%	5%	10%	5%	10%	5%
(1)	50	50	50	**0.472**	**0.357**	0.294	0.214	0.044	0.042
(1)	50	50	100	**0.574**	**0.476**	0.397	0.297	0.086	0.086
(1)	50	50	200	**0.668**	**0.589**	0.472	0.375	0.107	0.106
(2)	50	50	50	**1.000**	**1.000**	0.124	0.072	0.078	0.044
(2)	50	50	100	**1.000**	**1.000**	0.107	0.061	0.076	0.030
(2)	50	50	200	**1.000**	**1.000**	0.101	0.051	0.067	0.032
(3)	50	50	50	**1.000**	**1.000**	0.138	0.076	0.009	0.002
(3)	50	50	100	**1.000**	**1.000**	0.112	0.063	0.014	0.007
(3)	50	50	200	**1.000**	**1.000**	0.084	0.039	0.022	0.018

Summary of Different Distance-base Metrics

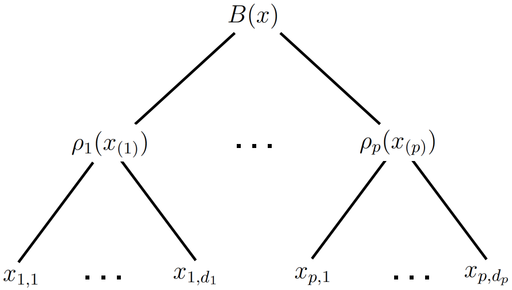
Choice of $\rho_i(x, x')$	Asymptotic behavior of homogeneity metrics	Asymptotic behavior of dependence metrics
the semi-metric $\|x - x'\|^2$	sum of squared Euclidean distances	sum of squared Pearson correlations
the Euclidean distance $\|x - x'\|$	sum of groupwise energy distances	sum of groupwise (squared) distance covariances
$k_i(x, x) + k_i(x', x') - 2k_i(x, x')$, where k_i is a characteristic kernel on $\mathbb{R}^{p_i} \times \mathbb{R}^{p_i}$	sum of groupwise MMD with the characteristic kernel k_i	sum of groupwise HSIC with the characteristic kernel k_i

More on Grouping

- The new distance corresponds to a semi-norm

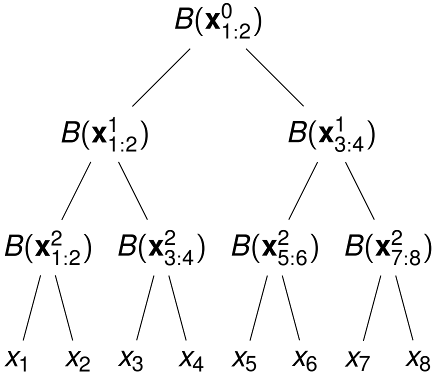
$$B(x) = \sqrt{\rho_1(x_{(1)}) + \ldots, \rho_p(x_{(p)})}.$$

- An interpretation of the semi-norm $B(\cdot)$ based on a tree:



More on Grouping

- Grouping allows us to detect a wider range of alternatives.
- Trees with more levels:



- The induced distance $B(\cdot - \cdot)$ is of strong negative type.

Summary

- Understand the curse of dimensionality for distance and kernel-based metrics.
- Provide a cautionary note on the use of classical metrics in high dimension.
- Propose a general class of metrics and a unified framework for studying them.
- Applications to statistical inference and learning, e.g., high-dimensional change-point detection, kernel-based learning and training Generative Adversarial Network.
- This talk is based on
  - Chakraborty, S., and Zhang, X. (2021) A New Framework for Distance and Kernel-based Metrics in High-dimension.
  - Zhu, C., Zhang, X., Yao, S., and Shao, X. (2020) Distance-based and RKHS-based Dependence Metrics in High-dimension.

THANK YOU!