The two-sample problem in high dimension: A ranking-based method

ETH-UCPH-TUM Workshop 2022

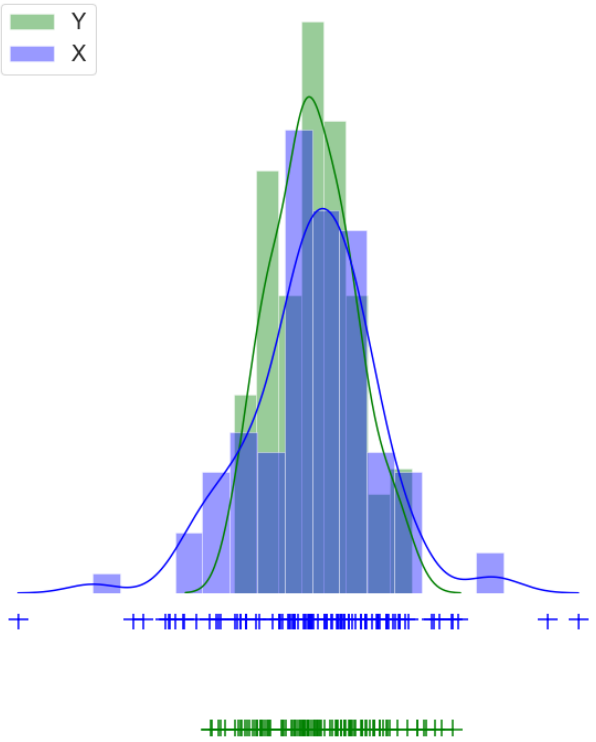
Myrto LIMNIOS

Postdoc @ Copenhagen Causality Lab, UCPH, Denmark

October 11, 2022

joint work with: Stephan Clémençon (Telecom Paris), Nicolas Vayatis (ENS Paris-Saclay)

Toy example



- Student's grades of the morning class

Marie	8.2
Antoine	14
...	

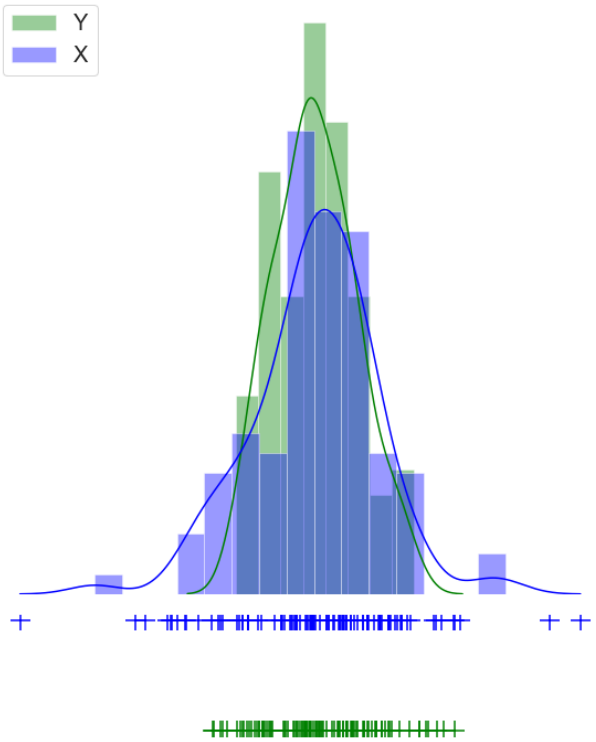
- Student's grades of the evening class

Clémence	12
Arthur	9
...	
X	0.5

- Are the students of similar level in the morning compared to the evening?
- How not to penalize the evening group because of X?

$\Longrightarrow$ Spearman (1904) introduced rank statistics to *"reduce the "accidental errors""* and as a response to the Gaussian assumption

Univariate two-sample linear rank statistics and linearization



- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} G$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} H$, independent
- $\mathrm{Rank}(X_i) = \sum_{k=1}^{n} \mathbb{I}\{X_k \leq X_i\} + \sum_{j=1}^{m} \mathbb{I}\{Y_j \leq X_i\}$
- $\phi : [0,1] \to \mathbb{R}$ *score-generating function*, nondecreasing

$$\widehat{W}^{\phi}_{n,m} = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(X_i)}{n+m+1}\right) \tag{1}$$

Hajèk projection: project a statistic $T - \mathbb{E}[T]$ into independent sums if $\mathbb{E}[T^2] < \infty$

$$\widehat{W}_{n,m} - \mathbb{E}[\widehat{W}_{n,m}] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(H(X_i) - \mathbb{E}[H(X_i)]) - \frac{1}{m}\sum_{j=1}^{m}(G(Y_j) - \mathbb{E}[G(Y_j)])}_{\text{i.i.d. sum of variables} \Rightarrow \text{classic theorems}} + \mathcal{O}_{\mathbb{P}}(1/n + 1/m)$$

The univariate two-sample problem

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} G$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} H$, independent, valued in $\mathcal{Z} \subset \mathbb{R}$

Hypothesis testing problem for sample comparison

$$\mathcal{H}_0 : G = H \quad \text{versus} \quad \mathcal{H}_1 : G \neq H$$

Example: Mann-Whitney-Wilcoxon or ranksum statistic (Wilcoxon (1945); Mann and Whitney (1947))

$$W_{n,m} = \sum_{i=1}^{n} \mathrm{Rank}(X_i)$$

Why rank statistics and not parametric statistics?

- Distribution-free under $\mathcal{H}_0$
- Competitive power compared to parametric tests (UMP for the location test, Lehmann and Romano (2005))
- Unbiased and consistent

The two-sample problem for high-dimensional data $\mathcal{Z} \subset \mathbb{R}^d$

Main idea: reject H_0 for "large $\mathcal{D}(\widehat{\mu}_n, \widehat{\nu}_m)$", $\mathcal{D}$ pseudo-distance:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{X}_i} \text{ and } \widehat{\nu}_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{\mathbf{Y}_j} \tag{2}$$

- Multivariate generalizations of classical statistics: Energy distances (Szekely and Rizzo (2004)), Kolmogorov-Smirnov statistics (Biau and Gyorfi (2005)), Wald-Wolfowitz (Friedman and Rafsky (1979)), Mann-Whitney-Wilcoxon (Clémençon et al. (2009))
- Statistics based on kernel methods (Gretton et al. (2012); Bach et al. (2008))
- Statistics based on optimal transport distances (Ramdas et al. (2015); Deb and Sen (2019a))

Main limitations

- Unknwon null distribution => data-driven testing threshold
- Depend on the definition of the statistic
- Asymptotic guarantees and usually only for the consistency

Generalization of two-sample linear rank statistics

Framework

- $p \in (0,1)$, $N \in \mathbb{N}^*$ such that $N \geq 1/p$
- $n = \lfloor pN \rfloor$ and $m = \lceil (1-p)N \rceil = N - n$

- $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} G$, $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{i.i.d.}{\sim} H$, independent

- G, H unknown (nonparametric) and continuous, defined on $\mathcal{Z}$ high-dimensional and measurable

- $\mathcal{S}$ class of *scoring functions* $s : \mathcal{Z} \to (-\infty, +\infty]$, measurable

Generalization of two-sample linear R-statistics[1]

- For $s \in \mathcal{S}$ we consider the univariate samples

$$\{s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)\},\ \{s(\mathbf{Y}_1),\ \ldots,\ s(\mathbf{Y}_m)\}$$

- $\mathrm{Rank}(s(\mathbf{X}_i)) = \sum_{k=1}^{n} \mathbb{I}\{s(\mathbf{X}_k) \leq s(\mathbf{X}_i)\} + \sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq s(\mathbf{X}_i)\}$

Definition

$$\widehat{W}_{n,m}^{\phi}(s) = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(s(\mathbf{X}_i))}{N+1}\right),\ \forall s \in \mathcal{S} \tag{3}$$

[1] Clémençon, Limnios, and Vayatis. *Electronic Journal of Statistics*, 2021.

Scalar performance measure[1]

- $\mathbf{X} \sim G$, $\mathbf{Y} \sim H$, independent, valued in $\mathcal{Z}$
- For $s \in \mathcal{S}$, $G_s(t) = \mathbb{P}\{s(\mathbf{X}) \leq t\}$, $H_s(t) = \mathbb{P}\{s(\mathbf{Y}) \leq t\}$
- $F_s = pG_s + (1-p)H_s$

Definition

The W_ϕ-**ranking performance criterion** based on F_s is

$$W_\phi(s) = \mathbb{E}[(\phi \circ F_s)(s(\mathbf{X}))] \tag{4}$$

Which is the *best* scoring function s?

[1] Clémençon, Limnios, and Vayatis. *Electronic Journal of Statistics*, 2021.

Oracle class of scoring functions

- $\mathcal{S}^*$= nondecreasing transforms of $\Psi(\mathbf{z}) = (dG/dH)(\mathbf{z})$, $\Psi(z) = (dG_\Psi/dH_\Psi)(\Psi(z))$
  - All elements of $\mathcal{S}^*$ maximize W_ϕ and we define $W_\phi^* := W_\phi(s^*)$

For a well-chosen scoring function

(i) a total preorder $\preccurlyeq_s$ on $\mathcal{Z}$ is induced

(ii) the ranking of $\mathbf{X}$'s are at the *top* of the list and the $\mathbf{Y}$'s are at the *bottom* of the list

[1] Proposition 2, Clémençon and Vayatis (2008)

Problem statement

Based on $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} G$, $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{i.i.d.}{\sim} H$, independent, $\mathcal{S}_0 \subset \mathcal{S}$, find

$$\widehat{s} \in \arg\max_{s \in \mathcal{S}_0} \widehat{W}^{\phi}_{n,m}(s) \tag{5}$$

Statistical learning guarantees: decompose the generalization error

$$W_\phi(s^*) - W_\phi(\widehat{s}) \leq \underbrace{2 \sup_{s \in \mathcal{S}_0} \left| \frac{1}{n} \widehat{W}^{\phi}_{n,m}(s) - W_\phi(s) \right|}_{\text{estimation error}} + \underbrace{W_\phi(s^*) - \sup_{s \in \mathcal{S}_0} W_\phi(s)}_{\text{approximation error/bias}} \tag{6}$$

Main result: generalization error bound

(A1) Let $M > 0$. For all $s \in \mathcal{S}_0$, the random variables $s(\mathbf{X})$ and $s(\mathbf{Y})$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathcal{W}^{2,\infty}$-norms bounded by $M < +\infty$

(A2) The score-generating function $\phi : [0,1] \mapsto \mathbb{R}$, is nondecreasing and twice continuously differentiable

(A3) The class of scoring functions $\mathcal{S}_0$ is a VC class of finite VC dimension $\mathcal{V} < +\infty$

Theorem (Corollary 7[1])

Suppose (A1-3) fulfilled. For any $\delta \in (0,1)$, with probability at least $1-\delta$

$$W_\phi^* - W_\phi(\hat{s}) \leq 2C_3\sqrt{\frac{\log(C_2/\delta)}{pN}} + \left(W_\phi^* - \sup_{s \in \mathcal{S}_0} W_\phi(s)\right) \tag{7}$$

for sufficiently large N, where the constants C_i, $i \leq 4$ depend only on ϕ, $\mathcal{V}$

[1] Clémençon, Limnios, and Vayatis. *Electronic Journal of Statistics*, 2021.

Theoretical ingredients for the proof

Main tool: linearization of R-processes with uniform control of the remainder w.p. $1-\delta$ over $\mathcal{S}_0$[1]

$$\widehat{W}_{n,m}^{\phi}(s) = \underbrace{n\widehat{W}_{\phi}(s)}_{\text{central statistic}} + \underbrace{\left(\widehat{V}_n^X(s) - \mathbb{E}\left[\widehat{V}_n^X(s)\right]\right) + \left(\widehat{V}_m^Y(s) - \mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right)}_{\text{empirical processes}} + \underbrace{\mathcal{R}_{n,m}(s)}_{\text{remainder}} \quad (8)$$

Additional ingredients for the proof

(i) Permanence results to control the complexity of classes of functions[2]

(ii) Linearization methods applied to one-/two-samples asymmetric U-processes using[3]:
  (a) Hájek projection method
  (b) Hoeffding decomposition results for generalized U-statistics

(iii) **Uniform concentration bounds for one-/two-samples asymmetric (degenerate) U-processes**[4] *(symmetrization, chaining, ...*[5]*)*

(iv) Maximal inequalities [6]

[1] **Proposition 5, Clémençon, Limnios and Vayatis. EJS, 2021**
[2] **Lemma 19-20, Clémençon, Limnios and Vayatis. EJS, 2021**
[3] Hájek (1968); Serfling (1980)
[4] Theorem 2 Major (2006), **Lemma 16, Clémençon, Limnios and Vayatis. EJS, 2021**
[5] van de Geer (2000); van der Vaart (1998); De la Pena and Giné (1999)
[6] Lemma 2.4 Neumeyer (2004), Theorem 6 Nolan and Pollard (1987), **Clémençon, Limnios and Vayatis. EJS, 2021**

Theoretical ingredients for the proof

Main tool: linearization of R-processes with uniform control of the remainder w.p. $1-\delta$ over $\mathcal{S}_0$[1]

$$\widehat{W}_{n,m}^{\phi}(s) = \underbrace{n\widehat{W}_{\phi}(s)}_{\text{central statistic}} + \underbrace{\left(\widehat{V}_n^X(s) - \mathbb{E}\left[\widehat{V}_n^X(s)\right]\right) + \left(\widehat{V}_m^Y(s) - \mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right)}_{\text{empirical processes}} + \underbrace{\mathcal{R}_{n,m}(s)}_{\text{remainder}} \quad (8)$$

Additional ingredients for the proof

(i) Permanence results to control the complexity of classes of functions[2]

(ii) Linearization methods applied to one-/two-samples asymmetric U-processes using[3]:
  (a) Hájek projection method
  (b) Hoeffding decomposition results for generalized U-statistics

(iii) **Uniform concentration bounds for one-/two-samples asymmetric (degenerate) U-processes**[4] *(symmetrization, chaining, ...*[5]*)*

(iv) Maximal inequalities [6]

[1] **Proposition 5, Clémençon, Limnios and Vayatis. EJS, 2021**
[2] **Lemma 19-20, Clémençon, Limnios and Vayatis. EJS, 2021**
[3] Hájek (1968); Serfling (1980)
[4] Theorem 2 Major (2006), **Lemma 16, Clémençon, Limnios and Vayatis. EJS, 2021**
[5] van de Geer (2000); van der Vaart (1998); De la Pena and Giné (1999)
[6] Lemma 2.4 Neumeyer (2004), Theorem 6 Nolan and Pollard (1987), **Clémençon, Limnios and Vayatis. EJS, 2021**

The two-sample problem with R-statistics

The two-sample problem formulated with R-statistics

At level $\alpha \in (0,1)$, test

$$\mathcal{H}_0 : \quad W_\phi^* := W_\phi(s^*) = \int_0^1 \phi(u)du \quad \textit{versus} \quad \mathcal{H}_1 : \quad W_\phi^* > \int_0^1 \phi(u)du$$

Our practical method: two-stage ranking-based method[1]. Split the independent samples $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d.}{\sim} G$, $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{i.i.d.}{\sim} H$ in two: $n' < n$, $m' < m$, $n = n' + n''$, $m = m' + m''$

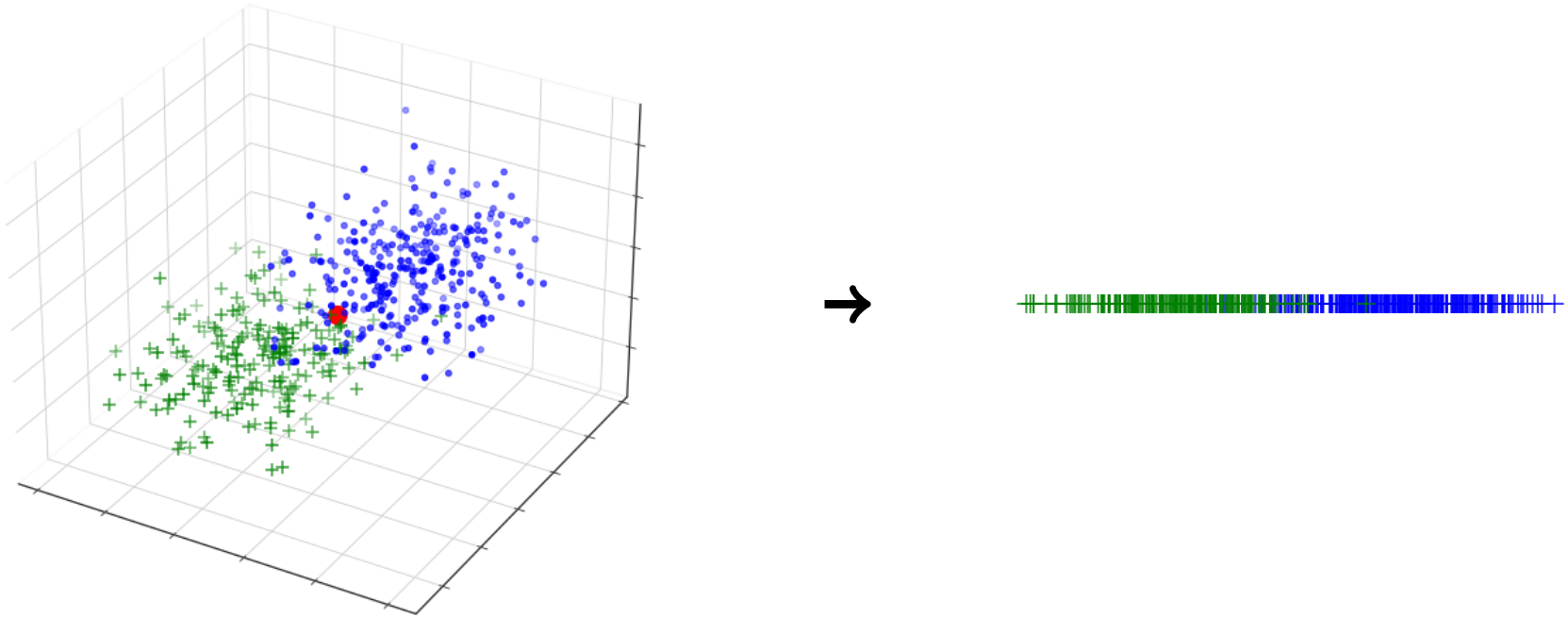
1. $\mathcal{D}_{n',m'} = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n'}\} \cup \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{m'}\}$
2. $\mathcal{D}_{n'',m''} = \{\mathbf{X}_{1+n'}, \ldots, \mathbf{X}_n\} \cup \{\mathbf{Y}_{1+m'}, \ldots, \mathbf{Y}_m\}$

[1] Clémençon, Limnios, Vayatis. WP.

Step 1: learn the optimal scoring function based on $\mathcal{D}_{n',m'}$

How to maximize the empirical W_ϕ-ranking performance criterion?

- Gradient-based algorithm to maximize a smoothed version of $\widehat{W}_{n,m}^{\phi}$ [1]
- Bipartite ranking algorithms



$\Longrightarrow$ Proved theoretical nonasymptotic guarantees[2]

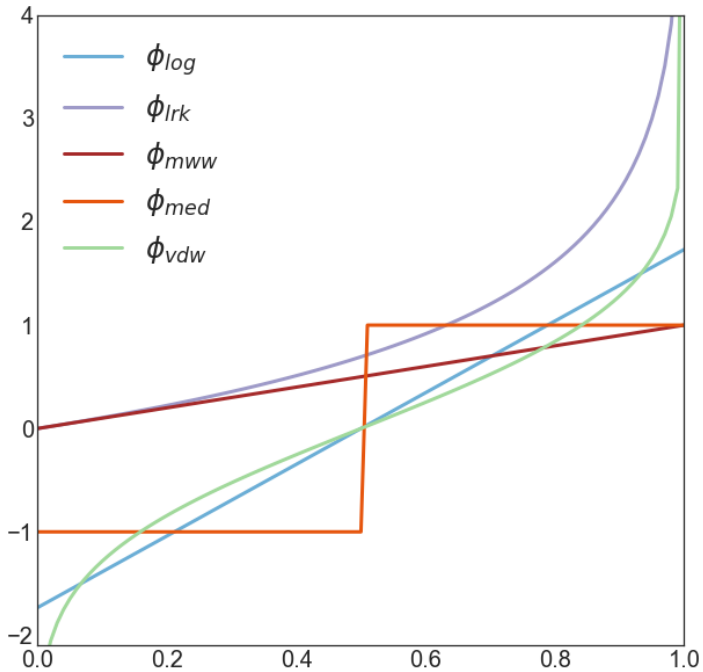
[1] Clémençon, Limnios, and Vayatis. *Electronic Journal of Statistics*, 2021.

[2] Clémençon, Limnios, and Vayatis. 2021 for GA, Menon and Williamson (2016) for BR

Step 2: univariate two-sample rank test

- $\mathcal{D}_{n'',m''}(\widehat{s}) = \{\widehat{s}(\mathbf{X}_{1+n'}), \ldots, \widehat{s}(\mathbf{X}_n)\} \cup \{\widehat{s}(\mathbf{Y}_{1+m'}), \ldots, \widehat{s}(\mathbf{Y}_m)\}$
- $q^{\phi}_{n'',m''}(\alpha)$ $(1-\alpha)$-quantile of the null distribution
- **Univariate test statistic**

$$\Phi^{\phi}_{\alpha}(\mathcal{D}_{n'',m''}(\widehat{s})) = \mathbb{I}\left\{ \frac{1}{n''} \widehat{W}^{\phi}_{n'',m''}(\widehat{s}) > \int_0^1 \phi(u)du + q^{\phi}_{n'',\, m''}(\alpha) \right\}$$



Statistic	ϕ
MWW (green)	$\phi(u) = u$
Logistic (blue)	$\phi_{log}(u) = 2\sqrt{3}(u - 1/2)$
Logrank (orange)	$\phi_{lrk}(u) = -\log(1-u)$
Median (red)	$\phi_{med}(u) = \mathrm{sgn}(u - 1/2)$
Van der Waerden (purple)	$\phi_{vdw}(u) = \Delta^{-1}(u)$

$\Longrightarrow$ Properties of the test statistic Φ^{ϕ}_{α}?

Analysis of statistical testing errors[1]

Let $\widehat{s}$ optimal element minimizing the bipartite ranking loss (*Step 1*) over $\mathcal{S}_0 \subset \mathcal{S}$

$$\widehat{W}^{\phi}_{n'',m''}(\widehat{s}) = \sum_{i=n'+1}^{n} \phi\left(\frac{\mathrm{Rank}(\widehat{s}(\mathbf{X}_i))}{N''+1}\right) \tag{9}$$

$$\begin{aligned}
\frac{1}{n''}\widehat{W}^{\phi}_{n'',m''}(\widehat{s}) - \int_0^1 \phi(u)du \quad &= \quad \left\{\frac{1}{n''}\widehat{W}^{\phi}_{n'',m''}(\widehat{s}) - W_\phi(\widehat{s})\right\} \\
&+ \quad \underbrace{\left\{W_\phi(\widehat{s}) - W^*_\phi\right\}}_{\leq\ 2\ \sup_{s\in\mathcal{S}_0}\left|\frac{1}{n'}\widehat{W}^{\phi}_{n',m'}(s) - W_\phi(s)\right| + \delta \quad \forall G,\ H \in \mathcal{B}(\delta) \quad \text{Step 1}} \\
&+ \quad \underbrace{\left\{W^*_\phi - \int_0^1 \phi(u)du\right\}}_{\geq\ \varepsilon \quad \forall G,\ H \in \mathcal{H}_1(\varepsilon) \quad \text{Deviation from } \mathcal{H}_0} \quad .
\end{aligned}$$

[1] Clémençon, Limnios, Vayatis. WP.

Concentration bound for the statistical type-I error

Theorem (Type-I error bound[1])

Let $\phi(u)$, $n,\ m \geq 2$, and fix $\alpha \in (0,1)$. Under the null hypothesis $\mathcal{H}_0$, the type-I error of the test is less than α

$$\mathbb{P}_{\mathcal{H}_0}\left\{\Phi_\alpha^\phi\left(\mathcal{D}_{n'',m''}(\widehat{s})\right)=1\right\} \leq \alpha \tag{10}$$

for all $1 \leq n'' < n$ and $1 \leq m'' < m$

[1] S. Clémençon, M. Limnios, N. Vayatis. WP.

Uniform concentration bound for the statistical type-II error

Theorem (Type-II error bound[2])

Let ϕ and $\varepsilon > \delta > 0$. Fix $\alpha \in (0,1)$. Suppose (A1-3) are fulfilled. Let $p \in (0,1)$ such that $N' \wedge N'' \geq 1/p$. Set $n' = \lfloor pN' \rfloor$ and $m' = \lceil (1-p)N' \rceil = N' - n'$, as well as $n'' = \lfloor pN'' \rfloor$ and $m'' = \lceil (1-p)N'' \rceil = N'' - n''$. Then, there exist constants C_1 and $C_2 \geq 24$, such that the type-II error of the test is uniformly bounded

$$\sup_{(H,G)\in\mathcal{H}_1(\varepsilon)\cap\mathcal{B}(\delta)} \mathbb{P}_{H,G}\left\{\Phi_\alpha^\phi(\mathcal{D}_{n'',m''}(\widehat{s})) = 0\right\} \leq 18\exp\left(-CN''(\varepsilon-\delta)^2/16\right) + C_2\exp\left(-\frac{N'}{8C_2}p(p\wedge(1-p))(\varepsilon-\delta)\log\left(1+\frac{\varepsilon-\delta}{32C_1(p\wedge(1-p))}\right)\right) \quad (11)$$

for sufficiently large N'', where the constants C (resp. C_1, , C_2) depend only on p, ϕ (resp. ϕ, $\mathcal{V}$)

[1] Clémençon, Limnios, Vayatis, WP.

Numerical applications

Numerical applications: experimental settings

- **Bipartite ranking algorithms:** `RankNN` (RNN, Burges et al. (2005)), linear `RankSVM` (see Joachims (2002)) with L_1 and L_2 losses (rSVM1, rSVM2), `RankBoost` (rBoost, Freund et al. (2003)), `Ranking Forest` (TreeRank, Clémençon et al. (2013))

- **Score-generating functions:** $\phi_{MWW}(u) = u$ (MWW, Wilcoxon (1945))

- **Two-sample statistics:**

  MMD: Maximum Mean Discrepancy Gretton et al. (2007, 2012), $\mathcal{F}$ unit ball of a Reproducing Kernel Hilbert Space (RKHS)

$$\mathrm{MMD}(G, H) = \sup_{f \in \mathcal{F}} |\mathbb{E}[f(\mathbf{X})] - \mathbb{E}[f(\mathbf{Y})]| \tag{12}$$

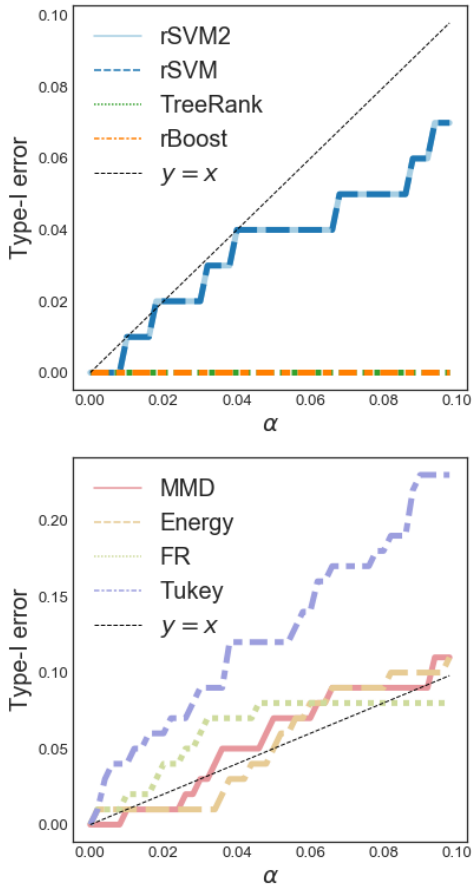
  Energy: metric-based Energy test Székely and Rizzo (2013), $\|\cdot\|$ Euclidean norm in $\mathbb{R}^d$

$$\mathcal{E}_{n,m} = \frac{mn}{m+n}\left(\frac{2}{nm}\sum_{i,j\leq n,m}\|\mathbf{X}_i - \mathbf{Y}_j\| - \frac{1}{n^2}\sum_{i,j\leq n}\|\mathbf{X}_i - \mathbf{X}_j\| - \frac{1}{m^2}\sum_{i,j\leq m}\|\mathbf{Y}_i - \mathbf{Y}_j\|\right) \tag{13}$$

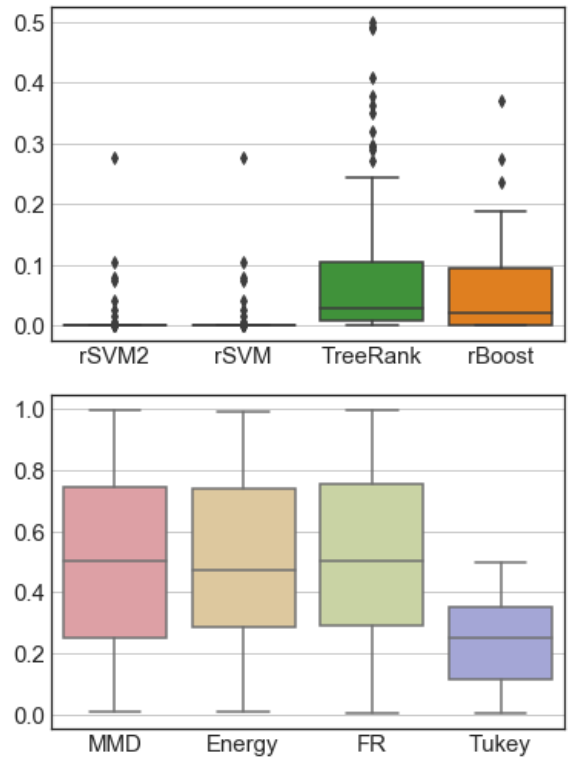
  FR: generalization of graph-based Wald-Wolfowitz Runs test Friedman and Rafsky (1979)

  Tukey: statisical depth-based generalization of ranks Tukey (1975) with method for testing of Liu and Singh (1993)

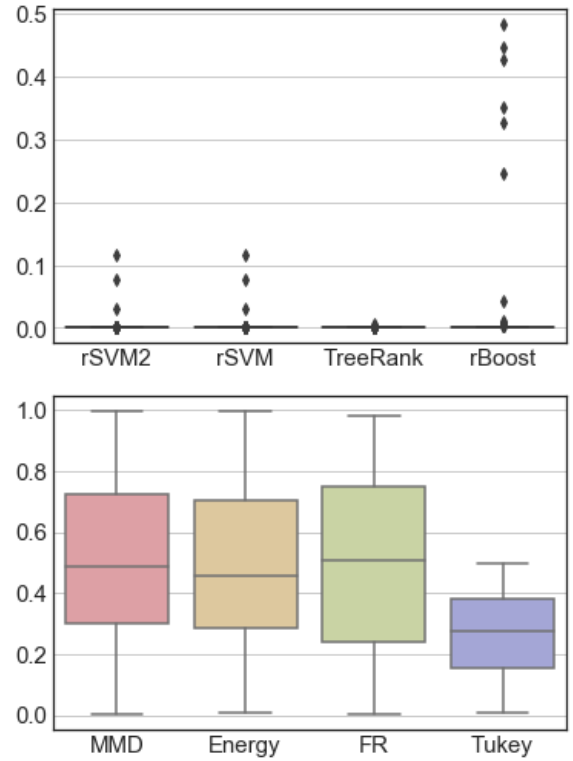
Control of the empirical type-I error and distribution of the p-values



a. $\mathcal{H}_0$



b. $\mathcal{H}_1$, $\varepsilon = 0.02$

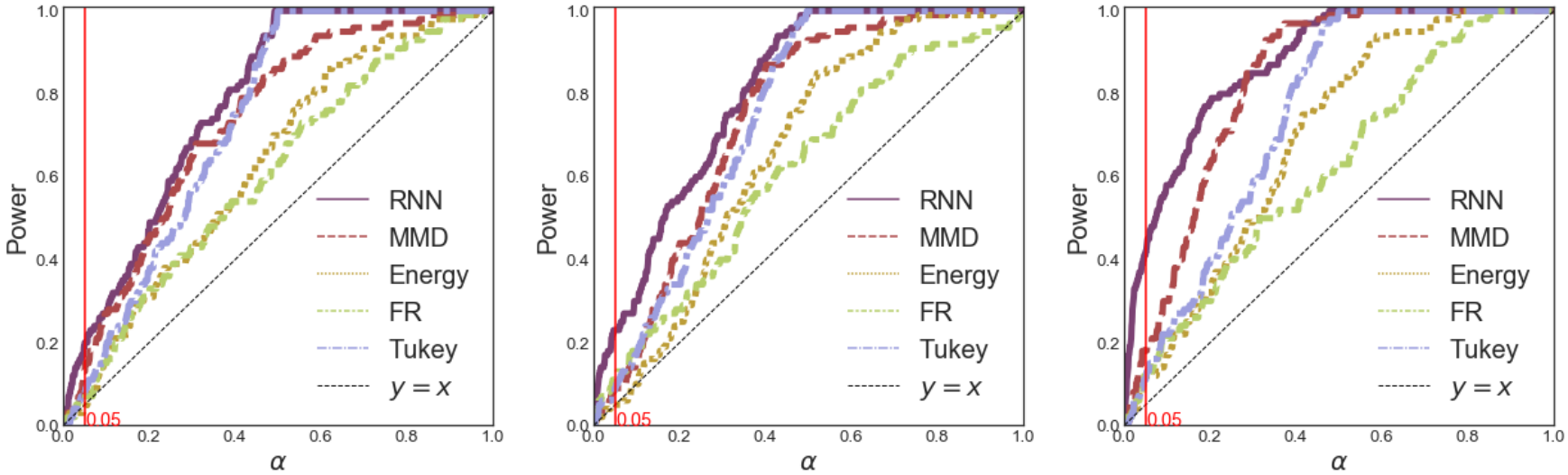


c. $\mathcal{H}_1$, $\varepsilon = 0.05$

Gaussian location model: $\mathbf{X} \sim \mathcal{N}_d((\varepsilon/\sqrt{d}) \times \mathbf{1}_d, \Sigma)$ and $\mathbf{Y} \sim \mathcal{N}_d(0_d, \Sigma)$ and $\Sigma \in S_d^+(\mathbb{R})$ s.t. the first marginal is negatively correlated with all the others and for $2 \leq k \leq d$ the coordinates are mutually independent

Numerical parameters: $n = m = 1000$, $n' = m' = 4N/5$, $d = 6$

Impact of the dimension of the feature space d on the power



a. $d = 30$ b. $d = 60$ c. $d = 100$

Gaussian scale model: $\mathbf{X} \sim \mathcal{N}_d(0_d, \Sigma_X)$, $\mathbf{Y} \sim \mathcal{N}_d(0_d, \Sigma_Y)$ with decreasing correlation matrix $\Sigma_{X,i,j} = (\beta + \varepsilon)^{|i-j|}$, $\Sigma_{Y,i,j} = \beta^{|i-j|}$, for $i, j \leq d$

Numerical parameters: $n = m = 1000$, $n' = m' = 4N/5$, $\varepsilon = 0.05$, $\beta = 0.2$

Summary

- Generic ranking-based method for two-sample comparison testing with R-statistics in high-dimensional spaces
- Nonasymptotic control of the statistical testing errors (type-I and type-II)
- Competitive method for small deviations from the null hypothesis and high dimensions

Thank you !

Further reading

- Theoretical analysis and application to bipartite ranking: Clémençon, Limnios, and Vayatis. Concentration inequalities for two-sample rank processes with application to bipartite ranking. *Electronic Journal of Statistics*, 2021.
  https://hal.archives-ouvertes.fr/hal-03190532

- Application to learning to rank anomalies: Limnios, Noiry, and Clémençcon. Learning to rank anomalies: Scalar performance criteria and maximization of two-sample rank statistics. *Proceedings of Machine Learning*, 2021
  https://proceedings.mlr.press/v154/limnios21a.html

- Application to the nonparametric two-sample problem with posturographic data: Bargiotas, Kalogeratos, Limnios, Vidal, Ricard, and Vayatis. Revealing posturographic profile of patients with Parkinsonian syndromes through a novel hypothesis testing framework based on machine learning. *PLOS ONE*, 2021
  https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0246790

References I

F. Bach, Z. Harchaoui, and E. Moulines. Testing for homogeneity with kernel Fisr discriminant analysis. In *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, page 89–96, 2005.

P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996.

S. Clémençon and N. Vayatis. Tree-structured ranking rules and approximation of the optimal ROC curve. In *ALT '08: Proceedings of the 2008 conference on Algorithmic Learning Theory*, 2008.

S. Clémençon, M. Depecker, and N. Vayatis. AUC maximization and the two-sample problem. In *Advances in Neural Information Processing Systems*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2009.

S. Clémençon, M. Depecker, and N. Vayatis. Ranking Forests. *Journal of Machine Learning Research*, 14:39–73, 2013.

V. De la Pena and E. Giné. *Decoupling: from dependence to independence*. Springer Science and Business Media, 1999.

References II

N. Deb and B. Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation, 2019a.

N. Deb and B. Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. 2019b.

Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

J. H. Friedman and L. C. Rafsky. Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annals of Statistics*, 7(4):697 – 717, 1979.

A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola. A kernel two-sample problem. *Journal of Machine Learning Research*, 13:723–773, 2012.

J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, 39:325–346, 1968.

M. Hallin and D. Paindaveine. Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *The Annals of Statistics*, 30(4):1103 – 1133, 2002a.

M. Hallin and D. Paindaveine. Optimal procedures based on interdirections and pseudo-Mahalanobis ranks for testing multivariate elliptic white noise against ARMA dependence. *Bernoulli*, 8(6):787 – 815, 2002b.

References III

M. Hallin and D. Paindaveine. Optimal rank-based tests for homogeneity of scatter. *The Annals of Statistics*, 36(3):1261 – 1298, 2008.

T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 133–142. Association for Computing Machinery, 2002.

E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

R. Y. Liu and K. Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260, 1993.

A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la société française de statistique*, 156(4): 133–162, 2015.

P. Major. An estimate on the supremum of a nice class of stochastic integrals and u-statistics. *Probability Theory and Related Fields*, 134(3):489–537, 2006.

H. Mann and D. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18:50–60, 1947.

A. Menon and R. Williamson. Bipartite ranking: A risk theoretic perspective. *Journal of Machine Learning Research*, 7:1–102, 2016.

J. Möttönen, H. Oja, and J. Tienari. On the efficiency of multivariate spatial sign and rank tests. *The Annals of Statistics*, 25(2):542–552, 1997.

J. Möttönen, H. Oja, and R. Serfling. Multivariate generalized spatial signed-rank methods. *Journal of Statistical Research*, 39(1):19–42, 2005. ISSN 0256-422X.

References IV

J. Möttönen and H. Oja. Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, 5(2):201–213, 1995.

N. Neumeyer. A central limit theorem for two-sample u-processes. *Statistics and Probability Letters*, 67(1):73 – 85, 2004.

D. Nolan and D. Pollard. *U*-Processes: Rates of Convergence. *The Annals of Statistics*, 15(2): 780 – 799, 1987.

H. Oja. Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1 (6):327–332, 1983.

A. Ramdas, N. Garcia, and M. Cuturi. On wasserstein two sample testing and related families of nonparametric tests, 2015.

R. Serfling. *Approximation theorems of mathematical statistics*. John Wiley and Sons, 1980.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

G. Szekely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 11 2004.

G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. ISSN 0378-3758.

J. W. Tukey. Mathematics and the picturing of data. In R. D. James, editor, *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531. Canadian Mathematical Congress, 1975.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

Concepts of ranks for multivariate observations

How to generalize a relation order in high-dimensional spaces?

Past works

- Component-wise ranks (Lung-Yut-Fong et al. (2015))
- Data-depth ranks with *e.g.* center-outward distribution, quantile functions (Chaudhuri (1996); Oja (1983); Deb and Sen (2019b))
- Spatial ranks (Möttönen and Oja (1995); Möttönen et al. (1997, 2005))
- Distance-based ranks (Hallin and Paindaveine (2002a,b, 2008))

Main limitations

- Strong assumptions on the distributions (semiparametric)
- Only asymptotic guarantees
- Dependent on the model specification (local representation of the data, definition of the statistic)
- Heavy computational cost when N increases

A quality measure for $\mathcal{S}$ with ROC analysis

- $\mathcal{S}$ class of scoring functions
- **Definition**

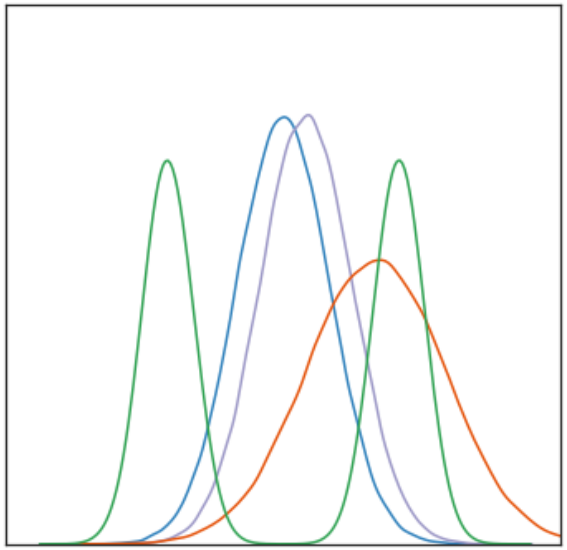
$$\mathrm{ROC}_s : t \in \mathbb{R} \mapsto \left(\underbrace{\mathbb{P}\{s(\mathbf{Y}) \geq t\}}_{\text{False Positive Rate}}, \underbrace{\mathbb{P}\{s(\mathbf{X}) \geq t\}}_{\text{True Positive Rate}} \right)$$

- **Bipartite ranking expected loss and** AUC

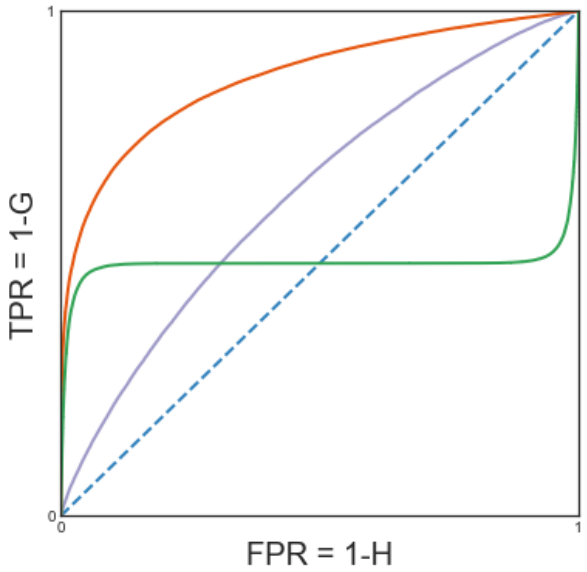
$$\begin{aligned} L(s) &= \mathbb{E}[\mathbb{I}\{s(\mathbf{Y}) > s(\mathbf{X})\}] + \frac{1}{2}\mathbb{P}\{s(\mathbf{Y}) = s(\mathbf{X})\} \\ &= 1 - \mathrm{AUC}(s) \end{aligned}$$

- **Empirical** AUC **and rank statistic**

$$W_{ld}(s) = \frac{(1-p)p}{2} \int_0^1 \mathrm{ROC}_s(\alpha)\, d\alpha + cst'(p)$$



(a) Probability distributions



(b) Corresponding ROC curves

Linearization of R-processes

Proposition (Proposition 5[1])

Suppose (A1-3) fulfilled. Then, for all $s \in S_0$

$$\widehat{W}_{n,m}^{\phi}(s) = \underbrace{n\widehat{W}_{\phi}(s)}_{\text{central statistic}} + \underbrace{\left(\widehat{V}_n^X(s) - \mathbb{E}\left[\widehat{V}_n^X(s)\right]\right) + \left(\widehat{V}_m^Y(s) - \mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right)}_{\text{empirical processes}} + \underbrace{\mathcal{R}_{n,m}(s)}_{\text{remainder}} \quad (14)$$

where

$$\widehat{W}_{\phi}(s) = \frac{1}{n}\sum_{i=1}^{n}(\phi \circ F_s)\left(s(\mathbf{X}_i)\right)$$

$$\widehat{V}_n^X(s) = \frac{n}{N+1}\sum_{i=1}^{n}\int_{s(\mathbf{X}_i)}^{+\infty}(\phi' \circ F_s)(u)dG_s(u)$$

$$\widehat{V}_m^Y(s) = \frac{n}{N+1}\sum_{j=1}^{m}\int_{s(\mathbf{Y}_j)}^{+\infty}(\phi' \circ F_s)(u)dG_s(u)$$

For any $\delta \in (0,1)$, there exist constants $c_1,\ c_3 > 0,\ c_2 \geq 2,\ c_4 > 6$, such that

$$\mathbb{P}\left\{\sup_{s \in \mathcal{S}_0}|\mathcal{R}_{n,m}(s)| < t\right\} \geq 1 - \delta \quad (15)$$

where $t = c_1 + (c_2/\sqrt{p(1-p)})\log(c_4/\delta)$, for sufficiently large N, with $d_1 > 0$ ($d_2 > 0$) constant depending on ϕ, $\mathcal{V}(\phi)$

[1] Clémençon, Limnios, and Vayatis. *Electronic Journal of Statistics*, 2021.