Empirical Maximization of R-statistics in the Two-sample Problem

Myrto Limnios, Stéphan Clemençon*, Nicolas Vayatis

CMLA, CNRS, ENS Paris-Saclay, Université Paris-Saclay, 94235, Cachan cedex, France
*LTCI, Telecom ParisTech, Université Paris Saclay, 46 rue Barrault, 75634 Paris, France

Objective: Derive optimal dissimilarity criteria by tayloring ranking random processes in the two-sample problem.

Two-sample problem: Let $\{\mathbf{X}_i\}_{i\leq n}$, $\{\mathbf{Y}_j\}_{j\leq m}$ be observations drawn from two samples of unknown probability distribution.
By ranking the first sample's data amongst the pooled sample, being able to distinguish possible differences between both distributions.

Contributions

- Express empirical performance criteria as *linear rank statistics* by using Hajek projection and Hoeffding decomposition of U-statistics technics.
- Analyse concentration properties of this novel class of *linear rank processes* when it is generalized with unknown scoring-generating function and optimized over the class of measurable scoring functions.
- In-depth understanding of both global and local dissimilarities criteria for the two-sample problem.
- Apply *linear rank processes* in the two-sample problem and nonparametric homogeneity tests in high dimension.

Related work

- *Linear rank statistics* were initially introduced in semi/nonparametric univariate framework by [7], [5].
- Empirical risk minimization of bivariate loss function has been shown to be equivalent with empirical maximization of the R-statistic associated with ([2]).
- Hypothesis testing has been widely studied in univariate and mostly parametric framework.
- Homogeneity testing for the two-sample problem has recently gained interest for multivariate distribution-free settings, especially through the work of Gretton [4] by introducing the Maximum Mean Discrepancy.

Notations and Framework

- Let $\mathbf{X}\sim G$, $\mathbf{Y}\sim H$ two independant absolute continuous *r.v.* in the probability space $(\mathcal{X},\mathcal{P}(\mathcal{X}))$ and consider $\{\mathbf{X}_i\}_{i\leq n}$, $\{\mathbf{Y}_j\}_{j\leq m}$ its realizations *s.t.* p proportion of $\mathbf{X}$ in the pooled sample. Denote by F the *c.d.f.* of the pooled sample *s.t.* : $F:=pG+(1-p)H$.
- Let $\mathcal{S}$ by the major class of *scoring functions s.t.* $\mathcal{S}:=\{s:\mathcal{X}\mapsto\mathbb{R}$ measurable$\}$ that maps observations into the real line where its natural relation order can be used. $\mathcal{S}$ has a VC-dimension denoted by $\mathcal{V}$.
- Once observations are mapped with $s\in\mathcal{S}$, denote by G_s (*resp.* H_s, F_s) the *c.d.f.* of resp. $s(\mathbf{X})$ (*resp.* $s(\mathbf{Y})$, $F_s:=pG_s+(1-p)H_s$).
- Denote by Ψ the likelihood ratio defined by $\Psi: x\in\mathcal{X}\mapsto\frac{dG(x)}{dH(x)}$.

Linear rank processes

Definitions: The *W-ranking performance measure for two samples* is defined by:

$$W_\phi(s)=\mathbb{E}[\phi(F_s(s(\mathbf{X})))],\ \forall s\in\mathcal{S}. \tag{1}$$

Let $n,m\in\mathbb{N}^*$, the *empirical W-ranking performance measure for two samples* $\{\mathbf{X}_i\}_{i\leq n}$, $\{\mathbf{Y}_j\}_{j\leq m}$ has the following empirical risk functional:

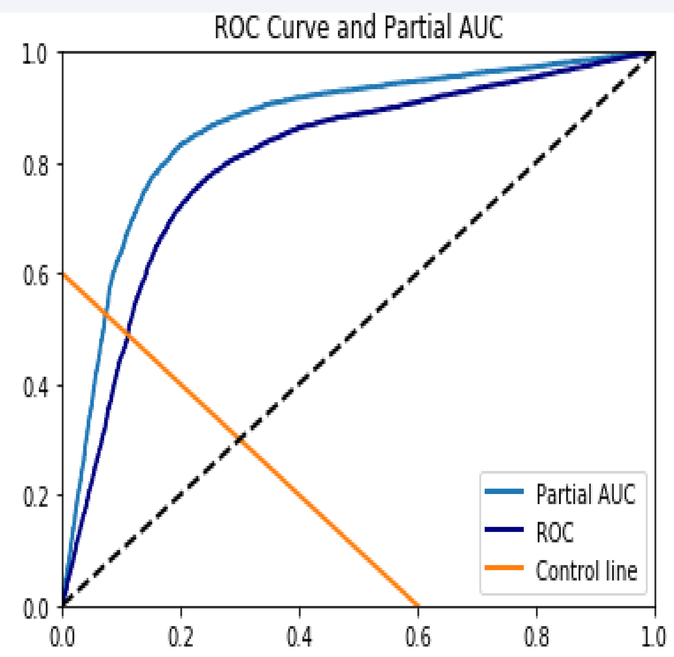
$$\widehat{W}_{n,m}(s)=\sum_{i=1}^{n}\phi\left(\frac{\mathrm{Rank}(s(\mathbf{X}_i))}{N+1}\right),\ \forall s\in\mathcal{S}. \tag{2}$$

The function $\phi:[0,1]\mapsto[0,1]$ is called the *score-generating function* of the *rank process* $\{\widehat{W}_{n,m}(s)\}_{s\in\mathcal{S}}$. It is supposed to be fixed, nondecreasing and continuously twice differentiable.

Choice of the score-generating function ϕ

Scoring-generating function	Empirical Ranking process	Related Statistic
$\phi=\mathrm{Id}_{[0,1]}$	$\widehat{W}_{n,m}(s)=\frac{1}{N+1}\sum_{i=1}^{n}\sum_{j=1}^{N}\mathbb{I}\{s(\mathbf{Z}_j)\leq s(\mathbf{X}_i)\}$	Mann-Whitney-Wilcoxon [1], $\widehat{W}_{n,m}(s)=nm\widehat{\mathrm{AUC}}_{n,m}(s)+\frac{n(n+1)}{2}$
$\phi:u\mapsto u\cdot\mathbb{I}_{\{u\geq u_0\}},\ u_0\in(0,1)$	$\widehat{W}_{n,m}(s)=\frac{1}{N+1}\sum_{i=1}^{n}\mathrm{Rank}(s(\mathbf{X}_i))\mathbb{I}\{\mathrm{Rank}(s(\mathbf{X}_i))\geq u_0(N+1)\}$	Local AUC, concentrates the decision rule on the "best" instances [3]
$\phi:u\mapsto u^q$	$\widehat{W}_{n,m}(s)=\frac{1}{(N+1)^q}\sum_{i=1}^{n}\mathrm{Rank}(s(\mathbf{X}_i))^q$	Related to q-norm push [6]

Table: Examples of different choices of scoring generating functions



Optimality

Let $n,m\in\mathbb{N}^*$, express:

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

where $\widehat{F}_{s,N}$ is the empirical *c.d.f.* of the scored pooled sample.

A widely used tool for measuring the performance of a scoring function s is the ROC curve defined by:

$$\mathrm{ROC}(s,.):\alpha\in[0,1]\mapsto 1-G_s\circ H_s^{-1}(1-\alpha) \tag{4}$$

Goal: Interpret the R-processes as optimal unbiased two-sample statistic through the ROC functional curve.

Consider $\mathcal{S}^*=\{s^*=T\circ\Psi\mid T:[0,1]\to\mathbb{R}$ strictly increasing$\}$.

Proposition: Assume that the score-generating function ϕ is strictly increasing. Then, we have:

$$\forall s\in\mathcal{S}\ ,\quad W_\phi(s)\leq W_\phi(\Psi)\ . \tag{5}$$

Moreover $W_\phi^*\doteq W_\phi(\Psi)=W_\phi(s^*)$ for any $s^*\in\mathcal{S}^*$.

Consequence: The optimal scoring function $s\in\mathcal{S}$ for the homogeneity two-sample problem is the solution of the empirical maximization of the R-process $\{\widehat{W}_{n,m}(s)\}_{s\in\mathcal{S}}$.

Linearization

Proposition: Let $\mathcal{S}_0\subset\mathcal{S}$ be a VC-major class of functions and suppose ϕ as in definition. Then:

$$\widehat{W}_{n,m}(s)=n\widehat{W}_\phi(s)+\widehat{V}_{n,m}(s)+\mathcal{O}_{\mathbb{P}}(1),\ \forall s\in\mathcal{S}_0, \tag{6}$$

up to a centering term for the random process, for n,m:

$$\widehat{V}_{n,m}(s)=\sum_{i=1}^{n}\Phi_s(s(\mathbf{X}_i))+\sum_{j=1}^{m}\Phi_s(s(\mathbf{Y}_j)), \tag{7}$$

where $\Phi_s:x\in\mathbb{R}\longmapsto p\int_x^{+\infty}\phi'(F_s(u))dG_s(u)$.

Uniform bound

Theorem: Under the same assumptions, at ϕ fixed, set the empirical W-ranking performance maximizer $\hat{s}_{n,m}=\mathrm{argmax}_{s\in\mathcal{S}_0}\widehat{W}_{n,m}(s)$. We have, for any $\delta\in(0,1)$, for $N\gg\max(\kappa_1+\kappa_2\log(c/\delta),\sqrt{\kappa_3})$, and with probability at least $1-\delta$:

$$W_\phi^*-W_\phi(\hat{s}_{n,m})\leq\kappa\sqrt{\frac{\mathcal{V}(\mathcal{S}_0)}{N}}+\kappa'\sqrt{\frac{\log(2/\delta)}{N}} \tag{8}$$

for some universal positive constants κ, κ', c depending on p, bounds of ϕ and its derivatives, $\mathcal{V}(\mathcal{S}_0)$. $(\kappa_i)_{i\in\{1,2,3\}}$ depend also on δ.

References

[1] 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.

[2] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[3] S. Clémençon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

[4] 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.

[5] J. Hajek. Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Stat.*, 39:325–346, 1968.

[6] C. Rudin. Ranking with a P-Norm Push. In H. Simon and G. Lugosi, editors, *Proceedings of COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 589–604, 2006.

[7] A. van de Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.