On extensions of rank correlation coefficients to multivariate spaces

Fang Han, University of Washington, Seattle
fanghan@uw.edu

Communicated by the Editor

This note summarizes ideas presented in a lecture for the *Bernoulli New Researcher Award 2021*. Rank correlations for measuring and testing against dependence of two random scalars are among the oldest and best-known topics in nonparametric statistics. This note reviews recent progress towards understanding and extending rank correlations to multivariate spaces through building connections to optimal transport and graph-based statistics.

Measuring the strength of dependence and testing independence for a pair of random scalars/vectors (X, Y) based on n independent realizations $\{(X_i, Y_i)\}_{i=1}^n$ is a century-old problem. In the univariate case, many correlation coefficients have been proposed and our interest is in those that meet (most of) the following four criteria.

(a) **Distribution-freeness**: the (limiting) distribution of the correlation coefficient under the hypothesis of independence should not depend on the marginal distributions of X and Y;

(b) **Consistency**: the correlation coefficient should consistently estimate a measure of dependence that is 0 if and only if X is independent of Y within a fairly large distribution family of (X, Y);

(c) **Statistical efficiency**: the test of independence based on the correlation coefficient should have nontrivial power over root-n neighborhoods of "smooth" parametric models;

(d) **Computational efficiency**: there should exist a nearly linear-time algorithm to compute the correlation coefficient.

In the above four criteria, we are particularly insistent on the first that was prescribed as the *genesis* of all rank tests (Hájek et al., 1999, Page 1). Our attention is thus restricted to rank correlation coefficients. A rank correlation coefficient estimates a certain measure of dependence only using the ranks of univariate margins. Distribution-freeness is then immediate given that the probability measure is continuous. Hoeffding (1948) introduced the first rank correlation coefficient that, in contrast to other popular ones, satisfies the consistency criterion. In addition, this rank correlation coefficient, ofter referred to as Hoeffding's D, can be computed in $O(n \log n)$ time. In this note we first present some recent results on Hoeffding's D and its variants and in particular, an identity between Hoeffding's D, Blum-Kiefer-Rosenblatt's R (Blum et al., 1961), and Bergsma-Dassios-Yanagimoto's τ^* (Bergsma and Dassios, 2014; Yanagimoto, 1970) that raises interesting connections to local structures in combinatorics (Even-Zohar and Leng, 2021).

Extending the aforementioned rank correlation coefficients to a multivariate setting when X and Y are multidimensional is a long-studied problem. Componentwise rank-based methods that simply rank univariate margins cannot eliminate within-group dependence and thus fail to be distribution-free in multivariate spaces, neither could other alternatives such as spatial, Mahalanobis, and cone ordering-based ranks (Hallin, 2021, Section 3.2). A recent breakthrough due to Chernozhukov et al. (2017) and Hallin et al. (2021) paved an ingenious path towards a solution. It relates multivariate ranks to an optimal transport (OT) problem that studies mappings between the data generating probability to a preset reference measure that is known to the user – noticing that the cumulative distribution function (CDF) is a univariate transport function to the Lebesgue measure over $[0, 1]$. In the second part of this note we will reveal that the corresponding notion of multivariate rank can lead to correlation coefficients that achieve the first three goals, yet are not computationally efficient.

In addition to the OT-based extension of rank correlations to higher dimensions, in recent years there has been a growing interest in connecting rank correlations to graph-based statistics. Some rather remarkable results are due to Mona Azadkia and Sourav Chatterjee in their two recent papers (Chatterjee, 2021; Azadkia and Chatterjee, 2021). Noticing that the univariate ranks could also be understood as a correspondence to a 1-nearest neighbor (1-NN) graph — although NN graphs are metric-based but not the univariate ranks — they built a measure of dependence and its estimates over 1-NN graphs. In the third part of this note we will show their proposal leads to multivariate rank correlation coefficients that successfully achieve the goals of (a), (b), (d), but are not statistically efficient, though ways to boost efficiency were recently proposed.

§1. Univariate rank correlation coefficients

Spearman's ρ and Kendall's τ, like Pearson's correlation coefficient, do not satisfy the consistency property; a canonical example is the bivariate-t distribution which cannot admit independent components. Letting $F(\cdot, \cdot), F_X(\cdot), F_Y(\cdot)$ be the bivariate

and marginal CDFs of (X, Y), X, and Y, respectively, Hoeffding (Hoeffding, 1948) introduced the following correlation measure,

$$D = \int \Big\{F(x,y) - F_X(x)F_Y(y)\Big\}^2 \mathrm{d}F(x,y);$$

assuming F is absolutely continuous, D is zero if and only F corresponds to a product measure. Furthermore, noticing that D is equal to $\mathrm{E}\mathbf{1}(X_1 \le X_3, Y_1 \le Y_3)\mathbf{1}(X_2 \le X_3, Y_2 \le Y_3) - 2\mathrm{E}\mathbf{1}(X_1 \le X_4, Y_1 \le Y_4)\mathbf{1}(X_2 \le X_4)\mathbf{1}(Y_3 \le Y_4) + \mathrm{E}\prod_{j=1}^{2}\mathbf{1}(X_j \le X_5)\prod_{k=3}^{4}\mathbf{1}(Y_k \le Y_5)$, an unbiased estimator of D could then be constructed as

$$\widehat{D}_n = \binom{n}{5}^{-1} \sum_{i_1<\ldots<i_5} h_D\{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_5}, Y_{i_5})\},$$

which constitutes a U-statistic of order 5. Since the kernel function $h_D(\cdot)$ only involves ordinal comparisons of the inputs, $\widehat{D}_n$ is a rank correlation coefficient. Two more such rank correlation coefficients were later proposed by Blum et al. (1961) (denoted as $\widehat{R}_n$) and Bergsma and Dassios (2014) (denoted as $\widehat{\tau}_n^*$); they are U-statistics of orders 6 and 4 separately, and are both rank-based. The following items document their advantages.

(1) All three rank correlation coefficients are rank-based and hence satisfy the criterion (a). As a matter of fact, under independence they all weakly converge to a convolution of weighted chi-square distributions of distribution-free weights; cf. Shi et al. (2021c, Proposition 4).

(2) All three satisfy the criterion (b) for absolutely continuous measures; cf. Shi et al. (2021c, Propositions 2 and 3).

(3) All three lead to tests of independence admiting nontrivial power over root-n neighborhoods within the class of quadratic mean differentiable alternatives and thus satisfy the criterion (c); cf. Shi et al. (2021c, Theorem 1).

(4) All three can be computed in $O(n \log n)$ time, which is via Hoeffding (1948, Section 5), Even-Zohar and Leng (2021, Corollary 4), and the following identity

$$3\widehat{D}_n + 2\widehat{R}_n = 5\widehat{\tau}_n^*$$

that is due to Drton et al. (2020, Equ. (6.1)), who traced it back to Yanagimoto (1970, Proposition 9).

(5) Technically speaking, under independence all three rank correlations are degenerate U-statistics. Our recent works have established Cramér-type moderate deviation theorems and Bernstein-type tail bounds in complex stochastic systems for such statistics; cf. Drton et al. (2020, Theorem 4.1) and Shen et al. (2020, Theorem 2.1).

§2. OT-based correlation coefficients

Starting from this section, let's consider either X or Y or both of them are multivariate. Since a canonical ordering in general does not exist in a multidimensional space, extending rank correlation coefficients to higher dimensions is non-trivial and all existing extensions available into the 2000s are either lacking distribution-freeness in general or hard to compute (Hallin, 2021, Section 3.2). A major breakthrough was made in 2017, when Chernozhukov, Galichon, Hallin, and Henry (Chernozhukov et al., 2017) successfully connected the notion of multivariate CDF, and accordingly the notion of multivariate rank, to optimal transport.

Thinking about the univariate CDF as a mapping or *transport* from the data generating probability to the Lebesgue measure over $[0, 1]$, their idea can be briefly described as follows. For any probability measure P in $\mathbb{R}^d$, set up a reference probability measure ν in $\mathbb{R}^d$ and then define the "multivariate CDF" $F^{\mathrm{P},\nu}$ as a transport from P to ν. As when $d \ge 2$ there generally exist multiple such mappings, define $F^{\mathrm{P},\nu}$ to be the *optimal transport* that minimizes the transportation cost under the squared Euclidean loss (analytically) or, more generally, is the gradient of a convex function $\psi : \mathbb{R}^d \to \mathbb{R}$ (geometrically). The celebrated McCann's theorem (McCann, 1995) guarantees the existence and uniqueness of such an $F^{\mathrm{P},\nu}$ as long as both P and ν are absolutely continuous (w.r.t the Lebesgue measure). Cafarelli-type regularity properties of $F^{\mathrm{P},\nu}$ (e.g., Lipschitz-ness and higher-order smoothness) further exist and were developed in, e.g., Figalli (2018, Theorem 1.1) and Hallin et al. (2021, Proposition 2.3), among many others.

Turning to statistical estimation of $F^{\mathrm{P},\nu}$, given that an empirical measure P_n of P has been observed, a natural idea is to "discretize" the reference distribution to some ν_n that will weakly converge to ν, and then define $\widehat{F}_n^{\mathrm{P},\nu}$ to be the corresponding optimal transport pushing P_n to ν_n. This is called *plug-in estimation* in optimal transport literature; the estimators' stochastic behavior (e.g., distribution-freeness and maximal ancillarity), uniform consistency as well as the rate of convergence for $\widehat{F}_n^{\mathrm{P},\nu}$ to estimate $F^{\mathrm{P},\nu}$ have already been established (Chernozhukov et al., 2017; Ghosal and Sen, 2019; Hallin et al., 2021; Deb et al., 2021; Manole et al., 2021).

Now let's set up two regular reference distributions ν_X and ν_Y as "couples" of P_X and P_Y (the marginal probability measures of X and Y), respectively. We are then ready to define OT-based correlation coefficients as extensions to rank-based ones. Think about a generic multivariate correlation coefficient such as a U-statistic of order m and kernel $H(\cdot)$; one canonical example is the distance covariance with more to be found in Shi et al. (2020, Section 2). We then introduce OT-based correlation coefficients

as those that admit the same U-statistic form but with the input changed from the original data to its multivariate ranks, $\{(\widehat{F}_n^{\mathrm{P}_X,\nu_X}(X_i), \widehat{F}_n^{\mathrm{P}_Y,\nu_Y}(Y_i))\}_{i=1}^n$. The following items summarize the proposal's properties.

(1) OT-based correlation coefficients satisfies the distribution-freeness criterion, and their limiting null distributions are only dependent on ν_X, ν_Y that are known to the user; cf. Shi et al. (2021a, Theorem 3.1) and Deb and Sen (2021, Theorem 4.1) for a special example of distance covariance, and Shi et al. (2020, Corollary 5.1) for a general one.

(2) As long as the original multivariate correlation coefficient is consistent, the corresponding OT-based extension is consistent. This is due to the measure-preserving nature of the optimal transport; cf. Shi et al. (2020, Proposition 5.3).

(3) Tests of independence built on OT-based correlation coefficients are statistically efficient, namely, they have nontrivial power over root-n neighborhoods within the class of quadratic mean differentiable alternatives; cf. Shi et al. (2020, Theorem 5.3).

(4) As long as one of X and Y is multidimensional, there does not exist an algorithm to compute any considered OT-based correlation coefficient in nearly linear time; the time complexity is normally between $O(n^2)$ and $O(n^4)$; cf. Shi et al. (2020, Section B.3).

(5) Technically speaking, the weak convergence results could be established using the permutation uniformity nature of OT-induced ranks and thus combinatorial inference tools; this route was explored in Shi et al. (2021a, Theorems 4.1 and 4.2). Different from that, we employed Hájek representation theorems, which facilitate local power analyses via invoking Le Cam's third lemma (Shi et al., 2020). Using either way, the limiting null distribution can be established without resorting to any sort of rate of convergence for $\widehat{F}_n^{\mathrm{P},\nu}$ but only consistency.

§3. Graph-based correlation coefficients

Graph-based inference encompasses a long and rich literature in nonparametric statistics and has been applied to test independence by, for example, drawing a data-driven (e.g., tree-structured) partition and then summarizing information across bins; cf. Heller et al. (2016). Chatterjee (2021) and later Azadkia and Chatterjee (2021) recently introduced an ingenious way to estimate the following measure of dependence between a random scalar Y and a random vector X whose format was first proposed in Dette et al. (2013):

$$\xi = \frac{\int \mathrm{Var}\big\{\mathrm{E}\big[\mathbf{1}\big(Y \geq y\big) \mid X\big]\big\}\mathrm{d}F_Y(y)}{\int \mathrm{Var}\big\{\mathbf{1}\big(Y \geq y\big)\big\}\mathrm{d}F_Y(y)}.$$

This dependence measure has some rather appealing properties including, in particular, the capability of being both consistent (i.e. $\xi = 0$ if and only if Y is independent of X) and able to detect **perfect dependence** (i.e., $\xi = 1$ if and only if Y is a measurable function of X).

To estimate ξ, let R_i represent the rank of Y_i among $Y_1, \ldots, Y_n$ and $N(i)$ index the nearest neighbor of X_i; the following graph-based correlation coefficient can be shown to be a strongly consistent estimator of ξ:

$$\xi_n = \frac{\sum_{i=1}^n \min(R_i, R_{N(i)}) - (n+1)(2n+1)/6}{(n^2-1)/6}.$$

To understand it, let's recall the law of total variance for ξ and that

$$\mathrm{E}[\mathrm{Var}\{\mathbf{1}(Y \geq t) \,|\, \boldsymbol{X}\}] \approx \frac{\mathrm{E}\big\{\mathbf{1}(Y_1 \geq t) - \mathbf{1}(Y_{N(1)} \geq t)\big\}^2}{2}$$

with

$$\begin{aligned}
&\frac{1}{2}\int \mathrm{E}\big\{\mathbf{1}(Y_1 \geq t) - \mathbf{1}(Y_{N(1)} \geq t)\big\}^2 \mathrm{d}\mathrm{P}_Y(t) \\
\approx\ &\int \mathrm{E}\big\{1(Y_1 \geq t) - 1(Y_1 \geq t)1(Y_{N(1)} \geq t)\big\}\mathrm{d}\mathrm{P}_Y(t) \\
=\ &\mathrm{E}\big[F_Y(Y_1) - \min\{F_Y(Y_1), F_Y(Y_{N(1)})\}\big] \\
\approx\ &\mathrm{E}[R_1/n - \min(R_1, R_{N(1)})/n],
\end{aligned}$$

so that $\mathrm{E}\xi_n$ is approximately ξ.

Azadkia and Chatterjee conjectured that under independence $\sqrt{n}\xi_n$ is asymptotically normal. In Shi et al. (2021b) we resolved this conjecture based on an elegant prior work of Deb et al. (2020).

Theorem [*Distribution-freeness of* ξ_n]. Assume Y in $\mathbb{R}$ is continuous and independent of X in $\mathbb{R}^p$, which is absolutely continuous. We then have, as $n \to \infty$,

$$\sqrt{n}\xi_n \longrightarrow N\Big(0, \frac{2}{5} + \frac{2}{5}\mathfrak{q}_p + \frac{4}{5}\mathfrak{o}_p\Big) \text{ in distribution,}$$

where $\mathfrak{q}_p$ and $\mathfrak{o}_p$ are explicitly defined in Shi et al. (2021b, Equ. (3.2) and (3.3)) and, in particular, are independent of P_X and P_Y.

The following items summarize ξ_n's properties.

(1) Azadkia-Chatterjee's graph-based correlation coefficient satisfies the criterion (a) and a test of independence built on it is directly implementable without recurring to permutational critical values; cf. Theorem 1 above.

(2) It is further consistent in view of ξ's property; cf. Shi et al. (2021b, Proposition 2.3).

(3) Tests of independence built on Azadkia-Chatterjee's graph-based correlation coefficient are statistically *inefficient*. In the case $p = 1$, this is proved in Cao and Bickel (2020) and Shi et al. (2021c) and the critical detection boundary is shown to be at the order of $n^{-1/4}$ in a regular model (Auddy et al., 2021); the higher dimensional result is first derived in Shi et al. (2021b, Theorem 4.1).

(4) For $p = 1$, we recently devised a revision of Azadkia-Chatterjee's original proposal that provably boosts the power to be nearly parametrically efficient; cf. Lin and Han (2021).

(5) The correlation coefficient can be computed in $O(n \log n)$ time due to the fast speed to conduct a nearest neighbor search.

(6) Technically speaking, the theoretical results are built on large-sample properties of nearest neighbor graphs, central limit theorems under local dependence, conditional central limit theorem, and Le Cam's third lemma.

Figure 1 gives a summary of the results.

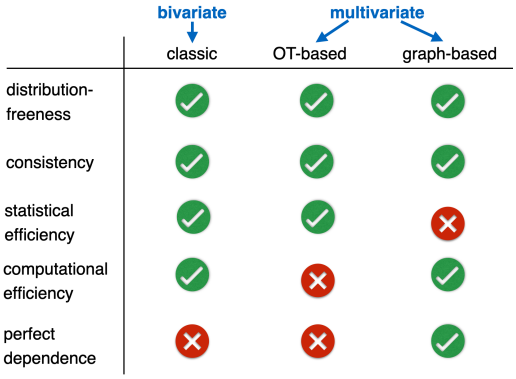

Figure 1: Summary of discussed correlation coefficients' properties.

Acknowledgement. The author is grateful to Peter Bickel, Mathias Drton, Marc Hallin, and Bodhisattva Sen for reading and commenting on a preliminary version of this note. He would also like to thank Sky Cao and Hongjian Shi for helpful discussions throughout. Research presented in this note is supported by National Science Foundation DMS-1712536 and SES-2019363.

References

Auddy, A., Deb, N., and Nandy, S. (2021). Exact detection thresholds for Chatterjee's correlation. Available at arXiv:2104.15140v1.

Azadkia, M. and Chatterjee, S. (2021+). A simple measure of conditional dependence. *Ann. Statist.* (in press).

Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20(2):1006–1028.

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.*, 32(2):485–498.

Cao, S. and Bickel, P. J. (2020). Correlations with tailored extremal properties. Available at arXiv:2008.10177v2.

Chatterjee, S. (2021+). A new coefficient of correlation. *J. Amer. Statist. Assoc.* (in press).

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.*, 45(1):223–256.

Deb, N., Ghosal, P., and Sen, B. (2020). Measuring association on topological spaces using kernels and geometric graphs. Available at arXiv:2010.01768v2.

Deb, N., Ghosal, P., and Sen, B. (2021). Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. Available at arXiv:2107.01718v1.

Deb, N. and Sen, B. (2021+). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *J. Amer. Statist. Assoc.* (in press).

Dette, H., Siburg, K. F., and Stoimenov, P. A. (2013). A copula-based non-parametric measure of regression dependence. *Scand. J. Stat.*, 40(1):21–41.

Drton, M., Han, F., and Shi, H. (2020). High-dimensional consistent independence testing with maxima of rank correlations. *Ann. Statist.*, 48(6):3206–3227.

Even-Zohar, C. and Leng, C. (2021). Counting small permutation patterns. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2288–2302, Philadelphia, PA. Society for Industrial and Applied Mathematics (SIAM).

Figalli, A. (2018). On the continuity of center-outward distribution and quantile functions. *Nonlinear Anal.*, 177(part B):413–421.

Ghosal, P. and Sen, B. (2019). Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing. Available at arXiv:1905.05340v2.

Hájek, J., Šidák, Z., and Sen, P. K. (1999). *Theory of Rank Tests* (2nd ed.). Probability and Mathematical Statistics. Academic Press, Inc., San Diego, CA.

Hallin, M. (2021). Measure transportation and statistical decision theory. *Annu. Rev. Stat. Appl.* (in press).

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *Ann. Statist.*, 49(2):1139–1165.

Heller, R., Heller, Y., Kaufman, S., Brill, B., and Gorfine, M. (2016). Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.*, 17(29):1–54.

Hoeffding, W. (1948). A non-parametric test of independence. *Ann. Math. Statist.*, 19(4):546–557.

Lin, Z. and Han, F. (2021). On boosting the power of Chatterjee's rank correlation. Available at arXiv:2108.06828v1.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2021). Plugin estimation of smooth optimal transport maps. Available at arXiv:2107.12364v1.

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2):309–323.

Shen, Y., Han, F., and Witten, D. (2020). Exponential inequalities for dependent V-statistics via random Fourier features. *Electron. J. Prob.*, 25:1–18.

Shi, H., Drton, M., and Han, F. (2021+a). Distribution-free consistent independence tests via center-outward ranks and signs. *J. Amer. Statist. Assoc.* (in press).

Shi, H., Drton, M., and Han, F. (2021b). On Azadkia-Chatterjee's conditional dependence coefficient. Available at arXiv:2108.06827v1.

Shi, H., Drton, M., and Han, F. (2021+c). On the power of Chatterjee's rank correlation. *Biometrika*. (in press).

Shi, H., Hallin, M., Drton, M., and Han, F. (2020). On universally consistent and fully distribution-free rank tests of vector independence. Available at arXiv:2007.02186v2.

Yanagimoto, T. (1970). On measures of association and a related problem. *Ann. Inst. Statist. Math.*, 22(1):57–63.