Independence testing in high dimension with empirical copulas

Cambyse Pakzad , https://cpakzad.github.io

Testing for stochastic independence

Problem: Let $\boldsymbol{X} = (X_1, \dots, X_d)^\top$ be a d-variate random vector. We are interested in testing the null

$$H_d : X_1, \dots, X_d \text{ are } \textit{mutually} \text{ independent}$$

based on an i.i.d. sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ of $\boldsymbol{X}$ with $\boldsymbol{X}_i = (X_{i1}, \dots, X_{id})^\top$,

Vast literature:

- The bivariate case $d = 2$: Hoeffding (1948), Feuerverger (1993), ...
- The fixed d-case: Blum, Kiefer, Rosenblatt (1961), Deheuvels (1979), Genest and Rémillard (2004), Székely et al. (2007), Genest et al. (2019), ...
- The $d = d(n) \to \infty$ case: ongoing research in recent years.

The high dimensional regime

Existing literature in the high dimensional regime typically uses the proxy hypothesis

$$H_2 : X_1, \ldots, X_d \text{ are pairwise independent,}$$

with $H_d \Rightarrow H_2$, but not vice versa.

(Testing for H_2 amounts to simultaneously testing $\binom{d}{2}$ different sub-hypotheses of H_d.)

Heuristic motivation:

- If $\boldsymbol{X}$ is Gaussian, then $H_2 \Leftrightarrow H_d$ (Schott, Biometrika 2005; Cai and Liang, AoS, 2011; ...)
- Practically relevant alternatives from $\neg H_d$ should typically involve pairwise dependencies (*main effect of joint dependence*).

  Hence, we should design test statistics that are sensitive towards deviations from H_2.

Testing for pairwise independence in high dimensions

Bivariate association/dependence measures: Pearson Correlation, Kendall's tau, Spearman's rho, Distance covariance, ..., e.g., for $1 \le p < q \le d$,

$$\hat{\tau}_{p,q} := \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \operatorname{sign}(X_{ip} - X_{jp}) \cdot \operatorname{sign}(X_{iq} - X_{jq}).$$

- Han, Chen, Liu (Biometrika, 2017): Maximum of linear rank statistics and (squared) non-degenerate rank-based U-statistics (e.g. Kendall's τ). Test is based on a Gumbel approximation:

$$\frac{9n(n-1)}{2(2n+5)} \max_{1 \le p < q \le d} \hat{\tau}^2_{p,q} - 4 \log d + \log \log d \rightsquigarrow_{H_2} \text{Gumbel}$$

- Leung, Drton (AoS, 2018): Sum of (squared) rank-based (possibly degenerate) U-statistics (e.g. Kendall's τ). Test is based on a normal approximation:

$$\frac{9n}{2d} \left(\sum_{1 \le p < q \le d} \hat{\tau}^2_{p,q} - \frac{2(2n+5)}{9n(n-1)} \right) \rightsquigarrow_{H_2} \mathcal{N}(0,1)$$

The tests are inconsistent for H_2, as $\tau = 0$ does not imply bivariate independence (Yao, Shang, Shao (JRSSB, 2018): similar as Leung, Drton (2018), but with distance covariances).

Goal: try to overcome the shortcomings from the pairwise methods by considering the problem of testing for independence from a copula perspective.

- Quite surprisingly: there is no copula-related asymptotic theory or specific methodology for the high dimensional regime $d = d(n) \to \infty$ (to the best of our knowledge).
- **Origin of statistics for copulas:** Deheuvels (1979, 1981a, 1981b) - **Independence testing!**

Sklar's Theorem: if $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \sim F$ has continuous marginal c.d.f.s $F_1, \ldots, F_d$, then there exists an unique copula $C : [0,1]^d \to [0,1]$ such that

$$F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d)), \qquad \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Moreover, $\boldsymbol{C}(\boldsymbol{u}) = \mathbb{P}(\boldsymbol{U} \leq u)$, where $U_p = F_p(X_p) \sim \mathrm{Unif}(0,1)$ for $p = 1, \ldots, d$.

Consequence for independence testing: in case of continuous marginal c.d.f.s,

$$H_d : X_1, \ldots, X_d \text{ are } \textit{mutually} \text{ independent} \quad \Longleftrightarrow \quad H_d : C = \Pi_d.$$

Here, $\Pi_d(\boldsymbol{u}) = \prod_{p=1}^d u_d$ denotes the independence copula.

Decomposing mutual independence H_d in terms of margins of C

For $k \in \{2, \ldots, d\}$, let

$$H_k : \begin{cases} X_1, \ldots, X_d \text{ are } k\text{-wise independent, i.e.,} \\ \qquad \text{any subvector of length } k \text{ has mutual independent components.} \end{cases}$$

Note that $H_d \Rightarrow \cdots \Rightarrow H_k \Rightarrow \cdots \Rightarrow H_3 \Rightarrow H_2$.

Characterization: In case of continuous marginal cdfs, we may rewrite

$$H_k : C_A = \Pi_k \text{ for all } A \subset \{1, \ldots, d\} \text{ with } |A| = k,$$

where $C_A(\boldsymbol{u}) = C(\boldsymbol{u}^A) = \mathbb{P}(U_p \leq u_p \text{ for all } p \in A)$ denotes the A-margin of C.

The validity of H_k may hence be assessed based on a nonparametric estimator of C.

Nonparametric estimation of C: the empirical copula

The empirical copula: Recalling $C(\boldsymbol{u}) = \mathbb{P}(\boldsymbol{U}_i \leq \boldsymbol{u})$ with $U_{ip} = F_p(X_{ip})$ suggests to define

$$\hat{C}_n(\boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^{n} 1\{\hat{\boldsymbol{U}}_i \leq \boldsymbol{u}\},$$

where $\hat{U}_{ip} = \hat{F}_{np}(X_{ip}) = \frac{1}{n+1} \sum_{j=1}^{n} 1\{X_{jp} \leq X_{ip}\} = \frac{R_{ip}}{n+1}$, with R_{ip} the rank of X_{ip} among $X_{1p}, \ldots, X_{np}$.

Fixed-d case: $\hat{C}_n$ is a strongly consistent estimator of any C as $n \to \infty$. Under smoothness conditions on C (Rüschendorf, 1976; ..., Segers, 2012; ...), with $\mathbb{G}_C$ a continuous Gaussian process on $[0,1]^d$ with covariance $\mathrm{Cov}(\mathbb{G}_C(\boldsymbol{u}), \mathbb{G}_C(\boldsymbol{v})) = C(\boldsymbol{u} \wedge \boldsymbol{v}) - C(\boldsymbol{u})C(\boldsymbol{v})$,

$$\left\{\sqrt{n}(\hat{C}_n(\boldsymbol{u}) - C(\boldsymbol{u}))\right\}_{\boldsymbol{u}} \rightsquigarrow \{\mathbb{G}'_C(\boldsymbol{u})\}_{\boldsymbol{u}} = \{\mathbb{G}_C(\boldsymbol{u}) - \sum_{j=1}^{d} \dot{C}_j(\boldsymbol{u})\mathbb{G}_C(\boldsymbol{u}^j)\}_{\boldsymbol{u}} \quad \text{in } (\ell^\infty([0,1]^d), \|\cdot\|_\infty)$$

Testing for independence of a subvector $\boldsymbol{X}_A$: use $S_{n,A} = \|\sqrt{n}((\hat{C}_n)_A - \Pi_A)\| \rightsquigarrow_H \|(\mathbb{G}'_\Pi)_A\|$. However, the weak limits will be dependent, which complicates aggregation over different sets A.

For a real-valued function H on $[0,1]^d$, the mapping $H \mapsto \big(\mathcal{M}_A(H)\big)_{A \subset \{1,\dots,d\}: 2 \le |A| \le d}$ with

$$\mathcal{M}_A(H)(\boldsymbol{u}) = \sum_{B \subset A} (-1)^{|A \setminus B|} H(\boldsymbol{u}^B) \prod_{j \in A \setminus B} u_j.$$

is called the **Moebius transformation** of H.

$$H_k \quad \Longleftrightarrow \quad \mathcal{M}_A(C) \equiv 0 \text{ for all } A \subset \{1,\dots,d\} \text{ with } 2 \le |A| \le k.$$

Significant deviations of $\mathcal{M}_A(\hat{C}_n)$ from zero indicate dependence in $\boldsymbol{X}_A$. Assessing significance (d fixed):

$$\sqrt{n}\mathcal{M}_A(\hat{C}_n) = \sqrt{n}\{\mathcal{M}_A(\hat{C}_n) - \mathcal{M}_A(\Pi)\} = \mathcal{M}_A(\sqrt{n}(\hat{C}_n - \Pi)) \rightsquigarrow_H \mathcal{M}_A(\mathbb{G}_\Pi) =: \mathbb{M}_A$$

The Moebius transformation of the empirical copula

In then the fixed-d case and under H,

$$\mathbb{M}_{n,A} = \sqrt{n}\mathcal{M}_A(\hat{C}_n) \rightsquigarrow_H \mathbb{M}_A$$

Here, by a straightforward calculation,

$$\mathbb{M}_{n,A} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \prod_{p \in A} \Big(1\{R_{ip} \leq (n+1)u_p\} - u_p \Big).$$

The weak convergence holds jointly in $A \subset \{2, \ldots, d\}$, and the Gaussian limit process satisfies

$$\mathrm{Cov}(\mathbb{M}_A(\boldsymbol{u}), \mathbb{M}_{A'}(\boldsymbol{v})) = 1_{\{A=A'\}} \prod_{p \in A} (u_p \wedge v_p - u_p v_p).$$

The weak limits are independent over A! Aggregation by sum/max functionals should yield feasible limits (Deheuvels, 1981; Genest and Rémillard, 2004).

As in Genest and Rémillard (2004), we assess non-independence of $\boldsymbol{X}_A$ by **Cramér-von Mises statistics**:

$$S_{n,A}^{\mathrm{M}} = \int_{[0,1]^{|A|}} \mathbb{M}_{n,A}^2(\boldsymbol{u})\,\mathrm{d}\Pi_A(\boldsymbol{u}) \rightsquigarrow_H \int_{[0,1]^{|A|}} \mathbb{M}_A^2(\boldsymbol{u})\,\mathrm{d}\Pi_A(\boldsymbol{u}).$$

After a slight redefinition of $\mathbb{M}_{n,A}$ (such that the process is centred), we obtain the representation

$$S_{n,A}^{\mathrm{M}} = \frac{1}{n}\sum_{i,j=1}^{n}\prod_{p\in A} I_{i,j}^{(p)},$$

where

$$I_{i,j}^{(p)} = \frac{2n+1}{6n} + \frac{R_{ip}(R_{ip}-1)}{2n(n+1)} + \frac{R_{jp}(R_{jp}-1)}{2n(n+1)} - \frac{\max(R_{ip},R_{jp})}{n+1}.$$

Deviations of H_k will be measured by **sum aggregation** (akin to Leung and Drton (2018) for $k=2$):

$$T_n(k) = \sum_{\substack{A\subset\{1,\dots,d\}\\ |A|=k}} S_{n,A}^{\mathrm{M}}, \qquad 2\le k\le d.$$

Towards asymptotics of sum aggregation

The heuristics from the fixed d case suggest a (joint) normal approximation for

$$T_n(k) = \sum_{\substack{A\subset\{1,\dots,d\}\\|A|=k}} \|\mathbb{M}_{n,A}\|^2_{L^2([0,1]^k)} = \sum_{\substack{A\subset\{1,\dots,d\}\\|A|=k}} S^{\mathrm{M}}_{n,A} = \sum_{\substack{A\subset\{1,\dots,d\}\\|A|=k}} \frac{1}{n}\sum_{i,j=1}^{n}\prod_{p\in A} I^{(p)}_{i,j}, \quad 2\le k\le d.$$

Proposition (Bücher and P., 2024)

Under H_k, we have, for all $A\subset\{1,\dots,d\}$ with $|A|=k$,

$$\mu_n(k) := \mathbb{E}[S^{\mathrm{M}}_{n,A}] = \left(\frac{1}{6}-\frac{1}{6n}\right)^k + (n-1)\left(\frac{-1}{6n}\right)^k,$$

$$\sigma_n^2(k) := \mathrm{Var}(S^{\mathrm{M}}_{n,A}) \sim \frac{2}{90^k}.$$

Explicit formulas for $\sigma_n^2(k)$ are available for $k\in\{2,3\}$.

Weak convergence of $T_n(k)$

Introduce scaling sequences:

$$T_n(k) := \sum_{|A|=k} \|\mathbb{M}_{n,A}\|^2_{L^2([0,1]^k)}, \quad \nu_n(k) := \binom{d}{k}\mu_n(k), \qquad \bar{\delta}^2_n(k) := \binom{d}{k}\sigma^2_n(k), \qquad \delta^2_n(k) := \binom{d}{k}\frac{2}{90^k}.$$

Theorem (Bücher and P., 2024)

Under H_d, if $d = d_n \to \infty$, we have

$$\frac{T_n(2) - \nu_n(2)}{\delta_n(2)} \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

Moreover, for fixed $m \in \{3, 4, \dots\}$, if $1 \ll d_n \ll n^{\frac{1}{m-1}}$, we have

$$\left(\frac{T_n(2) - \nu_n(2)}{\delta_n(2)}, \dots, \frac{T_n(m) - \nu_n(m)}{\delta_n(m)}\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1)^{\otimes(m-1)}.$$

As a consequence,

$$\bar{T}_n(m) = \frac{1}{\sqrt{m-1}} \sum_{k=2}^{m} \frac{T_n(k) - \nu_n(k)}{\delta_n(k)} \xrightarrow[n\to+\infty]{d} \mathcal{N}(0,1).$$

The same results are true for $\bar{\delta}_n$ instead of δ_n.

- Straightforward test: Reject H_m iff $\bar{T}_n(m) > u_{1-\alpha} = \Phi^{-1}_{\mathcal{N}(0,1)}(1-\alpha)$.
- The computational cost to calculate $\bar{T}_n(m)$ is $\Theta(mn^2 d^m)$.
- No growth conditions must be put on $d = d_n$ for $k = 2$; this is akin to Leung and Drton (2018).
- For the weak limit result, it actually suffices to be under H_{4m-3} instead of H_d.
- The proof is based on a reduction to centred summands $\tilde{S}^{\mathrm{M}}_{n,A}$, on a reduction to summands $I^{(p)}_{i,j}$ with $i \neq j$, and finally on a central limit theorem for martingale arrays; with the sum over A restricted to $\max A \leq r \in \{k, \ldots, d\}$ and with filtration $\mathcal{F}_{n,r} = \sigma(U_{ip} : 1 \leq i \leq n, 1 \leq p \leq r)$.
- Also works for $k := k_n$ for some regime derived from Stirling's approximation.

Finite sample results

We study finite-sample rejection probabilities of the following tests:

- For $k \in \{2, 3, 4\}$, let $\mathcal{S}_k$ denote the test

$$\text{reject } H_k \text{ if } \frac{T_n(k) - \nu_n(k)}{\delta_n(k)} > 1.645 = u_{0.95}.$$

- For $m \in \{3, 4\}$, let $\mathcal{T}_m$ denote the test

$$\text{rejects } H_m \text{ if } \bar{T}_n(m) = \frac{1}{\sqrt{m-1}} \sum_{k=2}^{m} \frac{T_n(k) - \nu_n(k)}{\delta_n(k)} > 1.645 = u_{0.95}.$$

Empirical rejections probabilities in % under mutual independence

		Asymptotic variance scaling						
Test	$n \setminus d$	4	8	16	32	64	128	256
$\mathcal{S}_2$	16	4.8	1.6	2.8	2.6	3.2	2.8	3.0
$\mathcal{S}_3$		2.2	3.6	6.2	9.4	19.8	29.0	31.4
$\mathcal{T}_3$		1.6	1.2	1.6	6.0	10.0	18.6	23.8
$\mathcal{S}_4$		2.2	18.2	29.4	34.4	43.4	46.0	42.4
$\mathcal{T}_4$		2.4	11.4	24.2	30.6	39.8	45.4	42.0
$\mathcal{S}_2$	32	4.0	5.4	3.0	5.0	3.2	4.2	4.0
$\mathcal{S}_3$		5.0	4.2	5.0	10.8	15.6	19.4	27.8
$\mathcal{T}_3$		4.0	3.4	2.6	6.2	8.0	9.8	19.0
$\mathcal{S}_4$		6.6	13.4	26.6	40.2	44.8	42.6	46.6
$\mathcal{T}_4$		4.4	9.6	18.0	36.0	42.4	40.8	45.8
$\mathcal{S}_2$	64	6.0	5.8	6.6	4.6	4.4	3.2	4.6
$\mathcal{S}_3$		5.0	4.2	6.2	6.2	9.6	13.6	24.0
$\mathcal{T}_3$		5.0	4.2	4.4	3.8	4.8	8.6	14.8
$\mathcal{S}_4$		4.8	9.0	25.6	33.4	41.8	45.0	47.2
$\mathcal{T}_4$		5.4	7.2	17.8	29.8	39.2	43.2	46.4
$\mathcal{S}_2$	128	6.8	4.0	5.8	4.8	5.2	5.8	4.4
$\mathcal{S}_3$		6.6	4.8	4.2	7.8	6.0	10.8	15.0
$\mathcal{T}_3$		6.2	3.6	4.2	5.6	4.6	6.0	9.2
$\mathcal{S}_4$		6.6	10.4	22.0	30.0	36.6	43.2	46.8
$\mathcal{T}_4$		5.8	6.8	13.6	25.2	33.2	40.2	44.6

The Romano-Siegel Model

- Let Z_1, Z_2, Z_3 iid standard normal random variables. Define

$$X_1 = |Z_1| \cdot \mathrm{sign}(Z_2 Z_3), \quad X_2 = Z_2, \quad X_3 = Z_3.$$

- (X_1, X_2, X_3) exhibits pairwise independence but not mutual independence.
- Generating Z_4, Z_5, Z_6 iid $\mathcal{N}(0,1)$ random variables independently of (Z_1, Z_2, Z_3), we duplicate step 1 to construct X_4, X_5, X_6. Etc...
- Example $d = 9$

$$X_1, X_2, X_3, \quad X_4, X_5, X_6, \quad X_7, X_8, X_9$$

$$\text{typical triplets:} \quad (X_1, X_2, X_3), \quad (X_1, X_4, X_5), \quad (X_1, X_4, X_7)$$

- Out of the $\binom{d}{3}$ triplets, only $d/3$ are not independent, which is a proportion of $O(d^{-2})$. The tests' power should hence be decreasing in d.

Empirical rejections probabilities in % in the Romano-Siegel Model

		Finite variance scaling						
Test	$n \setminus d$	3	6	15	30	63	126	255
$\mathcal{S}_2$	16	3.6	6.8	5.2	4.6	4.0	7.0	2.8
$\mathcal{S}_3$		16.6	23.4	26.4	29.4	32.4	33.8	53.4
$\mathcal{T}_3$		12.6	13.2	15.0	20.0	24.2	27.6	22.4
$\mathcal{S}_2$	32	2.2	4.8	5.0	3.6	4.2	5.4	2.4
$\mathcal{S}_3$		82.4	57.6	39.8	34.2	36.2	33.6	98.8
$\mathcal{T}_3$		45.8	32.6	22.2	20.4	24.2	24.4	98.2
$\mathcal{S}_2$	64	2.2	5.2	3.6	3.8	4.6	6.8	4.2
$\mathcal{S}_3$		100.0	100.0	84.8	63.8	46.2	37.2	100.0
$\mathcal{T}_3$		100.0	90.4	60.6	41.6	30.8	26.6	100.0
$\mathcal{S}_2$	128	0.8	3.6	4.2	4.0	5.0	4.8	3.4
$\mathcal{S}_3$		100.0	100.0	100.0	99.2	85.6	62.4	100.0
$\mathcal{T}_3$		100.0	100.0	99.2	91.8	66.8	46.4	100.0

Thank you!

Fixed d:

- P. Deheuvels (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis 11*, 102–113.
- C. Genest and B. Rémillard (2004). Tests of independence and randomness based on the empirical copula process. *Test 13*, 335–370.

Increasing d:

- D. Leung. and M. Drton (2018). Testing independence in high dimensions with sums of rank correlations. *Annals of Statistics 46*, 280–307.
- F. Han, S. Chen and H. Liu (2017). Distribution-free tests of independence in high dimensions. *Biometrika 104*, 813–828.
- S. Yao, X. Zhang and X. Shao (2018). Testing mutual independence in high dimension via distance covariance. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 80*, 455–480.

This talk:

- A. Bücher and C. Pakzad (2024). Testing for independence in high dimensions based on empirical copulas. *Annals of Statistics*.