1.2.3 Statistical estimation

In this section, we present the most classical estimators of the measures of association introduced in the previous section, including the linear correlation due to its widespread use, even if it is not a measure of association.

The sampling definition of the linear correlation coefficient is given by:

Definition 1.37. Let $\left((X_1^k, X_2^k)\right)_{k=1,\dots,N}$ be a sample of size N of the random vector $\boldsymbol{X} = (X_1, X_2)$. The sampling linear correlation coefficient $\hat{\rho}_N(X_1, X_2)$ is defined by

$$\hat{\rho}_N(\boldsymbol{X}) = \frac{\sum_{k=1}^{n}(X_1^k - \bar{X}_1)(X_2^k - \bar{X}_2)}{\sqrt{\sum_{k=1}^{N}(X_1^k - \bar{X}_1)^2 \sum_{k=1}^{N}(X_2^k - \bar{X}_2)^2}} \tag{1.31}$$

where $\bar{X}_1 = \frac{1}{N}\sum_{k=1}^{N} X_1^k$ and $\bar{X}_2 = \frac{1}{N}\sum_{k=1}^{N} X_2^k$.

The asymptotic properties of this estimator are given in the following theorems [Gay51, Equations 53 and 54]:

Theorem 1.38. *Let $\boldsymbol{X}$ be a bi-dimensional random vector with finite second moments $\mathbb{E}\left[X_1^2\right] < \infty$ and $\mathbb{E}\left[X_2^2\right] < \infty$. Then:*

$$\hat{\rho}_N(\boldsymbol{X}) \overset{a.s}{\to} \rho(\boldsymbol{X}) \text{ when } N \to \infty$$

Theorem 1.39. *Let $\boldsymbol{X}$ be a bi-dimensional random vector with finite fourth-order moments $\mathbb{E}\left[X_1^4\right] < \infty$ and $\mathbb{E}\left[X_2^4\right] < \infty$. Then:*

$$\sqrt{N}\left(\hat{\rho}_N(\boldsymbol{X}) - \rho(\boldsymbol{X})\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_\rho^2) \text{ when } N \to \infty$$

where the asymptotic variance σ_ρ^2 is given by:

$$\sigma_\rho^2 = \left(1 + \frac{\rho^2(\boldsymbol{X})}{2}\right)\frac{m_{22}}{m_{20}m_{02}} + \frac{\rho^2(\boldsymbol{X})}{4}\left(\frac{m_{40}}{m_{20}^2} + \frac{m_{04}}{m_{02}^2} - \frac{4}{m_{11}}\left(\frac{m_{31}}{m_{20}} + \frac{m_{13}}{m_{02}}\right)\right)$$

where $m_{k\ell} = \mathbb{E}\left[(X_1 - \mu_1)^k (X_2 - \mu_2)^\ell\right]$, $\mu_1 = \mathbb{E}[X_1]$ and $\mu_2 = \mathbb{E}[X_2]$.

The notion of rank plays a key role in the estimation of measures of association.

Definition 1.40. Let $(X^k)_{k=1,\dots,N}$ be a sample of size N of the random variable X and $\sigma \in \mathfrak{S}_N$ a random permutation such that $X_{\sigma(1)} \leq \dots \leq X_{\sigma(N)}$ a.s. (such a permutation is almost surely unique if X is continuous). The rank of X^k is defined by:

$$\text{rank}(X^k) = \sigma^{-1}(k)$$

It is the random position of X^k in the sorted sample $(X_{\sigma(k)})_{k=1,\dots,N}$.

The definition of the Spearman rho coupled with the expression of the linear correlation coefficient estimator given previously, we estimate the Spearman rho as being the linear correlation coefficient of the ranks of the observations. For the case where there is no tie in the observations, which is the case of interest for applications with continuous distributions, we are able to express this estimator in a more compact way:

Definition 1.41. Let $\left((X_1^k, X_2^k)\right)_{k=1,\ldots,N}$ be a sample of size N of the random vector $\boldsymbol{X} = (X_1, X_2)$. The Spearman rho estimator $\hat{\rho}_{S,N}(\boldsymbol{X})$ is the linear correlation coefficient estimator applied to the ranks of the given sample:

$$\hat{\rho}_{S,N}(\boldsymbol{X}) = \frac{\sum_{k=1}^{n}\left(\operatorname{rank}(X_1^k) - \overline{\operatorname{rank}}(X_1)\right)\left(\operatorname{rank}(X_2^k) - \overline{\operatorname{rank}}(X_2)\right)}{\sqrt{\sum_{k=1}^{N}\left(\operatorname{rank}(X_1^k) - \overline{\operatorname{rank}}(X_1)\right)^2 \sum_{k=1}^{N}\left(\operatorname{rank}(X_2^k) - \overline{\operatorname{rank}}(X_2)\right)^2}} \tag{1.32}$$

where $\overline{\operatorname{rank}}(X_1) = \frac{1}{N}\sum_{k=1}^{N}\operatorname{rank}(X_1^k)$ and $\overline{\operatorname{rank}}(X_2) = \frac{1}{N}\sum_{k=1}^{N}\operatorname{rank}(X_2^k)$. If there is no tie, i.e. $\forall i, j,\ (i \neq j) \Rightarrow (X_1^i \neq X_1^j \text{ or } X_2^i \neq X_2^j)$, the sampling Spearman rho $\hat{\rho}_{S,N}(X_1, X_2)$ is given by

$$\hat{\rho}_{S,N}(\boldsymbol{X}) = 1 - \frac{6\sum_{k=1}^{N}\left(\operatorname{rank}(X_1^k) - \operatorname{rank}(X_2^k)\right)^2}{N(N^2-1)} \tag{1.33}$$

The asymptotic properties of this estimator are given in the following theorems, deduced from the corresponding theorems for the linear correlation coefficient and the fact that $F_i(X_i)$ $(i = 1, 2)$ is uniformly distributed over $[0, 1]$ for continuous F_i:

Theorem 1.42. *Let $\boldsymbol{X}$ be a bi-dimensional continuous random vector. Then:*

$$\hat{\rho}_{S,N}(\boldsymbol{X}) \overset{a.s}{\to} \rho_S(\boldsymbol{X}) \textit{ when } N \to \infty$$

where $\rho_S(\boldsymbol{X})$ is the Spearman rho between X_1 and X_2, as defined in Definition 1.2.2.

Theorem 1.43. *Let $\boldsymbol{X}$ be a bi-dimensional continuous random vector. Then:*

$$\sqrt{N}\left(\hat{\rho}_{S,N}(\boldsymbol{X}) - \rho_S(\boldsymbol{X})\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_{\rho_S}^2) \textit{ when } N \to \infty$$

where the asymptotic variance $\sigma_{\rho_S}^2$ is given by:

$$\sigma_{\rho_S}^2 = \left(1 + \frac{\rho_S(\boldsymbol{X})^2}{2}\right)\frac{4(5 + 192\eta_{10})}{3(4\eta_{00} - 1)^2} + \frac{\rho_S(\boldsymbol{X})^2}{4}\left(\frac{342}{125} - \frac{12}{5}\left(\frac{24(\eta_{20} + \eta_{02}) - 1)}{4\eta_{00} - 1}\right)\right)$$

$$\textit{where } \eta_{k\ell} = \iint_{[0,1]^2}\left(u_1 - \frac{1}{2}\right)^k\left(u_2 - \frac{1}{2}\right)^\ell C(u_1, u_2)\,\mathrm{d}u_1\mathrm{d}u_2 \textit{ and } C \textit{ is the copula of } \boldsymbol{X}.$$

The definition of the Kendall tau leads to an estimator that can also be expressed easily in terms of the discordance and concordance of the observations when there is no tie. In this case, the estimator reads:

Definition 1.44. Let $\left((X_1^k, X_2^k)\right)_{k=1,\ldots,N}$ be a sample of size N of the random vector $\boldsymbol{X} = (X_1, X_2)$. The sampling Kendall tau $\hat{\tau}_N(X_1, X_2)$ is given by

$$\hat{\tau}_N(\boldsymbol{X}) = \frac{2}{N(N-1)}\sum_{1 \le i < j \le N}\operatorname{sgn}(X_1^i - X_1^j)\operatorname{sgn}(X_2^i - X_2^j) \tag{1.34}$$

The asymptotic properties of this estimator are given in the following theorems:

Theorem 1.45. *Let $\boldsymbol{X}$ be a bi-dimensional random vector. Then:*

$$\hat{\tau}_N(\boldsymbol{X}) \overset{a.s}{\to} \tau(\boldsymbol{X}) \text{ when } N \to \infty$$

Theorem 1.46. *Let $\boldsymbol{X}$ be a bi-dimensional random vector. Then:*

$$\sqrt{N}\left(\hat{\tau}_N(\boldsymbol{X}) - \tau(\boldsymbol{X})\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_\tau^2) \text{ when } N \to \infty$$

where the asymptotic variance σ_τ^2 is given by:

$$\sigma_\tau^2 = 4\,\mathbf{Var}\left[\mathbb{E}\left[\operatorname{sgn}(X_1 - X_1')\operatorname{sgn}(X_2 - X_2') \,|\, X_1, X_2\right]\right]$$

where $\boldsymbol{X}' = (X_1', X_2')$ is an independent copy of $\boldsymbol{X}$.

In contrast with the previous measures, no estimator for the upper or lower tail dependence coefficients has become standard, despite the large amount of research in this area, in relation with the estimation of **extrem values copulas** (see [KN00]). Being defined as a limit, these quantities are difficult to estimate, and except in fully parametric contexts, there will always be a trade-off between the bias (taking into account a large amount of the available data, including non-extreme ones) and the variance (taking into account only the most extreme data) of the estimator. We restrict the presentation to non-parametric estimators, based on the **empirical copula** defined here:

Definition 1.47. Let $\left((X_1^k, X_2^k)\right)_{k=1,\dots,N}$ be a sample of size N of the random vector $\boldsymbol{X} = (X_1, X_2)$. The **empirical copula** $\hat{C}_N$ of this sample is the bivariate function defined by:

$$\forall (u_1, u_2) \in [0,1]^2, \quad \hat{C}_N(u_1, u_2) = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{(\operatorname{rank}(X_1^k) \leq N u_1, \operatorname{rank}(X_2^k) \leq N u_2)} \tag{1.35}$$

We present a non-parametric estimators of the upper-tail coefficient based on the empirical copula of block maxima proposed in [SS04] and in [FJS05]:

Definition 1.48. Let $\left((X_1^k, X_2^k)\right)_{k=1,\dots,N}$ be a sample of size N of the random vector $\boldsymbol{X} = (X_1, X_2)$. Let m be a positive integer and $\ell = [N/m]$. We consider the sample $\left((x_1^{*j}, x_2^{*j})\right)_{j=1,\dots,m}$ of **componentwise block maxima**:

$$x_1^{*j} = \max\left\{X_1^i, i = 1 + (j-1)\ell, \dots, j\ell\right\}$$
$$x_2^{*j} = \max\left\{X_2^i, i = 1 + (j-1)\ell, \dots, j\ell\right\}$$

for $j = 1, \dots, m$. For a given integer threshold $0 < k(m) < m$, the upper tail coefficient λ_U can be estimated by:

$$\hat{\lambda}_{U,m}(\boldsymbol{X}) = 2 - \frac{1 - \hat{C}_m\left(\frac{m-k}{m}, \frac{m-k}{m}\right)}{1 - \frac{m-k}{m}}$$

The parameters m and k allow to deal with the bias/variance trade-off. The properties of this estimator are given in the following theorems, given in [SS04, Theorem 7] and [SS04, Corollary 2]:

Theorem 1.49. *Let $\boldsymbol{X}$ be a bi-dimensional random vector with continuous marginal distribution function. If the upper tail copula $\Lambda_U \neq 0$ exists and $k(m)$ is such that $k(m)/\log\log m \to 0$ as $m \to \infty$. Then:*

$$\hat{\lambda}_{U,m}(\boldsymbol{X}) \overset{a.s}{\to} \lambda_U(\boldsymbol{X}) \text{ when } m \to \infty$$

Theorem 1.50. *Let $\boldsymbol{X}$ be a bi-dimensional random vector with continuous marginal distribution function. If the upper tail copula $\Lambda_U \neq 0$ exists, possesses continuous partial derivatives, and satisfies the additional second order condition: it exists a function $A : \mathbb{R}_+ \to \mathbb{R}_+$ such that $A(t) \to 0$ as $t \to \infty$ and:*

$$\lim_{t\to\infty} \frac{\Lambda_U(\boldsymbol{u}) - (1-t)C(1-u_1/t, 1-u_2/t)}{A(t)} = g(\boldsymbol{u}) < \infty$$

locally uniformly for $\boldsymbol{u} \in [0,1]^2$ and some nonconstant function g.

Then, if $\sqrt{k(m)}A(m/k(m)) \to 0$ as $m \to \infty$:

$$\sqrt{k(m)}\left(\hat{\lambda}_{U,m}(\boldsymbol{X}) - \lambda_U(\boldsymbol{X})\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_U^2) \text{ when } m \to \infty$$

with

$$\begin{aligned}\sigma_U^2 =& \lambda_U(\boldsymbol{X}) + \left(\frac{\partial}{\partial x}\Lambda_U(1,1)\right)^2 + \left(\frac{\partial}{\partial y}\Lambda_U(1,1)\right)^2 + \\ & 2\lambda_U(\boldsymbol{X})\left(\left(\frac{\partial}{\partial x}\Lambda_U(1,1) - 1\right)\left(\frac{\partial}{\partial y}\Lambda_U(1,1) - 1\right) - 1\right).\end{aligned}$$

Conclusion

In this introductory chapter, we have introduced several concepts and measures linked with dependence modeling that will be used in the sequel of the manuscript. It covers both the probabilistic aspects linked with the distribution function of a random vector and the dependence quantification through scalar measures. We have also given some elements on statistical estimation of these measures given a set of multidimensional data.