

Modelling bivariate dependence - with application to efficiency of dependence tests

Modelli per dipendenza bivariata

Daniele De Martini¹
Dipartimento SEMEQ
Università del Piemonte Orientale
demartin@eco.unipmn.it

Erika Vespa
Dipartimento di Matematica
Università di Genova

Riassunto: In questo lavoro si considerano i test di indipendenza ed i metodi per valutarne l'efficienza. Si evidenzia dapprima come molti degli usuali test siano funzione della copula empirica, che risulta essere statistica sufficiente. Pertanto la potenza dei test non dipende dalle leggi marginali ma soltanto dalla copula teorica. In seguito, vengono considerati i test di dipendenza monotona per i quali si propone una modellizzazione parametrizzata della curva di potenza basata sulla copula di una generica legge congiunta, anche non parametrica. Infine, si dimostra che la curva di potenza ottenuta risulta monotona crescente per alternative di dipendenza monotona positiva.

Keywords: power of independence statistical tests; copula.

1. Introduction

In order to evaluate the efficiency on nonparametric tests, Asymptotic Relative Efficiency (ARE) techniques are usually applied. In the context of independence testing, a review of the most important results can be found in Nikitin (1995), where basic ARE techniques are also presented. Many results regard traditional Spearman's Rho and Kendall's Tau, or concern statistics based on differences between bivariate empirical distribution functions. Recently, Conti and Nikitin (1999) explored efficiency properties of generalized versions of Gini's cograduation coefficient and Spearman's footrule. Moreover, ARE for independence tests with random sample size are studied by Beghin and Nikitin (1999).

On the other hand, ARE is not an indisputable result, as there are also other useful techniques for comparing nonparametric statistical tests (Nikitin (1995)). Indeed, it is a common way of thinking that when among two different statistical tests, one presents ARE better than the other, this is not a good reason for preferring it in any circumstance.

In general, other measures based on the power function of the test can be used. Indeed, Modarres (1996) proposed a bootstrap method to compute the power curve of the generalized correlation coefficient when the distribution of the data is unknown, as far as in nonparametric frameworks it is natural to consider the distribution unknown. Although we agree with the methodology, we found the algorithm proposed by Modarres (1996) inadequate, because it did not provide an increasing power curve. This latter observation motivated this work, which firstly present a theorem on the power of the most used independence statistical tests. Then, this paper deals with a theoretical result on the power function of statistical tests for monotone dependence. Indeed, we propose a parametric

¹Il lavoro è stato svolto nell'ambito del COFIN99: Metodi di Inferenza Statistica per Problemi Complessi

model based on copulas which is useful to define the power function, that is shown to be strictly increasing.

2. Independence statistical tests, sufficient statistic, power

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be n independent couples of r.v. with joint distribution F_{XY} , whose marginal distributions F_X and F_Y are continuous. The copula related to F_{XY} is $C_{XY}(u, v) = F_{XY}(F_X^{-1}(u), F_Y^{-1}(v))$. Note also that $F_{XY}(x, y) = C_{XY}(F_X(x), F_Y(y))$. It is known that $C^-(u, v) = \max(u + v - 1, 0) \leq C_{XY}(u, v) \leq \min(u, v) = C^+(u, v)$, where C^- and C^+ are the copulas related to the maximum negative dependence and the maximum positive dependence respectively. The copula of the complete independence is $C^0 = uv$. More details on copulas can be found in Nelsen (1998). When F_X and F_Y are continuous the copula C_{XY} is unique (Schweizer and Wolff (1981)). The statistical hypotheses are $H_0 : F_{XY}(x, y) = F_X(x)F_Y(y), \forall (x, y) \in R^2$ and $H_1 : F_{XY}(x, y) \neq F_X(x)F_Y(y), (x, y) \in I_1 \neq \emptyset$. But, as the parameter under testing is the *Copula*, i.e. the functional parameter C_{XY} , equivalent hypotheses are: $H_0 : C_{XY}(u, v) = uv \forall (u, v) \in [0, 1]^2$ and $H_1 : C_{XY}(u, v) \neq uv, (u, v) \in I_1 \neq \emptyset$.

Let $X_{(i)}$ and $Y_{(j)}$ be i -th and j -th order statistics of X 's and Y 's respectively. We can now introduce the *Empirical Copula*, which is defined as

$$C_{XY}^n(i/n, j/n) = \sum_{l=1}^n I(X_l \leq X_{(i)}, Y_l \leq Y_{(j)})/n.$$

Theorem 1. The distribution of C_{XY}^n depends solely on C_{XY} (see also Deheuvels (1979)). Moreover, the distributions of the empirical marginal distributions do not depend on C_{XY} . Hence, C_{XY}^n is a sufficient statistic for C_{XY} .

Proof. (Sketch.) The distribution of C_{XY}^n can be computed calculating the probability of each ordering on X 's and Y 's, which is the integral of the joint density $f_{X,Y}$ over the particular ordering. Then, $f_{X,Y}$ can be written as a function of $C_{X,Y}, F_X, F_Y, f_X$ and f_Y , and with the change of variable $u = F_X(x)$ and $v = F_Y(y)$ it becomes a function solely of $C_{X,Y}$.

Many nonparametric statistical tests, among them Kolmogorov-Smirnov-type independence tests and tests based on the General Measure of Monotone Dependence (Cifarelli et al. (1996)), can be defined as functions of the empirical copula (see Schweizer and Wolff (1981), Nelsen (1998)). For example, Spearman's Rho is:

$$12 \sum_{i=1}^n \sum_{j=1}^n [C_{XY}^n(i/n, j/n) - ij/n^2] / (n^2 - 1).$$

Corollary 1. The power of statistical tests which are functions of C_{XY}^n depends solely on C_{XY} . (The proof is a direct consequence of Theorem 1.)

We remark that the power of statistical tests cited above does not depend on the marginal distributions.

3. Monotone dependence, copulas dependence model, power function

Modarres (1996) proposed a bootstrap algorithm to compute the power function of correlation tests. Such an algorithm presented some statistical errors and, most of all, was

inadequate as it provided a power curve not increasing, because it did not consider the copula as the fundamental parameter for the power. Moreover, Modarres studied non-parametric tests for monotone dependence, that is any tendency to concordance or to discordance. Then, in this context, correct hypotheses are $H_0 : F_{XY}$ is *indifferent* and $H_1 : F_{XY}$ has a tendency to *concordance* (see Cifarelli et al. (1996)).

Note that independence implies indifference but the converse is not true, as it can be easily shown using the copula given by the mean of the maximum positive and negative dependence $\tilde{C}^0 = (C^+ + C^-)/2$: indeed $\tilde{C}^0 \neq C^0$ but $R(\tilde{C}^0) = R(C^0) = 0$, where R is a functional meaning monotone dependence.

Now, we want to define a one-dimensional parametric model related to changes in monotone dependence providing a power function which must be increasing when the parameter varies from values meaning negative dependence to values of positive dependence, according to the concept of power function.

Note that, when F_{XY} is parametrized we can carry the parametrization over C_{XY} (for example with Normal distributions we find the Gaussian copula); otherwise, we can model a parametrization on F_{XY} . This is analogous to the one-dimensional location shift problem, when a density $g(x)$ with mean equal to 0 is parametrized by $g_\mu(x) = g(x - \mu)$, with a generic mean μ .

As far as the power of many nonparametric dependence tests depends solely on the copula, we model a continuous change between C^- and C^+ , through C^0 and by the nature of the dependence in F_{XY} , i.e. the copula C_{XY} . We make use of the parameter $\theta \in [-1, 1]$, obtaining $F_{XY}(\cdot, \cdot; \theta)$ such that: 1) for every θ , F_X and F_Y do not depend on θ ; 2) $F_{XY}(\cdot, \cdot; 0) = F_X(\cdot)F_Y(\cdot)$; 3) $F_{XY}(\cdot, \cdot; -1)$ and $F_{XY}(\cdot, \cdot; 1)$ are degenerate joint distributions related to maximum negative dependence and maximum positive dependence respectively.

In practice, we build a path of copulas from C^- to C^+ through C^0 , which is characterized by C_{XY} , the copula derived from a given F_{XY} . In order to define a path passing through C_{XY} , we consider a dependence parameter $\rho = R(C_{XY})$ (for example Spearman's Rho). We also need to define a symmetric copula C_{XY}^s ; then, we find that the copulas of $(-X, Y)$ and $(X, -Y)$, which are respectively $C_1(u, v) = v - C(1 - u, v)$ and $C_2(u, v) = u - C(u, 1 - v)$, both have dependence parameter $-\rho$. Hence, we set C_{XY}^s as the mean of the copulas given by the two symmetric relations: $C_{XY}^s(u, v) = [C_1(u, v) + C_2(u, v)]/2$, and we obtain $R(C_{XY}^s) = -\rho$. Without loss of generality, we suppose $\rho > 0$ and finally we define the parametrization $C_{XY}(\theta)$ related to the copula C_{XY} of F_{XY} , through a linear combination of copulas:

$$C_{XY}(\theta) = \begin{cases} ((1 + \theta)C_{XY}^s - (\rho + \theta)C^-)/(1 - \rho) & \text{if } \theta \in [-1, -\rho) \\ ((\rho + \theta)C^0 - \theta C_{XY}^s)/\rho & \text{if } \theta \in [-\rho, 0) \\ ((\rho - \theta)C^0 + \theta C_{XY})/\rho & \text{if } \theta \in [0, \rho) \\ ((1 - \theta)C_{XY} + (\theta - \rho)C^+)/ (1 - \rho) & \text{if } \theta \in [\rho, 1] \end{cases}$$

A simpler linear combination of copulas is also often used to model dependence in ARE computations for independence tests (see for example Conti and Nikitin (1999)). Now, note that: 1) $C_{XY}(\theta)$ is a copula; 2) low values of θ imply a dependence near to $Y = h(X)$, where h is a monotone decreasing function; 3) values near zero lead to an almost null dependence, as the copula $C_{XY}(\theta)$ is near the independence copula C^0 ; 4) high values of θ imply a dependence near to $Y = h'(X)$, where h' is a monotone increasing function; 5) we can define the joint distribution associated to θ , based on $C_{XY}(\theta)$, F_X and F_Y :

$F_{XY}(x, y; \theta) = C_{XY}(F_X(x), F_Y(y); \theta)$; 6) when $\theta = 0$, $F_{XY}(x, y; 0) = F_X(x)F_Y(y)$; 7) when $\theta = \rho$, the original joint distribution: $F_{XY}(x, y; \rho) = F_{XY}(x, y)$ is obtained.

We remark that such a parametrization $C_{XY}(\theta)$, can be defined “The Class of Copulas generated by C_{XY} ”, as $C_{XY}(\theta)$ contains a path of monotone dependence tendency between X and Y , induced by the “dependence nature” implicit in F_{XY} .

Now, let $(X_1, Y_1), \dots, (X_n, Y_n)$ be n independent couples of r.v. with joint distribution $F_{XY}(\cdot, \cdot; \theta)$. Then, defining the test statistic $r_n = R(C_{XY}^n)$ and the critical value of the test $r_{n,\alpha}$, the power function is: $pow_R(\theta) = P(r_n \geq r_{n,\alpha} | F_{XY}(\cdot, \cdot; \theta))$.

Theorem 2. The power function, when $\theta \in [-1, 1]$, is strictly increasing. Moreover, when α is such that $r_{n,\alpha} > \min\{r_n\}$, then $pow_R(-1) = 0$, and when $r_{n,\alpha} \leq \max\{r_n\}$, then $pow_R(1) = 1$.

Proof. (Sketch.) Consider $\rho \leq \theta' \leq \theta'' \leq 1$. $F_{XY}(\cdot, \cdot; \theta')$ and $F_{XY}(\cdot, \cdot; \theta'')$ are mixtures of $F_{XY}(\cdot, \cdot)$ and $F_{XY}(\cdot, \cdot; 1)$, with known weights. Split the power considering k pairs coming from $F_{XY}(\cdot, \cdot; 1)$, $k = 0, \dots, n$. Use $P(r_n \geq r_{n,\alpha} | k \text{ pairs from } F_{XY}(\cdot, \cdot; 1)) \leq P(r_n \geq r_{n,\alpha} | k + 1 \text{ pairs from } F_{XY}(\cdot, \cdot; 1))$.

This result can be extended to the general alternative hypothesis of *not indifference*, obtaining a power function strictly decreasing when $\theta \in [-1, 0]$ and strictly increasing when $\theta \in [0, 1]$.

Finally, it is worth noting that the power function depends on the given modelization, as the latter is only one of the paths which, through the use of copulas, walk from hypothesis of independence to monotone dependence. Other models may be considered, for example non linear combinations of copulas, which should bring other power functions. But, also ARE results often depend on the choice of copula-based models. Moreover, the linear convex combination of copulas we proposed appears as the most natural.

This work can be developed considering applications of the copula-based-dependence-model to comparisons among the power functions of the most used monotone dependence tests and to the estimation the power function of a monotone dependence test.

References

- Beghin, L. and Nikitin, Y. (1999). Approximate asymptotic bahadur efficiency of independence tests with random sample size, *J.Ital.Statist.Soc.* **8**(1), 1–23.
- Cifarelli, D., Conti, P. and Regazzini, E. (1996). On the asymptotic distribution of a general measure of monotone dependence, *Ann. Statist.* **25**, 1386–1399.
- Conti, P. and Nikitin, Y. (1999). Asymptotic efficiency of independence tests based on gini’s rank association coefficient, Spearman’s footrule and their generalizations, *Commun.Statist.Theory Meth.* **28**(2), 453–465.
- Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés - un test non paramétrique d’indépendance, *Acad.Roy.Belg.Bull.Cl.Sci. 5^e ser.* (65), 274–292.
- Modarres, R. (1996). Bootstrap power of the generalized correlation coefficient, *Stat.Comp.* **6**, 139–145.
- Nelsen, R. (1998). *An Introduction to Copulas*, Springer.
- Nikitin, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*, Cambridge University Press.
- Schweizer, B. and Wolff, E. (1981). On nonparametric measures of dependence for random variables, *Ann. Statist.* **9**(4), 879–885.