Chapter 3
Visualizing Association Structure in Bivariate Copulas Using New Dependence Function

Teresa Ledwina

Abstract Measuring a strength of dependence of random variables is an important problem in statistical practice. We propose a new function valued measure of dependence of two random variables. It allows one to study and visualize explicit dependence structure, both in some theoretical models and empirically, without prior model assumptions. This provides a comprehensive view of association structure and makes possible much detailed inference than based on standard numeric measures of association. In this contribution, we focus on copula-based variant of the measure. We present theoretical properties of the new measure of dependence and discuss estimation of it. Some artificial and real data examples illustrate the behavior and practical utility of the measure and its estimator.

3.1 Introduction

Measuring a strength of dependence of two random variables has long history and wide applications. Detailed information can be found in Drouet Mari and Kotz [9] as well as Balakrishnan and Lai [2], for example. Most of measures of dependence, introduced in vast literature on the subject, are scalar ones. Such indices are called global measures of dependence. However, nowadays there is strong evidence that an attempt to represent complex dependence structure via a single number can be misleading. To overcome this drawback, some local dependence functions have been introduced as well. In particular, Kowalczyk and Pleszczyńska [14] invented function valued measure of monotonic dependence, based on some conditional expectations and adjusted to detect dependence weaker than the quadrant one. Next, Bjerve and Doksum [5], Bairamov et al. [1] and Li et al. [17], among others, introduced local dependence measures based on regression concepts. See the last mentioned paper for more information. Holland and Wang [11] defined the local dependence function, which mimics cross-product ratios for bivariate densities and treats the two variables in a symmetrical way. This function valued measure has several appealing

T. Ledwina (✉)
Institute of Mathematics, Polish Academy of Sciences, ul. Kopernika 18, 51-617 Wrocław, Poland
e-mail: ledwina@impan.pan.wroc.pl

A. Steland et al. (eds.), *Stochastic Models, Statistics and Their Applications*,
Springer Proceedings in Mathematics & Statistics 122,
DOI 10.1007/978-3-319-13881-7_3

properties and received considerable attention in the literature; cf. Jones and Koch [13] for discussion and references. However, on the other hand, this measure has some limitations: it is not normalized, requires existence of densities of the bivariate distribution, and is intimately linked to strong form of dependence, the likelihood ratio dependence. Recently, Tjøstheim and Hufthammer [21] extensively discussed the role and history of local dependence measures in finance and econometrics. They also proposed the new local dependence measure, the local correlation function, based on approximating of bivariate density locally by a family of Gaussian densities. Similarly as the measure of Holland and Wang [11], this measure treats both variables on the same basis. Though the idea behind the construction of this measure is intuitive one its computation and estimation is a difficult and complex problem. In Berentsen et al. [3] this theory is applied to describe dependence structure of different copula models. In particular, this work strongly emphasizes a need for intuitive and informative diagnostic plots.

In this paper, we propose the new function valued measure of dependence of two random variables X and Y and present its properties. The measure has simple form and its definition exploits cumulative distribution functions (cdf's), only. In particular, we do not assume existence of a density of the observed vector. We focus here on copula-based variant of the measure which corresponds to some cdf on $[0, 1]^2$ with uniform marginals. General case is presented in Ledwina [15]. The measure takes values in $[-1, 1]$ and treats both variables in a symmetrical way. The measure preserves the correlation order, or equivalently the concordance order, which is the quadrant order restricted to the class of distributions with fixed marginals. In particular, it is non-negative (non-positive) if and only if X and Y are positively (negatively) quadrant dependent. Quadrant dependence is relatively weak, intuitive and useful dependence notion, widely used in insurance and economics; see Dhaene et al. [8] for an evidence and further references. The new measure obeys several properties formulated in the literature as useful or desirable. It allows for readable visualization of departures from independence. Simple and natural estimator of the copula-based measure in the i.i.d. case is proposed and its appealing properties are discussed. The estimator is simply standardized empirical copula. Due to theoretical results proved in Ledwina and Wyłupek [16], the estimator can be effectively exploited to assess graphically underlying bivariate dependence structure and to build some formal local and global tests. For some details see Sect. 3.3. Two illustrative examples are given in Sect. 3.3 to support utility of the new solution.

3.2 Copula-based measure of dependence and its estimate

Consider a pair of random variables X and Y with joint cdf H and marginals F and G. In this paper, to avoid technicalities and to concentrate on the main idea, we restrict attention to cdf's H with continuous marginals. Under such a restriction there exists a unique copula C such that $H(x, y) = C(F(x), G(y))$. In other words, C is the restriction to the unit square of the joint cdf of $U = F(X)$ and $V = G(Y)$.

The copula captures the dependence structure among X and Y, irrespective of their marginal cdf's. This is important in many applications. For the related discussion see Póczos et al. [19]. Below we show that properly standardized copula function can be seen to be well defined function valued dependence measure.

Namely, set

$$q(u,v) = q_C(u,v) = \frac{C(u,v) - uv}{\sqrt{uv(1-u)(1-v)}}, \quad (u,v) \in (0,1)^2, \tag{3.1}$$

and define additionally $w(u,v) = 1/\sqrt{uv(1-u)(1-v)}$.

Following Ledwina and Wyłupek [16], notice that the value of q at (u,v) can be interpreted as correlation coefficient of two specific functions of U and V. Namely, for $u \in (0,1)$ and $s \in [0,1]$ consider

$$\phi_u(s) = -\sqrt{(1-u)/u}\,\mathbb{1}_{[0,u]}(s) + \sqrt{u/(1-u)}\,\mathbb{1}_{(u,1]}(s).$$

Then

$$q(u,v) = E_C\big[\phi_u(U)\cdot\phi_v(V)\big] = Cov_C\big(\phi_u(U),\phi_v(V)\big) = Corr_C\big(\phi_u(U),\phi_v(V)\big). \tag{3.2}$$

Remark 3.1 The last expression in (3.2) shows indeed that the function q is based on aggregated local correlations. Moreover, the second expression in (3.2) implies that $q(u,v)$ can be interpreted as Fourier coefficient of C pertaining to the quasi-monotone function $\phi_u(s)\cdot\phi_v(t)$, $(s,t) \in [0,1]^2$.

The measure q fulfills natural postulates, motivated by the axioms formulated in Schweizer and Wolff [20] and updated in Embrechts et al. [10].

Proposition 3.1 *The copula based measure of dependence q, given by* (3.1), *has the following properties.*

1. $-1 \le q(u,v) \le 1$ *for all* $(u,v) \in (0,1)^2$.
2. *By Fréchet–Hoeffding bounds for copulas, the property* 1 *can be further sharpened to* $B_*(u,v) \le q(u,v) \le B^*(u,v)$, $(u,v) \in (0,1)^2$, *where* $B_*(u,v) = w(u,v) \times [\max\{u+v-1,0\} - uv]$ *and* $B^*(u,v) = w(u,v)[\min\{u,v\} - uv]$.
3. q *is maximal (minimal) if and only if* $Y = f(X)$ *and* f *is strictly increasing (decreasing) a.s. on the range of* X.
4. $q(u,v) \equiv 0$ *if and only if* X *and* Y *are independent.*
5. *The equation* $q(u,v) \equiv c$, c *a constant, can hold true if and only if* $c = 0$.
6. q *is non-negative (non-positive) if and only if* (X,Y) *are positively (negatively) quadrant dependent.*
7. q *is invariant to strictly increasing a.s. on ranges of* X *and* Y*, respectively, transformations.*
8. *If* X *and* Y *are transformed by strictly decreasing a.s. functions then* $q(u,v)$ *transforms to* $q(1-u,1-v)$.
9. *If* f *and* g *are strictly decreasing a.s. on ranges of* X *and* Y*, respectively, then* q*'s for the pairs* $(f(X),Y)$ *and* $(X,g(Y))$ *take the forms* $-q(1-u,v)$ *and* $-q(u,1-v)$*, accordingly.*

10. *q respects concordance ordering, i.e. for cdf's* H_1 *and* H_2 *with the same marginals and pertaining copulas* C_1 *and* C_2, $H_1(x, y) \le H_2(x, y)$ *for all* $(x, y) \in \mathbb{R}^2$ *implies* $q_{C_1}(u, v) \le q_{C_2}(u, v)$ *for all* $(u, v) \in (0, 1)^2$.
11. *If* (X, Y) *and* (X_n, Y_n), $n = 1, 2, \ldots,$ *are pairs of random variables with joint cdf's* H *and* H_n, *and the pertaining copulas* C *and* C_n, *respectively, then weak convergence of* $\{H_n\}$ *to* H *implies* $q_{C_n}(u, v) \to q_C(u, v)$ *for each* $(u, v) \in (0, 1)^2$.

Proof The property 1 follows from (3.2), 3 is a consequence of 2. To justify 5 observe that the equation is equivalent to $C(u, v) = C_c(u, v) = uv + c\sqrt{uv(1-u)(1-v)}$. Since C is quasi-monotone, then $C_c(u, v)$ should also possess such a property. Since $C_c(u, v)$ is absolutely continuous then quasi-monotonicity is equivalent to $\frac{\partial^2}{\partial u \partial v} C_c(u, v) \ge 0$ for almost all $(u, v) \in [0, 1]^2$ (in the Lebesgue measure); cf. Cambanis et al. [7]. However, $\frac{\partial^2}{\partial u \partial v} C_c(u, v) = 1 + c[u - 1/2][v - 1/2]w(u, v)$ and for $c \neq 0$ this expression can be negative on the set of positive Lebesgue measure. Properties 7–9 follow from Theorem 3 in Schweizer and Wolff [20]. The convergence in 11 is due to continuity of C. The remaining properties are immediate. □

Remark 3.2 The properties 4 and 7–9 provide some compromise to too demanding postulates P4 and P5 discussed in Embrechts et al. [10]. The property 5 is very different from respective property of the local dependence function of Holland and Wang [11] which is constant for the bivariate normal distribution and some other models; cf. Jones [12] for details.

Now, we discuss briefly estimation of q. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from cdf H. Furthermore, let R_i be the rank of X_i, $i = 1, \ldots, n$, in the sample $X_1, \ldots, X_n$ and S_i the rank of Y_i, $i = 1, \ldots, n$, within $Y_1, \ldots, Y_n$. Simple estimate of C has the form

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(\frac{R_i}{n+1} \le u, \frac{S_i}{n+1} \le v \right), \quad (u, v) \in [0, 1]^2. \tag{3.3}$$

We shall consider the following estimator of q.

$$Q_n(u, v) = w(u, v)\big[C_n(u, v) - uv\big] = \frac{C_n(u, v) - uv}{\sqrt{uv(1-u)(1-v)}}, \quad (u, v) \in (0, 1)^2. \tag{3.4}$$

Moreover, we set

$$L_n(u, v) = \sqrt{n}\, Q_n(u, v) \tag{3.5}$$

for the standardized version of this estimate. So, L_n is the standardized empirical copula. Simple algebra yields that for any $(u, v) \in (0, 1)^2$ it holds

$$L_n(u, v) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_u\left(\frac{R_i}{n+1} \right) \phi_v\left(\frac{S_i}{n+1} \right) + O\left(\frac{1}{\sqrt{n}} \right). \tag{3.6}$$

Table 3.1 Simulated critical values of the test rejecting independence for large values of $|L_n(u, v)|$ for two selected (u, v), versus n and α

(u, v)	$\alpha = 0.01$					$\alpha = 0.05$				
	n					n				
	200	300	400	500	600	200	300	400	500	600
$(\frac{1}{2}, \frac{1}{2})$	2.546	2.540	2.600	2.504	2.613	1.980	1.848	2.000	1.968	1.960
$(\frac{1}{16}, \frac{1}{16})$	2.753	2.879	2.933	2.349	2.591	1.546	1.894	2.080	1.586	1.894

Therefore, up to deterministic term of the order $O(1/\sqrt{n})$, the standardized estimator $L_n(u, v)$ is a linear rank statistic with the quasi-monotone score generating function $\phi_u \cdot \phi_v$. Moreover, the definition of L_n and Proposition 1 in Ledwina and Wyłupek [16] yield that

$$P_{C_1}\big(L_n(u, v) \geq c\big) \geq P_{C_2}\big(L_n(u, v) \geq c\big) \tag{3.7}$$

for any $(u, v) \in (0, 1)^2$, any c, any n, and any two copulas C_1 and C_2 such that C_1 has larger quadrant dependence than C_2. Summarizing the above mentioned results, let us note that under independence $L_n(u, v)$ is distribution free. So, given n, under independence, the significance of the obtained values of this statistic can be easily assessed on a basis of simple simulation experiment. For large n one can rely on asymptotic $N(0, 1)$ law of $L_n(u, v)$. Due to (3.7), similar conclusions follow if one likes to verify hypothesis asserting that $q_C(u, v) \geq 0$. In particular, given n, we are able to control the significance level over the whole set of positively quadrant dependent distributions. Moreover, (3.7) implies that different levels of strength of quadrant dependence of the underlying H's shall be adequately quantified by order preserving $L_n(u, v)$'s. These results make the values of $L_n(u, v)$, $(u, v) \in (0, 1)^2$, a useful diagnostic tool allowing for easy graphical presentation and precise evaluation of significance of different types of departures from independence. Heat map of $L_n(u, v)$'s helps also to recognize regions in $(0, 1)^2$ in which independence, positive quadrant dependence and, in consequence, some stronger forms of positive dependence, etc are invalidated. This is obviously not the case when ones relies on graphs of $C_n(u, v)$ or $C_n(u, v) - uv$, solely. Moreover, without using the 'magnifying glass' $w(u, v)$ departures from independence can be hardly seen in some cases.

To close, note that, given u and v, the score generating function $\phi_u \cdot \phi_v$, appearing in (3.6), is not smooth one and takes on at most four possible values, only. This causes that, under independence, the convergence of $L_n(u, v)$ to the limiting $N(0, 1)$ law is not very fast. Moreover, the rate of convergence is expected to depend on u and v, with the least favorable situation when (u, v) is close to the vertices of the unit square. We illustrate these aspects in Table 3.1, where simulated critical values of the test rejecting independence for large values of $|L_n(u, v)|$ are given under two choices of (u, v)'s, five different sample sizes, and two selected significance levels α.

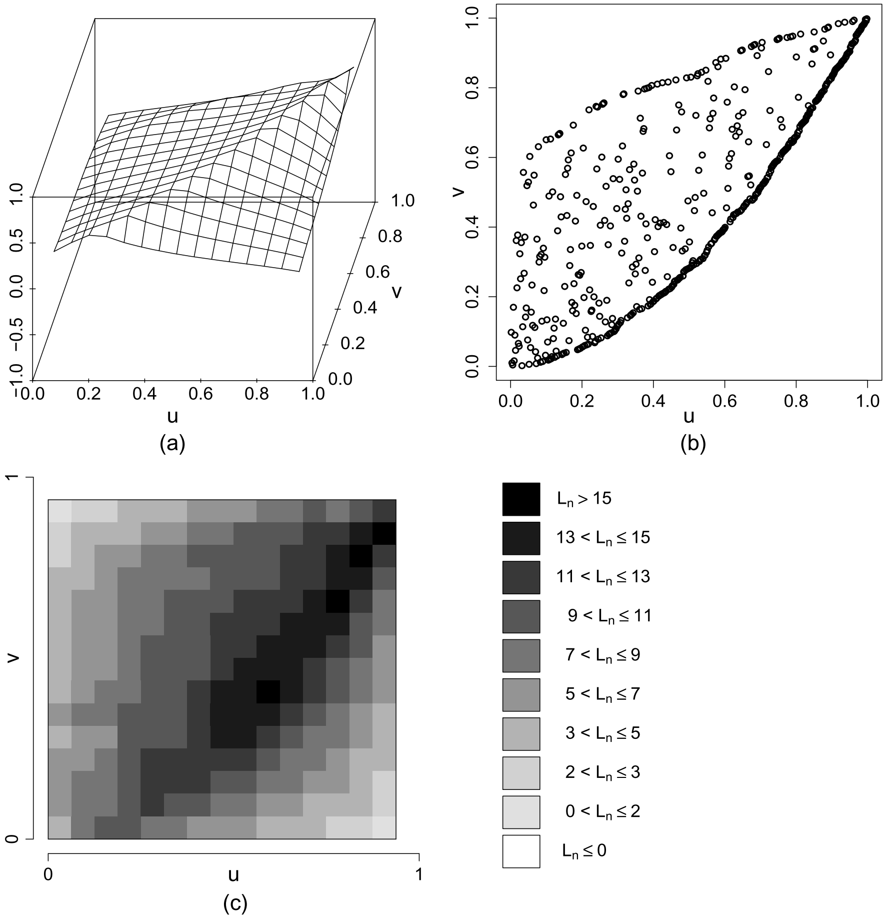

Fig. 3.1 (**a**): dependence function $q(u, v)$ for the Mai–Scherer copula; (**b**): scatter plot of $(R_i/(n+1), S_i/(n+1))$, $i = 1, \ldots, n$, $n = 500$, of simulated observations from the copula; (**c**): standardized estimator $L_n(u, v)$ of $q(u, v)$ on the grid $\mathbb{G}_{16}$. $L_* = 1.5$, $L^* = 16.1$

3.3 Illustration

3.3.1 Example 1: Extreme Value Copula

We start with simulated data set of size $n = 500$ from Mai–Scherer copula given by $C(u, v) = C_{a,b}(u, v) = \min\{u^a, v^b\}\min\{u^{1-a}, v^{1-b}\}$, $a = 0.9$, $b = 0.5$; cf. Mai and Scherer [18], p. 313. The copula possesses a singular part. In Fig. 3.1 we show dependence functions $q(u, v)$ for this model. The function is accompanied by scatter plots of pseudo-observations $(R_i/(n+1), S_i/(n+1))$, $i = 1, \ldots, 500$, from the simulated sample. The scatter plot nicely exhibits the singularity. Panel (c) in this figure displays respective heat map of standardized correlations $L_n(u, v)$'s calcu-

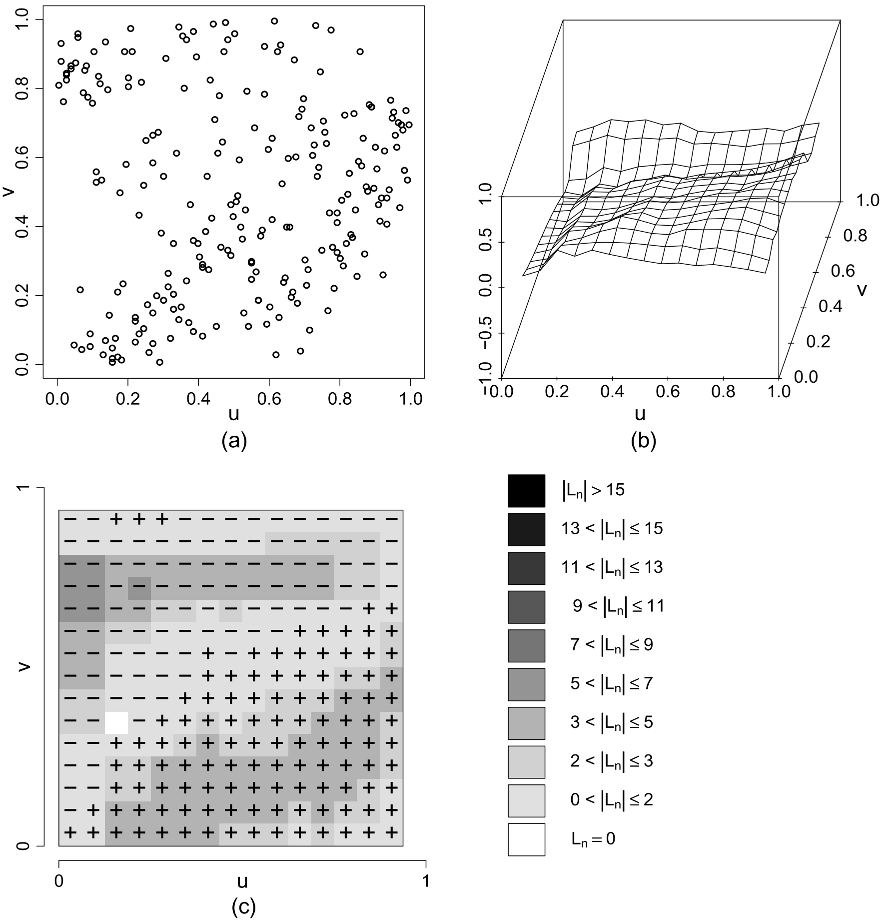

Fig. 3.2 **(a)**: scatter plot of $(R_i/(n+1), S_i/(n+1))$, $i = 1, \ldots, n$, $n = 230$, for aircraft data; **(b)**: estimator $Q_n(u, v)$ of $q(u, v)$ on the grid $\mathbb{G}_{16}$; **(c)**: standardized estimator $L_n(u, v) = \sqrt{n}Q_n(u, v)$ on the grid $\mathbb{G}_{16}$. $L_* = -6.5$, $L^* = 4.6$

lated on the grid $\mathbb{G}_{16} = \{(u, v) : u = i/16, v = j/16,\ i, j = 1, \ldots, 15\}$. Each of 225 squares of size 0.0625×0.0625 represents the respective value of L_n in its upper-right corner. To simplify reading, the heat map is accompanied with two numbers

$$L_* = \min_{1 \le i, j \le 15} L_n(i/16, j/16) \quad \text{and} \quad L^* = \max_{1 \le i, j \le 15} L_n(i/16, j/16).$$

The copula represents positively quadrant dependent distribution. Under such dependence large values of U tend to associate large values of V and similar pattern applies to small values. This tendency is nicely seen in the figure. The points of the grid $\mathbb{G}_{16}$ in which the estimated correlations Q_n are significant on the levels 0.05 and 0.01 can be easily identified; cf. Table 3.1 and related comments. Some possibility of testing for positive local and/or global dependence is sketched in Sect. 3.3.2.

Next example follows similar pattern. It concerns real data set considered earlier by Jones and Koch [13] and Berentsen and Tjøstheim [4]. However, in contrast to our approach based on scatter plots, they investigated the original bivariate observations. Below we use the same scale of intensity of colors in the heat map as above. This allows one to compare how different degrees of association are reflected by our estimators.

3.3.2 Example 2: Aircraft Data

Consider $n = 230$ aircraft span and speed data, on log scales, from years 1956–1984, reported and analyzed in Bowman and Azzalini [6]. We summarize the data in Fig. 3.2. Since in this example both negative and positive correlations appear, we added respective signs to the colors in the heat map. The figure exhibits that small and moderately large values of log speed are positively associated with log span, while for the remaining cases the relation is reversed. Two, approximately symmetrically located, regions of relatively strong dependence are seen. In general, in this example, the strength of dependence is weaker than in the previous case.

Bowman and Azzalini [6], p. 42, used these data to discuss some drawbacks of standard correlation measures when applied to invalidate independence. Indeed, for these data classical Pearson's, Spearman's and Blomqvist's rank statistics for assessing lack of association yield simulated p-values 0.81, 0.74, and 0.79, respectively. Kendall's rank correlation gives simulated p-value 0.31, which also seems to be too high, when one is looking at the magnitude of standardized local correlations in Fig. 3.2. Combining the local correlations into global statistic $L^o = \max_{1\le i,j\le 15} |L_n(i/16, j/16)|$, with large values being significant, basing on 10 000 Monte Carlo runs, we get p-value 0 for such global independence test. This shows that local correlations prove to be more informative than each of the above mentioned single classical global indices of association. Moreover, statistics L_* and L^* can be successfully applied to detect positive and negative quadrant dependence; cf. Ledwina and Wyłupek [16] for details on a very similar solution to L_*.

For further examples and more detailed discussion see Ledwina [15].

Acknowledgements Research was supported by the Grant N N201 608440 from the National Science Center, Poland. The author thanks Dr. Grzegorz Wyłupek for his help in preparing figures and the table of this article and for useful remarks. The author also thanks a referee for the careful reading of the manuscript and the useful comments.

References

1. Bairamov I, Kotz S, Kozubowski TJ (2003) A new measure of linear local dependence. Statistics 37:243–258
2. Balakrishnan N, Lai Ch-D (2009) Continuous bivariate distributions. Springer, Dordrecht

3. Berentsen G, Støve B, Tjøstheim D, Nordbø T (2014) Recognizing and visualizing copulas: an approach using local Gaussian approximation. Insur Math Econ 57:90–103
4. Berentsen G, Tjøstheim D (2014) Recognizing and visualizing departures from independence in bivariate data using local Gaussian correlation. Stat Comput 24:785–801
5. Bjerve S, Doksum K (1993) Correlation curves: measures of association as function of covariate values. Ann Stat 21:890–902
6. Bowman AW, Azzalini A (1997) Applied smoothing techniques for data analysis. Clarendon Press, Oxford
7. Cambanis S, Simons G, Stout W (1976) Inequalities for $Ek(X, Y)$ when marginals are fixed. Z Wahrscheinlichkeitstheor Verw Geb 36:285–294
8. Dhaene J, Denuit M, Vanduffel S (2009) Correlation order, merging and diversification. Insur Math Econ 45:325–332
9. Drouet Mari D, Kotz S (2001) Correlation and dependence. Imperial College Press, London
10. Embrechts P, McNeil A, Straumann D (2002) Correlation and dependency in risk management: properties and pitfalls. In: Dempster M, Moffatt H (eds) Risk management: value at risk and beyond. Cambridge University Press, Cambridge, pp 176–223
11. Holland PW, Wang YJ (1987) Dependence function for continuous bivariate densities. Commun Stat, Theory Methods 16:863–876
12. Jones MC (1998) Constant local dependence. J Multivar Anal 64:148–155
13. Jones MC, Koch I (2003) Dependence maps: local dependence in practice. Stat Comput 13:241–255
14. Kowalczyk T, Pleszczyńska E (1977) Monotonic dependence functions of bivariate distributions. Ann Stat 5:1221–1227
15. Ledwina T (2014) Dependence function for bivariate cdf's. arXiv:1405.2200v1 [stat.ME]
16. Ledwina T, Wyłupek G (2014) Validation of positive quadrant dependence. Insur Math Econ 56:38–47
17. Li R, Cheng Y, Fine JP (2014) Quantile association regression models. J Am Stat Assoc 109:230–242
18. Mai J-F, Scherer M (2011) Bivariate extreme-value copulas with discrete Pikands dependence measure. Extremes 14:311–324
19. Póczos B, Ghahramani Z, Schneider J (2012) Copula-based kernel dependency measures. In: Proceedings of the 29th international conference on machine learning. Omnipress, New York, pp 775–782
20. Schweizer B, Wolff EF (1981) On nonparametric measures of dependence for random variables. Ann Stat 9:879–885
21. Tjøstheim D, Hufthammer KO (2013) Local Gaussian correlation: a new measure of dependence. J Econom 172:33–48