Test (2011) 20:276–280
DOI 10.1007/s11749-011-0254-2

DISCUSSION

Comments on: Inference in multivariate Archimedean copula models

Weijing Wang · Takeshi Emura

Received: 9 April 2011 / Accepted: 13 May 2011 / Published online: 5 July 2011

1 Multivariate Archimedean copula models

A d-dimensional random vector $(X_1, \dots, X_d)$ belongs to the family of multivariate Archimedean (AC) copula models if its joint distribution H can be written as

$$H(x_1, \dots, x_d) = C_{\psi,d}\big(F_1(x_1), \dots, F_d(x_d)\big),$$

where $F_i(x_i) = \Pr(X_i \le x_i)$, $i = 1, \dots, d$,

$$C_{\psi,d}(u_1, \dots, u_d) = \psi\big\{\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)\big\} \tag{1}$$

for all $u_1, \dots, u_d \in (0, 1)$ and ψ^{-1} is the inverse of ψ which satisfies the conditions that $\psi(0) = 1$, $\psi(x) \to 0$ as $x \to \infty$ and ψ is d-monotone. McNeil and Nešlehová (2009) derived useful properties for multivariate AC models. One important result is the following stochastic representation

$$(X_1, \dots, X_d) =^d R_{\psi,d} \times (S_1, \dots, S_d), \tag{2}$$

where the radial variable $R_{\psi,d} > 0$ is independent of $(S_1, \dots, S_d)$ which is uniformly distributed on the standard simplex $\{(s_1, \dots, s_d) \in [0, 1]^d : s_1 + \cdots + s_d = 1\}$. Since

Communicated by Ricardo Cao.

This comment refers to the invited paper available at doi:10.1007/s11749-011-0250-6.

W. Wang (✉)
Institute of Statistics, National Chiao-Tung University, Hsin-Chu, Taiwan, ROC
e-mail: wjwang@stat.nctu.edu.tw

T. Emura
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC

$(S_1, \ldots, S_d)$ is distribution-free, all the information about ψ is contained in the radial distribution. The correspondence between ψ and $R_{\psi,d}$ is characterized by the so-called Williamson d-transform:

$$\psi(x) = \int_x^\infty \left(1 - \frac{x}{r}\right)^{d-1} dG_{\psi,d}(r) \quad (x \geq 0), \tag{3}$$

where $G_{\psi,d}(r) = \Pr(R_{\psi,d} \leq r)$. The representation in (2) provides a simple way to generate random samples from a multivariate AC model given that the form of ψ or $G_{\psi,d}(\cdot)$ is specified.

For practical applications, the interest usually comes from the other direction. Based on observed data $(X_{j1}, \ldots, X_{jd})$, $j = 1, \ldots, n$, a random sample from $(X_1, \ldots, X_d)$, what is the underlying generator function? The paper by Genest, Nešlehová and Ziegel serves this purpose. Specifically they aim to establish the result that the Kendall distribution $K_{\psi,d}(\cdot)$, which is the distribution function of $W = H(X_1, \ldots, X_d)$, uniquely determines ψ, up to a scaling factor. Although a formal proof for $d \geq 4$ is still not available, strong evidence has been provided for the generality to any dimension. As a result, nonparametric estimation of $K_{\psi,d}(\cdot)$ plays the crucial role of identifying the functional form of $\psi(\cdot)$ or $G_{\psi,d}(\cdot)$ based on (3).

2 Nonparametric estimation of Kendall's process

Modifying the idea of Genest and Rivest (1993) for the bivariate case, Genest, Nešlehová and Ziegel propose to estimate $K_{\psi,d}(\cdot)$ based on pseudo-observations of W, namely $W_1, \ldots, W_n$, where $W_j = \sum_k I(X_{k1} \leq X_{j1}, \ldots, X_{kd} \leq X_{jd})/(n+1)$. The corresponding empirical distribution of $K_{\psi,d}(\cdot)$ is given by

$$K_{n,d}(w) = \frac{1}{n}\sum_{j=1}^{n} I(W_j \leq w). \tag{4}$$

Note that the condition $K_{n,d}(w-) > w$ is essential for deriving the estimator of $G_{\psi,d}(\cdot)$ through the equation $K_{\psi,d}(w) = \Pr(\psi(R_{\psi,d}) \leq w)$.

Consider the following two representations of $K_{\psi,d}(\cdot)$:

$$K_{\psi,d}(w) = \int_0^1 \cdots \int_0^1 I\big(C_{\psi,d}(u_1, \ldots, u_d) \leq w\big) C_{\psi,d}(du_1, \ldots, du_d) \tag{5a}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I\big(H(x_1, \ldots, x_d) \leq w\big) H(dx_1, \ldots, dx_d). \tag{5b}$$

Accordingly $K_{\psi,d}(\cdot)$ can also be estimated by plugging in a nonparametric estimator of $C_{\psi,d}$ in (5a) or a nonparametric estimator of H in (5b). Based on the data $(X_{j1}, \ldots, X_{jd})$ $(j = 1, \ldots, n)$, the empirical copula is given by

$$\tilde{C}_n(u_1, \ldots, u_d) = \frac{1}{n}\sum_{j=1}^{n} 1\big(\hat{U}_{j1} \leq u_1, \ldots, \hat{U}_{jd} \leq u_d\big),$$

where $\hat{U}_{ji} = \sum_{k=1}^{n} 1(X_{ki} \le X_{ji})/(n+1)$ $(j = 1, \ldots, n)$, is the pseudo-observation of $U_i = F_i(X_i)$ for $i = 1, \ldots, d$ and, and the empirical estimator of H is given by

$$\hat{H}(x_1, \ldots, x_d) = \frac{1}{n} \sum_{i=1}^{n} I(X_{i1} \le x_1, \ldots, X_{id} \le x_d).$$

Thus $K_{\psi,d}(\cdot)$ can also be estimated by the following two estimators:

$$\tilde{K}_d(w) = \int_0^1 \cdots \int_0^1 I\big(\tilde{C}_n(u_1, \ldots, u_d) \le w\big)\tilde{C}_n(du_1, \ldots, du_d)$$

and

$$\hat{K}_d(w) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I\big(\hat{H}(x_1, \ldots, x_d) \le w\big)\hat{H}(dx_1, \ldots, dx_d). \tag{6}$$

In comparison with $K_{n,d}(w)$ in (4), $\tilde{K}_d(w)$ and $\hat{K}_d(w)$ involve numerical integrations and hence provide no obvious advantage. However Wang and Wells (2000) utilized a modified version of (6) for the inference of bivariate AC models based on right censored data in which pseudo-observations $W_1, \ldots, W_n$ or U_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, d$ are not available but nonparametric estimators of H exist. In light of Proposition 1 and Algorithm 1 of Genest, Nešlehová and Ziegel, one may still derive the estimator of $G_{\psi,d}(\cdot)$ from $\hat{K}_d(\cdot)$ if a d-dimensional nonparametric estimator of H for right censored data is available. The paper by Dabrowska (1988) described how to extend their bivariate product-limit estimator to higher dimensions. Simpler estimators may be obtained under special censoring structures.

For $d = 2$, the nonparametric estimator of $K_{\psi,d}(\cdot)$ can be used to estimate $\psi(\cdot)$ based their one-to-one relationships (up to a multiplicative scaling factor). It follows that

$$\psi^{-1}(t) = \exp\left\{\int_{t_0}^{t} \frac{dw}{w - K_{\psi,2}(w)}\right\}, \tag{7}$$

where $t_0 \in (0, 1)$ is an arbitrary constant. Due to the lack of explicit formulas for $d > 2$, it is harder to derive the estimator of $\psi(\cdot)$ directly from the nonparametric estimator of $K_{\psi,d}(\cdot)$. On the other hand, it is easier to estimate $\psi(\cdot)$ from $G_{\psi,d}(\cdot)$ using the Williamson d-transform in (3) for any dimension. Hence, the problem is how to retrieve the nonparametric estimator of $G_{\psi,d}(\cdot)$ from $K_{n,d}(\cdot)$. This problem is effectively solved by Proposition 1 and Algorithm 1 of Genest, Nešlehová and Ziegel, in which they utilize the key property $K_{n,d}(w-) > w$. However, generalization of the algorithm for right censored data is not trivial since the condition $\hat{K}_d(w-) > w$ may not hold for finite samples.

3 Procedures of goodness-of-fit

Practitioners often prefer using a parametric model of $\psi_\theta(\cdot)$, where θ is an unknown parameter, rather than estimating it nonparametrically. In this case, it becomes crucial to justify the goodness-of-fit for the imposed $\psi_\theta(\cdot)$ function based on the data at hand. The paper by Genest et al. (2009) reviews the ideas of different tests and provides a power study for the bivariate case ($d = 2$). Conjecture 1 implies that the Kendall's distribution $K_{\psi,d}(\cdot)$ contains all the information for verifying the underlying copula function for any dimension d. That is, $C_{\psi_1,d} \neq C_{\psi_2,d}$ implies $K_{\psi_1} \neq K_{\psi_2}$ for generator functions ψ_1 and ψ_2. Hence, one may use

$$S_n = \int_0^1 \left\{K_{n,d}(w) - K_{\hat{\theta},d}(w)\right\}^2 dw$$

or

$$T_n = \sup_{0<w<1} \left\| K_{n,d}(w) - K_{\hat{\theta},d}(w) \right\|$$

as the goodness-of-fit statistics, where $K_{\theta,d}(\cdot)$ is the model-based Kendall's distribution corresponding to $\psi_\theta(\cdot)$, and $\hat{\theta}$ is an estimator of θ. Note that the presence of $\hat{\theta}$ affects the distributional properties of the test statistics.

Note that Proposition 4.4 of McNeil and Nešlehová (2009) highlights two other ways of testing the goodness-of-fit. The first one utilizes the result that, under the correct form of ψ, $\sum_{j=1}^d \psi^{-1}(U_j)$ is independent of the random vector

$$\left(\frac{\psi^{-1}(U_1)}{\sum_{j=1}^d \psi^{-1}(U_j)}, \dots, \frac{\psi^{-1}(U_d)}{\sum_{j=1}^d \psi^{-1}(U_j)}\right).$$

The other is based on the fact that $(1 - \sum_{j=1}^d \psi^{-1}(U_j))^{d-1}$ follows a standard uniform distribution for all $j = 1, \dots, d$. These two tests, despite of their simplicity, may suffer from the identifiability issue. For example, $(1 - \sum_{j=1}^d \psi^{-1}(U_j))^{d-1}$ may be a standard uniform distribution even under the incorrect form of ψ. The test based on $K_{n,d}(w)$, which is a unique characterization of a copula, do not suffer the identifiability issue under Conjecture 1.

4 Concluding remarks

Genest, Nešlehová and Ziegel develop useful inference methods for the multivariate Archimedean family which have the potential of being applied to general dimensions. Furthermore a clever algorithm, stated in "Algorithm 1", is proposed to implement the rank-based inference procedure. Here we mention possible extension to right censored data. The proposed method, or other extended work, strongly rely on "Conjecture 1", which states that the copula function is uniquely determined by the Kendall's distribution for any dimension. A formal proof is important for justifying the generality of these methods.

References

Dabrowska DM (1988) Kaplan–Meir estimate on the plane. Ann Stat 16:1475–1489

Genest C, Rivest L-P (1993) Statistical inference procedures for bivariate Archimedean copulas. J Am Stat Assoc 88:1034–1042

Genest C, Rémillard B, Beaudoin D (2009) Goodness-of-fit tests for copulas: a review and a power study. Insur Math Econ 44:199–213

McNeil AJ, Nešlehová J (2009) Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. Ann Stat 37:3059–3097

Wang W, Wells MT (2000) Model selection and semiparametric inference for bivariate failure-time data. J Am Stat Assoc 95:62–72