Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Dependence patterns associated with the fundamental diagram: a copula function approach*

Jia LI[1], Yue-ping XU[†‡2]

([1]*Department of Civil and Environmental Engineering, University of Massachusetts Amherst, Amherst, MA 01002, USA*)

([2]*Department of Civil Engineering, Zhejiang University, Hangzhou 310058, China*)

[†]E-mail: yuepingxu@zju.edu.cn

Received Nov. 11, 2008; Revision accepted Aug. 20, 2009; Crosschecked Sept. 29, 2009

Abstract: Randomness plays a major role in the interpretation of many interesting traffic flow phenomena, such as hysteresis, capacity drop and spontaneous breakdown. The analysis of the uncertainty and reliability of traffic systems is directly associated with their random characteristics. Therefore, it is beneficial to understand the distributional properties of traffic variables. This paper focuses on the dependence relation between traffic flow density and traffic speed, which constitute the fundamental diagram (FD). The traditional model of the FD is obtained essentially through curve fitting. We use the copula function as the basic toolkit and provide a novel approach for identifying the distributional patterns associated with the FD. In particular, we construct a rule-of-thumb nonparametric copula function, which in general avoids the mis-specification risk of parametric approaches and is more efficient in practice. By applying our construction to loop detector data on a freeway, we identify the dependence patterns existing in traffic data. We find that similar modes exist among traffic states of low, moderate or high traffic densities. Our findings also suggest that highway traffic speed and traffic flow density as a bivariate distribution is skewed and highly heterogeneous.

Key words: Nonparametric copula, Dependence patterns, Traffic flow, Loop detector
doi:10.1631/jzus.A0800855 **Document code:** A **CLC number:** U49

1 Introduction

Although no consensus has been reached regarding the occurrence of some interesting traffic flow phenomena, such as hysteresis, capacity drop and spontaneous breakdown, it is generally agreed that randomness plays a significant role therein. Much effort has been devoted in the last two decades to either describing or interpreting traffic flows and jams probabilistically. For example, a model of probabilistic breakdown near on-ramps was postulated as an analogy of a nucleation, growth and condensation process (Mahnke and Kaupužs, 2001; Kühne *et al*., 2004). The capacity of a transportation facility is regarded as a random variable above which the traffic flow breaks into stop-and-go or even standing traffic, which enables the analysis of the reliability of a traffic network (Brilon *et al*., 2005). The dynamics of the breakdown probability have been modeled in a phenomenological manner which, coupled with a kinetic model, gives information on traffic breakdown possibility over a specified time-space domain (Hoogendoorn *et al*., 2008). It is also known that the car-following and cellular automata (CA) model can include stochastic ingredients, as randomization of certain parameters is introduced. Nagel *et al*. (2003) discussed these models in detail in the context of reproducing phase transition. As for the macroscopic model, with the assumption that flux function is driven by parameter uncertainty, the new model admits a stochastic process as a solution, which is interpreted as a prediction with imperfect knowledge of reality (Li *et al*., 2008). The inter-relations of flow speed and occupancy reflect the nature of traffic flow

‡ Corresponding author

* Project (No. 50809058) supported by the National Natural Science Foundation of China

and have been widely studied (Cassidy and Bertini, 1999).

The above list is not comprehensive but, clearly, a deep understanding of randomness associated with traffic variables is highly desirable. To this end, we provide a statistical approach based on the application of a copula function. With this approach, the traffic speed and density data from loop detector measurements can be analyzed. In particular, when modeled by bivariate distribution on a speed-density plane, the inter-dependence structure of the two variables can be recovered from data. Our approach can be regarded as a generalization of the traditional way to obtain the fundamental diagram (FD), and thus encompasses a broader perspective of the speed-density relation. We apply this approach using the literal freeway loop detector measurement of GA-400, and obtain new insights into the dataset. As for potential applications of this kind of analysis, we enumerate the validation of microscopic traffic flow models and reliability analysis of general traffic flow models among others.

2 Copula function: basics and estimation

The copula function is a statistical tool modeling the dependence relations among the elements of random vectors. The concept of copula was first introduced in the seminal work of Sklar (1959). In the last decade, its merits have been realized by researchers in the fields of operations research and engineering. In particular, the copula function assumes rich functional forms and thus presumably models the data more flexibly. Many copula functions find their counterparts in multivariate distributions, and thus are analytically appealing and reduce the difficulty of estimation. Moreover, a copula function is not restricted to a certain dimension, so its application to data of high dimension is promising. Finally, a copula provides a possible way to visualize the complex dependence relation of 2D data. The applications of copula include decision making and risk management (Clemen and Reilly, 1999; Junker and May, 2005; Chavez-Demoulin *et al*., 2006) and hydrological frequency analysis (Favre *et al*., 2004). For state-of-the-art reviews, we refer to Nelsen (1999), Trivedi and Zimmer (2007) and Genest and Favre (2007). In the following introduction to copula, we discuss a copula of general dimensions. When it comes to our proposed nonparametric copula, we assume two dimensions to make the statement concrete.

A copula is defined as a multivariate distribution function with uniform 1D marginal distributions: C:$(u_1, u_2, \ldots, u_d)\in[0, 1]^d \mapsto (u_1, u_2, \ldots, u_d)\in[0, 1]$. Firstly, Sklar (1959)'s theorem states that, for each continuous multivariate distribution function, a unique decomposition exists, namely

$$F(x_1, x_2, \ldots, x_d) = C_F(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)), \quad (1)$$

where $F_1(x_1)$, $F_2(x_2)$, …, $F_d(x_d)$ are marginal distribution functions of $F_1(x_1, x_2, \ldots, x_d)$, and $C_F(\cdot)$ is its copula function. A largely equivalent formulation of Eq. (1) is

$$\begin{aligned} f(x_1, x_2, \ldots, x_d) &= \frac{\partial C_F}{(\partial x_1 \ldots \partial x_d)} \cdot f_1(x_1)\ f_2(x_2) \ldots f_d(x_d) \\ &\equiv c_F(F_1, F_2, \ldots, F_d)\, f_1 f_2 \ldots f_d, \end{aligned} \quad (2)$$

where $c_F()$ is actually the density of $C_F()$. Decomposition of Eqs. (1) or (2) has a great advantage in that the marginal information and inter-dependence relation of all marginal distributions are split, which implies that a joint distribution may be constructed by modeling these two aspects independently.

A second property of a copula is that it is invariant under non-decreasing transformation of each margin, i.e., the copula corresponding to random vector $(X_1, X_2, \ldots, X_d)$ is the same as that of $(h_1(X_1), h_2(X_2), \ldots, h_d(X_d))$ when $h_1(\cdot), h_2(\cdot), \ldots, h_d(\cdot)$ are non-decreasing. This implies that, in realistic applications, the underlying copula remains invariant even when the metrics change. These properties make the copula function an appropriate measure of dependence for multivariate data.

For the real application of modeling data with a parametric copula family (e.g., Gaussian copula, t-copula), two successive steps are often involved, i.e., specification and estimation. The specification of the parametric form of a copula is generally a difficult task because it is not directly observable. The result of mis-specification could be disastrous when the subsequent model is sensitive to the specification error,

i.e., the specification error is non-decaying (Devroye, 1982). If the parametric form of a copula function is adopted, many techniques exist for estimating its parameters, although none of them is perfect. Usually, upon the determination of the functional form of a parametric copula, standard maximum likelihood estimation (MLE) procedures apply. Assuming that the parameter set of a particular copula is β, the corresponding log-likelihood function reads:

$$l(\beta)=\sum_{i=1}^{n}\log c_F(F_1(x_i^1),F_2(x_i^2),\ldots,F_d(x_i^d))+\sum_{i=1}^{n}\sum_{j=1}^{d}\log f_j(x_i^j), \tag{3}$$

and the MLE of parameter set β is defined as follows:

$$\beta_{\mathrm{MLE}}=\arg\max l(\beta). \tag{4}$$

The nonparametric copula estimator has attracted attention in recent years (Chen and Huang, 2007). Compared with the parametric family, the nonparametric copula estimator has two advantages and hence appears more promising. First, the step of model specification is avoided, so the risk of mis-specification no longer exists. Second, the nonparametric family is much more flexible than its parametric counterpart, so it is easier to extract the information regarding dependence patterns from the data under consideration. One popular approach for estimating a copula function nonparametrically is to use the methodology of kernel estimation (Silverman, 1986).

In any kernel estimation method, the kernel function and bandwidth are the two indispensable elements. The quality of kernel estimation depends mainly on the bandwidth (Silverman, 1986). Various optimal bandwidth selection methods exist but most of them are computation intensive and involve a tricky human selection step. In practice, for the purpose of data visualization and preliminary exploration, it is desirable to have a nonparametric estimator that is immediately applicable, i.e., tedious computation of optimal bandwidth should be avoided if possible. Rather, a reasonable approximation of optimal estimation with significantly less computational effort is preferred. Following this principle, we provide a very straightforward nonparametric estimator of the copula function. We assume that all the data involved below are bivariate. Nonetheless, most of the following statements could be generalized to a higher dimension with little difficulty.

We construct the nonparametric copula as follows. First, the available data are transformed:

$$x_i=(x_i^1,x_i^2)\mapsto(F_1(x_i^1),F_2(x_i^2)). \tag{5}$$

Owing to the first property of copulas mentioned above, the transformed data corresponds to the same copula function as the original data. Since the marginal distribution function $F_i(\cdot)$ is typically unknown in practice, one may use its empirical estimation, e.g., the empirical distribution function. Assuming that the data in use are already transformed, a kernel estimator of the copula function can be defined as follows:

$$f(x)=n^{-1}\sqrt{\det(\boldsymbol{H})}\sum_{i=1}^{n}k(\boldsymbol{H}^{-1/2}(x-x_i)), \tag{6}$$

here $k(\cdot)$ is the kernel function, conventionally taken to be the standard bivariate normal density function, and $\boldsymbol{H}$ is a 2×2 matrix controlling the smoothing degree. The optimal smoothing matrix is notoriously difficult to select. A rule of thumb from Silverman (1986) is adopted, which is actually the Cholesky factor of the covariance matrix of the data. That is, let $\boldsymbol{H}$ be the matrix $\boldsymbol{L}$ satisfying

$$\boldsymbol{L}\boldsymbol{L}^{\mathrm{t}}=(\mathrm{cov}(X^i,X^j)). \tag{7}$$

Since the copula function is defined only on a bounded region, it is also required that the estimated density function is bounded. Unfortunately, the above estimator does not have such a property automatically. To circumvent the problem, one can apply a scaling factor less than 1 to the original smoothing matrix, or adopt the bounded kernels such as the beta density function. The scaling factor approach is quite straightforward. The value of the scaling factor can be determined by visual inspection. That is, one reduces the value of the scaling factor from 1 gradually, with an appropriate step size, e.g., 0.1. This process is terminated when the estimation obtained has the least support region that covers the area [0,1]×[0,1].

3 Case study: freeway investigation

We used the estimation method presented above to analyze the dependence structures of traffic speed and density data. The data under investigation came from GA-400, which is a toll road to the north of the Atlanta metropolitan area, Georgia, USA. The length of road monitored by NaviGAtor, which is the intelligent transportation system (ITS) used in Georgia, is about 20 km. The site map is shown in Fig. 1. The data were extracted from video cameras that were operated automatically, by software running in the background. A virtual loop detector was placed over each of the lanes and traffic conditions were reported every 20 s. Therefore, the dataset can be regarded as equivalent to loop detector data. Contained in the original dataset were the time-mean speed and flow rate at various stations. From these two variables, the density of traffic flow could be calculated, as the ratio of flow rate to speed. The following analysis concentrates on the speed-density relationship at one fixed station rather than the speed-flow relation, which could be considered more natural from a data acquisition point of view. The speed-density relation has been one of the most popular and intriguing topics in traffic flow theory since the 1960s, when Greenshields' model and numerous subsequent variations were proposed. This is partly because the speed-density relation links directly to kinetic wave models, perhaps the most attested traffic flow models up to now. Our investigation aimed to further understand this fundamental relation. Moreover, according to the Highway Capacity Manual 2000 (HCM2000), the speed-density relation directly indicates the level of service (LoS) of a roadway, therefore, the inter-dependence of density and speed is more significant.

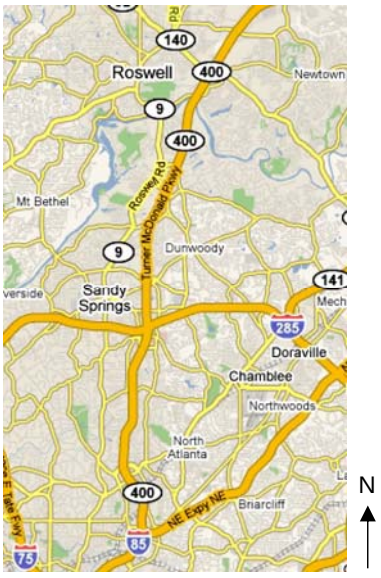

Fig.1 The geometry of the road under investigation (GA-400, Georgia, USA) (From Google Map)

Fig. 2 shows the mean and standard deviation of the traffic speed plotted against the corresponding traffic density. The mean speed decreased as a function of density and fluctuated significantly when the density value was high. The standard deviation varied over the range of density, reaching a peak of 25 km/h at about 30 veh/km. One interpretation of this phenomenon is that when the traffic flow is approaching the critical capacity, the system is very uncertain and driver speed spreads over a large range. These observations indicate that the traffic characteristics, especially the variability features, were quite heterogeneous. In this sense, the FD obtained by plain regression methods could be misleading, since such methods implicitly assume that the fluctuation of noise is well-behaved. Such an FD is not realistic enough to accommodate the real random features shown by data.

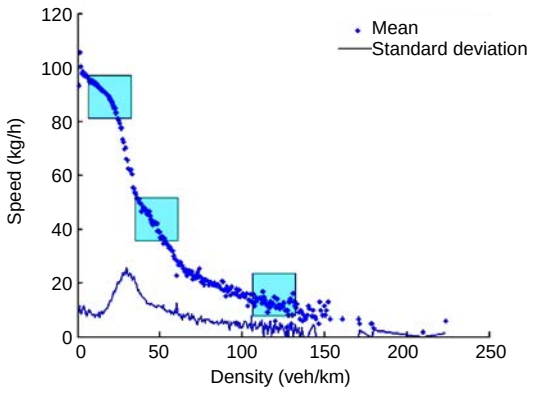

Fig. 2 The mean and standard deviation of speed vs. density (shaded areas indicate the data regions to be analyzed)

We performed our analysis in three representative regions. Regarding traffic dynamics, bivariate data (k, v) in three regions are typical. One is k~0, v~v_f, i.e., the free flow state. The second is k~k_j, v~0, i.e., the congestion state. Another is k_c~$k_j/2$ and the corresponding v_c, the state when the system is approximately critical. The three regions are indicated in

Fig. 2 by shaded squares. For simplicity, the free flow speed v_f and jam density k_j were identified by visual inspection as 100 km/h and 120 veh/km, respectively, for this dataset. In each region, a bandwidth of Δk=10 veh/km was predefined and the bivariate data whose first component fell within $[k-\Delta k/2,\ k+\Delta k/2]$ were selected, where k is one of k_f, k_j or k_c.

The descriptive statistics of the three selected datasets can be found in Table 1. The correlation, Kendall's τ and Spearman's ρ are all dependence measures between the speed and density, lying in the range of [−1, 1]. When a bivariate normal distribution is considered, there exists simple correspondence between the three measures. However, we found no straightforward correspondence among these measures from the calculated values. This implied that the dependence structure was beyond the descriptive ability of the three measures, and that the copula function should be employed to capture the dependence relation of the data. The normality of each dataset was checked. Assuming the dataset follows a bivariate normal distribution, Kendall's τ and Spearman's ρ corresponding to the correlation can be obtained directly (the values in parentheses). Normality was evaluated by comparing these values with those calculated from the data. The free flow state data was not even close to normal, while the critical state and congestion state data could be reasonably regarded to follow the normal distribution.

Table 1 Descriptive statistics of the original speed-density data

	Sample size	Correlation	τ	ρ
Free flow state	2721	−0.096	−0.102 (−0.061)	−0.135 (−0.092)
Critical state	3517	−0.138	−0.083 (−0.088)	−0.115 (−0.132)
Congestion state	317	−0.144	−0.100 (−0.092)	−0.142 (−0.138)

The nonparametric copula estimation procedure described above was applied to this dataset. We present the important findings in Fig. 3. The irregular shapes of the copula functions show that existing parametric families like the Gaussian copula, t-copula and Archimedean copula will fail to capture the dependence pattern. The negative values of the correlation, Kendall's τ and Spearman's ρ of the data confirm this finding. Close scrutiny within the plot reveals that speed was more likely to decrease as density increased. In addition, in all the contour plots, three very similar modes existed. Namely, the locations of the peaks in the contour plots tended to coincide. We interpret this as the existence of special traffic patterns around the equilibrium speed-density relation, possibly owing to different driver behaviors in accelerating, decelerating and constant traffic flow, which are also believed to be the cause of hysteresis. Thus, it seems that the fluctuation around an FD type plot does not distribute evenly. Many factors may influence this, and our study shows that the fluctuation (i.e., scattering) cannot simply be treated as statistical error. The copula approach presented here helps to unveil the underlying patterns, but a full understanding of the scattering phenomenon requires further investigation.

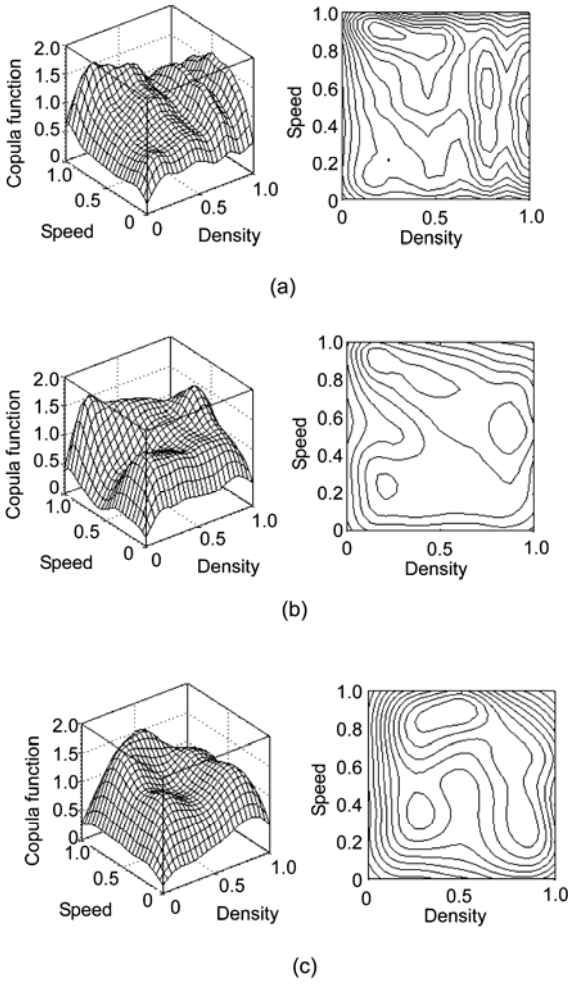

Fig. 3 The estimated copula function and corresponding contour plot for the traffic data. (a) Free flow state; (b) Critical state; (c) Congestion state

If empirical data are available, sampling via the nonparametric copula is straightforward, through simply adding noise to local samples (Hörmann *et al.*, 2004). Table 2 shows a comparison of the dependence measures of the original and simulated samples, where the values before '/' are for original samples and the remainder are for simulated samples. All variables were approximated quite well, with relative errors of no more than 25%.

Table 2 Comparison of dependence measures of original and simulated samples

	Correlation	τ	ρ
Free flow state	−0.096	−0.102	−0.135
	/−0.081	/−0.097	/−0.144
Critical state	−0.138	−0.083	−0.115
	/−0.171	/−0.096	/−0.142
Congestion state	−0.144	−0.100	−0.142
	/−0.179	/−0.115	/−0.174

4 Conclusion

In this paper, we developed a rule-of-thumb statistical analysis approach based on the concept of copula function, and used it to analyze the inter-dependence of time-mean speed and density data. The copula function enables the analysis and visualization of the dependence structure of multivariate data without making any artificial assumptions about the marginal distributions. Compared with parametric families, the nonparametric copula estimator proposed in this paper is flexible, free of the mis-specification risk, and moreover, immediately applicable without the tedious and tricky bandwidth selection issue. These advantages are especially appealing for realistic applications. To illustrate the application, we adopted the virtual loop detector data of speed and density, and compared the corresponding copula plots as well as summary statistics. The results showed that the speed-density data, when modeled by a bivariate copula function, are non-normal, multimodal, and heterogeneous. This is not surprising as traffic dynamics are presumably complex processes and essentially different in various states. Our study provides an approach which smoothes out the noise existing in the observations, independent of the selection of the marginal distribution, and can visualize the dependence patterns of speed and density at a very fine resolution. The copula is a promising tool for revealing features of traffic flow that were previously hidden. Therefore, we hope that this study will promote further investigation of the inter-dependence of traffic states. Traffic flow theory will be significantly advanced if our understanding of the FD is deepened.

References

Brilon, W., Geistefeldt, J., Regler, M., 2005. Reliability of Freeway Traffic Flow: A Stochastic Concept of Capacity. Proceedings of the 16th International Symposium on Transportation and Traffic Theory (ISTTT), Woods Hole, MA, USA. [doi:10.1016/B978-008044680-6/50009-X]

Cassidy, M.J., Bertini, R.L., 1999. Some traffic features at freeway bottlenecks. *Transportation Research Part B: Methodological*, **33**(1):25-42. [doi:10.1016/S0191-2615(98)00023-X]

Cassidy, M.J., Mauch, M., 2001. An observed traffic pattern in long freeway queues. *Transportation Research Part A: Policy and Practice*, **35**(2):143-156. [doi:10.1016/S0965-8564(99)00052-X]

Chavez-Demoulin, V., Embrechts, P., Nešlehová., 2006. Quantitative models for operational risk: extremes, dependence and aggregation. *Journal of Banking & Finance*, **30**(10):2635-2658. [doi:10.1016/j.jbankfin.2005.11.008]

Chen, S.X., Huang, T., 2007. Nonparametric estimation of copula functions for dependence modeling. *Canadian Journal of Statistics*, **35**(2):265-282. [doi:10.1002/cjs.5550350205]

Clemen, R.T., Reilly, T., 1999. Correlations and copulas for decision and risk analysis. *Management Science*, **45**(2):208-224. [doi:10.1287/mnsc.45.2.208]

Devroye, L., 1982. A note on approximations in random variate generation. *Journal of Statistical Computation and Simulation*, **14**(2):149-158. [doi:10.1080/00949658208810536]

Favre, A.C., El-Adlouni, S., Perreault, L., Thiemonge, N., Bobee, B., 2004. Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, **40**(1):W01101. [doi:10.1029/2003WR002456]

Genest, C., Favre, A.C., 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, **12**(4):347-368. [doi:10.1061/(ASCE)1084-0699(2007)12:4(347)]

Hoogendoorn, S.P., van Lint, H., Knoop, V.A., 2008. Stochastic Macroscopic Modeling Framework to Interpret the Fundamental Diagram. Symposium on the Fundamental Diagram: 75 Years, Woods Hole, MA, USA.

Hörmann, W., Leydold, J., Derflinger, J., 2004. Automatic Nonuniform Random Variate Generation. Springer, p.63-68.

Junker, M., May, A., 2005. Measurement of aggregate risk with copulas. *The Econometrics Journal*, **8**(3):428-454. [doi:10.1111/j.1368-423X.2005.00173.x]

Kühne, R., Lubashevsky, I., Mahnke, R., Kaupush, J., 2004. Probabilistic Description of Traffic Breakdowns Caused by On-ramp Flow. arXiv:cond-mat/0405163v1.

Li, J., Chen, Q.Y., Wang, H., Ni, D., 2008. Investigation of LWR Model with Flux Function Driven by Random Free Flow Speed. Symposium on the Fundamental Diagram: 75 Years, Woods Hole, MA, USA.

Mahnke, R., Kaupužs, J., 2001. Probabilistic description of traffic flow. *Networks and Spatial Economics*, **1**(1/2): 103-136. [doi:10.1023/A:1011581111761]

Nagel, K., Wagner, P., Woesler, R., 2003. Still flowing: approaches to traffic flow and traffic jam modeling. *Operations Research*, **51**(5):681-710. [doi:10.1287/opre.51.5.681.16755]

Nelsen, R.B., 1999. An Introduction to Copulas. Springer, p.54-79.

Silverman, B.W., 1986. Kernel Density Estimation for Statistics and Data Analysis. Chapman and Hall, p.100-150.

Sklar, A., 1959. Fonctions de Répartition à n-Dimensions et Leurs Marges. *Publications de l'Institut Statistique de l'Université de Paris*, **8**:229-231.

Trivedi, P.K., Zimmer, D.M., 2007. Copula Modeling: An Introduction for Practitioners. Now Publishers Inc., p.230-240.