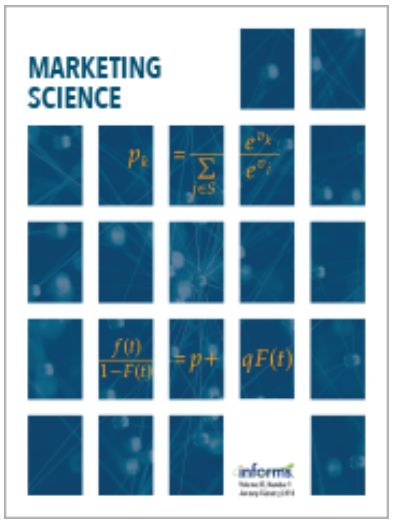

Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Commentary—A Latent Variable Perspective of Copula Modeling

Edward I. George, Shane T. Jensen,

To cite this article:
Edward I. George, Shane T. Jensen, (2011) Commentary—A Latent Variable Perspective of Copula Modeling. Marketing Science 30(1):22-24. https://doi.org/10.1287/mksc.1100.0579

MARKETING SCIENCE
CELEBRATING 30 YEARS
Vol. 30, No. 1, January–February 2011, pp. 22–24
ISSN 0732-2399 | EISSN 1526-548X | 11 | 3001 | 0022

informs

DOI 10.1287/mksc.1100.0579

Commentary

A Latent Variable Perspective of Copula Modeling

Edward I. George, Shane T. Jensen
Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania 19104
{edgeorge@wharton.upenn.edu, stjensen@wharton.upenn.edu}

The likelihood for copula modeling appears when both the data and the copula representations are seen as being driven by common uniform latent variables. This perspective facilitates Bayesian inference for prediction and copula selection.

Key words: Bayesian analysis; latent variable; likelihood
History: Received: March 12, 2010; accepted: April 9, 2010; Eric Bradlow served as the editor-in-chief for this article. Published online in *Articles in Advance* November 4, 2010.

The Essential Ideas

Let us begin by congratulating Danaher and Smith (2011) on an excellent contribution that serves as a lucid introduction to copula modeling, as well as providing a sensible Bayesian approach for its application to both continuous and discrete data. The essential concept we take away is that modeling dependence in multivariate data is facilitated by transforming the marginal data distributions to spaces where dependencies are more naturally represented.

This central idea is most clearly illustrated in the case where each component of the original multivariate data $(X_1, \ldots, X_p)$ is a realization of a continuous random variable. Each continuous X_j with cumulative distribution function (cdf) F_j can be transformed to a desired random variable $X_j^\star$ by first transforming X_j to a uniform random variable, $U_j = F_j(X_j)$, and then transforming this to $X_j^\star = G_j^\star(U_j)$, where $G_j^\star$ is the inverse of the cdf $F_j^\star$ of $X_j^\star$.

The goal is to select marginal distributions for $X_j^\star$ that have a natural dependence structure. For example, if $X_j^\star$ is chosen to have a Gaussian distribution, as Danaher and Smith recommend, then $(X_1^\star, \ldots, X_p^\star)$ are treated as a multivariate Gaussian vector with unknown covariance Γ, which could be estimated with the $X_j^\star$ observations. The resulting dependence structure implicitly imposed on $(U_1, \ldots, U_p)$ and hence on $(X_1, \ldots, X_p)$ is the Gaussian copula.

As Danaher and Smith point out, this idea extends to the general case by treating discrete X_j as corresponding to a latent uniform variable U_j, which takes values according to their Equation (5). Equivalently, any random variable X_j, with cdf F_j, can be considered as the realization of an underlying uniform U_j via $X_j = G_j(U_j)$, where

$$G_j(u) = \inf\{x:\ u \le F_j(x)\} \tag{1}$$

is the suitably defined inverse probability integral transform. From this perspective, the entire copula approach can be summarized by a unified probabilistic framework where a central set of uniforms $\mathbf{U} = (U_1, \ldots, U_p)$ simultaneously generate the original $\mathbf{X} = (X_1, \ldots, X_p)$ and transformed $\mathbf{X}^\star = (X_1^\star, \ldots, X_p^\star)$ variables:

$$\mathbf{X} \xleftarrow{G} \mathbf{U} \xrightarrow{G^*} \mathbf{X}^*. \tag{2}$$

Conditional on the original data $\mathbf{X}$, this framework provides a likelihood for inference about the newly introduced dependence structure. For instance, suppose $F_\theta^\star$ was the multivariate distribution contemplated for $\mathbf{X}^\star$, with parameter θ indexing the unknown dependence structure. Through (2), $F_\theta^\star$ provides a marginal likelihood

$$L(\theta \mid \mathbf{x}) = \int_D p(\mathbf{x} \mid \mathbf{x}^\star) p(\mathbf{x}^\star \mid \theta)\, d\mathbf{x}^\star, \tag{3}$$

which enables inference about θ. Here, integration is only required over D, the range of those $\mathbf{x}^*$ components corresponding to discrete components of $\mathbf{x}$.

This likelihood formulation is especially appealing since suitable F_θ^* can be coupled with a prior $p(\theta)$, and then the integration in (3) can be approximated using Markov chain Monte Carlo sampling from the posterior $p(\theta \mid \mathbf{x}) \propto L(\theta \mid \mathbf{x}) p(\theta)$. Danaher and Smith illustrate this methodology in the case of the Gaussian copula, employing a Gibbs sampler to simulate $\mathbf{x}^\star$ given covariance Γ and a random-walk Metropolis-Hastings algorithm to simulate Γ given $\mathbf{x}^\star$.

For the simulation of Γ, Danaher and Smith propose a clever method based on the decomposition (14) that reduces the problem to simulating an unconstrained upper triangular matrix R. This proposal appears to place a prior on R that is uniform over its upper nondiagonal elements, which implicitly places a prior on Γ. It would be useful if the authors could make $p(\Gamma)$ explicit and perhaps comment on its essential features.

Also, the random-walk Metropolis-Hastings procedure involves a tuning parameter for the proposal distribution, namely, the variance $= 0.01$. Presumably this value was chosen because it worked well in the examples presented. In general, however, we wonder if the authors would recommend diagnostics for the proper adjustment of this tuning parameter. Going further, potential improvements of this sampling procedure might be obtained by optimal or adaptive extensions (Rosenthal 2011) of the Metropolis-Hastings strategy.

As Danaher and Smith (2011) illustrate, the ability to simulate from $p(\Gamma \mid \mathbf{x})$ opens the floodgates for inference. In addition to obtaining the posterior mean of Γ as in (15), the simulated Γ values can also be used to obtain predictive samples of $\mathbf{X}^\star$, $\mathbf{U}$, and $\mathbf{X}$ values via (2). Such $\mathbf{U}$ values can be used for inference about Spearman correlation coefficients as in (16), and such $\mathbf{X}$ values used to infer characteristics like total exposure as in their §4. These predictive samples are also useful for model validation by comparison with the original $\mathbf{x}$ or holdout samples. We are impressed by the gains of the Bayesian Gaussian copula approach over competing alternatives in each presented example.

Going Further?

It is clear that we appreciate the contributions of Danaher and Smith (2011) and heartily endorse their methodology as a practical approach to modeling dependence, but there are a number of issues that merit further investigation. An important issue is the selection of a Gaussian copula to model the dependence structure in $X_1 \ldots, X_p$. Consider the website visit and spend analyses in Figure 1 of Danaher and Smith, where they illustrate the appeal of the Gaussian-copula model. Figure 1(a) shows that the Pearson correlation will be nearly useless as a result of the extreme skewness of the original data. Figure 1(b) shows that the transformation to bivariate uniform has worked beautifully to spread out the data remarkably evenly, and then Figure 1(c) shows that transformation back to bivariate Gaussian shows picture-perfect normality on which Pearson's correlation is certainly meaningful.

The remarkably uniform appearance of the marginals in Figure 1(b) suggests that the estimated transformations have worked perfectly. However, we worry that this evaluation is misleading because $\hat{F}_j(X_j)$ will tend to be more uniform than $F_j(X_j)$, a consequence of overfitting as a result of the use of "plug-in" maximum likelihood estimates $\hat{\theta}$ of the F_j parameters. The subsequent effect would be that the $X^\star$ data would agree even more with the copula. This overfitting could be remedied by introducing priors for θ and taking a fully Bayesian approach to their estimation. By focusing on point estimates of θ, Danaher and Smith are ignoring uncertainty in their marginal distributions $\hat{F}_j(X_j)$ and are subsequently ignoring uncertainty in the copula because the copula parameters are estimated conditional on fixed $\hat{F}_j(X_j)$.

Interestingly, Figure 1(d), featuring the transformation back to bivariate t, looks quite reasonable as well, and it is not clear from the plots whether the Gaussian copula is preferable to the t-copula. Figure 1(e) demonstrates the importance of the dependence revealed in Figures 1(c) and 1(d). We certainly agree with Danaher and Smith that the much larger Pearson's correlation for the transformed data is more likely to alert the analyst to dependence compared with the Pearson's correlation for the original data. Of course, any of the three correlation measures reported in their Table 3 would suffice for this purpose, and for screening purposes over many data sets, it may be most sensible to use the more robust Spearman's rank correlation.

We also agree that the Gaussian copula "does the job" in the examples presented in their paper and is clearly superior to simpler alternatives in terms of exploiting dependence. However, other copula-based models might provide an even better model for these applications. A promising direction for future research would be the development of alternative parametric copulas, such as mixtures of normals, which may have a suitable Bayesian implementation. Bayesian factors could then provide a natural justification for the model selection of one copula formulation over the others.

We end with a cautionary tale about Gaussian copulas in financial applications. Once seen as a "magic bullet" for evaluating portfolio risk in the financial industry, the Gaussian copula model is now seen as having failed miserably because of its inability to account for extreme correlations (Salmon 2009). One possible reason for this failure is that the level of correlation between financial asset values increases as their risk approaches extreme levels. This characteristic is directly at odds with the key Gaussian model assumption of independence between the mean levels and the variance-covariance structure. This independence is a blessing in many situations, but it is a curse in this case because the correlation between financial items cannot be tied to their level of risk.

Ultimately, Gaussian copula models may not be flexible enough for a host of applications in finance. Creating copula models that can more properly address extreme correlations between financial instruments is another potentially fruitful area for future research.

Acknowledgment

The authors thank the editor, Eric Bradlow, for giving them the opportunity to comment on this promising contribution to statistical methodology.

References

Danaher, P., M. S. Smith. 2011. Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Sci.* **30**(1) 4–21.

Rosenthal, J. S. 2011. Optimal proposal distributions and adaptive MCMC. S. Brooks, A. Gelman, G. Jones, X.-L. Meng, eds. *Handbook of Markov Chain Monte Carlo: Methods and Applications*. Chapman & Hall/CRC Press, Boca Raton, FL. Forthcoming. http://probability.ca/jeff/ftpdir/galinart.ps.

Salmon, F. 2009. Recipe for disaster: The formula that killed Wall Street. *Wired Magazine* (February 23), http://www.wired.com/techbiz/it/magazine/17-03/wp_quant.