Non-Gaussian Multivariate Statistical Models and their Applications

Narayanaswamy Balakrishnan (McMaster University),
Christopher Field (Dalhousie University),
Marc G. Genton (King Abdullah University of Science and Technology),
Harry Joe (University of British Columbia)

May 19–24, 2013

1 Overview of the Field

In recent years, the extension of classical Gaussian-based statistical models to non-Gaussian ones has received sustained attention and generated many research topics. There have been various approaches to develop multivariate non-Gaussian distributions. Two main directions to approach such extensions consist of multivariate skew-symmetric distributions and of copula models.

Skew-symmetric distributions are extensions of the multivariate Gaussian distribution to model both skewness and heavy tails; see the book edited by Genton (2004) and the review article by Azzalini (2005) for more details. They include the multivariate normal distribution as a particular case and have many appealing properties. Moreover they are flexible for modeling of data from real applications and can be used in practical settings. Indeed, those skew-symmetric multivariate models have been applied to problems such as: non-Gaussian Kalman filters to analyze climatic time series data; hierarchical Bayesian spatial models for rainfall data; mixed models with non-Gaussian random effects to study biomedical data; clinical trials where one of several doses or treatments is selected in a phase II study to be examined further in a phase III study and naturally give rise to skew-symmetric models; non-Gaussian asset pricing and portfolio selection in finance; non-Gaussian GARCH models in economics; non-Gaussian shape modeling in image analysis; non-Gaussian modeling of coastal flooding in oceanography; non-Gaussian distance determination in astronomy; non-Gaussian space time regression modeling of tree growth in forestry; sample selection bias in Heckman-type selection models in economics; perturbation of numerical confidential non-Gaussian data for database security in management science; non-Gaussian modeling of the distribution of pollutants in environmetrics; models for non-Gaussian distributions of wind speed and direction in wind power forecasting for renewable energy; and many more. In the past six years, many new results have been obtained and their applications have been expanding.

Somewhat in parallel, copula models have been developed to describe the joint dependence structure of multivariate distributions separately from their marginal distributions; see Joe (1997), McNeil et al. (2005), Nelsen (2006), Jaworski et al. (2010), Kurowicka and Cooke (2006), Kurowicka and Joe (2011) for overviews. Also there are other books aimed at readers in quantitative finance and insurance that have chapters on copulas. In applications in finance and insurance, models with more dependence in the joint tails than multivariate non-Gaussian are important. This explains why copula models which can allow for tail dependence have become especially important in finance and insurance. Modeling multivariate data with copulas has found many

applications in recent years, especially in finance and hydrology, but also in some of the applications mentioned above for skew-symmetric models; see Genest and Favre (2007) for a review and references therein. An important class of applications of copulas is to model risks, for example recently in spatial extremes related to climate. There are many parametric copula families available, which usually have parameters that control the strength of dependence. Some of the current research deals with the construction of new copula models, for example vine copulas, with testing and goodness-of-fit procedures, with nonparametric and semiparametric approaches for copula estimation, and with developing copulas for spatial statistics. The field of copulas is very active and developing rapidly.

This workshop intended for the first time to bring together those two communities, namely skew-symmetric multivariate distributions and copula models, discuss recent research results in each of these two groups of researchers, and study their connections in order to further advance the field of multivariate non-Gaussian statistical modeling and their applications. This workshop also intended to study extensions of classical Gaussian-based multivariate statistical models to non-Gaussian ones that are more realistic to solve important problems arising from many applications, including areas such as economics, finance, management science, oceanography, climatology, environmetrics, forestry, engineering, renewable energies, image processing, astronomy, clinical trials, and biomedical science. Moreover, several of the aforementioned applications were related to processes affecting Planet Earth. Therefore, they were relevant to the Mathematics of Planet Earth 2013 (MPE 13) initiative. The workshop contributed also with a post to the Planet Earth 2013 blog (http://mpe2013.org/2013/05/24/birs-workshop-non-gaussian-multivariate-statistical-models-and-their-applications/):

"A diverse group of 42 scholars from 15 countries converged this week at the Banff International Research station (BIRS) for a workshop on Non-Gaussian Multivariate Statistical Models and their Applications. The workshop consisted of a variety of talks and presentations on the theory and applications of copulas and skew-elliptical distributions when used as multivariate models. One of the aims was to generate intellectual discussion of the use of these statistical models for analyzing data arising from several disciplines. Applications varied from disciplines including but not limited to climate change, finance, insurance, and medicine. For example, a multivariate framework was constructed to understand the uncertainties resulting from expert opinions of future sea-level rise from ice sheets in East and West Antartica and Greenland. Multivariate spatial models were also used to analyze the brain activities of individuals with certain neurological disorders such as Down Syndrome. This first-of-a-kind workshop is expected to provide avenues for further advancements of research in this exciting topic."

2 Recent Developments and Open Problems

Since 2006, the biggest advances in copula models for high-dimensional data analysis include the vine pair-copula constructions (Aas, Czado, Brechmann, Cooke, Kurowicka, Joe) and factor copula models (Joe, Krupskii, Nikoloulopoulos); the latter are truncated vines rooted at some latent variables. These copula models have been implemented in software; regular vines (R package VineCopula) have been used for up to 50 dimensions and structured factor copula models (software to be released in 2014) have been tested in over 100 dimensions.

Because the vine pair-copula construction is based on a sequence of bivariate copulas, perhaps nonparametric approaches can be used in estimating the bivariate copulas. Several such recent approaches were mentioned in the presentation about vines. It remains to be seen with future research on whether any of the non-parametric approaches can adequately be used for tail inference. The quality of tail inference from the vine pair-copula construction including factor copula models is an ongoing research topic.

Here are a couple of open problems that came up:

1. In skew-symmetric modeling, there seemed to remain some issues about the parameterization. In particular should we strive to have orthogonality at least between classes of parameters. There is also the question of identifiability with certain choices of parameters. There seem to be difficulties with degrees of freedom and scale estimates.

2. There is a need for substantive settings in which either the skew symmetric or copula approaches give solutions and insights which are different from those obtained by more standard methods.

3. There seems to be some question about how to go from an empirical copula to a more smoothed version. One choice was to jitter the data but there was little enthusiasm for this. One objection was that smoothing a Gaussian copula can give you a non-Gaussian copula. It is not clear that this is a critical objection. The approach of working in neighborhoods of models is well established in the robustness literature and some of these concepts might well carry over.

4. There seems to be a need for more and better ways of assessing the goodness of fit of an estimated copula to the data whether the analysis is done via Bayesian or frequentist methods.

3 Presentation Highlights

The overview lectures on the first day were very helpful and important as there were researchers working on two different areas of research who were not always aware of research in the other area. In addition there were graduate students for which was their first introduction to the subject. All the presentations were of very high quality and the schedule was set up to allow ample time for discussion which was often quite lively.

For the skew-symmetric side of the program, Azzalini gave an overview of skew-symmetric and related models stressing skewness and tail behavior that deviate from the classical multivariate Gaussian model; he also included plots of several different bivariate skew-normal and skew-t distributions, and later in the week presented an R package called `sn` to fit such models.

For skew-symmetric research, some applications were the following: (a) Giorgi proposed moment-free measures for the multivariate skew-t distribution, (b) Jimenez-Gamero described testing procedures for skew-symmetric models, (c) Pewsey reviewed various inferential issues for skew-symmetric distributions, (d) Dey introduced state space models for binary response data with flexible skewed link functions, (e) Liseo discussed Bayesian inference for the multivariate skew-normal distribution, and in particular an approach based on population Monte-Carlo computations, (f) Loschi proposed nonparametric mixtures based on skew-normal distributions with application to density estimation, (g) Adcock discussed some challenges in portfolio theory and asset pricing for non-Gaussian distributions, in particular he motivated the need for a certain skew-t distribution and applications of Stein's lemma in this context, (h) Jones discussed alternative non-standard skew-symmetric distributions, (i) Branco presented a poster on Bayesian sensitivity analysis under a skew-normal class of prior distributions using concentration functions, (j) Kim introduced mixtures of skewed Kalman filters in a poster based on the closed skew-normal distribution and its extension to a scale mixture class of closed skew-normal distributions. The resulting family of distributions is skewed and has heavy tails too, so it is appropriate for robust analysis. Hence it is possible to handle skewed and heavy tailed data simultaneously, (k) Nathoo developed a skew-t space-varying regression model for the spectral analysis of resting state brain activity in a poster, (l) Scarpa fitted age-specific fertility rates by a flexible generalized skew-normal probability density function and studied in a poster the resulting improvements.

For the copula side of the program, Joe gave an overview of copula models stressing asymmetries and tail behavior that deviate from the classical multivariate Gaussian model; he also included plots of several different bivariate skew-normal distributions when used as copulas. Aas and Czado gave an overview of the pair-copula constructions and their applications. Brechmann gave a presentation on the `VineCopula` R package, which he has co-authored with Schepsmeier and Stöber. Nikoloulopoulos gave a presentation of factor copula models for item response data; the version of factor copulas corresponded to truncated canonical vines rooted at latent variables. Krupskii had a poster presentation of factor copula models for continuous data with an analogous construction.

For copula research not involving vines, some applications were the following: (a) Yoshiba used the skew-t copula for an application in risk aggregation (the t copula is commonly used in statistical finance applications if there is not much tail asymmetry), (b) Valdez applied multivariate negative binomial regression models (some of which were based on copulas) to insurance claim count data, (c) Cooke presented tail dependence elicitation in an application that is relevant to climate change, (d) Naveau used multivariate extreme value theory in the analysis of heavy rainfall over many locations in a spatial network. More theoretical presentations included the following: (e) Hua presented another approach for strength on bivariate tail dependence based on conditional tail expectations, (f) Genton presented tests for different structures for bivariate copulas, (g) Kojadinovic, Remillard and Volgushev in separate talks presented research involving

the empirial copula processes, (h) Neslehova used the empirical checkerboard copula to create tests of independence for variables of mixed types, (i) Kurowicka presented an approach to simulate uniformly from the set of correlation matrices with specified correlations satisfying chordal sparsity patterns — this potentially has applications to simulation from truncated vines, since these correspond to some high-order partial correlations set to 0. Also, Kojadinovic gave a presentation of the `copula` R package, which he has co-authored with Hofert, Maechler and Yan; this package is quite distinct from the `VineCopula` R package in terms of available copula models.

After the talks on day one of the workshop, it was clear that there are differences between the opportunities and challenges offered by different classes of non-normal distributions. The most notable of these is that copulas offer flexibility in the computation of marginal distributions, whereas skew-elliptical distributions have a very simple genesis. The hour-long session which took place on the second day of the workshop provided a forum for these and other differences between copulas and skew-elliptical distributions to be discussed in more detail.

There was a lengthy and lively period in which the theoretical basis of the dependence structure and its natural occurrence were covered in detail. Perhaps unsurprisingly, this part of the session highlighted a number of aspects of these models which are not specifically concerned with the theoretical basis of the dependence structure. The issues raised covered aspects shared by the two approaches and differences between them. Users of skew-elliptical distributions are often drawn to them by the way in which their parameterisation arises naturally. Copulas, by contrast are able to reflect dependence structures which may be severely non-normal or even non-elliptical. The failure of models based directly or indirectly on the normal distribution were cited as examples of the need for flexibility in the dependence structures.

Given the earlier talks, there was discussion of the use of Vines for high dimensional problems. The ability of these relatively recent developments to deal with many variables, hundreds or perhaps thousands, highlights problems shared by all approaches; notably the need for large data sets and problems of over-parameterization. The discussion continued with problems of estimation and inference, which are common to all multivariate models. Two other common themes are the use of latent variables and Bayesian methods for estimation. By contrast, researchers working in copula theory have and are developing non-parametric methods.

4 Scientific Progress Made

There will certainly be some new collaborations as a result of the meeting. Researchers have discussed visits between their institutions in order to develop such collaborations. For example, Azzalini and Genton have started a project on a comparison of two approaches for constructing multivariate skew-normal distributions, which has been continued the week after the workshop at the Canadian Statistical Society meeting in Edmonton.

At the Friday morning discussion, a proof of half-plane support, for the Azzalini and Capitanio multivariate skew-t distribution at a boundary of the parameter space, was given based on the form of the density in Joe (2006). The notation used here is a small modification of that in Joe (2006). For this distribution, $(Y_1, \ldots, Y_d)$ is distributed as $(X_1, \ldots, X_d) \mid X_0 > 0$ where $(X_0, X_1, \ldots, X_d)$ is jointly multivariate t with ν degrees of freedom and correlation matrix $R = [1, \boldsymbol{\delta}'; \boldsymbol{\delta}, \Omega]$. Let the univariate t density with ν degrees of freedom be denoted as t_ν and let its cumulative distribution function be denoted as T_ν. Also let $t_{d,\nu}(\cdot; \Sigma)$ denote the d-variate t density with ν degrees of freedom and covariance matrix Σ. The multivariate skewed t density of Azzalini and Capitanio (2003) is:

$$2t_{d,\nu}(\mathbf{y}; \Omega)\, T_{d+\nu}\Big(\boldsymbol{\alpha}'\mathbf{y}\Big[\frac{\nu + d}{\nu + \mathbf{y}'\Omega^{-1}\mathbf{y}}\Big]^{1/2}\Big)$$

where $\boldsymbol{\alpha}' = (1 - \boldsymbol{\delta}'\Omega^{-1}\boldsymbol{\delta})^{-1/2}\boldsymbol{\delta}'\Omega^{-1}$. Its univariate margins are

$$2t_\nu(y_j)\, T_{1+\nu}\Big(\zeta_j y_j\Big[\frac{\nu + 1}{\nu + y_j^2}\Big]^{1/2}\Big),$$

where $\zeta_j = (1 - \delta_j^2)^{-1/2}\delta_j$ can be interpreted as the skewness parameter and $\delta_j = \zeta_j/(1 + \zeta_j^2)^{1/2}$. The

bivariate (1,2) marginal density, with $\rho = \rho_{12}$, has the form

$$2t_{2,\nu}(y_1, y_2; \rho)\, T_{2+\nu}\Big(\frac{(\beta_1 y_1 + \beta_2 y_2)}{\sigma}\Big[\frac{\nu+2}{\nu + (y_1^2 + y_2^2 - 2\rho y_1 y_2)/(1-\rho^2)}\Big]^{1/2}\Big),$$

where

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = (1-\rho^2)^{-1}\begin{pmatrix}\delta_1 - \rho\delta_2\\ \delta_2 - \rho\delta_1\end{pmatrix},$$

and

$$\sigma^2 = \Big[1 - \frac{\delta_1^2 + \delta_2^2 - 2\rho\delta_1\delta_2}{(1-\rho^2)}\Big].$$

Note the constraint

$$\delta_1^2 + \delta_2^2 - 2\rho\delta_1\delta_2 \le 1 - \rho^2$$

from the positive definiteness condition, so the amount of univariate skewness is not independent of the dependence. If $d = 2$ and $\delta_1 = \delta_2$, then the two univariate margins are the same.

From contour plots of the bivariate density, one can see that the density has support on a half-plane when the matrix $R = \begin{pmatrix}1 & \delta_1 & \delta_2\\ \delta_1 & 1 & \rho\\ \delta_2 & \rho & 1\end{pmatrix}$ is singular non-negative definite. The proof of this property is as follows. Then $\sigma^2 \to 0^+$ (this is the conditional variance for $\nu \to \infty$ and a multiplier in the conditional scale for finite ν) when the boundary of the space of positive definite correlation matrices is reached. Hence the argument of $T_{2+\nu}$ in (3) goes to $+\infty$ (respectively, $-\infty$) is (y_1, y_2) is such that $\beta_1 y_1 + \beta_2 y_2 > 0$ (respectively, $\beta_1 y_1 + \beta_2 y_2 < 0$). Hence the density (3) is 0 for the half-plane $\{(y_1, y_2) : \beta_1 y_1 + \beta_2 y_2 < 0\}$. This behavior also occurs for dimension $d > 2$ and is similar to the half-normal limit of the univariate skew-normal distribution with skew parameter going to $\pm\infty$.

During the discussion session on Tuesday afternoon, it was mentioned that there were problems with maximum likelihood inference for the Azzalini-DallaValle multivariate skew-normal distribution. This may be partly caused by the lack of algebraic independence of univariate skewness and dependence parameters (that is, the range of dependence is not independent of the range of univariate skewness).

The Joe-Seshadri-Arnold construction does have algebraic independence of univariate skewness and dependence parameters, but this seems to be at the cost of density contour plots with bulges to all corners and hence not necessarily a good match to typical multivariate data.

With Yoshiba's presentation and other discussion during the workshop, it would be desirable to have a multivariate skew-t distribution and copula that has bivariate tail skewness, tail dependence and some linear properties of the elliptical distribution.

Hence there is an open problem of developing alternative skew-elliptical distributions which have some of the above properties but avoid the boundary problems of the Azzalini-Capitanio multivariate skew-t. An approach anologous to the generalized hyperbolic skew t construction of Demarta and McNeil (2005) is one possibility. Because of the advance in computer speed, some models which would have been computationally too difficult to use in the 1990s could be considered now.

5 Outcome of the Meeting

For the students and younger researchers, they will come away with a nice set of tools for statistical analysis and a good sense of the open questions in both areas.

Some further research problems that bridged the copula and skew-symmetric sides of the workshop were identified.

Overall, this workshop suggested a number of conclusions that can be drawn. The nature of an application may lead statisticians naturally to the choice of multivariate model. There are common methodological issues. Skew-elliptical distributions may be used as copulas, but equally skew-elliptical distributions can be constructed which have different margins.

References

[1] Azzalini, A. (2005), The skew-normal distribution and related multivariate families (with discussion by Marc G. Genton and a rejoinder by the author). *Scandinavian Journal of Statistics*, 32, 159 200.

[2] Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society* B, 65, 367–389.

[3] Demarta, S. and McNeil, A.J. (2005). The t copula and related copulas, *International Statistical Review*, 73, 111–129.

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

[5] Genton, M. G. (2004), Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality. Edited Volume, Chapman & Hall/CRC, Boca Raton, 416 pp.

[6] Jaworski, P., Durante, F., Haerdle, W., and Rychlik, T. (2010), Copula Theory and Its Applications. Lecture Notes in Statistics, Springer

[7] Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

[8] Joe, H. (2006). Discussion of "Copulas: tales and facts", by T Mikosch. *Extremes*, 9, 37–41.

[9] Joe, H., Seshadri, V. and Arnold, B.C. (2012). Multivariate inverse Gaussian and skew-normal densities. *Statistics and Probability Letters*, 82, 2244–2251.

[10] Kurowicka, D. and Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley, Chichester.

[11] Kurowicka, D. and Joe, H. (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific, Singapore.

[12] McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management*. Princeton University Press, Princeton, NJ.

[13] Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer Lecture Notes in Statistics. Springer, New York.

[14] Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York, second edition.