Mathematical Challenge April 2017

Conditional Dependence Modeling

References

- E. F. Acar, R. V. Craiu, and F. Yao. *Dependence calibration in conditional copulas: a nonparametric approach.* Biometrics, 67 (2): 445–453, June 2011.
- E. F. Acar, R. V. Craiu, and F. Yao. *Statistical testing of covariate effects in conditional copula models.* Electronic Journal of Statistics, 7: 2822–2850, 2013.
- R. V. Craiu and A. Sabeti. *In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes.* Journal of Multivariate Analysis, 110: 106–120, 2012.
- P. Embrechts, A. McNeil, and D. Straumann. *Correlation and dependence in risk management: properties and pitfalls.* Risk management: value at risk and beyond, pages 176–223, 2002.
- I. Gijbels, N. Veraverbeke, and M. Omelka. *Conditional copulas, association measures and their applications.* Computational Statistics and Data Analysis, 55 (5): 1919–1932, May 2011.
- H. Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall/CRC, 1997.
- R. B. Nelsen. *An Introduction to Copulas*. Springer, 1999.
- A. J. Patton. *Applications of Copula Theory in Financial Econometrics*. PhD thesis, University of California, San Diego, 2002.
- A. Sabeti, M. Wei, and R. V. Craiu. *Additive models for conditional copulas.* Stat, 3 (1): 300–312, 2014.
- A. Sklar. *Fonctions de répartition à n dimensions et leurs marges.* Publications de L'Institut de Statistique de L'Université de Paris, 8: 229–231, 1959.
- T. Vatter and V. Chavez-Demoulin. *Generalized additive models for conditional dependence structures.* Journal of Multivariate Analysis, 141: 147–167, 2015.

Description

Consider the following research questions:

- Cellular biologists study how predictor genes coordinate the expression of target genes. Can we further quantify how the association between the targets depends on the predictors?
- There are relationships between a population's life expectancy and the country's GDP, as well as between male and female life expectancy in a given country. Can we measure the effect of the GDP on the later while controlling for the former?
- Volatilities of intraday asset returns show periodicities due to the cyclical nature of market activity and macroeconomic news releases. Is this also true for their dependence structure?

To obtain a statistically sound answer, we need a framework to model the dependence structure between random variables *conditionally* on the realization of exogenous predictors (covariates).

As an introducing example, consider the conditional correlation between the random variables Y_1 and Y_2 given X, which is defined by

$$corr(Y_1, Y_2|X) = \frac{E[\{Y_1 - E(Y_1|X)\}\{Y_2 - E(Y_2|X)\} \mid X]}{E\left[\{Y_1 - E(Y_1|X)\}^2 \mid X\right]^{1/2} \times E\left[\{Y_2 - E(Y_2|X)\}^2 \mid X\right]^{1/2}}.$$

For instance, when $Y_1 = A + B\,X$ and $Y_2 = C + D\,X$ where A, B, C, D are independently distributed random variables with equal variance and correlation ρ, we have that

$$corr(Y_1, Y_2|X) = \frac{\rho(1+X)^2}{1+2\rho X+X^2}.$$

In this example, it is clear that the conditional correlation depends on the value of the conditioning variable explicitly. Hence, it is in general not equal to the so-called partial correlation. However, while the statistical literature abounds with models allowing the description of covariates effects on the behavior of a single random variable, letting the dependence structure itself be a function of exogenous predictors has only been recently explored.

Similarly as in the unconditional case, Pearson's conditional correlation detects linear (conditional) relationships between variables, but it has three drawbacks: it lacks robustness to outliers, it does not always exists (see Embrechts et al. 2002), and it depends on the marginal (conditional) distribution of each random variable. While the first issue can be alleviated using robust methods, the second and third are arguably more fundamental. In essence, the last two drawbacks teach us that the moments of a distribution are not always appropriate to describe the joint conditional behavior of the underlying random variables.

In this context, it is useful to distinguish between two closely related concepts, namely dependence and concordance. Loosely speaking, a dependence measure relates to any functional characteristic expressing how close the joint distribution is to the product of the margins. As for the concordance, it measures the degree of agreement between positive and negative comovements. Borrowing from Nelsen (1999), two desirable properties of a dependence (respectively concordance) measure are:

- invariance to monotone increasing transformations of the margins (respectively up to a sign change if one of the transformations is monotone decreasing);
- the existence and uniqueness of a minimum (respectively a zero), which is attained whenever the variables are independent.

For instance, rank correlation coefficients satisfy the properties above for a concordance measure (or a dependence measure for their absolute value). For two random variables X_1 and X_2 two of such measures are:

- Spearman's rho, which is simply the Pearson's correlation between the probability integral transforms $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$, where F_1 and F_2 are the marginal distributions,
- and Kendall's tau is the difference between probability of concordance and discordance, namely $\tau = P\{(X_1 - \widetilde{X_1})(X_2 - \widetilde{X_2}) > 0\} - P\{(X_1 - \widetilde{X_1})(X_2 - \widetilde{X_2}) < 0\}$, where $\widetilde{X_1}$ and $\widetilde{X_2}$ are independent copies following the same distribution than (X_1, X_2).

Note that copulas and dependence measures are closely related (see Joe 1997; Nelsen 1999 for textbook treatments). This is because the first of the two desirable properties of a dependence measure, namely invariance to monotone increasing transformations of the margins, essentially states that such a measure should depend on the copula only. For instance, if C is the copula of the two random variables X_1 and X_2, then their Kendall's tau is $4\iint_0^1 C(u_1, u_2)dC(u_1, u_2) - 1$, and their Spearman's rho can be written as $12\iint_0^1 u_1 u_2 dC(u_1, u_2) - 3$. Furthermore, in many cases, there is a simple mapping between the copula parameter and the dependence measure. As an example, letting θ be the natural copula parameter, if C is the Clayton copula, then we have that $\tau = \theta/(\theta + 2)$ similarly, for the Gumbel copula, the $\tau = 1 - 1/\theta$. As for the Gaussian copula (and all elliptical copulas such as the Student's *t*) with correlation parameter θ, then we have that $\tau = 2\sin^{-1}(\theta)/\pi$.

While copulas have been studied for more than fifty years, their formal extensions to conditional distributions have younger origins. To the best of our knowledge, it is Patton (2002) who first extended the standard Sklar's theorem (see Sklar 1959) by imposing a mutual conditioning algebra for each margin and the copula. Using the concept of conditional copula, namely a conditional distribution which is supported in the unit hypercube and has uniform conditional margins, Patton (2002)'s main theoretical contribution lies in the following:

Theorem:

Let $Y = (Y_1, \cdots, Y_d)$ and $X = (X_1, \cdots, X_q)$ be two random vectors such that the conditional margin distribution of $Y_i \mid X$ is $F_{Y_i \mid X}$. For all $x \in R^q$, F is a joint conditional distribution with conditional margins $F_{Y_i \mid X}$ for $i \in \{1, \cdots, d\}$ if and only if there exists a conditional copula C such that

$$F(y_1, \cdots, y_d \mid x) = C\{F_{Y_1 \mid X}(y_1 \mid x), \cdots, F_{Y_d \mid X}(y_d \mid x) \big| x\} \; (1)$$

for all $y \in R^d$. Moreover, if the conditional margins and F are continuous, then C is unique.

Similarly as in the unconditional case, C is the joint distribution of $U = (U_1, \cdots, U_d)$, where U is a random vector such that $U_i = F_{Y_i \mid X}(Y_i \mid X)$., namely the (conditional) probability integral transforms. Taking partial derivatives with respect to the conditioned vector on both sides of (1) (and provided that they exist), we also have that

$$f(y_1, \cdots, y_d \mid x) = c\{F_{Y_1 \mid X}(y_1 \mid x), \cdots, F_{Y_d \mid X}(y_d \mid x) \big| x\} \prod_{i=1}^{d} f_{Y_i \mid X}(y_i \mid x),$$

where $f_{Y_i \mid X}$ for $i \in \{1, \cdots, d\}$ are the conditional margins and c is the conditional copula density. Because the right-hand side is a product, the joint conditional log-likelihood can be written as a sum between the log-likelihood of each conditional margin and the log-likelihood of the conditional copula. This fact can be conveniently exploited in a two-step procedure: estimate each of the conditional margins separately first, and then take the probability integral transform of the data using those margins to estimate the conditional copula. Furthermore, as in the unconditional case, the first of the two desirable properties of a conditional dependence measure states that such a measure should depend on the conditional copula only. Hence, modeling conditional dependence and modeling conditional copulas are two closely related problems.

When modeling conditional copulas and dependence measures, as often in statistics, it is useful to distinguish between distribution-free and parametric methods. In Gijbels et al. (2011), the authors suggest two kernel-based estimators of conditional copulas and corresponding conditional association measures. While useful as descriptive statistics, their estimators are limited to a single predictor. On the parametric side, Acar et al. (2011) consider a copula parameter that varies with a single covariate. The authors estimate their model using the concept of local likelihood, and they further suggest a testing framework in Acar et al. (2013). In Craiu and Sabeti (2012), the authors develop Bayesian inference tools for a bivariate copula, conditional on a single covariate, coupling mixed or continuous outcomes. It is extended to multiple covariates in the continuous case by Sabeti et al. (2014). Alternatively, Vatter and Chavez-Demoulin (2015) tackle the issue of conditional dependence modeling using generalized additive models, a natural extension of linear and generalized linear models. Their framework is completely flexible, meaning that the dependence structure varies with an arbitrary set of covariates in a parametric, nonparametric or semiparametric way. Furthermore, it is fast and numerically stable, which means that it is suitable for exploratory data analysis and stepwise model building.

Questions:

- What are the pros and cons of the four approaches (non-parametric, local likelihood, Bayesian, generalized additive models)?
- Discuss how those approaches can be used to test for covariates effects in the dependence structure between random variables.
- The four approaches are limited to modeling the distribution of two random variables conditional on a set of predictors. Can you suggest extensions to the *d*-dimensional case?

We look forward to your opinions and insights.

Best Quant Regards,

swissQuant Group Leadership Team