Quantile regression for heavy tailed distributions
Non-crossing and C-vine copula for spatial modelling

El Adlouni Salaheddine
Université de Moncton, Moncton, Canada – salah-eddine.el.adlouni@umoncton.ca

Abstract

Spatial Quantile Regression model is proposed to estimate the quantile curve for a given probability of non-exceedance, as function of the covariates and the location. C-vines are considered to represent the spatial dependence structure. The marginal at each location is an Asymmetric Laplace distribution where the location parameter is a function of the covariates. The full conditional quantile distribution is given for the Joe-Clayton copula. The problem of crossing curves will be discussed and an approach for the regularly varying distributions, is proposed.

Keywords: *Spatial Quantile Regression; C-vine copula; Crossing; Heavy tailed distribution.*

1. Introduction

In several applied statistic fields, extreme events are due to a combination of several interrelated events. When the relationship between the variable of interest and covariates is linear with normally distributed errors, classical regression models allows to estimate the conditional distribution of the central part such as the mean (El Adlouni et al. 2016, Gòmez et al., 2016). However, for extremes, the quantile regression (QR) model is one of the most suited models. Such models represent the effect of covariates on different quantiles especially for low probability of exceedance. Koenker and Bassett (1978) presented the problem of QR as linear optimisation problem with asymmetric loss function.

Two main points will be discussed in the present study. The non-crossing problem due to the estimation of the quantile curves separately, and the implementation of the Quantile Regression approach in spatial modelling. For the first point, the proposed approach is based on the assumption that the residuals have a regularly varying distribution. For the second point, a spatial Quantile Regression model based on C-vines copula is considered different spatial dependence structure. The marginal at each location is Asymmetric Laplace distribution, the location parameter is a non-parametric function of the covariates and the spatial dependence has a C-vine copula structure. The full conditional quantile distribution is given for the Joe-Clayton copula.

2. C-vine copula

A bivariate copula C is a distribution $C: [0,1]^2 \rightarrow [0,1]$ with uniform marginal. Sklar theorem (Sklar, 1959) shows that, for any bivariate distribution F with continuous marginal F_1 and F_2, there exist a unique copula $C(.,.)$, such that :

$$F(x_1, x_2) = C_{12}\big(F_1(x_1), F_2(x_2)\big) \ , \qquad \forall (x_1, x_2) \in \mathbb{R}^2 \qquad (1)$$

Theoretically, multivariate copulas are available for high dimensional variables. However, for practical statistical inference, hierarchical structure based on bivariate copula offers more flexible procedure (Joe, 1996).

The pair-copula construction provides a procedure to decompose high dimensional copula into a set of bivariate copulas (Aas et al., 2009). This approach allows to combine different families of copulas for different pairs of margins and higher order dependencies. Prior dependence structure is selected with respect to the causality between the variables and the overarching goal of the study.
Vine copula approach is based on the similar procedure for some special hierarchical dependence structures (Bedford and Cooke, 2001; Kurowicka and Cooke, 2006; Heinen and Valdesogo, 2009; Erhardt et al., 2015a,b; Pham et al., 2015). Three main structures are developed in the literature; the Regular, Canonical and Drawable vine copula (R-vine, C-vine and D-vine respectively).
Note that a multivariate copula density is a product of bivariate copula marginal (Joe, 1996; Bedford et Cooke (2001); Aas et al. (2009) et Czado (2010)):

$$f(x_1, \dots, x_n) = \prod_{j=1}^{n_s-1} \prod_{i=1}^{n_s-j} C_{i,(i+j)|(i+1),\dots,(i+j-1)} \cdot \prod_{k=1}^{n_s} f_k(x_k) \qquad (2)$$

with $c_{i,j|i_1,\dots,i_k} := c_{i,j|i_1,\dots,i_k}\left(F\left(x_i|x_{i_1},\dots,x_{i_k}\right), F\left(x_j|x_{i_1},\dots,x_{i_k}\right)\right)$ for $i < j$ and $i_1 < \dots < i_k$.
The first term represents the conditional bivariate copulas whereas the second corresponds to the marginal densities.
The Directed Acyclic Graph (DAG) representation of R-vine copula proposed by Bedford and Cooke (2001) allows to illustrate the dependence structure and to simplify the implementation of the multivariate model (Brechmann and Schepsmeier, 2013).
The C-vine copula structure is the most appropriate to model spatial dependencies. It allows to develop a k-dimensional distribution based on the (k-1)-nearest neighbors in term of Euclidean distance.

C-vine copula

The C-vine corresponds to the case where each tree has one node connected to all others (Figure1).
The multivariate density function of a C-vine in dimension $d = 5$, could be deduced from Eq. (2) and is given by :

$$f_{12345} = \overbrace{f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5}\ \cdot \overbrace{c_{12} \cdot c_{13} \cdot c_{14} \cdot c_{15}}\ \cdot \overbrace{c_{23|1} \cdot c_{24|1} \cdot c_{25|1}}\ \cdot \overbrace{c_{34|12} \cdot c_{35|12}}\ \cdot \overbrace{c_{45|123}} \qquad (3)$$

The regrouped items correspond to, nodes T1, T2, T3, T4 and edge T4, respectively (Figure 1).
In the case of high-dimensional model, statistical inference may be very requesting in time. Indeed, the multivariate model identification is a combination of several steps: (a) selection of the dependence structure; (b) best bivariate copula model for each edge and (c) parameter estimation.

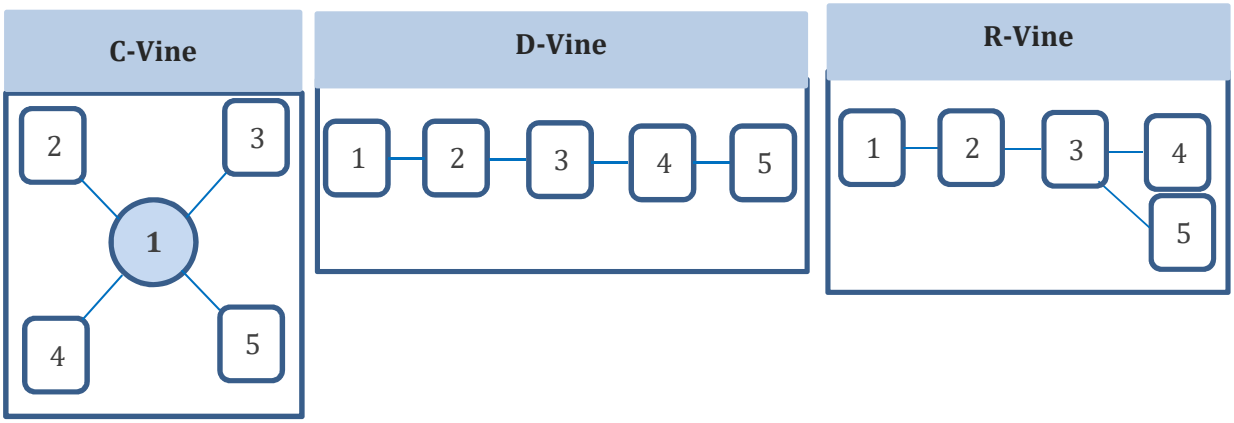

Figure1: Illustration of vine copula types for *d = 5*

3. Spatial quantile regression and C-vine copula

The main objective of the present study is to propose a spatial quantile regression model where the Canonical vine Copula (C-vine) represents the spatial dependence structure. The marginal distributions $Y^{(k)} \sim ALD\left(X^{(k)^t}\beta^{(k)}, \tau^{(k)}, p\right)$, $(k = 1, \dots, s)$, for s locations, depend on the at-site covariates $X^{(k)} = \left(X_k^1, \dots, X_k^c\right)$. The parameters of the local QR model are $\beta_k \in \mathbb{R}^c$ for a probability of non-exceedance p and $\tau^{(k)}$ is the scale parameter of the residual distribution. The parameters of the QR, at location $k, (k = 1, \dots, s)$ can be estimated by maximum likelihood or by regularized approach (Li et al. 2010) to ensure the oracle property. In the present work, the maximum likelihood with Lasso penalty is considered (El Adlouni et al. 2016). In the next section, the notion of vine copula, especially the C-vine representation, is introduced. Then the proposed methodology for spatial QR with C-vine copula is developed. The inference is detailed for Joe-Clayton copula and the conditional distribution for a given site and probability of non-exceedance is presented in explicit form.

Let $\left(Y^{(k)}\right)_{k=1,\dots,s}$ a spatial random field for a set of s sites; and $\left[y_1^{(k)}, \dots, y_n^{(k)}\right]^t$ the observed sample at location k; $(k = 1, \dots, s)$.

For each location k, the quantile $Q_p^{(k)}$ of order p is a linear combination of the local covariates, and the errors have an asymmetric Laplace distribution and can be estimated by $\widehat{Q_p^{(k)}} = X^{(k)^t}\widehat{\beta^{(k)}}$, where $\widehat{\beta^{(k)}}$ is the maximum likelihood estimator of the vector of parameters for the model:

$$\varepsilon_p^{(k)} = Y^{(k)} - X^{(k)^t}\beta^{(k)} \sim \text{ALD}\left(0, \tau^{(k)}, p\right) \quad (4)$$

In the case of neighbourhood of 4 locations, the likelihood is deduced from the multivariate density function of the vector $\left(\varepsilon_p^{(1)}, \varepsilon_p^{(2)}, \varepsilon_p^{(3)}, \varepsilon_p^{(4)}, \varepsilon_p^{(5)}\right)^t$. Indeed, the marginal distribution F_k of $Y^{(k)}$ is $\left(X^{(k)^t}\beta^{(k)}, \tau^{(k)}, p\right)$, $(k = 1, \dots, s)$ and the multivariate density is given by (Eq. 3).

Two main assumptions are considered for the proposed model.

H1: The parameters of the QR model for an ungauged site are the same as the nearest site,

H2: The multivariate dependence structure of the d-nearest sites, is the same for all location in the studied region.

The proposed Quantile Regression with C-vine model (QRCV) is based on C-vine copula of dimension $d = 5$ and ALD as marginal distributions. The Joe-Clayton copula is proposed for all the trees. The Joe-Clayton copula is an Archimedean copula with two parameters. A parameter related to the upper tail dependence and the second for the lower tail dependence. The Joe-Clayton (JC) copula has been widely used in modeling tail-dependence. It allows both positive and negative slopes for the quantile curves (Durrleman et al., 2000).

Conditional distribution and the h-function

In the bivariate case the h-function gives the expression of the conditional distributions. Czado et al. (2012) give the expressions of the h-function for some copulas such as Normal, Student, BB1 and BB7 (JC copula). For the JC copula, the h-function is given by:

$$h(u|v)=\left[1-\left(\left(1-(1-u)^{\theta}\right)^{-\delta}+\left(1-(1-v)^{\theta}\right)^{-\delta}-1\right)^{-\frac{1}{\delta}}\right]^{\frac{1}{\theta}-1} .\left[\left(1-(1-u)^{\theta}\right)^{-\delta}+\left(1-(1-v)^{\theta}\right)^{-\delta}-1\right]^{-\frac{1}{\delta}-1} .\left(1-(1-v)^{\theta}\right)^{-\delta-1}.(1-v)^{\theta-1} \quad (5)$$

The h-function is used recursively to deduce the conditional uniform distribution at the ungauged sites.

Conditional distribution at the ungauged station

In the case of 5 nodes C-vine the conditional distribution at an ungauged location is given by:

$$F(y_5|y_1,\dots,y_4)=\frac{\partial C_{45,123}\left(F(y_5|y_1,y_2,y_3),\ F(y_4|y_1,y_2,y_3)\right)}{\partial F(y_4|y_1,y_2,y_3)} \quad (6)$$

where

$$F(y_5|y_1,y_2,y_3)=\frac{\partial C_{35,12}(F(y_5|y_1,y_2),F(y_3|y_1,y_2)\ )}{\partial F(y_3|y_1,y_2)}$$

$$F(y_4|y_1,y_2,y_3)=\frac{\partial C_{34,12}(F(y_4|y_1,y_2),F(y_3|y_1,y_2)\ )}{\partial F(y_3|y_1,y_2)}$$

$$F(y_3|y_1,y_2)=\frac{\partial C_{23,1}(F(y_3|y_1),F(y_2|y_1)\ )}{\partial F(y_2|y_1)}$$

$$F(y_4|y_1,y_2)=\frac{\partial C_{24,1}(F(y_4|y_1),F(y_2|y_1)\ )}{\partial F(y_2|y_1)}$$

$$F(y_5|y_1,y_2)=\frac{\partial C_{25,1}(F(y_5|y_1),F(y_2|y_1)\ )}{\partial F(y_2|y_1)}$$

The h-function for the C-vine can be deduced recursively from the bivariate expression and is given by:

$$F(y_5|y_1,y_2,y_3,y_4)=h\left(h\left(T_{25,1}|T_{23,1};\alpha_{35,12}\right)\middle|\ h\left(T_{24,1}|T_{23,1};\ \alpha_{34,12}\right);\ \alpha_{45,123}\right) \quad (7)$$

where $T_{jk,1}=h\left(h(u_k|u_1;\ \alpha_{1k})\middle|\ h\left(u_j|u_1;\ \alpha_{1j}\right);\ \alpha_{jk,1}\right),\quad 2\le j<k\le 5$

$u_k=F_k(y_k)$, $F_k(.)$ the marginal of the residuals and $\alpha_{jk,.}=(\theta,\delta)$ corresponds to the parameter vector of the JC copula of the corresponding edge.

The proposed spatial model may have two main situations. The first one corresponds to the information transfer from the neighbours to site with small dataset. In this case, no assumption of stationary spatial random field is needed. However, for spatial interpolation, we will assume that the spatial random field is stationary in space and time. This is necessary to have the same dependence model between locations of the spatial region. Indeed, for a given location, even with no available dataset, the quantile estimates are obtained through the conditional local distribution (with ALD distribution as marginal), the spatial dependence structure (C-vine copula) and the values of the local covariates.

Under the assumption of spatial stationarity of the conditional distribution of temperature knowing covariates, the distribution of $Y(s_0)$ at an unobserved location s_0 corresponds to $F(Y(s_0)|y_1,y_2,y_3,y_4)$ where (y_1,y_2,y_3,y_4) is the vector of the nearest observed neighbours. This conditional distribution can be considered to estimate the p^{th} conditional quantile at the ungauged site s_0 , and is given by: $Q_p^{(s_0)} = F^{-1}_{Y_{s_0|y_1,\dots,y_4}}(p)$, where the inverse of the conditional distribution $F^{-1}_{Y_{s_0|y_1,\dots,y_4}}$ has an explicit form.

4. Non-crossing quantile regression

Concerning the crossing problem of the quantile regression curves, a new approach is developed. The proposed approach allows to estimate several quantiles simultaneously for regularly varying distributions (El Adlouni and Baldé, 2017). The non-crossing property is guaranteed by a constraint on the minimal distance that should separate quantiles' curves, which depend on the extreme index. The inference implemented in bayesian framework where the constraints are introduced as priors and the conditional posterior distributions are deduced explicitly. Simulations and comparison with previous studies (Bondell et al., 2010) will be presented.

5. Conclusions

In this study, a new approach for spatial Quantile Regression with C-vine copula (QRC-vine) is presented. The C-vine copula structure is compatible with the spatial dependence modelling. The first level represents the bivariate distributions of the target sites with the (d-1)-neighbours. The marginal at each location is an Asymmetric Laplace distribution, where the location parameter depends on the covariates. The maximum likelihood method is proposed to estimate the marginal parameters and the Joe-Clayton copula has been proposed for bivariate distributions of the C-vine levels. The QRC-vine model has the advantages of the C-vines where the dependence structure can be fixed independently of the marginal distributions. We discussed also the problem of crossing curves when the quantile regression parameters are estimated separately. A constraints based on the assumption of regularly varying errors are considered to ensure the non-crossing even for close probabilities of non-exceedance.

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. Insurance, Mathematics and Economics 44, 182-198.

Bedford, T. and R. M. Cooke (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. Annals of Mathematics and Artificial Intelligence 32, 245-268.

Bondell, H. D., Reich, B. J., and Wang, H. (2010). Non-crossing quantile regression curve estimation. Biometrika 97, 825-838.

Brechmann, E. C. and U. Schepsmeier (2013). Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine. Journal of Statistical Software, 52 (3), 1-27. http://www.jstatsoft. org/v52/i03/.

Czado, C., 2010. Pair-copula constructions of multivariate copulas. In: Jaworski, P., Durante, F., H¨ardle, W., Rychlik, T. (Eds.), Copula Theory and Its Applications. Springer, Berlin.

Czado, C., U. Schepsmeier, and A. Min (2012). Maximum likelihood estimation of mixed c-vine pair copula with application to exchange rates. Statistical Modeling 12, 229–255.

Clayton, D. G. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incident. Biometrika 65, 141-151.

Durrleman, V., Nikeghbali, A., and Roncalli, T. (2000). Which Copula is the right one? Technical report, Groupe de Recherche Opérationelle, Crédit Lyonnais.

El Adlouni, S., G. Salaou and A. St-Hilaire (2017). Regularized Bayesian Quantile Regression. Communication in Statistics Simulation and Computation. Accepted, January 2017.

El Adlouni, S. and I. Baldé (2017). Bayesian non-crossing Quantile Regression for Regularly Varying distributions. Submitted.

Erhardt, T. M., Czado, C., and Schepsmeier, U. (2015a). Spatial composite likelihood inference using local C-vines. Journal of Multivariate Analysis, 138, 74-88.

Erhardt, T. M., Czado, C. and Schepsmeier, U. (2015b). R-vine models for spatial time series with an application to daily mean temperature. Biometrics 71, 323-332.

Gòmez M., M. Concepcion Ausın and C. Dominguez (2016). Seasonal copula models for the analysis of glacier discharge at King George Island, Antarctica. Stoch Environ Res Risk Assess, DOI 10.1007/s00477-016-1217-7.

Heinen, A., A. Valdesogo (2009). Asymmetric CAPM dependence for large dimensions: The Canonical Vine Autoregressive Model. CORE discussion papers 2009069, Universit´e catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Joe, H. (1996) Families of m-variate distributions with given margins and m(m- 1)/2 bivariate dependence parameters. In L. Rüschendorf , B. Schweizer and M. D. Taylor, editor, Distributions with Fixed Marginals and Related Topics.

Koenker, R., and G. S. Bassett (1978): “Regression Quantiles,” Econometrica, 46, 33–50

Kurowicka D.and R. M. Cooke (2006). Uncertainty Analysis with High Dimensional Dependence Modelling, Chichester, John Wiley,

Pham M.T., H. Vernieuwe, B. De Baets, P. Willems and N. E. C. Verhoest (2015). Stochastic simulation of precipitation-consistent daily reference evapotranspiration using vine copulas. Stoch Environ Res Risk Assess, DOI 10.1007/s00477-015-1181-7.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris, 8:229–231.