Covariate-dependence copula model based on web semantics

Yu Liu

Beihang University

liuyu96@buaa.edu.cn

October 22, 2018

Outline

1. Idea
2. Covariate-dependent copula model
   - Sklar's theorem
   - Tail-dependence
   - Nonlinear correlation coefficient - Kendall's τ
   - The reparameterized Joe-Clayton copula
   - Covariate-dependent copula features
   - Marginal models
   - Covariates
3. News information
   - Data generation
   - Convert document into vector
4. Future work

Idea

- Measuring the dependence of financial market
- News will have an impact on financial market volatility

Sklar's theorem

In a bi-variate setting: let F_{xy} be a joint distribution with margins F_x and F_y. Then exist a function C, such that:

sklar's theorem

$$F_{xy}(x, y) = F(F_1^{-1}(u_1), F_2^{-1}(u_2)) = C(F_x(x), F_y(y))$$

If X and Y are continuous, then C is unique. Conversely if C is a copula function and F_x and F_y are distribution functions, then the function F_{xy} is a joint distribution with margins F_x and F_y.

Tail-dependence

$$\lambda_L = \lim_{u \to 0^+} p(X_1 < F_1^{-1}(u) | X_2 < F_2^{-1}(u)) = \lim_{u \to 0^+} \frac{C(u, u)}{u}$$

$$\lambda_U = \lim_{u \to 1^-} p(X_1 > F_1^{-1}(u) | X_2 > F_2^{-1}(u)) = \lim_{u \to 1^-} \frac{1 - 2u + C(u, u)}{1 - u}$$

Kendall's τ

As for a set of observations: $((x_1, y_1), (x_2, y_2), ..., (x_n, y_n))$:

- *concordant*: if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$
- *discordant*: if $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$

The kendall's τ coefficient is defined as:

$$\tau = \frac{n(concordant) - n(discordant)}{n(n-1)/2}$$

It also can be writed as :

$$\tau = 4\int\int F(x_1, x_2)dF(x_1, x_2) - 1 = 4\int\int C(u_1, u_2)dC(u_1, u_2) - 1$$

The reparameterized Joe-Clayton copula

Joe-Calyton copula

$$C(u, v|\theta, \delta) = \eta(\eta^{-1}(u) + \eta^{-1}(v)) = 1 - [1 - \{(1-\bar{u}^{\theta})^{-\delta} + (1-\bar{v}^{\theta}) - 1\}^{-1/\delta}]^{1/\theta}$$

where $\eta(s) = 1 - [1 - (1+s)^{-1/\delta}]^{1/\theta}$, $\theta \geq 1$, $\delta \geq 0$, $\bar{u} = 1 - u$, and $\bar{v} = 1 - v$.

Tail-dependece

$\lambda_L = 2^{-1/\delta}$and $\lambda_U = 2 - 2^{1/\theta}$

The reparameterized Joe-Clayton copula

$$\tau = \begin{cases} 1 - 2/[\delta(2-\theta)] + 4B(\delta+2, 2/\theta-1)/(\theta^2\delta) 1 \leq \theta < 2 \\ 1 - [\psi(2+\delta) - \psi(1) - 1]/\delta\theta = 2 \\ 1 - 2/[\delta(2-\theta)] - 4\pi/[\theta^2\delta(2+\delta) sin(2\pi\theta) B(1+\delta+2/\theta, 2-2/\theta)]\theta > \end{cases} \tag{1}$$

The reparameterized Joe-Clayton copula

Reparametrization

$$C(u, v|\lambda_L, \tau) = 1 - [1 - [[1 - \bar{u}^{log2/log(t-\tau^{-1}(\lambda_L))}]^{log2/log\lambda_L}$$

$$+[1 - \bar{v}^{log2/log(2-\tau^{-1}(\lambda_L))} - 1]^{log\lambda_L/log2}]^{log(2-\tau^{-1}(\lambda_L))/log2}$$

Where $\tau^{-1}(\lambda_L) = \lambda_U = 2 - 2^{1/\theta}$

Covariate-dependent copula features

The covariate-dependent copula model that allows the copula features to be linked to the observed covariates:

$$\tau = l_\tau^{-1}(\boldsymbol{X}\boldsymbol{\beta}_\tau) \qquad \textit{and} \qquad \lambda = l_\lambda^{-1}(\boldsymbol{X}\boldsymbol{\beta}_\lambda)$$

- λ without subscripts represents the dependences in the lower or upper tails
- τ is Kendall's τ
- $\mathbf{X}$ is the set of covariates matrix
- β with subscripts is the corresponding coefficients vector
- $l_\lambda(\cdot)$ and $l_\tau(\cdot)$ are suitable *link function* that connect λ and τ with $\mathbf{X}$

Marginal models

We assume the marginal models to be *split-t distributions*, then we allow the mean $\boldsymbol{\mu_k}$, the scale $\phi_{\boldsymbol{k}}$, the degree of freedom $\boldsymbol{\nu_k}$, the skewness $\boldsymbol{\kappa_k}$ of the split-t density in the kth to be linked to covariates:

$$\mu_k = \boldsymbol{X}\boldsymbol{\beta}_{\mu k}, \nu_k = exp(\boldsymbol{X}\boldsymbol{\beta}_{\nu k});$$

$$\phi_k = exp(\boldsymbol{X}\boldsymbol{\beta}_{\phi k}), \kappa_k = exp(\boldsymbol{X}\boldsymbol{\beta}_{\kappa k})$$

where $\boldsymbol{X}_k$ is the covariate matrix in the kth margin

Covariates

Covariates(X_s)	**Explanation**
LastDay	the returns from yesterday
LastWeek	the returns from the previous five trading days
LastMonth	the returns from the previous twenty trading days
CloseAbs95	$(1-\rho)\sum_{s=0}^{\infty}\rho^s\lvert y_{t-2-s}\rvert$
CloseSqr95	$(1-\rho)\sum_{s=0}^{\infty}\rho^s(y_{t-2-s})^2$
MaxMin95	$(1-\rho)\sum_{s=0}^{\infty}\rho^s(lnp_{t-1-s}^{(h)} - lnp_{t-1-s}^{(l)})$

Table: Seven variables

Data generation

Crawling the following information:

- Company: JD and BABA
- source: *https:caixin.com*
- date: 2017/05/23 - 2018/08/17

News data preprocessing:

- Combine news data from the same day
- Combine weekend news data with next Monday news data
- Remove news data from weekend

Convert document into vector

Our model require the input to be represented as a fixed-length feature vector. One of the most common fixed-length features is bag-of-words. Despite their popularity, bag-of-words features have two major weaknesses:

- they lose the ordering of the words
- they ignore semantics of the words

So, we use **Doc2vec** to convert document into vectore. Then we can convert each document into a 200-dimensional vector. After that, the function *umap()* can help us to reduce the dimension of the vector to two dimensions.

Future work

Experiment	Covariate	Variable selected
1	one	None
2	X_s	Y
3	X_s	N
4	X_n	Y
5	X_n	N
6	$X_s + X_n$	Y
7	$X_s + X_n$	N

Table: Experiment programe