Variational Inference for high dimensional structured factor copulas

Hoang Nguyen, María Concepcíon Ausín and Pedro Galeano

Department of Statistics, Universidad Carlos III de Madrid

Introduction

Factor copula model helps to explain the dependence structure of variables in term of a few latent variables. Due to the flexibility of copula functions, factor copula can capture well the correlation as well as the tail dependence in extreme events. Estimation of factor copula models in high dimensions is a challenging problem.

- Intractable likelihood.
- Unknown bivariate copula links.

Main contribution

In this paper, we propose a Bayesian procedure to make inference for multi-factor and structured factor copulas.

- We employ a VI approximation by [1] to estimate the different specifications of the factor copula models in high dimensions.
- We compare posterior estimation of VI and the MCMC approach.
- We derive an automated procedure to recover the dependence structure based on the posterior means of the latent factors.

Structure factor copula models

In one-factor copula model, the dependence structure is characterized through d bivariate copulas between variable U_i and a latent variable V_0 where $U_i, V_0 \sim U(0,1)$ and $i = 1, \dots, d$.

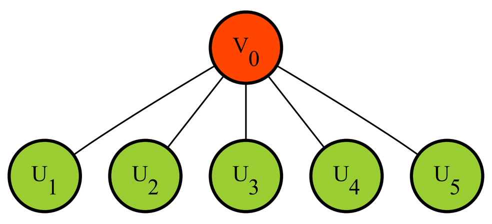

Figure 1: One-factor copula models ([2])

In order to account for more latent variables, [3] extend the one-factor copula to structured factor copulas by adding a hierarchical dependence structure for latent variables in case of nested factor copulas or by using the latent variables to capture the conditional dependence in higher tree layers in case of bi-factor copulas.

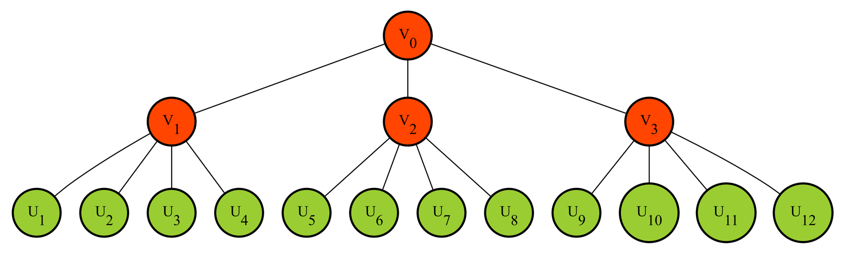

Figure 2: Nested factor copulas with $d = 12$ and $G = 3$ ([3])

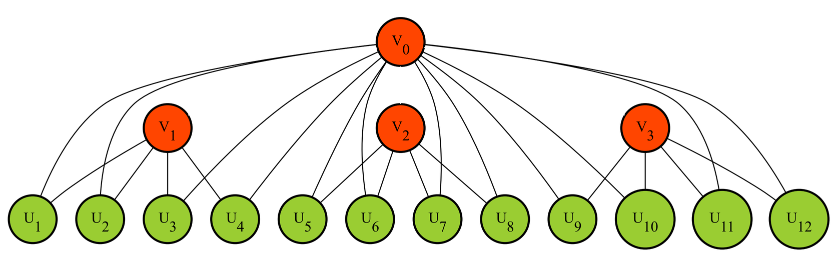

Figure 3: Bi-factor copulas with $d = 12$ and $G = 3$ ([3])

Automatic differentiation variational inference [1]

VI approach tries to minimize the Kullback-Leibler divergence between the proposal distribution $q(\Theta;\lambda)$ and the posterior distribution $p(\Theta|u)$ where Θ is the set of unknown, which is equivalent to maximize the evidence lower bound (ELBO)

$$\arg\min_{\lambda} \text{KL}\left(q(\Theta;\lambda)||p(\Theta|u)\right)$$

$$\arg\max_{\lambda} \text{ELBO}(q) = \mathbb{E}_q[\log p(u,\Theta)] - \mathbb{E}_q[\log q(\Theta;\lambda)]$$

$$= \log p(u) - KL(q(\Theta;\lambda)||p(\Theta|u)) \leq \log p(u)$$

ELBO is evaluated by sampling S samples from $q(\Theta;\lambda)$ and averaging over generated samples.

$$\text{ELBO}(q) \approx \frac{1}{S}\sum_{s=1}^{S}\left[\log p(u,\Theta_{(s)}) - \log q(\Theta_{(s)};\lambda)\right]$$

- Optimize ELBO based on its unbiased noisy gradient.
- Transform the space of the model parameters to the real coordinate space.
- Approximate the posterior in the real domain by a product of Gaussian density.

VI vs MCMC

Table 1: Time estimation using VI and NUTS (d = 100 series and T = 1000 obs)

Copula type	Gaussian	Student	Clayton	Gumbel	Frank	Joe	Mix
Time estimated (s) using VI							
One-factor	10	509	24	41	11	17	98
Nested factor	20	783	25	43	13	19	121
Bi-factor	90	3366	165	267	87	154	596
Time estimated (s) using NUTS							
One-factor	1097	491166	3762	5262	1083	2221	8535
Nested factor	1786	567862	5977	6098	1264	3478	20520
Bi-factor	17935	1758007	83680	177932	30345	236809	118729

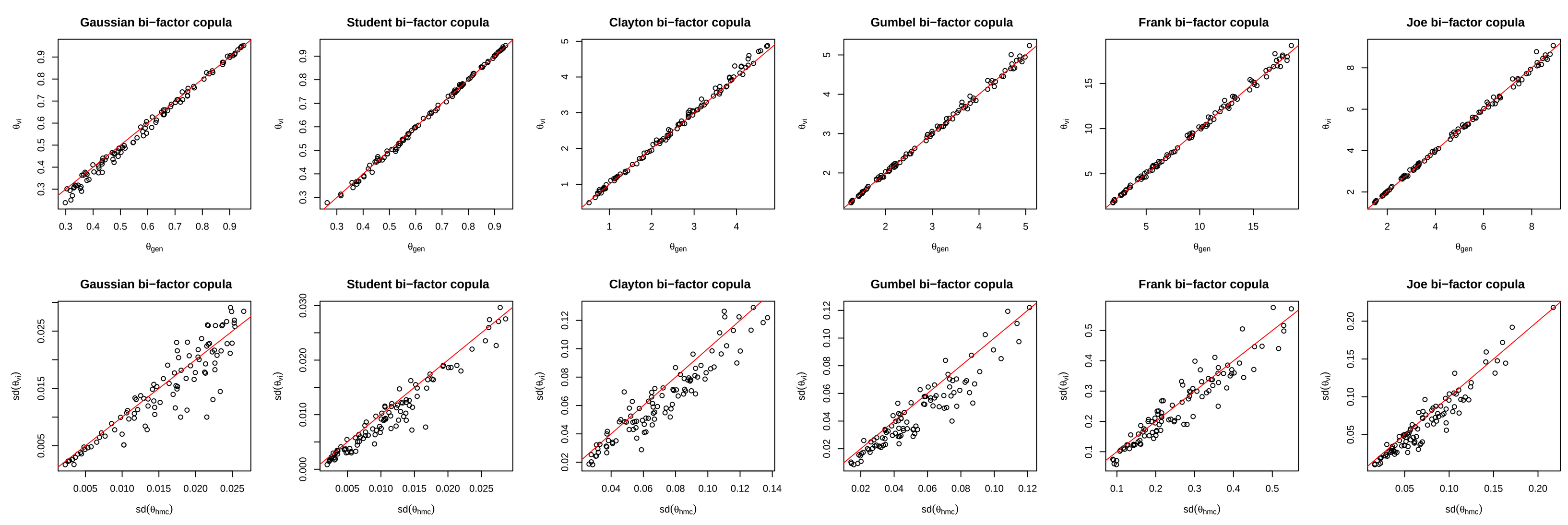

Figure 4: Posterior mean and standard deviation of VI estimation vs MCMC estimation

Model selection

Starting with an arbitrage bivariate links, we can estimate the common factors and obtain their posterior mean. We inspect the initial assumption of bivariate copula functions by fitting the best bivariate copula function for copula data u_i and estimated latent v based on BIC. We reestimate the new structure until convergence (unchanged dependence structure). This procedure assures to improve overal BIC of the factor copula model.

Table 2: Model comparison for Bi-factor copula models

Copula type	Gaussian	Student	Clayton	Gumbel	Frank	Joe	Mix
Initial at correct structure							
AIC	-115.1	-170.4	-275.0	-254.4	-214.7	-279.4	-212.7
BIC	-114.1	-168.9	-274.0	-253.4	-213.7	-278.5	-211.6
Initial at random structure							
% accuracy Tree 1	99	76	84	99	99	46	85
% accuracy Tree 2	97	83	19	66	92	2	67
AIC	-115.1	-170.1	-263.1	-250.6	-214.0	-266.8	-209.7
BIC	-114.1	-168.8	-262.1	-249.6	-213.0	-265.8	-208.7

Empirical Results

We estimate the joint dependence of daily temperature measured at 477 stations in Germany. The time is selected during the period from 06/04/2015 to 31/12/2017 which results in $T = 1000$ observations. First, we model the marginal distribution of daily temperature using ARMA(1,1) with skew Student distribution. We obtain the copula data as the cdf transformation of the residuals. Then, we cluster stations into 23 groups such that the distance of stations in each group is at most 200 kilometers and propose multi-factor and structured factor copula models.

Table 3: Model comparison of daily temperature dependence

Structure	One-factor	Two-factor	Nested factor	Bi-factor
ELBO	234.7	308.7	476.3	486.8
AIC	-469.9	-622.1	-956.0	-985.4
BIC	-465.7	-616.0	-951.4	-978.6
$\log p(u\|\theta)$	235.8	312.3	478.9	494.1
# bivariate links	477	896	500	954
# Gaussian	37	195	27	353
# Student	369	343	437	424
# Clayton (rotated)	0	24	0	1
# Gumbel (rotated)	70	118	36	118
# Frank (rotated)	1	216	0	57
# Joe (rotated)	0	0	0	1
# Independence	0	58	0	0

Discussion and Extension

- Accurate posterior mean
- Rotated Archimedean copula similarity
- Time varying factor copula models
- Clustering using factor copulas.

References

[1] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei, "Automatic differentiation variational inference," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 430–474, 2017.

[2] P. Krupskii and H. Joe, "Factor copula models for multivariate data," *Journal of Multivariate Analysis*, vol. 120, pp. 85 – 101, 2013.

[3] P. Krupskii and H. Joe, "Structured factor copula models: Theory, inference and computation," *Journal of Multivariate Analysis*, vol. 138, pp. 53–73, 2015.