Contributions to multivariate copula modelling

Gildas Mazo F. Forbes S. Girard

Inria & Laboratoire Jean Kuntzmann

Les rencontres des jeunes statisticiens
Aussois
26-30 août 2013

The starting problem

Let $\mathbf{X} = (X_1, \ldots, X_d) \sim F$.

What are the models $F(\theta)$ for F?

- $d = 1$ ✓
- $d = 2$ ✓
- $d \geq 3$ ✗

Can you think of some?

You may have came up with

- $N_d(0, \Sigma)$ (multivariate normal)
- $f(\mathbf{x}) = f_1(x_1) f_2(x_2) f_{3|12}(x_3|x_1, x_2) f_{4|2}(x_4|x_2) f_{5|1}(x_5|x_1)$ (Graphical models)
- $F(\mathbf{x}) = \exp\left[-\left(x_1^{-1/\theta} + \cdots + x_d^{-1/\theta}\right)^{\theta}\right]$ (One-parameter)
- $F(\mathbf{x}) = F_1(x_1) \ldots F_d(x_d)$ (independence)

You may have came up with

- $N_d(0, \Sigma)$ (multivariate normal)
- $f(\mathbf{x}) = f_1(x_1) f_2(x_2) f_{3|12}(x_3|x_1, x_2) f_{4|2}(x_4|x_2) f_{5|1}(x_5|x_1)$ (Graphical models)
- $F(\mathbf{x}) = \exp\left[-\left(x_1^{-1/\theta} + \cdots + x_d^{-1/\theta}\right)^{\theta}\right]$ (One-parameter)
- $F(\mathbf{x}) = F_1(x_1) \ldots F_d(x_d)$ (independence)

But we want

- tail dependence ($\Longrightarrow$ nonnormal models)
- a suitable number of parameters

Margins are known but not dependence

Example Rainfall maximum at station i

$$\text{Extreme value theory} \implies$$

$$\text{Maximum} \sim F_i(x) = \exp - \left((1 + \xi_i \frac{x - \mu_i}{\sigma_i} \right)^{-1/\xi_i}$$

However, to model the **dependencies** between stations are challenging...

Copula

- The copula C of F (or $\mathbf{X}$) is the d.f. of $\mathbf{U} = (U_1, \ldots, U_d) = ((F_1(X_1), \ldots, F_d(X_d))$,
- (Sklar 1959) unique d.f. such that

$$F(\mathbf{x}) = C\left(F_1(x_1), \ldots, F_d(x_d)\right)$$

Example $C(u) = \exp\left[-\left((-\log u_1)^{1/\theta} + \cdots + (-\log u_d)^{1/\theta}\right)^{\theta}\right]$ (Gumbel)

$CA(u) = \min(u_1, u_2)^{\theta}(u_1 u_2)^{1-\theta}$ (Cuadras-Augé)

To summarize

- Copula $\equiv$ dependence
- Copula + Margins $\equiv$ a multivariate model

Our problem: Copula modelling

Let $\mathbf{U} = (U_1, \ldots, U_d)$ be distributed according to some copula C.

What are the models $C(\theta)$ for C?

Entering the world of copulas...

Product of copulas

A product of copulas is not a copula. But (Liebscher construction 2008)

$$C(u) = \prod_{j=1}^{k} C_j \left(g_{j1}(u_1), \ldots, g_{jd}(u_d) \right) \text{ (and conditions for the } g_{ji})$$

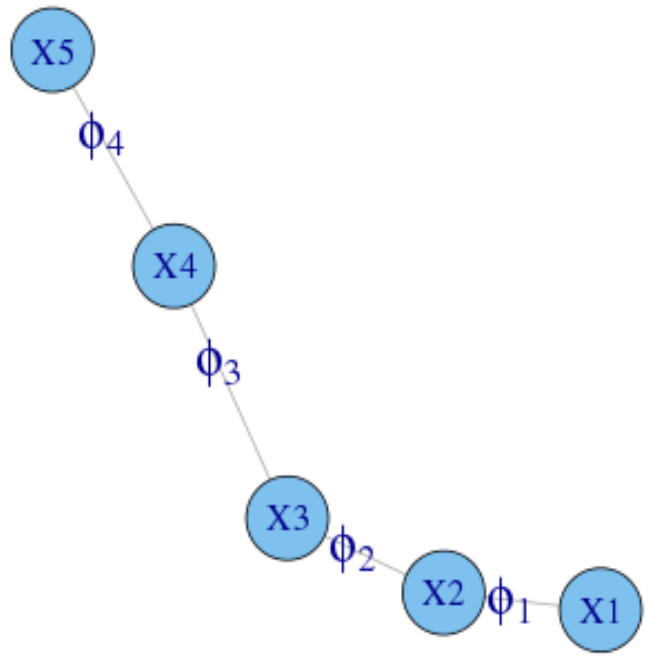
is a copula.

No inference available

Our contribution

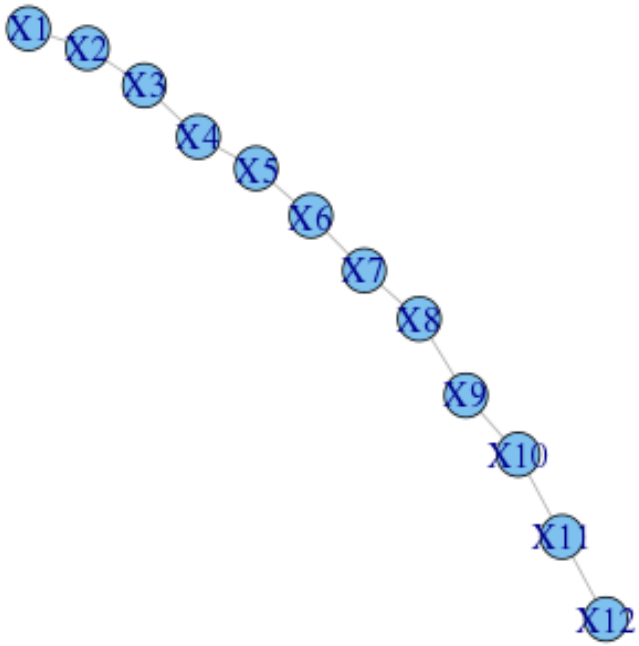
We have shown **equivalence** between Liebscher and Cumulative Distribution Networks (CDN's).

$$F(x) = \Phi_1(x_1, x_2)\Phi_2(x_2, x_3)$$
$$\Phi_3(x_3, x_4)\Phi_4(x_4, x_5)$$
$$C(u) = C_1(u_1, u_2^{1/2})C_2(u_2^{1/2}, u_3^{1/2})$$
$$C_3(u_3^{1/2}, u_4^{1/2})C_4(u_4^{1/2}, u_5)$$



The interest of CDN's is that **maximum likelihood inference** is available.

Inference performance



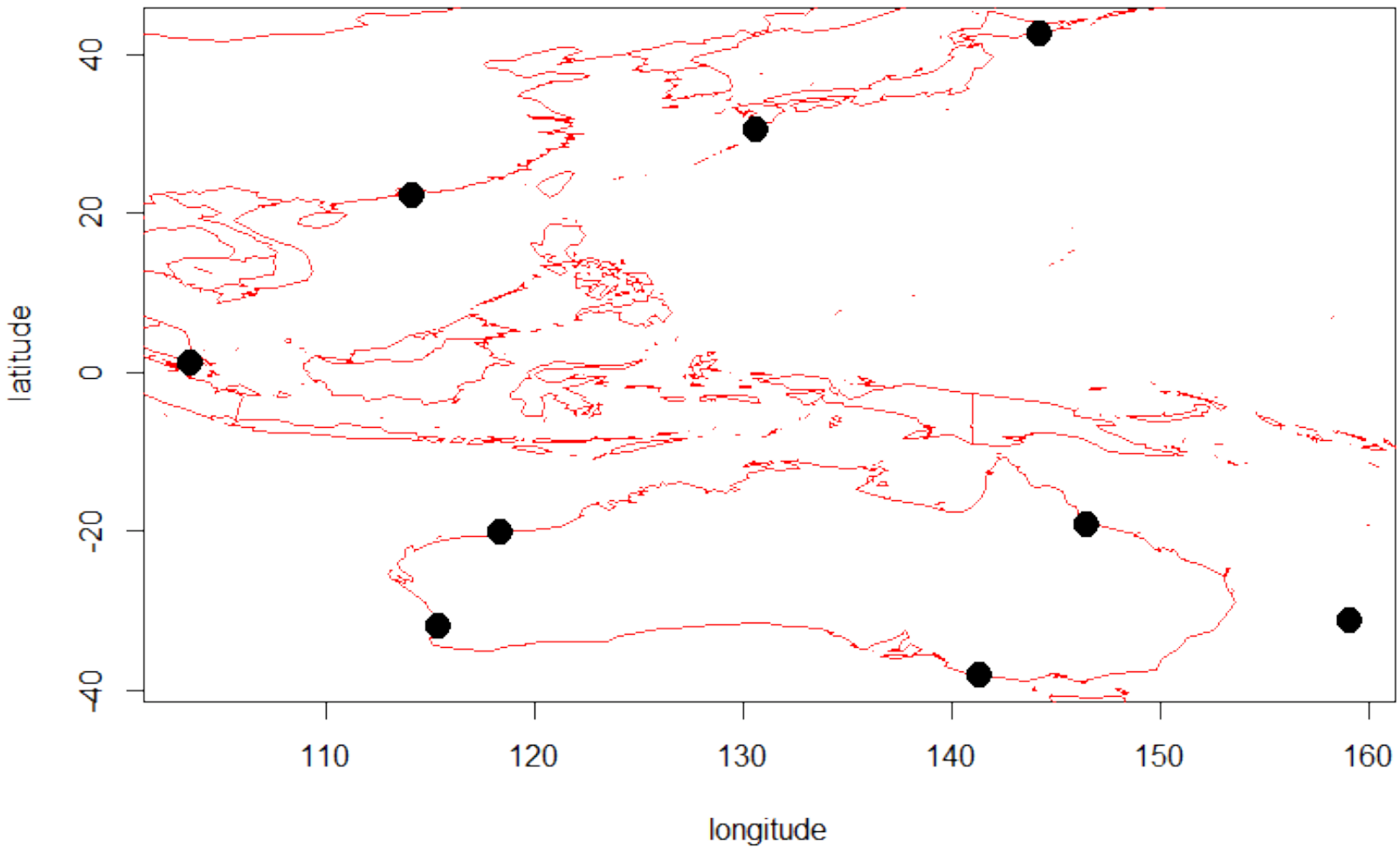
- Simulation of a 12 variables copula CDN with Gumbel linking copulas (70 observations)
- Estimate the parameters (maximum-likelihood inference)

	pairwise	multivariate
mse	0.0034	0.0019

44% improvement

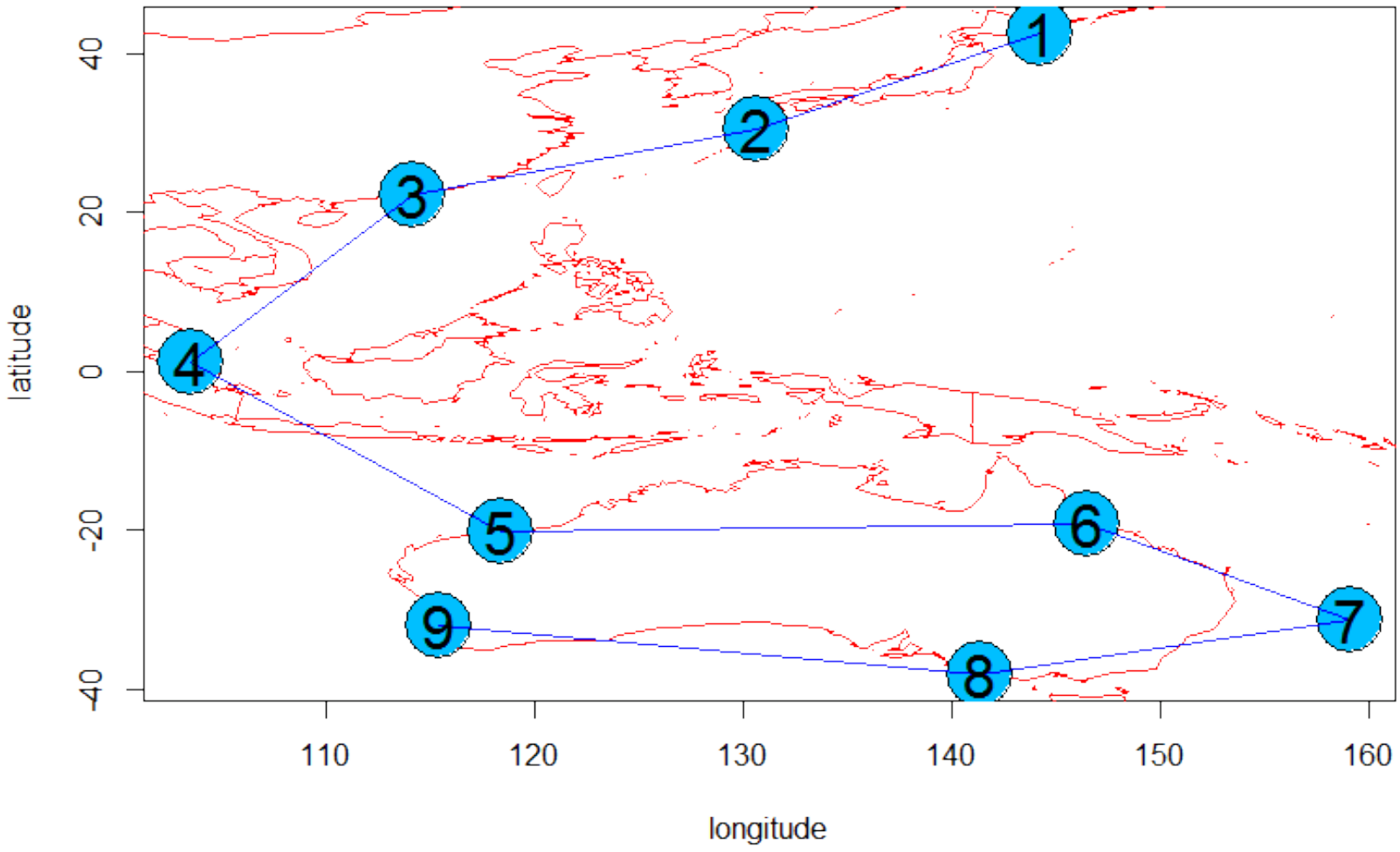
Application (Oceanography)

Sea level measurements

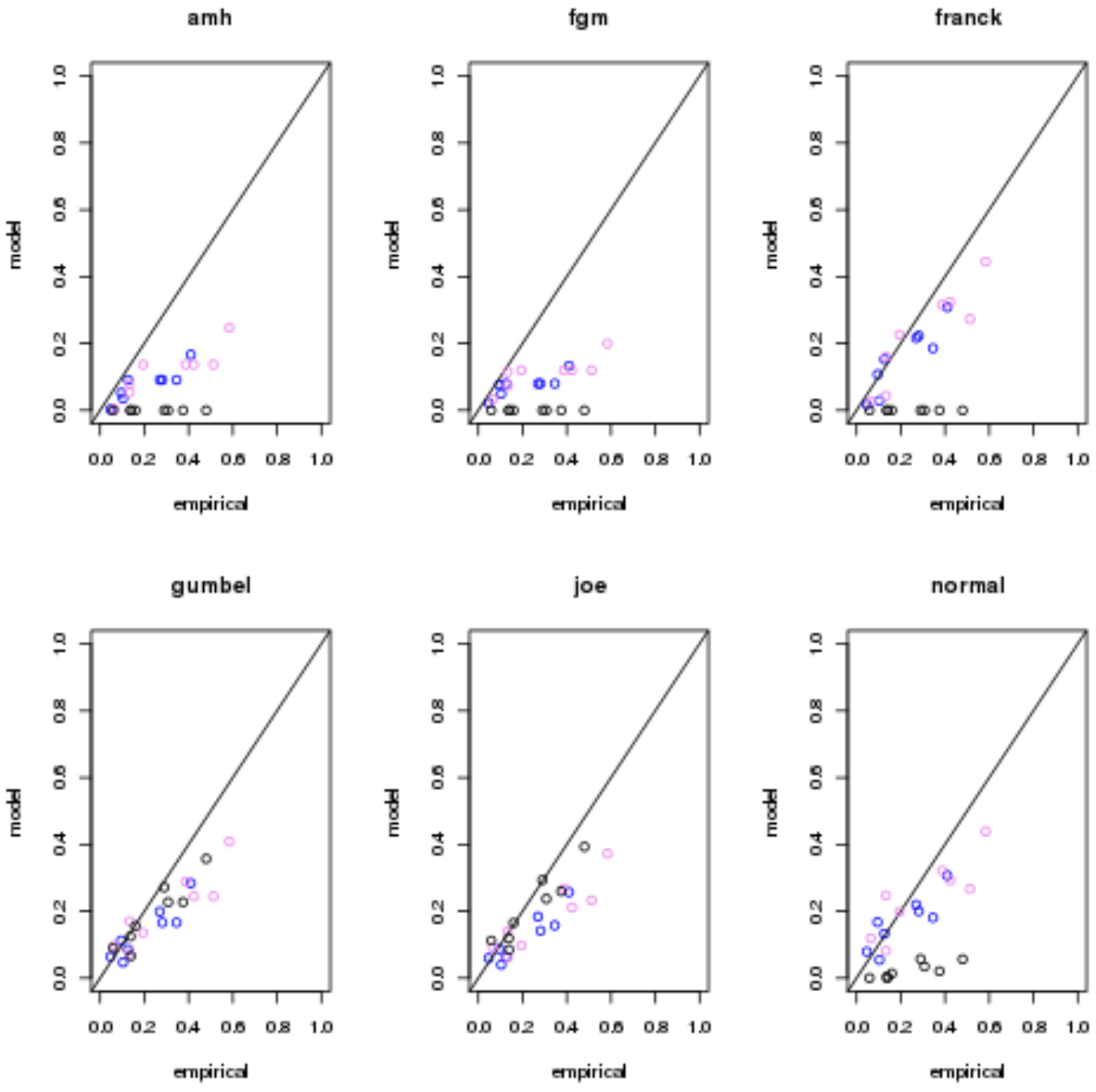


Application (Oceanography)

Sea level measurements



Results


amh
fgm
franck
gumbel
joe
normal
model
empirical
0.0
0.2
0.4
0.6
0.8
1.0

Our model proposition

Recall that $\mathbf{U} = (U_1, \ldots, U_d)$ is the vector of interest. Add a factor (latent variable) U_0 s.t.

- the U_i are independent given U_0
- $(U_0, U_j) \sim L_{0j}$

The copula is given by

$$C(\mathbf{u}) = \int_0^1 \prod_{j=1}^{d} L_{j|0}(u_j|u_0)du_0$$

$$L_{0j} \equiv \text{CA} \implies C(\mathbf{u}) \text{ is explicit}$$

Properties:

- ✓ Copula is explicit
- ✓ $O(d)$ parameters
- ✓ flexible

Inference

Pairwise moment strategy

$$\hat{\Theta}_n = \underset{\theta \in (0,1)^d}{\operatorname{argmin}} \sum_{i<j} \left(\theta_i \theta_j - \hat{\lambda}_{ij}^{(n)} \right)^2 .$$

Estimator Properties

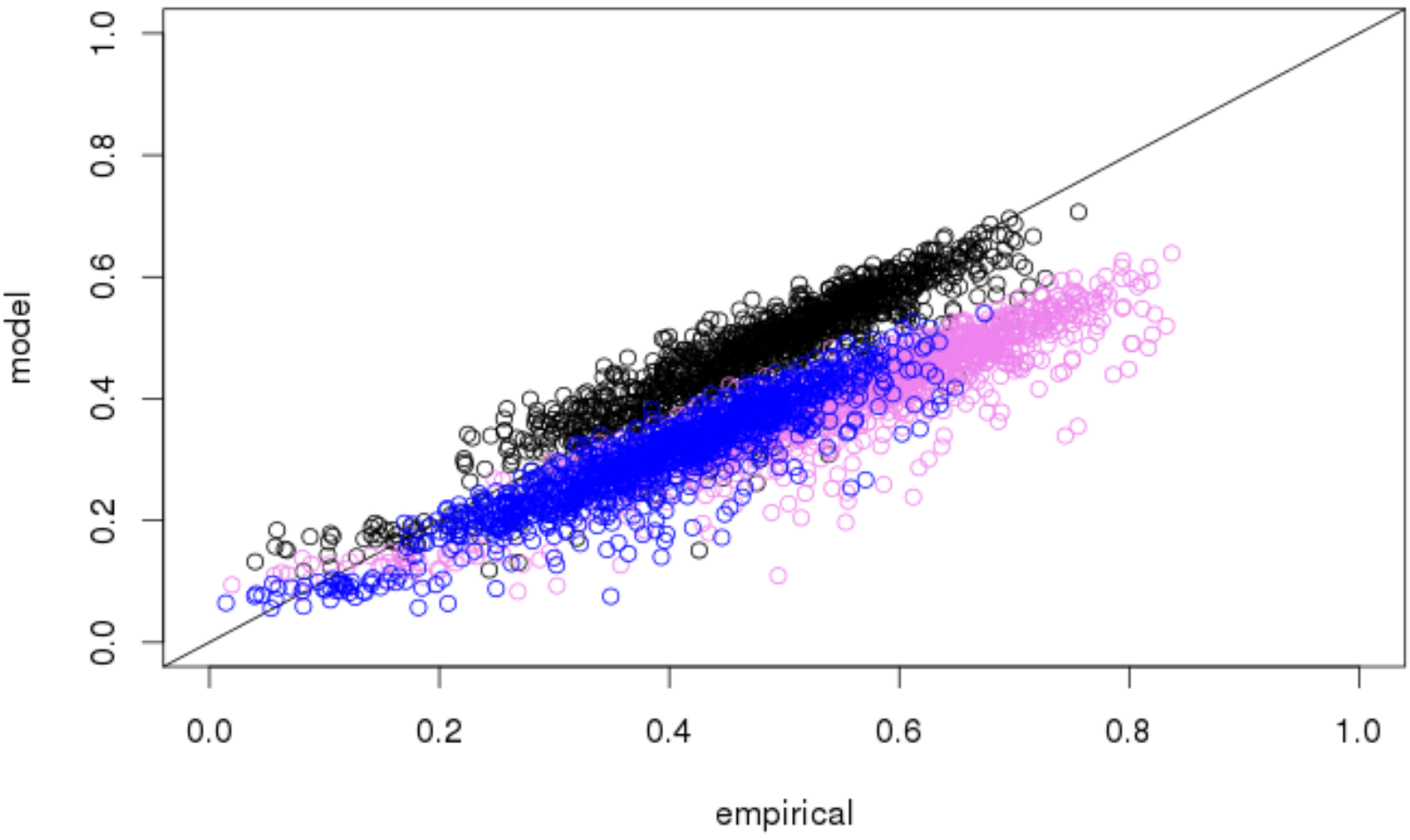
- $P\left[\hat{\Theta}_n \neq \emptyset \text{ and } \hat{\Theta}_n \subset \overline{B}_\varepsilon(\theta_0)\right] \to 1$ (existence)
- $\hat{\theta}_n \xrightarrow{P} \theta_0$ (consistency)
- $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_d(0, \cdot)$ (asymptotic normality)

Application (finance)

49 industry portfolio returns ($\simeq$500 observations)

	Agric	Food	Soda	Beer	Smoke
196907	0.06883365	0.04971319	0.1873805	0.2390057	0.14149140
196908	0.69980880	0.81835564	0.7896750	0.8699809	0.67304015
196909	0.54302103	0.31931166	0.7782027	0.3164436	0.68068834
196910	0.97896750	0.94072658	0.8527725	0.9713193	0.96940727
196911	0.30783939	0.31357553	0.1357553	0.1892925	0.41873805
196912	0.43594646	0.20076482	0.6615679	0.3479924	0.27151052
197001	0.51625239	0.15296367	0.2256214	0.2915870	0.07839388

Results



Summary and future work

Summary

- Bridge between CDN's and Liebscher copulas, thus providing ML inference for the former
- Our own model and inference possess good properties

Future work

- Real-world applications
- Compare our model to the literature
- Can we do ML inference for our model?
- Generalize to broader class than CA