# Data Science for Insurance

## Inference for copula models

Roberta Pappadà

a.a. 25-26

rpappada@units.it

# Table of Contents

# Table of Contents

# Dependence and ranks



Suppose that a random sample

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

is given from a pair of continuous rv's $(X, Y)$.

Consider the pairs of ranks

$$(R_1, S_1), \ldots, (R_n, S_n),$$

where $R_i$ stands for the rank of $X_i$ among $X_1, \ldots, X_n$ and $S_i$ stands for the rank of $Y_i$ among $Y_1, \ldots, Y_n$.

# Empirical copula



The empirical copula associated to the random sample $\{(X_i, Y_i)\}_{i=1,\dots,n}$ is defined by

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left( \frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v \right).$$

Notice that $C_n$ is a jump function, and therefore not a copula itself.

If the copula of $(X, Y)$ admits continuous first-order partial derivatives on $(0, 1)^2$

$$\mathbb{C}_n(u, v) = \sqrt{n}(C_n(u, v) - C(u, v)), \quad u, v \in [0, 1]$$

converges weekly as $n \to \infty$ to a centered Gaussian process.

Rüschendorf (1976), Segers (2012)

# Empirical copula

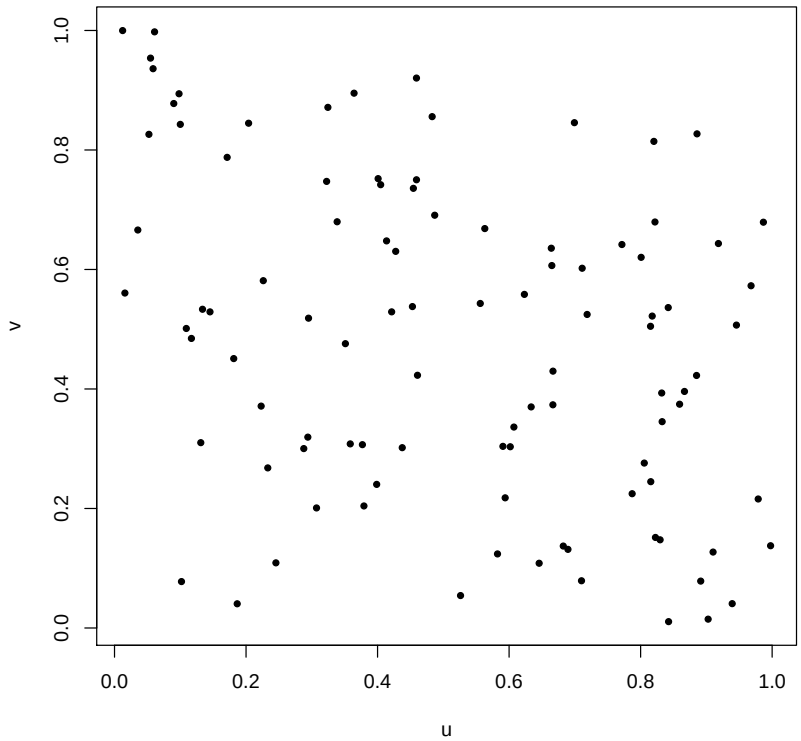


The empirical copula is used in a series of different tasks:

- perform goodness-of-fit testing and to assist in model selection by comparing $C_n$ and an estimated copula $C_{\hat{\theta}}$ fitted to the observations;
- test for some qualitative property, like lack of symmetry by comparing (e.g., by comparing $C_n(u, v)$ and $C_n(v, u)$ for every $u, v$)
- have a starting point for estimation of measures of association by applying the plug-in principle: e.g. $\tau(C_n)$ can be considered as an approximation of $\tau(C)$.

# Illustration





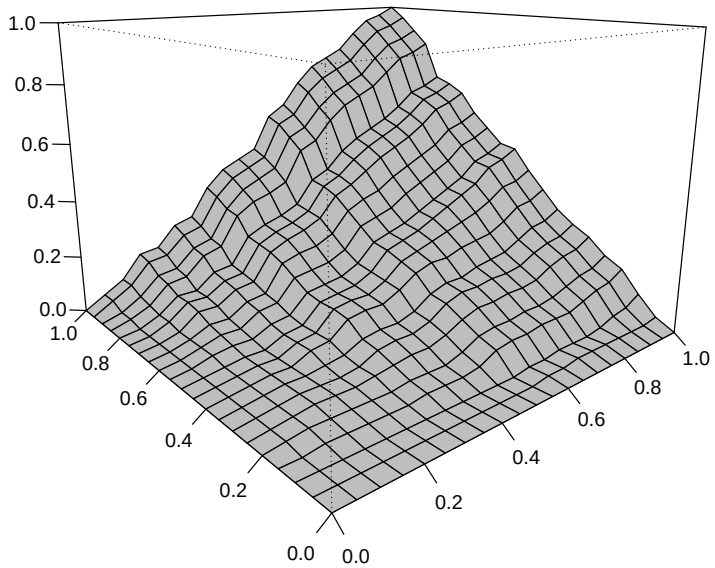




Figure: Scatter plot of $n = 100$ observations drawn from a Gaussian copula $C_\rho$ with $\rho = -0.5$ (left) and corresponding empirical copula $\hat{F}_n$ (right).

# Table of Contents

# Inference for copula models



The process of statistical inference for copula models consists of three main steps:

1. Model selection: descriptive statistics
2. Model fitting: estimation
3. Model validation: selection and goodness-of-fit tests

# Inference for copula models



The **marginal modelling** can be done in three ways:

- fitting parametric distributions to each margin

# Inference for copula models



The **marginal modelling** can be done in three ways:

- fitting parametric distributions to each margin
  $\rightarrow$ inference-functions-for-margins method by Joe (1997)

# Inference for copula models



The **marginal modelling** can be done in three ways:

- fitting parametric distributions to each margin
  $\rightarrow$ inference-functions-for-margins method by Joe (1997)
- modelling the margins nonparametrically using a version of the empirical distribution function

# Inference for copula models



The **marginal modelling** can be done in three ways:

- fitting parametric distributions to each margin
  $\rightarrow$ inference-functions-for-margins method by Joe (1997)
- modelling the margins nonparametrically using a version of the empirical distribution function
  $\rightarrow$ pseudo-likelihood method investigated by Genest et al. (1995);

# Inference for copula models



The **marginal modelling** can be done in three ways:

- fitting parametric distributions to each margin
  $\rightarrow$ inference-functions-for-margins method by Joe (1997)
- modelling the margins nonparametrically using a version of the empirical distribution function
  $\rightarrow$ pseudo-likelihood method investigated by Genest et al. (1995);
- using a hybrid of the parametric and nonparametric methods

# Inference for copula models



The **marginal modelling** can be done in three ways:

- fitting parametric distributions to each margin
  $\rightarrow$ inference-functions-for-margins method by Joe (1997)
- modelling the margins nonparametrically using a version of the empirical distribution function
  $\rightarrow$ pseudo-likelihood method investigated by Genest et al. (1995);
- using a hybrid of the parametric and nonparametric methods (for instance, the tails of the marginal distributions are modelled using a generalized Pareto distribution and the body of the distribution is described by the empirical distribution function)

# Inference for copula models



Let $d = 2$. Assume the marginal dfs have been estimated by one of the methods described above.
Suppose that a parametric family $\{C_\theta\}$ of absolutely continuous copulas is considered as a model for the dependence between $X$ and $Y$.
In order to estimate the copula parameter(s), the following methods can be considered:

- maximum likelihood (ML),
- method of moments (MM)

# Maximum likelihood



The ML method consists in maximizing the so-called *log-likelihood function*

$$\mathcal{L}(\theta, \alpha, \beta) = \sum_{i=1}^{n} \log\left(f_{\alpha,\beta,\theta}(X_i, Y_i)\right),$$

where $f$ is the density of the model, $(\alpha, \beta)$ is the parameter vector of the marginal laws and $\theta$ is the parameter (vector) of the copula with density $c_\theta$. The log–likelihood function can be rewritten as

$$\begin{aligned}\mathcal{L}(\theta, \alpha, \beta) &= \sum_{i=1}^{n} \log\left(c_\theta\left(F_\alpha(X_i), G_\beta(Y_i)\right) f_\alpha(X_i) g_\beta(Y_i)\right) \\ &= \sum_{i=1}^{n} \log f_\alpha(X_i) + \sum_{i=1}^{n} \log g_\beta(Y_i) + \sum_{i=1}^{n} \log c_\theta\left(F_\alpha(X_i), G_\beta(Y_i)\right)\end{aligned}$$

where $F_\alpha$, $f_\alpha$, $G_\beta$, $g_\beta$ are the cdf's and densities of $X$ and $Y$, respectively.

# ML method with known margins



Since $\alpha$ and $\beta$ are assumed to be known, one can apply the functions $F_\alpha(\cdot)$ and $G_\beta(\cdot)$ to the respective coordinate of the sample and obtain the new sample $(F_\alpha(X_i), G_\beta(Y_i)) = (U_i, V_i)$, distributed according to the copula $C_\theta$.

Then, the ML method consists in maximizing

$$\mathcal{L}(\theta, \alpha, \beta) = \sum_{i=1}^{n} \log f_\alpha(X_i) + \sum_{i=1}^{n} \log g_\beta(Y_i) + \sum_{i=1}^{n} \log c_\theta(U_i, V_i)$$

over $\theta$, which is equivalent to only maximizing the last sum over $\theta$.

# Pros and Cons



The ML approach:

- is easy to implement;
- yields an asymptotically Gaussian, unbiased estimate.

However, it may produce wrong estimation of dependence structure if the margins are incorrectly specified.

Moreover, without additional information about the marginal laws, it becomes a problem of joint maximization over the whole parameter space, which often requires the solution of a non-trivial optimization problem.

# Inference for margins (IFM) method



If the inference-functions-for-margins method (IFM) is adopted, the estimation consists of the following two steps:

- marginal laws are estimated separately and the resulting estimates are $\hat{\alpha}_n$ and $\hat{\beta}_n$.
- $\theta$ is obtained by maximization of

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log \left( c_\theta \left( \widehat{F}(X_i), \widehat{G}(Y_i) \right) \right),$$

with $\widehat{F} = F_{\hat{\alpha}_n}$ and $\widehat{G} = G_{\hat{\beta}_n}$

Warning: $(U_i, V_i) = (\hat{F}(X_i), \hat{G}(Y_i))$ is only approximately an iid sample from $C_\theta$.

# Inference for margins (IFM) method



The IFM method:

- is rooted in the decomposition of the joint distribution into marginal laws and copula;
- yields an asymptotically Gaussian, consistent estimate for $\theta$;
- allows reduction of numerical complexity by separate optimizations for each marginal law and the dependence parameters.

As an alternative, the pseudo-likelihood method involves estimation of the margins by the empirical df, to avoid parametric assumptions on the marginal laws (which are only assumed to be continuous).

# The MPL estimation method



Here, the marginal d.f.'s are estimated non-parametrically by their sample empirical distributions, $\hat{F}_n$ and $\hat{G}_n$, and the pseudo-copula data are:

$$(U_i, V_i) = (\hat{F}_n(X_i), \hat{G}_n(Y_i))', \; i = 1, \ldots, n$$

Then, the method of maximum pseudo–likelihood simply involves maximizing over $\theta$ the rank–based log-likelihood of the form

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log\left(c_\theta\left(U_i, V_i\right)\right).$$

This maximization is not particularly easy in higher dimensions. For this reason, the method-of-moments is of practical interest.

# Estimation based on method of moments



In general, suppose that the dependence of $(X, Y)$ is appropriately modelled by a family of copulas $\{C_\theta\}$ with continuous marginals $F_\alpha(x)$, $G_\beta(y)$, such that

$$\theta \to f(\theta) = \kappa$$

for some function $f$ and a copula-based measure of association $\kappa$
Then, the empirical version of this dependence measure can be estimated from the data, yielding the estimate $\hat{\kappa}_n$. Finally,

$$\hat{\theta}_n = f^{-1}(\hat{\kappa}_n)$$

may be refereed to as the estimator of $\theta$ based on inversion of $\kappa$. This procedure is usually adopted by considering $\kappa = \tau$ or $\kappa = \rho$.

# Sample version of Spearman's rho



The sample version of the Spearman's $\rho$ associated to the random sample $\{(X_i, Y_i)\}_{i=1,\dots,n}$ is defined by

$$\hat{\rho}_n = \frac{12}{(n+1)(n-1)} \sum_{i=1}^{n} R_i S_i - 3\frac{n+1}{n-1}$$

When the random sample comes from the copula $\Pi_2$, one has

$$\hat{\rho}_n \dot{\sim} \mathcal{N}(0, 1/(n-1)).$$

This is used to test the null hypothesis of independence.

# Sample version of Kendall's tau



The sample version of the Kendall's $\tau$ associated to $\{(X_i, Y_i)\}_{i=1,\dots,n}$ is defined by

$$\hat{\tau}_n = \frac{4}{n(n-1)} P_n - 1,$$

where $P_n$ is the number of concordant pairs.

When the random sample comes from the copula $\Pi_2$, one has

$$\tau_n \dot{\sim} \mathcal{N}(0, 2(2n+5)/9n(n-1)).$$

This is used to test the null hypothesis of independence.

# MM estimation: FGM copula



Suppose that the dependence of $(X, Y)$ is appropriately modelled by the FGM family of copulas given by

$$C(u, v) = uv + \theta uv(1 - u)(1 - v).$$

It is known that, for this family, the value of Kendall's tau is given by

$$\tau_C = \frac{2\theta}{9}.$$

Therefore, an appropriate estimation of $\theta$ is

$$\hat{\theta} = \frac{9\hat{\tau}_n}{2},$$

where $\hat{\tau}_n$ is the sample version of the Kendall's tau.

# MM estimation: Gumbel copula



Consider the Gumbel family

$$C(u,v) = \exp\left[-((-\log u)^\theta + (-\log v)^\theta)\right], \theta \in [1, +\infty)$$

The theoretical Kendall's $\tau$ is

$$f(\theta) := (\theta - 1)/\theta$$

Given a set of data we thus compute $\hat{\tau}_n$ and use it to estimate $\theta$:

$$\hat{\theta}_n = f^{-1}(\hat{\tau}_n) = \frac{1}{1 - \hat{\tau}_n}$$

# Pros and Cons



The method of moments presents some advantages:

- is easy to implement;
- has well-studied asymptotic properties;
- can be applied to arbitrary parametric family of copulas;

On the other hand, it is based on the assumption that Kendall's $\tau$ and Spearman's $\rho$ are known functions of the parameters.

In the case of a two-dimensional parameter space the method has to be adapted!

# Graphical Diagnostics and Tests

When dealing with bivariate data, the most intuitive way of checking the adequacy of the copula $C_\theta$ with respect to our model would be to compare the scatter plot (or contour-plot) of the pseudo-observations with a sample generated from the copula $C_\theta$

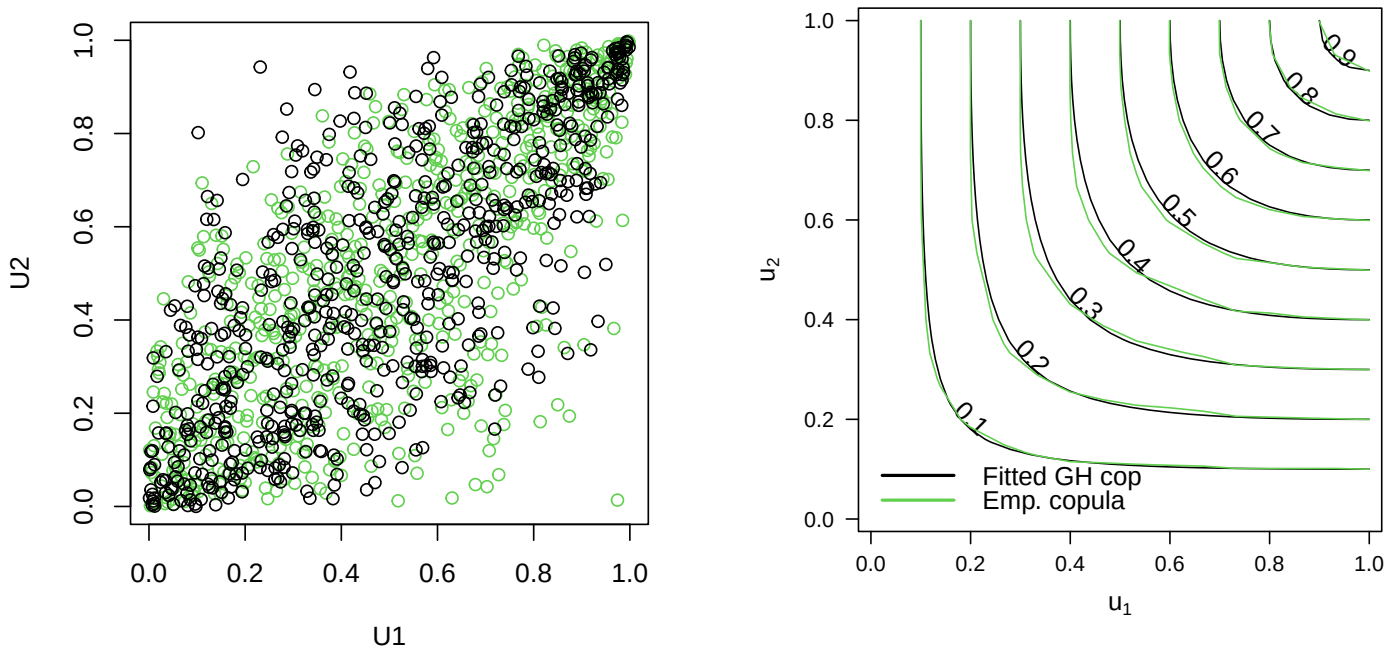



Figure: Scatter plot (left) and contour plot (right) of the fitted Khoudraji-Gumbel-Hougaard copula compared to those of the empirical copula of the data (green)

# Goodness–of–fit tests

In typical modeling, the user has a choice between several different dependence structures for the data at hand.

Suppose that different copulas $\{C_1, \ldots, C_k\}$, belonging to different families $\{C_\theta^1, \ldots, C_\theta^k\}$, were fitted by some arbitrary method. It is then natural to ask

*what is the copula that best fits the observations?*

To answer the question, we may perform a formal statistical test of type

$$H_0 \colon C \in (C_\theta),$$

which gives as output a *p*–value $p$ that, if small, may be interpreted as empirical evidence against the null hypothesis.

# GOF Tests based on the empirical copula process

The empirical copula is a consistent estimator of the unknown copula $C$ whether $H_0$ is true or not.

Hence, a natural goodness-of-fit test consists of comparing Cn with an estimate $C_{\hat{\theta}_n}$ of $C$ (computed from the pseudo-observations) obtained under the assumption that $C \in (C_\theta)$ holds

Most GOF statistics are based on distances between $C_n$ and the true copula $C_0$ or between $C_n$ and $C_{\hat{\theta}_n}$. The one based on the Cramér–von Mises distance is

$$S_n = \sum_{i=1}^{n} \left( C_n(U_i, V_i) - C_{\hat{\theta}_n}(U_i, V_i) \right)^2.$$

see Genest et al. (2009)

# GOF Tests based on the empirical copula process

For testing for independence in $d$ dimensions, $C_0 = \Pi$ and the test statisic is

$$S_n^{\Pi} = \int_{[0,1]^d} n(C(\mathbf{u}) - \Pi(\mathbf{u}))^2 \mathrm{d}\mathbf{u}$$

This test is implemented in the R copula package (see Genest and Rémillard (2008))

Several tools can guide the choice of an appropriate copula family:

- Tests of Exchangeability/Radial Symmetry
- Tests of Extreme-Value Dependence
- Goodness-of-fit tests based on pairwise Rosenblatt transforms
- Information Criteria (e.g. AIC) and cross-validation

See Hofert et al. (2018) for more details.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Genest, C. and Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques*, 44:1096–1127.

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2018). *Elements of Copula Modeling with R*. Springer Use R! Series.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC press.