Dependence modeling through copulas

Christian Genest & Johanna G. Nešlehová

McGill University, Montréal, Canada

ROBUST, September 11, 2016

Prelude

WIRED MAGAZINE: 17.03

Recipe for Disaster: The Formula That Killed Wall Street

By Felix Salmon   02.23.09

In the mid-'80s, Wall Street turned to the quants—brainy financial engineers—to invent new ways to boost profits. Their methods for minting money worked brilliantly... until one of them devastated the global economy.
Photo: Jim Krantz/Gallery Stock

Road Map for Financial Recovery: Radical Transparency Now!

A year ago, it was hardly unthinkable that a math wizard like David X. Li might someday earn a Nobel Prize. After all, financial economists—even Wall Street quants—have received the Nobel in economics before, and Li's work on measuring risk has had more impact, more quickly, than previous Nobel Prize-winning contributions to the field. Today, though, as dazed bankers, politicians, regulators, and investors survey the wreckage of the biggest financial meltdown since the Great Depression, Li is probably thankful he still has a job in finance at all. Not that his achievement should be dismissed. He took a notoriously tough nut—determining correlation, or how seemingly disparate events are related—and cracked it wide open with a simple and elegant mathematical formula, one that would become ubiquitous in finance worldwide.

Prelude

$$\Pr[T_A<1, T_B<1] = \phi_2(\phi^{-1}(F_A(1)), \phi^{-1}(F_B(1)), \gamma)$$

Here's what killed your 401(k) *David X. Li's Gaussian copula function as first published in 2000. Investors exploited it as a quick—and fatally flawed—way to assess risk. A shorter version appears on this month's cover of* Wired.

Probability
Specifically, this is a joint default probability—the likelihood that any two members of the pool (A and B) will both default. It's what investors are looking for, and the rest of the formula provides the answer.

Survival times
The amount of time between now and when A and B can be expected to default. Li took the idea from a concept in actuarial science that charts what happens to someone's life expectancy when their spouse dies.

Equality
A dangerously precise concept, since it leaves no room for error. Clean equations help both quants and their managers forget that the real world contains a surprising amount of uncertainty, fuzziness, and precariousness.

Copula
This couples (hence the Latinate term copula) the individual probabilities associated with A and B to come up with a single number. Errors here massively increase the risk of the whole equation blowing up.

Distribution functions
The probabilities of how long A and B are likely to survive. Since these are not certainties, they can be dangerous: Small miscalculations may leave you facing much more risk than the formula indicates.

Gamma
The all-powerful correlation parameter, which reduces correlation to a single constant—something that should be highly improbable, if not impossible. This is the magic number that made Li's copula function irresistible.

Prelude

PROFILE
Financial Meltdown

Was David Li the guy who 'blew up Wall Street?'

thespec.com http://www.thespec.com/News/Break

Canadian scholar scapegoat for global meltdown

Math whiz proposed applying this statistical formula to credit risk, and financial meltdown

NZZ Online

Donnerstag, 19. März 2009, 10:56:37 Uhr, NZZ Online

Nachrichten › Forschung und Technik

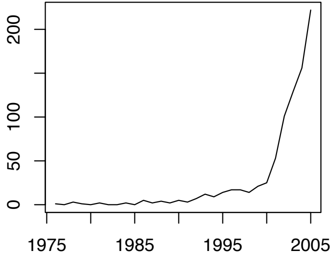
18. März 2009, Neue Zürcher Zeitung

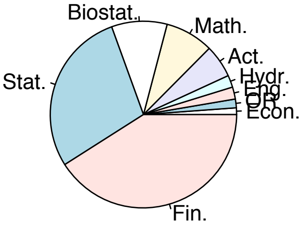
Eine falsch angewendete Formel und ihre Folgen

Unterschätzte Korrelation von Anlagewerten als Auslöser der Finanzkrise?

Prelude

The copula wave (1986–2005)





Source: Genest et al. (2009)

MathSciNet (2006–2014): 53, 53, 87, 106, 98, 120, 124, 159, 138.

Outline

1. Copulas
2. Copula models
3. Modeling strategies
4. Current issues

1. Copulas: Definition

What is a copula?

The joint distribution function of a random pair (U, V), where

$$U \sim \mathcal{U}(0,1), \quad V \sim \mathcal{U}(0,1).$$

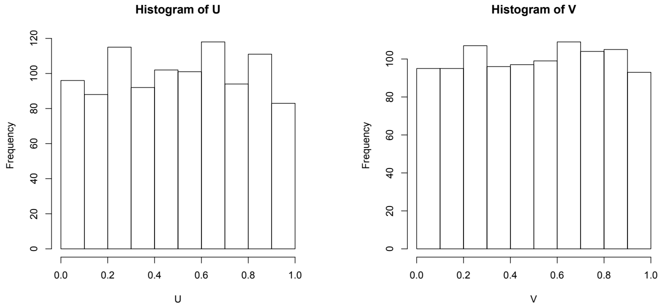
For all $u, v \in (0,1)$,

$$\begin{aligned} \Pr(U \le u) &= u, \\ \Pr(V \le v) &= v, \\ \Pr(U \le u, V \le v) &= C(u,v). \end{aligned}$$

The copula C describes how U and V vary together.

1. Copulas: Illustration

Draw 1000 observations from U and V independently:

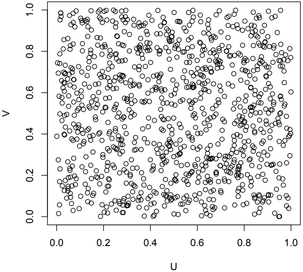


Theoretical model for U (or V):

$$F(u) = \Pr(U \leq u) = u, \quad \frac{d}{du} F(u) = f(u) = 1.$$

1. Copulas: Illustration

Plot the pairs (U, V):



Theoretical representation of the joint behavior of U and V:

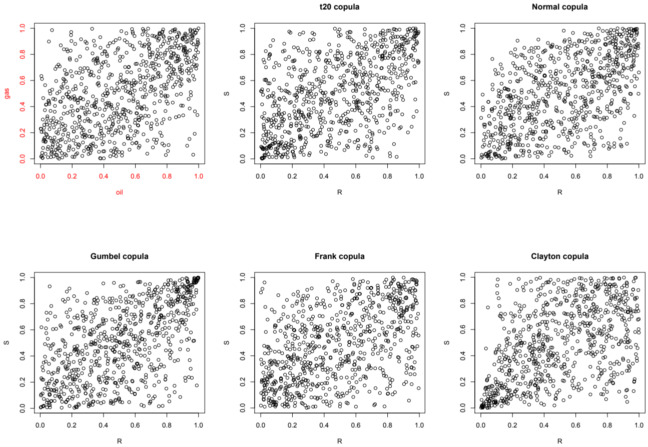
$$C(u, v) = \Pr(U \leq u, V \leq v) = \Pr(U \leq u) \Pr(V \leq v) = uv$$

and

$$\frac{\partial^2}{\partial u \partial v} C(u, v) = 1.$$

1. Copulas

Other sampling mechanisms



Analytical example

For arbitrary $\theta \in [-1, 1]$, set

$$C_\theta(u, v) = uv + \theta uv(1 - u)(1 - v), \quad u, v \in (0, 1).$$

This is the Farlie–Gumbel–Morgenstern *parametric* copula family.

Case $\theta = 0$:

$$C_\perp(u, v) = uv \quad \Leftrightarrow \quad U \perp V.$$

Other examples (there are book treatments of this)

- Archimedean copulas
- Elliptical copulas
- Extreme-value copulas
- Vine copulas

1. Copulas

Innumerable applications, e.g., in finance and insurance

✓ Analysis of CDOs	✓ Annuity valuation
✓ Asset modelling	✓ Capital assessment
✓ Credit risk	✓ Joint life tables
✓ Option pricing	✓ Portfolio modelling
✓ Premium calculations	✓ Reinsurance contracts
✓ Risk aggregation	✓ Risk assessment

Why is it so?

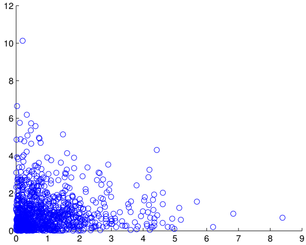
Because copulas...

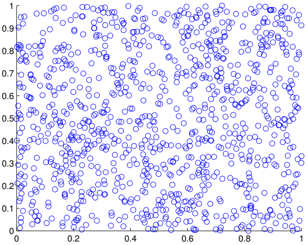
- reveal the true nature of dependence between variables;
- lead to flexible multivariate stochastic models.

Most existing models assume margins of the same form.

1. Copulas

“Reveal dependence”

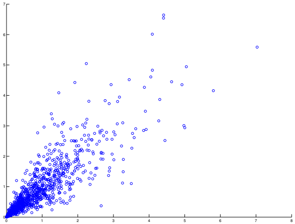


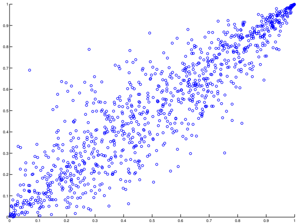


$$X, Y \sim F = \mathcal{E}(1) \quad \Rightarrow \quad (U, V) = (F(X), F(Y))$$

1. Copulas

"Reveal dependence"





$$X, Y \sim F = \mathcal{E}(1) \quad \Rightarrow \quad (U, V) = (F(X), F(Y))$$

Characterize dependence

Sklar (1959) showed that when H is continuous,

- one can always write

$$H(x, y) = C\{F(x), G(y)\}, \quad x, y \in \mathbb{R};$$

- C is unique.

The case where H can have jumps is trickier.

1. Copulas

Flexible models

If C is a copula and F, G are univariate distribution functions, then

$$H(x,y) = C\{F(x), G(y)\}, \quad x,y \in \mathbb{R}$$

is a joint distribution for the pair (X,Y) with margins F, G, viz,

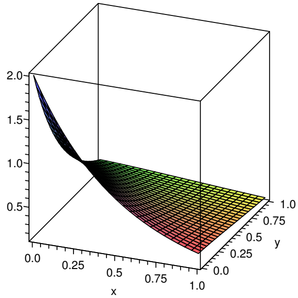
$$\Pr(X \le x, Y \le y) = C\{\Pr(X \le x), \Pr(Y \le y)\}.$$

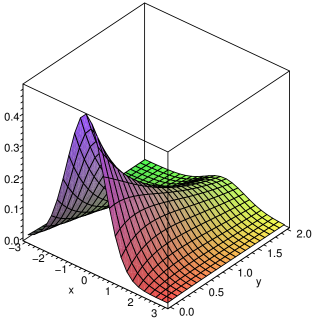
Ex.: The FGM model

$$H(x,y) = F(x)G(y) + \theta F(x)G(y)\{1 - F(x)\}\{1 - G(y)\}, \; x,y \in \mathbb{R}.$$

1. Copulas

Two FGM distributions

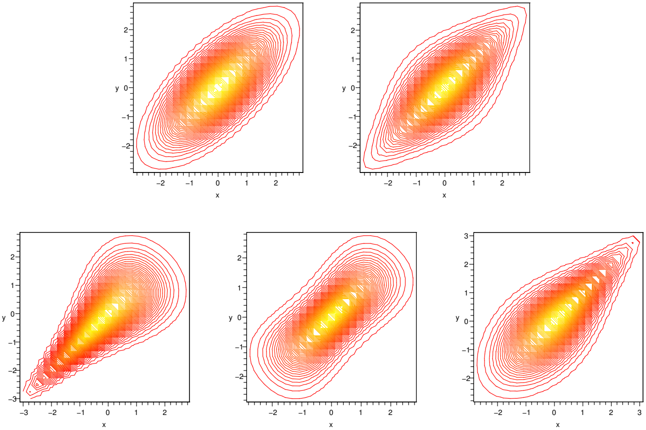




$X, Y \sim \mathcal{E}(1), \qquad X \sim \mathcal{N}(0,1), Y \sim \mathcal{E}(1)$

1. Copulas

$$X \sim \mathcal{N}(0, 1),\ Y \sim \mathcal{N}(0, 1),\ C = ?$$



2. Copula models

Definition

Regression model:

$$\mathrm{E}(Y|X = x) = \alpha + \beta x + \epsilon.$$

Copula model:

$$H(x, y) = C\{F(x), G(y)\}, \quad x, y \in \mathbb{R}$$

with

$$C \in \mathcal{C}_\theta, \quad F \in \mathcal{F}_\alpha, \quad G \in \mathcal{G}_\beta.$$

2. Copula models

Advantages

In the model

$$H(x, y) = C_\theta\{F_\alpha(x), G_\beta(y)\},$$

- F_α, G_β are arbitrary;
- can involve covariates;
- C_θ, F_α, G_β can be treated separately.

Illustration: Grégoire et al. (2008)

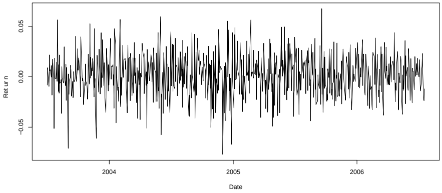
Variables:

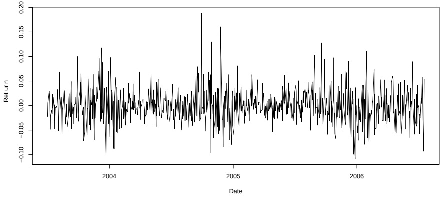
- Price of Oil (Light Sweet Crude)
- Price of Natural Gas (mmBTU)

Period: 2004–01–01 $\rightarrow$ 2006–08–31

Margins: GARCH

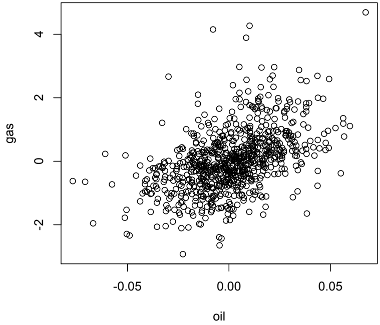
2. Copula models





Returns (2003–2006)

2. Copula models



Residuals

Can one see the copula?

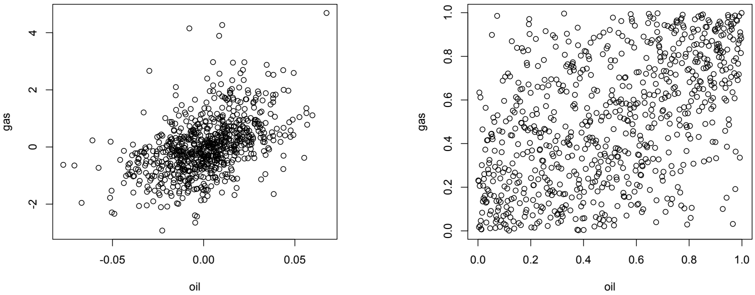
Estimate the margins by

$$F_n(x) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x), \quad G_n(y) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}(Y_i \leq y).$$

Plot the pairs

$$(\hat{U}_i, \hat{V}_i) = (F_n(X_i), G_n(Y_i)) = \left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right).$$

2. Copula models



Empirical copula (2003–2006)
Source: Grégoire et al. (2008)

2. Copula models

Empirical copula

For all $u, v \in [0, 1]$,

$$\hat{C}_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} 1(\hat{U}_i \leq u, \hat{V}_i \leq v),$$

Idea: Rüschendorf (1976), Deheuvels (1979)

Convergence: Fermanian et al. (2004), Segers (2012), etc.

For the discrete case: Genest, Nešlehová & Rémillard (2016)

Asymptotic result

If the copula C is "regular," then, as $n \to \infty$,

$$\hat{\mathbb{C}}_n = \sqrt{n}\,(\hat{C}_n - C) \rightsquigarrow \hat{\mathbb{C}},$$

where

$$\hat{\mathbb{C}}(u,v) = \mathbb{C}(u,v) - \frac{\partial C(u,v)}{\partial u}\,\mathbb{C}(u,1) - \frac{\partial C(u,v)}{\partial v}\,\mathbb{C}(1,v),$$

with $\mathbb{C}$ = Brownian sheet. In short, if n is large, then $\hat{C}_n$ is a good approximation of C, viz.

$$\hat{C}_n \approx C.$$

3. Inference

Steps to consider

3A Model selection

3B Model fitting

3C Model validation

Once a model has been validated, it can be used for prediction.

3A. Descriptive statistics

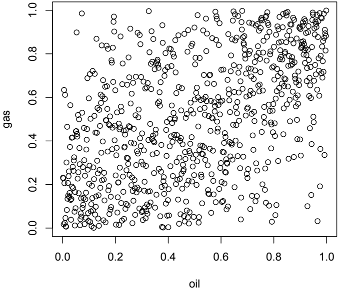
Doing what comes naturally

Compute the correlation ρ_n using the pairs

$$(\hat{U}_1, \hat{V}_1), \ldots, (\hat{U}_n, \hat{V}_n).$$

This is in contrast to the standard Pearson correlation, based on the pairs

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



3A. Descriptive statistics

Spearman's rho

When $n \to \infty$,

$$\begin{aligned} \rho_n &= \text{rank correlation} \\ &\to \rho = -3 + 12 \int_0^1 \int_0^1 C(u, v) \mathrm{d}v \mathrm{d}u \\ &= \mathrm{corr}\{F(X), G(Y)\}. \end{aligned}$$

Under $\mathcal{H}_0 : C = C_\perp$,

$$\rho_n \approx \mathcal{N}\left(0, \frac{1}{n-1}\right).$$

3A. Descriptive statistics

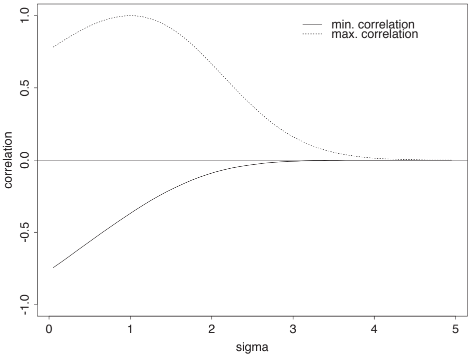
Pearson's correlation

This traditional measure of dependence has many faults:

- $\text{corr}(X, Y)$ measures linear dependence;
- is margin dependent;
- can be quite misleading;
- may not exist (e.g., Cauchy).

3A. Descriptive statistics

Bounds on corr(X, Y)



$$X \sim \mathrm{LN}(0,1), Y \sim \mathrm{LN}(0,\sigma^2)$$

The empirical copula: A unifying concept

- Linear functionals of $\hat{\mathbb{C}}_n = \sqrt{n}\,(\hat{C}_n - C)$ are Gaussian.
- Margins are nuisance parameters.
- Pairs of ranks are maximally invariant.

3A. Descriptive statistics

Example: Consistent tests of independence

A test of independence is designed to check whether

$$\mathcal{H}_0 : C = C_\perp$$

Consistent tests of this hypothesis can be based on

$$\hat{\mathbb{C}}_n = \sqrt{n}\,(\hat{C}_n - C_\perp)$$

and its Möbius decomposition.

See, e.g., Genest & Rémillard (2004).

3B. Parametric estimation

Suppose you believe that $C = C_\theta$ for some $\theta \in \Theta$. How can you estimate θ?

A popular solution (among others)

Maximize

$$\ell(\theta) = \sum_{i=1}^{n} \ln[c_\theta\{F_n(X_i), G_n(Y_i)\}]$$
$$= \sum_{i=1}^{n} \ln\{c_\theta(\hat{U}_i, \hat{V}_i)\},$$

called the pseudo-likelihood.

E1. Parametric likelihood inference

Log pseudo-likelihood

Under the assumption that $C \in \mathcal{C} = (C_\theta)$, the equation to solve is

$$\begin{aligned}
\dot{\ell}(\theta) &= \frac{\partial}{\partial\theta}\,\ell(\theta) \\
&= \sum_{i=1}^{n} \dot{c}_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) \Big/ c_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) \\
&= 0.
\end{aligned}$$

E1. Parametric likelihood inference

Example: FGM copula

Suppose that $C_\theta(u, v) = uv + \theta uv(1-u)(1-v)$.

In this case,

$$c_\theta(u, v) = 1 + \theta(1-2u)(1-2v)$$

and

$$\frac{\dot{c}_\theta(u, v)}{c_\theta(u, v)} = \frac{(1-2u)(1-2v)}{1 + \theta(1-2u)(1-2v)}$$

for all $u, v \in (0, 1)$.

E1. Parametric likelihood inference

FGM log pseudo-likelihood

It is given by

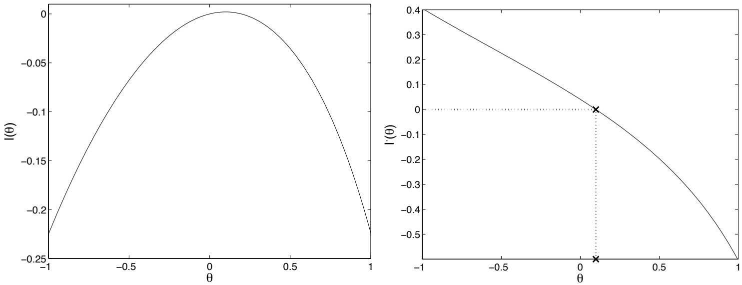
$$\ell(\theta) = \sum_{i=1}^{n} \ln \left\{ 1 + \theta \left(1 - \frac{2R_i}{n+1}\right)\left(1 - \frac{2S_i}{n+1}\right) \right\}$$

and the corresponding pseudo-score function is

$$\begin{aligned} \dot{\ell}(\theta) &= \sum_{i=1}^{n} \frac{\left(1 - 2\frac{R_i}{n+1}\right)\left(1 - 2\frac{S_i}{n+1}\right)}{1 + \theta\left(1 - 2\frac{R_i}{n+1}\right)\left(1 - 2\frac{S_i}{n+1}\right)} \\ &= \sum_{i=1}^{n} \frac{(n+1-2R_i)(n+1-2S_i)}{(n+1)^2 + \theta(n+1-2R_i)(n+1-2S_i)}. \end{aligned}$$

E1. Parametric likelihood inference

FGM log pseudo-likelihood (left) and score function (right)



3B. Parametric estimation

Asymptotic results

It can be shown that the maximum pseudo-likelihod estimator

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{\nu^2}{n}\right)$$

and one can also get an estimate of ν^2.

This procedure is

- margin-free;
- efficient near independence.

3B. Parametric estimation

Initial values

When $\theta \in \mathbb{R}$, classical nonparametric measures of dependence are often increasing in θ, e.g.

- ✓ Spearman's rho: $\rho = \psi(\theta)$;
- ✓ Kendall's tau: $\tau = \phi(\theta)$.

Simple moment-like estimates can then be obtained by inversion:

$$\breve{\theta}_n = \phi^{-1}(\tau_n), \quad \breve{\theta}_n = \psi^{-1}(\rho_n).$$

3B. Parametric estimation

Spearman's rho

$$\rho(X, Y) = -3 + 12 \int_0^1 \int_0^1 C(u, v) \mathrm{d}v \mathrm{d}u = \mathrm{corr}\{F(X), G(Y)\}.$$

Kendall's tau

$$\begin{aligned} \tau(X, Y) &= -1 + 4 \int_0^1 \int_0^1 C(u, v) \mathrm{d}C(u, v) \\ &= -1 + 4 \Pr(X_1 < X_2, Y_1 < Y_2) \end{aligned}$$

where (X_1, Y_1), (X_2, Y_2) are independent copies of (X, Y).

3B. Parametric estimation

Illustration

In the FGM model,

$$\rho = \theta/3, \quad \tau = 2\,\theta/9.$$

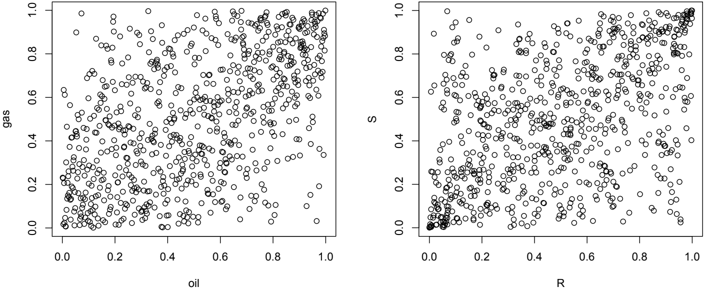
$\theta \in [-1, 1]$ can be estimated by

$$\check{\theta}_n = 3\,\rho_n, \quad \check{\theta}_n = 9\,\tau_n/2.$$

In general, consistency and asymptotic normality follow from U-statistic theory; see, e.g., Lee (1990).

3B. Estimation

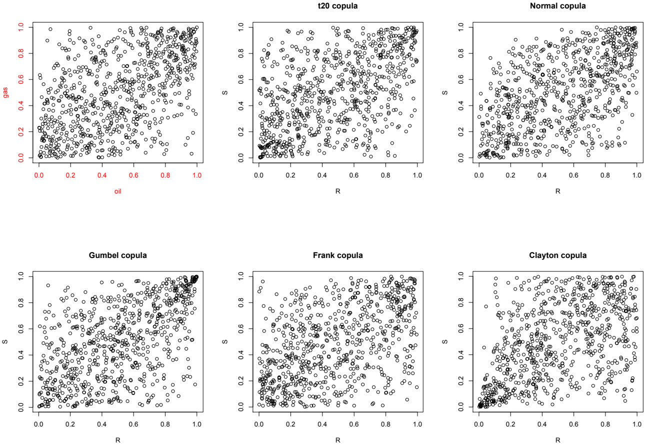
Illustration with oil-and-gas data



Fitting a $t_{(20)}$ copula using $\rho_n \approx 0.522$.

3C. Goodness-of-fit tests

Best model?



3C. Goodness-of-fit tests

General approach

- Assume $\mathcal{H}_0 : C \in C_\theta$.
- Compute θ_n: estimate of θ.
- Measure a "distance"

$$D_n = \mathcal{D}_n(\hat{C}_n, C_{\theta_n}).$$

- Find the distribution of D_n or approximate it as $n \to \infty$.
- Compute a p-value or an approximation thereof:

$$p = \Pr(\mathcal{D}_n > D_n | \mathcal{H}_0).$$

3C. Goodness-of-fit tests

Common approach

$$D_n = n \int_0^1 \int_0^1 \{C_{\theta_n}(u,v) - \hat{C}_n(u,v)\}^2 \mathrm{d}\hat{C}_n(u,v)$$

$$= \sum_{i=1}^{n} \{C_{\theta_n}(\hat{U}_i, \hat{V}_i) - \hat{C}_n(\hat{U}_i, \hat{V}_i)\}^2.$$

Other popular choice

$$T_n = \max_{1 \le i \le n} |C_{\theta_n}(\hat{U}_i, \hat{V}_i) - \hat{C}_n(\hat{U}_i, \hat{V}_i)|.$$

3C. Goodness-of-fit testing

Complication

The distribution of

$$D_n = n \int_0^1 \int_0^1 \{C_{\theta_n}(u, v) - \hat{C}_n(u, v)\}^2 \mathrm{d}\hat{C}_n(u, v)$$

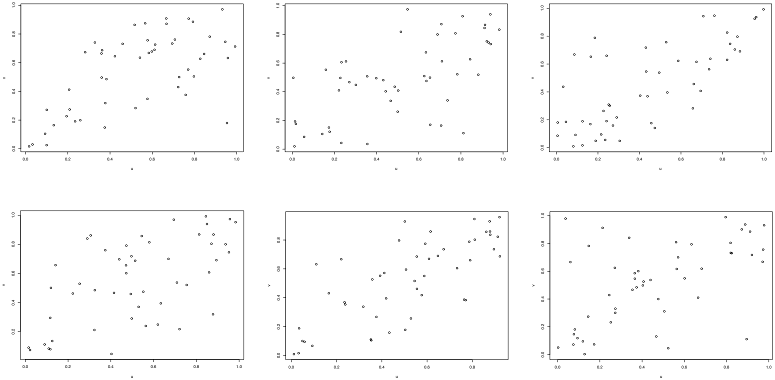
depends on the unknown value of θ_0.

Options

- Parametric bootstrap
- Multiplier method

3C. Goodness-of-fit testing

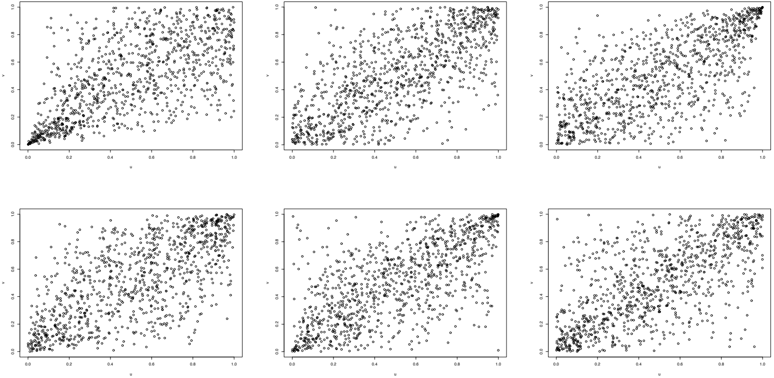
Warning



Clayton, Frank, Gumbel $\qquad n = 50$
Gaussian, Student (4), Plackett $\qquad \tau = 1/2$

3C. Goodness-of-fit testing

Warning

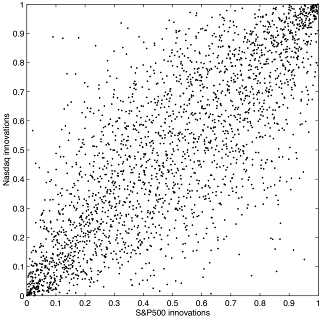


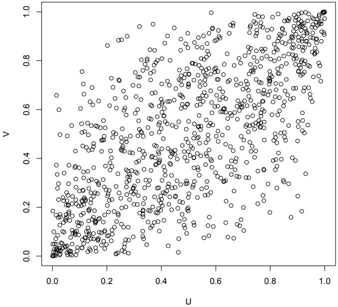
Clayton, Frank, Gumbel $\quad n = 100$

Gaussian, Student (4), Plackett $\quad \tau = 1/2$

3. Goodness-of-fit testing

"The Formula that killed Wall Street?"





4. Current issues

There are many, hence our presence here!

New challenges arise when data are...

- multivariate;
- discontinuous;
- incomplete;
- time series;
- dependent on covariates;
- etc.

Dealing with large-dimensional problems

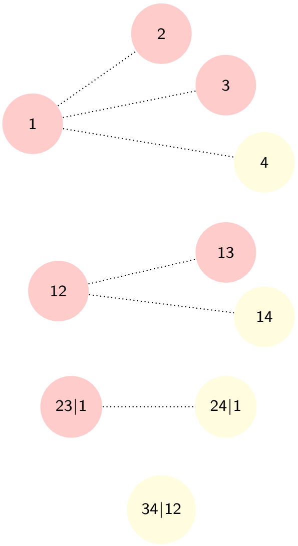
Idea: Proceed by successive conditionings

Advantages:

- There is already a rich class of bivariate copulas.
- Complete flexibility and no compatibility issues.

Ref.: Joe (1997), Bedford & Cooke (2002),
Aas et al. (2009), Kurowicka & Joe (2011)

4. Current issues



C-Vine

$$f(x_1, x_2, x_3, x_4) =$$

$$f_1(x_1)\ f_2(x_2)\ f_3(x_3)\ f_4(x_4)$$

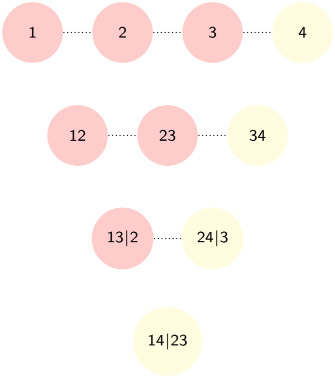
$$c_{12}\{F_1(x_1), F_2(x_2)\}$$
$$c_{13}\{F_1(x_1), F_3(x_3)\}$$
$$c_{14}\{F_1(x_1), F_4(x_4)\}$$

$$c_{23|1}\{F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1) \mid x_1\}$$
$$c_{24|1}\{F_{2|1}(x_2|x_1), F_{4|1}(x_4|x_1) \mid x_1\}$$

$$c_{34|12}\{F_{3|12}(x_3|x_1, x_2), F_{4|12}(x_4|x_1, x_2) \mid x_1, x_2\}$$

4. Current issues

D-Vine



$$f(x_1, x_2, x_3, x_4) =$$

$$f_1(x_1)\ f_2(x_2)\ f_3(x_3)\ f_4(x_4)$$

$$c_{12}\{F_1(x_1), F_2(x_2)\}$$
$$c_{23}\{F_2(x_2), F_3(x_3)\}$$
$$c_{34}\{F_3(x_3), F_4(x_4)\}$$

$$c_{13|2}\{F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2) \mid x_2\}$$
$$c_{24|3}\{F_{2|3}(x_2|x_3), F_{4|3}(x_4|x_3) \mid x_3\}$$

$$c_{14|23}\{F_{1|23}(x_1|x_2, x_3), F_{4|23}(x_4|x_2, x_3) \mid x_2, x_3\}$$

4. Current issues

Looking for models with interpretation

Models with an underlying structure, such as “factor models.”

For example, McNeil & Nešlehová (2009) suppose that

$$X = RAU$$

with $A = I_d$, $R > 0$ being independent from U, uniformly distributed on

$$\{(s_1, \ldots, s_d) \geq 0 : s_1 + \cdots + s_d = 1\}.$$

This leads to Archimedean copulas, viz.

$$C(u_1, \ldots, u_d) = \psi\{\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)\}.$$

Dealing with extreme events

Extreme-value copulas are used to predict the frequency and size of catastrophic events due to simultaneous events:

✓ Floods	✓ Droughts
✓ Hurricanes	✓ Heat waves, etc.

A bivariate extreme-value copula looks like this:

$$C(u, v) = \exp\left[\ln(uv) A\left\{\frac{\ln(v)}{\ln(uv)}\right\}\right]$$

where $A : [0, 1] \to [1/2, 1]$ must satisfy specific conditions.

4. Current issues

The Saguenay Flood: July 19, 1996

100 mm of rain in two days
1.5 billion dollars in damage in Chicoutimi and surroundings