Everything you always wanted to know about copula modelling but were afraid to ask

Christian Genest, PhD, PStat

McGill University, Montréal, Canada

NIPS 2011, Sierra Nevada, Spain

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

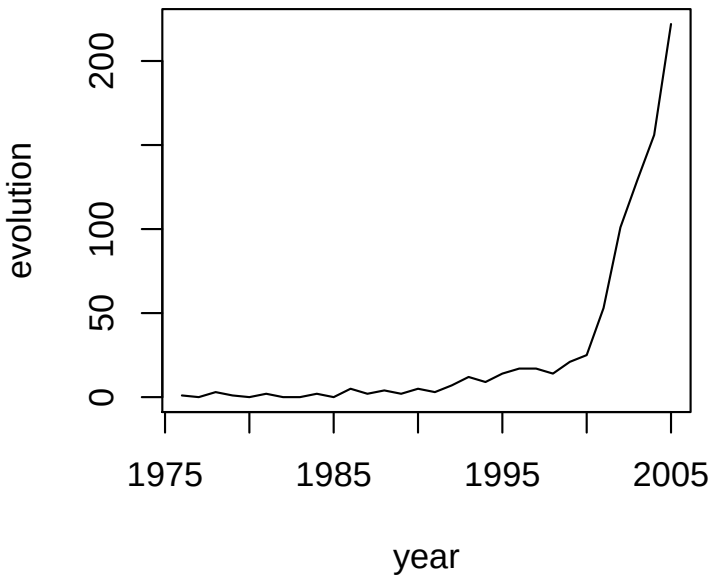
18. März 2009, Neue Zürcher Zeitung

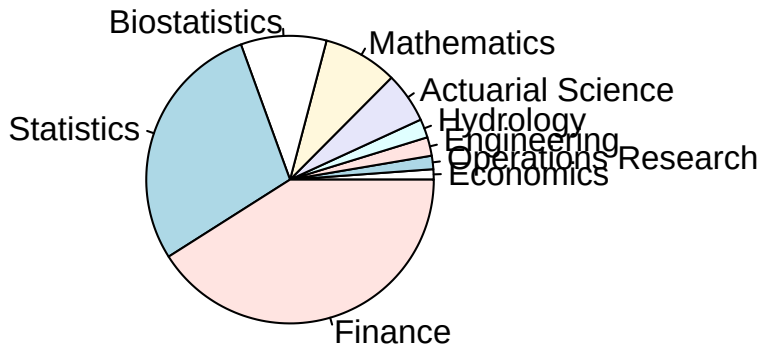
Eine falsch angewendete Formel und ihre Folgen

Unterschätzte Korrelation von Anlagewerten als Auslöser der Finanzkrise?

Prelude

The spread of copulas in data sciences: 1976–2005





Source: Genest et al. (2009)

MathSciNet counts, 2006–2010: 53, 53, 87, 106, 92

Outline

1. Copulas
2. Copula models
3. Inference for copula models
4. Strategies for constructing copula models

1. Copulas

What is a copula?

A copula is the distribution function of a pair (U, V), where

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

Thus if

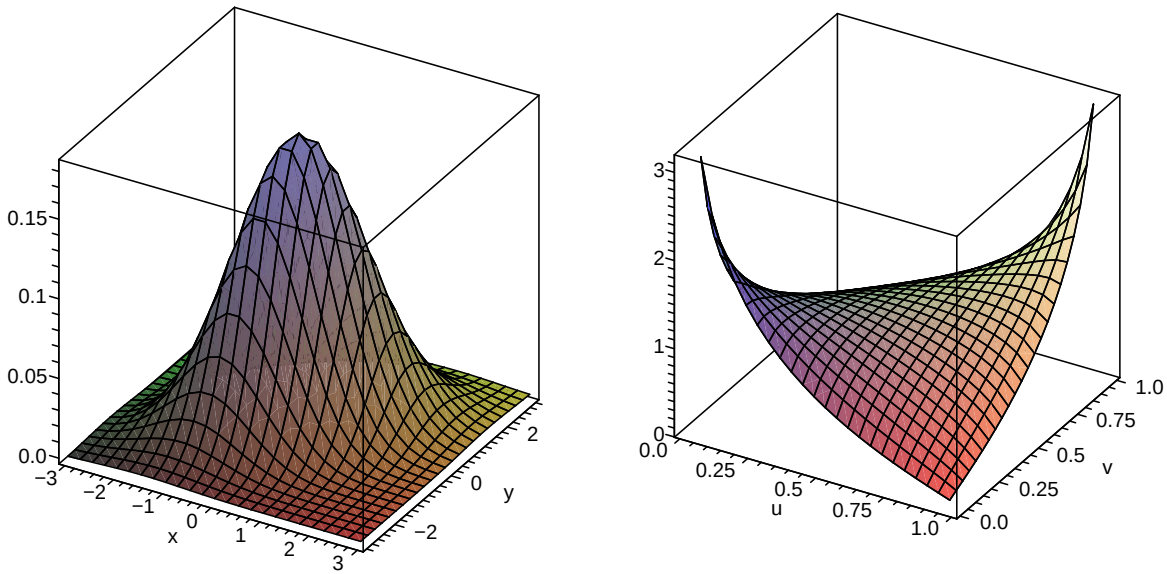
$$(X, Y) \sim H, \quad X \sim F, \quad Y \sim G,$$

and if H is continuous, then

$$(U, V) \equiv (F(X), G(Y)) \sim C\text{: a copula.}$$

1. Copulas

Example 1: $(X, Y) \sim \mathcal{N}_r(0, 1)$

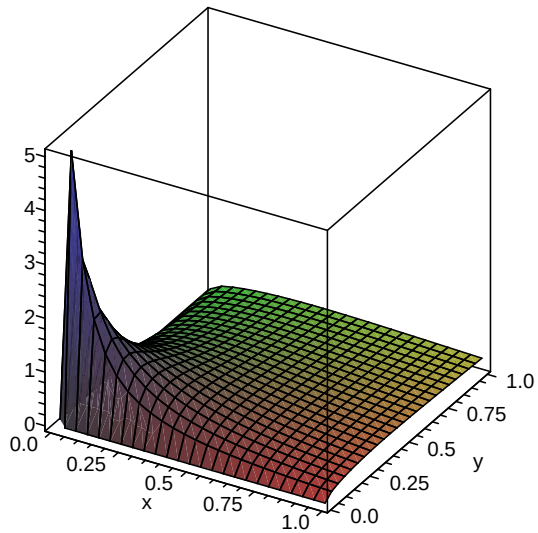


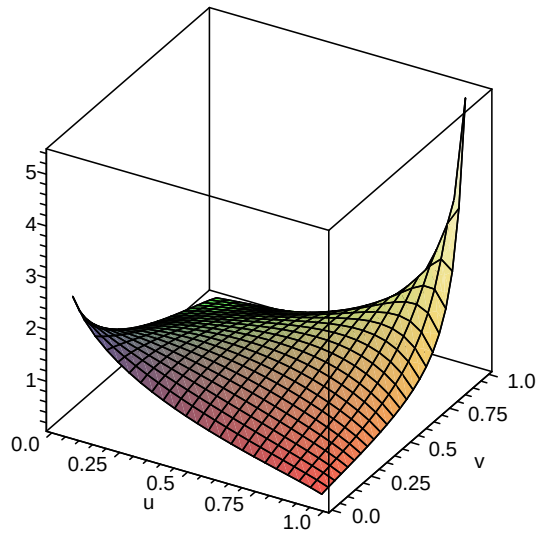
$$(X, Y) \mapsto (U, V) = (\Phi(X), \Phi(Y))$$

Bivariate Gaussian density and associated copula with $r = 0.5$

1. Copulas

Example 2: Gumbel (1960)





$$(X, Y) \mapsto (U, V) = (1 - e^{-X}, 1 - e^{-Y})$$

Bivariate exponential density and copula with $\theta = 1.5$

Example 3: An analytic example of a copula

A "toy example" of copula family is given by

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

The special case $\theta = 0$ corresponds to independence, viz.

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

Other common examples include Archimedean, extreme-value, meta-elliptical, and vine copulas.

1. Copulas

Main references for lists of copulas

- H. Joe (1997).
  Multivariate Models and Dependence Concepts.
  Chapman & Hall, London.

- R. B. Nelsen (1999, 2006).
  An Introduction to Copulas.
  Springer, New York.

- D. Drouet-Mari & S. Kotz (2001).
  Correlation and Dependence.
  Imperial College Press, London.

1. Copulas

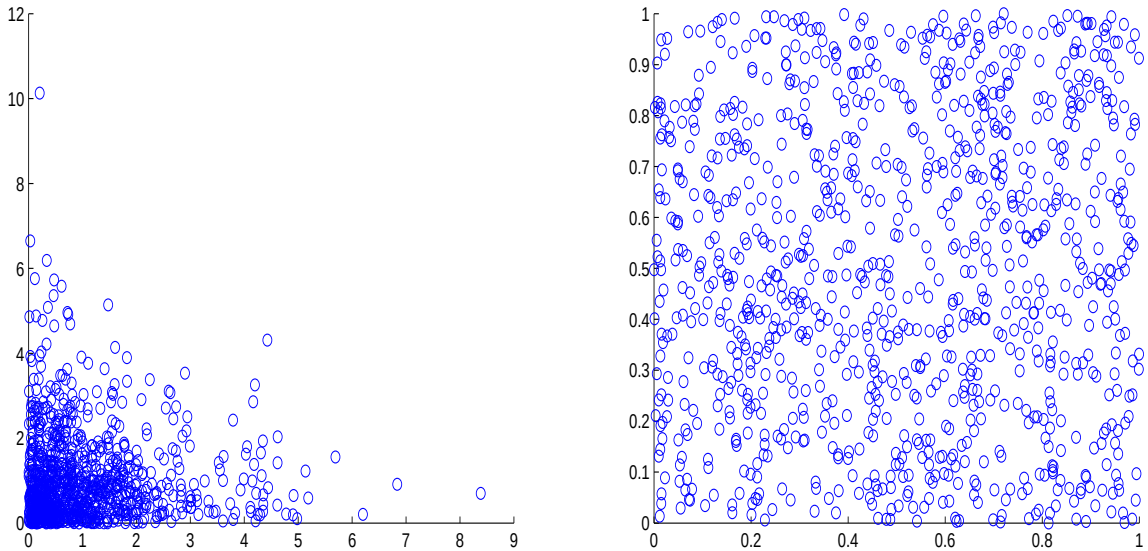
Why should we care about copulas?

Because copulas...

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

Most existing models assume rigid margins of the same form (e.g., Gaussian, Student, exponential, Gamma, Weibull, etc.).

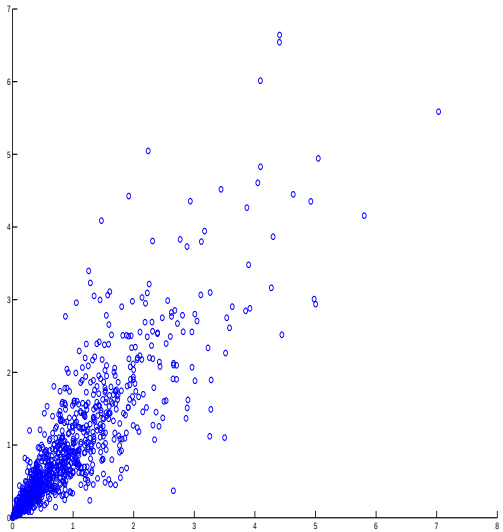
Copulas reveal the true nature of dependence

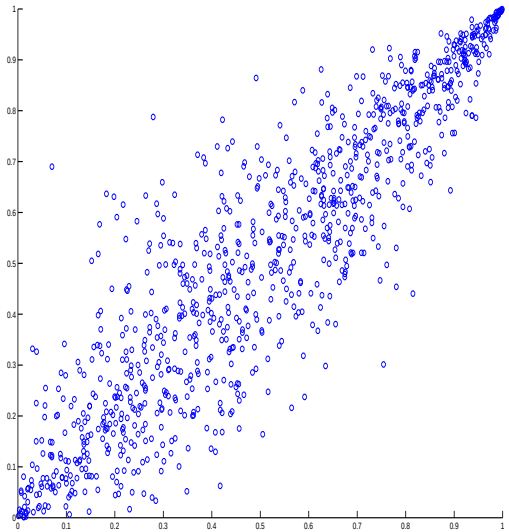


Transforming two independent exponentials into uniforms

1. Copulas

Copulas reveal the true nature of dependence





Transforming to uniforms the margins of
Gumbel's exponential distribution

Characterization of dependence through copulas

Sklar (1959) showed that

- Any distribution H can be written in the form

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

- The representation is unique when F and G are continuous.

If F or G is discontinuous, see Genest & Nešlehová (2007).

Models with arbitrary margins built from copulas

If C is an arbitrary copula, then

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

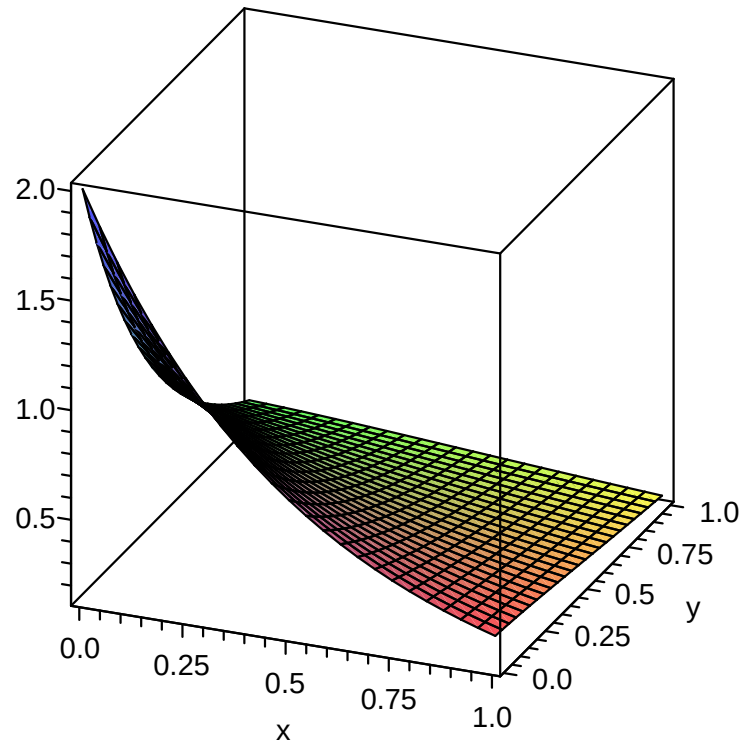
is a bivariate distribution function with margins F and G.

Example: The Farlie–Gumbel–Morgenstern (FGM) model

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

1. Copulas

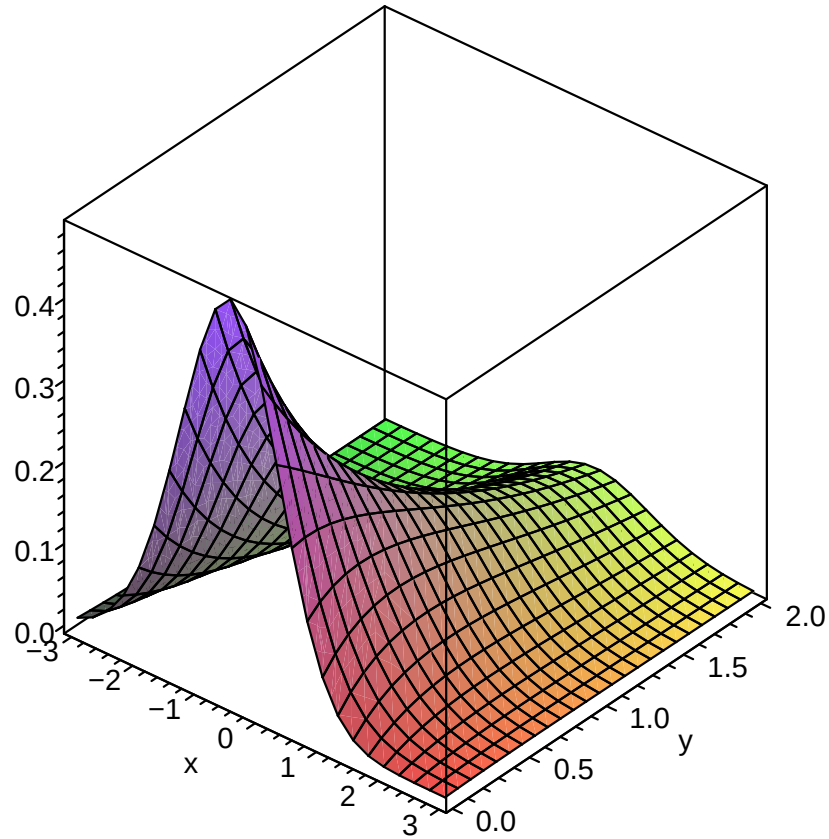
An FGM exponential model



FGM copula model with unit exponential margins

1. Copulas

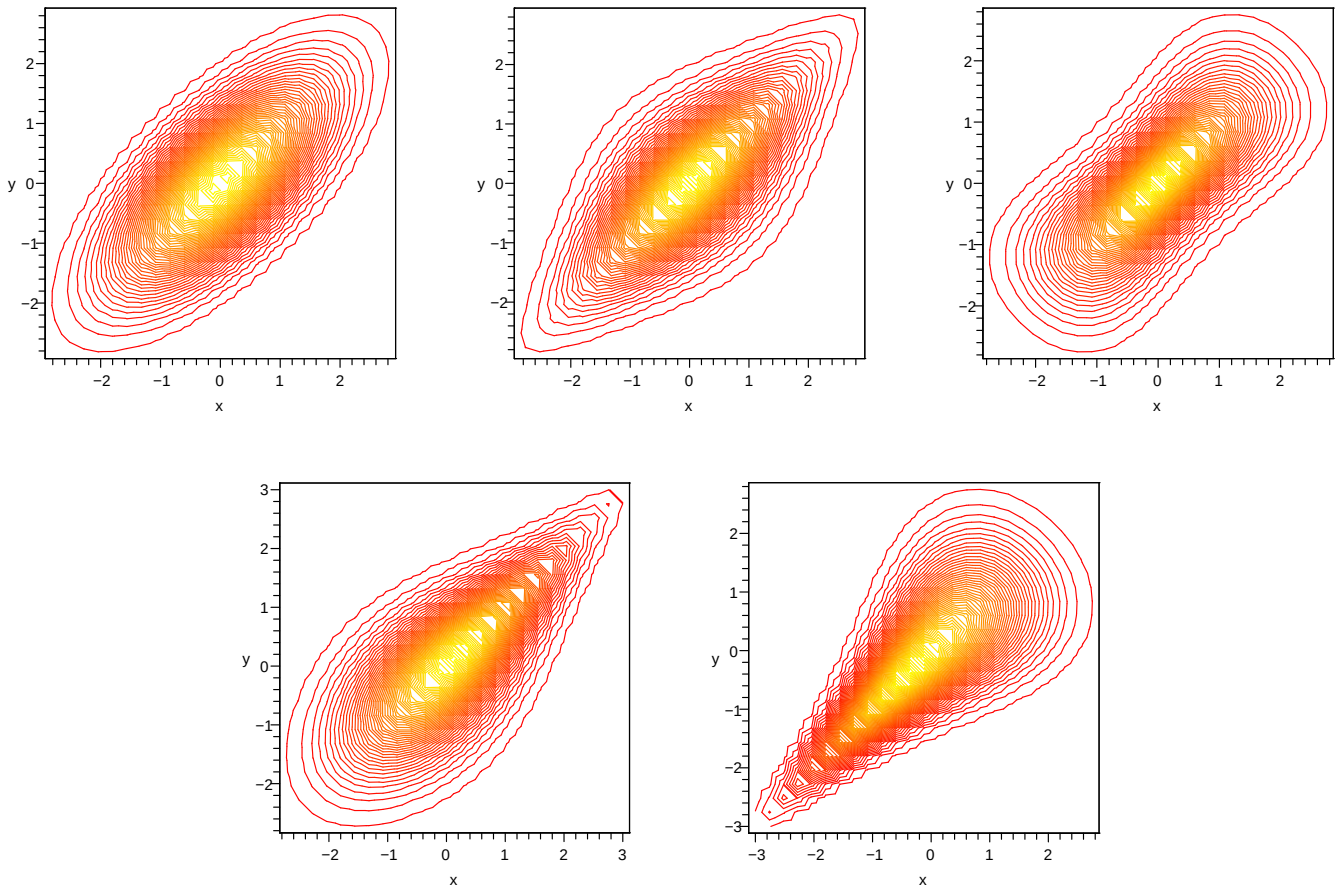
An FGM mixed model



FGM model with a unit exponential and a $\mathcal{N}(0,1)$ margin

1. Copulas

Five copulas with standard Gaussian margins



2. Copula models

What is a copula model?

Many stochastic models define the relation between X and Y through expectations, viz.

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

In a copula model, the joint distribution is written as

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

with

$$C \in (C_\theta), \quad F \in (F_\alpha), \quad G \in (G_\beta).$$

2. Copula models

Advantages

In the copula model

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

- F_α and G_β could take very different forms;
- they could involve covariables;
- the estimation of C_θ, F_α, and G_β can be done separately.

This type of modelling is non-restrictive because of Sklar's representation theorem.

Example 1 (van den Goorbergh et al., 2005)

Consider the following variables:

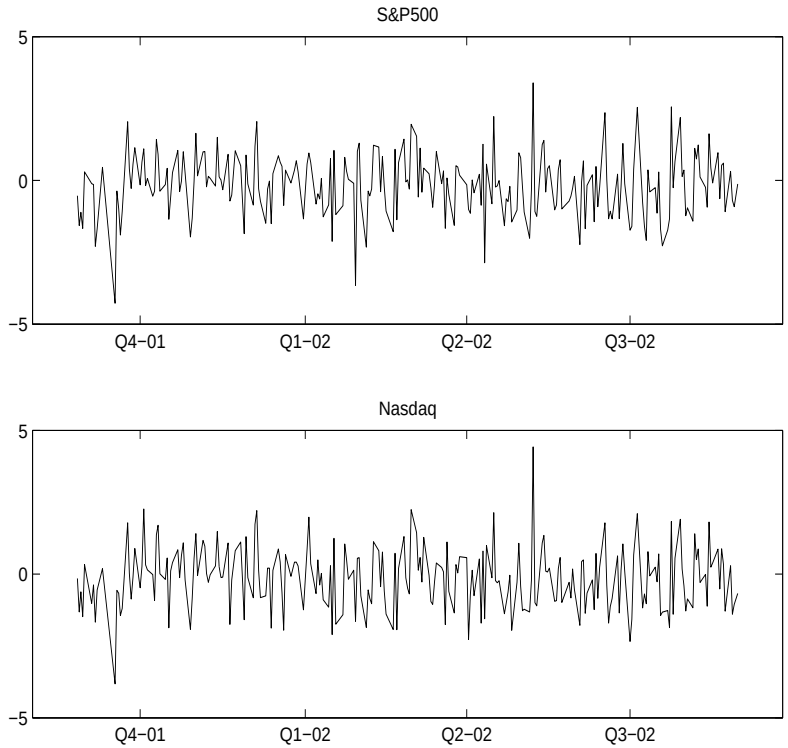
- Nasdaq from January 1, 1993 to August 30, 2002;
- S&P 500 from January 1, 1993 to August 30, 2002.

A standard (adequate) model for each log-return series is

$$\begin{aligned} r_{i,t+1} &= \mu_i + \eta_{i,t+1}, \\ \mathcal{L}(\eta_{i,t+1}|\mathcal{I}_t) &= \mathcal{L}(0, h_{i,t}), \\ h_{i,t+1} &= \omega_i + \beta_i h_{i,t} + \alpha_i \eta_{i,t+1}^2. \end{aligned}$$

2. Copula models

Estimated standardized GARCH innovations



Subsample of Nasdaq and S&P 500 indices
from January 1, 1993 to August 30, 2002

2. Copula models

How can we view the underlying copula?

Estimate the margins in the most conservative way possible, viz.

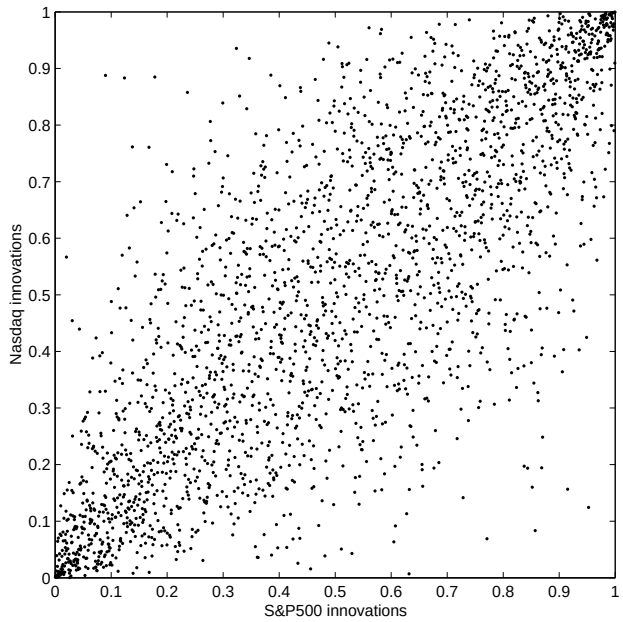
$$F_n(x) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x), \quad G_n(y) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}(Y_i \leq y).$$

Then 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

Example 1 (cont'd)



Empirical copula for estimated standardized GARCH innovations of Nasdaq vs S&P500 from January 1, 1993 to August 30, 2002

2. Copula models

The empirical copula (Deheuvels 1979)

It is defined for all $u, v \in [0,1]$ by

$$\hat{C}_n(u,v) = \frac{1}{n} \sum_{i=1}^{n} 1\left(\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v \right),$$

Its properties were studied by various authors, beginning with Rüschendorf (1976).

See also Gänßler & Stute (1987), van der Vaart & Wellner (1996), Fermanian et al. (2004), Genest & Rémillard (2004), Tsukahara (2005), and Segers (2012).

Note that $\hat{C}_n$ is a jump function, and therefore not a copula itself.

Theorem of Rüschendorf (1976)

If C admits continuous first-order partial derivatives on $(0,1)^2$,

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

converges weakly as $n \to \infty$ to a centered Gaussian process

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

where $\mathbb{C}$ is a centered Gaussian process with covariance function

$$\mathrm{cov}\{\mathbb{C}(u,v), \mathbb{C}(w,z)\} = C(u \wedge w, v \wedge z) - C(u,v)C(w,z).$$

Example 2 (Grégoire et al., 2008)

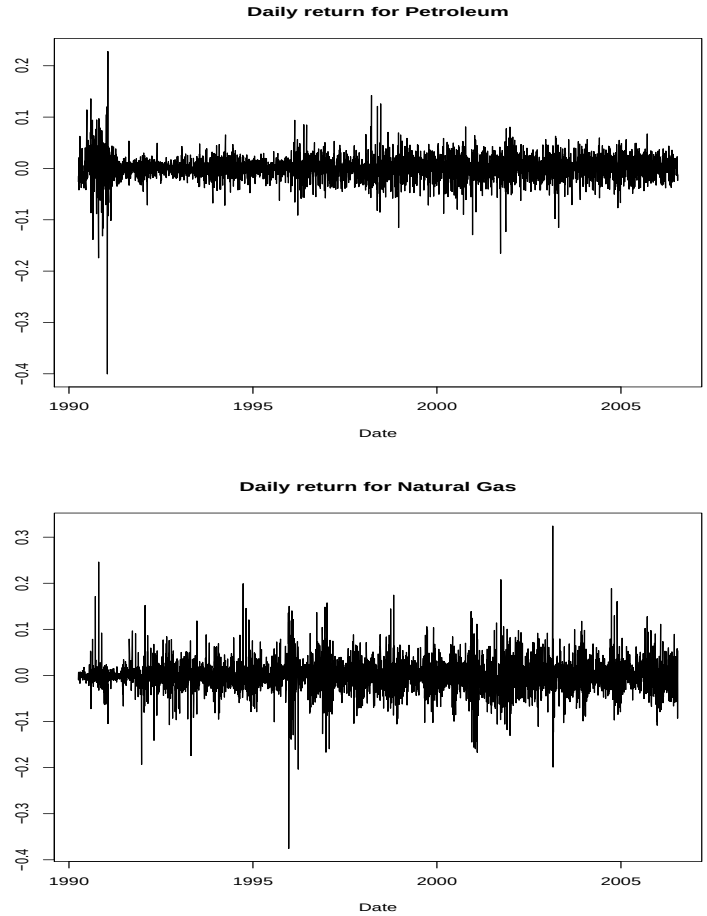
Consider the following variables

- the price of Light, Sweet Crude Oil per barrel
- the price of Natural Gas per mmBTU

based on data from January 1, 2004, to August 31, 2006.

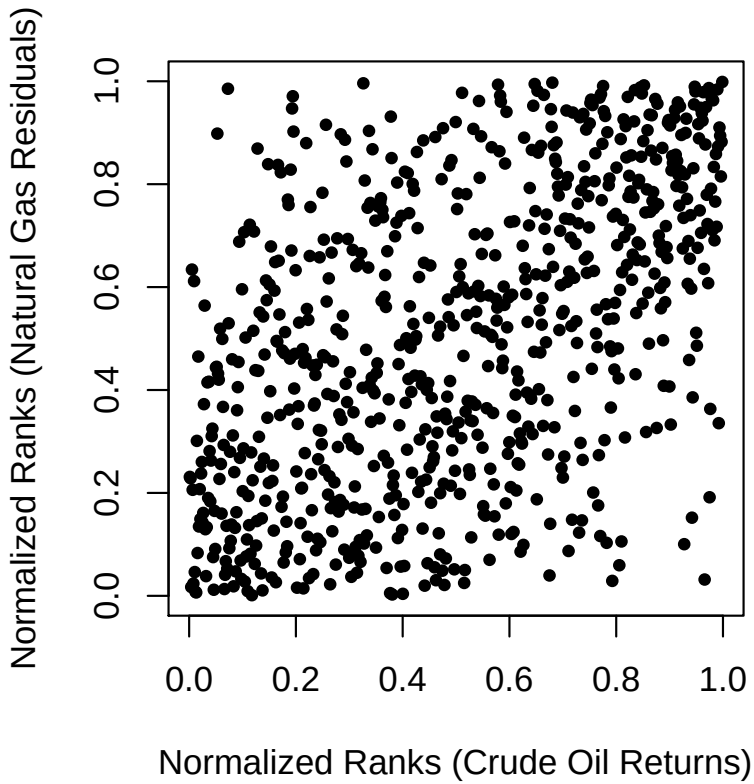
Here again, the time variation of the marginal series can be captured by GARCH models.

2. Copula models



Returns for Petroleum and Natural Gas, 2003–2006

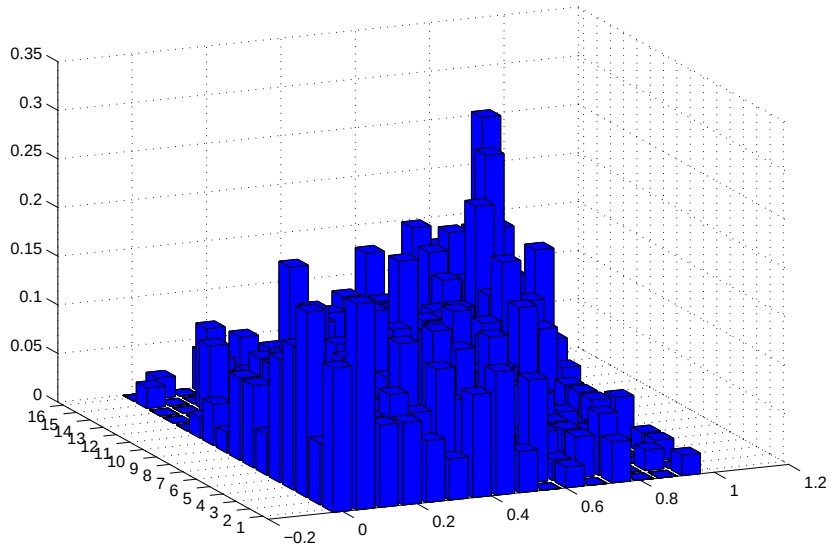
2. Copula models



Empirical copula for Petroleum and Natural Gas, 2003–2006
Source: Grégoire et al. (2008)

2. Copula models

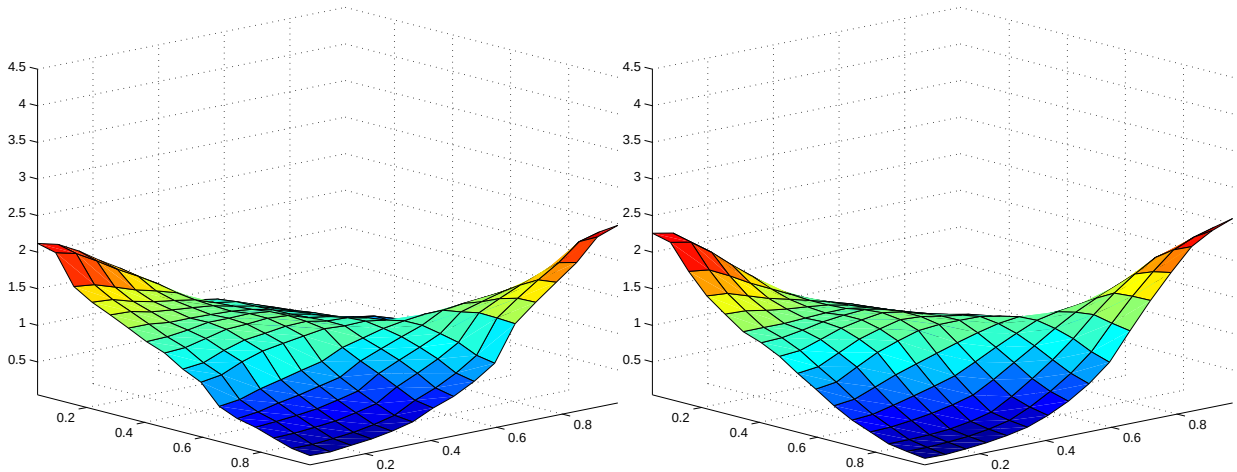
Alternative tools for copula visualization



A 3D-histogram

2. Copula models

Rank-based wavelet estimates are preferable



The choice of wavelets does not matter much:

Haar wavelets (left), Daubechies wavelets (right).

For more information, see, e.g., Genest et al. (2009).

3. Inference for copula models

Steps to consider

Statistical inference usually proceeds in three logical steps, viz.

3A Model selection: descriptive statistics

3B Model fitting: estimation

3C Model validation: goodness-of-fit testing

Once a model has been chosen and validated, it can be used for prediction or decision-making in the face of uncertainty.

Doing what comes naturally

A natural idea consists of measuring the dependence in $\hat{C}_n$ by computing the rank correlation, viz.

$$\rho_n = \frac{\sum_{i=1}^{n} \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2 \sum_{j=1}^{n} \left(S_j - \bar{S}\right)^2}}.$$

3A. Descriptive statistics

Spearman's rho

As it turns out, ρ_n is an asymptotically unbiased estimate of

$$\rho = -3 + 12 \int \int C(u, v) \mathrm{d}v \mathrm{d}u,$$

which is known as Spearman's rho.

When $C = C_\perp$, one has

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

This can be used to test the null hypothesis of independence.

3A. Descriptive statistics

Spearman's rho (cont'd)

It is easy to see (using Hoeffding's identity) that

$$\begin{aligned}
\rho &= -3 + 12 \int \int C(u, v) \mathrm{d}v \mathrm{d}u \\
&= -3 + 12 \int \int uv \mathrm{d}C(u, v) \\
&= -3 + 12 \int \int F(x)G(y) \mathrm{d}C\{F(x), G(y)\} \\
&= -3 + 12 \int \int F(x)G(y) \mathrm{d}H(x, y) \\
&= \mathrm{corr}\{F(X), G(Y)\}.
\end{aligned}$$

Other nonparametric measures of dependence

van der Waerden's coefficient:

$$\text{corr}\{\Phi^{-1}(U), \Phi^{-1}(V)\} = \text{corr}\{\Phi^{-1} \circ F(X), \Phi^{-1} \circ G(Y)\}$$

Blomqvist's beta:

$$-1 + 4C\left(\frac{1}{2}, \frac{1}{2}\right)$$

Other nonparametric measures of dependence

Kendall's tau (or coefficient of concordance):

$$\begin{aligned}\tau &= -1 + 4 \iint C(u,v)\,\mathrm{d}C(u,v) \\ &= \Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\} \\ &\qquad - \Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\},\end{aligned}$$

where (X_1, Y_1) and (X_2, Y_2) are independent observations from H.

Other options include Blest's coefficient, Gini's coefficient, etc.

3A. Descriptive statistics

Remarks

- All these statistics can be estimated by replacing C by $\hat{C}_n$ in the formulas, e.g.,

$$\tau_n = -1 + 4 \iint \hat{C}_n(u, v) \, \mathrm{d}\hat{C}_n(u, v);$$

- their asymptotic distribution is typically Gaussian;
- they can be used to construct various tests of independence.

Note: The marginal distributions are nuisance parameters and the vectors of ranks are maximally invariant for this problem.

Examples

Schucany et al. (1978) show that in the FGM model,

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

When

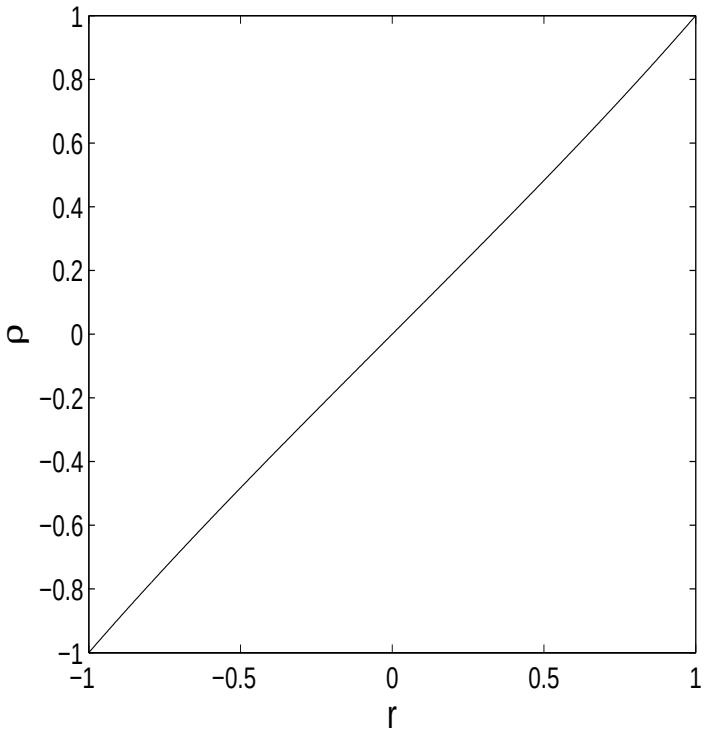
$$(X, Y) \sim \mathcal{N}_2 \left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & r\sigma_X\sigma_Y \\ r\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right],$$

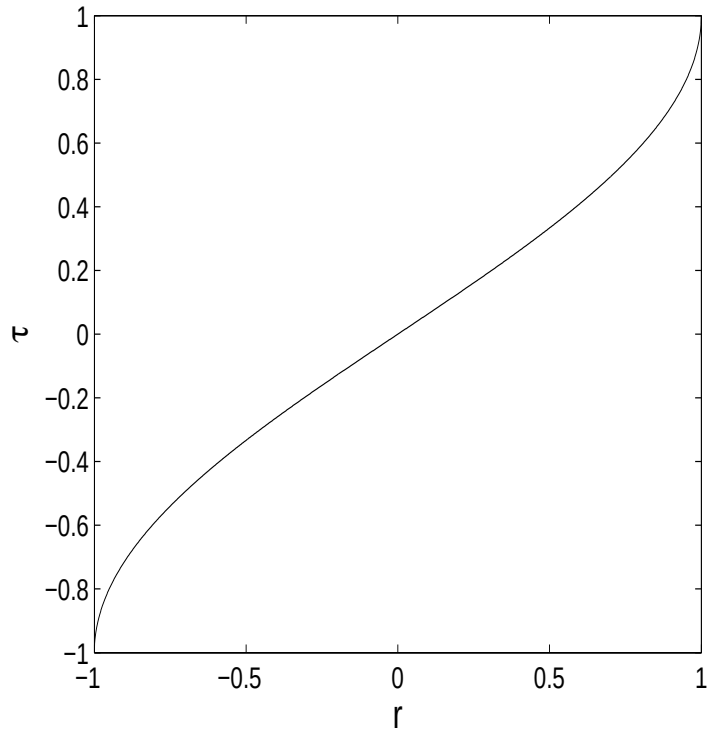
It has been known since the work of Esscher (1924) that

$$\rho(X, Y) = \frac{6}{\pi} \arcsin(r/2), \quad \tau(X, Y) = \frac{2}{\pi} \arcsin(r).$$

3A. Descriptive statistics

Spearman's and Kendall's tau in the Gaussian model





Criticism of Pearson's correlation coefficient

In the Gaussian model, Spearman's rho is almost the same as

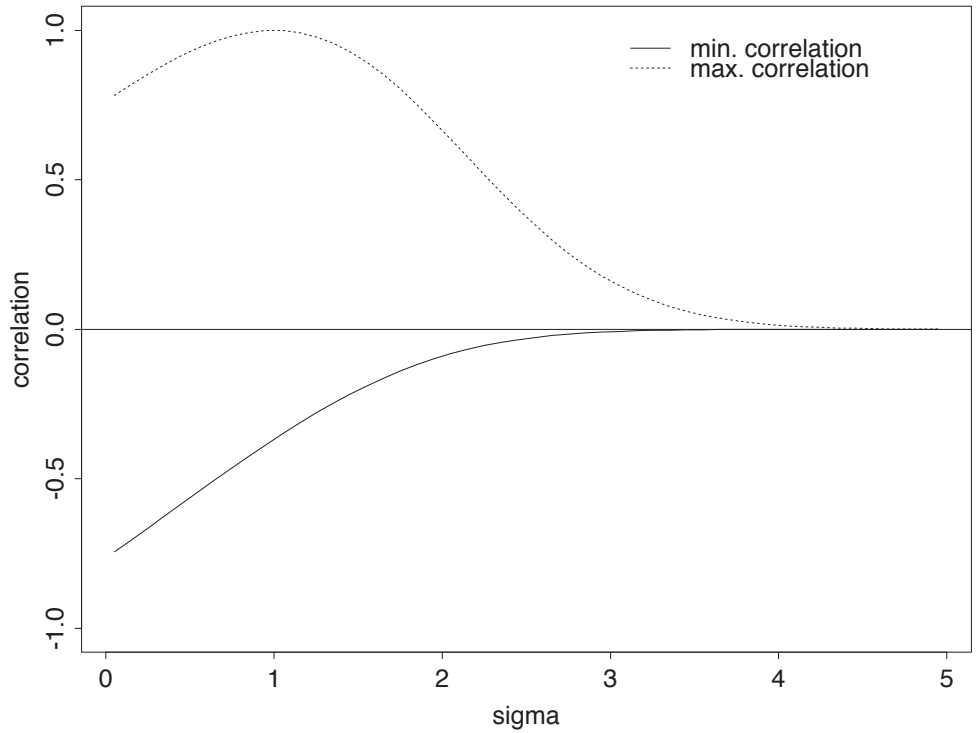
$$r(X, Y) = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}\,.$$

Outside the Gaussian paradigm, r has major drawbacks (Embrechts et al. 1999):

- it is only a measure of linear association;
- its value depends on the marginal distributions;
- it can be close to 0 even in case of strong dependence;
- it may not even exist (e.g., Cauchy).

3A. Descriptive statistics

Example (McNeil et al. 2005)



Bounds on corr(X, Y) when $X \sim \mathrm{LN}(0, 1)$ and $Y \sim \mathrm{LN}(0, \sigma^2)$

3A. Descriptive statistics

Limiting cases

Nonparametric measures of dependence reach 1 when

$$C(u, v) = M(u, v) \equiv \min(u, v).$$

They equal -1 when

$$C(u, v) = W(u, v) = \max(0, u + v - 1).$$

M and W are called the Fréchet–Hoeffding bounds because

$$W(u, v) \leq C(u, v) \leq M(u, v).$$

3B. Estimation

How to proceed with estimation?

Estimation can proceed as with any other multivariate model having a joint density, viz.

$$\frac{\partial^2}{\partial x \partial y} C_\theta\{F_\alpha(x), G_\beta(y)\} = c_\theta\{F_\alpha(x), G_\beta(y)\} f_\alpha(x) g_\beta(y).$$

Joe (2005) argues that it is more convenient (and nearly as efficient) to proceed in two steps:

- estimate the marginal parameters first;
- estimate the dependence parameter second.

3B. Estimation

Relative merits of this approach

- it is easy to implement;
- it yields an asymptotically Gaussian, unbiased estimate;
- typically, it is as efficient as one-stage maximum likelihood estimation at independence.

However, if the margins are incorrectly specified, the estimation of C_θ may be seriously affected (Kim et al., 2007).

3B. Estimation

Alternative approach

Estimate the margins conservatively using

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

Then replace these estimates into the log-likelihood and maximize

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

This approach was studied by Genest et al. (1995) and Shih & Louis (1995). It is usually referred to as the canonical maximum likelihood method.

3B. Estimation

Asymptotic results

Genest et al. (1995) showed that

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

and provided a consistent estimate of ν^2.

This procedure

- assumes nothing about the marginal distributions;
- provides estimators that are robust to misspecification in the marginal models.

3B. Estimation

Illustration (Genest & Favre, 2007)

Suppose it is desired to fit an FGM copula, viz.

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

The estimator θ_n is then found by solving

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

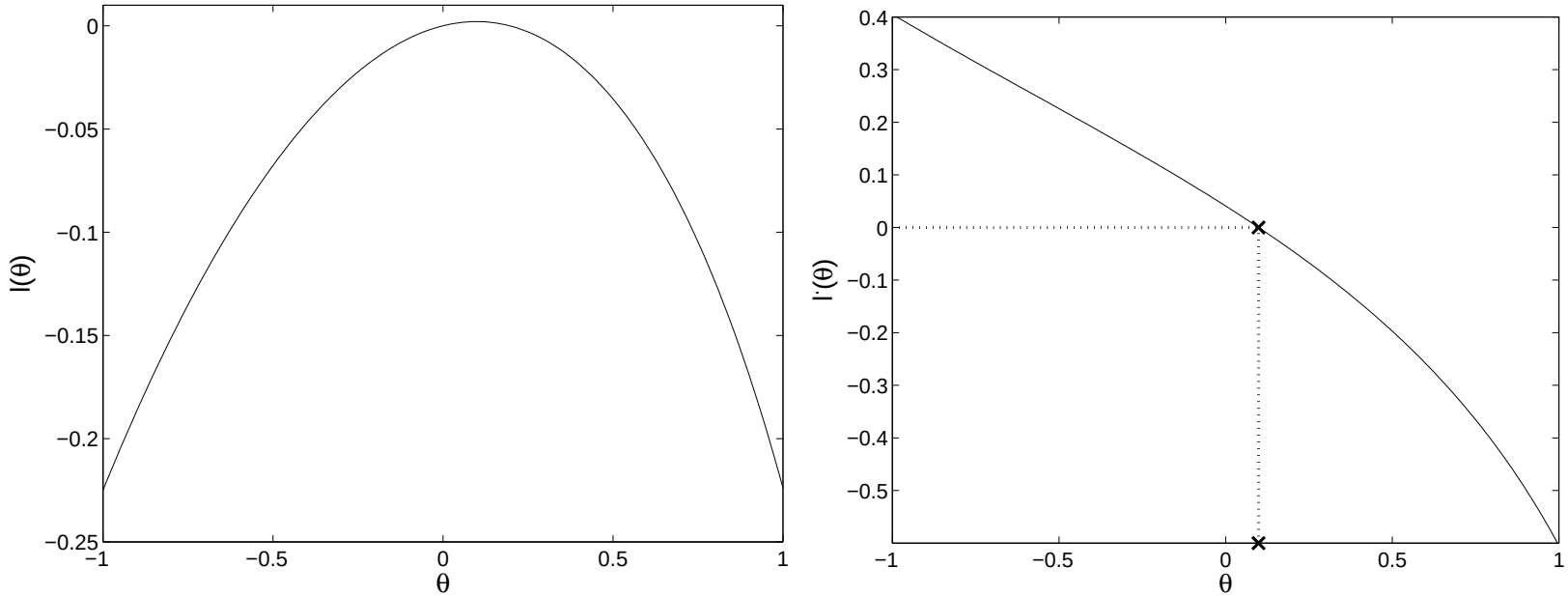
Illustration (cont'd)

The pseudo-score function is then

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

3B. Estimation

Illustration (cont'd)



Pseudo log-likelihood (left) and score function (right)

3B. Estimation

Moment-like estimators

When the dependence parameter θ is real, then

$$\tau = \psi(\theta) \quad \text{and} \quad \rho = \Psi(\theta)$$

are typically monotone increasing functions of θ.

It is thus natural to estimate θ either by

$$\check{\theta}_n = \psi^{-1}(\tau_n) \quad \text{or} \quad \breve{\theta}_n = \Psi^{-1}(\rho_n).$$

These estimators are similar to those obtained by the "method of moments" in classical statistics.

3B. Estimation

Example

In the FGM model, one has

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

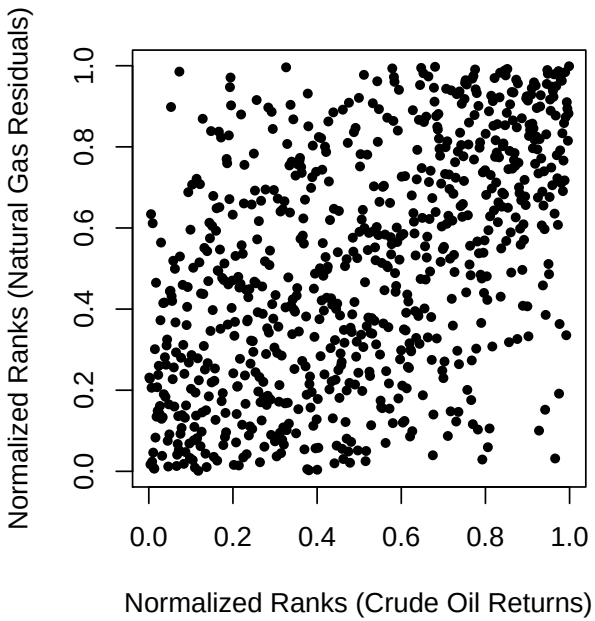
Therefore, two possible estimates of $\theta \in [-1, 1]$ are

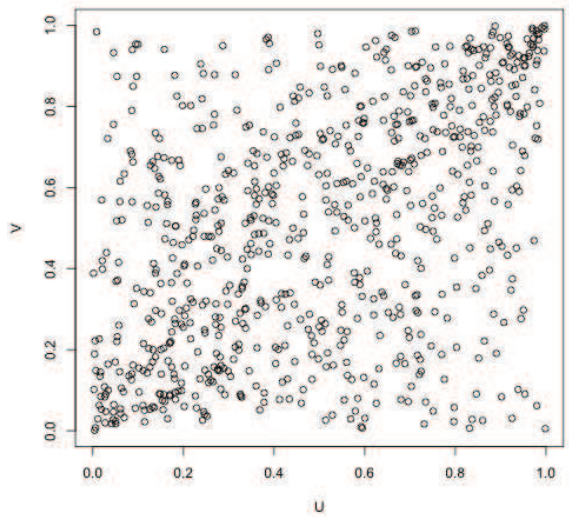
$$\breve{\theta}_n = 3\,\rho_n, \quad \breve{\theta}_n = 9\,\tau_n/2.$$

Both estimators are asymptotically unbiased and Gaussian, because ρ_n and τ_n are U-statistics and one can invoke Slutsky's Lemma.

3B. Estimation

Example: Oil vs Gas data





In this case, one can try to fit the Student t copula.

The right picture shows a sample of size 756 from this copula with 4 DoF and $\rho_n = 0.5221467$ corresponding to $\tau_n = 0.3497373$.

Example (cont'd): Oil vs Gas data

Student copula fit with the `QRMlib` package:

```
# Fit the Student t copula with
# the pseudo ML method


> fit.tcopula(roilgas)
$P
          [,1]      [,2]
[1,] 1.0000000 0.5041283
[2,] 0.5041283 1.0000000

$nu
[1] 45.6735

$converged
[1] TRUE

$ll.max
[1] 107.7697
```

```
# Fit Student t copula with the
# moment-based estimator
# of Kendall's tau

> fit.tcopula.rank(roilgas)
$P
           V2        V3
V2 1.0000000 0.5221467
V3 0.5221467 1.0000000

$nu
[1] 42.25870

$converged
[1] TRUE

$ll.max
[1] 107.5007
```

3B. Estimation

Example (cont'd): Oil vs Gas data

Gauss copula fit with the `QRMlib` package:

```
# Given that the degrees of freedom
# of the t copula are high, one can
# fit the Gaussian copula as well.

> fit.gausscopula(roilgas)
$P
          [,1]      [,2]
[1,] 1.0000000 0.5021058
[2,] 0.5021058 1.0000000

$converged
[1] TRUE

$ll.max
[1] 107.5049
```

Here, the likelihood of the Student t model is just a little higher. To choose between the two, goodness-of-fit tests are needed.

3C. Goodness-of-fit testing

How to go about goodness-of-fit testing?

Goodness-of-fit testing: Test the hypothesis that

$$H_0 : C \in (C_\theta)$$

from a random sample. The literature on the subject is exploding.

This is not the same as model selection treated, e.g., Palaro & Hotta (2006) using an Akaike or a BIC criterion.

3C. Goodness-of-fit testing

General strategy

- Assume $H_0 : C \in (C_\theta)$ holds true.
- Estimate θ by θ_n under this assumption.
- Measure a "distance" between C_{θ_n} and the empirical copula,

$$\hat{C}_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} \left(\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v \right).$$

- See whether that distance is large or small, taking into account its natural variability under H_0.
- Derive a p-value and conclude.

3C. Goodness-of-fit testing

Current approaches fall into four categories:

A) Copula-specific tests

B) Tests based on dichotomization of the data

C) General tests involving tuning parameters

D) Blanket tests

33C. Goodness-of-fit testing

A) Copula-specific procedures

These are rank-based tests, but designed to test specific dependence structures:

- Gaussian copula: Malevergne & Sornette (2003), etc.
- Clayton copula: Shih (1998), Glidden (1999), Cui & Sun (2004), etc.
- Archimedean copulas: Trede & Savu (2009), etc.

3C. Goodness-of-fit testing

B) Tests based on dichotomization of the data

Procedure:

- A frequency table with uniform margins is computed.
- Observed counts and expected counts are compared via a standard test statistic, e.g., the chi-square distance.
- The limiting null distribution is used to compute p-values.

3C. Goodness-of-fit testing

B) Tests based on dichotomization (cont'd)

Contributions along those lines include:

- Klugman & Parsa (1999)
- Andersen et al. (2005)
- Dobrić & Schmid (2005)
- Junker & May (2005)

Note:

- These tests depend on the number of categories selected;
- their distribution is not the same as in the classical context, due to dependence between ranks.

3C. Goodness-of-fit testing

C) General tests involving tuning parameters

These tests apply to any copula family but involve:

- an arbitrary parameter, as in Wang & Wells (2000);
- kernels, weight functions, and associated smoothing parameters, as in
  - Fermanian (2005);
  - Panchenko (2005);
  - Scaillet (2007);
  - Berg & Bakken (2009).

3C. Goodness-of-fit testing

D) Blanket tests

This term covers goodness-of-fit procedures that

- are applicable to all copula structures;
- require no strategic choice for their use.

Included in this category are the tests of:

- Breymann et al. (2003);
- Genest et al. (2006);
- Dobrić & Schmid (2007);
- Genest & Rémillard (2008);
- Genest et al. (2009).

3C. Goodness-of-fit testing

Tests based on $\hat{C}_n$

In my opinion, the most natural test is based on

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

where

- $\hat{C}_n$ is the empirical copula;
- θ_n is a rank-based estimate of θ, e.g., obtained by inversion of Kendall's tau.

3C. Goodness-of-fit testing

S_n and variants thereof

The computation of S_n is not as hard as it looks, because integration is with respect to $\hat{C}_n$, viz.

$$S_n = \sum_{i=1}^{n} \left\{ C_{\theta_n}\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) - \hat{C}_n\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) \right\}^2.$$

Other distances can be considered, e.g., L_1 or L_∞, viz.

$$T_n = \max_{1 \le i \le n} \left| C_{\theta_n}\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) - \hat{C}_n\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}\right) \right|.$$

3C. Goodness-of-fit testing

Advantages of S_n

- It is conceptually simple and based on ranks.
- It does not involve any "strategic choice" (dichotomization or tuning parameters).
- It yields a consistent test, as a result on the behaviour of the copula process

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

and the associated process

$$\sqrt{n}\,(\hat{C}_n - C_{\theta_n}).$$

3C. Goodness-of-fit testing

Parametric bootstrap

The limiting distribution of S_n is unwieldy and depends on the unknown value of θ under H_0.

However,

- Genest & Rémillard (2008) show that this distribution can be approximated efficiently under H_0.
- Their procedure relies on a "parametric bootstrap."
- Genest et al. (2009) assess the power of this test, along with seven other blanket procedures.

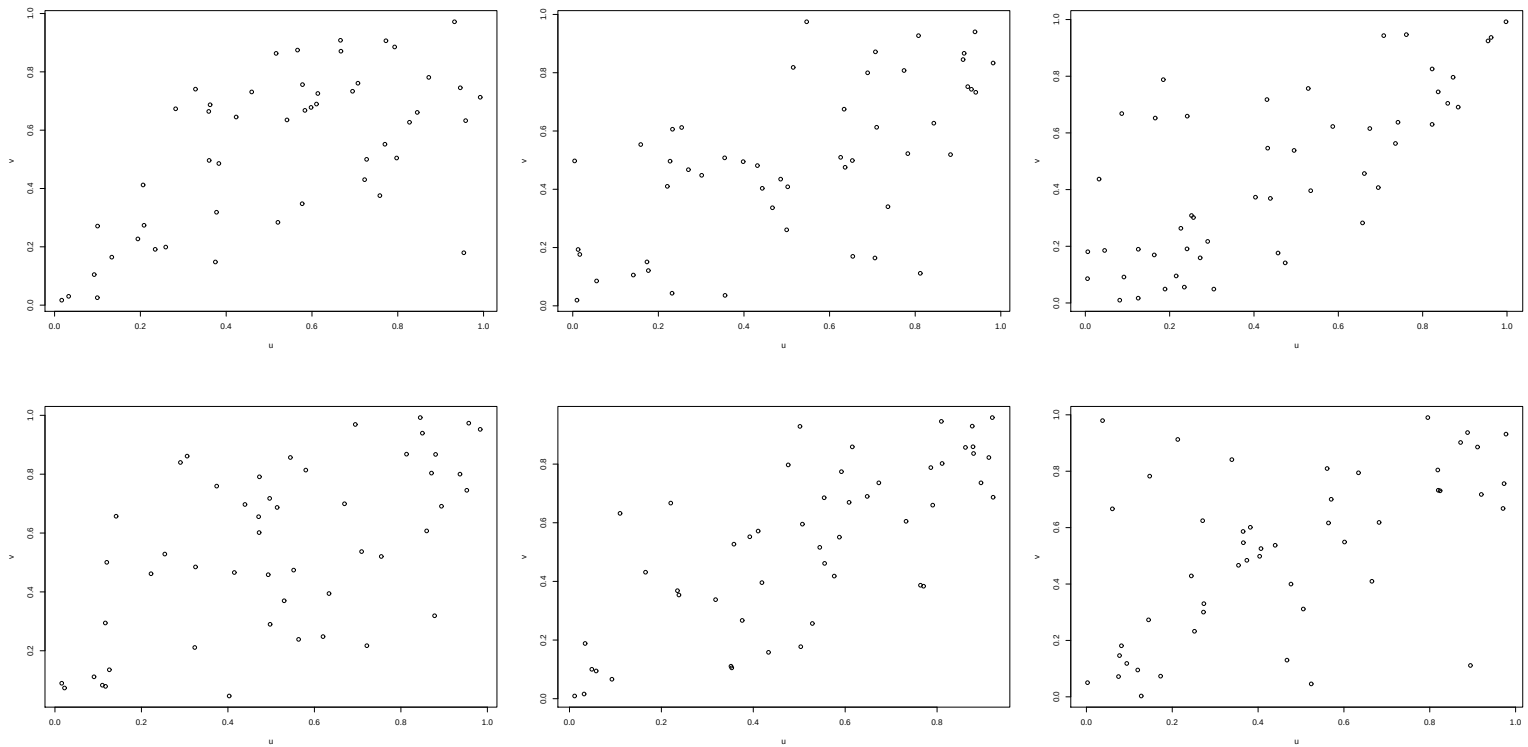
3C. Goodness-of-fit testing

Words of caution

- The jury is still out on the "best goodness-of-fit test."
- There is no perfect solution: some tests may be preferable for certain types of hypotheses or types of dependence.
- Working with ranks affects asymptotic distributions; cf. Breymann et al. (2003) and Dobrić & Schmid (2007).
- The Multiplier Method is often an efficient alternative to the parametric bootstrap; see, e.g., Kojadinovic & Yan (2011).

3C. Goodness-of-fit testing

Caution: large samples are required



Random samples of size $n = 50$ from 6 copulas with $\tau = 1/2$: Clayton, Frank, Gumbel, Gaussian, Student, and Plackett copula

4. Strategies for constructing copula models

Most common approaches

Broadly speaking, there are three:

- derive copulas from existing multivariate distributions, viz.

$$C(u, v) = H\{F^{-1}(u), G^{-1}(v)\};$$

- deduce them from probabilistic constructions;
- build them through successive conditionings, using bivariate copulas as blocks.

4. Strategies for constructing copula models

Copulas arising from well-known models

The prime example is the class of elliptical distributions, in which

$$X = \mu + RA\mathcal{U},$$

where, for the present purpose, one can take $\mu = 0$ while

- R is a positive random variable;
- $AA^\top$ is a Cholesky decomposition of Σ;
- $\mathcal{U}$ is uniformly distributed on $\mathcal{S}_d = \{u \in \mathbb{R}^d : ||u|| = 1\}$.

4. Strategies for constructing copula models

Table of generators of elliptical distributions

Copula	$R^2 \sim$	$g(t)$
Gaussian	$\chi^2_{(d)}$	$(2\pi)^{-d/2} \exp(-t/2)$
Student	$d \times \mathcal{F}(d, \nu)$	$\frac{(\pi\nu)^{-d/2}\Gamma(\frac{d+\nu}{2})}{\Gamma(\nu/2)}(1 + t/\nu)^{-(d+\nu)/2}$
Cauchy	$d \times \mathcal{F}(d, 1)$	$\frac{(\pi)^{-d/2}\Gamma(\frac{d+1}{2})}{\Gamma(1/2)}(1 + t)^{-(d+1)/2}$
Pearson type II	$\text{Beta}(d/2, \nu + 1)$	$\frac{\Gamma(d/2+\nu+1)}{\pi^{d/2}\Gamma(\nu+1)}(1 - t)^{\nu}$, $t \in [-1, 1], \quad \nu > -1$

4. Strategies for constructing copula models

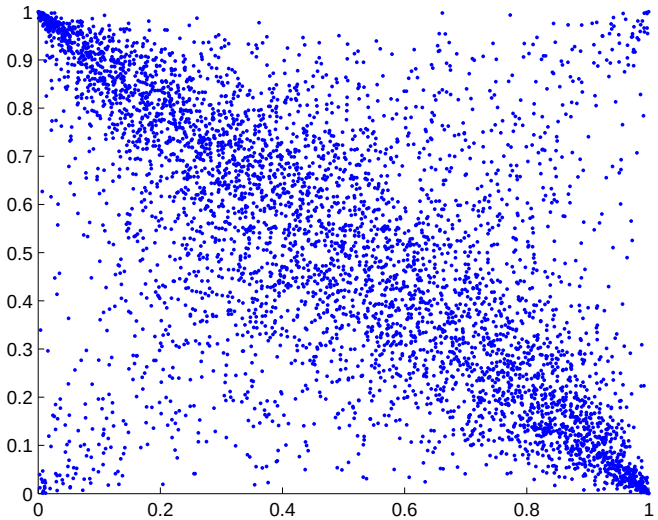
Advantages of elliptical (or meta-elliptical) copulas

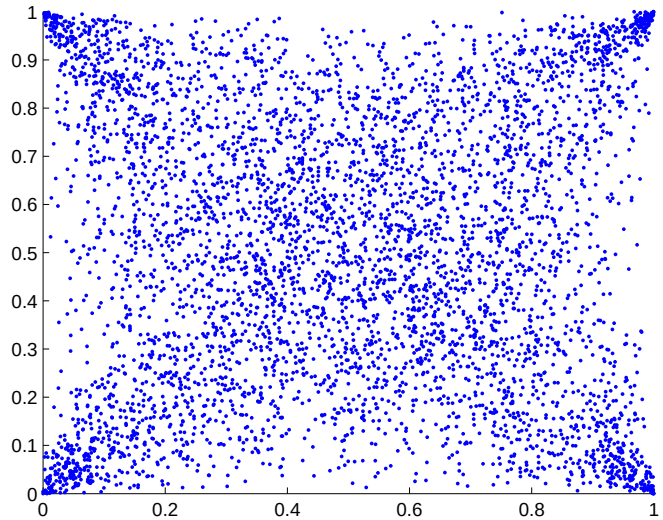
- Their properties are well known.
- They are easy to simulate using the `R copula` package:

```
# The multivariate normal copula

> nc3 <- normalCopula(c(0.7,0.9,0.7), dim=3, dispstr="un")
> sample <- rcopula(nc2, 1000)
> pairs(sample)

# The multivariate Student t copula

> tc <- tCopula(c(0.7,0.9,0.7), dim=3, df=2, dispstr="un")
> sample <- rcopula(tc, 1000)
> pairs(sample)
```

4. Strategies for constructing copula models

5000 pairs of the bivariate Cauchy copula





$r = -0.5$ (left) and $r = 0$ (right)

4. Strategies for constructing copula models

Elliptical copulas are...

- available in general dimension $d \geq 2$;
- can accommodate tail-dependence behaviour, e.g.,

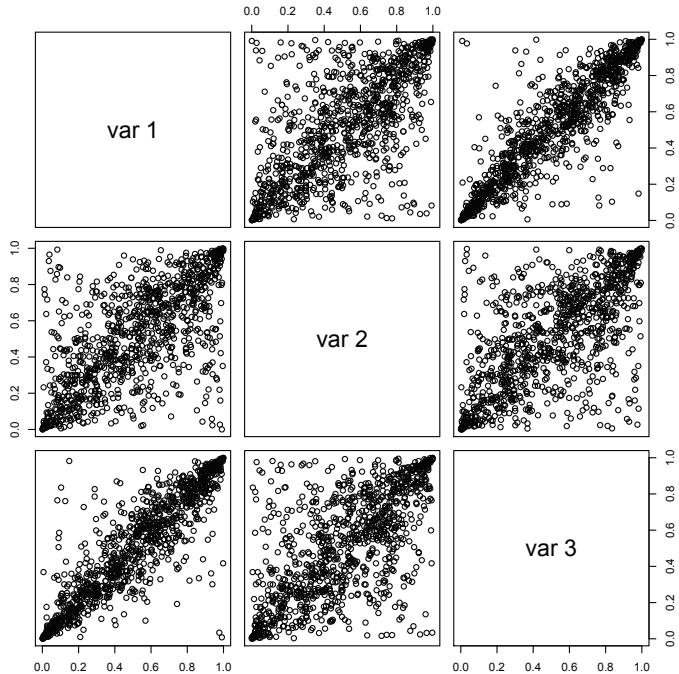
$$\lambda_U = \lim_{q \to 1} \frac{\bar{C}(q, q)}{1 - q} = \lim_{q \to 1} \Pr\{Y > G^{-1}(q) | X > F^{-1}(q)\}.$$

However,

- all lower-dimensional marginals are identical;
- $r = 0$ (or Σ diagonal) does not correspond to independence except in the Gaussian case (Kelker 1970).

4. Strategies for constructing copula models

1000 observations from the trivariate t copula



$r_1 = 0.7$, $r_2 = 0.9$ and 4 degrees of freedom

4. Strategies for constructing copula models

Copulas deduced from probabilistic constructions

Examples include

- Archimedean copulas, viz.

$$C(u, v) = \psi\{\psi^{-1}(u) + \psi^{-1}(v)\}$$

where ψ is a Laplace transform with inverse ψ^{-1};

- extreme-value copulas, viz.

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

where $A : [0, 1] \to [0, 1]$ is convex and

$$\max(t, 1 - t) \leq A(t) \leq 1, \quad t \in [0, 1].$$

4. Strategies for constructing copula models

Extreme-value copulas

The joint distribution of a pair (M_{n1}, M_{n2}) of maxima is

$$\Pr(M_{n1} \leq x, M_{n2} \leq y) = H^n(x, y), \quad x, y \in \mathbb{R}$$

and its margins are F^n and G^n.

The copula of the pair (M_{n1}, M_{n2}) is thus

$$C^n(u^{1/n}, v^{1/n}), \quad u, v \in [0, 1].$$

When $n \to \infty$, an extreme-value copula obtains; its form is characterized by the function A (Pickands 1981).

4. Strategies for constructing copula models

Archimedean copulas

They are the dependence structures of mixture models.

When $T_1, \ldots, T_d$ are survival times such that

$$\begin{aligned} \Pr(T_1 > t_1, \ldots, T_d > t_d | Z = z) &= \prod_{i=1}^{d} \Pr(T_i > t_i | Z = z) \\ &= B_1(t_1)^z \times \cdots \times B_d(t_d)^z, \end{aligned}$$

the copula C of $(T_1, \ldots, T_d)$ is then Archimedean with generator

$$\psi(t) = \int e^{-zt} \mathrm{d}G(z),$$

where G is the distribution function of the frailty Z.

4. Strategies for constructing copula models

Examples of Archimedean copulas

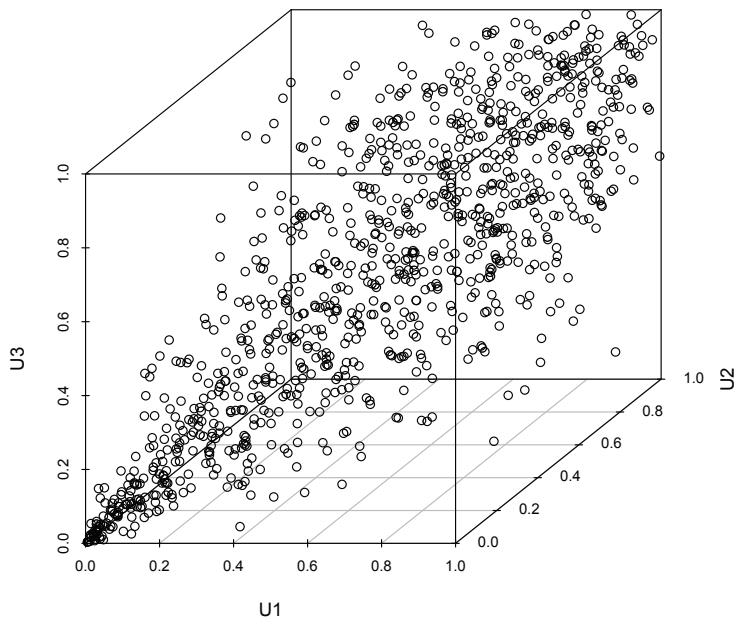
Statistical uses of Archimedean copulas go back by Genest & MacKay (1986); their mixture representation is due to Marshall & Olkin (1988); see also Oakes (1989).

Distribution of Z	Copula family
Gamma	Clayton (1978), Cook & Johnson (1981)
Logarithmic series	Frank (1979), Genest (1987)
Positive stable	Gumbel (1961)

These copulas are very popular in actuarial science and finance; see, e.g., Frees & Valdez (1998) or McNeil et al. (2005).

4. Strategies for constructing copula models

Bivariate Clayton Copula



Sample of size 500 from a Clayton copula with parameter $\theta = 2$

4. Strategies for constructing copula models

Archimedean copulas...

- are easy to simulate with the `R copula` package;
- can accommodate tail-dependence behaviour.

However,

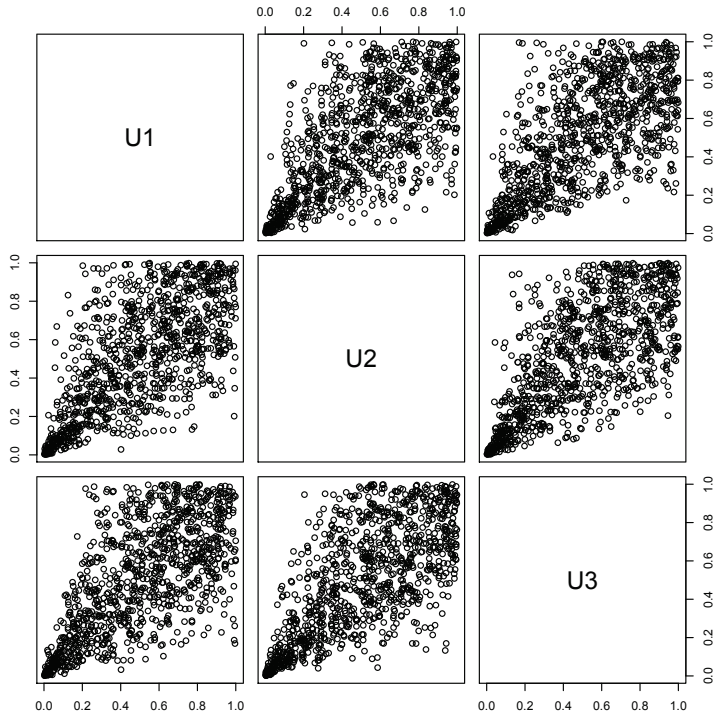
- they have highly symmetric dependence structures, viz.

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

  and in particular identical lower-dimensional margins;
- they cannot accommodate all degrees of negative dependence when $d > 2$ (in fact, less and less so as d increases).

4. Strategies for constructing copula models

Trivariate Clayton Copula



Sample of size 500 from a Clayton copula with parameter $\theta = 2$

4. Strategies for constructing copula models

Literature

- for properties of multivariate Archimedean copulas:

  A.J. McNeil & J. Nešlehová (2009).
  Multivariate Archimedean Copulas, d-monotone functions and ℓ_1-norm symmetric distributions.
  Ann. Statist., 37, 3059–3097.

- for asymmetrized Archimedean (Liouville) copulas:

  A.J. McNeil & J. Nešlehová (2010).
  From Archimedean to Liouville copulas.
  J. Multivariate Anal., 101, 1772–1790.

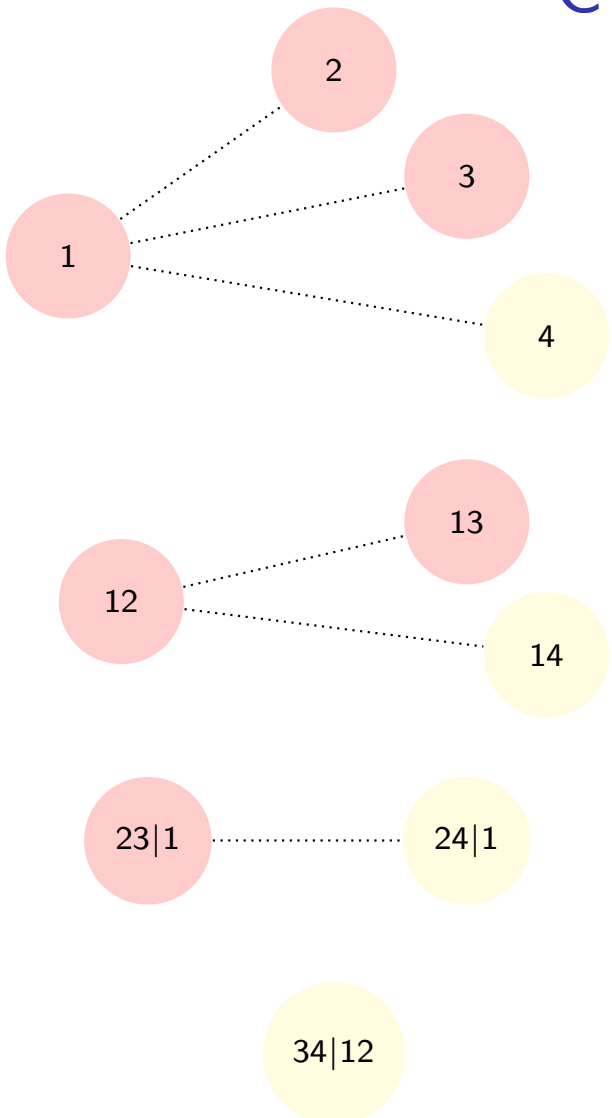
4. Strategies for constructing copula models

Pair Copula Constructions (PCCs)

- Idea: Decompose a multivariate copula into a cascade of pair copulas, some of which are conditional.
- Advantages
  - There is already a rich class of bivariate copulas.
  - Inference and model selection for bivariate copulas are well established.
  - PCCs allow different pairwise dependence patterns in model variables, hence more flexible.
- Vines define a systematic way to decompose multivariate distributions.
- Two commonly used regular vines are C-vines and D-vines.

4. Strategies for constructing copula models

C-Vine with four variables



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

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

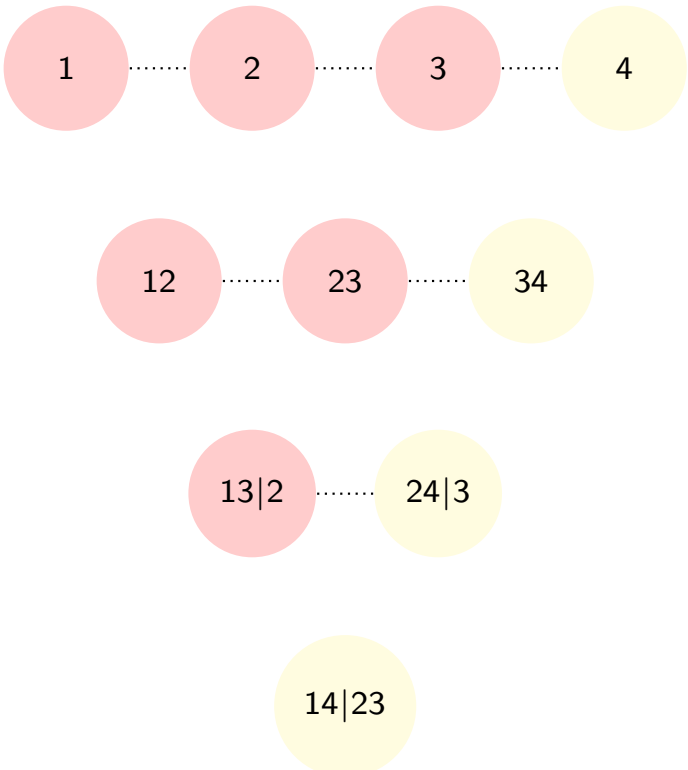
$$c_{12}\{F_1(x_1), F_2(x_2)\}$$
$$c_{13}\{F_1(x_1), F_3(x_3)\}$$
$$c_{14}\{F_1(x_1), F_4(x_4)\}$$

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

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

4. Strategies for constructing copula models

D-Vine with four variables



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

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

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

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

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

4. Strategies for constructing copula models

Research on PCCs is expanding quickly

Key references include

- K. Aas, C. Czado, A. Frigessi & H. Bakken (2009).
  Pair-copula constructions of multiple dependence.
  Insurance: Mathematics and Economics, 44, 182–198.
- I. Hobæk Haff, K. Aas & A. Frigessi (2010).
  On the simplified pair-copula construction - simply useful or too simplistic?
  J. Multivariate Anal., 101, 1296–1310.
- D. Kurowicka & H. Joe (2011).
  Dependence Modeling: Handbook on Vine Copulae.
  World Scientific Publishing.

Inference on PCCs is just beginning to emerge...

Acknowledgements

Funding in support of this work was provided by:

- the Canada Research Chairs Program;
- the Natural Sciences and Engineering Research Council of Canada;
- the Fonds québécois de la recherche sur la nature et les technologies.