A REVIEW OF COPULAS

by

Victor H. de la Peña

Department of Statistics

Columbia University

Theor. Appl. Climatol. (2008)
DOI 10.1007/s00704-007-0351-z
Printed in The Netherlands

Theoretical and Applied Climatology

[1] Department of Statistics, Columbia University, New York, USA
[2] Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York, USA

Detecting shifts in correlation and variability with application to ENSO-monsoon rainfall relationships

L. F. Robinson[1], V. H. de la Peña[1], Y. Kushnir[2]

With 3 Figures

Received 29 January 2007; Accepted 23 July 2007; Published online 6 February 2008

Summary

This paper addresses the retrospective detection of step changes at unknown time points in the correlation structure of two or more climate times series. Both the variance of individual series and the covariance between series are addressed. For a sequence of vector-valued observations with an approximate multivariate normal distribution, the proposed method is a parametric likelihood ratio test of the hypothesis of constant covariance against the hypothesis of at least one shift in covariance. The formulation of the test statistic and its asymptotic distribution are taken from Chen and Gupta (2000). This test is applied to the series comprised of the mean summer NINO3 index and the Indian monsoon rainfall index for the years 1871–2003. The most likely change point year was found to be 1980, with a resulting p-value of 0.12. The same test was applied to the series of NINO3 and Northeast Brazil rainfall observations from the years 1856–2001. A shift was detected in 1982 which is significant at the 1% level. Some or all of this shift in the covariance matrix can be attributed to a change in the variance of the Northeast Brazil rainfall. A variation of this methodology designed to increase power under certain multiple change point alternatives, specificallly when a shift is followed by a reversal, is also presented. Simulations to assess the power of the test under various alternatives are also included, in addition to a review of the literature on alternative methods.

Correspondence: Lucy F. Robinson, Department of Statistics, Columbia University, 1255 Amsterdam Ave. 10th flr. MC 4409, New York, NY 10027, USA, e-mail: lfr24@columbia.edu

1. Introduction

Assessing the stability over time of climate processes and the connections between them is crucial to our understanding of a changing climate. Changes in variability or connections between processes, if robust, can profoundly change our assessment of climate impacts and affect climate predictability. An area of great recent concern is the relationship between the Indian monsoon rainfall (IMR) and the El Niño/Southern Oscillation (ENSO) phenomenon. The existence of a significant negative correlation between time series has been long been observed (Walker and Bliss 1937), but whether the strength of the relationship has decreased in recent decades is a subject of current debate.

Running correlation analysis, in which correlations are computed in overlapping moving windows, has frequently been used in an attempt to document and understand changes in the correlation between two climate indices. In particular, the existence of low-frequency modes of variability is of current interest in many areas of climate research, and running correlations have been used to represent the multi-decadal evolution of the relationship between two processes.

INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. 27: 179–187 (2007)
Published online 10 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1394

RMetS
ROYAL METEOROLOGICAL SOCIETY

Indian summer monsoon rainfall and its link with ENSO and Indian Ocean climate indices

Chie Ihara,[a,*] Yochanan Kushnir,[a] Mark A. Cane[a] and Victor H. De La Peña[b]
[a] *Lamont-Doherty Earth Observatory of Columbia University, 61 Route 9W Palisades, NY 10964-8000, New York*
[b] *Department of Statistics, Columbia University, 1255 Amsterdam Avenue, NY 10027, New York*

Abstract:

We examine the relationship between the state of the equatorial Indian Ocean, ENSO, and the Indian summer monsoon rainfall using data from 1881 to 1998. The zonal wind anomalies and SST anomaly gradient over the equatorial Indian Ocean are used as indices that represent the condition of the Indian Ocean. Although the index defined by the zonal wind anomalies correlates poorly with Indian summer monsoon rainfall, the linear reconstruction of Indian summer monsoon rainfall on the basis of a multiple regression from the NINO3 and this wind index better specifies the Indian summer monsoon rainfall than the regression with only NINO3. Using contingency tables, we find that the negative association between the categories of Indian summer monsoon rainfall and the wind index is significant during warm years (El Niño) but not during cold years (La Niña). Composite maps of land precipitation also indicate that this relationship is significant during El Niño events. We conclude that there is a significant negative association between Indian summer monsoon rainfall and the zonal wind anomalies over the equatorial Indian Ocean during El Niño events. A similar investigation of the relationship between the SST index and Indian summer monsoon rainfall does not reveal a significant association. Copyright © 2006 Royal Meteorological Society

KEY WORDS Indian summer monsoon rainfall; ENSO; Indian Ocean Dipole Mode

Received 25 June 2005; Revised 25 March 2006; Accepted 27 June 2006

1. INTRODUCTION

The all-India summer monsoon rainfall (hereafter referred to as ISMR, Sontakke *et al.*, 1993), is defined as the rainfall received during the summer monsoon season (June, July, August, and September) over India. The ISMR has a large impact on the agriculture and related economic activities of the region, and prediction of the interannual variability of ISMR is thus a matter of great concern to society. Researchers have been working on this problem since the late 1800s. El Niño and Southern Oscillation (ENSO) has been known to exert the most important external forcing on ISMR (e.g., Krishna Kumar *et al.*, 1999; Rasmusson and Carpenter, 1983; Webster and Yang, 1992; Ropelewski and Halpert, 1987; Lau and Nath, 2000; Wang *et al.*, 2003). As Walker found out long ago, the anomalous high pressure over the western Pacific–eastern Indian Ocean and anomalous low pressure over the eastern and central Pacific associated with El Niño, influence the monsoon circulation. Krishna Kumar *et al.* (1999), among others, suggest that El Niño/La Niña shifts the location of the tropical Walker circulation and brings about deficit/excess of rainfall by suppressing/enhancing the convection over the Indian region.

The canonical patterns of atmospheric and oceanic variables over the Indo Pacific regions during El Niño and La Niña events were described by Reason *et al.* (2000). They showed that when an El Niño occurred during the summer, the Indian Ocean was characterized by a slight warming of SST compared to normal, which was associated with weaker wind magnitudes than normal and reduced cloudiness. After the summer monsoon season, they found a clear influence of El Niño over the Indian Ocean; the SST over the entire basin was significantly warmer than normal. At this time, large negative wind speed anomalies around the equator were seen in the Indian Ocean, with an increase in cloudiness over the western Indian Ocean, and a decrease over the eastern Indian Ocean. They also demonstrated that the opposite configuration occurred during La Niña events.

The state of ENSO does not explain all the interannual variability of ISMR (Kripalani and Kulkarni, 1997). For example, in spite of the occurrence of strong El Niño events in 1914, 1963, 1976, 1983, and 1997, these years did not experience deficiencies in ISMR (Figure 1). Kripalani and Kulkarni (1997) pointed out the existence of the interdecadal variability of ISMR and found that when ISMR was in the above normal interdecadal phase,

*Correspondence to: Chie Ihara, Lamont-Doherty Earth Observatory of Columbia University, 61 Route 9W Palisades, NY 10964-8000, New York. E-mail: cihara@ldeo.columbia.edu

Journal of Banking & Finance xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Journal of Banking & Finance

journal homepage: www.elsevier.com/locate/jbf

International diversification: A copula approach

Lorán Chollete [a,*], Victor de la Peña [b], Ching-Chih Lu [c]

[a] University of Stavanger, Stavanger N-4036, Norway
[b] Columbia University, New York, NY 10027, USA
[c] National Chengchi University, 64 Sec. 2, Zhi-Nan Road, Taipei 116, Taiwan

ARTICLE INFO

Article history:
Received 29 June 2009
Accepted 18 August 2010
Available online xxxx

JEL classification:
C14
F30
G15

Keywords:
Diversification
Copula
Correlation complexity
Downside risk
Systemic risk

ABSTRACT

The viability of international diversification involves balancing benefits and costs. This balance hinges on the degree of asset dependence. In light of theoretical research linking diversification and dependence, we examine international diversification using two measures of dependence: correlations and copulas. We document several findings. First, dependence has increased over time. Second, we find evidence of asymmetric dependence or downside risk in Latin America, but less in the G5. The results indicate very little downside risk in East Asia. Third, East Asian and Latin American returns exhibit some correlation complexity. Interestingly, the regions with maximal dependence or worst diversification do not command large returns. Our results suggest international limits to diversification. They are also consistent with a possible tradeoff between international diversification and systemic risk.

1. Introduction

The net benefit of international diversification is of great importance in today's economic climate. In general, the tradeoff between diversification's benefits and costs hinges on the degree of dependence across securities, as observed by Samuelson (1967), Ibragimov et al. (2009b), Shin (2009), Veldkamp and Van Nieuwerburgh (2010), and Bai and Green (2010), among others. Economists and investors often assess diversification benefits using a measure of dependence, such as correlation.[1] It is therefore vital to have accurate measures of dependence. There are several measures available in finance, including the traditional correlation and copulas. While each approach has advantages and disadvantages, researchers have rarely compared them in the same empirical study.[2] Such reliance on one dependence measure prevents easy assessment of the degree of international diversification opportunities, and how they differ over time or across regions.

The main goal of this paper is to assess diversification opportunities available in international stock markets, using both correlations and copulas. The recent history of international markets is interesting in itself, due to the large number of financial crises, increasingly globalized markets, and financial contagion.[3] We also examine some basic implications for international asset pricing. In particular, we investigate whether the diversification measures are related to international stock returns. This research is valuable because considerations of diversification and dependence should affect risk premia.

A secondary focus of our paper is the relation between diversification and systemic risk. We motivate this aspect by theoretical research such as Brumelle (1974), Ibragimov et al. (2009b), and Shin (2009), and it concerns two separate, distributional properties: heavy tails and tail dependence. The term 'heavy tail' refers to the tail mass of the marginal, univariate distributions, while 'tail dependence' refers to the connection between marginal distributions at extreme quantiles.[4] While no general theoretical results link

* Corresponding author. Tel.: +47 5183 1500; fax: +47 5183 1550.
E-mail addresses: loran.g.chollete@uis.no (L. Chollete), vp@stat.columbia.edu (V. de la Peña), cclu@nccu.edu.tw (C.-C. Lu).

[1] See Solnik (1974), Ingersoll (1987, Chapter 4); Carrieri et al. (2008); You and Daigler (2010). Moreover, asset prices, which reflect their diversification benefits in equilibrium, are assessed using dependence or covariance. See research on CAPM and stochastic discount methods, such as Sharpe (1964), Lintner (1965), Lucas (1978), and Hansen and Singleton (1982).

[2] Throughout, we use the word dependence as an umbrella to cover any situation where two or more variables move together. We adopt this practice because there are numerous words in use (e.g. correlation, concordance, co-dependency, comovement), and we wish to use a general term. We do not assume that any dependence measure is ideal, and throughout we indicate advantages and disadvantages as the case may be.

[3] See Dungey and Tambakis (2005), Reinhart (2008), Reinhart and Rogoff (2009), Markwat et al. (2009), and Dungey et al. (2010).

[4] We formally define tail dependence and tail indices in Eqs. (5) and (9). Further, we estimate both heavy tails and dependence in Tables 8 and 9, and Table 11, respectively.

0378-4266/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbankfin.2010.08.020

From: Embrechts, McNeil and Straumann (1999)

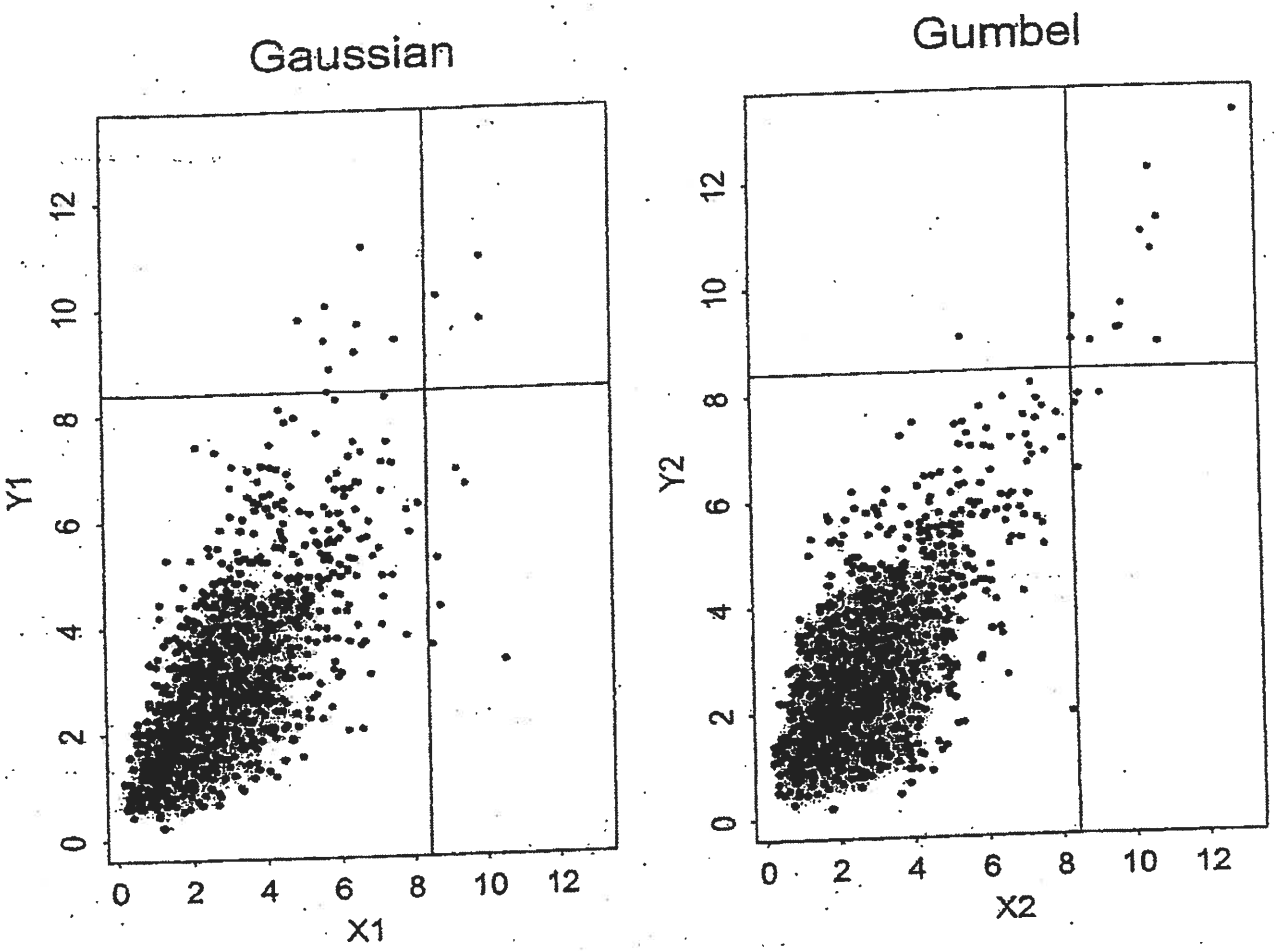


FIGURE 1. 1000 random variates from two distributions with identical Gamma(3,1) marginal distributions and identical correlation $\rho = 0.7$, but *different* dependence structures.

A review of Probability: Marginal and Joint Distributions

• **Assumptions** The random variables have continuous distributions and strictly increasing cumulative distribution functions and are real valued.

•**Definition of Probability Density Function $f_X(x)$ and cumulative distribution function (CDF) $F_X(x)$.** For a random variable **X**

$$F_X(x_a) = P(\mathbf{X} \leq x_a) = \int_{-\infty}^{x_a} f_X(x)dx.$$

• The density function and the cumulative distribution function are related by the formula $f_X(x) = \frac{d}{dx}F_X(x)$.

• It is common to drop the terms cumulative and function in the CDF and to speak generically by saying that a random variable has a given distribution.

Inverse Distribution Function and Value at Risk

Example (Value at Risk) For $0 < \alpha < 1$ the value at risk VaR_α is the value x_α for which $F_X(x_\alpha) = \alpha$. In statistics, x_α is called the $\alpha \cdot 100\%$ percentile of F_X

Since $F_X(x)$ is an increasing function in x, its inverse function $F_X^{-1}(x)$ exists and from the definition we get $x_\alpha = F_X^{-1}(\alpha)$.

Generating Random Variables

• The Uniform random variable is the key random variable in the development of copulas and the generation of variables.

Definition (Uniform Random Variable). Let U be a uniform random variable on the interval $[0, 1]$. Then, $f_U(u) = 1$ for $0 \leq u \leq 1$ and $f_U(u) = 0$ otherwise.

Example (CDF of U)

$$F_U(u) = \int_0^u f_U(x)dx = \int_0^u 1du = u$$

Simulating Random Variables

Let $\mathbf{U}$ be a uniform random variable on $[0,1]$. Let $\mathbf{X}$ be a random variable with distribution function $F_X(x)$.

- The random variable $\mathbf{Y} = F_X^{-1}(\mathbf{U})$ is a random variable with distribution function $F_X(x)$.

To see this observe that

$$F_Y(x_a) = P(\mathbf{Y} \leq x_a) = P(F_X^{-1}(\mathbf{U}) \leq x_a)$$

$$= P(\mathbf{U} \leq F_X(x_a)) = F_X(x_a).$$

And therefore the distribution of $F_X^{-1}(\mathbf{U})$ is the same as the distribution of $\mathbf{X}$.

• Key Fact

For any distribution function $F_X(x)$, the random variable $F_X(\mathbf{X})$ is a uniform random variable.

To verify the claim we use the properties of the inverse function.

$$P(F_X(\mathbf{X}) \leq u) = P(\mathbf{X} \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u.$$

Joint Distribution Functions

Given a vector of random variables $(\mathbf{X}, \mathbf{Y})$ there is an associated joint cumulative distribution function that describes all the probabilities associated with the outcomes of $(\mathbf{X}, \mathbf{Y})$. The joint distribution function of $(\mathbf{X}, \mathbf{Y})$ is given by

$$F_{X,Y}(x, y) = P(\mathbf{X} \leq x, \mathbf{Y} \leq y).$$

Associated to the joint cumulative distribution function there is a joint density function $f_{X,Y}(x, y) = \frac{d}{dxdy} F_{X,Y}(x, y)$. Then for any function $g(x, y)$

$$Eg(\mathbf{X}, \mathbf{Y}) = \int \int g(x, y) f_{X,Y}(x, y) dxdy.$$

Dependent Variables vs Independent Variables

- $\mathbf{X}$ and $\mathbf{Y}$ are said to be independent if for all x, y

$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

for all x, y. That is,

$$P(\mathbf{X} \leq x \text{ and } \mathbf{Y} \leq y) = P(\mathbf{X} \leq x)P(\mathbf{Y} \leq y).$$

- **Intuitive Interpretation.** Two variables are said to be independent if the availability of information on the outcome of one of them does not change the probabilities associated with the other variable.

This concept is more easily understood in the context of the probability of random events. Let $\mathbf{A}$ and $\mathbf{B}$ be two events of interest. For example, the event that the asset value $\mathbf{X}$ and the liabilities $\mathbf{Y}$ lie on specified intervals, $\mathbf{A} = (x_1 \leq \mathbf{X} \leq x_2)$ and $\mathbf{B} = (y_1 \leq \mathbf{Y} \leq y_2)$. The definition of independence then becomes. Two random sets are independent if

$$P(\mathbf{A} \text{ and } \mathbf{B}) = P(\mathbf{A})P(\mathbf{B}).$$

- Extension by measure theory

A Measure of Linear Dependence:

Pearson's Correlation Coefficient $\rho_P(X,Y)$

• $\rho_P(X,Y)$: The most widely used measure of dependence

• $\rho(\mathbf{X},\mathbf{Y})$ can take on all values between -1 and 1.

• $\rho_P(X,Y)$: a measure of linear association between variables

$\rho(\mathbf{X},\mathbf{Y}) = 1$ if one can find constants a, b so that $\mathbf{Y} = a\mathbf{X} + b$ with $a > 0$. If $a < 0$ $\rho(\mathbf{X},\mathbf{Y}) = -1$.

$$\rho_P(X,Y) = \frac{E(\mathbf{X} - E\mathbf{X})(\mathbf{Y} - E\mathbf{Y})}{\sqrt{Var(\mathbf{X})}\sqrt{Var\mathbf{Y}}}.$$

• If $(\mathbf{X},\mathbf{Y})$ is bivariate normal then $\rho_P(X,Y) = 0$ implies $\mathbf{X}$ and $\mathbf{Y}$ are independent.

• $\rho_P(X,Y)$: Its scope of applicability is frequently over-estimated.

If we go away from linear relationship to a quadratic relationship the picture changes. For example if $\mathbf{Y} = \mathbf{X}^2$ the knowledge of $\mathbf{X}$ completely determines the value of $\mathbf{Y}$ and hence they are fully dependent but $\rho_P(X,Y)$ is not equal to one.

• $\rho_P(X,Y) = 0$ does not in general imply that $\mathbf{X}$ and $\mathbf{Y}$ are independent.

Dependence and Risk Management

Example: Probability of Default

- $\mathbf{V}$ Asset value.

- $\mathbf{L}$ Liabilities.

- $P(\mathbf{V} \leq \mathbf{L})$ Probability of default.

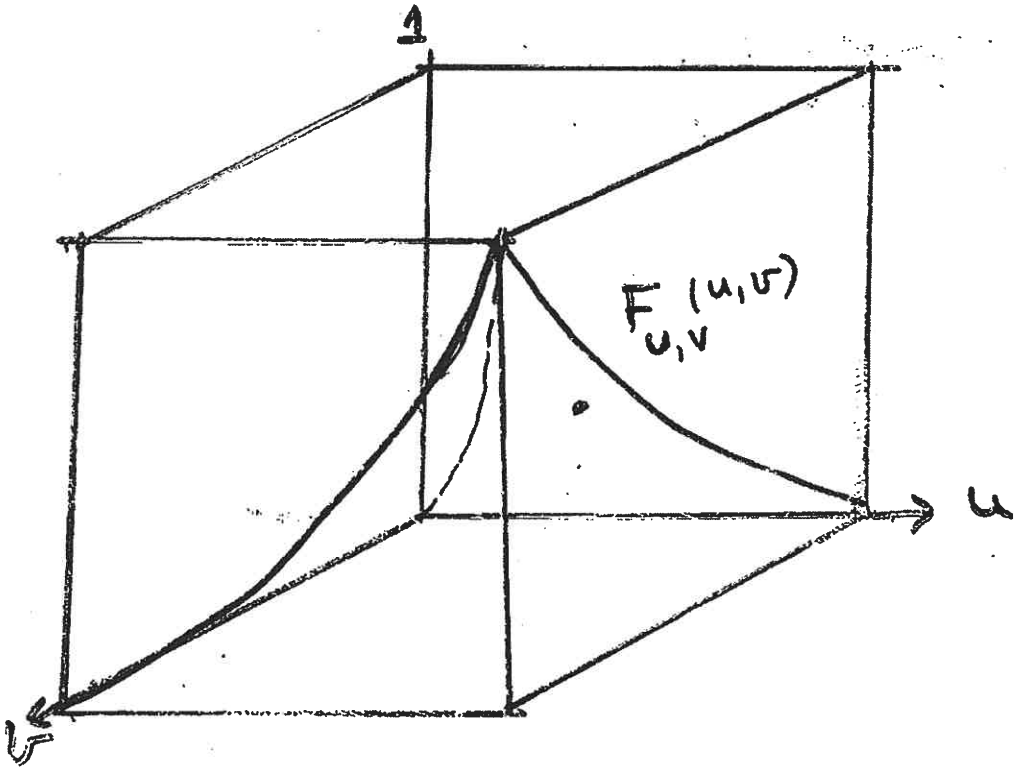
- Common Assumption: $\mathbf{V}$ and $\mathbf{L}$ are independent.

- If $\mathbf{V}$ and $\mathbf{L}$ are driven by a common factor $\mathbf{T}$ (e.g. currency or market fluctuations), the natural assumption is that they are dependent.

- The correlation coefficient frequently will not be a good measure of the dependence between $\mathbf{V}(\mathbf{T})$ and $\mathbf{L}(\mathbf{T})$ if $\mathbf{V}$ and $\mathbf{L}$ are driven by $\mathbf{T}$ in a non-linear way.

- The **copula** function is a function that captures (in a very general context) all the dependence in a vector of random variables when the marginal distributions are given.

Copula = Joint distribution of uniform random variables



$$F_{U,V}(u,v) = C_{U,V}(u,v)$$

$$P(U \le u, V \le v) = F_{U,V}(u,v)$$

What is a Copula?

Definition (Copula). A copula is a joint distribution of uniform random variables.

In the case of two variables **U**, **V** the copula is given by

$$C(u, v) = P(\mathbf{U} \leq u, \mathbf{V} \leq v) = F_{U,V}(u, v).$$

• Given any joint distribution of two random variables $F_{X,Y}(x, y) = P(\mathbf{X} \leq x, \mathbf{Y} \leq y)$ there is an associated copula.

Calculating The Copula of (X, Y)

Step 1. Find the marginal distributions of **X**, and **Y** $F_X(x)$ and $F_Y(y)$.

Step 2. Calculate the inverse transforms $F_X^{-1}(u)$ and $F_Y^{-1}(v)$.

Step 3. Plug in in joint CDF.

$$C(u, v) = F_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v)).$$

Let us check that this is indeed a copula.

$$F_{X,Y}(F_X^{-1}(u), F_Y^{-1}(v)) = P(\mathbf{X} \leq F_X^{-1}(u), \mathbf{Y} \leq F_Y^{-1}(v))$$

$$= P(F_X(\mathbf{X}) \leq u, F_Y(\mathbf{Y}) \leq v),$$

and as discussed earlier the random variables $F_X(\mathbf{X})$ and $F_Y(\mathbf{Y})$ are uniform random variables.

Properties of Copulas

1. The copula construction does not constrain the choice of marginal distributions.

2. The copula construction provides a way of obtaining joint distribution functions.

3. Sklar (1959) showed that any joint distribution function can be written in terms of a copula. If the marginals are continuous there is only one way of doing this.

4. If $(\mathbf{U}, \mathbf{V})$ has copula C then the vector $(F_X^{-1}(\mathbf{U}), F_Y^{-1}(\mathbf{V}))$ has a distribution $F(x, y)$ with marginals $F_X(x)$, $F_Y(y)$.

5. If $\mathbf{Y} = g(\mathbf{X})$ where g is a strictly de- creasing function then the copula is able to capture the relationship between $\mathbf{X}$ and $\mathbf{Y}$. More precisely, $C_{X,Y}(x, y) = \min\{x, y\}$.

Properties of Copulas

6. The joint density function can be obtained from the marginal distributions and and the copula in the following way.

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) C_{12}(F_X(x), F_Y(y)),$$

where $C_{12}(u,v) = \frac{d}{du}\frac{d}{dv} C_{X,Y}(u,v)$.

7. Schweizer and Wolf (1981) showed that the copula captures all the dependence contained in the joint distribution of $(\mathbf{X}, \mathbf{Y})$ in the sense that for any strictly increasing functions $g(x), h(y)$, the copula associated with $(g(\mathbf{X}), h(\mathbf{Y}))$ is the same as the copula associated with $(\mathbf{X}, \mathbf{Y})$.

- Moral

THE COPULA EXTRACTS THE WAY IN WHICH THE VARIABLES MOVE TOGETHER

NO MATTER WHAT SCALE IS USED TO MEASURE THEM

Copulas: Main Drawbacks

- The copula is an entire function, not a single number as the correlation.

- One needs to identify the appropriate copula to work with. This can be hard.

- The calculations required are frequently linked to the particular copula.

- The correlation coefficient ρ_P can not be calculated by using only the information provided by the copula.

The following transparencies will deal with methods of dealing with these problems, including a method for generating copulas, several examples of copulas and two new correlation coefficients that can be computed solely from the information contained by the copula.

A Method of Constructing $F(X, Y)$

Consider a fixed vector of uniform random variables $(\mathbf{U}, \mathbf{V})$. Then its distribution function is a copula and is given by $C(u, v) = P(\mathbf{U} \leq u, \mathbf{V} \leq v)$. To complete the construction pick arbitrary marginal distributions $F_X(x)$ and $F_Y(y)$. Then the function $C(F_X(x), F_Y(y))$ defines a bivariate distribution function with marginal distribution functions $F_X(x)$ and $F_Y(y)$ and

$$C(F_X(x), F_Y(y)) = F(x, y) = P(\mathbf{X} \leq x, \mathbf{Y} \leq y).$$

In order to construct a copula with particular marginals we could then follow the following steps.

- **Step 1.** Pick the marginal distributions of our problem. In the case of $\mathbf{V}$ (Asset Values) and $\mathbf{L}$ (Liabilities), $F_V(v)$ and $F_L(l)$.

- **Step 2.** Pick a particular dependence structure from a library of joint bivariate uniform distributions. Associated with this, is the copula $C(x, y)$.

- **Step 3.** Evaluate $C(F_V(v), F_L(l))$.

$C(F_V(v), F_L(l)) = F(v, l)$ is a cumulative distribution function with marginal probabilities $F_V(v) = P(V \leq v)$ and $F_L(l) = P(L \leq l)$.

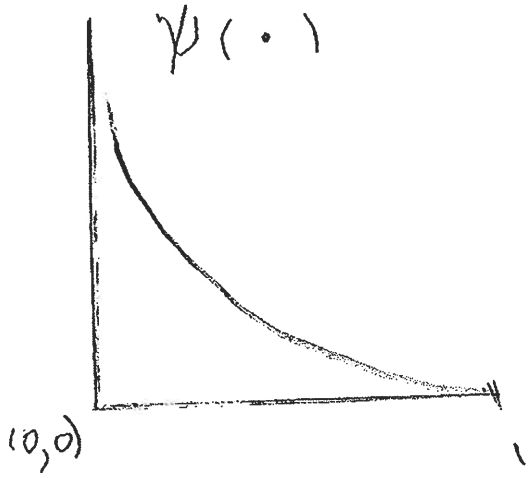
- Valdez and Frees (1997) present formal approaches to simulate and fit copulas based on empirical data and the above construction. The main difficulty arises in Step 2 which is considered next.

Examples of Copulas

1. If $\mathbf{X}$ and $\mathbf{Y}$ are independent then $C_{X,Y}(u,v) = u \cdot v$

2. Gaussian Copula. Let $(\mathbf{X}, \mathbf{Y})$ be a bivariate standard normal random vector with correlation ρ. Then the Gaussian copula is given by $C(u,v) = P(\Phi(\mathbf{X}) \leq u, \Phi(\mathbf{Y}) \leq v) = P(\mathbf{X} \leq \Phi^{-1}(u), \mathbf{Y} \leq \Phi^{-1}(v))$ where $\Phi(x) = \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{x} \exp\{-u^2/2\} du$.

3. Archimedean Copulas. A method of generating copulas on the basis of a single function. Let ψ be a function taking $[0,1] \to [0,\infty)$ such that a) ψ is convex and strictly decreasing b) $\psi(1) = 0$. Then, $C(u,v) = \psi^{-1}(\psi(u) + \psi(v))$ is a copula.



From Nelsen (1999)

Table 4.1. One-parameter

(4.2.#)	$C_\theta(u,v)$	$\varphi_\theta(t)$
1	$\max\left(\left[u^{-\theta}+v^{-\theta}-1\right]^{-1/\theta},0\right)$	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$
2	$\max\left(1-\left[(1-u)^\theta+(1-v)^\theta\right]^{1/\theta},0\right)$	$(1-t)^\theta$
3	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\ln\frac{1-\theta(1-t)}{t}$
✓ 4	$\exp\left(-\left[(-\ln u)^\theta+(-\ln v)^\theta\right]^{1/\theta}\right)$	$(-\ln t)^\theta$
5	$-\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$-\ln\frac{e^{-\theta t}-1}{e^{-\theta}-1}$
6	$1-\left[(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta\right]^{1/\theta}$	$-\ln\left[1-(1-t)^\theta\right]$
7	$\max(\theta uv+(1-\theta)(u+v-1),0)$	$-\ln[\theta t+(1-\theta)]$
8	$\max\left[\frac{\theta^2uv-(1-u)(1-v)}{\theta^2-(\theta-1)^2(1-u)(1-v)},0\right]$	$\frac{1-t}{1+(\theta-1)t}$
9	$uv\exp(-\theta\ln u\ln v)$	$\ln(1-\theta\ln t)$
10	$uv/\left[1+(1-u^\theta)(1-v^\theta)\right]^{1/\theta}$	$\ln\left(2t^{-\theta}-1\right)$
11	$\max\left(\left[u^\theta v^\theta-2(1-u^\theta)(1-v^\theta)\right]^{1/\theta},0\right)$	$\ln\left(2-t^\theta\right)$
12	$\left(1+\left[(u^{-1}-1)^\theta+(v^{-1}-1)^\theta\right]^{1/\theta}\right)^{-1}$	$\left(\frac{1}{t}-1\right)^\theta$
13	$\exp\left(1-\left[(1-\ln u)^\theta+(1-\ln v)^\theta-1\right]^{1/\theta}\right)$	$(1-\ln t)^\theta-1$
14	$\left(1+\left[(u^{-1/\theta}-1)^\theta+(v^{-1/\theta}-1)^\theta\right]^{1/\theta}\right)^{-\theta}$	$(t^{-1/\theta}-1)^\theta$

θ is usually estimated using ρ_K / ρ_S

From Chollete, de la Peña and Lu (2010)

Table 1: Distribution of Various Copulas

Copula	Distribution	Parameter Range	Complete Dependence	Independence
Normal	$C_N(u,v;\rho) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$	$\rho \in (-1,1)$	$\rho = 1$, or -1	$\rho = 0$
Student-t	$C_t(u,v;\rho,d) = t_{d,\rho}(t_d^{-1}(u), t_d^{-1}(v))$	$\rho \in (-1,1)$	$\rho = 1$,or -1	$\rho = 0$
Gumbel	$C_G(u,v;\beta) = \exp\{-[(-\ln(u))^{1/\beta} + (-\ln(v))^{1/\beta}]^\beta\}$	$\beta \in (0,1)$	$\beta = 0$	$\beta = 1$
RG	$C_{RG}(u,v;\alpha) = u + v - 1 + C_G(1-u, 1-v;\alpha)$	$\alpha \in (0,1)$	$\alpha = 0$	$\alpha = 1$
Clayton	$C_C(u,v;\theta) = \max\left((u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, 0\right)$	$\theta \in [-1,\infty) \setminus \{0\}$	$\theta \to \infty$	$\theta \to 0$
RC	$C_{RC}(u,v;\theta) = u + v - 1 + C_C(1-u, 1-v, \theta)$	$\theta \in [-1,\infty) \setminus \{0\}$	$\theta \to \infty$	$\theta \to 0$

RG and RC denote the Rotated Gumbel and Rotated Clayton copulas, respectively. The symbols $\Phi_\rho(x,y)$ and $t_{\nu,\rho}(x,y)$ denote the standard bivariate normal and Student-t cumulative distributions, respectively: $\Phi_\rho(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{1}{2\pi|\Sigma|}\exp\{-\frac{1}{2}(x\ y)\Sigma^{-1}(x\ y)'\}dxdy$, and $t_{\nu,\rho}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\nu/2)(\nu\pi)|\Sigma|^{1/2}}\{1 + (s\ t)\Sigma^{-1}(s\ t)'/\nu\}^{\frac{-(\nu+2)}{2}}dsdt$. The correlation matrix is given by $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Measures of Dependence Using Copulas

Spearman's $\rho_S(X,Y)$ and Kendall's $\rho_K(X,Y)$.

These coefficients are not affected by non-linear increasing transformations of X and Y unlike Pearson's correlation coefficient. Moreover, they can be computed using the copula associated to the vector $(\mathbf{X}, \mathbf{Y})$.

Spearman's Coefficient

$$\rho_S(X,Y) = 12E\{(F_X(\mathbf{X}) - \frac{1}{2})(F_Y(\mathbf{Y}) - \frac{1}{2})\}$$

$$= 12 \int \int \{C(u,v) - uv\} du \cdot dv$$

Kendall's Coefficient

Let $(\mathbf{X}, \mathbf{Y})$ and $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ be i.i.d. vectors with distribution function $F_{X,Y}$. Then,

$$\rho_K(X,Y) = P((\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{Y} - \tilde{\mathbf{Y}}) > 0) - P((\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{Y} - \tilde{\mathbf{Y}}) < 0).$$

Kendall's coefficient can be written in terms of the copula as well

$$\rho_K(X,Y) = 4 \int_0^1 \int_0^1 C(u,v) dC(u,v) - 1$$

Properties of Spearman's and Kendall's Coefficients

• $\rho_S(\mathbf{X}, \mathbf{Y})$ and $\rho_K(X, Y) = 1$ whenever $\mathbf{Y} = g(\mathbf{X})$ for a strictly increasing function g. (e.g. $\mathbf{Y} = \mathbf{X}^2$).

• ρ_S and $\rho_K = -1$ whenever $\mathbf{Y} = h(\mathbf{X})$ for a strictly decreasing function h.

• For arbitrary strictly increasing functions g_1, g_2

$$\rho_S(X, Y) = \rho_S(g_1(X), g_2(Y)),$$

and

$$\rho_K(X, Y) = \rho_K(g_1(X), g_2(Y)).$$

• Since Pearson's correlation $\rho_P(X, Y)$ depends both on the joint distribution (and hence the copula) and the marginal distributions, Pearson's correlation coefficient does not have the above property.

• Recall the example when asset value $\mathbf{V}$ and liabilities $\mathbf{L}$ are driven by a common factor $\mathbf{T}$ we see that the third property above could be helpful.

From de la Peña, Collete and Lu (2010)

Table 10: Tail Dependence and Kendall's τ for various Copulas

	Left Tail Dep	Right Tail Dep	Kendall's τ
Gaussian	0	0	$\frac{2}{\pi}\arcsin\rho$
Student-t	$2t_{d+1}\left(-\sqrt{\frac{(d+1)(1-\rho)}{1+\rho}}\right)$	$2t_{d+1}\left(-\sqrt{\frac{(d+1)(1-\rho)}{1+\rho}}\right)$	$\frac{2}{\pi}\arcsin\rho$
Gumbel	0	$2-2^{\alpha}$	$1-\alpha$
R. Gumbel	$2-2^{\alpha}$	0	$1-\alpha$
Clayton	0	$2^{-\frac{1}{\theta}}$	$\frac{\theta}{\theta+2}$
R. Clayton	$2^{-\frac{1}{\theta}}$	0	$\frac{\theta}{\theta+2}$

The table presents analytical formulas for tail dependence and Kendall's τ. Further information may be obtained from Chapter 5 of Embrechts et al. (2005).
R. Clayton and R. Gumbel denote the Rotated Clayton and Rotated Gumbel copulas. α and θ denote dependence parameters of the Gumbel and Clayton copulas, and d denotes the degrees of freedom of the Student-t copula.

$$d(u) = P(F_X(X) \le u \mid F_Y(Y) \le u)$$

$$= \frac{C(u,u)}{u}$$

$$\lambda_L = \lim_{\lambda \to 0^+} d(u)$$

(Left tail dependence)

Journal of Banking & Finance xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Journal of Banking & Finance

journal homepage: www.elsevier.com/locate/jbf

International diversification: A copula approach

Lorán Chollete [a,*], Victor de la Peña [b], Ching-Chih Lu [c]

[a] University of Stavanger, Stavanger N-4036, Norway
[b] Columbia University, New York, NY 10027, USA
[c] National Chengchi University, 64 Sec. 2, Zhi-Nan Road, Taipei 116, Taiwan

The main goal of this paper is to assess diversification opportunities available in international stock markets, using both correlations and copulas. The recent history of international markets is interesting in itself, due to the large number of financial crises, increasingly globalized markets, and financial contagion.[3] We also examine some basic implications for international asset pricing. In particular, we investigate whether the diversification measures are related to international stock returns. This research is valuable because considerations of diversification and dependence should affect risk premia.

1. Introduction

The net benefit of international diversification is of great importance in today's economic climate. In general, the tradeoff between diversification's benefits and costs hinges on the degree of dependence across securities, as observed by Samuelson (1967), Ibragimov et al. (2009b), Shin (2009), Veldkamp and Van Nieuwerburgh (2010), and Bai and Green (2010), among others. Economists and investors often assess diversification benefits using a measure of dependence, such as correlation.[1] It is therefore vital to have accurate measures of dependence. There are several measures available in finance, including the traditional correlation and copulas. While each approach has advantages and disadvantages, researchers have rarely compared them in the same empirical study.[2] Such reliance on one dependence measure prevents easy assessment of the degree of international diversification opportunities, and how they differ over time or across regions.

The main goal of this paper is to assess diversification opportunities available in international stock markets, using both correlations and copulas. The recent history of international markets is interesting in itself, due to the large number of financial crises, increasingly globalized markets, and financial contagion.[3] We also examine some basic implications for international asset pricing. In particular, we investigate whether the diversification measures are related to international stock returns. This research is valuable because considerations of diversification and dependence should affect risk premia.

A secondary focus of our paper is the relation between diversification and systemic risk. We motivate this aspect by theoretical research such as Brumelle (1974), Ibragimov et al. (2009b), and Shin (2009), and it concerns two separate, distributional properties: heavy tails and tail dependence. The term 'heavy tail' refers to the tail mass of the marginal, univariate distributions, while 'tail dependence' refers to the connection between marginal distributions at extreme quantiles.[4] While no general theoretical results link

* Corresponding author. Tel.: +47 5183 1500; fax: +47 5183 1550.
E-mail addresses: loran.g.chollete@uis.no (L. Chollete), vp@stat.columbia.edu (V. de la Peña), cclu@nccu.edu.tw (C.-C. Lu).

[1] See Solnik (1974), Ingersoll (1987, Chapter 4); Carrieri et al. (2008); You and Daigler (2010). Moreover, asset prices, which reflect their diversification benefits in equilibrium, are assessed using dependence or covariance. See research on CAPM and stochastic discount methods, such as Sharpe (1964), Lintner (1965), Lucas (1978), and Hansen and Singleton (1982).

[2] Throughout, we use the word dependence as an umbrella to cover any situation where two or more variables move together. We adopt this practice because there are numerous words in use (e.g. correlation, concordance, co-dependency, comovement), and we wish to use a general term. We do not assume that any dependence measure is ideal, and throughout we indicate advantages and disadvantages as the case may be.

[3] See Dungey and Tambakis (2005), Reinhart (2008), Reinhart and Rogoff (2009), Markwat et al. (2009), and Dungey et al. (2010).

[4] We formally define tail dependence and tail indices in Eqs. (5) and (9). Further, we estimate both heavy tails and dependence in Tables 8 and 9, and Table 11, respectively.

doi:10.1016/j.jbankfin.2010.08.020

Table 11: Tail Dependence, Kendall's τ from Different Copula Models

Panel A: G5

	Left Tail Dependence			Right Tail Dependence			Kendall's τ		
Model	Average	Max	Min	Average	Max	Min	Average	Max	Min
Gumbel	0.0000			0.4016	0.6177 (FR-DE)	0.2344 (JP-US)	0.3279	0.5329 (FR-DE)	0.1798 (JP-US)
R. Gumbel	0.4200	0.6398 (FR-DE)	0.2434 (JP-US)	0.0000			0.3447	0.5562 (FR-DE)	0.1872 (JP-US)
Clayton	0.4238	0.6949 (FR-DE)	0.1680 (JP-US)	0.0000			0.3000	0.4878 (FR-DE)	0.1627 (JP-US)
R. Clayton	0.0000			0.3579	0.6383 (FR-DE)	0.1359 (JP-US)	0.2625	0.4357 (FR-DE)	0.1479 (JP-US)
Normal	0.0000			0.0000			0.3534	0.5491 (FR-DE)	0.1949 (JP-US)
Student t	0.1276	0.4658 (FR-DE)	0.0036 (DE-US)	0.1276	0.4658 (FR-DE)	0.0036 (DE-US)	0.3591	0.5625 (FR-DE)	0.2000 (JP-US)

Panel B: East Asia

	Left Tail Dependence			Right Tail Dependence			Kendall's τ		
Models	Average	Max	Min	Average	Max	Min	Average	Max	Min
Gumbel	0.0000			0.2868	0.4147 (HK-SI)	0.2031 (TW-TH)	0.2242	0.3353 (HK-SI)	0.1545 (TW-TH)
R. Gumbel	0.3058	0.4449 (HK-SI)	0.2261 (TW-TH)	0.0000			0.2403	0.3630 (HK-SI)	0.1731 (TW-TH)
Clayton	0.2681	0.4773 (HK-SI)	0.1581 (TW-TH)	0.0000			0.2103	0.3191 (HK-SI)	0.1582 (TW-TH)
R. Clayton	0.0000			0.2681	0.4773 (HK-SI)	0.1581 (TW-TH)	0.1804	0.2603 (HK-SI)	0.1248 (TW-TH)
Normal	0.0000			0.0000			0.2499	0.3577 (HK-SI)	0.1836 (TW-TH)
Student t	0.0557	0.2246 (HK-SI)	0.0031 (KR-TH)	0.0557	0.2246 (HK-SI)	0.0031 (KR-TH)	0.2523	0.3678 (HK-SI)	0.1835 (TW-TH)

Panel C: Latin America

	Left Tail Dependence			Right Tail Dependence			Kendall's τ		
Models	Average	Max	Min	Average	Max	Min	Average	Max	Min
Gumbel	0.0000			0.3075	0.3615 (AR-ME)	0.2562 (AR-CH)	0.2411	0.2877 (AR-ME)	0.1978 (AR-CH)
R. Gumbel	0.3447	0.3912 (BR-ME)	0.2897 (AR-CH)	0.0000			0.2733	0.3140 (BR-ME)	0.2258 (AR-CH)
Clayton	0.3526	0.4330 (BR-ME)	0.2769 (AR-CH)	0.0000			0.2503	0.2928 (BR-ME)	0.2125 (AR-CH)
R. Clayton	0.0000			0.2144	0.3080 (AR-ME)	0.1418 (AR-CH)	0.1839	0.2274 (AR-ME)	0.1507 (AR-CH)
Normal	0.0000			0.0000			0.2719	0.3123 (AR-ME)	0.2331 (AR-CH)
Student t	0.1207	0.1527 (AR-BR)	0.0617 (AR-CH)	0.1207	0.1527 (AR-BR)	0.0617 (AR-CH)	0.2723	0.3167 (AR-ME)	0.2267 (AR-CH)

Maximum Likelihood Estimation

Remark: Let $(\mathbf{X}, \mathbf{Y})$ be a vector of random variables with cumulative distribution function $F_{X,Y}(x, y)$. One can recover the associated joint density function of $f_{X,Y}(x, y)$ from the knowledge of the marginal density functions and the copula $C_{X,Y}(u, \ )$ associated to $(\mathbf{X}, \mathbf{Y})$. This remark is important in developing maximum likelyhood estimators.

Let $C_{12}(u, v) = \frac{d}{\ } \frac{d}{\ } C_{X,Y}(\ , \ )$. Then,

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) C_{12}(F_X(x), F_Y(y)).$$

If $(\mathbf{x}_i, \mathbf{x}_i)$ are independent realizations of the vector (X, Y) with fixed copula C then the log-likelihood for this sequence is

$$\sum_{i=1}^{n} \log(f_X(x_i) f_Y(y_i) C_{12}(F_X(x_i), F_Y(y_i))).$$

Table 4: Comparing Dependence Structures using Information Criteria

Panel A: G5

Models	AIC	BIC
Gumbel	-269.17	-264.44
Rotated Gumbel	-312.37	-307.64
Clayton	-275.46	-270.73
Rotated Clayton	-206.26	-201.53
Normal	-302.82	-298.10
Student t	-316.20	-306.75
Mixed Copula	-318.18	-294.57

Panel B: East Asia

Models	AIC	BIC
Gumbel	-111.25	-106.53
Rotated Gumbel	-139.43	-134.71
Clayton	-122.70	-117.98
Rotated Clayton	-87.31	-82.59
Normal	-132.38	-127.66
Student t	-138.47	-129.02
Mixed Copula	-138.98	-115.36

Panel C: Latin America

Models	AIC	BIC
Gumbel	-121.23	-116.51
Rotated Gumbel	-183.97	-179.25
Clayton	-171.26	-166.54
Rotated Clayton	-86.50	-81.78
Normal	-153.02	-148.30
Student t	-167.56	-158.12
Mixed Copula	-179.22	-155.61

AIC and BIC are the average Akaike and Bayes Information Criteria for countries in each region.

Table 5: Comparing Dependence Structures using Mixture Weights

Weights	G5	East Asia	Latin America
W_{Gumbel}	0.097 (0.085)	0.145 (0.102)	0.099 (0.084)
$W_{\text{R. Gumbel}}$	0.517 (0.170)	0.384 (0.147)	0.787 (0.160)
W_{Normal}	0.386 (0.177)	0.471 (0.196)	0.114 (0.161)

W_i denotes the average weight on copula i in each region, where i = Gumbel, Rotated Gumbel (R. Gumbel), and normal. The average standard deviation of weights for each region is in parentheses.

Table 6: Comparing Dependence Structures using Likelihood Methods

A. G-5 Countries

	FR-DE	FR-JP	FR-UK	FR-US	DE-JP	DE-UK	DE-US	JP-UK	JP-US	UK-US
Normal vs. Clayton	-1.05	0.14	-2.72	-2.79	1.16	-0.71	-2.33	0.34	-0.57	-2.43
Normal vs. R. Clayton	-6.49	-4.36	-6.25	-4.30	-4.90	-7.01	-5.21	-4.21	-2.25	-4.91
Normal vs. Gumbel	-1.75	-3.19	-3.00	-2.50	-3.10	-3.37	-3.49	-2.39	-0.61	-3.00
Normal vs. R. Gumbel	1.89**	0.73	-0.02	-0.66	1.28*	1.18	-0.52	0.71	0.27	-0.65
Normal vs. *t*	0.00	0.44	0.19	0.88	0.13	0.05	0.25	0.11	0.19	0.62
Normal vs. Mixed	3.34**	1.16	1.85**	1.32*	1.52*	2.44**	1.19	1.46*	1.16	0.82
t vs. Clayton	-0.01	-0.24	-0.98	-2.94	0.07	-0.09	-2.44	-0.07	-0.29	-2.68
t vs. R. Clayton	-0.01	-3.68	-1.96	-4.50	-0.90	-0.37	-5.12	-0.57	-0.57	-5.01
t vs. Gumbel	-1.75	-3.19	-3.00	-2.50	-3.10	-3.37	-3.49	-2.39	-0.61	-3.00
t vs. R. Gumbel	0.00	0.22	-0.18	-0.86	0.14	0.00	-0.57	0.00	-0.12	-0.82
t vs. Mixed	0.00	0.43	0.07	1.06	0.14	0.03	1.17	0.04	0.02	0.50

B. Asian Countries

	HK-KR	HK-SI	HK-TW	HK-TH	KR-SI	KR-TW	KR-TH	SI-TW	SI-TH	TW-TH
Normal vs. Clayton	-1.39	0.12	-0.46	-0.25	-1.32	-0.93	-1.99	-0.28	-1.36	-0.02
Normal vs. R. Clayton	-3.55	-5.07	-3.50	-3.42	-3.31	-2.84	-2.88	-3.30	-3.46	-2.83
Normal vs. Gumbel	-2.71	-2.50	-2.62	-2.09	-2.46	-2.00	-2.11	-2.00	-1.77	-2.63
Normal vs. R. Gumbel	-0.25	1.75**	0.53	0.98	-0.28	0.17	-0.61	0.43	0.49	0.51
Normal vs. *t*	0.55	0.02	0.69	0.09	0.63	0.89	0.65	0.77	0.07	0.66
Normal vs. Mixed	0.94	2.61**	1.58*	1.95**	0.65	1.14	0.45	1.10	1.90**	0.98
t vs. Clayton	-1.60	-0.02	-0.81	-0.11	-1.53	-1.29	-2.18	-0.55	-0.16	-0.31
t vs. R. Clayton	-3.76	-0.08	-3.68	-0.39	-3.42	-3.09	-3.07	-3.78	-0.30	-3.04
t vs. Gumbel	-2.71	-2.50	-2.62	-2.09	-2.46	-2.00	-2.11	-2.00	-1.77	-2.63
t vs. R. Gumbel	-0.38	0.00	0.25	0.01	-0.39	-0.05	-0.73	0.18	-0.03	0.36
t vs. Mixed	0.78	0.01	1.11	0.06	0.36	0.79	0.13	0.73	0.02	0.81

C. Latin American Countries

	AR-BR	AR-CH	AR-ME	BR-CH	BR-ME	CH-ME
Normal vs. Clayton	1.53*	1.44*	-0.45	0.68	2.10**	1.78**
Normal vs. R. Clayton	-4.35	-4.53	-4.21	-3.84	-6.03	-4.61
Normal vs. Gumbel	-2.41	-3.34	-2.67	-2.58	-4.89	-3.54
Normal vs. R. Gumbel	1.91**	1.28*	1.04	1.49*	2.38**	1.85**
Normal vs. *t*	0.01	0.13	0.08	0.03	0.08	0.04
Normal vs. Mixed	2.27**	1.32*	1.97**	1.97**	2.53**	2.00**
t vs. Clayton	0.00	0.15	-0.11	-0.02	0.14	0.02
t vs. R. Clayton	-0.04	-1.05	-0.46	-0.13	-0.66	-0.17
t vs. Gumbel	-2.41	-3.34	-2.67	-2.58	-4.89	-3.54
t vs. R. Gumbel	0.01	0.20	0.03	0.01	0.18	0.04
t vs. Mixed	0.01	0.20	0.07	0.02	0.19	0.04

Test statistics are generated using the pseudo-likelihood ratio test of Chen and Fan (2006). * and ** denote significance at the 10% and 5% levels, respectively. R. Gumbel and R. Clayton represent the Rotated Gumbel and Rotated Clayton copulas, respectively.

Chen and Fan (2006)

Table 8: Tail Index Measured by the Hill Estimator

	Left Tail			Right Tail		
	5%	7.5%	10%	5%	7.5%	10%
FR	2.78	2.30	2.45	3.09	3.29	3.17
	(0.43)	(0.29)	(0.27)	(0.48)	(0.42)	(0.35)
DE	2.76	2.33	2.15	3.36	3.18	3.12
	(0.43)	(0.30)	(0.24)	(0.52)	(0.40)	(0.34)
JP	4.00	3.14	2.80	3.16	2.72	2.44
	(0.62)	(0.40)	(0.31)	(0.49)	(0.34)	(0.27)
UK	3.09	3.22	2.76	3.64	3.04	3.15
	(0.48)	(0.41)	(0.30)	(0.56)	(0.39)	(0.35)
US	3.31	3.05	2.25	3.48	3.00	2.37
	(0.51)	(0.39)	(0.25)	(0.54)	(0.38)	(0.26)
HK	2.42	2.17	2.07	3.82	3.14	3.39
	(0.37)	(0.28)	(0.23)	(0.59)	(0.40)	(0.37)
KR	2.86	2.60	2.49	2.79	2.53	2.58
	(0.44)	(0.33)	(0.27)	(0.43)	(0.32)	(0.28)
SI	2.79	2.11	2.21	3.71	2.97	2.62
	(0.43)	(0.27)	(0.24)	(0.57)	(0.38)	(0.29)
TW	2.67	2.80	2.59	2.62	2.62	2.43
	(0.41)	(0.36)	(0.28)	(0.40)	(0.33)	(0.27)
TH	3.44	2.69	2.08	3.14	3.16	2.37
	(0.53)	(0.34)	(0.23)	(0.48)	(0.40)	(0.26)
AR	3.51	2.92	2.55	3.18	2.52	2.08
	(0.54)	(0.37)	(0.28)	(0.49)	(0.32)	(0.23)
BR	2.26	2.36	1.95	3.00	2.60	2.79
	(0.35)	(0.30)	(0.21)	(0.46)	(0.33)	(0.31)
CH	2.92	2.74	2.65	3.23	2.99	2.41
	(0.45)	(0.35)	(0.29)	(0.50)	(0.38)	(0.26)
ME	2.62	2.50	2.26	2.94	2.70	2.41
	(0.40)	(0.32)	(0.25)	(0.45)	(0.34)	(0.26)

The table presents estimates of right and left tail indices for each series, corresponding to the 5%, 7.5%, and 10% most extreme observations in the distribution. The tail index is estimated using the non-parametric estimator of Hill (1975). Standard errors, in parentheses, are calculated using the asymptotic variance of the Hill estimator, and obtained by the Delta method.

Table 9: Tail Index Measured by OLS log-log rank-size regression

	Left Tail			Right Tail		
	5%	7.5%	10%	5%	7.5%	10%
FR	3.48	3.00	2.77	3.21	3.21	3.20
	(0.76)	(0.54)	(0.43)	(0.70)	(0.58)	(0.50)
DE	3.61	3.14	2.78	3.63	3.49	3.39
	(0.79)	(0.56)	(0.43)	(0.79)	(0.63)	(0.53)
JP	5.11	4.32	3.71	3.67	3.44	3.10
	(1.12)	(0.78)	(0.58)	(0.80)	(0.62)	(0.48)
UK	3.84	3.51	3.30	3.28	3.33	3.29
	(0.84)	(0.63)	(0.51)	(0.72)	(0.60)	(0.51)
US	3.79	3.51	3.12	3.89	3.57	3.15
	(0.83)	(0.63)	(0.48)	(0.85)	(0.64)	(0.49)
HK	3.26	2.74	2.52	4.44	4.05	3.76
	(0.71)	(0.49)	(0.39)	(0.97)	(0.73)	(0.58)
KR	2.78	2.74	2.71	3.77	3.14	2.94
	(0.61)	(0.49)	(0.42)	(0.82)	(0.56)	(0.46)
SI	3.13	2.78	2.52	3.71	3.64	3.34
	(0.68)	(0.50)	(0.39)	(0.81)	(0.65)	(0.52)
TW	3.17	3.02	2.88	3.33	3.06	2.89
	(0.69)	(0.54)	(0.45)	(0.73)	(0.55)	(0.45)
TH	4.33	3.57	3.01	3.34	3.28	3.05
	(0.94)	(0.64)	(0.47)	(0.73)	(0.59)	(0.47)
AR	3.73	3.40	3.21	3.50	3.18	2.80
	(0.81)	(0.61)	(0.50)	(0.76)	(0.57)	(0.43)
BR	2.88	2.60	2.46	4.01	3.28	3.06
	(0.63)	(0.47)	(0.38)	(0.87)	(0.59)	(0.47)
CH	3.23	3.06	2.98	3.75	3.49	3.19
	(0.71)	(0.55)	(0.46)	(0.82)	(0.63)	(0.49)
ME	2.83	2.70	2.59	3.49	3.21	2.94
	(0.62)	(0.49)	(0.40)	(0.76)	(0.58)	(0.46)

The table presents estimates of right and left tail indices for each series, corresponding to the 5%, 7.5%, and 10% most extreme observations in the distribution. The tail index is estimated using the log log rank-size estimator of Gabaix and Ibragimov (2009). Standard deviations are in parentheses.

IMS Lecture Notes–Monograph Series
2nd Lehmann Symposium – Optimality
Vol. 49 (2006) 183–209

DOI: 10.1214/074921706000000455

Characterizations of joint distributions, copulas, information, dependence and decoupling, with applications to time series

Victor H. de la Peña[1,*], Rustam Ibragimov[2,†] and Shaturgun Sharakhmetov[3]

Columbia University, Harvard University and Tashkent State Economics University

Abstract: In this paper, we obtain general representations for the joint distributions and copulas of arbitrary dependent random variables absolutely continuous with respect to the product of given one-dimensional marginal distributions. The characterizations obtained in the paper represent joint distributions of dependent random variables and their copulas as sums of U-statistics in independent random variables. We show that similar results also hold for expectations of arbitrary statistics in dependent random variables. As a corollary of the results, we obtain new representations for multivariate divergence measures as well as complete characterizations of important classes of dependent random variables that give, in particular, methods for constructing new copulas and modeling different dependence structures.

The results obtained in the paper provide a device for reducing the analysis of convergence in distribution of a sum of a double array of dependent random variables to the study of weak convergence for a double array of their independent copies. Weak convergence in the dependent case is implied by similar asymptotic results under independence together with convergence to zero of one of a series of dependence measures including the multivariate extension of Pearson's correlation, the relative entropy or other multivariate divergence measures. A closely related result involves conditions for convergence in distribution of m-dimensional statistics $h(X_t, X_{t+1}, \ldots, X_{t+m-1})$ of time series $\{X_t\}$ in terms of weak convergence of $h(\xi_t, \xi_{t+1}, \ldots, \xi_{t+m-1})$, where $\{\xi_t\}$ is a sequence of independent copies of $X_t's$, and convergence to zero of measures of intertemporal dependence in $\{X_t\}$. The tools used include new sharp estimates for the distance between the distribution function of an arbitrary statistic in dependent random variables and the distribution function of the statistic in independent copies of the random variables in terms of the measures of dependence of the random variables. Furthermore, we obtain new sharp complete decoupling moment and probability inequalities for dependent random variables in terms of their dependence characteristics.

*Supported in part by NSF grants DMS/99/72237, DMS/02/05791, and DMS/05/05949.

†Supported in part by a Yale University Graduate Fellowship; the Cowles Foundation Prize; and a Carl Arvid Anderson Prize Fellowship in Economics.

[1]Department of Statistics, Columbia University, Mail Code 4690, 1255 Amsterdam Avenue, New York, NY 10027, e-mail: vp@stat.columbia.edu

[2]Department of Economics, Harvard University, 1805 Cambridge St., Cambridge, MA 02138, e-mail: ribragim@fas.harvard.edu

[3]Department of Probability Theory, Tashkent State Economics University, ul. Uzbekistanskaya, 49, Tashkent, 700063, Uzbekistan, e-mail: tim001@tseu.silk.org

AMS 2000 subject classifications: primary 62E10, 62H05, 62H20; secondary 60E05, 62B10, 62F12, 62G20.

Keywords and phrases: joint distribution, copulas, information, dependence, decoupling, convergence, relative entropy, Kullback–Leibler and Shannon mutual information, Pearson coefficient, Hellinger distance, divergence measures.

Theorem : *A function* $F : \mathrm{R}^n \to [0,1]$ *is a joint cdf with one-dimensional marginal cdf's* $F_k(x_k)$, $x_k \in \mathrm{R}$, $k = 1, \ldots, n$, *absolutely continuous with respect to the product of marginal cdf's* $\prod_{k=1}^n F_k(x_k)$, *if and only if there exist functions* $g_{i_1,\ldots,i_c} : \mathrm{R}^c \to \mathrm{R}$, $1 \le i_1 < \cdots < i_c \le n$, $c = 2, \ldots, n$, *satisfying conditions*

A1 (integrability):

$$E|g_{i_1,\ldots,i_c}(\xi_{i_1}, \ldots, \xi_{i_c})| < \infty,$$

A2 (degeneracy):

$$E(g_{i_1,\ldots,i_c}(\xi_{i_1}, \ldots, \xi_{i_{k-1}}, \xi_{i_k}, \xi_{i_{k+1}}, \ldots, \xi_{i_c})|\xi_{i_1}, \ldots, \xi_{i_{k-1}}, \xi_{i_{k+1}}, \ldots, \xi_{i_c}) =$$

$$\int_{-\infty}^{\infty} g_{i_1,\ldots,i_c}(\xi_{i_1}, \ldots, \xi_{i_{k-1}}, x_{i_k}, \xi_{i_{k+1}}, \ldots, \xi_{i_c}) dF_{i_k}(x_{i_k}) = 0, \ (a.s.)$$

$1 \le i_1 < \cdots < i_c \le n$, $k = 1, 2, \ldots, c$, $c = 2, \ldots, n$,

A3 (positive definiteness):

$$U_n(\xi_1, \ldots, \xi_n) \equiv \sum_{c=2}^{n} \sum_{1 \le i_1 < \cdots < i_c \le n} g_{i_1,\ldots,i_c}(\xi_{i_1}, \ldots, \xi_{i_c}) \ge -1 \ (a.s.)$$

and such that the following representation holds for F :

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} (1 + U_n(t_1, \ldots, t_n)) \prod_{i=1}^{n} dF_i(t_i).$$

Moreover, $g_{i_1,\ldots,i_c}(\xi_{i_1}, \ldots, \xi_{i_c}) = f_{i_1,\ldots,i_c}(\xi_{i_1}, \ldots, \xi_{i_c})$ *(a.s.),* $1 \le i_1 < \cdots < i_c \le n$, $c = 2, \ldots, n$, *where*

$$f_{i_1,\ldots,i_c}(x_{i_1}, \ldots, x_{i_c}) = \sum_{k=2}^{c} (-1)^{c-k} \sum_{1 \le j_1 < \cdots < j_k \in \{i_1,\ldots,i_c\}} \left(\frac{dF(x_{j_1}, \ldots, x_{j_k})}{dF_{j_1} \cdots dF_{j_k}} - 1 \right).$$

Applications to copulas.

Let $U_1, ..., U_n$ denote independent r.v.'s uniformly distributed on $[0, 1]$.

Theorem A function $C : [0,1]^n \to [0,1]$ is an absolutely continuous $n-$dimensional copula if and only if there exist functions $\tilde{g}_{i_1,...,i_c} : \mathbf{R}^c \to \mathbf{R}$, $1 \leq i_1 < ... < i_c \leq n$, $c = 2, ..., n$, satisfying the integrability, degeneracy and non-negativity conditions

$$\int_0^1 ... \int_0^1 |\tilde{g}_{i_1,...,i_c}(t_{i_1}, ..., t_{i_c})| dt_{i_1} ... dt_{i_c} < \infty,$$

$$E(\tilde{g}_{i_1,...,i_c}(U_{i_1}, ..., U_{i_{k-1}}, U_{i_k}, U_{i_{k+1}}, ..., U_{i_c}) | U_{i_1}, ..., U_{i_{k-1}}, U_{i_{k+1}}, ..., U_{i_c}) =$$

$$\int_0^1 \tilde{g}_{i_1,...,i_c}(U_{i_1}, ..., U_{i_{k-1}}, t_{i_k}, U_{i_{k+1}}, ..., U_{i_c}) dt_{i_k} = 0 (a.s.),$$

$x_{i_j} \in \mathbf{R}$, $j = 1, 2, ..., c$, $j \neq k$, $1 \leq i_1 < ... < i_c \leq n$, $k = 1, 2, ..., c$, $c = 2, ..., n$,

$$\sum_{c=2}^{n} \sum_{1 \leq i_1 < ... < i_c \leq n} \tilde{g}_{i_1,...,i_c}(U_{i_1}, ..., U_{i_c}) \geq -1 (a.s.)$$

and such that

$$C(u_1, ..., u_n) = \int_0^{u_1} ... \int_0^{u_n} \left(1 + \sum_{c=2}^{n} \sum_{1 \leq i_1 < ... < i_c \leq n} \tilde{g}_{i_1,...,i_c}(t_{i_1}, ..., t_{i_c}) \right) \prod_{i=1}^{n} dt_i.$$

The above results provide a general device for constructing multivariate copulas and joint distributions. E.g., taking $n = 2$ in the representation, $\tilde{g}_{1,2}(t_1, t_2) = \alpha(1 - 2t_1)(1 - 2t_2)$, $\alpha \in [-1, 1]$, we get the family of bivariate Eyraud-Farlie-Gumbel-Mongenstern copulas

$$C_\alpha(u_1, u_2) = u_1 u_2 \left(1 + \alpha(1 - u_1)(1 - u_2)\right)$$

and corresponding bivariate distributions

$$F_\alpha(x_1, x_2) = F_1(x_1)F_2(x_2) \left(1 + \alpha(1 - F_1(x_1))(1 - F_2(x_2))\right).$$

More generally, taking $\tilde{g}_{i_1,...,i_c}(t_{i_1}, ..., t_{i_c}) = 0$, $1 \leq i_1 < ... < i_c \leq n$, $c = 2, ..., n-1$, $\tilde{g}_{1,2,...,n}(t_1, t_2, ..., t_n) = \alpha(1 - 2t_1)(1 - 2t_2)...(1 - 2t_n)$, we obtain the following multivariate Eyraud-Farlie-Gumbel-Mongenstern copulas

$$C_\alpha(u_1, u_2, ..., u_n) = \prod_{i=1}^{n} u_i \left(1 + \alpha \prod_{i=1}^{n} (1 - u_i)\right)$$

and corresponding multivariate distributions

$$F_\alpha(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} F_i(x_i) \left(1 + \alpha \prod_{i=1}^{n} (1 - F_i(x_i))\right).$$

Taking $n = 2$, $\tilde{g}_{1,2}(t_1, t_2) = \theta c(t_1, t_2)$, where c is a continuous function on the unit square $[0, 1]^2$ satisfying the properties

$$\int_0^1 c(t_1, t_2)dt_1 = \int_0^1 c(t_1, t_2)dt_2 = 0,$$

$$1 + \theta c(t_1, t_2) \geq 0$$

for all $0 \leq t_1, t_2 \leq 1$, one obtains the class of bivariate densities studied by Long and Krzysztofowicz (1995)

$$f(x_1, x_2) = f_1(x_1)f_2(x_2)(1 + \theta c(F_1(x_1), F_2(x_2))$$

with the covariance characteristic c and the covariance scalar θ. Furthermore, it follows that representation in fact holds for an arbitrary density function and the function $\theta c(t_1, t_2)$ is unique.

$$E\xi_i\xi_j g_{ij}(\xi_i,\xi_j)=0,\ 1\le i<j\le n.$$

Characterizations of classes of dependent random variables

The following Corollaries give characterizations of different classes of dependent r.v.'s in terms of functions g that appear in the representations for joint distributions obtained before. Completely similar results hold for the functions g that enter corresponding representations for copulas.

Corollary 0. R.v.'s $X_1, ..., X_n$ with the one-dimensional cdf's $F_k(x_k)$, $x_k \in \mathbf{R}$, $k = 1, ..., n$, are independent if and only if the functions $g_{i_1,...,i_c}$ in the representation satisfy the conditions $g_{i_1,...,i_c}(\xi_{i_1}, ..., \xi_{i_c}) = 0$ (a.s.), $1 \leq i_1 < ... < i_c \leq n$, $c = 2, ..., n$.

Definition 1 A sequence of r.v.'s $\{X_n\}$ is called strictly stationary if the vector $(X_{j_1}, ..., X_{j_k})$ has the same distribution as the vector $(X_{j_1+h}, ..., X_{j_k+h})$ for all $1 \leq j_1 < ... < j_k$, $k = 1, 2, ...$, and all $h = 0, 1, 2, ...$

Corollary 1 A sequence of r.v.'s $\{X_n\}$ is strictly stationary if and only if the functions $g_{i_1,...,i_c}$ in the representations for any finite-dimensional distribution satisfy the conditions $g_{i_1+h,...,i_c+h}(\xi_{i_1}, ..., \xi_{i_c}) = g_{i_1,...,i_c}(\xi_{i_1}, ..., \xi_{i_c})$ (a.s.) $1 \leq i_1 < ... < i_c \leq n$, $c = 2, 3, ...$, $h = 0, 1, ...$

Definition 2 A sequence of r.v.'s $\{X_n\}$ with $EX_k = 0$, $EX_k^2 < \infty$, $k = 1, 2, ...$, is called weakly stationary if the function $f(s,t) = cov(X_s, X_t)$ depends only on $|t - s|$, $t, s = 1, 2, ...$

Corollary 2 A sequence of r.v.'s $\{X_n\}$ with $EX_k = 0$, $EX_k^2 < \infty$, $k = 1, 2, ...$, is weakly stationary if and only if the functions g in the representations for any finite-dimensional distribution have the property that the function $h(s,t) = E\xi_s\xi_t g_{st}(\xi_s, \xi_t)$, depends only on $|t-s|$, $t, s = 1, 2, ...$ where $\{\xi_n\}$ is a sequence of independent r.v.'s such that ξ_k is identically distributed with X_k, $k = 1, 2, ...$

Definition 3 R.v.'s $X_1, ..., X_n$ with $EX_i = 0$, $i = 1, ..., n$, are called orthogonal if $EX_iX_j = 0$ for all $1 \leq i < j \leq n$.

Corollary 3 R.v.'s $X_1, ..., X_n$ with $EX_k = 0$, $k = 1, ..., n$, are orthogonal if and only if the functions g in representations satisfy the conditions

Definition 4 R.v.'s $X_1, ..., X_n$ are called exchangeable of all $n!$ permutations $(X_{\pi(1)}, ..., X_{\pi(n)})$ of the r.v.'s have the same joint distributions.

Corollary 4 Identically distributed r.v.'s $X_1, ..., X_n$ are exchangeable if and only if the functions $g_{i_1,...,i_c}$ in representations satisfy the conditions

$$g_{i_1,...,i_c}(\xi_{i_1}, ..., \xi_{i_c}) = g_{i_{\pi(1)},...,i_{\pi(c)}}(\xi_{i_{\pi(1)}}, ..., \xi_{i_{\pi(i_c)}})$$

(a.s.) for all $1 \leq i_1 < ... < i_c \leq n$, $c = 2, ..., n$, and all permutations π of the set $\{1, ..., n\}$.

Definition 5 R.v.'s $X_1, ..., X_n$ are called $m-$dependent $(1 \leq m \leq n)$ if any two vectors
$(X_{j_1}, X_{j_2}, ..., X_{j_{a-1}}, X_{j_a})$ and $(X_{j_{a+1}}, X_{j_{a+2}}, ..., X_{j_{l-1}}, X_{j_l})$, where
$1 \leq j_1 < ... < j_a < ... < j_l \leq n$, $a = 1, 2, ..., l-1$, $l = 2, ..., n$, $j_{a+1} - j_a \geq m$, are independent.

Corollary 5 R.v.'s $X_1, ..., X_n$ are $m-$dependent if and only if the functions g in representations satisfy the conditions
$g_{i_1,...,i_k,i_{k+1},...,i_c}(\xi_{i_1}, ..., \xi_{i_k}, \xi_{i_{k+1}}, ..., \xi_{i_c}) = g_{i_1,...,i_k}(\xi_{i_1}, ..., \xi_{i_k}) g_{i_{k+1},...,i_c}(\xi_{i_{k+1}}, ..., \xi_{i_c})$
for all $1 \leq i_1 < ... < i_k < i_{k+1} ... < i_c \leq n$, $i_{k+1} - i_k \geq m$, $k = 1, ..., c-1$, $c = 2, ..., n$.

COPULA-BASED CHARACTERIZATIONS FOR HIGHER-ORDER MARKOV PROCESSES

By Rustam Ibragimov[1]

Department of Economics, Harvard University

Address for manuscript correspondence:

Rustam Ibragimov
Department of Economics
Harvard University
Littauer Center
1805 Cambridge St.
Cambridge, MA 02138
Email: ribragim@fas.harvard.edu
Phone: +1-617-496-4795
Fax: +1-617-495-7730

ABSTRACT

In this paper, we obtain characterizations of higher-order Markov processes in terms of copulas corresponding to their finite-dimensional distributions. The results are applied to establish necessary and sufficient conditions for Markov processes of a given order to exhibit $m-$dependence, $r-$independence or conditional symmetry. The paper also presents a study of applicability and limitations of different copula families in constructing higher-order Markov processes with the above dependence properties. We further introduce new classes of copulas that allow one to combine Markovness with $m-$dependence or $r-$independence in time series.

Key words and phrases: copulas, dependence, time series, Markov processes, $m-$dependence, $r-$independence, conditional symmetry, martingales, stochastic differential equations, Fourier copulas

JEL Classification: C14, C22, C32, C51

[1]This paper was previously circulated under the title "Copula-based dependence characterizations and modeling for time series". An extended working paper version of the paper is available as Ibragimov (2005a). I thank three referees, Donald Andrews, Brendan Beare, Christian Gourieroux, George Lentzas, Jeremiah Lowin, Andrew Patton, Peter Phillips, Murray Rosenblatt, Yildiray Yildirim and the participants at seminars at the Departments of Economics at Boston University, Harvard University and Yale University, Whitman School of Management at Syracuse University and Harvard Statistics Summer Retreat on "Recent Advances in Computational Finance" (June 2006) for helpful comments and suggestions. A part of the paper was completed under the financial support from a Yale University Dissertation Fellowship and a Cowles Foundation Prize.

Copula-Based Models for Financial Time Series[1]

First version: 31 August 2006. This version: 19 November 2007.

Andrew J. Patton

Department of Economics and Oxford-Man Institute of Quantitative Finance, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom.

andrew.patton@economics.ox.ac.uk

Abstract This paper presents an overview of the literature on applications of copulas in the modelling of financial time series. Copulas have been used both in multivariate time series analysis, where they are used to charaterise the (conditional) cross-sectional dependence between individual time series, and in univariate time series analysis, where they are used to characterise the dependence between a sequence of observations of a scalar time series process. The paper includes a broad, brief, review of the many applications of copulas in finance and economics.

1 Introduction

The central importance of risk in financial decision-making directly implies the importance of dependence in decisions involving more than one risky asset. For example, the variance of the return on a portfolio of risky assets depends on the variances of the individual assets and also on the linear correlation between the assets in the portfolio. More generally, the distribution of the return on a portfolio will depend on the univariate distributions of the individual assets in the portfolio and on the dependence between each of the assets, which is captured by a function called a 'copula'.

The number of papers on copula theory in finance and economics has grown enormously in recent years. One of the most influential of the 'early' papers on copulas in finance is

[1]This paper was prepared for the forthcoming *Handbook of Financial Time Series*, T. G. Andersen, R. A. Davis, J.-P. Kreiss and T. Mikosch (eds.), Springer Verlag. I would particularly like to thank B. Beare, P. Embrechts, J.-D. Fermanian, T. Mikosch and J. Rosenberg for detailed comments and suggestions on this chapter. I would also like to thank Y. Fan, J.-C. J.-P. Kreiss, Rodriguez, C. Schleicher and T. Schuermann for helpful comments. Some Matlab code for copulas is available from http://www.economics.ox.ac.uk/members/andrew.patton/code.html.

assets will exhibit more extreme returns than identical assets with a Normal copula.

2 Copula-based models for time series

The application of copulas to time series modelling currently has two distinct branches. The first is the application to multivariate time series, where the focus is in modelling the joint distribution of some random vector, $\mathbf{X}_t = [X_{1t}, X_{2t}, ..., X_{nt}]'$, conditional on some information set $\mathcal{F}_{t-1}$. (The information set is usually $\mathcal{F}_{t-1} = \sigma(\mathbf{X}_{t-j}; j \geq 1)$, though this need not necessarily be the case.) This is an extension of some of the early applications of copulas in statistical modelling where the random vector of interest could be assumed to be independent and identically distributed (*iid*), see Clayton (1978) and Cook and Johnson (1981) for example. This application leads directly to the consideration of time-varying copulas.

The second application in time series is to consider the copula of a sequence of observations of a univariate time series, for example, to consider the joint distribution of $[X_t, X_{t+1}, ..., X_{t+n}]'$. This application leads us to consider Markov processes and general nonlinear time series models. We discuss each of these branches of time series applications of copulas below.

2.1 Copula-based models for multivariate time series

In this sub-section we consider the extension required to consider the conditional distribution of $\mathbf{X}_t$ given some information set $\mathcal{F}_{t-1}$. Patton (2006a) defined a "conditional copula" as a multivariate distribution of (possibly correlated) variables that are each distributed as $Uniform(0,1)$ conditional on $\mathcal{F}_{t-1}$. With this definition, it is then possible to consider an extension of Sklar's theorem to the time series case:

$$\mathbf{F}_t(\mathbf{x}|\mathcal{F}_{t-1}) = \mathbf{C}_t(F_{1,t}(x_1|\mathcal{F}_{t-1}), F_{2,t}(x_2|\mathcal{F}_{t-1}), ..., F_{n,t}(x_n|\mathcal{F}_{t-1})|\mathcal{F}_{t-1}), \quad \forall\, \mathbf{x} \in \mathbb{R}^n, \quad (4)$$

where $X_i|\mathcal{F}_{t-1} \sim F_{i,t}$ and $\mathbf{C}_t$ is the conditional copula of $\mathbf{X}_t$ given $\mathcal{F}_{t-1}$.

The key complication introduced when applying Sklar's theorem to conditional distributions is that the conditioning set, $\mathcal{F}_{t-1}$, must be the same for all marginal distributions and the copula. Fermanian and Wegkamp (2004) and Fermanian and Scaillet (2005) con-

The Annals of Applied Statistics
2011, Vol. 5, No. 3, 1752–1779
DOI: 10.1214/11-AOAS466

MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS[1]

BY QUNHUA LI, JAMES B. BROWN, HAIYAN HUANG AND PETER J. BICKEL

University of California at Berkeley

Reproducibility is essential to reliable scientific discovery in high-throughput experiments. In this work we propose a unified approach to measure the reproducibility of findings identified from replicate experiments and identify putative discoveries using reproducibility. Unlike the usual scalar measures of reproducibility, our approach creates a curve, which quantitatively assesses when the findings are no longer consistent across replicates. Our curve is fitted by a copula mixture model, from which we derive a quantitative reproducibility score, which we call the "irreproducible discovery rate" (IDR) analogous to the FDR. This score can be computed at each set of paired replicate ranks and permits the principled setting of thresholds both for assessing reproducibility and combining replicates.

Since our approach permits an arbitrary scale for each replicate, it provides useful descriptive measures in a wide variety of situations to be explored. We study the performance of the algorithm using simulations and give a heuristic analysis of its theoretical properties. We demonstrate the effectiveness of our method in a ChIP-seq experiment.

1. Introduction. High-throughput profiling technologies play an indispensable role in modern biology. By studying a large number of candidates in a single experiment and assessing their significance using data analytical tools, high-throughput technologies allow researchers to effectively select potential targets for further studies. Despite their ubiquitous presence in biological research, it is known that any single experimental output from a high-throughput assay is often subject to substantial variability. Reproducibility of high-throughput assays, such as the level of agreement between results from replicate experiments across (biological or technical) replicate

Received May 2010; revised January 2011.
[1]Supported in part by NIH 1U01HG004695-01, 1-RC2-HG005639-01 and R21EY019094.
Key words and phrases. Reproducibility, association, mixture model, copula, iterative algorithm, irreproducible discovery rate, high-throughput experiment, genomics.

the replicates to be independent, that is, $\rho_0 = 0$; whereas, since genuine signals usually are positively associated between replicates, we expect $\rho_1 > 0$, though ρ_1 is not required to be positive in our model. It also seems natural to assume that the underlying latent variables, reflecting replicates, have the same marginal distributions. Finally, we note that if the marginal scales are unknown, we can only identify the *difference* in means of the two latent variables and the *ratio* of their variances, but not the means and variances of the latent variables. Thus, the parametric model generating our copula can be described as follows:

Let $K_i \sim \text{Bernoulli}(\pi_1)$ and $(z_{i,1}, z_{i,2})$ be distributed as

$$\text{(2.5a)}\quad \begin{pmatrix} z_{i,1} \\ z_{i,2} \end{pmatrix} \Bigg| K_i = k \sim N\left(\begin{pmatrix} \mu_k \\ \mu_k \end{pmatrix}, \begin{pmatrix} \sigma_k^2 & \rho_k \sigma_k^2 \\ \rho_k \sigma_k^2 & \sigma_k^2 \end{pmatrix} \right), \qquad k = 0, 1,$$

where $\mu_0 = 0$, $\mu_1 > 0$, $\sigma_0^2 = 1$, $\rho_0 = 0$, $0 < \rho_1 \leq 1$.

Let

$$\text{(2.5b)}\quad \begin{aligned} u_{i,1} \equiv G(z_{i,1}) &= \frac{\pi_1}{\sigma_1} \Phi\left(\frac{z_{i,1} - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z_{i,1}), \\ u_{i,2} \equiv G(z_{i,2}) &= \frac{\pi_1}{\sigma_1} \Phi\left(\frac{z_{i,2} - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z_{i,2}). \end{aligned}$$

Our actual observations are

$$\text{(2.5c)}\quad \begin{aligned} x_{i,1} &= F_1^{-1}(u_{i,1}), \\ x_{i,2} &= F_2^{-1}(u_{i,2}), \end{aligned}$$

where F_1 and F_2 are the marginal distributions of the two coordinates, which are assumed continuous but otherwise unknown.

Thus, our model, which we shall call a copula mixture model, is a semiparametric model parametrized by $\theta = (\pi_1, \mu_1, \sigma_1^2, \rho_1)$ and (F_1, F_2). The corresponding mixture likelihood for the data is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} [\pi_0 h_0(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2}))) \\ &\qquad + \pi_1 h_1(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))] \qquad \text{(2.6a)} \\ &= \prod_{i=1}^{n} [c(F_1(x_{i,1}), F_2(x_{i,2})) g(G^{-1}(F_1(x_{i,1}))) g(G^{-1}(F_2(x_{i,2})))], \qquad \text{(2.6b)} \end{aligned}$$

where

$$\text{(2.7)}\quad c(u_1, u_2) = \frac{\pi_0 h_0(G^{-1}(u_1), G^{-1}(u_2)) + \pi_1 h_1(G^{-1}(u_1), G^{-1}(u_2))}{g(G^{-1}(u_1)) g(G^{-1}(u_2))}$$