Robust Estimation for Pair-Copula Models

Beatriz Vaz de Melo Mendes[1]

IM - Federal University at Rio de Janeiro, Brazil.

Abstract

In this paper we extend the previous work of Mendes, Melo and Nelsen (2007) and study robust estimation for pair-copula models. The extention is straightforward since a pair-copula construction is just a hierarchical decomposition of a multivariate copula into a cascade of bivariate copulas and estimation takes place at the level of the two-dimensional data. Robust Minimum Distance Estimators (MDE) are proposed for new copula families — BB7 and t-student — and the weighted version of the Maximum Likelihood Estimators (WMLE) proved again to be the best option for members of the elliptical family of copulas. We conditionally model the time-varying behavior of series of realized volatilities in the Brazilian equity market using pair-copulas. We robustly estimate and forecast their evolution.

Keywords: Robust estimation; Pair-copulas; Minimum Distance Estimators; Robust Covariance Estimators.

1 Introduction

Volatility plays important role in asset pricing and allocation and risk management. It is usually estimated using member of the GARCH family. As such, they are only valid under the assumptions of the model. Ex post realized volatilities may be constructed by summing the squares of evenly spaced intraday high-frequency returns computed from continuously recorded transaction prices (Andersen et al. 2001). Papers dealing with construction and applications of realized volatilities have focused on developed markets data (ref ref ref). Here we examine the six most traded Brazilian stocks volatilities.

The decomposition of a d-dimensional copula ($d > 2$) into a collection of potentially different bivariate copulas was originally proposed by Joe (1996), and later discussed in detail by Bedford and Cooke (2001, 2002), Kurowicka and Cooke (2006) and Aas, Czado, Frigessi, and Bakken (2007). The method of construction is hierarchical, where variables are sequentially incorporated into the conditioning sets as one moves from tree 1 to tree $d - 1$. Pair-copulas are flexible since all composing bivariate copulas may vary freely,

[1]Email: beatriz@im.ufrj.br.

covering any omplex pattern of dependence usually exhibited by multivariate data, being easy to estimate and simulate.

Pair-copulas estimation is usually performed in the context of independent and identically distributed observations by extending the IFM method, inference function for margins, initially proposed for copulas parameters' estimation. Under the IFM method, introduced by Joe and Xu (1996), the maximum likelihood estimates are obtained for the marginal and copula parameters. Joe (1997) argues that we can expect the IFM method to be quite efficient since fully based on maximum likelihood estimation. The success of this estimation procedure starts with good marginal fits (see Frahm, Junker, and Schmidt, 2004), which typically pose no difficulties. Alternatively, a semiparametric method may be applied, where the margins are fitted empirically and the dependence parameters are fitted by maximum likelihood (see Genest et al., 1995). Confidence intervals are usually obtained through bootstrap methods, or for some families the asymptotic variance may be computed.

Bayesian methods have also been applied to pair-copulas. Dalla Valle (2007) proposed Bayesian inference based on MCMC for multivariate elliptical copulas using the inverse Wishart distribution as a prior for the correlation matrix . Min and Czado (2008) also developed a Markov chain Monte Carlo algorithm which provides credibility intervals.

When all data points come from the same data generating process, the maximum likelihood estimates (MLE) possess the usual good statistical properties (see Genest, Ghoudi, and Rivest (1995), and Shih and Louis (1995)). However, contaminations may occur in many ways and atypical points may change the strength of association, resulting in distorted estimation of dependence measures and poor predictions. When dealing with high frequency data, contaminations may occur from heavy data manipulation (automatic checkings will fail if not all scenarios have been contemplated) and

Data manipulation may result in gross errors, or imply in data columns misalignment, which would not damage the marginal fits but would completely distort estimates of the dependence structure. We also note that the copula sample space $[0,1]^d$ poses difficulties for the graphical inspection of atypical points. We need thus an automatic robust procedure that would work well when data are, and when data are not contaminated.

However, to the best of our knowledge, no one has yet proposed robust estimates for pair-copulas.

Mendes, Melo and Nelsen (2007) proposed two classes of robust estimators for copulas, aiming to provide guidance when modeling real data. They are based on either a redescending weight function or on a hard rejection rule. The minimum distance estimators (MDE) minimize some selected weighted distance based statistics. The weighted

maximum likelihood estimates (WMLE) result from a two-step procedure. In the first step, outlying data points are identified by a robust covariance estimator and receive zero weights, and in the second step the MLE are computed for the reduced data. In this paper we extend this previous work and propose to robustly estimate pair-copulas using the MDE and the WMLE estimates. The extention is straightforward since estimation takes place at the level of the bivariate copulas.

Robust estimators are found for new copula families — BB7 and t-student — and the WMLE proved again to be the best option for members of the elliptical family of copulas. We carried on simulations and considered varying proportions of contaminating points located at different regions of the copula support. Among the 18 different types of MDE estimators we report those resulting in the smaller mean squared error. The simulation experiments indicate that for each copula family there is at least one robust estimate performing very well, in the sense of small mean squared error, despite the contamination percentage and location, and the sample size.

The remaining of this paper is organized as follows. In Section 2 we briefly review the definition of pair-copula. In Section 3 we define the robust estimates and report the results from the new simulations. In Section 4 we work out two convincing examples where robust estimation is needed and results in better inferences. To illustrate, we model and forecast the realized volatility of two Brazilian stocks. In Section 5 we summarize and discuss the results of this paper.

2 Pair-copulas

In this section we provide a brief review of copulas and pair-copulas.

2.1 Copulas

The most important theorem in copula theory dates back to the fifties (Sklar, 1959). It states that any multivariate distribution can be expressed by its copula function evaluated at its marginal distribution functions.

Consider a continuous random vector $X_1, \ldots, X_d$ with joint distribution function (c.d.f.) $H(x_1, \ldots, x_d)$ and marginal distributions $F_1, \ldots, F_d$. For every $(x_1, \ldots, x_d) \in [-\infty, \infty]^d$ consider the point in $[0,1]^{d+1}$ with coordinates $(F_1(x_1), \ldots, F_d(x_d), H(x_1, \ldots, x_d))$. This mapping from $[0,1]^d$ to $[0,1]$ is a *d-dimensional copula*.

Sklar's theorem ensures that there exists a d-copula C such that for all $(x_1, \ldots, x_d) \in [-\infty, \infty]^d$

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{1}$$

Conversely, if C is a d-copula and $F_1, \ldots, F_d$ are c.d.f.s, the function H defined by (1) is a d-dimensional distribution function with margins $F_1, \ldots, F_d$. Furthermore, if all marginal c.d.f.s are continuous, C is unique.

Given a joint c.d.f. H with continuous margins $F_1, \ldots, F_d$, as in Sklar's Theorem, it is easy to construct the corresponding copula:

$$C(u_1, \ldots, u_d) = H(F_1^-(u_1), \ldots, F_d^-(u_d)), \tag{2}$$

where F_i^- is the generalized inverse of F_i, that is, $F_i^-(u) = \sup\{t \in \Re : F_i(t) \leq u,\ 0 \leq u \leq 1\}$. (2) is the tool for extracting the copula pertaining to a multivariate distributions.

When C is absolutely continuous, taking partial derivatives of (1) one obtains

$$h(x_1, \cdots, x_d) = c(F_1(x_1), \cdots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i) \tag{3}$$

for some d-dimensional copula density c. This expression will prove useful later for parameter estimation. (3) allows for tailored marginal modeling considering all characteristics of each F_i, including the mean, standard deviation, skewness, kurtosis and any type of short and long memory serial dependence, plus a search for the best fit for the dependence structure through a large number of copula families that may be considered. This results in flexible multivariate distributions with any choice of margin distributions. An important example is the family of the *meta*-elliptical distributions (Fang, Fang and Kotz, 2002, 2005) which, unlike the family of elliptical distributions, do not impose any constraints on their margins. This flexibility have motivated applications of copulas in several research areas.

2.2 Pair-copulas

However, the copula approach for modeling high-dimensional copulas has also limitations. Firstly, the generalization of bivariate copulas to multivariate copulas of dimension larger than 2 is not straightforward. When high-dimensional copulas are available, there are significant obstacles to solving the required optimization problem over many dimensions, and most of the available softwares deal only with the bivariate case. Even if we are able to fit a d-dimensional copula, $d > 2$, parametric copula families usually restrict all pairs to having the same type or strength of dependence. For example, in the case of the t-copula, in addition to the correlation coefficients, a single parameter, the number of degrees of freedom, is used to compute the coefficient of tail dependence for *all pairs*. This is a serious restriction since the dependence structure among pairs of variables usually vary substantially, including changes in copula family.

Pair-copulas, being a collection of potentially different bivariate copulas, are flexible and very appealing. The method of construction is hierarchical, where variables are sequentially incorporated into the conditioning sets as one moves from level 1 (tree 1) to tree $d-1$. The composing bivariate copulas may vary freely, from the parametric family to the parameter values. Therefore, all types and strengths of dependence can be covered. Pair-copulas are easy to estimate and simulate, making them very appropriate for modeling large dimensional data sets.

The decomposition of a multivariate distribution in a cascade of pair-copulas was originally proposed by Joe (1996), and later discussed in detail by Bedford and Cooke (2001, 2002), Kurowicka and Cooke (2006) and Aas, Czado, Frigessi, and Bakken (2007).

Consider again the joint distribution H with density h with strictly continuous marginal c.d.f.s $F_1, \cdots, F_d$ with densities f_i. First note that any multivariate density function may be uniquely decomposed as

$$h(x_1, ..., x_d) = f_d(x_d) \cdot f(x_{d-1}|x_d) \cdot f(x_{d-2}|x_{d-1}, x_d) \cdots f(x_1|x_2, ..., x_d). \tag{4}$$

The conditional densities in Equation (4) may be written as functions of the corresponding copula densities. That is, for every j

$$f(x \mid v_1, v_2, \cdots, v_d) = c_{xv_j|\mathbf{v}_{-j}}(F(x \mid \mathbf{v}_{-j}), F(v_j \mid \mathbf{v}_{-j})) \cdot f(x \mid \mathbf{v}_{-j}), \tag{5}$$

where $\mathbf{v}_{-j}$ denotes the d-dimensional vector $\mathbf{v}$ excluding the jth component. Note that $c_{xv_j|\mathbf{v}_{-j}}(\cdot, \cdot)$ is a *bivariate* marginal copula density. For example, when $d = 3$,

$$f(x_1|x_2, x_3) = c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot f(x_1|x_2)$$

and

$$f(x_2|x_3) = c_{23}(F(x_2), F(x_3)) \cdot f(x_2)$$

and

$$c(u_1, u_2, u_3) = c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \cdot c_{13|2}(F(x_1 \mid x_2), F(x_3 \mid x_2)).$$

Expressing all conditional densities in Equation (4) by means of Equation (5), we derive a decomposition for $h(x_1, \cdots, x_d)$ that consists of only univariate marginal distributions and bivariate copulas. Thus we obtain the *pair-copula decomposition* for the d-dimensional copula $c_{1\cdots d}$, a factorization of a d-dimensional copula based only in bivariate copulas. This is a very flexible and natural way of constructing a higher dimensional copula. Note that, given a specific factorization, there are many possible reparametrizations.

The conditional c.d.f.s necessary for pair-copulas construction are given (Joe, 1996) by

$$F(x \mid \mathbf{v}) = \frac{\partial C_{x,v_j|\mathbf{v}_{-j}}(F(x \mid \mathbf{v}_{-j}), F(v_j \mid \mathbf{v}_{-j}))}{\partial F(v_j \mid \mathbf{v}_{-j})}.$$

For the special case (unconditional) when v is univariate, and x and v are standard uniform, we have

$$F(x \mid v) = \frac{\partial C_{xv}(x, v, \Theta)}{\partial v}$$

where Θ is the set of copula parameters.

For large d, the number of possible pair-copula constructions is very large. As shown in Bedford and Cooke (2001), there are 240 different decompositions when $d = 5$. These authors introduce a systematic way to obtain the decompositions, which involves graphical models that they call *regular vines*. They also aid in understanding the conditional specifications made for the joint distribution. Special cases are the hierarchical canonical vines (C-vines) and the D-vines. Each of these graphical models is a specific way of decomposing the density $h(x_1, \cdots, x_d)$. For example, for a C-vine, h is equal to

$$\prod_{k=1}^{d} f(x_k) \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{j,j+i|1,\ldots,j-1}(F(x_j|x_1, ..., x_{j-1}), F(x_{j+i}|x_1, ..., x_{j-1})).$$

In a D-vine, there are $d-1$ hierarchical *trees* with increasing conditioning sets, and there are $d(d-1)/2$ bivariate copulas. For a detailed description, see Aas, Czado, Frigessi, and Bakken (2007). In the real data illustrations we work out in Section 4, there is a key variable that interact with all others. In such a situation it is more convenient choose a C-vine decomposition and to place this variable at the root of the canonical vine. Figure 1 shows the C-vine decomposition for $d = 4$. The C-vine consists of 3 nested trees, with tree T_j having $5 - j$ nodes, and $4 - j$ edges corresponding to a bivariate copula. The copulas in tree 1 are unconditional, and all others are conditional.

Simulations from both C- and D-vine pair-copulas can be easily implemented and take very little time to run.

3 Robust estimates

We obtain robust estimates for pair-copulas by adapting to the pair-copulas environment, the concept of *Weighted Minimum Distance estimators* (WMDE) and *Weighted Maximum Likelihood Estimates* (WMLE) originally proposed for copulas. A comprehensive simulation study showed that for each copula family one can always find a specific weighted

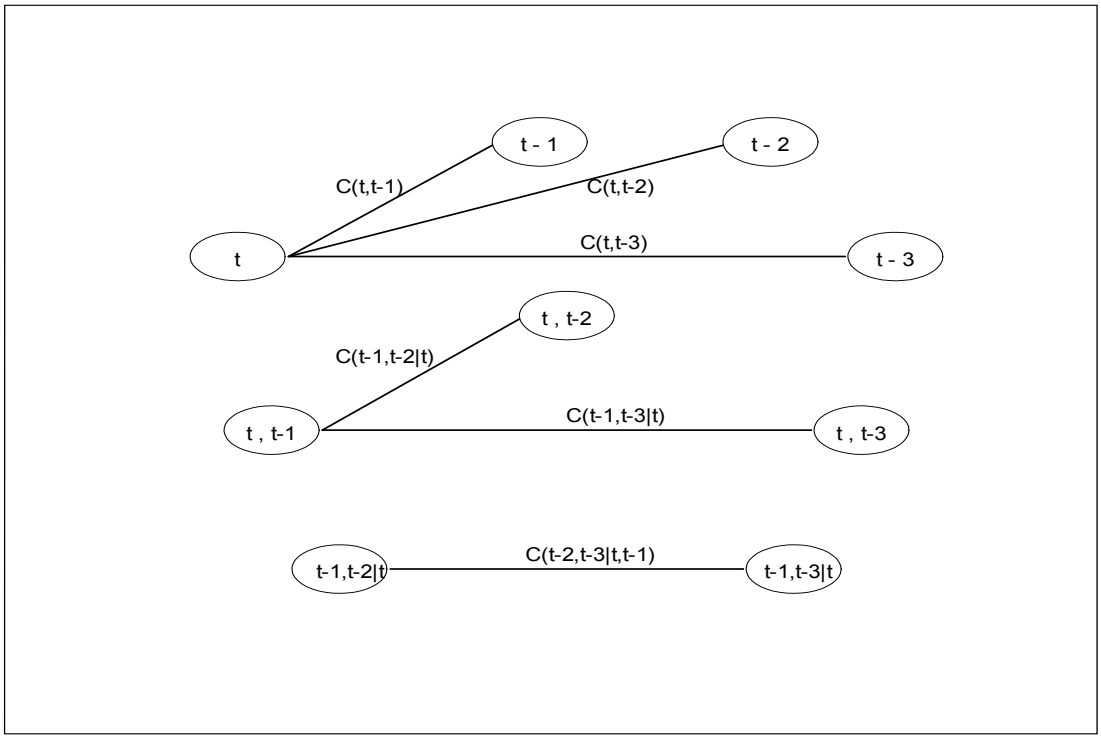

Figure 1: *The Canonical vine graphical hierarchical representation of a four-dimensional pair-copula.*

minimum distance estimator, which do not depend on the sample size and on the strength of dependence able to downweight the influence of contaminating points, introducing robustness. For elliptical copulas, as expected, the WMLE is always the best robust option.

Since estimation of a pair-copula model is performed at the level of bivariate data, we review the definitions of these estimators in the bivariate case. Details on these estimates may be found in Mendes, Melo and Nelsen (2007). Let $d = 2$ and let $(X_{1,t}, X_{2,t})$, $t = 1, ..., T$, be T independent copies of (X_1, X_2) with joint c.d.f. H. The *bivariate empirical distribution function* is given by

$$H_T(x_1, x_2) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{I}_{\{X_{1,t} \leq x_1, X_{2,t} \leq x_2\}}, \quad -\infty < x_1, x_2 < +\infty \ ,$$

where $\mathbf{I}_{\{A\}}$ is the indicator function of event A. Its associated *marginal empirical distribution functions* $F_{i,T}(x_1)$, $i = 1, 2$, are defined as

$$F_{1,T} = F_T(x_1, +\infty) \quad \text{and} \quad F_{2,T}(x_2) = F_T(+\infty, x_2) \ .$$

Let $F_{i,T}^{-}$ represent the generalized inverse of $F_{i,T}$. The *empirical copula function* $\widetilde{C}$ (De-

heuvels (1979), Deheuvels (1981a, 1981b), Fermanian, Radulović, and Wegkamp, 2004) is defined by

$$\widetilde{C}(u,v) = F_T(F^-_{1,T}(u), F^-_{2,T}(v)), \quad 0 \leq u, v \leq 1 \ .$$

According to Deheuvels (1979), $\widetilde{C}$ converges to C as T increases. $\widetilde{C}$ is computed on the lattice $\mathcal{L} = \{(\frac{t_1}{T}, \frac{t_2}{T})\}$, where t_1 and t_2 are integers, $1 \leq t_1, t_2 \leq T$:

$$\widetilde{C}(\frac{t_1}{T}, \frac{t_2}{T}) \ = \ F_T(F^{-1}_{1,T}(\frac{t_1}{T}), F^{-1}_{2,T}(\frac{t_2}{T})) \ \forall \ (\frac{t_1}{T}, \frac{t_2}{T}) \ \in \ \mathcal{L} \ . \tag{6}$$

Let C_θ represent a parametric copula parameterized by θ. The minimum distance estimate for θ is the solution θ^* which minimizes over all θ in the parameter space Θ, some distance between the empirical copula $\widetilde{C}$ and the parametric copula $\widehat{C} = C_{\widehat{\theta}}$ fitted to the data. These distance measures are further modified through the application of appropriate redescending weight functions, giving rise to 24 types of WMDE.

Examples of metrics defining the MDEs are the Kolmogorov statistic K

$$K = \max_{(\frac{t_1}{T}, \frac{t_2}{T}) \in \mathcal{L}} \quad |\widetilde{C}(\frac{t_1}{T}, \frac{t_2}{T}) - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})|, \tag{7}$$

and the Cramér-Von Mises statistic W^2

$$W^2 = \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} [\widetilde{C}(\frac{t_1}{T}, \frac{t_2}{T}) - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]^2. \tag{8}$$

By applying to (7) and (8) the weight function

$$w_{AK} = \frac{1}{\sqrt{[\widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})][1 - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]}}, \tag{9}$$

which emphasizes deviations in the *tails* (the corners of the unit square), one obtains the so called Anderson-Darling statistic (10)

$$AD_{AK} = \max_{(\frac{t_1}{T}, \frac{t_2}{T}) \in \mathcal{L}} \ \frac{|\widetilde{C}(\frac{t_1}{T}, \frac{t_2}{T}) - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})|}{\sqrt{[\widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})][1 - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]}}, \tag{10}$$

and the Integrated Anderson-Darling statistic (11)

$$IAD_{AK} = \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} \frac{[\widetilde{C}(\frac{t_1}{T}, \frac{t_2}{T}) - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]^2}{[\widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})][1 - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]}. \tag{11}$$

New redescending weight functions emphasizing all different regions of the $[0,1]^2$ were proposed in Mendes, Melo and Nelsen (2007) and resulted in other robust variations of

the Kolmogorov and of the Cramér-Von Mises statistics. For example, the weight function w_1

$$w_1(\frac{t_1}{T}, \frac{t_2}{T}) = \frac{1}{\sqrt{[\frac{t_1}{T} + \frac{t_2}{T} - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})][1 - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]}}\ , \tag{12}$$

emphasizes just the points in the lower left (LL) *and* the upper right (UR) corners. The factors $[\frac{t_1}{T} + \frac{t_2}{T} - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]$ and $[1 - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]$ correspond to the c.d.f. areas located at the LL and the UR quadrants of the unit square. By applying w_1 to (7) and (8) one obtains

$$AD_1 = \max_{1 \leq t_1, t_2 \leq T} \mid \tilde{C}(\frac{t_1}{T}, \frac{t_2}{T}) - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T}) \mid w_{1,LL-UR}(\frac{t_1}{T}, \frac{t_2}{T}), \tag{13}$$

$$IAD_1 = \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} [\tilde{C}(\frac{t_1}{T}, \frac{t_2}{T}) - \widehat{C}(\frac{t_1}{T}, \frac{t_2}{T})]^2\ [w_{1,LL-UR}(\frac{t_1}{T}, \frac{t_2}{T})]^2. \tag{14}$$

The AD_1 and the IAD_1 estimators proved useful for the new classes of copulas investigated in this paper.

The WMLE consists in a weighted robustification of the MLE computed in two steps. In the first step atypical points are identified by a high breakdown point covariance matrix estimator. There are many high breakdown point estimators of multivariate location and scatter that could be used in this preliminary phase. In the application worked out in Section 4 we use the robust Stahel-Donoho (SD) estimator based on projections (Stahel, 1981 and Donoho, 1982) which is implemented in the free *R* software. Those points with statistical distance to the center of the ellipsoid is greater than some cutoff point (the 0.95-quantile of a chisquare random variable) are given zero weight. In the second step, a parametric copula family is fitted to the points defined by the hard rejection rule. Note that no particular distributions (marginals or copula) were assumed for the data.

Table 1: *Summary of results from simulations. Winners under uncontaminated and at the contaminated models.*

Copula	*Copula type*	λ_L	λ_U	*No contamination*	*Contamination*
				Small SS — Large SS	*Small SS — Large SS*
t-student	Elliptical	$\surd$	$\surd$	MLE or AD_1	$WMLE$ or AD_1
Gumbel	Archimed./EV		$\surd$	MLE or IAD_1	$WMLE$ or AD_1 or $URAD_1$
BB7	Archimedean	$\surd$	$\surd$	MLE or IAD_1	$WMLE$ or IAD_1

Notation in table: SS, Sample Size.

We have ran simulations for assessing the performance of these estimates for two parametric copula families, the elliptical t-student and the archimedean BB7 (notation of Joe, 1997). These estimators are compared to the MLE and the relative efficiency among estimates are assessed by comparing their mean squared error from Monte Carlo simulations. Under the true model these estimates are expected to possess small bias but large variances when compared to the MLE. However, the simulations showed that for the majority of scenarios considered, they possess small bias and small variance, outperforming the MLE.

We considered ϵ-contaminated models, where a proportion ϵ of observations were replaced by atypical ones generated from a contaminating distribution F^*. We set ϵ equal to 0% , 5%, and 10%, and F^* as the bivariate normal distribution with correlation coefficient $\rho = 0.00$ and very small variances, acting as a point mass contamination. There were 5 possibilities for the location of the contaminating distribution F^*: the center of the unit square and the regions nearby the 4 corners. Contaminated data generation was monitored such that points falling outside the unit square were discarded. Of course, many other contamination schemes are possible. The way our experiments were designed covered the worst possible contaminating scenarios.

For each experiment considered we compute the MLE, the WMLE, and all 24 WMDE estimates, and report their mean value and mean-squared error. Three sample sizes (50, 100, and 300) were considered. The number of repetitions for each one was 1000.

Summary of results for all copula models are given in Table 1. For the Gumbel copula, as expected, accuracy and precision of all estimators increase with sample size. We have here a very nice result, since we are able to choose an *overall winner*, despite the sample size, contamination location, and strength of dependence. For the BB7 copula and under no contamination: The MLE is the best estimator. However, the WMDE estimator IAD_1 shows up as a good option under the set ups considered.

4 Forecasting volatility for Brazilian stocks

Volatility plays an important role when managing risks, pricing options, composing portfolios. However it must be estimated and there are many sources of uncertainty including model specification and estimation. The realized volatility is an unbiased and highly efficient model free measure of the daily return variability computed from high-frequency intraday returns. The realized variance may be simply defined as the sum of the squared high-frequency intraday returns over this interval (see theoretical details in Andersen, Bollerslev, Christoffersen and Diebold, 2006).

High frequency data possess unique characteristics, not present in low frequency data

(daily, weekly, monthly), calling for specific data treatments. Transactions (with variable volumes) occur in irregular time space, data show cicles and different market activity patterns along the day, there may be asynchronicity in the data, and the bid-ask bounce effect may distort inferences. All these features make their statistical analysis more interesting. On the other hand, the number of observations is huge, increasing the chances of many types of errors such as transaction and recording errors.

We examine the time-varying behavior of the daily equity return realized volatilities obtained from high-frequency intraday transaction prices in the BOVESPA. The temporal dynamics of the series are modeled using pair-copulas. That is, we model the temporal dependence of the realized volatility at day t, on the past values at days $t-1, t-2, \cdots$ using pair-copulas. To the best of our knowledge no one has already taken this approach for modeling series of realized volatilities. Moreover, we apply classical and robust estimation. Applying robust estimates in this context seems promising due to the large amount of data and the extensive data manipulation that may increase the proportion of atypical values, therefore increasing the bias of predictions.

Data are composed by high-frequency returns computed from continuously recorded transaction prices of the seven most traded Brazilian stocks (PETR4 (Petrobras), VALE5 (Vale), TNLP4(Telemar), USIM5 (Usiminas), BBDC4 (Bradesco), CSNA3 (Siderrgica Nacional) and ITAU4 (Itauunibanco)) was provided by BOVESPA and covers the 7-months period from October, 01, 2008 to April, 30, 2009. Liquid stocks are needed to avoid the negative autocorrelation induced by the small time intervals between records. Each data record is similar to the "Trades and Quotes" (TAQ) provided by the NYSE, and contains information on the price, volume traded, day and time of trading, dates and names of the trading firms (those buying and selling).

To eliminate the impacts caused by the changes in the BOVESPA closing time (summer time), data were firstly expressed using the Greenwich mean time format. For data alignment at fixed time intervals, instead of the common practice denominated *before*[2], we obtain the volume-weighted average price (VWAP). The VWAP gives rise to a smaller realized variance, since it is closest to the efficient price instead of the closing price.

We divide the trading period (7 hours) in 84 Δ-intervals, $\Delta = 5min$. Let $P_{t^*,k}$ and $Q_{t^*,k}$, $k = 1, \cdots, n$, represent the k-th price and volume for a Brazilian stock during some time interval Δ corresponding to time t^*, $t^* = 5, 10, \cdots, 420$.

The real value log-price p_{t^*} of this stock for a $5min$ time interval is given by

[2] "Before" makes use of the most recent observation, or the closest, with respect to the desired minute, and obtains the average of the bid-ask values through a linear interpolation of the log-price.

$$p_{t^*} = \ln\left(\frac{P_{t^*,1}\cdot Q_{t^*,1} + P_{t^*,2}\cdot Q_{t^*,2} + \cdots + P_n \cdot Q_{t^*,n}}{Q_{t^*,1} + Q_{t^*,2} + \cdots + Q_{t^*,n}}\right). \tag{15}$$

The $5min$ intra-day continuously compounded return r_{t^*} on this stock from time to $t^* - \Delta$ to t^* is computed as

$$r_{t^*} = p_{t^*} - p_{t^*-\Delta}. \tag{16}$$

Intra-day data may present seasonality and volatility clusters. To eliminate the negative autocorrelation present in the r_{t^*} due to microstructure effects, an ARMA(p, q) filter was applied to the intra-day data before computing the realized volatility.

The realized variance (RV_t) at day t, $t = 1, \ldots, T$ is defined as:

$$RV_t = \sum_{t^* \in \text{ day } t} r_{t^*}^2 \tag{17}$$

where T is the series length. We note that consistency of the estimator is attached to $\Delta \to 0$. The realized volatility $(RVOL_t)$ and the log-realized volatility $(RLVOL_t)$ at day t are

$$RVol_t = \sqrt{RV_t} \quad \text{and} \quad LRVol_t = \ln\left(RVol_t\right) \tag{18}$$

Now we investigate the temporal dependencies within the series of LRVol using pair-copulas. We split the data in two parts, one for estimation (6 months), and the other for the one-step-ahead out-of-sample predictions (1 month, April/2010). Before any modeling we explore the series' sample acf and pacf. They indicate that the RLVol at times $t-1$, $t-2$, and $t-3$ are able to explain the behavior of the series at time t. Thus, we will fit a 4-dimensional pair-copula to the original and lagged series of RLVol. The Canonical vine is interesting because the three bivariate unconditional copulas in tree 1 will be linking the original series on time t with the three other lagged ones.

The first step is to find the best unconditional distribution representing the univariate series of LRVol. The values of the skewness and kurtosis coefficients in Table (1) suggest using a skew-t distribution (Hansen (1994), Patton (2006), Fantazzini (2006)).

We fit by maximum likelihood a skew-t distribution to the 6-months series of LRVol and using the c.d.f. of the fitted models we obtain the pseudo uniform$(0, 1)$ data. Marginal fits should be carefully checked since a poor fit will result in probability integral transforms not being standard uniform or *i.i.d.*. As a consequence, any copula model will be mis-specified.

Table 2: *Basic statistics for the Brazilian stocks RLVol series.*

	Mean	Stdev	Skewness	Kurtosis	Maximum
$RLVol_{PETR4}$	1.8970	0.4148	0.9988	5.6532	4.6006
$RLVol_{VALE}$	1.8970	0.4148	0.9988	5.6532	4.6006

To estimate the bivariate copulas (3 unconditional and 3 conditional), we examined the uniform data scatter plots (see Figure 2) and considered as possible candidates five copula families: Normal, t-student, BB7, Gumbel, and the product copula. To find the best copula fit, we compared the penalized log-likelihood (AIC), examined the pp-plots based on the estimated and the empirical copula, and computed a GOF test statistic (Genest and Rémillard (2005) and Genest, Rémillard, and Beaudoin (2007)).

The chosen pair-copula decomposition, along with best classical and robust copula fits with parameter estimates, are shown in Table (2).

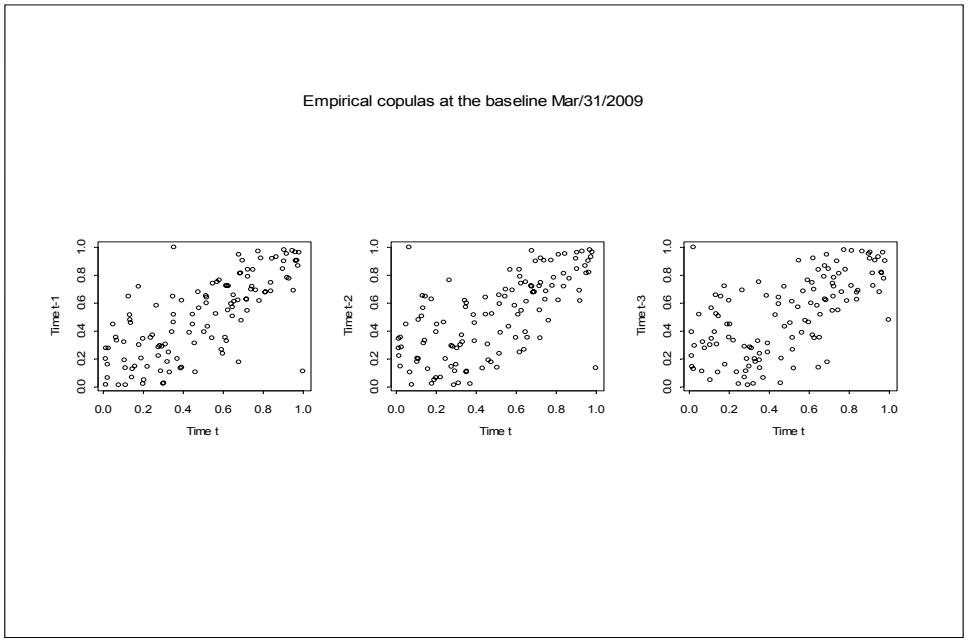

Figure 2: *Probability transformed pairs of PETROBRAS LRVol.*

Table 3: *Best classical and robust canonical vine fits to the PETROBRAS RLVol.*

	CLASSICAL			ROBUST		
Tree 1	$(t, t-1)$	$(t, t-2)$	$(t, t-3)$	$(t, t-1)$	$(t, t-2)$	$(t, t-3)$
	t-student	Gumbel	t-student	Gumbel	Gumbel	Normal
	(0.71)	(1.72)	(0.59)	(2.36)	(2.01)	(0.63)
Tree 2	$(t-1, t-2 \mid t)$		$(t-1, t-3 \mid t)$	$(t-1, t-2 \mid t)$		$(t-1, t-3 \mid t)$
	t-student		Gumbel	Gumbel		Gumbel
	(0.47)		(1.29)	(1.44)		(1.36)
Tree 3	$(t-2, t-3 \mid t, t-1)$			$(t-2, t-3 \mid t, t-1)$		
	t-student			Gumbel		
	(0.36)			(1.43)		

The outliers that can be seen in the LR and UL corners of plots of Figure 2 affect the copula fits. The changes are in copula family and in parameters estimates values, see Table 3. The robust fits reflect the pattern of the majority of days and are expected to provide better forecasts.

For predicting, we keep the estimated model and incorporate each new daily observation as it comes, updating the data. Twenty realized log-volatility predictions are obtained. We assess and compare the performance of the classical and robust forecast methods by computing the sum of the squared differences between both forecasts and the true realized volatilities. The robust method performed better providing a value of 0.4042 whereas the classical method yielded a sum of 1.4057. Figure **??** shows the twenty predictions ander

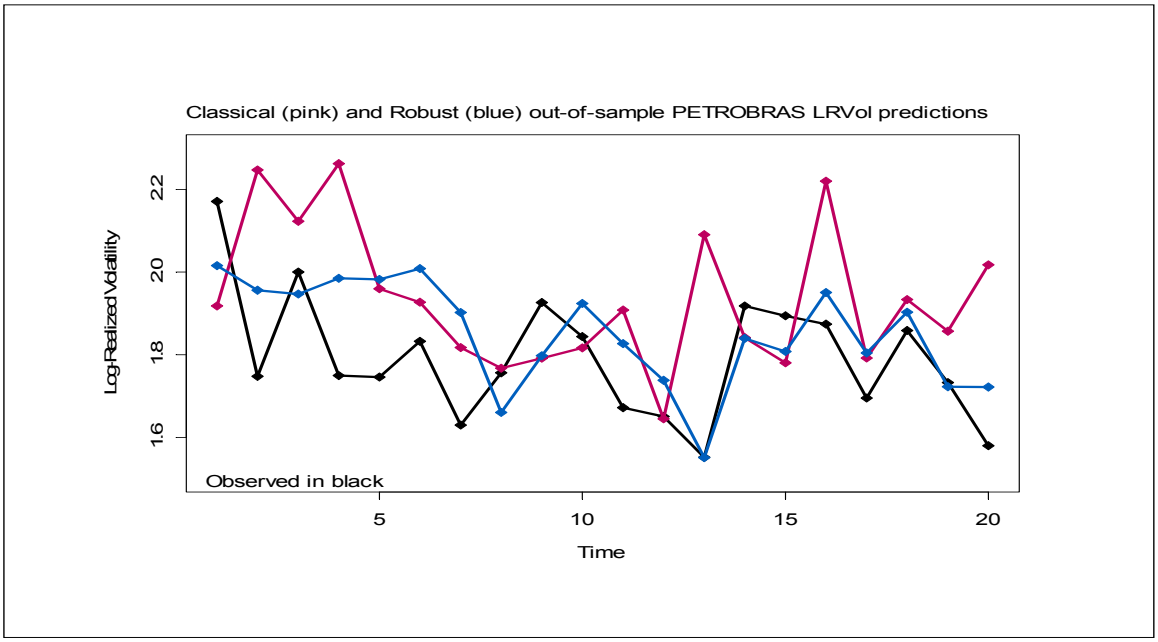

Figure 3: *PETROBRAS out-sample Classical and Robust PC volatility forecasts.*

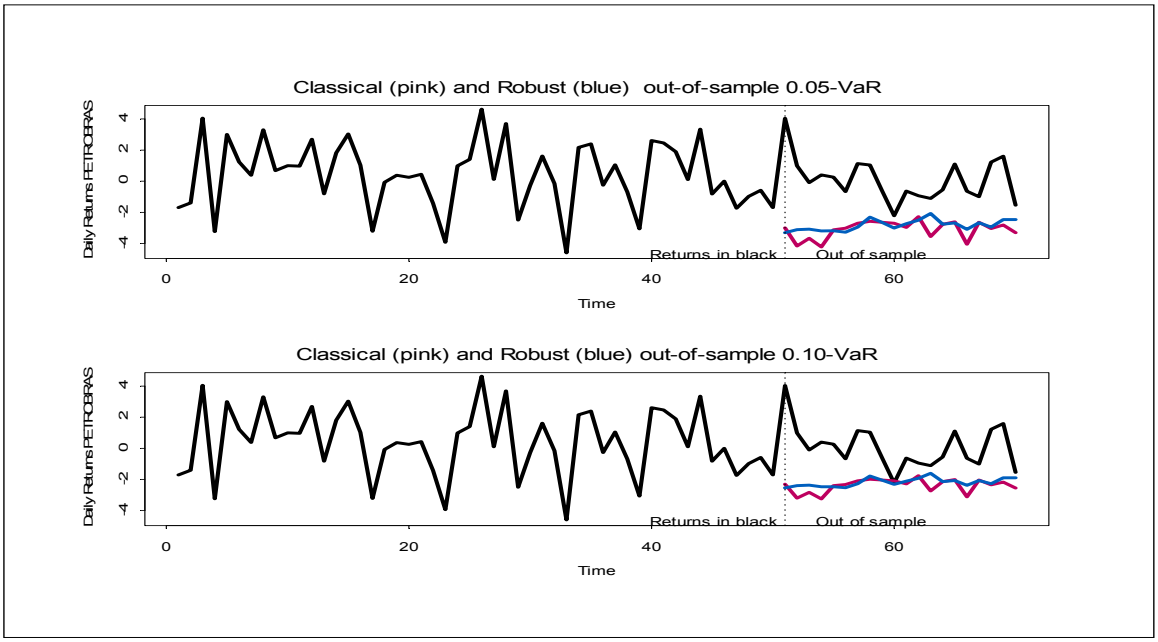

Figure 4: *PETROBRAS out-sample Classical and Robust PC Value-at-Risk estimates.*

5 Discussion

In this paper we extend the previous work of Mendes, Melo and Nelsen (2007) and study robust estimation of pair-copula models. The extention is straightforward since pair-copulas is just a hierarchical decomposition of a multivariate copula into a cascade of bivariate copulas, and estimation takes place at the level of two-dimensional data.

For each data set (contaminated or not) we are able to propose a robust estimator as good as the maximum likelihood estimates. When data are contaminated (or when this is an issue) we propose to compare the classical and robust estimates. We provide two convincing examples where robust estimates are needed. Different weight functions emphasize different regions on the unit cube where contaminations may be located. The resulting WMDE estimators are compared to the classical maximum likelihood estimators MLE, and to their weighted version WMLE, an estimator obtained in two steps. All estimators are compared in a comprehensive simulation study. For each epsilon-contaminated pair-copula model specified, we show that there is a robust estimator improving over the MLE and able to capture the correct strength of dependence of the data, despite the contamination percentual and location, and the sample size.

For any other copula family not considered here, the simulations may be easily implemented and run relatively fast. Estimators sample distributions may be assessed by simulations. We did some experimentation and found a well behaved distribution for some of them. Tables may be constructed for testing hypothesis. We are not addressing the

important problem *Which copula is the right one?* But we are indeed providing guidance for estimating several copula models. A rule of thumb: Always compare a classical and a robust fit. In summary, whenever one suspects there may exist contaminations, we would recommend the $WMLE$ for elliptical copulas, and the W^2 and the IAD_2 for all other copula families. In the case one is not sure whether or not there are contaminations, he/she should compare the MLE, W^2, AD_2, and IAD_2 estimators. Methodology seems promising for forecasting series of realized volatilities possessing short and long range dependence and different tail dependence at consecutive small and large values.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. 2007. "Pair-copula constructions of multiple dependence." *Insurance: Mathematics and Economics*, 2, 1, 1-25.

Andersen, T.G., Bollerslev, T., Christoffersen, P. F., Diebold, F.X. (2006). "Volatility and Correlation Forecasting" in G. Elliot, C.W.J. Granger, and Alan Timmermann (eds), *Handbook of Economic Forecasting*, Amsterdan: North Holland, 778-878.

Bedford, T., and Cooke, R. M. 2001. "Probability density decomposition for conditionally dependent random variables modeled by vines." *Annals of Mathematics and Artificial Intelligence*, 32, 245-268.

Bedford, T., and Cooke, R. M. 2002. "Vines - a new graphical model for dependent random variables." *Annals of Statistics*, 30, 4, 10311068.

Berg, D. 2008. "Copula Goodness-of-fit testing: An overview and power comparison." *European Journal of Finance*.

Berg, D., and Aas, K. 2008. "Models for construction of multivariate dependence: A comparison study." *European Journal of Finance*.

Cooke, R. M., and Bedford, T. 1995. "Reliability methods as management tools: Dependence modeling and partial mission success." *London: Chameleon Press*.

Demarta, S., and McNeil, A. 2005. "The t Copula and Related Copulas". *International Statistical Review*, 73(1).

Embrechts P., McNeil A., and Straumann D. 1999. "Correlation and Dependence in Risk Management: Properties and Pitfalls.", Preprint ETH Zurich. Available at http://www.math.ethz.ch/#embrechts; http://citeseer.ist.psu.edu/article/embrechts99correlation.html

Embrechts P., McNeil, A., and Straumann, D. 2001. "Correlation and Dependency in Risk Management: Properties and Pitfalls." *Value at Risk and Beyond*. Cambridge University Press.

Fantazzini, D. 2006. "Dynamic Copula Modelling for value-at-Risk." *Frontiers in Finance and Economics*. Available at SSRN: http://ssrn.com/abs=944172.

Fischer, M., Köck, C., Schlüter, S., and Weigert, F. 2008. "Multivariate copula models at work: outperforming the 'Desert Island Copula'?" http://www.statistik.wiso.uni-erlangen.de/forschung/d0079.p

Gatzert, N., Schmeiser, H., and Schuckmann, S. 2008. "Enterprise risk management in financial group: analysis of risk concentration and default risk." *Financial Markets and Portfolio Management*. Vol. 22, 3, 241-258.

Genest, C., Quessy, J.F., and Rémillard, B. 2006. "Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation". *Scandinavian J. of Statistics*, 33, 337-366.

Genest, C., and Rémillard, B. 2005. "Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models". Technical Report G-2005-51, GERAD, Montreal, Canada.

Genest, C., Rémillard, B., and Beaudoin, D. 2007. "Omnibus goodness-of-fit tests for copulas. A review and a power study". Working paper, Université Laval.

Genest, C., and Rivest, L. P. 1993. "Statistical inference procedures for bivariate Archimedian copulas", *Journal of the American Statistical Association*, 88, 423, 1034-1043.

Hampel, F. R., Ronchetti, E. M., Rousseeaw, P. J. 1986. "Robust Statistics: The Approach based on influence functions". *J. Willey and Sons, Inc.*

Hansen, B. 1994. "Autoregressive Conditional Density Estimation." *International Economic Review*, 35,3.

Joe, H. 1996. "Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters". In L. Rüchendorf, B. Schweizer, and M. D. Taylor (Eds.), *Distributions with fixed marginals and Related Topics.*

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

Joe, H., Li, H., and Nikoloulopoulos, A. K. 2008. "Tail dependence functions and vine copulas." *Math technical report*, 2008-3, Washington State University.

Kassberger, S., and Kiesel, R. 2006. "A fully parametric approach to return modeling and risk management of hedge funds." *Financial Markets and Portfolio Management.* V. 20, 472491.

Kurowicka, D., and Cooke, R. M. 2000. "Conditional and partial for graphical uncertainty models". *Recent Advances in Reliability Theory. Methodology, Practice, and Inference.* By Nikolaos Limnios, Mikhail Stepanovich Nikulin, Birkhäuser.

Kurowicka, D., and Cooke, R. M. 2001. "Conditional, partial, and rank correlation for the elliptical copula; Dependence modeling in uncertainty analysis". *Proceedings ESREL.*

Kurowicka, D., and Cooke, R. M. 2006. "Completion Problem with Partial correlation Vines". *Linear Algebra and its Applications.* Vol. 418, No. 1, pp. 188-200.

Markowitz, H. M. 1959. "Portfolio Selection: Efficient Diversification of Investments". *New York: J. Willey.*

Mendes, B. V. M., and Leal, R. P. C. 2005. "Robust Multivariate Modeling in Finance". *International Journal of Managerial Finance*. V.1, N. 2, pp. 95-106.

Mendes, B. V. M., and Leal, R. P. C. 2009. "Portfolio Management with Semi-Parametric Bootstrapping". Forthcoming in *Journal of Risk Management in Financial Institutions*.

Mendes, B. V. M., Melo, E. F. L., e Nelsen, R. B. 2007. "Robust fits for copula models". *Communication in Statistics*, 36, pp.997-1017.

Min, A., and C. Czado 2008. "Bayesian inference for multivariate copulas using pair-copula constructions". Preprint, available under http://www-m4.ma.tum.de/Papers/index.html.

Misiewicz, J., Kurowicka, D., and Cooke, R.M. 2000. "Elliptical copulae". To appear.

Nelsen, R.B. 2007. "An introduction to copulas." *Lectures Notes in Statistics*. New York: Springer.

Patton, A. 2001. "On the out-of-sample importance of skewness and asymmetric dependence for asset allocation". *Journal of Financial Econometrics*, 2, 1, 130-168.

Patton, A. 2006. "Modeling asymmetric exchange rate dependence." *International Economic Review*, 47, 2.

Ragea, V. 2003. "Testing correlation stability during hetic financial markets". *Financial Markets and Portfolio Management*. V. 17, 3, 289308.

Rockinger, M., and Jondeau, E. 2001. "Conditional dependency of financial series: An application of copulas". *HEC Paris DP 723*.

Rosenblatt, M. 1952. "Remarks on a multivariate transformation". *Annals of Mathematical Statistics*, 23, 470-472.

Rousseeuw, P.J., and Leroy, A.M. 1987. "Robust Regression and Outlier Detection". *New York: John Wiley & Sons*.

Scott, D.W. 1992. "Multivariate Density Estimation." *New York: Wiley*.

Sklar, A. 1959. "Fonctions de répartition á n dimensions et leurs marges." *Publ. Inst. Statist. Univ. Paris*, 8, 229-231.

Sklar, A. 1996. "Random variables, distribution functions, and copulas (a personal look backward and forward)." *Distributions with Fixed Marginals and Related Topics*, ed. by L. Rüschendor, B. Schweizer, and M. Taylor,. 1-14. IMS, Hayward, CA.

Yule, G. U., and Kendall, M. G. 1965. "An Introduction to the Theory of Statistics." 14th ed. *London: Charles Griffin & Co.*