Hidden Markov structures for dynamic copulae *

Weining Wang[†], Ostap Okhrin [‡], Wolfgang Karl Härdle[§]

February 2, 2014

Abstract

Understanding the time series dynamics of a multivariate dimensional dependency structure is a challenging task. A multivariate covariance driven Gaussian or mixed normal time varying models are limited in capturing important data features such as heavy tails, asymmetry, and nonlinear dependencies. This research aims at tackling this problem by proposing and analysing a hidden Markov model (HMM) for hierarchical Archimedean copulae (HAC). The HAC constitute a wide class of models for multivariate dimensional dependencies, and HMM is a statistical technique for describing regime switching dynamics. HMM applied to HAC flexibly models multivariate dimensional non-Gaussian time series.

We apply the expectation maximization (EM) algorithm for parameter estimation. Consistency results for both parameters and HAC structures are established

*The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 Ökonomisches Risiko, Humboldt-Universität zu Berlin is gratefully acknowledged. We thank Prof. Cheng-Der Fuh for his comments. We remain responsible for errors and omission.

[†]Hermann-Otto-Hirschfeld Junior Professor in Nonparametric Statistics and Dynamic Risk Management at the Ladislaus von Bortkiewicz Chair of Statistics of Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany. Email:wangwein@cms.hu-berlin.de

[‡]Professor at the Ladislaus von Bortkiewicz Chair of Statistics of Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany. Email:ostap.okhrin@wiwi.hu-berlin.de

[§]Professor at Humboldt-Universität zu Berlin and Director of CASE - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany. Singapore Management University, 50 Stamford Road, Singapore 178899. Email:haerdle@wiwi.hu-berlin.de

in an HMM framework. The model is calibrated to exchange rate data with a VaR application. This example is motivated by a local adaptive analysis that yields a time varying HAC model. We compare the forecasting performance with other classical dynamic models. In another, second, application we model a rainfall process. This task is of particular theoretical and practical interest because of the specific structure and required untypical treatment of precipitation data.

Keywords: Hidden Markov Model, Hierarchical Archimedean Copulae, Multivariate Distribution, Dynamic Dependency Structure

JEL classification: C13, C14, G50

1 Introduction

Modeling multi-dimensional time series is an often underestimated exercise of routine econometrical and statistical work. This slightly pejorative attitude towards day to day statistical analysis is unjustified since actually the calibration of time series models in multi dimensions for standard data sizes is not only difficult on the numerical side but also on the mathematical side. Computationally speaking, integrated models for multi dimensional time series become more involved when the parameter space is too large. Consequently the mathematical and econometrical aspect become more difficult since the parameter space become to complex, especially when time variation of these is added. An example is the multivariate GARCH(1,1) BEKK model that for even two dimensions has an associated parameter space of dimension 12. For moderate sample sizes, the parameter space dimension might be in the range of the sample size or even bigger. This data situation has evoked a new strand of literature on dimension reduction via penalty methods.

In this paper we take a different route, by calibrating an integrated dynamic model with unknown dependency structure among the d dimensional time series variables. More precisely, the unknown dependency structure may vary within a set of given dependencies.

These dependency structures might have been selected via a preliminary study as described e.g. in Härdle, Herwartz and Spokoiny (2003). The specific dependence at each time t is unknown to the data analyst, but depends on the dependency pattern at time $t-1$. Therefore, hidden Markov models (HMM) naturally come into play. This leaves us with the task of specifying the set of dependencies.

An approach based on assuming a multivariate Gaussian or mixed normals is handicapped in capturing important types of data features such as heavy tails, asymmetry, and nonlinear dependencies. Such a simplification is certainly in practical questions concerning the tails too restrictive and might lead to biased results. Copulae are one possible approach to solving these problems. Moreover, copulae allow us to separate the marginal distributions and the dependency model, see Sklar (1959). In recent decades, copula-based models have gained popularity in various fields like finance, insurance, biology, hydrology, etc. Nevertheless, many basic multivariate copulae are still too restrictive and the extension to many parameters leads initially to a nonparametric density estimation problem that suffers of course from the curse of dimensionality. A natural compromise is the class of hierarchical Archimedean copulae (HAC). An HAC allows a rich copula structure with a finite number of parameters. Recent research has demonstrated their flexibility see McNeil and Nešlehová (2009), Okhrin, Okhrin and Schmid (2013), Whelan (2004).

Insights into the dynamics of a copulae has been offered by Chen and Fan (2005) assuming an underlying Markovian structure, and a specific class of copulae functions for the temporal dependence; Patton (2004) considers an asset-allocation problem with a time-varying parameter of bivariate copulae; while Rodriguez (2007) studies financial contagion using switching-parameter bivariate copulae. Similarly, Okimoto (2008) provides strong empirical evidence that a Markov switching multivariate normal model is not appropriate for the dependence structures in international equity markets.

Moreover, an adaptive method isolating time varying dependency structure via a local change point method (LCP), has been proposed in Giacomini, Härdle and Spokoiny (2009), Härdle, Okhrin and Okhrin (2013). Figure 1 presents an analysis of HAC for exchange rate data using LCP on a moving window, where the window sizes are adaptively

selected by the LCP algorithm. It plots the changes of estimated structure (upper panel) and parameters (lower panel) in each window over time. In particular the upper panel y-axis corresponds to dependency structures picked by estimation of a three dimensional copulae, and the lower panel y-axis shows the estimated two dependency parameters (value converted to Kendall's τ) corresponding to the estimated structure. In more details, we have three dimensional exchange rates series: P (GBP/EUR), Y(JPY/EUR), D(USD/EUR), and the label P(DY) means then the structure D and Y have stronger dependency than other possible pairs. For a more detailed introduction to HAC and their structures, please refer to Section 2.1. One observes that the structure very often remains the same for a long time, and the parameters are only slowly varying over time. This indicates that the dynamics of HAC functions is likely to be driven by a Markovian sequence seemingly determining the structures and parameter values. This observation motivate us to pursue a different path of modeling the dynamics. Instead of taking a local point of view, we adopt a global dynamic model HMM for the change of both the tree structure and the parameters of the HAC along the time horizon. In this situation, the not directly observable underlying Markov process X determines the state of distributions of Y.

HMM has been widely applied to speech recognition, see Rabiner (1989), molecular biology, and digital communications over unknown channels. Markov switching models are introduced to economic literature by Hamilton (1989), where the trend component of a univariate nonstationary time series is changing according to an underlying Markov chain. Later it was extended and combined with many different time series model, see e.g. Pelletier (2006). For estimation and inference issues in HMM, see Bickel, Ritov and Rydén (1998) and Fuh (2003), among others.

In this paper, we propose a new type of dynamic model, called HMM HAC, by incorporating HAC into an HMM framework. The theoretical problems such as parameter consistency and structure consistency are solved. The expectation maximization (EM) algorithm is developed in this framework for parameter estimation. See Section 2 for the model description, Section 3 for theorems about consistency and asymptotic normality. EM algorithm and computation issues are in Section 4. Section 5 is for the simulation

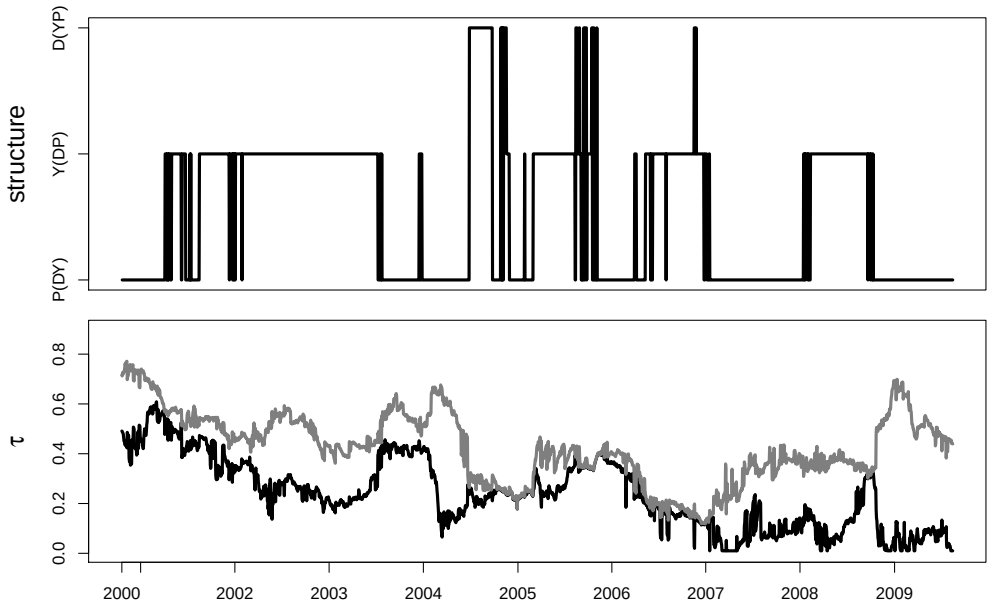


Figure 1: LCP for exchange rates: structure (upper) and parameters (lower, θ_1(grey) and θ_2(black)) for Gumbel HAC. $m_0 = 40$ (starting value for the window size in the algorithm).

study, and Section 6 is for applications. The technical details are put into the Appendix.

2 Model Description

In this section, we introduce our model and estimation method. Subsection 2.1 briefly introduces the definition and properties of HAC, and Subection 2.2 introduces the HMM HAC. Finally in the last subsection, we describe the estimation and algorithm we used.

2.1 Copulae

Let $Z_1, \ldots, Z_d$ be r.v. with continuous cumulative distribution function (cdf) $F(\cdot)$. The Sklar theorem guarantees the existence and uniqueness of copula functions by stating:

Theorem 2.1 (Sklar's theorem)**.** *Let F be a multivariate distribution function with margins $F_1^{\mathfrak{m}}, \ldots, F_d^{\mathfrak{m}}$, then a copula C exists such that*

$$F(z_1, \ldots, z_d) = C\{F_1^{\mathfrak{m}}(z_1), \ldots, F_d^{\mathfrak{m}}(z_d)\}, \quad z_1, \ldots, z_d \in \mathbb{R}.$$

If $F_i^{\mathfrak{m}}(\cdot)$ are continuous for $i = 1, \ldots, d$ then $C(\cdot)$ is unique. Otherwise $C(\cdot)$ is uniquely determined on $F_1^{\mathfrak{m}}(\mathbb{R}) \times \cdots \times F_d^{\mathfrak{m}}(\mathbb{R})$.

Conversely, if $C(\cdot)$ is a copula and $F_1^{\mathfrak{m}}, \ldots, F_d^{\mathfrak{m}}$ are univariate distribution functions, then the function F defined above is a multivariate distribution function with margins $F_1^{\mathfrak{m}}, \ldots, F_d^{\mathfrak{m}}$.

The family of Archimedean copulae is very flexible and captures tail dependency, has an explicit form, and are simple to estimate,

$$C(u_1, \ldots, u_k) = \phi\{\phi^{-1}(u_1) + \cdots + \phi^{-1}(u_d)\}, \quad u_1, \ldots, u_d \in [0, 1], \tag{1}$$

where $\phi(\cdot)$ is defined as the generator of the copula and depends on a parameter θ, see Nelsen (2006). $\phi(\cdot)$ is d monotone, and $\phi(\cdot) \in \mathfrak{L} = \{\phi(\cdot) : [0; \infty) \to (0, 1] \,|\, \phi(0) = 1, \phi(\infty) = 0; (-1)^j \phi^{(j)} \geq 0; j = 1, \ldots, d-2\}$. As an example, the Gumbel generator is given by $\phi(x) = \exp(-x^{1/\theta})$ for $0 \leq x < \infty,\ 1 \leq \theta < \infty$.

In this work we consider less restrictive compositions of simple Archimedean copulae leading to a Hierarchical Archimedean Copula (HAC) $C(u_1, \ldots, u_d; \boldsymbol{\theta}, s)$, where $s = \{(\ldots(i_1 \ldots i_{j_1}) \ldots (\ldots) \ldots)\}$ denotes the structure of HAC, with $i_\ell \in \{1, \ldots, d\}$ being a reordering of the indices of the variables and s_j the structure of the subcopulae with $s_d = s$, and $\boldsymbol{\theta}$ is the set of copula parameters. For example, the fully nested HAC (see Figure 2, left) can be expressed by

$$\begin{aligned} & C(u_1, \ldots, u_d; \boldsymbol{\theta}, s = s_d) = C\{u_1, \ldots, u_d; (\theta_1, \ldots, \theta_{d-1})^\top, ((s_{d-1})d)\} \\ = \ & \phi_{d-1,\theta_{d-1}}(\phi^{-1}_{d-1,\theta_{d-1}} \circ C\{u_1, \ldots, u_{d-1}; (\theta_1, \ldots, \theta_{d-2})^\top, ((s_{d-2})(d-1))\} + \phi^{-1}_{d-1,\theta_{d-1}}(u_d)), \end{aligned}$$

where $s = \{(\ldots(12)3)\ldots)d)\}$. On the RHS of Figure 2 we have the partially nested HAC with $s = ((12)(34))$ in dimension $d = 4$. For more details of HAC, see Joe (1997), Whelan (2004), Savu and Trede (2010), Okhrin et al. (2013).

It is worth noting though that not all generator functions can be mixed within one HAC. We therefore concentrate on one single generator family within one HAC. This boils

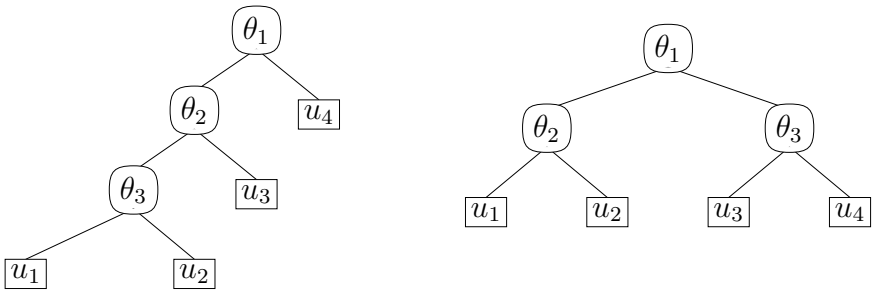


Figure 2: Fully and partially nested copulae of dimension $d = 4$ with structures $s = (((12)3)4)$ on the left and $s = ((12)(34))$ on the right

down to binary structures, i.e., at each level of the hierarchy only two variables are joined together. This makes in fact the architecture very flexible and yet parsimonious.

Note that for each HAC not only are the parameters unknown, but also the structure has to be determined. We adopt the modified computational steps of Okhrin et al. (2013) to estimate the HAC structure and parameters. One estimates the marginal distributions either parametrically or nonparametrically. Then (assuming that the marginal distributions are known) one selects the couple of variables with the strongest fit and denotes the corresponding estimator of the parameter at the first level by $\hat{\theta}_1$ and the set of indices of the variables by I_1. The selected couple is joined together to define the pseudo-variables $z_1 = C\{(I_1); \hat{\theta}_1, \phi_1\}$. Next, one proceeds in the same way by considering the remaining variables and the new pseudo-variable. At every level, the copula parameter is estimated by assuming that the margins as well as the copula parameters at lower levels are known. This algorithm allows us to determine the estimated structure of the copula recursively.

2.2 Incorporating HAC into HMM

A hidden Markov model is a parameterized time series model with an underlying Markov chain viewed as missing data, as in Leroux (1992), Bickel et al. (1998), and Gao and Song (2011). Specifically, in the HMM HAC framework, let $\{X_t, t \geq 0\}$ be a stationary Markov chain of order one on a finite state space $D = \{1, 2, \ldots, M\}$, with transition probability

matrix $P = \{p_{ij}\}_{i,j=1,\ldots,M}$ and initial distribution $\pi = \{\pi_i\}_{i=1,\ldots,M}$.

$$\begin{aligned} \mathrm{P}(X_0 = i) &= \pi_i, && (2) \\ \mathrm{P}(X_t = j | X_{t-1} = i) &= p_{ij} && (3) \\ &= \mathrm{P}(X_t = j | X_{t-1} = i, X_{t-2} = x_{t-2}, \ldots, X_1 = x_1, X_0 = x_0), \\ &\quad i, j = 1, \ldots, M \end{aligned}$$

Let $\{Y_t, t \geq 0\}$ be the associated observations, and they are adjoined with $\{X_t, t \geq 0\}$ in such a way that given $X_t = i, i = 1, \ldots, M$, the distribution of Y_t is fixed:

$$\begin{aligned} (X_t | X_{0:(t-1)}, Y_{0:(t-1)}) &\overset{\mathcal{L}}{=} (X_t | X_{t-1}), && (4) \\ (Y_t | Y_{0:(t-1)}, X_{(0:t)}) &\overset{\mathcal{L}}{=} (Y_t | X_t), && (5) \end{aligned}$$

where $Y_{0:(t-1)}$ stands for $\{Y_0, \ldots, Y_{t-1}\}$, $t < T$.

Let $f_j\{\cdot\}$ be the conditional density of Y_t given $X_t = j$ with $\boldsymbol{\theta} \in \Theta, \mathbf{s} \in S,\ j = 1, \ldots, M$ being the unknown parameters. Here, $\{X_t, t \geq 0\}$ is the Markov chain, and given $X_0, X_1, \ldots, X_T$, the $Y_0, Y_1, \ldots, Y_T$ are independent. Note that $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}) \in \mathbb{R}^{(d-1)M}$ are the unknown dependency parameters, $\mathbf{s} = (s^{(1)}, \ldots, s^{(M)})$ are the unknown HAC structure. Denote their true values by $\boldsymbol{\theta}^*$ and $\mathbf{s}^*$ respectively. Please see appendix 7.2 for more details.

For the time series $y_1, \ldots, y_T \in \mathbb{R}^d$ ($y_t = (y_{1t}, y_{2t}, y_{3t}, \ldots, y_{dt})^\top$) and the unobservable (or missing) $x_1, \ldots, x_T$ from the given hidden Markov model, define π_{x_t} as the π_i for $x_0 = i, i = 1, \ldots, M$, and $p_{x_{t-1}x_t} = p_{ji}$ for $x_{t-1} = j$ and $x_t = i$. The full likelihood for $\{x_t, y_t\}_{t=1}^T$ is then:

$$p_T(y_{0:T}; x_{0:T}) = \pi_{x_0} \prod_{t=1}^{T} p_{x_{t-1}x_t} f_{x_t}(y_t), \tag{6}$$

and the likelihood for only the observations $\{y_t\}_{t=1}^T$ is calculated by marginalization:

$$p_T(y_{0:T}) = \sum_{x_0=1}^{M} \cdots \sum_{x_T=1}^{M} \pi_{x_0} \prod_{t=1}^{T} p_{x_{t-1}x_t} f_{x_t}(y_t). \tag{7}$$

The HAC is a parametrization of $f_{x_t}(y_t)(x_t = i)$, which helps to properly understand the dynamics of a multivariate distribution. Up to now, typical parameterizations have been mixtures of log-concave or elliptical symmetric densities, such as those from Gamma or Poisson families, which are not flexible enough to model multi dimensional time series. The advantage of the copula is that it splits the multivariate distribution into its margins and a pure dependency component. In other words, it captures the dependency between variables eliminating the impact of the marginal distributions as introduced in the previous subsection.

Furthermore, we incorporate this procedure into the HMM framework. We denote the underlying Markov variable X_t as a dependency type variable. If $x_t = i$, the parameters $(\boldsymbol{\theta}^{(i)}, s^{(i)})$ determined by state $i = 1, \ldots, M$ take values on $\Theta \times S$, where S is a set of discrete candidate states corresponding to different dependency structures of the HAC, and Θ is a compact set in $\mathbb{R}^{d-1}$ wherein the HAC parameters take their values. Therefore,

$$f_i(\cdot) = c\{F_1^{\mathfrak{m}}(y_1), F_2^{\mathfrak{m}}(y_2), \ldots, F_d^{\mathfrak{m}}(y_d), \boldsymbol{\theta}^{(i)}, s^{(i)}\} f_1^{\mathfrak{m}}(y_1) f_2^{\mathfrak{m}}(y_2) \cdots f_d^{\mathfrak{m}}(y_d), \tag{8}$$

with $f_i^{\mathfrak{m}}(y_i)$ the marginal densities, $F_i^{\mathfrak{m}}(y_i)$ the marginal cdf and $c(\cdot)$ the copula density, which is defined in assumption A.4 in Section 3.

Let $\boldsymbol{\theta}^{(i)} = (\theta_{i1}, \ldots, \theta_{i,d-1})^\top$ be the dependency parameters of the copulae starting from the lowest up to the highest level connected with a fixed state $x_t = i$ and corresponding density $f_i(.)$. Refining the algorithm of Okhrin et al. (2013), the multistage maximum likelihood estimator $(\hat{\boldsymbol{\theta}}^{(i)}, \hat{s}^{(i)})$ solves the system

$$\left(\frac{\partial \mathcal{L}_1}{\partial \theta_{i1}}, \ldots, \frac{\partial \mathcal{L}_{d-1}}{\partial \theta_{id-1}}\right)^\top = \mathbf{0}, \tag{9}$$

$$\begin{aligned} \text{where} \quad \mathcal{L}_j &= \sum_{t=1}^{T} w_{it} l_{ij}(Y_t), \text{ for } j = 1, \ldots, d-1, \\ l_{ij}(Y_t) &= \log\Big(c\big[\{\hat{F}_m^{\mathfrak{m}}(y_{tm})\}_{m\in\{1,\ldots,j\}}; \{\theta_{i\ell}\}_{\ell=1,\ldots,j-1}, s_m^{(i)}\big] \prod_{m\in\{1,\ldots,j\}} \hat{f}_m^{\mathfrak{m}}(y_{tm})\Big) \\ &\quad \text{for } t = 1, \ldots, T. \end{aligned}$$

where j denote the layers of a tree structure, and $\hat{F}_m^{\mathfrak{m}}(\cdot)$ is an estimator (either nonparametric with $\hat{F}_m^{\mathfrak{m}}(x) = (T+1)^{-1}\sum_{t=1}^{T} \mathbf{1}(Y_{tm} \le x)$ or parametric $\hat{F}_m^{\mathfrak{m}}(x) = F_m^{\mathfrak{m}}(x, \hat{\boldsymbol{\alpha}}_m)$) of the marginal cdf $F_m^{\mathfrak{m}}(\cdot)$. Note that nonparametric estimation of margins would lead to semi-parametric nature of our estimation. The marginal densities $\hat{f}_m^{\mathfrak{m}}(\cdot)$ are estimated parametrically or nonparametrically (kernel density estimation) corresponding to the estimation of the marginal distribution functions, and w_{it} is the weight associated with state i and time t, see (14). Chen and Fan (2006) and Okhrin et al. (2013) provide the asymptotic behavior of the estimates.

2.3 Likelihood estimation

For the estimation of the HMM HAC model, we adopt the EM algorithm, see Dempster, Laird and Rubin (1977), also known as the Baum–Welch algorithm in the context of HMM. Recall the full likelihood $p_T(y_{0:T}; x_{0:T})$ in (6) and the partial likelihood $p_T(y_{0:T})$ in (7), and the log likelihood:

$$\log\{p_T(y_{0:T})\} = \log\Big\{\sum_{x_0=1}^{M} \cdots \sum_{x_n=1}^{M} \pi_{x_0} \prod_{t=1}^{T} p_{x_{t-1}x_t} f_{x_t}(y_t; \boldsymbol{\theta}^{(x_t)}, s^{(x_t)})\Big\}. \tag{10}$$

The EM algorithm suggests estimating a sequence of parameters $\mathfrak{g}_{(r)} \stackrel{\text{def}}{=} (P_{(r)}, \boldsymbol{\theta}_{(r)}, \mathbf{s}_{(r)})$ (for the rth iteration) by iterative maximization of $\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(r)})$ with

$$\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(r)}) \stackrel{\text{def}}{=} \mathsf{E}_{\mathfrak{g}_{(r)}}\{\log p_T(Y_{0:T}; X_{0:T}) | Y_{0:T} = y_{0:T}\}.$$

Namely, one carries out the following two steps:

- (a) E-step: compute $\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(r)})$,
- (b) M-step: choose the update parameters $\mathfrak{g}_{(r+1)} = \arg\max_{\mathfrak{g}} \mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(r)})$.

The essence of the EM algorithm is that $\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(r)})$ can be used as a surrogate for $\log p_T(y_{0:T}; x_{0:T}; \theta)$, see Cappé, Moulines and Rydén (2005).

In our setting, we may write $\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(r)})$ as:

$$\begin{aligned}
\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(r)}) &= \sum_{i=1}^{M} \mathsf{E}_{\mathfrak{g}_{(r)}}[\mathbf{1}\{X_0 = i\} \log\{\pi_i f_i(y_0)\} | Y_{0:T} = y_{0:T}] \qquad (11) \\
&+ \sum_{t=1}^{T} \sum_{i=1}^{M} \mathsf{E}_{\mathfrak{g}_{(r)}}[\mathbf{1}\{X_t = i\} \log f_i(y_t) | Y_{0:T} = y_{0:T}] \\
&\quad + \sum_{t=1}^{T} \sum_{i=1}^{M} \sum_{j=1}^{M} \mathsf{E}_{\mathfrak{g}_{(r)}}[\mathbf{1}\{X_t = j\}\mathbf{1}\{X_{t-1} = i\} \log\{p_{ij}\} | Y_{0:T} = y_{0:T}] \\
&= \sum_{i=1}^{M} \mathrm{P}_{\mathfrak{g}_{(r)}}(X_0 = i | Y_{0:T} = y_{0:T}) \log\{\pi_i f_i(y_0)\} \\
&+ \sum_{t=1}^{T} \sum_{i=1}^{M} \mathrm{P}_{\mathfrak{g}_{(r)}}(X_t = i | Y_{0:T} = y_{0:T}) \log f_i(y_t) \\
&\quad + \sum_{t=1}^{T} \sum_{i=1}^{M} \sum_{j=1}^{M} \mathrm{P}_{\mathfrak{g}_{(r)}}(X_{t-1} = i, X_t = j | Y_{0:T} = y_{0:T}) \log\{p_{ij}\}, \qquad (12)
\end{aligned}$$

where $f_i(\cdot)$ is as in (8). The E-step, in which $\mathrm{P}_{\mathfrak{g}_{(r)}}(X_t = i | Y_{0:T}), \mathrm{P}_{\mathfrak{g}_{(r)}}(X_{t-1} = i, X_t = j | Y_{0:T})$ are evaluated, is carried out by the forward-backward algorithm and the M-step is explicit in the p_{ij} and the π_i. Adding constraints to (12) yields

$$\mathfrak{L}(\mathfrak{g}, \lambda; \mathfrak{g}') = \mathcal{Q}(\mathfrak{g}; \mathfrak{g}') + \sum_{i=1}^{M} \lambda_i (1 - \sum_{j=1}^{M} p_{ij}). \qquad (13)$$

For the M-step, we need to take the first order partial derivative, and plug into (13). So the dependency parameters $\boldsymbol{\theta}$ and the structure parameters $\mathbf{s}$ need to be estimated

iteratively, for $\boldsymbol{\theta}^{(i)}$ ($\boldsymbol{\theta}^{(i)} = \{\theta_{i1}, \ldots, \theta_{i(d-1)}\}$):

$$\frac{\partial \mathfrak{L}(\mathfrak{g}, \lambda; \mathfrak{g}')}{\partial \theta_{ij}} = \sum_{t=1}^{T} \mathrm{P}_{\mathfrak{g}'}(X_t = i | Y_{0:T}) \partial \log f_i(y_t) / \partial \theta_{ij}. \tag{14}$$

To simplify the procedure, we adopt the HAC estimation method (9) with weights in terms of $w_{it} \stackrel{\text{def}}{=} \mathrm{P}_{\mathfrak{g}'}(X_t = i | Y_{0:T})$. We also fix $\pi_i, i = 1, \ldots, M$ as it influences only the first observation x_0 which may be considered also as given and fixed. Maximizing (12) w.r.t. π_i would return the first derivative with one observation y_0. Also as the previous information for the distribution of x_0 is not available, our EM algorithm would not involve updating π_i. The estimation of the transition probabilities p_{ij} follows:

$$\frac{\partial \mathfrak{L}(\mathfrak{g}, \lambda; \mathfrak{g}')}{\partial p_{ij}} = \sum_{t=1}^{T} \frac{\mathrm{P}_{\mathfrak{g}'}(X_{t-1} = i, X_t = j | Y_{0:T})}{p_{ij}} - \lambda_i, \tag{15}$$

$$\frac{\partial \mathfrak{L}(\mathfrak{g}, \lambda; \mathfrak{g}')}{\partial \lambda_i} = 1 - \sum_{j=1}^{M} p_{ij}. \tag{16}$$

Equating (15) and (16) yields:

$$\hat{p}_{ij} = \frac{\sum_{t=1}^{T} \mathrm{P}_{\mathfrak{g}'}(X_{t-1} = i, X_t = j | Y_{0:T})}{\sum_{t=1}^{T} \sum_{l=1}^{M} \mathrm{P}_{\mathfrak{g}'}(X_{t-1} = i, X_t = l | Y_{0:T})}. \tag{17}$$

3 Theoretical Results

In this section, we discuss the conditions needed for deriving consistency and asymptotic properties of copulae function. We then state our main theoretical theorems. Throughout the theory we concentrate on the most interesting case: semi-parametric estimation with nonparametric margins.

Assumptions

A.1 $\{X_t\}$ is stationary, strictly irreducible, and aperiodic Markov process of order one with final discrete state, and $\{Y_t\}_{t=1}^{T}$ are conditionally independent given $\{X_t\}_{t=1}^{T}$ and generated from an HAC HMM model with parameters $\{s^{*(i)}, \theta^{*(i)}, \pi^*, \{p^*_{ij}\}_{i,j}\}, i, j = 1, \cdots, d$.

It is worth noting that A.1 is the basic assumption on the evolution of a hidden Markov chain. One key property is that given one realization of the path $\{X_t\}$, the conditional distributions of $\{Y_t\}_{t=1}^T$ are totally fixed. But $\{Y_t\}$ will be dependent and even being observed have a finite mixture distribution from the given parametric family. The envolvement of $\{X_t\}$ will later be linked to the dependency parameters of the state space distribution of $\{Y_t\}$.

A.2 The family of mixtures of at most M elements $\{f(y;\boldsymbol{\theta}^{(i)}, s^{(i)}) : \boldsymbol{\theta}^{(i)} \in \Theta, s^{(i)} \in S\}$ is identifiable w.r.t. the parameters and structures:

$$\sum_{i=1}^{M} \alpha_i f(y;\boldsymbol{\theta}^{(i)}, s^{(i)}) = \sum_{i=1}^{M} \alpha'_i f(y;\boldsymbol{\theta}'^{(i)}, s'^{(i)}) \quad a.e. \tag{18}$$

$$then, \sum_{i=1}^{M} \alpha_j \delta_{\boldsymbol{\theta}^{(i)}, s^{(i)}} = \sum_{i=1}^{M} \alpha'_i \delta_{\boldsymbol{\theta}'^{(i)}, s'^{(i)}}, \tag{19}$$

defining $\delta_{\boldsymbol{\theta}^{(i)}, s^{(i)}}$ as the distribution function for a point mass in Θ associated with the structure $s^{(i)}$, noting that $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}'^{(i)}$ is only meaningful when $s^{(i)} = s'^{(i)}$.

The property of identifiability is nothing else than the construction of a finite mixture model, McLanchlan and Peel (2000). As a copula is a special form of a multivariate distribution, similar techniques may be applied to get identifiability also in the case of copulae. The family of copula mixtures has been thoroughly investigated in Caia, Chen, Fan and Wang (2006) while developing estimation techniques. In that general case, one should be careful, as the general copula class is very wide and its mixture identification may cause some problems because of the different forms of the densities. The very construction of the HAC narrows this class. Imposing the same generator functions on all levels of the HAC, we restrict the family to the vector of parameters and the tree structure, see also Okhrin et al. (2013). Moreover, we restrict the classes to only binary trees with distinct parameters to avoid identifiability issues induced by the case of the same parameter values on each layers of a tree. Our preliminary numerical analysis shows that the HAC fulfills the identifiability property for all the structures and parameters used in this study.

A.3 The true marginal distribution $f_m^{\mathfrak{m}}(\cdot) \in \mathbf{C}^2$, and the derivatives up to second order are

bounded for all $m = 1, \cdots, d$. Also $\sqrt{f^{\mathfrak{m}}}$ are absolute continuous. In case of nonparametric estimation for $f_i^{\mathfrak{m}}(\cdot) \in \mathbf{C}^2$, one needs to also ensure the kernel function $K(\cdot) \in C^2$ has support on a compact set B, is symmetric and has integrable first derivative. $\int_B K(u)du = 1$.

We would like to focus on the dependency parameter, therefore in the following setting, we simply assume that the marginal processes $y_{t1}, y_{t2}, \cdots, y_{td}$ are identically distributed respectively.

A.4 $\mathsf{E}\{|\log f_i(y)|\} < \infty$, for $i = 1, \ldots, M$, $\forall s^{(i)} \in S$. Define the copulae density function $c(u_1, u_2, \ldots, u_d, \boldsymbol{\theta}^{(i)}, s^{(i)}) \stackrel{\text{def}}{=} \partial^d C(u_1, u_2, \cdots, u_d, , \boldsymbol{\theta}^{(i)}, s^{(i)})/\partial u_1 \partial u_2 \cdots \partial u_d$, then $\log c(u_1, u_2, \cdots, u_d, \boldsymbol{\theta}^{(i)}, s^{(i)})$ as well as its first and second partial derivatives w.r.t. v_is and $\boldsymbol{\theta}$ are well defined for $((0,1)^d \times \Theta)$. Also, their supreme in a compact set $((E^d) \times \Theta)$ $(E^d \in [0,1]^d)$ has finite moments up to the order four.

A.5 For every $\boldsymbol{\theta} \in \Theta$, and any particular structure considered $s \in S$,

$$\mathsf{E}[\sup_{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| < \delta} \{f_i(Y_1, \boldsymbol{\theta}', s)\}^+] < \infty,$$

for some $\delta > 0$.

A.6 The true point $\boldsymbol{\theta}^*$ is an interior point of Θ

A.7 Exist a constant δ^0, such that $\mathrm{P}(\sup_{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| < \delta^0} \max_{i,j} \mathsf{E} \frac{\{f_i(Y_1, \boldsymbol{\theta}', s)\}}{\{f_j(Y_1, \boldsymbol{\theta}', s)\}} = \infty | X_1 = i) < 1$

Denote as $p_T(y_{0:T}; v, \omega)$ the density in (7) with parameters $\{v, \omega\} \in \{V, \Omega\}$ as described in the Appendix 7.2. Define $\hat{\boldsymbol{\theta}}^{(i)}, \hat{s}^{(i)}$ as $\hat{\boldsymbol{\theta}}^{(i)}(\hat{v}, \hat{\omega})$ and $\hat{s}^{(i)}(\hat{v}, \hat{\omega})$ with $(\hat{v}, \hat{\omega})$ being the point where $p_T(y_{0:T}; v, \omega)$ achieves its maximum value over the parameter space $\{V, \Omega\}$.

It is known that HMM is not itself identifiable as the permutation of states would yield the same value for $p_T(y_{0:T}; v, \omega)$. We assume therefore $\boldsymbol{\theta}^{*(j)}$s and $s^{*(j)}$s to be distinct in the sense that for any $s^{*(i)} = s^{*(j)}, i \neq j$ we have $\boldsymbol{\theta}^{*(i)} \neq \boldsymbol{\theta}^{*(j)}$.

Theorem 3.1. *Under A.1–A.6, we find the corresponding structure:*

$$\lim_{T \to \infty} \min_{i \in 1, \ldots, M} \mathrm{P}(\hat{s}^{(i)} = s^{*(i)}) = 1. \tag{20}$$

Moreover,

Theorem 3.2. *Assume A.1–A.6. The parameter* $\hat{\boldsymbol{\theta}}^{(i)}$ *satisfies,* $\forall \varepsilon > 0$*:*

$$\lim_{T\to\infty} \max_{i\in 1,\ldots,M} \mathrm{P}(|\hat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{*(i)}| > \varepsilon | \hat{s}^{(i)} = s^{*(i)}) = 0. \tag{21}$$

In addition, we can also establish asymptotic normality results for parameters.

Theorem 3.3. *Assume A.1–A.6, given* $s^{*(i)}$ *is correctly estimated, which an event with probability tending to* 1*, we have*

$$\sqrt{n}\{\hat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{*(i)}\} \to \mathsf{N}(0, \Sigma^*), \tag{22}$$

where Σ^* *is the asymptotic covariance function, defined as* $\Sigma^* \stackrel{\text{def}}{=} B^{-1}\mathsf{Var}(\sqrt{n}A)B^{-1}$*, where* B, A *are defined in the Appendix in (45).*

Proofs are presented in the Appendix.

4 Simulation

The estimation performance of HMM HAC is evaluated in this section: subsection I aims to investigate whether the performance of estimation is affected by 1. adopting a nonparametric or parametric margins; 2. by introducing GARCH dependency in the marginal time series. Subsection II presents results for a five dimensional time series model. In subsection III we perform a horse race between the DCC method and our HMM HAC method. All the simulations show that our algorithm converges after a few iterations with moderate estimation errors, and the results are robust with respect to different estimation methods for margins. Moreover our method dominates the DCC one. Regarding the selection of orders, in both simulations and applications, we have started with a model with three states, which is suggested by the initial moving window analysis described later. In applications the number of states will even be degenerated to two or

one for some windows. This suggests us three states are sufficient for model estimations. However, one can consider general BIC or AIC criteria for selecting the number of states.

4.1 Simulation I

In this subsection, a three dimensional generating process has fixed marginal distributions: $Y_{t1}, Y_{t2}, Y_{t3} \sim \mathsf{N}(0,1)$. For studying the effect of deGARCH step in our application, we simulated also according to GARCH(1,1) model,

$$Y'_{tj} = \mu_{tj} + \sigma_{tj}\varepsilon_{tj} \text{ with } \sigma^2_{tj} = \omega_j + \alpha_j \sigma^2_{t-1j} + \beta_j (Y_{t-1j} - \mu_{t-1j})^2, \tag{23}$$

with parameters $\omega_j = 10^{-6}, \alpha_j = 0.1, \beta_j = 0.8$, which standard normal residuals $\varepsilon_{t1}, \varepsilon_{t2}, \varepsilon_{t3} \sim \mathsf{N}(0,1)$. The dependence structure is modeled through HAC with Gumbel generators. Let us consider now a Monte Carlo setup where the setting employs realistic models. The three states with $M = 3$ are taken as follows:

$$\begin{aligned}
&C\{u_1, C(u_2, u_3; \theta_1^{(1)} = 1.3); \theta_2^{(1)} = 1.05\} \text{ for } i = 1,\\
&C\{u_2, C(u_3, u_1; \theta_1^{(2)} = 2.0); \theta_2^{(2)} = 1.35\} \text{ for } i = 2,\\
&C\{u_3, C(u_1, u_2; \theta_1^{(3)} = 4.5); \theta_2^{(3)} = 2.85\} \text{ for } i = 3,
\end{aligned}$$

where the dependency parameters corresponds to the Kendall's τ ranging between 0.05 and 0.78 which is typical for financial data. The transition matrix is chosen as:

$$P = \begin{pmatrix} 0.982 & 0.010 & 0.008 \\ 0.008 & 0.984 & 0.008 \\ 0.003 & 0.002 & 0.995 \end{pmatrix},$$

and initial probabilities as $\pi = (0.2, 0.1, 0.7)$ and sample size $T = 2000$. Figure 3 represents the underlying states and a marginal plot of the generated three dimensional time series. No state switching patterns are evident from the marginal plots. Figure 4, however, clearly displays the switching of dependency patterns. The circles, triangles, and

crosses correspond to the observations from states $i = 1, 2, 3$ respectively.

Generally, the iteration procedure stops after around ten steps. Figure 5 presents the deviations of the estimated states, the transition matrix, and the parameters from their true values for first ten iterations of one sample. Since the starting values may influence the results, a moving window estimation is proposed to decide the initial parameters. The dashed black and solid black line show, respectively, how the estimators behave with the initial values close to the true (dashed) and initial values obtained from the proposed rolling window algorithm (solid). For the "close to the true initial states", we mean true structures with parameters all shifted up by 0.5 from the true ones. For "rolling window algorithm" we estimate HAC for overlapping windows of width 100, and then took M most frequent structures with averaged parameters as initial states. The left panel of Figure 5 shows the (L_1) difference ($\sum_{i,j=1}^{d} |\hat{p}_{ij} - p_{ij}|$) of the true transition matrix from the estimated ones at each iteration, we see that for the three particular samples, the values all converge to around 0.4, which are moderately small; the middle panel is the sum of the estimated parameter errors of the four states with the correctly estimated states, we see that the accumulated errors are different w.r.t different starting values; the right panel represents the percentage of wrongly estimated states, in all cases the percentage of wrongly estimated states are smaller than 8. One can see that our choice of initial values can perform as well as the true one through showing small differences, and our results from more iterations further confirms this.

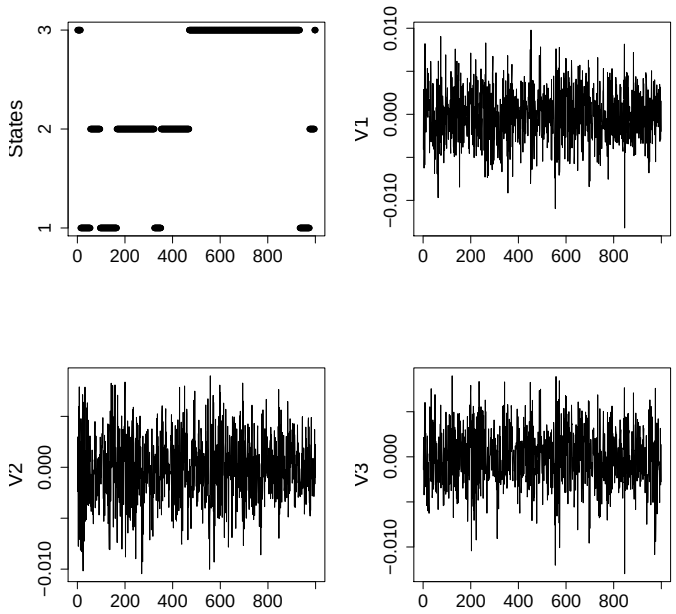


Figure 3: The underlying sequence x_t (upper left panel), marginal plots of $(y_{t1}, y_{t2}, y_{t3})(t = 0, \ldots, 1000)$.

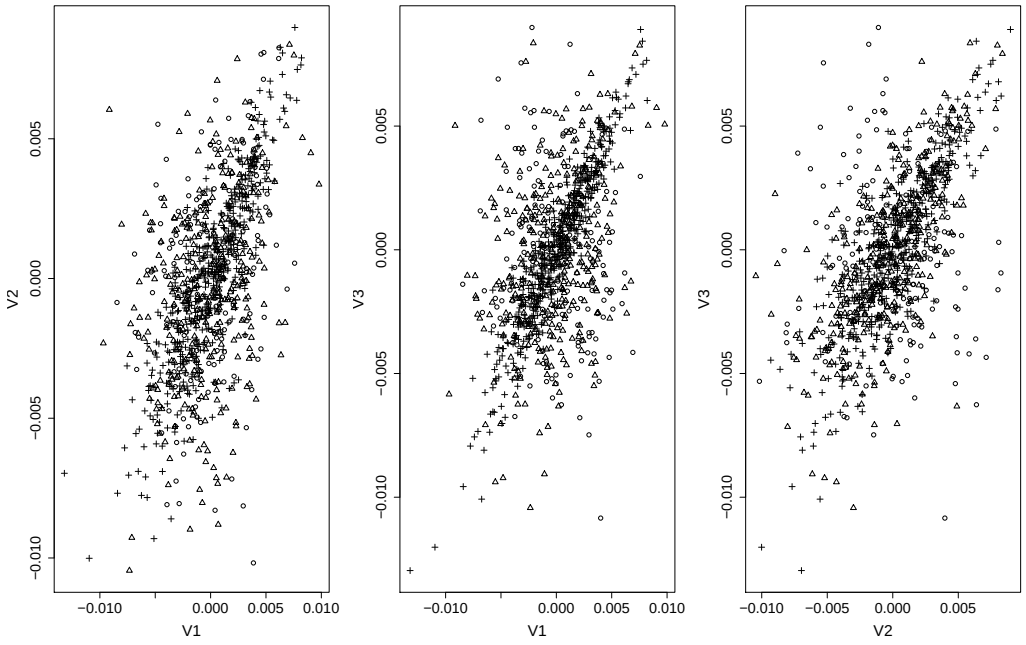
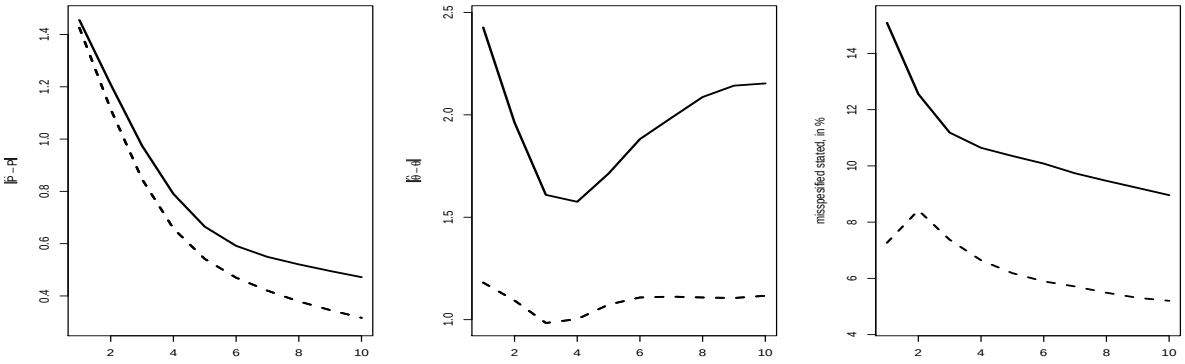
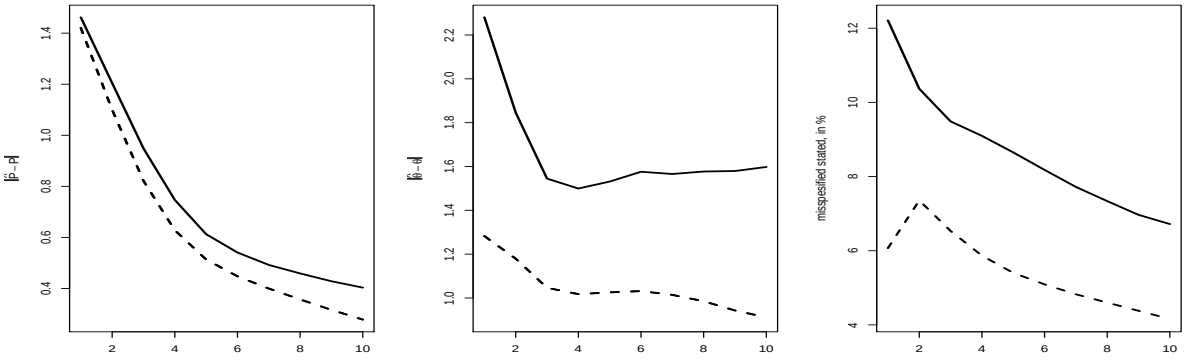


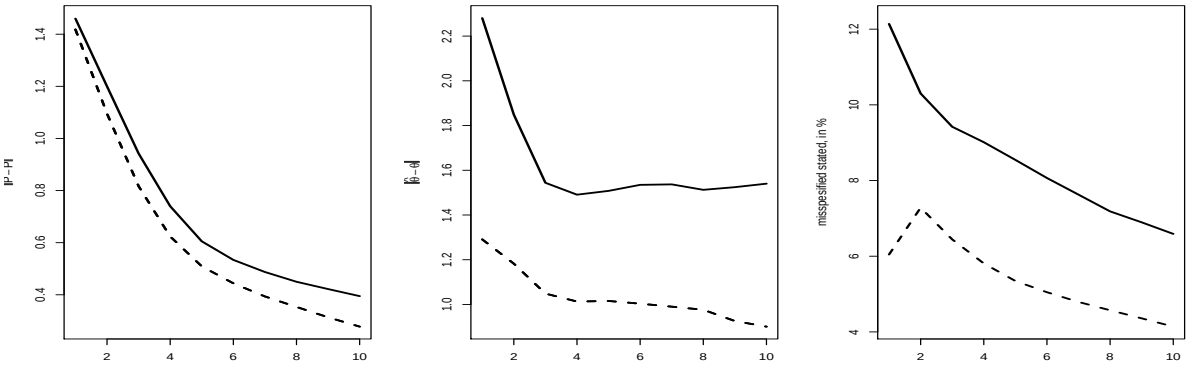
Figure 4: Snapshots of pairwise scatter plots of dependency structures $(t = 0, \ldots, 1000)$, the (y_{t1}) vs. (y_{t2}) (left), the (y_{t1}) vs. (y_{t3}) (middle), and the (y_{t2}) vs. (y_{t3})(right).



(a) With GARCH effects in the margins, nonparametrically estimated margins



(b) Without GARCH effects in the margins, nonparametrically estimated margins



(c) Without GARCH effects in the margins, parametric margins

Figure 5: The transition matrix (left panel), parameters (middle panel), convergence of states (right panel). Estimation starts from near true value (dashed); starts from values obtained by rolling window (solid).

				True DGP is ID		True DGP is GARCH(1.1)	
			True	Rol. Win.	True Str.	Rol. Win.	True Str.
Nonparametric Margins	C_1	$\theta_1^{(1)}$	1.05	1.030 (0.046, 0.003)	1.057 (0.068, 0.005)	1.04 (0.068, 0.005)	1.081 (0.073, 0.006)
		$\theta_2^{(1)}$	1.30	1.313 (0.156, 0.025)	1.308 (0.083, 0.007)	1.31 (0.292, 0.085)	1.294 (0.073, 0.005)
	C_2	$\theta_1^{(2)}$	1.35	1.366 (0.121, 0.015)	1.346 (0.182, 0.033)	1.38 (0.179, 0.033)	1.351 (0.199, 0.040)
		$\theta_2^{(2)}$	2.00	2.556 (1.052, 1.416)	3.212 (1.991, 5.433)	2.64 (1.174, 1.783)	3.662 (2.882, 11.068)
	C_3	$\theta_1^{(3)}$	2.85	2.854 (0.073, 0.005)	2.854 (0.073, 0.005)	2.8 (0.088, 0.010)	2.799 (0.081, 0.009)
		$\theta_2^{(3)}$	4.50	4.497 (0.133, 0.018)	4.496 (0.130, 0.017)	4.38 (0.150, 0.038)	4.374 (0.150, 0.038)
	rat. of correct states			0.958 (0.029)	0.933 (0.056)	0.948 (0.035)	0.910 (0.064)
	$\sum_{i,j=1}^{d} \lvert\hat{p}_{ij} - p_{ij}\rvert$			0.278 (0.230)	0.404 (0.307)	0.316 (0.237)	0.472 (0.323)
	rat. of correct structures			0.949	0.918	0.966	0.912
Parametric Margins	C_1	$\theta_1^{(1)}$	1.05	1.030 (0.041, 0.002)	1.056 (0.066, 0.004)	1.046 (0.077, 0.006)	1.084 (0.071, 0.006)
		$\theta_2^{(1)}$	1.30	1.310 (0.154, 0.024)	1.306 (0.087, 0.008)	1.302 (0.247, 0.061)	1.288 (0.074, 0.006)
	C_2	$\theta_1^{(2)}$	1.35	1.365 (0.130, 0.017)	1.344 (0.173, 0.030)	1.382 (0.174, 0.031)	1.364 (0.208, 0.043)
		$\theta_2^{(2)}$	2.00	2.544 (0.962, 1.221)	3.157 (1.906, 4.971)	2.567 (1.109, 1.551)	3.480 (2.270, 7.343)
	C_3	$\theta_1^{(3)}$	2.85	2.855 (0.074, 0.006)	2.854 (0.074, 0.005)	2.807 (0.092, 0.010)	2.807 (0.086, 0.009)
		$\theta_2^{(3)}$	4.50	4.513 (0.133, 0.018)	4.513 (0.132, 0.018)	4.407 (0.161, 0.035)	4.406 (0.160, 0.034)
	rat. of correct states			0.959 (0.029)	0.934 (0.056)	0.946 (0.034)	0.911 (0.061)
	$\sum_{i,j=1}^{d} \lvert\hat{p}_{ij} - p_{ij}\rvert$			0.278 (0.232)	0.395 (0.297)	0.294 (0.220)	0.442 (0.296)
	rat. of correct structures			0.955	0.921	0.954	0.908
deGARCHing	C_1	$\theta_1^{(1)}$	1.05	1.030 (0.045, 0.002)	1.056 (0.067, 0.005)	1.031 (0.042, 0.002)	1.064 (0.070, 0.005)
		$\theta_2^{(1)}$	1.30	1.320 (0.264, 0.070)	1.307 (0.081, 0.007)	1.315 (0.408, 0.167)	1.302 (0.075, 0.006)
	C_2	$\theta_1^{(2)}$	1.35	1.367 (0.123, 0.015)	1.345 (0.166, 0.028)	1.372 (0.136, 0.019)	1.349 (0.213, 0.045)
		$\theta_2^{(2)}$	2.00	2.577 (1.273, 1.953)	3.180 (1.976, 5.297)	2.619 (1.344, 2.189)	3.485 (2.691, 9.447)
	C_3	$\theta_1^{(3)}$	2.85	2.852 (0.074, 0.005)	2.852 (0.074, 0.005)	2.848 (0.074, 0.005)	2.847 (0.074, 0.005)
		$\theta_2^{(3)}$	4.50	4.489 (0.133, 0.018)	4.488 (0.130, 0.017)	4.469 (0.132, 0.018)	4.468 (0.133, 0.019)
	rat. of correct states			0.958 (0.029)	0.933 (0.056)	0.956 (0.029)	0.927 (0.059)
	$\sum_{i,j=1}^{d} \lvert\hat{p}_{ij} - p_{ij}\rvert$			0.280 (0.234)	0.399 (0.299)	0.297 (0.245)	0.436 (0.332)
	rat. of correct structures			0.950	0.919	0.956	0.917

Table 1: Simulation results for different DGPs, sample size $T = 2000$, 1000 repetitions, standard deviations and MSEs are provided in brackets.

Finally, we summarize our estimation results over 1000 repetitions. In Table 1, we report the averaged estimation values with standard deviations (in bracket) and MSE (in brackets) for the estimated states, the transition matrix, and the parameters. For the impact of estimating the copula model on estimated standardized residuals (after GARCH fitting, for example), we have also included the comparison of the estimation on the de-GARCHed residuals. Also the estimation for different ways of deciding starting values are shown: "close to the true initial states" (True str), rolling window algorithm (Rol. Win.). Apparently, nonparametric or parametric estimation of the margins does not make big differences, this is also true for the pre-whitening step. Regarding the preciseness of the estimation, we see for all parameters in different states have small estimation errors except for the parameter $\theta_2^{(2)}$. The standard deviations of design matrix are also relatively high. This is due to our selected design matrix having very small off-diagonal values, so for some realizations we have too few observations for state 2 to achieve accurate estimated. One sees in our simulation II nicer results with a different transition matrix.

4.2 Simulation II

In this subsection, we consider a five dimensional model. The marginal distributions are taken as: $Y_{t1}, Y_{t2}, Y_{t3}, Y_{t4}, Y_{t5} \sim \mathsf{N}(0,1)$. The dependence structure is modeled through HAC with Gumbel generators as well. We set also three states ($M = 3$) :

$$C\{C(u_1, C(u_2, C(u_3, C(u_5, u_4; \theta_1 = 3.15); \theta_2 = 2.45); \theta_3 = 1.75); \theta_4 = 1.05)\} \text{ for } i = 1,$$
$$C\{C(u_3, C(u_5, C(u_2, C(u_1, u_4; \theta_1 = 3.15); \theta_2 = 2.45); \theta_3 = 1.75); \theta_4 = 1.05)\} \text{ for } i = 2,$$
$$C\{C(u_5, C(u_4, C(u_3, C(u_1, u_2; \theta_1 = 3.15); \theta_2 = 2.45); \theta_3 = 1.75); \theta_4 = 1.05)\} \text{ for } i = 3,$$

the transition matrix is chosen as:

$$P = \begin{pmatrix} 0.82 & 0.10 & 0.08 \\ 0.08 & 0.84 & 0.08 \\ 0.03 & 0.02 & 0.95 \end{pmatrix},$$

and the initial probabilities are $\pi = (0.2, 0.1, 0.7)$, $T = 2000$. Figure 6 shows again the pairwise scatterplots of observations generated from the above mentioned model. Similarly, Table 2 and Figure 7 presents the estimation accuracy for this model. Although the computation is more demanding when the dimension gets higher, we still can achieve the same degree of accuracy as it is in the three dimensional case.

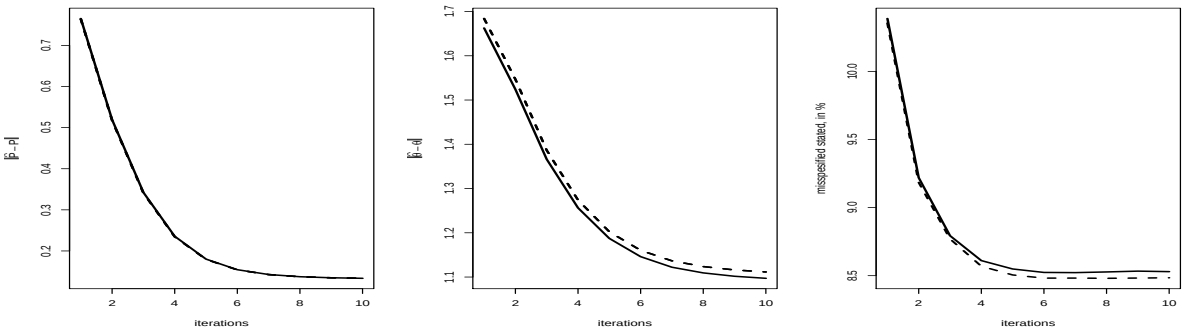


Figure 6: The transition matrix (left panel), parameters (middle panel), convergence of states (right panel). Estimation starts from near true value (dashed); starts from values obtained by rolling window(solid).

		True	Param. Margins	deGARCHing
C_1	$\theta_1^{(1)}$	1.05	1.019 (0.020, 0.001)	1.019 (0.020, 0.001)
	$\theta_2^{(1)}$	1.75	1.739 (0.077, 0.006)	1.741 (0.078, 0.006)
	$\theta_3^{(1)}$	2.45	2.584 (0.126, 0.034)	2.583 (0.126, 0.034)
	$\theta_4^{(1)}$	3.15	3.328 (0.194, 0.069)	3.318 (0.194, 0.066)
C_2	$\theta_1^{(2)}$	1.05	1.017 (0.021, 0.002)	1.017 (0.021, 0.002)
	$\theta_2^{(2)}$	1.75	1.795 (0.084, 0.009)	1.797 (0.084, 0.009)
	$\theta_3^{(2)}$	2.45	2.499 (0.120, 0.017)	2.499 (0.122, 0.017)
	$\theta_4^{(2)}$	3.15	3.381 (0.216, 0.100)	3.369 (0.215, 0.094)
C_3	$\theta_1^{(3)}$	1.05	1.044 (0.017, 0.000)	1.045 (0.018, 0.000)
	$\theta_2^{(3)}$	1.75	1.745 (0.041, 0.002)	1.747 (0.041, 0.002)
	$\theta_3^{(3)}$	2.45	2.492 (0.065, 0.006)	2.492 (0.065, 0.006)
	$\theta_4^{(3)}$	3.15	3.189 (0.094, 0.010)	3.185 (0.095, 0.010)
rat. of correct states			0.915 (0.011)	0.915 (0.011)
$\sum_{i,j=1}^{d} \lvert\hat{p}_{ij} - p_{ij}\rvert$			0.133 (0.054)	0.133 (0.054)
rat. of correct structures			1	1

Table 2: The summary of estimation accuracy in five dimensional model.

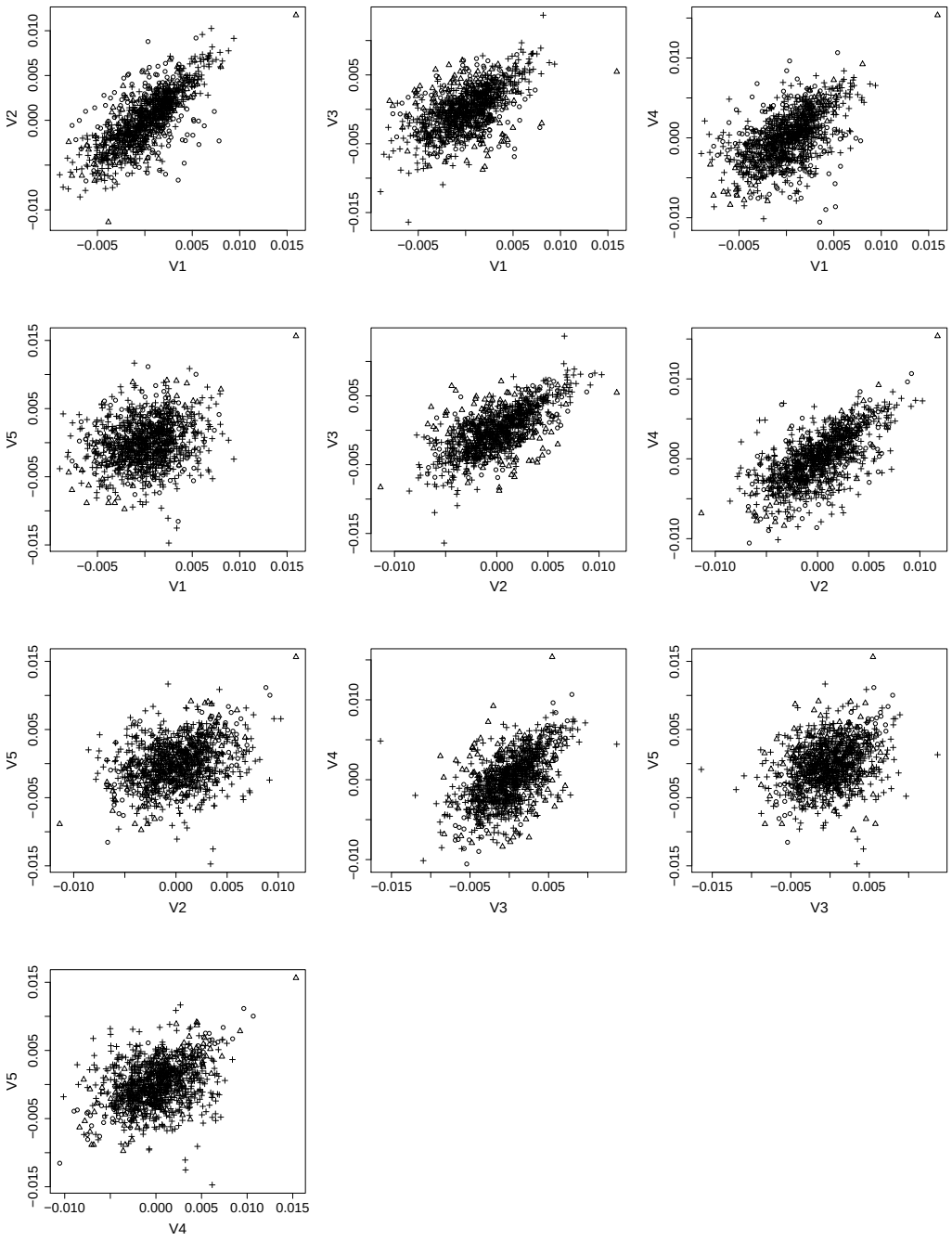


Figure 7: Snapshots of pairwise scatter plots of dependency structures ($t = 0, \ldots, 1000$).

True \ Estimated	Sample size	HMMGARCH	HMM ID	DCC
HMM GARCH		**0.0911 (0.0362)**	0.0912 (0.0382)	0.1958 (0.0929)
DCC	250	**0.0607 (0.0241)**	0.0723 (0.0320)	0.0782 (0.0309)
HMM ID		0.0908 (0.0359)	**0.0867 (0.0345)**	0.1424 (0.0271)
HMMGARCH		**0.0935 (0.0334)**	0.0968 (0.0386)	0.1724 (0.0455)
DCC	500	**0.0541 (0.0194)**	0.0672 (0.0325)	0.0774 (0.0254)
HMM ID		**0.0936 (0.0331)**	0.0924 (0.0326)	0.1515 (0.0239)
HMM GARCH		**0.0923 (0.0331)**	0.0934 (0.0374)	0.1663 (0.0319)
DCC	1000	**0.0494 (0.0166)**	0.0659 (0.0320)	0.0823 (0.0392)
HMM ID		**0.0919 (0.0331)**	0.0907 (0.0322)	0.1509 (0.0213)

Table 3: The estimated mean KS test statistics (standard deviation) of the forecast distribution from the true model and the estimated model. Number of repetitions is 1000.

4.3 Simulation III

To compare the forecasting performance difference models, we simulate data from different true models: HMM GARCH, HMM id and DCC, from which we simulated 3 dimensional time series with $T-1$ observations. Then we fit different models (HMM GARCH, HMM id, HAC GARCH, HAC id and DCC) with $T-1$ observations at hand, and compare the distribution forecast one step ahead for the true and the estimated models. More specific, for the distribution forecast comparison we calculated the sum $y_{T1}+y_{T2}+y_{T3}$ (may be thought as the return of an equally weighted portfolio).

Simulation of 1000 observations $y_{T1}+y_{T2}+y_{T3}$ allows us to compare the forecast distribution between the true model and estimated models. Further, we calculate Kolmogorov-Smirnov (KS) test statistics to measure the difference between the forecast distribution of observations from the true and the estimated model. The comparison has been done by $T = 200, 500, 1000$ Table 3 reports the mean and the standard deviation of the KS test statistics for different model w.r.t. to the true one. We see obvious advantages of our method over the DCC model in the sense that our HMM GARCH model are in all cases closer to the forecast distribution of the true on average than the DCC model. Especially when the data generating processes are HMM GARCH or HMM ID .

5 Applications

To see how HMM HAC performs on a real data set, applications to financial and rainfall data are offered. A good model for the dynamics of exchange rates gives insights into exogenous economic conditions, such as the business cycle. It is also helpful for portfolio risk management and decisions on asset allocation. We demonstrate the performance of our proposed technique by applying it to forecasting the VaR of a portfolio and compare it with multivariate GARCH models (DCC, BEKK, etc.) The backtesting results show that the VaR calculated from HMM HAC performs significantly better.
The second application is on modeling a rainfall process. HMM is a conventional model for rainfall data, however, bringing HMM and HAC together for modeling the multivariate rainfall process is an innovative modeling path.

5.1 Application I

5.1.1 Data

The data set consists of the daily values for the exchange rates JPY/EUR, GBP/EUR and USD/EUR. The covered period is [4.1.1999; 14.8.2009], resulting in 2771 observations.

To eliminate intertemporal conditional heteroscedasticity, we fit to each marginal time series of log-returns a univariate GARCH(1,1) process

$$Y_{j,t} = \mu_{j,t} + \sigma_{j,t}\varepsilon_{j,t} \text{ with } \sigma^2_{j,t} = \omega_j + \alpha_j \sigma^2_{j,t-1} + \beta_j (Y_{j,t-1} - \mu_{j,t-1})^2 \tag{24}$$

and $\omega > 0$, $\alpha_j \geq 0$, $\beta_j \geq 0$, $\alpha_j + \beta_j < 1$.

The residuals exhibit the typical behavior: they are not normally distributed, which motivates nonparametric estimation of the margins. From the results of the Box–Ljung test, whose p-values are $0.73, 0.01$, and 0.87 for JPY/EUR, GBP/EUR and USD/EUR, we conclude that the autocorrelation of the residuals is strongly significant only for the GBP/EUR rate. After this intertemporal correction, we work only with the residuals.

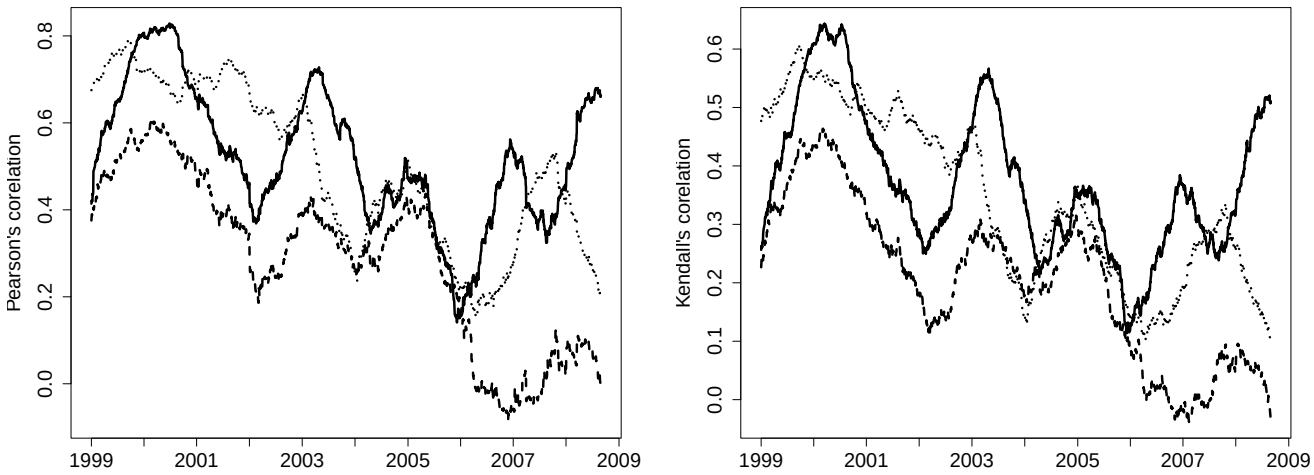


Figure 8: Rolling window estimators of Pearson's (left) and Kendall's (right) correlation coefficients between the GARCH(1,1) residuals of exchange rates: JPY and USD (solid line), JPY and GBP (dashed line), GBP and USD (dotted line). The width of the rolling window is set to 250 observations.

The dependency variation is measured by Kendall's and Pearson's correlation coefficients: Figure 8 shows the variation of both coefficients calculated in a rolling window of width $r = 250$. Their dynamic behavior is similar, but not identical. This motivates once more a time varying copula based model.

5.1.2 Fitting an HMM model

Figures 1, 9, and 10 summarize the analysis using three methods: moving window, LCP, and HMM HAC. LCP uses moving windows, with varying sizes. To be more specific, LCP is a scaling technique which determines a local homogeneous window at each time point Härdle et al. (2013). In contrast to LCP, HMM HAC is based on a global modeling concept rather than a local one. One observes relatively smooth changes of the parameters, see Figures 1 and 9. HMM HAC is as flexible as LCP, as can be seen from Figures 1, 9, and 10, since the structure estimated also takes three values and is confirmed by the variations of structures estimated from LCP. Moreover, the moving window analysis or LCP can serve as a guideline for choosing the initial values for our HMM HAC. Figure 11 displays the number of states for HMM HAC for rolling windows with a length of 500 observations.

A VaR estimation example is to show the good performance of HMM HAC. We generate

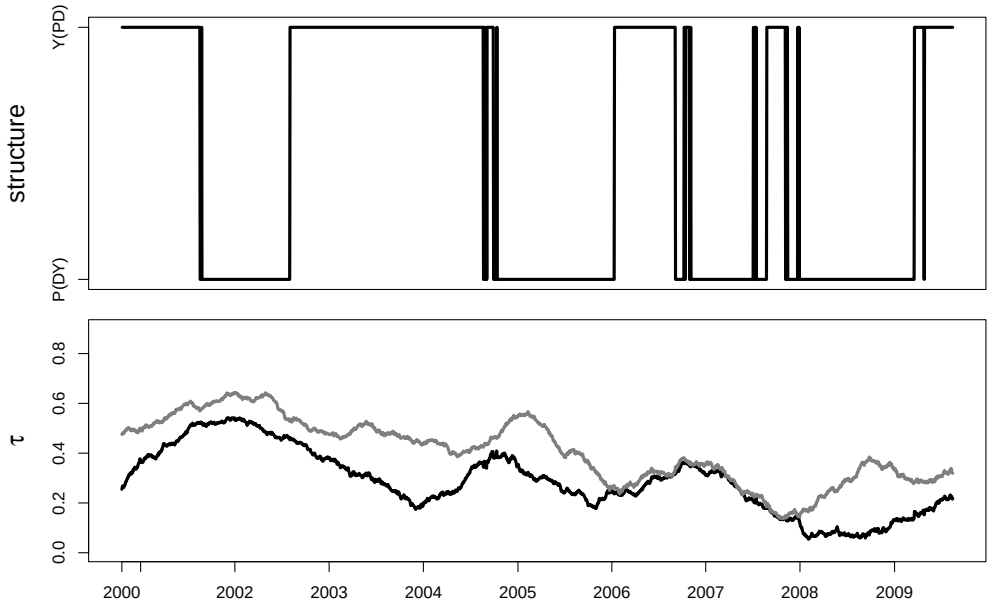


Figure 9: Rolling window for exchange rates: structure (upper) and dependency parameters (lower, θ_1 (grey) and θ_2 (black)) for Gumbel HAC. Rolling window size $win = 250$.

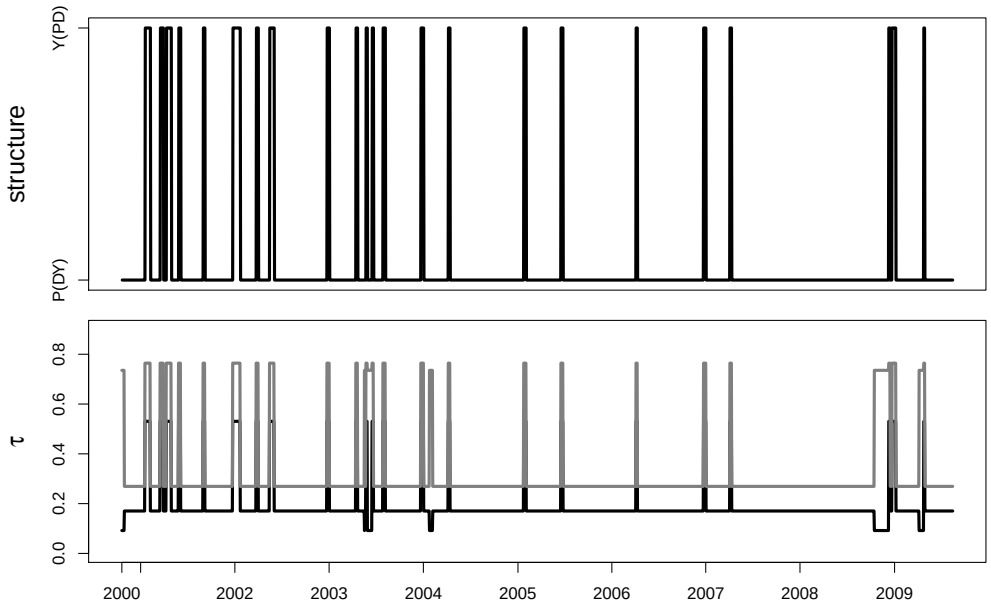


Figure 10: HMM for exchange rates: structure (upper) and dependency parameters (lower, θ_1 (grey) and θ_2 (black)) for Gumbel HAC.

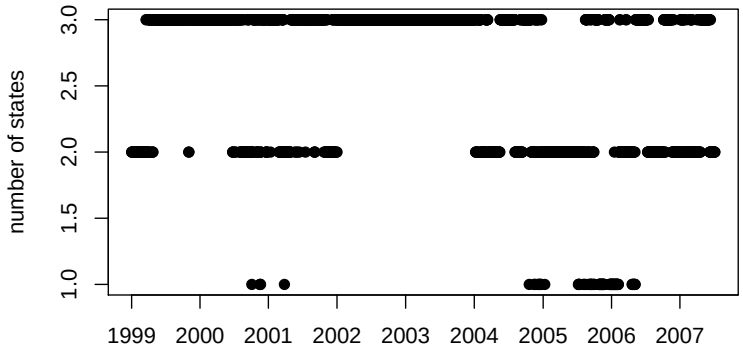


Figure 11: Plot of estimated number of states for each window

$N = 10^4$ paths with $T = 2219$ observations, and $|W| = 1000$ combinations of different portfolios, where $W = \{(1/3, 1/3, 1/3)^\top \bigcup [\mathbf{w} = (w_1, w_2, w_3)^\top]\}$, with $w_i = w_i' / \sum_{i=1}^3 w_i'$, $w_i' \in U(0, 1)$. The Profit Loss (P&L) function of a weighted portfolio based on assets y_{td} is $L_{t+1} \stackrel{\text{def}}{=} \sum_{d=1}^3 w_i(y_{t+1d} - y_{td})$, with weights $\mathbf{w} = (w_1, w_2, w_3) \in W$. The VaR of a particular portfolio at level $0 < \alpha < 1$ is defined as $VaR(\alpha) \stackrel{\text{def}}{=} F_L^{-1}(\alpha)$, where the $\hat{\alpha}_{\mathbf{w}}$ is estimated as a relative fraction of violations, see Table 4:

$$\hat{\alpha}_{\mathbf{w}} \stackrel{\text{def}}{=} T^{-1} \sum_{t=1}^{T} \mathbf{I}\{L_t < \widehat{VaR}_t(\alpha)\},$$

and the distance between $\hat{\alpha}_{\mathbf{w}}$ and α is

$$e_{\mathbf{w}} \stackrel{\text{def}}{=} (\hat{\alpha}_{\mathbf{w}} - \alpha)/\alpha.$$

If the portfolio distribution is i.i.d., and a well calibrated model is properly mimicking the true underlying asset process, $\hat{\alpha}_{\mathbf{w}}$ is close to its nominal level α. The performance is measured through an average of $\alpha_{\mathbf{w}}$ over all $|W|$ portfolios, see Table 4.

We considered four main models: HMM HAC for 500 observation windows for Gumbel and rotated Gumbel; multiple rolling window with 250 observations windows; LCP with $m_0 = 20$ and $m_0 = 40$ with Gumbel copulae (the LCP finds the optimal length of window in the past by a sequence of tests on windows of increasing sizes, m_0 is a starting window

Window\α		0.1	0.05	0.01
HMM, RGum	500	0.0980	**0.0507**	**0.0128**
HMM, Gum	500	**0.0981**	0.0512	0.0135
Rolwin, RGum	250	0.1037	0.0529	0.0151
Rolwin, Gum	250	0.1043	0.0539	0.0162
LCP, $m_0 = 40$	468	0.0973	0.0520	0.0146
LCP, $m_0 = 20$	235	0.1034	0.0537	0.0169
DCC	500	0.0743	0.0393	0.0163

Table 4: VaR backtesting results, $\bar{\hat{\alpha}}$, where "Gum" denotes the Gumbel copula and "RGum" the rotated survival Gumbel one.

Window\α		0.1	0.05	0.01
HMM, RGum	500	-0.0204 (0.013)	**0.0147** (0.012)	**0.2827** (0.064)
HMM, Gum	500	**-0.0191** (0.008)	0.0233 (0.018)	0.3521 (0.029)
Rolwin, RGum	250	0.0375 (0.009)	0.0576 (0.012)	0.5076 (0.074)
Rolwin, Gum	250	0.0426 (0.009)	0.0772 (0.030)	0.6210 (0.043)
LCP, $m_0 = 40$	468	-0.0270 (0.010)	0.0391 (0.018)	0.4553 (0.037)
LCP, $m_0 = 20$	235	0.0344 (0.009)	0.0735 (0.026)	0.6888 (0.050)
DCC	500	-0.2573 (0.015)	-0.2140 (0.015)	0.6346 (0.091)

Table 5: Robustness relative to $A_W(D_W)$

size); and DCC, see Engle (2002), based on 500 observation windows. For all the models we made an out of sample forecast. To better evaluate the performance, we calculated the average and SD of e_W:

$$A_W = \frac{1}{|W|}\sum_{\mathbf{w}\in W} e_{\mathbf{w}}, \qquad D_W = \left\{\frac{1}{|W|}\sum_{\mathbf{w}\in W}(e_{\mathbf{w}} - A_W)^2\right\}^{1/2}.$$

Tables 4 and 5 show the backtesting performance for the described models. One concludes that HMM HAC performs better than the concurring moving window, LCP, or DCC, as A_w and D_w are typically smaller in absolute value.

5.2 Application II

Rainfall models are used to forecast, to simulate and to price weather derivatives. The difficulty in precipitation data is the nonzero point mass at zero and spatial relationships, see Ailliot, Thompson and Thomson (2009) for Gaussian dependency among locations

with HMM application.

In this application we extend it to a copula framework. Different from application I, the marginal distribution here will be varying over states. We propose two methods for modeling the marginal distributions: one is to take y_{tk} to be censored normal distributions, with the following equation:

$$f_k^{\mathfrak{m}}\{y_{tk}\} = \begin{cases} 1 - p_k^{x_t} & y_{tk} = 0, \\ p_k^{x_t}\varphi[\{y_{tk} - \mu^{x_t}(k)\}/\{\sigma^{x_t}(k)\}]/\sigma^{x_t}(k) & y_{tk} > 0; \end{cases}$$

with $k = 1, \ldots, d$ as the location, $\varphi(\cdot)$ as the standard normal density, $p_k^{x_t}$ as the rainfall occurrence probability for the location k and state x_t, and $\mu^{x_t}(k), \sigma^{x_t}(k)$ the mean and standard deviation parameters at time t for location k.

A second proposal for the marginal distributions are the gamma distributions:

$$f_k^{\mathfrak{m}}\{y_{tk}\} = \begin{cases} 1 - p_k^{x_t} & y_{tk} = 0, \\ p_k^{x_t}\gamma\{y_{tk}; \alpha(k)^{x_t}, \beta(k)^{x_t}\} & y_{tk} > 0; \end{cases}$$

where again the $\alpha(k)^{x_t}, \beta(k)^{x_t}$ are the shape and scale parameters for state x_t and location k. We take the joint distribution function to be a truncated version of a continuous copula function, with the copula density $c_d(\cdot)$ denoted by

$$c_d(\mu, \theta) = \begin{cases} c_c(\mu, \theta), & y_{tk} > 0, \forall k, \\ \partial C_c(\mu, \theta)/\partial\mu_{k_1} \ldots \partial\mu_{k_B}, & k_i \in \{y_{tk_i} > 0\}, i \in 1, \ldots, E; \end{cases} \tag{25}$$

where E denotes the number of wet places among the d locations, the C_c are the continuous copula functions, and c_c are the continuous copula densities.

Assume that the daily rainfall observations from the same month are yearly independent realizations of a common underlying hidden Markov model, whose states represents

Figure 12: Map of Guangxi, Guangdong, Fujian in China

different weather types. As an example, we take every June's daily rainfall.

$$\begin{aligned} & \log p_T(y_{0:T}, x_{0:T}; v \times \omega) \\ = & \sum_{i=1}^{M} \mathbf{1}\{x_0 = i\} \log\{\pi_i f_i(y_0)\} + \sum_{t=1}^{T} \sum_{i=1}^{M} \sum_{j=1}^{M} \mathbf{1}\{x_t = j\} \mathbf{1}\{x_{t-1} = i\} \log\{p_{ij} f_j(y_t)\} \\ & + \sum_{t \in B} \sum_{i=1}^{M} \left[\mathbf{1}\{x_t = i\}\{\log(\pi_i)\} - \sum_{j=1}^{M} \mathbf{1}\{x_t = i\} \mathbf{1}\{x_{t-1} = j\} \log(p_{ji}) \right], \end{aligned}$$

with B is the set of days which are the first day of June for each year. We use here 50 years of rainfall data from three locations in China: Guangxi, Guangdong, and Fujian (Figure 12). The graphical correlation can naturally be captured by the fitting of different copulae state parameters.

Table 6 presents (with a truncated Gumbel) the estimated three states, the corresponding different marginal distributions and copula parameters, with estimated initial probability: $\hat{\pi}_{X_t} = (0.298, 0.660, 0.042)$ and estimated transition probability matrix:

$$\hat{P} = \begin{pmatrix} 0.590 & 0.321 & 0.298 \\ 0.188 & 0.742 & 0.660 \\ 0.329 & 0.271 & 0.042 \end{pmatrix}.$$

In our data situation, gamma distributions fit better as marginals. The states filtered out represent different weather types. The third states are the most humid states, with a high

X_t	Shape	Scale	Occur Prob
1	(0.442,0.429,0.552)	(139.33,116.70,169.66)	(0.252,0.256,0.439)
2	(0.671,0.618,0.561)	(273.83,253.25,427.46)	(0.806,0.786,0.683)
3	(0.636,1.125,0.774)	(381.09,264.83,514.08)	(0.667,1.000,0.944)

Table 6: Rainfall occurrence probability and shape, scale parameters estimated from HMM (data 1957–2006) .

Location	True	$\widehat{\text{Corr}}(Y_{t,1}, Y_{t,2})$
$1--2$	0.308	0.300 $(0.235, 0.373)$
$2--3$	0.261	0.411 $(0.256, 0.586)$
$1--3$	0.203	0.130 $(0.058, 0.215)$

Table 7: True correlations, simulated averaged correlations from 1000 samples and their 5% confidence intervals. 1 Fujian, 2 Guangdong, 3 Guangxi.

rainfall occurrence probabilities, while the second states are drier, and the first are the driest. From the parameters of the gamma distributions, one sees the variance increases from the first to the third states, which indicates a higher chance for heavy rainfall for the humid states.

To validate our model, 1000 samples of artificial time series of 1500 observations were generated from the fitted model and compared with the original data. Table 7 presents the true Pearson correlation compared with the estimated ones from the generated time series. The 5% confidence intervals of the estimators cover the true correlation, which implies that the simulated rainfall can describe the real correlation of the data quite well. Figure 13 shows a marginal plot of the log survival function derived from the empirical cdf of the real data and generated data. The log survival function is a transformation of the marginal cdf $F_k^{\mathrm{m}}(y_{tk})$:

$$\log\{1 - F_k^{\mathrm{m}}(y_{tk})\}. \tag{26}$$

Again we see that the 95% confidence interval can cover the true curve fairly well.

Table 7 contains the autocorrelations and cross-correlations of the real data and the generated time series. Unfortunately, our generated time series does not show a similar autocorrelation or cross-correlation. Since there is usually more than one significant lag of autocorrelation or cross-correlation, the simulated time series mostly only have one lag.

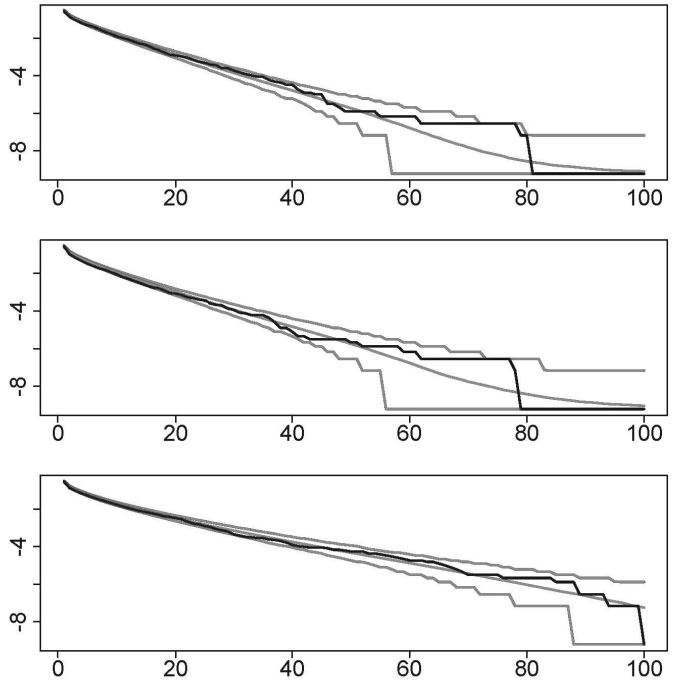


Figure 13: Log-survivor-function (grey) and 95% prediction intervals (grey) of the simulated distribution for the fitted model with sample log-survivor-function superimposed (black)

This is an issue also observed in Ailliot et al. (2009). The precipitation can be modelled first by a vector autoregressive (VAR) type model, adjusted for zero observations. An alternative could be imposing additional dependency structure on $\{Y_t\}$.

6 Conclusion and Discussions

We propose a dynamic model for multivariate time series with non-Gaussian dependency. Applying an HMM for general copulae leads to a rich clan of dynamic dependency structures. The proposed methodology is helpful in studying financial contagion at an extreme level over time, and naturally it can help in deriving conditional risk measures, such as CoVaR, Adrian and Brunnermeier (2011). We have shown that dynamic copula models are well fitting financial markets as well as rainfall patterns.

In the finalcial application, we performed degarching to remove the second order dependencies in the marginal time series. As this is a $\sqrt{n}$ step, it will not contaminate the final

estimation, and our simulation study confirms this. In the rainfall application, we extend our to model to allow the marginal distribution's parameters also varying over states. Typically it will adapt to nonstationary marginal time series with trend.

7 Appendix

7.1 Proof of Theorems 3.1 and 3.2

In the HMM HAC framework, let $\{X_t, t \geq 0\}$ with transition probability matrix $P^{v,\omega} = [p_{ij}^{v,\omega}]_{i,j=1,\ldots,M}$ and initial distribution $\pi^{v,\omega} = \{\pi_i^{v,\omega}\}_{i=1,\ldots,M}$, where $\{v, \omega\} \in \{V, \Omega\}$ denotes an element in the parameter space $\{V, \Omega\}$ which parametrizes this model, and q is the number of continuous parameters (note that our parameter space is partly discrete (V), and partly continuous (Ω)). We introduce the event $\{v, \omega\}$ because Ω correspond to events induced by continuous parameters $\boldsymbol{\theta}, s_j, p_{ij}, \pi_i$. Suppose that a real-valued additive component $B_{t,j} = \sum_{k=0}^{t} Y_{k,j}, j \in 1, \ldots, d$, with $B_t = (B_{t,1}, B_{t,2}, \ldots, B_{t,d})^\top$ and with $Y_k = (Y_{k,1}, Y_{k,2}, \ldots, Y_{k,d})^\top$ a r.v. taking values on $\mathbb{R}^d$, is adjoined to the chain such that $\{(X_t, B_t), t \geq 0\}$ is a Markov chain on $D \times \mathbb{R}^d$ and

$$
\begin{aligned}
&\mathrm{P}\{(X_t, B_t) \in A \times (B + b) | (X_{t-1}, B_{t-1}) = (i, b)\} \qquad (27)\\
&= \mathrm{P}\{(X_1, B_1) \in A \times B | (X_0, B_0) = (i, 0)\}\\
&= \mathrm{P}(i, A \times B) = \sum_{j \in A} \int_{b \in B} p_{ij}^{v \times \omega} f_j\{b; \boldsymbol{\theta}^{(j)}(v \times \omega), s^{(j)}(v \times \omega)\} \mu(db),
\end{aligned}
$$

where $B, b \subseteq \mathbb{R}^d$, $A \subseteq D$, $f_j\{b; \boldsymbol{\theta}^{(j)}(v, \omega), s^{(j)}(v, \omega)\}$ is the conditional density of Y_t given X_{t-1}, X_t with respect to a σ-finite measure μ on $\mathbb{R}^d$, and $\boldsymbol{\theta}(v, \omega) \in \Theta, s(v, \omega) \in S$, $j = 1, \ldots, M$ are the unknown parameters. That is, $\{X_t, t \geq 0\}$ is a Markov chain, given $X_0, X_1, \ldots, X_T$, with $Y_1, \ldots, Y_T$ being independent.$\{B_t, t \geq 0\}$ is called a hidden Markov model if there is a Markov chain $\{X_t, t \geq 0\}$ such that the process $\{(X_t, B_t), t \geq 0\}$ satisfies (27). Note that in (27), the usual parameterization $\boldsymbol{\theta}^{(j)}(v, \omega) = \boldsymbol{\theta}^{(j)}$, and $s^{(j)}(v, \omega) = s^{(j)}$. Recall the associated parameter space $\{V, \Omega\}$, where V consists of a set of discrete finite

elements and Ω is associated with the parameters $\boldsymbol{\theta}, [p_{ij}]_{i,j}$. Define $\mathbf{s}^*$ and $\boldsymbol{\theta}^*$ associated with the point $\{v^0, \omega^0\}$ in the parameter space, as in the following definitions:

$$q_T(Y_{0:T}; v^0, \omega^0) \stackrel{\text{def}}{=} \max_{j \in 0,\ldots,M} p_T(Y_{0:T} | x_1 = j; v^0, \omega^0) \tag{28}$$

$$H(v^0, \omega^0) \stackrel{\text{def}}{=} \mathsf{E}_{v^0,\omega^0}\{-\log p(Y_0 | Y_{-1}, Y_{-2}, \ldots; v^0, \omega^0)\},$$

where $Y_{-1}, \ldots, Y_{-T}$ are a finite number of past values of the process.

$$H(v^0, \omega^0, v, \omega) \stackrel{\text{def}}{=} \mathsf{E}_{v^0,\omega^0}\{\log p_T(Y_{0:T}; v, \omega)\}$$

Theorem 7.1 (Leroux (1992))**.** *Under A.1–A.5,*

$$\begin{aligned} \lim_{T\to\infty} T^{-1} \mathsf{E}_{v^0,\omega^0}\{\log p_T(Y_{0:T}; v^0, \omega^0)\} &= -H(v^0, \omega^0) \\ \lim_{T\to\infty} T^{-1} \log p_T(Y_{0:T}; v^0, \omega^0) &= -H(v^0, \omega^0), \end{aligned}$$

with probability 1*, under* (v^0, ω^0)*, and*

$$\begin{aligned} \lim_{T\to\infty} T^{-1} \mathsf{E}_{v^0,\omega^0}\{\log p_T(Y_{0:T}; v, \omega)\} &= H(v^0, \omega^0, v, \omega) \\ \lim_{T\to\infty} T^{-1} \log p_T(Y_{0:T}; v, \omega) &= H(v^0, \omega^0, v, \omega), \end{aligned}$$

with probability 1*, under* (v^0, ω^0)*.*

Lemma 7.2. *$\forall v_i, u_j, i, j \in 1, \ldots, M$ as weights, the difference between M linear combination of states would lead to*

$$\sum_{i=1}^{M} v_i f(y, \boldsymbol{\theta}_{s^{(i)}}, s^{(i)}) \neq \sum_{j=1}^{M} \mu_j f(y, \boldsymbol{\theta}_{s'^{(j)}}, s'^{(j)}). \tag{29}$$

Proof. For each $s^{(i)}, i \in 1, \ldots, M$ associated with dependency parameter $\boldsymbol{\theta}_{s^{(i)}} \in \mathbb{R}_+^d$.

So

$$\sum_{i=1}^{M} v_i \delta_{s^{(i)}} \neq \sum_{j=1}^{M} \mu_j \delta_{s'^{(j)}}, a.e. \tag{30}$$

implies

$$\sum_{i=1}^{M} v_i \delta_{s^{(i)}} \delta_{\boldsymbol{\theta}_{s^{(i)}}} \neq \sum_{j=1}^{M} \mu_j \delta_{s'^{(j)}} \delta_{\boldsymbol{\theta}_{s'^{(j)}}}, a.e.. \tag{31}$$

□

Also if (30), then the corresponding point in the parameter space (v, ω) would lead to $\mathcal{K}(v^0, \omega^0; v, \omega)$, and (v, ω) would not be in the equivalent class of (v^0, ω^0) as long as the point v and v^0 are different as (30) , (the equivalence class of v^0 is defined in Leroux (1992)), and the divergence between (v, ω) and (v^0, ω^0) is defined as $\mathcal{K}(v^0, \omega^0; v, \omega) \stackrel{\text{def}}{=} H(v^0, \omega^0, v^0, \omega^0) - H(v^0, \omega^0, v, \omega)$. It is connected to the log likelihood ratio process, and one can prove that if either (30) or (31) holds (A.2), (29) will hold, and it will lead to $\mathcal{K}(v^0, \omega^0; v, \omega) > 0$. Namely, the divergence can distinguish points from different equivalent classes.

Next, we study whether plugging in nonparametric estimated margins would affect the consistency results by analyzing the uniform convergence of $\hat{f}_i(y)$.

Recall $\hat{f}_i(y) \stackrel{\text{def}}{=} c\{\hat{F}_1^{\mathfrak{m}}(y_1), \hat{F}_2^{\mathfrak{m}}(y_2), \ldots, \hat{F}_d^{\mathfrak{m}}(y_d), \hat{\boldsymbol{\theta}}^{(i)}, \hat{s}^{(i)}\} \hat{f}_1^{\mathfrak{m}}(y_1) \hat{f}_2^{\mathfrak{m}}(y_2) \cdots \hat{f}_d^{\mathfrak{m}}(y_d)$, we have according to the uniform consistency of copulae density, for all $t \in 1, \ldots, T, i \in 1, \ldots, M$

$$\begin{aligned} & \max_{s^{(i)}} \sup_{y_{t1}, \ldots, y_{td} \in B^d, \boldsymbol{\theta}^{(i)} \in \Theta} |\hat{c}(\hat{F}_1^{\mathfrak{m}}(y_{t1}), \hat{F}_2^{\mathfrak{m}}(y_{t2}), \ldots, \hat{F}_d^{\mathfrak{m}}(y_{td}), \boldsymbol{\theta}^{(i)}, s^{(i)}) \qquad (32) \\ & -c(F_1^{\mathfrak{m}}(y_{t1}), F_2^{\mathfrak{m}}(y_{t2}), \ldots, F_d^{\mathfrak{m}}(y_{td}), \boldsymbol{\theta}^{(i)}, s^{(i)})| \\ \leq & \sum_{j=1}^{d} |c(F_{1,\eta_1}^{\mathfrak{m}}(y_{t1}), F_{2,\eta_2}^{\mathfrak{m}}(y_{t2}), \ldots, F_{d,\eta_d}^{\mathfrak{m}}(y_{td}))\{\hat{F}_j^{\mathfrak{m}}(y_{tj}) - F_j^{\mathfrak{m}}(y_{tj})\}|, \qquad (33) \end{aligned}$$

where $F_{j,\eta_j}^{\mathfrak{m}}(\cdot) \stackrel{\text{def}}{=} F_j^{\mathfrak{m}}(\cdot) + \eta_j[F(\cdot) - F_j^{\mathfrak{m}}(\cdot)]$, $\eta_j = [0, 1]$, and $F_{j,\eta_j}^{\mathfrak{m}}(\cdot)$ lies in the set of admission function for $F_j^{\mathfrak{m}}$.

Bickel et al. (1998) states that as $\{X_t\}$ is ergodic, then it following $\{Y_t\}$ is also ergodic. It is known that any strictly irreducible and aperiodic Markov chain is going to be $\beta-$ mixing, Bradley (1986). Then the marginal distribution of $Y_{tm}, m = 1, \cdots, M$ follows a process that is β mixing with the exponential decay rate, namely $\beta_t = \mathcal{O}\{t^{-b}\}$ for some constant a. The temporal dependence of the marginal univariate time series Y_{tm} be only

inherited only from the underlying Markov chain as it is a measurable transformation of X_t. $\{Y_t\}$ follow HMM HAC, then the mariginal distribution of Y_{tm} would follow a process that is β mixing with decay rate $\beta_t = \mathcal{O}(b^{-t})$ for some constant b. Then following the results of Liu and Wu (2010), under assumptions A1-A5, we will have Bickel and Rosenblatt (1973) type of uniform consistency for the marginal kernel density estimation.

$$\sup_{y \in B} |\hat{f}_i^{\mathfrak{m}}(y) - f_i^{\mathfrak{m}}(y)| = \mathcal{o}_p(1), \tag{34}$$

Also according to Chen and Fan (2005),

$$\sqrt{n} \sup_{y \in B} |\hat{F}_m^{\mathfrak{m}}(y) - F_m^{\mathfrak{m}}(y)| = \mathcal{o}_p(1). \tag{35}$$

Finally we will have,

$$\begin{aligned} \max_s \sup_{y_1,\ldots,y_d \in B^d, \boldsymbol{\theta} \in E} \quad & |\hat{c}(\hat{F}_1^{\mathfrak{m}}(y_1), \hat{F}_2^{\mathfrak{m}}(y_2), \ldots, \hat{F}_d^{\mathfrak{m}}(y_d), \boldsymbol{\theta}^{(i)}, s^{(i)}) \\ & - c(F_1^{\mathfrak{m}}(y_1), F_2^{\mathfrak{m}}(y_2), \ldots, F_d^{\mathfrak{m}}(y_d), \boldsymbol{\theta}^{(i)}, s^{(i)})| = \mathcal{o}_p(1) \end{aligned}$$

Therefore, the multivariate distribution at each state would have the following property,

$$\sup_{y \in B^d} |\hat{f}_j(y) - f_j(y)| = \mathcal{o}_p(1),$$

where B, B^d are compact sets. So the plug in estimation would not contaminate the consistency results.

To prove the consistency of our estimation of this parameter, we restate the theorems of consistency in Leroux (1992) for our parameter space. One needs to show that first for the discrete subspace V^c which does not contain any point of the equivalence class of v^0, for $v \in V^c$ and any arbitrary value of $\omega \in \Omega$, it holds, with probability 1,

$$\lim_{T \to \infty} \left[\min_{v \in V^c} \log \sup_{\omega \in \Omega} p_T(Y_{0:T}; v, \omega) - \log p_T(Y_{0:T}; v^0, \omega^0) \right] \to -\infty. \tag{36}$$

The fact follows directly from lemma 7.2 (the identifiability of the states parameters), and its consequence $\mathcal{K}(v^0, \omega^0; v, \omega) > 0$. Theorem 3.1 is proved.

To prove Theorem 3.2, note that $\lim_{T\to\infty} \max_{i\in 1,\ldots,M} \mathrm{P}(|\hat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{*(i)}| > \varepsilon | \hat{s}^{(i)} = s^{*(i)})$ is conditioning on the event $\{\hat{s}^{(i)} = s^{*(i)}\}$ which asymptotically holds with probability 1. Therefore it is suffice to prove, for any $\hat{s}^{(i)} = s^{(i)}$

$$\lim_{T\to\infty} \min_{i\in 1,\ldots,M} \mathrm{P}(|\hat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{*(i)}| > \varepsilon) = 0. \tag{37}$$

To show (37), one needs to show that for (V^c, Ω^c) which does not contain any point of the equivalence class of (v^0, ω^0), we have, with probability 1,

$$\lim_{T\to\infty} \{\log \sup_{\omega\in\Omega^c} p_T(Y_{0:T}; v^0, \omega) - \log p_T(Y_{0:T}; v^0, \omega^0)\} \to -\infty, \tag{38}$$

which is implied from the following statement: for any closed subset C of Ω^c, there exists a sequence of open subsets of $\mathcal{O}_{\omega_h}$ with $h = 1, \ldots, H$ with $C \subseteq \cup_{h=1}^{H} \mathcal{O}_{\omega_h}$, such that

$$\lim_{T\to\infty} \{\max_h \log \sup_{\omega\in\mathcal{O}_{\omega_h}} p_T(Y_{0:T}; v^0, \omega) - \log p_T(Y_{0:T}; v^0, \omega^0)\} \to -\infty. \tag{39}$$

To prove (39), we have the modified definition:

$$H(v^0, \omega^0, v^0, \omega; \mathcal{O}_{\omega_h}) \stackrel{\text{def}}{=} \lim_T \log \sup_{\omega'\in\omega^0} q_T(Y_{0:T}, v^0, \omega')/T. \tag{40}$$

It can be derived that

$$H(v^0, \omega^0, v^0, \omega) < H(v^0, \omega^0, v^0, \omega^0), \tag{41}$$

for (v^0, ω) and (v^0, ω^0) does not lie in the same equivalence class. Then (41) is a consequence of the identifiability condition A.2, and this leads to: $\exists \varepsilon > 0$, T_ε and $\mathcal{O}_\omega$ such

that

$$\mathsf{E}\log\sup_{\omega'\in\mathcal{O}_\omega} q_{T_\varepsilon}(v^0,\omega')/T_\varepsilon < \mathsf{E}\log q_{T_\varepsilon}(v^0,\omega)/T_\varepsilon + \varepsilon < H(v^0,\omega^0,v^0,\omega^0) - \varepsilon.$$

Also because $\log\sup_{\omega'\in\mathcal{O}_\omega} p_T(Y_{0:T},v^0,\omega')/T$ and $\log\sup_{\omega'\in\mathcal{O}_\omega} q_T(Y_{0:T},v^0,\omega')/T$ have the same limit value, there exists a constant $\varepsilon>0$,

$$\lim_{T\to\infty}\log\sup_{\omega'\in\mathcal{O}_{\omega_h}} p_T(y_{0:T},v^0,\omega')/T = H(v^0,\omega^0,v^0,\omega;\mathcal{O}_{\omega_h}) \le H(v^0,\omega^0,v^0,\omega^0)-\varepsilon.$$

Now (39) follows.

7.2 Proof of Theorem 3.3

Recall from the last subsection, under A.3,

$$\sup_{y\in B}|\hat{f}_i^{\mathfrak{m}}(y) - f_i^{\mathfrak{m}}(y)| = o_p(1) \tag{42}$$

$$\sqrt{n}\sup_{y\in B}|\hat{F}_m^{\mathfrak{m}}(y) - F_m^{\mathfrak{m}}(y)| = o_p(1). \tag{43}$$

Let $U_{tm}\overset{\text{def}}{=}F_m^{\mathfrak{m}}(Y_{tm})$, $\tilde{U}_{tm}\overset{\text{def}}{=}\hat{F}_m^{\mathfrak{m}}(Y_{tm})$, and $\mathbf{U}_t\overset{\text{def}}{=}(U_{t1},\cdots,U_{td})$. Define the log likelihood $L_T(\boldsymbol{\theta})=L_T(\boldsymbol{\theta},\mathbf{U}_{0:T})\overset{\text{def}}{=}\log p_T(y_{0:T})$, in our case we are working with $L_T(\boldsymbol{\theta},\tilde{\mathbf{U}}_{0:T})$. Relying on the LAN property proved in Bickel et al. (1998), under A.1-A.7,

$$L_T(\boldsymbol{\theta}^*+T^{-1/2}\boldsymbol{\theta},\mathbf{U}_{1:T}) - L_T(\boldsymbol{\theta}^*,\mathbf{U}_{1:T}) = T^{-1/2}\boldsymbol{\theta}^\top\partial L_T(\boldsymbol{\theta}^*) + T^{-1}\boldsymbol{\theta}^\top\partial^2 L_T(\boldsymbol{\theta}^*)\boldsymbol{\theta}/2 + R_T(\boldsymbol{\theta}), \tag{44}$$

where $R_T(\boldsymbol{\theta})$ tends to zero in probability uniformly of the compact subset of parameter space of $\boldsymbol{\theta}$.

Next we need to prove that uniformly over $\boldsymbol{\theta}$

$$\begin{aligned}&L_T(\boldsymbol{\theta}^*+T^{-1/2}\boldsymbol{\theta},\mathbf{U}_{1:T}) - L_T(\boldsymbol{\theta}^*,\mathbf{U}_{1:T}) - L_T(\boldsymbol{\theta}^*+n^{-1/2}\boldsymbol{\theta},\tilde{\mathbf{U}}_{1:T}) + L_T(\boldsymbol{\theta}^*,\tilde{\mathbf{U}}_{1:T})\\ &-T^{-1/2}\boldsymbol{\theta}^\top\sum_t\sum_m W_m(U_{tm}) = o_p\{R_T(\boldsymbol{\theta})\},\end{aligned}$$

where

$$W_m(U_{tm}) \stackrel{\text{def}}{=} \int_{v_1,\cdots,v_d} \{\mathbf{1}(U_{tm} \le v_m) - v_m\} (\mathsf{E}\, \partial a_t b_m / \partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}) c(v_1, \cdots, v_d, \boldsymbol{\theta}^{*(m)}, s^{*(m)}) dv_1 \cdots dv_d$$

. $a_t(\cdot)$ and $b_m(\cdot)$ are functions defined later in the proof.

Similarly, we have

$$\begin{aligned}
& L_T(\boldsymbol{\theta}^*, \tilde{\mathbf{U}}_{1:T}) - L_T(\boldsymbol{\theta}^*, \mathbf{U}_{1:T}) \\
= \;& \log\Big(\frac{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \prod_{t=1}^T p_{x_{t-1}x_t} \tilde{f}_{x_t}(y_t)}{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)}\Big) \\
= \;& \frac{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \prod_{t=1}^T p_{x_{t-1}x_t} \tilde{f}_{x_t}(y_t) - \sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)}{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)} + \mathcal{O}_p(1) \\
\stackrel{\text{def}}{=} \;& \sum_t \sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \tilde{a}_t(\boldsymbol{\theta}^*) \{\tilde{f}_{x_t}(y_t) - f_{x_t}(y_t)\} + \mathcal{O}_p(1),
\end{aligned}$$

where $\tilde{a}_t(\boldsymbol{\theta}^*) = \frac{\pi_{x_0} \prod_{t=0}^{t_0-1} p_{x_{t-1}x_t} \tilde{f}_{x_t}(y_t) \prod_{t=t_0}^T p_{x_{t-1}x_t} f_{x_t}(y_t)}{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)}$.

As

$$\begin{aligned}
& \tilde{f}_{x_t}(y_t) - f_{x_t}(y_t) = c(\tilde{\mathbf{U}}_{0:T}, \boldsymbol{\theta}^{*(x_t)}, s^{*(x_t)}) \prod_{m=1}^d f_m^{\mathfrak{m}} - c(\mathbf{U}_{0:T}, \boldsymbol{\theta}^{*(x_t)}, s^{*(x_t)}) \prod_{j=1}^d f_j^{\mathfrak{m}} \\
= \;& \sum_m c_{u_m} \{F_1^{\mathfrak{m}}(y_{1t}), F_2^{\mathfrak{m}}(y_{2t}), \cdots, F_d^{\mathfrak{m}}(y_{dt}), \boldsymbol{\theta}^{*(x_t)}, s^{*(x_t)}\} \{\hat{F}_m^{\mathfrak{m}}(y_{mt}) - F_m^{\mathfrak{m}}(y_{mt})\} \prod_{j=1}^d f_j^{\mathfrak{m}} + \mathcal{O}_p(1) \\
\stackrel{\text{def}}{=} \;& \sum_m \tilde{b}_m(\boldsymbol{\theta}^{(x_t)}) \{\hat{F}_m^{\mathfrak{m}}(y_{mt}) - F_m^{\mathfrak{m}}(y_{mt})\} + \mathcal{O}_p(1),
\end{aligned}$$

where $\tilde{b}_m(\boldsymbol{\theta}^{(x_t)}) \stackrel{\text{def}}{=} c_{u_m}\{F^{\mathfrak{m}}(y_{1t}), F^{\mathfrak{m}}(y_{2t}), \cdots, F^{\mathfrak{m}}(y_{dt}), \boldsymbol{\theta}^{(x_t)}, s^{(x_t)}\} \prod_{j=1}^d f_j^{\mathfrak{m}}$, and c_{u_m} denotes the partial derivative of the copulae density w.r.t. u_m.

Then it follows that

$$\begin{aligned}
& L_T(\boldsymbol{\theta}^* + T^{-1/2}\boldsymbol{\theta}, \mathbf{U}_{1:T}) - L_T(\boldsymbol{\theta}^*, \mathbf{U}_{1:T}) - L_T(\boldsymbol{\theta}^* + T^{-1/2}\boldsymbol{\theta}, \tilde{\mathbf{U}}_{1:T}) + L_T(\boldsymbol{\theta}^*, \tilde{\mathbf{U}}_{1:T}) \\
= \; & T^{-1/2}\boldsymbol{\theta}^\top \sum_{x_0=1}^{M} \cdots \sum_{x_T=1}^{M} \sum_t \left[\sum_m \partial \tilde{a}_t \tilde{b}_m / \partial \boldsymbol{\theta} \{ \hat{F}^{\mathfrak{m}}(y_{mt}) - F^{\mathfrak{m}}(y_{mt}) \} \right] \\
& + \mathcal{O}_p(T^{-1/2}) \\
= \; & T^{-1/2}\boldsymbol{\theta}^\top \sum_t \sum_m W_m(U_{tm}) + \mathcal{O}_p(T^{-1/2})
\end{aligned}$$

So let

$$\begin{aligned}
B &\stackrel{\text{def}}{=} \mathsf{E}\{\partial^2 L_T(\boldsymbol{\theta}^*, \mathbf{U}_{1:T})\} \\
A &\stackrel{\text{def}}{=} \{\partial L_T(\boldsymbol{\theta}^*, \mathbf{U}_{1:T}) + \sum_t \sum_m W_m(U_{tm})\}, \qquad (45)
\end{aligned}$$

Finally we have the estimated $\hat{\boldsymbol{\theta}}$ can be represented by $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = B^{-1}A + \mathcal{O}_p(T^{-1/2})$ coming from Bickel et al. (1998).

References

Adrian, T. and Brunnermeier, M. K. (2011). CoVaR, *Staff Reports 348*, Federal Reserve Bank of New York.

Ailliot, P., Thompson, C. and Thomson, P. (2009). Space-time modeling of precipitation by using a hidden markov model and censored gaussian distributions, *Journal of the Royal Statistical Society* **58**: 405–426.

Bickel, P. J., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models, *Annals of Statistics* **26**(4): 1614–1635.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *The Annals of Statistics* **1**: 1071–1095.

Bradley, R. (1986). Basic properties of strong mixing conditions, *In: Eberlein, E., Taqqu, M.S. (Eds.), Dependence in Probability and Statistics. Birkhauser, Boston* pp. 165–192.

Caia, Z., Chen, X., Fan, Y. and Wang, X. (2006). Selection of copulas with applications in finance, *Working paper.* available at http://www.economics.smu.edu.sg/femes/2008/papers/219.pdf.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*, Springer Verlag.

Chen, X. and Fan, Y. (2005). Estimation of copula-based semiparametric time series models, *Journal of Econometrics* **130**(2): 307–335.

Chen, X. and Fan, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspesification, *Journal of Econometrics* **135**: 125–154.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion), *J. Roy. Statistical Society B* **39**: 1–38.

Engle, R. (2002). Dynamic conditional correlation, *Journal of Business and Economic Statistics* **20**(3): 339–350.

Fuh, C.-D. (2003). SPRT and CUSUM in hidden Markov Models, *Ann. Statist.* **31**(3): 942–977.

Gao, X. and Song, P. X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden markov model, *Statistica Sinica* **21**: 165–185.

Giacomini, E., Härdle, W. K. and Spokoiny, V. (2009). Inhomogeneous dependence modeling with time-varying copulae, *Journal of Business and Economic Statistics* **27**(2): 224–234.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* **57**(2): 357–384.

Härdle, W., Herwartz, H. and Spokoiny, V. (2003). Time inhomogeneous multiple volatility modeling, *Journal of Financial econometrics* **1**(1): 55–95.

Härdle, W. K., Okhrin, O. and Okhrin, Y. (2013). Dynamic structured copula models, *Statistics & Risk Modeling* **30**(4).

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

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden markov models, *Stochastic Processes and their Applications* **40**: 127–143.

Liu, W. and Wu, W. (2010). Simultaneous nonparametric inference of time series, *The Annals of Statistics* **38**: 2388–2421.

McLanchlan, G. and Peel, D. (2000). *Finite Mixture Models*, Wiley.

McNeil, A. J. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1 norm symmetric distributions, *Annals of Statistics* **37**(5b): 3059–3097.

Nelsen, R. B. (2006). *An Introduction to Copulas*, Springer Verlag, New York.

Okhrin, O., Okhrin, Y. and Schmid, W. (2013). On the structure and estimation of hierarchical Archimedean copulas, *Journal of Econometrics* **173**: 189–204.

Okimoto, T. (2008). Regime switching for dynamic correlations, *Journal of Financial and Quantitative Analysis* **43**(3): 787–816.

Patton, A. J. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation, *Journal of Financial Econometrics* **2**: 130–168.

Pelletier, D. (2006). Regime switching for dynamic correlations, *Journal of Econometrics* **131**: 445–473.

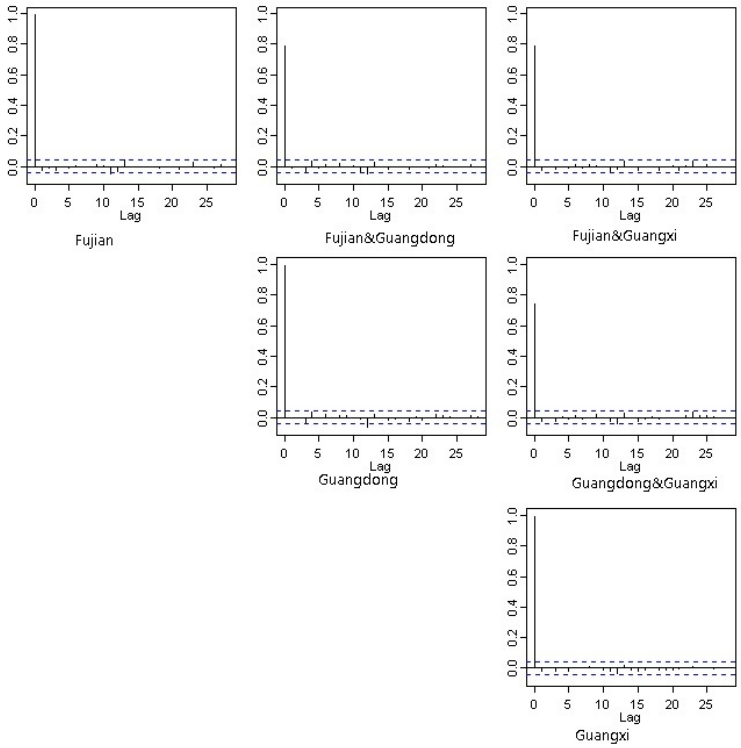
Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proceedings of IEEE* **77**(2).

Rodriguez, J. C. (2007). Measuring financial contagion: a copula approach, *Journal of Empirical Finance* **14**: 401–423.

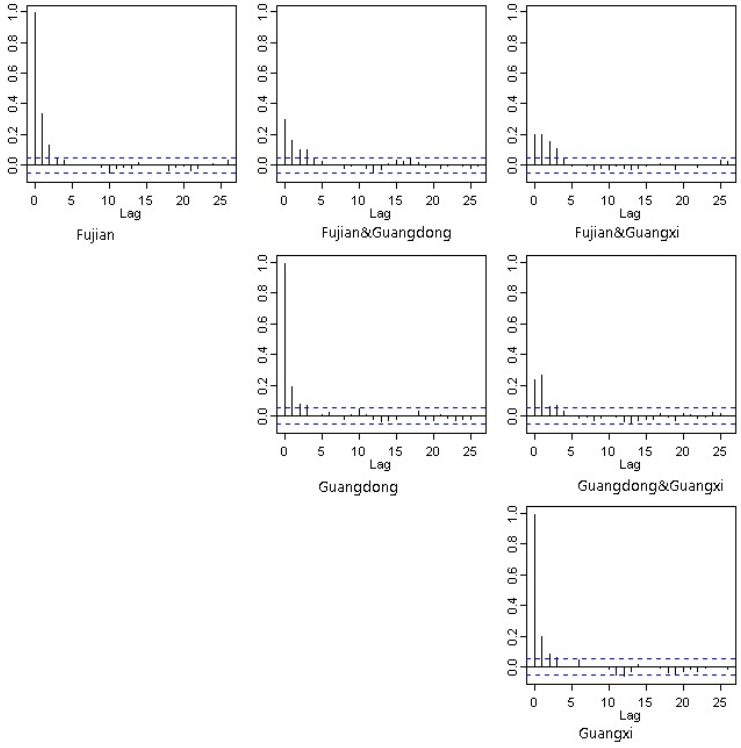
Savu, C. and Trede, M. (2010). Hierarchical Archimedean copulas, *Quantitative Finance* **10**: 295–304.

Sklar, A. (1959). Fonctions dé repartition á n dimension et leurs marges, *Publ. Inst. Stat. Univ. Paris* **8**: 299–231.

Whelan, N. (2004). Sampling from Archimedean copulas, *Quantitative Finance* **4**: 339–352.



(a) the simulated rainfall time series.



(b) the original rainfall time series.

Figure 14: Autocorrelations and cross-correlations of the simulated rainfall and original time series