High Dimension Copula-Based Distributions with Mixed Frequency Data

Dong Hwan Oh

Federal Reserve Board

Andrew J. Patton

Duke University

February 2015

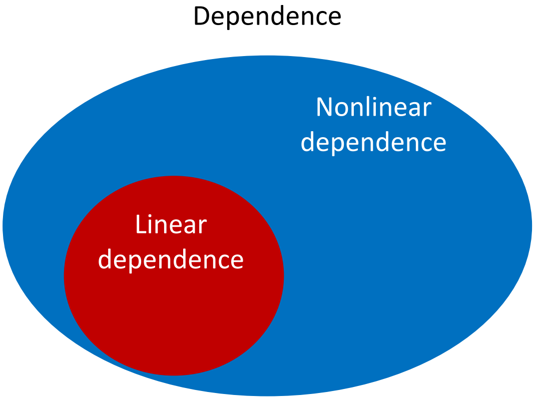
Motivation

- A model for the distribution of returns on a collection of financial assets is crucial for risk management and asset allocation
  - And these collections tend to be **large**: eg, median number of stocks held by US mutual funds is **94** (25/75 percentiles are 46 and 208)
- But there are relatively few dynamic, high-dimension models available
  - Many are based on multivariate Normality, despite its limitations
  - Almost all use data from a common sampling frequency
- We propose a new approach for **constructing and estimating** high dimension distribution models, drawing on two areas of recent research:
  1. **High frequency data** is very useful for estimating lower-frequency second moments (eg, correlation)
  2. **Copula-based distributions** are useful for constructing flexible models in high dimensions

High frequency data in lower frequency copula models

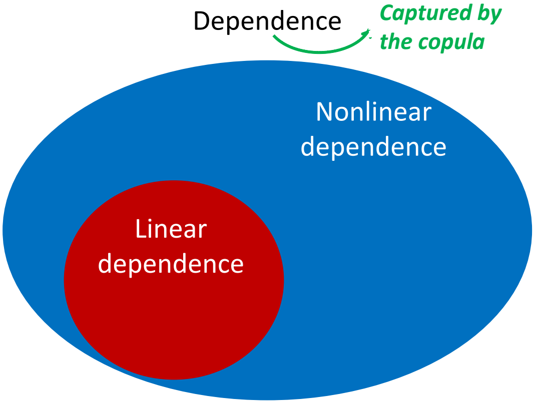
- Exploiting high frequency data in lower-frequency copula-based models is not straightforward:
  - Unlike covariances, the copula of daily returns is not generally a known function of the copula of high frequency returns
  - So most of the nice theory from high frequency financial econometrics cannot be used directly
- We propose decomposing the dependence structure of daily returns into **linear** and **nonlinear** components:
  - **High frequency data** is used to accurately model the linear dependence
  - **Low frequency data** and a new class of copulas is used to capture the remaining nonlinear dependence

Decomposition of dependence



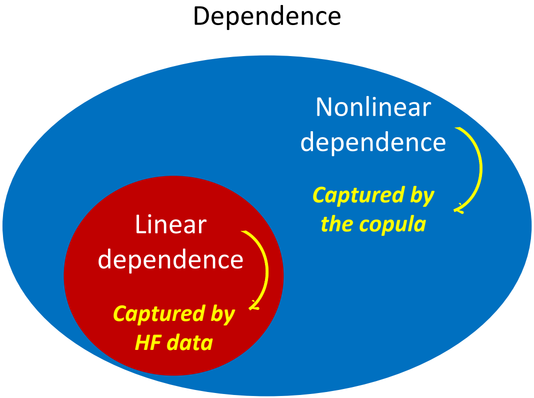
- **Linear** dependence: Captured by **correlation**
- **Nonlinear** dependence: Any dependence **beyond linear correlation**

Standard use of copulas in the literature



- Chasing two rabbits with only one tool
- A heavy burden for the copula model

Our approach: in pictures



- Chasing two rabbits with two tools: high frequency data and copulas
- High frequency data shares the heavy burden with the copula model

Our approach: in equations

- We construct a model for a N-vector of daily returns $\mathbf{r}_t$ as follows. Let:

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu}_t + \mathbf{H}_t^{1/2}\mathbf{e}_t \\ \text{where} \quad \mathbb{E}_{t-1}\left[\mathbf{e}_t\right] &= 0, \ \ \mathbb{E}_{t-1}\left[\mathbf{e}_t\mathbf{e}_t'\right] = \mathbf{I} \end{aligned}$$

Our approach: in equations

- We construct a model for a N-vector of daily returns $\mathbf{r}_t$ as follows. Let:

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu}_t + \mathbf{H}_t^{1/2}\mathbf{e}_t \\ \text{where} \quad \mathbb{E}_{t-1}[\mathbf{e}_t] &= 0, \;\; \mathbb{E}_{t-1}[\mathbf{e}_t\mathbf{e}_t'] = \mathbf{I} \end{aligned}$$

- Use standard methods to estimate $\boldsymbol{\mu}_t$
- Use **high frequency data** to obtain improved estimates of $\mathbf{H}_t$
    - We propose a HAR-type model for $\mathbf{H}_t$ (more details below)

Our approach: in equations

- We construct a model for a N-vector of daily returns $\mathbf{r}_t$ as follows. Let:

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu}_t + \mathbf{H}_t^{1/2}\mathbf{e}_t \\ \text{where} \quad \mathbb{E}_{t-1}\left[\mathbf{e}_t\right] &= 0, \ \mathbb{E}_{t-1}\left[\mathbf{e}_t\mathbf{e}_t'\right] = \mathbf{I} \end{aligned}$$

- Use standard methods to estimate $\boldsymbol{\mu}_t$
- Use **high frequency data** to obtain improved estimates of $\mathbf{H}_t$
  - We propose a HAR-type model for $\mathbf{H}_t$ (more details below)
- Decompose the distribution of the **uncorrelated residuals** as

$$\mathbf{e}_t \sim iid \ \mathbf{F}\left(\cdot;\boldsymbol{\eta}\right) = \mathbf{C}\left(F_1\left(\cdot;\boldsymbol{\eta}\right), ..., F_N\left(\cdot;\boldsymbol{\eta}\right);\boldsymbol{\eta}\right)$$

Our approach: in equations

- We construct a model for a N-vector of daily returns $\mathbf{r}_t$ as follows. Let:

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu}_t + \mathbf{H}_t^{1/2}\mathbf{e}_t \\ \text{where} \quad \mathbb{E}_{t-1}[\mathbf{e}_t] &= 0, \;\; \mathbb{E}_{t-1}[\mathbf{e}_t\mathbf{e}_t'] = \mathbf{I} \end{aligned}$$

- Use standard methods to estimate $\boldsymbol{\mu}_t$
- Use **high frequency data** to obtain improved estimates of $\mathbf{H}_t$
  - We propose a HAR-type model for $\mathbf{H}_t$ (more details below)
- Decompose the distribution of the **uncorrelated residuals** as

$$\mathbf{e}_t \sim iid\ \mathbf{F}(\cdot;\boldsymbol{\eta}) = \mathbf{C}(F_1(\cdot;\boldsymbol{\eta}), ..., F_N(\cdot;\boldsymbol{\eta});\boldsymbol{\eta})$$

  - Can easily choose F_i to ensure that $\mathbb{E}[e_{it}] = 0$ and $\mathbb{E}\left[e_{it}^2\right] = 1$
  - But also need to ensure that $\mathbf{F}$ is such that $\mathbb{E}[e_{it}e_{jt}] = 0 \ \forall \ i \neq j$

Contributions of this paper

- This paper makes four main contributions. We:

1. Propose a new class of "**jointly symmetric**" copulas, useful in MV density models that contain a covariance matrix model (eg, DCC, HAR, SV, etc.)
2. Show that **composite likelihood** methods can be used to estimate these new models, and verify good finite sample properties via simulations
3. Propose a new, simple model for **high-dimension covariance matrices**, drawing on the HAR and DCC models of Corsi (2009) and Engle (2002)
4. Apply these news models to a detailed study of **104 US equity returns**, and show that they outperform existing approaches both in- and out-of-sample

Outline

1. Introduction
2. Models of linear and nonlinear dependence
   - Jointly symmetric copulas
   - A new covariance matrix model
3. Estimation and comparison via composite likelihood
4. Simulation study
5. Analysis of S&P 100 equity returns

Outline

A model for uncorrelated residuals

- A key building block for our model is an N-dim distribution **F** that guarantees an identity correlation matrix
- There are very few existing copulas that do this
  - Normal copula with identity correlation matrix (ie, independence copula)
  - t copula with identity correlation matrix, when combined with symmetric marginals
- The idea in this paper is to exploit the fact that multivariate distributions that satisfy a certain symmetry condition automatically ensure zero correlation

Joint symmetry and lack of correlation

Definition: Let $\mathbf{X}$ be a vector of N variables and let $\mathbf{a} \in R^N$. Then $\mathbf{X}$ is **jointly symmetric** about $\mathbf{a}$ if the following 2^N vectors of N random variables have the same joint distribution

$$\begin{aligned} \tilde{\mathbf{X}}^{(i)} &= \left[\tilde{X}_1^{(i)}, ..., \tilde{X}_1^{(N)}\right], \quad i = 1, 2, ..., 2^N \\ \text{where } \tilde{X}_j^{(N)} &= (X_j - a_j) \text{ or } (a_j - X_j) \quad \text{for } j = 1, 2, .., N \end{aligned}$$

Joint symmetry and lack of correlation

Definition: Let $\mathbf{X}$ be a vector of N variables and let $\mathbf{a} \in R^N$. Then $\mathbf{X}$ is **jointly symmetric** about $\mathbf{a}$ if the following 2^N vectors of N random variables have the same joint distribution

$$\begin{aligned} \tilde{\mathbf{X}}^{(i)} &= \left[\tilde{X}_1^{(i)}, \ldots, \tilde{X}_1^{(N)}\right], \quad i = 1, 2, \ldots, 2^N \\ \text{where } \tilde{X}_j^{(N)} &= (X_j - a_j) \text{ or } (a_j - X_j) \quad \text{for } j = 1, 2, .., N \end{aligned}$$

Lemma 1: If $\mathbf{X}$ is jointly symmetric and has finite second moments, then it has an **identity correlation matrix**.

Joint symmetry and lack of correlation

Definition: Let $\mathbf{X}$ be a vector of N variables and let $\mathbf{a} \in R^N$. Then $\mathbf{X}$ is **jointly symmetric** about $\mathbf{a}$ if the following 2^N vectors of N random variables have the same joint distribution

$$\begin{aligned} \tilde{\mathbf{X}}^{(i)} &= \left[\tilde{X}_1^{(i)}, ..., \tilde{X}_1^{(N)}\right], \quad i = 1, 2, ..., 2^N \\ \text{where } \tilde{X}_j^{(N)} &= (X_j - a_j) \text{ or } (a_j - X_j) \text{ for } j = 1, 2, .., N \end{aligned}$$

Lemma 1: If $\mathbf{X}$ is jointly symmetric and has finite second moments, then it has an **identity correlation matrix**.

Lemma 2: Let $\mathbf{X} \sim \mathbf{F} = \mathbf{C}(F_1, ..., F_N)$, where X_i is symmetric about a_i $\forall$ i. Then $\mathbf{X}$ is jointly symmetric iff $\mathbf{C}$ is jointly symmetric.

Joint symmetry and lack of correlation

Definition: Let $\mathbf{X}$ be a vector of N variables and let $\mathbf{a} \in R^N$. Then $\mathbf{X}$ is **jointly symmetric** about $\mathbf{a}$ if the following 2^N vectors of N random variables have the same joint distribution

$$\begin{aligned} \tilde{\mathbf{X}}^{(i)} &= \left[\tilde{X}_1^{(i)}, ..., \tilde{X}_1^{(N)}\right], \quad i = 1, 2, ..., 2^N \\ \text{where } \tilde{X}_j^{(N)} &= (X_j - a_j) \text{ or } (a_j - X_j) \text{ for } j = 1, 2, .., N \end{aligned}$$

Lemma 1: If $\mathbf{X}$ is jointly symmetric and has finite second moments, then it has an **identity correlation matrix**.

Lemma 2: Let $\mathbf{X} \sim \mathbf{F} = \mathbf{C}(F_1, ..., F_N)$, where X_i is symmetric about $a_i \ \forall \ i$. Then $\mathbf{X}$ is jointly symmetric iff $\mathbf{C}$ is jointly symmetric.

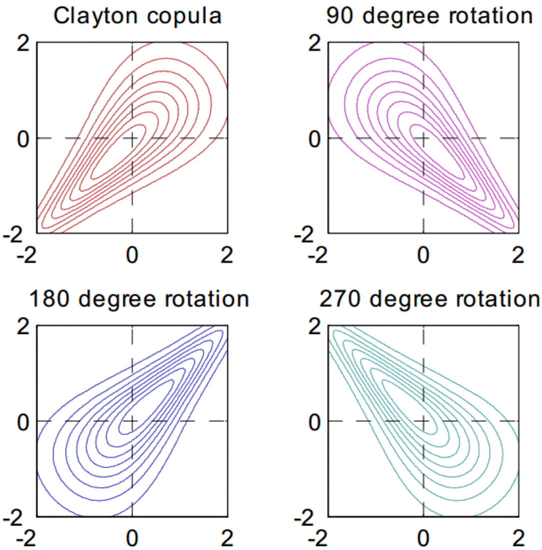
Result: Any combination of symmetric marginals and jointly symmetric copula yields a jointly symmetric joint distribution, implying an identity correlation matrix

Jointly symmetric copula models

- There are numerous interesting/useful copula models in the literature, almost none of which are jointly symmetric.
- We overcome this lack of choice by proposing a novel way to obtain a jointly symmetric copula: **rotate existing copulas**

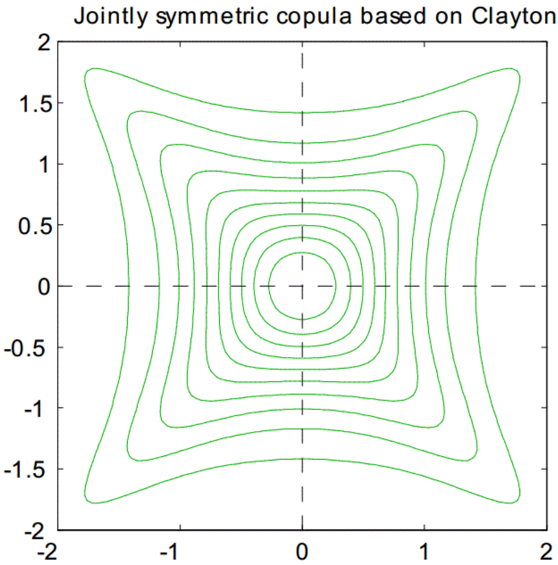
Example: rotations of the Clayton copula

Bivariate distributions with rotated Clayton copulas and N(0,1) margins



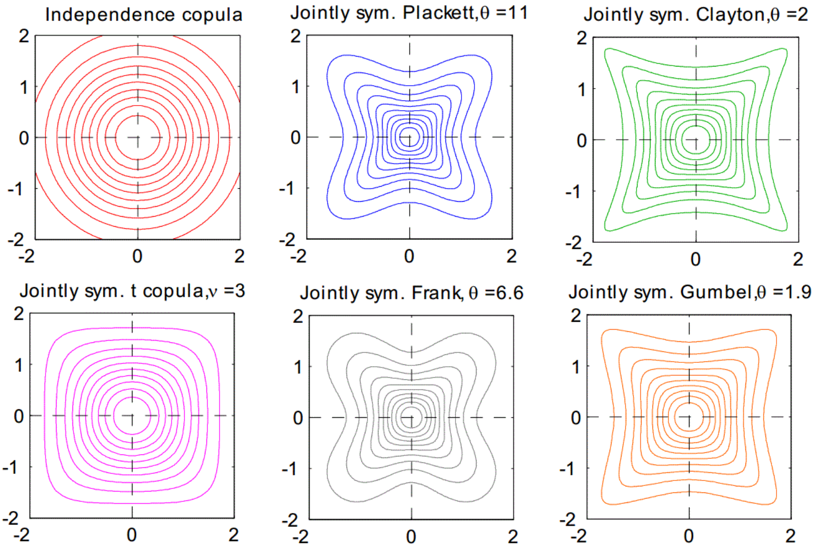
Example: a jointly symmetric Clayton copula

Bivariate distributions with jointly symmetric Clayton copula and N(0,1) margins



Example: other jointly symmetric distributions

Bivariate distributions with jointly symmetric copulas and N(0,1) margins



N-dimensional jointly symmetric copulas

Theorem: Given any N-dimensional copula $\mathbf{C}$ with density $\mathbf{c}$, then

(i) The following copula $\mathbf{C}^{JS}$ is jointly symmetric:

$$\mathbf{C}^{JS}(u_1, \ldots, u_N) = \frac{1}{2^N}\left[\sum_{k_1=0}^{2} \cdots \sum_{k_N=0}^{2} (-1)^R \cdot \mathbf{C}(\widetilde{u}_1, \ldots, \widetilde{u}_N)\right]$$

$$\text{where} \quad \widetilde{u}_i = \begin{cases} 1, & k_i = 0 \\ u_i, & k_i = 1 \\ 1 - u_i, & k_i = 2 \end{cases}, \quad \text{and } R = \sum_{i=1}^{N} \mathbf{1}\{k_i = 2\}$$

(ii) The probability density function $\mathbf{c}^{JS}$ implied by $\mathbf{C}^{JS}$ is

$$\mathbf{c}^{JS}(u_1, \ldots, u_N) = \frac{1}{2^N}\left[\sum_{k_1=1}^{2} \cdots \sum_{k_N=1}^{2} \mathbf{c}(\widetilde{u}_1, \ldots, \widetilde{u}_N)\right]$$

Outline

1. Introduction
2. Models of linear and nonlinear dependence
   - Jointly symmetric copulas
   - **A new covariance matrix model**
3. Estimation and comparison via composite likelihood
4. Simulation study
5. Analysis of S&P 100 equity returns

A new, simple covariance matrix model I

- Let Δ be the sampling frequency (eg, five minutes), yielding $1/\Delta$ observations per trade day, and define the realized covariance matrix as

$$\begin{aligned} RCov_t^{\Delta} &= \sum_{j=1}^{1/\Delta} \mathbf{r}_{t-1+j\Delta}\mathbf{r}'_{t-1+j\Delta} = \sqrt{\mathbf{RV}_t^{\Delta}}\, RCorr_t^{\Delta} \sqrt{\mathbf{RV}_t^{\Delta}} \\ \text{where } \mathbf{RV}_t^{\Delta} &= diag\left\{\left[RV_{1t}^{\Delta}, ..., RV_{Nt}^{\Delta}\right]\right\} \end{aligned}$$

A new, simple covariance matrix model I

- Let Δ be the sampling frequency (eg, five minutes), yielding $1/\Delta$ observations per trade day, and define the realized covariance matrix as

$$\begin{aligned} RCov_t^{\Delta} &= \sum_{j=1}^{1/\Delta} \mathbf{r}_{t-1+j\Delta}\mathbf{r}'_{t-1+j\Delta} = \sqrt{\mathbf{RV}_t^{\Delta}}\, RCorr_t^{\Delta} \sqrt{\mathbf{RV}_t^{\Delta}} \\ \text{where } \mathbf{RV}_t^{\Delta} &= diag\left\{\left[RV_{1t}^{\Delta}, \ldots, RV_{Nt}^{\Delta}\right]\right\} \end{aligned}$$

- We suggest using a HAR model (Corsi, 2009) for the log realized variances:

$$\begin{aligned} \log RV_{i,t}^{\Delta} &= \phi_i^{(const)} + \phi_i^{(day)} \log RV_{i,t-1}^{\Delta} \\ &\quad + \phi_i^{(week)} \frac{1}{4} \sum_{j=2}^{5} \log RV_{i,t-j}^{\Delta} \\ &\quad + \phi_i^{(month)} \frac{1}{15} \sum_{j=6}^{20} \log RV_{i,t-j}^{\Delta} + \xi_{it} \end{aligned}$$

estimated via OLS for each variance.

A new, simple covariance matrix model II

- We next propose a HAR-type model for the realized correlation matrix, imposing parameter constraints similar to the DCC model of Engle (2002):

$$
\begin{aligned}
vech\left(RCorr_t^{\Delta}\right) &= (1-a-b-c)\, vech\left(\overline{RCorr_T^{\Delta}}\right) \\
&\quad + a \cdot vech\left(RCorr_t^{\Delta}\right) \\
&\quad + b \cdot \frac{1}{4}\sum_{k=2}^{5} vech\left(RCorr_{t-k}^{\Delta}\right) \\
&\quad + c \cdot \frac{1}{15}\sum_{k=6}^{20} vech\left(RCorr_{t-k}^{\Delta}\right) + \xi_t
\end{aligned}
$$

where $(a, b, c) \in \mathbb{R}^3$.

- This parsimonious model can easily be estimated via OLS, and guarantees positive definiteness if $(a, b, c) > 0$ and $a + b + c < 1$.

Outline

1. Introduction
2. Models of linear and nonlinear dependence
   - Jointly symmetric copulas
   - A new covariance matrix model
3. **Estimation and comparison via composite likelihood**
4. Simulation study
5. Analysis of S&P 100 equity returns

Multi-stage estimation

- Our model for the vector of asset returns is

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu}_t + \mathbf{H}_t^{1/2}\mathbf{e}_t \\ \text{where} \quad \mathbf{e}_t &\sim iid \ \mathbf{F}(\cdot;\boldsymbol{\eta}) = \mathbf{C}(F_1(\cdot;\boldsymbol{\eta}), ..., F_N(\cdot;\boldsymbol{\eta}); \boldsymbol{\eta}) \end{aligned}$$

  and $\mathbf{F}$ is constrained so that $\mathbb{E}[\mathbf{e}_t] = 0$ and $\mathbb{E}[\mathbf{e}_t\mathbf{e}_t'] = \mathbf{I}$.

- We will first discuss estimation of $\mathbf{C}$, and then consider estimation of the rest of the model (in stages)
  - Inference methods will take into account the multi-stage estimation method

Composite likelihood estimation of the copula I

- Our method for constructing a JS cpoula requires 2^N evaluations of a given original copula density. Even for moderate dimensions this can be very slow
- Eg: computation time for **one evaluation** of density of JS Clayton:

N	10	20	30	50	100
Time	0.23 sec	4 min	70 hours	10^6 years	10^{17} years

Composite likelihood estimation of the copula I

- Our method for constructing a JS cpoula requires 2^N evaluations of a given original copula density. Even for moderate dimensions this can be very slow
- Eg: computation time for **one evaluation** of density of JS Clayton:

N	10	20	30	50	100
Time	0.23 sec	4 min	70 hours	10^6 years	10^{17} years

- We propose overcoming this difficulty by using **composite likelihood** methods (Lindsay 1988)
  - Estimate parameters of the full model by maximizing the likelihoods of submodels
  - Consistent if submodels are sufficient to identify parameter of full model
  - Less efficient, though loss need not be great

Composite likelihood estimation of the copula II

- Composite likelihood is particularly attractive for jointly symmetric copulas:

Proposition: For an N-dimensional jointly symmetric copula generated using Theorem 1, the (i, j) bivariate marginal copula density is obtained as

$$\mathbf{c}_{ij}^{JS}(u_i, u_j) = \frac{1}{4}\left\{\mathbf{c}_{ij}(u_i, u_j) + \mathbf{c}_{ij}(1\text{-}u_i, u_j) + \mathbf{c}_{ij}(u_i, 1\text{-}u_j) + \mathbf{c}_{ij}(1\text{-}u_i, 1\text{-}u_j)\right\}$$

where $\mathbf{c}_{ij}$ is the (i, j) marginal copula density of the original N-dimensional copula.

- Thus while the full copula model requires 2^N rotations of the original density, bivariate marginal copulas only require 2^2 rotations.

Composite likelihood estimation of the copula III

- Similar to Engle, *et al.* (2008), we consider CL based either on all pairs, adjacent pairs, or just one pair of variables:

$$
\begin{aligned}
CL_{all}\left(u_1,\ldots,u_N\right) &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\log \mathbf{c}_{i,j}\left(u_i,u_j\right)\\
CL_{adj}\left(u_1,\ldots,u_N\right) &= \sum_{i=1}^{N-1}\log \mathbf{c}_{i,i+1}\left(u_i,u_{i+1}\right)\\
CL_{first}\left(u_1,\ldots,u_N\right) &= \log \mathbf{c}_{1,2}\left(u_1,u_2\right)
\end{aligned}
$$

Composite likelihood estimation of the copula III

- Similar to Engle, *et al.* (2008), we consider CL based either on all pairs, adjacent pairs, or just one pair of variables:

$$
\begin{aligned}
CL_{all}(u_1,\ldots,u_N) &= \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \log \mathbf{c}_{i,j}(u_i,u_j) \\
CL_{adj}(u_1,\ldots,u_N) &= \sum_{i=1}^{N-1} \log \mathbf{c}_{i,i+1}(u_i,u_{i+1}) \\
CL_{first}(u_1,\ldots,u_N) &= \log \mathbf{c}_{1,2}(u_1,u_2)
\end{aligned}
$$

- Comparison of compuation times for single evaluation of log-likelihood:

N	10	20	30	50	100
Full likelihood	0.23 sec	4 min	70 hours	10^6 years	10^{17} years
All pairs CL	0.05 sec	0.21 sec	0.45 sec	1.52 sec	5.52 sec
Adjacent pairs CL	0.01 sec	0.02 sec	0.03 sec	0.06 sec	0.11 sec
First pair CL	0.001 sec	0.001 sec	0.001 sec	0.001 sec	0.001 sec

Composite likelihood estimation of the copula IV

- The maximum composite likelihood estimator (MCLE) is then obtained as:

$$\hat{\boldsymbol{\theta}}_{MCLE} = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{T} CL\left(u_{1t}, .., u_{Nt}; \boldsymbol{\theta}\right)$$

- Under standard regularity conditions, Cox and Reid (2004) show that

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{MCLE} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(0, \mathcal{H}_0^{-1}\mathcal{J}_0\mathcal{H}_0^{-1}\right)$$

- A key condition for CL to work is that the submodels used are rich enough to identify the parameters
  - This needs to be verified on a case by case basis
  - Is easily satisfied for the jointly symmetric copulas we consider: all have just a single unknown parameter, which appears in all bivariate submodels

Model selection tests with composite likelihood I

- We first define the composite Kullback-Leibler information criterion (cKLIC) following Varin and Vidoni (2005).

Definition (Varin and Vidoni, 2005): Given an N-dimension random variable $\mathbf{Z}$ with true density $\mathbf{g}$, the composite Kullback-Leibler information criterion (cKLIC) of a density $\mathbf{h}$ relative to $\mathbf{g}$ is

$$I_c(\mathbf{g}, \mathbf{h}) = E_{\mathbf{g}}\left[\log \prod_{i=1}^{N-1} \mathbf{g}_i(z_i, z_{i+1}) - \log \prod_{i=1}^{N-1} \mathbf{h}_i(z_i, z_{i+1})\right]$$

where $\prod_{i=1}^{N-1} \mathbf{g}_i(z_i, z_{i+1})$ and $\prod_{i=1}^{N-1} \mathbf{h}_i(z_i, z_{i+1})$ are adjacent-pair composite likelihoods using the true density $\mathbf{g}$ and a competing density $\mathbf{h}$.

- Above uses CL with adjacent pairs, but other cKLICs can be defined

Model selection tests with composite likelihood II

- Note that the expectation is with respect to the (complete) true density $\mathbf{g}$ rather than the CL of the true density, so it possible to interpret cKLIC as a linear combination of the KLICs of submodels:

$$I_c(\mathbf{g},\mathbf{h}) = \sum_{i=1}^{N-1} E_{\mathbf{g}}\left[\log \frac{\mathbf{g}_i(z_i, z_{i+1})}{\mathbf{h}_i(z_i, z_{i+1})}\right] = \sum_{i=1}^{N-1} E_{\mathbf{g}_i}\left[\log \frac{\mathbf{g}_i(z_i, z_{i+1})}{\mathbf{h}_i(z_i, z_{i+1})}\right]$$

- This implies that existing in-sample model selection tests, such as those of Vuong (1989) and Rivers and Vuong (2002) can be applied to model selection using cKLIC.

Model selection tests with composite likelihood III

- We may also wish to select the best model in terms of out-of-sample (OOS) forecasting performance measured by some scoring rule, $\mathcal{S}$, for the model.
- Gneiting and Raftery (2007) define "proper" scoring rules as those which ensure that the true density always receives a higher score than other densities
  - The log density, i.e. $\mathcal{S}\left(\mathbf{h}\left(\mathbf{Z}_{t+1}\right)\right) = \log \mathbf{h}\left(\mathbf{Z}_{t+1}\right)$ is proper.
- We may consider a similar scoring rule based on **log composite density**:

$$\mathcal{S}\left(\mathbf{h}\left(\mathbf{Z}_{t+1}\right)\right) = \sum_{i=1}^{N-1} \log \mathbf{h}_i\left(Z_{i,t+1}, Z_{i+1,t+1}\right)$$

- We show that this scoring rule is also proper.
- Thus OOS tests based on CL are related to the cKLIC, analogous to OOS tests based on the (full) likelihood being related to the KLIC.

Multi-stage estimation of the complete model

- In our empirical work we use an AR(1) for the mean: $\hat{\boldsymbol{\theta}}_i^{mean}$ $\forall$ i
- Estimate the individual variance models using the HAR model: $\hat{\boldsymbol{\theta}}_i^{var}$ $\forall$ i
- Estimate the HAR-correlation model: $\hat{\boldsymbol{\theta}}^{corr}$
- Compute the standardized uncorrelated residuals

$$\hat{\mathbf{e}}_t = \hat{\mathbf{H}}_t^{-1/2}\mathbf{r}_t$$

  and estimate their (symmetric) marginal distributions: $\hat{\boldsymbol{\theta}}_i^{mar}$ $\forall$ i
- Estimate the jointly symmetric copula model: $\hat{\boldsymbol{\theta}}^{cop}$.
- Define

$$\hat{\boldsymbol{\theta}}_{MSML} = \left[\hat{\boldsymbol{\theta}}_1^{mean}, ..., \hat{\boldsymbol{\theta}}_N^{mean}, \hat{\boldsymbol{\theta}}_1^{var}, \ldots, \hat{\boldsymbol{\theta}}_N^{var}, \hat{\boldsymbol{\theta}}^{corr}, \hat{\boldsymbol{\theta}}_1^{mar}, \ldots, \hat{\boldsymbol{\theta}}_N^{mar}, \hat{\boldsymbol{\theta}}^{cop}\right]$$

- Multi-stage ML estimation (including with a composite likelihood stage) is a form of multi-stage GMM estimation, and under standard regularity conditions it can be shown (see Newey and McFadden, 1994) that

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{MSML} - \boldsymbol{\theta}^*\right) \xrightarrow{d} N\left(0, V_{MSML}^*\right) \text{ as } T \to \infty$$

Outline

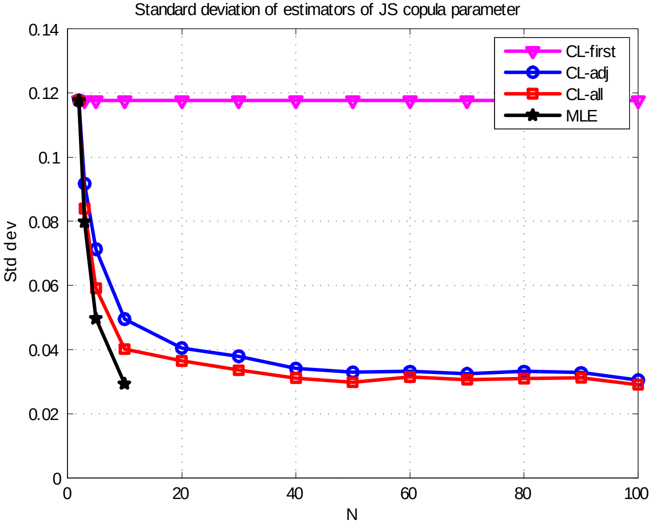
1. Introduction
2. Models of linear and nonlinear dependence
   - Jointly symmetric copulas
   - A new covariance matrix model
3. Estimation and comparison via composite likelihood
4. **Simulation study**
5. Analysis of S&P 100 equity returns

Finite-sample properties of CL estimators

- We consider the estimation of jointly symmetric copula parameters via composite likelihood, compared with maximum likelihood (where feasible).
- We use the JS Clayton and JS Gumbel copulas
- Dimension of problem varies: $N \in \{2, 3, 5, 10, ..., 100\}$.
- Sample size is $T = 1000$.

Standard deviation of ML and CL as a function of N

ML is best, but infeasible for N>10; CL-adj gets close to CL-all for N>50



Outline

1. Introduction
2. Models of linear and nonlinear dependence
   - Jointly symmetric copulas
   - A new covariance matrix model
3. Estimation and comparison via composite likelihood
4. Simulation study
5. **Analysis of S&P 100 equity returns**

Data used in our empirical analysis

- We study daily returns on all constituents of the S&P 100 index ($N = 104$) over the period January 2006–December 2012 ($T = 1761$)
- High frequency data is from the NYSE TAQ database, cleaned following Barndorff-Nielsen, Hansen, Lunde and Shephard (2009)
  - We adjust for stock splits and dividends using the adjustment factor from CRSP
- We use 5-minute sampling to compute the realized covariance matrix

Summary stats and mean models

	Mean	5%	25%	Median	75%	95%
	Cross-sectional distribution					
Summary statistics						
Skewness	-0.07	-0.66	-0.32	-0.03	0.18	0.56
Kurtosis	11.86	6.92	8.47	10.50	13.40	20.02
Corr	0.47	0.33	0.40	0.46	0.52	0.63
Conditional mean model						
Constant	0.00	-0.00	0.00	0.00	0.00	0.00
AR(1)	-0.05	-0.13	-0.08	-0.06	-0.03	0.01

Tests for skewness, kurtosis, and correlation

	# of rejections
$H_0 : Skew\left[r_{it}\right] = 0$	3 out of 104
$H_0 : Kurt\left[r_{it}\right] = 3$	104 out of 104
$H_0 : Corr\left[r_{it}, r_{jt}\right] = 0$	5356 out of 5356

Volatility and correlation models

We also consider a GJR-GARCH/DCC model (details in paper)

	Mean	5%	25%	Median	75%	95%
	Cross-sectional distribution					
Variance model						
Constant $\phi_i^{(const)}$	-0.00	-0.08	-0.04	-0.01	0.02	0.10
HAR day $\phi_i^{(day)}$	0.38	0.32	0.35	0.38	0.40	0.44
HAR week $\phi_i^{(week)}$	0.31	0.23	0.28	0.31	0.35	0.39
HAR month $\phi_i^{(mth)}$	0.22	0.16	0.20	0.21	0.24	0.30

Correlation model

	Est	Std Err
HAR day (a)	0.12	0.01
HAR week (b)	0.32	0.02
HAR month (c)	0.38	0.03

Marginal distribution models

	Cross-sectional distribution					
	Mean	5%	25%	Median	75%	95%
HAR standardized residuals						
Mean	0.00	-0.01	-0.00	0.00	0.01	0.02
Std dev	1.09	0.96	1.02	1.08	1.14	1.29
Skewness	-0.16	-1.58	-0.47	-0.08	0.34	0.72
Kurtosis	13.12	5.06	6.84	9.87	16.03	32.72
Correlation	0.00	-0.04	-0.02	0.00	0.02	0.05
Marginal t distribution parameter estimates						
HAR	5.30	4.12	4.75	5.12	5.87	6.88

Tests for skewness, kurtosis, and correlation

	# of rejections	
	HAR	DCC
$H_0 : Skew\left[e_{it}\right] = 0$	4 out of 104	6 out of 104
$H_0 : Kurt\left[e_{it}\right] = 3$	104 out of 104	104 out of 104
$H_0 : Corr\left[e_{it}, e_{jt}\right] = 0$	497 out of 5356	1 out of 5356

Jointly symmetric copula models

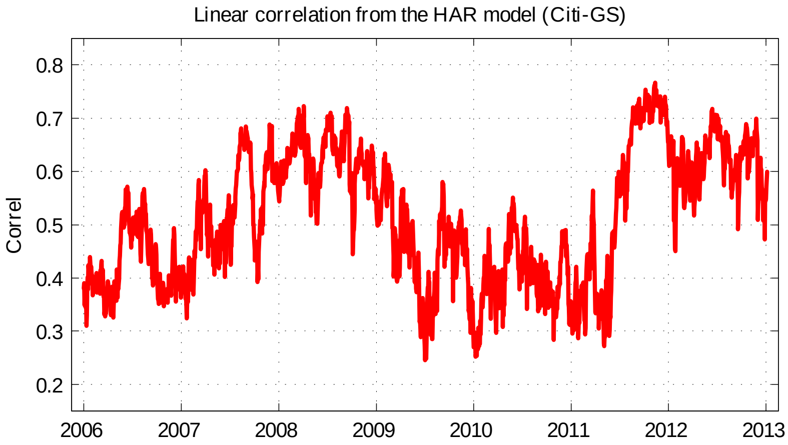
- We consider three classes of models for the standardized residuals $(\mathbf{e}_t)$:
  - **Jointly symmetric copula** models (Clayton, Gumbel, Frank and t) combined with N Student's t distributions for the marginals
  - The **independence** copula model, with N Student's t dist'ns for the marginals
  - A **jointly symmetric multivariate** t distribution
- The first two are copula-based approaches, allowing for separate specification of the marginals and copula
- The third corresponds to existing "best practice" for this problem
  - We do not even bother considering the MV Normal distribution...

Jointly symmetric copula model results

		Jointly symmetric copula models				Benchmarks	
		t	Clayton	Frank	Gumbel	Indep	MV t dist
HAR	Est.	39.44	0.09	1.27	1.02	-	$6.43^{\dagger}$
	s.e.	4.35	0.01	0.09	0.01	-	0.14
t-test of indep		8.45	10.07	13.43	5.25	-	45.72
Rank of LogL		*1*	*2*	*3*	*4*	*5*	*6*
DCC	Est.	28.21	0.11	1.60	1.03	-	$7.10^{\dagger}$
	s.e.	5.50	0.02	0.15	0.01	-	0.36
t-test of indep		6.13	7.36	10.36	4.40	-	17.80
Rank of LogL		*7*	*8*	*9*	*10*	*11*	*12*

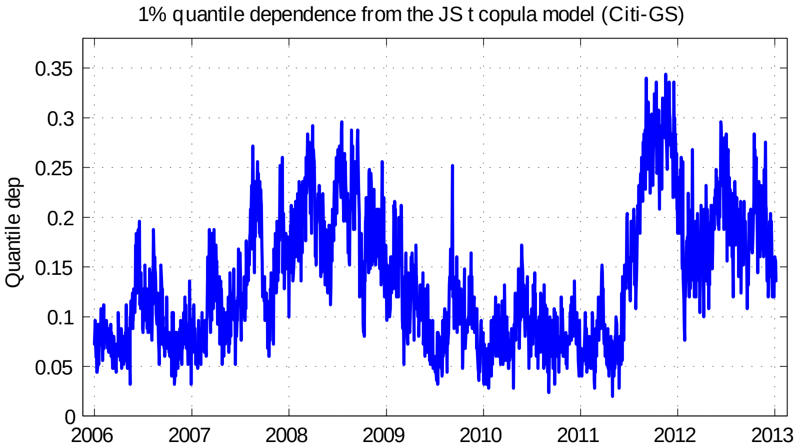
Linear correlation from the HAR model (Citi-GS)

Correlation varies from 0.25 to 0.75 over the sample period



Quantile dependence (1%) from the JS t model (Citi-GS)

Quant dep(q) = C(q,q)/q. Ranges from 0.03 to 0.35



Model comparison tests

- We use the composite likelihood KLIC (cKLIC) to compare these models:

$$\begin{aligned} & H_0 &:& \quad E\left[CL_t^A - CL_t^B\right] = 0 \\ \text{vs.} \quad & H_1 &:& \quad E\left[CL_t^A - CL_t^B\right] > 0 \\ & H_2 &:& \quad E\left[CL_t^A - CL_t^B\right] < 0 \end{aligned}$$

- Rivers and Vuong (2002) provide a method for testing this null (in-sample) for the non-nested models
- We use Giacomini and White (2006) to test the out-of-sample analogue of this null
  - OOS comparisons involve a penalty for **excess parameters**
  - We use a rolling window estimation scheme, with the last two years as the OOS period

In-sample model comparison t-statistics: HAR vs HAR

A positive value indicates the column model beats the row model

	t^{JS}	Clayton^{JS}	Frank^{JS}	Gumbel^{JS}	Indep	MV t
t^{JS}	–					
Clayton^{JS}	**2.92**	–				
Frank^{JS}	**2.16**	1.21	–			
Gumbel^{JS}	**5.38**	6.02	1.75	–		
Indep*	**8.45**	10.07	13.43	5.25	–	
MV t	**19.70**†	**19.52**	**19.45**	**19.23**	**18.40**‡	–

- The jointly symmetric t copula model significantly beats all competitors
- The multivariate t distribution is beaten by all competitors

In-sample model comparison t-statistics: DCC vs DCC

A positive value indicates the column model beats the row model

	t^{JS}	Clayton^{JS}	Frank^{JS}	Gumbel^{JS}	Indep	MV t
t^{JS}	–					
Clayton^{JS}	**4.48**	–				
Frank^{JS}	**2.69**	1.27	–			
Gumbel^{JS}	**6.74**	7.47	1.74	–		
Indep*	**6.13**	7.36	10.36	4.40	–	
MV t	**18.50**†	**18.11**	**17.94**	**17.60**	**15.69**‡	–

- The jointly symmetric t copula model significantly beats all competitors
- The multivariate t distribution is beaten by all competitors

In-sample model comparison t-statistics: HAR vs DCC

A positive value indicates the column model beats the row model

		HAR models					
		t^{JS}	Clayton^{JS}	Frank^{JS}	Gumbel^{JS}	Indep	MV t
	t^{JS}	**7.86**	7.85	7.85	7.84	7.82	**6.92**
	Clayton^{JS}	7.86	**7.86**	7.85	7.85	7.83	6.93
DCC	Frank^{JS}	7.85	7.85	**7.84**	7.83	7.82	6.91
models	Gumbel^{JS}	7.88	7.87	7.87	**7.86**	7.84	6.94
	Indep*	7.90	7.90	7.90	7.89	**7.87**	6.97
	MV t	8.95	8.95	8.94	8.94	8.92	**8.03**

- HAR models beat DCC equivalents for all choices of copula model
- Even the worst HAR model significantly beats the best DCC model

Out-of-sample model comparison t-statistics: HAR vs HAR

A positive value indicates the column model beats the row model

	t^{JS}	Clayton^{JS}	Frank^{JS}	Gumbel^{JS}	Indep	MV t
t^{JS}	–					
Clayton^{JS}	**1.50**	–				
Frank^{JS}	**0.89**	**0.44**	–			
Gumbel^{JS}	**2.88**	3.09	1.21	–		
Indep	**2.57**	2.60	2.34	1.84	–	
MV t	**10.75**	**10.63**	**10.65**	**10.48**	**10.00**	–

- The jointly symmetric t, Clayton and Frank copula models are signif better than all others, but not signif diff from each other
- The multivariate t distribution is still beaten by all competitors

Out-of-sample model comparison t-statistics: DCC vs DCC

A positive value indicates the column model beats the row model

	t^{JS}	Clayton^{JS}	Frank^{JS}	Gumbel^{JS}	Indep	MV t
t^{JS}	–					
Clayton^{JS}	**1.55**	–				
Frank^{JS}	**1.79**	**1.34**	–			
Gumbel^{JS}	**2.96**	3.31	0.01	–		
Indep	**3.10**	3.12	2.38	2.44	–	
MV t	**14.65**	**14.33**	**14.56**	**13.88**	**12.80**	–

- The jointly symmetric t, Clayton and Frank copula models are signif better than all others, but not signif diff from each other
- The multivariate t distribution is still beaten by all competitors

Out-of-sample model comparison t-statistics: HAR vs DCC

A positive value indicates the column model beats the row model

		HAR models					
		t^{JS}	Clayton^{JS}	Frank^{JS}	Gumbel^{JS}	Indep	MV t
	t^{JS}	**5.23**	5.23	5.23	5.23	5.22	**4.55**
	Clayton^{JS}	5.23	**5.23**	5.23	5.23	5.22	4.55
DCC models	Frank^{JS}	5.23	5.22	**5.23**	5.22	5.21	4.55
	Gumbel^{JS}	5.24	5.24	5.24	**5.23**	5.22	4.56
	Indep	5.24	5.24	5.24	5.23	**5.22**	4.56
	MV t	6.05	6.05	6.05	6.05	6.04	**5.41**

- HAR models beat DCC equivalents for all choices of copula model
- Even the worst HAR model significantly beats the best DCC model

Summary and conclusion

- We propose a new class of **dynamic**, **high-dimensional** distribution models
  - We exploit **high frequency data** to accurately measure and model linear dependence (correlation)
  - We use a new class of **jointly symmetric copulas** to capture any remaining nonlinear dependence
  - We consider **composite likelihood** estimation and model comparison to overcome the computational burden of estimating our JS copulas
- In an application to daily returns on 104 US equities, we find:
  - Significant gains to using high frequency data for estimating linear dependence
  - Significant gains from capturing the remaining nonlinear dependence using a jointly symmetric copula
  - Both of the above conclusions hold both **in- and out-of-sample**

Appendix

Related literature: dynamic high dimension distributions

- Use copula model to capture entire dependence structure
  - Patton (2006), Rodriguez (2007), Hafner and Manner (2010), Creal, et al. (2013), De Lira and Patton (2014), and others
- Model covariance matrix and combine with a Normal or Student's t distribution
  - Jondeau and Rockinger (2012), Hautsch et al. (2013), Jin and Maheu (2013), and others
- Combine a covariance matrix model for returns and a copula model for the uncorrelated residuals
  - This paper, and Lee and Long (2009)

Comparison with Lee and Long (2009)

- Lee and Long also suggest a linear/nonlinear decomposition:

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu}_t + \mathbf{H}_t^{1/2} \Sigma^{-1/2} \mathbf{w}_t \\ \text{where} \quad \mathbf{w}_t &\sim iid \ \mathbf{G} = \mathbf{C}_{\mathbf{w}} \left(G_1, ..., G_N \right) \end{aligned}$$

$$\mathbb{E}_{t-1} \left[w_{it} \right] = 0, \ \mathbb{E}_{t-1} \left[w_{it}^2 \right] = 1 \quad \text{and} \quad \Sigma \equiv \mathbb{E}_{t-1} \left[\mathbf{w}_t \mathbf{w}_t' \right]$$

- Key differences from our approach:
  - LL allow for **any model G**, and impose the zero correlation constraint by rotating the variables, $\mathbf{w}_t$, by their covariance matrix, Σ.
  - This step **rules out multistage estimation** of **G**, as all marginals and the coupla are needed to compute Σ
  - The covariance matrix Σ usually requires **numerical methods** for computation
  - Smaller: LL use a GARCH model for $\mathbf{H}_t$, while we exploit recent work in **high frequency** methods to estimate this

Where the bodies are buried...

- Our model:

$$\begin{aligned} \mathbf{r}_t &= \boldsymbol{\mu}_t + \mathbf{H}_t^{1/2}\mathbf{e}_t \\ \text{where } \mathbf{e}_t &\sim iid\ \mathbf{F} = \mathbf{C}^{JS}(F_1, ..., F_N) \end{aligned}$$

- All components of this model are parametric: covariance, marginals, copula
  - X All are thus subject to model misspecification
  - ✓ In high dimension applications some parametric structure is needed
- Residuals $\mathbf{e}_t$ are *iid* $\Rightarrow$ all dynamics in this model come from $\mathbf{H}_t$ (and $\boldsymbol{\mu}_t$)
  - X Rules out separate variation in higher-order moments or other dep measures
  - ✓ Second-moment variation is easily most prominent in financial data
- Joint symmetry assumption implies returns are conditionally symmetric
  - X Will not be plausible in some applications
  - ✓ Can use Lee-Long method if needed (computationally difficult)