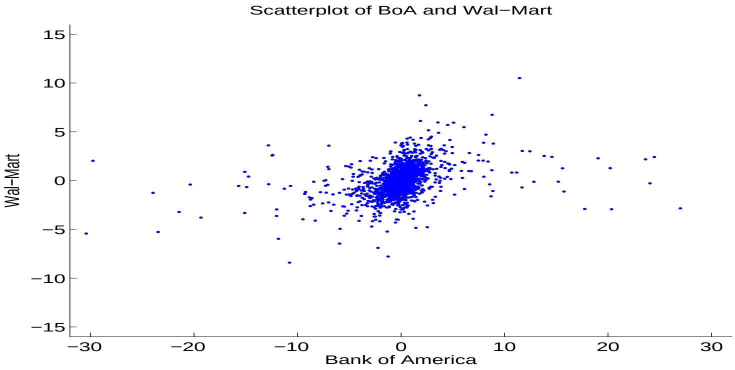
ALRIGHT: Asymmetric LaRge-Scale (I)GARCH with Hetero-Tails

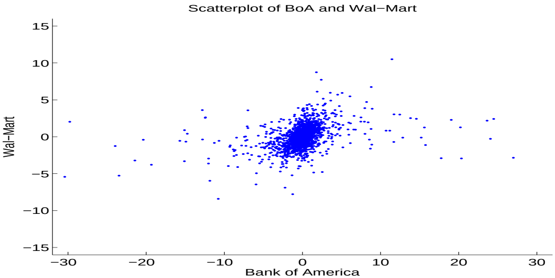
Marc S. Paolella

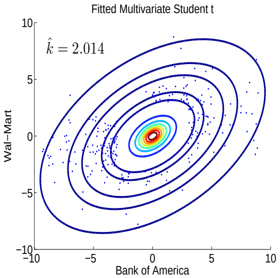
Swiss Banking Institute, University of Zürich

Do Asset Returns Have Different Tail Indices?

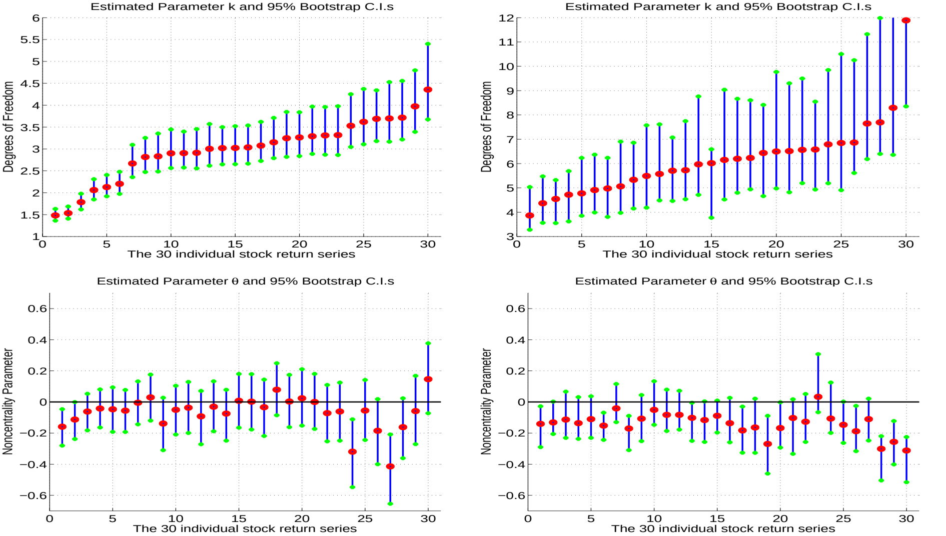


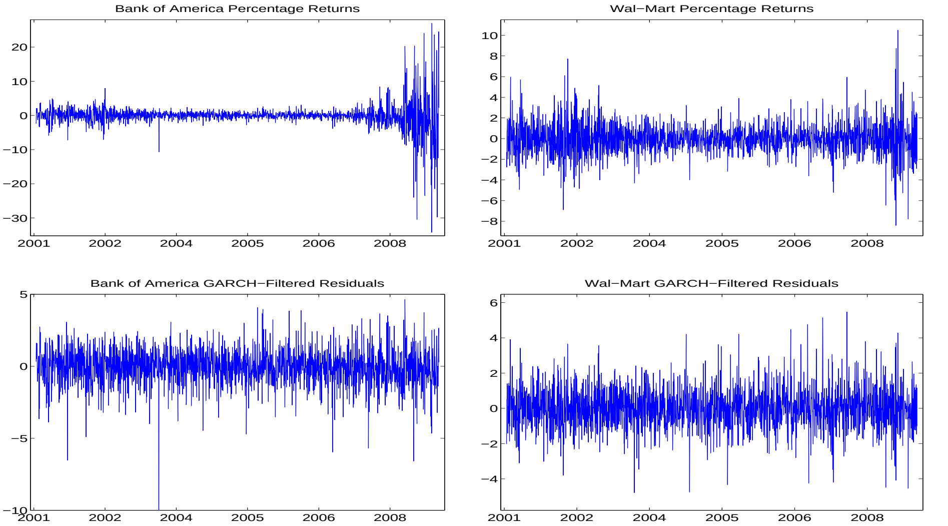
Asset Returns Have Different Tail Indices





Asset Returns Have Different Tail Indices




Bank of America Percentage Returns
Wal−Mart Percentage Returns
Bank of America GARCH−Filtered Residuals
Wal−Mart GARCH−Filtered Residuals
2001
2002
2004
2005
2006
2008

Meta-Elliptical t Distribution

- The pdf of the **meta-elliptical** t distribution is given by

$$f_{\mathbf{X}}(\mathbf{x}; \mathbf{k}, \mathbf{R}) = \psi\big(\Phi_{k_0}^{-1}(\Phi_{k_1}(x_1)), \ldots, \Phi_{k_0}^{-1}(\Phi_{k_d}(x_d)); \mathbf{R}, k_0\big) \prod_{i=1}^{d} \phi_{k_i}(x_i), \tag{1}$$

  where
  $\mathbf{x} = (x_1, \ldots, x_d)' \in \mathbb{R}^d$;
  $\mathbf{k} = (k_0, k_1, \ldots, k_d)' \in \mathbb{R}_{>0}^{d+1}$;
  $\phi_k(x)$ and $\Phi_k(x)$ denote, respectively, the univariate Student's t pdf and cumulative distribution function (cdf) with k degrees of freedom, evaluated at $x \in \mathbb{R}$;
  $\mathbf{R}$ is a d-dimensional correlation matrix, ...

Meta-Elliptical t Distribution

and, with $\mathbf{z} = (z_1, z_2, \ldots, z_d)' \in \mathbb{R}^d$, the copula density function $\psi(\cdot;\cdot) = \psi\left(z_1, z_2, \ldots, z_d; \mathbf{R}, k\right)$ multiplicatively relating the joint distribution of $\mathbf{X}$ to their distribution under independence is given by

$$\psi(\cdot;\cdot) = \frac{\Gamma\{(k+d)/2\}\{\Gamma(k/2)\}^{d-1}}{\left[\Gamma\{(k+1)/2\}\right]^{d} |\mathbf{R}|^{1/2}} \left(1 + \frac{\mathbf{z}'\mathbf{R}^{-1}\mathbf{z}}{k}\right)^{-(k+d)/2} \times \prod_{i=1}^{d} \left(1 + \frac{z_i^2}{k}\right)^{(k+1)/2} .$$

FaK (Fang, Fang Kotz)

- We express a random variable $\mathbf{T}$ with location parameter $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)' \in \mathbb{R}^d$, scale terms $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)' \in \mathbb{R}^d_{>0}$, and correlation matrix $\mathbf{R}$, as $\mathbf{T} \sim \mathrm{FaK}(\mathbf{k}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{R})$, with FaK a reminder of the involved authors, and density

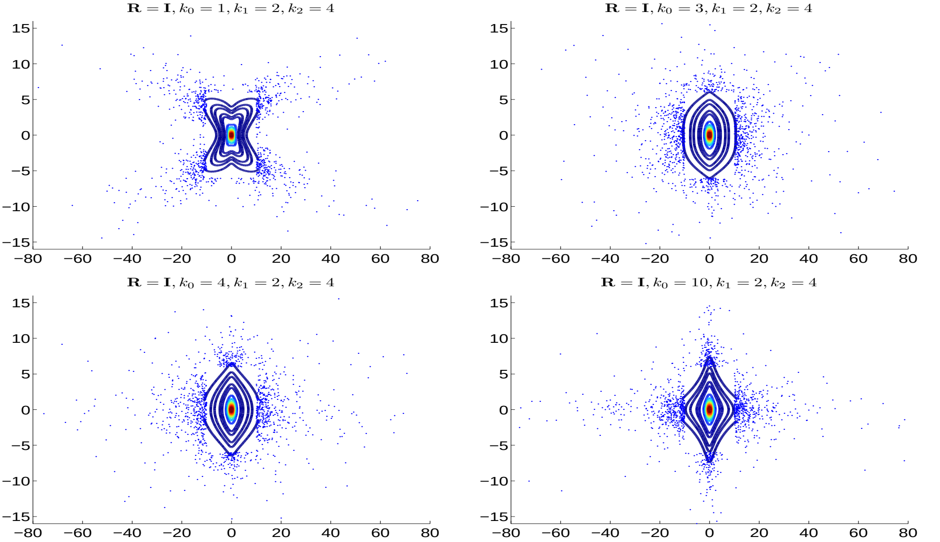
$$f_{\mathbf{T}}(\mathbf{y}; \mathbf{k}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{R}) = \frac{f_{\mathbf{X}}(\mathbf{x}; \mathbf{k}, \mathbf{R})}{\sigma_1 \sigma_2 \cdots \sigma_d}, \quad \mathbf{x} = \left(\frac{y_1 - \mu_1}{\sigma_1}, \ldots, \frac{y_d - \mu_d}{\sigma_d} \right), \tag{2}$$

  where $f_{\mathbf{X}}(\mathbf{x}; \mathbf{k}, \mathbf{R})$ is given in (1).
- From its construction as a copula, the marginal distribution of each $(T_i - \mu_i)/\sigma_i$ is a standard Student's t with k_i degrees of freedom, irrespective of k_0.
- If second moments exist for each T_i, then the variance-covariance matrix of $\mathbf{T}$ is given by $\boldsymbol{\Sigma} = \mathbb{V}(\mathbf{T}) = \mathbf{MRM}$, where $\mathbf{M} = \mathrm{diag}(\boldsymbol{\sigma} \odot \boldsymbol{\kappa})$, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_d)'$, and $\kappa_i = \sqrt{k_i/(k_i - 2)}$, $i = 1, \ldots, d$. In particular, $\mathbb{E}[T_i] = \mu_i$ and $\mathbb{V}(T_i) = \sigma_i^2 \kappa_i^2$.

FaK Parameter k_0

- While the marginals are not influenced by k_0, its value does alter the dependency structure of the distribution.
- Via comparison with scatterplots of actual financial returns data, one might speculate that only values of $k_0 \geq \max_i k_i$, $i = 1, \ldots, d$, are of interest, and one could entertain just setting $k_0 = \max_i k_i$.
- In the empirical comparison, we indeed find that $\hat{k}_0$ is very close to $\max(\hat{k}_1, \hat{k}_2)$ when it is freely estimated jointly with all other model parameters; and its attained maximum log-likelihood is statistically indistinguishable from that of the model which imposes the restriction $k_0 = \max_i k_i$.

Effect of Parameter k_0



FaK with Asymmetric Marginals: AFaK

- Introduce noncentrality parameters $\theta_i \in \mathbb{R}$, $i = 1, 2, \ldots, d$, so that, with $\phi_{k,\theta}(x)$ and $\Phi_{k,\theta}(x)$ the pdf and cdf of the noncentral t distribution at $x \in \mathbb{R}$, $f_{\mathbf{X}}(\mathbf{x}; \mathbf{k}, \mathbf{R}, \boldsymbol{\theta})$ is

$$\psi\left(\Phi^{-1}_{k_0,\theta_0}(\Phi_{k_1,\theta_1}(x_1)), \ldots, \Phi^{-1}_{k_0,\theta_0}(\Phi_{k_d,\theta_d}(x_d)); \mathbf{R}, k_0\right) \prod_{i=1}^{d} \phi_{k_i,\theta_i}(x_i),$$

  still in conjunction with (2), and with $\theta_0 = 0$.

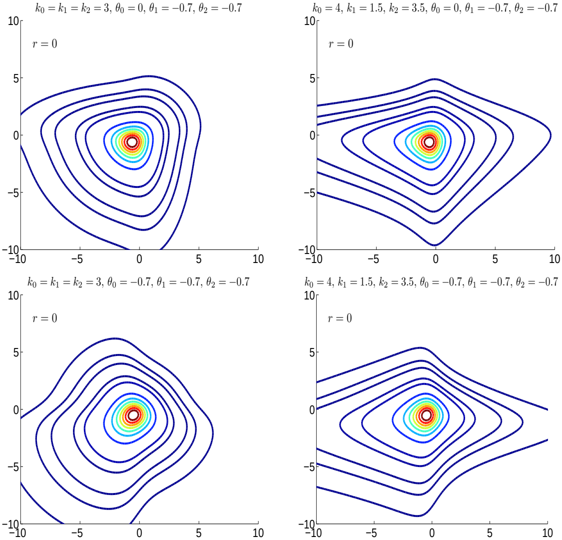
- The location-scale variant $f_{\mathbf{T}}(\mathbf{y}; \mathbf{k}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{R}, \boldsymbol{\theta})$ is analogous to (2), and we write $\mathbf{T} \sim \mathrm{AFaK}(\mathbf{k}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{R}, \boldsymbol{\theta})$, for **asymmetric FaK**.

- We have $\mathbb{V}(\mathbf{T}) = \mathbf{MRM}$, where $\mathbf{M} = \mathrm{diag}(\boldsymbol{\sigma} \odot \mathbf{v}^{1/2})$, where $\mathbf{v} = (\mathbb{V}(S_1), \ldots, \mathbb{V}(S_d))'$, for $S_i = (T_i - \mu_i)/\sigma_i \sim t'(k_i, \theta_i, 0, 1)$, with the variance of S_i computed from

$$\mathbb{E}[S_i] = \theta_i \left(\frac{k_i}{2}\right)^{1/2} \frac{\Gamma(k_i/2 - 1/2)}{\Gamma(k_i/2)}, \quad k_i > 1, \tag{3}$$

$\mathbb{E}[S_i^2] = [k_i/(k_i - 2)](1 + \theta_i^2)$ for $k_i > 2$, $\mathbb{V}(S) = \mathbb{E}[S^2] - (\mathbb{E}[S])^2$.

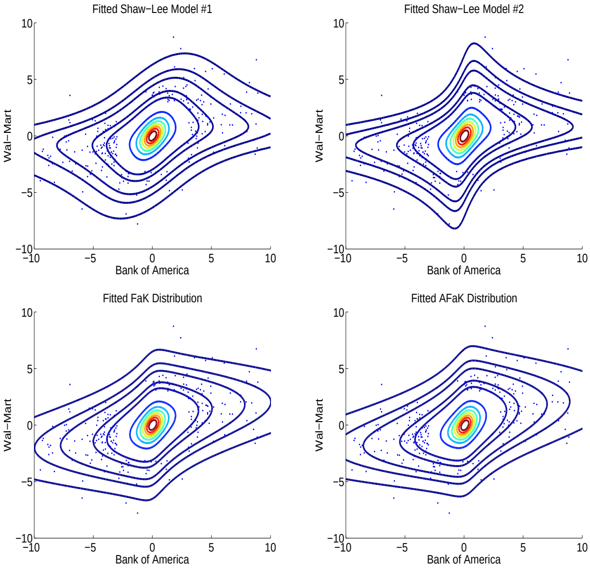
Examples of Bivariate AFaK



Bivariate Example: BoA and Wal-Mart

FaK	loglik[a]	k_0	k_1	k_2			μ_1	μ_2	scal
MLE	−7086.1	3.975	1.464	3.873			0.0331	0.0027	0.8
std err Hess		(0.497)	(0.067)	(0.344)			(0.026)	(0.028)	(0.0
std err NPB		(0.562)	(0.058)	(0.376)			(0.025)	(0.031)	(0.0
std err PB		(0.526)	(0.068)	(0.349)			(0.026)	(0.028)	(0.0
AFaK		k_0	k_1	k_2	θ_1	θ_2	μ_1	μ_2	scal
MLE	−7079.1	3.903	1.472	3.879	−0.165	0.136	0.190	−0.192	0.8
std err Hess		(0.481)	(0.068)	(0.344)	(0.055)	(0.094)	(0.057)	(0.119)	(0.0
std err NPB		(0.551)	(0.059)	(0.374)	(0.060)	(0.094)	(0.062)	(0.115)	(0.0
std err PB		(0.486)	(0.081)	(0.330)	(0.049)	(0.096)	(0.051)	(0.122)	(0.0
S-L 1			v_1	v_2			μ_1	μ_2	scal
MLE	−7092.2		1.618	3.731			0.0275	−0.0068	0.9
std err Hess			(0.074)	(0.306)			(0.027)	(0.028)	(0.0
std err NPB			(0.078)	(0.317)			(0.026)	(0.029)	(0.0
std err PB			(0.082)	(0.337)			(0.035)	(0.036)	(0.0
S-L 2			v_1	v_2			μ_1	μ_2	scal
MLE	−7142.7		1.601	4.813			0.0313	−0.0057	0.9
std err Hess			(0.077)	(0.491)			(0.027)	(0.030)	(0.0
std err NPB			(0.072)	(0.508)			(0.027)	(0.028)	(0.0
std err PB			(0.067)	(0.498)			(0.024)	(0.024)	(0.0

Data Scatterplot and the Fitted Densities



Two-Step Unconditional Estimation

We propose the following, essentially obvious, two-step procedure:

1. The three (or four) parameters k_i, μ_i and σ_i (and θ_i) based on the univariate data set corresponding to the ith variable are estimated via maximum likelihood, $i = 1, \ldots, d$. Observe that only three (or four) parameters need to be estimated simultaneously. Set $\hat{k}_0$ to $\max_i(\hat{k}_i)$.
2. Parameter $\mathbf{R}$ is estimated as the sample correlation matrix, $\widetilde{\mathbf{R}}$, or a shrinkage-based variant of it; see below.

Remarks on: Two-Step Unconditional Estimation

1. Unlike with maximum likelihood, application of this two step procedure (in particular, the second step) only makes sense if $\min(k_i) > 2$. Have a solution... In the more realistic case that a conditional model via GARCH will be used, the conditional tail index k_i is, in all probability, larger than two.

Remarks on: Two-Step Unconditional Estimation

2. Observe that step 1 will be extremely fast in the symmetric (FaK) case, as only the usual univariate Student's t density is required for the likelihood.
   For the asymmetric case, computing the density of the noncentral t distribution at each point involves either a univariate numeric integration, or evaluation of an infinite sum, and will thus be massively slower than computing the usual Student's t distribution.
   This bottleneck can be overcome by using the second-order closed-form saddlepoint approximation to the density, which is extremely accurate (even, and especially, in the tails) and about 1200 times faster to compute.
   The derivation and relevant formulae are given in Broda and Paolella (2007) and the references therein. Crucially, *there is virtually no difference in the estimates when using either the true or the saddlepoint density.*

Remarks on: Two-Step Unconditional Estimation

3. It is well-known that shrinkage of the estimated covariance matrix in the traditional portfolio optimization setup is highly beneficial. They could be shrunk towards their mean value. We can express this algebraically as, with $a = \mathbf{1}'\big(\tilde{\mathbf{R}} - \mathbf{I}\big)\mathbf{1} / \left[d\left(d-1\right)\right]$ and $\mathbf{1}$ a d-length column of ones,

$$\widehat{\mathbf{R}} = (1 - s_c)\widetilde{\mathbf{R}} + s_c\big((1-a)\mathbf{I} + a\mathbf{1}\mathbf{1}'\big). \tag{4}$$

Remarks on: Two-Step Unconditional Estimation

4. One might consider robust estimation of the covariance matrix, say $\widehat{\boldsymbol{\Sigma}}$, from which $\widehat{\mathbf{R}} = \widehat{\mathbf{D}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}}^{-1}$ can be computed, where $\mathbf{D} = \mathrm{diag}(\boldsymbol{\sigma})$, and the scale terms σ_i are estimated in step one.

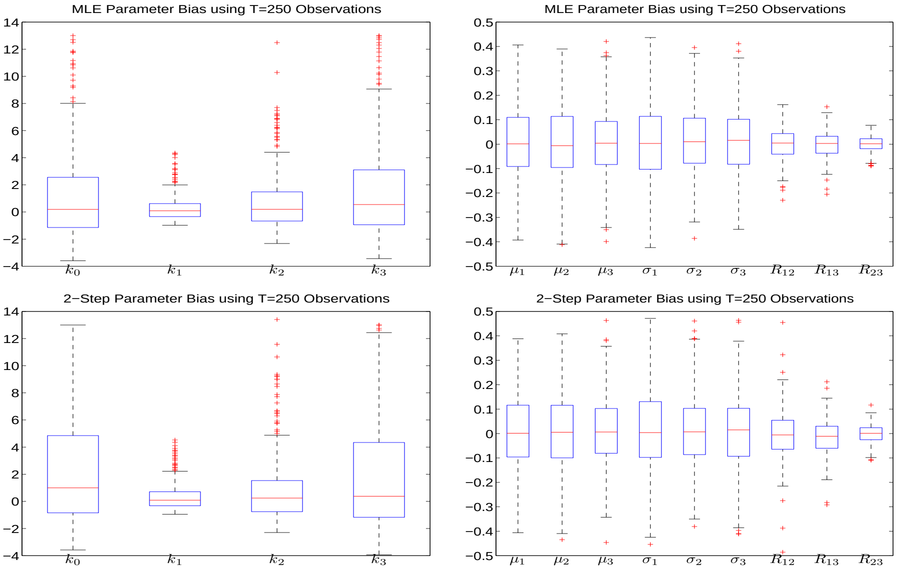
Simulation to Assess Quality: FaK

- Consider the tri-dimensional FaK distribution with parameters

$$\begin{aligned} &k_1 = 3,\ k_2 = 5,\ k_3 = 7, \quad k_0 = \max(k_i) = 7, \\ &\mu_1 = 0.2,\ \mu_2 = 0,\ \mu_3 = -0.2, \\ &\sigma_1 = \sigma_2 = \sigma_3 = 2, \quad R_{12} = 0.25, \quad R_{13} = 0.5, \quad R_{23} = 0.75, \end{aligned}$$

  (and $\boldsymbol{\theta} = \mathbf{0}$).
- We assess, via simulation, the differences in the quality (bias and spread) of the estimated parameters when using joint maximum likelihood and the two-step procedure.
- This is conducted for the sample size $T = 250$, and based on 500 replications.

Simulation to Assess Quality: FaK



Simulation to Assess Quality: FaK

- The average time for joint parameter estimation of this FaK model (using a 3GHz PC, Matlab) is 34 seconds. The two-step method requires 0.050 seconds.
- Observe that, by design, the required estimation time for the two-step method increases linearly in d, but will increase exponentially in d for the joint parameter estimation.
- Furthermore, as the number of parameters to be simultaneously estimated increases, the problems associated with avoiding inferior local maxima of the log-likelihood become exacerbated.

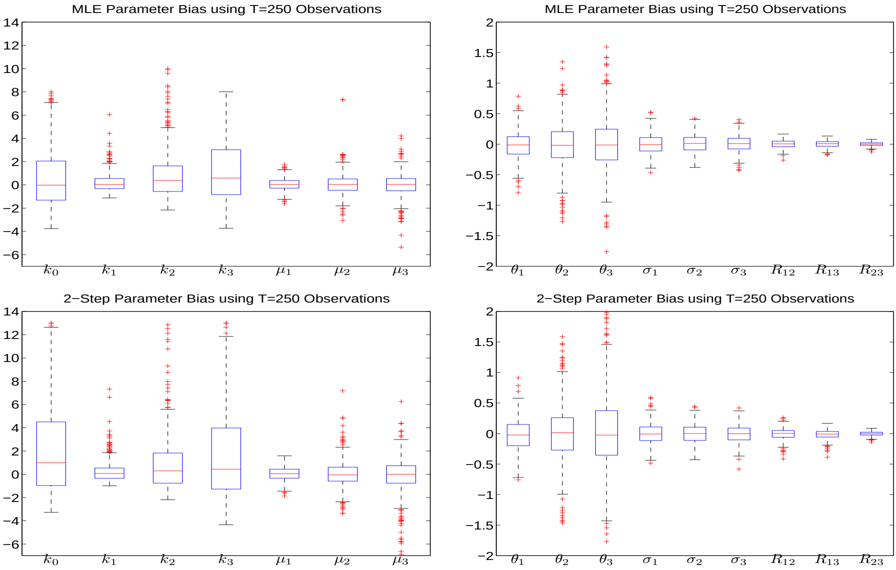
Simulation to Assess Quality: AFaK

We use the tri-dimensional AFaK distribution with the parameters as given above, but additionally take the noncentrality parameters to be $\theta_1 = -0.2$, $\theta_2 = 0$, $\theta_3 = 0.2$.

A distinction can be seen for parameters k_3, θ_3 and μ_3, for which the joint MLE does indeed perform noticeably better, albeit not demonstrably so.

With regard to estimation time, using the same computing platform mentioned above, joint maximum likelihood (AFaK for $d = 3$ and $T = 250$) takes, on average, 14.0 minutes, while the two-step procedure, using the saddlepoint approximation, takes on average 0.82 seconds, i.e., it is over 1,000 times faster.

Simulation to Assess Quality: FaK



CCC-GARCH

- We extend the model to CCC-GARCH. The 2-step procedure applies.
- Each marginal distribution is a (noncentral) Student's t with its own degree of freedom (and asymmetry parameter).
- They are linked via the t-copula as the (A)FaK distribution, but such that each univariate time series is endowed with a time-varying scale term via a t'-(I)GARCH model.
- The correlation matrix is estimated from the multivariate set of t'-(I)GARCH residuals and is not time-varying.
- We will refer to this as the (A)FaK-(I-)CCC model.

Density Forecasting

- Good in-sample fit is nice... Good simulation results are good... but what counts is the ability to forecast.
- We forecast the entire multivariate density.
- The measure of interest is what we will call the **(realized) predictive log-likelihood**, given by

$$\pi_t(\mathcal{M}, v) = \log f^{\mathcal{M}}_{t|I_{t-1}}(\mathbf{y}_t; \widehat{\psi}), \tag{5}$$

  where v denotes the size of the rolling window used to determine I_{t-1} (and the set of observations used for estimation of ψ) for each time point t.

Density Forecasting

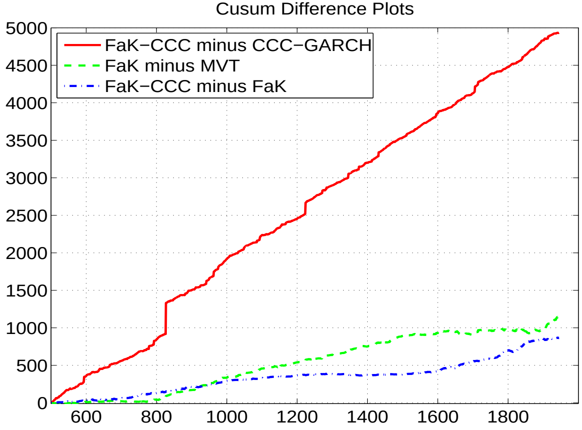
- We suggest to use what we refer to as the **normalized sum of the realized predictive log-likelihood**, given by

$$S_{\tau_0,T}(\mathcal{M}, v) = \frac{1}{(T - \tau_0)\, d} \sum_{t=\tau_0+1}^{T} \pi_t(\mathcal{M}, v), \qquad (6)$$

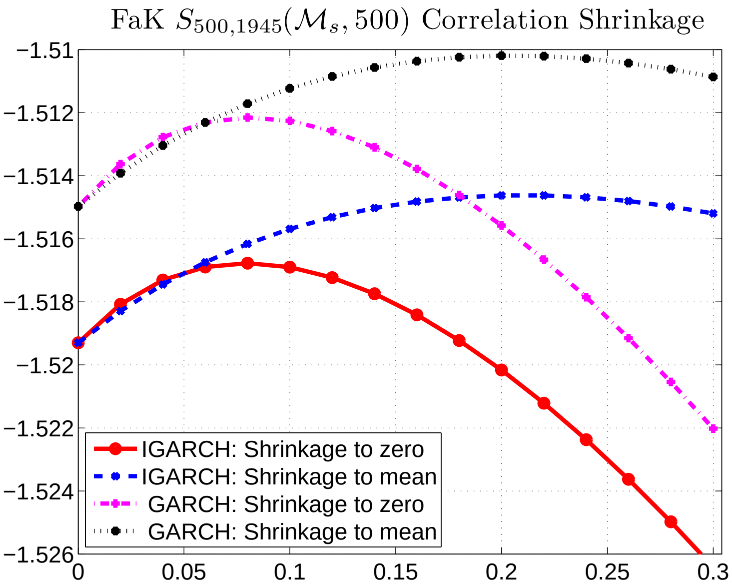
  where d is the dimension of the data.
- It is thus the average realized predictive log-likelihood, averaged over the number of time points used and the dimension of the random variable under study. This facilitates comparison over different d, τ_0 and T.
- In our setting, we use the $d = 30$ daily return series of the DJ-30, with $v = \tau_0 = 500$, which corresponds to two years of data, and $T = 1,945$.

Forecast Cusum Plots

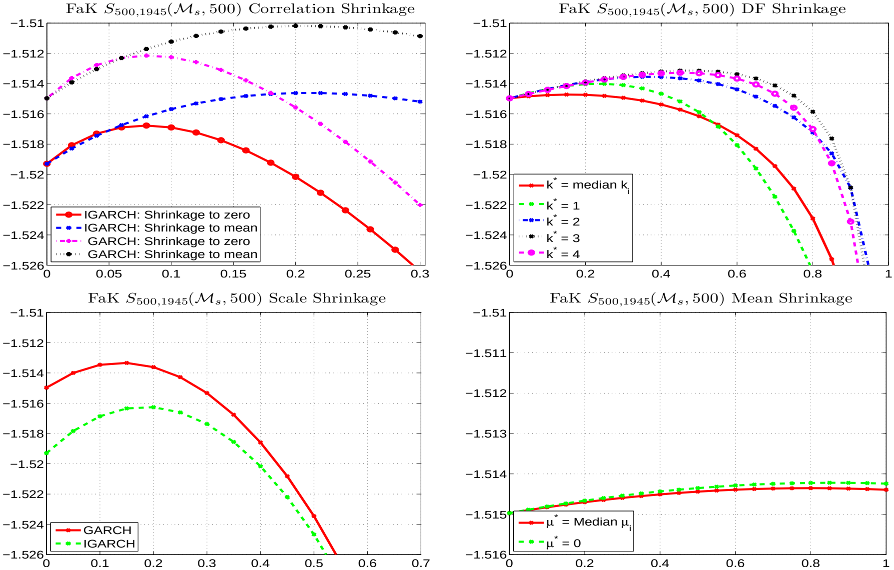
To nicely **illustrate the differences among the models** and to **contrast their sources of forecast improvement**, plot difference of the cumulative sum (cusum) of the $\pi_t(\mathcal{M}_i, 500)$, for two models i, and does so for 3 combinations of interest.



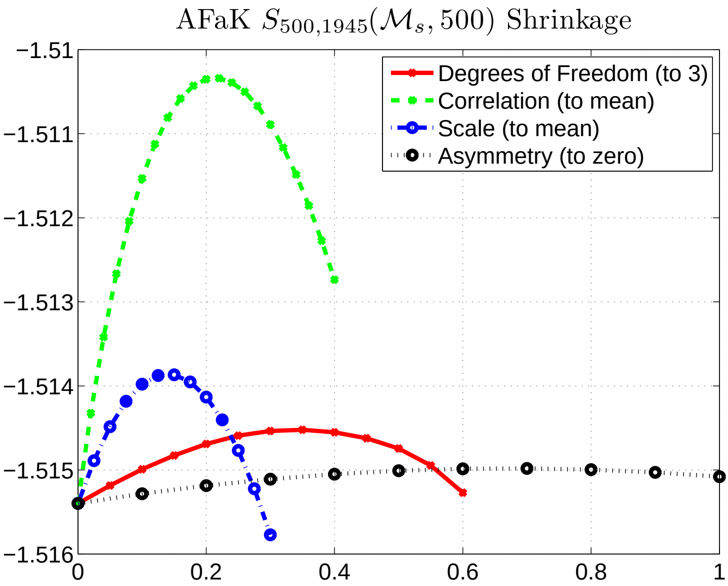
Shrinkage for the FaK-CCC Model



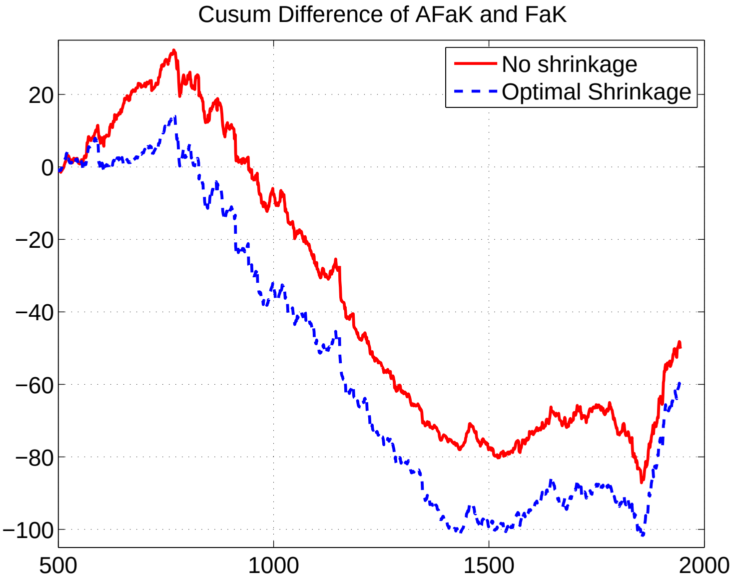
Shrinkage for the FaK-CCC Model



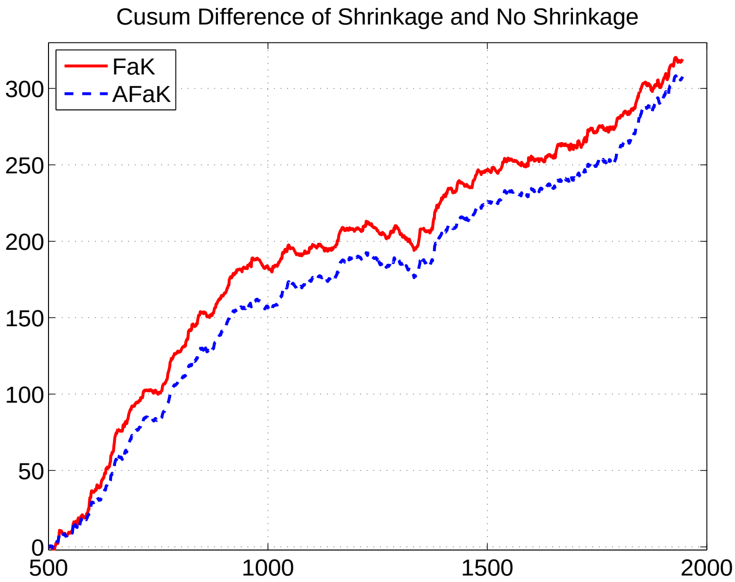
Asymmetry: Shrinkage for the AFaK-CCC Model



How Much Does Asymmetry Help? AFaK vs. FaK

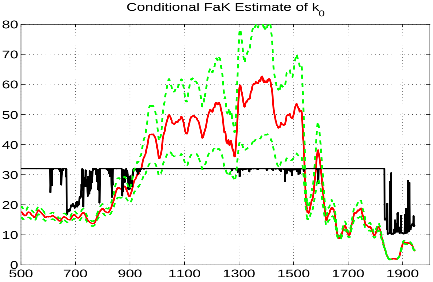


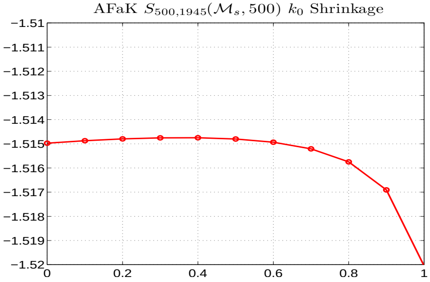
How Much Does Shrinkage Help?



Estimating (Time Varying) k_0

Now use a three-step procedure, with the final step allowing the incorporation of a time-varying copula into the model, by estimating the value of k_0, conditional on all other model parameters.





Weighted Likelihood

- The model is wrong w.p.1, but has value as a simplified filter, so use weighted likelihood to put more weight on recent observations.
- We use $w_t \propto (T - t + 1)^{\rho - 1}$, where the single parameter ρ dictates the shape of the weighting function, and the actual weights are just re-normalized such that they sum to one.
- When researchers choose a window length (usually an arbitrary multiple of 100), an implicit decision is made to weight all the observations in the window equally likely, and observations which came (right) before it receive zero weight. Such a scheme should appear rather crude and primitive!
- The procedure applied to step one helps significantly with univariate density forecasting, but not with $d = 30$ assets. However, it does help with the correlation matrix.
- The weighted correlation matrix is formed in a natural way by taking the sample means, covariances, and correlations for assets i and j as

$$m_i = T^{-1} \sum_{t=1}^{T} w_t r_{i,t}, \; v_{i,j} = T^{-1} \sum_{t=1}^{T} w_t (r_{i,t} - m_i)(r_{j,t} - m_j), \; R_{i,j} = \frac{v_{i,j}}{\sqrt{v_{i,i} v_{j,j}}},$$

Weighted Likelihood for the Correlation Matrix