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?

Scatterplot of BoA and Wal-Mart
Asset Returns Have Different Tail Indices

Scatterplot of BoA and Wal-Mart

Fitted Multivariate Student t

\( \hat{k} = 2.014 \)
Asset Returns Have Different Tail Indices

Estimated Parameter $k$ and 95% Bootstrap C.I.s

The 30 individual stock return series

Estimated Parameter $\theta$ and 95% Bootstrap C.I.s

Noncentrality Parameter
ALRIGHT: Asymmetric LaRge-Scale (I)GARCH with Hetero-Tails
The pdf of the **meta-elliptical** *t* distribution is given by

\[
f_{\mathbf{x}}(\mathbf{x}; \mathbf{k}, \mathbf{R}) = \psi\left(\Phi_{k_0}^{-1}(\Phi_{k_1}(x_1)), \ldots, \Phi_{k_0}^{-1}(\Phi_{k_d}(x_d)); \mathbf{R}, k_0\right) \prod_{i=1}^{d} \phi_{k_i}(x_i),
\]

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

\[
\psi(\cdot; \cdot) = \frac{\Gamma\{ (k + d)/2 \} \{ \Gamma(k/2) \}^{d-1}}{[\Gamma\{ (k + 1)/2 \}]^d |\mathbf{R}|^{1/2}} \left( 1 + \frac{\mathbf{z}' \mathbf{R}^{-1} \mathbf{z}}{k} \right)^{-(k+d)/2} \\
\times \prod_{i=1}^{d} \left( 1 + \frac{z_i^2}{k} \right)^{(k+1)/2}.
\]
We express a random variable $T$ with location parameter

$$
\mu = (\mu_1, \ldots, \mu_d)' \in \mathbb{R}^d,
$$

scale terms $\sigma = (\sigma_1, \ldots, \sigma_d)' \in \mathbb{R}_{>0}^d$, and correlation matrix $R$, as $T \sim \text{FaK}(k, \mu, \sigma, R)$, with FaK a reminder of the involved authors, and density

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

where $f_X(x; k, R)$ 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 $T$ is given by $\Sigma = \nabla(T) = \text{MRM}$, where $M = \text{diag}(\sigma \circ \kappa)$, $\kappa = (\kappa_1, \ldots, \kappa_d)'$, and $\kappa_i = \sqrt{k_i/(k_i - 2)}$, $i = 1, \ldots, d$. In particular, $\mathbb{E}[T_i] = \mu_i$ and $\nabla(T_i) = \sigma_i^2 \kappa_i^2$. 
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 $k_0$

\[ \mathbf{R} = \mathbf{I}, k_0 = 1, k_1 = 2, k_2 = 4 \]

\[ \mathbf{R} = \mathbf{I}, k_0 = 3, k_1 = 2, k_2 = 4 \]

\[ \mathbf{R} = \mathbf{I}, k_0 = 4, k_1 = 2, k_2 = 4 \]

\[ \mathbf{R} = \mathbf{I}, k_0 = 10, k_1 = 2, k_2 = 4 \]
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 $x \in \mathbb{R}$, $f_X(x; k, R, \theta)$ is

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

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

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

- We have $\mathbb{V}(T) = \text{MRM}$, where $M = \text{diag}(\sigma \otimes v^{1/2})$, where $v = (\mathbb{V}(S_1), \ldots, \mathbb{V}(S_d))'$, 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

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

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

\[
k_0 = k_1 = k_2 = 3, \theta_0 = 0, \theta_1 = -0.7, \theta_2 = -0.7
\]

\[
k_0 = k_1 = k_2 = 4, \theta_0 = 0, \theta_1 = -0.7, \theta_2 = -0.7
\]
<table>
<thead>
<tr>
<th>FaK</th>
<th>loglik&lt;sup&gt;a&lt;/sup&gt;</th>
<th>$k_0$</th>
<th>$k_1$</th>
<th>$k_2$</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>-7086.1</td>
<td>3.975</td>
<td>1.464</td>
<td>3.873</td>
<td>0.0331</td>
<td>0.0027</td>
<td>0.82</td>
</tr>
<tr>
<td>std err Hess</td>
<td>(0.497)</td>
<td>(0.067)</td>
<td>(0.344)</td>
<td>(0.376)</td>
<td>(0.026)</td>
<td>(0.028)</td>
<td></td>
</tr>
<tr>
<td>std err NPB</td>
<td>(0.562)</td>
<td>(0.058)</td>
<td>(0.349)</td>
<td>(0.349)</td>
<td>(0.025)</td>
<td>(0.031)</td>
<td></td>
</tr>
<tr>
<td>std err PB</td>
<td>(0.526)</td>
<td>(0.068)</td>
<td></td>
<td></td>
<td>(0.026)</td>
<td>(0.028)</td>
<td></td>
</tr>
<tr>
<td>AFaK</td>
<td></td>
<td>$k_0$</td>
<td>$k_1$</td>
<td>$k_2$</td>
<td>$\theta_1$</td>
<td>$\theta_2$</td>
<td>$\mu_1$</td>
</tr>
<tr>
<td>MLE</td>
<td>-7079.1</td>
<td>3.903</td>
<td>1.472</td>
<td>3.879</td>
<td>-0.165</td>
<td>0.136</td>
<td>0.190</td>
</tr>
<tr>
<td>std err Hess</td>
<td>(0.481)</td>
<td>(0.068)</td>
<td>(0.344)</td>
<td>(0.055)</td>
<td>(0.094)</td>
<td>(0.057)</td>
<td>(0.119)</td>
</tr>
<tr>
<td>std err NPB</td>
<td>(0.551)</td>
<td>(0.059)</td>
<td>(0.374)</td>
<td>(0.060)</td>
<td>(0.094)</td>
<td>(0.062)</td>
<td>(0.115)</td>
</tr>
<tr>
<td>std err PB</td>
<td>(0.486)</td>
<td>(0.081)</td>
<td>(0.330)</td>
<td>(0.049)</td>
<td>(0.096)</td>
<td>(0.051)</td>
<td>(0.122)</td>
</tr>
<tr>
<td>S-L 1</td>
<td></td>
<td>$\nu_1$</td>
<td>$\nu_2$</td>
<td></td>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
<td>scale</td>
</tr>
<tr>
<td>MLE</td>
<td>-7092.2</td>
<td>1.618</td>
<td>3.731</td>
<td></td>
<td>0.0275</td>
<td>-0.0068</td>
<td>0.90</td>
</tr>
<tr>
<td>std err Hess</td>
<td>(0.074)</td>
<td>(0.306)</td>
<td>(0.317)</td>
<td></td>
<td>(0.027)</td>
<td>(0.028)</td>
<td></td>
</tr>
<tr>
<td>std err NPB</td>
<td>(0.078)</td>
<td>(0.317)</td>
<td>(0.337)</td>
<td></td>
<td>(0.026)</td>
<td>(0.029)</td>
<td></td>
</tr>
<tr>
<td>std err PB</td>
<td>(0.082)</td>
<td></td>
<td></td>
<td></td>
<td>(0.035)</td>
<td>(0.036)</td>
<td></td>
</tr>
<tr>
<td>S-L 2</td>
<td></td>
<td>$\nu_1$</td>
<td>$\nu_2$</td>
<td></td>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
<td>scale</td>
</tr>
<tr>
<td>MLE</td>
<td>-7142.7</td>
<td>1.601</td>
<td>4.813</td>
<td></td>
<td>0.0313</td>
<td>-0.0057</td>
<td>0.90</td>
</tr>
<tr>
<td>std err Hess</td>
<td>(0.077)</td>
<td>(0.491)</td>
<td>(0.508)</td>
<td></td>
<td>(0.027)</td>
<td>(0.030)</td>
<td></td>
</tr>
<tr>
<td>std err NPB</td>
<td>(0.072)</td>
<td>(0.508)</td>
<td>(0.498)</td>
<td></td>
<td>(0.027)</td>
<td>(0.028)</td>
<td></td>
</tr>
<tr>
<td>std err PB</td>
<td>(0.067)</td>
<td></td>
<td></td>
<td></td>
<td>(0.024)</td>
<td>(0.024)</td>
<td></td>
</tr>
</tbody>
</table>
Data Scatterplot and the Fitted Densities

Fitted Shaw–Lee Model #1

Bank of America
Wal−Mart

Fitted Shaw–Lee Model #2

Bank of America
Wal−Mart

Fitted FaK Distribution

Bank of America
Wal−Mart

Fitted AFaK Distribution

Bank of America
Wal−Mart

ALRIGHT: Asymmetric LaRge-Scale (I)GARCH with Hetero-Tails
We propose the following, essentially obvious, two-step procedure:

1. The three (or four) parameters $k_i$, $\mu_i$ and $\sigma_i$ (and $\theta_i$) based on the univariate data set corresponding to the $i$th 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, $\tilde{\mathbf{R}}$, or a shrinkage-based variant of it; see below.
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.
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.
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 $a = 1'(\hat{R} - I)1 / [d(d - 1)]$ and $1$ a $d$-length column of ones,

$$\hat{R} = (1 - s_c)\hat{R} + s_c((1 - a)I + a11').$$  \hspace{1cm} (4)
4. One might consider robust estimation of the covariance matrix, say \( \hat{\Sigma} \), from which \( \hat{R} = \hat{D}^{-1} \hat{\Sigma} \hat{D}^{-1} \) can be computed, where \( \hat{D} = \text{diag}(\hat{\sigma}) \), and the scale terms \( \sigma_i \) are estimated in step one.
Consider the tri-dimensional FaK distribution with parameters

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

(and \( \theta = 0 \)).

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

MLE Parameter Bias using $T=250$ Observations

2-Step Parameter Bias using $T=250$ Observations

Marc S. Paolella

ALRIGHT: Asymmetric LaRge-Scale (I)GARCH with Hetero-Tails
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.
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 $k_3$, $\theta_3$ and $\mu_3$, 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

MLE Parameter Bias using T=250 Observations

2–Step Parameter Bias using T=250 Observations

Marc S. Paolella

ALRIGHT: Asymmetric LaRge-Scale (I)GARCH with Hetero-Tails
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.
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 \textbf{(realized) predictive log-likelihood}, given by

$$\pi_t(M, \nu) = \log f_{t|t-1}(y_t; \hat{\psi}), \quad (5)$$

where $\nu$ denotes the size of the rolling window used to determine $I_{t-1}$ (and the set of observations used for estimation of $\psi$) for each time point $t$. 
We suggest to use what we refer to as the normalized sum of the realized predictive log-likelihood, given by

\[ S_{\tau_0, T}(\mathcal{M}, \nu) = \frac{1}{(T - \tau_0) d} \sum_{t=\tau_0+1}^{T} \pi_t(\mathcal{M}, \nu), \quad (6) \]

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 \( d \), \( \tau_0 \) and \( T \).

In our setting, we use the \( d = 30 \) daily return series of the DJ-30, with \( \nu = \tau_0 = 500 \), which corresponds to two years of data, and \( T = 1,945 \).
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

FaK $S_{500,1945}(M_s, 500)$ Correlation Shrinkage

IGARCH: Shrinkage to zero
IGARCH: Shrinkage to mean
GARCH: Shrinkage to zero
GARCH: Shrinkage to mean

Marc S. Paolella

ALRIGHT: Asymmetric LaRge-Scale (I)GARCH with Hetero-Tails
Shrinkage for the FaK-CCC Model

FaK $S_{500,1945}(M_s, 500)$ Correlation Shrinkage

FaK $S_{500,1945}(M_s, 500)$ DF Shrinkage

FaK $S_{500,1945}(M_s, 500)$ Scale Shrinkage

FaK $S_{500,1945}(M_s, 500)$ Mean Shrinkage
Asymmetry: Shrinkage for the AFaK-CCC Model

AFaK $S_{500,1945}(M_s, 500)$ Shrinkage

- Degrees of Freedom (to 3)
- Correlation (to mean)
- Scale (to mean)
- Asymmetry (to zero)

Marc S. Paolella

ALRIGHT: Asymmetric LaRge-Scale (I)GARCH with Hetero-Tails
How Much Does Asymmetry Help? AFaK vs. FaK

Cusum Difference of AFaK and FaK

No shrinkage
Optimal Shrinkage
How Much Does Shrinkage Help?

Cusum Difference of Shrinkage and No Shrinkage

FaK

AFaK

Marc S. Paolella
ALRIGHT: Asymmetric Large-Scale (I)GARCH with Hetero-Tails
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 $\rho$ 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

$$m_i = T^{-1} \sum_{t=1}^{T} w_t r_{i,t}, \quad v_{i,j} = T^{-1} \sum_{t=1}^{T} w_t (r_{i,t} - m_i)(r_{j,t} - m_j), \quad R_{i,j} = \frac{v_{i,j}}{\sqrt{v_{i,i} v_{j,j}}}.$$
Weighted Likelihood for the Correlation Matrix

Weighted Likelihood for FaK–CCC Correlation

Marc S. Paolella
ALRIGHT: Asymmetric LaRge-Scale (I)GARCH with Hetero-Tails