TI 11-078/2/DSF22

Modeling Dynamic Volatilities and Correlations under Skewness and Fat Tails

Xin Zhang[1]
Drew Creal[2]
Siem Jan Koopman[1]
André Lucas[1,3]

[1] Faculty of Economics and Business, VU University Amsterdam, and Tinbergen Institute;
[2] University of Chicago;
[3] Duisenberg school of finance.

Modeling Dynamic Volatilities and Correlations under Skewness and Fat Tails

Xin Zhang[*]   Drew Creal[†]   Siem Jan Koopman[‡]   André Lucas[§]

May 11, 2011

Abstract

We propose a new model for dynamic volatilities and correlations of skewed and heavy-tailed data. Our model endows the Generalized Hyperbolic distribution with time-varying parameters driven by the score of the observation density function. The key novelty in our approach is the fact that the skewed and fat-tailed shape of the distribution directly affects the dynamic behavior of the time-varying parameters. It distinguishes our approach from familiar alternatives such as the generalized autoregressive conditional heteroskedasticity model and the dynamic conditional correlation model where distributional assumptions affect the likelihood but not the parameter dynamics. We present a modified expectation-maximization algorithm to estimate the model. Simulated and empirical evidence shows that the model outperforms its close competitors if skewness and kurtosis are relevant features of the data.

Keywords: Dynamic conditional correlations, Generalized Hyperbolic distributions, Observation driven models.

JEL Code: C10, C16, C22, C32.

---

[*]Department of Finance, VU University Amsterdam, and Tinbergen Institute.
[†]Booth School of Business, University of Chicago.
[‡]Department of Econometrics, VU University Amsterdam and Tinbergen Institute.
[§]Department of Finance, VU University Amsterdam, Duisenberg school of finance and Tinbergen Institute.

1   Introduction

We propose a new dynamic observation driven model for correlations and volatilities based on the class of multivariate Generalized Hyperbolic (GH) distributions. The GH distribution was introduced by Barndorff-Nielsen (1977) and further explored in Barndorff-Nielsen (1978) and Blæsild (1981). The distribution's flexible form accommodates many of the relevant features in financial time series data, such as excess kurtosis, skewness, and time-varying volatilities and correlations; see McNeil, Frey, and Embrechts (2005), Eberlein and Keller (1995), Franses and Van Dijk (2000), Engle (2002) and others.

The dynamics of the time-varying parameters in our GH distributions are driven by the scaled score of the local observation density. This is a distinguishing feature of our approach. By using the density scores, the skewed and fat-tailed nature of the observation distribution not only affects the likelihood, but also the dynamics of the volatilities and correlations. This differentiates our approach from other well-known models where the distributional assumptions affect the likelihood only, but not the parameter dynamics, e.g., the Dynamic Conditional Correlation (DCC) model of Engle (2002) with normal or Student's t distributed innovations.

The literature on time-varying parameter models for volatilities and correlations is vast. Our model follows the literature on observation driven rather than parameter driven models. For surveys of the latter in the current context, see for example Shephard (2005) and Asai and McAleer (2009). Within the observation driven class of volatility and correlation models, many multivariate extensions of the seminal generalized autoregressive conditional heteroskedasticity (GARCH) model have been proposed. Bollerslev (1990) introduced the idea of having dynamically evolving variance matrices with the individual variances specified as GARCH processes but with the corresponding conditional correlations treated as unknown constants. This specification has become known as the CCC model. Engle (2002) generalized the CCC model by introducing a simple and parsimonious observation driven mechanism for the conditional correlations. The parsimony of the DCC model combined with its time-varying full conditional correlation matrix makes the DCC model attractive to empirical researchers. This feature is retained in the new model proposed in this paper as well. Other multivariate extensions of the GARCH model include the VEC model of Engle and Kroner (1995), the BEKK model of Engle and Kroner (1995), the Orthogonal GARCH model of Alexander (1998) and Alexander

(2001), and the Generalized Orthogonal GARCH model of van der Weide (2002) and Boswijk and Van der Weide (2006).

Most of the above models were originally derived under the assumption of a (conditionally) normal distribution for the underlying data. Since then, generalizations have been proposed to accommodate alternative distributions, including the Student's t, the skewed t, and the GH distribution; see, for example, Bauwens and Laurent (2005), Fiorentini et al. (2003), Hu (2005), Mencia and Sentana (2004), Peters (2001), and Prause (1999). In all of these models, the likelihood changes but the dynamic specifications for volatilities and correlations are unaffected.

It is rather surprising that the form of the distribution should have no impact on the specification of volatility and correlation dynamics. If, for example, the distribution is leptokurtic, we expect to see large (absolute) observations from time to time. The occurrence of a large observation should not automatically be attributed to a recent increase in volatility, as is done in a standard GARCH specification. Similarly, if the data are drawn from a skewed distribution, we would expect large negative or positive observations to convey different signals about current volatility levels. Again, this would imply a link between the shape of the observation distribution and the specification of the volatility and correlation dynamics. No such direct link is present in the standard GARCH and DCC models.

Our main contribution in this paper is to provide a general model for time-varying variances and correlations in which the form of the error distribution governs the specification of volatility and correlation dynamics. For this purpose we extend the framework of Creal et al. (2008) and Creal et al. (2011) to a multivariate setting with skewed and heavy-tailed data. Creal, Koopman, and Lucas (2011) treat the special case where time series are drawn from a multivariate Student's t distribution. Nelson and Foster (1994)) and Harvey and Chakravarty (2008) give treatments of the univariate version of this model. Similar to these approaches, our model provides an automatic mechanism that limits the impact of outlying or aberrant observations on future correlations and volatilities. At an intuitive level, the new model attributes part of the sign and magnitude of each observation to the skewed and fat-tailed nature of the data generating process rather than to direct changes in volatilities or correlations.

Our results provide a full treatment of skewness and kurtosis effects on volatilities and correlations in a multivariate setting. We show that the volatility and correlation updating mechanism includes a natural asymmetry effect to allow for a different impact of negative

3

versus positive realizations. For example, if the distribution is left-skewed, large negative real-izations are more likely and should not automatically be attributed to local volatility increases. A large positive realization for a left-skewed distribution, however, is extremely unlikely unless volatilities or correlations have increased recently. Via the density score, our dynamic specifi-cation for volatilities and correlations includes an interaction between the skewness coefficient and past observations. In this way, the possibly asymmetric impact of past observations on future volatilities and correlations enters the dynamic specification in a natural way.

Parameter estimation is straightforward for our model, since the model is defined in con-ditional terms similar to the standard GARCH model and its multivariate counterparts. This implies that the likelihood function can be specified in closed analytical form and computed using a prediction error decomposition. In the literature, maximum likelihood estimation of the parameters or the GH distribution is often carried out using the Expectation-Maximization (EM) algorithm of Dempster, Laird, and Rubin (1977), see Mencia and Sentana (2004). EM estimation, however, is not straightforward for our new model due to the highly non-linear func-tions of the data that are used to drive the volatility and correlation dynamics. We show how to modify the standard EM algorithm to our specific setting to make estimation by EM feasible again. The key step is to replace the density score as a driving mechanism by a conditional density score that runs in parallel to the conditional expectations taken in the expectations step of the EM algorithm.

In a simulation experiment, we compare the performance of our new model to its direct competitors, including versions of the DCC model. We carry out simulations with different correlation dynamics and a variety of error distributions. We also consider the DCC model with GH distributed observations. Although it is not our primary focus, the DCC model with GH errors can also be regarded as a contribution of our paper to the current literature. If the true error distribution is normal, differences in performance between the different statistical models are limited. For fat-tailed error distributions, our model with the GH distribution has superior performance. If in addition the error distributions are skewed, our model performs best.

We provide an empirical illustration of the new model to investigate the volatilities and correlations between four blue-chip stocks from different industries. The sample period includes the recent financial crisis. We find that the estimated correlation dynamics differ substantially

4

between our new approach and a traditional DCC models. The new approach seems much less influenced by incidental influential observations. Accounting for the skewness and fat-tailed nature of the data, we show that volatilities for all series are relatively smaller and that the overall persistence of volatilities and correlations is generally higher.

The remainder of the paper is organized as follows. Section 2 introduces the model. Section 3 discusses some alternative model parameterizations. Section 4 extends the model for the scale rather than the covariance matrix and proposes a modified EM algorithm for parameter estimation. Section 5 provides Monte Carlo evidence on the performance of the model compared to some of its competitors. Section 6 presents the empirical illustration. Section 7 concludes.

2   The dynamic GH model

We assume our data generating process is given by

$$y_t = L_t \varepsilon_t, \qquad \Sigma_t = L_t L_t', \tag{1}$$

where $y_t, \varepsilon_t \in \mathbb{R}^k$ for $t = 1, \ldots, n$, L_t is a $k \times k$ lower triangular matrix giving rise to a time-varying $k \times k$ covariance matrix Σ_t, and ε_t follows a Generalized Hyperbolic (GH) distribution with zero mean and unit covariance matrix. The specification (1) can easily be extended to include a conditional or unconditional non-zero and possibly time-varying mean for y_t. In line with Engle (2002), we further decompose the covariance matrix Σ_t as

$$\Sigma_t = L_t L_t' = D_t R_t D_t, \tag{2}$$

with D_t a diagonal matrix containing the standard deviations of the elements in y_t, and R_t the correlation matrix of y_t.

The Generalized Hyperbolic (GH) distribution introduced by Barndorff-Nielsen (1977) is a flexible distribution that accommodates both thin and fat-tailed as well as positively and negatively-skewed distributions. We present the GH class as the normal mean-variance mixture model

$$\varepsilon_t = \mu_\varepsilon + \zeta_t T \gamma + \sqrt{\zeta_t} T z_t, \qquad z_t \sim \mathrm{N}(0, \mathrm{I}_k), \tag{3}$$

where $\zeta_t \in \mathbb{R}^+$ is a positively valued random scalar that is independent of z_t, $\mu_\varepsilon \in \mathbb{R}^k$ is the location parameter, $k \times k$ matrix TT' is the scaling matrix and $\gamma \in \mathbb{R}^k$ is the skewness parameter.

The GH class includes distributions such as the normal ($\gamma = 0$ and $\zeta_t = 1$), the (skewed) multivariate Student's t (for which ζ_t has an inverse Gamma distribution with $\gamma = 0$ for the symmetric case and $\gamma \neq 0$ for the asymmetric case), the (skewed) variance-gamma distribution (for which ζ_t has a Gamma distribution) and the Generalized Hyperbolic distribution (for which ζ_t has a Generalized Inverse Gaussian (GIG) distribution with parameters λ, χ, and ψ).

Since we assume that ε_t has zero mean and unit covariance matrix, we obtain from (3) that

$$0 = \mathrm{E}[\varepsilon_t] = \mu_\varepsilon + \mu_\zeta T\gamma \quad \Leftrightarrow \quad \mu_\varepsilon = -\mu_\zeta T\gamma, \tag{4}$$

and

$$\mathrm{I}_k = \mathrm{E}[\varepsilon_t \varepsilon_t'] = T\left(\mu_\zeta \mathrm{I} + \sigma_\zeta^2 \gamma\gamma'\right) T' \quad \Leftrightarrow \quad (T'T)^{-1} = \mu_\zeta \mathrm{I} + \sigma_\zeta^2 \gamma\gamma', \tag{5}$$

where μ_ζ and σ_ζ^2 denote the mean and variance of ζ_t, respectively. The mean and variance of ε_t exist if the mean and variance of ζ_t exist, respectively. The density of y_t for our specification of ε_t in (3) is given in the appendix.

We let the variances and correlations for y_t be time-varying by assuming that both D_t and R_t in (2) depend on a time-varying parameter f_t, such that $D_t = D(f_t)$ and $R_t = R(f_t)$. This accommodates a setting where correlations and volatilities have their own dynamics, as well as a setting where correlations and volatilities are driven by a smaller set of time-varying common factors such as in the factor GARCH literature.

We model the dynamics of f_t using the framework of Creal, Koopman, and Lucas (2008, 2011). Their updating equation for the time-varying factor f_t is given by

$$f_{t+1} = \sum_{i=0}^{p-1} A_i s_{t-i} + \sum_{j=0}^{q-1} B_j f_{t-j}, \tag{6}$$

where matrices A_i and B_j, with appropriate dimensions, depend on a static parameter vector θ, that is $A_i = A_i(\theta)$ and $B_j = B_j(\theta)$. The innovation variable in (6) is s_t and is specified as a function of current and past values of y_t and f_t. For example, in the univariate case with $f_t = D_t^2$ and normally distributed y_t, the model in (6) embeds the standard GARCH model by setting $s_t = y_t^2$. The simplicity of this choice for s_t is appealing, but it is generally hard to extend it to a natural candidate s_t in other, more complicated cases. For example, in our current setting we want (6) to account for the possibly fat-tailed and skewed nature of the GH distribution, as well as for the adopted parameterization $D(f_t)$ and $R(f_t)$.

6

Creal et al. (2008, 2011) demonstrate that a good choice for s_t in a general non-linear time series context is the scaled density score, as given by

$$s_t = S_t \nabla_t, \tag{7}$$

$$\nabla_t = \partial \ln p(y_t | \mathcal{F}_{t-1}; f_t, \theta) / \partial f_t, \tag{8}$$

where S_t is an $\mathcal{F}_{t-1}$-adapted scaling matrix, and $\mathcal{F}_t = \{y_t, \ldots, y_1\}$. By using the density score, the time-varying parameters are changed in the direction that increases the model's local fit as measured by the log-density. For our standard GH distributed ε_t in (3), we rely on the following result.

Result 1. *Let $\otimes$ denote the Kronecker product, and let $\mathrm{vec}(\cdot)$ denote the operator that stacks the columns of a matrix into a column vector. If ε_t is modeled as in (3) with zero mean and unit covariance matrix, we have*

$$\nabla_t = \Psi_t' H_t' \left(w_t (y_t \otimes y_t) - \mathrm{vec}(\tilde{\Sigma}_t) - (1 - w_t \mu_\zeta)(y_t \otimes \tilde{L}_t \gamma) \right), \tag{9}$$

where $\Psi_t = \partial \mathrm{vech}(\Sigma_t)/\partial f_t'$, $\tilde{L}_t = L_t T$, $\tilde{\Sigma}_t = \tilde{L}_t \tilde{L}_t'$, w_t is a scalar weight, and H_t is a $k \times k$ matrix. We define w_t in (A6) and H_t in (A9) in the appendix.

Our current model generalizes some of the well-known univariate and multivariate GARCH models. If ε_t is normally distributed, i.e., $\gamma = 0$ and $T = I_k$, the weight w_t reduces to $w_t = 1$ and equation (9) reduces to $\nabla_t = \Psi_t' H_t' \mathrm{vec}(y_t y_t' - \Sigma_t)$. This is the usual expression for a multivariate GARCH model for time-varying volatilities and correlations. The matrix H_t captures the relation between $\tilde{L}_t$ and Σ_t. The matrix Ψ_t is determined by the parameterization of $D(f_t)$ and $R(f_t)$ in terms of the time-varying parameter vector f_t.

There are two interesting differences between a standard multivariate GARCH model, that is driven by squares and cross-products of vector y_{t-1}, and our model, that is driven by the score function (9) of the GH distribution. The first difference is the presence of the weighting factor w_t, which is fully defined in (A6) in the Appendix. The second difference is the presence of the asymmetry term $y_t \otimes \gamma$. These differences are the result of the fat-tailedness and skewness of the distribution of y_t, respectively. For the case of a symmetric Student's t distribution, Creal et al. (2011) also obtain a weight effect but not the asymmetry term.

We learn from Appendix A.2 that the weight w_t is generally a decreasing function of $d_{x_t}^\chi$, where

$$d_{x_t}^\chi = \chi + x_t' x_t, \qquad x_t = \tilde{L}_t^{-1} y_t + \mu_\varsigma \gamma,$$

for fat-tailed distributions in the GH class, where x_t is the standardized version of the original observation y_t. As a result, the impact of lagged (cross)-products in $y_t \otimes y_t$ on future values of f_t (and thus on volatilities and correlations) is mitigated by w_t, if y_t is large in the sense that $d_{x_t}^\chi$ is large. The intuition is as follows. If y_t is drawn from a fat-tailed distribution, large values of y_t are not necessarily due to local volatility or correlation increases. Instead, large y_t's may be due to the fat-tailed nature of the distribution. The dynamics of f_t (volatilities and correlations) should therefore only partly reflect the large value of y_t. The remainder is then attributed to the fat-tailed nature of the distribution and should not affect the volatility and correlation dynamics.

The second difference in (9) is the asymmetry term. The term takes a different role than the usual leverage effect in volatility models, which captures increases in volatilities if recent returns have been negative. Such a leverage effect can still be included in (9) in the usual way. Our asymmetry term $y_t \otimes \gamma$ is due to the skewness of the distribution. If, for example, y_t is univariate and right-skewed ($\gamma > 0$), a large positive value of y_t is more likely and is not necessarily attributable to a local volatility increase. However, a large negative value of y_t should be taken as a very strong signal of a volatility increase, because large negative observations are extremely unlikely for a right-skewed distribution (unless the volatility has increased). This is precisely the effect of the asymmetry term $y_t \otimes \gamma$ in (9): for a right skew ($\gamma > 0$), the term mitigates the volatility increase if $y_t > 0$, and reinforces the volatility increase if $y_t < 0$.

Since both the shape and the parameterization of the distribution affect the dynamics of f_t, our current model is clearly different from the GARCH class of models with non-Gaussian observations. For the GARCH class of models, the non-Gaussian assumption only affects the likelihood function; it does not affect the dynamic behavior of f_t. In our framework, the distributional properties of y_t affect both the likelihood and the dynamic evolution of f_t at the same time.

Our model retains many of the convenient properties of the GARCH and DCC type models. For example, s_t is $\mathcal{F}_t$-adapted and therefore parameter estimation in model (7)–(9) is carried out in the same convenient way as in GARCH models. Indeed, our likelihood function can be

expressed in closed-form via the prediction error decomposition and the basic recursion (6). It leads to fast likelihood evaluation. The interpretation of the model is also intuitive. Depending on the choice of the scaling matrix S_t, the driver s_t can be interpreted as a local Gauss-Newton or Steepest-Ascent improvement to the likelihood at time t. The score of the observation density at time t, evaluated at the current estimate f_t of the time-varying parameter, determines in what direction f_t is best updated to improve the fit of the model. The additional lags and dynamics in (6) add further flexibility to the size and speed of these adjustments as time progresses.

We collect all static parameters of the model, such as γ, μ_ζ, σ_ζ^2, $A_1, \ldots, A_p$, $B_1, \ldots, B_q$, into the vector θ. The parameter vector θ is estimated by the method of Maximum Likelihood (ML). Inference on θ is carried out in the usual way by taking the negative inverse Hessian of the log likelihood function at the optimum as the covariance matrix of the ML estimator.

3   Model parameterizations

The GH distribution has a considerable number of parameters from which a selection cannot be identified simultaneously. In particular, χ and ψ are not separately identified; only their product $\chi\psi$ is identified. Identification can be achieved in several ways. For example, we can set $|\Sigma_t|$ to a fixed constant, say unity, such that Σ_t is normalized. Alternatively, we can simply fix χ or ψ and estimate the other parameter in an unrestricted way. In our implementation, we estimate $\kappa = (\chi\psi)^{1/2}$ and extract χ and ψ separately through the identifying assumption $\mu_\zeta = 1$. This normalization turns out to be particularly useful when estimating the GH model using the Expectation Maximization (EM) algorithm of Section 4. Given the identifying restriction $\mu_\zeta = 1$, we can obtain χ and ψ for a fixed value of κ by the equality

$$1 = \mu_\zeta = \frac{\sqrt{\chi\psi}K_{\lambda+1}\left(\sqrt{\chi\psi}\right)}{\psi K_\lambda\left(\sqrt{\chi\psi}\right)} \quad \Leftrightarrow \quad \psi = \frac{\kappa \cdot K_{\lambda+1}\left(\kappa\right)}{K_\lambda\left(\kappa\right)}, \tag{10}$$

with $\chi = \kappa^2/\psi$.

Following Creal et al. (2011), we can consider two obvious choices for the parameterization of both the diagonal matrix of variances D_t^2 and the correlation matrix R_t in (2). We can take the variances $\text{diag}(D_t^2)$ themselves or the log-variances $\ln(\text{diag}(D_t^2))$ as parameters. The advantage of taking log-variances as parameters is that the resulting variances are always positive.

9

When the variances themselves are taken as parameters, we need to impose restrictions on the coefficient matrices A_i and B_j in (6) to ensure positive variances at all times. In higher dimensional models with more lags in the updating equation (6), such restrictions become rather complicated. We therefore take log-variances as parameters.

The specification of the correlation matrix R_t is subject to the constraints that R_t is a positive definite matrix with diagonal elements equal to one, for all t. A possible parameterization of R_t is similar to the DCC model of Engle (2002). Let $Q_t = Q(f_t)$ be an auxiliary time-varying parameter matrix, and set

$$R_t = \Delta_t^{-1} Q_t \Delta_t^{-1}, \tag{11}$$

where Δ_t^2 is a diagonal matrix holding the diagonal elements of Q_t. The matrix Q_t has k redundant elements compared to the correlation matrix R_t. As a result, only $k(k-1)/2$ independent signals in ∇_t are distributed over the k^2 elements of Q_t. The details of this specification and its implication for Ψ_t in (9) are presented in Creal et al. (2011).

An alternative specification for the correlation matrix is given by the hypersphere transformation as adopted by, for example, Jaeckel and Rebonato (2000), van der Weide (2002), and Creal et al. (2011). The correlation matrix is obtained from the Choleski decomposition $R_t = X_t X_t'$ where X_t is a lower triangular $k \times k$ matrix that is constructed from a set of $k(k-1)/2$ time-varying angles ϕ_{ijt} in $[0, \pi]$ and is given by

$$X_t = \begin{pmatrix} 1 & c_{12t} & c_{13t} & \cdots & c_{1kt} \\ 0 & s_{12t} & c_{23t}s_{13t} & \cdots & c_{2kt}s_{1kt} \\ 0 & 0 & s_{23t}s_{13t} & \cdots & c_{3kt}s_{2kt}s_{1kt} \\ 0 & 0 & 0 & \cdots & c_{4kt}s_{3kt}s_{2kt}s_{1kt} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c_{k-1,kt}\prod_{\ell=1}^{k-2} s_{\ell kt} \\ 0 & 0 & 0 & \cdots & \prod_{\ell=1}^{k-1} s_{\ell kt} \end{pmatrix}, \tag{12}$$

with $c_{ijt} = \cos(\phi_{ijt})$ and $s_{ijt} = \sin(\phi_{ijt})$. For the 2-dimensional case, we have the Choleski and correlation matrices given by

$$X_t = \begin{pmatrix} 1 & \cos(\phi_{12,t}) \\ 0 & \sin(\phi_{12,t}) \end{pmatrix}, \qquad R_t = X_t' X_t = \begin{pmatrix} 1 & \cos(\phi_{12,t}) \\ \cos(\phi_{12,t}) & 1 \end{pmatrix}, \tag{13}$$

10

with the correlation given by $\cos(\phi_{12,t})$. The second column of X_t in (13) expresses a two-dimensional unit-length vector in terms of its polar rather than its Cartesian coordinates. The generalization to the k-dimensional setting is given by the kth column of X_t in (12).

The number of unknown coefficients in X_t equals the number of correlations in the matrix R_t such that there are no redundancies as in the specification (11). We collect all angles ϕ_{ijt} in the vector ϕ_t which is specified as a function of f_t. For any value of ϕ_t, the matrix $R_t = X_t'X_t$ satisfies the properties of a correlation matrix. The specification of Ψ_t in (9) when using the hypersphere parameterization of R_t is provided in Creal et al. (2011).

The definition of s_t in (6) is completed by the choice of a scaling matrix S_t. Creal et al. (2008) discuss a number of possible choices, all of which are based on the local curvature of the model density at time t via the (local) information matrix. Computing the information matrix for the general GH distribution, however, is analytically intractable. Therefore, we consider the computationally feasible alternative by setting the scaling matrix equal to the inverse information matrix for the symmetric Student's t distribution as a special member of the GH class. This information matrix is known analytically and is derived in Creal et al. (2011). This choice accommodates both the possible fat-tailed nature of the distribution and the time-variation in the volatilities and correlations. The form of scaling can be implemented efficiently and has shown to work well for both simulated and empirical data, see also the results in the following sections. Finally, this choice also makes our current model directly comparable to the familiar multivariate GARCH models if the distribution is Gaussian.

An interesting final feature of our model is that one can easily impose a factor structure on the volatilities and correlations. This can be done by picking the dimension of f_t to be lower than the number of elements in $\text{diag}(D_t)$ and Q_t or X_t. This approach can be used if the same factors drive more correlations, or if correlations and volatilities are driven by the same factors. The model allows the dynamic factors f_t to adapt automatically via the specification of Ψ_t in (9). Through the score of the density function, our framework naturally weights and combines the different sources of information in y_t to improve the current estimates of volatilities and correlations.

4   Time-varying scale matrix and an EM algorithm

The time-varying covariance matrix $\Sigma_t = L_t L_t'$ is specified by means of the factor f_t which is modeled by (6) – (9). We assume that the variance of the multivariate GH distribution exists and therefore we must constrain the fat-tailedness of the mixing variable ζ_t. For example, in the case of a skewed Student's t distribution, we require the degrees of freedom parameter to be higher than 4, rather than the usual 2 for the symmetric case. This constraint may not be realistic for financial data, especially returns on individual equities that have many jumps and outliers. As an alternative, we can specify the time-varying scaling matrix $\tilde{\Sigma}_t$ in (9) rather than the time-varying covariance matrix Σ_t. Moment restrictions are then no longer needed since the scaling matrix $\tilde{\Sigma}_t$ always exists.

The GH distribution relies on many parameters. This can complicate parameter estimation, particularly when the dimension of y_t is high; see the discussion in Hu (2005). This is one of the reasons why maximum likelihood estimation for the GH distribution is usually carried out by the Expectation-Maximization (EM) algorithm of Dempster, Laird, and Rubin (1977). A basic introduction of the EM algorithm for the GH distribution with a time-invariant covariance matrix is provided in McNeil et al. (2005). Parameter estimation for a multivariate GARCH model with a GH distribution is considered by Hu (2005). A key simplification in the EM algorithm is that the parameters for the mixing distribution can be separated from the location, skewness, and scale parameters. This convenient property does not hold for the model specification with the covariance matrix Σ_t. However, if we consider the model specification in terms of the scale matrix $\tilde{\Sigma}_t$, we are able to develop a newly modified EM algorithm for estimation. The usual advantages of EM estimation then again apply to our setting of a GH distribution with time-varying parameters.

First, we reformulate the model in terms of the scaling matrix $\tilde{\Sigma}_t = \tilde{L}_t \tilde{L}_t'$. Second, we develop the modified EM algorithm for estimating the static parameter vector θ. The mean-variance normal mixture model for the observations y_t using the square root scaling matrix $\tilde{L}_t$ is given by

$$y_t = \mu_{ty} + \zeta_t \tilde{L}_t \gamma + \sqrt{\zeta_t} \tilde{L}_t z_t, \tag{14}$$

with $\mu_{ty} = -\mu_\zeta \tilde{L}_t \gamma$. This specification follows from (3). Since $\tilde{\Sigma}_t = \tilde{L}_t \tilde{L}_t'$ is a covariance matrix for the normal variable in the mixture specification (3), we can use similar matrices as developed

in the previous section, that is $\tilde{\Sigma}_t = \tilde{D}_t \tilde{R}_t \tilde{D}_t$ and $\tilde{\Psi}_t = \partial \text{vech}(\tilde{\Sigma}_t)/\partial f_t'$.

In the implementation of the EM algorithm for the GH distribution, as proposed by McNeil et al. (2005), estimation of parameters governing the mixing variable specification (3) can be separated from estimation of the other parameters. The main difficulty in our current context is the dynamic process for f_t that is driven by the scaled score s_t of the GH distribution and depends on the parameters of the mixing variable. It appears difficult to split the parameter vector and to reduce a high-dimensional likelihood optimization into two lower dimensional optimization problems. Our modification of the EM algorithm, however, circumvents this problem on the basis of Result 2.

Result 2. *We can express the score function of the conditional observation density by*

$$\nabla_t = \frac{\partial \ln p(y_t|f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t} = \text{E}\left[\left.\frac{\partial \ln p(y_t|\zeta_t, f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t}\right| \mathcal{F}_t\right]. \tag{15}$$

The result enables us to partition the parameter vector as $\theta' = (\theta_1', \theta_2')$ where θ_2 contains the parameters associated with the distribution of the mixing variable ζ_t, in particular λ, χ, and ψ. The remaining parameters are collected in θ_1. We define the joint log likelihood of the observation y_t and the unobserved mixing variable ζ_t as

$$\sum_{t=1}^{n} \ln p(y_t, \zeta_t|f_t, \mathcal{F}_{t-1}; \theta) = \mathcal{L}_{1n}(\theta) + \mathcal{L}_{2n}(\theta_2), \tag{16}$$

with

$$\mathcal{L}_{1n}(\theta) = \sum_{t=1}^{n} \ln p(y_t|\zeta_t, f_t, \mathcal{F}_{t-1}; \theta), \qquad \mathcal{L}_{2n}(\theta_2) = \sum_{t=1}^{n} \ln p(\zeta_t; \theta_2), \tag{17}$$

where the conditional density $p(y_t|\zeta_t, f_t, \mathcal{F}_{t-1}; \theta)$ is Gaussian and the marginal density $p(\zeta_t; \theta_2)$ is Generalized Inverse Gaussian (GIG) denoted by $\text{N}^-(\lambda, \chi, \psi)$. For the implementation of the E-step in the EM algorithm, we define

$$Q_1(\theta, \hat{\theta}) = \int \dots \int \mathcal{L}_{1n}(\theta) \left(\prod_{t=1}^{n} p(\zeta_t|y_t, \mathcal{F}_{t-1}; \hat{\theta})\right) \mathrm{d}\zeta_n \dots \mathrm{d}\zeta_1 = \text{E}_{\hat{\theta}}\left[\mathcal{L}_{1n}(\theta)|\mathcal{F}_n\right], \tag{18}$$

and, similarly,

$$Q_2(\theta_2, \hat{\theta}) = \text{E}_{\hat{\theta}}\left[\mathcal{L}_{2n}(\theta_2)|\mathcal{F}_n\right]. \tag{19}$$

In Appendix A.3 we show that under the normalization constraint $\mu_\zeta = 1$, $Q_1(\theta, \hat{\theta})$ depends on θ_1 only. Consequently, we write $Q_i(\theta_i, \hat{\theta})$ for $i = 1, 2$, with a slight abuse of notation. The EM algorithm for parameter estimation is given follows.

Modified EM algorithm for a the dynamic GH model for the scale matrix

1. Start with an initial guess of the parameters, $\hat{\theta}^{(0)}$, and set $\ell = 0$.

2. Given a trial value of the parameters $\hat{\theta}^{(\ell)}$, define the modified transition equation for the scaling matrix as

$$f_{t+1} = \sum_{i=0}^{p-1} A_i \tilde{s}_{t-i} + \sum_{j=0}^{q-1} B_j f_{t-j}, \tag{20}$$

   where $\tilde{s}_t = S_t \tilde{\nabla}_t$, and

$$\tilde{\nabla}_t^{(\ell)} = \mathrm{E}_{\hat{\theta}^{(\ell)}} \left[\partial p(y_t | \zeta_t, f_t, \mathcal{F}_{t-1}; \theta) / \partial f_t \mid \mathcal{F}_t \right], \tag{21}$$

   with $\tilde{\nabla}_t^{(\ell)}$ fully specified in the appendix.

3. Given the modified dynamics, compute $Q_1(\theta_1, \hat{\theta}^{(\ell)})$ and maximize it numerically with respect to θ_1. The maximum is obtained at $\tilde{\theta}_1$.

4. Update $\hat{\theta}^{(\ell)}$ to $\tilde{\theta}^{(\ell)} = (\tilde{\theta}_1', (\hat{\theta}_2^{(\ell)})')$, compute $Q_2(\theta_2, \tilde{\theta}^{(\ell)})$ and maximize it numerically with respect to θ_2. The maximum is obtained at $\tilde{\theta}_2$.

5. Update $\hat{\theta}^{(\ell)}$ to $\hat{\theta}^{(\ell+1)} = (\tilde{\theta}_1', \tilde{\theta}_2')'$, increase ℓ by one, and iterate steps 2–5 until convergence.

Steps 3–5 are standard for the GH-EM algorithm; see, for example, McNeil et al. (2005). The E-step is developed in Appendix A.3. An important feature of our modified EM algorithm is that the optimization can still be split into two lower dimensional problems in steps 3 and 4, even though we have a GH model with time-varying parameters governed by complex dynamics. The key to this result is that step 3 of the algorithm is effectively based on fitting a standard multivariate Gaussian GARCH model with the updating equation (20). The crucial part that enables this is our modification to the standard EM algorithm in step 2. In this step, the updating equation that depends on θ_2 only via the score ∇_t is replaced by a simple equation that does not depend on θ_2. The intuition follows from Result 2. In the same way as in the E-step of the EM algorithm, the score function is replaced by a conditional expectation of a score function that depends on parameter values from the previous iteration, that is $\hat{\theta}^{(\ell)}$. As this score function is conditional on ζ_t, it is the score of a Gaussian density and therefore takes a very simple form. It follows that as the EM iterations converge to the ML estimates, the

14

score $\tilde{\nabla}_t^{(\ell)}$ in the EM algorithm converges to the score ∇_t of the full GH distribution via (15). This is confirmed by numerical experiments, where the ML parameter estimates are obtained by the modified EM algorithm and by directly maximizing the likelihood.

5 Monte Carlo evidence

To study the behavior of the new model, we carry out a Monte Carlo study. In the next section, we investigate the model's performance in an empirical study. In both settings, we benchmark the model's performance to the well-known DCC model. The simulations test the accuracy of the different models in estimating correlation patterns, similar to the experiments in Engle (2002). We describe the set-up in Subsection 5.1 and present the results in Subsection 5.2.

5.1 Simulation design

The design of our Monte Carlo experiments are similar to the original experiments for correlations as described in Engle (2002). We take the same deterministic functions as in Engle's paper, namely

(1) Constant: $f_t = 0.9$,

(2) Sine: $f_t = 0.5 + 0.4\cos(2\pi t/200)$,

(3) Fast Sine: $f_t = 0.5 + 0.4\cos(2\pi t/20)$,

(4) Step: $f_t = 0.9 - 0.5(t > 500)$,

(5) Ramp: $f_t = \mod(t/200)$.

This allows us to study the properties of competing statistical models under a range of correlation dynamics, such as slow and fast oscillations, and structural breaks.

The simulation experiment concentrates on recovering dynamic correlation patterns. We consider a bivariate series y_t with zero mean and unit variances, such that we can fully concentrate on the correlations. Using the five deterministic patterns for correlations described above, we generate bivariate time series y_t as

$$y_t \sim GHST(0, D_t R_t D_t, \gamma, \nu), \qquad D_t = I_2, \qquad R_t = \begin{pmatrix} 1 & f_t \\ f_t & 1 \end{pmatrix}. \tag{22}$$

Given the five different correlation patterns, we consider three different GHST distributions in our experiments. The GHST distribution contains the symmetric Student's t and the normal distribution as special cases. In particular, the GHST collapses to the symmetric Student's t distribution if the skewness parameter γ goes to zero. It further reduces to the normal distribution if the degrees of freedom parameter ν goes to infinity. As a benchmark, we start with the normal distribution. Then we introduce moderate kurtosis by setting $\nu = 5$. Finally, we introduce mild skewness by setting $\gamma = (-0.03, -0.03)'$.

In the experiment, we take the DCC model of Engle (2002) as our benchmark. Again, for each simulated DGP we use the correct class of distributions when computing the likelihood. We like to emphasize that we are not aware of an earlier application that considers a DCC model with GHST or GH distributed error terms. The DCC models are compared to our new model with a diagonal structure for the 3×3 matrices A_1 and B_1. To model the correlation, we use the hypersphere parameterization. The performance of the different statistical models is measured using the Mean Absolute Error (MAE) based on the difference between the estimated correlation and its true value. The MAEs are averaged across time and across simulations. We generate samples of size $T = 1,100$, discarding the first 100 observations to avoid dependence on initial conditions, and use 100 Monte Carlo replications.

5.2 Simulation results

Table 1 contains the results for our experiment. For the normal distribution, the performance of both models is roughly the same. There appears to be no noticeable loss in efficiency in this case of using the over-parameterized GHST distribution in the new model. Again we note that as the error distribution becomes more complex, the MAEs of the DCC model increases, whereas the MAEs of the new model remain rather stable. We also see that the new model outperforms the DCC in four out of five cases for the fat-tailed and skewed DGPs. The improved performance is mainly due to the weighting function and asymmetry effect in the updating equations (6) and (9) for the factor f_t. Due to this weighting incidentally large observations result in less distortions for the estimated correlation dynamics.

6 Empirical application

In an empirical study, we examine the correlations in a multivariate dataset with four blue-chip stocks from different industries: Coca-Cola, IBM, Merck and J.P. Morgan. All four stocks are part of the Dow Jones 30 index. We use daily log returns from January 1989 to December 2009 from CRSP. The final dataset contains 5295 daily observations. Descriptive statistics are provided in Table 2. It is clear that the series exhibit significant excess kurtosis and skewness, warranting the use of the GH distribution.

To estimate the volatilities and correlations, we use the DCC model with a Gaussian and GHST error distribution as our benchmark. We also implement our own model using the GHST, the GH Variance-Gamma (GHVG, with $\psi = 2\lambda$ and $\chi = 0$), and the GH error distribution. Our model has ten factors: four volatilities, and six correlations. We estimated the model using both the DCC and hypersphere parameterization for the correlation matrix. The estimation results for the dynamic parameters were similar, so we only report the results obtained under the DCC parameterization to maximize comparability with the DCC model. We use $p = q = 1$ in (6) and impose the same parsimony as in the DCC model. This means that we use diagonal matrices A_1 and B_1 in (6), and that the diagonal elements corresponding to the correlation equations have the same value.

The estimation results are presented in Table 3. The parameters governing the dynamics are statistically significant for all models. For the DCC model with a normal distribution, the persistence parameters for the volatilities $(A + B)$ are high. All the standard stationarity conditions are satisfied. Changing the specification to a DCC model with a GHST distribution has several effects. First, the likelihood increases by more than 1,200 points by adding only four parameters. The GHST distribution, therefore, provides a much better fit to the data. Second, the volatilities of the first two stocks (Coca Cola, IBM) are less affected by lagged squared errors. This can be seen from the reduced values for the A coefficients. By contrast, the persistence of the volatility dynamics of Merck (B_{d3}) goes up substantially. This is due to some highly influential observations during the sample period for this stock.

The skewness parameters in the DCC model with the GHST distribution are mostly insignificant. The only exception is the positive skewness for J.P. Morgan (stock 4). The signs of the skewness coefficients γ_i in Table 3 are compatible with the descriptive statistics in Table 2.

The degrees of freedom is estimated at 6.4 with a relatively small standard error.

For the new model with a GHST distribution, we see a further increase in the likelihood of more than 100 points. This increase in the likelihood is obtained without adding any parameters relative to the DCC model with a GHST distribution. The persistence parameters B and the degrees of freedom ν are estimated at similar values as for the DCC-GHST model. Note that the B parameters for the new model must be compared to the $A + B$ parameter of the DCC model. The primary reason for the increase in the likelihood is the effect of the fat-tailed and skewed GHST distribution on the volatility and correlation dynamics.

The effect of the altered dynamic specification under fat tails on the correlation dynamics can be clearly illustrated in Figure 1. The figure shows the estimated correlations of the different statistical models for the sub-sample 2002–2005 for the two pairs Coca Cola-IBM and Merck-J.P. Morgan. During this period, we note that several influential observations caused abrupt shifts in the estimated correlation levels based on the DCC model. For Coca Cola-IBM, clear examples of this behavior are seen at the end of the first quarter of 2002, mid 2003, mid 2004 and September 2004, and April 2005. For Merck-JP Morgan, similar patterns are observed around October 2003, November 2003, October 2004 (very clear), and February, March, and October 2005. During all these episodes, the evolution of the correlations for the new model is much more stable and in line with expectations that correlations should behave rather smoothly. We also note that the estimated dynamics of the correlations for the DCC models repeatedly take a long time to revert to their old pattern. For example, for Coca Cola-IBM, it takes roughly three months starting from the big drop in mid-2003 before the DCC model and the new model exhibit similar correlation levels. The same is true for mid-2004. This holds even more strongly for Merck-JP Morgan after October 2004, when Merck experienced a large incidental drop in its stock price after it announced a major worldwide withdrawal of its products. As seen from the DCC-GHST model, the use of the GHST model alone does not remedy the distortive impact of such an influential observation. To adequately cope with such outliers, changes in the dynamic equation of the correlations are needed in addition to a fat-tailed observation density. Our current score driven modeling framework provides such corrections in a natural way.

The correlation differences between the DCC and score driven models over the entire sample and for all six pairs of stocks are presented in Figure 2. A positive value in the graph indicates that the DCC estimate of the correlation is lower than that of the score driven model. We

see that the differences in the correlations can be substantial at times. When the underlying observations causing the differences are extreme, the differences can persist for months and in some cases even years. In particular, we note that during the dotcom crash in 2000 the IBM correlations in the new model are estimated at a higher level than the estimates in the DCC models. Also, for the pairs involving J.P. Morgan, the correlations from the score driven models during the Financial Crisis are larger than for the DCC models. Such differences can have important implications for diversification, risk management, and asset allocation.

The skewness parameters for the new model with a GHST distribution are significant, except for Merck. The signs are in line with the descriptive statistics from Table 2. We note again that the skewness parameters also contribute to the different correlation dynamics via (6).

Figure 3 shows the volatility estimates of both models. The volatility patterns are at first sight much more in line between the two different volatility specifications. However, closer inspection shows that the same effect of large innovations affects the volatility dynamics. This is most clearly seen for Merck. For example, in October 2004 the stock price drops significantly for reasons explained earlier. This causes a large spike in the volatility estimate of the GARCH-DCC model, despite the use of a GHST error distribution. The spike in volatility only decreases very slowly to normal levels over a period of almost a year. The score driven model, by contrast, also shows an increase in volatility since October 2004, but on a much more modest and realistic scale. Though this is one of the most striking differences between the two models, there are many more. Particularly the IBM stock shows over the entire sample period various cases where the volatility as estimated by the DCC model first jumps and then gradually recedes to normal levels. This results in small reverse saw-tooth like patterns in the graph. The corresponding volatility dynamics for the score driven models do not exhibit such peculiar behavior and are much more stable.

To conclude the empirical analysis, we also estimate a specification based on the GH Variance Gamma (GHVG) distribution, and on the general GH distribution. The GHVG has $\chi = 0$ and $\psi = 2\lambda$ and has a clear link to Lévy driven stochastic processes. The likelihood of the GHVG model is lower than that of the GHST model, and even lower than the DCC-GHST specification. This is confirmed by the model using the unrestricted GH distribution. Interestingly, the unconstrained GH estimates reveal that the GHST model is a good model for the data set at hand. We see that the parameter $\kappa = (\psi\chi)^{1/2}$ is very close to zero, and that the

value of λ is negative. For the GHST, we have $\lambda = -\nu/2$, and this is precisely the value that is estimated under the GH specification. We conclude that the GHST distribution provides sufficient flexibility to accommodate the current levels of fat-tailedness and skewness combined with correlation and volatility dynamics.

7 Conclusion

We have proposed a new time-varying conditional correlation model that accounts for skewness and fat tails through the use of the Generalized Hyperbolic (GH) distribution with time-varying parameters. The distinguishing feature of the model is that the shape of the observation distribution directly affects the mechanism by which volatilities and correlations are updated. The key mechanism for this is the use of the local density score to update volatilities and correlations. As a result, large observations are reweighted before they enter the updating equation. Because of this, the model is much less sensitive to outliers and incidental influential observations. The new model also includes a natural asymmetry term if the GH distribution is skewed.

We showed that the model is easy to estimate by standard maximum likelihood and Expectation-Maximization procedures. In a simulation experiment, we demonstrated that the model does a better job at estimating the unknown correlation dynamics than competing models if the error distribution is fat-tailed and skewed. When applied to real data, we showed that the model yields a more robust assessment of local volatility and correlation dynamics. Because the new model accounts for fat tails and skewness in the volatility and correlation dynamics, it is less affected by aberrant observations and therefore produces clearer picture of actual volatilities and correlations.

A Appendix

A.1 Skewness of the GH distribution

Define $\bar{y}_t = \tilde{L}_t^{-1} y_t$ and $m_{i\zeta} = \mathrm{E}[(\zeta_t - \mu_\zeta)^i]$ for integer i. Let e_i denote the ith column of I_k. We obtain

$$\mathrm{E}[\bar{y}_t] = 0, \tag{A1}$$

$$\mathrm{E}[\bar{y}_t \bar{y}_t'] = \mu_\zeta \mathrm{I} + m_{2\zeta} \gamma \gamma' = (T'T)^{-1}, \tag{A2}$$

$$\mathrm{E}[\bar{y}_t \otimes \bar{y}_t \bar{y}_t'] = m_{3\zeta} \gamma \otimes \gamma \gamma' + m_{2\zeta} \begin{pmatrix} \gamma_1 \mathrm{I}_k + \gamma e_1' + e_1 \gamma' \\ \vdots \\ \gamma_k \mathrm{I}_k + \gamma e_k' + e_k \gamma' \end{pmatrix}, \tag{A3}$$

such that the skewness of $\bar{y}_t$ only depends on γ and on the variance and skewness of the mixing variable ζ_t.

A.2 The Score of the GH distribution

Define the matrix $\mathrm{vec}(L) = \mathcal{D}_k^0 \mathrm{vech}(L)$ for a lower triangular matrix L. Note that $\mathcal{D}_k^0$ is different from the standard duplication matrix $\mathcal{D}_k$ for a symmetric matrix S, i.e., $\mathrm{vec}(S) = \mathcal{D}_k \mathrm{vech}(S)$ with $\mathcal{B}_k = (\mathcal{D}_k' \mathcal{D}_k)^{-1} \mathcal{D}_k'$. Also note that $\mathcal{D}_k^{0'} \mathcal{D}_k^0 = \mathrm{I}_k$, such that $\mathcal{B}_k^0 = \mathcal{D}_k^{0'}$. Finally, let $\mathcal{C}_k$ be the commutation matrix, $\mathrm{vec}(S') = \mathcal{C}_k \mathrm{vec}(S)$ for an arbitrary matrix S. For completeness, we mention that $\tilde{L}_t = L_t T$, and $\tilde{\Sigma}_t = \tilde{L}_t \tilde{L}_t'$.

An intermediate result is

$$
\begin{aligned}
\mathrm{d}\Sigma_t &= \mathrm{d}(L_t L_t') &\Leftrightarrow \\
\mathrm{vec}(\mathrm{d}\Sigma_t) &= (\mathrm{I}_{k^2} + \mathcal{C}_k)(L_t \otimes \mathrm{I}_k) \mathrm{vec}(\mathrm{d}L_t) &\Leftrightarrow \\
\mathcal{D}_k \mathrm{vech}(\mathrm{d}\Sigma_t) &= (\mathrm{I}_{k^2} + \mathcal{C}_k)(L_t \otimes \mathrm{I}_k) \mathcal{D}_k^0 \mathrm{vech}(\mathrm{d}L_t) &\Leftrightarrow \\
\mathrm{vech}(\mathrm{d}L_t) &= \left(\mathcal{B}_k (\mathrm{I}_{k^2} + \mathcal{C}_k)(L_t \otimes \mathrm{I}_k) \mathcal{D}_k^0 \right)^{-1} \mathrm{vech}(\mathrm{d}\Sigma_t).
\end{aligned}
\tag{A4}
$$

First define the standardized y_t as $x_t = \tilde{L}_t^{-1} y_t + \mu_\zeta \gamma$. The random variable x_t has a GH distribution with location 0 and scaling matrix I_k. Let $d_z^\nu = \nu + z'z$ for a scalar ν and a vector z. With this notation, the density of the GH distribution of y_t is given by

$$p^{GH}(y_t | f_t; \lambda, \chi, \psi, \mu_\zeta, \sigma_\zeta^2, \gamma, \Sigma_t) = \frac{e^{\gamma' x_t}}{|2\pi \tilde{L}_t \tilde{L}_t'|^{1/2}} \cdot \frac{\left(\sqrt{d_{x_t}^\chi / d_\gamma^\psi}\right)^{\lambda - k/2} \cdot K_{\lambda - k/2}\left(\sqrt{d_{x_t}^\chi d_\gamma^\psi}\right)}{(\sqrt{\chi/\psi})^\lambda \cdot K_\lambda\left(\sqrt{\chi\psi}\right)}, \tag{A5}$$

21

Let $k_\lambda(\cdot) = \ln K_\lambda(\cdot)$ with first derivative $k'_\lambda(\cdot)$. Define the scalar weight

$$w_t = -\frac{\lambda - k/2}{d^\chi_{x_t}} - \frac{k'_{\lambda - k/2}\left(\sqrt{d^\chi_{x_t} d^\psi_\gamma}\right)}{\sqrt{d^\chi_{x_t}/d^\psi_\gamma}}. \tag{A6}$$

We obtain

$$\nabla_t \;=\; \frac{\partial \text{vech}(\Sigma_t)'}{\partial f_t}\, \frac{\partial \text{vech}(L_t)'}{\partial \text{vech}(\Sigma_t)}\, \frac{\partial \text{vec}(\tilde{L}_t)'}{\partial \text{vech}(L_t)}\, \frac{\partial \ln p^{GH}(y_t|f_t)}{\partial \text{vec}(\tilde{L}_t)} \;=\; \Psi'_t \bar{H}'_t \frac{\partial \ln p^{GH}(y_t|f_t)}{\partial \text{vec}(\tilde{L}_t)},$$

with $\Psi_t = \partial \text{vech}(\Sigma_t)/\partial f'_t$ and

$$\bar{H}_t = \left((T' \otimes \mathrm{I}_k)\mathcal{D}^0_k\right)\left(\mathcal{B}_k\left(\mathrm{I}_{k^2} + \mathcal{C}_k\right)(L_t \otimes \mathrm{I}_k)\mathcal{D}^0_k\right)^{-1} \tag{A7}$$

using the intermediate result (A4).

Taking the derivative of the log-density with respect to $\text{vec}(\tilde{L}_t)$ and then via the chain rule with respect to f_t, we get

$$\begin{aligned}
\frac{\partial \ln p^{GH}(y_t|f_t)}{\partial \text{vec}(\tilde{L}_t)} &=\; \frac{\partial x'_t}{\partial \text{vec}(\tilde{L}_t)}\left(-0.5 w_t \frac{\partial d^\chi_{x_t}}{\partial x_t} + \gamma\right) - \text{vec}((\tilde{L}'_t)^{-1}) \\
&=\; (\tilde{L}^{-1}_t y_t \otimes (\tilde{L}'_t)^{-1})(w_t x_t - \gamma) - \text{vec}((\tilde{L}'_t)^{-1}) \\
&=\; (\tilde{L}^{-1}_t \otimes (\tilde{L}'_t)^{-1})(y_t \otimes \mathrm{I})(w_t \tilde{L}^{-1}_t y_t + w_t \mu_\zeta \gamma - \gamma) - \text{vec}((\tilde{L}'_t)^{-1}) \\
&=\; (\tilde{L}'_t \otimes \mathrm{I})(\tilde{\Sigma}^{-1}_t \otimes \tilde{\Sigma}^{-1}_t)\left(w_t y_t \otimes y_t - \text{vec}(\tilde{\Sigma}_t) - (1 - w_t \mu_\zeta)(y_t \otimes \tilde{L}_t \gamma)\right). \tag{A8}
\end{aligned}$$

The main result is now obtained by defining

$$H'_t = \bar{H}'_t(\tilde{L}'_t \otimes \mathrm{I})(\tilde{\Sigma}^{-1}_t \otimes \tilde{\Sigma}^{-1}_t). \tag{A9}$$

A.3   EM algorithm for time-varying scale matrix $\tilde{\Sigma}_t$

We first prove Result 2. It is easy to check that

$$\begin{aligned}
\nabla_t = \frac{\partial \ln p(y_t|f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t} &=\; \frac{1}{p(y_t|f_t, \mathcal{F}_{t-1}; \theta)} \int \frac{\partial p(y_t, \zeta_t|f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t} \mathrm{d}\zeta_t \\
&=\; \int \frac{\partial p(y_t|\zeta_t, f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t}\, \frac{p(\zeta_t; \theta_2)}{p(y_t|f_t, \mathcal{F}_{t-1}; \theta)} \mathrm{d}\zeta_t \\
&=\; \int \frac{\partial \ln p(y_t|\zeta_t, f_t, \mathcal{F}_{t-1}; \theta_1)}{\partial f_t}\, \frac{p(y_t, \zeta_t|f_t, \mathcal{F}_{t-1}; \theta)}{p(y_t|f_t, \mathcal{F}_{t-1}; \theta)} \mathrm{d}\zeta_t \\
&=\; \mathrm{E}\left[\frac{\partial \ln p(y_t|\zeta_t, f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t}\bigg|\mathcal{F}_t\right] \;=\; \tilde{\nabla}_t. \tag{A10}
\end{aligned}$$

22

Throughout, we impose the normalization constraint $\mu_\zeta = 1$. We note that

$$\ln p(y_t|\zeta_t, f_t, \mathcal{F}_{t-1}; \theta) = -\frac{1}{2}\ln|\tilde{\Sigma}_t| - \frac{k}{2}\ln(\zeta_t) - \frac{k}{2}\ln(2\pi) - \frac{1}{2\zeta_t}(y_t - (\zeta_t - \mu_\zeta)\tilde{L}_t\gamma)'\tilde{\Sigma}_t^{-1}(y_t - (\zeta_t - \mu_\zeta)\tilde{L}_t\gamma),$$
(A11)

and

$$\ln p(\zeta_t; \theta_2) = -\frac{\lambda}{2}\ln(\chi/\psi) - \ln(2) - \ln K_\lambda\left(\sqrt{\chi\psi}\right) + (\lambda - 1)\ln(\zeta_t) - \frac{1}{2}(\chi\zeta_t^{-1} + \psi\zeta_t),$$
(A12)

where $\tilde{L}_t = \tilde{L}(f_t)$ and $\tilde{\Sigma}_t = \tilde{\Sigma}(f_t)$, and where the mapping from f_t to $\tilde{\Sigma}_t$ does not depend on θ_2.

We define $\tilde{x}_t = \tilde{L}_t^{-1}y_t + \gamma$. From (A11) and (A12) and the properties of the Generalized Inverse Gaussian distribution (see Appendix A.2 of McNeil et al. (2005)), we get

$$\delta_{1t}^{(\ell)} = \mathrm{E}_{\hat{\theta}^{(\ell)}}\left[\zeta^{-1}\middle|\mathcal{F}_n\right] = \left(\frac{d_{\tilde{x}_t}^\chi}{d_\gamma^\psi}\right)^{-1/2}\frac{K_{\lambda-1-k/2}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right)}{K_{\lambda-k/2}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right)},$$
(A13)

$$\delta_{2t}^{(\ell)} = \mathrm{E}_{\hat{\theta}^{(\ell)}}\left[\zeta\middle|\mathcal{F}_n\right] = \left(\frac{d_{\tilde{x}_t}^\chi}{d_\gamma^\psi}\right)^{1/2}\frac{K_{\lambda+1-k/2}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right)}{K_{\lambda-k/2}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right)},$$
(A14)

$$\delta_{3t}^{(\ell)} = \mathrm{E}_{\hat{\theta}^{(\ell)}}\left[\ln(\zeta)\middle|\mathcal{F}_n\right] = \frac{\partial}{\partial\xi}\left(\frac{d_{\tilde{x}_t}^\chi}{d_\gamma^\psi}\right)^{\xi/2}\frac{K_{\lambda+\xi-k/2}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right)}{K_{\lambda-k/2}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right)}\Bigg|_{\xi=0},$$
(A15)

where $d_\cdot$ is defined below (A4).

From (A11) and using $\mu_\zeta = 1$, we obtain

$$\frac{\partial\ln p(y_t|\zeta_t, f_t, \mathcal{F}_{t-1}; \theta)}{\partial f_t} = \tilde{\Psi}_t'\tilde{H}_t'\mathrm{vec}\left(\zeta_t^{-1}y_t(y_t + \tilde{L}_t\gamma)' - \tilde{\Sigma}_t\right)$$
(A16)

with $\tilde{\Psi}_t = \partial\mathrm{vech}(\tilde{\Sigma}_t)/\partial f_t'$ and

$$\tilde{H}_t = (\tilde{\Sigma}_t^{-1} \otimes \tilde{\Sigma}_t^{-1})(\tilde{L}_t \otimes \mathrm{I})\mathcal{D}_k^0(\mathcal{B}_k(\mathrm{I}_{k^2} + \mathcal{C}_k)(\tilde{L}_t \otimes \mathrm{I}_k)\mathcal{D}_k^0)^{-1},$$

and with $\tilde{L}_t$ a lower triangular matrix. Taking conditional expectations, we obtain

$$\tilde{\nabla}_t^{(\ell)} = \tilde{\Psi}_t'\tilde{H}_t'\left(\delta_{1t}^{(\ell)}y_t \otimes y_t - \mathrm{vec}(\tilde{\Sigma}_t) - (1 - \delta_{1t}^{(\ell)})(y_t \otimes \tilde{L}_t\gamma)\right),$$
(A17)

which only depends on $\hat{\theta}^{(\ell)}$, γ, and $\tilde{\Sigma}_t$, and therefore not on θ_2. As a result, the modified model for y_t conditional on ζ_t depends on θ_1 only.

Using these results, it is clear that $Q_1(\cdot)$ only depends on θ_1 and $\hat{\theta}^{(\ell)}$. We have

$$Q_1(\theta_1, \hat{\theta}^{(\ell)}) = -\frac{1}{2}\ln|\tilde{\Sigma}_t| - \frac{k}{2}\delta_{3t}^{(\ell)} - \frac{k}{2}\ln(2\pi) - \frac{1}{2}\delta_{1t}^{(\ell)}\tilde{x}_t'\tilde{x}_t + \tilde{x}_t'\gamma - \frac{1}{2}\delta_{2t}^{(\ell)}\gamma'\gamma. \tag{A18}$$

For expositional purposes, we restrict our attention to the model with order (1,1) dynamics,

$$f_{t+1} = A_1 S_t \tilde{\nabla}_t^{(\ell)} + B_1 f_t. \tag{A19}$$

Optimizing (A18) using the dynamics in (A19) now becomes similar to estimating a Gaussian multivariate GARCH in Mean model. The transition equation uses weighted (by $\delta_{1t}^{(\ell)}$) rather than unweighted squared observations to drive volatilities and correlations, see (A17). Similarly, there is weighting by $\delta_{1t}^{(\ell)}$ in the objective (A18). When optimizing over θ_1, however, these weights are fixed. Numerical optimization should therefore be faster than direct ML estimation of the full θ vector due to the less complicated likelihood and the lower dimensional parameter space.

Using the new estimate of θ_1 obtained by maximizing (A18), we update the parameter estimate to $\tilde{\theta}^{(\ell)}$, and use this new estimate to update the weights $\delta_{it}^{(\ell)}$. The second part of the EM maximization step then follows from

$$Q_2(\theta_2, \tilde{\theta}^{(\ell)}) = -\frac{\lambda}{2}\ln(\chi/\psi) - \ln(2) - \ln K_\lambda\left(\sqrt{\chi\psi}\right) + (\lambda - 1)\delta_{3t}^{(\ell)} - \frac{1}{2}(\chi\delta_{1t}^{(\ell)} + \psi\delta_{2t}^{(\ell)}), \tag{A20}$$

which can be optimized numerically with respect to θ_2 under the constraint $\mu_\zeta = 1$. This can be achieved by optimizing over $\kappa = \chi\psi > 0$ and λ, and using (10).

The similarity of (A17) and (A8) can be taken a step further by noting that $w_t = \delta_{1t}^{(\infty)}$, where $\delta_{1t}^{(\infty)}$ is evaluated using the true parameters. This follows from the fact that

$$w_t - \delta_{1t}^{(\infty)} = -\frac{\lambda - k/2}{d_{\tilde{x}_t}^\chi} + 0.5\frac{K_{\lambda-k/2+1}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right) - K_{\lambda-k/2-1}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right)}{K_{\lambda-k/2}\left(\sqrt{d_{\tilde{x}_t}^\chi d_\gamma^\psi}\right)\sqrt{d_{\tilde{x}_t}^\chi/d_\gamma^\psi}} \tag{A21}$$

and the properties of the modified Bessel function of the second kind,

$$K_{\lambda+1}(\kappa) = 2\lambda \cdot \kappa^{-1} \cdot K_\lambda(\kappa) + K_{\lambda-1}(\kappa),$$

and

$$\frac{\partial \ln K_\lambda(\kappa)}{\partial \kappa} = \frac{K_{\lambda+1}(\kappa) + K_{\lambda-1}(\kappa)}{2 \cdot K_\lambda(\kappa)},$$

such that from (A21) it follows that $w_t - \delta_{1t}^{(\infty)} = 0$.

References

Alexander, C. (1998). Volatility and correlation: Methods, models and applications. *Risk Manangement and Analysis: Measuring and Modelling Financial Risk*.

Alexander, C. (2001). Orthogonal garch. *Mastering risk 2*, 21–38.

Asai, M. and M. McAleer (2009). The structure of dynamic correlations in multivariate stochastic volatility models. *Journal of Econometrics 150*(2), 182–192.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 353*(1674), 401–419.

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics 5*(3), 151–157.

Bauwens, L. and S. Laurent (2005). A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. *Journal of Business & Economic Statistics 23*(3), 346–355.

Blæsild, P. (1981). The two-dimensional hyperbolic distribution and related distributions, with an application to Johannsen's bean data. *Biometrika 68*(1), 251.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. *The Review of Economics and Statistics 72*(3), 498–505.

Boswijk, H. and R. Van der Weide (2006). Wake me up before you GO-GARCH. *Discussion Papers*, 06–079.

Creal, D., S. Koopman, and A. Lucas (2008). A general framework for observation driven time-varying parameter models.

Creal, D., S. Koopman, and A. Lucas (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, (forthcoming).

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*(1), 1–38.

Eberlein, E. and U. Keller (1995). Hyperbolic distributions in finance. *Bernoulli 1*(3), 281–299.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics 20*(3), 339–350.

Engle, R. and K. Kroner (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory 11*(1), 122–150.

Fiorentini, G., E. Sentana, and G. Calzolari (2003). Maximum likelihood rstimation and inference in multivariate conditionally heteroscedastic dynamic regression models with Student's T innovations. *Journal of Business & Economic Statistics 21*(4), 532–547.

Franses, P. and D. Van Dijk (2000). *Nonlinear time series models in empirical finance.* Cambridge Univ Pr.

Harvey, A. and T. Chakravarty (2008). Beta-t-(E) GARCH. *Cambridge Working Papers in Economics*.

Hu, W. (2005). *Calibration Of Multivariate Generalized Hyperbolic Distributions Using The EM Algorithm, With Applications In Risk Management, Portfolio Optimization And Portfolio Credit Risk.* Ph. D. thesis.

Jaeckel, P. and R. Rebonato (1999/2000, Winter). The most general methodology for creating a valid correlation matrix for risk management and option pricing purposes. *Journal of Risk 2*(2), 17–28.

McNeil, A., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton university press Princeton, NJ.

Mencia, F. and E. Sentana (2004). Estimation and testing of dynamic models with generalised hyperbolic innovations. *FMG Discussion Papers*.

Nelson, D. B. and D. P. Foster (1994). Asymptotic filtering theory for univariate ARCH models. *Econometrica 62*(1), 1–41.

Peters, J. (2001). Estimating and forecasting volatility of stock indices using asymmetric GARCH models and (Skewed) Student-t densities. *Preprint, University of Liege, Belgium, 20 pp. pdf*.

Prause, K. (1999). The generalized hyperbolic model: estimation, financial derivatives, and risk measures. *University of Freiburg, Doctoral Thesis*.

Shephard, N. (2005). *Stochastic Volatility: selected readings.* Oxford University Press.

van der Weide, R. (2002). GO-GARCH: a multivariate generalized orthogonal GARCH model. *Journal of Applied Econometrics 17*(5), 549–564.
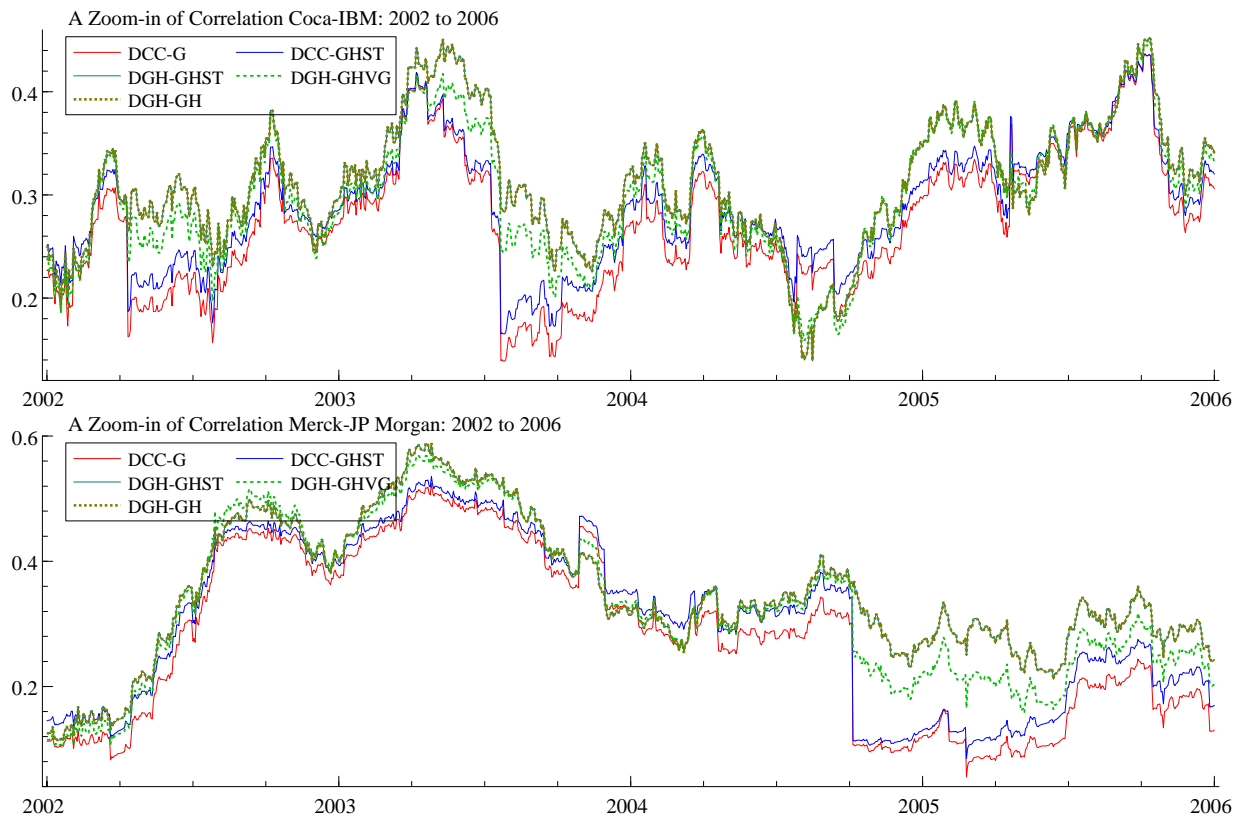
Figure 1: Estimated correlation subsample for the DCC-GHST, and DGH-GHST models.

To look into the difference of the DGH(1,1) and the DCC(1,1) model, we have zoom-in plot of the estimated correlations under different models: DCC-GHST and DGH-GHST. We can see that the DGH(1,1) is less volatile than the estimates from DCC.
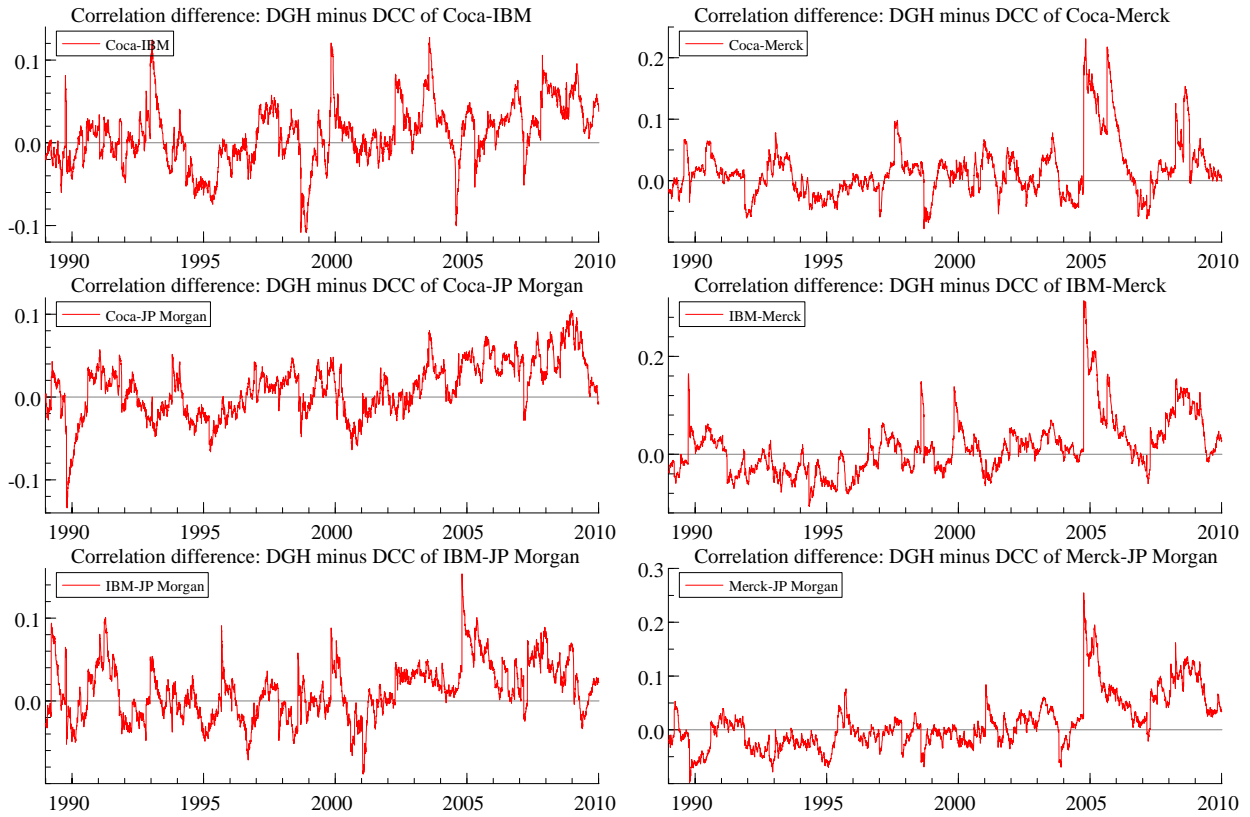
Figure 2: Estimated correlation difference for the DCC-GHST, and DGH-GHST models.

To look into the difference of the DGH(1,1) and the DCC(1,1) model, we have the plot of the difference in estimation under GHST distributions. It appears that the DCC differs from DGH model even under the same parametric assumption.
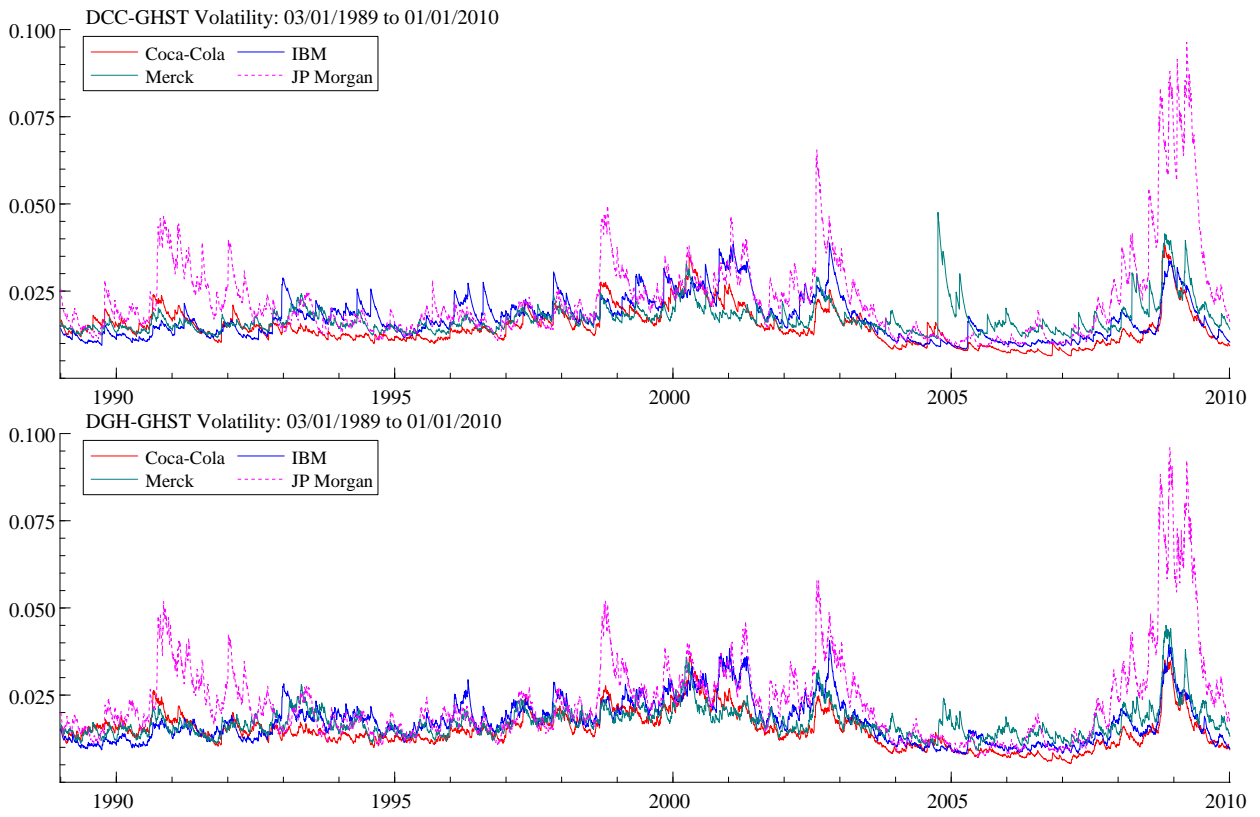
Figure 3: Volatility estimation from DCC-GHST and DGH-GHST: stock returns

The volatility estimates from the DCC under GHST distributions and DGH(1,1)-GHST with stock return data. From the graphs, we can see that the volatility from DGH-GHST is smoother than the DCC estimates.

Table 1: Mean Absolute Errors for Correlation Estimates

The table presents the average Mean Absolute Error (MAE) over 100 Monte Carlo replications and 1,000 time series observations for the correlation estimates for three different distributions (in pairs of columns) and five different correlation patterns. The distributions used are the normal, Student's $t(5)$, and GHST$(0, \Sigma_t, -0.03, \nu)$. The boldface numbers show the models with the smallest MAE for a given DGP.

| | Dynamic Correlations | | | | | |
| | normal | | $t(5)$ | | GHST | |
	DCC	model (6)–(9)	DCC	model (6)–(9)	DCC	model (6)–(9)
Constant	0.004	0.004	0.006	**0.005**	0.005	0.005
Sine	0.135	**0.133**	0.139	**0.129**	0.151	**0.133**
Fast Sine	**0.225**	0.226	0.255	**0.219**	0.254	**0.221**
Step	0.066	**0.065**	0.068	**0.067**	0.094	**0.070**
Ramp	0.159	**0.157**	**0.159**	0.165	0.168	**0.164**

Table 2: Data descriptive statistics.

The descriptive statistics for the empirical CRSP stock returns between January 1989 and December 2009. All observations are daily log returns. All four stocks are part of the Dow Jones 30 composite index. All skewness and excess kurtosis statistics have p-values below 10^{-4}.

	Coca-Cola	IBM	Merck	JP Morgan
Mean $\times 10^4$	6.33	5.27	5.69	7.56
Median	0.00	0.00	0.00	0.00
Standard Deviation$\times 100$	1.56	1.89	1.82	2.62
Skewness	0.23	0.29	-0.12	0.74
Excess Kurtosis	2.05	3.89	3.00	8.33
Minimum	-0.10	-0.16	-0.15	-0.21
Maximum	0.14	0.13	0.13	0.25

Table 3: Empirical Estimation Results

Empirical results based on stock return data between January 1989 and December 2009 for Coca Cola, IBM, Merck, and JP Morgan. The DCC model is defined as in Engle (2002) and uses a normal or GH skewed t (GHST) likelihood. The new model uses the GH, GHST, and GH Variance Gamma (GHVG) distribution for the likelihood and the parameter dynamics. Intercepts are not reported to save space. A_{d1} to A_{d4} and B_{d1} to B_{d4} contain the diagonal elements of A_1 and B_1 from (6) corresponding to the volatilities, and A_ρ and B_ρ the parameter corresponding to the correlations. γ_i is the skewness parameter for series i (1: Coca Cola, 2: IBM, 3: Merck, 4: JP Morgan), $\kappa = (\chi\psi)^{1/2}$, with χ, ψ and λ the GH parameters. For the GHST, we report $\nu = -2\lambda$.

| | DCC | | New Model (6)–(9) | | |
	Gaussian	GHST	GHST	GHVG	GH
A_{d1}	0.037^a	0.029^a	0.032^a	0.119^a	0.032^a
	(0.004)	(0.003)	(0.004)	(0.017)	(0.004)
A_{d2}	0.035^a	0.026^a	0.034^a	0.109^a	0.034^a
	(0.005)	(0.003)	(0.004)	(0.016)	(0.004)
A_{d3}	0.038^a	0.030^a	0.038^a	0.110^a	0.038^a
	(0.007)	(0.005)	(0.005)	(0.020)	(0.005)
A_{d4}	0.057^a	0.053^a	0.053^a	0.174^a	0.052^a
	(0.005)	(0.005)	(0.007)	(0.024)	(0.006)
A_ρ	0.010^a	0.010^a	0.010^a	0.030^a	0.010^a
	(0.001)	(0.001)	(0.001)	(0.005)	(0.001)
B_{d1}	0.959^a	0.969^a	0.996^a	0.999^a	0.996^a
	(0.004)	(0.004)	(0.001)	(0.002)	(0.001)
B_{d2}	0.959^a	0.969^a	0.994^a	1.000^a	0.994^a
	(0.005)	(0.004)	(0.002)	(0.001)	(0.002)
B_{d3}	0.913^a	0.956^a	0.989^a	0.985^a	0.989^a
	(0.018)	(0.008)	(0.003)	(0.004)	(0.003)
B_{d4}	0.939^a	0.944^a	0.994^a	0.996^a	0.994^a
	(0.005)	(0.006)	(0.002)	(0.001)	(0.002)
B_ρ	0.986^a	0.985^a	0.996^a	0.997^a	0.996^a
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
γ_1		0.045	0.089^a	0.131^a	0.089^a
		(0.028)	(0.029)	(0.038)	(0.029)
γ_2		0.026	0.074^a	0.099^a	0.074^a
		(0.027)	(0.028)	(0.037)	(0.028)
γ_3		-0.038	-0.026	-0.025	-0.026
		(0.028)	(0.028)	(0.037)	(0.028)
γ_4		0.056^a	0.083^a	0.126^a	0.083^a
		(0.028)	(0.028)	(0.038)	(0.028)
ν		6.434^a	6.318^a		
		(0.235)	(0.245)		
λ				3.738^a	-3.160^a
				(0.128)	(0.123)
κ					0.027
					(0.467)
Log-lik	-39991	-38787	-38684	-38994	-38684