Journal of Econometrics 213 (2019) 493–515

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Regime switching dynamic correlations for asymmetric and fat-tailed conditional returns☆

Marc S. Paolella [a,b], Paweł Polak [c,*], Patrick S. Walker [a]

[a] Department of Banking and Finance, University of Zurich, Switzerland
[b] Swiss Finance Institute, Zurich, Switzerland
[c] Department of Statistics, Columbia University, NY, United States

ARTICLE INFO

Article history:
Received 23 March 2018
Received in revised form 25 May 2019
Accepted 4 July 2019
Available online 20 August 2019

JEL classification:
C01
C32
C51
C53
C58
G11
G17

Keywords:
GARCH
Markov switching
Multivariate generalized hyperbolic distribution
Portfolio optimization
Value-at-risk

ABSTRACT

A non-Gaussian multivariate regime switching dynamic correlation model for financial asset returns is proposed. It incorporates the multivariate generalized hyperbolic law for the conditional distribution of returns. All model parameters are estimated consistently using a new two-stage expectation–maximization algorithm that also allows for incorporation of shrinkage estimation via quasi-Bayesian priors. It is shown that use of Markov switching correlation dynamics not only leads to highly accurate risk forecasts, but also potentially reduces the regulatory capital requirements during periods of distress. In terms of portfolio performance, the new regime switching model delivers consistently higher Sharpe ratios and smaller losses than the equally weighted portfolio and all competing models. Finally, the regime forecasts are employed in a new dynamic risk control strategy that avoids most losses during the financial crisis and vastly improves risk-adjusted returns.

1. Introduction

Within the field of financial time series analysis, models for conditional heteroskedasticity of asset returns now have a long history and are widely applied to account for clustering effects in the time-varying volatility of asset returns; see, e.g., Bauwens et al. (2006) and Bollerslev (2010) for surveys. Several GARCH-type models in the multivariate setting, hereafter MGARCH, have been proposed, some of which are, in terms of parameter estimation, suitable for large

☆ This research was partially conducted while Walker was a visiting scholar at the Statistics Department at Columbia University. He is grateful for the financial support by the Swiss National Science Foundation (SNSF) through the Doc.Mobility fellowship no. P1ZHP1171760.

The authors are grateful to Richard Davis, Walter Pohl and Karl Schmedders, as well as the participants of the 10th International Conference on Computational and Financial Econometrics 2016 in Sevilla the Swiss Finance Institute PhD in Finance Workshop 2016 in Zurich the 1st International Conference on Econometrics and Statistics 2017 in Hong Kong the 4th Konstanz-Lancaster Workshop on Finance and Econometrics 2018 in Konstanz and research seminars at the University of Zurich and Columbia University for helpful remarks and suggestions. Moreover we wish to thank two anonymous referees and the editor Yacine Aï t-Sahalia for excellent and instructive comments that led to a significant improvement in our initial submitted work.

* Corresponding author.
E-mail address: pawel.polak@columbia.edu (P. Polak).

https://doi.org/10.1016/j.jeconom.2019.07.002
0304-4076/

dimensions. This set includes the arguably most popular construction, namely the constant conditional correlation (CCC) model of Bollerslev (1990). For CCC, the filtered innovations from individual Gaussian univariate GARCH models applied to each constituent series are used via the usual plug-in estimator to estimate the correlation matrix. This assumes the correlations to be constant, but via the time-varying volatility from the individual fitted GARCH recursions, the covariance matrix changes through time. Based on the stylized facts of asset returns, of considerable interest are models that address the strong non-Gaussianity of the filtered GARCH innovation sequences. MGARCH models that allow for this include Aas et al. (2005), using the multivariate normal inverse Gaussian (NIG); Jondeau et al. (2007, Sec. 6.2) and Wu et al. (2015), using the multivariate skew-Student density; Santos et al. (2013), using a multivariate Student-t; and Virbickaite et al. (2016), using a Dirichlet location-scale mixture of multivariate normals. Multivariate generalizations of the univariate mixed normal GARCH model have been proposed and investigated by Bauwens et al. (2007) and Haas et al. (2009). Paolella and Polak (2015b) capitalize on the continuous Gaussian mixture property of the multivariate generalized hyperbolic (hereafter MGHyp) distribution to allow for very flexible non-Gaussianity, equation-by-equation GARCH modeling, and full parameter maximum-likelihood estimation via an EM algorithm. Models that can be related to the former framework using other non-Gaussian distributional families include Bianchi et al. (2016) and Slim et al. (2017).

One of the critiques of the popular CCC model of Bollerslev (1990) is that conditional covariances are time-varying only because of the univariate GARCH dynamics, while the correlations are constant. Addressing the stylized fact of time-varying correlations has become a highly active research area, with the seminal dynamic conditional correlation, or DCC, model of Engle (2002, 2009), and the varying correlation, or VC, model of Tse and Tsui (2002). These authors augment the baseline CCC model with a simple, two-parameter structure that allows for motion in the correlation coefficients. The fact that only two parameters are invoked to model the evolution of a correlation matrix is obviously a limitation, but can also be seen as its strong point: Estimation is straightforward, parameter proliferation in high dimensions is avoided, and the DCC and VC models can still pick up enough signal to improve forecasts over the CCC. Clearly, such a model (like GARCH itself) is a simplistic representation of a more complicated underlying process, and its efficacy is judged not on an economic modeling justification, but rather by an improved forecasting ability. Pelletier (2006) took a different approach, suggesting an MGARCH model with Markov switching between two CCC structures. We refer to this model subsequently as Gaussian-RSDC and build on this approach, addressing the non-Gaussianity of the underlying innovation process, and doing so in a coherent stochastic framework, as opposed to an ad-hoc two step approach commonly used to (possibly incorrectly) cobble together, e.g., a Gaussian GARCH-DCC or -RSDC model with a, say, multivariate Student-t distribution; see Paolella and Polak (2017) for a discussion of this issue.

In this paper, our goal is to introduce a model that lends itself to feasible estimation in high dimensions, addresses the aforementioned stylized facts, and, crucially, leads to superior out-of-sample forecasting ability. Before discussing this new model, we continue with some relevant literature review upon which our model is inspired. So and Yip (2012) extend the DCC framework by grouping similar correlations into clusters and allowing different dynamics for the different clusters. Similarly, a clustering of variances in the DCC model is investigated by Aielli and Caporin (2014), whereas Kasch and Caporin (2013) extend the DCC framework by allowing the correlation dynamics to depend on the variances through a threshold structure. Billio and Caporin (2005) generalize the DCC framework through Markov switching in both the parameters of the DCC specification and the unconditional correlations. Models that allow for excess kurtosis and asymmetry of the distribution of innovations in the DCC model have been proposed in Paolella and Polak (2017) and Urga et al. (2011). The so-called DECO equicorrelation construction of Engle and Kelly (2012) models the correlation matrices as a parsimonious time-varying convex combination of the unity matrix and a matrix of ones. With our proposed model we build upon the strand of literature that invokes Markov switching. These models continue to gain prominence in their ability to disentangle differing states in the data generating process of financial asset returns; see e.g. Haas et al. (2004), Haas and Paolella (2012) and Henry (2009) in the univariate setting; and Billio and Pelizzon (2000), Chevallier and Goutte (2015), Chollete et al. (2009), Fink et al. (2017), Haas and Liu (2018) and Pelletier (2006) in the multivariate case. Our model is closest to the Gaussian-RSDC of the latter article, in which a superior in-sample fit, compared to the GARCH-DCC model, is demonstrated. While improved in-sample fit does not necessarily imply improved (let alone economically significant) out-of-sample forecasting ability, it suggests that the changing correlations could be better modeled as being constant over short periods of time, and allowed to change according to a discrete Markov chain. With respect to out-of-sample performance of both density and risk prediction, we find, arguably unsurprisingly, that the Gaussianity assumption plays a large and detrimental role. The goal of this paper is to (i) address this factor in a coherent, non-ad-hoc way, such that parameter estimation in large dimensions is computationally feasible, straightforward, and consistent; and (ii) conduct extensive out-of-sample forecasting exercises to measure and compare its effectiveness to the DCC model, as well as the various models nested within our proposed structure. This includes the Gaussian-RSDC, as well as multivariate unconditional models that use non-Gaussian special cases of the generalized hyperbolic distributional assumption; see McNeil et al. (2015, Ch. 15), and references therein.

The proposed model also generalizes the class of so-called COMFORT models introduced in Paolella and Polak (2015b). An interesting feature of the MGHyp distribution employed in the COMFORT model is the introduction of a univariate stochastic term that elegantly and coherently gives rise to the non-Gaussianity. It can be interpreted as a common market factor, responsible for modeling news arrivals that jointly affect the distributions of all current period returns. While COMFORT is a very powerful model that outperforms all its special cases in terms of density forecasts, its dynamics cannot handle changes in the correlation structure because it is also endowed with a CCC structure. In this paper, we

relax this assumption and allow for regime switches in the dependency matrix, suggesting COMFORT-RSDC as the name for our new model class. The regime switching correlation dynamics are empirically motivated by the well documented leverage or down market effect, referring to the negative correlation between volatility and stock returns, as investigated, e.g., in the classical work of Black (1976); and the contagion effect, which describes the tendency of correlations between assets to shoot up during market crashes.

Crucially, the newly proposed model class is straightforward to estimate, even with a large number of assets, owing to a new two-step expectation–maximization (EM) algorithm for likelihood maximization. This lends itself to straightforward estimation of the multivariate predictive density, as discussed below, and this for potentially large dimensions.

As detailed in the empirical section below, the COMFORT-RSDC model with two regimes significantly outperforms the single-regime case and all Gaussian-based competitors in an extensive in- and out-of-sample analysis. This includes a forecasting exercise for the daily return distribution and financial risk measures, as well as an investigation of out-of-sample portfolio optimization. To further motivate the use of our likelihood-based estimation algorithm, issues related to its convergence properties are studied.

The remainder of the paper is structured as follows. Section 2 presents the new model, while Section 3 discusses the proposed method of estimation and some of its properties. Section 4 provides a detailed empirical study of the performance of the various models. Section 5 concludes and discusses future potential work based on the new RSDC model. Finally, mathematical details on the new estimation algorithm are provided in an Appendix.

2. Model

Let $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,K})'$ denote a return vector of K financial assets at time t, for $t = 1, \ldots, T$. The equally spaced realization of the return vector is denoted $\mathbf{Y} = [\mathbf{Y}_1|\mathbf{Y}_2|\cdots|\mathbf{Y}_T]$ and the information set at time t, defined as the sigma algebra generated by the history of returns $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$, is denoted $\mathbf{\Phi}_t$. In what follows, we assume that $\mathbf{Y}_t$ has a time varying conditional distribution with the COMFORT representation, from Paolella and Polak (2015b), given by

$$\begin{aligned} \mathbf{Y}_t|\mathbf{\Phi}_{t-1} &\overset{d}{=} \boldsymbol{\mu} + \boldsymbol{\gamma} G_t + \boldsymbol{\varepsilon}_t, \quad \text{with} \\ \boldsymbol{\varepsilon}_t &= \mathbf{H}_t^{1/2}\sqrt{G_t}\mathbf{Z}_t, \end{aligned} \tag{1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$ are column vectors in $\mathbb{R}^K$; $\mathbf{H}_t$ is a symmetric, positive definite dispersion matrix of order K; $\mathbf{Z}_t \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{I}_K)$ is a sequence of independent and identically distributed (i.i.d.) normal random variables; and $G_t \sim \mathrm{GIG}(\lambda, \chi, \psi)$ is an i.i.d. mixing random variable, independent of $\mathbf{Z}_t$, with the generalized inverse Gaussian (GIG) density given by

$$f_G(x; \lambda, \chi, \psi) = \frac{\chi^{-\lambda}\left(\sqrt{\chi\psi}\right)^{\lambda}}{2\mathcal{K}_\lambda\left(\sqrt{\chi\psi}\right)} x^{\lambda-1}\exp\left(-\frac{1}{2}\left(\chi x^{-1} + \psi x\right)\right), \quad x > 0;$$

$\mathcal{K}_\lambda(x)$ is the modified Bessel function of the third kind, given by

$$\mathcal{K}_\lambda(x) = \frac{1}{2}\int_0^\infty t^{\lambda-1}\exp\left(-\frac{x}{2}\left(t + t^{-1}\right)\right)\mathrm{d}t, \quad x > 0;$$

and $\chi > 0$, $\psi \geq 0$ if $\lambda < 0$; $\chi > 0$, $\psi > 0$ if $\lambda = 0$; and $\chi \geq 0$, $\psi > 0$ if $\lambda > 0$. The same MGHyp distribution arises from the parameter constellation $(\lambda, \chi/c, c\psi, \boldsymbol{\mu}, c\mathbf{H}_t, c\boldsymbol{\gamma})$ for any $c > 0$; hence for identification purposes, we fix either χ or ψ. The model assumes $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$, as well as the GIG parameters (λ, χ, ψ) to be time invariant. For the GIG distribution to be well defined, $\mathbb{E}[G_t|\mathbf{\Phi}_{t-1}]$ and $\mathbb{E}\left[G_t^2|\mathbf{\Phi}_{t-1}\right]$ have to be positive; sufficient conditions are given in Paolella and Polak (2015b).

The symmetric, positive definite, conditional dispersion matrix $\mathbf{H}_t$ is decomposed as

$$\mathbf{H}_t = \mathbf{S}_t\mathbf{\Gamma}_t\mathbf{S}_t, \tag{2}$$

where $\mathbf{S}_t$ is a diagonal matrix holding the strictly positive conditional scale terms $s_{k,t}$, $k = 1, \ldots, K$, and $\mathbf{\Gamma}_t$ is a positive definite dependency matrix. The univariate scale terms $s_{k,t}$ are each modeled by a GARCH-type process. To describe the volatility clustering we use the GARCH(1,1) model defined by

$$s_{k,t}^2 = \omega_k + \alpha_k\varepsilon_{k,t-1}^2 + \beta_k s_{k,t-1}^2, \tag{3}$$

where $\varepsilon_{k,t} = y_{k,t} - \mu_k - \gamma_k G_t$ is the kth element of the $\boldsymbol{\varepsilon}_t$ vector in (1), and $\omega_k > 0$, $\alpha_k \geq 0$, $\beta_k \geq 0$, for $k = 1, \ldots, K$.

In our model, $\boldsymbol{\mu}$ and $\mathbf{H}_t$ are the location vector and the dispersion matrix of the conditional distribution of $\mathbf{Y}_t$, respectively, while the mean and the covariance matrix are given by

$$\begin{aligned} \mathbb{E}[\mathbf{Y}_t|\mathbf{\Phi}_{t-1}] &= \boldsymbol{\mu} + \mathbb{E}[G_t|\mathbf{\Phi}_{t-1}]\boldsymbol{\gamma}, \\ \mathrm{Cov}(\mathbf{Y}_t|\mathbf{\Phi}_{t-1}) &= \mathbb{E}[G_t|\mathbf{\Phi}_{t-1}]\mathbf{H}_t + \mathbb{V}(G_t|\mathbf{\Phi}_{t-1})\boldsymbol{\gamma}\boldsymbol{\gamma}', \end{aligned}$$

where $\mathbb{V}(G_t|\mathbf{\Phi}_{t-1}) = \mathbb{E}\left[G_t^2|\mathbf{\Phi}_{t-1}\right] - (\mathbb{E}[G_t|\mathbf{\Phi}_{t-1}])^2$. Furthermore, by definition of the mean–variance mixture distribution of $\mathbf{Y}_t|\mathbf{\Phi}_{t-1}$, the matrix $\mathbf{\Gamma}_t$ is a correlation matrix, conditionally on the realization of the mixing process G_t. For convenience

and to make the analogy to the Gaussian literature, we use the names dependency matrix and correlation matrix interchangeably for $\mathbf{\Gamma}_t$. The CCC model entails using the time invariant dependency matrix $\mathbf{\Gamma}_t = \mathbf{\Gamma}$, as is also the case in the COMFORT model of Paolella and Polak (2015b). Under a Markov switching model with N regimes, the stochastic dependency matrix $\mathbf{\Gamma}_t$ at time t is given by

$$\mathbf{\Gamma}_t = \sum_{n=1}^{N} \mathbf{1}_{\{\Delta_t = n\}} \mathbf{\Gamma}_n, \tag{4}$$

where $\mathbf{1}_{\{\}}$ is an indicator function; Δ_t is a latent random variable governed by a first order, homogeneous Markov chain, independent of $\mathbf{Z}_t$ and G_t, which can take one of N possible values; and $\mathbf{\Gamma}_n$, $n = 1, \ldots, N$, are state-specific $K \times K$ symmetric, positive definite correlation matrices such that $\mathbf{\Gamma}_n \neq \mathbf{\Gamma}_m$ for $n \neq m$. The probability law governing Δ_t, and thereby $\mathbf{\Gamma}_t$, is defined by its time-invariant transition probability matrix, denoted by $\mathbf{\Pi} = \left[\pi_{n,m}\right]_{n,m=1,\ldots,N}$, where $\pi_{n,m}$ is the probability of going from state n in period t to state m in period $t+1$.

Dynamics in the dependency matrix can also be incorporated into the COMFORT model with the DCC approach of Engle (2002), as shown in Paolella and Polak (2015). However, these authors demonstrate that the lion's share of the improved gains in out-of-sample forecasts are generated from relaxing the Gaussian assumption (with or without GARCH), followed by use of a GARCH-type process for the margins, as compared to an i.i.d. assumption. While in a Gaussian framework, use of DCC does offer improvements to forecasting compared to CCC, when using (several special cases of) the MGHyp, the added benefit of DCC over CCC is marginal. We conjecture that, in a relatively heavily misspecified model such as the Gaussian-CCC, any reasonable additional structure to the data generating process will help improve forecasts. However, when the non-Gaussianity aspect is addressed, the sparsely parametrized DCC structure no longer is able to convey a distinct advantage. This is in no way to be interpreted as either that the returns do not have time-varying correlations, or that the DCC model should be discarded. Quite on the contrary, not only do asset returns display prominent time-varying correlation, but also the DCC structure is admirable for its use of only two additional parameters to account for some of this movement. In light of these observations, we believe that a richer DCC structure, such as the matrix-variate versions proposed in Billio et al. (2006) and Cappiello et al. (2006), would be able to extract further signal that leads to enhanced forecasting ability. In this work, instead of devising a model that allows continuous changes in correlations, we advocate the use of a small set of fixed constant correlation matrices incorporated into a Markov switching framework, as suggested in Pelletier (2006) in the Gaussian setting.

The predictive multivariate probability density function of returns, i.e., the density of the distribution of $\mathbf{Y}_{t+1}|\Phi_t$, is a finite mixture of MGHyp densities with different correlation matrices, weighted by the forecasted probabilities of next period's regimes

$$f_{\mathbf{Y}_{t+1}|\Phi_t}(\mathbf{y}) = \sum_{n=1}^{N} \xi_{n,t+1|t}\, f_{\mathbf{Y}_{t+1}|\Phi_t}(\mathbf{y}|\Delta_{t+1} = n),$$

where $\xi_{n,t+1|t}$, for $n = 1, \ldots, N$, is the forecasted probability that regime n will manifest itself at time $t+1$, given the information up to time t.

In the remainder of this paper, we restrict our attention to three non-Gaussian special cases of the MGHyp family, namely the multivariate Laplace, normal inverse Gaussian, and the Student-t distribution, all subsequently defined. While all the parameters of the MGHyp distribution are identified, in agreement with Prause (1999) and Protassov (2004), Paolella and Polak (2015b) find that the MGHyp exhibits a relatively flat likelihood in some of the parameters, and advocate fixing some parameters to overcome numerical issues. These special cases are faster and numerically more reliable to estimate, but still retain the flexibility required for modeling asset returns. Crucially, they allow for higher kurtosis than the normal distribution and also for individual asset asymmetry parameters. The multivariate asymmetric Laplace (MALap) or variance-gamma distribution is obtained for $G_t \sim \text{GIG}(\lambda, \chi, \psi)$ with $\lambda > 0$, $\chi = 0$ and $\psi = 2$; the multivariate asymmetric normal inverse Gaussian (NIG) distribution for $\lambda = -1/2$, $\chi > 0$ and $\psi = 1$; and the multivariate asymmetric Student-t (MAt) distribution for $\lambda = -\chi/2$ and $\psi = 0$. The corresponding symmetric distributions with $\boldsymbol{\gamma} = \mathbf{0}$ are abbreviated as MLap, SNIG, and Mt, respectively. Note that all univariate margins in the MGHyp distribution possess the same tail thickness, which is a potential concern of this distributional assumption when the constituent series have markedly different tail behaviors. It is not obvious how to generalize the current framework to allow for different tail behaviors and still preserve the analytic tractability of the portfolio distribution (weighted sum of the margins) arising from the multivariate predictive density. As an example, copula structures can easily address the heterogeneous tail behavior of the margins issue, but at the price of losing tractability of the portfolio distribution. Alternatively, as shown in Näf et al. (2019), making use of a Cholesky decomposition allows to incorporate tail-heterogeneity in multivariate normal mean–variance mixture distributions and still retain joint maximum likelihood estimation of all parameters.

3. Two-stage maximum likelihood estimation

Paolella and Polak (2015b) develop an expectation conditional maximization either (ECME) algorithm appropriate for estimating the model under the assumption of a constant conditional correlation matrix, i.e., the single-component case.

To generalize this to the N-component case, in principle, one could extend that algorithm into a nested double ECME algorithm, with the inner loop estimating the regime switching correlations. However, the rapid increase of the number of parameters combined with the dynamic structure in the conditional correlation matrix of the RSDC model makes this nested ECME impractical for large K. To overcome this, a new, two-stage estimation procedure is proposed. It replaces a double iterative procedure by two separate procedures applied iteratively, whereby the second procedure is conducted conditional on the results from the first one.

For notational convenience later, we collect the parameters of the model into four vectors (process, distribution, correlation matrices and Markov transition matrix):

$$\begin{aligned}&\boldsymbol{\theta}_P = \left(\boldsymbol{\mu}', \boldsymbol{\gamma}', \boldsymbol{\omega}', \boldsymbol{\alpha}', \boldsymbol{\beta}'\right)', \quad \boldsymbol{\theta}_D = (\lambda, \chi, \psi)', \\ &\boldsymbol{\theta}_C = \left(\text{vech}\,(\boldsymbol{\Gamma}_1)', \ldots, \text{vech}\,(\boldsymbol{\Gamma}_N)'\right)', \text{ and } \boldsymbol{\theta}_M = (\text{vec}_{N-1}\,(\boldsymbol{\Pi}))',\end{aligned} \tag{5}$$

where $\boldsymbol{\omega}, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are K-dimensional vectors of GARCH(1,1) parameters from (3); (λ, χ, ψ) denote the constant GIG parameters; $\text{vech}\,(\boldsymbol{\Gamma}_n)$, for $n = 1, \ldots, N$, denotes a column vector of the elements above the main diagonal of matrix $\boldsymbol{\Gamma}_n$; and $\text{vec}_{N-1}\,(\boldsymbol{\Pi})$ is a column vector of first $N-1$ columns of the transition probabilities matrix $\boldsymbol{\Pi}$. For the CCC model, $\boldsymbol{\theta}_C$ reduces to $\text{vech}\,(\boldsymbol{\Gamma})$ and $\boldsymbol{\theta}_M = \emptyset$ because there is no Markov chain governing regimes. For brevity in this case, we omit $\boldsymbol{\theta}_M$ in the formulas. Moreover, let $\mathbf{G} = (G_1, \ldots, G_T)'$ denote the latent univariate random variables that we deem the common market factor, and let $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_T)'$ denote the latent Markov chain governing the correlations.

The estimation algorithm is based on the decomposition of the complete conditional log-likelihood function into a sum of three terms: the normal log-likelihood (i.e., conditionally on the realization of the mixing random variable and the state of the Markov chain), the log-likelihood function of the mixing random variable (which is state independent), and the log-likelihood of the Markov chain variable, i.e.,

$$\log L_{\mathbf{Y},\mathbf{G},\boldsymbol{\Delta}}\,(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M) = \log L_{\mathbf{Y}|\mathbf{G},\boldsymbol{\Delta}}\,(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M) + \log L_{\mathbf{G}}\,(\boldsymbol{\theta}_D) + \log L_{\boldsymbol{\Delta}}(\boldsymbol{\theta}_M). \tag{6}$$

Next, the normal log-likelihood can be split as in Bollerslev (1990) and Pelletier (2006), i.e.,

$$\log L_{\mathbf{Y}|\mathbf{G},\boldsymbol{\Delta}}\,(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M) = \log L^{\text{MV}}_{\mathbf{Y}|\mathbf{G}}\,(\boldsymbol{\theta}_P) + \log L^{\text{Corr}}_{\mathbf{Y}|\mathbf{G},\boldsymbol{\Delta}}\,(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)\,, \tag{7}$$

where $L^{\text{MV}}_{\mathbf{Y}|\mathbf{G}}\,(\boldsymbol{\theta}_P)$ is the mean-volatility term given by

$$\begin{aligned}\log L^{\text{MV}}_{\mathbf{Y}|\mathbf{G}}\,(\boldsymbol{\theta}_P) = -\frac{1}{2}\sum_{t=1}^{T}\Big[&K\log(2\pi) + \log|\mathbf{S}_t|^2 \\ &+ g_t^{-1}\,(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\gamma} g_t)'\,\mathbf{S}_t^{-1}\mathbf{S}_t^{-1}\,(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\gamma} g_t) + \log g_t\Big],\end{aligned} \tag{8}$$

and $L^{\text{Corr}}_{\mathbf{Y}|\mathbf{G},\boldsymbol{\Delta}}\,(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)$ is the correlation term given by

$$\log L^{\text{Corr}}_{\mathbf{Y}|\mathbf{G},\boldsymbol{\Delta}}\,(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M) = -\frac{1}{2}\sum_{n=1}^{N}\left(\sum_{t:\Delta_t=n}\left[\log|\boldsymbol{\Gamma}_n| + \boldsymbol{e}_t'\boldsymbol{\Gamma}_n^{-1}\boldsymbol{e}_t - \boldsymbol{e}_t'\boldsymbol{e}_t\right]\right),$$

where $\boldsymbol{e}_t = g_t^{-1/2}\mathbf{S}_t^{-1}\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\varepsilon}_t = \mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\gamma} g_t$ are the residuals from (1). Owing to the mixture structure of the MGHyp, $L_{\mathbf{Y}|\mathbf{G},\boldsymbol{\Delta}}\,(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)$ is a multivariate Gaussian likelihood with a GARCH structure for the scales and a given conditional correlation model. As such, maximization of $L_{\mathbf{Y}|\mathbf{G},\boldsymbol{\Delta}}\,(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_P)$ can be done in two steps: First, with the correlation structure ignored, the GARCH parameters in (3) are estimated for each of the K assets separately (or concurrently with parallel computing) by maximizing $L_{\mathbf{Y}|\mathbf{G}}\,(\boldsymbol{\theta}_P|\boldsymbol{\theta}_C = \mathbf{I}_K)$, and, in the second step, the correlation parameters $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$ are estimated from the first-step standardized residuals.

In case of a constant dependency matrix, as in Paolella and Polak (2015b), the $\boldsymbol{\Gamma}$ matrix is estimated by the usual empirical correlation estimator (the MLE under normality) of the standardized residuals $g_t^{-1/2}\mathbf{S}_t^{-1}\widehat{\boldsymbol{\varepsilon}}_t$, with the unobserved realizations of $g_t^{-1/2}$ replaced by their conditional expectations from the E-step below. When $\boldsymbol{\Gamma}_t$ has a regime switching structure, the correlation step would be considerably slower. Therefore, we propose to keep the correlation estimates fixed at some $\boldsymbol{\theta}_C^*$ and $\boldsymbol{\theta}_M^*$ and turn directly to the maximization of the second term in the decomposition (6). Given the $\boldsymbol{\theta}_P$ estimates, denoted as usual by $\widehat{\boldsymbol{\theta}}_P$, we estimate the mixing process parameters $\boldsymbol{\theta}_D$ by maximizing $L_{\mathbf{Y}}\big(\boldsymbol{\theta}_D|\widehat{\boldsymbol{\theta}}_P, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\big)$. This implies that, in each iteration, we update $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_D$, and keep $\boldsymbol{\theta}_C^*$ and $\boldsymbol{\theta}_M^*$ fixed. Given all these estimates, we proceed with the next E-step update of the unobserved mixing random variable $\mathbf{G}$ (still with fixed $\boldsymbol{\theta}_C^*$ and $\boldsymbol{\theta}_M^*$) and continue to iterate until convergence. This completes the Stage-I ECME algorithm.

The updates of the correlation dynamics, $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$, are obtained in the second stage by a separate EM algorithm, conditional on the estimates in the first stage. By iterating the two algorithms we obtain the proposed two-stage EM. We now present the details of our algorithm.

Stage-I ECME algorithm

Stage I-1: Set the iteration count to $\ell = 1$ and choose appropriate starting values for $\boldsymbol{\theta}_D^{[1]} = \big(\lambda^{[1]}, \chi^{[1]}, \psi^{[1]}\big)$, and $\boldsymbol{\theta}_P^{[1]} = \Big(\boldsymbol{\mu}^{[1]'}, \boldsymbol{\gamma}', ^{[1]'}, \boldsymbol{\omega}^{[1]'}, \boldsymbol{\alpha}^{[1]'}, \boldsymbol{\beta}^{[1]'}\Big)'$.

Stage-I E-step: Calculate $\mathbb{E}\left[\log L_{\mathbf{Y},\mathbf{G}}|\Phi_t;\boldsymbol{\theta}_P^{[\ell]},\boldsymbol{\theta}_D^{[\ell]},\boldsymbol{\theta}_C^{*},\boldsymbol{\theta}_M^{*}\right]$. The log-likelihood (8) is linear with respect to g_t and g_t^{-1}, so that we can replace the unobserved realizations of G_t and G_t^{-1} in (6) by their conditional expectations $\mathbb{E}\left[G_t^{\pm1}|\Phi_t;\boldsymbol{\theta}_P^{[\ell]},\boldsymbol{\theta}_D^{[\ell]},\boldsymbol{\theta}_C^{*},\boldsymbol{\theta}_M^{*}\right]$, where

$$\left(G_t|\Phi_t;\boldsymbol{\theta}_P^{[\ell]},\boldsymbol{\theta}_D^{[\ell]},\boldsymbol{\theta}_C^{*},\boldsymbol{\theta}_M^{*}\right)\sim \text{GIG}\left(\lambda^{[\ell]}-K/2,\,\chi^{[\ell]}+\|\mathbf{y}_t-\boldsymbol{\mu}^{[\ell]}\|^2_{\mathbf{H}_t^{[\ell]}},\,\psi^{[\ell]}+\|\boldsymbol{\gamma}^{[\ell]}\|^2_{\mathbf{H}_t^{[\ell]}}\right) \tag{9}$$

and $\|\mathbf{x}\|^2_{\mathbf{H}}=\mathbf{x}'\mathbf{H}^{-1}\mathbf{x}$, for any vector $\mathbf{x}$ and a square matrix $\mathbf{H}$ of proper dimension.

Stage-I CM1-step: Update $\boldsymbol{\theta}_P$ by computing

$$\boldsymbol{\theta}_P^{[\ell+1]}=\arg\max_{\boldsymbol{\theta}_P}\log L^{MV}_{\mathbf{Y}|\mathbf{G}}\left(\boldsymbol{\theta}_P\right), \tag{10}$$

where $L^{MV}_{\mathbf{Y}|\mathbf{G}}\left(\boldsymbol{\theta}_P\right)$ is a Gaussian likelihood with zero correlations. This allows to estimate the parameters $(\mu_k,\gamma_k,\omega_k,\alpha_k,\beta_k)$ separately for each asset by maximizing the corresponding GARCH likelihood.

Stage-I CM2-step: Given the CM1-step estimates of $\boldsymbol{\theta}_P$, update the estimates of $\boldsymbol{\theta}_D$ by maximizing the incomplete data log-likelihood

$$\boldsymbol{\theta}_D^{[\ell+1]}=\arg\max_{\boldsymbol{\theta}_D}\log L_{\mathbf{Y}}\left(\boldsymbol{\theta}_D|\boldsymbol{\theta}_P^{[\ell+1]},\boldsymbol{\theta}_C^{*},\boldsymbol{\theta}_M^{*}\right).$$

Iterate these E- and CM-steps until convergence, to get $\hat{\boldsymbol{\theta}}_P$ and $\hat{\boldsymbol{\theta}}_D$.

Then proceed with Stage-II to update the regime switching parameters $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$. This is done conditional on $\hat{\boldsymbol{\theta}}_P$ and $\hat{\boldsymbol{\theta}}_D$ by the maximization

$$\arg\max_{\boldsymbol{\theta}_C,\boldsymbol{\theta}_M}\log L_{\mathbf{Y}}\left(\boldsymbol{\theta}_C,\boldsymbol{\theta}_M|\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\theta}}_D\right).$$

The Stage-II algorithm is itself an EM algorithm, conditional on the Stage-I estimates, and generalizes the algorithm of Hamilton (1993, 1994) to the case of MGHyp innovations. The smoothing formula for inference of the regime probabilities of Kim (1994) is adopted. A quasi-Bayesian approach is utilized for shrinkage estimation of the correlation matrices as in Hamilton (1991) and Paolella (2015) to reduce the estimation error in finite samples and to increase the forecasting performance of the model (see below).

Stage-II ECME algorithm

Stage II-1: Set the iteration count to $\ell=1$ and choose appropriate starting values for $\boldsymbol{\theta}_C^{[1]}=\left[\boldsymbol{\Gamma}_1^{[1]},\ldots,\boldsymbol{\Gamma}_N^{[1]}\right]$, and $\boldsymbol{\theta}_M^{[1]}=\boldsymbol{\Pi}^{[1]}$.

Stage-II E-step-1: Recursively calculate the forecasted and inferred state probabilities $\boldsymbol{\xi}_{t+1|t}^{[\ell]}$ and $\boldsymbol{\xi}_{t|t}^{[\ell]}$, as well as the smoothed inference $\boldsymbol{\xi}_{t|T}^{[\ell]}$, as derived in Kim (1994), from iterating on

$$\boldsymbol{\xi}_{t|t}^{[\ell]}=\frac{\boldsymbol{\xi}_{t|t-1}^{[\ell]}\odot\boldsymbol{\eta}_t^{[\ell]}}{\mathbf{1}_N'\left(\boldsymbol{\xi}_{t|t-1}^{[\ell]}\odot\boldsymbol{\eta}_t^{[\ell]}\right)},\quad \boldsymbol{\xi}_{t+1|t}^{[\ell]}=\boldsymbol{\Pi}^{[\ell]}\boldsymbol{\xi}_{t|t}^{[\ell]},\quad \boldsymbol{\xi}_{t|T}^{[\ell]}=\boldsymbol{\xi}_{t|t}^{[\ell]}\odot\left\{\boldsymbol{\Pi}^{[\ell]'}\left[\boldsymbol{\xi}_{t+1|T}^{[\ell]}\div\boldsymbol{\xi}_{t+1|t}^{[\ell]}\right]\right\}, \tag{11}$$

where $\odot$ and $\div$ denote the element by element product and division, respectively, and

$$\boldsymbol{\eta}_t^{[\ell]}=\begin{bmatrix} f\left(\mathbf{y}_t|\boldsymbol{\Phi}_{t-1},\Delta_t=1;\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\theta}}_D,\boldsymbol{\Gamma}_1^{[\ell]}\right)\\ f\left(\mathbf{y}_t|\boldsymbol{\Phi}_{t-1},\Delta_t=2;\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\theta}}_D,\boldsymbol{\Gamma}_2^{[\ell]}\right)\\ \vdots\\ f\left(\mathbf{y}_t|\boldsymbol{\Phi}_{t-1},\Delta_t=N;\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\theta}}_D,\boldsymbol{\Gamma}_N^{[\ell]}\right)\end{bmatrix}.$$

Here, $\boldsymbol{\xi}_{s|t}^{[\ell]}$ denotes the N-dimensional vector of probabilities of observing each of the N states at time s, based on the data obtained through date t and based on knowledge of the population parameters (5). Moreover, $\mathbf{1}_N$ is an $N\times1$ vector of ones. The starting values $\hat{\boldsymbol{\xi}}_{1|0}$ are chosen via maximum likelihood estimation by extending the parameter space accordingly.

Stage-II E-step-2: Calculate $\mathbb{E}\left[\log L_{\mathbf{Y},\mathbf{G},\boldsymbol{\Delta}}|\Phi_t;\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\theta}}_D,\boldsymbol{\theta}_C^{[\ell]},\boldsymbol{\theta}_M^{[\ell]}\right]$. The log-likelihood (8) is linear with respect to g_t and g_t^{-1}, so we replace the unobserved realizations of G_t and G_t^{-1} in (6) by their conditional expectations $\mathbb{E}\left[G_t^{\pm1}|\boldsymbol{\Phi}_t,\boldsymbol{\Delta}_t;\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\theta}}_D,\boldsymbol{\theta}_C^{[\ell]},\boldsymbol{\theta}_M^{[\ell]}\right]$. In particular, for $n=1,\ldots,N$,

$$g_{n,t}^{\pm1}=\mathbb{E}\left[G_t^{\pm1}|\boldsymbol{\Phi}_t,\Delta_t=n;\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\theta}}_D,\boldsymbol{\theta}_C^{[\ell]},\boldsymbol{\theta}_M^{[\ell]}\right], \tag{12}$$

where

$$\left(G_t|\boldsymbol{\Phi}_t;\hat{\boldsymbol{\theta}}_P,\hat{\boldsymbol{\theta}}_D,\boldsymbol{\theta}_C^{[\ell]},\Delta_t=n\right)\sim\text{GIG}\left(\lambda^{[\ell]},\chi_{n,t}^{[\ell]},\psi_{n,t}^{[\ell]}\right),$$

with

$$\lambda^{[\ell]} = \hat{\lambda} - K/2, \quad \chi_{n,t}^{[\ell]} = m_{n,t}^{[\ell]} + \hat{\chi}, \quad \psi_{n,t}^{[\ell]} = \hat{\psi} + \hat{\boldsymbol{\gamma}}' \left(\mathbf{S}_{n,t} \boldsymbol{\Gamma}_n^{[\ell]} \mathbf{S}_{n,t}\right)^{-1} \hat{\boldsymbol{\gamma}},$$
$$m_{n,t}^{[\ell]} = \left(\mathbf{y}_t - \hat{\boldsymbol{\mu}}\right)' \left(\mathbf{S}_{n,t} \boldsymbol{\Gamma}_n^{[\ell]} \mathbf{S}_{n,t}\right)^{-1} \left(\mathbf{y}_t - \hat{\boldsymbol{\mu}}\right),$$

and we define $\mathbf{S}_{n,t}$ to be the diagonal matrices collecting the estimated scale terms generated by Eq. (3) using the first-step estimates of $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_D$. The $\mathbf{S}_{n,t}$ also require regime specific conditional expectations of G_t and G_t^{-1} from (12), which can be calculated from the explicit formula for the moments of a GIG random variable, see, e.g., Paolella (2007, Sec. 9.4).

Stage-II CM-1-step: Update the estimates of the correlation matrices by a modification of the quasi-Bayesian estimator of Hamilton (1991), analogous to Paolella (2015),

$$\tilde{\boldsymbol{\Gamma}}_n^{[\ell+1]} = \frac{a_n \mathbf{B}_n + \sum_{t=1}^{T} g_{n,t}^{-1} \mathbf{S}_{n,t}^{-1} \left(\mathbf{y}_t - \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\gamma}} g_{n,t}\right) \left(\mathbf{y}_t - \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\gamma}} g_{n,t}\right)' \mathbf{S}_{n,t}^{-1} \xi_{n,t|T}^{[\ell]}}{a_n + \sum_{t=1}^{T} \xi_{n,t|T}^{[\ell]}}, \tag{13}$$

where the tilde indicates that the matrix must be standardized to be a correlation matrix, and fixed quantities $a_n \geq 0$ and $\mathbf{B}_n$ represent the strength and the target of the nth prior correlation matrix of the shrinkage estimator. Every $\mathbf{B}_n$ must be symmetric, positive definite with 1 on the diagonal and all off-diagonal elements between -1 and 1. By $\xi_{n,t|T}^{[\ell]}$ we denote the nth element of the probability vector $\boldsymbol{\xi}_{t|T}^{[\ell]}$ from (11). Hatted parameters stem from the Stage-I ECME algorithm.

The matrix (13) must be rescaled to be a correlation matrix. This is done as in Pelletier (2006) by setting

$$\boldsymbol{\Gamma}_n^{[\ell+1]} = \mathbf{D}_n^{-1[\ell+1]} \tilde{\boldsymbol{\Gamma}}_n^{[\ell+1]} \mathbf{D}_n^{-1[\ell+1]},$$

for $n = 1, \ldots, N$, with

$$\mathbf{D}_n^{[\ell+1]} = \operatorname{diag}\left(\sqrt{\tilde{\boldsymbol{\Gamma}}_{1,1,n}^{[\ell+1]}}, \ldots, \sqrt{\tilde{\boldsymbol{\Gamma}}_{K,K,n}^{[\ell+1]}}\right).$$

Stage-II CM-2-step: The Markov transition matrix $\boldsymbol{\Pi} = \left(\pi_{n,m}\right)_{n,m=1,\ldots,N}$ is then updated using the formula in Hamilton (1994, eq. (22.4.16)) by

$$\pi_{n,m}^{[\ell+1]} = \sum_{t=2}^{T} \xi_{m,t|T}^{[\ell]} \frac{\pi_{n,m}^{[\ell]} \xi_{n,t-1|t-1}^{[\ell]}}{\xi_{m,t|t-1}^{[\ell]}} \Bigg/ \sum_{t=2}^{T} \xi_{n,t-1|T}^{[\ell]}.$$

These steps are iterated until convergence. This closes the Stage-II EM algorithm.

In every iteration of the two-stage algorithm, $\boldsymbol{\theta}_C^*$ and $\boldsymbol{\theta}_M^*$ are set in Stage-I to the correlation estimates from Stage-II in the previous iteration. In particular, $\boldsymbol{\theta}_P$ still can be estimated from (10) but the Stage-I E-step and Stage-I CM2-step need to be updated with the new $\boldsymbol{\theta}_C^*$, and $\boldsymbol{\theta}_M^*$. As demonstrated in the Appendix, each stage of the algorithm preserves the monotonic increase of the incomplete data likelihood function. Therefore, when iterating between the two stages until convergence, the self-consistency of the EM algorithm guarantees, under standard regularity conditions, that our estimates converge to the maximum likelihood estimates of all parameters. In particular, desirable large sample properties such as consistency, efficiency, and asymptotic normality hold for our two-stage EM estimator.

The gains in computational speed compared with the aforementioned nested EM algorithm are due to the fact that, in our empirical analysis, we find that just the first iteration using $\boldsymbol{\theta}_C^* = \mathbf{I}_K$, and ignoring the correlation structure in the Stage-I algorithm, is usually enough to obtain precise parameter estimates in finite samples. Intuitively, neither the filtered G_t from Stage-I E-step, nor the estimates of $\boldsymbol{\theta}_D$ in the Stage-I CM2-step are strongly impacted by ignoring the correlations. The former, because it depends on the correlation matrix only through the shape of ellipsoids in (9), and the latter because of the aforementioned flat likelihood function problem with respect to the GIG parameters.

We note that, even under the normality assumption, the efficiency and asymptotic normality of the two-stage procedure in Pelletier (2006) are not guaranteed. Instead, the author uses the two-stage estimates as initial values for a numerical optimizer of the full likelihood function, as suggested earlier in Pagan (1986).

4. Empirical application

Our data set consists of the daily returns of $K = 29$ components of the Dow Jones Industrial Index (DJ30) from June 1st 1999 until December 31st 2014. The constituents are based on the DJ30 composition in July 2015, with VISA being removed because the stock was not listed before March 18th 2008. Returns are computed as continuously compounded percentage returns, given by $y_{k,t} = 100 \log\left(p_{k,t}/p_{k,t-1}\right)$, where $p_{k,t}$ is the split- and dividend-adjusted price of asset k at time t.

We compare various models, namely the CCC model of Bollerslev (1990), the DCC model of Engle (2002), and the RSDC model of Pelletier (2006), all denoted with a prefix MN- for Multivariate Normal distribution, against numerous MGHyp-based models proposed in this paper: MALap and MLap, both under CCC and RSDC correlation structures, and the same for NIG, SNIG, MAt and Mt. The scale term dynamics from (3) are used in all models except for MGHyp-IID, which we include in the analysis for completeness. We begin in Section 4.1 by comparing the in-sample fit of several models based on the

full DJ30 data set with 3923 daily observations. In Section 4.2 we consider the probability density forecasting performance across different models based on a large number of rolling windows to create one-day-ahead forecasts. Section 4.3 inspects the forecasts of which market regime will manifest in the next period. An economic application is given in Section 4.4, where the rolling window approach is employed to derive forecasts of Value-at-Risk (VaR). Finally, in Section 4.5, we turn to a financial application by conducting out-of-sample portfolio optimization based on the density forecasts of the competing models.

Summarizing the empirical results, the COMFORT-RSDC model yields a valuable prediction of which market regime will get realized next period, and delivers the best in-sample fit and one-step-ahead density forecasts among all considered models. In particular, we find (i) a large improvement moving from the normal to the MGHyp distribution, both under CCC and RSDC; (ii) the best performing subclass of the MGHyp distribution is the symmetric Student-t (Mt) (as defined near the end of Section 2); (iii) shrinkage estimation of the correlation matrices improves density forecasts; (iv) forecasting performance is rather insensitive to the choice of the shrinkage strength above a certain threshold; (v) asymmetry in the MGHyp distribution does not improve forecasts: In almost all cases, the elliptical distributions outperform the asymmetric ones in density and risk measure forecasting. As expected, the accuracy of the VaR forecasts is largely determined by the heavy-tailed nature of the return distribution and the volatility clustering effect, but not by the correlation dynamics. However, use of the latter results in lower VaR forecasts during turbulent times as compared to the CCC models while causing the same violation rates. Such results could have meaningful economic benefits for financial institutions, e.g., potentially lower regulatory capital requirements for positions exposed to market risk. In the portfolio optimization context, combining the MGHyp distribution with regime switching correlations improves risk-adjusted returns and reduces extreme losses. The new model outperforms all its special cases and the equally weighted portfolio. Moreover, the regime forecasts of the RSDC models can be used to effectively avoid heavy losses during market downturns.

4.1. In-sample performance

The in-sample fit of the various models under consideration for the entire DJ30 data set of 3923 daily returns is presented in Table 1. The second column compares the attained maximum of the log-likelihood function, whereas the fourth and fifth columns report the Akaike and the Bayesian information criteria (AIC and BIC, respectively) to account for model complexity. The table is sorted by AIC in descending order. When considering either AIC or BIC, we find that all non-Gaussian models are superior to their Gaussian counterparts. Noteworthy is that the three symmetric special cases of the MGHyp distribution considered here only have one additional parameter compared to the normal distribution, this being enough to add great flexibility by means of the mixture structure. Based on AIC, the Mt-RSDC model is best, and all COMFORT-RSDC models outperform the corresponding CCC models. The ranking changes under the BIC measure, which penalizes large numbers of parameters more heavily. In particular, under BIC, the COMFORT-CCC models are now preferred over their RSDC counterparts, due to having almost half as many parameters.

It is remarkable that all COMFORT-CCC and -RSDC models are superior to the Gaussian-RSDC model of Pelletier (2006) under any measure of in-sample fit. Moving from the normal distribution to any special case of the MGHyp distribution strongly enhances model fit; this finding being in line with those of McNeil et al. (2015) and Paolella and Polak (2015). While in terms of likelihood and AIC, all RSDC models are consistently better than the corresponding CCC models, this improvement is largest for the Gaussian distribution. Once we use the more flexible MGHyp distribution, it becomes more difficult to improve in-sample fit by better modeling the correlations. In all COMFORT-CCC and -RSDC models, the best MGHyp-subclass is the Student-t family, whereas the NIG family is superior when assuming i.i.d. returns.

Finally, the in-sample model fit facilitates the choice of the number of regimes to optimally explain and forecast the dynamics of conditional correlations for the given data set. Theoretically most problematic is the case where the true number of regimes is $N = 1$, and one attempts to estimate a model with $N = 2$ or higher. This concern, in turn, is mitigated by the fact that, first, the entire model is – as with all models for financial asset returns – misspecified, and our paradigm should instead be viewed as an approximation, judged on its ability for efficacy in terms of out-of-sample predictive performance and practicality in terms of numerical optimization stability. Second, the improvement in terms of in-sample fit and out-of-sample performance (see below) when moving from $N = 1$ to $N = 2$ is very strong, suggesting that $N = 1$ is more misspecified than the $N = 2$ case. Moreover, the stylized facts of multivariate financial asset returns data over substantial periods of time strongly suggest movements in the conditional correlation, so that any model that allows for time-varying conditional correlations, whether it be DCC or our RSDC model, should be preferred to the simpler CCC setting.

To investigate the question whether $N = 3$ would be an even better choice, we fit the Gaussian-RSDC and (one special case of) the COMFORT-RSDC model also with three correlation regimes, abbreviated MN-RSDC-3 and Mt-RSDC-3, respectively, in Table 1. In both cases, as expected, the attained maximum-likelihood is higher under $N = 3$, but the AIC is improved only for the Gaussian-RSDC model and not for the COMFORT-RSDC model. Referring to the BIC, using $N = 3$ regimes in both RSDC models generates a poorer fit than using $N = 2$. We conclude that there is no clear improvement in using $N = 3$ regimes, and thus focus on $N = 2$ in the following out-of-sample analysis.

Table 1
Comparison of the in-sample fit using 3923 daily returns of 29 companies of the DJ30 index, from 01.06.1999 to 31.12.2014.

Model	*Log-Likelihood*	*Parameters*	*AIC*	*BIC*
Mt-RSDC	−182404.3	931	**366670.6**	372512.1
Mt-RSDC-3	−**182042.5**	1341	366767.0	375181.3
MAt-RSDC	−182495.4	960	366910.9	372934.2
SNIG-RSDC	−182658.0	931	367177.9	373019.3
NIG-RSDC	−182723.4	960	367366.8	373390.2
MLap-RSDC	−182883.3	931	367628.7	373470.1
MALap-RSDC	−182873.7	960	367667.4	373690.8
Mt-CCC	−183382.3	523	367810.6	**371092.1**
MAt-CCC	−183354.6	552	367813.1	371276.6
SNIG-CCC	−183697.9	523	368441.8	371723.2
NIG-CCC	−183677.5	552	368459.0	371922.5
MLap-CCC	−183809.5	523	368665.1	371946.5
MALap-CCC	−183793.8	552	368691.6	372155.0
SNIG-IID	−186337.7	**465**	373605.3	374346.4
NIG-IID	−186329.8	494	373647.7	374434.9
MLap-IID	−186859.6	**465**	374649.3	375390.3
MALap-IID	−186851.7	494	374691.3	375478.6
MN-RSDC-3	−186799.5	1340	376279.1	384687.0
MN-RSDC	−187264.6	930	376389.2	382224.4
Mt-IID	−187944.8	**465**	376819.6	377560.7
MAt-IID	−187938.1	494	376864.1	377651.4
MN-CCC	−190673.4	522	382390.7	385665.9

4.2. Out-of-sample density forecasting

The main practical relevance centers on the quality of one-step-ahead predictions of the multivariate probability density function of returns, as this is a crucial input for risk prediction and portfolio optimization. Furthermore, considering out-of-sample predictions naturally overcomes the problem of data overfitting because the model performance is evaluated at a data point that is not part of the training set.

Evaluating the quality of forecasts of entire probability distributions instead of estimates of, say, the mean, has gained in popularity both in the statistics literature, as well as in numerous applications; see Elliott and Timmermann (2008), Timmermann (2000) and Tay and Wallis (2000) for an overview of economic and financial applications, as well as Amisano and Giacomini (2007) for an associated statistical test. To assess the out-of-sample predictive ability of a model we use the normalized sum of the realized predictive log-likelihood (short RPLL), also called the predictive log-score, as proposed and discussed in Weigend and Shi (2000), Paolella and Polak (2015a) and Paolella (2015). It is given by

$$S_T(\mathcal{M}) = \frac{1}{T}\sum_{t=1}^{T} \log f_{t+1|t}^{\mathcal{M}}\left(\mathbf{Y}_{t+1}|\widehat{\boldsymbol{\theta}}\right), \tag{14}$$

where $f_{t+1|t}^{\mathcal{M}}$ indicates the forecasted density of $\mathbf{Y}_{t+1}$ using information up to and including time t, $\mathcal{M}$ indicates the model, and thus $f_{t+1|t}^{\mathcal{M}}\left(\mathbf{Y}_{t+1}|\widehat{\boldsymbol{\theta}}\right)$ is the forecasted density evaluated at the realized return at time $t+1$, $\mathbf{Y}_{t+1}$, based on model $\mathcal{M}$. This measure puts equal emphasize on the entire distribution, whereas using weighted log-likelihoods, one can put the focus on certain regions of the density, for instance the tails, as discussed in Amisano and Giacomini (2007) and Gneiting and Ranjan (2011). We abstain from using such weighting schemes, as they require to integrate over the multivariate density function, which, in high dimensions, is computationally costly.

For the RSDC models, the correlation matrix of next day's returns is stochastic because the latent market regime follows a Markov chain. Hence, we extend the above definition of $S_T(\mathcal{M})$ by replacing (14) with a conditional expectation analogue, namely

$$S_T(\mathcal{M}) = \frac{1}{T}\sum_{t=1}^{T} \log\left(\sum_{n=1}^{N} \xi_{n,t+1|t}\, f_{t+1|t}^{\mathcal{M}}\left(\mathbf{Y}_{t+1}|\Delta_{t+1}=n;\widehat{\boldsymbol{\theta}}\right)\right), \tag{15}$$

where $\xi_{n,t+1|t}$ is the nth element of the vector $\boldsymbol{\xi}_{t+1|t}$, given in (11). Using the predictive log-likelihood (15), we compare the forecasting performance of the aforementioned models. Using the DJ30 data set, we estimate each model on 2922 daily rolling windows, each consisting of 1000 data points, such that the forecasting period begins on May 23rd 2003.

In line with numerous forecasting applications that advocate the use of shrinkage estimation to enhance out-of-sample forecasting performance, see, e.g., Diebold and Li (2006), we do so in the Stage-II EM algorithm, where estimation of the regime-specific correlation matrices $\boldsymbol{\Gamma}_n$, $n=1,\ldots,N$, is enhanced by the quasi-Bayesian shrinkage estimator (13).

For the shrinkage targets $\mathbf{B}_n$, $n=1,\ldots,N$, in (13), we refer to the one-step estimated COMFORT-CCC model, by using the resulting CCC correlation matrix estimate $\widehat{\boldsymbol{\Gamma}}_{\text{CCC}}$ as our target. For parsimony, we use the same target $\mathbf{B}_n = \widehat{\boldsymbol{\Gamma}}_{\text{CCC}}$ for

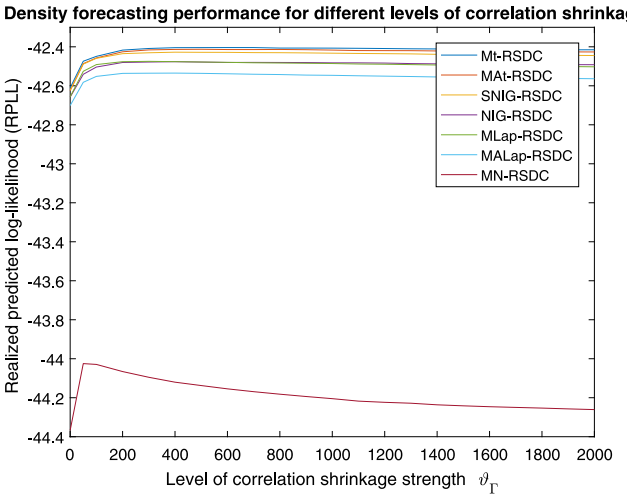

Fig. 1. Realized predictive log-likelihood measures $S_T(\mathcal{M})$ for different models $\mathcal{M}$, plotted as a function of the correlation shrinkage strength parameter $0 \leq \vartheta_\Gamma \leq 2000$, data set consisting of 29 stocks of DJ30, out-of-sample period from 23.05.2003 to 31.12.2014.

all correlation matrices $\mathbf{\Gamma}_n$, but we impose different levels of shrinkage strength $a_n \geq 0$ for the various regimes. In our empirical analysis we consider the case $N = 2$ and couple the shrinkage strengths using a hyper-parameter: We set $a_1 = \omega\vartheta_\Gamma$ and $a_2 = \vartheta_\Gamma/\omega$, with $\vartheta_\Gamma \geq 0$ a scalar strength hyper-parameter and $\omega > 0$ a scalar parameter determining the level of similarity of both regimes. Less prior strength is assigned to the second regime correlation matrix, allowing more flexibility in the correlation matrix and potentially resulting in higher correlations during turbulent market periods. In a multivariate finite mixture model, Paolella (2015) uses the value $\omega = 2$. We do not optimize over this parameter, but instead choose a priori $\omega = 3$ to allow for larger spreads in correlations between regimes and to accommodate the fact that we use the same target for both correlation matrices (which is not the case in Paolella, 2015). Other choices of shrinkage targets are feasible and might well be superior to our target, e.g., one could shrink towards the unit matrix or a matrix of ones, akin to the DECO model Engle and Kelly (2012).

Thus the only parameter to choose is the shrinkage strength hyper-parameter $\vartheta_\Gamma \geq 0$. Our objective for the optimal value ϑ_Γ^* is to maximize the out-of-sample density forecasting performance, measured by the RPLL, denoted $S_T(\mathcal{M})$ in (15). This is done in a computationally intense grid-search, spanning a wide range of values $\vartheta_\Gamma \geq 0$. Fig. 1 presents the results by plotting the RPLL against the level of correlation shrinkage ϑ_Γ for all RSDC models. We find that the optimal shrinkage strength ϑ_Γ^* depends on the specific distribution but the difference in optimal strengths across the various non-Gaussian MGHyp distributions is small. Furthermore, the forecasting ability is quite robust to levels of $\vartheta_\Gamma \geq 400$, as shown by the concave shape of the curves flattening out very slowly for higher levels of shrinkage. This allows the statistician to choose a priori a well-functioning, albeit not optimal, level of correlation shrinkage strength in order to avoid data snooping. The CCC structure constitutes the limiting case for $\vartheta_\Gamma \to \infty$ only when the CCC correlation matrix is chosen for both shrinkage targets $\mathbf{B}_1$ and $\mathbf{B}_2$. In addition to improving density forecasts, correlation shrinkage has the pleasant effect of significantly reducing computational time of the Stage-II EM algorithm. For example, estimation time can be reduced by up to a factor of three by using shrinkage.

Using the optimal level of correlation shrinkage ϑ_Γ^* for each RSDC model according to the grid search, we compare the density prediction capability in Table 2. We also include the i.i.d. case of all MGHyp distributions considered herein and the Gaussian-DCC model of Engle (2002). Overall the best model is the symmetric Student-RSDC and we observe that the biggest gain is achieved by replacing the normal with any of the MGHyp distributions. Within the MGHyp distribution family, the Student-t subclass is best for the conditional models CCC and RSDC. Most notably, all RSDC models deliver significantly better density forecasts than the corresponding CCC models. The improvement from the RSDC extension is largest for the Gaussian case but the Gaussian-RSDC model is still vastly inferior to the worst performing COMFORT-CCC model. Moreover, the weakest COMFORT-RSDC model outperforms the best COMFORT-CCC model, demonstrating the importance of dynamic correlation modeling for multivariate density forecasting. Comparing the two Gaussian models with dynamic conditional correlations, we see that the Gaussian-RSDC is better suited for density forecasting than the Gaussian-DCC model, which performs equally as the Gaussian-CCC due to its restrictive functional form of correlation dynamics.

The fact that all MGHyp-IID models are more potent for density forecasting than the Gaussian-CCC, -DCC and -RSDC models, underlines the utmost importance of non-Gaussian distributions for asset return modeling. Among the MGHyp-IID models, the SNIG distribution performs best, which is contrary to the conditional setting. This can be explained by the well-known fact that GARCH processes induce excess kurtosis, even when a thin-tailed innovation distribution is used. Hence the optimal tail thickness of the innovation process will differ between unconditional and conditional models.

Table 2
Comparison of out-of-sample density prediction performance measured by realized predictive log-likelihood, data set consisting of 29 stocks of DJ30, out-of-sample period 23.05.2003 - 31.12.2014; RSDC models with optimal levels of correlation shrinkage $\vartheta^*_{\mathbf{\Gamma}}$.

Rank	*Model*	*RPLL*	$\vartheta^*_{\mathbf{\Gamma}}$
1	Mt-RSDC	**−42.4038**	600
2	MAt-RSDC	−42.4127	500
3	SNIG-RSDC	−42.4276	400
4	MLap-RSDC	−42.4746	300
5	NIG-RSDC	−42.4765	400
6	MALap-RSDC	−42.5349	400
7	Mt-CCC	−42.6861	–
8	MAt-CCC	−42.7166	–
9	SNIG-CCC	−42.7251	–
10	MLap-CCC	−42.7601	–
11	NIG-CCC	−42.7617	–
12	MALap-CCC	−42.8010	–
13	SNIG-IID	−43.1385	–
14	NIG-IID	−43.1608	–
15	MLap-IID	−43.3109	–
16	MALap-IID	−43.3337	–
17	Mt-IID	−43.5238	–
18	MAt-IID	−43.5448	–
19	MN-RSDC	−44.0250	50
20	MN-DCC	−44.6677	–
21	MN-CCC	−44.6721	–

Noteworthy is also that all elliptical MGHyp distributions outperform their asymmetric variants in this out-of-sample analysis. For out-of-sample forecasting in the MGHyp class, the simpler elliptical distributions appear to be at least as capable as those allowing for asymmetric margins. In fact, several ways of shrinkage estimation of the vector of asymmetry parameters $\boldsymbol{\gamma}$ did not improve the forecasting quality beyond that of the elliptical MGHyp distributions.

Comparison of the COMFORT-CCC models in Table 2 with the left-end points in the plots in Fig. 1 shows that, even without correlation shrinkage, the COMFORT-RSDC models do better than their CCC counterparts. This means that regime switching correlation dynamics improve the out-of-sample forecasts even without the quasi-Bayesian prior (13), but the predictive performance can be further improved with it; the latter result also holds for the Gaussian-RSDC model of Pelletier (2006), as seen from Fig. 1. Hence as a byproduct of our work, we provide a straightforward method to improve the Gaussian-RSDC model considerably by shrinkage estimation of the correlation matrices.

Referring to the discussion of the optimal number of regimes in the RSDC models, for completeness, we compare the forecasting performance of the Gaussian-RSDC and COMFORT-RSDC models with $N = 2$ and $N = 3$ under different levels of correlation shrinkage. We find that three regimes are able to forecast equally well, but not better, as two regimes, with the important drawback of $N = 3$ being that one has to strongly increase the correlation shrinkage strength. Hence, while we use more regimes, these must be forced to be more similar to each other. As $N = 3$ does not increase the forecasting performance, but requires heavier shrinkage and considerably more computation time, we again conclude that the $N = 2$ case is sufficient. In fact, the improved density forecasts, when the number of free parameters in the case $N = 2$ is restricted by means of shrinkage estimation, indicate that the model complexity under $N = 2$ is sufficiently high.

The main findings from this out-of-sample density forecasting exercise support the conclusion that the best forecasts are achieved by combining the distributional flexibility of the MGHyp family with the correlation dynamics of a Markov regime switching model, along with judicious use of shrinkage estimation.

4.3. Market regime forecasting

Another object of interest are the out-of-sample forecasts of regime probabilities, i.e., the one-step-ahead forecasts of which regime will occur next. More precisely, we make a prediction at time t, based on a rolling window of the last 1000 observations, of the probability that a given market regime will occur at time $t + 1$. Using the optimal level of correlation shrinkage, Fig. 2 presents the out-of-sample forecasts $\boldsymbol{\xi}_{t+1|t}$ of the probabilities governing which state is going to realize next in the Gaussian-RSDC and the symmetric Student-RSDC model. The actual estimates $\hat{\boldsymbol{\xi}}_{t+1|t}$ are plotted as dots and a local regression smoother is added to facilitate interpretation. Comparing the two figures, we see that the forecasted state probabilities under the Gaussian-RSDC and the COMFORT-RSDC models are similar, but the two models predict different regimes towards the end of the sample period.

Remarkably, the forecasts of changes in market regimes inferred from the COMFORT-RSDC model are closely linked to historical economic events in the US stock markets. This validates the interpretation that the first regime represents a calm market and the second regime a volatile or crisis market period. In the right plot of Fig. 2 we can identify seven time frames that match specific periods of market conditions. The first such period starts in mid-2003, lasts until summer

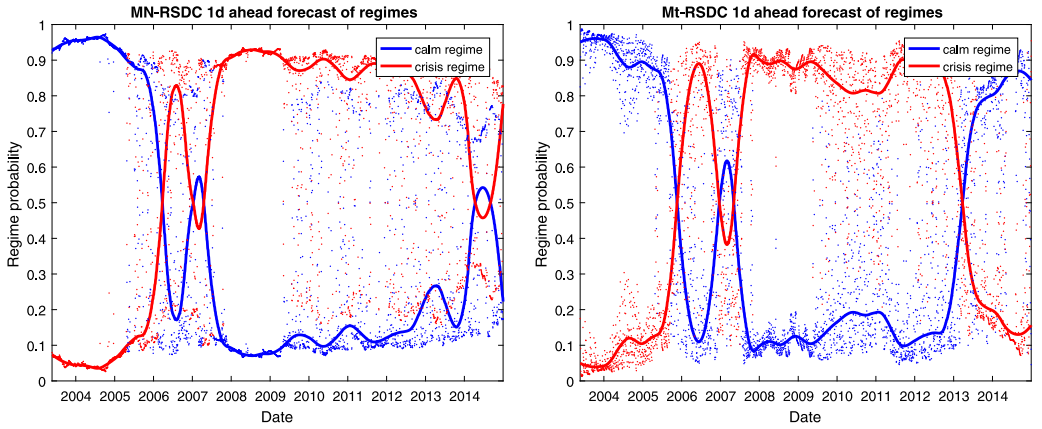

Fig. 2. *Left*: One-day-ahead forecasts of regimes of the Gaussian-RSDC model for the data set consisting of 29 stocks of the DJ30, out-of-sample period from 23.05.2003–31.12.2014. The plot shows the forecasted probabilities of the two regimes under optimal correlation shrinkage $\vartheta_{\Gamma}^{*} = 50$. *Right*: Same plot for the symmetric Student-RSDC model with optimal correlation shrinkage $\vartheta_{\Gamma}^{*} = 600$.

2005, and is characterized by a low probability of a crisis, due to the ongoing bull market during that time. From the end of 2005 until mid 2006, the probability of a crisis goes up, reflecting an increasing nervousness in the market. From mid 2006 until mid 2007, market volatility decreases amid a strong bull market. Then, from mid 2007, when Bear Stearns had to close down several of its hedge funds, until mid 2009, the global financial crisis manifests itself in a strongly dominant and stable crisis regime. The probability of the crisis ending the following day is (except for the two isolated dots in mid 2008), always estimated at less than 20% during these two years. The next time frame is spanned from fall 2009 until the end of 2011, and is characterized by an increase in the probability of a recovery and a less clear separation of regimes. The high uncertainty of this period stems from the recovery in the financial markets, whereas at the same time the real economy was still struggling heavily, casting doubt on the further performance of equity markets. During the years 2011 and 2012, the regime forecasts predict a resurgence of the financial crisis. This fear was spurred by the massive debt crisis in Europe, which reached its peak when Ireland had to be bailed out in November 2010 and Portugal had to be rescued in April 2011. The struggle to keep Greece solvent and within the euro zone went on until fall 2012, when the ECB signaled to undertake unlimited debt purchases, and the euro zone members agreed on a joint liability for Greece's debt. Finally, starting in spring 2013, the calm regime dominates again, because of the strong bull markets fueled by central bank money worldwide and the fading of the euro crisis. Interestingly, the Gaussian-RSDC model does not capture this bull market and instead predicts a resurgence of the volatile regime that did not take place. The plots for other non-Gaussian MGHyp distributions used in the COMFORT-RSDC model look very similar to one presented here.

4.4. Value-at-risk forecasting and backtesting

We now apply the new COMFORT-RSDC model to forecasting risk measures, in particular the Value-at-Risk (VaR). One major use of GARCH models is to produce concise short-term forecasts of tail risk measures. Santos et al. (2013) compares univariate and multivariate models for this task and favors the multivariate approach. Various regime switching models, both in the univariate and the multivariate case, are successfully applied to VaR estimation in Billio and Pelizzon (2000). Concentrating on the prevalent risk measure of VaR, we analyze the quality of daily VaR forecasts based on the conditional density forecasts derived from the various models. The forecast quality is evaluated by means of the ubiquitous Christoffersen (1998) backtest, which is still the industry standard despite the development of newer backtest procedures, such as Christoffersen et al. (2001), Christoffersen and Pelletier (2004), Haas (2005), and recently Pelletier and Wei (2016).

Much of the earlier literature on VaR forecasting made use of univariate models applied to the returns induced by a given portfolio strategy or, quite often, just to a stock market index; see, e.g., Kuester et al. (2006), Santos et al. (2013), Pelletier and Wei (2016) and Slim et al. (2017), and the references therein. More recently, non-Gaussian multivariate models have also been used; see, e.g., Santos et al. (2013), who show that, among the various univariate and multivariate models they considered, the latter tended to perform better; while Härdle et al. (2015) propose and study a copula-based Markov switching multivariate model and illustrate its favorable performance with VaR prediction.

Herein, we compare the VaR prediction performance of the newly proposed COMFORT-RSDC model to its various special cases, such as the CCC and Gausian-based models. Although it is well-known that unconditional models are not suitable for risk measure forecasting, for completeness we include the SNIG-IID model, which performed well for density forecasting. Our goal is not a horse race among all competing VaR prediction methods, and thus is less ambitious than Kuester et al. (2006) and Santos et al. (2013), as our emphasis is on the study and efficacy of a new model. Anticipating

Table 3
Backtest results of 95% daily VaR forecasts of the equally weighted portfolio, DJ30 data set with out-of-sample period spanning 23.05.2003 - 31.12.2014, rolling window size of 1000, RSDC models without correlation shrinkage.

95%**VaR**								
Model	*Stat. CC*	*p-value CC*	*Stat. UC*	*p-value UC*	*Stat. IND*	*p-value IND*	*Violations*	*Coverage Rate*
MALap-CCC	**0.2414**	**0.8863**	0.0095	0.9222	0.2318	0.6302	145	0.0496
MLap-CCC	0.2697	0.8738	**0.0002**	**0.9898**	0.2696	0.6036	**146**	**0.0499**
NIG-CCC	0.4384	0.8032	0.3740	0.5408	0.0643	0.7998	139	0.0476
MLap-RSDC	0.4575	0.7955	0.0581	0.8095	0.3994	0.5274	149	0.0510
NIG-RSDC	0.7434	0.6896	0.6844	0.4081	0.0590	0.8080	156	0.0534
SNIG-CCC	0.7781	0.6777	0.7589	0.3837	**0.0192**	**0.8897**	136	0.0465
MAt-CCC	0.7972	0.6713	0.2434	0.6217	0.5537	0.4568	152	0.0520
Mt-CCC	0.9436	0.6239	0.3331	0.5639	0.6106	0.4346	153	0.0523
SNIG-RSDC	1.0385	0.5950	0.6844	0.4081	0.3541	0.5518	156	0.0534
MAt-RSDC	1.3825	0.5010	1.1576	0.2820	0.2248	0.6354	159	0.0544
MN-CCC	1.5109	0.4698	0.1059	0.7449	1.4050	0.2359	150	0.0513
MN-DCC	1.5109	0.4698	0.1059	0.7449	1.4050	0.2359	150	0.0513
MN-RSDC	1.5109	0.4698	0.1059	0.7449	1.4050	0.2359	150	0.0513
MALap-RSDC	1.5620	0.4579	0.6844	0.4081	0.8776	0.3489	156	0.0534
Mt-RSDC	2.3981	0.3015	1.9745	0.1600	0.4236	0.5152	163	0.0558
SNIG-IID	13.7050	0.0011	3.0613	0.0802	10.6437	0.0011	126	0.0431

the results discussed in more detail below, we find that the largest gains in VaR forecasting performance are due to the heavier tails of the MGHyp distribution, irrespective of the correlation structure used. While COMFORT-RSDC does not generally improve the outcome of the Christoffersen (1998) VaR backtest compared to COMFORT-CCC, it is economically more frugal by producing lower VaR forecasts during periods of high market volatility.

We use our data set of 29 stocks from the DJ30 and rolling windows consisting of 1000 observations each. The out-of-sample period starts on 23.05.2003. Robustness checks on other data sets and subsamples lead to similar findings.

For the RSDC models, we do not make use of correlation shrinkage estimation discussed in Section 4.2 because we find this to reduce the quality of VaR forecasts. This does not contradict the improved density forecasts under correlation shrinkage because VaR forecasts concentrate solely on the left tail of the univariate portfolio return distribution, while the predicted log-likelihood measure considers the full probability distribution of the multivariate asset returns.

The backtest is performed for the one-step-ahead daily 95% and 99% VaR forecasts of the equally weighted (or $1/K$) portfolio. The results are presented in Tables 3 and 4, which are sorted by the p-values of the conditional coverage test, as this is a critical measure, e.g., to fulfill regulatory requirements. The p-values of the independence, unconditional coverage, and conditional coverage tests are calculated from the asymptotic chi-square distribution with one, one, and two degrees of freedom, respectively. The asymptotic distribution is a decent approximation because we use a large sample size of 2922 out-of-sample VaR forecasts.

For the 95% VaR, the MALap- and MLap-CCC models perform best, while all models except the SNIG-IID, pass the backtest for all conventional confidence levels. The RSDC specification performs slightly worse than CCC but all RSDC models still easily pass the test. Even for the 95% VaR, the Gaussian assumption is not recommended as most MGHyp-based models perform better; nevertheless all three Gaussian models pass the test easily and perform identically well irrespective of the correlation structure. The superiority of the non-Gaussian models becomes much more striking for 99% VaR forecasts. The Gaussian models, as well as the SNIG-IID, drastically fail the backtest with p-values of the conditional coverage test close to zero due to the vast number of violations. The three Gaussian models deliver identical backtest results because they produce the exact same sequence of VaR violations. In contrast, all COMFORT-CCC and -RSDC models pass the backtest with ease. Overall, the MLap-RSDC, MAt-CCC, MAt-RSDC, and Mt-CCC models perform best. For completeness we also performed the backtests for 99.5% and 99.9% VaR, with the findings being very similar to the 99% VaR case. The results are detailed in the working paper version Paolella et al. (2019).

It is not surprising that the i.i.d. assumption is worst in the independence test for all quantiles; even for high quantiles, the Gaussian-CCC, -DCC and -RSDC models perform better in this respect because of the sensitivity of the GARCH based VaR forecasts to increases in market volatility. Despite its excess kurtosis, the SNIG-IID model fails the backtest for all quantiles considered herein. We find the same for all other MGHyp-based i.i.d. models. We conclude, which is nothing new, that any reasonable model for risk forecasts has to account for volatility clustering, and therefore the i.i.d. assumption should be avoided irrespective of the probability distribution used. The predominance of conditional models, especially those that feature the stylized fact of heavy-tailed financial returns, for risk measure forecasting has been noted, amongst others, in Christoffersen et al. (2001), Kuester et al. (2006) and Righi and Ceretta (2015).

Summarizing the VaR backtest, we find that only those models that combine the flexible, semi-heavy-tailed MGHyp class with an MGARCH structure to account for volatility clustering pass the backtest for all quantiles. Overall, both the COMFORT-CCC and -RSDC class deliver superb forecasts of daily VaR, and the MAt-RSDC, Mt-CCC, and MALap-RSDC models appear to be particularly suitable.

Based on the backtest results in Tables 3 and 4, there is no clear reason to prefer the COMFORT-RSDC model over the simpler CCC structure. All COMFORT-RSDC and CCC models pass the backtest but the best model can either be of the RSDC

Table 4
Same as Table 3 but for 99% daily VaR forecasts.

99% **VaR**								
Model	*Stat. CC*	*p-value CC*	*Stat. UC*	*p-value UC*	*Stat. IND*	*p-value IND*	*Violations*	*Coverage Rate*
MLap-RSDC	**0.6427**	**0.7252**	**0.0203**	**0.8867**	0.6224	0.4302	**30**	**0.0103**
MAt-CCC	**0.6427**	**0.7252**	**0.0203**	**0.8867**	0.6224	0.4302	**30**	**0.0103**
MAt-RSDC	**0.6427**	**0.7252**	**0.0203**	**0.8867**	0.6224	0.4302	**30**	**0.0103**
Mt-CCC	**0.6427**	**0.7252**	**0.0203**	**0.8867**	0.6224	0.4302	**30**	**0.0103**
MLap-CCC	0.7710	0.6801	0.1062	0.7446	0.6648	0.4149	31	0.0106
NIG-CCC	0.8413	0.6566	0.3745	0.5406	0.4669	0.4944	26	0.0089
MALap-CCC	0.9659	0.6170	0.2572	0.6120	0.7087	0.3999	32	0.0109
SNIG-CCC	1.0817	0.5823	0.6502	0.4200	**0.4315**	**0.5113**	25	0.0086
SNIG-RSDC	1.9308	0.3808	1.0822	0.2982	0.8487	0.3569	35	0.0120
MALap-RSDC	2.3733	0.3052	1.4751	0.2245	0.8982	0.3433	36	0.0123
Mt-RSDC	2.3733	0.3052	1.4751	0.2245	0.8982	0.3433	36	0.0123
NIG-RSDC	3.4297	0.1800	2.4283	0.1192	1.0014	0.3170	38	0.0130
SNIG-IID	15.2984	0.0005	7.3781	0.0066	7.9203	0.0049	45	0.0154
MN-CCC	22.2573	0.0000	22.2357	0.0000	0.0216	0.8831	58	0.0198
MN-DCC	22.2573	0.0000	22.2357	0.0000	0.0216	0.8831	58	0.0198
MN-RSDC	22.2573	0.0000	22.2357	0.0000	0.0216	0.8831	58	0.0198

or the CCC type, depending on the VaR level and most likely the data set. However, the Christoffersen (1998) backtest considers merely the number and timing of violations but does not explicitly account for the level of VaR forecasts. Two competing models that produce the same series of VaR violations can have different VaR forecasts and hence could potentially imply different regulatory capital requirements. Naturally, given that a model passes the regulatory backtest, financial institutions prefer a model that is more capital efficient and consequently less costly. We make the simplifying assumption that lower VaR numbers translate directly into lower regulatory capital requirements, i.e., that capital costs are reduced when VaR forecasts are lower. This cost saving is expressed on the same scale as the VaR numbers, so that, for example, a VaR forecast of 7% versus 8% translates into a 1% reduction in regulatory capital. Furthermore, it might be desirable when VaR forecasts are less volatile over time (while still passing the backtest), because a more volatile VaR might imply more variation in regulatory capital requirements or could be disadvantageous for internal risk budgeting.

Fig. 3 shows that the Gaussian-CCC model produces huge VaR forecasts during the financial crisis due to the variance shooting up. Because of the low market liquidity at that time, this can be economically very costly and thus the Gaussian distribution for VaR calculation should be avoided even for lower quantiles, such as the 95% VaR, where the failure rate is still acceptable according to the backtest. More crucially, Fig. 3 shows that our COMFORT-RSDC model produces lower VaR forecasts throughout the financial crisis of 2008–2009 and at the peak of the euro zone crisis in 2011, and hence lower capital costs than the CCC models of approximately the same backtest quality. This comes at the cost of the RSDC model having slightly higher VaR forecasts during calm markets, best seen in the lower plot of Fig. 3. A second advantage of the COMFORT-RSDC model is that it produces smoother VaR forecasts over time compared the CCC model.

Fig. 4 investigates in more detail the timing when use of the COMFORT-RSDC reduces capital requirements over the CCC model. The upper plot shows the relationship between the average GARCH-implied market volatility and the capital savings of the NIG-RSDC over the NIG-CCC model for the 99% VaR. Positive values of capital saving stand for the RSDC model being more capital efficient than the CCC model. The scatter plot and the linear regression line depict a strong positive relationship between the two variables. The linear regression has an adjusted R^2 of 0.798 and the regression coefficient of the market volatility is 1.176. Hence, in periods of volatile markets, the RSDC model helps to save costs by reducing capital requirements. This can be explained by the superior capability of the RSDC model to capture the correlation dynamics, especially in times of distress. By switching to the volatile regime, the RSDC model is able to better account for changes in the dependencies between assets during a crisis. The capital requirement reduction is even more important because, during distressed market periods, liquidity is scarcest and borrowing capital is most costly.

The bottom plot of Fig. 4 compares the capital savings of the RSDC model with the evolution of the costs of borrowing capital on the interbank market. We approximate these costs by considering the three month USD LIBOR rate, which is one of the most important interbank market rates, and accommodate the general level of interest rates by calculating the spread between this LIBOR rate and the effective FED funds rate. All data is gathered from The Federal Reserve Bank of St. Louis. The plot shows a striking congruence of time points when the RSDC model reduces capital requirements and when borrowing capital is costly for a bank, most notably the financial crisis 2008–2009 and the climax of the euro crisis in 2011. When performing an Engle–Granger test, the null hypothesis of no cointegration between the two time series is strongly rejected with a p-value below 0.001. If additionally we consider the St. Louis FED Financial Stress Index instead of the interbank rates, we find that the sharp increases in capital savings of the COMFORT-RSDC model at the end of 2003 and in the second half of 2011 correspond to jumps in this stress index that are not fully reflected in the interbank rates.

It should be noted that, for all non-Gaussian distributions, the COMFORT-RSDC implies capital savings with the same time pattern as depicted in Fig. 4. Moreover, the average level of VaR forecasts over the whole sample period of almost all COMFORT-RSDC models is lower than those of the corresponding CCC models. Hence, even irrespective of different

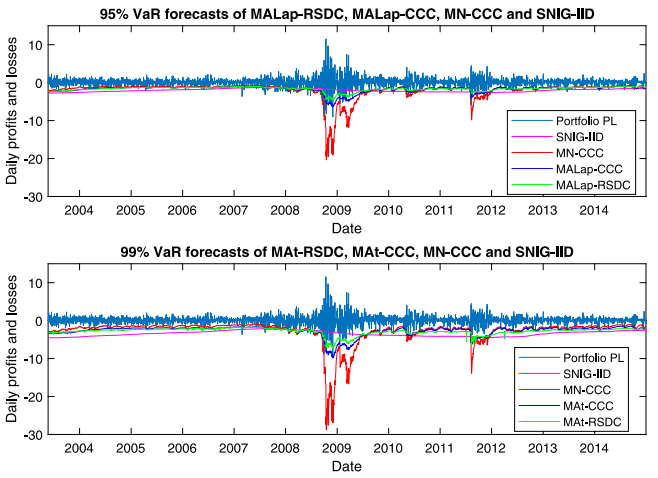

Fig. 3. Daily profits and losses of the $1/K$ portfolio and daily VaR forecasts for several models, *Top:* 95% VaR forecasts of MALap-RSDC, MALap-CCC, SNIG-IID and MN-CCC; *Bottom:* 99% VaR forecasts of MAt-RSDC, MAt-CCC, SNIG-IID and MN-CCC.

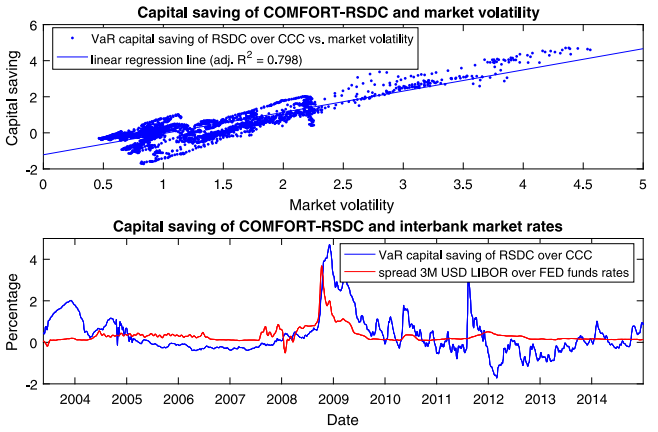

Fig. 4. *Top:* Scatter plot and linear regression line of market volatility vs. cost savings of the NIG-RSDC over the NIG-CCC model for 99% VaR forecasting, *Bottom:* Time variation of cost savings of the NIG-RSDC over the NIG-CCC model for 99% VaR forecasting and of the costs of borrowing capital on the interbank market measured by the spread between the three month USD LIBOR rate over the effective FED funds rate.

market levels of capital costs, the COMFORT-RSDC models can reduce the overall costs associated with regulatory capital requirements, both compared to the COMFORT-CCC class and, more dramatically, compared to Gaussian-based models.

Summarizing our findings, we suggest the COMFORT-RSDC model as a new method for highly accurate VaR forecasts that in addition appears to be particularly capital efficient during times of financial distress and causes less variation in daily VaR forecasts. The most distinct finding, and which is not new, is that the level of excess kurtosis and the ability to account for clustering in volatility are pivotal for good risk measure forecasts.

4.5. Out-of-sample portfolio optimization

Finally, in this section we apply the COMFORT-RSDC model in the context of out-of-sample portfolio optimization. The importance of accounting for market downturns in correlation modeling and portfolio allocation has been nicely highlighted by Greenspan (1999): "[...] joint distributions estimated over periods without panics will misestimate the degree of correlation between asset returns during panics. [...] Consequently, the benefits of portfolio diversification will tend to be overestimated when the rare panic periods are not taken into account."

Again, we form 2922 rolling windows, each of length 1000 days, and use the one-step-ahead density predictions to compute the optimal portfolio weights. Our portfolio optimization methodology is the minimum-expected shortfall (short: min-ES) framework, i.e., in every time step we minimize the expected portfolio shortfall over a one-day holding period, subject to the constraint of being fully invested in the stock market. With respect to the choice of the risk measure, expected shortfall has many desirable properties and takes the tail of the distribution into account; the latter argument being of importance when leaving the Gaussian world. In Paolella and Polak (2015), the min-ES portfolio based on a broad

range of multivariate models is shown to achieve slightly higher Sharpe ratios than the corresponding minimum-variance portfolio.

In our implementation we make use of the approach of Rockafellar and Uryasev (2000, 2002) to translate the task of computing the min-ES portfolio into a convex optimization problem. Let the K-dimensional space of no-short-sale, fully invested portfolios be

$$\mathbf{X} := \{\mathbf{w} : w_i \geq 0 \ \forall i = 1, \ldots, K, |\mathbf{w}| = 1\} \subseteq \mathbb{R}^K,$$

and define an auxiliary function on $\mathbf{X} \times \mathbb{R}$ by

$$H_\alpha(\mathbf{w}, l) = l + \frac{1}{1-\alpha} \int_{-\infty}^{\infty} (-p - l)^+ f_{t+1|t}^P(p; \mathbf{w}) \mathrm{d}p, \tag{16}$$

where $f_{t+1|t}^P(p, \mathbf{w})$ is the one-step-ahead density forecast of the portfolio return distribution based on observations up to time t and under the weight vector $\mathbf{w}$. The auxiliary function (16) is convex and continuously differentiable in $(\mathbf{w}, l)$ so that optimization is computationally fast and delivers the global extremum. The min-ES portfolio is then computed as the global minimum of H_α, i.e.,

$$\min_{\mathbf{w} \in \mathbf{X}} \mathrm{ES}_\alpha(\mathbf{w}) = \min_{(\mathbf{w}, l) \in \mathbf{X} \times \mathbb{R}} H_\alpha(\mathbf{w}, l),$$

see Rockafellar and Uryasev (2000, Theorem 2). This approach does not require computing the VaR in advance, and minimization of this well-behaved auxiliary function is fast even for large K, allowing us to perform an extensive comparison of out-of-sample portfolio performance of the various models under consideration.

The results of the portfolio analysis, sorted by Sharpe ratio, are summarized in Table 5. Comparing risk adjusted returns, we find that all MGHyp-based CCC and RSDC models outperform the corresponding Gaussian models but the gain in Sharpe ratio is moderate. In particular, the Gaussian-RSDC beats the MAt-CCC and SNIG-CCC models, despite having inferior density forecasts. It is noteworthy that the min-ES portfolios based on even the worst performing models, the Gaussian-DCC and -CCC, beat the passive $1/K$ strategy in terms of higher Sharpe ratio, lower portfolio volatility and lower maximum drawdown. Compared to COMFORT-CCC, all COMFORT-RSDC models lead to higher expected returns and Sharpe ratios. An important finding is that the average increase in Sharpe ratio when moving from CCC to RSDC is higher (around 0.1) than the gain from using one of the non-Gaussian MGHyp special cases over a Gaussian distribution. Other measures of risk adjusted returns such as the Sortino and Starr ratios lead to the same conclusion. Another clear advantage of the COMFORT-RSDC over the CCC model is the increase in realized total return of around 15%. Moreover, the COMFORT-RSDC model leads to a reduction in the maximum drawdown of about $3 - 4\%$. These findings indicate that, for portfolio allocation, modeling the dynamic dependency structure between assets is potentially more important than using a non-normal distribution. We point out that the Markov regime switching model used in this study, albeit delivering better portfolio results, is a rather simplistic approach that is likely to leave room for further improvements.

While the benchmark $1/K$ strategy has the highest expected daily return and realized total return, it also has the highest volatility and maximum drawdown, as well as the lowest Sharpe-, Sortino- and Starr-ratios. Passive strategies, such as the equally weighted portfolio or value-weighted market ETFs, have large downside risk in market turmoils and the $1/K$ portfolio underperforms all min-ES portfolios in our analysis in terms of risk-adjusted returns on a long time horizon. In contrast, the min-ES portfolios based on the non-Gaussian COMFORT-CCC and -RSDC models can be regarded as safe investment strategies suitable for risk-averse investors.

Finally and more generally, we observe that better out-of-sample density forecasts do not necessarily translate into better portfolio performance. For example, the Gaussian-RSDC model attains a higher Sharpe ratio and total return than the MAt-CCC and the SNIG-CCC models, although it is vastly inferior in forecasting the return density. The same result is found for the impact of correlation shrinkage on the resulting portfolio performance; in contrast to positive impact on density prediction, which is rather insensitive to the level ϑ_Γ of shrinkage strength, we observe no clear pattern in the impact of correlation shrinkage on the performance as measured by the Sharpe ratio. A plot, analogous to Fig. 1 but having the Sharpe ratio on the vertical axis, shows an erratic, non-monotone relationship between shrinkage strength ϑ_Γ and Sharpe ratio. However, using no shrinkage at all seems to be inferior to using any positive shrinkage strength. To avoid data overfitting we do not search for an optimal, but highly data dependent, shrinkage strength and instead use the level ϑ_Γ^* that is optimal for density forecasting.

Dynamic Risk Control Based on Regime Forecasts

The probability forecasts of regimes from the COMFORT-RSDC model can be utilized for creating a portfolio strategy with the so-called Dynamic Risk Control (DRC) mechanism, where the stock portfolio is liquidated and cash is held if the volatile regime is likely to arise. For this analysis we focus on the Mt-RSDC model but the other cases of the MGHyp distribution perform similarly. In order to avoid excessive switching in and out of the stock market we compute the exponentially weighted moving average (EWMA) of the forecasts of regime probabilities $\hat{\boldsymbol{\xi}}_{t+1|t}$ and use this as signal for the market timing strategy. We hold the min-ES portfolio (computed as above) if and only if the smoothed forecast of the volatile regime is below a threshold $\tau \in [0, 1]$, otherwise we hold cash without interest. In practice one would move to high quality bonds or buy put options as protection. Both the number of lags of the EWMA smoother and the threshold

Table 5
Comparison of min-ES portfolios at 99%-ES; $K = 29$ stocks from DJ30; out-of-sample trading period 23.05.2003 - 31.12.2014; rolling window size 1000 data points; RSDC models with same levels of correlation shrinkage $\vartheta_{\mathbf{\Gamma}}$ as in density forecasting analysis; all returns measured in percent.

Model	*Exp. Daily Return*	*Volatility*	*Total Return*	*Max. Drawdown*	*Sharpe*	*Sortino*	*Starr*
Mt-RSDC	0.0425	0.8725	124.2199	22.8742	**0.7732**	**1.1058**	**0.0115**
MAt-RSDC	0.0426	0.8760	124.5814	23.1305	0.7724	1.1052	**0.0115**
MLap-RSDC	0.0423	0.8750	123.6263	22.5707	0.7673	1.0955	0.0113
NIG-RSDC	0.0425	0.8838	124.2862	22.8413	0.7638	1.0945	0.0114
MALap-RSDC	0.0420	0.8831	122.8223	**22.5410**	0.7554	1.0839	0.0112
SNIG-RSDC	0.0409	0.8715	119.6801	22.6054	0.7458	1.0653	0.0110
NIG-CCC	0.0383	0.8660	111.9449	25.9797	0.7020	1.0025	0.0104
Mt-CCC	0.0373	0.8657	108.9755	27.7583	0.6836	0.9669	0.0100
MALap-CCC	0.0370	**0.8624**	108.2771	25.5303	0.6819	0.9684	0.0101
MLap-CCC	0.0370	0.8648	108.1175	27.7149	0.6790	0.9621	0.0100
MN-RSDC	0.0369	0.8668	107.9421	26.0883	0.6763	0.9543	0.0098
MAt-CCC	0.0364	0.8642	106.4670	26.4901	0.6691	0.9490	0.0098
SNIG-CCC	0.0363	0.8661	106.1463	27.2155	0.6656	0.9416	0.0097
MN-CCC	0.0341	0.8736	99.5647	30.5057	0.6190	0.8752	0.0090
MN-DCC	0.0332	0.8698	96.9519	32.2004	0.6053	0.8538	0.0088
1/K	**0.0443**	1.1816	**129.4294**	42.1190	0.5949	0.8410	0.0085

Table 6
Comparison of min-ES portfolios with and without dynamic risk control (DRC); $K = 29$ stocks from DJ30; out-of-sample trading period 13.10.2004 - 31.12.2014.

Model	*Exp. Daily Return*	*Volatility*	*Total Return*	*Max. Drawdown*	*Sharpe*	*Sortino*	*Starr*
Mt-RSDC + DRC	0.0309	**0.4661**	79.5669	**5.8851**	**1.0527**	**1.5597**	**0.0156**
Mt-RSDC	0.0396	0.8995	102.0194	26.5049	0.6997	0.9987	0.0103
MN-RSDC + DRC	0.0196	0.4670	50.5603	10.8500	0.6676	0.9437	0.0094
MN-RSDC	0.0352	0.8913	90.5023	29.4612	0.6265	0.8833	0.0091
1/K	**0.0412**	1.1691	**120.3210**	48.7481	0.5589	0.7885	0.0079

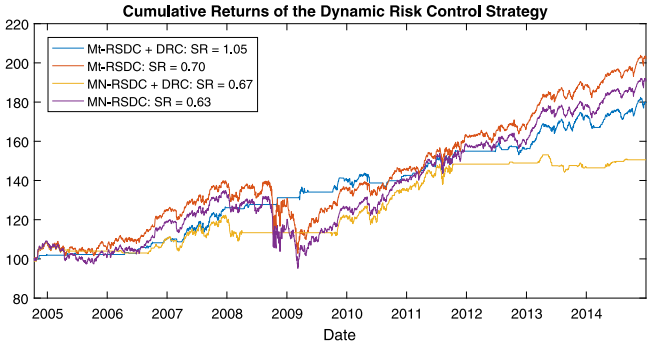

Fig. 5. Cumulative returns of the min-ES portfolio based on COMFORT-RSDC with and without dynamic risk control, same for Gaussian-RSDC, out-of-sample trading period from 13.10.2004 to 31.12.2014.

τ are chosen by cross-validation, using a window size of 250 daily returns, such that the (in-sample) Sharpe ratio is maximized. We emphasize the important aspect here that no future data points are used to compute the DRC signal.

Fig. 5 presents the out-of-sample rolling-window cumulative returns of the COMFORT-RSDC-based min-ES strategy with and without DRC, and the same results for the Gaussian-RSDC model, and Table 6 reports the performance measures. The regime-based DRC mechanism allows to circumvent most losses in the financial crisis, drastically reducing the portfolio volatility and maximum drawdown. The DRC mechanism works best for the non-Gaussian COMFORT-RSDC model because the forecasts of regimes are enhanced compared to the Gaussian setup. In particular, the Gaussian-based models switch slower into the risk-free asset when market troubles arise, thereby not avoiding losses as effectively as the COMFORT-RSDC model. The gain in risk-adjusted returns in the Gaussian case is therefore only around 6%, whereas under the MGHyp-based model the use of the DRC increases the Sharpe ratio by 50%. The improvement of the risk-return profile stems purely from risk reduction, while the cumulative returns under DRC are lower than those of the baseline strategies. It should be noted that the DRC strategy does not use any external information about the stress level of the market, but is based solely on past return data, making the method applicable in markets where a stress index is not available. Finally, the improvements due to DRC are not limited to the acute financial crisis of 2008 but also help to avoid smaller market downturns, as in mid 2010 and in early 2012.

Summarizing the portfolio optimization applications of the COMFORT-RSDC model, we find that using our new model boosts portfolio performance in terms of consistently higher Sharpe ratios and total returns while at the same time reduces the extreme losses during the financial crisis. Overall, the influence of replacing the CCC by an RSDC structure appears to be at least as important for portfolio performance than replacing the Gaussian distribution by a fat-tailed one. This underscores the finding that modeling the dependencies between assets is of utmost importance for asset allocation, and that the best results are achieved by combining the RSDC dynamics with a non-Gaussian MGHyp distribution to account for the excess kurtosis of returns. This reasoning also applies to our dynamic risk control mechanism. It is able to avoid heavy losses in the financial crisis, based on the regime forecasts of the COMFORT-RSDC model.

In order to validate the robustness of our empirical findings with respect to other data sets, we consider three further data sets: (i) the top $K = 100$ firms by market capitalization of the Standard & Poor's 500 (SP500) index from January 2nd 1997 until December 31st 2014; (ii) the $K = 19$ stocks from the Swiss Market Index (SMI) from January 5th 2004 until December 30th 2015, where Julius Baer Group is removed due to the company's restructuring on October 1st 2009; (iii) $K = 6$ foreign exchange (FX) rates against the Swiss Franc from July 1st 2004 until December 31st 2015. All sample periods are chosen as to include the global financial crisis for the out-of-sample analysis, and all data sets are obtained from Bloomberg Professional or from Wharton Research Data Services. The empirical results gained from these additional data sets are fully in line with those obtained from the DJ30 data, confirming the finding that the COMFORT-RSDC class surpasses the popular Gaussian models, as well as the COMFORT-CCC class, in terms of in-sample fit and out-of-sample forecasting ability, and that the regime forecasts of the COMFORT-RSDC model provide genuine value for financial risk and portfolio management. We refer to the working paper version Paolella et al. (2019) of this article for further details.

5. Conclusion

We have introduced a new, parametric, multivariate model class for financial asset returns that incorporates univariate GARCH-type dynamics, non-Gaussian conditional returns, and hidden Markov regime switching dynamics for the correlations. The model is coherent in the sense of being a well-defined stochastic process, as opposed to an ad-hoc assemblage of models estimated sequentially, such as a Gaussian-DCC or -RSDC overlaid with a multivariate Student-t with the latter estimated based on the residuals of the former.

We proposed and tested a likelihood-based estimation procedure that splits the estimation problem into two stages, both of which are amenable to estimation via an EM algorithm. Necessary and sufficient conditions for the consistency and asymptotic normality of the resulting iterative two-stage estimator are given. Out-of-sample model performance is enhanced by shrinkage estimation using a quasi-Bayesian prior on the regime-specific correlation matrices. A welcome side effect of our proposed methodology is, in addition to improved out-of-sample predictive performance, correlation shrinkage often greatly reduces the computation time.

In the empirical analysis, we investigated two non-Gaussian models of conditional correlations: (i) the Constant Conditional Correlation (COMFORT-CCC), and (ii) the Regime Switching Conditional Correlation model (COMFORT-RSDC), for various non-elliptic and elliptic special cases of the MGHyp distribution, including the Gaussian special case. On the basis of in-sample fit and out-of-sample density forecasting performance, all non-Gaussian models outperformed their Gaussian counterparts, including Gaussian-CCC and Gaussian-RSDC. In turn and importantly, all COMFORT-RSDC models outperformed their COMFORT-CCC counterparts, both in- and out-of-sample, with the best overall model being the symmetric Student-RSDC. There is some robustness to our empirical findings, in that various data sets, of different dimensionality and asset classes, were used, and very similar conclusions were found.

In an economic application, we have demonstrated that the new model leads to improved VaR forecasting, in terms of lower capital requirements during financial distress while easily passing regulatory backtests. When compared to the Gaussian-CCC model, the vastly superior risk measure forecasting performance is due to two factors: First, and obviously, addressing the blatant non-Gaussianity of the GARCH-filtered innovation sequences, and, secondly, via improved correlation prediction. The latter implies that applications explicitly centering on correlation predictions could also benefit from our model. We address this in an out-of-sample analysis of portfolio optimization—which is of interest to major financial institutions and large investors such as pension funds.

In a financial application, the COMFORT-RSDC model is shown to deliver superior results in portfolio optimization. Using the min-ES methodology, the new model delivers higher risk-adjusted returns while also minimizing losses in times of financial distress compared to the COMFORT-CCC and Gaussian models. Importantly, the model easily outperforms the equally weighted benchmark portfolio and has preferable risk characteristics for risk-averse investors. Our results highlight that the best results are achieved when a sophisticated model for correlation dynamics is supplemented with a non-Gaussian distribution that accounts for the excess kurtosis of returns. The regime forecasts generated by the COMFORT-RSDC model can be used to detect market crashes and to avoid most losses, even in severe market downturns such as the financial crisis, by shifting the stock portfolio to a risk-free asset. While returns are inevitably lowered, the risk-adjusted performance of this dynamic risk control strategy is vastly enhanced.

New light is also shed on the importance of modeling the skewness of multivariate return series. Comparing the likelihoods of the COMFORT model, the asymmetric cases of the MGHyp distribution perform slightly better in-sample than the elliptical variants, but with respect to AIC and BIC this advantage is lost; similar findings are reported for the i.i.d. case in McNeil et al. (2015). Additionally, in accordance with our findings, the best distribution overall in Hu and

Kercheval (2010) and McNeil et al. (2015) is the symmetric Student-t. In our out-of-sample analysis, the elliptical MGHyp models perform consistently better than their asymmetric counterparts. While returns are almost surely asymmetric (or more generally non-elliptical), e.g., exhibit different and non-zero asymmetry coefficients and tail coefficients (Paolella and Polak, 2015a; Näf et al., 2019), these properties are difficult to capture and more so to forecast. In a multivariate parametric model, due to the proliferation of the parameters and the bias–variance tradeoff, modeling these characteristics can induce too much estimation variance that harms the forecasting performance. For our MGHyp-based model, several approaches of shrinkage estimation of the asymmetry parameters were tried but could not overcome this issue. As such, it seems reasonable to forgo the use of non-elliptical MGHyp models, i.e., set all asymmetry parameters to zero, in this context because meaningful estimation of this parameter vector simply asks too much from the data.

Future research could investigate more general scale term dynamics than the GARCH(1,1)-model used herein, e.g., using APARCH or GJR-GARCH models. While this should improve modeling performance, we expect the advantage to be relatively small compared to the gains achieved when moving from the Gaussian to the MGHyp distribution, and from CCC to RSDC. As the regimes tend to be aligned with low and high volatility markets, another plausible future generalization would be to assume other features of the DGP also to be different between regimes, such as the MGHyp shape parameters, or possibly the persistence of volatility as dictated by the GARCH parameters. Endowing the tail shape parameter(s) with a regime-switching mechanism is shown in Liu (2017) to successfully capture non-linearities in systemic risk contributions of institutions to a financial system, and is likely to be helpful also for modeling asset returns. As such, one could entertain a more general structure allowing for this, though note that doing so is not an immediate, trivial generalization of the framework herein, and the applicability of the two-stage EM algorithm would be jeopardized, thus rendering the model inapplicable for large scale problems.

It should be emphasized that, like Engle (2002) and Pelletier (2006), our model does not explicitly use any exogenous (say, macroeconomic) information. Future work (as kindly and astutely suggested by an anonymous referee on an earlier draft) could explore if the regime shifts are related to the business cycle, policy regimes, or other such structural macroeconomic measures. Having now established the theoretical framework of model estimation, properties, and also demonstration of its efficacy in terms of risk prediction, such considerations will be of interest for future, possibly more applied, work.

Appendix. Convergence of the two-step EM algorithm

We show here that, under the usual assumption that the model reflects the true DGP except for the unknown parameters, the proposed two-stage EM algorithm monotonically increases the incomplete-data likelihood function. The proof is done in three steps. First, in Lemma A.1, we show that ignoring the correlation structure does not change estimates of $\boldsymbol{\theta}_P$ in the Stage-I ECME algorithm. Second, in Theorem A.1, we show that the monotonic increase of the incomplete-data likelihood is preserved when the Stage-I is performed on the subspace of the parameters space with fixed $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$. Finally, in Theorem A.2, we show that the Stage-II EM algorithm also monotonically increases the incomplete-data likelihood function.

The first round of the Stage-I ECME algorithm is performed on the zero correlations subspace. Hence, the Markov chain Δ is omitted from the likelihood functions, $\boldsymbol{\theta}_M$ is not in the parameter space, and $\boldsymbol{\theta}_C = \mathbf{I}_K$ is fixed.

Let the incomplete-data likelihood function be denoted by $L_{\mathbf{Y}}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)$. Then, by Bayes' rule and the complete log-likelihood function decomposition (7),

$$\begin{aligned}\log L_{\mathbf{Y}}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M) &= \log L_{\mathbf{Y},G}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M) - \log k_{G|\mathbf{Y}}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)\\ &= \log L^{\mathrm{MV}}_{\mathbf{Y}|G}(\boldsymbol{\theta}_P) + \log L^{\mathrm{Corr}}_{\mathbf{Y}|G}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)\\ &+ \log L_G(\boldsymbol{\theta}_D) - \log k_{G|\mathbf{Y}}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M), \end{aligned} \tag{17}$$

where $k_{G|\mathbf{Y}}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)$ denotes the conditional probability density function of $G|\mathbf{Y}$. In the Stage-I E-step of the algorithm we take the conditional expectation under the probability measure given by last step estimates of $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_D$ and with fixed $\boldsymbol{\theta}^*_C$ and $\boldsymbol{\theta}^*_M$. Thus we start by taking the expectation of both sides of (17) with respect to the conditional distribution of G_t given $\boldsymbol{Y}_t$, using the ℓth fit of parameters $(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D)$, together with $\boldsymbol{\theta}^*_C$ and $\boldsymbol{\theta}^*_M$ fixed. This results in

$$\begin{aligned}\log L_{\mathbf{Y}}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M) &= \mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}^*_C, \boldsymbol{\theta}^*_M\right)}\left[\log L^{\mathrm{MV}}_{\mathbf{Y}|G}(\boldsymbol{\theta}_P) + \log L^{\mathrm{Corr}}_{\mathbf{Y}|G}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)\,|\Phi_t\right]\\ &+ \mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}^*_C, \boldsymbol{\theta}^*_M\right)}\left[\log L_G(\boldsymbol{\theta}_D)\,|\Phi_t\right]\\ &- \mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}^*_C, \boldsymbol{\theta}^*_M\right)}\left[\log k_{G|\mathbf{Y}}(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M)\,|\Phi_t\right]. \end{aligned} \tag{18}$$

Lemma A.1. *Let ℓ and $\ell+1$ be the consecutive steps in the Stage-I ECME algorithm performed on the fixed correlation subspace $(\boldsymbol{\theta}^*_C, \boldsymbol{\theta}^*_M)$. Then*

$$\boldsymbol{\theta}_P^{[\ell+1]} = \arg\max \mathcal{L}^{\mathrm{MV}}_{\mathbf{Y}|G}(\boldsymbol{\theta}_P) := \arg\max \mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}^*_C, \boldsymbol{\theta}^*_M\right)}\left[\log L^{\mathrm{MV}}_{\mathbf{Y}|G}(\boldsymbol{\theta}_P)\,|\Phi_t\right], \tag{19}$$

maximizes also

$$\mathcal{L}_{\mathbf{Y}|G}\left(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right) := \mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right)}\left[\log L_{\mathbf{Y}|G}^{\mathrm{MV}}\left(\boldsymbol{\theta}_P\right) + \log L_{\mathbf{Y}|G}^{\mathrm{Corr}}\left(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right)|\Phi_t\right]. \tag{20}$$

Proof. We show that both sets of first order conditions with respect to $\boldsymbol{\theta}_P$ are satisfied for the same vector of parameters. Hence the $\ell+1$ step estimates of $\boldsymbol{\theta}_P$ from the fixed correlation model and the full model coincide. Define $\tilde{\varepsilon}_{k,t} = \left(Y_{k,t} - \mu_k - \gamma_k g_t\right)$ and vectors $\tilde{\boldsymbol{\varepsilon}}_t = \left(\mathbf{Y}_t - \boldsymbol{\mu} - \boldsymbol{\gamma} g_t\right)$ and $M_{k,t} = \left[0, \ldots, \left(Y_{k,t} - \mu_k - \gamma_k g_t\right) s_{k,t}^{-1}, \ldots, 0\right]'$. Then $\boldsymbol{\theta}_P^{[\ell+1]}$ are chosen so that:

$$\frac{\partial \mathcal{L}_{\mathbf{Y}|G}^{\mathrm{MV}}\left(\boldsymbol{\theta}_P\right)}{\partial \theta_{1,k,j}} = 0 \quad \Leftrightarrow$$

for $\theta_{1,k,j}$ equal ω_k, α_k or β_k we have

$$\mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right)}\left[-\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \theta_{1,k,j}} + g_t^{-1}\tilde{\varepsilon}_{k,t} s_{k,t}^{-1} s_{k,t}^{-1}\tilde{\varepsilon}_{k,t}\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \theta_{1,k,j}}\right] = 0;$$

for $\theta_{1,k,j}$ equal μ_k we have

$$\mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right)}\left[-\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \mu_k} + g_t^{-1}\tilde{\varepsilon}_{k,t} s_{k,t}^{-1} s_{k,t}^{-1}\left(\tilde{\varepsilon}_{k,t}\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \mu_k} - 1\right)\right] = 0;$$

for $\theta_{1,k,j}$ equal γ_k we have

$$\mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right)}\left[-\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \gamma_k} + g_t^{-1}\tilde{\varepsilon}_{k,t} s_{k,t}^{-1} s_{k,t}^{-1}\left(\tilde{\varepsilon}_{k,t}\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \gamma_k} - g_t\right)\right] = 0.$$

On the other hand, the $\boldsymbol{\theta}_P^{[\ell+1]}$ that maximizes (20) must satisfy

$$\frac{\partial \mathcal{L}_{\mathbf{Y}|G}\left(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right)}{\partial \theta_{1,k,j}} = 0 \quad \Leftrightarrow$$

for $\theta_{1,k,j}$ equal ω_k, α_k or β_k we have

$$\mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right)}\left[-\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \theta_{1,k,j}} + g_t^{-1}\tilde{\boldsymbol{\varepsilon}}_t \mathbf{S}_t^{-1}\boldsymbol{\Gamma}_t^{-1} M_{k,t}\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \theta_{1,k,j}}\right] = 0;$$

for $\theta_{1,k,j}$ equal μ_k we have

$$\mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right)}\left[-\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \mu_k} + g_t^{-1}\tilde{\boldsymbol{\varepsilon}}_t \mathbf{S}_t^{-1}\boldsymbol{\Gamma}_t^{-1}\left(M_{k,t}\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \mu_k} - s_{k,t}^{-1}\right)\right] = 0;$$

for $\theta_{1,k,j}$ equal γ_k we have

$$\mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right)}\left[-\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \gamma_k} + g_t^{-1}\tilde{\boldsymbol{\varepsilon}}_t \mathbf{S}_t^{-1}\boldsymbol{\Gamma}_t^{-1}\left(M_{k,t}\frac{1}{s_{k,t}}\frac{\partial s_{k,t}}{\partial \gamma_k} - s_{k,t}^{-1} g_t\right)\right] = 0.$$

When $\theta_{1,k,j}$ equals ω_k, α_k, or β_k, by applying the trace operator, we get that

$$g_t^{-1}\left(\mathbf{Y}_t - \boldsymbol{\mu} - \boldsymbol{\gamma} g_t\right)\mathbf{S}_t^{-1}\boldsymbol{\Gamma}_t^{-1}\left[0, \ldots, \left(Y_{k,t} - \mu_k - \boldsymbol{\gamma}_k g_t\right) s_{k,t}^{-1}, \ldots, 0\right]'$$

is a random variable with the same mean as $g_t^{-1}\left(Y_{k,t} - \mu_k - \gamma_k g_t\right) s_{k,t}^{-1} s_{k,t}^{-1}\left(Y_{k,t} - \mu_k - \gamma_k g_t\right)$. The analogous result follows for $\theta_{1,k,j}$ equal to μ_k or γ_k.

From this we conclude that the value of $\boldsymbol{\theta}_P^{[\ell+1]}$ that maximizes (19) also maximizes (20). □

Next, we show the monotonicity of the Stage-I ECME algorithm.

Theorem A.1 (*Monotonicity of Stage-I ECME*). *Let ℓ and $\ell+1$ be the consecutive steps in the ECME algorithm performed on the subspace with fixed correlation parameters ($\boldsymbol{\theta}_C^*$, $\boldsymbol{\theta}_M{}^*$), then the incomplete-data likelihood function increases for each iteration of the algorithm, i.e.,*

$$L_{\mathbf{Y}}\left(\boldsymbol{\theta}_P^{[\ell+1]}, \boldsymbol{\theta}_D^{[\ell+1]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right) \geq L_{\mathbf{Y}}\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right). \tag{21}$$

Proof. The CM1-step of the algorithm maximizes the first term in (18) and does not change the second term. The CM2-step finds $\boldsymbol{\theta}_D^{[\ell+1]}$ such that

$$\log L_{\mathbf{Y}}\left(\boldsymbol{\theta}_P^{[\ell+1]}, \boldsymbol{\theta}_D^{[\ell+1]}, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right) \geq \log L_{\mathbf{Y}}\left(\boldsymbol{\theta}_P^{[\ell+1]}, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C^*, \boldsymbol{\theta}_M^*\right), \text{ for all } \boldsymbol{\theta}_D.$$

For the last term in (18), by Jensen's inequality and the concavity of the logarithmic function, for all $\boldsymbol{\theta} = (\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C)$,

$$\begin{aligned}
&\mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right)}\left[\log k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}\right)|\boldsymbol{\Phi}_t\right] - \mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right)}\left[\log k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right)|\boldsymbol{\Phi}_t\right] \\
&= \mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right)}\left[\log \frac{k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}\right)}{k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right)}|\boldsymbol{\Phi}_t\right] \\
&\leq \log \mathbb{E}_{\left(\boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right)}\left[\frac{k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}\right)}{k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right)}|\boldsymbol{\Phi}_t\right] \\
&= \log \int \frac{k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}\right)}{k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right)} k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}_P^{[\ell]}, \boldsymbol{\theta}_D^{[\ell]}, \boldsymbol{\theta}_C^{*}, \boldsymbol{\theta}_M^{*}\right) \mathrm{d}g \\
&= \log \int k_{G_t|\boldsymbol{\Phi}_t}\left(g|\mathbf{y}_t; \boldsymbol{\theta}\right) \mathrm{d}g = 0.
\end{aligned}$$

Combining these three arguments proves (21) and establishes monotonicity of the Stage-I ECME algorithm. □

Finally, we show that the Stage-II EM algorithm also monotonically increases the incomplete-data log-likelihood. This EM algorithm estimates $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$, and is performed conditionally on Stage-I ECME estimates $\widehat{\boldsymbol{\theta}}_P$ and $\widehat{\boldsymbol{\theta}}_D$. In the E-step of the algorithm we take the conditional expectation under the probability measure given by the estimates of $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_D$ and with the last update of $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$. Therefore, we start by taking the expectation of both sides of (17) with respect to the conditional distribution of G_t given $\boldsymbol{Y}_t$, $\widehat{\boldsymbol{\theta}}_P$, and $\widehat{\boldsymbol{\theta}}_D$, and using the ℓth fit of parameters in $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$. This results in

$$\begin{aligned}
\log L_{\mathbf{Y}}\left(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right) &= \mathbb{E}\left[\log L_{\mathbf{Y}|G}^{\mathrm{MV}}\left(\boldsymbol{\theta}_P\right)|\boldsymbol{\Phi}_t\right] + \mathbb{E}\left[\log L_{\mathbf{Y}|G,\Delta}^{\mathrm{Corr}}\left(\boldsymbol{\theta}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right)|\boldsymbol{\Phi}_t\right] \\
&+ \mathbb{E}\left[\log L_G\left(\boldsymbol{\theta}_D\right)|\boldsymbol{\Phi}_t\right] - \mathbb{E}\left[\log k_{G|\mathbf{Y},\Delta}\left(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right)|\boldsymbol{\Phi}_t\right]. \\
&+ \mathbb{E}\left[\log L_\Delta\left(\boldsymbol{\theta}_M\right)|\boldsymbol{\Phi}_t\right] - \mathbb{E}\left[\log k_{\Delta|\mathbf{Y}}\left(\boldsymbol{\theta}_P, \boldsymbol{\theta}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right)|\boldsymbol{\Phi}_t\right],
\end{aligned} \tag{22}$$

where all the expectations are taken conditionally on $\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right)$.

Theorem A.2 (*Monotonicity of Stage-II EM*). *Let ℓ and $\ell+1$ be the consecutive steps in the second-stage EM algorithm performed conditionally on the Stage-I estimates $\widehat{\boldsymbol{\theta}}_P$ and $\widehat{\boldsymbol{\theta}}_D$. Then the incomplete-data likelihood function increases for each iteration of the algorithm, i.e.,*

$$L_{\mathbf{Y}}\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell+1]}, \boldsymbol{\theta}_M^{[\ell+1]}\right) \geq L_{\mathbf{Y}}\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right).$$

Proof. In each iteration, conditional on the Stage-I estimates $\widehat{\boldsymbol{\theta}}_P$ and $\widehat{\boldsymbol{\theta}}_D$, we update only the parameters in $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$. From (22), $\log L_{\mathbf{Y}}\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right)$ depends on $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$ only through

$$\begin{aligned}
&\mathbb{E}_{\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right)}\left[\log L_{\mathbf{Y}|G,\Delta}^{\mathrm{Corr}}\left(\widehat{\boldsymbol{\theta}}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right) + \log L_\Delta(\boldsymbol{\theta}_M)|\boldsymbol{\Phi}_t\right], \\
&\mathbb{E}_{\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right)}\left[\log k_{G|\mathbf{Y},\Delta}\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right)|\boldsymbol{\Phi}_t\right], \text{ and} \\
&\mathbb{E}_{\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right)}\left[\log k_{\Delta|\mathbf{Y}}\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right)|\boldsymbol{\Phi}_t\right].
\end{aligned}$$

First, note that, under no correlation shrinkage,

$$\boldsymbol{\theta}_C^{[\ell+1]} = \arg\max_{\boldsymbol{\theta}_C} \mathbb{E}_{\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right)}\left[L_{\mathbf{Y}|G,\Delta}^{\mathrm{Corr}}\left(\widehat{\boldsymbol{\theta}}_P, \boldsymbol{\theta}_C, \boldsymbol{\theta}_M\right)|\boldsymbol{\Phi}_t\right],$$

where $\hat{\mathbf{e}}_{n,t} = g_{n,t}^{-1/2}\mathbf{S}_{n,t}^{-1}\boldsymbol{\varepsilon}_{n,t}$, and $\boldsymbol{\varepsilon}_{n,t} = \mathbf{y}_t - \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\gamma}} g_{n,t}$ are the regime-specific residuals with $g_{n,t}$ filtered in the E-step-2 of Stage-II algorithm.

Second, following the derivation in Hamilton (1990),

$$\boldsymbol{\theta}_M^{[\ell+1]} = \arg\max_{\boldsymbol{\theta}_M} \mathbb{E}_{\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right)}\left[\log L_{\mathbf{Y}|\mathbf{G},\boldsymbol{\Delta}}\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M\right) + \log L_{\boldsymbol{\Delta}}\left(\boldsymbol{\theta}_M\right)|\boldsymbol{\Phi}_t\right].$$

Then, using Jensen's inequality and the same argument as in the proof of the Stage-I ECME algorithm,

$$\mathbb{E}_{\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right)}\left[\log k_{G|\mathbf{Y},\Delta}\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell+1]}, \boldsymbol{\theta}_M^{[\ell+1]}\right) - \log k_{G|\mathbf{Y},\Delta}\left(\widehat{\boldsymbol{\theta}}_P, \widehat{\boldsymbol{\theta}}_D, \boldsymbol{\theta}_C^{[\ell]}, \boldsymbol{\theta}_M^{[\ell]}\right)|\boldsymbol{\Phi}_t\right] \leq 0$$

for all $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$, and

$$\mathbb{E}_{\left(\widehat{\boldsymbol{\theta}}_P,\widehat{\boldsymbol{\theta}}_D,\boldsymbol{\theta}_C^{[\ell]},\boldsymbol{\theta}_M^{[\ell]}\right)}\left[\log k_{\Delta|\mathbf{Y}}\left(\widehat{\boldsymbol{\theta}}_P,\widehat{\boldsymbol{\theta}}_D,\boldsymbol{\theta}_C^{[\ell+1]},\boldsymbol{\theta}_M^{[\ell+1]}\right)-\log k_{\Delta|\mathbf{Y}}\left(\widehat{\boldsymbol{\theta}}_P,\widehat{\boldsymbol{\theta}}_D,\boldsymbol{\theta}_C^{[\ell]},\boldsymbol{\theta}_M^{[\ell]}\right)|\boldsymbol{\Phi}_t\right]\leq 0$$

for all $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_M$. This completes the proof. □

The consequence of Theorems A.1 and A.2 is that, if one performs Stage-I and Stage-II of our two-stage EM algorithm iteratively, if $\vartheta_{\mathbf{\Gamma}}=0$, and if the starting values for both algorithms, $\left(\boldsymbol{\theta}_P^{[1]},\boldsymbol{\theta}_D^{[1]}\right)$ and $\left(\boldsymbol{\theta}_C^{[1]},\boldsymbol{\theta}_M^{[1]}\right)$, are sufficiently close to the true parameters, or if the log likelihood is unimodal in the parameter space, then monotonicity in the likelihood values of the estimates $\left(\boldsymbol{\theta}_P^{[\ell]},\boldsymbol{\theta}_D^{[\ell]}\right)$ and $\left(\boldsymbol{\theta}_C^{[\ell]},\boldsymbol{\theta}_M^{[\ell]}\right)$ guarantees their convergence to the corresponding maximum likelihood estimates. Thus, the iterative two-stage EM algorithm, under the aforementioned assumptions, converges to the global maximum of the incomplete data likelihood function with respect to $(\boldsymbol{\theta}_P,\boldsymbol{\theta}_D,\boldsymbol{\theta}_C,\boldsymbol{\theta}_M)$. If correlation shrinkage is employed, i.e., in case $\vartheta_{\mathbf{\Gamma}}>0$, then the objective function is altered along the lines of Hamilton (1991), and the proof follows with the likelihood $L_{\mathbf{Y}}$ replaced by the new objective function as in Hamilton (1994, 22.3.11).

References

Aas, K., Haff, I.H., Dimakos, X.K., 2005. Risk estimation using the multivariate normal inverse Gaussian distribution. J. Risk 8 (2), 39–60.

Aielli, G.P., Caporin, M., 2014. Variance clustering improved dynamic conditional correlation MGARCH estimators. Comput. Statist. Data Anal. 76, 556–576.

Amisano, G., Giacomini, R., 2007. Comparing density forecasts via weighted likelihood ratio tests. J. Bus. Econom. Statist. 25 (2), 177–190.

Bauwens, L., Hafner, C.M., Rombouts, J.V.K., 2007. Multivariate mixed normal conditional heteroskedasticity. Comput. Statist. Data Anal. 51 (7), 3551–3566.

Bauwens, L., Laurent, S., Rombouts, J.K.V., 2006. Multivariate GARCH models: A survey. J. Appl. Econometrics 21 (1), 79–109.

Bianchi, M.L., Tassinari, G.L., Fabozzi, F.J., 2016. Riding with the four Horsemen and the multivariate normal tempered stable model. Int. J. Theor. Appl. Finance 19 (04), 1650027.

Billio, M., Caporin, M., 2005. Multivariate Markov switching dynamic conditional correlation GARCH representations for contagion analysis. Stat. Methods Appl. 14 (2), 145–161.

Billio, M., Caporin, M., Gobbo, M., 2006. Flexible dynamic conditional correlation multivariate GARCH models for asset allocation. Appl. Financ. Econ. Lett. 2 (2), 123–130.

Billio, M., Pelizzon, L., 2000. Value-at-risk: A multivariate switching regime approach. J. Empir. Financ. 7 (5), 531–554.

Black, F., 1976. Studies of stock price volatility changes, In: Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section, pp. 177–181.

Bollerslev, T., 1990. Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH approach. Rev. Econ. Stat. 72, 498–505.

Bollerslev, T., 2010. Glossary to ARCH (GARCH). In: Bollerslev, T., Russell, J., Watson, M. (Eds.), Volatility and Time Series Econometrics: Essays in Honor of Robert Engle. Oxford University Press, Oxford, pp. 137–163.

Cappiello, L., Engle, R.F., Sheppard, K., 2006. Asymmetric dynamics in the correlations of global equity and bond returns. J. Financ. Econ. 4 (4), 537–572.

Chevallier, J., Goutte, S., 2015. Detecting jumps and regime switches in international stock markets returns. Appl. Econ. Lett. 22 (13), 1011–1019.

Chollete, L., Heinen, A., Valdesogo, A., 2009. Modeling international financial returns with a multivariate regime-switching copula. J. Financ. Econ. 7 (4), 437–480.

Christoffersen, P.F., 1998. Evaluating interval forecasts. Internat. Econom. Rev. 841–862.

Christoffersen, P.F., Hahn, J., Inoue, A., 2001. Testing and comparing value-at-risk measures. J. Empir. Financ. 8 (3), 325–342.

Christoffersen, P.F., Pelletier, D., 2004. Backtesting value-at-risk: A duration-based approach. J. Financ. Econ. 2 (1), 84–108.

Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. J. Econometrics 130 (2), 337–364.

Elliott, G., Timmermann, A., 2008. Economic forecasting. J. Econ. Lit. 46 (1), 3–56.

Engle, R.F., 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. J. Bus. Econom. Statist. 20, 339–350.

Engle, R.F., 2009. Anticipating Correlations: A New Paradigm for Risk Management. Princeton University Press, Princeton.

Engle, R.F., Kelly, B., 2012. Dynamic equicorrelation. J. Bus. Econom. Statist. 30 (2), 212–228.

Fink, H., Klimova, Y., Czado, C., Stöber, J., 2017. Regime switching vine copula models for global equity and volatility indices. Econometrics 5 (1), 3.

Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold-and quantile-weighted scoring rules. J. Bus. Econom. Statist. 29 (3), 411–422.

Greenspan, A., 1999. New challenges for monetary policy - A Symposium Sponsored by the Federal Reserve Bank of Kansas City, Opening Remarks of the Chairmain of the Board of Governors of the Federal Reserve System.

Haas, M., 2005. Improved duration-based backtesting of value-at-risk. J. Risk 8 (2), 17—38.

Haas, M., Liu, J.-C., 2018. A multivariate regime-switching GARCH model with an application to global stock market and real estate equity returns. Stud. Nonlinear Dyn. Econ. 22 (3), 1–27.

Haas, M., Mittnik, S., Paolella, M.S., 2004. A new approach to Markov switching GARCH models. J. Financ. Econ. 2 (4), 493–530.

Haas, M., Mittnik, S., Paolella, M.S., 2009. Asymmetric multivariate normal mixture GARCH. Comput. Statist. Data Anal. 53 (6), 2129–2154.

Haas, M., Paolella, M.S., 2012. Mixture and regime–switching GARCH models. In: Bauwens, L., Hafner, C.M., Laurent, S. (Eds.), Handbook of Volatility Models and their Applications. John Wiley & Sons, Inc., Hoboken, New Jersey.

Hamilton, J.D., 1990. Analysis of time series subject to changes in regime. J. Econometrics 45 (1), 39–70.

Hamilton, J.D., 1991. A quasi-Bayesian approach to estimating parameters for mixtures of normal distributions. J. Bus. Econom. Statist. 9 (1), 21–39.

Hamilton, J.D., 1993. In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Estimation, Inference, and Forecasting of Time Seris Subject to Changes in Regime. In: Handbook of Statistics, vol. 11, North-Holland, New York.

Hamilton, J.D., 1994. Time Series Analysis. Princton University Press.

Härdle, W.K., Okhrin, O., Wang, W., 2015. Hidden Markov structures for dynamic copulae. Econometric Theory 31 (05), 981–1015.

Henry, O.T., 2009. Regime switching in the relationship between equity returns and short-term interest rates in the UK. J. Bank. Financ. 33 (2), 405–414.

Hu, W., Kercheval, A.N., 2010. Portfolio optimization for student t and skewed t returns. Quant. Finance 10 (1), 91–105.

Jondeau, E., Poon, S.-H., Rockinger, M., 2007. Financial Modeling Under Non-Gaussian Distributions. Springer, London.
Kasch, M., Caporin, M., 2013. Volatility threshold dynamic conditional correlations: An international analysis. J. Financ. Econ. 11 (4), 706–742.
Kim, C.-J., 1994. Dynamic linear models with Markov-switching. J. Econometrics 60 (1–2), 1–22.
Kuester, K., Mittnik, S., Paolella, M.S., 2006. Value–at–risk prediction: A comparison of alternative strategies. J. Financ. Econ. 4, 53–89.
Liu, X., 2017. Measuring systemic risk with regime switching in tails. Econ. Model. 67, 55–72.
McNeil, A.J., Frey, R., Embrechts, P., 2015. Quantitative Risk Management: Concepts, Techniques and Tools, Revised ed. In: Princeton Series in Finance, Princeton University Press.
Näf, J., Paolella, M.S., Polak, P., 2019. Heterogeneous tail generalized COMFORT modeling via Cholesky decomposition. J. Multivariate Anal. 172, 84–106.
Pagan, A., 1986. Two stage and related estimators and their applications. Rev. Econom. Stud. 53 (4), 517–538.
Paolella, M.S., 2007. Intermediate Probability: A Computational Approach. John Wiley & Sons, Ltd.
Paolella, M.S., 2015. Multivariate asset return prediction with mixture models. Eur. J. Financ. 21 (13–14), 1214–1252.
Paolella, M.S., Polak, P., 2015. Portfolio selection with active risk monitoring, Swiss Finance Institute Research Paper Series No. 15–17.
Paolella, M.S., Polak, P., 2015a. ALRIGHT: Asymmetric large-scale (I) GARCH with hetero-tails. Int. Rev. Econ. Finance 40, 282–297.
Paolella, M.S., Polak, P., 2015b. COMFORT: A common market factor non-Gaussian returns model. J. Econometrics 187 (2), 593–605.
Paolella, M.S., Polak, P., Density and risk prediction with non-gaussian COMFORT models, 2017, Mimeo.
Paolella, M.S., Polak, P., Walker, P.S., 2019. A flexible regime switching model for asset returns, Swiss Finance Institute Research Paper No. 19–27.
Pelletier, D., 2006. Regime switching for dynamic correlations. J. Econometrics 131, 445–473.
Pelletier, D., Wei, W., 2016. The geometric-VaR backtesting method. J. Financ. Econ. 14 (4), 725.
Prause, K., 1999. The Generalized Hyperbolic Model: Estimation, Financial Derivatives, and Risk Measures (Ph.D. thesis). University of Freiburg.
Protassov, R., 2004. EM-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed lambda. Stat. Comput. 14 (1), 67–77.
Righi, M.B., Ceretta, P.S., 2015. A comparison of expected shortfall estimation models. J. Econ. Bus. 78, 14–47.
Rockafellar, R., Uryasev, S., 2000. Optimization of conditional value-at-risk. J. Risk 2, 21–42.
Rockafellar, R., Uryasev, S., 2002. Conditional value-at-risk for general loss distributions. J. Bank. Financ. 26 (7), 1443–1471.
Santos, A.A.P., Nogales, F.J., Ruiz, E., 2013. Comparing univariate and multivariate models to forecast portfolio value–at–risk. J. Financ. Econ. 11 (2), 400–441.
Slim, S., Koubaa, Y., BenSaida, A., 2017. Value-at-risk under Lévy GARCH models: Evidence from global stock markets. J. Int. Financ. Mark. Inst. Money 46, 30–53.
So, M.K.P., Yip, I.W.H., 2012. Multivariate GARCH models with correlation clustering. J. Forecast. 31 (5), 443–468.
Tay, A.S., Wallis, K.F., 2000. Density forecasting: A survey. J. Forecast. 19 (4), 124–143.
Timmermann, A., 2000. Density forecasting in economics and finance. J. Forecast. 19 (4), 231–234.
Tse, Y.K., Tsui, A.K.C., 2002. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. J. Bus. Econom. Statist. 20 (3), 351–362.
Urga, G., Cajigas, J.-P., Ghalanos, A., 2011. Dynamic conditional correlation models with asymmetric multivariate laplace innovations, 2011, Working Paper.
Virbickaite, A., Ausin, M.C., Galeano, P., 2016. A Bayesian non-parametric approach to asymmetric dynamic conditional correlation model with application to portfolio selection. Comput. Statist. Data Anal. 100, 814–829.
Weigend, A.S., Shi, S., 2000. Predicting daily probability distributions of S&P500 returns. J. Forecast. 19 (4), 375–392.
Wu, L., Meng, Q., Velazquez, J.C., 2015. The role of multivariate skew-student density in the estimation of stock market crashes. Eur. J. Financ. 21 (13–14), 1144–1160.