Dynamic Adaptive Mixture Models with an Application to Volatility and Risk

Leopoldo Catania[a,*]

[a] *Department of Economics and Business Economics, Aarhus University and CREATES*

Abstract

In this paper we propose a new class of dynamic mixture models (DAMMs) being able to sequentially adapt the mixture components as well as the mixture composition using information coming from the data. The information driven nature of the proposed class of models allows to exactly compute the full likelihood and to avoid computer intensive simulation schemes. Specific models for financial data are developed starting from the general specification. These models nest many specifications already available in the literature. The properties of the new class of models are discussed through the paper and a large–scale application in quantitative risk management using US equity data is reported.

Keywords: Dynamic Mixture Models, Score–Driven models, Adaptive Models, Quantitative Risk Management.

1. Introduction

Mixtures of distributions are extremely diffused parametric tools used to model non–Gaussian shapes that usually characterize empirical data. A great level of flexibility can be achieved in Mixture Models by appropriate choices of the mixture components distributions. Moreover, the mixture components can also be adaptive with respect to the new information as in linear and generalized Mixture Models, see e.g. Bishop (2006). Within the context of Mixture Models, also the mixture composition can be allowed to evolve over time, this class of models is usually identified as Dynamic Mixture Models, see e.g. Yu (2012). Dynamic Mixture Models have been successfully applied in process monitoring (Yu, 2012), intervention detections (Gerlach et al., 2000), insurance losses (Frigessi et al., 2002), and graphical engineering (KaewTraKulPong and Bowden, 2002; Xie et al., 2005). A drawback of these models is that, when nonlinear non–Gaussian specifications are assumed for the mixture components and for the evolution of the mixture composition, classical inference cannot be applied anymore, see e.g. Gerlach et al. (2000). Usual solutions relay on computer intensive Markov Chain Monte Carlo (MCMC) simulation schemes to carry out Bayesian inference which highly reduces the attractiveness of such models and their implementation in commercial softwares, see e.g. Gerlach et al. (2000), Yu (2012), and Billio et al. (2013).

In this paper, we follow a different approach to model the time evolution of the mixture component distributions as well as the mixture composition in a fully observation–driven framework (Cox, 1981).

*Department of Economics and Business Economics, Aarhus BSS, Fuglesangs Allé 4, DK–8210, Aarhus V. leopoldo.catania@econ.au.dk, Phone: +45 87165536, web page: http://pure.au.dk/portal/en/leopoldo.catania@econ.au.dk.

We build our model starting from recent advances in Score–Driven models for which the likelihood function is available in closed form as the product of conditional densities, see e.g. Creal et al. (2013) and Harvey (2013). Within the class of Score–Driven models, the latent dynamic parameters are updated using a forcing variable based on the score of the conditional distribution. In our context, the mixture components can be chosen to be any parametric distribution with the possibility of allowing for time variation in the full set of parameters of each component. We also allow for the mixture composition to be sequentially updated using the information contained in data. We call this class of models Dynamic Adaptive Mixture Models (DAMMs) given their high flexibility in terms of possible dynamic parametric assumptions and their ability to sequentially adapt the mixture composition. Starting from the general DAMM specification, we derive several models particularly suited for financial econometrics applications. We show that, in the case of dynamic Gaussian mixture components, some specifications already introduced in the literature such as the Mixture of ARCH models proposed by Wong and Li (2001), and the Mixture of GARCH models independently proposed by Haas et al. (2004a), Alexander and Lazar (2006), and Zhang et al. (2006) are obtained as special cases. However, once we depart from the mixture of Gaussian specifications, a variety of new models is obtained. In order to demonstrate the flexibility of DAMMs, we perform an extensive Monte Carlo experiment composed by two parts with an emphasis on financial econometrics applications. The first part focuses on filtering several artificial patterns assumed for the correlation of a bivariate random variable. The second experiment is analogous to the first one but focuses on time–varying mixture weights. We find that DAMM exhibits very high filtering abilities compared to alternative specifications. The supplementary material accompanying this paper reports additional Monte Carlo experiments including an extensive study on the finite sample properties of the Maximum Likelihood estimator for DAMM specifications with Student's t mixture components.

The paper also contributes from an empirical perspective. Indeed, we report a large scale empirical study on the log–returns time series of 403 US domiciled companies covering the last sixteen years of data from early 2000 to mid–2016. The analysis is divided in two parts. In the first part, we report an in–sample study where we compare the new models with some alternatives recently proposed in the financial econometrics literature. We also study the impact of allowing for time–varying mixture composition. The second part focuses on quantitative risk management. Specifically, we predict the Value–at–Risk and the Expected–Shortfall risk measures associated with one day, one week, and one month ahead conditional distributions of financial returns. Results are reported also for volatility prediction of each series as well as for risk measure and density predictions of 403 portfolios of two assets.

The paper is organized as follows. Section 2 illustrates why dynamic mixtures can be relevant for financial time series. Section 3 describes the general DAMM model and details the updating mechanism for the mixture component distributions and the mixture composition. Section 4 focuses on DAMM specifications for financial time series. Section 5 reports the Monte Carlo experiments. Sections 6 reports the empirical application. Finally, Section 7 concludes and reports some suggestions for future research.

2. Dynamic Mixtures for Financial Time Series

Most of the stylized facts characterizing financial time series can be summarized as: "The conditional distribution of financial data exhibits a high degree of variability over time." The Autoregressive Conditional Density model of Hansen (1994) has been one of the first models that directly addressed this issue in a unified framework. A different approach has been to employ a

mixture of distributions. The first model of this kind specifically formulated for financial data has been proposed by Vlaar and Palm (1993) in their application to the weekly exchange rates in the European monetary system. More recent examples are the Mixture Autoregressive Conditional Heteroscedastic Model of Wong and Li (2001) and Wong et al. (2009), the Mixed Normal Conditional Heteroscedastic model of Haas et al. (2004a), Alexander and Lazar (2006) and Zhang et al. (2006),[1] its extension with time–varying mixture probabilities (Haas et al., 2013) as well as the multivariate versions of Bauwens et al. (2007) and Galeano and Ausín (2010). As will be clear from Section 4, the DAMM with Gaussian mixture components nests most of these models, however, a much higher flexibility is achieved when mixtures of non–Gaussian distributions are employed.

It is also worth to mention that the Markov–switching Autoregressive Conditional Heteroscedasticity model, introduced by Cai (1994) and Hamilton and Susmel (1994) and generalized by Gray (1996) and Klaassen (2002) and finally reformulated by Haas et al. (2004b), implicitly assumes that the conditional distribution of financial returns is a mixture of distributions with time–varying weights given by the model–implied predictive probabilities, see Hamilton (1989) for details on Markov–switching models. Further details about the relation between Mixture and Markov–switching Autoregressive Conditional Heteroscedastic Models are reported in Haas and Paolella (2012). With regards to applications in finance, one may ask if these kinds of very flexible models are required by the data. To empirically answer this question, let consider the time series of financial returns for all the components of the Standard & Poors 500 index from January 3, 2000 to October 21, 2016.[2] This data set is used for the empirical application in risk management in Section 6 where further details are reported. Consider the GARCH(1,1) model of Bollerslev (1986) with mixture innovations for the univariate return at time t, y_t:

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned} \tag{1}$$

where ε_t is distributed as a mixture of two standardized Student's t distributions, i.e.:

$$\varepsilon_t \sim \omega \mathcal{T}(0, 1, \nu_1) + (1 - \omega)\mathcal{T}(0, 1, \nu_2),$$

where $\omega \in (0, 1)$, and $\nu_1 > \nu_2$ is imposed for identification purposes. Note that the mixture components are parametrized such that $\mathbb{E}[\varepsilon_t^2] = 1$. This model incorporates the two very well known stylized facts of financial returns: i) conditional heteroscedasticity and ii) fat tails of the conditional distribution, see e.g. McNeil et al. (2015). However, it assumes that the shape of the returns innovation, ε_t, is constant over time. This implies that the tails of the conditional distribution are equal during, say, crisis and non crisis periods. This assumption is evidently violated for many empirical series such as the one reported in Figure 1 which plots the log–returns series in percentage points for C.H. Robinson Worldwide Inc. from January 3, 2000 to October 21, 2016. As evident from the figure, the second part of the series is characterized by a higher number of large negative returns, suggesting that the returns distribution has become more left skewed and with fatter tails.

[1] See also the Mixture Memory GARCH model of Li et al. (2013) and the Mixture Double Autoregressive model of Li et al. (2017). Note that, in both cases time–varying mixture composition is considered.

[2] We remove the assets for which the time series is not complete, leaving us with 403 series.

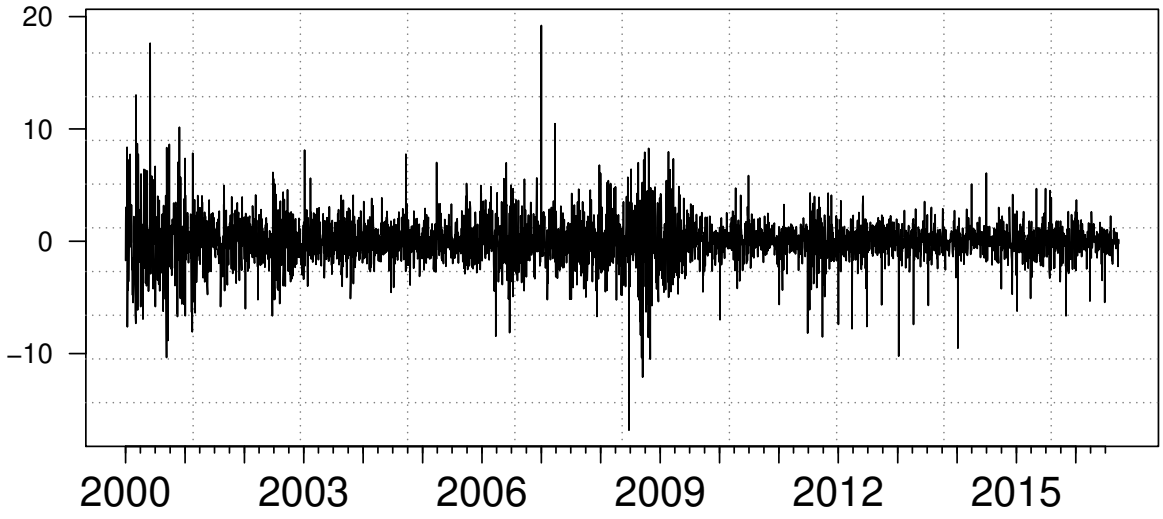


Figure 1: Series of log–returns in percentage points for C.H. Robinson Worldwide Inc. from January 3, 2000 to October 21, 2016 for a total of 4,210 observations. Log–returns are constructed starting from the time series of prices obtained from Datastream.

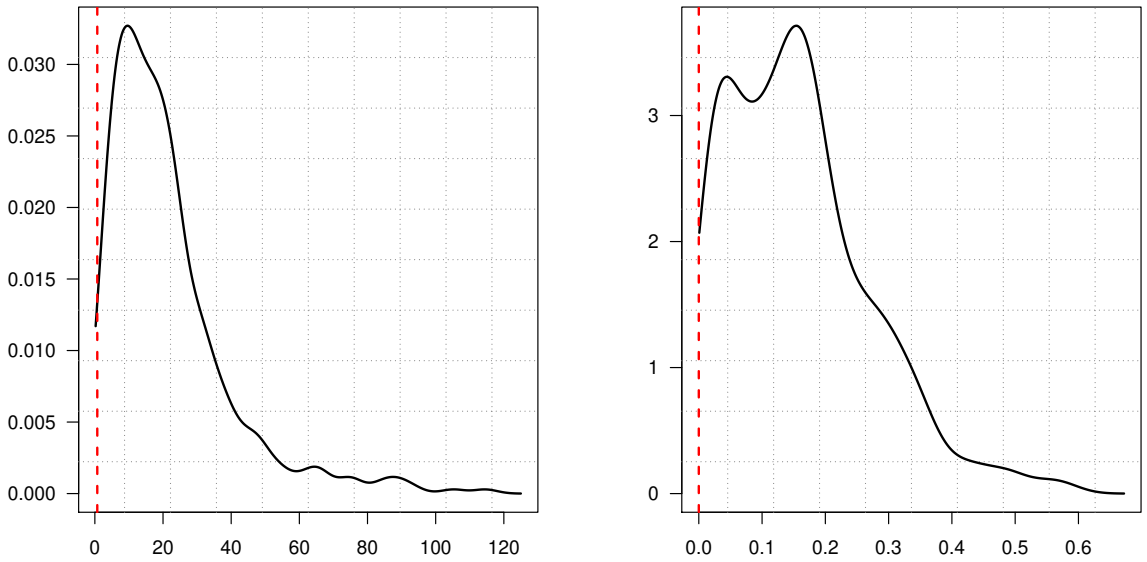
Skewness cannot be accounted for by model (1), hence we only focus on the tails of the distribution. A possible way to account for this empirical evidence is by allowing the shape parameters ν_1 and ν_2 to change over time. However, this strategy can be problematic since a very low portion of the data is informative about these parameters, see for example the discussion in Zhu and Galbraith (2010). A more natural way is to allow ω to evolve over time, a strategy implemented by Haas et al. (2013) in their generalization of the Mixture of GARCH models proposed by Haas et al. (2004a) to which model (1) relates, see also Li et al. (2013) and Li et al. (2017). Let us consider the alternative distributional assumption for ε_t:

$$\varepsilon_t \sim \omega_t \mathcal{T}(0,1,\nu_1) + (1-\omega_t)\mathcal{T}(0,1,\nu_2) \tag{2}$$

with:

$$\begin{aligned}\omega_t = \; & \omega^{(1)}\mathbb{1}_{\{t\leq\lfloor T/5\rfloor\}} + \omega^{(2)}\mathbb{1}_{\{\lfloor T/5\rfloor<t\leq\lfloor 2T/5\rfloor\}} + \omega^{(3)}\mathbb{1}_{\{\lfloor 2T/5\rfloor\leq t<\lfloor 3T/5\rfloor\}} + \\ & \omega^{(4)}\mathbb{1}_{\{\lfloor 3T/5\rfloor\leq t<\lfloor 4T/5\rfloor\}} + \omega^{(5)}\mathbb{1}_{\{t>\lfloor 4T/5\rfloor\}},\end{aligned}$$

where $\mathbb{1}_{\{A\}}$ is the indicator function for the event A and $\lfloor x\rfloor$ is the greatest integer less than or equal to x. Model (2) allows the mixture composition to change at fixed time periods and



(a) Likelihood Ratio (b) Mean Absolute Deviation

Figure 2: Figure 2a depicts a Gaussian kernel density estimated over the Likelihood Ratio statistics between the GARCH model with static and dynamic mixture of Student's t innovations. Parameters are estimated by Maximum Likelihood using the log–returns series of the S&P500 constituents for the period spanning from January 3, 2000 to October 21, 2016. The red vertical dashed line indicates the value of the critical region according to the asymptotic χ^2_4 distribution of the test statistics. In three cases the test favors the null specification of static mixture, associated companies ticker are: "R.N", "NEM.N", and "EOG.N". Figure 2b depicts a Gaussian kernel density estimated over the mean absolute deviations between the mixture weights of the static and dynamic mixture model reported in Section 2. The red vertical dashed line indicates the value 0, which coincides with no differences between the mixture composition estimated from the static and dynamic specifications.

hence, to adapt the tails of the distribution to different economic phases.[3] Evidently, when $\omega^{(1)} = \omega^{(2)} = \omega^{(3)} = \omega^{(4)} = \omega^{(5)} = \omega$, we recover the equivalence between the static and dynamic specifications. We estimate both models on the log–returns series of the S&P 500 components and perform a standard likelihood ratio test.

Figure 2a reports the distribution of the test statistics computed over the cross section of assets. The red vertical line indicates the critical level implied by the χ^2_4 asymptotic distribution of the test. We find that for almost all the series we observe statistical evidence of change in the mixture composition. To investigate the magnitude of the changes in ω, in Figure 2b, we report the distribution of $\frac{1}{5}\sum_{j=1}^{5} |\widehat{\omega}^{(j)} - \widehat{\omega}|$, where $\widehat{\omega}^{(j)}$ and $\widehat{\omega}$ are the ML estimate of $\omega^{(j)}$, $\quad j = 1, \ldots, 5$ and ω across the cross section of log–returns. This quantity indicates the average absolute deviation from the static case of the mixture composition across the S&P500 constituents. As reported in Figure 2b, we find that this quantity is remarkably large for many of the S&P500 constituents.

[3]Here we have arbitrarily chosen to select 5 subperiods. This choice implies that the mixture composition is allowed to change approximately every 3 years. The selected time frame roughly corresponds to two changes of the mixture composition per business cycle according to the NBER's Business Cycle Dating Committee. Results selecting 2, 3, 4, and 6 subperiods are qualitatively the same.

3. Dynamic Adaptive Mixture Models

Let $\mathbf{y}_t \in \Re^d$ be a d–dimensional random vector with conditional density $p\left(\mathbf{y}_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}_t\right)$, where $\mathcal{F}_{t-1}$ is the filtration generated by the process $\{\mathbf{y}_s, s > 0\}$ up to time $t-1$, and $\boldsymbol{\theta}_t$ is a vector of time–varying conditional parameters. We assume $p\left(\cdot\right)$ to be a finite mixture of J real valued conditional distributions, i.e.:

$$p\left(\mathbf{y}_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}_t\right) = \sum_{j=1}^{J} \omega_{j,t} p_j\left(\mathbf{y}_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}_{j,t}\right), \tag{3}$$

with $\omega_{j,t} \in (0,1)$ and $\sum_{j=1}^{J} \omega_{j,t} = 1 \quad \forall \quad t = 1, \ldots$ and $\boldsymbol{\theta}_t = \left(\boldsymbol{\theta}'_{j,t}, \omega_{j,t}, j = 1, \ldots, J\right)'$. Within the class of Dynamic Mixture Models, the mixture component density parameters, $\boldsymbol{\theta}_{j,t} \in \Omega^j$, generally follow a stochastic process. Convenient choices are first order nonlinear autoregression (Billio et al., 2012; Casarin et al., 2015) and Markov Switching processes (Kim, 1994; Kim and Nelson, 1999; Harrison and West, 1999; Ardia, 2008). The latter are usually estimated by particle filters in a Bayesian context, while for the former the Expectation–Maximisation algorithm of Dempster et al. (1977) is frequently employed. In this paper, we follow a different approach and rely on the Score–Driven framework introduced by Creal et al. (2013) and Harvey (2013) by letting the full set of parameters to be updated using the score of the conditional distribution, $\mathbf{y}_t|\mathcal{F}_{t-1}$.

Formally, let $\boldsymbol{\omega}_t = (\omega_{j,t}, j = 1, \ldots, J)'$ be the vector containing the mixture weights at time t, and $\widetilde{\boldsymbol{\omega}}_t \in \Re^{J-1}$ be a $(J-1)$–dimension vector such that $\Lambda^{\omega}\left(\widetilde{\boldsymbol{\omega}}_t\right) = \boldsymbol{\omega}_t$, for an $\mathcal{F}_{t-1}$ measurable differentiable mapping function $\Lambda^{\omega} : \Re^{J-1} \to \mathcal{S}^J$ where $\mathcal{S}^J$ is the standard unit J–simplex. Similarly, let $\widetilde{\boldsymbol{\theta}}_{j,t} \in \Re^{d_j}$ be a d_j–dimension vector such that for each time t we have $\Lambda^j\left(\widetilde{\boldsymbol{\theta}}_{j,t}\right) = \boldsymbol{\theta}_{j,t}$, where $\Lambda^j : \Re^{d_j} \to \Omega^j$ for all $j = 1, \ldots, J$. In order to avoid complicated nonlinear constraints on the parameters dynamic, in this paper, instead of directly modeling the vector $\boldsymbol{\theta}_t$ defined on $\mathcal{S}^J \times \Omega^1 \times \cdots \times \Omega^J$, we model the unconstraint vector of parameters $\widetilde{\boldsymbol{\theta}}_t = \left(\widetilde{\boldsymbol{\omega}}'_t, \widetilde{\boldsymbol{\theta}}'_{j,t}, j = 1, \ldots, J\right)'$ defined on $\Re^{J-1} \times \Re^{d_1} \times \cdots \times \Re^{d_J}$. To this end, we reparametrize the conditional density (3) into $p\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right)$, where henceforth the dependence from $\mathcal{F}_{t-1}$ has been omitted for notational purposes.

In the Score–Driven framework, the quantity of interest is the score of the conditional distribution given by:

$$\begin{aligned}\widetilde{\nabla}\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right) &= \left.\frac{\partial \ln p\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}\right)}{\partial \widetilde{\boldsymbol{\theta}}}\right|_{\widetilde{\boldsymbol{\theta}}=\widetilde{\boldsymbol{\theta}}_t} \\ &= \mathcal{J}\left(\widetilde{\boldsymbol{\theta}}_t\right)' \nabla\left(\mathbf{y}|\boldsymbol{\theta}_t\right),\end{aligned}$$

where $\mathcal{J}\left(\widetilde{\boldsymbol{\theta}}_t\right)$ is the Jacobian of the full mapping function and $\nabla\left(\mathbf{y}|\boldsymbol{\theta}_t\right)$ is the conditional score with respect to $\boldsymbol{\theta}_t$ evaluated in $\mathbf{y}_t$. Here with "full mapping function" we refer to the vector–valued function $\Lambda : \Re^{J-1} \times \Re^{d_1} \times \cdots \times \Re^{d_J} \to \mathcal{S}^J \times \Omega^1 \times \cdots \times \Omega^J$ that incorporates $\Lambda^{\omega}\left(\cdot\right)$ and $\Lambda^j\left(\cdot\right), \quad j = 1, \ldots, J$, such that $\Lambda\left(\widetilde{\boldsymbol{\theta}}_t\right) = \boldsymbol{\theta}_t$, $\forall t$.

The term $\widetilde{\nabla}\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right)$ enters linearly as a forcing variable into the dynamic updating equation of $\widetilde{\boldsymbol{\theta}}_t$ as follow:

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\kappa} + \mathbf{A}\boldsymbol{\Xi}_t \widetilde{\nabla}\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right) + \mathbf{B}\widetilde{\boldsymbol{\theta}}_t, \tag{4}$$

where $\boldsymbol{\kappa}$ is a $\left(J-1+\sum_{j=1}^{J} d_j\right) = L$–dimension vector and $\mathbf{A}$ and $\mathbf{B}$ are $L \times L$ diagonal matrices of coefficients that need to be estimated. Positivity of the elements of $\mathbf{A}$ should be imposed in order to ensure that parameters are updated following the direction indicated by the score. Weak stationarity of the process $\{\widetilde{\boldsymbol{\theta}}_s, s > 0\}$ is obtained by constraining all the elements of $\mathbf{B}$ to be smaller than one in modulus since $E_{t-1}\left[\boldsymbol{\Xi}_t \widetilde{\nabla}\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right)\right]$ for all t.[4] The term $\boldsymbol{\Xi}_t$ is an $L \times L$ positive definite matrix which scales the score. Creal et al. (2013) suggest to set $\boldsymbol{\Xi}_t = \mathbb{E}_{t-1}\left[\widetilde{\nabla}\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right)\widetilde{\nabla}\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right)'\right]^{-\delta}$, with $\delta \in \{0, 1/2, 1\}$, that is, scaling the score with the Fisher information matrix ($\delta = 1$), its pseudo inverse ($\delta = 1/2$), or not scaling ($\delta = 0$). Unfortunately, for DAMMs the Fisher information matrix is generally not available in closed form. In the following we detail a different scaling mechanism similar to the one employed by Bernardi and Catania (2018). The dynamic updating equation (4) can be divided into $J+1$ individual dynamics as follows:

$$\begin{aligned}\widetilde{\boldsymbol{\omega}}_{t+1} &= \boldsymbol{\kappa}^{\omega} + \mathbf{A}^{\omega}\boldsymbol{\Xi}_t^{\omega}\mathcal{J}^{\omega}\left(\widetilde{\boldsymbol{\omega}}_t\right)'\nabla^{\omega}\left(\mathbf{y}_t|\boldsymbol{\omega}_t\right) + \mathbf{B}^{\omega}\widetilde{\boldsymbol{\omega}}_t \\ \widetilde{\boldsymbol{\theta}}_{j,t+1} &= \boldsymbol{\kappa}^{j} + \mathbf{A}^{j}\boldsymbol{\Xi}_t^{j}\mathcal{J}^{j}\left(\widetilde{\boldsymbol{\theta}}_{j,t}\right)'\nabla^{j}\left(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}\right) + \mathbf{B}^{j}\widetilde{\boldsymbol{\theta}}_{j,t}, \quad j = 1, \ldots, J,\end{aligned} \tag{5}$$

where all the symbols have the same interpretation as before, but are now related to each specific quantity of the model. For DAMMs, unreported extensive simulations and empirical analysis have shown that setting $\boldsymbol{\Xi}_t^{\omega} = \mathbb{I}$ and:

$$\begin{aligned}\boldsymbol{\Xi}_t^{j} &= \left[\mathcal{J}^{j}\left(\widetilde{\boldsymbol{\theta}}_{j,t}\right)'\left(\int_{\Re^d}\left[\nabla_{p_j}\left(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}\right)\nabla_{p_j}\left(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}\right)'\right]p_j\left(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}\right)\mathrm{d}\mathbf{y}_t\right)\mathcal{J}^{j}\left(\widetilde{\boldsymbol{\theta}}_{j,t}\right)\right]^{-1/2} \\ &= \left[\mathcal{J}^{j}\left(\widetilde{\boldsymbol{\theta}}_{j,t}\right)\mathcal{I}_{p_j}\left(\boldsymbol{\theta}_{j,t}\right)\mathcal{J}^{j}\left(\widetilde{\boldsymbol{\theta}}_{j,t}\right)'\right]^{-1/2},\end{aligned}$$

where $\nabla_{p_j}\left(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}\right) = \frac{\partial \ln p_j\left(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}\right)}{\partial \boldsymbol{\theta}_{j,t}}$ and $\mathcal{I}_{p_j}\left(\boldsymbol{\theta}_{j,t}\right)$ are the score and the Fisher information matrix of the j–th mixture component, provides best results.[5] For the rest of the paper, where not explicitly stated, results are reported according to this parametrization of $\boldsymbol{\Xi}_t^{\omega}$ and $\boldsymbol{\Xi}_t^{j}$.

3.1. Update the mixture composition

Several different choices are available in order to reparameterize and update the mixture weights $\boldsymbol{\omega}_t$. For example, Billio et al. (2013) use the logistic–transformed Gaussian (LTG) weights, i.e. their mapping function is the vector valued function $\Lambda^{\mathrm{LTG}} : \Re^J \to \mathcal{S}^J$, with j–th component given by $\Lambda_j^{\mathrm{LTG}}\left(\widetilde{\omega}_{j,t}\right) = \frac{\exp(\widetilde{\omega}_{j,t})}{\sum_{i=1}^{J}\exp(\widetilde{\omega}_{i,t})}$, and $\widetilde{\boldsymbol{\omega}}_t \sim \mathcal{N}_J\left(\widetilde{\boldsymbol{\omega}}_t|\widetilde{\boldsymbol{\omega}}_{t-1}, \Sigma\right)$, with Σ be a proper covariance matrix. This mapping and updating scheme assumes that the weights do not depend on the observable data. We propose a convenient choice for the function $\Lambda^{\omega}(\cdot)$, given by:

$$\Lambda^{\omega}\left(\widetilde{\boldsymbol{\omega}}_t\right) := \begin{cases}\omega_{j,t} = \lambda_{[0,b_{j,t}]}\left(\widetilde{\omega}_{j,t}\right), & j = 1, \ldots, J-1 \\ \omega_{J,t} = 1 - \sum_{h=1}^{J-1}\omega_{h,t},\end{cases}$$

[4] Provided the existence of $\mathbb{E}\left[\boldsymbol{\Xi}_t\widetilde{\nabla}\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right)\widetilde{\nabla}\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right)'\boldsymbol{\Xi}_t'\right]$.

[5] Results of this analysis are not reported in order to save space. We tried to set $\boldsymbol{\Xi}_t^{\omega}$ and $\boldsymbol{\Xi}_t^{j}$ to the correct Fisher information matrices by approximating them using simulation schemes and numerical integration techniques. Although having sensibly increased the computational burden during the model estimation, the results have shown that the choice reported in the paper provides higher likelihood values.

where $b_{j,t} = b_{j-1,t} - \omega_{j-1,t}$ with $b_{1,t} = 1$ and $\lambda_{[L,U]} : \Re \to [L, U] \subset \Re$ is the modified logistic function $\lambda_{[L,U]}(x) = L + \frac{(U-L)}{1+\exp(-x)}$.[6] With these choices of $\Lambda^{\omega}(\cdot)$ and $\lambda_{[L,U]}(\cdot)$, the (j,h)–th element of the $J \times J-1$ Jacobian matrix $\mathcal{J}^{\omega}(\cdot)$ is given by:

$$\mathcal{J}^{\omega}\left(\widetilde{\boldsymbol{\omega}}_t\right)_{(j,h)} = \begin{cases} \frac{b_{j,t}\exp(-\widetilde{\omega}_{j,t})}{(1+\exp(-\widetilde{\omega}_{j,t}))^2}, & \text{if } h = j \\ \frac{-\sum_{k=1}^{j-1}\mathcal{J}^{\omega}\left(\widetilde{\boldsymbol{\omega}}_t\right)_{(k,h)}}{1+\exp(-\widetilde{\omega}_{j,t})}, & \text{if } h < j \quad \wedge \quad j \neq J \\ -\sum_{k=1}^{J-1}\mathcal{J}^{\omega}\left(\widetilde{\boldsymbol{\omega}}_t\right)_{(k,h)}, & \text{if } j = J, \\ 0, & \text{if } h > j. \end{cases}$$

The score of (3) with respect to the mixture weights parameters is given by $\nabla^{\omega}(\mathbf{y}_t|\boldsymbol{\omega}_t) = \left(\frac{p_j(\mathbf{y}_t|\boldsymbol{\theta}_{j,t})}{p(\mathbf{y}_t|\boldsymbol{\theta}_t)}, \quad j = 1, \ldots, J\right)'$. Note that the j–th component of $\nabla^{\omega}(\mathbf{y}_t|\boldsymbol{\omega}_t)$ is given by the ratio between j–th component density and the mixture density. This mechanism naturally suggests to update more the weights associated to the components of the mixture from which it has been more likely to sample previous observations.

3.2. Update the mixture components

As for the update of the mixture composition, the reparameterized mixture component parameters $\widetilde{\boldsymbol{\theta}}_{j,t} \quad j = 1, \ldots, J$ are updated using the conditional score of the mixture distribution $p\left(\mathbf{y}_t|\widetilde{\boldsymbol{\theta}}_t\right)$ with respect to the vector $\widetilde{\boldsymbol{\theta}}_{j,t}$, given by $\mathcal{J}^j\left(\widetilde{\boldsymbol{\theta}}_{j,t}\right)'\nabla^j(\mathbf{y}_t|\boldsymbol{\theta}_{j,t})$. Straightforward calculations show that:

$$\nabla^j(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}) = \omega_{j,t}\frac{p_j(\mathbf{y}_t|\boldsymbol{\theta}_{j,t})}{p(\mathbf{y}_t|\boldsymbol{\theta}_t)}\nabla_{p_j}(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}) \quad , j = 1, \ldots, J,$$

where $\nabla_{p_j}(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}) = \frac{\partial \ln p_j(\mathbf{y}_t|\boldsymbol{\theta}_{j,t})}{\partial \boldsymbol{\theta}_{j,t}}$ is the score of the j–th mixture component. The updating equation for the j–th mixture component parameters can be written as:

$$\widetilde{\boldsymbol{\theta}}_{j,t+1} = \boldsymbol{\kappa}^j + \xi_{j,t}\mathbf{A}^j\boldsymbol{\Xi}_t^j\mathcal{J}^j\left(\widetilde{\boldsymbol{\theta}}_{j,t}\right)'\nabla_{p_j}(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}) + \mathbf{B}^j\widetilde{\boldsymbol{\theta}}_{j,t}, \tag{6}$$

where

$$\xi_{j,t} = \omega_{j,t}\frac{p_j(\mathbf{y}_t|\boldsymbol{\theta}_{j,t})}{p(\mathbf{y}_t|\boldsymbol{\theta}_t)} \tag{7}$$

is the relative contribution of the j–th component to the mixture density at time t conditionally on past information. It is worth to comment that equation (6) is very similar to that usually found in Score–Driven processes. Indeed, if $J = 1$, we recover the Generalized Autoregressive Score (GAS) and the Dynamic Conditional Score (DCS) models of Creal et al. (2013) and Harvey (2013), respectively. However, in our context, the mixture assumption naturally suggests to scale the score contribution in

[6] We tried different specifications for $\Lambda^{\omega}(\cdot)$ along with: i) the multinomial logistic transformation detailed in Bishop (2006) for which $\omega_{j,t} = \exp(\widetilde{\omega}_{j,t})/(1 + \sum_{k=1}^{J-1}\exp(\widetilde{\omega}_{k,t}))$ for $j = 1, \ldots, J-1$ and $\omega_{J,t} = 1 - \sum_{h=1}^{J-1}\omega_{h,t}$ and ii) the transformation employed by Frühwirth-Schnatter (2006) such that $\omega_{j,t} = \text{logit}^{-1}(\widetilde{\omega}_{j,t})\prod_{h=1}^{j-1}(1 - \text{logit}^{-1}(\widetilde{\omega}_{h,t}))$ for $j = 1, \ldots, J-1$ and $\omega_{J,t} = \prod_{h=1}^{J-1}(1 - \text{logit}^{-1}(\widetilde{\omega}_{h,t}))$. According to our tests we found that the one reported provides the best results both in simulations and real applications.

a way to account for the relative importance each mixture component has at time t. More formally, if we interpret the mixture weight $\omega_{j,t}$ as the prior probability of sampling from the j–th component and the ratio $p_j\left(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}\right)/p\left(\mathbf{y}_t|\boldsymbol{\theta}_t\right)$ as the period–t likelihood for component j given information up to time $t-1$, it follows immediately that $\xi_{j,t}$ as given in equation (7) is the posterior probability of sampling from the j–th mixture component. That is, in equation (6) the score $\nabla_{p_j}\left(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}\right)$ is premultiplied by the posterior probability of sampling from the j–th component at time t. Interestingly, a similar result has been found by Bernardi and Catania (2018) in their Switching Generalized Autoregressive Score Copula model and by Bazzi et al. (2017) in their Time–Varying Hidden Markov Model.

Estimation of DAMMs can be performed using the Maximum Likelihood (ML) estimator as detailed by Blasques et al. (2014), see also Blasques et al. (2016). To reduce the possibility of converging to a local optimum of the likelihood, good starting values and multiple tries should be implemented. We suggest to initialize the algorithm using the estimates delivered by a static mixture model estimated by the EM algorithm, as suggested by Yu (2012) in a similar context. The supplementary material accompanying this paper reports an extensive simulation study on the finite sample properties of the ML estimator for selected DAMMs with Student's t and Gaussian mixtures. Results indicate that, depending on the number of time–varying parameters, a sample size between $T = 1000$ and $T = 5000$ is appropriate in order to estimate a DAMM with time–varying locations, scales, and mixture composition.[7]

4. DAMMs for Financial Returns

Financial returns exhibit well known stylized facts such as: i) zero mean and absence of serial autocorrelation, ii) skewness and fat tails, and iii) conditional heteroscedasticity, see e.g. McNeil et al. (2015). The aim of this section is to develop DAMM specifications that are able to match the empirical regularities of financial returns. Consider the case of a DAMM with J univariate Gaussian components, labeled as Gaussian DAMM (G–DAMM):

$$p(y_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{J} \omega_{j,t}\phi(y_t|\mu_j, \sigma^2_{j,t}),$$

where $\phi(\cdot|\mu_j, \sigma^2_{j,t})$ is the density of a univariate Gaussian distribution with mean μ_j and time–varying variance $\sigma^2_{j,t}$. To ensure that y_t is a zero mean uncorrelated process, we follow Haas et al. (2004a) and set $\mu_{J,t} = -\sum_{j=1}^{J-1}(\omega_{j,t}/\omega_{J,t})\mu_j$, such that now $\mathbb{E}_{t-1}[y_t] = 0$. Note that, if we use an identity mapping function for $\Lambda^j(\cdot)$, i.e., $\widetilde{\sigma}^2_{j,t} = \sigma^2_{j,t}$, the j–th conditional variance is given by:[8]

$$\sigma^2_{j,t+1} = \kappa^j + A^j \Xi^j_t \xi_{j,t} \nabla_{p_j}(y_t|\sigma^2_{j,t}) + B^j \sigma^2_{j,t}\,, \tag{8}$$

where

[7] Estimation of DAMMs will be included in the future release of the **GAS** package for R of Catania et al. (2017), see also Ardia et al. (2016).

[8] Note that here we are imposing an identity mapping function just to show the relation between DAMM and other models available in the literature. The identity mapping also implies $\mathcal{J}^j(\widetilde{\boldsymbol{\theta}}_{j,t}) = 1$ such that this quantity is not reported in equation (9). In the empirical application we employ the G–DAMM model with an exponential link function for the volatility which is detailed later in this section.

$$\nabla_{p_j}(y_t|\sigma_{j,t}) = -\frac{1}{2\sigma_{j,t}^2}\left(1 - \frac{(y_t-\mu_j)^2}{\sigma_{j,t}^2}\right),$$

is the score of a Gaussian distribution with respect to the variance parameter, and κ^j, A^j, B^j are now scalar coefficients. Consider the following choice of the scalar Ξ_t^j:

$$\Xi_t^j = \left(\xi_{j,t}\int_{\Re}\nabla_{p_j}(y_t|\sigma_{j,t})^2\phi_j(y_t;\mu_j,\sigma_{j,t}^2)\mathrm{d}y_t\right)^{-1} \tag{9}$$

$$= \frac{2\sigma_{j,t}^4}{\xi_{j,t}}, \tag{10}$$

that is, Ξ_t^j is defined as the inverse of the Fisher information quantity[9] of $\sigma_{j,t}^2$ where the expectation is taken with respect to the j–th Gaussian component and multiplied by the probability of sampling from that component at time t, $\xi_{j,t}$. Substituting Ξ_t^j in (8) and rearranging the terms we obtain:

$$\sigma_{j,t+1}^2 = \kappa^j + \alpha^j\varepsilon_{j,t}^2 + \beta^j\sigma_{j,t}^2,$$

where $\alpha^j = A^j$, $\beta^j = B^j - A^j$ and $\varepsilon_{j,t}^2 = (y_t-\mu_j)^2$, i.e., we recover a GARCH dynamic for each mixture component variance. If we further impose that $\omega_{j,t} = \omega_j$ by setting $\mathbf{A}^\omega = \mathbf{B}^\omega = \mathbf{0}$ in (5), we recover the Mixed Normal GARCH model independently proposed by Haas et al. (2004a), Alexander and Lazar (2006) and Zhang et al. (2006). A very similar version of the The Time–varying Normal mixture GARCH model with likelihood–driven mixing weights of Haas et al. (2013) can also be recovered by a different choice of the weights mapping function $\Lambda^\omega(\cdot)$ and setting $\mathbf{A}^\omega = \alpha^\omega\mathbb{I}$, $(\alpha^\omega > 0)$ and $\mathbf{B}^\omega = \mathbf{0}$.[10] Furthermore, under the additional constraints $B^j = A^j$ for all $j = 1,\ldots,J$ we recover the Mixture Autoregressive Conditional Heteroscedastic model of Wong and Li (2001).

It is worth to mention that by the definition of Ξ_t^j given in equation (9), we have ruled out the very appealing mechanism of DAMMs that redistributes new information across the variances of the J components via the $\xi_{j,t}$ term. On the contrary, if we define $\Xi_t^j = \left(\int_{\Re}\nabla_{p_j}(y_t|\sigma_{j,t})^2\phi_j(y_t;\mu_j,\sigma_{j,t}^2)\mathrm{d}y_t\right)^{-1}$ as described in Section 3, we obtain the following updating equation:

$$\sigma_{j,t+1}^2 = \kappa^j + \alpha^j\xi_{j,t}\varepsilon_{j,t}^2 + \beta^j\sigma_{j,t}^2,$$

which retain this appealing mechanism but, unless $J = 1$, does not nest other models previously proposed in the literature.

Similarly, we can define the Student's t DAMM (t–DAMM) as a dynamic mixture of J Student's t distributions:

$$p(y_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{J}\omega_{j,t}t(y_t|\mu_j,\sigma_{j,t}^2,\nu_j), \tag{11}$$

[9] The Fisher information matrix for scalars is defined as the Fisher information quantity.

[10] See Appendix A for details about the specification of Haas et al. (2013). The reason why DAMM does not nest exactly this specification is that in the denominator of equation A.1 where the authors use the sum of the component densities which of course differs from the density of the mixture we use in DAMM, see Section 3.2.

where $t(\cdot|\mu_j, \sigma^2_{j,t}, \nu_j)$ is the density of a Student's t distribution with $\nu_j > 2$ degrees of freedom, parameterized such that its variance is $\sigma^2_{j,t}$. According to model (11), we define the following updating equation for $\sigma^2_{j,t}$:

$$\sigma^2_{j,t} = \exp(2\widetilde{\sigma}_{j,t}) \tag{12}$$

$$\widetilde{\sigma}_{j,t+1} = \kappa^j + A^j \xi_{j,t} \mathcal{J}(\widetilde{\sigma}_{j,t}) \Xi^j_t \nabla_{p_j}(y_t|\sigma_{j,t}) + B^j \widetilde{\sigma}_{j,t}, \tag{13}$$

where, if we use $\Xi^j_t = \mathcal{I}(\widetilde{\sigma}_{j,t})^{-1/2}$ as detailed in Section 3, the term $\mathcal{J}(\widetilde{\sigma}_{j,t})\Xi^j_t \nabla_{p_j}(y_t|\sigma_{j,t})$ turns out to be

$$\mathcal{J}(\widetilde{\sigma}_{j,t})\Xi^j_t \nabla_{p_j}(y_t|\sigma_{j,t}) = \sqrt{\frac{\nu_j + 3}{2\nu_j}} \left(\frac{(\nu_j + 1) z^2_{j,t}}{(\nu_j - 2) + z^2_{j,t}} - 1 \right) \tag{14}$$

with $z_{j,t} = (y_t - \mu_j)/\sigma_{j,t}$. From (14) we see that t–DAMM does not nest other models apart from the case $J = 1$, for which we recover the tGAS model of Creal et al. (2013). If we let $\nu_j \to \infty$ for all j, we recover the G–DAMM specification with exponential link function for the conditional variance for which:

$$\mathcal{J}(\widetilde{\sigma}_{j,t})\Xi^j_t \nabla_{p_j}(y_t|\sigma_{j,t}) = \frac{1}{\sqrt{2}} \left(z^2_{j,t} - 1 \right). \tag{15}$$

Hereafter we refer to the t–DAMM and G–DAMM specifications with exponential link function for the conditional variance reported in equations (14) and (15), respectively. In the empirical application reported in Section 6, we impose the same parametrization of Haas et al. (2004a) in order to constraint the first conditional moment of financial returns to zero: $\mu_{J,t} = -\sum_{j=1}^{J-1} (\omega_{j,t}/\omega_{J,t})\mu_j$. Given the mixture parametrization, it follows that conditional moments are easily available. For instance, the conditional variance of y_t is given as:

$$\sigma^2_t = \sum_{j=1}^{J} \omega_{j,t}(\sigma^2_{j,t} + \mu^2_j),$$

the time–varying skewness by:

$$\mathrm{Sk}_t = \frac{1}{\sigma^2_t} \sum_{j=1}^{J} \omega_{j,t}\mu_j(3\sigma^2_{j,t} + \mu^2_j),$$

and excess kurtosis coefficient by:

$$\mathrm{Ku}_t = \frac{1}{\sigma^4_t} \sum_{j=1}^{J} \omega_{j,t}(\mu^4_j + 6\mu_j\sigma^2_{j,t} + 3\sigma^4_{j,t}) - 3$$

for G–DAMM, and by:

$$\mathrm{Ku}_t = \frac{1}{\sigma^4_t} \sum_{j=1}^{J} \omega_{j,t}\mu_j(\mu^4_j + 6\mu_j\sigma^2_{j,t} + \sigma^4_{j,t}\psi(\nu_j)) - 3,$$

where $\psi(\nu_j) = (\nu_j - 2)^2\Gamma(5/2)\Gamma((\nu_j - 4)/2)/(\sqrt{\pi}\Gamma(\nu_j/2))$ for t–DAMM, respectively.

4.1. Multivariate DAMM specifications for financial returns

Multivariate financial returns are also characterized by well known stylized facts such as time–varying correlation and the presence of tail dependence, see for example McNeil et al. (2015). The Dynamic Conditional Correlation (DCC) model of Engle (2002) and Tse and Tsui (2002) is widely employed to model changes in the linear correlation coefficient of multivariate financial returns. An extension of the DCC model with mixture of Gaussians distributed innovations has been proposed by Galeano and Ausín (2010), however, their model formulation implies that the dynamic correlation coefficient is the same among the mixture components. A Score–Driven model with time–varying correlations and time–varying individual variances has been proposed by Creal et al. (2011). When the conditional distribution in DAMM is multivariate Gaussian or multivariate Student's t, we obtain an extension of the model proposed by Creal et al. (2011), when the number of mixture components is one ($J = 1$), we obtain the equivalence between the two specifications. An important difference from Creal et al. (2011) is that DAMM with a conditional multivariate Student's t distribution allows the model to account for time–varying upper and lower tail dependence among different financial securities,[11] see for example Patton (2012) for a review about this literature.

To simplify the exposition in this section and in the empirical application, we focus on bivariate DAMM specifications. The supplementary material accompanying this paper reports examples for the general multivariate case.[12] Consider the case $\mathbf{y}_t = (y_{1,t}, y_{2,t})' \in \Re^2$ where:

$$p(\mathbf{y}_t|\mathcal{F}_{t-1}) = \sum_{j=1}^{J} \omega_{j,t}\phi_2(\mathbf{y}_t|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{j,t}),$$

and $\phi_2(\cdot|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{j,t})$ indicates the bivariate Gaussian density with vector of mean $\boldsymbol{\mu}_j$, and covariance matrix $\boldsymbol{\Sigma}_{j,t}$. Similar to the univariate case, we set $\boldsymbol{\mu}_{J,t} = -\sum_{j=1}^{J-1}(\omega_{j,t}/\omega_{J,t})\boldsymbol{\mu}_j$ to ensure that $\mathbb{E}_{t-1}[y_{i,t}] = 0$ for $i = 1, 2$. We use the usual covariance decomposition $\boldsymbol{\Sigma}_{j,t} = \mathbf{D}_{j,t}\mathbf{R}_{j,t}\mathbf{D}_{j,t}$ where $\mathbf{D}_{j,t} = \text{diag}(\sigma_{1,j,t}, \sigma_{2,j,t})$ and:

$$\mathbf{R}_{j,t} = \begin{pmatrix} 1 & \rho_{j,t} \\ \rho_{j,t} & 1 \end{pmatrix},$$

and $\rho_{j,t} \in (-1, 1)$ is the linear correlation coefficient between $y_{1,t}$ and $y_{2,t}$ at time t in the j–th mixture component. Assuming $J = 2$, we can collect all the time–varying parameters in $\boldsymbol{\theta}_t = (\boldsymbol{\theta}'_{1,t}, \boldsymbol{\theta}'_{2,t})'$, where $\boldsymbol{\theta}_{j,t} = (\sigma_{1,j,t}, \sigma_{2,j,t}, \rho_{j,t})'$, for $j = 1, 2$. The score with respect to $\boldsymbol{\theta}_{j,t}$ is given by $\nabla_{p_j}(\mathbf{y}_t|\boldsymbol{\theta}_{j,t}) = (\nabla_t^{\sigma_{1,j}}, \nabla_t^{\sigma_{2,j}}, \nabla_t^{\rho_j})'$, where:

$$\nabla_t^{\sigma_{i,j}} = \frac{1}{\sigma_{i,j,t}}\left[\frac{z_{i,j,t}^2}{1-\rho_{j,t}^2} - \frac{\rho_{j,t}z_{1,j,t}z_{2,j,t}}{1-\rho_{j,t}^2} - 1\right] \quad \text{for} \quad i = 1, 2$$

$$\nabla_t^{\rho_j} = \frac{1}{1-\rho_{j,t}^2}\left[\rho_{j,t} + z_{1,j,t}z_{2,j,t}\frac{1+\rho_{j,t}^2}{1-\rho_{j,t}^2} - \frac{\rho_{j,t}}{1-\rho_{j,t}^2}(z_{1,j,t}^2 + z_{2,j,t}^2)\right],$$

[11] This is true only if the mixture weights are allowed to evolve over time.

[12] In the case of DAMMs with a conditional multivariate Gaussian or multivariate Student's t distribution the only additional difficulty concerns the definition of a mapping function for the correlation matrix. We suggest to follow the approach of Creal et al. (2011) and employ the hyperspherical coordinate transformation originally proposed by Pinheiro and Bates (1996). More details are reported in the supplementary material.

where $z_{i,j,t} = (y_{i,t} - \mu_{i,j})/\sigma_{i,j,t}$, for $i = 1, 2$ and $j = 1, 2$. If we employ an exponential link mapping for the conditional standard deviations: $\sigma_{i,j,t} = \exp(\widetilde{\sigma}_{i,j,t})$ and the modified logistic transformation for the conditional correlations: $\rho_{j,t} = (1-\exp(-\widetilde{\rho}_{j,t}))/(1+\exp(-\widetilde{\rho}_{j,t}))$, the Jacobian matrix $\mathcal{J}^j(\widetilde{\boldsymbol{\theta}}_{j,t})$ is given by:

$$\mathcal{J}^j(\widetilde{\boldsymbol{\theta}}_{j,t}) = \begin{pmatrix} \exp(\widetilde{\sigma}_{1,j,t}) & 0 & 0 \\ 0 & \exp(\widetilde{\sigma}_{2,j,t}) & 0 \\ 0 & 0 & -2\frac{\exp(-\widetilde{\rho}_{j,t})}{(1+\exp(-\widetilde{\rho}_{j,t}))^2} \end{pmatrix},$$

for $j = 1, 2$. The Fisher information matrix associated to the bivariate Gaussian distribution for the j–th component is given by:

$$\mathcal{I}_{p_j}(\boldsymbol{\theta}_{j,t}) = \begin{pmatrix} \iota_{j,t}^{1,1} & \cdot & \cdot \\ \iota_{j,t}^{1,2} & \iota_{j,t}^{2,2} & \cdot \\ \iota_{j,t}^{1,3} & \iota_{j,t}^{2,3} & \iota_{j,t}^{3,3} \end{pmatrix},$$

where $\iota_{j,t}^{l,l} = -\left[(\rho_{j,t} - 2)/(1 - \rho_{j,t}^2)\right]/\sigma_{l,j,t}^2$ for $l = 1, 2$, $\iota_{j,t}^{1,2} = -\rho_{j,t}^2/[\sigma_{1,j,t}\sigma_{2,j,t}(1 - \rho_{j,t}^2)]$,

$$\iota_{j,t}^{l,3} = -\frac{\rho_{j,t}}{\sigma_{l,j,t}(1 - \rho_{j,t}^2)}\left[\frac{2}{(1 - \rho_{j,t}^2)} - \frac{1 + \rho_{j,t}^2}{1 - \rho_{j,t}^2}\right], \quad \text{for} \quad l = 1, 2,$$

and $\iota_{j,t}^{3,3} = (1 + \rho_{j,t}^2)/(1 - \rho_{j,t}^2)^2$. Knowing $\nabla_{p_j}(\mathbf{y}_t|\boldsymbol{\theta}_{j,t})$, $\mathcal{J}^j(\widetilde{\boldsymbol{\theta}}_{j,t})$, and $\mathcal{I}_{p_j}(\boldsymbol{\theta}_{j,t})$, the updating equation (5) can be implemented for all j.[13] In the empirical application we label this model as MG–DAMM.

5. Simulation Studies: Filtering with DAMM

In order to assess the filtering ability of DAMMs, we report two simulation studies. The aim of these studies is to investigate the ability of DAMMs to perform filtering on dynamic properties of the observed time series. The two experiments are constructed in the spirit of those reported by Engle (2002) and Creal et al. (2011). Specifically, a number of artificial processes are assumed as the true Data Generating Process (DGP) from which observations are sampled. Alternative models are estimated on the simulated series and their precision in recovering the true signal is compared. Note that in this framework all the competing models are misspecified by construction.

The supplementary material accompanying this paper reports additional simulation experiments. One of these experiments investigates the cost of model misspecification in terms of the filtering precision. This analysis indicates that if the sample size is too small, it is better to not assume a highly parameterized DAMM specification. On the contrary, if sample size is large, the cost of misspecification is generally low.

5.1. Time–varying correlations

We first focus on time–varying correlations by simulating $B = 500$ time series of length $T = 1000$ from a standardized bivariate Gaussian distribution with time–varying correlation ρ_t. The dynamics we assume for the correlation parameter are:

[13] Also for the multivariate DAMM specificaion we set $\boldsymbol{\Xi}_t^j = \left[\mathcal{J}^j(\widetilde{\boldsymbol{\theta}}_{j,t})'\mathcal{I}_{p_j}(\boldsymbol{\theta}_{j,t})\mathcal{J}^j(\widetilde{\boldsymbol{\theta}}_{j,t})\right]^{-1/2}$.

- Constant: $\rho_t = 0.9$
- Sine: $\rho_t = 0.5 + 0.4\cos(2\pi t/200)$
- Fast Sine: $\rho_t = 0.5 + 0.4\cos(2\pi t/20)$
- Step: $\rho_t = 0.9 - 0.5\,(t > 500)$
- Ramp: $\rho_t = \mod(t/200)/200$
- $\mathcal{M}_1$: $\rho_t = \exp(\widetilde{\rho}_t)/[1+\exp(\widetilde{\rho}_t)]$ where $\widetilde{\rho}_t = -0.4\,(1-0.99) + 0.99\widetilde{\rho}_{t-1} + 0.14\eta_t^{\rho}$, $\quad \eta_t^{\rho} \sim \mathcal{N}(0,1)$
- $\mathcal{M}_2$: $\rho_t = \omega_t\rho_{1,t} + (1-\omega_t)\,\rho_{2,t}$ where $\rho_{i,t} = \exp(\widetilde{\rho}_t)/[1+\exp(\widetilde{\rho}_t)]$, $\widetilde{\rho}_{i,t} = \bar{\rho}_i\,(1-0.99) + 0.99\widetilde{\rho}_{i,t-1} + 0.14\eta_t^{\rho_i}$, $\quad i = 1,2$, and $\bar{\rho}_1 = -0.4$, $\bar{\rho}_2 = 0.4$, where $\omega_t = [1+\exp(\widetilde{\omega}_t)]^{-1}$, $\widetilde{\omega}_t = 0.98\widetilde{\omega}_{t-1} + \eta_t^{\omega}$ and $\eta^{\rho_1}, \eta^{\rho_2}, \eta_t^{\omega}$ are iid $\mathcal{N}(0,1)$.

We estimate by Maximum Likelihood the MG–DAMM with two mixture components detailed in Section 4.1 to each series of simulated data. This model has time–varying mixture weights and time–varying correlations. We also specify the two constraint MG–DAMM versions: MG–DAMM–$\bar{\rho}$ and MG–DAMM–$\bar{\omega}$. The MG–DAMM–$\bar{\rho}$ specification is defined as the MG–DAMM but with constant correlations, i.e. $\rho_{j,t} = \rho_j$, $\quad j = 1,2$. Analogously, the MG–DAMM–$\bar{\omega}$ is defined as the MG–DAMM but with static mixture composition, i.e. $\omega_t = \omega$. For comparative purposes we benchmark MG–DAMM with the Dynamic Conditional Correlation (DCC) model of Engle (2002) and with the Exponentially Weighed Moving Average (EWMA) model with forgetting factor equal to 0.96. DCC is the benchmark model in financial econometrics for time–varying correlation while EWMA is inspired by the RiskMetrics (Morgan et al., 1996) approach and widely used by practitioners. For all models, we set all parameters other than the correlation one to their true values.[14] The filtering comparison is reported in terms of the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) between the filtered and the simulated dynamics. Specifically, let $\rho_t^{(b,l)}$ be the correlation filtered at time t for model $l \in$ (MG–DAMM, MG–DAMM–$\bar{\rho}$, MG–DAMM–$\bar{\omega}$, DCC, EWMA) relative to the b–th simulation, $b = 1,\ldots,B$. The MAE and MSE for the b–th simulation are defined as:

$$\text{MAE}^{(b,l)} = \frac{1}{T}\sum_{t=1}^{T}|\rho_t^{(b,l)} - \rho_t|,$$

and

$$\text{MSE}^{(b,l)} = \frac{1}{T}\sum_{t=1}^{T}(\rho_t^{(b,l)} - \rho_t)^2,$$

where ρ_t is the true correlation value, respectively.

Table 1 reports the median MAE and MSE across the B replications for the seven artificial correlation patterns. Results are reported relative to the DCC model representing the benchmark. Values lower than one indicate outperformance with respect to the benchmark and viceversa. We

[14] For example, the conditional variances in the DCC specification are set to one and the conditional mean to 0. The same is done for the MG–DAMM specification detailed in Section 4.1 where the mixture component means are set to zero $\mu_{i,j} = 0$, and the static mixture component variances are set to one, $\sigma_{i,j} = 1$, for $i = 1,2$ and $j = 1,2$.

Specification	Const	Sine	FastSine	Step	Ramp	$\mathcal{M}_1$	$\mathcal{M}_2$
MAE							
MG–DAMM	0.94	0.91	0.86	0.92	0.36	0.96	1.00
MG–DAMM–$\bar{\rho}$	0.81	0.98	1.08	1.05	0.28	1.02	1.09
MG–DAMM–$\bar{\omega}$	0.90	0.98	0.98	0.98	0.76	0.96	0.97
DCC	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EWMA	4.32	1.29	1.11	1.09	0.99	1.05	1.03
MSE							
MG–DAMM	0.95	0.84	0.75	0.85	0.25	0.95	0.99
MG–DAMM–$\bar{\rho}$	0.59	0.94	1.06	1.02	0.21	1.02	1.13
MG–DAMM–$\bar{\omega}$	0.79	0.91	0.93	0.89	0.62	0.91	0.93
DCC	1.00	1.00	1.00	1.00	1.00	1.00	1.00
EWMA	23.97	1.46	1.28	1.09	1.01	1.12	1.11

Table 1: Mean Absolute Error (MAE) and Mean Squared Error (MSE) across the true patterns assumed for the correlation of a bivariate Gaussian distribution and the filtered correlations according to the Dynamic Conditional Correlation (DCC) model of Engle (2002), the Exponential Weighted Moving Average (EWMA) model with forgetting factor 0.96 inspired by the RiskMetrics methodology detailed in Morgan et al. (1996) and MG–DAMM. The two additional specifications labeled as MG–DAMM–$\bar{\rho}$ and MG–DAMM–$\bar{\omega}$ are constrained versions of MG–DAMM with constant correlations parameters and constant mixture composition, respectively. For each artificial correlation pattern, $B = 500$ bivariate time series of length $T = 1000$ are simulated. For each $b = 1, \ldots, B$, the MAE and MSE quantities are evaluated according to the different models, median values over the number of replicates are then compute. Results are reported relative to the DCC model acting as a benchmark. Values lower than one indicate outperformance with respect to the benchmark and viceversa. Gray cells indicate the best performing model.

note that DAMM specifications are preferred over DCC and EWMA in all cases. More precisely, the unrestricted MG–DAMM specification is preferred for the Sine, FastSine, and Step cases, the MG–DAMM–$\bar{\rho}$ for the Const and Ramp, while the MG–DAMM–$\bar{\omega}$ for $\mathcal{M}_1$ and $\mathcal{M}_2$. Interestingly, comparing the results of the MG–DAMM and the MG–DAMM–$\bar{\omega}$ specifications we note that the inclusion of a time–varying mixture composition helps when the underlying signal alternates between two "regimes" such as in the Sine, FastSine, Step, and Ramp patterns. Generally, we find MG–DAMM specifications to outperform the DCC and the EWMA models both under the MAE and MSE criteria. Performance gains with respect to the DCC model are in the range of 3%–72% depending on the complexity of the true DGP.

5.2. Time–varying mixture composition

The second simulation experiment focuses on time–varying mixture weights. The structure is analogous to the previous experiment: i) we simulate from an artificial model B time series, ii) we estimate alternative models on the simulated draws and, iii) we compare the filtered values with the true ones. In this experiment we specify a mixture of two univariate Gaussian distributions with fixed means and variances, i.e. the DGP is of the form:

$$y_t \sim \omega_t \mathcal{N}\left(y_t | -4, 6\right) + \left(1 - \omega_t\right) \mathcal{N}\left(y_t | 1, 3\right) ,$$

where ω_t evolves according to one of the following patterns:

- Constant: $\omega_t = 0.9$

- Sine: $\omega_t = \cos(2\pi t/200)$
- Fast Sine: $\omega_t = \cos(2\pi t/20)$
- Step: $\omega_t = 0.9 - 0.5\,(t > 500)$
- Ramp: $\omega_t = \mod(t/100)/100$
- $\mathcal{M}_1$: $\omega_t = [1 + \exp(\widetilde{\omega}_t)]^{-1}$ where $\widetilde{\omega}_t = -0.015 + 0.98\widetilde{\omega}_{t-1} + 0.1\eta_t^{\omega}, \quad \eta_t^{\omega} \sim \mathcal{N}(0,1)$
- $\mathcal{M}_2$: $\omega_t = [1 + \exp(\widetilde{\omega}_t)]^{-1}$ where $\widetilde{\omega}_t = -0.015 + 0.98\widetilde{\omega}_{t-1} + 0.5\eta_t^{\omega}, \quad \eta_t^{\omega} \sim \mathcal{N}(0,1)$.

$\mathcal{M}_1$ and $\mathcal{M}_2$ are nonlinear first order autoregressions with different standard deviations assumed for the innovations (0.1 for $\mathcal{M}_1$ and 0.5 for $\mathcal{M}_2$). According to the selected values of the innovation standard deviations, the ω_t process evolves more smoothly in the interval $(0,1)$ for $\mathcal{M}_1$, and displays abrupt changes from 0 to 1 for $\mathcal{M}_2$. The length of the simulated time series is fixed to $T = 1000$ and the number of replicates is $B = 500$. For each replicate, we estimate the G–DAMM with two components to the simulated observations and we store the filtered values of ω_t. The means and variances of the two Gaussian components are set equal to the true values such that results only depend on the mixture composition. As for the previous experiment, DAMM is compared with two alternative models. The first model is based on the "Time–varying mixture GARCH with likelihood–driven mixing weights" introduced by Haas et al. (2013) which assumes that:

$$\omega_t = \frac{W_{1,t}}{W_{1,t} + W_{2,t}},$$

where:

$$W_{j,t} = \kappa_j + \gamma \frac{\frac{1}{\sigma_j}\phi\left(\frac{y_t - \mu_j}{\sigma_j}\right)}{\frac{1}{\sigma_1}\phi\left(\frac{y_t - \mu_1}{\sigma_1}\right) + \frac{1}{\sigma_2}\phi\left(\frac{y_t - \mu_2}{\sigma_2}\right)},$$

with $\kappa_j > 0$ for $j = 1, 2$, and $\gamma > 0$.[15] Model parameters other than the mixture weights are set to their true values: $\mu_1 = -4$, $\sigma_1^2 = 6$, $\mu_2 = 1$, and $\sigma_2^2 = 3$. We label this model as Time–Varying Mixture (TVM). The second benchmark is a static Mixture Model (SMM) estimated using a rolling window of 100 observations. Similar to DAMM, means and variances of the alternative models are fixed to the true DGP values.

Table 2 reports the median MAE and the median MSE across the B replications for the three models. Results are reported relative to the SMM model. Similar to the correlation study, we find that DAMM reports superior results for all the patterns assumed for the mixture composition parameter ω_t. The only exception is when ω_t follows the "FastSine" dynamic where the SMM model marginally outperfoms DAMM. Gains with respect to the SMM ranges between 7% and 94% depending on the complexity of the true DGP. Interestingly, we find that the TVM model is not able to correctly recover the true signal.

[15] Among the various updating schemes proposed by the authors, the one we employ here reports best results in our experiment. Further details are reported in Appendix A.

Spec	Const	Sine	FastSine	Step	Ramp	$\mathcal{M}_1$	$\mathcal{M}_2$
MAE							
G–DAMM	0.27	0.88	0.74	1.01	0.54	0.93	0.63
SMM	1.00	1.00	1.00	1.00	1.00	1.00	1.00
TVM	0.87	1.59	1.27	7.46	1.34	1.88	1.98
MSE							
G–DAMM	0.06	0.95	0.74	0.98	0.38	0.87	0.43
SMM	1.00	1.00	1.00	1.00	1.00	1.00	1.00
TVM	0.87	2.20	1.62	33.28	1.82	3.49	3.03

Table 2: Mean Absolute Error (MAE) and Mean Squared Error (MSE) across the true patterns assumed for the weight of a mixture of two univariate Gaussian distributions and the filtered wights according to: i) the Time–Varying Mixture model constructed using the mixture weights updating mechanism detailed in Haas et al. (2013) (see Appendix A), ii) the static Mixture Model estimated on a rolling basis (SMM) using the last 100 observations and, iii) the DAMM with two Gaussian distributions detailed in Section 4 (G–DAMM). For each artificial weight pattern, $B = 500$ bivariate time series of length $T = 1000$ are simulated. For each $b = 1, \dots, B$, the MAE and MSE quantities are evaluated according to the different models, median values over the number of replicates are then compute. Results are reported relative to the SMM model acting as a benchmark. Values lower then one indicate outperformance with respect to the benchmark and viceversa. Gray cells indicate the best performing model.

6. Empirical Application

Our empirical application is divided in three parts. The first part consists in an in–sample analysis where we investigate which DAMM specification better fits financial time series. A comparison with alternative models via likelihood information criteria is also reported. In this part we also investigate whether the inclusion of fat–tailed distributions as mixture components as well as if time–variation in the mixture composition is required by the data. The second part of the analysis deals with the important task of predicting future risk measures at different forecast horizons $h > 0$ in the form of Value–at–Risk (VaR) and Expected–Shortfall (ES) levels, see e.g. Jorion (2006). The last part extends the previous analysis to the case of a portfolio composed of two assets. VaR and ES are the most commonly used risk measures in finance and also play a central role in nowadays financial regulation, see Basel (2012). Practically, VaR indicates the maximum loss of a long investment strategy for a given probability level $\alpha \in (0, 1)$. From a statistical point of view, the h–step ahead VaR at confidence level α, $VaR^{\alpha}_{T+h|T}$, coincides with the α–quantile of the h–step ahead conditional density indicated as $p_{T+h|T}(y_{t+h})$, where T is the length of the in–sample period in which the models' parameters are estimated. Denoting by $F_{T+h|T}(\cdot)$ the predictive cumulative density function associated to $p_{T+h|T}(\cdot)$, the VaR risk measure is evaluated as:

$$VaR^{\alpha}_{T+h|T} = F^{-1}_{T+h|T}(\alpha),$$

where $F^{-1}_{T+h|T}(\alpha)$ is the inverse of the cumulative density function, i.e. the quantile function, evaluated in α. ES measures the average loss after the VaR threshold and can be evaluated by integrating the VaR as follows:

$$ES^{\alpha}_{T+h|T} = \frac{1}{\alpha}\int_0^{\alpha} VaR^{\gamma}_{T+h|T} d\gamma.$$

Further details about the evaluation of these two quantities are reported in Section 6.3 where a study on volatility prediction is also reported. We now first describe the data set used for the empirical study.

6.1. Data set

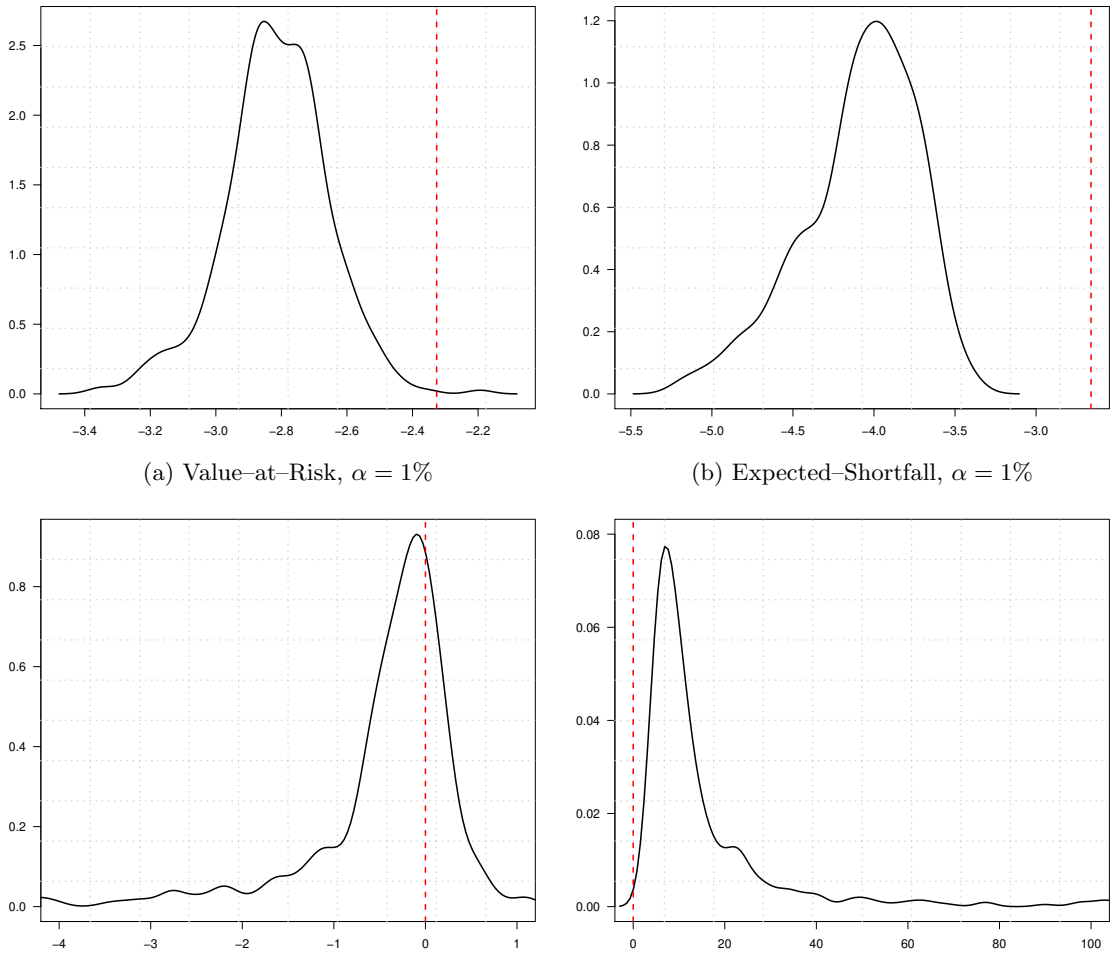
Our data set is composed by log–returns of the S&P 500 index constituents in percentage points recorded from January 2, 2000 to October 21, 2016. The length of the series ranges between 4,005 and 4,227. We consider the S&P 500 composition from January 2016 which consists of 506 US domiciled companies. Starting from the whole data set, series with less than 4,000 observations have been removed leaving us with a cross section of 403 companies. If present, serial autocorrelation has been removed using an autoregressive filter, finally all series are demeaned. Data is retrieved from Datastream.

From Section 2 we know that the conditional distribution of these series varies over time. Here we briefly report additional results regarding their unconditional distribution. For this part of the analysis we standardize the series by their unconditional standard deviation to improve comparability. We compute the VaR and ES measures at confidence level $\alpha = 1\%$ using the empirical distribution of the data. Gaussian kernel densities are estimated over the cross section of VaR and ES and reported in Figures 3a and 3b, respectively. Reference values according to a standardized Gaussian distribution are indicated by vertical dashed lines. We clearly observe that the tail of the unconditional distribution of the considered time series strongly departs from the Gaussian one. Indeed, VaR and ES are on average -2.8 and -4.1 which is far from their reference Gaussian values of -2.3 and -2.6, respectively. These findings suggest that a fat–tailed distribution is required by the data.

With regards to the overall shape of the unconditional distribution of returns, Figures 3c and 3d display the kernel densities evaluated on the skewness and excess of kurtosis coefficients, respectively. Again, the discrepancies between the empirical distribution of returns and the reference Gaussian distribution are well documented. Interestingly, we note that the majority of time series exhibit negative skewness indicating that negative returns (losses) are on average of bigger magnitude than positive returns (gains). We also computed the Jarque–Bera test statistic and the ARCH–LM test (not reported) of Jarque and Bera (1980) and Engle (1982), respectively. We find that the null hypothesis of Gaussianity and absence of ARCH effects cannot be rejected for all series at the usual confidence levels.

6.2. In–sample Analysis

We estimate a series of alternative models using all the available observations. Competing models are then compared in terms of their Akaike (AIC), Bayesian (BIC), and Hannan–Quinn (HQC) information criteria. Models with lower values of AIC, BIC, and HQC are preferred. We consider the Gaussian and Student's t DAMM specifications detailed in Section 4 labeled as G–DAMM and t–DAMM, respectively. For the whole analysis, we focus on the case of $J = 2$ mixture components. Models with $J = 3$ and $J = 4$ are rarely selected according to the considered information criteria. The two DAMM specifications are compared with related models proposed in the literature. As detailed in Section 4, these models can be obtained as special cases of the G–DAMM specification after a different choice of the score scaling matrix and the mapping function. The alternative models we consider are: i) the Mixture of ARCH models proposed by Wong and Li (2001) (MixARCH), ii) the Mixture of GARCH models independently proposed by Haas et al. (2004a), Alexander and Lazar (2006) and Zhang et al. (2006) (MixGARCH), and iii) the Time–Varying Mixture of GARCH



(a) Value–at–Risk, $\alpha = 1\%$

(b) Expected–Shortfall, $\alpha = 1\%$

(c) Skewness

(d) Excess of kurtosis

Figure 3: Gaussian kernel densities of the VaR (a), ES (b), skewness (c), and excess of kurtosis (d) coefficients evaluated on the empirical standardized distribution the S&P500 constituents using the full sample. VaR and ES are evaluated at the confidence level $\alpha = 1\%$. Red vertical dashed lines report the reference values according to a standard Gaussian distribution.

models with likelihood–driven mixing weights of Haas et al. (2013) (TVMixGARCH).[16] These models assume that returns are conditionally distributed according to a mixture of Gaussian distributions. Haas et al. (2004a) also investigate whether the specification of a mixture of Student's t instead of Gaussian distributions in the MixGARCH model is required for financial data. Their results indicate that a mixture of Gaussian distribution is a better choice. Below we report a similar analysis for DAMMs.

All models are estimated by Maximum Likelihood using the full sample of the data.[17] Table

[16] Additional details on models formulation are reported in Appendix A.

[17] Global optimization via the differential evolution algorithm of Storn and Price (1997) has been employed via the R implementation provided by the **DEoptim** package of Ardia et al. (2015), see also Mullen et al. (2011). The initial population for the algorithm has been selected after perturbation of the starting values obtained as detailed in Section 3.1.

	MixARCH	MixGARCH	TVMixGARCH	G–DAMM	t–DAMM
Panel A					
AIC	0.0	6.4	11.0	82.7	–
BIC	0.0	17.3	6.4	76.3	–
HQC	0.0	9.5	11.7	78.8	–
Panel B					
AIC	0.0	1.1	2.8	7.8	88.3
BIC	0.0	8.8	3.5	12.7	74.9
HQC	0.0	2.8	4.6	9.2	83.4

Table 3: Percentage of times when models are selected according to the Akaike (AIC), Bayesian (BIC), and Hannan–Quinn (HQC) information criteria. The cross section of time series is composed by the log–returns of 403 US domiciled companies spanning the period from January 2, 2000 to October 21, 2016. The length of the series ranges between 4005 and 4227. All models are estimated with two mixture components ($J = 2$) by Maximum Likelihood. Along with the G–DAMM and t–DAMM specifications, the following models are considered: i) the Mixture of ARCH models proposed by Wong and Li (2001) (MixARCH), ii) the Mixture of GARCH models independently proposed by Haas et al. (2004a), Alexander and Lazar (2006) and Zhang et al. (2006) (MixGARCH), and iii) the Time–Varying Mixture of GARCH models with likelihood–driven mixing weights of Haas et al. (2013) (TVMixGARCH). Further details on these models are reported in Appendix A.

3 reports the percentage of times each specification is preferred according to the three information criteria. Panel A of Table 3 compares only those specifications that assume a mixture of Gaussian distributions. In Panel B, we also add the t–DAMM model to the comparison.

Looking at Panel A of Table 3, we note that among those models which assume a mixture of Gaussian distributions, G–DAMM is preferred 83% of the time according to AIC, 76% times according to BIC and 79% of the time according to HQC. That is, G–DAMM is preferred most of times according to the three information criteria. Looking at Panel B we find that t–DAMM is selected 88% of the time according to AIC, 75% of times according to BIC, and 83% of times according to HQC. Interestingly, we find that, differently from the Mixture of GARCH models of Haas et al. (2004a), for DAMM the specification of a mixture of Student's t distributions is required by the data. This result is directly linked to the use of the Score–Driven methodology where a change of the conditional distribution assumption also implies a different filter for the volatility process, see Section 4. As a final result, we find that the Mixture of ARCH models (MixARCH) of Wong and Li (2001) is never selected according to all the information criteria we consider.

6.2.1. Time–varying Mixture Composition for DAMMs

We conclude the in–sample analysis by looking at the implication of specifying time–varying mixtures for DAMMs. We would like to remark that time–varying mixtures are common specifications in (financial) time series econometrics. Indeed, with regards to the financial econometrics literature, along with the Time–Varying Mixture of GARCH models of Haas et al. (2013), all Markov–switching models implicitly specify the predictive distribution as a time–varying mixture. The difference between Markov–switching models and, for instance, the model of Haas et al. (2013) as well as DAMM, is that in the latter case time–variation in the mixture composition arises from the predictive probabilities implied by the Hamilton filter. Differently, in the model of Haas et al. (2013) and in DAMM, time–variation in the mixture composition is explicitly modeled.

We repeat the same in–sample analysis as before by looking at constrained versions of DAMM

	G–DAMM	G–DAMM–$\bar{\omega}$	t–DAMM	t–DAMM–$\bar{\omega}$
Panel A				
AIC	89.4	10.6	–	–
BIC	64.7	35.3	–	–
HQC	82.7	17.3	–	–
Panel B				
AIC	–	–	87.3	12.7
BIC	–	–	67.8	32.2
HQC	–	–	80.9	19.1
Panel C				
AIC	8.5	0.4	80.6	10.6
BIC	12.4	7.4	55.8	24.4
HQC	10.6	1.4	73.1	14.8

Table 4: Percentage of times models are selected according to the Akaike (AIC), Bayesian (BIC), and Hannan–Quinn (HQC) information criteria. The cross section of time series is composed by the log–returns of 403 US domiciled companies spanning the period from January 2, 2000 to October 21, 2016. The length of the series ranges between 4005 and 4227. All models are estimated with two mixture components ($J = 2$) by Maximum Likelihood. The table reports a comparison between the DAMMs with mixture of Gaussian (G–DAMM) and mixture of Student's t (t–DAMM) along with their constrained versions with constant mixture compositions G–DAMM–$\bar{\omega}$ and t–DAMM–$\bar{\omega}$, respectively. Further details on the models are reported in Appendix A.

with constant mixture composition. As detailed in Section 4, such constrained versions are obtained by setting $\mathbf{A}^{\omega} = \mathbf{B}^{\omega} = \mathbf{0}$ in equation (5). Table 4 reports the AIC, BIC, and HQC criteria for the G–DAMM and t–DAMM specifications along with their constrained versions labeled as G–DAMM–$\bar{\omega}$ and t–DAMM–$\bar{\omega}$, respectively.

As before, Panel A compares mixture of Gaussian specifications, Panel B compares mixture of Student's t specifications, while Panel C reports an overall comparison. Results are very clear and show that time–varying mixtures are required by the data. Indeed, the unrestricted G–DAMM specification is selected 89% of the time according to AIC, 65% according to BIC, and 83% of the time according to HQC. Results are similar for the t–DAMM case. Overall, as emerged from the previous analysis, we find that t–DAMM is the most frequently selected model according to the three information criteria.

6.3. Out–of–sample Analysis

The out–of–sample analysis focuses on prediction of future VaR, ES, and volatility levels. We select three forecast horizons: i) one day ahead, $h = 1$, ii) one week ahead, $h = 5$, and iii) one month ahead, $h = 20$. The estimation period is indicated with T and is fixed to 2,000 observations, i.e. around half of the full sample. As for the Mixture of GARCH models, in DAMM only the one–step ahead predictive distribution, $y_{T+1}|\mathcal{F}_T$, is available in closed form. The multi–step ahead distribution ($h > 1$) is estimated by standard Monte Carlo simulation techniques.[18] The analysis is

[18]Specifically, we start by sampling 10,000 observations from the one–step ahead distribution, $y_{T+1}|\mathcal{F}_T$. This is a simple mixture of distributions and standard routines can be used. Subsequently, for each $l = 1, \ldots, h$ we update the time–varying model's parameters using the updating equations reported in Section 3 and we iterate sampling from

performed using a rolling window approach keeping the length of the estimation period constant to 2,000 observations. The models' parameters are updated every 10 observations.

For $h > 1$, $VaR^{\alpha}_{T+h|T}$ is evaluated as the empirical quantile at level α of the simulated draws, $ES^{\alpha}_{T+h|T}$ is evaluated as the average of the simulated draws being below the $VaR^{\alpha}_{T+h|T}$ level. Using the same methodology, the h–step ahead predictive volatility level, $\sigma_{T+h|t}$, is evaluated as the standard deviation of the simulated observations. For VaR and ES we consider the two probability levels $\alpha = 1\%$ and $\alpha = 5\%$ as these are the most common choices in similar studies.

Model comparison is performed via the definition of a loss function. The loss function is specified such that it compares the predicted quantity with the realized observation and maps in a real positive value. Values are subsequently averaged over the forecasting period and models with lower average loss are preferred. Different loss functions for the same quantity of interest are available in the literature, here we employ the most common ones. The usual loss function for VaR predictions is the so–called *tick* function defined as:

$$\ell^{VaR}_{T+h|T}(y_{T+h}, VaR^{\alpha}_{T+h|T}) = (\alpha - \mathbb{1}_{\{y_{T+h} < VaR^{\alpha}_{T+h|T}\}})(y_{T+h} - VaR^{\alpha}_{T+h|T}).$$

Evidently, $\ell^{VaR}_{T+h|T}(\cdot,\cdot)$ is an asymmetric loss function which penalizes more heavily observations below the predicted VaR level. It can easily be shown that $\min_{u}\{\mathbb{E}[\ell^{VaR}_{T+h|T}(y_{T+h}, u)]\} = 0$ when u coincides with the true h–step ahead quantile, see Koenker (2005). This property implies that VaR is an *elicitable* risk measure, which means that it can be obtained as the unique minimizer of a loss function. Unfortunately, ES does not share the elicitability properties with VaR. However, recently Fissler and Ziegel (2016) show that even if ES is not elicitable, it is a component of a higher order elicitable functional also defined in terms of VaR. This property implies that there exists a family of functions so that their unique minimizer equals to the true pair of VaR and ES. Among these functions, we employ the one used by Patton et al. (2017) which assumes that $VaR^{\alpha}_{T+h|T} < 0$ and $ES^{\alpha}_{T+h|T} < VaR^{\alpha}_{T+h|T}$; two conditions that hold by construction in our case. The joint loss function for ES and VaR is defined as:

$$\begin{aligned}\ell^{VaR,ES}_{T+h|T}(y_{T+h}, VaR^{\alpha}_{T+h|T}, ES^{\alpha}_{T+h|T}) =& \frac{\mathbb{1}_{\{y_{T+h} < VaR^{\alpha}_{T+h|T}\}}(y_{T+h} - VaR^{\alpha}_{T+h|T})}{\alpha ES^{\alpha}_{T+h|T}} \\ &+ \frac{VaR^{\alpha}_{T+h|T}}{ES^{\alpha}_{T+h|T}} + \log(-ES^{\alpha}_{T+h|T}) - 1.\end{aligned}$$

The reader is referred to Patton et al. (2017) for additional details. Finally, with regards to volatility prediction, we employ the QLIKE loss function defined as:

$$\ell^{\sigma}_{T+h|T}(\sigma_{T+h|T}, \sigma^{*}_{T+h|T}) = \log\left(\frac{\sigma^2_{T+h|T}}{\sigma^{*2}_{T+h|T}}\right) + \frac{\sigma^2_{T+h|T}}{\sigma^{*2}_{T+h|T}} - 1,$$

where $\sigma^{*}_{T+h|T}$ is a volatility proxy at time $T+h$. The use of imperfect proxies for volatility forecast comparison can have dramatic consequences for the ranking of alternative models, see Andersen and

the one–step ahead distribution conditional on previous simulated draws until the end of the forecast horizon. Last simulated observations are draws from the distribution $y_{T+h}|\mathcal{F}_T$.

	ℓ^{σ}	ℓ^{VaR}		$\ell^{VaR,ES}$	
		$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 5\%$
$h = 1$					
MixARCH	1.00	1.00	1.00	1.00	1.00
MixGARCH	0.78	0.79	0.82	0.88	0.86
TVMixGARCH	0.78	0.79	0.83	0.89	0.86
G–DAMM	0.77	0.76	0.82	0.87	0.84
t–DAMM	0.78	0.79	0.87	0.89	0.91
$h = 5$					
MixARCH	1.00	1.00	1.00	1.00	1.00
MixGARCH	0.52	0.57	0.63	0.75	0.69
TVMixGARCH	0.52	0.58	0.63	0.75	0.71
G–DAMM	0.52	0.56	0.62	0.74	0.68
t–DAMM	0.52	0.57	0.66	0.75	0.73
$h = 20$					
MixARCH	1.00	1.00	1.00	1.00	1.00
MixGARCH	0.50	0.57	0.64	0.74	0.70
TVMixGARCH	0.50	0.57	0.65	0.75	0.71
G–DAMM	0.49	0.56	0.62	0.74	0.68
t–DAMM	0.50	0.57	0.66	0.76	0.75

Table 5: This table reports the median average loss over the cross section of 403 US domiciled companies. The second column reports results for the QLIKE volatility loss function, ℓ^{σ}. The third and fourth columns report results for the VaR loss function, ℓ^{VaR}, at confidence levels $\alpha = 1\%$ and $\alpha = 5\%$, respectively. Last two columns report results for the joint loss function defined for VaR and ES, $\ell^{VaR,ES}$, at confidence levels $\alpha = 1\%$ and $\alpha = 5\%$, respectively. Outputs are evaluated as follows: for each asset, we compute the average loss over the out–of–sample period, then we compute the median over the cross section dimension. Results are reported relative to the MixARCH model which acts as a benchmark. Values lower than one indicate outperformance with respect to the benchmark. Gray cells indicate best performing models. The out–of–sample period spans from January 14, 2008 to October 21, 2016. The length of the out–of–sample period ranges between 2,005 and 2,227. The length of the estimation period is fixed to $T = 2000$ observations, models' parameters are re–estimated each 10 observations during the forecast exercise. The models' details are reported in Appendix A.

Bollerslev (1998) and Patton (2011). Given the large cross section of data we use in this paper, the best volatility proxy we can obtain is given by the squared returns $\sigma^{*2}_{T+h|T} = y^2_{T+h}$. Squared returns are known to be bad volatility proxies, however, as shown by Patton (2011), QLIKE is robust with respect to this choice and still provides an appropriate ranking of alternative models.

6.3.1. Forecasting Results

The comparison across models is reported in Table 5. The table reports the median average loss over the cross section of 403 US domiciled companies for the three loss functions and the two quantile confidence levels. The entries of Table 5 are evaluated as follow: for each asset, we compute the average loss over the out–of–sample period, then we compute the median over the cross section dimension. Results are reported relative to the MixARCH model which acts as a benchmark. Values lower than one indicate outperformance with respect to the benchmark. Gray cells indicate the best

	ℓ^{σ}	ℓ^{VaR}		$\ell^{VaR,ES}$	
		$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 5\%$
Panel A					
$h = 1$	29.1	27.6	35.8	8.5	11.7
$h = 5$	23.1	35.3	26.4	7.2	9.2
$h = 20$	35.1	40.5	17.4	11.2	13.4
Panel B					
$h = 1$	10.7	33.3	10.2	3.7	4.7
$h = 5$	13.9	33.1	13.7	6.7	9.2
$h = 20$	14.7	32.8	24.6	9.5	10.4

Table 6: This table reports the percentage of times the bilateral null hypothesis of equal predictive ability between the G–DAMM and MixGARCH models is rejected at the 5% confidence level. Panel A reports the percentage of times the null hypothesis is rejected and the test statistics favors G–DAMM. Panel B reports the percentage of times the null hypothesis is rejected and the test statistics favors MixGARCH. Results are reported for the three forecast horizons $h = 1, 5, 20$ as well as for the different loss functions, ℓ^{σ}, ℓ^{VaR} and $\ell^{VaR,ES}$. The VaR and the joint VaR, ES loss functions are evaluated for the two quantile levels $\alpha = 1\%$ and $\alpha = 5\%$.

performing model. The out–of–sample period spans from January 14, 2008 to October 21, 2016. The length of the out–of–sample period ranges between 2,005 to 2,227, which is the length of the full sample minus the length of the estimation period fixed to $T = 2,000$.

Results are very clear and indicate G–DAMM as the best performing model. Indeed, it always reports lower loss values for all loss functions and forecast horizons. Gains with respect to the benchmark model are huge suggesting that the MixARCH specification should not be used for risk management purposes. Interestingly, we find that G–DAMM outperforms t–DAMM in our large scale analysis. The reason for the outperformance of G–DAMM over t–DAMM may be related to the additional estimation error which is transmitted to the predictions. Probably, the use of a longer estimation period would increase the performance of t–DAMM, however, we leave this aspect for future research.

It is worth to be mentioned that in Table 5 we have not provided any measure of uncertainty across the ranking of the models. Indeed, we report a simple loss comparison without performing any statistical test. For the last part of the analysis, we test the null hypothesis of equal predictive ability between the forecasts provided by the G–DAMM and MixGARCH models. Hypothesis testing for each series and loss function is carried out via a bilateral Diebold–Mariano (DM) test between the forecasts delivered by the two competing models. Heteroskedasticity– and autocorrelation–consistent standard errors are computed according to the estimator proposed by West and Newey (1987).

Table 6 reports the percentage of times the bilateral null hypothesis of equal predictive ability between the G–DAMM and MixGARCH models is rejected at the 5% confidence level. Panel A reports the percentage of times the null hypothesis is rejected and the test statistics favors G–DAMM. Panel B reports the percentage of times the null hypothesis is rejected and the test statistics favors MixGARCH. We find that generally G–DAMM reports more often average losses which are statistically lower than MixGARCH. For instance, looking at Panel A of Table 6 we see that the VaR losses at $\alpha = 5\%$ for G–DAMM are about 36% of the time lower than those delivered by MixGARCH. Analogously, from Panel B we see that the frequency of times MixGARCH reports lower VaR losses is only 10%. In the remaining 54% of the time the null hypothesis of equal predictive ability cannot be

	ℓ^{σ}	ℓ^{VaR}		$\ell^{VaR,ES}$	
		$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 5\%$
Panel A					
$h = 1$	18.3	35.2	23.9	9.9	7.0
$h = 5$	18.6	45.7	25.7	5.7	10.0
$h = 20$	15.7	47.1	17.1	5.7	7.1
Panel B					
$h = 1$	7.0	14.1	4.2	5.6	5.6
$h = 5$	15.7	14.3	11.4	5.7	11.4
$h = 20$	21.4	14.3	18.6	11.4	18.6

Table 7: This table reports the percentage of times the bilateral null hypothesis of equal predictive ability between the G–DAMM and its constrained version with constant mixture composition labeled as G–DAMM–$\bar{\omega}$, is rejected at the 5% confidence level. Panel A reports the percentage of times the null hypothesis is rejected and the test statistics favors G–DAMM. Panel B reports the percentage of times the null hypothesis is rejected and the test statistics favors G–DAMM–$\bar{\omega}$. Results are reported for the three forecast horizons $h = 1, 5, 20$ as well as for the different loss functions, ℓ^{σ}, ℓ^{VaR} and $\ell^{VaR,ES}$. The VaR and the joint VaR, ES loss functions are evaluated for the two quantile levels $\alpha = 1\%$ and $\alpha = 5\%$.

rejected at the 5% confidence level. Similar results are found for the volatility prediction: G–DAMM outperforms MixGARCH 29% of the time for $h = 1$, 23% for $h = 5$ and, 35% of the time for $h = 20$. Percentages of outperformance for MixGARCH over G–DAMM are 11%, 14%, and 15% for $h = 1$, $h = 5$, and $h = 20$, respectively.

6.3.2. Do We Need Time–varying Mixtures?

In Section 6.2.1 we have shown that allowing for time–variation in the mixture composition strongly improves the fit of the data according to the three information criteria we consider. We now investigate if these improvements also translate in better out–of–sample forecasts. We consider again the constrained version of the G–DAMM specification with constant mixture composition labeled as G–DAMM–$\bar{\omega}$. We perform the same pairwise comparison of Section 6.3.1 reporting the percentage of times the null hypothesis of equal predictive ability between the time–varying and static mixture specifications is rejected at the 5% confidence level according to the DM test statistics.

Results are reported in Table 7. Analogously to the previous analysis, in Panel A we report the percentage of times the null hypothesis is rejected and the test statistics favors G–DAMM while Panel B reports the percentage of times the null hypothesis is rejected and the test statistics favors G–DAMM–$\bar{\omega}$. We find that for short forecast horizons ($h = 1$ and $h = 5$), the specification of time–varying mixtures generally improves the prediction of volatility and risk levels. For instance, G–DAMM outperforms G–DAMM–$\bar{\omega}$ 46% of the time for five days ahead VaR predictions at confidence level 1% while G–DAMM–$\bar{\omega}$ outperforms G–DAMM only 14% of the time. Evidence is less pronounced for the joint loss function for VaR and ES. Interestingly, when we look at the long forecast horizon, $h = 20$, we note that G–DAMM–$\bar{\omega}$ is generally preferred suggesting that this specification may be more adequate to capture long–term characteristics of the data.[19]

[19]Results for the t–DAMM specification are similar and are not reported to save space. They are available upon request.

6.4. Portfolio Risk Levels Predictions

	ℓ^{VaR}		$\ell^{VaR,ES}$	
	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 5\%$
$\lambda = 0.2$				
DCC	1.00	1.00	1.00	1.00
MG–DAMM	0.94	1.00	0.94	1.02
MG–DAMM–$\bar{\omega}$	0.97	1.00	0.98	1.02
$\lambda = 0.5$				
DCC	1.00	1.00	1.00	1.00
MG–DAMM	0.92	1.02	0.92	1.05
MG–DAMM–$\bar{\omega}$	0.95	1.02	0.96	1.04
$\lambda = 0.8$				
DCC	1.00	1.00	1.00	1.00
MG–DAMM	0.89	1.06	0.92	1.08
MG–DAMM–$\bar{\omega}$	0.93	1.06	0.95	1.08

Table 8: This table reports the one–step ahead median average loss over the portfolios which assign weight $(1 - \lambda)$ to the S&P 500 and λ to one company included in the cross section of 403 US domiciled companies. The second and third columns report results for the VaR loss function, ℓ^{VaR}, at confidence levels $\alpha = 1\%$ and $\alpha = 5\%$, respectively. The last two columns report results for the joint loss function defined for VaR and ES, $\ell^{VaR,ES}$, at confidence levels $\alpha = 1\%$ and $\alpha = 5\%$, respectively. Outputs are evaluated as follows: for each asset we compute the average loss over the out–of–sample period, then we compute the median over the cross section dimension. Results are reported relative to the DCC model which acts as a benchmark. Values lower than one indicate outperformance with respect to the benchmark. Gray cells indicate best performing models. The out–of–sample period spans from January 14, 2008 to October 21, 2016. The length of the out–of–sample period ranges between 2005 and 2227. The length of the estimation period is fixed to $T = 2,000$ observations, models' parameters are re–estimated each 10 observations during the forecast exercise.

So far we have concentrated on univariate DAMM specifications. In the remaining of this empirical analysis we only focus on the MG–DAMM specification as detailed in Section 4.1. We do so by considering the VaR and ES predictions associated to a portfolio of two assets. The first asset is one company of our data set and the second asset is the S&P 500 index. We assume to invest one dollar and assign weight λ to the company, "c" with $c = 1, \ldots, 403$, and $(1 - \lambda)$ to the S&P 500 index, "sp". The vector of two assets, $\mathbf{y}_t = (y_t^c, y_t^{sp})'$, is assumed to follow the MG–DAMM specification with two mixture components ($J = 2$). The constrained specification with constant mixture weights is also considered and labeled as MG–DAMM–$\bar{\omega}$. We follow the same forecast design of the previous analysis: i) the estimation period is fixed to $T = 2000$, ii) static models' parameters are updated every 10 observations, and iii) one–step ahead predictions for the portfolio's VaR and ES at levels $\alpha = 1\%$ and $\alpha = 5\%$ are computed.

VaR and ES predictions of the one–step ahead portfolio's distribution can easily be computed for MG–DAMM. We first note that the one–step ahead distribution of (y_t^c, y_t^{sp}) is a mixture of two bivariate Gaussian distributions. It follows that the one–step ahead portfolio's distribution is a mixture of two univariate Gaussian distributions with mean $\mu_j^p = \boldsymbol{\lambda}'\boldsymbol{\mu}_j$ and variance ${\sigma_{j,t}^p}^2 = \boldsymbol{\lambda}'\boldsymbol{\Sigma}_{j,t}\boldsymbol{\lambda}$, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_{j,t}$ are the mean vector and covariance matrix of the j–th mixture component and $\boldsymbol{\lambda} = (\lambda, 1 - \lambda)'$. The portfolio's VaR at time $T + 1$ given information up to time T at risk level α,

$VaR^{\alpha p}_{T+1|T}$, is then obtained by inverting the predictive portfolio's distribution while the portfolio's ES is obtained as

$$ES^{\alpha p}_{T+1|T} = \sum_{j=1}^{2} \omega_{j,T+1} \frac{\Phi(h_{j,T+1})}{\alpha} \left[\mu^p_j - \sigma^p_{j,T+1} \frac{\phi(h_{j,T+1})}{\Phi(h_{j,T+1})} \right],$$

where $h_{j,T+1} = (VaR^{\alpha p}_{T+1|T} - \mu^p_j)/\sigma^p_{j,T+1}$, see Broda and Paolella (2011).

We consider the two loss functions ℓ^{VaR} and $\ell^{VaR,ES}$ and compare the MG–DAMM and MG–DAMM–$\bar{\omega}$ predictions with those delivered by the Dynamic Conditional Correlation (DCC) model of Engle (2002) with multivariate Gaussian innovations for three portfolio compositions: $\lambda \in \{0.2, 0.5, 0.8\}$. Table 8 reports the median average loss values for the three competing models. As before, losses are averaged over the forecast period and then the median is computed over the cross section dimension. Results are reported relative to the DCC model. Unlike the univariate application, we find that DAMM specifications are preferred over the simpler DCC model only when $\alpha = 1\%$ and not when $\alpha = 5\%$. This result suggests that a more sophisticated model is required when the risk manager focuses on high–risk levels ($\alpha = 1\%$) and not on moderate risk levels ($\alpha = 5\%$). Indeed, when $\alpha = 1\%$, we observe gains in the range 11% – 6% for MG–DAMM and 7% – 2% for MG–DAMM–$\bar{\omega}$. Consistent with previous findings, we observe that allowing the mixture weights to change over time implies better risk level predictions.

6.5. Density Predictions with MG–DAMM

	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$	Joint
DCC	1.00	1.00	1.00	1.00
MG–DAMM	1.00	0.99	0.97	0.98
MG–DAMM–$\bar{\omega}$	1.01	0.99	0.98	0.98

Table 9: This table reports the one–step ahead median average negative log score over the portfolios which assign weight $(1 - \lambda)$ to the S&P 500 and λ to one company included in the cross section of 403 US domiciled companies as well as for the joint distribution of the company and S&P 500 index. Outputs are evaluated as follows: for each portfolio and couple (y^c_t, y^{sp}_t), we compute the average NLS over the out–of–sample period, then we compute the median over the cross section dimension. Results are reported relative to the DCC model which acts as a benchmark. Values lower than one indicate outperformance with respect to the benchmark. Gray cells indicate best performing models. The out–of–sample period spans from January 14, 2008 to October 21, 2016. The length of the out–of–sample period ranges between 2,005 and 2,227. The length of the estimation period is fixed to $T = 2,000$ observations, models' parameters are re–estimated each 10 observations during the forecast exercise.

We conclude our analysis by comparing MG–DAMM and DCC in the prediction of the one–step ahead distribution of the data. We concentrate on the joint distribution of the pair (y^c_t, y^{sp}_t), as well as on the distribution of the portfolio $y^p_t = \lambda y^c_t + (1 - \lambda) y^{sp}_t$ for all $c = 1, \ldots, 403$ and $\lambda \in \{0.2, 0.5, 0.8\}$. Both distributions can be computed in closed form as detailed in the previous subsection. Results are compared with DCC using the negative log score (NLS) as a measure of predictive ability.[20] NLS for a density prediction at time $T + 1$ conditional on information available up to time T, $NLS_{T+1|T}$ is defined as:

[20] NLS is a proper scoring rule which defines a complete ordering over different models, see Gneiting et al. (2007) for additional details.

$$NLS_{T+1|T} = -\log p(\mathbf{y}_{T+1}|\mathcal{F}_T, \widehat{\boldsymbol{\theta}}_{T+1}),$$

where we emphasize that $\widehat{\boldsymbol{\theta}}_{T+1}$ is the prediction of the parameter value at time $T+1$. Table 9 reports the median average NLS for the three portfolios and the joint distribution of each company and the S&P 500 index. Results are reported relative to the DCC specifications. Values lower than one indicates outperformance of MG–DAMM or MG–DAMM–$\bar{\omega}$ over DCC. We find that the two MG–DAMM specifications improve density predictions by a factor of 3% - 1% in all cases excluding the portfolio with $\lambda = 0.2$. We also note that when the investor assigns weight $\lambda = 0.8$ to the company (and thus 0.2 to the S&P 500 index), MG–DAMM reports higher gains than in the case $\lambda = 0.5$. The result from this analysis is thus that for a portfolio that gives a higher weight to a more volatile asset (like the single company over the S&P 500 index), it is better to employ a more flexible specification. Results for the joint distribution also favor MG–DAMM over DCC and are reported in the last column of Table 9. Finally, also for this analysis, we observe that having time–varying mixture weights improves predictions of financial returns.

7. Conclusion

In this paper, we proposed a new general class of dynamic mixture models named DAMM. We allow the mixture components as well as the mixture composition to change over time, exploiting the information contained in the scaled score of the conditional distribution assumed for the data as suggested by Creal et al. (2013) and Harvey (2013). Starting from the general model formulation, three models for financial econometrics applications are developed. We show that when Gaussian distributions are chosen as mixture components, and the conditional score is properly scaled, the proposed model nests other models previously proposed in the literature. Additional multivariate DAMM specifications are reported in the supplementary appendix accompanying this paper.

Two Monte Carlo experiments investigate the filtering ability of DAMMs in terms of time–varying correlation and time–varying mixture composition. We find that DAMM outperforms alternative specifications available in the literature in terms of filtering. The supplementary material accompanying this paper also investigates the finite sample properties of the Maximum Likelihood estimator for a DAMM with Student's t mixture components. We find that a sample size between $T = 1,000$ and $T = 5,000$ is enough in order to estimate very flexible DAMMs specifications.

The paper also contributes from an empirical perspective by reporting a large scale empirical study on the log–returns time series of 403 US domiciled companies covering the period from January 2, 2000 to October 21, 2016. The in–sample and out–of–sample analyses favour the new specification over alternative models such as: i) the Mixture of ARCH models proposed by Wong and Li (2001), ii) the Mixture of GARCH models independently proposed by Haas et al. (2004a), Alexander and Lazar (2006) and Zhang et al. (2006), and iii) the Time–Varying Mixture of GARCH models with likelihood–driven mixing weights of Haas et al. (2013).
In contrast to Haas et al. (2004a), we found that the specification of mixture of Student's t distributions for financial returns is required according to our in–sample analysis. However, we find that for the out–of–sample analysis in quantitative risk management, the simpler specification with a mixture of Gaussian distributions reports better results.

As also noted by an anonymous referee, we believe that DAMMs might be of interest also for other fields of application different from financial econometrics. For instance, applications in: i) process monitoring, ii) intervention detections, iii) insurance losses and, iv) graphical engineering where

alternative Dynamic Mixture Models have been successfully applied, are worth to be investigated. See the introduction of this paper for related references. The same referee also suggested to investigate the implications of allowing the mixture composition to be time–varying on the dynamic of other models' parameters. For instance, a comparison between the filtered volatility (or conditional mean) of the static and time–varying mixture specifications could be of interest.

Acknowledgments

I would like to thank the participants of the Statistical Methods session of the MAF 2016 conference held in Paris and in particular Anders Rahbek for useful comments. I would like to also thank participants of the Time Series Models session of the IAAE 2016 Annual Conference held in Milan, participants of the Nonlinear Models session of the 7th ICEEE conference held in Messina and, participants of the Time Series Models session of the 10th SoFiE conference held in New York. Feedbacks received during seminars at Vrije Universiteit Amsterdam, Aarhus University, Erasmus University, Carlos III university and Monash University are very valuable. Special thanks go to: Heather Anderson, George Athanasopoulos, Francisco Blasques, Dianne Cook, Fulvio Corsi, Davide Delle Monache, Dick van Dijk, Antoni Espasa, Niels Haldrup, Andrew Harvey, Rob Hyndman, Erik Kole, Siem Jan Koopman, Alessandra Luati, Jury Marcucci, Maura Mezzetti, Paolo Santucci de Magistris, Tommaso Proietti, Esther Ruiz, Soren Schwartz, Solveig Nygaard Sørensen, Timo Teräsvirta, and Helena Veiga.

Appendix A. Models Used in the Empirical Analysis

In order to allow for reproducibility of our results, in this appendix we report the exact specifications of the models we use in Section 6. This is required since some of these models were independently proposed by different authors using different formulations, as well as because within a single reference different model specifications may be provided. Along with the G–DAMM and t–DAMM specifications detailed in Section 4, the models we use in the empirical application are:

i) the Mixture of ARCH models proposed by Wong and Li (2001) (MixARCH).

ii) the Mixture of GARCH models independently proposed by Haas et al. (2004a), Alexander and Lazar (2006) and Zhang et al. (2006) (MixGARCH).

iii) the Time–Varying Mixture of GARCH models with likelihood–driven mixing weights of Haas et al. (2013) (TVMixGARCH).

We first introduce the TVMixGARCH model and then we recover the MixGARCH and MixARCH as constrained specifications. Let y_t be the log–return in percentage points at time t, the TVMixGARCH model with two components assumes that:

$$p(y_t|\mu_1, \sigma^2_{1,t}, \sigma^2_{2,t}, \omega_{1,t}, \omega_{2,t}) = \omega_{1,t}\frac{1}{\sigma_{1,t}}\phi\left(\frac{y_t - \mu_1}{\sigma_{1,t}}\right) + \omega_{2,t}\frac{1}{\sigma_{2,t}}\phi\left(\frac{y_t - \mu_{2,t}}{\sigma_{2,t}}\right),$$

where $\phi(\cdot)$ is the density of a standardized Gaussian distribution and $\mu_{2,t} = -(\omega_{1,t}/\omega_{2,t})\mu_1$ is imposed order to ensure $\mathbb{E}_{t-1}[y_t] = 0$. The mixture components variances evolve as standard GARCH(1,1) processes:

$$\sigma^2_{j,t+1} = \omega_j + \alpha_j y_t^2 + \beta_j \sigma^2_{j,t}, \quad j = 1, 2,$$

and the following likelihood–driven mechanism is assumed for the mixture weights:

$$\omega_{j,t} = \frac{W_{j,t}}{W_{1,t} + W_{2,t}}, \quad j = 1, 2,$$

where:

$$W_{j,t} = \kappa_j + \gamma \frac{\frac{1}{\sigma_{j,t}} \phi\left(\frac{y_t - \mu_{j,t}}{\sigma_{j,t}}\right)}{\frac{1}{\sigma_{1,t}} \phi\left(\frac{y_t - \mu_1}{\sigma_{1,t}}\right) + \frac{1}{\sigma_{2,t}} \phi\left(\frac{y_t - \mu_{2,t}}{\sigma_{2,t}}\right)}, \tag{A.1}$$

with $\kappa_j > 0$[21] for $j = 1, 2$, $\gamma > 0$, and $\mu_{j,t} = \mu_1$ when $j = 1$. Starting from the TVMixGARCH model, the MixGARCH is obtained by setting $\gamma = 0$. The MixARCH model is obtained by adding the following constraints: $\beta_1 = \beta_2 = 0$.

References

Alexander, C. and Lazar, E. (2006). Normal mixture garch (1, 1): Applications to exchange rate modelling. *Journal of Applied Econometrics*, 21(3):307–336.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, 39(4):885–905.

Ardia, D. (2008). *Financial risk management with Bayesian estimation of GARCH models*. Springer.

Ardia, D., Boudt, K., and Catania, L. (2016). Generalized Autoregressive Score Models in R: The GAS Package. *Journal of Statistical Software (forecaming)*.

Ardia, D., Mullen, K. M., Peterson, B. G., and Ulrich, J. (2015). ***DEoptim**: Differential Evolution in R*. version 2.2-3.

Basel (2012). 99th annual report. Technical report, Board of Governors of the Federal Reserve Systems.

Bauwens, L., Hafner, C., and Rombouts, J. V. (2007). Multivariate mixed normal conditional heteroskedasticity. *Computational Statistics & Data Analysis*, 51(7):3551–3566.

Bazzi, M., Blasques, F., Koopman, S. J., and Lucas, A. (2017). Time-varying transition probabilities for markov regime switching models. *Journal of Time Series Analysis*, 38(3):458–478.

Bernardi, M. and Catania, L. (2018). Switching-GAS Copula Models With Application to Systemic Risk. *Journal of Applied Econometrics*, 34(1):43–65.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2012). Combination schemes for turning point predictions. *The Quarterly Review of Economics and Finance*, 52(4):402–412.

Billio, M., Casarin, R., Ravazzolo, F., and Van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213–232.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blasques, F., Koopman, S. J., Łasak, K., and Lucas, A. (2016). In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation-driven models. *International Journal of Forecasting*, 32(3):875–887.

Blasques, F., Koopman, S. J., and Lucas, A. (2014). Maximum likelihood estimation for generalized autoregressive score models. Technical report, Tinbergen Institute Discussion Paper.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307 – 327.

Broda, S. A. and Paolella, M. S. (2011). Expected shortfall for distributions in finance. In *Statistical tools for finance and insurance*, pages 57–99. Springer.

Cai, J. (1994). A markov model of switching-regime arch. *Journal of Business & Economic Statistics*, 12(3):309–316.

Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H. K. (2015). Dynamic predictive density combinations for large data sets in economics and finance. *Tinbergen Institute Discussion Paper 15-084/III*.

[21] For identification we impose that $\kappa_1 + \kappa_2 = 1$.

Catania, L., Boudt, K., and Ardia, D. (2017). ***GAS**: Generalised Autoregressive Score Models*. R package version 0.2.4.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 8(2):93–115.

Creal, D., Koopman, S. J., and Lucas, A. (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, 29(4):552–563.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

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

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

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and osband's principle. *The Annals of Statistics*, 44(4):1680–1707.

Frigessi, A., Haug, O., and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3):219–235.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer.

Galeano, P. and Ausín, M. C. (2010). The gaussian mixture dynamic conditional correlation model: Parameter estimation, value at risk calculation, and portfolio selection. *Journal of Business & Economic Statistics*, 28(4):559–571.

Gerlach, R., Carter, C., and Kohn, R. (2000). Efficient bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, 95(451):819–828.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1):27 – 62.

Haas, M., Krause, J., Paolella, M. S., and Steude, S. C. (2013). Time-varying mixture GARCH models and asymmetric volatility. *The North American Journal of Economics and Finance*, 26:602 – 623.

Haas, M., Mittnik, S., and Paolella, M. S. (2004a). Mixed normal conditional heteroskedasticity. *Journal of Financial Econometrics*, 2(2):211–250.

Haas, M., Mittnik, S., and Paolella, M. S. (2004b). A new approach to markov-switching garch models. *Journal of Financial Econometrics*, 2(4):493–530.

Haas, M. and Paolella, M. S. (2012). Mixture and regime-switching garch models. *Handbook of volatility models and their applications*, pages 71–102.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 77(2):357–384.

Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of econometrics*, 64(1):307–333.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35(3):705–730.

Harrison, J. and West, M. (1999). *Bayesian Forecasting & Dynamic Models*. Springer.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. Cambridge University Press.

Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, 6(3):255–259.

Jorion, P. (2006). *Value at Risk – The New Benchmark for Managing Financial Risk*. McGraw–Hill, third edition.

KaewTraKulPong, P. and Bowden, R. (2002). An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems*, pages 135–144. Springer.

Kim, C.-J. (1994). Dynamic linear models with markov-switching. *Journal of Econometrics*, 60(1):1 – 22.

Kim, C.-J. and Nelson, C. R. (1999). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*, volume 2. MIT press Cambridge.

Klaassen, F. (2002). Improving garch volatility forecasts with regime-switching garch. In *Advances in Markov-Switching Models*, pages 223–254. Springer.

Koenker, R. (2005). *Quantile regression*. Number 38. Cambridge university press.

Li, G., Zhu, Q., Liu, Z., and Li, W. K. (2017). On mixture double autoregressive time series models. *Journal of Business*

& Economic Statistics, 35(2):306–317.

Li, M., Li, W. K., and Li, G. (2013). On mixture memory garch models. *Journal of Time Series Analysis*, 34(6):606–624.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools: Concepts, Techniques and Tools*. Princeton university press.

Morgan, J. et al. (1996). Riskmetrics technical document.

Mullen, K., Ardia, D., Gil, D., Windover, D., and Cline, J. (2011). DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software, Articles*, 40(6):1–26.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246 – 256.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18.

Patton, A. J., Ziegel, J. F., and Chen, R. (2017). Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk). *ArXiv e-prints*.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3):289–296.

Storn, R. and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359.

Tse, Y. K. and Tsui, A. K. C. (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business & Economic Statistics*, 20(3):351–362.

Vlaar, P. J. and Palm, F. C. (1993). The message in weekly exchange rates in the european monetary system: mean reversion, conditional heteroscedasticity, and jumps. *Journal of Business & Economic Statistics*, 11(3):351–360.

West, K. D. and Newey, W. K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Wong, C., Chan, W., and Kam, P. (2009). A student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika*, 96(3):751–760.

Wong, C. S. and Li, W. K. (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association*, 96(455):982–995.

Xie, L., Kennedy, L., Chang, S.-F., Divakaran, A., Sun, H., and Lin, C.-Y. (2005). Layered dynamic mixture model for pattern discovery in asynchronous multi-modal streams [video applications]. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 2, pages ii–1053. IEEE.

Yu, J. (2012). A particle filter driven dynamic gaussian mixture model approach for complex process monitoring and fault diagnosis. *Journal of Process Control*, 22(4):778–788.

Zhang, Z., Li, W. K., and Yuen, K. C. (2006). On a mixture garch time-series model. *Journal of Time Series Analysis*, 27(4):577–597.

Zhu, D. and Galbraith, J. W. (2010). A generalized asymmetric student-t distribution with application to financial econometrics. *Journal of Econometrics*, 157(2):297–305.