The European Journal of Finance

ISSN: 1351-847X (Print) 1466-4364 (Online) Journal homepage: http://www.tandfonline.com/loi/rejf20

Multivariate asset return prediction with mixture models

Marc S. Paolella

To cite this article: Marc S. Paolella (2015) Multivariate asset return prediction with mixture models, The European Journal of Finance, 21:13-14, 1214-1252, DOI: 10.1080/1351847X.2012.760167

To link to this article: http://dx.doi.org/10.1080/1351847X.2012.760167

Published online: 03 May 2013.

The European Journal of Finance, 2015
Vol. 21, Nos. 13–14, 1214–1252, http://dx.doi.org/10.1080/1351847X.2012.760167

Multivariate asset return prediction with mixture models

Marc S. Paolella[a,b]*

[a]*Swiss Banking Institute, University of Zurich, Rämistrasse 71, 8006 Zurich, Switzerland;* [b]*Swiss Finance Institute, Walchestrasse 9, CH-8006 Zurich, Switzerland*

(*Received 6 November 2012; final version received 8 December 2012*)

The use of mixture distributions for modeling asset returns has a long history in finance. New methods of demonstrating support for the presence of mixtures in the multivariate case are provided. The use of a two-component multivariate normal mixture distribution, coupled with shrinkage via a quasi-Bayesian prior, is motivated, and shown to be numerically simple and reliable to estimate, unlike the majority of multivariate GARCH models in existence. Equally important, it provides a clear improvement over use of GARCH models feasible for use with a large number of assets, such as constant conditional correlation, dynamic conditional correlation, and their extensions, with respect to out-of-sample density forecasting. A generalization to a mixture of multivariate Laplace distributions is motivated via univariate and multivariate analysis of the data, and an expectation–maximization algorithm is developed for its estimation in conjunction with a quasi-Bayesian prior. It is shown to deliver significantly better forecasts than the mixed normal, with fast and numerically reliable estimation. Crucially, the distribution theory required for portfolio theory and risk assessment is developed.

Keywords: density forecasting; expectation–maximization algorithm; fat tails; mixture distributions; multivariate Laplace distribution; quasi-Bayesian estimation; shrinkage estimation; weighted likelihood

JEL Classification: C51; C53; G11; G17

1. Introduction

The race is on to develop a numerically practical model for accurately predicting the multivariate distribution of a set of asset returns, and the number of contestants is constantly increasing. While the univariate case is well-served by GARCH-type models (Kuester, Mittnik, and Paolella 2006 and the references therein), the general multivariate case has proved to be demonstrably more difficult, with numerous multivariate-GARCH (MGARCH) models being proposed, although no single model has established itself as being overall satisfactory with respect to predictive power, ease of estimation, and tractability of weighted sums of the univariate marginal distributions (for portfolio optimization). Currently, many popular models are infeasible for even a modest number of assets. Bollerslev, Engle, and Wooldridge (1988) introduced an MGARCH specification for modeling conditional covariances. This is a very flexible model, though the number of parameters for d assets is $d(d+1)(d(d+1)+1)/2$ (e.g. for $d=3$ it is 78), so, as mentioned in Bauwens, Laurent, and Rombouts (2006), in practice this model can only be used in the bivariate case. To overcome this restriction, a diagonal version was proposed by Bollerslev, Engle, and Wooldridge (1988), and for which Ledoit, Santa-Clara, and Wolf (2003) developed an algorithm which increases estimation speed. But, as reported in Sheppard (2012), even that algorithm can be

*Email: marc.paolella@bf.uzh.ch

slow if d is large. Another specification is the BEKK-GARCH model of Engle and Kroner (1995). It has the benefit of guaranteeing positive definitiveness of the covariance matrix without imposing strong restrictions on the parameters, but otherwise it exhibits the same dimensionality problems as the Bollerslev, Engle, and Wooldridge (1988) model and cannot be applied to more than a few assets (see, e.g. the survey articles of Bauwens, Laurent, and Rombouts 2006; Silvennoinen and Teräsvirta 2009).

The goal of this paper is to advocate a model for large-scale multivariate asset returns which (i) is very easily estimated, (ii) adequately deals with non-normality, (iii) lends itself to tractable distributions of weighted sums of the marginal random variables, and (iv) outperforms competing models in terms of density forecasting, such as the constant conditional correlation (CCC) and dynamic conditional correlation (DCC) models, and their extensions. This is done by using a two-component mixture of multivariate normals (MVNs), and, new to this paper, a two-component mixture of multivariate Laplace distributions. In both cases, In both cases, owing to the large number of parameters to estimate, parameter shrinkage will be seen to play a fundamental role in the success of the model. The models do not use any GARCH structure, and so are exceptionally fast to estimate, even, and particulary with, hundreds of assets. By avoiding use of multivariate GARCH-type conditional structures to allow time-varying volatilities and correlations, numerous difficult theoretical and numerical problems are avoided; see McAleer (2005) and Caporin and McAleer (2010) for a clear discussion of such issues.

Our stated goal for fast parameter estimation is accomplished by circumventing the primary problem inherent in many conventional MGARCH models that the number of parameters to be simultaneously estimated becomes prohibitive even for a moderate number, say d, of assets. One way around this is to conduct a two-step estimation procedure whereby on the first step, d separate estimations, each with a small number of parameters, are conducted, followed by estimation of the remaining model parameters. This is done for the CCC and DCC models, as well as copula-based approaches. The other approach, which we take in this paper, is to simultaneously estimate the thousands of parameters of the model, but use a structure which is amenable to use of an expectation–maximization (EM) algorithm such that the maximization step involves a trivially computed closed-form expression. As an example for one of the models considered below, the two-component MVN mixture with d assets has $d^2 + 3d + 1$ parameters to estimate; with $d = 30$, this is 991. In a general optimization setting with so many parameters, this would be impossible, but in this context, the EM algorithm is applicable, and takes about one-tenth of a second to estimate.

We will use as our primary demonstration data set the $T = 1945$ daily returns on each of the $d = 30$ stocks composing the Dow Jones Industrial Index from 13 June 2001 to 11 March 2009 (hereafter DJ-30). The returns for each asset are computed in the usual way; as continuously compounded percentage returns, given by $r_{i,t} = 100 \log(p_{i,t}/p_{i,t-1})$, where $p_{i,t}$ is the price of a share of asset i at time t.

The rest of this paper is outlined as follows. Section 2 discusses the motivation for using mixture distributions and reviews the relevant literature. Section 3 discusses shrinkage estimation of the multivariate mixed normal distribution and confirms the presence of the features one would expect from data coming from a mixture. Section 4 assesses the model fit by splitting the data into two components, and argues why a multivariate Laplace distribution is a viable candidate. Section 5 investigates the optimal amount of shrinkage as a function of the number of assets under study. Section 6 illustrates the use of weighted likelihood and moving averages of the mixing parameter to provide further improvements in forecast quality. The mixture multivariate Laplace distribution is introduced and developed in Section 7, and its forecasting performance is shown to outperform

the use of the normal mixture. Section 8 quickly entertains use of a different data set, confirming the results found for the primary data set under study. After brief conclusions in Section 9, an appendix collects several technical results of the paper.

2. Mixtures in finance

The motivation for the proposed model comes from several sources of evidence presented below that multivariate returns data are reasonably well-described by a two-component mixture. The use of a univariate normal mixture seems to have been first considered by Fama (1965) in the context of modeling financial asset returns, and then subsequently by many authors; see, e.g. the discussion and references in Haas, Mittnik, and Paolella (2004a). Kim and White (2004, 72) provide further evidence of the appropriateness of mixtures for financial data, stating '. . . it may be more productive to think of the S&P500 index returns studied here as being better described as a mixture containing a predominant component that is nearly symmetric with mild kurtosis and a relatively rare component that generates highly extreme behaviour'. This can be contrasted with Neftci (2000), who argues that the extreme movements in asset prices are caused by mechanisms which are 'structurally different' from the 'routine functioning of markets'.

Based on such findings, it is not surprising that the coupling of GARCH structures with the components of a univariate normal mixture distribution would be entertained and found fruitful. This was indeed the case, and a general framework for this, nesting several particular ideas in the literature and characterizing the model's stationarity properties, was put forth independently and concurrently by Haas, Mittnik, and Paolella (2004a) and Alexander and Lazar (2006), and further augmented in Haas, Mittnik, and Paolella (2004b) and Alexander and Lazar (2005). These authors show that the model is very well-suited for capturing the complex dynamics and long memory of univariate asset return volatility. Other authors have subsequently shown that it delivers highly accurate out-of-sample risk predictions, and is instrumental in utility-based asset pricing and option pricing; see Kuester, Mittnik, and Paolella (2006), Wu and Lee (2007), Bauwens and Rombouts (2007), Giannikis, Vrontos, and Dellaportas (2008), Rombouts and Stentoft (2009), and the references therein.

The aforementioned univariate model was extended to the symmetric and asymmetric multivariate cases by Bauwens, Hafner, and Rombouts (2007) and Haas, Mittnik, and Paolella (2009), respectively. These authors show that these models clearly identify the existence of two components with distinctly different volatility dynamics, and that the low (high) volatility component is associated with positive (negative) means, implying that the low and high volatility components can be interpreted as bull and bear markets, respectively. In the asymmetric multivariate mixture model, a leverage effect (see below) is shown to be present in the high-volatility component. While these models are extremely rich in their ability to model multivariate asset returns, their obvious drawback is, similar to many multivariate GARCH models, the curse of dimensionality: In both of these multivariate extensions, only bivariate series were used in applications. The goal of this paper is to develop a model which is suitable for use with d assets, where d could be in the tens or hundreds, necessitating that its parameters can be quickly and reliably estimated, and which delivers accurate multivariate density forecasts on par with, or, as we find, better than, existing MGARCH models. This is achieved by accounting for the appearance of two markedly different components in the data, and the use of shrinkage and weighted likelihood, instead of a GARCH structure. An extensive forecasting exercise demonstrates that our model outperforms Bollerslev's (1990) CCC and Engle's (2002, 2009) DCC, as well as two recenteneralizations of these frameworks discussed below.

While, like with the univariate case, there are many multivariate distributions which nest (or yield as a limiting case) the normal, and otherwise allow for thicker tails and asymmetry, a mixture distribution has been empirically found by numerous authors in both the unconditional and conditional (GARCH) framework to be of particular relevance for this kind of financial data, as discussed above. Unlike other types of fat-tailed multivariate distributions, a mixture can capture (along with the two major stylized facts of excess kurtosis and asymmetry) the two further stylized facts of

- the *leverage* or *down market* effect – the negative correlation between volatility and asset returns, the economic reason for which is usually attributed to Black (1976); and
- *contagion effects* – the tendency of the dependency between asset returns to increase during pronounced market downturns, as well as during periods of higher volatility.

For example, with a two-component MVN mixture model with component parameters (in obvious notation) $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_2$, we would expect to have the higher-weighted, or primary, component, say the first, capturing the more typical, 'business as usual' stock return behavior, with a near-zero mean vector $\boldsymbol{\mu}_1$, and the second component capturing the more volatile, 'crisis' behavior, with

- (much) higher variances in $\boldsymbol{\Sigma}_2$ than in $\boldsymbol{\Sigma}_1$,
- larger correlations in $\boldsymbol{\Sigma}_2$ than in $\boldsymbol{\Sigma}_1$, reflecting the contagion effect,
- and a predominantly negative $\boldsymbol{\mu}_2$, reflecting the down market effect.

A distribution with only a single mean vector and covariance (or, more generally, dispersion) matrix cannot capture this behavior, no matter how many additional shape parameters the distribution possesses, even if each marginal were to be endowed with its own set of shape (tail and skew) parameters, as is possible when using a copula. We will see below that these aforementioned features are precisely apparent in the DJ-30 data. The ability to capture the contagion effect was also mentioned by Buckley, Saunders, and Seco (2008) as a motivation for using the MVN mixture distribution, along with its ability to adequately address the leptokurtosis and asymmetry in asset returns. Those authors derive conditions for efficiency of optimal portfolios under this distributional assumption, as well as investigating other issues of relevance for portfolio management problems.

3. Multivariate mixtures of normals and shrinkage

Let $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,d})' \overset{\text{iid}}{\sim} \text{Mix}_k N_d(\boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$, $t = 1, \ldots, T$, where $\text{Mix}_k N_d$ denotes the k-component, non-singular d-dimensional multivariate mixed normal distribution, with

$$\boldsymbol{M} = [\boldsymbol{\mu}_1 \mid \boldsymbol{\mu}_2 \mid \cdots \mid \boldsymbol{\mu}_k], \quad \boldsymbol{\mu}_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{dj})', \quad \boldsymbol{\Psi} = [\boldsymbol{\Sigma}_1 \mid \boldsymbol{\Sigma}_2 \mid \cdots \mid \boldsymbol{\Sigma}_k],$$

$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$, $\boldsymbol{\Sigma}_j > 0$ (i.e. positive definite), $j = 1, \ldots, k$, and

$$f_{\text{Mix}_k N_d}(\mathbf{y}; \boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\lambda}) = \sum_{j=1}^{k} \lambda_j f_N(\mathbf{y}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad \lambda_j \in (0,1), \quad \sum_{j=1}^{k} \lambda_j = 1, \tag{1}$$

with f_N denoting the d-variate normal distribution. The class of $\text{Mix}_k N_d$ distributions is identified (Yakowitz and Spragins 1968).

3.1 *Estimation with shrinkage*

Denote the latent, or hidden, variable associated with the tth observation $\mathbf{Y}_t$ as $\mathbf{H}_t = (H_{t,1}, \ldots, H_{t,k})$, $t = 1, \ldots, T$, where $H_{t,j} = 1$ if $\mathbf{Y}_t$ came from the jth component, and zero otherwise, $j = 1, \ldots, k$.

We state the parameter updating equations not in the usual way, but rather augmented by the quasi-Bayesian prior of Hamilton (1991), which is explained below, and of which we will make judicious use. They are

$$\hat{\lambda}_j = \frac{1}{T}\sum_{t=1}^{T} H_{t,j}, \quad j = 1, \ldots, k, \tag{2}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{c_j \mathbf{m}_j + \sum_{t=1}^{T} H_{t,j}\mathbf{Y}_t}{c_j + \sum_{t=1}^{T} H_{t,j}}, \tag{3}$$

and

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\mathbf{B}_j + \sum_{t=1}^{T} H_{t,j}(\mathbf{Y}_t - \hat{\boldsymbol{\mu}}_j)(\mathbf{Y}_t - \hat{\boldsymbol{\mu}}_j)' + c_j(\mathbf{m}_j - \hat{\boldsymbol{\mu}}_j)(\mathbf{m}_j - \hat{\boldsymbol{\mu}}_j)'}{a_j + \sum_{t=1}^{T} H_{t,j}}, \tag{4}$$

$j = 1, \ldots, k$. Fixed quantities $\mathbf{m}_j \in \mathbb{R}^d$, $a_j \geq 0$, $\mathbf{B}_j$ a $d \times d$ positive definite matrix, and $c_j \geq 0$, indicate the prior information, as introduced and discussed in Hamilton (1991). Briefly, a_j and c_j represent the weights of the jth prior, with $\mathbf{m}_j$ its mean, and, if $a_j > 0$, $\mathbf{B}_j/a_j$ its variance. Values a_j and c_j need not be the same, with a relatively smaller c_j corresponding to less information on $\boldsymbol{\mu}_j$ than on $\boldsymbol{\Sigma}_j$. One natural prior would be to take $\mathbf{m}_j$ to be a d-vector of zeros, and $\mathbf{B}_j$ the d-dimensional identity matrix, corresponding to shrinkage to the standard normal. Our choice will be similar to this, but altered in such a way to be more meaningful in the context of modeling daily equity returns in general, and specifically for the DJ-30 during the period under study.

The top panels in Figure 1 show the $d = 30$ values contained in each of the vectors $\hat{\boldsymbol{\mu}}_1$ (left panel) and $\hat{\boldsymbol{\mu}}_2$ (right panel), obtained from fitting the Mix_2N_{30} model to the DJ-30 data set via maximum likelihood, without incorporating any prior information. The resulting maximum likelihood estimator (MLE) values are in accordance with the discussion in the Introduction of the two regimes at work in the financial market. While the means in $\hat{\boldsymbol{\mu}}_1$ are closely centered around zero, those from $\hat{\boldsymbol{\mu}}_2$ are nearly all negative, and with a much higher magnitude than those from $\hat{\boldsymbol{\mu}}_1$. From the middle row of panels, we see that the variances from $\hat{\boldsymbol{\Sigma}}_2$ of the 30 components are about 10 times the size of those from $\hat{\boldsymbol{\Sigma}}_1$. Thus, the second component is picking up the relatively higher volatility returns and, notably, is associated with a highly negative mean term, thus capturing the leverage effect. Finally, we see from the bottom panels that the correlations between the 30 assets (all of which are positive) are also higher in the second component, indicating the presence of contagion effects. As already mentioned, while being fat-tailed and asymmetric, a non-mixture distribution, such as the multivariate generalized hyperbolic or a copula construction, has only one location vector and dispersion matrix, and so cannot capture these two separate types of market behavior.

Based on these findings, and in line with the usual motivation for the James–Stein estimator for the mean vector of a MVN population with independent components, our prior is one which shrinks the means, variances, and covariances from each of the two components toward their average values over the 30 series, as shown in Figure 1. Let $\mathbf{m}_i$ and $\mathbf{B}_i$ be the shrinkage target for the mean vector and variance–covariance matrix, respectively, for the ith component. The

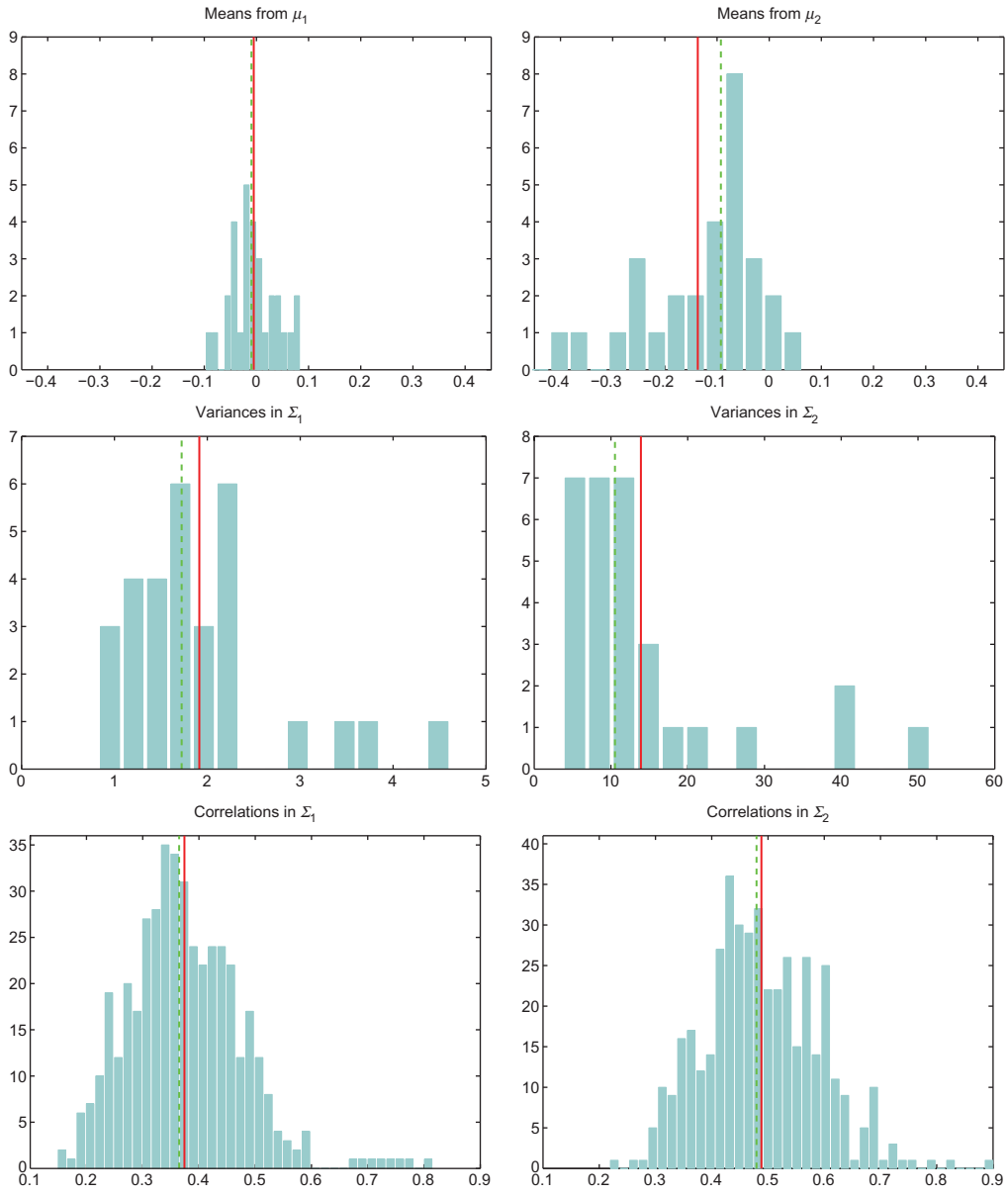

Figure 1. The estimated 30 means (top), 30 variances (middle) and 435 correlations (bottom) for the first and second components of the normal mixture. Solid and dashed vertical lines show the mean and median.

shrinkage target, as a function of a scalar hyper-parameter ω, is

$$\begin{aligned} &\mathbf{m}_1 = \mathbf{0}_d, \quad \mathbf{m}_2 = (-0.1)\mathbf{1}_d, \quad a_1 = 2\omega, \quad a_2 = \omega/2, \quad c_1 = c_2 = 20\omega, \\ &\mathbf{B}_1 = a_1[(1.5 - 0.6)\mathbf{I}_d + 0.6\mathbf{J}_d], \quad \mathbf{B}_2 = a_2[(10 - 4.6)\mathbf{I}_d + 4.6\mathbf{J}_d], \end{aligned} \tag{5}$$

where $\mathbf{J}_d$ is the $d \times d$ matrix of ones.

Prior weight a_j reflects our strength in the prior of the covariance matrix $\mathbf{\Sigma}_j$. In our $k = 2$ setting, we found that taking $a_1 = 2\omega$ and $a_2 = \omega/2$ was beneficial, compared to having them equal, which might have been anticipated as $\mathbf{\Sigma}_2$ embodies more uncertainty and magnitude in the variances than does $\mathbf{\Sigma}_1$. Prior weight c_j reflects our strength in the prior of mean vector $\boldsymbol{\mu}_j$, $j = 1, 2$. These should be higher than the a_j for two reasons. First, an appeal to the efficient market hypothesis provides some justification for shrinking the means in the first, primary component of the mixture toward zero, while the blatant down-market effect in financial crises lends support for shrinking the mean in the second component of the mixture toward a negative value. The second reason is that the errors in estimation of the mean vector are considered far more consequential in asset allocation and portfolio management (Chopra and Ziemba 1993), so that the benefits of shrinkage could be quite substantial. In light of this, we take $c_j = 20\omega, j = 1, 2$. (The large factor of 20 was determined simply by some trial and error based on the simulation exercise discussed next, and could, if desired, be taken as a further hyperparameter.)

The only tuning parameter which remains to be chosen is ω. This will be done below in Section 5, in which it is chosen to maximize forecast quality.

3.2 *Portfolio construction*

Let $\mathbf{Y} \sim \text{Mix}_k N_d(\boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$, with $\boldsymbol{M} = [\boldsymbol{\mu}_1 \mid \boldsymbol{\mu}_2 \mid \cdots \mid \boldsymbol{\mu}_k]$, $\boldsymbol{\Psi} = [\mathbf{\Sigma}_1 \mid \mathbf{\Sigma}_2 \mid \cdots \mid \mathbf{\Sigma}_k]$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$. Interest centers on the distribution of the portfolio $P = \mathbf{a}'\mathbf{Y}$, $\mathbf{a} \in \mathbb{R}^d$. We show in the appendix that

$$f_P(x) = \sum_{c=1}^{k} \lambda_c \phi(x; \mathbf{a}'\boldsymbol{\mu}_c, \mathbf{a}'\mathbf{\Sigma}_c\mathbf{a}), \tag{6}$$

where $\phi(x; \mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2 evaluated at x. With such a simple analytic result, portfolio optimization is straightforward, and simulation (as is required when copula-based methods are used) is not necessary. With $\mu_c = \mathbf{a}'\boldsymbol{\mu}_c$ and $\sigma_c^2 = \mathbf{a}'\mathbf{\Sigma}_c\mathbf{a}$, $c = 1, \ldots, k$, it is elementary to show that

$$\mu_P = \mathbb{E}[P] = \sum_{c=1}^{k} \lambda_c \mu_c, \quad \sigma_P^2 = \mathbb{V}(P) = \sum_{c=1}^{k} \lambda_c(\sigma_c^2 + \mu_c^2) - \mu_P^2. \tag{7}$$

It is common when working with non-Gaussian portfolio distributions to use a measure of downside risk, such as the value at risk, or the expected shortfall (ES). The former involves the γ-quantile of P, denoted $q_{P,\gamma}$, for $0 < \gamma < 1$, with γ typically 0.01 or 0.05, and which can be found numerically by using the cumulative distribution function (cdf) of P, easily seen to be $F_P(x) = \sum_{c=1}^{k} \lambda_c \Phi((x - \mu_c)/\sigma_c)$, with Φ the standard normal cdf. The γ-level ES of P is given by Acerbi and Tasche (2002) as

$$\text{ES}_\gamma(P) = \frac{1}{\gamma} \int_0^\gamma Q_P(p)\, \mathrm{d}p, \tag{8}$$

where Q_P is the quantile function of P. We prove in the appendix that this integral can be represented analytically as

$$\text{ES}_\gamma(P) = \sum_{j=1}^{k} \frac{\lambda_j \Phi(c_j)}{\gamma} \left\{ \mu_j - \sigma_j \frac{\phi(c_j)}{\Phi(c_j)} \right\}, \quad c_j = \frac{q_{P,\gamma} - \mu_j}{\sigma_j}, \quad q_{P,\gamma} = Q_P(\gamma). \tag{9}$$

4. Model fit

In this section, we examine several issues associated with assessing the fit of the model. Section 4.1 makes the brief but important point that estimated component separation is a necessary, but not sufficient, condition for data arising from a mixture distribution. Section 4.2 uses the fact that we can reasonably accurately separate the two components, so that single-component, univariate analysis can be done to assess model fit. This is then done in two ways.

4.1 *Use of filtered H_t values*

We wish to assess the appropriateness of using a mixture of two normal distributions for modeling the DJ-30 data, for which $T = 1945$. It is natural to plot the imputed values of $H_{t,1}$ computed from the EM algorithm, versus the time ordering t. These are shown in the left panel of Figure 2, as returned from the EM algorithm after it converged. The right panel is the same, but just showing the last 250 values. It is apparent that the two components are reasonably well-separated, with most values being either very close to zero, or to one.

While one might be tempted to use this finding as further evidence for the claim that there exist two, reasonably distinct, 'regimes' in financial markets, this is actually not the case, because the same effect would occur if the data had arisen from a fat-tailed single-component multivariate distribution, as simulations easily show. Thus, the separation apparent in Figure 2 is necessary, but not sufficient, to declare that the data were generated by a mixture distribution.

4.2 *Component separation and univariate analysis*

Returning now to the DJ-30 data, the separation apparent from Figure 2 is also highly advantageous, because it allows us to assign each $\mathbf{Y}_t$ to one of the two components, in most cases with very high confidence. Once done, we can assess how well each of the two estimated MVN distributions fits the observations assigned to its component. While we could use the rule to assign the tth observation $\mathbf{Y}_t$ to component 1 if $H_{t,1} > 0.5$, and to component 2 otherwise (which would result in 1490 observations assigned to component 1, or 76.6%, which is nearly the same as $\hat{\lambda}_1 = 0.763$), we instead use the criteria $H_{t,1} > 0.99$, choosing to place those $\mathbf{Y}_t$ whose corresponding values

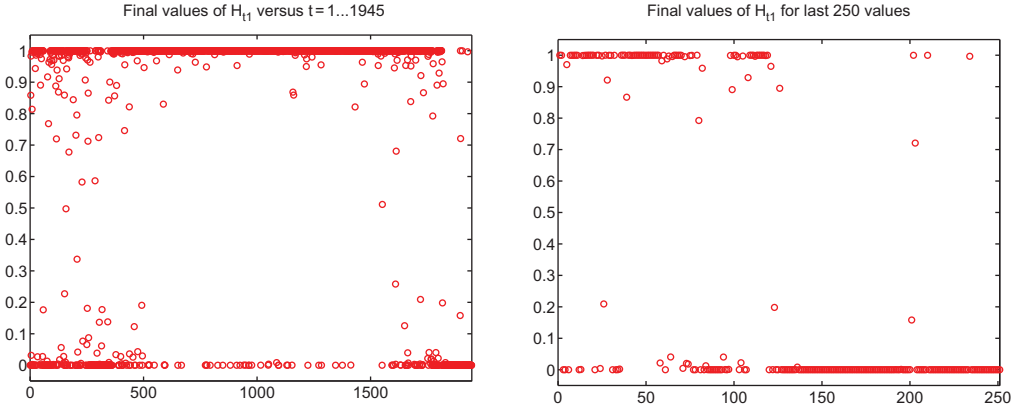

Figure 2. Final values of $H_{t,1}$ returned from the EM algorithm based on the two-component Mix$N$ for the DJ-30 return series.

of $H_{t,1}$ suggest even a slight influence from component 2, into this more volatile component. This results in 1373 observations assigned to component 1, or 70.6% of the observations, and 572 to the second component.

4.2.1 Heuristic testing I: use of the GAt distribution

There are many tests for composite univariate normality. These could all be applied to each of the $k \cdot d = 60$ marginal distributions, after having split the data into the k components in some reasonable way such as done above. While perhaps of some value, this is not of particular interest for two reasons. First, we essentially *know* that the data, after splitting into two components (not to mention the unsplit data), are not precisely normally distributed, so that if we somehow had a very powerful test for normality, then we know it would reject the null. Second, a more substantial concern is how close the data in, say, the first component are to being *jointly* normally distributed. For the purposes of constructing a financial portfolio of the assets, it might be of far greater importance to know that, while the data set deviates slightly from normality because it contains a small handful of (with respect to the normal distribution) outliers, the overall structure is close to MVN (marginals, conditionals and ellipticity).

Unfortunately, testing composite multivariate normality is not trivial; see the detailed discussion in Thode (2002, Chap. 9) and the survey article from Mecklin and Mundfrom (2004). Part of the reason for the complexity of testing multivariate normality is that there are many ways for the true distribution to depart from it, so that no single test will be optimal. So, based on this, instead of pursuing multivariate normality tests, we content ourselves with examining how the d marginals in each of the two components deviate from normality, so that we can suggest a more suitable multivariate distribution which at least accounts for the empirical behavior of the d marginals and, ideally, is still algebraically tractable and easy to estimate.

Based on the split into the $k = 2$ components, we will estimate, for each of the d univariate series in each of those two components, a flexible, asymmetric, fat-tailed distribution (which includes the normal and Laplace as limiting cases), and inspect the parameters to learn about the marginal distributions. For this, we use the generalized asymmetric t, or GAt, distribution, with location-zero, scale-one probability density function (pdf) given by

$$f_{\mathrm{GAt}}(z; p, \nu, \theta) = C_{p,\nu,\theta} \times \begin{cases} \left(1 + \dfrac{(-z \cdot \theta)^p}{\nu}\right)^{-(\nu+1/p)}, & \text{if } z < 0, \\ \left(1 + \dfrac{(z/\theta)^p}{\nu}\right)^{-(\nu+1/p)}, & \text{if } z \geq 0, \end{cases} \tag{10}$$

$p, \nu, \theta \in \mathbb{R}_{>0}$, where $C^{-1}_{p,\nu,\theta} = (\theta^{-1} + \theta) p^{-1} \nu^{1/p} B(1/p, \nu)$ and $B(\cdot, \cdot)$ is the beta function. This distribution has five parameters; three shape, and a location and scale. Of the three shape parameters, p measures the peakedness of the density, with values near one indicative of the Laplace-type behavior, while values near two indicate a peak similar to that of the Gaussian and Student's t distributions. Parameter ν indicates the tail thickness, and is analogous to the degrees of freedom parameter in the Student's t, except that moments of order $\nu \cdot p$ and higher do not exist, so that, if $p = 2$, then we would double the value of ν to make it comparable to the degrees of freedom parameter in the usual Student's t distribution. As $\nu \to \infty$, the GAt approaches the asymmetric generalized exponential distribution which, in the symmetric case, is normal for $p = 2$ and Laplace for $p = 1$. The third parameter, θ, controls the asymmetry, with values less than 1 indicating negative skewness. The necessary constraints during estimation are that $\hat{p} > 0$, $\hat{\nu} > 0$, $\hat{\theta} > 0$, and scale $\hat{c} > 0$.

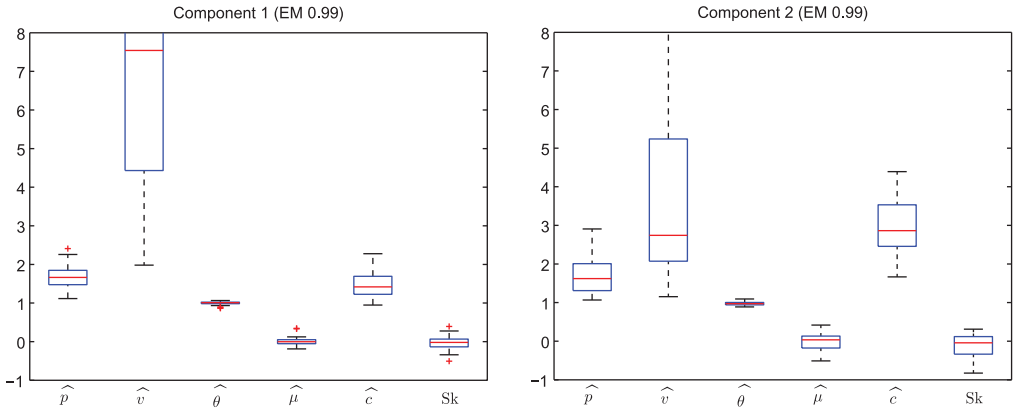

Figure 3. Truncated boxplots of the fitted GAt parameters of the 30 return series in components 1 and 2.

The results are shown as $k = 2$ boxplots over the $d = 30$ time series in Figure 3. Within each boxplot, the estimates for parameters p, v, θ, location μ and scale c are shown, along with the sample skewness, shown in the last boxplot. For component 1, the sample skewness is virtually centered around zero and has a much lower variation than those for component 2, indicating that we can assume symmetry in the marginals for component 1. In both components 1 and 2, the estimated value of the asymmetry parameter θ barely deviates from unity, lending support that the asymmetry exhibited in the asset returns is well-explained by using two symmetric components in a mixture distribution. This is interesting in and of itself, because a substantial literature draws attention to the asymmetry in asset returns (see, e.g. Harvey and Siddique 2000; Patton 2004; Kuester, Mittnik, and Paolella 2006, and the references therein), but as we see, once the multi-component nature of the data is recognized, the components themselves appear symmetric. The scale terms in component 1 are, as expected, much lower than those in component 2.

While the $\hat{v}_i$ (the tail thickness parameters, with $v_i = \infty$ corresponding to exponential tails as with the normal and GED distributions) for component 1 are, on average, quite high, and far higher than the $\hat{v}_i$ for component 2, some of them are still rather small, the smallest, corresponding to the stock returns of McDonald's corporation, being 1.98. (Recall that this does not suggest that moments of two and higher do not exist; in this case, $\hat{p} = 2.41$, so that $\hat{v} \cdot \hat{p} = 4.8$. This is also the stock with the lowest product $\hat{v} \cdot \hat{p}$.) This fact adds considerable weight against the multivariate normality hypothesis for each of the two components, though there are very few stocks, such as McDonald's, which have such aberrant behavior, and ending the story there would be premature.

4.2.2 Heuristic testing II: outlier removal and LRT

To investigate this further, consider conducting the following heuristic testing procedure, which appears to be new and of potentially more general use. For each of the 30 series, but *not* separating them into the two components, we fit the GAt, first with no parameter restrictions (other than the required ones mentioned above), and second, with the restriction that $90 < \hat{v} < 100$, which essentially forces normality if $p = 2$ and $\theta = 1$, or Laplace if $p = 1$ and $\theta = 1$, though it is important to emphasize that $\hat{p}$ and $\hat{\theta}$ were not constrained in this way. Then, we compute the asymptotically valid p-value of the likelihood ratio test. If that value is less than 0.05, then we remove the largest value (in absolute terms) from the series, and recompute the estimates and the

p-value. This is repeated until the *p*-value exceeds 0.05, and we report the smallest number of observations required to be removed in order to achieve this.

The results are given in Table 1, in the rows labeled 'All'. The other rows are the same, but having used the observations allocated to components 1 and 2 using the splitting procedure discussed above. Thus, for example, stock number 5 (Bank of America) is such that the 65 most extreme values had to be removed from the series to get the *p*-value above 0.05, but *no observations* from

Table 1. Number of observations to be removed such that normality cannot be rejected.

Stock #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
All	19	19	5	2	65	21	11	27	60	10	30	19	28	31	8
Comp1	0	0	0	0	0	3	0	0	1	1	1	0	0	0	0
Comp2	0	1	0	0	3	1	1	6	8	2	0	8	0	1	1
Stock #	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
All	7	12	8	5	19	7	15	28	3	14	10	11	6	8	10
Comp1	0	0	2	2	0	0	7	0	0	0	0	0	0	0	0
Comp2	0	0	0	1	0	0	1	2	0	2	1	3	0	2	1

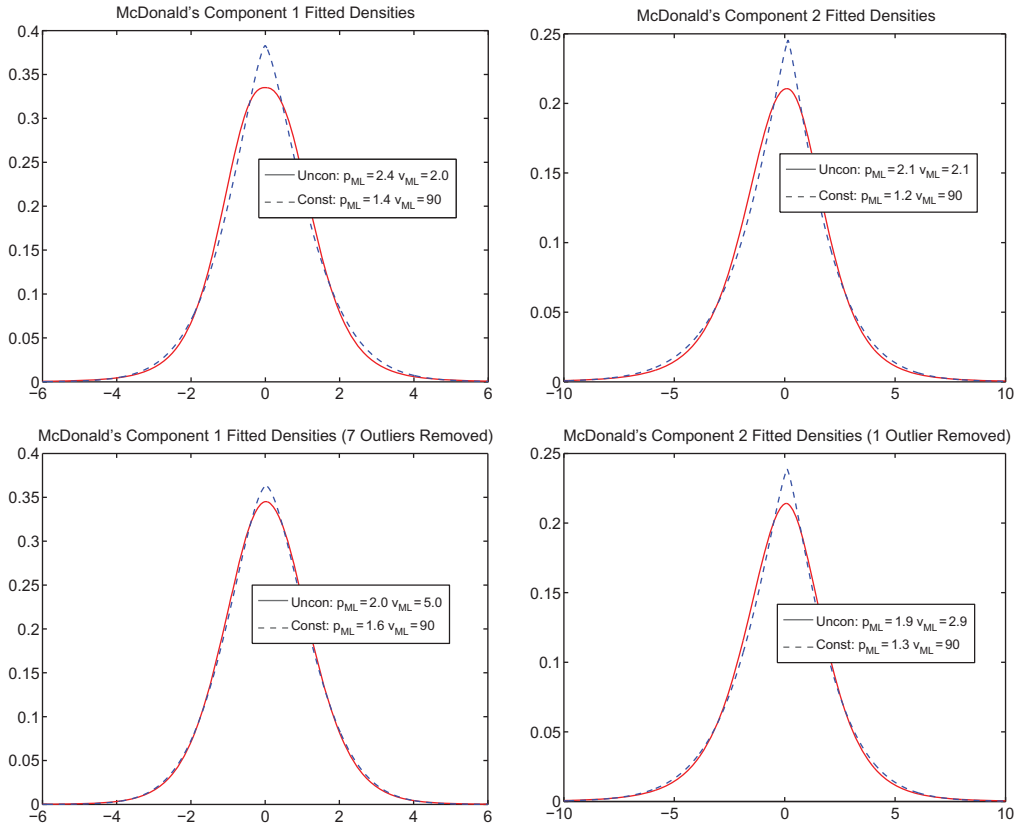

Figure 4. The first (left) and second (right) components of the McDonald's returns, with unrestricted and restricted GAt densities, without and with outlier removal.

component 1 needed to be removed, and only 3 from component 2. Except for stock numbers 6 (Boeing Co) and 22 (McDonald's corporation), either zero or one extreme value, or two in two cases, had to be removed from component one to render a thin-tailed distribution plausible for the data. Thus, we see that a mixture of two normal distributions can account for nearly, but not all, of the fat-tailed behavior in the returns, as well as the asymmetry, and, as already mentioned, has the advantage over other asymmetric, fat-tailed multivariate distributions in that the two components can capture highly distinct behavior, which would otherwise have to be averaged over when using only a single component.

4.2.3 Motivation for Laplace

Now briefly consider the McDonald's results in more detail, this being the most non-Gaussian case. Figure 4 shows their unrestricted and restricted fitted densities (top panels) and the fitted densities after having removed the 7 and 1 most extreme values from components 1 and 2, respectively (bottom). The unrestricted and restricted densities are surprisingly close, with their differences only observable in the far tails of the distribution. Once the extreme values are removed, the unrestricted and restricted densities are nearly indistinguishable. In all cases, but particularly the middle panels, in which the extreme values are not removed, the value of $\hat{p}$ decreases substantially when going from the unrestricted to the restricted model. This is because a lower value of p implies a higher kurtosis, and so it is able to offset the restriction that $\hat{\nu}$ is constrained to lie above 90. Recalling that $p = 1$, $\theta = 1$ and $\nu = \infty$ in Equation (10) corresponds to the Laplace distribution, this finding motivates the study and use of a mixture of two multivariate Laplace distributions instead of MVN. We return to this idea below in Section 7.

5. Density forecasting

After introducing and justifying the measure to be used to judge forecasting quality in Section 5.1, Sections 5.2 and 5.3 discuss the forecasting quality as a function of the amount of imposed shrinkage (for all 30, and less than 30 assets, respectively). Section 5.4 details the model performance in comparison to CCC, DCC, and two of their extensions, while Section 5.5 (i) discusses the merit of abandoning the typical assumption of strict (or covariance) stationarity of a time series model for asset returns, and (ii) why the Laplace mixture is superior to a mixture of Student t.

5.1 *Introduction*

Let $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,d})$ be the vector of returns at time t from the d assets. Interest centers on making a prediction about (functions of) moments or probabilities associated with a future, unobserved, value of $\mathbf{Y}$, which we currently restrict to being one-step. That is, interest centers on $\mathbf{Y}_{t+1}$, based on $\mathbf{Y}_1, \ldots, \mathbf{Y}_t$. As all such (measurable) functions can be ascertained from the distribution of $\mathbf{Y}_{t+1}$ conditional on the available information up to time t, it makes sense to deliver an entire density forecast for $\mathbf{Y}_{t+1}$, instead of just the, say, first two moments of $\mathbf{Y}_{t+1}$, or a particular tail probability.

To assess the predictive quality of a particular model, we fit it to a subset of the data, say $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_\nu$, and record a measure of the quality of the forecast of the density of $\mathbf{Y}_{\nu+1}$. Then either a growing window, or a moving window of length ν, is used to predict the density of $\mathbf{Y}_{\nu+2}$, etc., up to $\mathbf{Y}_T$. This can be done for several models, and the one which yields the best overall 'performance' would be chosen. The relevant question here is obviously the criteria for judging the quality of the density forecasts for $\mathbf{Y}_t$, $t = \nu + 1, \ldots, T$. In light of the prominent role of

the likelihood in statistical inference, a natural idea would be to use the log of the value of the forecast density itself, evaluated at $\mathbf{y}_t$, the observed value of $\mathbf{Y}_t$. In particular, denote the relevant information set up to time t as I_t (which could be a growing or moving window of the $\mathbf{Y}_i$, $i \leq t$, as well as possibly including other variables) and the predictive density of $\mathbf{Y}_t$, conditional on I_{t-1}, based on model $\mathcal{M}$, using the distribution $f^{\mathcal{M}}$, as $f^{\mathcal{M}}_{t|I_{t-1}}(\cdot; \hat{\boldsymbol{\Theta}})$, where $\boldsymbol{\Theta}$ denotes the parameter vector of model $\mathcal{M}$.

The measure of interest is what we will call the (*realized*) *predictive log-likelihood*, given by

$$\pi_t(\mathcal{M}, v) = \log f^{\mathcal{M}}_{t|I_{t-1}}(\mathbf{y}_t; \hat{\boldsymbol{\theta}}), \tag{11}$$

where $v \in \mathbb{N}$ denotes the size of the rolling window used to determine I_{t-1} (and the set of observations used for estimation of $\boldsymbol{\theta}$) for each t. Observe that the information set I_{t-1} is a function of v. We can then compute $\pi_t(\mathcal{M}_i, v)$, for each $t = \tau_0 + 1, \ldots, T$, where τ_0 indicates where the forecasting exercise starts (and is usually equal to v), and set of models $\mathcal{M}_1, \ldots, \mathcal{M}_m$, and compare them. To facilitate comparison for different values of τ_0 and d for a given model $\mathcal{M}$ and window size v, we take the *normalized sum of the realized predictive log-likelihood* to be

$$S_{\tau_0,T}(\mathcal{M}, v) = \frac{1}{(T - \tau_0)d} \sum_{t=\tau_0+1}^{T} \pi_t(\mathcal{M}, v), \tag{12}$$

where d is the dimension of the data. It is thus the average realized predictive log-likelihood, averaged over the number of time points used and the dimension of the random variable under study.

Other methods of density forecast evaluation exist. In particular, if the forecasted distribution coincides with the true one, then via the probability integral transform, the predictive cdf evaluated at the corresponding future observed value yields a uniform random variate. One can then test the sequence of such values for iid uniformity; see, e.g. Diebold, Gunther, and Tay (1998), Kuester, Mittnik, and Paolella (2006), and Broda et al. (2013) for application to density forecasts of financial return series using various graphical and statistical methods. The method is a special case of the transformation from Rosenblatt (1952), which maps a d-variate random vector with a continuous distribution to one with a uniform distribution on the d-dimensional hypercube. However, testing uniformity in higher dimensions is less studied: This is the primary reason for using method (12) instead.

5.2 *Determining the optimal shrinkage: all assets*

We can compute, for some set of models indexed by ω, $\pi_t(\mathcal{M}_\omega, v)$ from Equation (11), taking $\{\mathcal{M}_\omega\}$ to be the Mix_2N_{30} model estimated with shrinkage prior (5) for a given value of hyperparameter ω. We do this using a moving window of size $v = 250$, starting at observation $\tau_0 = v = 250$, and updating parameter vector $\hat{\boldsymbol{\theta}}$ at each time increment

It is possible that there exist multiple plausible local maxima of the likelihood function. Ideally, via use of many starting values, a local optimum would be located which is, with high probability, the global one. We instead use just two starting values, as follows. For a given window of observations, the first starting value is simply the final value obtained from the previous window. As these two data sets just differ by two observations, we expect the MLEs from both of them to be close, so that this should be a very reasonable starting value. Nevertheless, it is possible (and occurs with non-negligible frequency) that this leads precisely to an inferior local maximum. As such, our second starting value is the simple, default one.

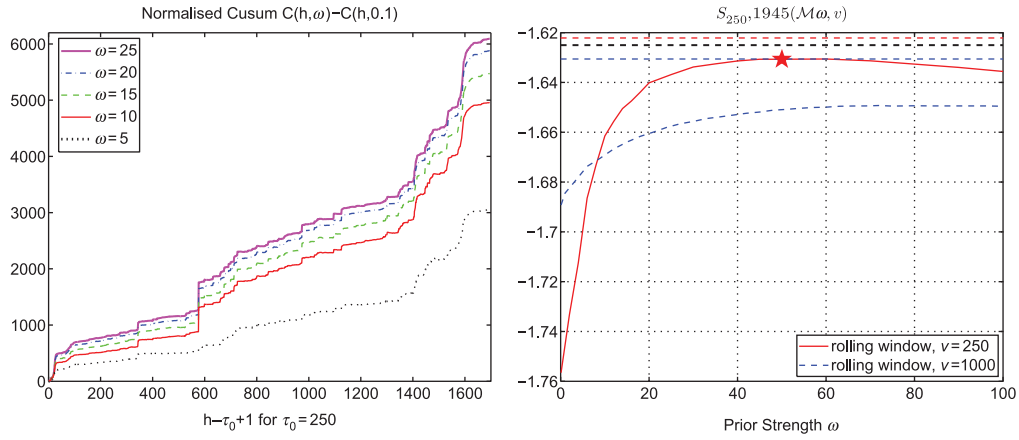

Figure 5. Left: Plot of the standardized cusum $C(h,\omega) - C(h,0.1)$, where C is given in Equation (13), versus $h - \tau_0 + 1$, for $\tau_0 = 250$ and several $\omega = 5, 10, 15, 20, 25$. Right: The normalized sum of the realized predictive log-likelihood (12) as a function of prior strength hyper-parameter ω, and based on estimation with a moving window of length $v = 250$ (solid line) and $v = 1000$ (dashed line). For the latter, τ_0 is still 250, and we use the convention in Equation (14). The star shows the best obtained value, corresponding to a prior weight of $\omega = 50$, and is the same star in the both panels of Figure 12 below, while the top-most horizontal line is the same line in the right panel of Figure 12, showing the additional improvement from the methods discussed in Section 6.

The left panel of Figure 5 plots the cumulative sum (cusum) of the $\pi_t(\mathcal{M}_\omega, 250)$, normalized by subtracting the cusum of $\pi_t(\mathcal{M}_{0.1}, 250)$; that is, we plot $C(h,\omega) - C(h,0.1)$, where

$$C(h,\omega) = \sum_{t=\tau_0+1}^{h} \pi_t(\mathcal{M}_\omega, 250), \quad h = \tau_0 + 1, \ldots, T, \tag{13}$$

versus $h - \tau_0 + 1$, for $\omega = 5, 10, 15, 20, 25$ and $T = 1945$. The improvement in forecast accuracy is virtually monotonically increasing with both increasing h and increasing ω, providing very strong evidence that shrinkage estimation vastly outperforms the use of the MLE in this context. In fact, the gains from using the shrinkage estimator compared to the MLE are higher than what are shown because we used the benchmark model $\mathcal{M}_{0.1}$ instead of the MLE $\mathcal{M}_0$.

The right panel of Figure 5 plots $S_{250,1945}(\mathcal{M}_\omega, 250)$ (solid line) as a function of ω, the weight of the shrinkage prior, from which we can see that the optimal amount of shrinkage based on $v = 250$ is, say, $\omega^*(250) = 50$. The plot also shows $S_{250,1945}(\mathcal{M}_\omega, 1000)$, i.e. based on a moving window of size $v = 1000$, and $\omega^*(1000) \approx 65$. Observe that $\tau_0 < v$, which appears to not make sense: We use the convention that, for computing $\pi_t(\mathcal{M}, v)$, we use the previous

$$\min(v, t-1) \tag{14}$$

observations. So, for example, with $\tau_0 = 250$ and $v = 1000$, at observation $t = 251$, we use the past 250 observations; at $t = 252$, we use the past 251 observations; etc., up to $t = 1000$, for which we use the past 999 observations; and for $t \geq 1001$, we use the previous $v = 1000$ observations in the information set I_{t-1}.

From the right panel of Figure 5, we immediately see three facts.

(1) When using the MLE ($\omega = 0$ in the plot), use of a larger sample size v for estimation (in this case 1000 versus 250) leads to improvement in the density forecasts.
(2) The effect of shrinkage (or the strength of the prior in a Bayesian or quasi-Bayesian method) decreases as the sample size v increases.
(3) When using $\omega^*(v)$, the optimal amount of shrinkage for a given sample size v, the quality of the density forecasts are not necessarily better as v increases.

When we take these three observed facts together, it might seem somewhat puzzling, if not disturbing, that forecast accuracy improves so much by using a shrinkage prior, and, when using it, the accuracy gets *worse* as the window size is increased. *This is because the assumed data generating process (DGP) is wrong.* One blatant reason it is wrong is because we assume an iid model, and so are ignoring volatility clustering, one of the most fundamental stylized facts of asset returns data. However, part of our modeling strategy is precisely to do away with GARCH or other stochastic volatility filters in order to simplify the model, avoid complex or even infeasible estimation problems, and capitalize on the resulting structure to allow for fast and reliable estimation via an EM algorithm. While it might at first blush appear that not accounting for volatility clustering via a GARCH-type model is equivalent to a gross neglect of information in the information set, and would lead to comparatively inferior forecasts, this is not precisely the case. The fact that the volatility is highly persistent and roughly constant over short time intervals implies that use of an iid model *with a small window* could be at least reasonably accurate, and quite possibly superior to a simplistic, autoregressive law of motion for the scale term via a GARCH filter. In doing so, we are faced with the all-too-common situation of having to make a tradeoff between bias and variance, and seek the optimal window length to minimize some criterion such as mean squared error. In light of this, we could entertain finding the optimal window size, say v^*, given by

$$v^* = \arg\max_v S_{\tau_0,T}(\mathcal{M}_{\omega^*(v)}, v), \tag{15}$$

which, in this case, would be somewhere between $v = 1$ and $v = 1000$. (In general, we require at least $k(d+1)$ observations to estimate a $\text{Mix}_k N_d$ model, but because we use the shrinkage prior (5) with $\omega > 0$, we can use less than this number of observations and still not land on a singularity in the likelihood.)

Figure 6 illustrates the idea by showing (12) versus window size v, for three values of hyperparameter ω and two values of τ_0, 100 (left) and 600 (right). From the right panel, we see that use of $\omega = 50$ dominates the other two values of ω for all v, and that, irrespective of ω (at least for the three values considered), the optimal choice of v is between 250 and 275. Comparison with the left panel shows that the effect of the choice of prior strength has a far greater impact on the performance when using a smaller τ_0, even though the density forecasts for observations beyond $t = 600$ in the time series are the same for both values of τ_0. Comparing the two graphs corresponding to $\omega = 50$ shows that Equation (12) is lower for the $\tau_0 = 100$ case for all v, in particular, $v = 100$. Recalling that Equation (12) standardizes by the number of π_t in the sum, this indicates that the density prediction of observations $t = 101, \ldots, 600$ was, relative to the remaining observations, less successful.

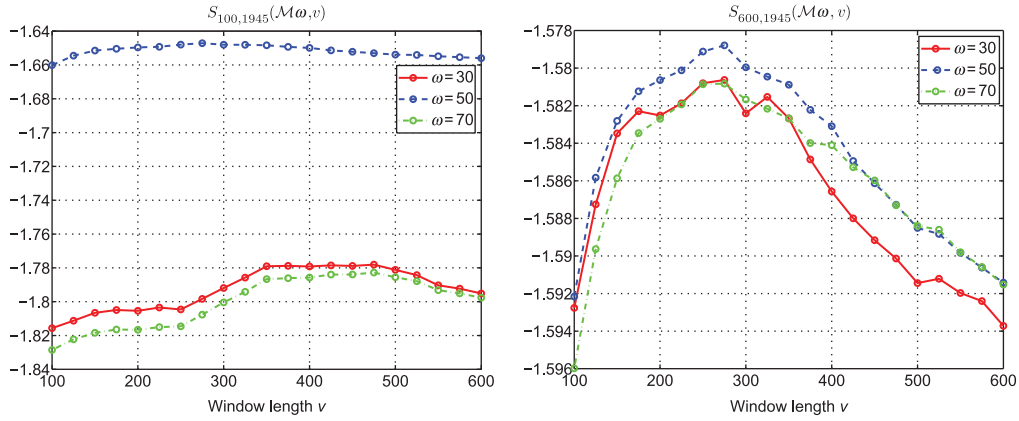

Figure 6. Both panels show the normalized sum of the realized predictive log-likelihood (12) as a function of moving window size v, $v = 100, 125, 150, \ldots, 575, 600$, for three values of prior strength hyper-parameter ω. The left uses $\tau_0 = 100$, while the right uses $\tau_0 = 600$. In the left panel, the plot for $\omega = 50$ (the one close to the top) has the same shape as the corresponding one in the right panel when the plot is magnified, with its maximum also at $v = 275$.

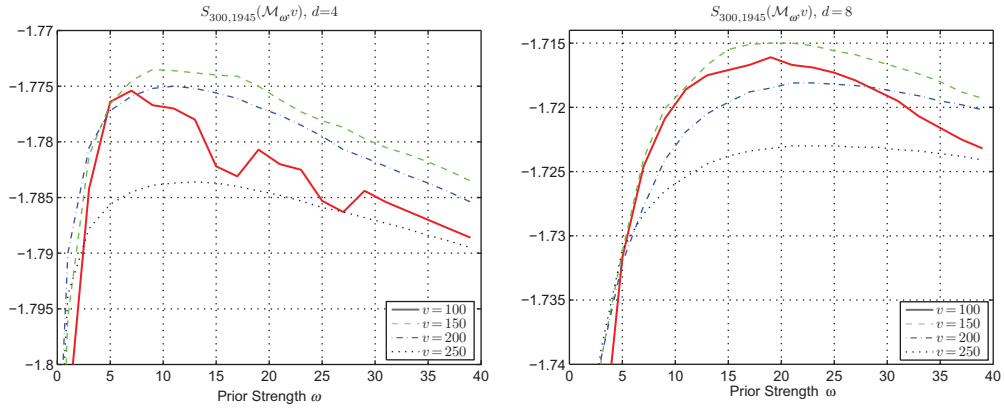

Figure 7. Similar to the right panel of Figure 5, except that here, the left panel shows the results for only $d = 4$ assets (and four values of v), while the right is for $d = 8$. Observe that $\tau_0 = 300$ for all four window sizes, so that the density predictions were based on precisely the same data points, namely $t = \tau_0 + 1 = 301$ to $t = T = 1945$.

5.3 *Optimal shrinkage as a function of d*

In the empirical analyses up to this point, all $d = 30$ series which constitute the DJ-30 were used. However, it is of obvious interest to investigate the performance using different values of d. With $d = 4$ and arbitrarily using the first four series in the DJ-30: 3M Company (MMM), Alcoa Inc. (AA), American Express Company (AXP), and AT&T Inc. (T), the left panel of Figure 7 shows that a window size of about $v = 150$ is optimal, and for which $\omega^*(150) \approx 10$. Similarly, we take $d = 8$, adding the next four stocks to the set: Bank of America Corporation (BAC), Boeing Company (BA), Caterpillar Inc. (CAT), and Chevron Corporation (CVX).

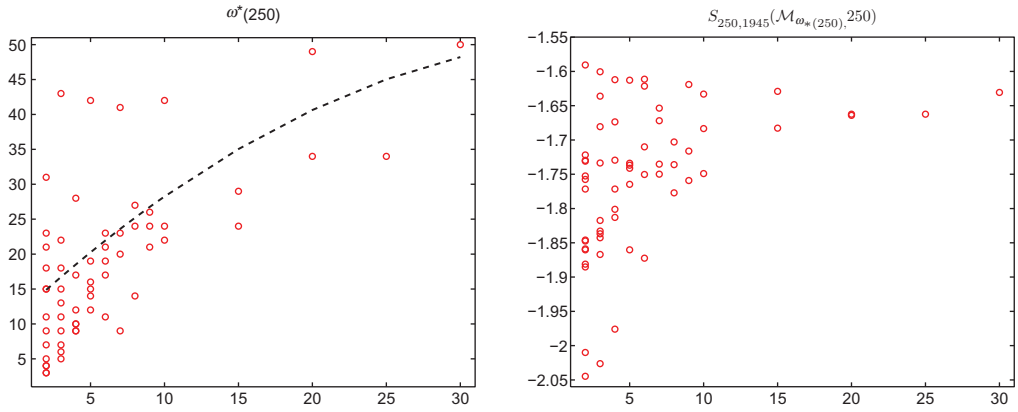

Figure 8. Left: The optimal value of ω for various subsets of the DJ-30 assets under study, and the fitted regression line of those points given by round($\mathbf{X}\hat{\boldsymbol{\beta}}$), where $\mathbf{X} = [1, d, d^2]$ and $\hat{\boldsymbol{\beta}} = [8.4272, 1.9604, -0.0240]'$. Right: The values of Equation (12) corresponding to the optimal value of ω.

Again, $v = 150$ is preferred, with $\omega^*(150) \approx 20$. This would appear to temper our previous finding regarding the optimality of using approximately one year of daily trading data. However, below in Section 6 when we use weighted likelihood and moving averages of λ, we will find that use of $v = 250$ is still better than $v = 150$ (Figure 13). This is because the gains from using a moving average of λ with $v = 250$ far exceed those based on $v = 150$. The reason for this is very appealing: While there *is* more information in a window of 250 observations than there is with 150, emphasis (via weighted likelihood and moving averages of λ) needs to be placed on more recent observations. Without this, the $v = 150$ case will outperform the $v = 250$ case simply because relatively less valuable observations have been removed from the window.

From the plots in Figure 7, $\omega^*(250) \approx 12$ for $d = 4$, while for $d = 8$, $\omega^*(250) \approx 22$. One might thus conjecture that $\omega^*(250)$ is increasing in d. To investigate this, we conducted this exercise with all seven non-overlapping subsets of $d = 4$ series (omitting series 29 and 30). Continuing, we do this for all 15 subsets of $d = 2$; all 10 subsets of $d = 3$, etc., for $d = 2, 3, \ldots, 10$, as well as $d = 15$; and also $d = 20$ (using series 1 through 20, and 11 through 30, so that in this case, there is overlap), $d = 25$ (series 1 through 25), and all $d = 30$ series. For each data set, values $\omega = 1, 2, \ldots, 60$ were tried.

The results are shown in Figure 8. The left panel shows the $\omega \in \{1, 2, \ldots, 60\}$ which yielded the highest value of Equation (12), denoted $\omega^*(250)$, as well as a fitted regression line in d and d^2. The right panel shows the corresponding maximal values of the predictive log-likelihoods (12). Recalling that Equation (12) divides by d, the values are comparable; we see that, as d increases, the quality of the forecasts tends to increase, and also the variability decreases. Thus, for this model and this data set, more assets tend to lead to better density forecasts.

The left panel of Figure 9 is the same as the right one in Figure 8, but uses the value of ω obtained from the aforementioned fitted regression line, denoted $\omega^{\text{reg}}(250)$. There is virtually no difference in quality. As a comparison, the right panel of Figure 9 is similar to the left, but uses the fixed value $\omega = 1$ for all the data sets. In this case, the differences are far more pronounced, and they increase in d, as expected, given that $\omega^{\text{reg}}(250)$ increases in d.

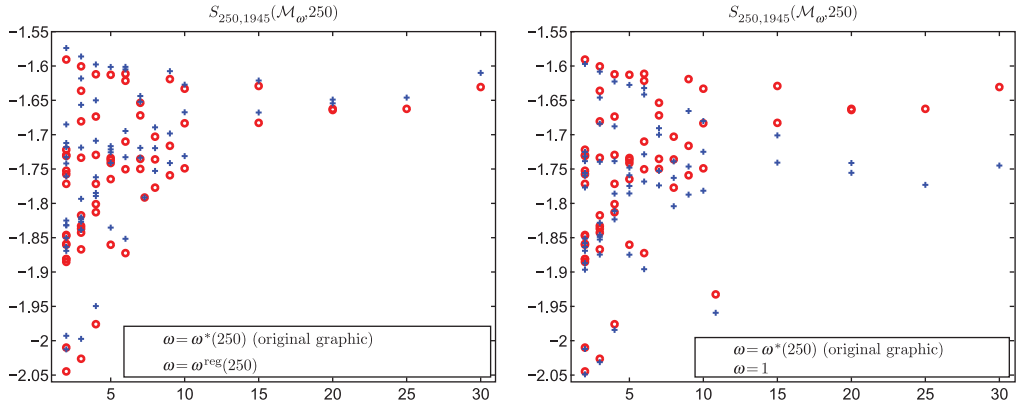

Figure 9. Left: Overlays same plot in the right panel of Figure 8, and additionally shows, as crosses, the result when taking ω to be from the regression line depicted in the left panel of Figure 8. The resulting values based on $\omega^{*}(250)$ and $\omega^{\mathrm{reg}}(250)$ are virtually identical. Right: Same as left, but based on a fixed value of $\omega = 1$.

5.4 *Model comparison*

Interest now centers on comparing its performance to that of CCC, DCC, and two of their extensions. Such models are valuable because, like the $\mathrm{Mix}_k N_d$ and its extension $\mathrm{Mix}_k \mathrm{Lap}_d$ introduced below, they are feasible to estimate for more than a small handful of assets. The first extension considered is the asymmetric generalized DCC of Cappiello, Engle, and Sheppard (2006, Eq. 5), which generalizes the DCC model in two regards, the primary one of which being the inclusion of parameters to capture asymmetries in the equation dictating the law of motion for the correlations. Those authors demonstrate the statistical significance of the asymmetry terms based on real data. Our concern is the extent to which this significance translates into improved density forecasts. We consider only the scalar parameter case, denoted A-DCC by those authors, because it is still feasible to estimate for more than a small number of assets.

The left panel of Figure 10 shows the same forecasting results as obtained previously for the $\mathrm{Mix}_k N_d$ model (namely, the contents of the right panel of Figure 8), and overlays this with the parallel results based on use of the A-DCC model. The scaling of all the plots in Figures 9 and 10 are the same. We immediately see that, like with the $\mathrm{Mix}_2 N_d$ model, forecast quality based on the A-DCC tends to improve as d increases, but also that the $\mathrm{Mix}_2 N_{30}$ model with shrinkage outperforms A-DCC. The results for CCC and DCC are not explicitly shown: It turns out that the corresponding values are graphically nearly indistinguishable from the A-DCC model, with only a barely visible preference of DCC over CCC, and A-DCC over DCC. In light of what appears to be a general consensus in the literature that DCC is significantly superior to CCC, and the encouraging in-sample results favoring asymmetry, as reported by Cappiello, Engle, and Sheppard (2006), these results were somewhat surprising.

An extension of the CCC framework which also recognizes the presence of more than one component in the data is the Regime Switching Dynamic Correlation (RSDC–)GARCH model, introduced by Pelletier (2006). Like the CCC model, a univariate GARCH model describes the scale parameter of each time series, but instead of CCCs, it uses a mixture of k distributions, where, like with our $\mathrm{Mix}_k N_d$ model, k is a fixed value, and such that the positive definite correlation matrices, $\mathbf{P}_{(i)}$, $i = 1, \ldots, k$, are regime-specific, and thus allowed to be different among the k components.

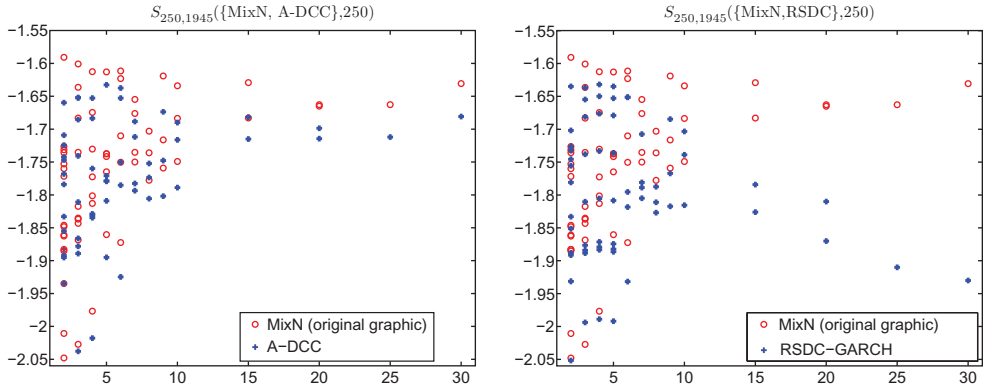

Figure 10. Left: Shows the same values for the Mix_2N_{30} model as depicted in the right panel of Figure 8, and with the same scaling to enable comparison, along with results from the A-DCC model. Three values for the A-DCC model are not shown (2 for $d = 2$; 1 for $d = 3$). Right: Same, but showing the results based on the RSDC–GARCH model from Pelletier (2006).

The right panel of Figure 10 is similar to the left, but displays the comparison with the performance of the two-component RSDC–GARCH model. As with the A-DCC, it is also outperformed by $\text{Mix}_k N_d$. Comparing the small solid circles in the left and right panels, we see that RSDC–GARCH is superior to A-DCC (and, thus, to the CCC and DCC) for small values of d, but begins to fail as the number of assets increases. The good results for small d can be attributed to the evidence provided above on the presence of (at least) two components in the data, a feature which the RSDC–GARCH model can capture. Nevertheless, and despite its incorporation of GARCH and CCC structures, it is still inferior to the unconditional normal mixture with shrinkage for small values of d. To explain its failure as d increases, evidence from simulation studies suggests that, for the sample size in use here ($T = 250$), as d increases, estimation of the transition probability matrix becomes quite inaccurate.

Given the appeal of the RSDC–GARCH model and its success for small d, several ways suggest themselves for improving it, such as finding the optimal sample size for maximizing predictive accuracy, and/or use of much larger sample sizes but with weighted likelihood; ways of enhancing estimation of all the parameters, such as use of shrinkage; and use of the multivariate Laplace distribution in place of the MVN.

5.5 *Comparison to use of a covariance stationary process*

As a reference point of comparison, if not a straw man, consider using the ordinary MVN distribution, with the MLE for the mean vector and covariance matrix. The resulting plot, which parallels those previously, is shown in the left panel of Figure 11. Unsurprisingly, the quality of the density forecasts is demonstrably lower than all models so far considered, but we also see that, unlike the other models, the quality substantially *decreases* as the number of assets is increased.

Returning to our setting (with daily data), we see by comparing Figures 9 and 10 with Figure 11 that, upon relaxing the assumption of (unconditional) covariance stationarity and assuming a strictly stationary process modeled with a CCC–GARCH model, density forecast accuracy increases with d. Alternatively, and as we advocate in this paper, the assumptions of strict stationarity and ergodicity should not be assumed, but rather only a certain 'smoothness' on the true DGP, in the sense that its underlying parameters change gradually enough through time, and/or

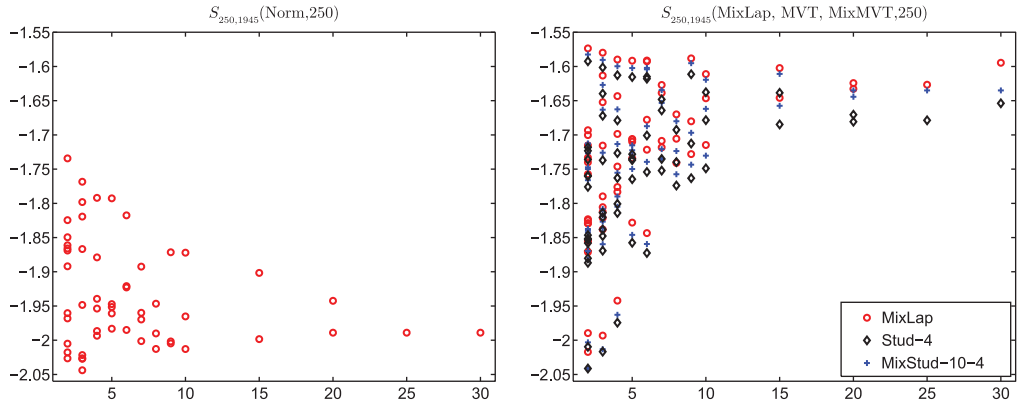

Figure 11. Left: Similar to the plot in the right panel of Figure 8, but based on the plain (single component, no shrinkage) MVN distribution. Right: Overlaid are the results based on the Mix_2Lap_d distribution (22) in Section 7.2; the multivariate Student's *t* distribution (MVT) with fixed degrees of freedom 4 and the two-component multivariate Student *t* mixture (using 10 and 4 degrees of freedom, for the two components).

with a small but non-zero probability at each time point of encountering a structural break in the parameters or the model structure. With this assumption, it makes sense to entertain a relatively simple model which is feasible to estimate and still captures the salient features of the data adequate for serving as a local approximation to the true DGP. It should be based on use of relatively small windows of data and/or with weighted likelihood for estimation, in order to emphasize more recent observations in the window, which carry, unlike in a strict-stationary ergodic setting, relatively more information about the time-varying process than observations further in the past.

The poor results in the left panel of Figure 11 are certainly at least partially due also to the normality assumption, which is well-known to be untenable. The right panel is similar, but having used the MVT distribution with fixed degree of freedom $\nu = 4$. The EM algorithm required for its estimation is given in McLachlan and Krishnan (2008). The result is an enormous improvement in fit, quite similar in quality to the use of the Mix_2N_d, and also such that, like the previously entertained models, density forecast accuracy increases with d. Thus, the normality assumption is not just partially, but primarily, responsible for the poor fit. Anticipating our results below for the multivariate mixture Laplace distribution, we see from the overlaid values in the plot that the values from use of the Mix_2Lap_d are superior to those of the MVT.

The results based on the multivariate *t* are in fact so good when compared only to the Mix_2N_d model that one might be inclined to do away with the latter, despite the evidence for the presence of a mixture discussed in the previous sections. As we show below, one key to further model improvement will be to maintain the mixture structure, but replace the Gaussian distribution with the multivariate Laplace. Combining these lines of thinking, one is behooved to entertain a mixture of (say two) multivariate *t* distributions, each with its own degree of freedom, thus reaping benefits from both types of structures. We conducted this using two components and fixed values of the two degrees of freedom parameters, say k_1 and k_2. Some trials suggest that taking $k_1 = 10$ and $k_2 = 4$ is essentially optimal. The results, also overlaid in the right panel of Figure 11, indicate that it is not only inferior to the Laplace mixture in terms of forecasting performance, but also to the single component MVT. In fact, from a numerical point of view, it was found that the EM algorithm for its estimation (with, as mentioned, fixed degrees of freedom, which is faster and numerically more reliable), as detailed in McLachlan and Peel (2000), takes over 15 times as long

to estimate as that for the multivariate Laplace mixture. If the two degrees of freedom parameters were jointly estimated with the others, then the procedure would presumably take yet even longer, and also be numerically problematic, because the likelihood in k_1 near its maximum would be quite flat, and might also tend toward infinity, given the results shown above in Section 4.2 on the data in the first component.

From a forecasting point of view, the Laplace mixture resulted in superior performance. We conjecture the reason for this to be that the tails of the *t*, *when used in a mixture context*, are too fat; recall the results in Section 4 and that random variable $X \sim t_\nu$ (the univariate Student's *t* distribution with ν degrees of freedom) has genuine power tails, so that $\mathbb{E}[|X|^m]$ does not exist for $m \geq \nu$. While higher kurtosis is still required for the second component, the Laplace offers this without power tails and the ensuing upper bound on finite absolute positive moments. Moreover, as demonstrated in Section 4.2, via use of the GAt distribution and the fitted values of peakedness parameter *p*, there is evidence that the two components, particularly the second, have a more peaked distribution than the Student's *t*. In summary, when used as a single component model, the MVT is superior to the Laplace (a result which makes sense, given the very fat tails of asset returns), but with two components, the reverse is true.

6. Weighted likelihood and moving averages of λ

Returning to the $\text{Mix}_k N_d$ model, and now satisfied with a window size of $\nu = 250$ and a shrinkage weight of $\omega = 50$ (for the full, $d = 30$ case), we entertain methods for extracting more information out of the data to further improve the density forecasts. Two ways are considered, both of which add a relatively small amount of improvement on par with that gained by replacing the multivariate Student's *t* distribution with a normal mixture distribution with shrinkage estimation, and less than the substantial gains achieved by replacing the mixed normal assumption by mixed Laplace, as demonstrated in the right panel of Figure 14. The first way is via the use of weighted likelihood; see Paolella and Steude (2008) and the references therein for a detailed account.

Briefly, for a set of ν observations, to implement the weighting scheme, a vector of weights $\boldsymbol{\varpi} = (\varpi_1, \ldots, \varpi_\nu)$ is used such that it is standardized to sum to a constant, such as ν (as with the conventional MLE), or to one, which is what we choose. The model parameters are then estimated by maximizing the weighted likelihood, whereby the likelihood component associated with period *t* is multiplied by ϖ_t. While there are many possible weighting schemes, two essentially obvious ones are the geometric and hyperbolic, given, respectively, by

$$\varpi_t \propto \rho^{\nu-t}, \quad \varpi_t \propto (\nu - t + 1)^{\rho-1}, \quad \sum_{t=1}^{\nu} \varpi_t = 1, \tag{16}$$

where the single parameter ρ dictates the shape of the weighting function. In both cases, values of $\rho < 1$ ($\rho > 1$) cause more recent observations to be given relatively more (less) weight than those values further in the past; while $\rho = 1$ corresponds to the standard, equally weighted likelihood. It is important to observe that the optimal value of the tuning parameter ρ cannot be estimated with the model parameters by maximizing the likelihood, but must be obtained with respect to some criterion outside of the likelihood function. In particular, as our goal is future density prediction, we wish to maximize Equation (12).

As mentioned, to apply the weighting scheme, the *t*th term entering into the log-likelihood gets multiplied by its corresponding weight ϖ_t, $t = 1, \ldots, \nu$. However, as we use the EM algorithm for the $\text{Mix}_2 N_{30}$ model, this direct implementation is not available. To accommodate this, we

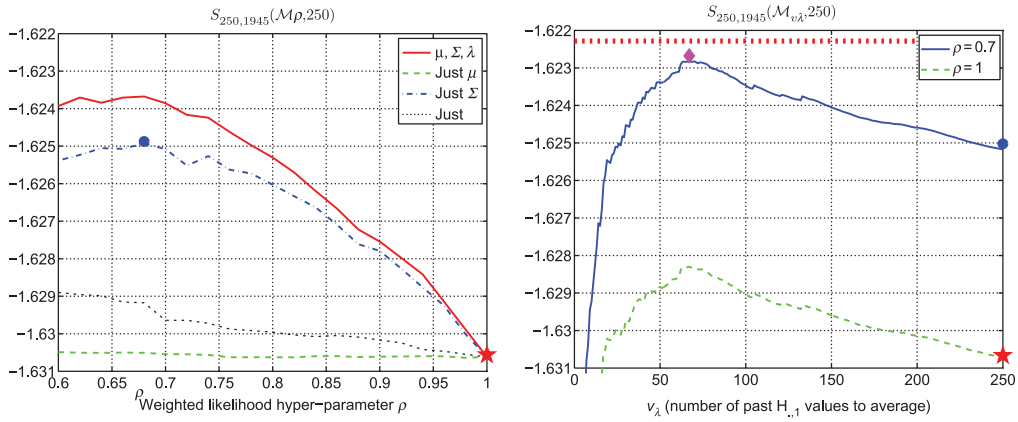

Figure 12. Left: Equation (12) for $\nu = 250$ and shrinkage hyper-parameter $\omega = 50$, as a function of hyper-parameter ρ, which controls the shape of the weights used in the weighted likelihood calculation. The star at the bottom right of the left and right panels is the same star shown in the right panel of Figure 5. Right: Equation (12) for $\nu = 250$ and shrinkage hyper-parameter $\omega = 50$, as a function of ν_λ, which dictates how many of the latest ν_λ values of $\{H_{\cdot,1}\}$ are averaged to form the value of $\hat{\lambda}$. The big circle in both plots represents the same value. Both plots have the same *y*-axis range, so that it is easy to see the improvement in using $\nu_\lambda = 70$ with weighted likelihood for $\rho = 0.7$ applied just to the $\hat{\boldsymbol{\Sigma}}_j$, compared to using $\nu_\lambda = 250$ with weighted likelihood applied to all model parameters, including λ. Finally, the horizontal line at the top of the graph is the result of taking $\hat{\lambda}$ to be $0.75\hat{\lambda}_{250,70} + 0.25\hat{\lambda}_{250,1}$.

suggest the following: Multiply each $\mathbf{Y}_t$ appearing in the mean updating Equation (3) by ϖ_t; multiply each $\mathbf{Y}_t$ appearing in the variance updating Equation (4) by ϖ_t; and multiply each $H_{t,j}$ in the component weight updating Equation (2) by ϖ_t. (Another possibility is just to multiply H_{tj} by ϖ_t, so that, in Equations (3) and (4), the weights appear in both the numerator and also the denominator. However, this was found to be less satisfactory, and is not subsequently used.) In fact, this structure allows us to apply the weights to any subset of these three equations. We consider the four possibilities of applying the weight to all three of them, and to just one of them. We use the hyperbolic weighting scheme in Equation (16) and compute Equation (12) as a function of ρ, based, as usual, on a moving window of $\nu = 250$ observations, and using a shrinkage parameter of $\omega = 50$ for the $d = 30$ asset case.

Figure 12 shows the results. Most of the improvement comes from applying the weights to the $\hat{\boldsymbol{\Sigma}}_j$, whereas virtually no improvement is obtained from weighting the $\hat{\boldsymbol{\mu}}_j$.

The optimal value of ρ is about 0.675, which we take to be just 0.7. The large circle in the graphic indicates the maximal obtained normalized sum of the realized predictive log-likelihood when just applying the weighted likelihood to the $\hat{\boldsymbol{\Sigma}}_j$. We use this in the subsequent method of improvement instead of the slightly better result obtained by applying weighted likelihood to all the parameters because we will require the unweighted values of the $H_{t,1}$.

As just alluded to, consider now a related approach involving the $H_{t,1}$ which can, and will, be used in conjunction with the method of weighted likelihood. Recall Figure 2, showing the final values of $H_{t,1}, t = 1, \ldots, 1945$, output from the EM algorithm, for the Mix_2N_{30} model. Particularly from the right panel, it is apparent that the $H_{t,1}$ are highly correlated, indicating that today's value of $H_{t,1}$ might be a good predictor of tomorrow's. In the estimation schemes used up to this point, we ignored the information in the sequence $\{H_{t,1}\}$ and just used the MLE $\hat{\lambda}_1 = \hat{\lambda}$, which is just the mean of the $H_{t,1}$, recalling Equation (2). One natural suggestion would be to take $\hat{\lambda}$ used in $\mathcal{M}$

for calculating the predictive density of $\mathbf{Y}_t$ based on a rolling window of $v = 250$ observations to be the average of the latest, say, v_λ values of $\{H_{\cdot,1}\}$, which we denote as $\hat{\lambda}_{v,v_\lambda}$, i.e.

$$\hat{\lambda}_{v,v_\lambda} = v_\lambda^{-1} \sum_{t=v-v_\lambda+1}^{v} H_{t,1}, \quad 1 \le v_\lambda \le v. \tag{17}$$

If $v_\lambda = v$, then $\hat{\lambda}_{v,v}$ is just the usual $\hat{\lambda}$, while $\hat{\lambda}_{v,1}$ is just the last value of $\{H_{\cdot,1}\}$ in the window of observations.

The solid line in the right panel of Figure 12 shows $S_{\tau_0,T}(\mathcal{M}_{v_\lambda}, v)$, for $v = \tau_0 = 250$, as a function of v_λ, based on the weighted likelihood parameter $\rho = 0.7$ (applied only to the $\hat{\mathbf{\Sigma}}_j$), and where $\mathcal{M}_{v_\lambda}$ is the Mix_2N_{30} model but such that $\hat{\lambda}$ is replaced by $\hat{\lambda}_{v,v_\lambda}$ in Equation (17). The dashed line is similar, but corresponds to $\rho = 1$, showing that virtually the same amount of improvement is gained with this method, irrespective of ρ, and that the optimal value of v_λ is not dependent on ρ. Indeed, we see that there are nearly monotone gains in forecast accuracy obtained as v_λ is decreased, and a maximum is reached at about $v_\lambda = 70$, after which the quality drops off rather abruptly. As $v_\lambda \to 1$, i.e. as we approach the strategy of taking $\hat{\lambda}$ to be the last value of $\{H_{\cdot,1}\}$, the performance turns abysmal, and the graph was truncated. So, it appears that use of v_λ corresponding to about 14 weeks of daily data is superior to use of the whole year. While this is useful, the strong correlation among the $H_{t,1}$ apparent in Figure 2 would suggest that the previous day's value of $H_{t,1}$ should still somehow be of value. This is the case: taking $\hat{\lambda}$ to be

$$\hat{\lambda}_{\text{mix}} := 0.75\hat{\lambda}_{250,70} + 0.25\hat{\lambda}_{250,1} \tag{18}$$

results in further improvement, shown as the horizontal dotted line in the right panel of Figure 12.

While the graphics in Figure 12 display the increase in density forecast performance from using weighted likelihood with $\rho = 0.7$ and, on top of that, from using $v_\lambda = 70$, the right panel of Figure 5 shows that these gains are relatively small compared to what is achieved from shrinkage. In that plot, the horizontal dashed lines at the top show the incremental gains from the weighted likelihood for the $\hat{\mathbf{\Sigma}}_j$ with $\rho = 0.7$, and the additional gain obtained by further taking $\hat{\lambda}$ to be Equation (18).

Thus, these additional tools provide only modest improvements relative to what is achieved with shrinkage. A more substantial improvement can be made by improving upon the normality assumption, as detailed below in Section 7. On top of all this, one might expect use of a GARCH structure for the $\hat{\mathbf{\Sigma}}_j$ to extract yet more information out of the signal hidden in the data. This is an interesting prospect, because with two or more mixture components, each can be endowed with a (possibly different) GARCH-type structure to better capture the rich dynamics of multivariate asset returns data. We do not pursue this here because it would only be practical, in its most general setting, for a small number of assets, in contradiction to our stated goal here of applicability to large d.

In Section 5.2, we also considered the case with $d = 4$ and $d = 8$ assets. Figure 13 is similar to Figure 12, but is for the $d = 4$ case, and having used two window sizes; $v = 250$ (with shrinkage parameter $\omega = 12$, in accordance with the results from the left panel of Figure 7), and $v = 150$ (with $\omega = 10$). The results are very similar qualitatively to the $d = 30$ case, particularly the optimal value of ρ, which we take to be 0.65. However, observe that the optimal value of v_λ from Equation (17) is 10, though interestingly, we see from the right panels of Figure 13 that there is a local maximum at $v_\lambda = 70$, precisely the optimal value for the $p = 30$ case. Using values $0.75\hat{\lambda}_{v,70} + 0.25\hat{\lambda}_{v,1}$ and $0.75\hat{\lambda}_{v,10} + 0.25\hat{\lambda}_{v,1}$ gave results (not shown) which were very close to

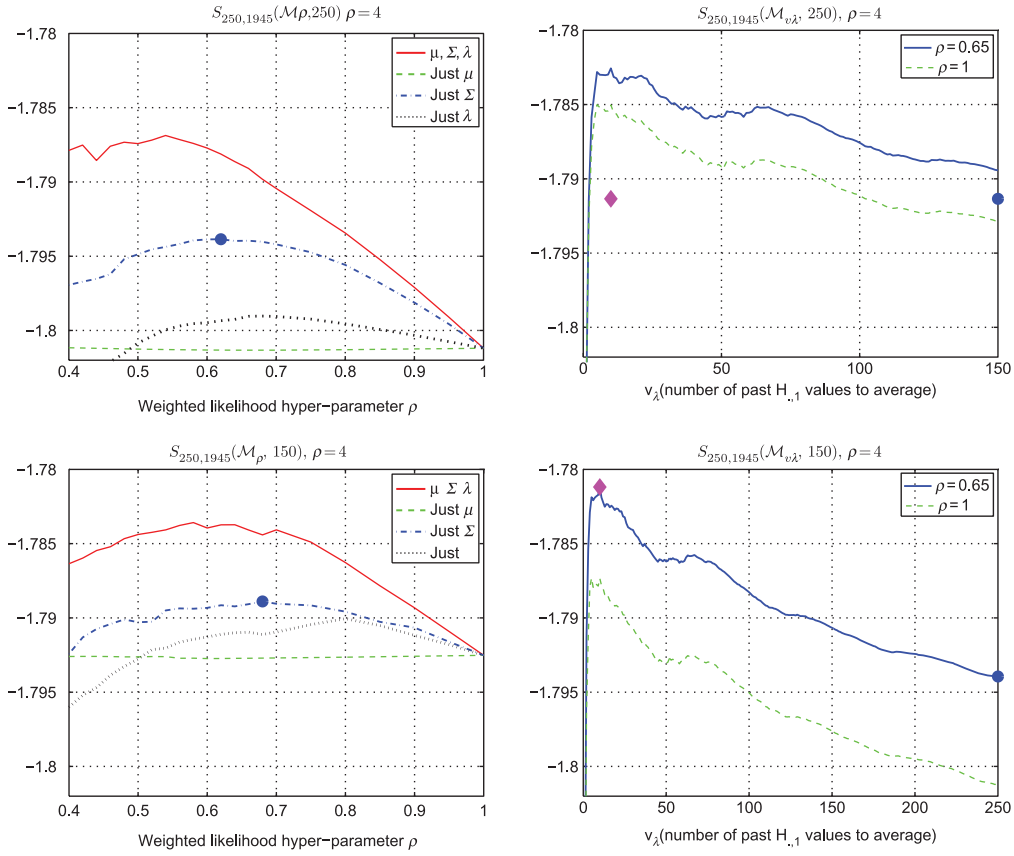

Figure 13. Similar to Figure 12 but using $d = 4$ assets instead of $d = 30$. Top is window size $v = 250$ and $\omega = 12$ (as ascertained from the left panel of Figure 7); bottom is window size $v = 150$ and $\omega = 10$. For both window sizes, $\tau_0 = 250$ so that the results are directly comparable.

(and below) the optimal value shown in the right panels as the large diamond. The rapid decline of the quality as v_λ decreases from 10 to 1 is alarming, so that use of $v_\lambda = 70$ might be a safer choice in practice. (A similar result, not shown, was found using $v = 500$ observations: a global, and sharp, peak at $v_\lambda = 10$ and a local maximum at $v_\lambda = 70$.)

Finally, comparing the two cases of $v = 250$ and $v = 150$ (note that the scaling of the y-axis is the same in all four plots), we see that, while the $v = 150$ case is superior without weighted likelihood and just using the default of $v_\lambda = v$, the $v = 250$ case at its optimal value, with weighted likelihood and $v_\lambda = 10$, is better than the $v = 150$ case at its optimal value (albeit not by much). It is less a matter that $v = 250$ with weighted likelihood and $v_\lambda = 10$ is *better* than $v = 150$; it is enough that they are close: This indicates, as already briefly discussed in the previous section, that there is indeed more information about the model parameters in the last 250 observations than in the last 150, but because the model is (without question in nearly every application) misspecified, we need to weight more recent observations and information relatively heavier than those further in the past. This is precisely what weighted likelihood and use of Equation (17) are accomplishing.

The left panel of Figure 14 illustrates the improvement from weighted likelihood and use of $\hat{\lambda}_{\text{mix}}$ in Equation (18) for the same groups of assets used in Figures 8 and 9. The right panel of Figure 14

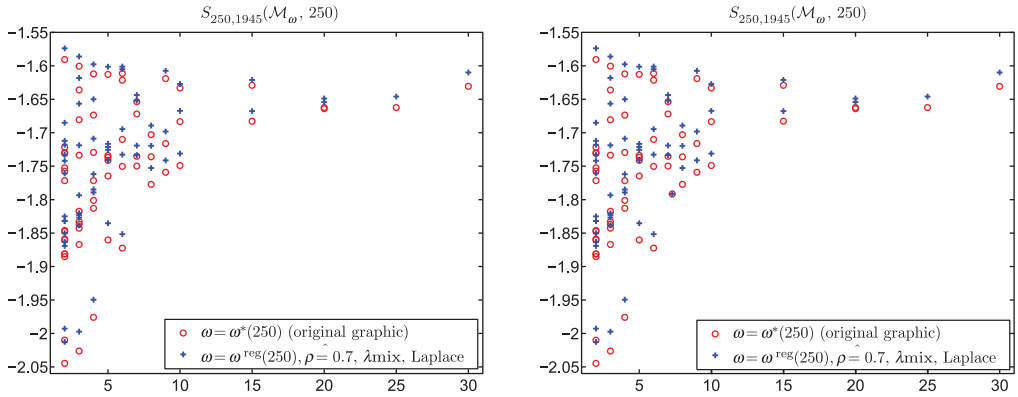

Figure 14. Left: Overlays same plot in the right panel of Figure 8, and additionally shows, as crosses, the result when taking ω to be from the regression line from the left panel of Figure 8 and additionally with (i) weighted likelihood, as discussed in Section 6, with $\rho = 0.7$ and just applied to the $\hat{\mathbf{\Sigma}}_j$, and (ii) a moving average of λ from Equation (18), i.e. $0.75\hat{\lambda}_{250,70} + 0.25\hat{\lambda}_{250,1}$. Right: Same as left, except that instead of the $\text{Mix}_2 N_d$ distribution, we use the Mix_2Lap_d distribution (22) from Section 7.2.

shows the results when using the two-component Laplace mixture distribution, introduced below in Section 7, instead of the normal mixture. Its use bestows a considerable improvement in forecast quality, particularly as d increases.

7. Mixtures of multivariate Laplace

7.1 *The multivariate Laplace distribution*

Section 4.2.3 motivated the use of the Laplace distribution as an alternative to the normal, while Section 5.5 showed why a mixture of Student t distributions is inferior to the use of Laplace. The derivation of the multivariate Laplace requires use of the integral form of the modified Bessel function of the third kind, which is given by

$$K_z(x) = \frac{1}{2}\int_0^\infty u^{z-1}\exp\left[-\frac{x}{2}\left(\frac{1}{u}+u\right)\right]\mathrm{d}u. \tag{19}$$

The appendix shows that

$$f_{\mathbf{Y}}(\mathbf{y},\boldsymbol{\mu},\mathbf{\Sigma},b) = \frac{1}{|\mathbf{\Sigma}|^{1/2}(2\pi)^{d/2}}\frac{2}{\Gamma(b)}\left(\frac{m}{2}\right)^{b/2-d/4}K_{b-d/2}(\sqrt{2m}), \tag{20}$$

where $m = (\mathbf{y}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$. This distribution is a (symmetric) multivariate variance gamma mixture of normals, but for convenience, we refer to it just as the multivariate Laplace, and write $\mathbf{Y} \sim \text{Lap}(\boldsymbol{\mu}, \mathbf{\Sigma}, b)$. See also Podgórski and Kozubowski (2001), in which a similar distribution is given.

It is straightforward to confirm that $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}$, while

$$\mathbb{V}(\mathbf{Y}) = \mathbb{E}[\mathbb{V}(\mathbf{Y} \mid G)] + \mathbb{V}(\mathbb{E}[\mathbf{Y} \mid G]) = \mathbb{E}[G]\mathbf{\Sigma} = b\mathbf{\Sigma}, \tag{21}$$

recalling that $\mathbb{E}[G] = b$.

The EM algorithm for estimating Equation (20) is applicable. Conditional on G_i, the distribution of $\mathbf{Y}_i$ is normal, so let the set of G_i be the latent, unobserved variables, which are iid $\text{Gam}(b, 1)$,

where we assume that b is known. Assuming known b simplifies matters considerably, and is anyway not a major restriction because we can just perform a similar exercise as we did in Figures 5 and 12, by varying b and assessing the density forecasting criteria. The appendix derives the iterative scheme for the EM algorithm. It requires repeated evaluation of the Bessel function $K_z(\cdot)$, which turns out to be its most time-consuming aspect. This is addressed by developing a method of its computation for the relevant set of values of z which is about 10 times faster than the existing methods; see the Appendix.

7.2 *Multivariate Laplace mixture distribution*

We say the d-dimensional random variable $\mathbf{Y}$ follows a k-component mixture of multivariate Laplace distributions, or Mix_kLap_d, if its distribution is given by

$$f_{\text{Mix}_k\text{Lap}_d}(\mathbf{y}; \boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \mathbf{b}) = \sum_{j=1}^{k} \lambda_j f_{\text{Lap}}(\mathbf{y}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, b_j), \quad \lambda_j \in (0,1), \quad \sum_{j=1}^{k} \lambda_j = 1, \tag{22}$$

with f_{Lap} denoting the d-variate multivariate Laplace distribution given in Equation (20), and where, similar to our notation for the $\text{Mix}_k N_d$ distribution in Section 3,

$$\boldsymbol{M} = [\boldsymbol{\mu}_1|\boldsymbol{\mu}_2|\cdots|\boldsymbol{\mu}_k], \quad \boldsymbol{\mu}_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{dj})', \quad \boldsymbol{\Psi} = [\boldsymbol{\Sigma}_1|\boldsymbol{\Sigma}_2|\cdots|\boldsymbol{\Sigma}_k],$$

$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)'$, and $\mathbf{b} = (b_1, \ldots, b_k)'$. We observe the set of d-variate random variables $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,d})'$, $t = 1, \ldots, T$, with $\mathbf{Y}_t \overset{\text{iid}}{\sim} \text{Mix}_k\text{Lap}_d(\boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \mathbf{b})$, and take the values of $\mathbf{b} = (b_1, \ldots, b_k)$ to be known constants (see below). Interest centers on estimation of the remaining parameters,

$$\boldsymbol{\theta} = [\text{vec}(\boldsymbol{M})\text{vech}(\boldsymbol{\Sigma}_1)'\text{vech}(\boldsymbol{\Sigma}_2)' \cdots \text{vech}(\boldsymbol{\Sigma}_k)'\boldsymbol{\lambda}']'. \tag{23}$$

This is conducted via an EM algorithm, the derivation of which is given in the Appendix. In addition, the EM recursions are extended there to support use of a quasi-Bayesian paradigm analogous to that used for the $\text{Mix}_k N_d$ distribution in Section 3.

What remains is to determine the choice of parameter $\mathbf{b}$ and the amount of shrinkage. These are done in the next two subsections. Based on their optimal values, the forecasting exercise conducted above under the mixed normal model is conducted using the mixed Laplace, with the results already having been shown in the right panel of Figure 11.

7.2.1 Estimation of Parameter $\boldsymbol{b}$

There are (at least) three ways of doing this. The first way is to augment the EM algorithm such that, at each iteration, a new step maximizes the likelihood with respect to the k unknown values in $\mathbf{b}$, conditional on the other parameters, similar to the way the EM algorithm for a multivariate Student's t, or mixtures of them, can be augmented to estimate the unknown degrees of freedom parameter (McLachlan and Peel 2000, Section 7.5). While possible, this step requires use of generic optimization routines, and will thus substantially reduce the speed of estimation. The second way is to use the EM algorithm as we present it, namely with a fixed value of $\mathbf{b}$, and repeatedly conduct this for various $\mathbf{b}$ (realistically using a generic optimization routine), until the optimal value is found. In other words, we maximize the profile log-likelihood in $\mathbf{b}$. This is equivalent to the first method mentioned – and equally time-consuming, though possibly with better numerical properties. We do not investigate this in detail, instead using the third way, which

is the fastest, but is *not* equivalent to the previous two methods, and is most certainly inferior from an estimation efficiency point of view, but is adequate for our purposes.

This third way entails splitting the data into two groups (discussed below), similar to what we did in Section 4 with the MVN mixture, and then examining the profile log-likelihood of the single component multivariate Laplace distribution in b_i, $i = 1, 2$. The benefit of this is that the EM algorithm for the single component multivariate Laplace distribution with known b is very fast, and only a univariate optimum needs to be located. The downside is that the split into two data sets is imperfect – there is loss of information and introduction of bias, and the resulting chosen b_i values will not be optimal, and not share the asymptotic properties of the MLE. However, the uncertainty (statistical error) involved in the estimated b_i is substantial, and so the impact of using this much faster method is minimal. Moreover, forecasting exercises using a range of b_i values indicate that our method is adequate – no substantial gain in forecast quality is achieved by choosing alternative values of the b_i, while choosing values which are far from those we select result in inferior forecasts.

To conduct the split, we proceed as follows: (i) arbitrary values of b_1 and b_2 are chosen; (ii) the EM algorithm for the two-component multivariate Laplace distribution is run; (iii) those observations for which $H_{t,1} > 0.99$ are assigned to component 1, otherwise to component 2. (The value of 0.99 of course represents a tuning parameter, which will affect the final results. Several reasonable values were tried, and the results were not sensitive to its choice.) Finally, (iv), single component multivariate Laplace distributions, for each of the two components, and over a grid of b_1 and b_2 values, are estimated; and (v), the optimal values of the b_i are obtained from the two profile log-likelihoods. These optimal values of the b_i are then used as the fixed values instead of those chosen in step (i), and the iterative method is conducted again, starting with step (ii). This could be done 'until convergence', but we used just two iterations.

Figure 15 depicts the profile log-likelihood (divided by 10,000) of b_1 and b_2, for each of the two components (which have, respectably, 1312 and 633 observations). The estimations also use a shrinkage prior with weight $\omega = 50$, to be discussed subsequently. The maximum is approximately $\hat{\mathbf{b}} = (7.5, 3.5)$. From the plots, it is obvious that the sampling error associated with the b_i is far higher in the first component, with little difference being achieved for values of $b_1 > 5$. Recall that, as b increases, the multivariate Laplace approaches the normal. Thus, we have further

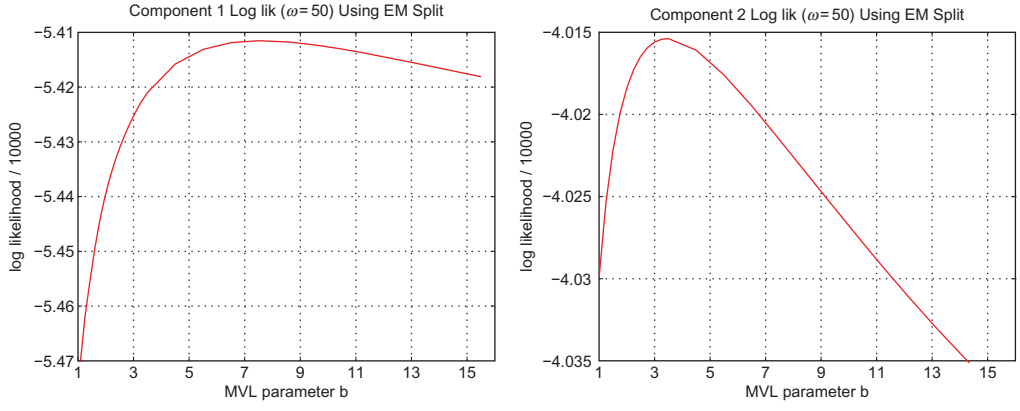

Figure 15. Profile log-likelihood of the fitted single-component multivariate Laplace distribution to each of the two components, decomposed using the EM algorithm for the two-component multivariate Laplace distribution, as a function of the parameter b, for the DJ-30 data.

evidence which augments the results in Figure 3 that the first component does not differ greatly from normality. The profile log-likelihood of b_2 is far more peaked, with a maximum around 3, confirming our expectations that the second component has higher kurtosis than the normal.

If instead of using all the available data, we use only the last 500 observations (so that the components have, respectably, 321 and 179 observations), we find the optimal values of b_1 and b_2 to be each 5.5. This two year period of observations (March 2007 to March 2009) occurs during a full unfolding of the recent liquidity crisis, massive market downturn and relatively enormous volatility, and so it is not surprising that even the first component is picking up non-normality of the data. This limited exercise also provides evidence that it might be valuable to allow the b_i to change through time, an idea which we do not pursue here.

7.2.2 Determining the shrinkage amount

We now turn to determining the optimal value of ω for the shrinkage prior for the two-component multivariate Laplace mixture distribution, using Equation (12) versus ω, for a moving window of length $v = 250$. (This was done for the two sets of **b**-values discussed in the previous section, (7.5, 3.5) and (5.5, 5.5); they lead to the same conclusions.) In addition to the prior structure given in Equation (A27) in the appendix detailing the EM algorithm for its estimation and the quasi-Bayesian extension, we entertain two other ones, with different weights on the covariance matrices. All three are as follows:

$$\begin{aligned} &\text{prior 1:} \quad a_1 = 2\omega, \quad a_2 = \omega/2, \quad c_1 = c_2 = 20\omega, \\ &\text{prior 2:} \quad a_1 = 1\omega, \quad a_2 = 1\omega, \quad c_1 = c_2 = 20\omega, \\ &\text{prior 3:} \quad a_1 = 1.5\omega, \quad a_2 = 1.5\omega, \quad c_1 = c_2 = 20\omega. \end{aligned} \tag{24}$$

The left panel of Figure 16 is similar to the right panel of Figure 5, and shows (12), having used the three priors, and a grid of ω-values. From the figure, we see that priors 2 and 3 not only yield better forecasts for all ω in the relevant range, but, particularly for prior 3, their plots are also much smoother, indicating more stable and reliable estimation. The optimal amount of shrinkage for priors 2 and 3 are $\omega^*(250) = 55$ and $\omega^*(250) = 40$, respectively, and yield nearly the same normalized sum of the realized predictive log-likelihood. In what follows, we use prior 3 and $\omega = 40$.

Observe that the improvement as ω increases from 0.1 to 40 is far less than obtained with the normal distribution. This makes perfect sense: The misspecified normal distribution is such that its tails are too thin for the data, and so its MLE without shrinkage attempts to compensate by increasing the variance parameters. Furthermore, under a Gaussian assumption, the MLE of the mean is the sample average, but we know this is not optimal for fat-tailed distributions, so that it will tend to be unduly influenced by outliers – which shrinkage is admirably able to counteract. The Laplace distribution is surely also misspecified, but significantly less so than the normal, so that less 'tinkering' via shrinkage is required.

The method of weighted likelihood via (16) is applicable. The solid horizontal line at the top of the left panel of Figure 16 shows the result; it is based on weighted likelihood applied just to the $\hat{\mathbf{\Sigma}}_j$, with parameter $\rho = 0.75$ (and using prior 3 from Equation (24) with $\omega = 40$). The right panel of Figure 16 is similar to that in Figure 12, showing the normalized sum of the realized predictive log-likelihood $S_{\tau_0,T}(\mathcal{M}_{\nu_\lambda}, v)$, for $v = \tau_0 = 250$, as a function of ν_λ, where $\mathcal{M}_{\nu_\lambda}$ is the $\text{Mix}_2\text{Lap}_{30}$ model with $\hat{\lambda}$ replaced by $\hat{\lambda}_{\nu,\nu_\lambda}$ (and having used prior 3 given in Equation (24), with $\omega = 40$). Interestingly, just as with the Mix_2N_{30} model, we again obtain nearly monotone gains in forecast accuracy as ν_λ is decreased, and a maximum is reached at about $\nu_\lambda = 70$. Also, again

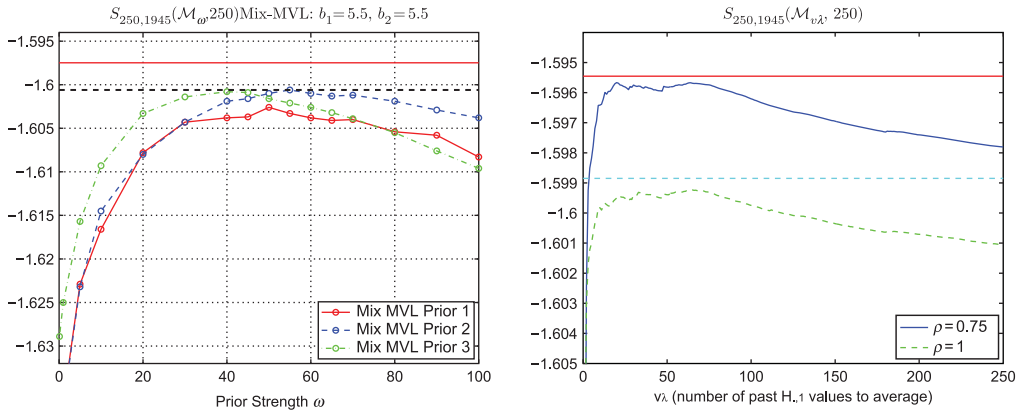

Figure 16. Left: (12) versus ω, based on the two-component multivariate Laplace mixture distribution, using a moving window of length $v = 250$, for $b_1 = b_2 = 5.5$, and three prior structures, as given in Equation (24). The horizontal solid line at the top of the graph shows the value obtained based on prior 3, $\omega = 40$, when using weighted likelihood, with $\rho = 0.75$ and applied to just the $\mathbf{\Sigma}_j$. Right: This is an analog to the right panel of Figure 12: With $\mathcal{M}_{v_\lambda}$ the $\text{Mix}_2\text{Lap}_{30}$ model with $\hat{\lambda}$ replaced by $\hat{\lambda}_{v,v_\lambda}$ the figure shows the normalized sum of the realized predictive log-likelihood versus v_λ, based on prior 3 for $\omega = 40$, with two weighted likelihood values $\rho = 1$ (corresponding to equal weights) and $\rho = 0.75$, applied to just the $\mathbf{\Sigma}_j$. The two horizontal lines are the result of taking $\hat{\lambda}$ to be $0.75\hat{\lambda}_{250,70} + 0.25\hat{\lambda}_{250,1}$, for each of the two values of ρ.

taking $\hat{\lambda}$ to be $0.75\hat{\lambda}_{250,70} + 0.25\hat{\lambda}_{250,1}$ results in further (albeit minor) improvement, as indicated by the horizontal line.

7.3 *Portfolio construction*

This section parallels Section 3.2. Let $\mathbf{L} \sim \text{Lap}(\boldsymbol{\mu}, \mathbf{\Sigma}, b)$ with density (20). Then, for $\mathbf{a} \in \mathbb{R}^d$, $P = \mathbf{a}'\mathbf{L} \sim \text{Lap}(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\mathbf{\Sigma}\mathbf{a}', b)$, which is a special case of the general result for normal mixture distributions, as shown in, e.g. McNeil, Frey, and Embrechts (2005, 76). Now let $\mathbf{Y} \sim \text{Mix}_k\text{Lap}_d(\boldsymbol{M}, \mathbf{\Psi}, \boldsymbol{\lambda}, \mathbf{b})$, with density (22), for $\boldsymbol{M} = [\boldsymbol{\mu}_1|\boldsymbol{\mu}_2|\cdots|\boldsymbol{\mu}_k]$, $\mathbf{\Psi} = [\mathbf{\Sigma}_1|\mathbf{\Sigma}_2|\cdots|\mathbf{\Sigma}_k]$, $\mathbf{b} = (b_1, \ldots, b_k)'$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$. Let $P = \mathbf{a}'\mathbf{Y}$. Then, analogous to result (6) as proved in the appendix, and using the same format of proof, we find that

$$f_P(x) = \sum_{c=1}^{k} \lambda_c \text{Lap}(x; \mathbf{a}'\boldsymbol{\mu}_c, \mathbf{a}'\mathbf{\Sigma}_c\mathbf{a}, b_c). \tag{25}$$

Similar to Equation (7), and in light of Equation (21), we have, with $\mu_c = \mathbf{a}'\boldsymbol{\mu}_c$ and $\sigma_c^2 = \mathbf{a}'\mathbf{\Sigma}_c\mathbf{a}$, $c = 1, \ldots, k$,

$$\mu_P = \mathbb{E}[P] = \sum_{c=1}^{k} \lambda_c \mu_c, \quad \sigma_P^2 = \mathbb{V}(P) = \sum_{c=1}^{k} \lambda_c (b_c\sigma_c^2 + \mu_c^2) - \mu_P^2. \tag{26}$$

Similar to Equation (9), and denoting the density and distribution function of the Laplace distribution as f and F, respectively, calculation shows that

$$\text{ES}_\gamma = \sum_{j=1}^{k} \frac{\lambda_j F(c_j, 0, 1, b_j)}{\gamma} \left(\mu_j - \sigma_j b_j \frac{f(c_j, 0, 1, b_j + 1)}{F(c_j, 0, 1, b_j)} \right), \quad c_j = \frac{Q_P(\gamma) - \mu_j}{\sigma_j}. \tag{27}$$

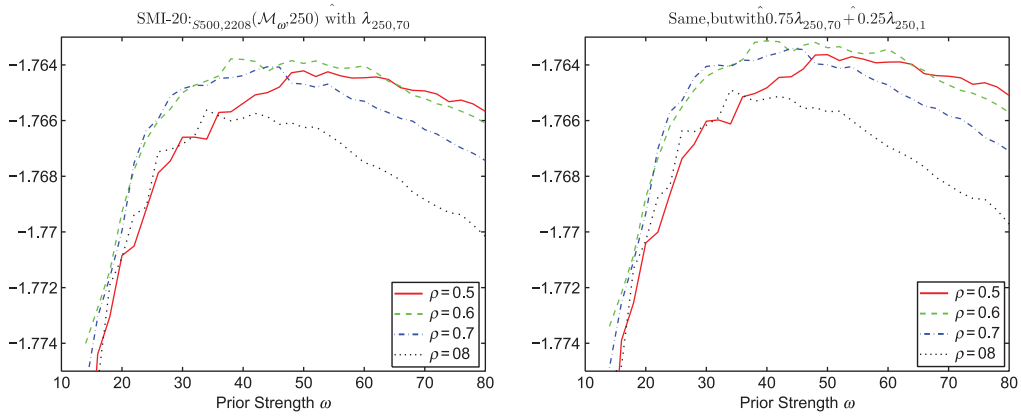

Figure 17. Both panels show (12) for the two-component mixed normal model based on the 2208 daily returns for 20 stocks in the SMI, from 10 November 2000 to 31 August 2009, as a function of shrinkage hyper-parameter ω, for four values of weighted likelihood parameter ρ (applied just to the $\hat{\mathbf{\Sigma}}_j$), and based on moving windows of length $\nu = 250$, with $\tau_0 = 500$. The left uses a moving average of the estimated component weight λ $\hat{\lambda}_{\nu,\nu_\lambda}$ from Equation (17) with $\nu_\lambda = 70$. The right panel is similar, but uses Equation (18).

8. Further data sets

So far, only the DJ-30 data, and several of its subsets, were investigated. More confidence in the new procedure is gained if it is applied, and demonstrated to be effective, with an entirely different data set. Figure 17 is similar to those above but based on $d = 20$ stocks from the Swiss Market Index (SMI). We see that the optimal value of weighted likelihood parameter ρ (applied just to the $\hat{\mathbf{\Sigma}}_j$) is between 0.6 and 0.7, precisely in agreement with the corresponding value for the DJ-30 data, as shown in Figure 12, and that use of the moving average (18) is superior to use of $\hat{\lambda}_{\nu,\nu_\lambda}$ from Equation (17) with $\nu_\lambda = 70$. Finally, $\omega^*(25) = 38$, which, interestingly, is precisely the value obtained from the regression line shown in the left panel of Figure 8, which refers to the DJ-30 data.

Figure 18 shows graphs for the SMI which parallel the ones used for the DJ-30 data. In particular, the top panels are similar to those in Figure 8, showing the performance of the multivariate mixed normal distribution for various subsets of the SMI data. The bottom panels are similar to those in Figure 14, and show the incremental improvement by using weighted likelihood and moving averages of λ and by using the mixture Laplace distribution. As with the DJ-30 data, the relative improvement gained by using the Laplace distribution is substantial.

9. Conclusions

The emphasis in this paper is on fast, simple, numerically reliable methods of generating density forecasts which increase in quality as the number of assets are added, and are competitive with (or outperform, as is the case here) the common MGARCH models which can be estimated with tens or hundreds of assets. This was achieved using mixture models, which, as argued in the Introduction and demonstrated throughout the paper, are useful structures for modeling particular stylized facts of the data. While it is well-known that the distribution of asset returns is fat-tailed and asymmetric, a normal mixture with just two components is capable of capturing most of these

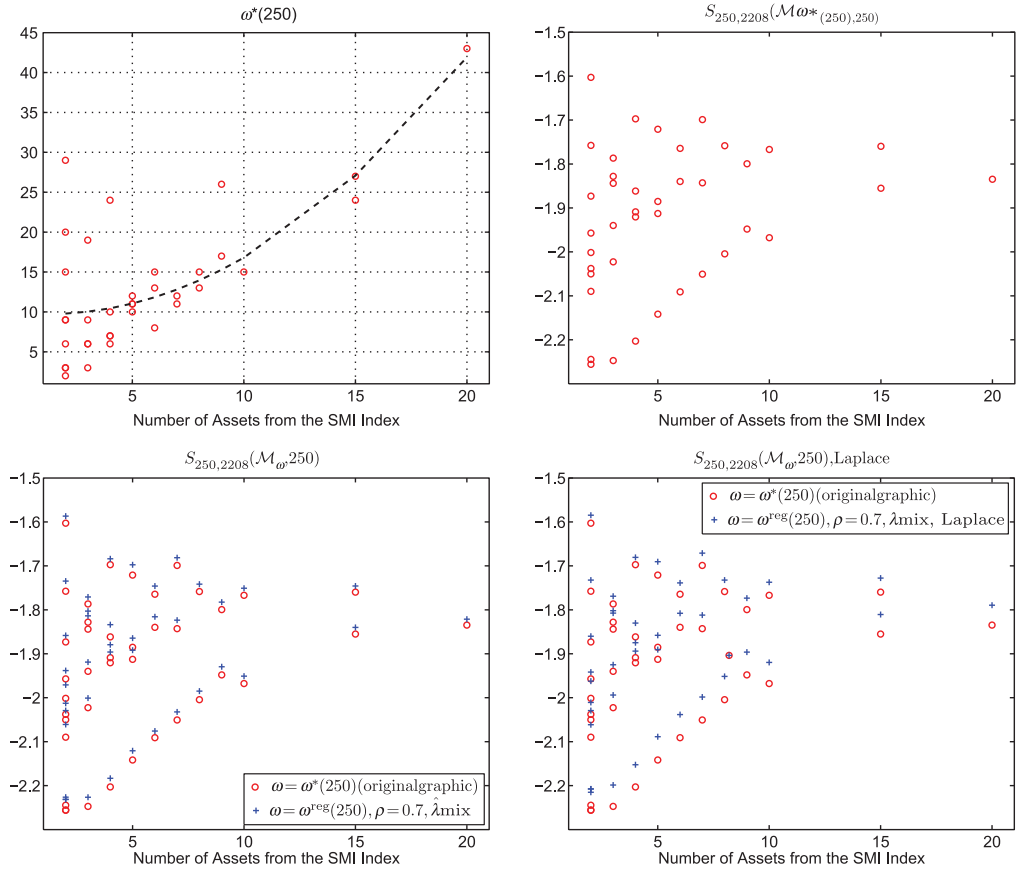

Figure 18. Top panels parallel those in Figure 8, based on the multivariate mixed normal distribution, but using the SMI-20 data (with the two points corresponding to 15 assets refer to stocks 1–15, and 6–20, so that they do contain overlap). The bottom panels are similar to those in Figure 14, but using the SMI data, and showing (left) the incremental improvement by using weighted likelihood and moving averages of λ and, additionally, (right) by using the mixture Laplace distribution.

features, and a Laplace mixture is even better. In both cases, shrinkage and weighted likelihood were found to enhance forecasting performance. Of course, it remains to be seen if good out-of-sample density forecasts translate into superior measures of portfolio performance, a topic to be pursued in future work.

Acknowledgements

Part of this research has been carried out within the National Centre of Competence in Research 'Financial Valuation and Risk Management' (NCCR FINRISK), which is a research program supported by the *Swiss National Science Foundation*. The author thanks Pawel Polak and Maria Putintseva for assistance with programming the RSDC–GARCH and ADCC models, as well as the detailed discussions which led to a significantly improved paper. The paper has also benefited substantially from the extensive comments and suggestions provided by Christian Brownlees, Michael McAleer, Eric Renault and David Veredas on an earlier draft of this paper, as well as those from anonymous referees.

References

Abadir, K. M., and J. R. Magnus. 2005. *Matrix Algebra*. Cambridge: Cambridge University Press.

Acerbi, C., and D. Tasche. 2002. "Expected Shortfall: A Natural Coherent Alternative to Value at Risk." *Economic Notes* 31 (2): 379–388.

Alexander, C., and E. Lazar. 2005. "Asymmetries and Volatility Regimes in the European Equity Market." ICMA Centre Discussion Papers in Finance 2005–14, The Business School for Financial Markets at the University of Reading.

Alexander, C., and E. Lazar. 2006. "Normal Mixture GARCH(1,1): Applications to Exchange Rate Modelling." *Journal of Applied Econometrics* 21 (3): 307–336.

Bauwens, L., C. M. Hafner, and J. V. K. Rombouts. 2007. "Multivariate Mixed Normal Conditional Heteroskedasticity." *Computational Statistics & Data Analysis* 51 (7): 3551–3566.

Bauwens, L., S. Laurent, and J. K. V. Rombouts. 2006. "Multivariate GARCH Models: A Survey." *Journal of Applied Econometrics* 21 (1): 79–109.

Bauwens, L., and J. V. K. Rombouts. 2007. "Bayesian Inference for the Mixed Conditional Heteroskedasticity Model." *Econometrics Journal* 10 (2): 408–425.

Black, F. 1976. "Studies of Stock Price Volatility Changes." Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section, 177–181.

Bollerslev, T. 1990. "Modeling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Approach." *Review of Economics and Statistics* 72 (3): 498–505.

Bollerslev, T., R. F. Engle, and J. M. Wooldridge. 1988. "A Capital Asset Pricing Model with Time-Varying Covariances." *Journal of Political Economy* 96 (1): 116–131.

Broda, S. A., M. Haas, J. Krause, M. S. Paolella, and S. C. Steude. 2013. "Stable Mixture GARCH Models." *Journal of Econometrics* 172 (2): 292–306.

Buckley, I., D. Saunders, and L. Seco. 2008. "Portfolio Optimization When Asset Returns Have the Gaussian Mixture Distribution." *European Journal of Operational Research* 185 (3): 1434–1461.

Caporin, M., and M. McAleer. 2010. "Do We Really Need Both BEKK and DCC? A Tale of Two Multivariate GARCH Models." Available at SSRN.

Cappiello, L., R. F. Engle, and K. Sheppard. 2006. "Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns." *Journal of Financial Econometrics* 4 (4): 537–572.

Chopra, V., and W. Ziemba. 1993. "The Effect of Errors in Means, Variances, and Covariances on Optimal Portfolio Choice." *Journal of Portfolio Management* 19 (2): 6–12.

Diebold, F. X., T. A. Gunther, and A. S. Tay. 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review* 39 (4): 863–883.

Engle, R. F. 2002. "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models." *Journal of Business and Economic Statistics* 20 (3): 339–350.

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

Engle, R. F., and K. F. Kroner. 1995. "Multivariate Simultaneous Generalized ARCH." *Econometric Theory* 11 (1): 122–150.

Fama, E. F. 1965. "The Behavior of Stock Market Prices." *Journal of Business* 38 (1): 34–105.

Giannikis, D., I. D. Vrontos, and P. Dellaportas. 2008. "Modelling Nonlinearities and Heavy Tails via Threshold Normal Mixture GARCH Models." *Computational Statistics & Data Analysis* 52 (3): 1549–1571.

Haas, M., S. Mittnik, and M. S. Paolella. 2004a. "Mixed Normal Conditional Heteroskedasticity." *Journal of Financial Econometrics* 2 (2): 211–250.

Haas, M., S. Mittnik, and M. S. Paolella. 2004b. "A New Approach to Markov Switching GARCH Models." *Journal of Financial Econometrics* 2 (4): 493–530.

Haas, M., S. Mittnik, and M. S. Paolella. 2009. "Asymmetric Multivariate Normal Mixture GARCH." *Computational Statistics & Data Analysis* 53 (6): 2129–2154.

Hamilton, J. D. 1991. "A Quasi-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions." *Journal of Business and Economic Statistics* 9 (1): 21–39.

Harvey, C. R., and A. Siddique. 2000. "Conditional Skewness in Asset Pricing Tests." *The Journal of Finance* 55 (3): 1263–1295.

Ingrassia, S., and R. Rocci. 2007. "Constrained Monotone EM Algorithms for Finite Mixture of Multivariate Gaussians." *Computational Statistics & Data Analysis* 51 (11): 5339–5351.

Kim, T.-H., and H. White. 2004. "On More Robust Estimation of Skewness and Kurtosis." *Finance Research Letters* 1 (1): 56–73.

Kuester, K., S. Mittnik, and M. S. Paolella. 2006. "Value-at-Risk Prediction: A Comparison of Alternative Strategies." *Journal of Financial Econometrics* 4 (1): 53–89.
Ledoit, O., P. Santa-Clara, and M. Wolf. 2003. "Flexible Multivariate GARCH Modeling with an Application to International Stock Markets." *Review of Economics and Statistics* 85 (3): 735–747.
McAleer, M. 2005. "Automated Inference and Learning in Modeling Financial Volatility." *Econometric Theory* 21 (1): 232–261.
McLachlan, G. J., and T. Krishnan. 2008. *The EM Algorithm and Extensions*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
McLachlan, G. J., and D. Peel. 2000. *Finite Mixture Models*. New York: John Wiley & Sons.
McNeil, A. J., R. Frey, and P. Embrechts. 2005. *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton: Princeton University Press.
Mecklin, C. J., and D. J. Mundfrom. 2004. "An Appraisal and Bibliography of Tests for Multivariate Normality." *International Statistical Review* 72 (1): 123–138.
Neftci, S. 2000. "Value at Risk Calculation, Extreme Events, and Tail Estimation." *Journal of Derivatives* 7 (3): 23–37.
Paolella, M. S. 2007. *Intermediate Probability: A Computational Approach*. Chichester: John Wiley & Sons.
Paolella, M. S., and S.-C. Steude. 2008. "Risk Prediction: A DWARF-like Approach." *The Journal of Risk Model Validation* 2 (1): 25–43.
Patton, A. J. 2004. "On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation." *Journal of Financial Econometrics* 2 (1): 130–168.
Pelletier, D. 2006. "Regime Switching for Dynamic Correlations." *Journal of Econometrics* 131 (1–2): 445–473.
Podgórski, T. J., and T. Kozubowski. 2001. "Asymmetric Laplace Laws and Modeling Financial Data." *Mathematical and Computer Modelling* 34 (9–11): 1003–1021.
Rombouts, J. V. K., and L. Stentoft. 2009. "Bayesian Option Pricing Using Mixed Normal Heteroskedasticity Models." CIRPÉE Working Paper 09-26, Institute of Applied Economics, HEC Montréal, Canada.
Rosenblatt, M. 1952. "Remarks on a Multivariate Transformation." *Annals of Mathematical Statistics* 23 (3): 470–472.
Schott, J. R. 2005. *Matrix Analysis for Statistics*. 2nd ed. New York: John Wiley & Sons.
Sheppard, K. 2012. "Forecasting High Dimensional Covariance Matrices." In *Handbook of Volatility Models and Their Applications*, edited by L. Bauwens, C. Hafner, and S. Laurent, 103–127. Hoboken, New Jersey: John Wiley & Sons.
Silvennoinen, A., and T. Teräsvirta. 2009. "Multivariate GARCH Models." In *Handbook of Financial Time Series*, edited by T. G. Andersen, R. A. Davis, J.-P. Kreiß, and T. Mikosch, 201–229. Berlin: Springer Verlag.
Thode Jr., H. C. 2002. *Testing for Normality*. Basel: Marcel Dekker.
Watson, G. N. 1922. *A Treatise on the Theory of Bessel Functions*. Cambridge: Cambridge University Press.
Wu, C., and J. C. Lee. 2007. "Estimation of a Utility-Based Asset Pricing Model Using Normal Mixture GARCH(1, 1)." *Economic Modelling* 24 (2): 329–349.
Yakowitz, S. J., and J. D. Spragins. 1968. "On the Identifiability of Finite Mixtures." *The Annals of Mathematical Statistics* 39 (1): 209–214.

Appendix

Derivation of (6)

Let $\mathbf{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, whose density we denote as $f_{\mathrm{N}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and whose characteristic function (cf) is given by

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}'\mathbf{X})] = \exp(i\mathbf{t}'\boldsymbol{\mu} - \tfrac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}) =: \varphi(\mathbf{t}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{A1}$$

for $\mathbf{t} \in \mathbb{R}^d$. As scalar $S = \mathbf{a}'\mathbf{X} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$ for $\mathbf{a} = (a_1, \ldots, a_d)' \in \mathbb{R}^d$, Equation (A1) implies that

$$\varphi(t; \mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}) = \varphi_S(t) = \mathbb{E}[\exp(it\mathbf{a}'\mathbf{X})] = \int_{\mathbb{R}^d} \exp(it\mathbf{a}'\mathbf{x})\, \mathrm{d}F_{\mathrm{N}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{A2}$$

Recall the inversion formula for the pdf of continuous scalar random variable X is given by $f_X(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} \exp(-itx)\varphi_X(t)\, \mathrm{d}t$. Let $\mathbf{Y} \sim \mathrm{Mix}_k\, N_d(\boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$. With discrete random variable C such that $f_C(c) = \lambda_c$, $\lambda_c \in (0, 1)$, $\sum_{c=1}^{k} \lambda_c = 1$, we can express the mixed normal density as

$$f_{\mathbf{Y}}(\mathbf{y}) = \int f_{\mathbf{Y}|C}(\mathbf{y} \mid c)\, \mathrm{d}F_C(c) = \sum_{c=1}^{k} \lambda_c f_{\mathrm{N}}(\mathbf{y}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c). \tag{A3}$$

Then, from Equations (A1) and (A3),

$$\varphi_{\mathbf{Y}}(\mathbf{t}) = \int_{\mathbb{R}^d} \exp(i\mathbf{t}'\mathbf{y})\,\mathrm{d}F_{\mathbf{Y}}(\mathbf{y}) = \sum_{c=1}^{k} \lambda_c \exp(i\mathbf{t}'\boldsymbol{\mu}_c - \tfrac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}_c\mathbf{t}),$$

and interest centers on the distribution of the portfolio $P = \mathbf{a}'\mathbf{Y}$. Its cf $\varphi_P(t)$ is, from Equation (A2),

$$\begin{aligned}\mathbb{E}[\exp(itP)] &= \int_{\mathbb{R}^d} \exp(it\mathbf{a}'\mathbf{y})\,\mathrm{d}F_{\mathbf{Y}}(\mathbf{y}) = \sum_{c=1}^{k} \lambda_c \int_{\mathbb{R}^d} \exp(it\mathbf{a}'\mathbf{y})\,\mathrm{d}F_{\mathrm{N}}(\mathbf{y};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\\ &= \sum_{c=1}^{k} \lambda_c \varphi(t;\mathbf{a}'\boldsymbol{\mu}_c,\mathbf{a}'\boldsymbol{\Sigma}_c\mathbf{a}),\end{aligned}$$

and applying the inversion theorem gives

$$\begin{aligned}f_P(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)\varphi_P(t)\,\mathrm{d}t = \sum_{c=1}^{k} \lambda_c \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)\varphi(t;\mathbf{a}'\boldsymbol{\mu}_c,\mathbf{a}'\boldsymbol{\Sigma}_c\mathbf{a})\,\mathrm{d}t\\ &= \sum_{c=1}^{k} \lambda_c f_{\mathrm{N}}(x;\mathbf{a}'\boldsymbol{\mu}_c,\mathbf{a}'\boldsymbol{\Sigma}_c\mathbf{a}),\end{aligned}$$

which is Equation (6).

Derivation of (9)

Expression (8) is equivalent to the tail conditional expectation in the continuous case, obtained by substituting $u = Q_P(p)$ into Equation (8) to get, with $q_{P,\gamma} = Q_P(\gamma)$, $\mathrm{ES}_\gamma(P) = \mathbb{E}[X \mid X \le q_{P,\gamma}]$. To derive $\mathrm{ES}_\gamma(P)$, the γ-level ES for random variable P, where $P \sim \mathrm{Mix}_k N_1$ with $f_P(x) = \sum_{c=1}^{k} \lambda_c \phi(x;\mu_c,\sigma_c^2)$, we require the following two, easily verified facts. First, if $Y = \sigma Z + \mu$ for $\sigma > 0$ and $\mathrm{ES}_\gamma(Z)$ exists, then $\mathrm{ES}_\gamma(Y) = \mu + \sigma\mathrm{ES}_\gamma(Z)$. Second, for $R \sim N(0,1)$ with pdf ϕ and cdf Φ, a simple integration shows that

$$\mathrm{ES}_\gamma(R) = \frac{-\phi\{\Phi^{-1}(\gamma)\}}{\gamma}. \tag{A4}$$

Let $q_{P,\gamma}$ be the γ-quantile of P, $X_j \sim N(\mu_j,\sigma_j^2)$, $c_j := (q_{P,\gamma} - \mu_j)/\sigma_j$, $Z \sim N(0,1)$. Using substitution $z = (x-\mu_j)/\sigma_j$,

$$\begin{aligned}\mathrm{ES}_\gamma(P) &= \frac{1}{\gamma}\int_{-\infty}^{q_{P,\gamma}} x f_P(x)\,\mathrm{d}x = \frac{1}{\gamma}\sum_{j=1}^{k}\lambda_j \int_{-\infty}^{q_{P,\gamma}} x\,\sigma_j^{-1} f_Z\left(\frac{x-\mu_j}{\sigma_j}\right)\mathrm{d}x\\ &= \frac{1}{\gamma}\sum_{j=1}^{k}\lambda_j \int_{-\infty}^{(q_{P,\gamma}-\mu_j)/\sigma_j} (\sigma_j z + \mu_j)\sigma_j^{-1} f_Z(z)\sigma_j\,\mathrm{d}z\\ &= \frac{1}{\gamma}\sum_{j=1}^{k}\lambda_j \left[\sigma_j \int_{-\infty}^{c_j} z f_Z(z)\,\mathrm{d}z + \mu_j \int_{-\infty}^{c_j} f_Z(z)\,\mathrm{d}z\right].\end{aligned} \tag{A5}$$

Using Equations (A4) and (A5), we obtain

$$\mathrm{ES}_\gamma(P) = \sum_{j=1}^{k} \frac{\lambda_j \Phi(c_j)}{\gamma}\left\{\mu_j - \sigma_j \frac{\phi(c_j)}{\Phi(c_j)}\right\}, \quad c_j = \frac{q_{P,\gamma}-\mu_j}{\sigma_j},$$

which is Equation (9).

Derivation of (20)

With $(\mathbf{Y} \mid G = g) \sim N(0, g\boldsymbol{\Sigma})$ and $G \sim \text{Gam}(b, 1)$, recalling that $|g\boldsymbol{\Sigma}| = g^d|\boldsymbol{\Sigma}|$, and setting $m = \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$ and, as $m \geq 0$, $u = \sqrt{m/2}/g$, the density function $f_{\mathbf{Y}}(\mathbf{y}, \boldsymbol{\Sigma}, b) = \int f_{\mathbf{Y}|G}(\mathbf{y}; g) f_G(g)\, \mathrm{d}g$ is given by

$$\begin{aligned}
&\int_0^\infty \frac{1}{|g\boldsymbol{\Sigma}|^{1/2}(2\pi)^{d/2}} \exp\left\{-\frac{1}{2}\mathbf{y}'(g\boldsymbol{\Sigma})^{-1}\mathbf{y}\right\} \frac{1}{\Gamma(b)} g^{b-1}\exp(-g)\,\mathrm{d}g \\
&= \frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{d/2}} \frac{1}{\Gamma(b)} \int_0^\infty g^{-d/2+b-1} \exp\left\{-\frac{m}{2g} - g\right\} \mathrm{d}g \\
&= \frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{d/2}} \frac{2}{\Gamma(b)} \left(\frac{m}{2}\right)^{b/2-d/4} \frac{1}{2} \int_0^\infty u^{d/2-b-1} \exp\left\{-\frac{\sqrt{2m}}{2}(u + u^{-1})\right\} \mathrm{d}u \\
&= \frac{1}{|\boldsymbol{\Sigma}|^{1/2}(2\pi)^{d/2}} \frac{2}{\Gamma(b)} \left(\frac{m}{2}\right)^{b/2-d/4} K_{b-d/2}(\sqrt{2m}),
\end{aligned} \tag{A6}$$

which is equivalent to Equation (20), recalling that $K_\nu(x) = K_{-\nu}(x)$.

EM algorithm for estimation of (20)

The log joint density of $\mathbf{Y}_t$ and G_t is, from Equation (A6), and omitting constants which do not depend on $\boldsymbol{\theta} = [\boldsymbol{\mu}'\ \text{vech}(\boldsymbol{\Sigma})']'$, and multiplying by two,

$$\begin{aligned}
\log f_{\mathbf{Y}_t, G_t}(\mathbf{y}_t, g_t; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \log f_{\mathbf{Y}_t|G_t}(\mathbf{y}_t; g_t) + \log f_{G_t}(g_t) \\
&\propto -\log|\boldsymbol{\Sigma}| - g_t^{-1}(\mathbf{y}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}),
\end{aligned}$$

so that the complete data log-likelihood is, with $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$, $\mathbf{G} = (G_1, \ldots, G_T)$, and using the fact that $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$,

$$\ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{G}) = -T\log|\boldsymbol{\Sigma}| - \sum_{t=1}^{T} G_t^{-1}\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_t - \boldsymbol{\mu})(\mathbf{Y}_t - \boldsymbol{\mu})'. \tag{A7}$$

The derivation of the MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is a well-known exercise involving matrix differentiation; see, e.g. Schott (2005, 373–4) or Abadir and Magnus (2005, 387–9). This yields

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{t=1}^{T} G_t^{-1}\mathbf{Y}_t}{\sum_{t=1}^{T} G_t^{-1}}, \quad \hat{\boldsymbol{\Sigma}} = T^{-1}\sum_{t=1}^{\mathrm{T}} G_t^{-1}(\mathbf{Y}_t - \hat{\boldsymbol{\mu}})(\mathbf{Y}_t - \hat{\boldsymbol{\mu}})'. \tag{A8}$$

Next, with $m_t = (\mathbf{y}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})$,

$$f_{G_t|\mathbf{Y}_t}(g_t; \mathbf{y}_t) = \frac{f_{\mathbf{Y}_t, G_t}(\mathbf{y}_t, g_t; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{f_{\mathbf{Y}_t}(\mathbf{y}_t; \boldsymbol{\mu}, \boldsymbol{\Sigma})} = \frac{g_t^{-d/2+b-1}\exp\{-m_t/(2g_t) - g_t\}}{2(m_t/2)^{b/2-d/4}K_{b-d/2}(\sqrt{2m_t})}\mathbb{I}_{(0,\infty)}(g_t). \tag{A9}$$

This is the generalized inverse Gaussian distribution (Paolella 2007, Section 9.4.1), with $G_t \mid \mathbf{Y}_t \sim \text{GIG}(b - d/2, m_t, 2)$, and (Paolella 2007, Eq. 9.18)

$$\mathbb{E}[G_t^{-1} \mid \mathbf{y}_t] = \frac{K_{b-d/2-1}(\sqrt{2m_t})}{(m_t/2)^{1/2}K_{b-d/2}(\sqrt{2m_t})}, \tag{A10}$$

which, for $b = (d+1)/2$, simplifies to $(m_t/2)^{-1/2}$. As g_t^{-1} enters linearly in Equation (A7), the conditional expectation of the complete data log-likelihood, with respect to the hidden variables $\mathbf{G}$, given the observed data $\mathbf{Y}$, and the value of parameter $\boldsymbol{\theta}$ at the sth iteration, $\mathbb{E}_{\boldsymbol{\theta}^{(s)}}[\ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{G}) \mid \mathbf{Y}]$ just involves substituting Equation (A10) in place of g_t^{-1} in Equation (A7). The EM algorithm then consists of iterating between Equations (A8) and (A10) until convergence.

Analogous to the Hamilton (1991) estimators for the $\text{Mix}_k N_d$ model given in Equations (3) and (4), and placing subscripts on the parameters, anticipating our use of mixtures of Laplace, we augment Equation (A8) by taking

$$\hat{\boldsymbol{\mu}}_j = \frac{c_j \mathbf{m}_j + \sum_{t=1}^{T} G_t^{-1} \mathbf{Y}_t}{c_j + \sum_{t=1}^{T} G_t^{-1}}, \tag{A11}$$

and

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\mathbf{B}_j + \sum_{t=1}^{T} G_t^{-1} (\mathbf{Y}_t - \boldsymbol{\mu}_j)(\mathbf{Y}_t - \boldsymbol{\mu}_j)' + c_j (\mathbf{m}_j - \hat{\boldsymbol{\mu}}_j)(\mathbf{m}_j - \hat{\boldsymbol{\mu}}_j)'}{a_j + T}, \tag{A12}$$

$j = 1, \ldots, k$, where, as with the $\text{Mix}_k N_d$ prior, $\mathbf{m}_j$ is the location prior for the jth component. Taking $k = 1$ results in the quasi-Bayesian prior for the single component multivariate Laplace.

Fast Bessel function evaluation

An asymptotic expansion of $K_\nu(z)$ as given in Watson (1922, 202) is

$$K_\nu(z) = \sqrt{\frac{\pi}{2z}} e^{-z} \times E(\nu, z), \quad \text{where} \tag{A13}$$

$$E(\nu, z) = 1 + \frac{4\nu^2 - 1^2}{1!8z} + \frac{(4\nu^2 - 1^2)(4\nu^2 - 3^2)}{2!(8z)^2} + \tag{A14}$$

so that $E(\frac{1}{2}, z) = 1$, $E(\frac{3}{2}, z) = (1 + 1/z)$, $E(\frac{5}{2}, z) = (1 + 3/z + 3/z^2)$, $E(\frac{5}{2}, z) = 1 + 6/z + 15/z^2 + 15/z^3$, and, in general, after a bit of trial and error, we find, for $\nu = k + \frac{1}{2}$, $k \in \mathbb{N}$,

$$E(\nu, z) = 1 + \sum_{i=1}^{k} \frac{(k+i)!}{2^i (k-i)! i!} \frac{1}{s^i}. \tag{A15}$$

This is about 10 times faster to evaluate than Matlab's implementation, but only when having pre-computed the coefficients in Equation (A15). This code is available upon request from the author. Using Equation (A15) thus yields a near 10-fold increase in the (already fast) EM algorithm for the estimation for those values of b for which it is applicable; these are, for $d = 30$, $b = 3.5, 5.5, \ldots, 27.5$; and for $d = 4$, $b = 0.5, 1.5, \ldots, 14.5$. These ranges are obtained as follows. Recalling (i) that $K_\nu(x) = K_{-\nu}(x)$ (the proof of which is trivial), (ii) that b must be positive (from the mixture construction of the multivariate Laplace distribution), (iii) that we need to evaluate the Bessel functions occurring in Equations (20) and (A10), and (iv) that, in our function we pre-computed the coefficients just for k in Equation (A15) only up to and including 12, we require, for d even, $b - \frac{1}{2}$ (and, for d odd, b) to be a non-negative integer such that $|b - d/2 - 1| - \frac{1}{2}$ and $|b - d/2| - \frac{1}{2}$ are both less than or equal to 12, for d even.

EM Algorithm for $\mathbf{Mix}_k\mathbf{Lap}_d$

As with the development of the EM algorithm for the $\text{Mix}_k N_d$ distribution, denote the hidden variable associated with the tth observation $\mathbf{Y}_t$ as $\mathbf{H}_t = (H_{t,1}, \ldots, H_{t,k})$, where $H_{t,j} = 1$ if $\mathbf{Y}_t$ came from the jth component, and zero otherwise, so that

$$f_{\mathbf{H}_t}(\mathbf{h}_t) = \prod_{j=1}^{k} \lambda_j^{h_{t,j}} \mathbb{I}_{\{0,1\}}(h_{t,j}) \mathbb{I}\left(\sum_{j=1}^{k} h_{t,j} = 1 \right). \tag{A16}$$

Then, analogous to the $\text{Mix}_k N_d$ case, the joint density of $\mathbf{Y}_t$ and $\mathbf{H}_t$ is

$$f_{\mathbf{Y}_t|\mathbf{H}_t}(\mathbf{y}, \mathbf{h}_t; \boldsymbol{\theta}) f_{\mathbf{H}_t}(\mathbf{h}_t; \boldsymbol{\theta}) = \prod_{j=1}^{k} [\lambda_j f_{\text{Lap}}(\mathbf{y}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, b_j)]^{h_{t,j}} \mathbb{I}_{\{0,1\}}(h_{t,j}) \mathbb{I}\left(\sum_{j=1}^{k} h_{t,j} = 1 \right),$$

where $\boldsymbol{\theta}$ is given in Equation (23); and $\mathbb{E}[H_{t,j} \mid \mathbf{Y}_t] = \Pr(H_{t,j} = 1 \mid \mathbf{Y}_t = \mathbf{y}_t)$ is

$$\hbar_{t,j} := \frac{\lambda_j f_{\text{Lap}}(\mathbf{y}_t; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, b_j)}{\sum_{j=1}^{k} \lambda_j f_{\text{Lap}}(\mathbf{y}_t; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, b_j)}, \quad t = 1, \ldots, T, \;\; j = 1, \ldots, k. \tag{A17}$$

Recall that, from the construction of the single component multivariate Laplace distribution, $(\mathbf{Y}_t \mid G_t = g_t) \sim N(0, g_t \boldsymbol{\Sigma})$, where $G_t \sim \text{Gam}(b, 1)$, and $\mathbf{G} = (G_1, \ldots, G_T)$ was part of the complete data log-likelihood. In the mixture context, G_t can come from one of k distributions, $\text{Gam}(b_j, 1)$, with pdf $f_{\text{Gam}}(g_t; b_j) = g_t^{b_j} \exp(-g_t)/\,\Gamma(b_j)\mathbb{I}_{(0,\infty)}(g_t), j = 1, \ldots, k$, so its specification requires conditioning on $\mathbf{H}_t$. That is,

$$(G_t \mid \mathbf{H}_t) = (G_t \mid H_{t,j} = 1) \sim \text{Gam}(b_j, 1) \quad \text{and} \quad (\mathbf{Y}_t \mid H_{t,j} = 1, G_t = g_t) \sim N(0, g_t \boldsymbol{\Sigma}).$$

Then $f_{\mathbf{Y}_t, G_t, \mathbf{H}_t}(\mathbf{y}_t, g_t, \mathbf{h}_t; \boldsymbol{\theta}) = f_{\mathbf{Y}_t | G_t, \mathbf{H}_t}(\mathbf{y}_t; g_t, \mathbf{h}_t, \boldsymbol{\theta}) f_{G_t | \mathbf{H}_t}(g_t; \mathbf{h}_t, \boldsymbol{\theta}) f_{\mathbf{H}_t}(\mathbf{h}_t; \boldsymbol{\theta})$, where, noting that conditional on $\mathbf{H}_t$ being $\mathbf{h}_t$ with jth element one,

$$f_{\mathbf{Y}_t | G_t, \mathbf{H}_t}(\mathbf{y}, g_t, \mathbf{h}_t; \boldsymbol{\theta}) = \prod_{j=1}^{k} [f_N(\mathbf{y}; \boldsymbol{\mu}_j, g_t \boldsymbol{\Sigma}_j)]^{h_{t,j}}. \tag{A18}$$

Similarly,

$$f_{G_t | \mathbf{H}_t}(g_t; \mathbf{h}_t) = \prod_{j=1}^{k} [f_{\text{Gam}}(g_t; b_j)]^{h_{t,j}}, \tag{A19}$$

so that, from Equations (A16), (A18) and (A19),

$$\begin{aligned} f_{\mathbf{Y}_t, G_t, \mathbf{H}_t}(\mathbf{y}_t, g_t, \mathbf{h}_t; \boldsymbol{\theta}) &= f_{\mathbf{Y}_t | G_t, \mathbf{H}_t}(\mathbf{y}_t; g_t, \mathbf{h}_t, \boldsymbol{\theta}) f_{G_t | \mathbf{H}_t}(g_t; \mathbf{h}_t) f_{\mathbf{H}_t}(\mathbf{h}_t; \boldsymbol{\theta}) \\ &= \prod_{j=1}^{k} [\lambda_j f_N(\mathbf{y}; \boldsymbol{\mu}_j, g_t \boldsymbol{\Sigma}_j) f_{\text{Gam}}(g_t; b_j)]^{h_{t,j} \mathbb{I}_{\{0,1\}}(h_{t,j}) \mathbb{I}\left(\sum_{j=1}^{k} h_{t,j} = 1\right)}. \end{aligned}$$

The complete data log-likelihood, $\ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{G}, \mathbf{H}) = \sum_{t=1}^{T} \log f_{\mathbf{Y}_t, \mathbf{G}_t, \mathbf{H}_t}(\mathbf{y}_t, g_t, \mathbf{h}_t; \boldsymbol{\theta})$, of $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$, $\mathbf{G}$, and $\mathbf{H} = (\mathbf{H}_1, \ldots, \mathbf{H}_T)$ is, recalling that $|g\boldsymbol{\Sigma}| = g^d |\boldsymbol{\Sigma}|$ for $g \in \mathbb{R}_{>0}$ and $d \in \mathbb{R}$, given by

$$\begin{aligned} &\sum_{t=1}^{T} \log f_{\mathbf{Y}_t | \mathbf{G}_t, \mathbf{H}_t}(\mathbf{y}_t; g_t, \mathbf{h}_t, \boldsymbol{\theta}) + \sum_{t=1}^{T} \log f_{\mathbf{G}_t | \mathbf{H}_t}(g_t; \mathbf{h}_t, \boldsymbol{\theta}) + \sum_{t=1}^{T} \log f_{\mathbf{H}_t}(\mathbf{h}_t; \boldsymbol{\theta}) \\ &\quad = \sum_{t=1}^{T} \sum_{j=1}^{k} h_{t,j} \left\{ -\frac{d}{2} \log(2\pi) - \frac{d}{2} \log(g_t) - \frac{1}{2} \log |\boldsymbol{\Sigma}_j| - \frac{1}{2} (\mathbf{y}_t - \boldsymbol{\mu}_j)' (g_t \boldsymbol{\Sigma}_j)^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_j) \right\} \\ &\qquad + \sum_{t=1}^{T} h_{t,j} \{ b_j \log(g_t) - g_t - \log \Gamma(b_j) \} + \sum_{t=1}^{T} \sum_{j=1}^{k} h_{t,j} \log \lambda_j. \end{aligned}$$

Omitting terms which do not depend on $\boldsymbol{\theta}$ from Equation (23), keeping in mind that the b_j are known, multiplying by two, and (less consequentially) using capital letters for the $\mathbf{Y}_t$, $\mathbf{G}_t$, and $\mathbf{H}_t$ to indicate them as random variables, we can write

$$\begin{aligned} \ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{G}, \mathbf{H}) = &\sum_{t=1}^{T} \sum_{j=1}^{k} H_{t,j} \{ -\log |\boldsymbol{\Sigma}_j| - G_t^{-1} (\mathbf{Y}_t - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{Y}_t - \boldsymbol{\mu}_j) \} \\ &+ 2 \sum_{t=1}^{T} \sum_{j=1}^{k} H_{t,j} \log \lambda_j. \end{aligned} \tag{A20}$$

As in the mixed normal case, the λ_j are disjoint from the other parameters in $\boldsymbol{\theta}$, so that their MLEs can be determined separately, and yield the same result as Equation (2), namely

$$\hat{\lambda}_j = \frac{1}{T}\sum_{t=1}^{T} H_{t,j}, \quad j = 1, \ldots, k. \tag{A21}$$

Estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the complete data log-likelihood follow easily because, firstly, of the binary nature of the $H_{t,j}$ and that $\sum_{j=1}^{k} H_{t,j} = 1$, so that we have k independent multivariate Laplace populations, each with $\sum_{t=1}^{T} H_{t,j}$ observations, and, secondly, we can then apply Equation (A8), i.e.

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{t=1}^{T} H_{t,j} G_t^{-1} \mathbf{Y}_t}{\sum_{t=1}^{T} H_{t,j} G_t^{-1}}, \quad \hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{t=1}^{T} H_{t,j} G_t^{-1} (\mathbf{Y}_t - \hat{\boldsymbol{\mu}}_j)(\mathbf{Y}_t - \hat{\boldsymbol{\mu}}_j)'}{\sum_{t=1}^{T} H_{t,j}}, \tag{A22}$$

$j = 1, \ldots, k$. As a direct generalization of the quasi-Bayesian estimator in Equations (3) and (4) for the $\text{Mix}_k N_d$ model, and Equations (A11) and (A12) for the single-component multivariate Laplace, we take

$$\hat{\boldsymbol{\mu}}_j = \frac{c_j \mathbf{m}_j + \sum_{t=1}^{T} H_{t,j} G_t^{-1} \mathbf{Y}_t}{c_j + \sum_{t=1}^{T} H_{t,j} G_t^{-1}}, \tag{A23}$$

and

$$\boldsymbol{\Sigma}_j = \frac{\mathbf{B}_j + \sum_{t=1}^{T} H_{t,j} G_t^{-1} (\mathbf{Y}_t - \hat{\boldsymbol{\mu}}_j)(\mathbf{Y}_t - \hat{\boldsymbol{\mu}}_j)' + c_j (\mathbf{m}_j - \hat{\boldsymbol{\mu}}_j)(\mathbf{m}_j - \hat{\boldsymbol{\mu}}_j)'}{a_j + \sum_{t=1}^{T} H_{t,j}}, \tag{A24}$$

$j = 1, \ldots, k$, and use the prior values in Equation (A27).

As usual in our EM algorithm derivations, computation of Equations (A21) and (A22) is not feasible because $\mathbf{G}$ and $\mathbf{H}$ are not observed. Hence, the E-step: Given what we do observe, the $\mathbf{Y}_t$, and the value of $\hat{\boldsymbol{\theta}}$, say $\boldsymbol{\theta}^{(s)}$ in the sth step of the iterative scheme, we compute $Ql(\boldsymbol{\theta}; \boldsymbol{\theta}^{(s)} r) = \mathbb{E}_{\boldsymbol{\theta}^{(s)}}[\ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{G}, \mathbf{H}) \mid \mathbf{Y} = \mathbf{y}]$, the conditional expectation of the complete data log-likelihood with respect to the hidden random variables, given the observed data $\mathbf{y}$, and using as parameter $\boldsymbol{\theta}$ the value current value $\boldsymbol{\theta}^{(s)}$. For this, we need $\mathbb{E}_{\boldsymbol{\theta}^{(s)}}[H_{t,j} \mid \mathbf{Y}_t]$, which is just (A17).

Next, for the tth observation but using the parameters of the jth component, let

$$m_{t,j} = (\mathbf{y}_t - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_j), \quad t = 1, \ldots, T, \;\; j = 1, \ldots, k,$$

and recall that, when conditioning on $\mathbf{H}_t = \mathbf{h}_t$, the jth element of $\mathbf{h}_t$ is one (and the rest zero). Then, to compute $\mathbb{E}_{\boldsymbol{\theta}^{(s)}}[G_t^{-1} \mid \mathbf{Y}_t = \mathbf{y}_t, \mathbf{H}_t = \mathbf{h}_t]$, we require $f_{G_t|\mathbf{Y}_t,\mathbf{H}_t}(g_t; \mathbf{y}_t, \mathbf{h}_t)$, or

$$\begin{aligned}
\frac{f_{\mathbf{Y}_t,\mathbf{G}_t,\mathbf{H}_t}(\mathbf{y}_t, g_t, \mathbf{h}_t; \boldsymbol{\theta})}{f_{\mathbf{Y}_t,\mathbf{H}_t}(\mathbf{y}_t, \mathbf{h}_t; \boldsymbol{\theta})} &= \frac{f_{\mathbf{Y}_t|\mathbf{G}_t,\mathbf{H}_t}(\mathbf{y}_t; g_t, \mathbf{h}_t, \boldsymbol{\theta}) f_{\mathbf{G}_t|\mathbf{H}_t}(g_t; \mathbf{h}_t, \boldsymbol{\theta})}{f_{\mathbf{Y}_t|\mathbf{H}_t}(\mathbf{y}, \mathbf{h}_t, \boldsymbol{\theta})} \frac{f_{\mathbf{H}_t}(\mathbf{h}_t; \boldsymbol{\theta})}{f_{\mathbf{H}_t}(\mathbf{h}_t; \boldsymbol{\theta})} \\
&= \frac{\prod_{j=1}^{k} [f_N(\mathbf{y}; \boldsymbol{\mu}_j, g_t \boldsymbol{\Sigma}_j) f_{\text{Gam}}(g_t; b_j)]^{h_{t,j}}}{\prod_{j=1}^{k} [f_{\text{Lap}}(\mathbf{y}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, b_j)]^{h_{t,j}}} = \frac{f_N(\mathbf{y}; \boldsymbol{\mu}_j, g_t \boldsymbol{\Sigma}_j) f_{\text{Gam}}(g_t; b_j)}{f_{\text{Lap}}(\mathbf{y}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, b_j)},
\end{aligned}$$

but this is the same density as given in Equation (A9), so that we can use the result from Equation (A10), just changing b to b_j and m_t to $m_{t,j}$ to get

$$\zeta_{t,j} := \mathbb{E}_{\boldsymbol{\theta}^{(s)}}[G_t^{-1} \mid \mathbf{Y}_t = \mathbf{y}_t, \mathbf{H}_t = \mathbf{h}_t] = \frac{K_{b_j - d/2 - 1}(\sqrt{2m_{t,j}})}{(m_{t,j}/2)^{1/2} K_{b_j - d/2}(\sqrt{2m_{t,j}})}, \tag{A25}$$

$t = 1, \ldots, T, j = 1, \ldots, k$. Then, with $W = (\ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{G}, \mathbf{H}) \mid \mathbf{Y} = \mathbf{y})$, we have $\mathbb{E}[W] = \mathbb{E}[\mathbb{E}[W \mid \mathbf{H}_t]]$, or, using Equations (A20) and (A25),

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}^{(s)}}[\ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{G}, \mathbf{H}) \mid \mathbf{Y} = \mathbf{y}] \\
&\quad = \mathbb{E}_{\boldsymbol{\theta}^{(s)}}[\mathbb{E}_{\boldsymbol{\theta}^{(s)}}[\ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{G}, \mathbf{H}) \mid \mathbf{Y} = \mathbf{y}; \mathbf{H}_t = \mathbf{h}_t] \mid \mathbf{Y} = \mathbf{y}] \\
&\quad = \mathbb{E}_{\boldsymbol{\theta}^{(s)}}\left[\sum_{t=1}^{T}\sum_{j=1}^{k} H_{t,j}\{-\log|\boldsymbol{\Sigma}_j| - \zeta_{t,j}(\mathbf{Y}_t - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{Y}_t - \boldsymbol{\mu}_j)\}\right. \\
&\qquad\qquad \left. + 2\sum_{t=1}^{T}\sum_{j=1}^{k} H_{t,j}\log\lambda_j \mid \mathbf{Y} = \mathbf{y}\right],
\end{aligned} \tag{A26}$$

which follows because $\mathbb{E}_{\boldsymbol{\theta}^{(s)}}[H_{t,j}G_t^{-1} \mid \mathbf{Y}_t, \mathbf{H}_t] = H_{t,j}\mathbb{E}_{\boldsymbol{\theta}^{(s)}}[G_t^{-1} \mid \mathbf{Y}_t, \mathbf{H}_t]$. As Equation (A26) is linear in the $H_{t,j}$, we need only the expectation of the $H_{t,j}$, conditional on $\mathbf{Y} = \mathbf{y}$, and using $\boldsymbol{\theta}^{(s)}$, which is given by Equation (A17).

Similar to Equation (5), and in light of the variance, as given in Equation (21), we take

$$\begin{aligned}
&a_1 = 2\omega, \quad a_2 = \frac{\omega}{2}, \quad c_1 = c_2 = 20\omega, \quad \mathbf{m}_1 = \mathbf{0}_d, \quad \mathbf{m}_2 = -0.1\mathbf{1}_d, \\
&\mathbf{B}_1 = \frac{a_1}{b_1}[(1.5 - 0.6)\mathbf{I}_d + 0.6\mathbf{J}_d], \quad \mathbf{B}_2 = \frac{a_2}{b_2}[(10 - 4.6)\mathbf{I}_d + 4.6\mathbf{J}_d],
\end{aligned} \tag{A27}$$

where b_1 and b_2 are the Laplace parameter b in the density (20).

Thus, with a starting value $\boldsymbol{\theta}^{(0)}$, the EM algorithm iterates between Equations (A17), (A21) and (A22), where it is understood that, in Equation (A17), λ_j, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the most current values from $\boldsymbol{\theta}^{(s)}$, and in Equations (A21) and (A22), the expectations of $H_{t,j}$ and G_t^{-1} are used, these being $\hbar_{t,j}$ in Equation (A17) and $\zeta_{t,j}$ in Equation (A25), respectively.

There are some issues related to computation which apply to both the normal and Laplace cases. During the EM iterations, roundoff error can induce $\hat{\boldsymbol{\Sigma}}_1$ or $\hat{\boldsymbol{\Sigma}}_2$ to deviate slightly from symmetry, which would invalidate their status as covariance matrices. A simple and effective solution is just to set $\hat{\boldsymbol{\Sigma}}_j = (\hat{\boldsymbol{\Sigma}}_j + \hat{\boldsymbol{\Sigma}}_j')/2$, $j = 1, \ldots, k$. Less trivial is the possibility that one or more of the $\hat{\boldsymbol{\Sigma}}_j$ are rank deficient, which is analogous to one or more of the scale terms σ_j in the univariate case approaching zero. It was found that simply setting eigenvalues of $\hat{\boldsymbol{\Sigma}}_j$ lower than some threshold to that threshold during the iteration was enough to prevent the algorithm from ceasing, and allowing it to either find a more plausible solution, or, on occasion, to return a rank-deficient $\hat{\boldsymbol{\Sigma}}_j$. This latter case is preventable with the use of $a_j > 0$, such as $\mathbf{B}_j = \mathbf{I}_d$ and $a_j = 0.1$ in Equation (4), whereby an extremely small amount of prior information is enough to prevent the optimizer from landing on a singularity point of the likelihood. (Another approach for ensuring full rank of the $\hat{\boldsymbol{\Sigma}}_j$ during estimation in the normal mixture case is proposed in Ingrassia and Rocci 2007.)

As with the mixture of normals (both univariate and multivariate), we can safely presume that the likelihood surface of the Mix_kLap_d model (with its 991 parameters for $k = 2$, $d = 30$ and fixed b_1 and b_2) has more than one local maxima. The results shown in this paper have all used the following strategy. Two starting values were used: (i) The values corresponding to the final values of the previous window (except on the first iteration); and (ii) the prior values as given in Equation (A27), i.e. $\boldsymbol{\mu}_i = \mathbf{m}_i$, $\boldsymbol{\Sigma}_i = \mathbf{B}_i/a_i$, $i = 1, 2$. The one returning the higher likelihood value is used.

In about 4% of cases considered for our data exercise involving estimation for each of the 1695 moving windows of length $v = 250$, the EM algorithm failed to converge. In such cases, the starting values are taken to be the prior ones, plus some random noise. This is repeated until the EM algorithm converges. In the few cases in which this was required, one or two further attempts were all that were necessary. As such, fitting the Mix_kLap_d model via the EM algorithm and with a non-zero quasi-Bayesian prior is numerically extremely reliable and, owing to the fast evaluation of the Bessel function developed above, is also nearly as fast as fitting a mixture of normals.