Higher Moment Constraints for Predictive Density Combinations

Laurent Pauwels, Peter Radchenko, and Andrey L. Vasnev
University of Sydney, New South Wales, Australia

November 22, 2018

Abstract: The majority of financial data exhibit asymmetry and heavy tails, which makes forecasting the entire density critically important. Recently, a forecast combination methodology has been developed to combine predictive densities. We show that combining individual predictive densities that are skewed and/or heavy-tailed results in significantly reduced skewness and kurtosis. We propose a solution to overcome this problem by deriving optimal log score weights under Higher-order Moment Constraints (HMC). The statistical properties of these weights are investigated theoretically and through a simulation study. Consistency and asymptotic distribution results for the optimal log score weights with and without high moment constraints are derived. An empirical application that uses the S&P 500 daily index returns illustrates that the proposed HMC weight density combinations perform very well relative to other combination methods.

Keywords: Forecast combination, Predictive densities, Optimal weights, Skewness, Kurtosis.

JEL Codes: C53, C58

March 2019

http://sydney.edu.au/business/business_analytics/research/working_papers
Higher Moment Constraints for Predictive Density Combinations

Laurent Pauwels, Peter Radchenko, and Andrey L. Vasnev

University of Sydney, New South Wales, Australia

November 22, 2018

Abstract:
The majority of financial data exhibit asymmetry and heavy tails, which makes forecasting the entire density critically important. Recently, a forecast combination methodology has been developed to combine predictive densities. We show that combining individual predictive densities that are skewed and/or heavy-tailed results in significantly reduced skewness and kurtosis. We propose a solution to overcome this problem by deriving optimal log score weights under Higher-order Moment Constraints (HMC). The statistical properties of these weights are investigated theoretically and through a simulation study. Consistency and asymptotic distribution results for the optimal log score weights with and without high moment constraints are derived. An empirical application that uses the S&P 500 daily index returns illustrates that the proposed HMC weight density combinations perform very well relative to other combination methods.

Keywords: Forecast combination, Predictive densities, Optimal weights, Skewness, Kurtosis.

JEL Codes: C53, C58

Corresponding author:
Laurent Pauwels
The University of Sydney Business School,
Room 4090, Abercrombie Building (H70),
Sydney, NSW 2006,
Australia
E-mail: laurent.pauwels@sydney.edu.au
1 Introduction

For risk managers, investors, and regulators alike, forecasting financial risk and asset returns is central to their market activities. When forecasting asset returns and the risk of a financial portfolio, point forecasts rarely suffice, and the entire density is often required. A predictive density allows for one to capture all of its characteristics, including its tails. For example, measures of downside risk for investments, such as Value-at-Risk (VaR), require information on the left tail of the distribution of asset returns. This requirement implies that when modeling the entire density, preserving characteristics such as the degree of asymmetry and the thickness of the tails measured by higher moments, such skewness and kurtosis, respectively, is crucial.

A multitude of financial density forecasts exists that can easily be produced using a variety of models, leaving the forecaster to choose a predictive density. Rather than restricting the choice to one density, a popular strategy is to combine the forecasts into a consensus forecast. Empirical applications of forecast combination often produce significant improvements in forecast accuracy. Concerning the recent M4 competition that included 100,000 series, Makridakis et al. (2018) found that out of the 17 most accurate methods, 12 were combinations. Since the introduction forecast combination by Bates and Granger (1969), the literature on combination has grown substantially. Timmermann (2006) provides an extensive overview. Until recently, most of the literature focused on point forecasts, and the treatment of predictive density combinations was sparse.

One of the earliest contributions addressing the problem of combining predictive densities is discussed in Wallis (2005) and Hall and Mitchell (2007). Hall and Mitchell (2007) proposed a practical way to select optimal weights by maximizing the average logarithmic score of the combined density forecast to minimize the “distance” between the forecasted and true (unknown) density, as measured by the Kullback–Leibler information criterion (KLIC). Geweke and Amisano (2011) used Bayesian methods and provided some theoretical justification for using optimal weighting schemes in pooling linear models. Recent contributions and applications to density forecast combinations include Geweke and Amisano (2010), Billio et al. (2013), Del Negro et al. (2016), and Crisóstomo and Couso (2017).

Although some theoretical advances have been provided by Elliott (2011) and Chan and Pauwels (2018) for forecast combinations, the statistical properties of the optimal weights proposed by Hall and Mitchell (2007) and Geweke and Amisano
(2011), such as consistency and asymptotic distribution, remain relatively unexplored. Recently, Pauwels and Vasnev (2016) has drawn attention to the fact that the optimal weights can have undesired properties when the number of forecasting periods is relatively small. Opschoor et al. (2017) demonstrates that using the censored likelihood scoring rule proposed by Diks et al. (2011) outperforms other methods if the tail of the distribution is the main feature of interest.

Despite these recent contributions on the optimal combination of predictive densities, little is known about the statistical properties of such combinations. In particular, what happens to the moments of the combination when the densities are combined? Specifically, what are the implications for higher moments such as the skewness and kurtosis of the combination? These questions have remained unanswered in the literature. However, the question is very important because the majority of financial returns on assets exhibits asymmetry and heavy tails. This phenomenon is illustrated by the sample moments of some of the main stock market indices shown in Table 1 (a similar table is reported in Jondeau and Rockinger, 2009). These higher moments are also crucial for VaR and Expected Shortfall forecasting (Polanski and Stoja, 2010).

We answer the question by empirically analyzing the impact of combining densities on higher moments, and then, we provide an asymptotic theory as justification for the observed results. We find that combinations with equal weights or optimal log score weights significantly reduce the skewness and kurtosis of the combination when the individual densities are skewed and/or fat-tailed.

We overcome this issue by restricting higher-order moments when maximizing the average log score. We provide a general method for combining predictive densities by maximizing the average logarithmic score subject to constraints that allow one to focus on specific characteristics of the combined density, such as the thickness of the tails or the asymmetry. In other words, we propose computing the optimal weights

<table>
<thead>
<tr>
<th></th>
<th>S&amp;P 500</th>
<th>DJIA 30</th>
<th>Nikkei 225</th>
<th>FTSE 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewness</td>
<td>-0.249</td>
<td>-0.121</td>
<td>-0.232</td>
<td>-0.357</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>11.358</td>
<td>11.500</td>
<td>7.901</td>
<td>4.424</td>
</tr>
</tbody>
</table>

Notes: The values reported are the daily returns of the market indices from January 3, 2000, until July 20, 2018. The data are from Yahoo! Finance.
under additional higher moments restrictions. We name these optimal weights derived under high moment constraints \textit{HMC weights}. The benefit of this approach is that the resulting combined density is suitable not only for the tails but also for the entire support of the distribution.

We show the validity of this approach both theoretically and numerically. First, we derive the statistical properties, namely, the consistency and the asymptotic distribution of both HMC weights and the optimal log score weights proposed by Hall and Mitchell (2007) and Geweke and Amisano (2011). Second, we run a series of simulations to compare the performance of the HMC weights with that of the optimal log score weights without such constraints and the commonly used equal weighting approach. Third, we provide an empirical illustration in forecasting the densities of the conditional returns of the S&P 500 index. The conditional returns are forecasted using several GARCH and EGARCH models, which are regularly employed in the financial econometrics literature. This illustration is especially relevant as the S&P 500 exhibits heavy tails (see Table 1). In both numerical studies, we evaluate the combined predictive densities on its overall performance in terms log score and its performance in the tails by forecasting Value-at-Risk. The simulations and empirical results strongly support the proposed methodology.

The remainder of this paper is organized as follows. Section 2 investigates the impact of combining densities on the moments of the combination. Section 3 proposes optimal weights under higher-order moment constraints and studies their statistical properties. Section 4 provides an empirical application for the S&P 500 index. Section 5 concludes.

\section{Moments of the combination}

\subsection{Behavior of the moments}

We start by describing the behavior of the moments of the density combination. A priori, the impact that combining \(k\) densities (or models) would have on the higher moments of the resulting combined density is not obvious. A simple way to combine \(k\) densities is to aggregate them linearly into one density as follows:

\[
p_c(\cdot; \omega, \theta) = \sum_{j=1}^{k} \omega_j p_j(\cdot; \theta_j), \quad (1)
\]

2
where $\mathbf{\omega} = (\omega_1, \ldots, \omega_k)^\top \in \mathbb{R}^k$ is the vector of weights, $\mathbf{\theta} = (\theta_1^\top, \ldots, \theta_k^\top)^\top$ is the combined vector of all parameters, and $\theta_j$ is a vector of parameters of the $j^{th}$ density, $p_j(\cdot; \theta_j)$. For $p_c(\cdot; \mathbf{\omega}, \mathbf{\theta})$ to be a density, the weights need to be nonnegative, $\omega_j \geq 0$, and sum up to one, $\sum_{j=1}^{k} \omega_j = 1$.

Whereas the first moment of the combination, $\mu_c$, is simply a linear combination of $k$ individual density means, other moments are more complicated and depend on higher-order polynomials of the weights $\omega_j$. Suppose that the $j$-th density has mean $\mu_j$, variance $\sigma_j^2$, skewness $\gamma_j$, kurtosis $\kappa_j$, and $s$-th centered moment $m_{j,s}$. The following proposition uses the definition of the moments and provides formulas for the moments of the aggregate density.

**Proposition 2.1.** The moments of the combined density $p_c(\cdot; \mathbf{\omega}, \mathbf{\theta})$ are

(a) the mean:
$$
\mu_c = \sum_{j=1}^{k} \omega_j \mu_j,
$$

(b) the variance:
$$
\sigma_c^2 = \sum_{j=1}^{k} \omega_j (\sigma_j^2 + (\mu_j - \mu_c)^2),
$$

(c) the skewness:
$$
\gamma_c = \sum_{j=1}^{k} \omega_j \left[ \gamma_j \frac{\sigma_j^3 + 3(\mu_j - \mu_c) \sigma_j^2 + (\mu_j - \mu_c)^3}{\sigma_c^3} \right], \quad (2)
$$

(d) the kurtosis:
$$
\kappa_c = \sum_{j=1}^{k} \omega_j \left[ \frac{\kappa_j \sigma_j^4 + 4(\mu_j - \mu_c) \kappa_j + 6 (\mu_j - \mu_c)^2 \sigma_j^2 + (\mu_j - \mu_c)^4}{\sigma_c^4} \right], \quad (3)
$$

(e) the $s$-th centered moment:
$$
m_{c,s} = \sum_{j=1}^{k} \omega_j \sum_{l=0}^{s} \binom{s}{l} (\mu_j - \mu_c)^l m_{j,s-l}, \quad (4)
$$

where $\binom{s}{l} = \frac{s!}{l!(s-l)!}$ is the binomial coefficient.

A simple numerical illustration shows that higher moments of the density combination that are relevant in empirical finance, such as skewness and kurtosis, can change considerably even when combining models with the same skewness and kurtosis. Figure 1 demonstrates the behavior of skewness, $\gamma_c$ in equation (2), for different
values of the weight $\omega_1$ when combining two similar distributions, such as a skewed normal. In Figure 1, the individual density parameters are set to $\sigma_1 = \sigma_2 = 1$, $\gamma_1 = \gamma_2 = 1$, and $\kappa_1 = \kappa_2 = 3$, but feature different means, $\mu_1$ and $\mu_2$. If $\mu_1 = 0.1$ and $\mu_2 = 1$, then for $\omega_1 = 0.35$, the skewness of the combination is approximately 0.75. If $\mu_1 = -1$ and $\mu_2 = 1$, $\gamma_c$ is lower than 0.5 for $\omega_1$ between 0.10 and 0.65.

Similarly, Figure 2 displays the behavior of the kurtosis, $\kappa_c$, in equation (3) for different values of the weight $\omega_1$ when combining two $t_5$ distributions. The parameters of the individual $t_5$ are set to $\sigma_1 = \sigma_2 = \sqrt{5/3}$, $\gamma_1 = \gamma_2 = 0$, and $\kappa_1 = \kappa_2 = 9$, and the means, $\mu_1$ and $\mu_2$, differ. In Figure 2 (a), the means are $\mu_1 = -1$ and $\mu_2 = 1$; moreover, when $\omega_1 = 0.5$, the kurtosis of the combination reduces to approximately 5. Additionally, in Figure 2 (b), when $\mu_1 = -5$, $\mu_2 = 1$ and the same weight, $\omega_1 = 0.5$, removes heavy tails altogether. Obviously, when $\omega_1$ is close to the boundary (0 or 1), only one density is selected, and the skewness and kurtosis of the combination are essentially those of the individual density.

### 2.2 Simulation

We illustrate the aforementioned effect of a significant change in the skewness and kurtosis when densities are combined in a systematic simulation experiment. Consider the data generating process given by the linear regression

$$ y_t = x_t^\top \beta + \varepsilon_t, $$

with $x_t = (x_{1t}, \ldots, x_{kt})^\top$ and $\beta = (1, \ldots, 1)^\top$. The regressors are standard normal random variables, $x_{jt} \sim N(0, 1)$, that are independent from each other. The error
Figure 2: $\kappa_c$ as a function of $\omega_1$ if $\sigma_1 = \sigma_2 = \sqrt{5/3}$, $\gamma_1 = \gamma_2 = 0$, and $\kappa_1 = \kappa_2 = 9$.

term is heavy-tailed, $\varepsilon_t \sim t_5$, and is generated independently of the regressors.

We observe the data for $t = 1, \ldots, T-1$ and produce $k$ forecasts for the conditional mean of $y_T$:

$$\hat{\theta}_j = \hat{\beta}_j x_{jT}; \quad j = 1, \ldots, k,$$

where $\hat{\beta}_j$ is the estimate of the slope coefficient in the simple linear regression model that only uses the $j$-th predictor. The distribution of $\varepsilon_t$ is known to belong to the $t$-distribution family, but the degrees of freedom are unknown. To predict $y_T$, we use a combination of densities, $p_j(\cdot; \hat{\theta}_j)$, where the densities are $t_5$ for $j \leq \lfloor \frac{k}{2} \rfloor$ and are $t_6$ for $j > \lfloor \frac{k}{2} \rfloor$, with mean $\hat{\theta}_j$.

The density combination is given by

$$p_c(\cdot; \hat{\theta}) = \sum_{j=1}^{k} \omega_j p_j(\cdot; \hat{\theta}_j)$$

with weights $\omega_j$ satisfying the restrictions $\sum_{j=1}^{k} \omega_j = 1$ and $\omega_j \geq 0$. We consider 6 different sets of ad hoc weights. The first set starts with weight 1 on the first model and 0 on all others. In the second set, the first weight decremented to 0.75 when weighting the rest of the models equally. More weight is distributed gradually to the remaining models at a step of 0.25 until the equal weight set is achieved. The last set of weights is subsequently described.

We consider the optimal weights of Hall and Mitchell (2007) and Geweke and Amisano (2011), which are based on the idea that, in practice, the combination being close to the true but unknown density $f(\cdot)$ of the predicted outcome $y_T$ is desirable. The Kullback–Leibler information criterion (KLIC) can be employed to
measure the distance of the combined density to the true density:

$$\text{KLIC}(\omega, \theta) = E \left[ \log \left( \frac{f(y_T)}{p_c(y_T; \omega, \theta)} \right) \right],$$  \hspace{1cm} (8)

and can be estimated with its sample analogue:

$$\hat{\text{KLIC}}(\omega, \theta) = \frac{1}{T} \sum_{t=1}^{T} \log \left( \frac{f(y_t)}{p(y_t; \omega, \theta)} \right),$$  \hspace{1cm} (9)

using the actual realizations $y_t$. Because the true density $f(\cdot)$ does not depend on $\omega$, the weight that minimizes $\text{KLIC}$ can be found by solving the following optimization problem

$$\begin{align*}
\text{maximize} & \quad \sum_{t=1}^{T} \log \left( \sum_{j=1}^{k} \omega_j p_j(y_t; \theta_j) \right) \\
\text{subject to} & \quad \sum_{j=1}^{k} \omega_j = 1, \\
& \quad \omega_j \geq 0, \quad j = 1, \ldots, k
\end{align*}$$  \hspace{1cm} (10)

For convenience, the optimal weights that solve equation (10) are named \textit{log score (optimal) weights.}

Table 2 presents the numerical results based on 5000 replications. Panel A shows the impact of increasing the number of models ($k$) on the skewness of the combination ($\gamma_c$), whereas Panel B shows the corresponding effect on the kurtosis of the combination ($\kappa_c$). The skewness of the $t_5$ error density is set to 1, and the kurtosis of the $t_5$ error density is 9. Both the skewness and the kurtosis of the combination decrease with the increasing number of models combined. This phenomenon is also evident in Figures 3 and 4, which depict histograms of the skewness and the kurtosis of the combination based on the optimal weights obtained by solving (10). Figure 3 shows the shift of the kurtosis toward 3 when the number of models used in combination increases, whereas Figure 4 illustrates the corresponding shift of the skewness of the combination toward 0.
### Table 2: Skewness ($\gamma_c$) and kurtosis ($\kappa_c$) of the combination

<table>
<thead>
<tr>
<th>Weights &amp; # of densities</th>
<th>$k = 2$</th>
<th>$k = 5$</th>
<th>$k = 10$</th>
<th>$k = 20$</th>
<th>$k = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${1, \ldots, 0}$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$\left{ 0.75, \frac{1-0.75}{k-1}, \ldots, \frac{1-0.75}{k-1} \right}$</td>
<td>0.792</td>
<td>0.746</td>
<td>0.717</td>
<td>0.705</td>
<td>0.684</td>
</tr>
<tr>
<td>$\left{ 0.50, \frac{1-0.50}{k-1}, \ldots, \frac{1-0.50}{k-1} \right}$</td>
<td>0.749</td>
<td>0.639</td>
<td>0.597</td>
<td>0.573</td>
<td>0.547</td>
</tr>
<tr>
<td>$\left{ 0.25, \frac{1-0.25}{k-1}, \ldots, \frac{1-0.25}{k-1} \right}$</td>
<td>0.786</td>
<td>0.594</td>
<td>0.539</td>
<td>0.503</td>
<td>0.474</td>
</tr>
<tr>
<td>$\left{ \frac{1}{k}, \ldots, \frac{1}{k} \right}$</td>
<td>0.749</td>
<td>0.592</td>
<td>0.527</td>
<td>0.481</td>
<td>0.449</td>
</tr>
<tr>
<td>Log score weights</td>
<td>0.758</td>
<td>0.621</td>
<td>0.541</td>
<td>0.452</td>
<td>0.386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weights &amp; # of densities</th>
<th>$k = 2$</th>
<th>$k = 5$</th>
<th>$k = 10$</th>
<th>$k = 20$</th>
<th>$k = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${1, \ldots, 0}$</td>
<td>9.000</td>
<td>9.000</td>
<td>9.000</td>
<td>9.000</td>
<td>9.000</td>
</tr>
<tr>
<td>$\left{ 0.75, \frac{1-0.75}{k-1}, \ldots, \frac{1-0.75}{k-1} \right}$</td>
<td>6.981</td>
<td>6.861</td>
<td>6.890</td>
<td>6.876</td>
<td>6.870</td>
</tr>
<tr>
<td>$\left{ 0.50, \frac{1-0.50}{k-1}, \ldots, \frac{1-0.50}{k-1} \right}$</td>
<td>6.158</td>
<td>5.759</td>
<td>5.771</td>
<td>5.727</td>
<td>5.714</td>
</tr>
<tr>
<td>$\left{ 0.25, \frac{1-0.25}{k-1}, \ldots, \frac{1-0.25}{k-1} \right}$</td>
<td>5.894</td>
<td>5.177</td>
<td>5.168</td>
<td>5.097</td>
<td>5.080</td>
</tr>
<tr>
<td>$\left{ \frac{1}{k}, \ldots, \frac{1}{k} \right}$</td>
<td>6.158</td>
<td>5.107</td>
<td>4.989</td>
<td>4.865</td>
<td>4.840</td>
</tr>
<tr>
<td>Log score weights</td>
<td>6.187</td>
<td>5.385</td>
<td>5.274</td>
<td>4.848</td>
<td>4.466</td>
</tr>
</tbody>
</table>

Notes: The optimal weights are obtained by solving (10). The combined $k$ densities are described in equation (7). The individual densities are constructed using the estimated parameters of the linear regression in (5). In Panel A, the skewness of the $t_5$ error distribution is set to 1. In Panel B, the kurtosis of the $t_5$ error distribution is 9.

### 2.3 Asymptotic results

The previous section demonstrated the undesirable effect that combining densities can have on the skewness and kurtosis. Here, we examine a setting in which the densities are combined with the simple equal weights, and the number of models grows toward infinity.

Consider the general simulation setup in Section 2.2. We again focus on the linear regression model (5) but let $\beta = (\beta, \ldots, \beta)^\top$. The regressors $x_t$ are i.i.d. zero mean random vectors independent from the errors $\varepsilon_t$, which are also i.i.d. with a zero mean. We observe the data for $t = 1, \ldots, T - 1$ and produce $k$ forecasts for the conditional mean of $y_T$ using the same approach as in (6). For each fixed $y$ and $j$, we let $p_j(y; \theta) = p(y - \theta)$, where $p$ is a known density. For example, $p$ could be the density of the errors in model (5).

We focus on the equally weighted combination of these $k$ densities and let $\gamma_c$ and
Figure 3: Distribution of the kurtosis of the density combinations ($\kappa_c$) for optimal log score weights that result from solving (10).

Figure 4: Distribution of the skewness of the density combinations ($\gamma_c$) for optimal log score weights that result from solving (10).
\( \kappa_c \) denote its skewness and kurtosis, respectively. Denote the standard deviation, skewness, and kurtosis of density \( p \) by \( \sigma_p, \gamma_p, \) and \( \kappa_p, \) respectively. Suppose that the predictors are independent and can be split into finitely many (asymptotically) equally sized groups, such that the predictors within each group are identically distributed. Assume that the number of predictor groups, \( G, \) stays constant as the number of predictors tends to infinity. We write \( \sigma^2_{X,g} \) for the variance of each predictor in group \( g \in \{1,\ldots,G\} \) and let \( \sigma^2_X \) denote the average variance across the predictor groups: \( \sigma^2_X = \sum_{g=1}^{G} \sigma^2_{X,g}/G. \) We define \( \gamma_X \) and \( \kappa_X \) by analogy, as the average predictor skewness and kurtosis, respectively. Note that if \( G = 1, \) then \( \sigma^2_X, \gamma_X, \) and \( \kappa_X \) are simply the variance, skewness, and kurtosis of each individual predictor.

Let \( R = \sigma^2_X/\sigma^2_p, \) and assume that all of the quantities defined in this paragraph are finite. The following result is proven in the Appendix.

**Theorem 2.2.** Suppose that \( T \to \infty, \ k \to \infty \) and \( k/\sqrt{T} \to 0. \) Then:

\[
\begin{align*}
\gamma_c \xrightarrow{P} \gamma_p \left[1 + \beta R \right]^{-3/2} + \gamma_X \left[1 + (\beta R)^{-1} \right]^{-3/2} \\
\kappa_c \xrightarrow{P} \kappa_p \left[1 + \beta R \right]^{-2} + \kappa_X \left[1 + (\beta R)^{-1} \right]^{-2} + 6 \left[2 + \beta R + (\beta R)^{-1} \right]^{-1}.
\end{align*}
\]

In addition, if we also let \( \beta \to 0 \) as \( T \to \infty, \) then \( \gamma_c \xrightarrow{P} \gamma_p \) and \( \kappa_c \xrightarrow{P} \kappa_p. \) Alternatively, if \( \beta \to \infty, \) then \( \gamma_c \xrightarrow{P} \gamma_X \) and \( \kappa_c \xrightarrow{P} \kappa_X. \)

Table 3 further illustrates Theorem 2.2 in the cases \( \beta \to 0 \) and \( \beta \to \infty. \) It uses the same simulation framework as in Section 2.2, but with different values of \( \beta. \) When \( \beta = (1/\sqrt{k}, \ldots, 1/\sqrt{k})^\top, \) the amount of the signal in the model is small relative to the variance of the error term; consequently, the kurtosis of the combination approaches the average kurtosis of the individual predictive densities. Alternatively, when \( \beta = (3, \ldots, 3)^\top, \) the amount of signal increases in relation to the noise, and hence, the kurtosis of the combination approaches the kurtosis of the individual predictors, that is, 3, because the predictors are normally distributed.

## 3 Higher moments constraints

### 3.1 Optimization problem

Because the optimal weights do not preserve characteristics of the combined densities, such as the thickness of tails, additional restrictions on weights are required.
Table 3: Combined kurtosis ($\kappa_c$) for different values of $\beta$

<table>
<thead>
<tr>
<th>Weights &amp; # of densities</th>
<th>$\beta = (1/\sqrt{k}, \ldots, 1/\sqrt{k})^\top$</th>
<th>$\beta = (3, \ldots, 3)^\top$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal weights</td>
<td>$k = 2$</td>
<td>$k = 10$</td>
</tr>
<tr>
<td></td>
<td>6.674</td>
<td>7.370</td>
</tr>
<tr>
<td>Log score weights</td>
<td>$k = 2$</td>
<td>$k = 10$</td>
</tr>
<tr>
<td></td>
<td>6.680</td>
<td>7.009</td>
</tr>
</tbody>
</table>

Notes: Section 2.2 provide the detailed setup of this simulation.

if the combination is to keep those properties. The higher moments constrained optimization is given by

 maximize $\sum_{t=1}^{T} \log \left[ \sum_{j=1}^{k} \omega_j p_j(y_t; \theta_j) \right]$

subject to $\sum_{j=1}^{k} \omega_j = 1,$

$\omega_j \geq 0, \quad j = 1, \ldots, k$

$\kappa_c \geq \kappa$ and/or $\gamma_c \geq \gamma,$

(11)

where the kurtosis of the combination, $\kappa_c$, is given by equation (2) and the skewness of the combination, $\gamma_c$, is given by equation (3). Without loss of generality, the constraints can be modified to suit the problem. The additional constraints are nonlinear, and bounds $\kappa$ and $\gamma$ must be selected carefully to avoid empty feasible sets. The optimal weights obtained by solving the log score objective function (11) under high moment constraints are named $HMC$ weights for brevity.

3.2 Simulation

Whereas optimizing (10) yields the best possible log score among density combinations, optimizing (11) results in kurtosis and skewness of the combination that cannot be lower than the lower bounds imposed by the corresponding constraints. Hence, log score optimal weights and HMC optimal weights will generally yield different density combinations. We can compare the two combinations by considering both the overall performance captured by log scoring together with the performance in the tails, which we evaluate by examining the Value-at-Risk (VaR) predictions.

As in Section 2, we continue using the simple regression framework given by (5) with a heavy-tailed error term now set to $\varepsilon_t \sim t_6$ and compute the combined pre-
dictive density \( p_c(:; \theta) \). Two sets of models (densities) are considered for combining. The first set features \( k = 2 \) densities, \( t_6 \) and \( t_{30} \). The second combination features two \( t_6 \) and three \( t_{30} \). The individual densities differ in means as previously because they are estimated from the regression model (5).

We compute 99% and 95% VaR forecasts for the combined predictive densities. The average of the VaR forecasts from individual predictive densities does not necessarily equate to the VaR forecast of the combined predictive densities. Hence, individual densities need to be simulated. The simulations draw from \( t_6 \) and \( t_{30} \) random variables proportionally to the optimal weights of the combination and amounting to 10,000 realizations in total.

We construct 3000 VaR forecasts that are compared with the 1% and 5% left tailed quantiles from the simulated distribution of the combination. We compute the number of times that \( y_T \) is to the left of the corresponding VaR forecasts. We experiment with several constraints on the kurtosis, \( \kappa_c \geq 5, 5.5 \) and 6, effectively treating the kurtosis constraint as a tuning parameter. Furthermore, we also present the log score optimal weights as defined in (10) and equal weight density combining for comparison. In addition to VaR, we also report the log score for the overall performance of the different combinations and their corresponding average kurtosis (\( \bar{\kappa}_c \)).

The results of the experiments can be found in Table 4. The optimal combination with a constraint on the kurtosis (HMC weights) performs best in predicting the 95% and 99% VaR over the other combinations. The average kurtosis of the combinations, \( \bar{\kappa}_c \), shows that the constraint is met. Furthermore, with every increase in \( \kappa \), the percentage violation at both the 1% and the 5% levels moves closer to the intended target, whether combining 2 or 5 models. Meanwhile, the kurtosis of both the log score weights and equal weights combination is effectively close to the average of the kurtosis of the individual densities. The log score performance of the HMC weights combination also tends to deteriorate with the increasing \( \kappa \).

### 3.3 Asymptotic Properties of HMC Weights

In this section, we establish consistency, rate of convergence, and the limiting distribution of the solution to the optimization problem (11). Our results cover the asymptotics of the corresponding unconstrained estimator as a special case. To the best of our knowledge, such results have not yet appeared in the literature, even for
### Table 4: VaR experiment

#### Panel A: Combination of models (k=2)

<table>
<thead>
<tr>
<th>Combination</th>
<th>Log score</th>
<th>$\kappa_c$</th>
<th>% viol. at 1%</th>
<th>% viol. at 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMC weights ($\kappa = 5$)</td>
<td>0.989</td>
<td>5.016</td>
<td>1.358</td>
<td>7.536</td>
</tr>
<tr>
<td>HMC weights ($\kappa = 5.5$)</td>
<td>0.977</td>
<td>5.501</td>
<td>1.076</td>
<td>6.355</td>
</tr>
<tr>
<td>HMC weights ($\kappa = 6$)</td>
<td>0.951</td>
<td>6.000</td>
<td>1.005</td>
<td>6.095</td>
</tr>
<tr>
<td>Log score weights</td>
<td>1.000</td>
<td>4.493</td>
<td>1.665</td>
<td>7.468</td>
</tr>
<tr>
<td>Equal weights</td>
<td>0.998</td>
<td>4.630</td>
<td>1.461</td>
<td>7.162</td>
</tr>
</tbody>
</table>

#### Panel B: Combination of models (k=5)

<table>
<thead>
<tr>
<th>Combination</th>
<th>Log score</th>
<th>$\kappa_c$</th>
<th>% viol. at 1%</th>
<th>% viol. at 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMC weights ($\kappa = 5$)</td>
<td>0.997</td>
<td>5.008</td>
<td>1.300</td>
<td>6.967</td>
</tr>
<tr>
<td>HMC weights ($\kappa = 5.5$)</td>
<td>0.989</td>
<td>5.500</td>
<td>1.000</td>
<td>6.969</td>
</tr>
<tr>
<td>HMC weights ($\kappa = 6$)</td>
<td>0.965</td>
<td>6.000</td>
<td>0.536</td>
<td>4.523</td>
</tr>
<tr>
<td>Log score weights</td>
<td>1.000</td>
<td>4.471</td>
<td>1.567</td>
<td>7.500</td>
</tr>
<tr>
<td>Equal weights</td>
<td>0.997</td>
<td>4.542</td>
<td>1.433</td>
<td>6.900</td>
</tr>
</tbody>
</table>

Notes: The table reports the proportion of times the VaR forecast exceeds the 1% and 5% quantiles. The considered simulations have $T = 3000$ VaR forecasts. $\kappa_c$ is the average kurtosis of the combinations. The log score is relative to the log score optimal weights. The log score optimal weights are obtained by solving (10) and the HMC optimal weights by solving (11).

We impose the following mild continuity and regularity assumptions. We note that if the constraint in optimization problem (11) involves only the skewness of the density combination, then assumption A3 can be relaxed to only concern the first three moments. In what follows, $\mathcal{B}(\theta^*)$ is a closed ball around $\theta^*$, whose radius we are allowed to have as arbitrarily small but positive. The vector $\theta^*$ is defined in assumption A2 and can be thought of as the “population” vector of the model parameters.

A1: $\{y_t\}_{t=1}^{\infty}$ is a stationary ergodic sequence.

A2: The estimates of the model parameters converge in probability as $T$ tends to infinity: $\hat{\theta} \overset{P}{\rightarrow} \theta^*$, for some fixed finite vector $\theta^*$.

A3: For $\theta \in \mathcal{B}(\theta^*)$ and all $j \leq k$, the first four moments of densities $p_j(\cdot; \theta_j)$ are well-defined continuous functions of $\theta_j$, and the corresponding variances are nonzero.

A4: For $\theta \in \mathcal{B}(\theta^*)$, $j \leq k$ and each fixed $y$, functions $p_j(y; \theta_j)$ are continuous in $\theta_j$. 

14
A5: 
\[ E \sup_{\theta \in B(\theta^*)} \left| \log p_j(y_1; \theta) \right| < \infty \quad \text{for } j = 1, \ldots, k. \]

A6: 
\[ E \sup_{\theta \in B(\theta^*)} p_j(y_1; \theta) < \infty \quad \text{for } j = 1, \ldots, k. \]

We define \( C(\theta) \) as the constraint set for the weights \( \omega \) in the optimization problem (11). We denote by \( \tilde{\omega} \) the HMC optimal weights, that is, the solution to the optimization problem (11) but with \( \theta \) replaced by \( \hat{\theta} \). The corresponding population solution is:
\[ \omega^* = \arg \min_{\omega \in C(\theta^*)} \text{KLIC}(\omega, \theta^*), \tag{12} \]
where KLIC(\( \omega, \theta \)) is defined in (8). Theorem 3.1 establishes the consistency of \( \tilde{\omega} \). The proof is in the Appendix.

**Theorem 3.1.** Suppose that \( \omega^* \) is the unique solution to the population problem (12). If assumptions A1–A6 are satisfied, then \( \tilde{\omega} \overset{P}{\rightarrow} \omega^* \) as \( T \to \infty \).

From the proof, it follows that if the convergence of \( \hat{\theta} \) to \( \theta^* \) is almost sure rather than in probability, then the result of Theorem 3.1 can be strengthened to the almost sure convergence as well.

We now establish a result of the limiting distribution of \( \tilde{\omega} \). For the simplicity of the exposition, we focus on the case \( \hat{\theta} = \theta^* \), which allows us to avoid imposing specific assumptions on the form of \( \hat{\theta} \) as a function of the data. Consequently, we change assumption A2 by setting \( \hat{\theta} = \theta^* \) and relax assumptions A3–A6 by setting \( B(\theta^*) = \{ \theta^* \} \). We denote the modified assumptions by A2’–A6’. We also impose additional dependence and regularity conditions. In what follows, the “unconstrained” minimizer of KLIC is defined under the restriction that the weights are nonnegative and sum to one.

A7: \( \{ y_t \}_{t=1}^{\infty} \) is an \( m \)-dependent sequence for some finite \( m \).

A8: All of the elements of the vector \( \omega^* \) are positive.

A9: The unconstrained minimizer of KLIC(\( \cdot, \theta^* \)) lies in \( C(\theta^*) \).
Let \( \ell^*_y \left( y \right) = \left( p_2(y; \theta^*) - p_1(y; \theta^*), \ldots, p_k(y; \theta^*) - p_1(y; \theta^*) \right)^\top / p(y; \omega^*, \theta^*) \), and define

\[
\Sigma^* = E \ell^*(y_1) \ell^*(y_1)^\top + 2 \sum_{i=2}^{m+1} E \ell^*(y_1) \ell^*(y_i)^\top \quad \text{and} \quad V_* = E \ell^*(y_1) \ell^*(y_1)^\top.
\]

Because the weights in all \( \omega \) that we consider are required to sum to one, we can write \( \omega_1 = 1 - \sum_{j=2}^{k} \omega_j \), and thus, every function of \( \omega \) can be expressed as a function of \( \omega_{-1} = (\omega_2, \ldots, \omega_k)^\top \).

Treating the constraint set \( C(\theta^*) \) as a set in the space of reduced vectors \( \omega_{-1} \), we let \( S^* \) denote the tangent cone of \( C(\theta^*) \) at the point \( \omega_{-1}^* \). More specifically, a vector \( v \) lies in \( S^* \) if and only if there exists a sequence \( \tau_n \) decreasing to 0 and a sequence \( \omega_n \in C(\theta^*) \) converging to \( \omega^* \), such that \( [(\omega_n)_{-1} - \omega_{-1}^*]/\tau_n \rightarrow v \). For a given convex set \( A \) and point \( x \), we write \( \text{Proj}_A x \) for the projection of \( x \) onto \( A \).

**Theorem 3.2.** Suppose that \( \omega^* \) is the unique solution to the population problem (12) and assumptions A1, A2–A6, and A7–A9 are satisfied. If \( \omega^* \) lies in the interior of \( C(\theta^*) \), then

\[
\sqrt{T} (\hat{\omega}_{-1} - \omega_{-1}^*) \overset{d}{\rightarrow} N \left( 0, V_*^{-1} \Sigma^* V_*^{-1} \right).
\]

If \( \omega^* \) lies on the boundary of \( C(\theta^*) \) and \( \tilde{Z} \sim N \left( 0, V_*^{-1/2} \Sigma^* V_*^{-1/2} \right) \), then

\[
\sqrt{T} (\hat{\omega}_{-1} - \omega_{-1}^*) \overset{d}{\rightarrow} V_*^{-1/2} \text{Proj}_{V_*^{1/2} S} \tilde{Z}.
\]

4 **Empirical illustration**

In this section, we illustrate the benefits of using the HMC optimal weights when combining density forecasts based on real data. We use the daily percent log returns of the Standard and Poors 500 index (S&P 500). The sample covers the S&P 500 returns from January 3, 2000, until July 20, 2018.

The returns at time \( t \) can be expressed as

\[
y_t = \mu + \sqrt{h_t} \eta_t, \quad \eta_t | \mathcal{F}_{t-1} \sim F(0, 1),
\]

where \( F(\cdot) \) is a distribution with mean 0 and variance 1, and \( \mathcal{F}_{t-1} \) is a filtration up to \( t - 1 \). We use two main volatility models to forecast the returns and the conditional volatility of the S&P 500 returns. The first set of models are the workhorse of
volatility models, the GARCH model introduced by Bollerslev (1986):

$$h_t = \omega + \alpha(y_{t-1} - \mu)^2 + \beta \log h_{t-1}. \tag{15}$$

The statistical properties relevant to GARCH models are discussed in Ling and McAleer (2003). We also used the EGARCH model of Nelson (1991):

$$\log h_t = \omega + \gamma \varepsilon_{t-1} + \alpha (|\eta_{t-1}| - \mathbb{E}|\eta_{t-1}|) + \beta \log h_{t-1} \tag{16}$$

The interpretation and regularity conditions of the EGARCH model are discussed in McAleer and Hafner (2014) and Chang and McAleer (2017).

We consider several distributions, $F(\cdot)$, for the GARCH and EGARCH conditional returns. We use not only a Gaussian distribution but also fat-tailed distributions, including the Student-$t$, Laplace, and Hansen (1994) skewed-$t$.

We use rolling samples of 1250 trading days, which correspond to 5 years of trading data, to estimate all of the parameters and produce a one-step-ahead forecast of the conditional returns and conditional volatility models (14)–(16). Furthermore, we construct one-step-ahead predictive densities for each model over the remaining sample. We combine these predictive densities by solving optimization problems in (10) and (11) for 750 observations (3 years of trading data), which yields a one-step-ahead combined density forecast. This last step is repeated using a rolling window of 750 observations. We then evaluate these one-step-ahead combined predictive densities. The first combined predictive densities correspond to December 13, 2007, and the last one corresponds to July 20, 2018, which means that a total of 2667 predictive densities are evaluated over the sample. The constraint in (11) is imposed on the kurtosis of the combined density forecasts and takes the values of $\kappa = 5, 5.5$ and 6, respectively.

Recently, Diks et al. (2011) and Opschoor et al. (2017) proposed a censored likelihood (CSL) scoring rule that focuses on the left tail of the distribution of asset returns. Optimal weights can be derived by maximizing the following censored likelihood function over the data history:

$$S^{CSL} = \sum_{t=1}^{T} \log \left[ \sum_{j=1}^{k} \omega_j \left( I[y_t \in B_t^c] p_j(y_t; \theta_j) + I[y_t \in B_t] \int_{B_t^c} p_j(y_t; \theta_j) \, dy \right) \right], \tag{17}$$

where $B_t$ is a specific region of the distribution, $B_t^c$ is its complement, and $I[\cdot]$ is an
indicator function equal to 1 whenever the data $y_t$ are outside the support region $B_t$. Following the practical recommendation of Opschoor et al. (2017), we set the region to 0.15. Several alternative scoring functions exist that have been proposed in the literature, including from Gneiting and Ranjan (2011) and Jore et al. (2010). In this empirical illustration, both equal weights and the CSL score-based weights are used to construct a predictive density combination for comparison with the HMC optimal weights.

The accuracy and performance of combining density forecasts are assessed in two primary ways. First, we evaluate the entire density using the log score function. Second, we focus on evaluating the performance of the forecast combination in the left tail of the distribution by considering both the 99% and 95% 1-day Value-at-Risk (VaR) estimates:

$$\hat{\text{VaR}}^{-q}_{t} = \hat{\mu}_t + \sqrt{\hat{h}_t} \eta_q,$$

where $\eta_q$ is the $q^{th}$ quantile of the assumed conditional distribution. Moreover, $\hat{\mu}_t$ is the forecasted conditional mean return as expressed in (14), and $\hat{h}_t$ is the forecasted conditional variance as expressed in equations (15) and (16) for the GARCH and EGARCH models, respectively. When combining models, the VaR of the combination needs to be evaluated with simulations as discussed in Section 3. The daily returns are simulated from the individual distributions in proportions corresponding to the estimated weights of the combination. The 99% VaR is computed from the 1% quantile of distribution of the simulated returns, and the 95% VaR, from the 5% quantile.

In turn, VaR forecasts are evaluated using two methods. First, we evaluate the VaR violation whenever the actual return is smaller than the 1% or 5% quantile of distribution of the simulated returns. Second, we report the Christoffersen (1998) conditional coverage test, which assesses whether violations are happening in clusters.

Figure 5 shows the distribution of the implied combined kurtosis according to the HMC and the log score optimal combination methods. The constraint implies trivially that most of the combined kurtosis is at its boundary of $\kappa = 5.5$ and higher. In contrast, a log score optimal combination produces a combined kurtosis between 3.59 and 8.06, with an average of 4.46 (see Table 5).

Tables 5 and 6 are summarized as follows. The HMC optimal weights produce VaR performance with lower 1% violation numbers compared with the log score combination and the CSL combination. The latter, however, returns the lowest 5%
violation numbers. Whereas, expectedly, the log score optimal combination has the best average log score performance, it is followed very narrowly by the HMC optimal weights combination, unlike the CSL and Equal Weights combinations. The HMC optimal weights combination also outperform individual models in both overall log score performance and number of violations at the 1% level. The results at the 5% level, however, are mixed relative to the CSL and equal weights but clearly superior to the log score optimal weights.

As observed in the simulations, adjusting the constraint $\kappa$ upward results in a stronger focus on the tails. Specifically, the number of violations declines; however, conversely, the log scoring performance also decreases. The performance of the HMC weights at $\kappa = 6$ deteriorates relative to the two other constraints. This deterioration can be easily explained as follows. First, setting the constraint affects the entire distribution, not just the tails, which impacts performance. Second, $\kappa_c = 6$ is the highest constraint level in the current set of models. Both the Laplace GARCH and EGARCH models have a kurtosis of 6. Occasionally, the $t$-GARCH and $t$-EGARCH models produce an estimated kurtosis higher than 6. Higher than $\kappa = 6$, no guarantee exists that the optimization will converge. This computational limitation can be remedied by including fatter tailed models than the ones included in this illustration.

The optimal log score weights tend to favor Gaussian models, whereas the equal-weighted combination gives relatively more weight to fat-tail models since 6 out of 8 models of the combination are heavy-tailed. Therefore, not surprisingly, the equal-weighted combination performs well in terms of VaR forecasting and, hence, produces a low number of violations but performs poorly in its average log score.
performance. The VaR accuracy and performance can be improved empirically by modifying all of the optimal weights according to Jore et al. (2010). These techniques help outperform equal weight combinations, as shown in Opschoor et al. (2017).

Table 5: Evaluation of 1-day forecast for S&P 500 index

<table>
<thead>
<tr>
<th>Combination</th>
<th>Log score</th>
<th>min($\kappa_c$)</th>
<th>$\bar{\kappa}_c$</th>
<th>max($\kappa_c$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMC weights ($\kappa = 5$)</td>
<td>0.998</td>
<td>5.00</td>
<td>5.18</td>
<td>10.33</td>
</tr>
<tr>
<td>HMC weights ($\kappa = 5.5$)</td>
<td>0.995</td>
<td>5.50</td>
<td>5.62</td>
<td>10.33</td>
</tr>
<tr>
<td>HMC weights ($\kappa = 6$)</td>
<td>0.990</td>
<td>6.00</td>
<td>6.11</td>
<td>10.26</td>
</tr>
<tr>
<td>Log score weights</td>
<td>1.000</td>
<td>3.59</td>
<td>4.46</td>
<td>8.06</td>
</tr>
<tr>
<td>CSL weights</td>
<td>0.962</td>
<td>3.66</td>
<td>5.51</td>
<td>10.65</td>
</tr>
<tr>
<td>Equal weights</td>
<td>0.983</td>
<td>4.15</td>
<td>5.27</td>
<td>9.37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Individual models</th>
<th>Log score</th>
<th>min($\kappa_c$)</th>
<th>$\bar{\kappa}_c$</th>
<th>max($\kappa_c$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GARCH (Gaussian)</td>
<td>0.962</td>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>GARCH ($t$)</td>
<td>0.961</td>
<td>4.12</td>
<td>4.74</td>
<td>9.78</td>
</tr>
<tr>
<td>GARCH (Laplace)</td>
<td>0.929</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td>GARCH (Skew-$t$)</td>
<td>0.865</td>
<td>3.17</td>
<td>4.69</td>
<td>7.64</td>
</tr>
<tr>
<td>EGARCH (Gaussian)</td>
<td>0.973</td>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
</tr>
<tr>
<td>EGARCH ($t$)</td>
<td>0.978</td>
<td>3.89</td>
<td>4.14</td>
<td>8.09</td>
</tr>
<tr>
<td>EGARCH (Laplace)</td>
<td>0.935</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td>EGARCH (Skew-$t$)</td>
<td>0.859</td>
<td>3.25</td>
<td>4.10</td>
<td>8.35</td>
</tr>
</tbody>
</table>

Notes: $\bar{\kappa}_c$ is the average kurtosis of the combinations and min $\kappa_c$ and max $\kappa_c$ are the minimum and maximum kurtosis produced by the combinations. The log scores are relative to the log score optimal weights. The log score optimal weights are obtained by solving (10), the HMC optimal weights by solving (11), and the CSL weights from optimizing (17).
Table 6: 1-day forecast 95% and 99% VaR estimates for S&P 500 index

<table>
<thead>
<tr>
<th>Combination</th>
<th># viol. at 1%</th>
<th>CC test</th>
<th># viol. at 5%</th>
<th>CC test</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMC weights ($\kappa = 5$)</td>
<td>45 (1.69%)</td>
<td>0.363</td>
<td>187 (6.45%)</td>
<td>0.248</td>
</tr>
<tr>
<td>HMC weights ($\kappa = 5.5$)</td>
<td>44 (1.65%)</td>
<td>0.363</td>
<td>164 (6.15%)</td>
<td>0.270</td>
</tr>
<tr>
<td>HMC weights ($\kappa = 6$)</td>
<td>47 (1.76%)</td>
<td>0.363</td>
<td>171 (6.37%)</td>
<td>0.318</td>
</tr>
<tr>
<td>Log score weights</td>
<td>58 (2.17%)</td>
<td>0.363</td>
<td>189 (7.09%)</td>
<td>0.275</td>
</tr>
<tr>
<td>CSL weights</td>
<td>53 (1.99%)</td>
<td>0.358</td>
<td>136 (5.17%)</td>
<td>0.358</td>
</tr>
<tr>
<td>Equal weights</td>
<td>32 (1.20%)</td>
<td>0.358</td>
<td>146 (5.47%)</td>
<td>0.358</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Individual models</th>
<th># viol. at 1%</th>
<th>CC test</th>
<th># viol. at 5%</th>
<th>CC test</th>
</tr>
</thead>
<tbody>
<tr>
<td>GARCH (Gaussian)</td>
<td>64 (2.40%)</td>
<td>0.358</td>
<td>153 (5.74%)</td>
<td>0.294</td>
</tr>
<tr>
<td>GARCH ($t$)</td>
<td>45 (1.69%)</td>
<td>0.358</td>
<td>123 (4.61%)</td>
<td>0.358</td>
</tr>
<tr>
<td>GARCH (Laplace)</td>
<td>109 (4.09%)</td>
<td>0.294</td>
<td>309 (11.59%)</td>
<td>0.271</td>
</tr>
<tr>
<td>GARCH (Skew-$t$)</td>
<td>68 (2.55%)</td>
<td>0.358</td>
<td>161 (6.04%)</td>
<td>0.358</td>
</tr>
<tr>
<td>EGARCH (Gaussian)</td>
<td>58 (2.17%)</td>
<td>0.363</td>
<td>159 (5.96%)</td>
<td>0.187</td>
</tr>
<tr>
<td>EGARCH ($t$)</td>
<td>50 (1.87%)</td>
<td>0.358</td>
<td>126 (4.72%)</td>
<td>0.358</td>
</tr>
<tr>
<td>EGARCH (Laplace)</td>
<td>99 (3.71%)</td>
<td>0.274</td>
<td>304 (11.40%)</td>
<td>0.120</td>
</tr>
<tr>
<td>EGARCH (Skew-$t$)</td>
<td>65 (2.44%)</td>
<td>0.358</td>
<td>156 (5.85%)</td>
<td>0.358</td>
</tr>
</tbody>
</table>

Notes: The table reports both the number and the proportion of times that the VaR forecast exceeds the 1% and 5% quantiles. The percentage violations are in brackets. The considered sample has $T = 2667$ VaR forecasts. The CC tests are the $p$-value for the Christoffersen (1998) conditional coverage test. The log score optimal weights are obtained by solving (10), the HMC optimal weights by solving (11), and the CSL weights from optimizing (17).

5 Concluding remarks

In this paper, we show that combining many density forecasts tends to have a significant impact on higher moments of the combination, namely, skewness and kurtosis, even when the individual densities are skewed and/or heavy-tailed. We propose a solution that preserves the characteristics of the distribution, such as fat tails or asymmetry, by constraining the weights of the combination to achieve a minimum level of kurtosis or a certain level of skewness.

We provide a general methodology to combine multiple density forecasts based on optimizing the average sample Kullback–Leibler information criterion subject to a constraint on the skewness and/or kurtosis of the combination. The high moment constraint (HMC) optimal weights deliver a solution that is accurate in forecasting the overall distribution, including characteristics such as heavy tails. Moreover, we derive the statistical properties, namely, consistency and the asymptotic distribution, of both the proposed HMC optimal weights and the log score optimal weights.
We conduct a simulation to evaluate the HMC optimal weights on both the overall performance of the forecasted density and the performance in the tails. We also evaluate the weights through an empirical illustration in forecasting the conditional returns of the S&P 500 index. Not surprisingly, the HMC optimal weights outperform the log score optimal weights counterpart in the tails, as measured by the 99% VaR forecasts. Naturally, the overall performance of HMC weights, as measured by log scoring, is somewhat worse than that of the optimal weights without high moments constraints. However, HMC weights attain better log scoring performance than the equally weighted density combinations.
References


**Acknowledgements**

The authors would like to thank Manabu Asai, Graham Elliott, Christian Hafner, Michael McAleer, Shiqing Ling, George Tauchen and the participants of the Time Series Analysis of Higher Moments and Distributions of Financial Data, May 2018, Hong Kong; SETA2018, June 2018, Sydney; and EcoSta2018, June 2018, Hong Kong.

**Appendix**

**Proof of Theorem 2.2**

*Proof.* For brevity of the exposition, we focus on the skewness and derive the result for $G = 1$ and $\theta$ fixed at a positive value. The remaining cases and the derivations for the kurtosis follow by analogous arguments with only minor modifications. Define

$$
\bar{\mu} = \frac{1}{k} \sum_{j=1}^{k} \hat{\mu}_j T, \quad \sigma^2_{\mu} = \frac{1}{k} \sum_{j=1}^{k} \left( \hat{\mu}_j T - \bar{\mu} \right)^2, \quad \gamma_{\mu} = \frac{1}{k} \sum_{j=1}^{k} \frac{\left( \hat{\mu}_j T - \bar{\mu} \right)^3}{\sigma^3_{\mu}}
$$

and $\tilde{R} = \sigma^2_{\mu} / \sigma^2$. Note that $\bar{\mu} = o_p(1)$ by the law of large numbers. It follows from (2) that

$$
\gamma_c = \gamma_p \left[ 1 + \tilde{R} \right]^{-3/2} + \gamma_{\mu} \left[ 1 + (\tilde{R})^{-1} \right]^{-3/2}.
$$

(19)
Define \( \bar{x}_j = \sum_{t=1}^{T-1} x_{jt}/[T-1] \) and \( \eta_{jT} = [T - 1](\sum_{t=1}^{T-1}(x_{jt} - \bar{x}_j)^2)^{-1} \). Write \( \hat{\mu}_{jT} \) in the form \( \hat{\mu}_{jT} = \theta X_{jT} + \eta_{jT} \zeta_{jT} \) and note that \( \max_{j \leq k} E\xi_{jT}^2 = O(k/T) \). The last bound implies (for example, by Lemma 2.2.2 in van der Vaart and Wellner, 1996) that \( \max_{j \leq k} |\eta_{jT}| = O_p(1) \) by the assumptions on \( k \) and \( T \). A similar argument, together with the law of large numbers, gives \( \max_{j \leq k} |\xi_{jT}| = O_p(1) \). It follows that

\[
\sigma^2_{\mu} = \theta^2 (1/k) \sum_{j=1}^{k} X_{jT}^2 + o_p(1).
\]

Another application of the law of large numbers gives \( \sigma^2_{\mu} = \theta^2 \sigma_X^2 + o_p(1) \), which implies \( \bar{R} = R + o_p(1) \). Similarly,

\[
\gamma_{\mu} = (1/k) \sum_{j=1}^{k} \left( \frac{\theta X_{jT}}{\sigma_{\mu}} \right)^3 + o_p(1) = (1/k) \sum_{j=1}^{k} \left( \frac{X_{jT}}{\sigma_X} \right)^3 + o_p(1) = \gamma_{X} + o_p(1).
\]

We conclude the proof by combining the expressions for \( \bar{R} \) and \( \gamma_{\mu} \) with (19).

**Proof of Theorem 3.1**

**Proof.** For simplicity of the exposition, we use the notation from the empirical process theory: \( P_T h = (1/T) \sum_{t=1}^{T} h(y_t) \) for every function \( h \). Similarly, we write \( P h \) for \( E h(y_1) \). Also, for the remainder of the proof, all of the \( \omega \) are assumed to lie in the set \( \mathcal{W} = \{ \omega : \sum_{j \leq k} \omega_j = 1, \omega_j \geq 0, j = 1, \ldots, k \} \).

Let \( p_{\omega, \theta} \) denote the function \( p_c(\cdot ; \omega, \theta) \) and define

\[
G(\omega, \theta) = P \log \left[ \frac{p_{\omega, \theta}}{p_{\omega^*, \theta^*}} \right], \quad G_T(\omega, \theta) = P_T \log \left[ \frac{p_{\omega, \theta}}{p_{\omega^*, \theta^*}} \right].
\]

Note that \( \omega^* \) maximizes the function \( G(\cdot , \theta^*) \) over the constraint set \( C(\theta^*) \), while \( \tilde{\omega} \) maximizes \( G_T(\cdot , \hat{\theta}) \) over \( C(\hat{\theta}) \). Let \( \omega_{\theta} \) denote a projection of \( \omega^* \) onto the constraint set \( C(\theta) \). Note that \( \hat{\theta} \in B(\theta^*) \) with probability tending to one and define

\[
\Delta_T = \sup_{\omega \in \mathcal{W}, \theta \in B(\theta^*)} |G_T(\omega, \theta) - G(\omega, \theta)|.
\]
It follows from parts (i) and (ii) of Lemma 5.1 that, with probability tending to one,

$$G(\hat{\omega}, \hat{\theta}) \geq G_T(\hat{\omega}, \hat{\theta}) - \Delta_T \geq G_T(\omega_\theta, \hat{\theta}) - \Delta_T = o_p(1).$$  \hspace{1cm} (20)

We now argue that the above stochastic bound implies convergence of $\hat{\omega}$ to $\omega^*$, which is a zero of the function $G(\cdot, \theta^*)$, as well as its maximum over the constraint set $C(\theta^*)$. Fix an arbitrary positive $\delta$ and let $B_\delta(\omega^*)$ denote an open ball of radius $\delta$ around $\omega^*$. It follows from part (iii) of Lemma 5.1 that there exists a positive constant $c_\delta$, such that

$$\max_{\omega \in C(\theta) \setminus B_\delta(\omega^*)} G(\omega, \hat{\theta}) < -c_\delta$$

with probability tending to one. However, stochastic bound (20) implies $G(\hat{\omega}, \hat{\theta}) > -c_\delta$ with probability tending to one. Hence, with probability tending to one, $\hat{\omega} \in B_\delta(\omega^*)$. As this argument holds for every positive $\delta$, we have established that $\hat{\omega}$ converges to $\omega^*$ in probability. \qed

The next result is used in the proof of Theorem 3.1.

**Lemma 5.1.** The following holds under the assumptions and notation in the statement and the proof of Theorem 3.1:

(i) $\Delta_T = o_p(1)$

(ii) $G_T(\omega_\theta, \hat{\theta}) = o_p(1)$

(iii) Given a positive $\delta$, there exists a positive constant $r_\delta$, such that

$$\max_{\omega \in C(\theta) \setminus B_\delta(\omega^*)}, \theta \in B_{r_\delta}(\theta^*) \ G(\omega, \theta) < 0.$$

**Proof.** We start with part (i). For convenience, we denote functions $\log[p_{\omega, \theta}/p_{\omega^*, \theta}]$ by $m_{\omega, \theta}$ and functions $p_j(\cdot; \theta)$ by $p_{j, \theta}$. We first show that the class of functions $m_{\omega, \theta}$ is pointwise compact in the sense of Example 19.8 in van der Vaart (2000). Specifically, (a) the map $(\omega, \theta) \mapsto m_{\omega, \theta}(y)$ is continuous for each fixed $y$; (b) $(\omega, \theta)$ belong to a compact set; (c) this class has an integrable envelope. Parts (a) and (b) hold by the imposed assumptions. Using the fact that the largest element in $\omega$ lies in $[1/k, 1]$ and taking into account the general inequality $\log x \leq x - 1$, we derive the following pointwise bound for function $m_{\omega, \theta}$:

$$\sup_{\theta \in B(\theta^*)} \ |m_{\omega, \theta}| \leq \max_{j \leq k} \ |p_{j, \theta}/k| + \max_{j \leq k} \ 2k p_{j, \theta}. \hspace{1cm} 27$$
As expected value of the function on the right-hand side is finite by assumptions A5 and A6, we have established part (c). Thus, as shown in the aforementioned Example 19.8, the $L_1$-bracketing numbers of the class of functions $m_{\omega, \theta}$ are finite. Also note that for each fixed $(\omega, \theta)$, convergence in probability of $G_T(\omega, \theta)$ to $G(\omega, \theta)$ follows from the law of large numbers. This “pointwise” convergence, together with the finiteness of the $L_1$-bracketing numbers, yields uniform convergence (as it is shown, for example, in the proof of Theorem 2.4.1 in van der Vaart and Wellner, 1996).

To establish part (ii), we first note that the imposed continuity assumptions imply that $C(\theta)$ converges to $C(\theta^*)$, with respect to the Hausdorff distance, as $\theta \to \theta^*$. Consequently, $\omega_\theta \to \omega^*$ as $\theta \to \theta^*$. For convenience, we define $W_T(\theta) = G_T(\omega_\theta, \theta)$ and $W(\theta) = G(\omega_\theta, \theta)$. By part (i), established in the previous paragraph, $W_T$ converges to $W$ uniformly over $\theta \in \mathcal{B}(\theta^*)$. Moreover, an application of the dominated convergence theorem establishes that function $W$ is continuous at $\theta^*$, due to the pointwise continuity of the functions $m_{\omega_\theta, \theta}$ and the existence of an integrable envelope, which was established in the previous paragraph. As $W(\theta^*) = 0$, an application of the continuous mapping theorem yields $W(\hat{\theta}) \to 0$, and thus, $W_T(\hat{\theta}) \to 0$ as $T$ goes to infinity.

We now move to part (iii). Arguments similar to the ones in the previous paragraph, involving the dominated convergence theorem, establish that function $G(\cdot, \theta^*)$ is continuous, and, thus, uniformly continuous on the compact set $\mathcal{W}$. As $G(\omega^*, \theta^*) = 0$ and $\omega^*$ is the unique maximum of $G(\cdot, \theta^*)$ over the closed set $C(\theta^*)$, the maximum of $G(\cdot, \theta^*)$ over the closed set $\omega \in C(\theta^*) \setminus B_\delta(\omega^*)$ is negative. By uniform continuity of $G(\cdot, \theta^*)$, we can slightly increase the constraint set while keeping the negativity of the maximum. Recall that $C(\theta)$ converges to $C(\theta^*)$ with respect to the Hausdorff distance as $\theta \to \theta^*$. Consequently, for a sufficiently small $r_\delta$,

$$\max_{\omega \in C(\theta) \setminus B_\delta(\omega^*) \setminus B_{r_\delta}(\theta^*)} G(\omega, \theta^*) < 0.$$  

Taking advantage of the dominated convergence theorem once again, we establish that, as $\theta \to \theta^*$, function $G(\omega, \theta)$ converges to $G(\omega, \theta^*)$ uniformly over $\omega$. Thus, we can replace $G(\omega, \theta^*)$ with $G(\omega, \theta)$ in the above display (reducing the $r_\delta$ if needed) and still maintain the strict inequality.

$\square$
Proof of Theorem 3.2

Proof. We continue to borrow notation from the empirical process theory, and denote $T^{1/2}(P_T h - P h)$ by $\nu_T h$ for every function $h$. Given expressions $E_1$ and $E_2$, we write $E_1 \lesssim E_2$ to mean that there exists a finite universal constant $c$, such that $E_1 \leq c E_2$.

For simplicity of the notation, we denote densities $p_j(\cdot; \theta^*)$ by $p_j$. We restrict our attention to a closed ball around $\omega^*$, denoted by $B(\omega^*)$, whose radius is chosen to be positive, but sufficiently small to ensure $\omega_j > 0$ for every $j$ and every $\omega \in B(\omega^*)$.

We write $\dot{h}_\omega(y)$, $\ddot{h}_\omega(y)$ and $\ldots h_\omega(y)$ for the first, second and third derivative, respectively, of the function $\omega \mapsto h_\omega(y)$, evaluated at $\omega - 1$. As a consequence of the definition of $B(\omega^*)$,

$$\sup_{\omega \in B(\omega^*)} \| \dot{h}_\omega \|_\infty = \sup_{\omega \in B(\omega^*)} \max_{2 \leq j \leq k} |p_j - p_1| \lesssim 1.$$  

A similar calculation shows $\sup_{\omega \in B(\omega^*)} \| \ddot{h}_\omega \|_\infty \lesssim 1$ and $\sup_{\omega \in B(\omega^*)} \| \ldots h_\omega \|_\infty \lesssim 1$. We also have $P \ddot{h}_{\omega^*} = -V$. Note that $V$ is nonsingular, because otherwise one of the densities $p_j$ could be expressed as a linear combination of the rest of the densities, which, in view of assumption A8, would contradict the uniqueness $\omega^*$ as the solution to the population problem (12). Consequently, function $P h_\omega$ has the following two term Taylor expansion around $\omega^*$:

$$P h_\omega = P h_{\omega^*} - \frac{1}{2} (\omega - \omega^* - 1)^\top V (\omega - \omega^* - 1) + o(\|\omega - \omega^* - 1\|^2).$$  

The linear term in the above expansion disappears, because, by assumptions A8 and A9, vector $\omega^*$ is a local maximum of $P h_\omega$.

We now establish the $T^{-1/2}$ rate of convergence for $\tilde{\omega}$. Define $h_\omega = \log[p_{\omega, \theta^*}/p_{\omega^*, \theta^*}]$. According to Theorem 5.52 in van der Vaart (2000), in view of the consistency of $\tilde{\omega}$, Taylor expansion (21) and non-singularity of $V$, we only need to derive

$$E \sup_{\|\omega - \omega^*\| \leq \delta} |\nu_T(h_\omega - h_{\omega^*})| \lesssim \delta.$$  

By the $m$-dependence of $\{y_t\}$, we can write the empirical process $\nu_T$ as a sum of $m + 1$ empirical processes, where each one is based on i.i.d. random variables, such as $\{y_{t+s(m+1)}, s = 0, 1, \ldots\}$. It is sufficient to establish the above bound (and similar bounds that follow) for each such process. Taking advantage of the bound established
for $\sup_{\omega \in B(\omega^*)} \|h_\omega\|_\infty$, we derive that, for every $\omega_1 \in B(\omega^*)$ and $\omega_2 \in B(\omega^*)$,

$$\|h_{\omega_1} - h_{\omega_2}\|_\infty \lesssim \|(\omega_1)_{-1} - (\omega_2)_{-1}\|.$$

Corollary 5.53 in van der Vaart (2000) then gives bound (22) as a consequence of the inequality above (the specific bound is established in the proof of Corollary 5.53). Thus, we have proved that $\hat{\omega} = \omega^* + O_p(T^{-1/2})$.

Lemma 19.31 in van der Vaart (2000) yields $\nu_T[T^{1/2}(h_{\omega^* + v_T T^{-1/2} - h_{\omega^*}}) - v_T^T \dot{h}_{\omega^*}] = o_p(1)$ for every stochastically bounded random sequence of $(k - 1)$-dimensional vectors $v_T$. Consequently, taking advantage of the Taylor expansion of $Ph_{\omega}$ at $\omega^*$, we conclude that

$$nP_n(h_{\omega^* + v_T T^{-1/2} - h_{\omega^*}}) = -\frac{1}{2} v_T^T \tilde{V}_* v_T + v_T^T \nu_T \dot{h}_{\omega^*} + o_p(1).$$

(23)

We derive the limiting distribution for $T^{1/2}(\hat{\omega}_{-1} - \omega_{-1}^*)$ by applying Theorem 4.4 in Geyer (1994). An analysis of the proof shows that for the conclusion of the aforementioned theorem to hold, the only required assumptions are: (i) stochastic bound (23) holds for every $O_p(1)$ random sequence $v_T$; (ii) $\hat{\omega}_{-1} = \omega_{-1}^* + O_p(1)$; (iii) the constraint set $C(\theta^*)$ is Chernoff regular at $\omega_{-1}^*$. We have already established (i) and (ii). Condition (iii) is only needed to rule out pathological cases. It is satisfied in our setting, because the constraint set is determined by finitely many fourth-order polynomial inequalities. Note that $V_*^{-1/2} v_T \dot{h}_{\omega^*}$ converges in distribution to $\tilde{Z}$ by the central limit for $m$-dependent sequences. We apply the aforementioned result in Geyer (1994) to conclude that $T^{1/2}(\hat{\omega}_{-1} - \omega_{-1}^*)$ converges in distribution to the minimizer of $\frac{1}{2} v^T V_* v - v^T V_*^{1/2} \tilde{Z}$ over $v \in S^*$. The result of Theorem 3.2 follows after completing the square.

\[ \square \]