Australian National University

Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

Higher Moment Constraints for Predictive Density Combination

CAMA Working Paper 45a/2020
May 2020

Laurent Pauwels
University of Sydney
Centre for Applied Macroeconomic Analysis, ANU

Peter Radchenko
University of Sydney

Andrey L. Vasnev
University of Sydney

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, such as consistency and asymptotic distribution, are investigated theoretically and through a simulation study. 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 combinations, Predictive densities, Moment constraints, Financial data.

JEL Classification

C53, C58

Address for correspondence:

(E) cama.admin@anu.edu.au

ISSN 2206-0332

Higher Moment Constraints for Predictive Density Combinations[*]

Laurent Pauwels
University of Sydney
CAMA (ANU)

Peter Radchenko
University of Sydney

Andrey L. Vasnev
University of Sydney

April 29, 2020

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, such as consistency and asymptotic distribution, are investigated theoretically and through a simulation study. 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 combinations, Predictive densities, Moment constraints, Financial data.

JEL Codes: C53, C58

[*] We would like to thank Manabu Asai, Chia-Ling Chang, Graham Elliott, Domenico Giannone, Christian Hafner, Michael McAleer, Shiqing Ling, George Tauchen and Shaun Vahey for detailed comments and suggestions. We are also grateful to the participants of the Time Series Analysis of Higher Moments and Distributions of Financial Data conference, the 14th Symposium on Econometric Theory and Applications, and the 2019 Conference on Real-Time Data Analysis, Methods and Applications.

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 Genest and Zidek (1986), DeGroot and Mortera (1991), 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. The linear pool approach has recently been generalised to nonlinear transformations of linear pools, with beta transformations in Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013), beta-mixtures for calibration and combination in Bassetti et al. (2018), and nonlinear pools and generalised weights in Kapetanios et al. (2015). Furthermore, Billio et al. (2013) and Del Negro et al. (2016) allow the weights of the combination to account for time instabilities and estimation uncertainty. Some theoretical advances have been provided by Elliott (2011) and Chan and Pauwels (2018) for forecast point combinations. For forecast density combinations, Kapetanios et al. (2015) establishes asymptotic normality for the proposed generalised weights, however, this result also covers the case of fixed weights considered in Hall and Mitchell (2007) and Geweke and Amisano (2011). Diks et al. (2011) proposes a censored likelihood scoring rule, which is demonstrated by Opschoor et al. (2017) to outperform other methods if the tail of the distribution is the main feature of interest. Smith and Vahey (2016) investigates methodologies to forecast densities by using a copula model with asymmetric margins. These asymmetric margins are produced from empirical and skew-t distributions.

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).

Table 1: Sample skewness and kurtosis in market returns

	S&P 500	DJIA 30	Nikkei 225	FTSE 100
Skewness	-0.249	-0.121	-0.232	-0.357
Kurtosis	11.358	11.500	7.901	4.424

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.

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 under additional higher moments restrictions. We name these optimal weights derived under high moment constraints *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 of the HMC weights, namely, consistency and the asymptotic distribution. These results are also applicable to the 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.

2 Moments of the combination

2.1 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:

$$p_c(\cdot;\boldsymbol{\omega},\boldsymbol{\theta}) = \sum_{j=1}^{k} \omega_j p_j(\cdot;\boldsymbol{\theta}_j), \tag{1}$$

where $\boldsymbol{\omega} = (\omega_1,\dots,\omega_k)^\top \in \mathbb{R}^k$ is the vector of weights, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top,\dots,\boldsymbol{\theta}_k^\top)^\top$ is the combined vector of all parameters, and $\boldsymbol{\theta}_j$ is a vector of parameters of the j^{th} density, $p_j(\cdot;\boldsymbol{\theta}_j)$. For $p_c(\cdot;\boldsymbol{\omega},\boldsymbol{\theta})$ to be a density, the weights need to be nonnegative, $\omega_j \geq 0$, and sum up to one, $\sum_{j=1}^k \omega_j = 1$. The restrictions on weights are necessary when combining densities but for point forecasts the restrictions can be relaxed, see Vasnev and Wang (2019) that investigates negative weights and Granger and Ramanathan (1984) that does not require summation to one.

Whereas the first moment of the combination, μ_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 ω_j. Suppose that the j-th density has mean μ_j, variance σ_j^2, skewness γ_j, kurtosis κ_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 $p_c(\cdot; \boldsymbol{\omega}, \boldsymbol{\theta})$ 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:

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

(d) the kurtosis:

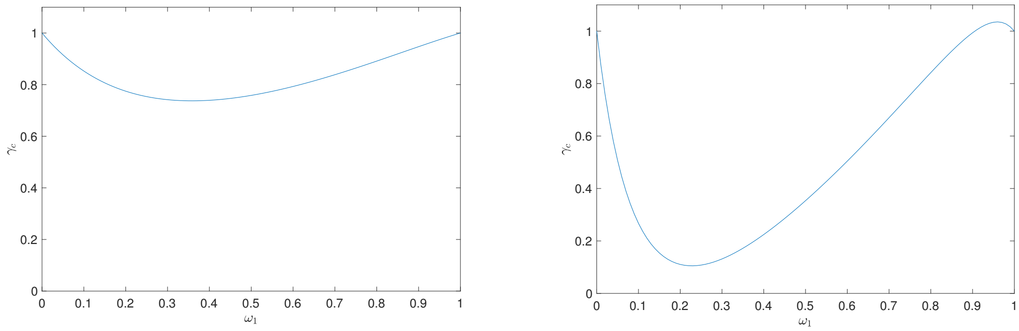
$$\kappa_c = \sum_{j=1}^{k} \omega_j \, \left[\frac{\kappa_j \, \sigma_j^4 + 4(\mu_j - \mu_c) \, \gamma_j + 6 \, (\mu_j - \mu_c)^2 \, \sigma_j^2 + (\mu_j - \mu_c)^4}{\sigma_c^4} \right], \tag{3}$$

(e) the s-th centered moment:

$$m_{c,s} = \sum_{j=1}^{k} \omega_j \sum_{l=0}^{s} \binom{s}{l} \, (\mu_j - \mu_c)^l \, m_{j,s-l}, \tag{4}$$

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

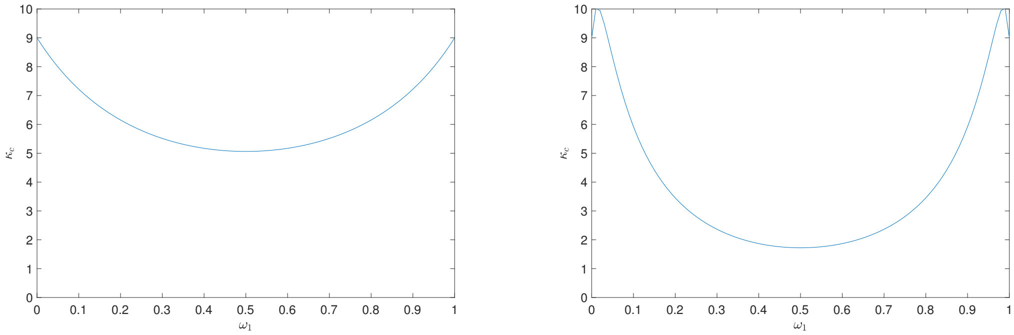
A simple numerical illustration shows that higher moments of the density combination that are relevant in empirical finance, such as skewness and kurtosis, can change considerably even when combining models with the same skewness and kurtosis. Figure 1 demonstrates the behavior of skewness, γ_c in equation (2), for different values of the weight ω_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, μ_1 and μ_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 $\mu_2 = 1$, γ_c is lower than 0.5 for ω_1 between 0.10 and 0.65.



(a) $\mu_1 = 0.1$, $\mu_2 = 1$ (b) $\mu_1 = -1$, $\mu_2 = 1$

Figure 1: γ_c as a function of ω_1 if $\sigma_1 = \sigma_2 = 1$, $\gamma_1 = \gamma_2 = 1$, and $\kappa_1 = \kappa_2 = 3$.

Similarly, Figure 2 displays the behavior of the kurtosis, κ_c, in equation (3) for different values of the weight ω_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, μ_1 and μ_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 ω_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.



(a) $\mu_1 = -1$, $\mu_2 = 1$ (b) $\mu_1 = -5$, $\mu_2 = 1$

Figure 2: κ_c as a function of ω_1 if $\sigma_1 = \sigma_2 = \sqrt{5/3}$, $\gamma_1 = \gamma_2 = 0$, and $\kappa_1 = \kappa_2 = 9$.

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

$$y_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \varepsilon_t, \tag{5}$$

with $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{kt})^\top$ and $\boldsymbol{\beta} = (1, \ldots, 1)^\top$. The regressors are standard normal random variables, $x_{jt} \sim \mathrm{N}(0, 1)$, that are independent from each other. The error 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:

$$\hat{\theta}_j = \hat{\beta}_j x_{jT}, \quad j = 1, \ldots, k, \tag{6}$$

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 ε_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 $j \leq \left\lfloor \frac{k}{2} \right\rfloor$ and are t_6 for $j > \left\lfloor \frac{k}{2} \right\rfloor$, with mean $\hat{\theta}_j$.

The density combination is given by

$$p_c(\cdot; \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{k} \omega_j p_j(\cdot; \hat{\theta}_j) \tag{7}$$

with weights ω_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:

$$\mathrm{KLIC}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \mathrm{E}\left[\log\left[\frac{f(y_T)}{p_c(y_T; \boldsymbol{\omega}, \boldsymbol{\theta})}\right]\right], \tag{8}$$

and can be estimated with its sample analogue:

$$\overline{\text{KLIC}}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \log\left[\frac{f(y_t)}{p_c(y_t; \boldsymbol{\omega}, \boldsymbol{\theta})}\right], \tag{9}$$

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

$$\begin{aligned} \text{maximize} \quad & \sum_{t=1}^{T} \log\left[\sum_{j=1}^{k} \omega_j p_j(y_t; \boldsymbol{\theta}_j)\right] \\ \text{subject to} \quad & \sum_{j=1}^{k} \omega_j = 1, \\ & \omega_j \geq 0, \quad j = 1, \ldots, k \end{aligned} \tag{10}$$

For convenience, the optimal weights that solve equation (10) are named *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 (γ_c), whereas Panel B shows the corresponding effect on the kurtosis of the combination (κ_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.

2.3 Asymptotic results

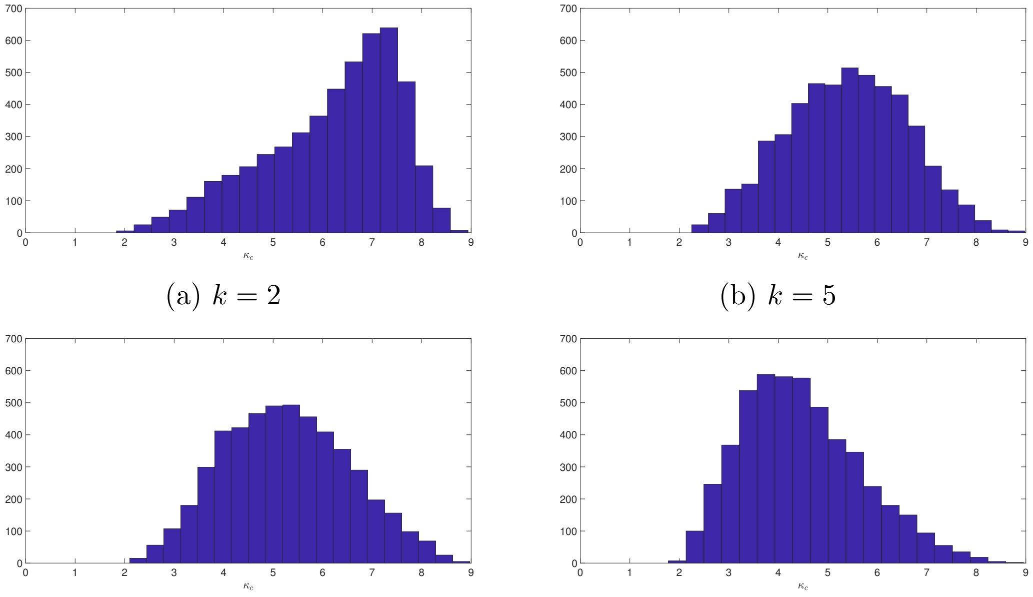
The previous section demonstrated the undesirable effect that combining densities can have on the skewness and kurtosis. Here, we examine a setting in which the densities are combined with the simple equal weights, and the number of models grows toward infinity.

Consider the general simulation setup in Section 2.2. We again focus on the linear regression model (5) but let $\boldsymbol{\beta} = (\beta, \ldots, \beta)^\top$. The regressors $\boldsymbol{x}_t$ are i.i.d. zero mean random vectors independent from the errors ε_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 $p_j(y; \boldsymbol{\theta}) = p(y - \boldsymbol{\theta})$,

Table 2: Skewness (γ_c) and kurtosis (κ_c) of the combination

	Panel A: γ_c				
Weights & # of densities	$k=2$	$k=5$	$k=10$	$k=20$	$k=30$
$\{1,\ldots,0\}$	1.000	1.000	1.000	1.000	1.000
$\left\{0.75, \frac{1-0.75}{k-1}, \ldots, \frac{1-0.75}{k-1}\right\}$	0.792	0.746	0.717	0.705	0.684
$\left\{0.50, \frac{1-0.50}{k-1}, \ldots, \frac{1-0.50}{k-1}\right\}$	0.749	0.639	0.597	0.573	0.547
$\left\{0.25, \frac{1-0.25}{k-1}, \ldots, \frac{1-0.25}{k-1}\right\}$	0.786	0.594	0.539	0.503	0.474
$\left\{\frac{1}{k}, \ldots, \frac{1}{k}\right\}$	0.749	0.592	0.527	0.481	0.449
Log score weights	0.758	0.621	0.541	0.452	0.386
	Panel B: κ_c				
Weights & # of densities	$k=2$	$k=5$	$k=10$	$k=20$	$k=30$
$\{1,\ldots,0\}$	9.000	9.000	9.000	9.000	9.000
$\left\{0.75, \frac{1-0.75}{k-1}, \ldots, \frac{1-0.75}{k-1}\right\}$	6.981	6.861	6.890	6.876	6.870
$\left\{0.50, \frac{1-0.50}{k-1}, \ldots, \frac{1-0.50}{k-1}\right\}$	6.158	5.759	5.771	5.727	5.714
$\left\{0.25, \frac{1-0.25}{k-1}, \ldots, \frac{1-0.25}{k-1}\right\}$	5.894	5.177	5.168	5.097	5.080
$\left\{\frac{1}{k}, \ldots, \frac{1}{k}\right\}$	6.158	5.107	4.989	4.865	4.840
Log score weights	6.187	5.385	5.274	4.848	4.466

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.



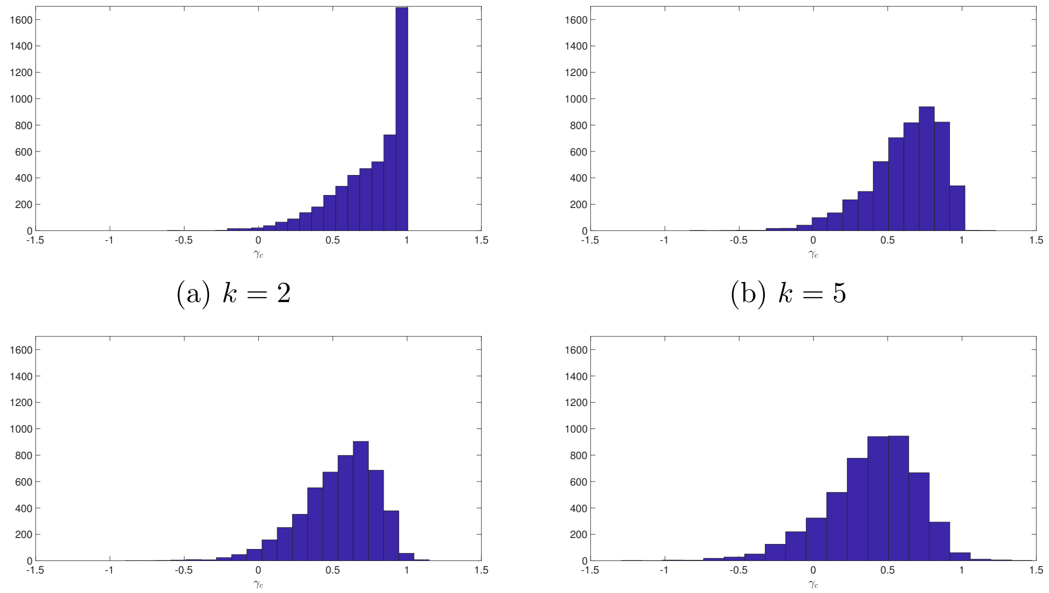
(a) $k=2$

(b) $k=5$

(c) $k=10$

(d) $k=30$

Figure 3: Distribution of the kurtosis of the density combinations (κ_c) for optimal log score weights that result from solving (10).



(a) $k = 2$

(b) $k = 5$

(c) $k = 10$

(d) $k = 30$

Figure 4: Distribution of the skewness of the density combinations (γ_c) for optimal log score weights that result from solving (10).

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 γ_c and κ_c denote its skewness and kurtosis, respectively. Denote the standard deviation, skewness, and kurtosis of density p by σ_p, γ_p, and κ_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 $g \in \{1, ..., G\}$ and let σ^2_X denote the average variance across the predictor groups: $\sigma^2_X = \sum_{g=1}^{G} \sigma^2_{X,g}/G$. We define γ_X and κ_X by analogy, as the average predictor skewness and kurtosis, respectively. Note that if $G = 1$, then σ^2_X, γ_X, and κ_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:*

$$\begin{aligned}
\gamma_c &\xrightarrow{P} \gamma_p \Big[1 + \beta R\Big]^{-3/2} + \gamma_X \Big[1 + (\beta R)^{-1}\Big]^{-3/2} \\
\kappa_c &\xrightarrow{P} \kappa_p \Big[1 + \beta R\Big]^{-2} + \kappa_X \Big[1 + (\beta R)^{-1}\Big]^{-2} + 6\Big[2 + \beta R + (\beta R)^{-1}\Big]^{-1}.
\end{aligned}$$

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 $\boldsymbol{\beta}$. When $\boldsymbol{\beta} = (1/\sqrt{k}, \ldots, 1/\sqrt{k})^\top$, 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 $\boldsymbol{\beta} = (3, \ldots, 3)^\top$, 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.

Table 3: Combined kurtosis (κ_c) for different values of β

	$\boldsymbol{\beta} = (1/\sqrt{k}, \ldots, 1/\sqrt{k})^\top$		$\boldsymbol{\beta} = (3, \ldots, 3)^\top$	
Weights & # of densities	$k = 2$	$k = 10$	$k = 2$	$k = 10$
Equal weights	6.674	7.370	3.929	2.972
Log score weights	6.680	7.009	4.013	2.899

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

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

$$\begin{aligned} \text{maximize} \quad & \sum_{t=1}^{T} \log \left[\sum_{j=1}^{k} \omega_j p_j(y_t; \boldsymbol{\theta}_j) \right] \\ \text{subject to} \quad & \sum_{j=1}^{k} \omega_j = 1, \\ & \omega_j \geq 0, \quad j = 1, \ldots, k \\ & \kappa_c \geq \underline{\kappa} \text{ and/or } \gamma_c \geq \underline{\gamma}, \end{aligned} \tag{11}$$

where the kurtosis of the combination, κ_c, is given by equation (2) and the skewness of the combination, γ_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 $\underline{\kappa}$ and $\underline{\gamma}$ 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 predictive density $p_c(\cdot;\theta)$. 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 $\underline{\kappa}$, 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 $\underline{\kappa}$.

Table 4: VaR experiment

	Panel A: Combination of models (k=2)			
Combination	Log score	$\bar{\kappa}_c$	% viol. at 1%	% viol. at 5%
HMC weights ($\underline{\kappa} = 5$)	0.989	5.016	1.358	7.536
HMC weights ($\underline{\kappa} = 5.5$)	0.977	5.501	1.076	6.355
HMC weights ($\underline{\kappa} = 6$)	0.951	6.000	1.005	6.095
Log score weights	1.000	4.493	1.665	7.468
Equal weights	0.998	4.630	1.461	7.162
	Panel B: Combination of models (k=5)			
Combination	Log score	$\bar{\kappa}_c$	% viol. at 1%	% viol. at 5%
HMC weights ($\underline{\kappa} = 5$)	0.997	5.008	1.300	6.967
HMC weights ($\underline{\kappa} = 5.5$)	0.989	5.500	1.000	6.969
HMC weights ($\underline{\kappa} = 6$)	0.965	6.000	0.536	4.523
Log score weights	1.000	4.471	1.567	7.500
Equal weights	0.997	4.542	1.433	6.900

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. $\bar{\kappa}_c$ 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).

3.3 Asymptotic Properties of HMC Weights

In this section, we establish novel results on 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. We do not require that the true predictive density is represented as a linear combination of the densities under consideration. All the proofs are provided in the Appendix.

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, $\mathcal{B}(\boldsymbol{\theta}^*)$ is a closed ball around $\boldsymbol{\theta}^*$, whose radius we are allowed to have as *arbitrarily small* but positive. The vector $\boldsymbol{\theta}^*$ 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: $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}^*$, for some fixed finite vector $\boldsymbol{\theta}^*$.

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

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

A5:

$$E \sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} \left|\log p_j(y_1;\boldsymbol{\theta})\right| < \infty \qquad \text{for } j = 1,\dots,k.$$

A6:

$$E \sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} p_j(y_1;\boldsymbol{\theta}) < \infty \qquad \text{for } j = 1,\dots,k.$$

We define $C(\boldsymbol{\theta})$ as the constraint set for the weights $\boldsymbol{\omega}$ in the optimization problem (11). We denote by $\widehat{\boldsymbol{\omega}}$ the HMC optimal weights, that is, the solution to the optimization problem (11) but with $\boldsymbol{\theta}$ replaced by $\widehat{\boldsymbol{\theta}}$. The corresponding population solution is:

$$\boldsymbol{\omega}^* = \operatorname*{arg\,min}_{\boldsymbol{\omega}\in C(\boldsymbol{\theta}^*)} \mathrm{KLIC}(\boldsymbol{\omega},\boldsymbol{\theta}^*), \tag{12}$$

where $\mathrm{KLIC}(\boldsymbol{\omega},\boldsymbol{\theta})$ is defined in (8). Theorem 3.1 establishes the consistency of $\widehat{\boldsymbol{\omega}}$.

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

From the proof, it follows that if the convergence of $\widehat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}^*$ 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 $\widehat{\boldsymbol{\omega}}$. For the simplicity of the exposition, we focus on the case $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$, which allows us to avoid imposing specific assumptions on the form of $\widehat{\boldsymbol{\theta}}$ as a function of the data. Consequently, we change assumption A2 by setting $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$ and relax assumptions A3–A6 by setting $\mathcal{B}(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta}^*\}$. 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 $\boldsymbol{\omega}^*$ are positive.

A9: The unconstrained minimizer of $\mathrm{KLIC}(\cdot,\boldsymbol{\theta}^*)$ lies in $C(\boldsymbol{\theta}^*)$.

Let $\ell^*(y) = \big(p_2(y;\boldsymbol{\theta}^*) - p_1(y;\boldsymbol{\theta}^*), ..., p_k(y;\boldsymbol{\theta}^*) - p_1(y;\boldsymbol{\theta}^*)\big)^\top / p(y;\boldsymbol{\omega}^*,\boldsymbol{\theta}^*)$, and define

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

Because the weights in all $\boldsymbol{\omega}$ 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 $\boldsymbol{\omega}$ can be expressed as a function of $\boldsymbol{\omega}_{-1} = (\omega_2, ..., \omega_k)^\top$.

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

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

$$\sqrt{T}(\widehat{\boldsymbol{\omega}}_{-1} - \boldsymbol{\omega}^*_{-1}) \xrightarrow{d} \mathcal{N}\left(0, V_*^{-1}\Sigma^* V_*^{-1}\right).$$

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

$$\sqrt{T}(\widehat{\boldsymbol{\omega}}_{-1} - \boldsymbol{\omega}^*_{-1}) \xrightarrow{d} V_*^{-1/2} Proj_{V_*^{1/2}\mathcal{S}^*} \tilde{Z}. \tag{13}$$

4 Empirical illustration

In this section, we illustrate the benefits of using the HMC optimal weights when combining density forecasts based on real data. Density forecast combination methods are often applied to financial data. Example include Geweke and Amisano (2010), Geweke and Amisano (2011), Kapetanios et al. (2015), Crisóstomo and Couso (2017), and Bassetti et al. (2018). In our application, 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

$$y_t = \mu + \sqrt{h_t}\eta_t, \quad \eta_t|\mathcal{F}_{t-1} \sim F(0,1), \tag{14}$$

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):

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

One of the main problems with EGARCH models is that they have no established analytical asymptotic properties that are independent of the error distributions considered. Specifically, the statistical properties of the (quasi-) maximum likelihood estimator of the EGARCH parameters are not available under general conditions. This issue is discussed in McAleer and Hafner (2014) and also in Chang and McAleer (2017). Typically the properties of EGARCH models have to be investigated empirically, as, for example, in Anyfantaki and Demos (2016). Some of the recent theoretical advances are made in Martinet and McAleer (2018), who show that the EGARCH(p, q) model can be derived from a stochastic process, and sufficient invertibility conditions can be stated in simple form.

Despite these known problems, EGARCH models have remained popular in empirical finance. In our empirical analysis, we include the EGARCH approach whilst acknowledging its pitfalls. Furthermore, we note that in the context of the current application, the predictive performance of the density forecasts based on the estimated EGARCH parameters is adequate in simulations (results are available upon request), on par with that of GARCH, and the same is found in the empirical results reported in Table 5.

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 $\underline{\kappa} = 5,\ 5.5$ 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:

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

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.

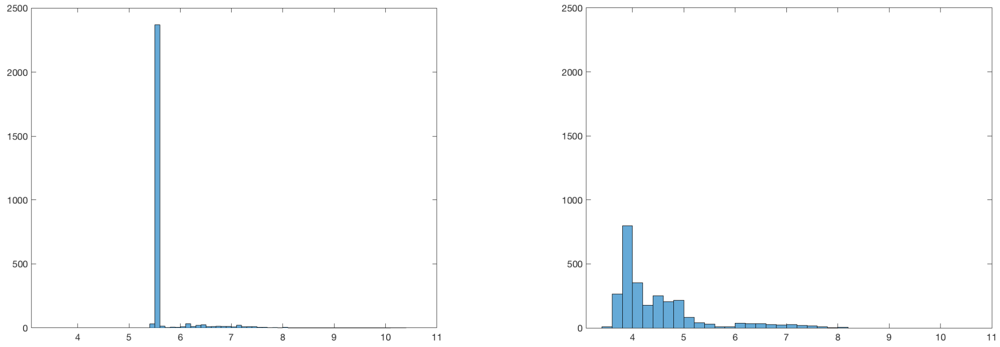
The accuracy and performance of combining density forecasts are assessed in two primary ways. First, we evaluate the entire density using the log score function. Second, we focus on evaluating the performance of the forecast combination in the left tail of the distribution by considering both the 99%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

$$\widehat{\text{VaR}}_t^{1-q} = \hat{\mu}_t + \sqrt{\hat{h}_t}\, \eta_q, \tag{18}$$

where η_q is the qth 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 $\underline{\kappa} = 5.5$ 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). The log score optimal combination produces a combined kurtosis above 5.5 (the minimum guaranteed by the HMC combination method) only 9.4% of the time. This number is remarkable as it means that the log score combination method is not able to produce a density that matches the kurtosis observed in the actual data (see Table 1).



(a) HMC ($\underline{\kappa} = 5.5$)

(b) Log score

Figure 5: Implied kurtosis of the optimal combinations

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 $\underline{\kappa}$ 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 $\underline{\kappa} = 6$

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 $\underline{\kappa} = 6$, 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

Combination	Log score	min(κ_c)	$\bar{\kappa}_c$	max(κ_c)
HMC weights ($\underline{\kappa} = 5$)	0.998	5.00	5.18	10.33
HMC weights ($\underline{\kappa} = 5.5$)	0.995	5.50	5.62	10.33
HMC weights ($\underline{\kappa} = 6$)	0.990	6.00	6.11	10.26
Log score weights	1.000	3.59	4.46	8.06
CSL weights	0.962	3.66	5.51	10.65
Equal weights	0.983	4.15	5.27	9.37
Individual models	**Log score**	**min(κ_c)**	**$\bar{\kappa}_c$**	**max(κ_c)**
GARCH (Gaussian)	0.962	3.00	3.00	3.00
GARCH (t)	0.961	4.12	4.74	9.78
GARCH (Laplace)	0.929	6.00	6.00	6.00
GARCH (Skew-t)	0.865	3.17	4.69	7.64
EGARCH (Gaussian)	0.973	3.00	3.00	3.00
EGARCH (t)	0.978	3.89	4.14	8.09
EGARCH (Laplace)	0.935	6.00	6.00	6.00
EGARCH (Skew-t)	0.859	3.25	4.10	8.35

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

Combination	# viol. at 1%	CC test	# viol. at 5%	CC test
HMC weights ($\underline{\kappa} = 5$)	45 (1.69%)	0.363	187 (6.45%)	0.248
HMC weights ($\underline{\kappa} = 5.5$)	44 (1.65%)	0.363	164 (6.15%)	0.270
HMC weights ($\underline{\kappa} = 6$)	47 (1.76%)	0.363	171 (6.37%)	0.318
Log score weights	58 (2.17%)	0.363	189 (7.09%)	0.275
CSL weights	53 (1.99%)	0.358	136 (5.17%)	0.358
Equal weights	32 (1.20%)	0.358	146 (5.47%)	0.358
Individual models	**# viol. at 1%**	**CC test**	**# viol. at 5%**	**CC test**
GARCH (Gaussian)	64 (2.40%)	0.358	153 (5.74%)	0.294
GARCH (t)	45 (1.69%)	0.358	123 (4.61%)	0.358
GARCH (Laplace)	109 (4.09%)	0.294	309 (11.59%)	0.271
GARCH (Skew-t)	68 (2.55%)	0.358	161 (6.04%)	0.358
EGARCH (Gaussian)	58 (2.17%)	0.363	159 (5.96%)	0.187
EGARCH (t)	50 (1.87%)	0.358	126 (4.72%)	0.358
EGARCH (Laplace)	99 (3.71%)	0.274	304 (11.40%)	0.120
EGARCH (Skew-t)	65 (2.44%)	0.358	156 (5.85%)	0.358

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 of the proposed HMC optimal weights, including consistency and the asymptotic distribution.

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

Anyfantaki, S. and Demos, A. (2016), "Estimation and Properties of a Time-Varying EGARCH(1,1) in Mean Model," *Econometric Reviews*, 35, 293–310.

Bassetti, F., Casarin, R., and Ravazzolo, F. (2018), "Bayesian Nonparametric Calibration and Combination of Predictive Distributions," *Journal of the American Statistical Association*, 113, 675–685.

Bates, J. M. and Granger, C. W. J. (1969), "The combination of forecasts," *Operational Research Quarterly*, 20, 451–468.

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

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

Chan, F. and Pauwels, L. L. (2018), "Some theoretical results on forecast combinations," *International Journal of Forecasting*, 34, 64 – 74.

Chang, C.-L. and McAleer, M. (2017), "The correct regularity condition and interpretation of asymmetry in EGARCH," *Economics Letters*, 161, 52 – 55.

Christoffersen, P. F. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39, 841–862.

Crisóstomo, R. and Couso, L. (2017), "Financial density forecasts: A comprehensive comparison of risk-neutral and historical schemes," *Journal of Forecasting*, 37, 589–603.

DeGroot, M. and Mortera, J. (1991), "Optimal Linear Opinion Pools," *Management Science*, 37, 546–558.

Del Negro, M., Hasegawa, R. B., and Schorfheide, F. (2016), "Dynamic prediction pools: An investigation of financial frictions and forecasting performance," *Journal of Econometrics*, 192, 391–405.

Diks, C., Panchenko, V., and van Dijk, D. (2011), "Likelihood-based scoring rules for comparing density forecasts in tails," *Journal of Econometrics*, 163, 215 – 230.

Elliott, G. (2011), "Averaging and the Optimal Combination of Forecasts," University of California, San Diego.

Genest, C. and Zidek, J. (1986), "Combining Probability Distributions: A critique and an annotated bibliography," *Statistical Science*, 1, 114–135.

Geweke, J. and Amisano, G. (2010), "Comparing and evaluating Bayesian predictive distributions of asset returns," *International Journal of Forecasting*, 26, 216 – 230, special Issue: Bayesian Forecasting in Economics.

— (2011), "Optimal prediction pools," *Journal of Econometrics*, 164, 130 – 141, annals Issue on Forecasting.

Geyer, C. J. (1994), "On the asymptotics of constrained M-estimation," *The Annals of Statistics*, 22, 1993–2010.

Gneiting, T. and Ranjan, R. (2011), "Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules," *Journal of Business & Economic Statistics*, 29, 411–422.

— (2013), "Combining Predictive Distributions," *Electronic Journal of Statistics*, 7, 1747–1782.

Granger, C. W. J. and Ramanathan, R. (1984), "Improved Methods of Combining Forecasts," *Journal of Forecasting*, 3, 197–204.

Hall, S. G. and Mitchell, J. (2007), "Combining density forecasts," *International Journal of Forecasting*, 23, 1 – 13.

Hansen, B. E. (1994), "Autoregressive Conditional Density Estimation," *International Economic Review*, 35, 705–730.

Jondeau, E. and Rockinger, M. (2009), "The Impact of Shocks on Higher Moments," *Journal of Financial Econometrics*, 7, 77–105.

Jore, A. S., Mitchell, J., and Vahey, S. P. (2010), "Combining forecast densities from VARs with uncertain instabilities," *Journal of Applied Econometrics*, 25, 621–634.

Kapetanios, G., Mitchell, J., Price, S., and Fawcett, N. (2015), "Generalised density forecast combinations," *Journal of Econometrics*, 188, 150–165.

Knight, K. and Fu, W. (2000), "Asymptotics for Lasso-type Estimators," *Annals of statistics*, 1356–1378.

Ling, S. and McAleer, M. (2003), "Asymptotic theory for a vector ARMA-GARCH model," *Econometric Theory*, 19, 278–308.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018), "The M4 Competition: Results, findings, conclusion and way forward," *International Journal of Forecasting*, 34, 802 – 808.

Martinet, G. G. and McAleer, M. (2018), "On the invertibility of EGARCH(p, q)," *Econometric Reviews*, 37, 824–849.

McAleer, M. and Hafner, C. M. (2014), "A One Line Derivation of EGARCH," *Econometrics*, 2, 92–97.

Nelson, D. B. (1991), "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59, 347–70.

Opschoor, A., van Dijk, D., and van der Wel, M. (2017), "Combining density forecasts using focused scoring rules," *Journal of Applied Econometrics*, 32, 1298–1313.

Polanski, A. and Stoja, E. (2010), "Incorporating higher moments into value-at-risk forecasting," *Journal of Forecasting*, 29, 523–535.

Radchenko, P. (2005), "Reweighting the Lasso," *2005 Proceedings of the American Statistical Association.*

Ranjan, R. and Gneiting, T. (2010), "Combining Probability Forecasts," *Journal of the Royal Statistical Society Series B*, 72, 71–91.

Smith, M. S. and Vahey, S. P. (2016), "Asymmetric Forecast Densities for U.S. Macroeconomic Variables from a Gaussian Copula Model of Cross-Sectional and Serial Dependence," *Journal of Business & Economic Statistics*, 34, 416–434.

Timmermann, A. (2006), "Forecast Combinations," in *Handbook of Economic Forecasting*, eds. Elliott, G., Granger, C. W. J., and Timmermann, A., Amsterdam, Netherlands: Elsevier, chap. 4, pp. 135–196.

van der Vaart, A. W. (2000), *Asymptotic statistics*, Cambridge university press.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak convergence and Empirical Processes*, Springer Series in Statistics, Springer-Verlag, New York.

Vasnev, A. L. and Wang, W. (2019), "What to do with negative weights when combining forecasts?" *mimeo.*

Wallis, K. (2005), "Combining density and interval forecasts: a modest proposal," *Oxford Bulletin of Economics and Statistics*, 67, 983–994.

Appendix: Proofs

Proof of Theorem 2.2

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

$$\bar{\mu} = (1/k)\sum_{j=1}^{k}\hat{\mu}_{jT}, \qquad \sigma_{\mu}^{2} = (1/k)\sum_{j=1}^{k}\Big(\hat{\mu}_{jT} - \bar{\mu}\Big)^{2}, \qquad \gamma_{\mu} = (1/k)\sum_{j=1}^{k}\frac{\Big(\hat{\mu}_{jT} - \bar{\mu}\Big)^{3}}{\sigma_{\mu}^{3}}$$

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

$$\gamma_{c} = \gamma_{p}\Big[1 + \tilde{R}\Big]^{-3/2} + \gamma_{\mu}\Big[1 + (\tilde{R})^{-1}\Big]^{-3/2}. \tag{19}$$

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 $\hat{\mu}_{jT} = \theta X_{jT} + \eta_{jT}\xi_{jT}$ 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 $\max_{j\leq k}|\xi_{jT}|$ is $O_{p}(kT^{-1/2})$, which simplifies to $o_{p}(1)$ by the assumptions on k and T. A similar argument, together with the law of large numbers, gives $\max_{j\leq k}|\eta_{jT}| = O_{p}(1)$. It follows that

$$\sigma_{\mu}^{2} = \theta^{2}(1/k)\sum_{j=1}^{k}X_{jT}^{2} + o_{p}(1).$$

Another application of the law of large numbers gives $\sigma_{\mu}^{2} = \theta^{2}\sigma_{X}^{2} + o_{p}(1)$, which implies $\tilde{R} = \theta R + o_{p}(1)$. 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 $\tilde{R}$ and γ_{μ} 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 Ph for $Eh(y_1)$. Also, for the remainder of the proof, all of the $\boldsymbol{\omega}$ are assumed to lie in the set $\mathcal{W} = \{\boldsymbol{\omega} : \sum_{j\le k}\omega_j = 1,\, \omega_j \ge 0, j = 1, ..., k\}$.

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

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

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

$$\Delta_T = \sup_{\boldsymbol{\omega}\in\mathcal{W},\,\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} |G_T(\boldsymbol{\omega},\boldsymbol{\theta}) - G(\boldsymbol{\omega},\boldsymbol{\theta})|.$$

It follows from parts (i) and (ii) of Lemma 5.1 that, with probability tending to one,

$$G(\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}) \ge G_T(\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}) - \Delta_T \ge G_T(\boldsymbol{\omega}_{\widehat{\boldsymbol{\theta}}},\widehat{\boldsymbol{\theta}}) - \Delta_T = o_p(1). \tag{20}$$

We now argue that the above stochastic bound implies convergence of $\widehat{\boldsymbol{\omega}}$ to $\boldsymbol{\omega}^*$, which is a zero of the function $G(\cdot,\boldsymbol{\theta}^*)$, as well as its maximum over the constraint set $C(\boldsymbol{\theta}^*)$. Fix an arbitrary positive δ and let $B_\delta(\boldsymbol{\omega}^*)$ denote an open ball of radius δ around $\boldsymbol{\omega}^*$. It follows from part (iii) of Lemma 5.1 that there exists a positive constant c_δ, such that $\max_{\boldsymbol{\omega}\in C(\widehat{\boldsymbol{\theta}})\setminus B_\delta(\boldsymbol{\omega}^*)} G(\boldsymbol{\omega},\widehat{\boldsymbol{\theta}}) < -c_\delta$ with probability tending to one. However, stochastic bound (20) implies $G(\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}) > -c_\delta$ with probability tending to one. Hence, with probability tending to one, $\widehat{\boldsymbol{\omega}} \in B_\delta(\boldsymbol{\omega}^*)$. As this argument holds for every positive δ, we have established that $\widehat{\boldsymbol{\omega}}$ converges to $\boldsymbol{\omega}^*$ in probability. □

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) $G_T(\boldsymbol{\omega}_{\widehat{\boldsymbol{\theta}}},\widehat{\boldsymbol{\theta}}) = o_p(1)$

(iii) Given a positive δ, there exists a positive constant r_δ, such that

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

Proof. We start with part (i). For convenience, we denote functions $\log[p_{\boldsymbol{\omega},\boldsymbol{\theta}}/p^*_{\boldsymbol{\omega},\boldsymbol{\theta}}]$ by $m_{\boldsymbol{\omega},\boldsymbol{\theta}}$ and functions $p_j(\cdot;\boldsymbol{\theta})$ by $p_{j,\boldsymbol{\theta}}$. We first show that the class of functions $m_{\boldsymbol{\omega},\boldsymbol{\theta}}$ is pointwise compact in the sense of Example 19.8 in van der Vaart (2000). Specifically, (a) the map $(\boldsymbol{\omega},\boldsymbol{\theta}) \mapsto m_{\boldsymbol{\omega},\boldsymbol{\theta}}(y)$ is continuous for each fixed y; (b) $(\boldsymbol{\omega},\boldsymbol{\theta})$ 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 $\boldsymbol{\omega}$ 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 $m_{\boldsymbol{\omega},\boldsymbol{\theta}}$:

$$\sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} \left|m_{\boldsymbol{\omega},\boldsymbol{\theta}}\right| \leq \max_{j\leq k} \sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} 2\left|\log[p_{j,\boldsymbol{\theta}}/k]\right| + \max_{j\leq k} \sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} 2kp_{j,\boldsymbol{\theta}}.$$

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 $m_{\boldsymbol{\omega},\boldsymbol{\theta}}$ are finite. Also note that for each fixed $(\boldsymbol{\omega},\boldsymbol{\theta})$, convergence in probability of $G_T(\boldsymbol{\omega},\boldsymbol{\theta})$ to $G(\boldsymbol{\omega},\boldsymbol{\theta})$ 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 $C(\boldsymbol{\theta})$ converges to $C(\boldsymbol{\theta}^*)$, with respect to the Hausdorff distance, as $\boldsymbol{\theta} \to \boldsymbol{\theta}^*$. Consequently, $\boldsymbol{\omega}_\theta \to \boldsymbol{\omega}^*$ as $\theta \to \theta^*$. For convenience, we define $W_T(\boldsymbol{\theta}) = G_T(\boldsymbol{\omega}_{\boldsymbol{\theta}},\boldsymbol{\theta})$ and $W(\boldsymbol{\theta}) = G(\boldsymbol{\omega}_{\boldsymbol{\theta}},\boldsymbol{\theta})$. By part (i), established in the previous paragraph, W_T converges to W uniformly over $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^*)$. Moreover, an application of the dominated convergence theorem establishes that function W is continuous at $\boldsymbol{\theta}^*$, due to the pointwise continuity of the functions $m_{\boldsymbol{\omega}_{\boldsymbol{\theta}},\boldsymbol{\theta}}$ and the existence of an integrable envelope, which was established in the previous paragraph. As $W(\boldsymbol{\theta}^*) = 0$, an application of the continuous mapping theorem yields $W(\widehat{\boldsymbol{\theta}}) \to 0$, and thus, $W_T(\widehat{\boldsymbol{\theta}}) \to 0$ 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 $G(\cdot,\boldsymbol{\theta}^*)$ is continuous, and, thus, uniformly continuous on the compact set $\mathcal{W}$. As $G(\boldsymbol{\omega}^*,\boldsymbol{\theta}^*) = 0$ and $\boldsymbol{\omega}^*$ is the unique maximum of $G(\cdot,\boldsymbol{\theta}^*)$ over the closed set $C(\boldsymbol{\theta}^*)$, the maximum of $G(\cdot,\boldsymbol{\theta}^*)$ over the closed set $\boldsymbol{\omega} \in C(\boldsymbol{\theta}^*) \setminus B_\delta(\boldsymbol{\omega}^*)$ is negative. By uniform continuity of $G(\cdot,\boldsymbol{\theta}^*)$, we can slightly increase the constraint set while keeping the negativity of the maximum. Recall that $C(\boldsymbol{\theta})$ converges to $C(\boldsymbol{\theta}^*)$ with respect to the Hausdorff distance as $\boldsymbol{\theta} \to \boldsymbol{\theta}^*$. Consequently, for a sufficiently small r_δ,

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

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

□

Proof of Theorem 3.2

Proof. We continue to borrow notation from the empirical process theory, and denote $T^{1/2}(P_T h - Ph)$ 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 $E_1 \leq cE_2$. For simplicity of the notation, we denote densities $p_j(\cdot; \boldsymbol{\theta}^*)$ by $p_j(\cdot)$. We restrict our attention to a closed ball around $\boldsymbol{\omega}^*$, denoted by $\mathcal{B}(\boldsymbol{\omega}^*)$, whose radius is chosen to be positive, but sufficiently small to ensure $\boldsymbol{\omega}_j > 0$ for every j and every $\boldsymbol{\omega} \in \mathcal{B}(\boldsymbol{\omega}^*)$.

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

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

A similar calculation shows $\sup_{\boldsymbol{\omega} \in \mathcal{B}(\boldsymbol{\omega}^*)} \|\ddot{h}_{\boldsymbol{\omega}}\|_\infty \lesssim 1$ and $\sup_{\boldsymbol{\omega} \in \mathcal{B}(\boldsymbol{\omega}^*)} \|\dddot{h}_{\boldsymbol{\omega}}\|_\infty \lesssim 1$. We also have $P\ddot{h}_{\boldsymbol{\omega}^*} = -V_*$. 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 $\boldsymbol{\omega}^*$ as the solution to the population problem (12). Consequently, function $Ph_{\boldsymbol{\omega}}$ has the following two term Taylor expansion around $\boldsymbol{\omega}^*$:

$$Ph_{\boldsymbol{\omega}} = Ph_{\boldsymbol{\omega}^*} - \frac{1}{2}(\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1})^\top V_* (\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1}) + o(\|\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1}\|^2). \tag{21}$$

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

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

$$E \sup_{\|\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1}\| \leq \delta} \left| \nu_T (h_{\boldsymbol{\omega}} - h_{\boldsymbol{\omega}^*}) \right| \lesssim \delta. \tag{22}$$

By the m-dependence of $\{y_t\}$, we can write the empirical process ν_T as a sum of $m + 1$ empirical processes, where each one is based on i.i.d. random variables, such as

$\{y_{1+s(m+1)}, s = 0, 1, ...\}$. It is sufficient to establish the above bound (and similar bounds that follow) for each such process. Taking advantage of the bound established for $\sup_{\boldsymbol{\omega}\in\mathcal{B}(\boldsymbol{\omega}^*)} \|\dot{h}_{\boldsymbol{\omega}}\|_\infty$, we derive that, for every $\boldsymbol{\omega}_1 \in \mathcal{B}(\boldsymbol{\omega}^*)$ and $\boldsymbol{\omega}_2 \in \mathcal{B}(\boldsymbol{\omega}^*)$,

$$\|h_{\boldsymbol{\omega}_1} - h_{\boldsymbol{\omega}_2}\|_\infty \lesssim \|(\boldsymbol{\omega}_1)_{-1} - (\boldsymbol{\omega}_2)_{-1}\|\,.$$

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 $\widehat{\boldsymbol{\omega}} = \boldsymbol{\omega}^* + O_p(T^{-1/2})$.

We establish the limiting distribution by the standard approach of applying a uniform limit theorem to the appropriately rescaled and reparametrized criterion function (van der Vaart, 2000; Knight and Fu, 2000; Radchenko, 2005). Lemma 19.31 in van der Vaart (2000) yields $\nu_T[T^{1/2}(h_{\boldsymbol{\omega}^*+v_T T^{-1/2}} - h_{\boldsymbol{\omega}^*}) - v_T^\top \dot{h}_{\boldsymbol{\omega}^*}] = o_p(1)$ for every stochastically bounded random sequence of $(k-1)$-dimensional vectors v_T. Consequently, taking advantage of the Taylor expansion of $Ph_{\boldsymbol{\omega}}$ at $\boldsymbol{\omega}^*$, we conclude that

$$nP_n(h_{\boldsymbol{\omega}^*+v_T T^{-1/2}} - h_{\boldsymbol{\omega}^*}) = -\frac{1}{2}v_T^\top V_* v_T + v_T^\top \nu_T \dot{h}_{\boldsymbol{\omega}^*} + o_p(1). \tag{23}$$

We derive the limiting distribution for $T^{1/2}(\widehat{\boldsymbol{\omega}}_{-1} - \boldsymbol{\omega}^*_{-1})$ 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) $\widehat{\boldsymbol{\omega}}_{-1} = \boldsymbol{\omega}^*_{-1} + O_p(1)$; (iii) the constraint set $C(\boldsymbol{\theta}^*)$ is Chernoff regular at $\boldsymbol{\omega}^*_{-1}$. 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 $V_*^{-1/2}\nu_T \dot{h}_{\boldsymbol{\omega}^*}$ 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 $T^{1/2}(\widehat{\boldsymbol{\omega}}_{-1} - \boldsymbol{\omega}^*_{-1})$ converges in distribution to the minimizer of $\frac{1}{2}v^\top V_* v - v^\top V_*^{1/2}\tilde{Z}$ over $v \in S^*$. The result of Theorem 3.2 follows after completing the square.

□