Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

High Moment Constraints for Predictive Density Combination

CAMA Working Paper 45/2020
May 2020
(Original Title: *Higher Moment Constraints for Predictive Density Combination*)

Revised in June 2023

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

Peter Radchenko
University of Sydney

Andrey L. Vasnev
University of Sydney

Abstract

Financial data typically exhibit asymmetry and heavy tails, which makes forecasting the entire density of the returns critically important. We investigate the effects of aggregating, or combining, predictive densities and find that even if the individual densities are skewed and/or heavy-tailed, the combined density often has significantly reduced skewness and kurtosis. This phenomenon has important implications for measuring downside risk in financial assets. When forecasting financial risk, recently proposed combination methods have focused on specific regions of the density support. We propose an alternative approach, which modifies the popular Log-Score weighting scheme by introducing data-driven constraints on the combination weights that control the skewness and kurtosis of the resulting predictive density. An empirical application using S&P 500 daily index returns demonstrates that the corresponding skewness and kurtosis successfully track the respective sample characteristics of the returns over time. Moreover, the proposed approach outperforms its natural competitors at forecasting the 1% Value-at-Risk for a broad range of estimation-window sizes.

Keywords

Forecasting, Forecast combinations, Predictive densities, Moment constraints, Financial data

JEL Classification

C53, C58

Address for correspondence:

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

ISSN 2206-0332

High Moment Constraints for Predictive Density Combinations[*]

Laurent Pauwels
NYUAD, University of Sydney,
and CAMA (ANU)

Peter Radchenko
University of Sydney

Andrey L. Vasnev
University of Sydney

June 30, 2023

Abstract

Financial data typically exhibit asymmetry and heavy tails, which makes forecasting the entire density of the returns critically important. We investigate the effects of aggregating, or combining, predictive densities and find that even if the individual densities are skewed and/or heavy-tailed, the combined density often has significantly reduced skewness and kurtosis. This phenomenon has important implications for measuring downside risk in financial assets. When forecasting financial risk, recently proposed combination methods have focused on specific regions of the density support. We propose an alternative approach, which modifies the popular Log-Score weighting scheme by introducing data-driven constraints on the combination weights that control the skewness and kurtosis of the resulting predictive density. An empirical application using S&P 500 daily index returns demonstrates that the corresponding skewness and kurtosis successfully track the respective sample characteristics of the returns over time. Moreover, the proposed approach outperforms its natural competitors at forecasting the 1% Value-at-Risk for a broad range of estimation-window sizes.

Keywords: Forecasting, 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.

1 Introduction

For risk managers, investors, and regulators alike, forecasting financial risk and asset returns is central to their market activities. When forecasting risk, 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) and Expected Shortfall (ES) forecasting (Polanski and Stoja, 2010), 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 high moments, such skewness and kurtosis, respectively, is crucial.

As density forecasts can be produced from a large range of financial models, forecasters are typically faced with multiple options to construct 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) and Wang et al. (2023) provide an extensive overview.

We ask the question: what happens to the moments of the combination when multiple predictive densities are combined? Specifically, what are the implications for high moments such as the skewness and kurtosis of the combination? The question is very important because the majority of financial returns on assets exhibits asymmetry and heavy tails as shown in Table 1 with some sample moments of some of the main stock market indices.[1]

Table 1: Sample skewness and kurtosis in market returns

	S&P 500	DJIA 30	Nikkei 225	FTSE 100
Skewness	-0.215	-0.063	-0.522	-0.366
Kurtosis	11.315	11.389	14.171	10.223

Notes: The values reported are for the daily returns of the market indices from January 3, 2000, until December 4, 2020. The data are from the "Realized Library" of the Oxford-Man Institute.

We answer the question by analyzing the impact of combining densities on high moments theoretically and numerically. 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.

[1] A similar table is reported in Jondeau and Rockinger (2009).

We propose to overcome this issue by restricting high-order moments when estimating the combination weights. 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 high 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 combination density and the HMC weights, namely, consistency and the rate of convergence. These results are also applicable to the weights proposed by Hall and Mitchell (2007) and Geweke and Amisano (2011). Second, we provide an empirical illustration in forecasting the density of the conditional returns of the S&P 500 index. The conditional returns are forecast using several GARCH-type models, which are regularly employed in the applied financial econometrics literature. This illustration is especially relevant as the S&P 500 exhibits heavy tails and skewness (see Table 1). We evaluate the proposed combined predictive density on its overall performance in terms the log-score and demonstrate that the skewness and kurtosis of this density are successful at tracking their sample counterparts over time. We also evaluate the performance of the proposed HMC approach in the tails, in terms of forecasting Value-at-Risk, demonstrating that HMC outperforms its natural competitors with respect to the accuracy of the 1% VaR forecasts. Overall, the empirical results provide convincing support for the proposed methodology.

Until recently, most of the literature focused on point forecasts, and the treatment of predictive density combinations was sparse. Some of the earliest contributions addressing the problem of combining predictive densities are 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 the (unknown) true density, as measured by the Kullback–Leibler information criterion (KLIC). Geweke and Amisano (2011) used Bayesian methods and provided some theoretical justifications for using optimal weighting schemes in linear pools of models. The linear pool approach has recently been generalised and extended, 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. On the theoretical front, Kapetanios et al. (2015) establish asymptotic normality for the proposed generalised weights, however, their 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 the empirical and skew-t distributions.

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 a new approach for constructing a forecast density combination under higher-order moment constraints and studies its statistical properties. Section 4 discusses an empirical application for the S&P 500 index, and Section 5 provides a conclusion.

2 Motivation

2.1 Behavior of high moments in combinations

We start by describing the behavior of the moments of a density combination. A simple way to combine k densities is to aggregate them linearly into one density as follows:

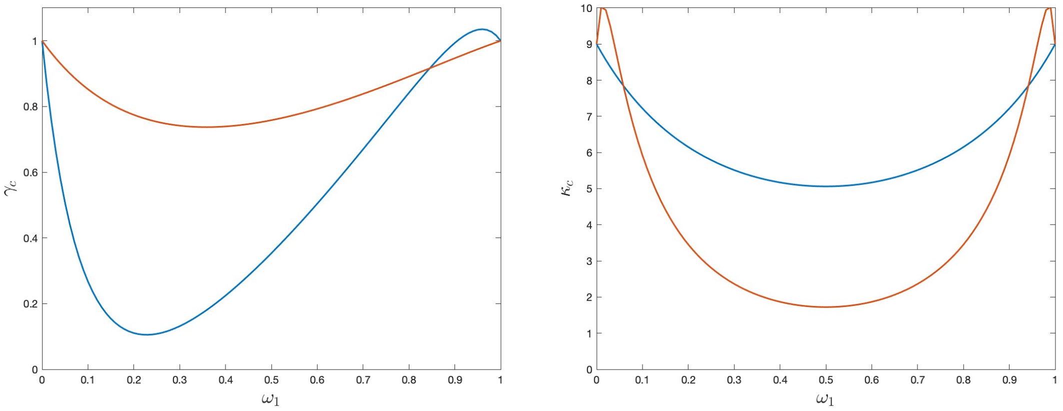
$$p_c(\cdot;\boldsymbol{\omega},\boldsymbol{\theta}) = \sum_{j=1}^{k} \omega_j p_j(\cdot;\boldsymbol{\theta}_j), \tag{1}$$

where $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_k)^\top \in \mathbb{R}^k$ is the vector of weights, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top,\ldots,\boldsymbol{\theta}_k^\top)^\top$ 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 Radchenko et al. (2023), which investigates negative weights, and Granger and Ramanathan (1984), which does not require summation to one.

A priori, the kind of impact that combining k densities (or models) would have on the higher moments of the resulting combined density is not obvious. Whereas the first moment of the combination, μ_c, is simply a linear combination of k individual density means, other moments have a more complicated nonlinear dependence on the parameters of the individual densities.[2] The simple numerical illustration below shows that the high moments of the density combination relevant in empirical finance, such as skewness and kurtosis, can change considerably even when combining models with the same skewness and kurtosis. Suppose that the j-th density has mean μ_j, variance σ_j^2, skewness γ_j and kurtosis κ_j. Figure 1a demonstrates the behavior of combination skewness, γ_c, for different

[2] Proposition A.1 in the online supplement provides formulas for the moments of the aggregate density.

values of the weight ω_1 when combining two similar distributions, such as a skewed normal. 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 $\mu_1 = -1$ and $\mu_2 = 1$, γ_c is lower than 0.5 for ω_1 between 0.10 and 0.65. If $\mu_1 = 0.1$ and $\mu_2 = 1$, then for $\omega_1 = 0.35$, the skewness of the combination is approximately 0.75.



(a) Skewness

(b) Kurtosis

Figure 1: The values of γ_c are depicted as a function of ω_1 for $\mu_1 = -1$, $\mu_2 = 1$ (blue) and $\mu_1 = 0.1$, $\mu_2 = 1$ (orange) with $\sigma_1 = \sigma_2 = 1$, $\gamma_1 = \gamma_2 = 1$, and $\kappa_1 = \kappa_2 = 3$. The values of κ_c are graphed as a function of ω_1 for $\mu_1 = -1$, $\mu_2 = 1$ (blue) and $\mu_1 = -5$, $\mu_2 = 1$ (orange) with $\sigma_1 = \sigma_2 = \sqrt{5/3}$, $\gamma_1 = \gamma_2 = 0$, and $\kappa_1 = \kappa_2 = 9$.

Similarly, Figure 1b displays the behavior of the kurtosis, κ_c, 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. When the means are $\mu_1 = -1$, $\mu_2 = 1$ and $\omega_1 = 0.5$, the kurtosis of the combination reduces to approximately 5. Additionally, if $\mu_1 = -5$, $\mu_2 = 1$, then the same weight as before, $\omega_1 = 0.5$, removes the heavy tails altogether. Naturally, when ω_1 is close to the boundary (0 or 1), only one density is selected, and the moments of the combination essentially equal those of the individual density.

2.2 Equally weighted combinations

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

Suppose that we have k density forecasts. The mean, variance, skewness, and kurtosis parameters of these forecasts are μ_j, v_j, γ_j, and κ_j, respectively, for $j = 1, ..., k$ (for compactness, we use v instead of σ^2 to denote the variance in this section.) We write μ^*, v^*, γ^*, and κ^* for the corresponding parameters of the true density. We assume that $\{\mu_j\}$, $\{v_j\}$, $\{\gamma_j\}$, and $\{\kappa_j\}$ are independent collections of i.i.d. random variables, such that $E[\mu_j] = \mu^*$ and $E[v_j] = v^*$ (we do not require that γ_j and κ_j are unbiased). We let μ_μ, v_μ, γ_μ, and κ_μ denote the mean, variance, skewness, and kurtosis, respectively, of the underlying distribution for μ_j, and we define the corresponding quantities for v_j, γ_j and κ_j analogously. We define $R = v_\mu/v^*$ and write ξ for a random variable with the same distribution as $\sqrt{v_j/v^*}$. We let γ_c and κ_c denote the skewness and kurtosis, respectively, of the equally weighted combination of the k density forecasts. The following result on the behavior of γ_c and κ_c is proved in the online supplement.

Theorem 2.1. *As $k \to \infty$,*

$$\gamma_c \xrightarrow{P} \mu_\gamma \, E\big[\xi^3\big]\big[1+R\big]^{-3/2} + \gamma_\mu\big[1+R^{-1}\big]^{-3/2}$$
$$\kappa_c \xrightarrow{P} \mu_\kappa\big[1+v_v/v^{*2}\big]\big[1+R\big]^{-2} + \kappa_\mu\big[1+R^{-1}\big]^{-2} + 6R\big[1+R\big]^{-2}.$$

Theorem 2.1 implies that that the limiting skewness and kurtosis of the combination can be significantly different from the corresponding parameters of the true density, γ^* and κ^*. To illustrate this point, we now consider two specific asymptotic scenarios.

Corollary 2.2. *If $v_\mu/v^* \to \infty$ as $k \to \infty$, then $\gamma_c \xrightarrow{P} \gamma_\mu$ and $\kappa_c \xrightarrow{P} \kappa_\mu$. Alternatively, if $v_\mu/v^* \to 0$ and $v_v/v^{*2} \to 0$, then $\gamma_c \xrightarrow{P} \mu_\gamma$ and $\kappa_c \xrightarrow{P} \mu_\kappa$.*

The above result shows that when the variance of the mean forecasts, v_μ, is large relative to the variance of the true density, v^*, the kurtosis of the combination is close to the kurtosis of the mean forecasts, κ_μ, rather than the kurtosis of the true density, κ^*. Alternatively, when the variances of the forecasts for both the mean and the variance are small relative to the variance of the true density, the kurtosis of the combination is close to the average kurtosis of the individual density forecasts. This average will be different from the true kurtosis unless the collection of individual densities is curated carefully. Similar observations hold for the skewness of the combination.

Numerical example. Figure 2 illustrates the discussion above by presenting the distributions (i.e., estimated densities) of the skewness (γ_c) and kurtosis (κ_c) resulting from combining $k = 10$ densities using three sets of weights $\boldsymbol{\omega}$. In the first set, $\omega_1 = 0.5$, while in the second set, $\omega_1 = 0.25$. The remaining weights, $j = 2, \ldots, k$, are spread equally: $\omega_j = \frac{1-\omega_1}{k-1}$. The third set weights all k densities equally: $\omega_j = 1/k$. The estimated densities are produced based on 5000 replications. For the kurtosis experiment, half of the individual distributions are mean zero t_ν distributions, where the degrees of freedom ν are drawn uniformly from the interval $[5, 6]$, and the other half are mean zero Normal with variances drawn uniformly from $[0.5, 1.5]$. For the skewness experiment, we use the skewed t distribution of Hansen (1994), with mean zero, variance 1, and skewness that uniformly varies between 0.5 and 1. The left panel of Figure 2 illustrates how including symmetric densities in the combination decreases the resulting skewness; the right panel of Figure 2 displays a similar phenomenon for the kurtosis. These observations are in line with the results of Corollary 2.2.

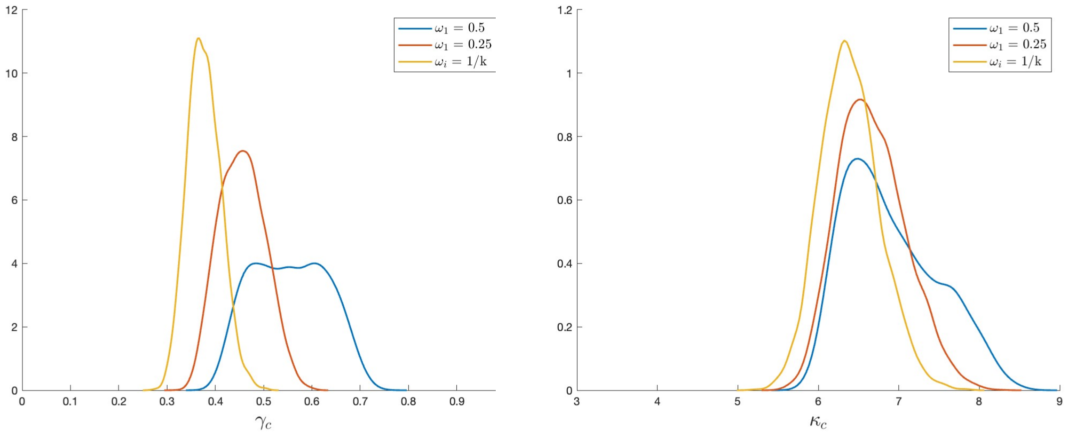


Figure 2: Distribution of the skewness (γ_c) and kurtosis (κ_c) of $k = 10$ combined densities with three sets of weights: $\omega_1 = 0.50,\ 0.25$, and equal weights, where $\omega_i = \frac{1-\omega_1}{k-1}$ for $i = 2, \ldots, k$. The distributions are drawn with 5000 replications.

Special case: regression setup. We now focus on the linear regression model

$$y_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \varepsilon_t, \qquad t = 1, \ldots, T-1, \tag{2}$$

with $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{kt})^\top$ and $\boldsymbol{\beta} = (\beta, \ldots, \beta)^\top$. 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 produce k conditional mean forecasts for y_T using $\hat{\mu}_j = \hat{\beta}_j x_{jT},\ j = 1, \ldots, k$, where $\hat{\beta}_j$ is the estimate of the slope coefficient in the simple linear regression model with only the j-th predictor. Then, we let $p_j(y) = p(y - \hat{\mu}_j)$ be the j-th individual density forecast, where p is a known density with a zero mean. For example, p could be the density of the errors in model (2). We consider the equally weighted combination of these k density forecasts and let γ_c and κ_c denote its skewness and kurtosis, respectively.

We denote the standard deviation, skewness, and kurtosis of density p by v_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. We assume that the number of predictor groups, G, stays constant as the number of predictors tends to infinity. We write $v_{X,g}$ for the variance of each predictor in group $g \in \{1, ..., G\}$ and let v_X denote the average variance across the predictor groups: $v_X = \sum_{g=1}^{G} v_{X,g}/G$. We define γ_X and κ_X by analogy, as the average predictor skewness and kurtosis, respectively. Note that if $G = 1$, then v_X, γ_X, and κ_X are simply the variance, skewness, and kurtosis of each individual predictor. Let $R = v_X/v_p$, and assume that all of the quantities defined in this paragraph are finite. The following result is proved in the online supplement.

Theorem 2.3. *Suppose that $T \to \infty$, $k \to \infty$ and $k/\sqrt{T} \to 0$. Then:*

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

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 $\kappa_c \xrightarrow{P} \kappa_X$.

Theorem 2.3 shows that when the true regression coefficients are large, the kurtosis of the combination is close to the average kurtosis of the predictors rather than the kurtosis of the true density. A similar observation holds for the skewness of the combination.

Numerical example. We now numerically illustrate the aforementioned effect of significant changes in the skewness and kurtosis when densities are combined. We focus on the linear regression setup described above, where x_{jt} are i.i.d. N$(0, 1)$ and ε_t are i.i.d. errors, generated independently of the regressors. When forecasting, we suppose that the distribution of ε_t is known, so we take p as the true density of the errors in the formula $p_j(y) = p(y - \hat{\mu}_j)$. We run this experiment 5000 times, for three different values of k. We estimate the parameters with a sample size of $T = 100$, and produce one-step ahead forecasts for the moments of the predicitive densities. For the kurtosis part of the experiment, we use (symmetric) t_5 as the density of the errors, while for the skewness part we use skewed-t_5, with the skewness value of 1. Figure 3 depicts the distributions (i.e., estimated densities) of the skewness and the kurtosis of the combination that uses equal weights. Both the skewness and the kurtosis of the combination decrease when the number of predictors, k, increases: the skewness shifts towards zero and the kurtosis shifts towards 3, which is the Normal kurtosis.

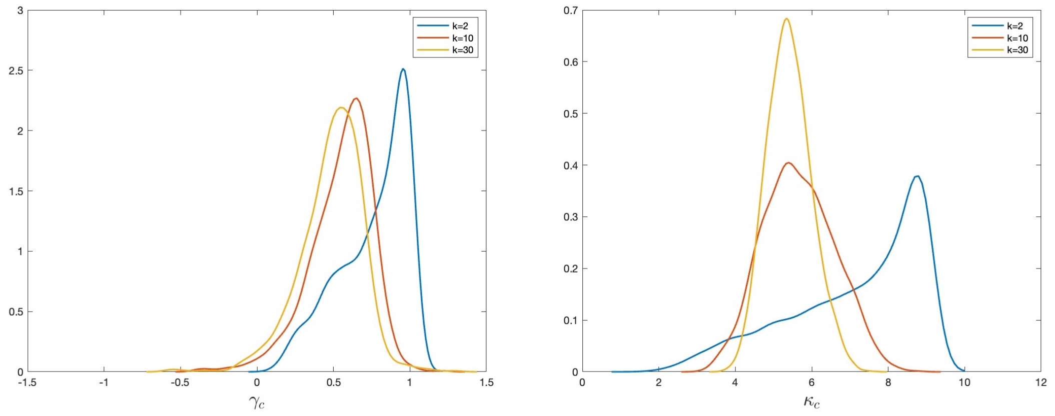


Figure 3: Distribution of the skewness (γ_c) and kurtosis (κ_c) of the equally weighted density combination.

3 Methodology

We now return to the general case, where the density combination is given by equation (1) and the weights are not necessarily equal. As a starting point for our proposed methodology, we consider the weights of Hall and Mitchell (2007) and Geweke and Amisano (2011), which are based on the idea that, in practice, the combination close to the true but unknown density f of the predicted outcome y_T is desirable. The Kullback–Leibler information criterion (KLIC) can be employed to gauge the distance from the combined to the true density,

$$\text{KLIC}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \text{E}\left[\log\left[\frac{f(y_T)}{p_c(y_T; \boldsymbol{\omega}, \boldsymbol{\theta})}\right]\right]. \tag{3}$$

Given a density function of the form $g(y) = p_c(y; \boldsymbol{\omega}, \boldsymbol{\theta})$, we extend the notation by sometimes writing $\text{KLIC}(g)$ in place of $\text{KLIC}(\boldsymbol{\omega}, \boldsymbol{\theta})$. The KLIC criterion can be estimated by its sample analogue,

$$\overline{\text{KLIC}}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \log\left[\frac{f(y_t)}{p_c(y_t; \boldsymbol{\omega}, \boldsymbol{\theta})}\right], \tag{4}$$

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

$$\max_{\boldsymbol{\omega}} \sum_{t=1}^{T} \log\left[\sum_{j=1}^{k} \omega_j p_j(y_t; \boldsymbol{\theta}_j)\right] \quad \text{subject to} \quad \sum_{j=1}^{k} \omega_j = 1, \;\; \omega_j \geq 0, \; j = 1, \ldots, k. \tag{5}$$

For convenience, the optimal weights that solve equation (5) are named *Log-Score* weights.

3.1 HMC optimization problem

In Section 2.2 we observed the undesirable effects that equally weighted density combinations have on skewness and kurtosis. In Appendix C of the online supplement we demonstrate similar effects for the Log-Score weights. However, we can modify optimization problem (5) by introducing additional restrictions on weights that provide control over the skewness and kurtosis of the combination. To this end, we propose solving the following *High Moment Constraints* (HMC) optimization problem:

$$\begin{aligned} &\text{maximize} && \sum_{t=1}^{T} \log\left[\sum_{j=1}^{k} \omega_j p_j(y_t; \boldsymbol{\theta}_j)\right] \\ &\text{subject to} && \sum_{j=1}^{k} \omega_j = 1, \quad \omega_j \geq 0, \quad j = 1, \ldots, k \\ & && \kappa_c \geq \underline{\kappa} \quad \text{and/or} \quad \gamma_c \geq \underline{\gamma} \quad \text{or} \quad \gamma_c \leq \underline{\gamma}, \end{aligned} \tag{6}$$

where the skewness and kurtosis of the combination, κ_c and γ_c, are functions of either the first three or the first four moments, respectively, of the individual densities. Closed-form expressions for κ_c and γ_c are provided, respectively, in equations (A.18) and (A.19) of the online supplement. The exact structure of the constraints can be selected to suit the problem. The optimal weights obtained by solving the HMC optimization problem (6) are named *HMC weights* for brevity.

We use the data to guide our selection of the thresholds $\underline{\kappa}$ and $\underline{\gamma}$. We take the kurtosis threshold as $\underline{\kappa} = \hat{\kappa}_T - \delta_T$, where $\hat{\kappa}_T$ is the sample kurtosis and δ_T is a small positive constant that depends on T. More specifically, we set the kurtosis threshold $\underline{\kappa}$ somewhat below the sample kurtosis, so that the true kurtosis is above $\underline{\kappa}$ with high probability. In practice, we let $\delta_T = z_{0.995} \times \mathrm{SE}(\hat{\kappa})$, where

$$\mathrm{SE}(\hat{\kappa}) = \sqrt{24n(n-2)(n-3)(n+1)^{-2}(n+3)^{-1}(n+5)^{-1}}$$

is the standard error of the sample kurtosis (Wright and Herrington, 2011) and $z_{0.995}$ is the 99.5% quantile of the standard normal distribution. If the sample skewness, $\hat{\gamma}_T$, is positive, we impose the constraint $\gamma_c \geq \underline{\gamma}$ and set the threshold as $\underline{\gamma} = \hat{\gamma}_T - \epsilon_T$, where $\epsilon_T = z_{0.999} \times \mathrm{SE}(\hat{\gamma})$ and

$$\mathrm{SE}(\hat{\gamma}) = \sqrt{6(n-2)(n+1)^{-1}(n+3)^{-1}}$$

is the standard error of the sample skewness (Wright and Herrington, 2011). Alternatively, if $\hat{\gamma}_T$ is negative, we impose the constraint $\gamma_c \leq \underline{\gamma}$ with $\underline{\gamma} = \hat{\gamma}_T + \epsilon_T$.

3.2 Comparison with existing methods

If one is interested only in a particular region of the density B_t, e.g., the tails, then Diks et al. (2011) and Opschoor et al. (2017) offer an alternative to our HMC approach. Instead of maximizing the complete log score, one can use the censored likelihood (CSL) scoring rule

$$S^{\mathrm{csl}} = \sum_{t=1}^{T} I(y_t \in B_t) \log \left[\sum_{j=1}^{k} \omega_j p_j(y_t; \boldsymbol{\theta}_j) \right] + I(y_t \in B_t^c) \log \int_{B_t^c} \left[\sum_{j=1}^{k} \omega_j p_j(y; \boldsymbol{\theta}_j) \right] dy \quad (7)$$

instead of imposing the restriction $\kappa_c \geq \underline{\kappa}$. The CSL approach produces an accurate density estimate for the focus region B_t but might be inadequate outside of it as demonstrated by Opschoor et al. (2017) in their Figure 1. The advantage of our proposed approach is that it aims to produce a density that is accurate on its entire support.

Kapetanios et al. (2015) model the weights using a known basis $\{\eta_s(y, \boldsymbol{\zeta}_s)\}_{s=1}^{\infty}$, which is parameterized by $\boldsymbol{\zeta}_s$ and is non-negative: $\eta_s(y, \boldsymbol{\zeta}_s) \geq 0$. In our notation, they consider

$$\omega_j(y_t) = \nu_{j0} + \sum_{s=1}^{\infty} \nu_{js}\eta_s(y_t, \boldsymbol{\zeta}_s) \tag{8}$$

with positive parameters $\nu_{js} \geq 0$ that still guarantee a well-defined combined density, i.e.,

$$\int \sum_{j=1}^{k} \omega_j(y)p_j(y; \boldsymbol{\theta}_j)dy = 1. \tag{9}$$

This setup allows estimation of the parameters ν_{js} and $\boldsymbol{\zeta}_s$ with the usual Normal asymptotics as given by Theorem 1 in Kapetanios et al. (2015). In contrast, our constraints on the kurtosis and the skewness restrict the parameter space and result in non-standard asymptotics (see the online supplement for the details). We do not model the weights as functions of y_t but rather focus on the effects of the high moment restrictions.

Bassetti et al. (2018) give another alternative for correcting the distortion produced by the combination. They introduce a beta mixture transformed linear pool and use Bayesian methods to estimate the parameters. We cannot compare their approach directly to our optimization problem (6); however, their aggregated cdf admits the following pdf representation (see equation 5 in Bassetti et al., 2018),

$$p_B(y; \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \sum_{s=1}^{S} w_s p_c(y; \boldsymbol{\omega}, \boldsymbol{\theta}) B_{\mu_s,\nu_s}(P_c(y; \boldsymbol{\omega}, \boldsymbol{\theta})), \tag{10}$$

where $B_{\mu,\nu}(x) = B(\mu\nu, (1-\mu)\nu)^{-1}x^{\mu\nu-1}(1-x)^{(1-\mu)\nu-1}1_{[0,1]}(x)$ is the pdf of a beta function with mean μ and precision ν, P_c is the cdf of the combination corresponding to p_c, w_s is the weight assigned to the s-th function in the beta mixture, and $\boldsymbol{w}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$ are the vectors that collect the additional parameters across all mixtures. The correction (10) is flexible enough to recover the underlying distribution, including a heavy-tailed distribution, however, it comes at the expense of the tractability of the combination weights $\boldsymbol{\omega}$.

3.3 Theoretical results: consistency and rate of convergence

In this section, we establish consistency and the rate of convergence of the HMC density combination, which is defined by solving optimization problem (6). Our results cover the corresponding unconstrained estimator as a special case. All the proofs are provided in the online supplement.

Suppose that the estimates of the model parameters converge in probability as T tends to infinity: $\widehat{\boldsymbol{\theta}}_T \xrightarrow{P} \boldsymbol{\theta}^*$, for some fixed finite vector $\boldsymbol{\theta}^*$, which can be thought of as the "population" vector of the model parameters. We first focus on the case where $\underline{\kappa}$ and $\underline{\gamma}$

are constants. We define $C(\boldsymbol{\theta})$ as the constraint set for the weights $\boldsymbol{\omega}$ in optimization problem (6) and denote by $\widehat{\boldsymbol{\omega}}_T$ the HMC optimal weights, that is, the solution to (6) but with $\boldsymbol{\theta}$ replaced by $\widehat{\boldsymbol{\theta}}_T$. The corresponding population solution is:

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}\in C(\boldsymbol{\theta}^*)} \text{KLIC}(\boldsymbol{\omega}, \boldsymbol{\theta}^*), \tag{11}$$

where KLIC is defined in (3). Estimated weights $\widehat{\boldsymbol{\omega}}_T$ and population weights $\boldsymbol{\omega}^*$ result in the following estimated and population versions of the density combination:

$$\widehat{f}_T(y_T) = \sum_{j=1}^{k} \widehat{\omega}_{Tj} p_j(y_T; \widehat{\theta}_{Tj}) \qquad \text{and} \qquad f^*(y_T) = \sum_{j=1}^{k} \omega_j^* p_j(y_T; \boldsymbol{\theta}_j^*). \tag{12}$$

We let $\|\cdot\|_1$ denote the functional L_1 norm. Theorem 3.1 establishes consistency of $\widehat{f}_T$ under mild regularity and continuity assumptions, which we provide and discuss in the appendix. This result does not require that the true density is represented as a linear combination of the individual densities under consideration.

Theorem 3.1. *Suppose that assumptions A1–A6 in the appendix are satisfied. Then,* $\text{KLIC}(\widehat{f}_T) \overset{P}{\to} \text{KLIC}(f^*)$ *as* $T \to \infty$. *Moreover, if the solution to the population problem* (11) *is unique, then* $\|\widehat{f}_T - f^*\|_1 \overset{P}{\to} 0$ *as* $T \to \infty$.

We now consider the setting where the true density can be represented as a combination of the individual densities. More specifically, suppose that

$$f(y_T) = \sum_{j=1}^{k} \omega_j^* p_j(y_t; \boldsymbol{\theta}_j^*), \tag{13}$$

for some weight vector $\boldsymbol{\omega}^*$. We also allow $\underline{\kappa}$ and $\underline{\gamma}$ to change with T, matching the implementation of the HMC approach in Section 3.1: $\underline{\kappa} = \hat{\kappa}_T - \delta_T$ and $\underline{\gamma} = \hat{\gamma}_T - \text{sign}(\hat{\gamma}_T)\epsilon_T$. The following result demonstrates that HMC can asymptotically recover the true density, and establishes the $T^{-1/2}$ rate of convergence.

Theorem 3.2. *Suppose that the true density f admits a unique representation* (13). *If assumptions A1–A6 in the appendix are satisfied, then* $\|\widehat{f}_T - f\|_1 \overset{P}{\to} 0$ *as* $T \to \infty$. *If assumptions A7–A12 are also satisfied, then* $\|\widehat{f}_T - f\|_1 = O_P\big(T^{-1/2}\big)$.

In the online supplement we provide additional results on the limiting distribution of the HMC weights. The constraints on kurtosis and skewness imply non-standard asymptotics, where the limiting normal distribution is projected onto a tangent cone.

4 An application to forecasting volatility

Density forecast combination methods are often applied to financial data due to its size and availability. Recent examples include Geweke and Amisano (2010), Geweke and Amisano (2011), Kapetanios et al. (2015), Crisóstomo and Couso (2017), and Bassetti et al. (2018). We follow their suit and use S&P 500 returns in our empirical application, which illustrates the benefits of the proposed HMC approach.

4.1 Empirical methodology

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 December 4, 2020. Daily returns and realized volatility measures are from the "Realized Library" of the Oxford-Man Institute.[3] As recommended in Shephard and Sheppard (2010), we use the realized kernel as the realized volatility measure. For more information on realized kernels, see Barndorff-Nielsen et al. (2008) and Barndorff-Nielsen et al. (2009).

Volatility models. The returns at time t can be expressed as

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

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

$$v_t^{\text{GARCH}} = \omega + \alpha(y_{t-1} - \mu)^2 + \beta v_{t-1}^{\text{GARCH}}, \tag{15}$$

which is the workhorse of volatility models. The statistical properties relevant to GARCH models are discussed in Ling and McAleer (2003). Our second model set is based on the EGARCH approach of Nelson (1991):

$$\log v_t^{\text{EGARCH}} = \omega + \gamma\varepsilon_{t-1} + \alpha(|\eta_{t-1}| - \mathrm{E}\,|\eta_{t-1}|) + \beta \log v_{t-1}^{\text{EGARCH}}. \tag{16}$$

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). Typically the properties of EGARCH models have to be investigated empirically, as, for example, in Anyfantaki and Demos (2016). Despite these

[3] All the data for this section were obtained from the Oxford-Man Institute's Realized Library.

known problems, EGARCH models have remained popular in empirical finance. In our empirical analysis, we include the EGARCH approach whilst acknowledging its pitfalls.

We also consider two more recent models for v_t that explicitly include a model-free estimator of the variance, namely, the realized measure of daily volatility: the HEAVY model of Shephard and Sheppard (2010) and the realized GARCH (RGARCH) model of Hansen et al. (2012). The reason for incorporating realized measures in the GARCH model is that they tend to provide a better estimate of daily volatility than the traditional squared daily returns. The dynamics for the conditional variance v_t are similar in the HEAVY and RGARCH models: both feature the realized measure instead of the squared realized returns in the variance equation (15) of the GARCH model. More specifically,

$$\begin{aligned} v_t^{\text{RGARCH}} &= c + \alpha(y_{t-1} - \mu)^2 + \beta v_{t-1}^{\text{RGARCH}} + \delta\, rv_{t-1}, \quad \text{and} \\ v_t^{\text{HEAVY}} &= c_h + \beta_h v_{t-1}^{\text{HEAVY}} + \delta_h rv_{t-1}, \end{aligned}$$

where rv_{t-1} is the realized measure of volatility. The above equaitons are sufficient for producing one-step-ahead forecasts that we need for our analysis. The two models differ in their treatment of the realized measure's dynamics: HEAVY features GARCH-type dynamics for the expectation of realized volatility whereas RGARCH models realized volatility as a function of the conditional variance, v_t, plus a leverage component.

Distributions. We consider several distributions F for the GARCH, EGARCH, HEAVY, and RGARCH models. In addition to the Gaussian, we also use fat-tailed distributions: Student-t, Laplace, and skewed-t (Hansen, 1994). In all the volatility models, we set the mean, μ, of the distribution to zero. The standard deviation of the distribution is $\sigma_t = \sqrt{v_t}$, where v_t are estimated for each model. For the Gaussian, t, and Laplace distributions the skewness is zero, while the kurtosis is 3, $6/[(df - 4) + 3]$, or 6, respectively, where df are the degrees of freedom for each t-distribution used in our GARCH-type models. We take the skewness and kurtosis for each model with the Hansen skewed-t distribution directly from Hansen (1994). The moments described above are needed for two purposes: first, they are used to specify the constraint set in the HMC optimization problem (6); and second, they are used to compute the moments of the density combinations.[4]

HMC Estimation. We compute three versions of the HMC weights by maximizing the criterion in optimization problem (6) subject to either a skewness constraint (HMCS), a kurtosis constraint (HMCK), or both constraints simultaneously (HMC). To speed up the computation of the solution to problem (6), we supply the optimization solver with "warm start" solutions - HMC0, HMCK0, and HMCS0 - one for each of the three versions

[4] Given an individual model and a weight-estimation window, we average the corresponding moments over the window in order to arrive at one set of moments for the model.

of the HMC weights.[5] For a given choice of the constraint set in problem (6), we produce the "warm start" solution using the following approach:

Step 1: From the set available individual models, we identify the ones that satisfy the constraint set in (6).

Step 2: On the reduced set of models identified in Step 1, we compute the optimal combination weights according to the CSL scoring rule of Opschoor et al. (2017) using a 15% tail threshold.[6]

Step 3: We assign zero weights to the models that were not selected in Step 1.

Forecasting method. We use rolling samples of 1250 trading days (5 years of trading data) to estimate all of the volatility model parameters and produce one-step-ahead forecasts. In total, we consider 16 forecasting models: as described above, we use four conditional volatility models and for each we use four possible distributions. We construct one-step-ahead individual predictive densities with the parameter estimates of each model over the remaining sample, thus covering the period from January 2005 to December 2020. Each evaluated density combination method combines these individual predictive densities by estimating the weights over windows with $T_{\text{win}} = 250$, 500, 750, and 1000 observations, which roughly correspond to between 1 and 4 years of trading data. This last step is repeated by moving the weight-estimation window over the entire forecasting sample. For example, with the 250-observation window, the first combined predictive density combination is produced for January 5, 2006, and the last – for December 4, 2020. The sample skewness and kurtosis required to specify the thresholds are computed using the weight-estimation window.

4.2 Empirical results

We compare our methodology with the equally weighted approach (EW), JMV weights of Jore et al. (2010), Log-Score combination weights of Geweke and Amisano (2011) computed using the Conflitti et al. (2015) algorithm, and two version (CSL15 and CSL25) of the censored likelihood approach proposed in Diks et al. (2011) and Opschoor et al. (2017). Armed with these weights, we compute the corresponding log-score, measure Value-at-Risk (VaR), as well as plot the skewness and kurtosis of the density combinations and compare them with the corresponding sample estimates.

First, the log-score of each of density combination is computed for every forecast time period. The second column of Table 2 reports the average log-score for each approach (relative to the maximum) over all the out-of-sample forecast periods. Naturally, the Log-Score weights perform the best when evaluated using their own objective criterion.

[5] We use MATLAB's nonlinear programming solver *fmincon* with the default 'interior-point' algorithm.

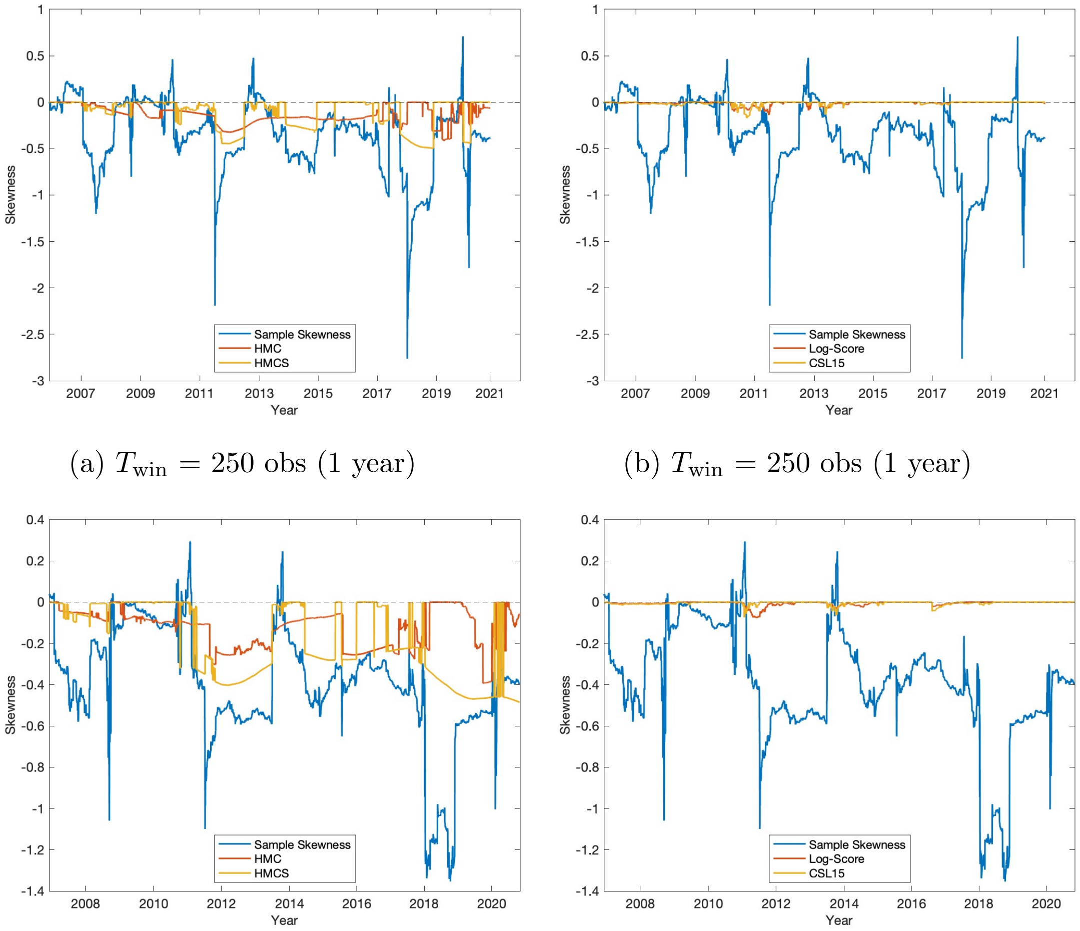
[6] We compute the optimal weights using the Conflitti et al. (2015) algorithm.

Table 2: Relative log-score, skewness and kurtosis of 1-day density forecasts for S&P 500

	Average	Skewness			Kurtosis		
Weights	log-score	$\min(\gamma_c)$	$\bar{\gamma}_c$	$\max(\gamma_c)$	$\min(\kappa_c)$	$\bar{\kappa}_c$	$\max(\kappa_c)$
			$T_{\text{win}} = 250$ obs. (1 year)				
Sample	–	-2.76	-0.36	0.71	2.83	5.78	25.72
Log-Score	1	-0.14	-0.01	0	3.20	4.17	6.44
HMC	0.962	-0.44	-0.08	0	2.99	6.23	10.83
HMCK	0.984	-0.30	-0.03	0	3.00	6.06	10.85
HMCS	0.964	-0.49	-0.07	0	3.00	5.14	10.83
CSL15	0.992	-0.17	-0.01	0	2.99	4.86	7.90
CSL25	0.994	-0.21	-0.01	0	3.01	5.09	8.69
JMV	0.972	-0.12	-0.07	-0.01	4.00	5.66	6.94
EW	0.968	-0.14	-0.08	-0.01	3.99	5.66	7.14
			$T_{\text{win}} = 500$ obs. (2 years)				
Sample	–	-1.35	-0.40	0.29	3.45	6.85	15.85
Log-Score	1	-0.08	-0.01	0	3.34	4.18	6.31
HMC	0.950	-0.39	-0.10	0	3.07	6.73	9.21
HMCK	0.987	-0.23	-0.03	0	3.07	6.56	11.30
HMCS	0.922	-0.49	-0.17	0	3.24	6.01	9.33
CSL15	0.994	-0.07	0.00	0	3.05	4.83	7.03
CSL25	0.994	-0.05	0.00	0	3.16	5.04	7.15
JMV	0.975	-0.12	-0.07	-0.01	4.02	5.65	6.79
EW	0.972	-0.12	-0.08	-0.01	4.00	5.63	6.52
			$T_{\text{win}} = 750$ obs. (3 years)				
Sample	–	-1.16	-0.38	0.06	4.03	7.71	18.56
Log-Score	1	-0.03	0	0	3.54	4.19	5.61
HMC	0.934	-0.35	-0.11	0	3.25	7.32	9.85
HMCK	0.984	-0.24	-0.05	0	3.25	6.72	9.85
HMCS	0.934	-0.49	-0.15	0	3.52	6.40	9.42
CSL15	0.991	-0.03	0	0	3.25	4.76	6.50
CSL25	0.993	-0.04	0	0	3.47	4.95	6.83
JMV	0.971	-0.12	-0.07	-0.02	4.07	5.67	6.57
EW	0.968	-0.12	-0.08	-0.02	4.06	5.65	6.30
			$T_{\text{win}} = 1000$ obs. (4 years)				
Sample	–	-1.18	-0.37	-0.03	4.42	8.42	16.11
Log-Score	1	-0.04	0	0	3.65	4.18	5.84
HMC	0.915	-0.33	-0.13	0	3.44	7.26	10.13
HMCK	0.975	-0.21	-0.05	0	3.43	6.78	10.13
HMCS	0.913	-0.46	-0.17	0	3.73	6.14	9.07
CSL15	0.991	-0.04	0	0	3.41	4.82	6.44
CSL25	0.993	-0.04	0	0	3.80	5.01	6.62
JMV	0.970	-0.10	-0.07	-0.02	4.59	5.72	6.45
EW	0.967	-0.11	-0.08	-0.03	4.58	5.70	6.13

Notes: T_{win} is the number of observations used to estimate the combination weights.

The HMC weights, specifically HMCK, perform better than the EW and JMV weighting schemes. However, the average log-scores of CSL15 and CSL25 are closer to that of the Log-Score combination than the HMC ones. This phenomenon can be explained by the greater (in magnitude) average skewness and kurtosis produced by the HMC combinations relative to the other weighting schemes, as seen in columns 4 and 7 of Table 2. As shown by in columns 3 and 8 in Table 2, relative to the CSL and Log-Score approaches, HMC and HMCS have lower (negative) minimum skewness, $\min(\gamma_c)$, while HMC and HMCK have higher maximum kurtosis, $\min(\kappa_c)$. Furthermore, their averages, $\bar{\gamma}_c$ and $\bar{\kappa}_c$, are consistently closer to the corresponding sample averages. Compared to the HMC approach, weighting schemes such as EW and Log-Score tend to produce less skewness and kurtosis. These observations hold for all windows T_{win}.



(a) $T_{\text{win}} = 250$ obs (1 year)

(b) $T_{\text{win}} = 250$ obs (1 year)

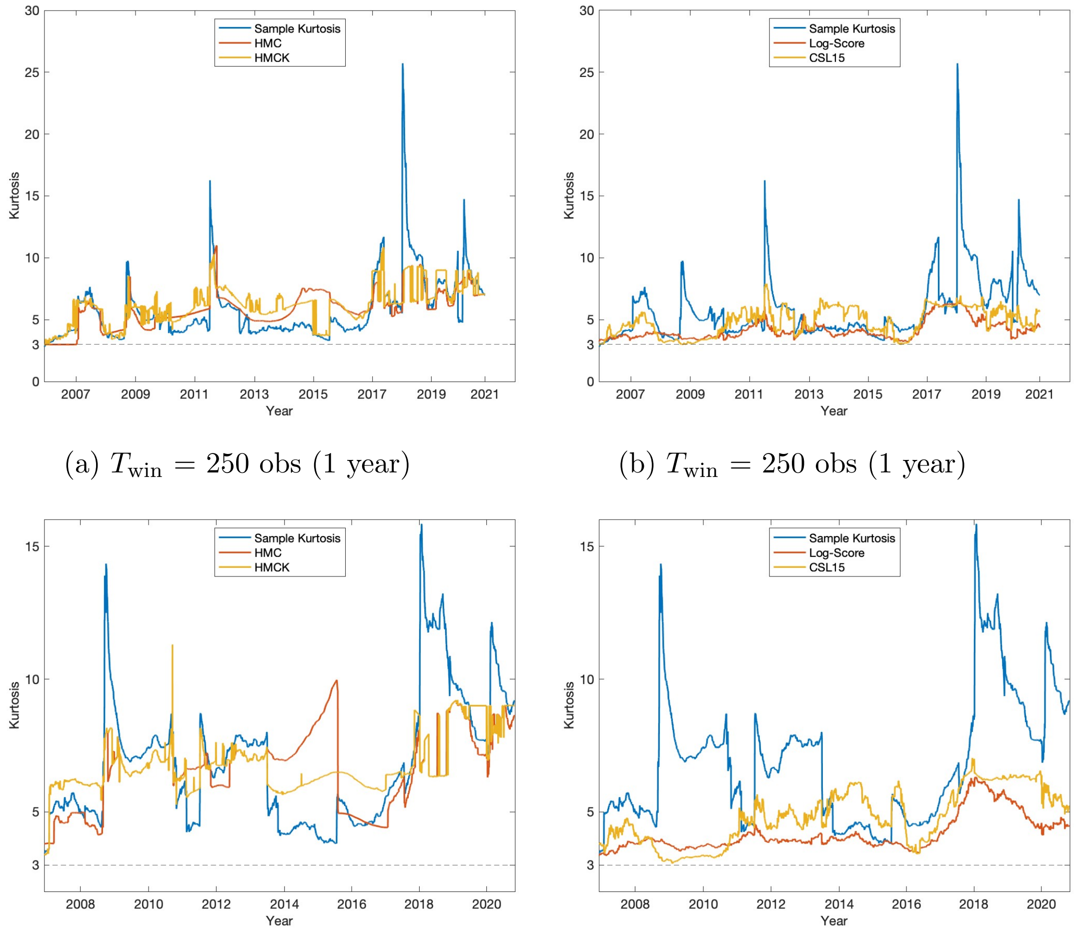
(c) $T_{\text{win}} = 500$ obs (2 years)

(d) $T_{\text{win}} = 500$ obs (2 years)

Figure 4: Sample skewness versus skewness of the combination for different weight-estimation windows T_{win}. Graphs in the left panel illustrate the skewness of the HMC and HMCS combinations. Graphs in the right panel illustrate the skewness of the Log-Score and CSL15 combinations. The sample skewness is estimated over T_{win} data points.

Next, we construct the skewness and kurtosis of each combined density forecast, which are implied by the estimated weights and the moments of the individual model densities.

Figures 4 and 5 plot the sample skewness and kurtosis, respectively, and compare them to the skewness and kutosis of the density combinations corresponding to the HMC(K/S) weights, the Log-Score weights, and the CSL15 weights.[7]



(a) $T_{\text{win}} = 250$ obs (1 year)

(b) $T_{\text{win}} = 250$ obs (1 year)

(c) $T_{\text{win}} = 500$ obs (2 years)

(d) $T_{\text{win}} = 500$ obs (2 years)

Figure 5: Sample kurtosis versus kurtosis of the combination for different weight estimation windows T_{win}. The left panel illustrates the kurtosis of the HMC and HMCK combinations. The right panel illustrates the kurtosis of the Log-Score and CSL15 combinations. The sample kurtosis is estimated over T_{win} data points.

HMC weighting schemes produce densities with skewness and kurtosis that closely track their sample counterparts. As expected, HMCK is better at tracking kurtosis while HMCS is better at tracking skewness, however, the HMC approach does well at tracking both. The tracking performance of the HMC approach is much better than that of other weighting schemes such as CSL15, Log-Score, or JMV. As illustrated by Figure 4, this difference in performance is especially striking in the case of skewness, where CSL15 and Log-Score combined densities are nearly symmetric. Whilst Log-Score and CSL15 combined densities do produce some kurtosis, it is lower than that of HMC and HMCK.

[7] For brevity, we focus on $T_{\text{win}} = 250$ and 500. Similar results hold for $T_{\text{win}} = 750$ and 1000.

Importantly, the observations presented above hold regardless of the sample size used to estimate the weights, demonstrating that constrained log-score optimization provides a solid density combination methodology that can achieve a comparable level of skewness and kurtosis to what is observed in the data.

Finally, we evaluate the performance of the forecast combinations in the left tail of the distribution by considering the 99% 1-day Value-at-Risk (VaR) estimates:

$$\widehat{\text{VaR}}_t^{i,1-q} = \hat{\mu} + \sqrt{\hat{v}_t^i}\ \eta_q, \tag{17}$$

where η_q is the qth quantile of the assumed conditional distribution. Here, $\hat{\mu}$ is the forecast conditional mean return, as expressed in (14), and $\hat{v}_t^i$ is the forecast conditional variance with i = GARCH, EGARCH, HEAVY, and RGARCH. We set $\hat{\mu}$ to zero for all models and distributions. The weighted average of the VaR forecasts from individual predictive densities is not generally equal to the VaR forecast of the combination density. We overcome this difficulty by simulating ten thousand returns in proportion to the combination weights. Each return is randomly generated using the relevant forecast moments of the corresponding model for one of the four distributions considered. The 1% quantile of the distribution of the simulated returns is compared to the actual returns for each period in the forecasting sample. We compute the number of violations, i.e., the number of times that the actual returns are to the left of the corresponding VaR forecasts. We also report the Christoffersen (1998) conditional coverage test, which assesses whether violations are happening in clusters.

The results for the 1% VaR are presented in Table 3. HMC combination weights outperform the other weighting schemes, exhibiting violation rates that are consistently the closest to the 1% target. This observation is true whether the weights are optimized over a small or large sample (T_{win}). When the weight-estimation window is small, i.e., 1 to 2 years, the best performing HMC scheme is HMCK. When the weight-estimation window is large, the HMC scheme with both constraints performs the best. This observation reinforces the notion that fat-tails caused by excess kurtosis are a dominant problem in financial returns time-series, more so than skewness. As seen from the earlier graphs, there is much more variation in excess kurtosis than in skewness.

The CLS weighting scheme consistently overestimates the 1% target violation rate for small weight-estimation windows, with values between 1.25 and 1.34 for $T_{\text{win}} = 1$ and 2 years, respectively. JMV and EW forecast combinations produce violation rates that are consistently below the 1% target threshold, while the violation rates of the Log-Score weighting scheme are consistently above 1.45%. All weighting schemes pass the conditional coverage test for all values of T_{win}.

We also consider the 2.5% VaR. The HMC weights, especially HMCK, perform consistently well relative to the other combination methods, and outperform the competitors

Table 3: 1-day forecast 99% VaR estimates for S&P 500

Combinations	# viol. at 1%	CC test	# viol. at 1%	CC test
	$T_{\text{win}} = 250$ (1 year)		$T_{\text{win}} = 500$ (2 years)	
Log-Score	59 (1.57)	0.76	57 (1.63)	0.74
HMC	46 (1.23)	0.18	52 (1.49)	0.63
HMCK	41 (**1.09**)	0.75	38 (**1.09**)	0.74
HMCS	64 (1.71)	0.58	59 (1.69)	0.16
CSL15	47 (1.25)	0.76	47 (1.34)	0.83
CSL25	48 (1.28)	0.78	47 (1.34)	0.87
JMV	24 (0.64)	0.87	30 (0.86)	0.85
EW	27 (0.72)	0.87	33 (0.94)	0.79
	$T_{\text{win}} = 750$ (3 years)		$T_{\text{win}} = 1000$ (4 years)	
Log-Score	48 (1.48)	0.84	46 (1.53)	0.86
HMC	43 (1.32)	0.65	30 (**1.00**)	0.72
HMCK	35 (**1.08**)	0.75	29 (0.97)	0.74
HMCS	38 (1.17)	0.75	32 (1.07)	0.76
CSL15	36 (1.11)	0.92	29 (0.97)	0.94
CSL25	35 (**1.08**)	0.90	27 (0.90)	0.92
JMV	23 (0.71)	0.88	19 (0.63)	0.92
EW	23 (0.71)	0.88	19 (0.63)	0.92

Notes: T_{win} is the number of observations used to estimate the combination weights. The table reports both the number and the proportion of times that the VaR forecast exceeds the 1% quantile (the corresponding percentage of violations is provided in parentheses). The CC tests report the p-value for the conditional coverage test (Christoffersen, 1998).

for most estimation-window sizes T_{win} (see Table 6 in the online supplement).

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 proposed High Moment Constraint (HMC) approach delivers a solution that is accurate in forecasting the overall distribution, including characteristics such as heavy tails. Moreover, we derive the statistical properties of the resulting density combinations, including consistency and the rate of convergence.

We evaluate the weights through an empirical illustration of forecasting the conditional returns of the S&P 500 index. We observe that our proposed HMC approach is consistently on par with the competing density combination methods, outperforming them for smaller weight-estimation windows. Moreover, our approach produces densities with skewness and kurtosis that closely track their sample counterparts.

Appendix: theoretical assumptions

We impose mild continuity and regularity assumptions for the consistency result in Theorem 3.1. We write $\mathcal{B}(\boldsymbol{\theta}^*)$ for a closed ball around $\boldsymbol{\theta}^*$ whose positive radius is allowed to be arbitrarily small. 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: $\widehat{\boldsymbol{\theta}}_T \overset{P}{\to} \boldsymbol{\theta}^*$, 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 $\log p_j(y;\boldsymbol{\theta}_j)$ are continuous in $\boldsymbol{\theta}_j$.

A5: $\mathrm{E}\,\sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} \big|\log p_j(y_1;\boldsymbol{\theta})\big| < \infty$ for $j = 1,\ldots,k$.

A6: $\mathrm{E}\,\sup_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} p_j(y_1;\boldsymbol{\theta}) < \infty$ for $j = 1,\ldots,k$.

We note that A3 and A4 are standard continuity assumptions relating to function $\mathrm{KLIC}(\boldsymbol{\omega},\boldsymbol{\theta})$ and constraint set $C(\boldsymbol{\theta})$. If the constraint in optimization problem (6) involves only the skewness of the density combination, then assumption A3 can be relaxed to only concern the first three moments. Assumptions A5 and A6 are needed to control the behavior of function $\overline{\mathrm{KLIC}}(\boldsymbol{\omega},\boldsymbol{\theta})$ near its expected value. For our rate of convergence result in Theorem 3.2, we impose additional assumptions.

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

A8: $\widehat{\boldsymbol{\theta}}_T = \boldsymbol{\theta}^* + O_p(T^{-1/2})$.

A9: $\min\{\delta_T, \epsilon_T\} \geq 0,\ \max\{\delta_T, \epsilon_T\} = o(1),\ \text{ and }\ T^{-1/2} = o\big(\min\{\delta_T, \epsilon_T\}\big)$.

A10: 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 continuously differentiable functions of $\boldsymbol{\theta}_j$.

A11: All of the elements of the vector $\boldsymbol{\omega}^*$ are positive.

A12: Function KLIC$(\boldsymbol{\omega}, \boldsymbol{\theta})$ admits a quadratic Taylor approximation at $(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*)$.

We note that A7-A12 are standard dependence, regularity and smoothness assumptions that are needed to establish the $T^{-1/2}$ rate of convergence.

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.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008), "Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76, 1481–1536.

— (2009), "Realized kernels in practice: trades and quotes," *The Econometrics Journal*, 12, C1–C32.

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.

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

Conflitti, C., Mol, C. D., and Giannone, D. (2015), "Optimal combination of survey forecasts," *International Journal of Forecasting*, 31, 1096 – 1103.

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.

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. (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.

Hansen, P. R., Huang, Z., and Shek, H. H. (2012), "Realized GARCH: a joint model for returns and realized measures of volatility," *Journal of Applied Econometrics*, 27, 877–906.

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.

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.

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., Vasnev, A. L., and Wang, W. (2023), "Too similar to combine? On negative weights in forecast combination," *International Journal of Forecasting*, 39, 18–38.

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

Shephard, N. and Sheppard, K. (2010), "Realising the future: forecasting with high-frequency-based volatility (HEAVY) models," *Journal of Applied Econometrics*, 25, 197–231.

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.

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

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2023), “Forecast combinations: An over 50-year review,” *International Journal of Forecasting*, forthcoming.

Wright, D. B. and Herrington, J. (2011), “Problematic standard errors and confidence intervals for skewness and kurtosis,” *Behavior Research Methods*, 8–17.

ONLINE SUPPLEMENT

Appendix A

Whereas the first moment of the combination, μ_c, is simply a linear combination of the k individual density means, other moments have more compicated expressions. 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 A.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:

$$\gamma_c = \sum_{j=1}^{k} \omega_j \; \left[\gamma_j \; \sigma_j^3 + 3(\mu_j - \mu_c) \; \sigma_j^2 + (\mu_j - \mu_c)^3\right] \sigma_c^{-3}, \tag{A.18}$$

(d) the kurtosis:

$$\kappa_c = \sum_{j=1}^{k} \omega_j \; \left[\kappa_j \; \sigma_j^4 + 4(\mu_j - \mu_c) \; \gamma_j + 6 \; (\mu_j - \mu_c)^2 \; \sigma_j^2 + (\mu_j - \mu_c)^{-4}\right] \sigma_c^{-4}, \tag{A.19}$$

(e) the s-th centered moment: $m_{c,s} = \sum_{j=1}^{k} \omega_j \sum_{l=0}^{s} \binom{s}{l} \; (\mu_j - \mu_c)^l \; m_{j,s-l}$*, where* $\binom{s}{l}$ *is the binomial coefficient given by* $s!/\left[l!(s-l)!\right]$.

Appendix B

In this section, we provide proofs for all of the theoretical results in the paper. We also establish an additional result on the limiting distribution of the HMC weights.

B.1 Proof of Theorem 2.1 and Corollary 2.2

Proof. We start with the result for γ_c, the skewness of the combination. We define

$$\bar{\mu} = \frac{1}{k}\sum_{j=1}^{k}\mu_j, \quad \hat{v}_\mu = \frac{1}{k}\sum_{j=1}^{k}\Big(\mu_j - \bar{\mu}\Big)^2, \quad \hat{\gamma}_\mu = \frac{1}{k}\sum_{j=1}^{k}\frac{\Big(\mu_j - \bar{\mu}\Big)^3}{\hat{v}_\mu^{3/2}}, \quad \hat{v}^* = \frac{1}{k}\sum_{j=1}^{k}v_j$$

and let $\hat{R} = \hat{v}_\mu/\hat{v}^*$. It follows from (A.18) that

$$\gamma_c = \frac{1}{k}\sum_{j=1}^{k}\gamma_j\big[v_j/\hat{v}^*\big]^{3/2}\Big[1+\hat{R}\Big]^{-3/2} + \frac{4}{k}\sum_{j=1}^{k}(\mu_j-\bar{\mu})v_j + \hat{\gamma}_\mu\Big[1+\hat{R}^{-1}\Big]^{-3/2}. \tag{B.20}$$

By the law of large numbers and the continuous mapping theorem, we have

$$\begin{aligned} &\hat{R} = R + o_p(1), \quad \hat{\gamma}_\mu = \gamma_\mu + o_p(1), \quad \hat{v}^* = v^* + o_p(1), \\ &\frac{1}{k}\sum_{j=1}^{k}\gamma_j\big[v_j/\hat{v}^*\big]^{3/2} = \mu_\gamma\, E\big[\xi^3\big] + o_p(1) \quad \text{and} \quad \frac{1}{k}\sum_{j=1}^{k}(\mu_j-\bar{\mu})v_j = o_p(1), \end{aligned} \tag{B.21}$$

as $T \to \infty$. Thus, we can rewrite (B.20) as

$$\gamma_c = \mu_\gamma\, E\big[\xi^3\big]\big[1+R\big]^{-3/2} + \gamma_\mu\big[1+R^{-1}\big]^{-3/2} + o_p(1),$$

which gives the desired result.

We now move to κ_c, the kurtosis of the combination. It follows from (A.19) that

$$\begin{aligned} \kappa_c \;=\; & \Big(\frac{1}{k}\sum_{j=1}^{k}\kappa_j\big[v_j/\hat{v}^*\big]^2\Big)\Big[1+\hat{R}\Big]^{-2} + 4\Big(\frac{1}{k}\sum_{j=1}^{k}(\mu_j-\bar{\mu})\gamma_j\Big) \\ & + 6\Big(\frac{1}{k}\sum_{j=1}^{k}(\mu_j-\bar{\mu})^2 v_j\Big)\big[\hat{v}_\mu\hat{v}^*\big]^{-1}\hat{R}\Big[1+\hat{R}\Big]^{-2} + \hat{\kappa}_\mu\Big[1+\hat{R}^{-1}\Big]^{-2}. \end{aligned} \tag{B.22}$$

By the law of large numbers and the continuous mapping theorem, we have

$$\begin{aligned} &\hat{v}_\mu = v_\mu + o_p(1), \quad \hat{\kappa}_\mu = \kappa_\mu + o_p(1), \quad \hat{v}^* = v^* + o_p(1), \\ &\frac{1}{k}\sum_{j=1}^{k}\kappa_j v_j^2 = \mu_\kappa\big(v^{*2} + v_v\big) + o_p(1), \quad \frac{1}{k}\sum_{j=1}^{k}(\mu_j-\bar{\mu})\gamma_j = o_p(1), \\ &\text{and} \quad \frac{1}{k}\sum_{j=1}^{k}(\mu_j-\bar{\mu})^2 v_j = v_\mu v^* + o_p(1). \end{aligned}$$

Combining these stochastic bounds with those in (B.21), and applying the continuous mapping theorem again, we can rewrite (B.22) as

$$\kappa_c = \mu_\kappa \Big[1 + v_v/v^{*2}\Big] \Big[1 + R\Big]^{-2} + \kappa_\mu \Big[1 + R^{-1}\Big]^{-2} + 6R\Big[1 + R\Big]^{-2} + o_p(1),$$

which yields the desired result.

The first result of Corollary 2.2 follows from Theorem 2.1 and the fact that $v_\mu/v^* \to \infty$ implies $R \to \infty$. To establish the second result, we note that

$$\Big(E[\xi^2]\Big)^{3/2} \leq E[\xi^3] \leq \Big(E[\xi^4]\Big)^{3/4},$$

which we can rewrite as

$$1 \leq E[\xi^3] \leq \Big(\frac{v_v + v^{*2}}{v^{*2}}\Big)^{3/4}.$$

Consequently, when $v_v/v^{*2} \to 0$, we have $E[\xi^3] \to 1$. The second result of Corollary 2.2 then follows from Theorem 2.1 and the fact that $v_\mu/v^* \to 0$ implies $R \to 0$. □

B.2 Proof of Theorem 2.3

Proof. For brevity of the exposition, we focus on the skewness and derive the result for $G = 1$ and β fixed at a positive value. The remaining cases and the derivations for the kurtosis follow by analogous arguments with only minor modifications. We 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 $\tilde{R} = \sigma_\mu^2/\sigma^2$. We note that $\bar{\mu} = o_p(1)$ by the law of large numbers. It follows from (A.18) that

$$\gamma_c = \gamma_p \Big[1 + \tilde{R}\Big]^{-3/2} + \gamma_\mu \Big[1 + (\tilde{R})^{-1}\Big]^{-3/2}. \tag{B.23}$$

We define $\bar{x}_j = \sum_{t=1}^{T-1} x_{jt}/[T-1]$ and $\eta_{jT} = [T-1](\sum_{t=1}^{T-1}(x_{jt} - \bar{x}_j)^2)^{-1}$. We write $\hat{\mu}_{jT}$ in the form $\hat{\mu}_{jT} = \beta X_{jT} + \eta_{jT}\xi_{jT}$ and note that $\max_{j\leq k} E\xi_{jT}^2 = O(k/T)$. The last bound implies that $\max_{j\leq k} |\xi_{jT}| = O_p(kT^{-1/2})$, for example, by Lemma 2.2.2 in van der Vaart and Wellner (1996). The above stochastic bound 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

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

Another application of the law of large numbers gives $\sigma_\mu^2 = \beta^2\sigma_X^2 + o_p(1)$, which implies $\tilde{R} = \beta R + o_p(1)$. Similarly,

$$\gamma_\mu = (1/k)\sum_{j=1}^{k}\Big(\frac{\beta X_{jT}}{\sigma_\mu}\Big)^3 + o_p(1) = (1/k)\sum_{j=1}^{k}\Big(\frac{X_{jT}}{\sigma_X}\Big)^3 + o_p(1) = \gamma_X + o_p(1).$$

We conclude the proof by combining the expressions for $\tilde{R}$ and γ_μ with (B.23). □

B.3 Proof of Theorem 3.1

Proof. To simplify the presentation, we will write $\widehat{\boldsymbol{\theta}}$ instead of $\widehat{\boldsymbol{\theta}}_T$. We will also use notation from the empirical process theory: given a function h, we let $P_T h = (1/T)\sum_{t=1}^{T} h(y_t)$. Similarly, we will write Ph for $Eh(y_1)$, i.e., we let P denote the underlying marginal distribution of the observed y_i. For the remainder of the proof, all of the $\boldsymbol{\omega}$ are assumed to lie in the set $\mathcal{W} = \{\boldsymbol{\omega} : \sum_{j\le k}\omega_j = 1,\, \omega_j \ge 0, j = 1, ..., k\}$.

To simplify the notation, we let $p_{\boldsymbol{\omega},\boldsymbol{\theta}}$ denote the function $p_c(\cdot;\boldsymbol{\omega},\boldsymbol{\theta})$ and define

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

We note that

$$\begin{aligned} G(\boldsymbol{\omega},\boldsymbol{\theta}) &= \mathrm{KLIC}(\boldsymbol{\omega}^*,\boldsymbol{\theta}^*) - \mathrm{KLIC}(\boldsymbol{\omega},\boldsymbol{\theta}) \quad \text{and} \\ G_T(\boldsymbol{\omega},\boldsymbol{\theta}) &= \overline{\mathrm{KLIC}}(\boldsymbol{\omega}^*,\boldsymbol{\theta}^*) - \overline{\mathrm{KLIC}}(\boldsymbol{\omega},\boldsymbol{\theta}). \end{aligned} \tag{B.24}$$

Thus, $\boldsymbol{\omega}^*$ maximizes function $G(\cdot,\boldsymbol{\theta}^*)$ over the constraint set $C(\boldsymbol{\theta}^*)$, while $\widehat{\boldsymbol{\omega}}$ maximizes function $G_T(\cdot,\widehat{\boldsymbol{\theta}})$ over $C(\widehat{\boldsymbol{\theta}})$. Noting that $\widehat{\boldsymbol{\theta}} \in \mathcal{B}(\boldsymbol{\theta}^*)$ with probability tending to one and $G(\boldsymbol{\omega}^*,\boldsymbol{\theta}^*) = 0$, and taking into account parts (i) and (ii) of Lemma B.1, we derive

$$\begin{aligned} G(\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}) &= G_T(\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}) + o_p(1) \;=\; \max_{\boldsymbol{\omega}\in C(\widehat{\boldsymbol{\theta}})} G_T(\boldsymbol{\omega},\widehat{\boldsymbol{\theta}}) + o_p(1) \;=\; \max_{\boldsymbol{\omega}\in C(\widehat{\boldsymbol{\theta}})} G(\boldsymbol{\omega},\widehat{\boldsymbol{\theta}}) + o_p(1) \\ &= \max_{\boldsymbol{\omega}\in C(\boldsymbol{\theta}^*)} G(\boldsymbol{\omega},\boldsymbol{\theta}^*) + o_p(1) \;=\; G(\boldsymbol{\omega}^*,\boldsymbol{\theta}^*) + o_p(1) \\ &= o_p(1). \end{aligned} \tag{B.25}$$

Equality (B.24) then implies $\mathrm{KLIC}(\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}) = \mathrm{KLIC}(\boldsymbol{\omega}^*,\boldsymbol{\theta}^*) + o_p(1)$ or, equivalently, $\mathrm{KLIC}(\widehat{f}_T) = \mathrm{KLIC}(f^*) + o_p(1)$, which establishes the first result of Theorem 3.1.

We now move to the part of Theorem 3.1 where the solution to problem (11) is assumed to be unique. We 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 B.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 (B.25) 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.

Pointwise continuity of functions $y \mapsto p_j(y, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and integrability of the density functions $p_j(\cdot, \boldsymbol{\theta})$ imply, by the dominated convergence theorem, that function $(\boldsymbol{\omega}, \boldsymbol{\theta}) \mapsto \int |\sum_{j=1}^k \omega_j p_j(y, \boldsymbol{\theta}) - \sum_{j=1}^k \omega_j^* p_j(y, \boldsymbol{\theta}^*)|dy$ is continuous in $(\boldsymbol{\omega}, \boldsymbol{\theta})$. Since $(\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\theta}})$ converges to $(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*)$ in probability, we derive that

$$\int \Big| \sum_{j=1}^k \widehat{\omega}_j p_j(y, \widehat{\boldsymbol{\theta}}) - \sum_{j=1}^k \omega_j^* p_j(y, \boldsymbol{\theta}^*) \Big| dy = o_p(1)$$

by the continuous mapping theorem. This establishes $\|\widehat{f}_T - f^*\|_1 = o_p(1)$, which is the second result of Theorem 3.1. □

B.4 Proof of Theorem 3.2

Proof. Let γ^* and κ^* denote the skewness and kurtosis of the true density f. Because $\underline{\kappa}$ and $\underline{\gamma}$ depend on T, we will write $C_T(\widehat{\boldsymbol{\theta}})$ for the constraint set in the HMC optimization problem (6). We will also write $C(\boldsymbol{\theta}^*)$ for the "population" version of the constraint set where $\underline{\kappa}$ and $\underline{\gamma}$ are replaced by κ^* and γ^*, respectively. We note that $\underline{\kappa} \le \kappa^*$ and $|\underline{\gamma}| \le |\gamma^*|$ by the classical rate of convergence results for the sample skewness and kurtosis, together with the assumption imposed on δ_T and ϵ_T. Thus, the weights of the true density combination (13) are also the solution to optimization problem (11). We also note that $C_T(\widehat{\boldsymbol{\theta}})$ converges to $C(\boldsymbol{\theta}^*)$ in probability with respect to the Hausdorff distance. Consequently, despite the dependence of the constraint set $C_T(\widehat{\boldsymbol{\theta}})$ on T, the proof of proof of Theorem 3.1 still goes through, leading to $\widehat{\boldsymbol{\omega}} = \boldsymbol{\omega}^* + o_p(1)$ and $\|\widehat{f}_T - f\|_1 = o_p(1)$.

As before, we only consider weights $\boldsymbol{\omega}$ in the set $\mathcal{W}$, so that the elements of $\boldsymbol{\omega}$ are nonnegative and sum to one. Note that for *all* the $\boldsymbol{\omega}$ under consideration, we can write $\omega_1 = 1 - \sum_{j=2}^k \omega_j$, and thus, every function of $\boldsymbol{\omega}$ can be expressed in terms of $\boldsymbol{\omega}_{-1} = (\omega_2, ..., \omega_k)^\top$. Because $\boldsymbol{\omega}_{-1}^*$ is a maximum of the function $\boldsymbol{\omega}_{-1} \mapsto G(\boldsymbol{\omega}, \boldsymbol{\theta}^*)$, we have $\frac{\partial G}{\partial \boldsymbol{\omega}_{-1}}(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*) = 0$. Hence, restricting out attention to $\boldsymbol{\omega} \in \mathcal{W}$, we can write a Taylor expansion for $G(\boldsymbol{\omega}, \boldsymbol{\theta})$ at $(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*)$ in the following form:

$$\begin{aligned} G(\boldsymbol{\omega}, \boldsymbol{\theta}) \;=\;& G(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \Big[\frac{\partial G}{\partial \boldsymbol{\theta}}(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*)\Big] \\ & + \frac{1}{2}(\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}_{-1}^*)^\top \Big[\frac{\partial^2 G}{\partial \boldsymbol{\omega}_{-1}^2}(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*)\Big](\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}_{-1}^*) \\ & + O\Big(\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\| \, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|\Big) + O\Big(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2\Big) + o\Big(\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|^2\Big). \end{aligned} \tag{B.26}$$

Because $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + O_p(T^{-1/2})$, we can then derive

$$\begin{aligned} G(\boldsymbol{\omega}, \widehat{\boldsymbol{\theta}}) - G(\boldsymbol{\omega}^*, \widehat{\boldsymbol{\theta}}) &= +\frac{1}{2}(\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1})^\top \Big[\frac{\partial^2 G}{\partial \boldsymbol{\omega}^2_{-1}}(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*)\Big](\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1}) \\ & \quad + O_p(T^{-1/2}\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|) + O_p(T^{-1}) + o(\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|^2). \end{aligned} \tag{B.27}$$

For simplicity of the notation, we denote densities $p_j(\cdot; \boldsymbol{\theta}^*)$ by $p_j(\cdot)$. We observe that

$$\frac{\partial^2 G}{\partial \boldsymbol{\omega}^2_{-1}}(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*) = -P\Big[\big(p_2 - p_1, \dots, p_k - p_1\big)^\top \big(p_2 - p_1, \dots, p_k - p_1\big)\, f^{-2}\Big]. \tag{B.28}$$

The above matrix 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 A11, would contradict the uniqueness $\boldsymbol{\omega}^*$ as the solution to the population problem (11).

Given expressions E_1 and E_2, we will write $E_1 \lesssim E_2$ to mean that there exists a finite universal constant c, such that $E_1 \le cE_2$. We again borrow notation from the empirical process theory, and denote $T^{1/2}(P_T h - Ph)$ by $\nu_T h$ for every function h.

We now establish the $T^{-1/2}$ rate of convergence for $\widehat{\boldsymbol{\omega}}$. Let $h_{\boldsymbol{\omega},\boldsymbol{\theta}} = \log[p_{\boldsymbol{\omega},\boldsymbol{\theta}}/p_{\boldsymbol{\omega}^*,\boldsymbol{\theta}^*}]$. According to Theorem 5.52 in van der Vaart (2000), in view of consistency of $\widehat{\boldsymbol{\omega}}$, approximation (B.27), and non-singularity of $[\partial^2 G/\partial \boldsymbol{\omega}^2_{-1}](\boldsymbol{\omega}^*, \boldsymbol{\theta}^*)$, it is only sufficient to derive

$$E \sup_{\|\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1}\| \le \delta,\, \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^*)} \big|\nu_T(h_{\boldsymbol{\omega},\boldsymbol{\theta}} - h_{\boldsymbol{\omega}^*,\boldsymbol{\theta}})\big| \lesssim \delta. \tag{B.29}$$

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 for each such process.

We restrict our attention to a small closed ball around $\boldsymbol{\omega}^*_{-1}$, which we denote by $\mathcal{B}(\boldsymbol{\omega}^*_{-1})$. We choose the radius of $\mathcal{B}(\boldsymbol{\omega}^*_{-1})$ to be positive but sufficiently small to ensure $\boldsymbol{\omega}_j > 0$ for every j and every $\boldsymbol{\omega} \in \mathcal{W}$ such that $\boldsymbol{\omega}_{-1} \in \mathcal{B}(\boldsymbol{\omega}^*_{-1})$ – this can be achieved because of assumption A11. We write $\dot{h}_{\boldsymbol{\omega},\boldsymbol{\theta}}(y)$ for the first derivative of the function $\boldsymbol{\omega}_{-1} \mapsto h_{\boldsymbol{\omega},\boldsymbol{\theta}}(y)$, evaluated at $\boldsymbol{\omega}_{-1}$, and note that

$$\sup_{\boldsymbol{\omega} \in \mathcal{W},\, \boldsymbol{\omega}_{-1} \in \mathcal{B}(\boldsymbol{\omega}^*_{-1}),\, \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^*)} \left\|\dot{h}_{\boldsymbol{\omega},\boldsymbol{\theta}}\right\|_\infty \le \max_{1 \le j \le k} \sup_{\boldsymbol{\omega} \in \mathcal{W},\, \boldsymbol{\omega}_{-1} \in \mathcal{B}(\boldsymbol{\omega}^*_{-1}),\, \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^*)} \left\|\frac{p_j(\cdot; \boldsymbol{\theta})}{p_{\boldsymbol{\omega},\boldsymbol{\theta}}}\right\|_\infty \lesssim 1,$$

where the last inequality follows from the definition of $\mathcal{B}(\boldsymbol{\omega}^*_{-1})$. Consequently, for every $\boldsymbol{\omega}_1 \in \mathcal{W}$ and $\boldsymbol{\omega}_2 \in \mathcal{W}$, such that $(\boldsymbol{\omega}_1)_{-1} \in \mathcal{B}(\boldsymbol{\omega}^*_{-1})$ and $(\boldsymbol{\omega}_2)_{-1} \in \mathcal{B}(\boldsymbol{\omega}^*_{-1})$, we have

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^*)} \|h_{\boldsymbol{\omega}_1,\boldsymbol{\theta}} - h_{\boldsymbol{\omega}_2,\boldsymbol{\theta}}\|_\infty \lesssim \|(\boldsymbol{\omega}_1)_{-1} - (\boldsymbol{\omega}_2)_{-1}\|\,.$$

Corollary 5.53 in van der Vaart (2000) then gives bound (B.29) as a consequence of the inequality above (the specific bound is established in the proof of Corollary 5.53). Thus, we have proved that $\widehat{\boldsymbol{\omega}} = \boldsymbol{\omega}^* + O_p(T^{-1/2})$. The corresponding stochastic bound for $\|\widehat{f}_T - f\|_1$ follows from the same argument as the one at the end of the proof of Theorem 3.1. □

B.5 Supporting results

The following lemma is used in the proof of Theorem 3.1.

Lemma B.1. *The following stochastic bounds hold under the assumptions and notation in the statement and proof of Theorem 3.1:*

(i) $\sup_{\boldsymbol{\omega}\in\mathcal{W},\, \boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*)} |G_T(\boldsymbol{\omega},\boldsymbol{\theta}) - G(\boldsymbol{\omega},\boldsymbol{\theta})| = o_p(1)$*;*

(ii) $\max_{\boldsymbol{\omega}\in C(\widehat{\boldsymbol{\theta}})} G(\boldsymbol{\omega},\widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\omega}\in C(\boldsymbol{\theta}^*)} G(\boldsymbol{\omega},\boldsymbol{\theta}^*) + o_p(1)$*;*

(iii) If $\boldsymbol{\omega}^*$ *is a unique solution to problem* (11)*, then, 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 of Lemma B.1. We start with part (i), denoting functions $y \mapsto p_j(y;\boldsymbol{\theta})$ by $p_{j,\boldsymbol{\theta}}$ and functions $y \mapsto \log[p_{\boldsymbol{\omega},\boldsymbol{\theta}}(y)/p_{\boldsymbol{\omega}^*,\boldsymbol{\theta}^*}(y)]$ by $m_{\boldsymbol{\omega},\boldsymbol{\theta}}$, to simplify the notation. We will first show that the class $\mathcal{M}$ of functions $m_{\boldsymbol{\omega},\boldsymbol{\theta}}$ is pointwise compact in the sense of Example 19.8 in van der Vaart (2000). Specifically, we will show that (a) map $(\boldsymbol{\omega},\boldsymbol{\theta}) \mapsto m_{\boldsymbol{\omega},\boldsymbol{\theta}}(y)$ is continuous for each fixed y; (b) parameter vectors $(\boldsymbol{\omega},\boldsymbol{\theta})$ belong to a compact set; (c) functional class $\mathcal{M}$ has an integrable envelope.

Parts (a) and (b) hold by the imposed assumptions. To establish part (c), we need to bound all members of the class $\mathcal{M}$ by a function that is integrable with respect to P. Using the fact that the largest element in $\boldsymbol{\omega}$ lies in $[1/k, 1]$ and taking into account the general inequality $\log x \le x - 1$, we derive the following pointwise bound for functions $m_{\boldsymbol{\omega},\boldsymbol{\theta}}$:

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

Because the expected value of the function on the right-hand side is finite by assumptions A5 and A6, part (c) follows from the above bound. Thus, as shown in the aforementioned example of van der Vaart (2000), 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. Such "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). This completes the proof of part (i) of Lemma B.1.

To prove part (ii) of the lemma, we first note that the imposed continuity assumptions imply that if $\boldsymbol{\theta} \to \boldsymbol{\theta}^*$, then $C(\boldsymbol{\theta})$ converges to $C(\boldsymbol{\theta}^*)$, with respect to the Hausdorff distance. Moreover, an application of the dominated convergence theorem establishes that function $G(\boldsymbol{\omega}, \boldsymbol{\theta})$ is continuous, due to the pointwise continuity of the functions $m_{\boldsymbol{\omega},\boldsymbol{\theta}}$ and the existence of an integrable envelope, which was established in the previous paragraph. Thus, function G is uniformly continuous for $\boldsymbol{\omega} \in \mathcal{W}$, $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^*)$. Uniform continuity of G together with the continuity of $C(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$ imply that function $W(\boldsymbol{\theta}) = \max_{\boldsymbol{\omega} \in C(\boldsymbol{\theta})} G(\boldsymbol{\omega}, \boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}^*$. The continuous mapping theorem then yields $W(\widehat{\boldsymbol{\theta}}) = W(\boldsymbol{\theta}^*) + o_p(1)$, which completes the proof of part (ii).

We now move to part (iii) of the lemma. Because G is continuous, $\boldsymbol{\omega}^*$ is the unique maximum of $G(\cdot, \boldsymbol{\theta}^*)$ over $C(\boldsymbol{\theta}^*)$, and $G(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*) = 0$, we conclude that the maximum of $G(\cdot, \boldsymbol{\theta}^*)$ over the compact set $\boldsymbol{\omega} \in C(\boldsymbol{\theta}^*) \setminus B_\delta(\boldsymbol{\omega}^*)$ is negative. We can replace $\boldsymbol{\theta}^*$ with a sufficiently close $\boldsymbol{\theta}$ and still keep the negativity of the above maximum, because of the uniform continuity of $G(\boldsymbol{\omega}, \boldsymbol{\theta})$ and the continuity of $C(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$. Consequently, for a sufficiently small but positive r_δ, we have

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

This completes the proof of part (iii). □

B.6 Additional theoretical results: limiting distribution

We now return to the setting where $\underline{\kappa}$ and $\underline{\gamma}$ are constants and 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 an additional regularity condition:

A13: The unconstrained minimizer of $\text{KLIC}(\cdot, \boldsymbol{\theta}^*)$ lies in $C(\boldsymbol{\theta}^*)$.

Above, "unconstrained" minimizer is still required to have nonnegative weights that sum to one.

We 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

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

As noted in the proof of Theorem 3.2, matrix V_* is nonsingular under the imposed assumptions.

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 orthogonal projection of x onto A.

Theorem B.2. *Suppose that $\boldsymbol{\omega}^*$ is the unique solution to the population problem (11) and assumptions A1, A2′–A6′, A7, and A10–A13 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

$$\sqrt{T}(\widehat{\boldsymbol{\omega}}_{-1} - \boldsymbol{\omega}^*_{-1}) \xrightarrow{d} V_*^{-1/2} Proj_{V_*^{1/2}\mathcal{S}^*} \tilde{Z}. \tag{B.30}$$

Proof of Theorem B.2. Consistency of $\widehat{\boldsymbol{\omega}}$ is a consequence of Theorem 3.1. The $T^{-1/2}$ rate of convergence for $\widehat{\boldsymbol{\omega}}$ follows by repeating the arguments in the proof of Theorem 3.2, with minor simplifications and notational adjustments. We will now establish the limiting distribution for $\widehat{\boldsymbol{\omega}}$.

As a direct consequence of Taylor expansion (B.26) and equation (B.28), we have

$$G(\boldsymbol{\omega}, \boldsymbol{\theta}^*) = G(\boldsymbol{\omega}^*, \boldsymbol{\theta}^*) - \frac{1}{2}(\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1})^\top V_*(\boldsymbol{\omega}_{-1} - \boldsymbol{\omega}^*_{-1}) + o\Big(\|\boldsymbol{\omega} - \boldsymbol{\omega}^*\|^2\Big). \tag{B.31}$$

We define $h_{\boldsymbol{\omega}} = \log[p_{\boldsymbol{\omega},\boldsymbol{\theta}^*}/p_{\boldsymbol{\omega}^*,\boldsymbol{\theta}^*}]$ and write $\dot{h}_{\boldsymbol{\omega}}(y)$ for the first derivative of the function $\boldsymbol{\omega}_{-1} \mapsto h_{\boldsymbol{\omega}}(y)$, evaluated at $\boldsymbol{\omega}_{-1}$. 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, noting that $G(\boldsymbol{\omega}, \boldsymbol{\theta}^*) = Ph_{\boldsymbol{\omega}}$ and using (B.31), we conclude that

$$nP_n(h_{\boldsymbol{\omega}^*+v_T T^{-1/2}} - h_{\boldsymbol{\omega}^*}) = -\frac{1}{2}v_T^\top V_* v_T + v_T^\top \nu_T \dot{h}_{\boldsymbol{\omega}^*} + o_p(1). \tag{B.32}$$

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 (B.32) holds for every $O_p(1)$ random sequence v_T; (ii) $\widehat{\boldsymbol{\omega}}_{-1} = \boldsymbol{\omega}^*_{-1} + O_p(T^{-1/2})$; (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 inequalities involving smooth functions of $\boldsymbol{\omega}$. We 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 B.2 follows after completing the square for the quadratic expression above. □

Appendix C

We use the simulation setup of the linear regression numerical example described at the end of Section 2.2 and consider five sets of ad-hoc weights together with the Log-Score weights. The first set of ad-hoc weights starts with weight 1 on the first model and 0 on all the 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 (EW) is achieved.

Table 4: Skewness (γ_c) and kurtosis (κ_c) of the combination

	Panel A: γ_c					Panel B: κ_c				
Weights (ω_1)	$k = 2$	$k = 5$	$k = 10$	$k = 20$	$k = 30$	$k = 2$	$k = 5$	$k = 10$	$k = 20$	$k = 30$
1	1.000	1.000	1.000	1.000	1.000	9.000	9.000	9.000	9.000	9.000
0.75	0.797	0.745	0.723	0.718	0.701	7.505	7.215	7.163	7.118	7.113
0.50	0.759	0.644	0.605	0.591	0.570	7.146	6.353	6.204	6.108	6.087
0.25	0.801	0.604	0.550	0.526	0.503	7.505	6.018	5.741	5.592	5.551
EW	0.759	0.603	0.542	0.507	0.482	7.146	6.006	5.660	5.466	5.406
Log-Score	0.767	0.618	0.528	0.441	0.376	7.264	6.218	5.609	4.991	4.587

Notes: The Log-Score weights are obtained by solving (5). The individual densities are constructed using the estimated parameters of the linear regression in (2). 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. $\omega_i = \frac{1-\omega_1}{k-1}$ for $i = 2, \ldots, k$.

Table 4 presents the numerical results based on 5000 replications. The parameters are estimated with 100 data points and the predictive densities are based on a one-step ahead forecast for all weight-sets except Log-Score. For the latter, we produce 100 out-of-sample forecasts as the corresponding optimization problem requires a series of forecast errors. 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 skewed-t_5 error density is set to 1, and the kurtosis of the error density is 9. We observe that both the skewness and the kurtosis of the combination decrease when the number of predictors increases.

Table 5 further illustrates Theorem 2.3 in the cases $\beta \to 0$ and $\beta \to \infty$. We continue the previous simulation setup 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; a similar phenomenon is observed for the skewness. Alternatively, when $\boldsymbol{\beta} = (3, \ldots, 3)^\top$, the amount of signal increases in relation

Table 5: γ_c and κ_c for different values of β

	$\boldsymbol{\beta} = (1/\sqrt{k}, \ldots, 1/\sqrt{k})^\top$		$\boldsymbol{\beta} = (3, \ldots, 3)^\top$	
Weights	$k = 2$	$k = 10$	$k = 2$	$k = 10$
	Panel A: γ_c			
EW	0.847	0.917	0.407	0.101
Log-Score	0.855	0.910	0.412	0.098
	Panel B: κ_c			
EW	7.798	8.349	4.337	3.006
Log-Score	7.947	8.351	4.398	2.862

Notes: k are the number of densities. β are the regression parameters used for one-step ahead forecasting.

to the noise, and hence, the skewness/kurtosis of the combination approaches the skewness/kurtosis of the individual predictors; the limiting values are 0 and 3, respectively, because the predictors are Normally distributed.

Figure 6 depicts the distributions of the skewness and the kurtosis of the combination based on the Log-Score weights. Both the skewness and the kurtosis of the combination decrease when the number of predictors increases: the skewness shifts towards zero and the kurtosis shifts toward 3.

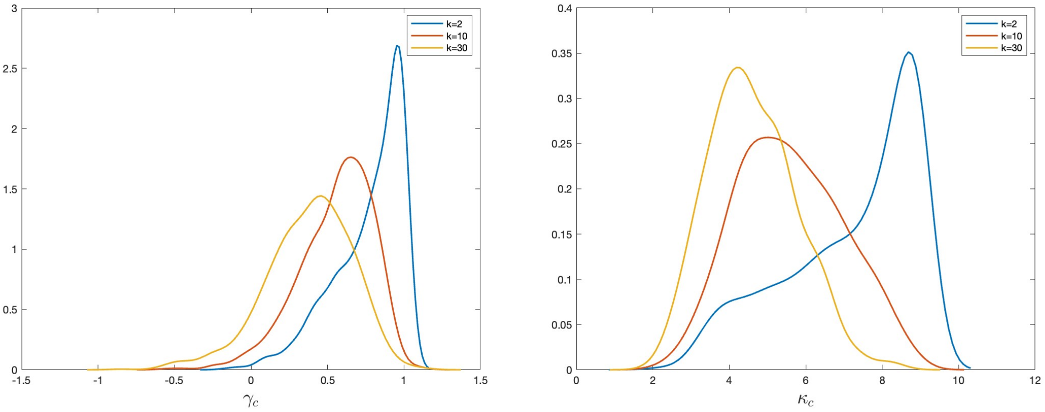


Figure 6: Distribution of the skewness of the density combinations (γ_c) and kurtosis (κ_c) for the Log-Score weights that result from solving (5).

Appendix D

Table 6, provided below, presents the 2.5% VaR results. The HMC weights, especially HMCK, perform on par with the other methods.

Table 6: 1-day forecast 97.5% VaR estimates for S&P 500

Combinations	# viol. at 2.5%	CC test	# viol. at 2.5%	CC test
	T_{win} = 250 (1 year)		T_{win} = 500 (2 years)	
Log-Score	118 (3.23)	0.26	115 (3.29)	0.34
HMC	109 (2.80)	0.70	105 (3.00)	0.95
HMCK	94 (**2.51**)	0.35	94 (2.69)	0.26
HMCS	120 (3.07)	0.13	116 (3.32)	0.43
CSL15	108 (2.94)	0.35	99 (2.83)	0.40
CSL25	113 (3.07)	0.33	103 (2.95)	0.36
JMV	81 (2.03)	0.46	87 (**2.49**)	0.46
EW	77 (2.11)	0.36	87 (**2.49**)	0.40
	T_{win} = 750 (3 years)		T_{win} = 1000 (4 years)	
Log-Score	104 (3.20)	0.28	93 (3.10)	0.38
HMC	100 (3.08)	1.00	92 (3.07)	0.20
HMCK	86 (**2.65**)	0.31	87 (2.90)	0.30
HMCS	91 (2.80)	0.55	74 (**2.47**)	0.51
CSL15	87 (2.68)	0.45	73 (2.44)	0.59
CSL25	98 (3.02)	0.28	80 (2.67)	0.22
JMV	72 (2.22)	0.58	59 (1.97)	0.64
EW	69 (2.13)	0.55	60 (2.00)	0.59

Notes: T_{win} is the number of observations required to estimate the combination weights. The table reports both the number and the proportion of times that the VaR forecast exceeds the 2.5% target (the percentage of violations is provided in parentheses).