Australian National University

Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

Optimal Forecast Combination with Mean Absolute Error Loss

CAMA Working Paper 59/2023
November 2023

Felix Chan
Curtin University

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

Abstract

The optimal aggregation of forecasts produced either from models or expert judgements presents an interesting challenge for managerial decisions. Mean absolute error (MAE) and mean squared error (MSE) losses are commonly employed as criteria of optimality to obtain the weights that combine multiple forecasts. While much is known about MSE in the context of forecast combination, less attention has been given to MAE. This paper shows that the optimal solutions from minimizing either MAE or MSE loss functions, i.e., the optimal weights, are equivalent provided that the weights sum to one. The equivalence holds under mild assumptions and includes a wide class of symmetric and asymmetric error distributions. The theoretical results are supported by a numerical study that features skewed and fat-tailed distributions. The practical implications of combining forecasts with MAE and MSE optimal weights are investigated empirically with a small sample of data on expert forecasts on inflation, growth, and unemployment rates for the European Union. The results show that MAE weights are less sensitive to outliers, and MSE and MAE weights can be close to equivalent even when the sample is small.

Keywords

Forecasting, forecast combination, optimization, mean absolute error, optimal weights

JEL Classification

C53, C61

Address for correspondence:

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

ISSN 2206-0332

Optimal forecast combination with mean absolute error loss

Felix Chan[1] and Laurent Pauwels[*2]

[1]School of Accounting, Economics and Finance, Curtin University.
[2]Business Analytics, The University of Sydney Business School and Centre for Applied Macroeconomic Analysis (CAMA), Australian National University.

November 2023

Abstract: The optimal aggregation of forecasts produced either from models or expert judgements presents an interesting challenge for managerial decisions. Mean absolute error (MAE) and mean squared error (MSE) losses are commonly employed as criteria of optimality to obtain the weights that combine multiple forecasts. While much is known about MSE in the context of forecast combination, less attention has been given to MAE. This paper shows that the optimal solutions from minimizing either MAE or MSE loss functions, i.e., the optimal weights, are equivalent provided that the weights sum to one. The equivalence holds under mild assumptions and includes a wide class of symmetric and asymmetric error distributions. The theoretical results are supported by a numerical study that features skewed and fat-tailed distributions. The practical implications of combining forecasts with MAE and MSE optimal weights are investigated empirically with a small sample of data on expert forecasts on inflation, growth, and unemployment rates for the European Union. The results show that MAE weights are less sensitive to outliers, and MSE and MAE weights can be close to equivalent even when the sample is small.

Keywords: *Forecasting, forecast combination, optimization, mean absolute error, optimal weights.*

[*]The authors would like to thank Eddie Anderson, Peter Fuleky, Róbert Lieli, Dmytro Matsypura, László Mátyás, Daniel Oron, Les Oxley, Anastasios Panagiotelis, Andrew Patton, Oleg Prokopyev, Peter Radchenko and the seminar participants at the Department of Economics and Business, Central European University and Waikato University's Management School for their comments and suggestions. The first author wishes to acknowledge the financial support from Curtin University, and the second author acknowledges the financial support from the University of Sydney Business School. Corresponding Author: Laurent Pauwels. Address: The University of Sydney Business School, Darlington, NSW 2008, Australia. Email: laurent.pauwels@sydney.edu.au

1 Introduction

Forecasting economic activity is one of the fundamental ingredients for decision analysis, whether it is for public policy or strategic planning. Typically, forecasts are produced either from models with a single predictor or with multiple predictors or from managerial expert judgments. Models produce different forecasts just as managers offer varied judgments, and it can be challenging to reconcile this information into a decision. One way to approach decision-making in this context is to aggregate forecasts, that is, to combine forecasts into a consensus forecast. Aside from facilitating decision-making, there are numerous statistical advantages to combining forecasts. Bates and Granger (1969) demonstrate that the optimal combination of forecasts for the mean squared error (MSE) loss function (or scoring rule) outperforms an individual forecast. Furthermore, the vector of the optimal linear combination has a simple closed form solution, which explains its popularity.

There is extensive literature that explores the combination of forecasts methodologically and empirically; see de Menezes et al. (2000) or Timmermann (2006) for surveys. One of the main focuses of the forecast combination literature is to investigate the forecasting performance of various weighting schemes, including simple averaging or "equal weights" (for example, see Larrick and Soll, 2006, Soll and Larrick, 2009, and Lichtendahl et al., 2013). The superiority of averaging forecasts over combining forecasts optimally has been coined a "forecast combination puzzle." The reasons for this puzzle are discussed in detail in Clemen and Winkler (1986), Capistrán and Timmerann (2009), Smith and Wallis (2009), Bjørnland et al. (2012), and Claeskens et al. (2016).

Practitioners must often forecast data that span over a short horizon, but combining forecasts in small samples by minimizing MSE loss is challenging. This is because it depends on the sample variance-covariance matrix of the forecast errors, which can be sensitive to outliers and extreme observations, especially in small samples. It is well known, however, that mean absolute error (MAE) loss is less sensitive to outliers. While combination methods based on MSE are prominent in theory and in practice, less attention has been devoted to other loss functions such as MAE.[1] This paper provides some theoretical results for combining forecasts under MAE loss and shows that it is equivalent to combining forecasts under MSE loss.

Our main contributions can be summarized as follows:

1. We start by providing the first-order condition for the MAE loss function, which is necessary to derive the optimal combination of forecasts under MAE loss. Then, we

[1] Some statistical properties of the MAE loss function are discussed in Gastwirth (1974) and Bassett Jr. and Koenker (1978). See Jose (2017) for an investigation of the advantages and properties of scale-free forecast accuracy measures, such as mean absolute percentage errors (MAPE).

prove that the optimal solutions, i.e., the optimal weights, are asymptotically equivalent when minimizing either MSE or MAE loss functions, provided that the weights sum to one. Hence, the equivalence of optimal weights implies the same optimal forecast combination. The equivalence holds even when the forecast errors are asymmetrically distributed or contain outliers. The literature, however, has emphasized the equivalence in optimization problems between classes of loss functions rather than the equivalence in optimal solutions. The latter is the subject of this paper. Fung and Mangasarian (2011) and Peng et al. (2015) provide the results of equivalence between optimization problems, namely, ℓ_p and ℓ_0 norm minimization. Patton (2019) provides a discussion for a set of loss functions that belong to the Bregman class, including MSE. See Bregman (1967), Savage (1971), and Banerjee et al. (2005) for further details.

2. We provide simulation evidence to support this equivalence by simulating fat-tailed and asymmetric forecast errors. The forecast errors are drawn from either skew normal or t_3 distributions. The Fat-tailed distributions are included to evaluate the impact of outliers in forecast combinations, as outliers usually imply poor forecasting performance in small samples. For example, see Gupta and Wilton (1987) and Winkler and Clemen (1992).

3. We investigate the practical implications of combining forecasts optimally under MAE and MSE loss with real-world data on inflation rate, unemployment rates, and growth rate forecasts. The forecasts are taken from the quarterly European Central Bank (ECB) Survey of Professional Forecasters (SPF). The ECB SPF is often used in empirical applications of forecast combination methods, including Genre et al. (2013), Conflitti et al. (2015), Matsypura et al. (2018), and Diebold and Shin (2019).[2] In this small sample of data, we show that the MAE optimal weights are less susceptible to bias and outliers than its MSE counterpart. Moreover, when there are no outliers, the MAE and MSE optimal weights are close to equivalent.

How are these results useful practically? In large samples, the optimal weights of the forecast combination are equivalent whether minimizing MSE loss or MAE loss when the weights sum to one. In this case, MSE appears to be more convenient due to its closed form minimization solution, i.e., there is no need for numerical computation. This is especially relevant in large-scale combination problems. However, estimating MSE-based weights requires a variance-covariance matrix that is subjected to finite sample bias. Furthermore, this can be exacerbated in the presence of outliers, whereas minimizing MAE loss is robust to them.

[2] For detailed discussion of the ECB Survey of Professional Forecasters, see García (2003), Kenny et al. (2007), and Bowles et al. (2010).

In small samples, we recommend checking for outliers in the dataset when choosing between MAE and MSE, whereas in a large sample, the choice is a matter of convenience.

The rest of the paper is organized as follows. Section 2 provides a motivating example and intuition for equivalence. Section 3 presents the main theoretical results of the paper. Section 4 conducts a simulation study to support the theory. Section 5 provides an empirical illustration. Section 6 concludes. All proofs and additional results are provided in Appendix A and in an Online Supplement available here. The computer code, data, and framework for all the figures and tables presented in the paper are available here.[3]

2 Motivation

This section provides an intuition with a motivating example for why forecast combinations with MAE loss provide the same optimal solutions, or optimal weights, as MSE loss under the constraint that the weights sum to one, i.e., a budget constraint. Suppose that an expert forecaster or a model i produces a point forecast, f_{it}, for some random variable of interest, say y_t, at a given time t. Then, a simple expression of the forecast errors, ν_{it}, is

$$\nu_{it} = y_t - f_{it}. \tag{1}$$

Without loss of generality, assume that f_{0t} is the best forecast in the sense that

$$\Pr\left[h\left(\nu_{0t}\right) < \varepsilon\right] > \Pr\left[h\left(\nu_{it}\right) < \varepsilon\right] \tag{2}$$

for $\varepsilon > 0$ and $i = 1, \ldots, k$, where h denotes some loss function or forecasting criterion and ν_{0t} essentially denotes the maximum bound on how accurately y_t can be forecast. The forecast errors can be rewritten as the sum of the random errors associated with the best forecast and a forecast specific random error, i.e., $\nu_{it} = \nu_{0t} + u_{it}$.

Consider the simple case of combining two forecasts, f_{1t} and f_{2t}, into

$$f_{ct} = a_1 f_{1t} + a_2 f_{2t} \tag{3}$$

with $a_1, a_2 \in \mathbb{R}$ representing the combination weights associated with the two forecasts. The optimization problem of combining the two forecasts at time t is expressed as

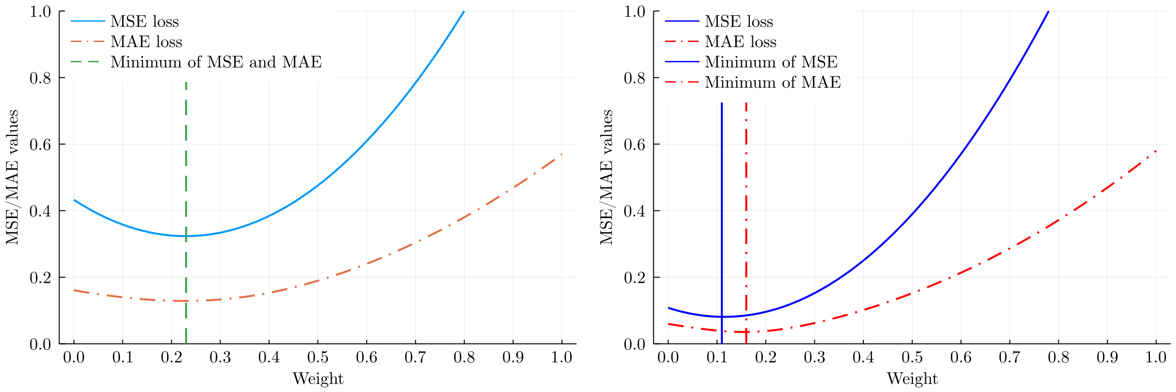
$$\begin{aligned} \underset{a_1, a_2}{\text{minimize}} \quad & \mathbb{E}\ \left[h\left(a_1 \nu_{1t} + a_2 \nu_{2t}\right)\right] \\ \text{subject to} \quad & a_1 + a_2 = c, \end{aligned} \tag{4}$$

[3] The URL for the GitLab repository is https://gitlab.com/chansta/mae-forecast/-/tree/master.

where $\mathbb{E}$ denotes the expected value, the combined forecast error is $a_1\nu_{1t}+a_2\nu_{2t} = y_t - a_1 f_{1t} - a_2 f_{2t}$ and is assumed to be second-order stationary, and $c \in \mathbb{R}$ is a constant that is typically set to one. The optimal weights are found by minimizing problem (4). Note that the best forecast is not included in the combination set of forecasts because if it were included, the optimal solution to the combination would weight the best forecast exclusively.[4] Moreover, the setup in this paper does not exclude forecasts from any combination of forecasts.

When combining multiple forecasts, h is usually specified to be the mean squared error such that the optimal solution minimizes the forecast error variance, i.e., $(a_1\nu_{1t} + a_2\nu_{2t})^2$. MSE is a convenient loss function to minimize because of its close form solution and interpretability (minimum variance). Optimal weights obtained by minimizing the forecast error variance were first introduced by Bates and Granger (1969). Since then, the optimality properties of combining forecasts by minimizing MSE have been investigated theoretically and empirically throughout the literature. See, among others, Smith and Wallis (2009), Elliott (2011), Claeskens et al. (2016), and Chan and Pauwels (2018). To our knowledge, the same cannot be said about mean absolute error loss. In the MAE case, h is expressed in absolute value form as such $|a_1\nu_{1t} + a_2\nu_{2t}|$.

For the purpose of demonstration, suppose that the two forecast error distributions are asymmetric specifically, they follow the skew normal distribution as defined in Azzalini (1985) with $\nu_1 \sim SN(\mu = 0, \sigma = 1.5, \lambda = 0.4)$ and $\nu_2 \sim SN(\mu = 0, \sigma = 0.4, \lambda = -0.3)$. Figure 1 depicts the MAE and MSE loss curves specified in problem (4) for a range of values of the forecast combination weight, for example, a_1. The optimal points are shown by the vertical lines cutting through the MAE and MSE loss curves.[5]



(a) Weights sum to 1

(b) Weights sum to 0.5

Figure 1: MAE and MSE optimal weights

[4] This is demonstrated formally in Corollary S.1, which can be found in Online Supplement.

[5] The exact framework and the computer code for Figure 1 are available in this Jupyter Notebook.

The left-hand side panel of Figure 1 shows the MAE and MSE loss curves for forecast combinations under the constraint that the weights sum to one ($c = 1$). The optimal weights of minimizing MAE and MSE loss under this constraint are the same and equal to 0.23. The right-hand side panel of Figure 1 depicts the same MAE and MSE loss but under the constraint that the weights sum to half ($c = 0.5$). Interestingly, in this case the optimal weight of minimizing the MAE loss is 0.16, whereas the optimal weight of minimizing the MSE loss is 0.11. This finding remains true for any value of the constraint not equal to one, i.e., $c \neq 1$. This simple example shows that there is an equivalence between MAE and MSE optimal weights when $c = 1$. However, this does *not* imply equivalence in the loss functions. Importantly, this equivalence seems to suggest that the 'sum to one' constraint is instrumental to this result.

A natural question that arises is what is special about $c = 1$? One implication of the budget constraint is that it ensures that forecast combinations are unbiased. To see this, let f_{1t} and f_{2t} be unbiased forecasts, i.e., $\mathbb{E}(f_{1t}) = \mathbb{E}(f_{2t}) = \mathbb{E}(y_t)$, combined as in equation (3) with the constraint $a_1 + a_2 = c$. When $c = 1$, it is clearly a special case, implying that

$$\begin{aligned}\mathbb{E}\left(f_{ct}\right) &= a_1\,\mathbb{E}\left(f_{1t}\right) + a_2\,\mathbb{E}\left(f_{2t}\right)\\ &= a_1\,\mathbb{E}(y_t) + (c - a_1)\ \mathbb{E}(y_t)\\ &= c\,\mathbb{E}(y_t).\end{aligned}$$

Thus $\mathbb{E}(f_{ct}) = \mathbb{E}(y_t)$ if and only if $c = 1$. When $c \neq 1$, the forecast combination is biased by a factor of c. The unbiasedness implication of the budget constraint imposes a restriction on the loss function, which leads to the equivalence result of this paper. The relaxation of this restriction invokes the possibility of a *bias-variance* trade-off. While it is possible to have a lower MSE with higher absolute values of c and hence more bias, there is no reason to believe that MAE will follow suit. MAE and MSE loss functions do not have to behave in the same way around such a trade-off; the bias will be handled differently. Hence, there are no guarantees that the minimum points on the MAE and MSE losses could coincide when $c \neq 1$.

Of course this simple example has its limitations. In practice, forecast combinations often involve more than two forecasts. Moreover, forecasting requires samples of data, which often involve estimation and random errors, outliers, and other data-related problems. The next two sections of the paper generalize and formalize the findings in this example with statistical theory and a numerical study.

3 Theory

3.1 Minimizing mean absolute error

The aim of combining the set of k forecasts is to optimize the forecast accuracy by minimizing the combined forecast errors, i.e., u_{it}. The k forecast equations (1) are written in matrix form as

$$\mathbf{f}_t = (y_t - \nu_{0t})\,\mathbf{1} - \mathbf{u}_t \tag{5}$$

where $\mathbf{1}$ denotes a $k \times 1$ vector of ones, $\mathbf{f}_t = (f_{1t}, \cdots, f_{kt})^\top$ and $\mathbf{u}_t = (u_{1t}, \cdots, u_{kt})^\top$. The linear combination of the k forecasts at time t can be written as follows:

$$\mathbf{f}_t^\top \mathbf{a} = (y_t - \nu_{0t})\,\mathbf{1}^\top \mathbf{a} - \mathbf{u}_t^\top \mathbf{a}. \tag{6}$$

where $\mathbf{a} \in \mathbb{R}^n$ is a vector of weights. If the combination of forecasts is linear, i.e. $\mathbf{a}^\top \mathbf{1} = 1$, then $y_t - \mathbf{f}_t^\top \mathbf{a} = \nu_{0t} + \mathbf{u}_t^\top \mathbf{a}$ is a $T \times 1$ vector containing the forecast errors from the forecast combination.

Several results on the optimality of minimizing the mean absolute error are derived in this section, which is then used to derive the equivalence results central to this paper. The optimization problem in (4) can be generalized to combining k forecasts at time t and formulated for mean absolute error loss as such

$$\begin{aligned} \underset{\mathbf{a}}{\text{minimize}} \quad & \mathbb{E}\left|\nu_{0t} + \mathbf{u}_t^\top \mathbf{a}\right| \\ \text{subject to} \quad & \mathbf{1}^\top \mathbf{a} = 1. \end{aligned} \tag{7}$$

The solution that minimizes the MAE loss function in (7), denoted as $\mathbf{a}^*_{\text{MAE}}$, is often found by linear programming. Standard techniques to solve optimization problems with a nondifferentiable function include transforming the problem into a linear programming problem or utilizing other sophisticated methods such as generalized gradient (for further details see Chapter 2 in Clarke, 1990). In this case, the optimal solution can be found by using the standard Lagrangian technique.

The optimization problem (7) can then be restated by grouping the forecast errors $(\nu_{0t} + \mathbf{u}_t^\top \mathbf{a})$ according to their signs. Let $\left(\mathbb{R}^{k+1}, \Im, G\right)$ be a probability space and define the positive set $X_{\mathbf{a}}^+ = \left\{(\nu_{0t}, \mathbf{u}_t) : \nu_{0t} + \mathbf{u}_t^\top \mathbf{a} > 0\right\}$ and the negative set $X_{\mathbf{a}}^- = \left\{(\nu_{0t}, \mathbf{u}_t) : \nu_{0t} + \mathbf{u}_t^\top \mathbf{a} < 0\right\}$. Then the MAE loss function in (7) is rewritten as

$$\mathbb{E}\left|\nu_{0t} + \mathbf{u}_t^\top \mathbf{a}\right| = \int_{X_{\mathbf{a}}^+} \left(\nu_{0t} + \mathbf{u}_t^\top \mathbf{a}\right) G\left(dv_t\right) - \int_{X_{\mathbf{a}}^-} \left(\nu_{0t} + \mathbf{u}_t^\top \mathbf{a}\right) G\left(dv_t\right) + \int_{X_{\mathbf{a}}^0} \left(\nu_{0t} + \mathbf{u}_t^\top \mathbf{a}\right) G\left(dv_t\right),$$

where $dv_t = d\nu_{0t} du_{1t} \ldots du_{kt}$. Note that $X_{\mathbf{a}}^0 = \left\{ (\nu_{0t}, \mathbf{u}_t) : \nu_{0t} + \mathbf{u}_t^\top \mathbf{a} = 0 \right\}$ and the last integral is 0 because $\nu_{0t} + \mathbf{u}_t^\top \mathbf{a} = 0$. Hence, the optimization problem (7) can be restated as

$$\begin{aligned} \underset{\mathbf{a}}{\text{minimize}} \quad & \int_{X_{\mathbf{a}}^+} \left(\nu_0 + \mathbf{u}^\top \mathbf{a} \right) G(dv) - \int_{X_{\mathbf{a}}^-} \left(\nu_0 + \mathbf{u}^\top \mathbf{a} \right) G(dv) \\ \text{subject to} \quad & \mathbf{1}^\top \mathbf{a} = 1. \end{aligned} \tag{8}$$

Under the assumption of stationarity, $G\left(dv\right) = G\left(dv_t\right)$ for all t, so the subscript t can be omitted from this point on.

Lemma 1 proposes mapping the optimization problem in (8) to a differentiable function. While the loss function is nondifferentiable at 0, the expectation of the loss function is differentiable for all values. Lemma 1 establishes the existence of the derivative, which is required to obtain the first-order conditions of problem (8):

Lemma 1. *Let $F(\mathbf{a}) = \int_{X_{\mathbf{a}}^+} \left(\nu_0 + \mathbf{u}^\top \mathbf{a} \right) G(dv) - \int_{X_{\mathbf{a}}^-} \left(\nu_0 + \mathbf{u}^\top \mathbf{a} \right) G(dv)$; then,*

$$\frac{\partial F}{\partial \mathbf{a}} = \int_{X_{\mathbf{a}}^+} \mathbf{u} \, G(dv) - \int_{X_{\mathbf{a}}^-} \mathbf{u} \, G(dv). \tag{9}$$

Proof. See Appendix A. □

Using the first derivative obtained in Lemma 1, Theorem 1 shows the optimal solution of problem (8):

Theorem 1. *Let $\omega_i(\mathbf{a}) = \int_{X_{\mathbf{a}}^+} u_i \, G(dv) - \int_{X_{\mathbf{a}}^-} u_i \, G(dv)$; then, the solution to the optimization problem, weights $\mathbf{a}^*$, as stated in equation (8), satisfies*

$$\omega_i \left(\mathbf{a}^* \right) = \omega_j \left(\mathbf{a}^* \right) \quad \forall i, j = 1, \ldots, k. \tag{10}$$

Proof. See Appendix A. □

If $\mathbf{1}^\top \mathbf{a} = 1$ is a binding constraint, then $\boldsymbol{\omega}(\mathbf{a}^*) = 0$, where $\boldsymbol{\omega}(\mathbf{a}) = \int_{X_{\mathbf{a}}^+} \mathbf{u} G(d\nu) - \int_{X_{\mathbf{a}}^-} \mathbf{u} G(d\nu)$.[6] Equation (10) of Theorem 1 has a more intuitive representation in terms of the expectation of $\mathbf{u}$ conditional on the sign of $\nu_{0t} + \mathbf{u}_t^\top \mathbf{a}$, namely,

$$\mathbb{E}\left[\mathbf{u} | \nu_0 + \mathbf{u}^\top \mathbf{a}^* > 0 \right] \Pr\left(\nu_0 + \mathbf{u}^\top \mathbf{a}^* > 0 \right) = \mathbb{E}\left[\mathbf{u} | \nu_0 + \mathbf{u}^\top \mathbf{a}^* < 0 \right] \Pr\left(\nu_0 + \mathbf{u}^\top \mathbf{a}^* < 0 \right). \tag{11}$$

[6] It is worth pointing out that $\omega_i(\mathbf{a}^*) = -\lambda$ for all i, where λ is the Lagrange multiplier associated with the budget constraint $\mathbf{1}^\top \mathbf{a} = 1$.

The geometric interpretation of equation (11) can be explained as follows.[7] First, consider the case where all forecast errors follow symmetric distributions, that is, $\Pr\left(\nu_0 + \mathbf{u}^\top \mathbf{a}^* > 0\right) = \Pr\left(\nu_0 + \mathbf{u}^\top \mathbf{a}^* < 0\right)$; then, equation (11) reduces to

$$\mathbb{E}\left[\mathbf{u}|\nu_0 + \mathbf{u}^\top \mathbf{a}^* > 0\right] = \mathbb{E}\left[\mathbf{u}|\nu_0 + \mathbf{u}^\top \mathbf{a}^* < 0\right]. \tag{12}$$

This means that the optimal solution, weights $\mathbf{a}^*$, is a hyperplane that satisfies $\mathbf{a}^{*\top}\mathbf{u} = -\nu_0$ and divides the space in such a way that the conditional expectation of $\mathbf{u}$ in both half spaces are equal. Second, in the case where the forecast errors do not follow symmetric distributions, the equality of the conditional expectations from the two half spaces is compensated by the skewness of the error distribution for the optimal combination of forecasts.

3.2 Equivalence in optimal solutions

In this section, we show that the optimal combination of forecasts obtained from minimizing the constrained MAE loss in (8) produces the same optimal combination of forecasts when minimizing an MSE loss function under the same constraint, i.e., the weights sum to one. As an exploratory exercise, this equivalence can be demonstrated intuitively under the assumption that forecast errors are normally distributed.

Let the combined forecast errors be $z_t = \nu_{0t} + \mathbf{u}_t^\top \mathbf{a}$ and $z_t \sim N(0, \sigma_z^2)$ such that $\sigma_z^2 = \sigma_\nu^2 + \mathbf{a}^\top \mathbf{\Omega} \mathbf{a}$, where $\sigma_\nu^2 = \mathbb{E}(\nu_{0t}^2)$ and $\mathbb{E}(\mathbf{u}_t \mathbf{u}_t^\top) = \mathbf{\Omega}$. Note that the expected value of the absolute combined error is

$$\begin{aligned}\mathbb{E}|z_t| =& \sigma_z \int_0^\infty w\phi(w)dw - \sigma_z \int_{-\infty}^0 w\phi(w)dw \\ =& \frac{2}{\sqrt{2\pi}}\sigma_z,\end{aligned}$$

where $\phi(w)$ denotes the standard normal density function. The last line above suggests that minimizing $\mathbb{E}|z_t|$ subject to the constraint that the weights sum to one is the same as minimizing the standard error of z_t subject to the same constraint. Since minimizing the standard deviation gives the same results as minimizing the variance under the same constraint, the optimization problem under the MAE loss has the same solution, i.e., it produces the same weight $\mathbf{a}^*$, as the optimization problem under the MSE loss when the combined forecast errors are normal.[8]

The result is not limited to normally distributed errors. A more general result can be found in Proposition 1 below.

[7] See the Online Supplement for full derivation.

[8] This intuition is formalized in Proposition S.1 in the Online Supplement.

Proposition 1. *Define $\nu_t = \nu_{0t} + \mathbf{u}_t^\top a$ and under the assumptions that (i) ν_t is independently distributed over t and that (ii) $\mathbb{E}|\nu_t|$, $\mathbb{E}|\nu_{0t}|$ and $\mathbb{E}|\mathbf{u}_t|$ exist and are finite for all t, then $\mathbf{a}^*_{\text{MAE}} = \mathbf{a}^*_{\text{MSE}}$ where $\mathbf{a}^*_{\text{MAE}}$ is defined in (7) and*

$$\begin{aligned} \underset{\mathbf{a}}{\textit{minimize}} \quad & \mathbb{E}\left(\nu_{0t} + \mathbf{u}_t^\top \mathbf{a}\right)^2 \\ \textit{subject to} \quad & \mathbf{1}^\top \mathbf{a} = 1, \end{aligned} \tag{13}$$

with closed form solution[9]

$$\mathbf{a}^*_{\text{MSE}} = \boldsymbol{\Omega}^{-1}\mathbf{1}\left(\mathbf{1}^\top\boldsymbol{\Omega}^{-1}\mathbf{1}\right)^{-1}. \tag{14}$$

Proof. See Appendix A. □

The strength of Proposition 1 is that there is no assumption on the distribution of the forecast errors. Providing that the two mild assumptions are satisfied and the sample considered is large enough, the optimal weight vector that combines forecasts from minimizing the MAE is the same as the one that minimizes the MSE. This equivalence result can be extended to other loss functions. Lemmas 3 and 4 in Appendix A provide a sufficient condition for the equivalence in Proposition 1 to hold for any loss function with unique minima under the constraint that the weights must sum to one. While the result can be generalized, the sufficient conditions may be too restrictive for the more general case. However, relaxing such sufficient conditions is beyond the scope of the current paper but may be an interesting direction for future research.

Thus far, the results have focused on the expected value. However, in practice, $\mathbf{a}^*_{\text{MAE}}$ is estimated by solving the following optimization problem using a sample of T observations:

$$\begin{aligned} \underset{\mathbf{a}}{\text{minimize}} \quad & \sum_{t=1}^{T} |y_t - \mathbf{f}_t\mathbf{a}| \\ \text{subject to} \quad & \mathbf{a}^\top\mathbf{1} = 1. \end{aligned} \tag{15}$$

The solution to (15) is denoted as $\hat{\mathbf{a}}^*_{\text{MAE}}$. Under the assumptions of Proposition 1, it is trivial to show that $\hat{\mathbf{a}}^*_{\text{MAE}} - \mathbf{a}^*_{\text{MAE}} = o_p(1)$, that is, that the difference converges in probability to zero. Note that the weights are not restricted to be positive, as is sometimes required in this literature.

It is important to note that the equivalence result is an asymptotic result, which means that the finite sample forecasting performance of MAE or MSE weights will vary from sample to sample. Chan et al. (2020) provides a discussion that $\hat{\mathbf{a}}^*_{\text{MSE}}$ and $\hat{\mathbf{a}}^*_{\text{MAE}}$ share the same finite

[9] See Elliott (2011) and Chan et al. (2020) for a derivation and discussion of the optimal MSE weights, $\mathbf{a}^*_{\text{MSE}}$.

sample approximation to their distributions, as they both converge to the same asymptotic distribution with the same variance-covariance matrix. The distribution of the MAE optimal weights is derived from this paper's Proposition 1 and the literature of epiconvergence, specifically, Corollary 2 in Knight (2001). The basic idea of epiconvergence is that, if $\hat{\mathbf{a}}^*_{\text{MSE}}$ is a solution to minimizing the MAE loss function, then $\hat{\mathbf{a}}^*_{\text{MAE}}$ converges in distribution to $\hat{\mathbf{a}}^*_{\text{MSE}}$. Therefore, it is sufficient to show that $\hat{\mathbf{a}}^*_{\text{MAE}}$ converges to $\hat{\mathbf{a}}^*_{\text{MSE}}$, which is given under the conditions of Proposition 1.

4 Simulations

The simulation study provides supporting evidence that the optimal weights obtained from minimizing either the MAE or the MSE loss functions are equivalent, as shown in Proposition 1. The simulation framework can be summarized as follows. We simulate, in turn, five forecast error series that come only from skew normal distributions and then only from t_3 distributions. These error distributions are chosen because the skew normal distribution is asymmetric and the t_3 distribution is fat-tailed, which means that the forecast errors contain outliers. Both types of distributions provide a nontrivial simulation framework to generate supporting evidence in favor of equivalence.

We minimize the combined errors under MAE loss and under MSE loss and collect sets of five weights for each optimization sequence and for each sample size N. The sample sizes considered are $N = \{20, 30, 50, 100\}$ and then in increments of 100 up to 1000, and the number of replications per sample is 5000.

We provide a brief description of how the skew normal forecast errors are simulated below. More details are available at Online Supplement.[10] We simulate $p = 5$ correlated skew normal random variates $\mathbf{Z}_{SN} \sim SN(\mathbf{0}, \mathbf{\Sigma}, \mathbf{\Lambda})$. The multivariate skew normal variates $\mathbf{Z}_{SN}$ depend on parametrizing two matrices, namely, a $p \times p$ matrix $\mathbf{\Lambda}$ and a $p \times p$ matrix $\mathbf{\Sigma}$. The $\mathbf{\Lambda}$ matrix gathers the skewness parameters for each of the p variates, and $\mathbf{\Sigma}$ is a semipositive definite matrix containing the correlation information between the p variates. We provide two parametrizations of $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$ as it produces different sets of optimal weights: one set with positive weights only and the other with a mix of positive and negative weights. The optimal weights come from the closed form solution $\mathbf{a}^* = \mathbf{\Omega}^{-1}\mathbf{i}\big(\mathbf{i}'\mathbf{\Omega}^{-1}\mathbf{i}\big)^{-1}$, where the variance-covariance matrix is $\mathbf{\Omega} = \mathbf{\Sigma} + \big(1 - \frac{2}{\pi}\big)\mathbf{\Lambda}\mathbf{\Lambda}'$.

Set 1: $\mathbf{a}^*$ such that $a_1 = 0.290$, $a_2 = 0.193$, $a_3 = 0.182$, $a_4 = 0.079$, and $a_5 = 0.255$.

Set 2: $\mathbf{a}^*$ such that $a_1 = 0.259$, $a_2 = -0.311$, $a_3 = 0.805$, $a_4 = 0.374$, and $a_5 = -0.128$.

[10] The Julia code for the simulations is available in this Jupyter Notebook. The code is also available for MatLab. The simulation results presented in this section are available in csv files.

The same two $\mathbf{\Omega}$ matrices shown in Sets 1 and 2 are used to generate a multivariate t distribution with three degrees of freedom, i.e., $\mathbf{Z}_{t3} \sim t_3(\mathbf{0}, \mathbf{\Omega})$. A convenient implication is that the set of weights a_i are the same because the skew normal and the t_3 variates have the same variance-covariance matrix.[11]

Figures 2 - 5 present a summary of the simulation results for the two forecast error distributions. Each figure presents the difference between the MAE weights and the MSE weights, i.e., $\hat{\mathbf{a}}_{\text{MAE}} - \hat{\mathbf{a}}_{\text{MSE}}$, and the plus minus one standard deviation of $\hat{\mathbf{a}}_{\text{MAE}} - \hat{\mathbf{a}}_{\text{MSE}}$. Only weight a_1 is shown for each case to keep the presentation concise. The complete simulation results displaying all the weights are shown in the Online Supplement.[12] The conclusions found for a_1 also apply to the four other weights.

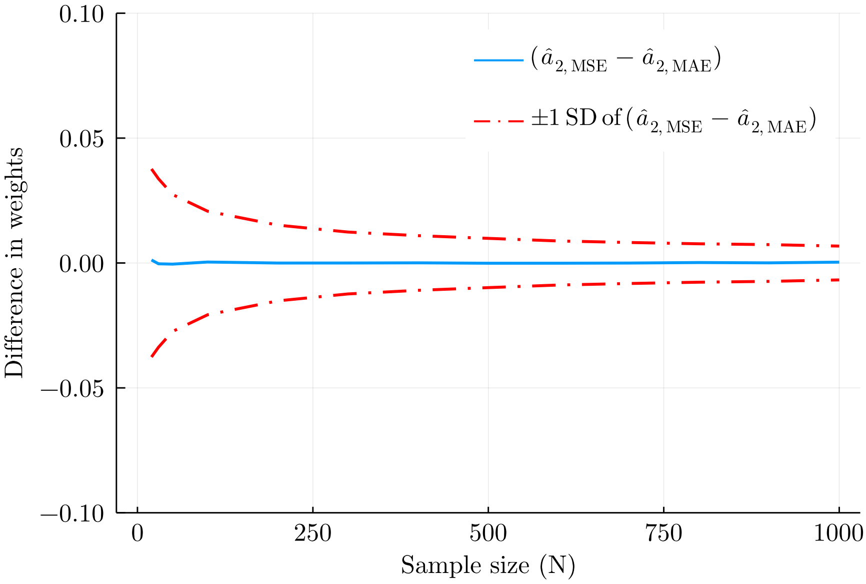


Figure 2: Skew Normal (Set 1)

The simulation results clearly support the theoretical findings in Proposition 1 in the variety of cases presented. Figures 2 and 3 show that the difference in the MAE and the MSE combination weight a_1 is very close to zero on average for all samples when the forecast error distributions are skew normal. The standard deviation of the difference also narrows as the sample size N grows large. The simulations show that in large samples, the difference between the MAE and MSE weights becomes negligible, supporting the equivalence presented in the theory. The same conclusion regarding the equivalence between the MAE and MSE

[11] Technically, the variance-covariance of $\mathbf{Z}_{t3}$ is $r/(r-2) \times \mathbf{\Omega}$, where $r > 2$ and r is the degrees of freedom of the t_r distribution. However, the theoretical weights remain the same as the $r/(r-2)$ cancel out.

[12] Note that only weights a_1 to a_4 are shown since $a_5 = 1 - \sum_{i=1}^{4} a_i$, and if all first four weights converge, so will the fifth.

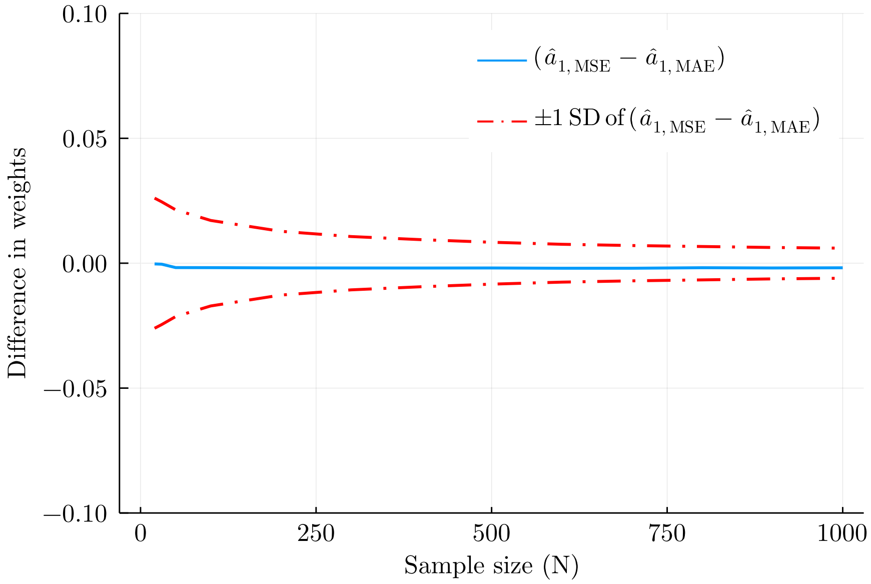


Figure 3: Skew Normal (Set 2)

weights can be drawn in Figure 4 and 5. However, the slightly wider standard deviation for smaller values of N arises from the fact that the standard deviation of the MSE weight is higher than MAE. The estimated variance-covariance matrix for the MSE weight is naturally more sensitive to outliers frequently encountered in very fat-tailed distributions, such as t_3, especially in smaller sample sizes. That difference reduces as N grows large. This observation is in line with the common view that MAE produces estimates that are less sensitive to outliers (see, for example, Gupta and Wilton, 1987, and Winkler and Clemen, 1992). We also observe this property with real-world data, which we turn to next.

5 Illustration

In this section, we turn our attention to a real-world small sample data application. The purpose of this empirical illustration is to gain insights into the practical use and implications of MAE and MSE optimal weights when aggregating expert forecasts. Each expert forecast represents an independent prediction of a key economic indicator. The results highlight that MAE optimal weights offer better forecasting performance than MSE optimal weights when the forecast horizons are short and outliers are present. Furthermore, in some cases, the optimal weights from both weighting schemes show signs of equivalence in small data samples.

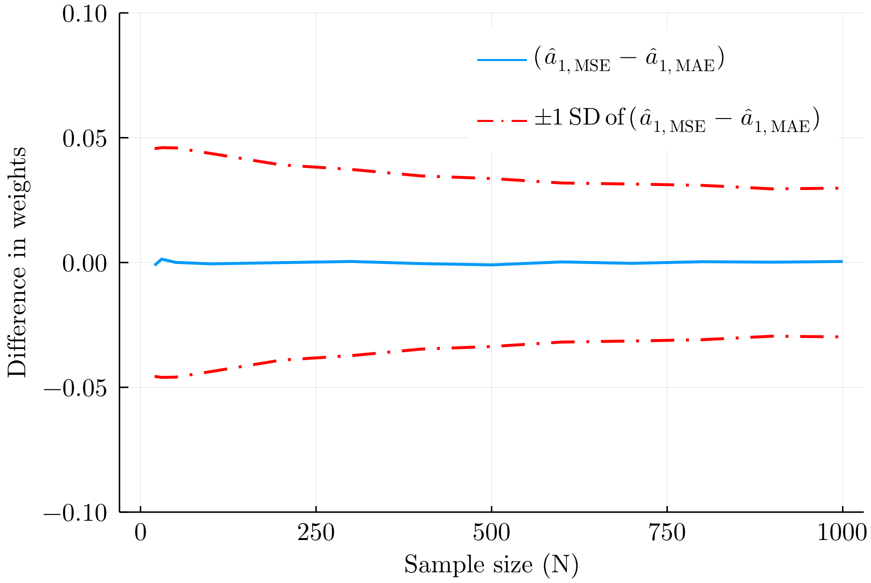


Figure 4: t_3 (Set 1)

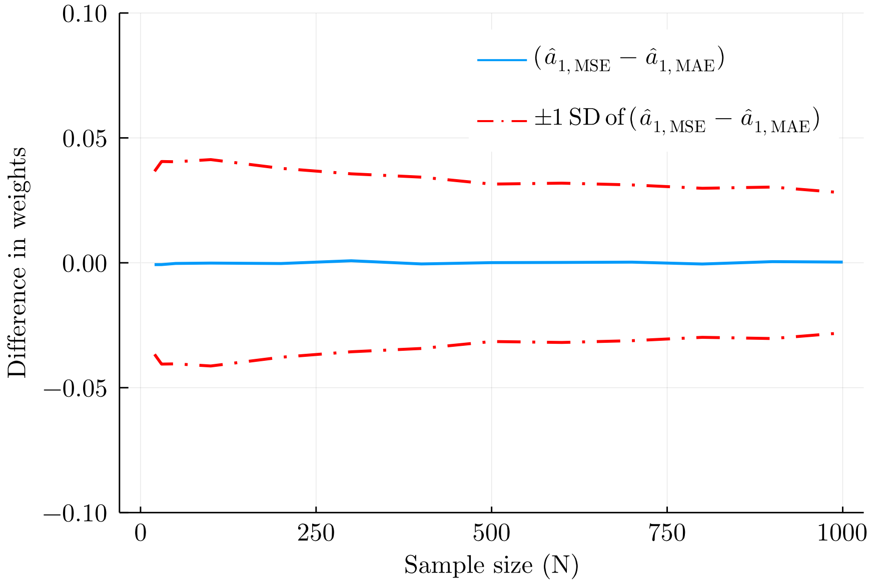


Figure 5: t_3 (Set 2)

5.1 Data and preliminary analysis

The expert forecasts for this illustration are obtained from the European Central Bank Survey of Professional Forecasters, which provides a comprehensive database of forecasts from various European Union-based experts affiliated with financial or nonfinancial institutions at a variety of horizons for key macroeconomic indicators.[13] In this study, we focus on the expert-predicted rate of inflation, real GDP growth, and unemployment for the European Union for the coming year (one year ahead).

The survey is conducted quarterly starting from Q3 1999 and ends in Q4 2018 for this illustration, covering approximately 19 years of economic activity in the European Union.[14] There are approximately 100 forecasters in the survey, but as is often the case in long-term surveys, there are many instances of nonresponse. Since the optimal weights need to be estimated over the entire historical period, only the forecasters that respond the most consistently and in the same surveyed periods are kept for analysis.[15] This amounts to four forecasters for each of the three economic indicators over the entire sample period.[16]

Figure 6 presents the predictions of the four forecasters for a given economic indicator.[17] Not surprisingly, all economic indicators exhibit larger fluctuations around the Global Financial Crisis of 2008. As a result, we are likely to observe larger individual forecast errors around this period, that is, a larger difference between the economic indicator and its respective forecasts. Furthermore, under these conditions, we could expect that the estimated MAE optimal weights provide better forecasting performance, as MAE weights are robust to outliers, which in this case would manifest as very large forecast errors. If, on the other hand, such outliers are not present, we can expect that either MSE or MAE weights would perform adequately.

How does one capture large forecast errors or outliers? A simple way to detect the presence of outliers in forecast errors is to measure the sample kurtosis. Table 1 presents the sample kurtosis of the forecast errors for each of the four forecasters and for the corresponding economic indicator. The kurtosis value for a standard normal distribution is 3. When the sample kurtosis values are less than 3, it indicates fewer expected outliers than when it is

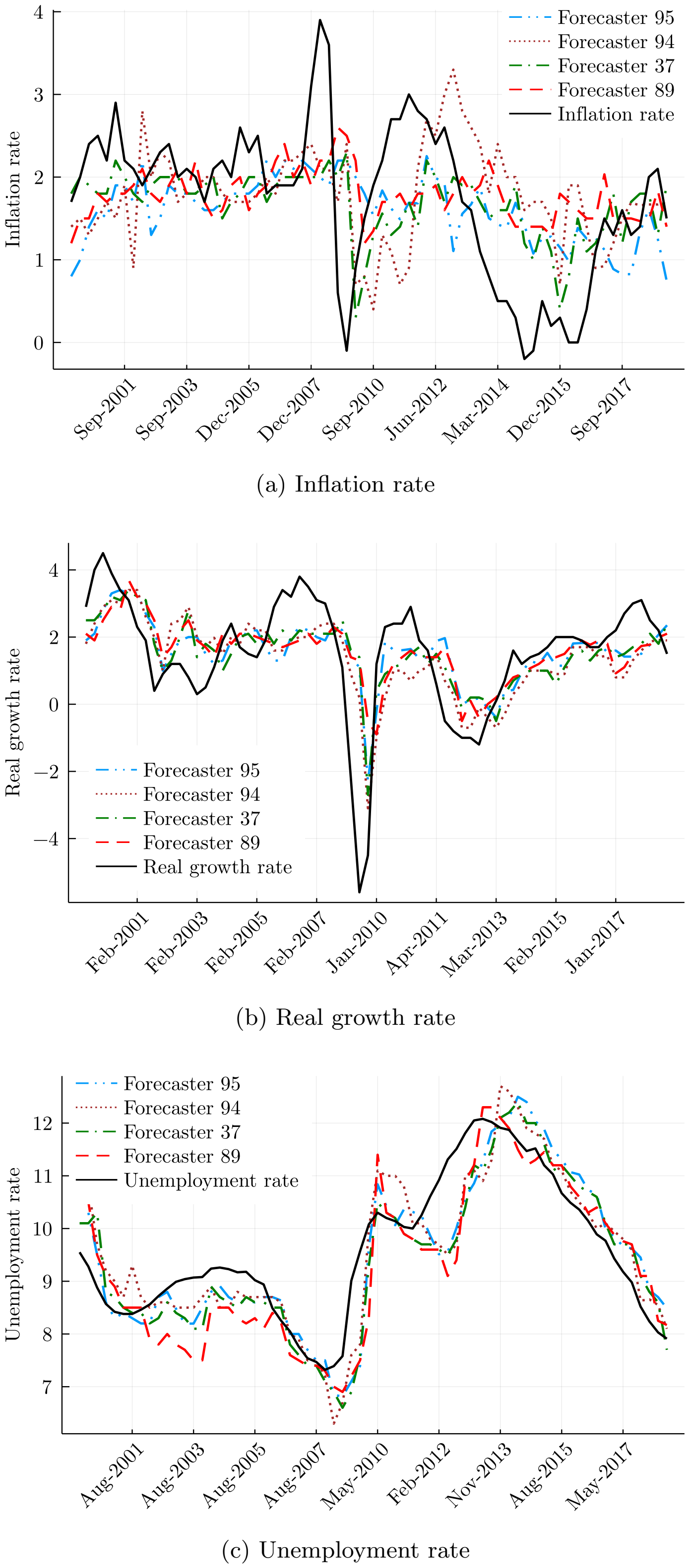
[13] The data are available publicly at `https://www.ecb.europa.eu/stats/ecb_surveys/html/index.en.html`. See Matsypura et al. (2018) for a recent comprehensive analysis of the data in the context of the aggregation of forecasts.

[14] The exact sample size is 68 data points for the inflation and unemployment rates and 69 for the growth rate. Each indicator's exact time coverage varies slightly.

[15] The actual values of the variables are occasionally revised by the ECB, but the revisions do not impact the performance of different forecasting methods, as discussed in Genre et al. (2013).

[16] The ECB SPF forecaster IDs correspond to 37, 89, 94 and 95. All four forecasters are mostly consistent over the time periods considered.

[17] The Julia code for the analysis in this section is available in a Jupyter Notebook with the relevant data. The code is also available for MatLab.



(a) Inflation rate

(b) Real growth rate

(c) Unemployment rate

Figure 6: Forecasters and economic indicators.

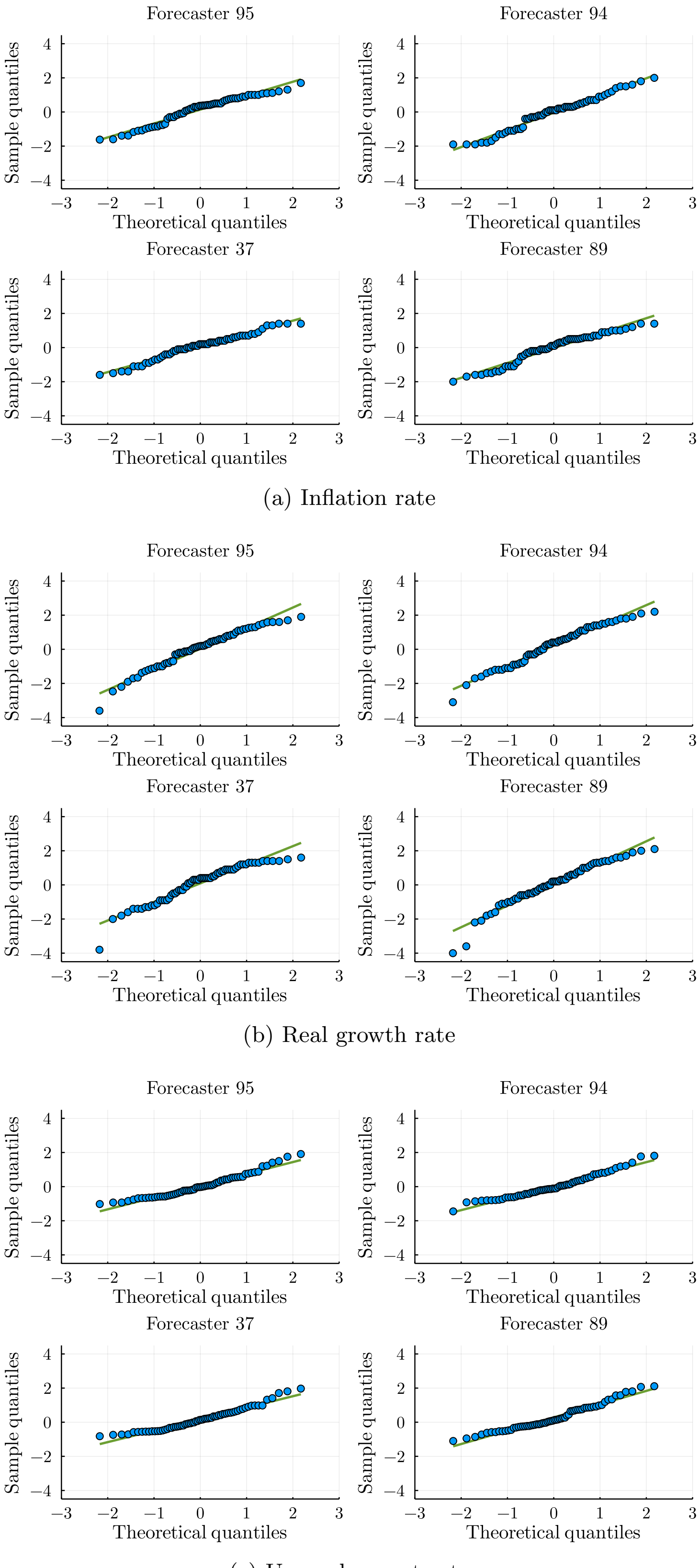
Table 1: Sample kurtosis & Jarque-Bera normality tests for forecast errors

Macro.		Professional Forecasters			
indicator	Statistic	Fcst 95	Fcst 94	Fcst 37	Fcst 89
Inflation	Kurtosis	2.86	2.61	3.45	2.78
	JB p-value	0.15	0.62	0.22	0.11
Growth	Kurtosis	9.32	6.85	11.34	8.45
	JB p-value	$<$ **0.01**	$<$ **0.01**	$<$ **0.01**	$<$ **0.01**
Unemployment	Kurtosis	3.33	3.03	3.24	2.71
	JB p-value	**0.03**	0.22	0.10	0.17

Notes: The table presents the sample kurtosis for the forecast errors of each of the four professional forecasters from the European Central Bank Survey of Professional Forecasters, namely forecasters 37, 89, 94 and 95. The three macroeconomic indicators forecast are the real growth rate, the inflation rate and the unemployment rate in the European Union. The sample spans from 1999 Q3 until 2018 Q4. The table also reports the p-values of the Jarque-Bera test for normality. The values in bold font reject the null hypothesis at the 5% level.

above 3. This is the case for the forecast errors of the inflation rate, which displays the lowest kurtosis of the three economic indicators and the unemployment rate. Both indicators have sample kurtosis around 3 or less. The real growth rate, however, presents the largest kurtosis values between 6.85 and 11.34, which indicates that outliers in the forecast errors are very likely.

We go one step further and test for normality in the individual forecast errors for each economic indicator. Table 1 also presents the p-value of the Jarque-Bera test. The results are consistent with what we observe with the sample kurtosis. Normality is systematically rejected at the 5% level for all four forecast errors of the real growth rate, whereas it is never rejected in the case of the inflation rate. The Jarque-Bera test results and the conclusions based on the sample kurtosis are further corroborated in Figure 7, showing quantile-quantile (QQ) plots for each individual forecast error series. The QQ plots of the individual forecast errors for the real growth rate in Figure 7 (b) show that the reason for non-normality lies in the tail due to the presence of outliers and, hence, can be attributed to the large kurtosis (rather than skewness). These findings have further use in explaining the empirical results below.



(a) Inflation rate

(b) Real growth rate

(c) Unemployment rate

Figure 7: Q-Q plots of individual forecast errors

Table 2: Forecast combination evaluation

Macro. indicator	Weighting scheme	Forecast evaluation MSFE	MAFE
Inflation	MSE	**0.691**	**0.667**
	MAE	0.707	0.705
Growth	MSE	1.404	1.013
	MAE	**1.023**	**0.842**
Unemployment	MSE	**0.528**	**0.582**
	MAE	0.675	0.679

Notes: The table presents the forecast combination of four professional forecasters from the ECB SPF, namely forecasters 37, 89, 94 and 95. The three macroeconomic indicators forecast are the real growth rate, the inflation rate and the unemployment rate in the European Union. The sample spans from 1999 Q3 until 2018 Q4. The two weighting schemes minimize MSE and minimize MAE. MSFE and MAFE are evaluated over the second half of the forecasting sample. The number presented are the average.

5.2 Methodology and results

All four forecasts are aggregated into two consensus forecasts: one with estimated MAE optimal weights and the other with MSE optimal weights. The weights are estimated recursively, starting with half of the sample, and then expanded by one data point until the end of the sample. We then compare mean square forecast error (MSFE) of the forecast combination with the estimated MSE weights and estimated MAE weights. We also do the same with mean absolute forecast error (MAFE). MSFE and MAFE are computed over the second half of the sample by using the estimated weights at time t to combine the forecasts at time $t + 1$ and construct forecast errors with the actual value of the economic indicator at time $t + 1$. This forecast combination exercise is conducted for all three economic indicators.

Table 2 presents the average MSFE and MAFE values for the inflation rate, real growth rate, and unemployment rate. The findings from these three forecast combination exercises can be succinctly described as follows. The MSFE and MAFE with the estimated MAE weights are consistently lower than those with the estimated MSE weights for the real growth rate. These results are consistent with the large sample kurtosis values in Table 1 and the QQ plots in Figure 7 (b), which pointed to the presence of outliers in the individual forecast errors. The estimated MAE weights are robust to such outliers. In contrast, the estimated MSE weights provide the best forecast combination performance in terms of both MSFE and MAFE for the unemployment and inflation indicators. Again, this is not surprising in light of the kurtosis measure in Table 1 and especially the QQ plots in Figure 7 (a) and (c).

The largest difference in MSFE between the estimated MAE and MSE weights comes from the real growth rate forecast combination, whereas the smallest difference is from the inflation rate forecast combination. The same is true for MAFE. While the discrepancy in forecasting evaluations can be explained by the presence of outliers in the case of the real growth rate, it is also manifested in the estimated weights themselves, as seen in Figure 8. The inflation rate, which shows the lowest MSFE and MAFE difference, also displays the smallest difference between the estimated MAE and MSE weights, as shown in Figure 8 (a). Moreover, Figure 8 (a) shows that the MAE and MSE optimal weights are close to equivalent, as predicted by theory, when the forecast errors are normally distributed.[18] Note that normality is supported empirically for the four sets of forecast errors for the inflation rate indicator, as seen from the Jarque-Bera test results in Table 1 and the QQ plot in Figure 7 (a). The equivalence is not as apparent for the unemployment rate and even less so for the real growth rate, which is most affected by the small sample and outliers in this exercise.

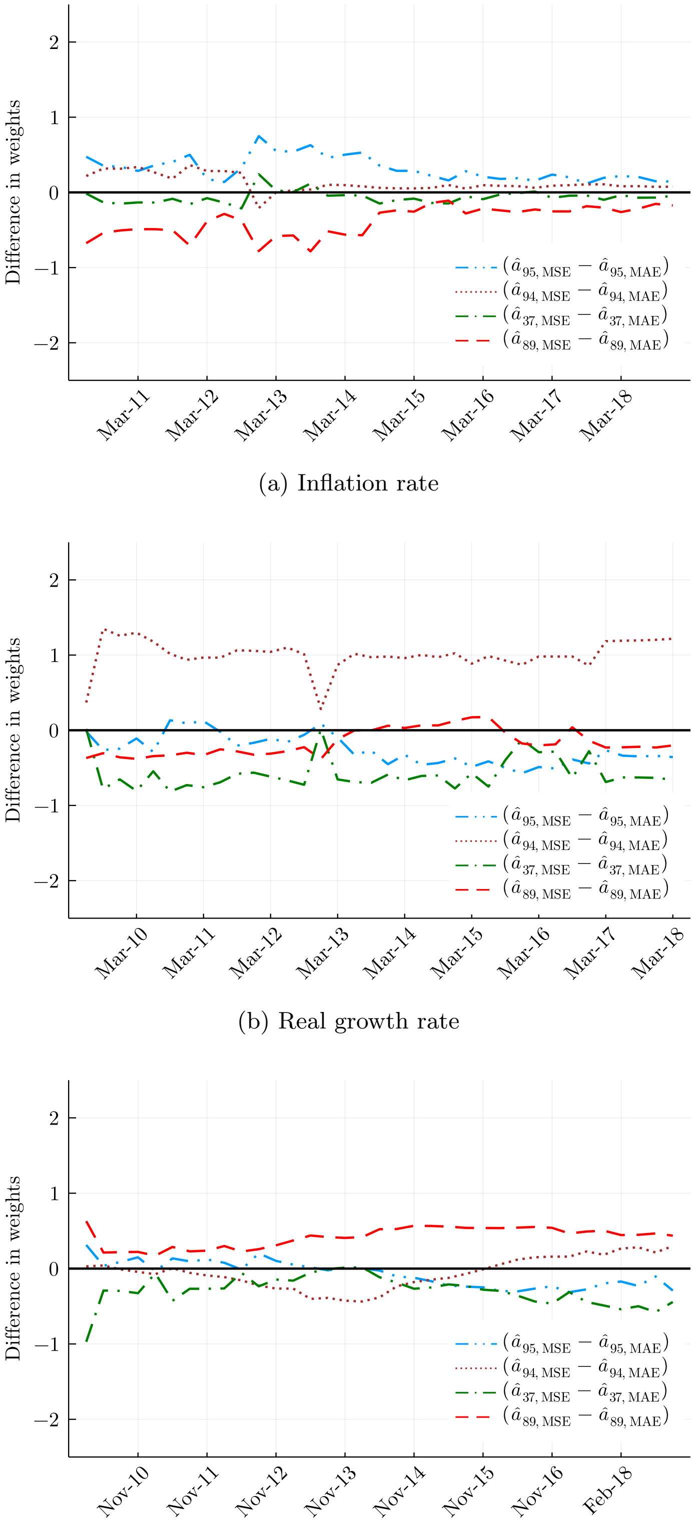
The results for the inflation rate are not unexpected. The ECB SPF's main purpose is to collect information on inflation expectations, as it is the ECB's mandate to ensure price stability with its policy tools. It is often perceived that inflation expectations convey the private sector's views on the macroeconomy (see García, 2003). While not being the primary focus, the forecasts from the two other economic indicators, the unemployment rate and the real growth rate, are also collected to give some Euro Area economy context. Both of these indicators tend to be harder to forecast, as they are not part of the official policy mandate of the ECB, unlike the inflation rate, but rather dependent on the economic climate of individual Euro Area countries.

6 Concluding comments

This paper demonstrates that, under a mild set of assumptions, there is an equivalence in optimal forecast combination between minimizing MAE and MSE loss functions, providing that the sample of data is large enough and the weights of the combination sum to one. This is demonstrated by theory derived in this paper. It is also supported by a simulation study featuring asymmetric and fat-tailed forecast error distributions.

In large samples, the advantages of using one weighting scheme over the other should mainly be guided by the convenience of computation and by the nature of the problem or dataset used. However, when dealing with small samples, the forecasting performance of weighting schemes is prone to estimation errors and outliers depending on the data. As shown with the ECB SPF of macroeconomic indicators for the European Union, the MAE

[18] See section 3.2 and Proposition S.1 in the Online Supplement.



(a) Inflation rate

(b) Real growth rate

(c) Unemployment rate

Figure 8: Difference between MAE and MSE forecast combination weights

optimal weighting scheme provides a more consistent strategy in small samples when outliers are present. In the case of the inflation rate, the equivalence in the MAE and MSE weights is observable even though the sample is small.

REFERENCES

Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, 171–178.

Banerjee, A., X. Guo, and H. Wang (2005, July). On the optimality of conditional expectation as a bregman predictor. IEEE Transactions on Information Theory 51(7), 2664–2669.

Bassett Jr., G. and R. Koenker (1978, September). Asymptotic theory of least absolute error regression. Journal of the American Statistical Association 73(363), 618–622.

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

Bjørnland, H. C., K. Gerdrup, A. S. Jore, C. Smith, and L. A. Thorsrud (2012). Does forecast combination improve norges bank inflation forecasts?*. Oxford Bulletin of Economics and Statistics 74(2), 163–179.

Bowles, C., R. Friz, V. Genre, G. Kenny, A. Meyler, and T. Rautanen (2010). An evaluation of the growth and unemployment forecasts in the ECB Survey of Professional Forecasters.

Bregman, L. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics 7(3), 200 – 217.

Capistrán, C. and A. Timmerann (2009). Forecast combination with entry and exit of experts. Journal of Business and Economic Statistics 27, 428–440.

Chan, F., L. Pauwels, and S. Soltyk (2020). Frequentist Averaging, pp. 329–357. Cham: Springer International Publishing.

Chan, F. and L. L. Pauwels (2018). Some theoretical results on forecast combinations. International Journal of Forecasting 34(1), 64–74.

Claeskens, G., J. R. Magnus, A. L. Vasnev, and W. Wang (2016). A simple theoretical explanation of the forecast combination puzzle. International Journal of Forecasting 32(3), 754–62.

Clarke, F. H. (1990). Optimization and Nonsmooth Analysis. Society for Industrial and Applied Mathematics.

Clemen, R. and R. Winkler (1986). Combining economic forecasts. Journal of Business and Economic Statistics 4, 39–46.

Conflitti, C., C. De Mol, and D. Giannone (2015). Optimal combination of survey forecasts. International Journal of Forecasting 31(4), 1096 – 1103.

de Menezes, L. M., D. W. Bunn, and J. W. Taylor (2000). Review of guidelines for the use of combined forecasts. European Journal of Operational Research 120(1), 190 – 204.

Diebold, F. X. and M. Shin (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. International Journal of Forecasting 35(4), 1679 – 1691.

Elliott, G. (2011, September). Averaging and the optimal combination of forecasts. University of California, San Diego.

Fung, G. and O. L. Mangasarian (2011, 10). Equivalence of minimal 0- and p-norm solutions of linear equalities, inequalities and linear programs for sufficiently small p. Journal of Optimization Theory and Applications 151, 1–10.

García, J. A. (2003, September). An introduction to the ECB's survey of professional forecasters. Occasional Paper Series 8, European Central Bank.

Gastwirth, J. L. (1974). Large sample theory of some measures of income inequality. Econometrica 42(1), 191–196.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013). Combining expert forecasts: Can anything beat the simple average? International Journal of Forecasting 29(1), 108 – 121.

Gupta, A. and P. C. Wilton (1987). Combination of forecasts: An extension. Management Science 33, 356–372.

Jose, V. R. R. (2017). Percentage and relative error measures in forecast evaluation. Operations Research 65(1), 200–211.

Kenny, G., V. Genre, C. Bowles, R. Friz, A. Meyler, and T. Rautanen (2007, April). The ECB survey of professional forecasters (SPF) - A review after eight years' experience. Occasional Paper Series 59, European Central Bank.

Knight, K. (2001). Limiting distributions of linear programming estimators. Extremes 2, 87–103.

Larrick, R. P. and J. B. Soll (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. Management Science 52(1), 111–127.

Lichtendahl, K. C., Y. Grushka-Cockayne, and R. L. Winkler (2013, July). Is It Better to Average Probabilities or Quantiles? Management Science 59(7), 1594–1611.

Matsypura, D., R. Thompson, and A. L. Vasnev (2018). Optimal selection of expert forecasts with integer programming. Omega 78(C), 165–175.

Patton, A. J. (2019). Comparing possibly misspecified forecasts. Journal of Business & Economic Statistics 0(0), 1–23.

Peng, J., S. Yue, and H. Li (2015, July). NP/CMP equivalence: A phenomenon hidden among sparsity models l_0 minimization and l_p minimization for information processing. IEEE Transactions on Information Theory 61(7), 4028–4033.

Rudin, W. (1976). Principles of Mathematical Analysis. Third Edition. McGraw-Hill.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. Journal of the American Statistical Association 66(336), 783–801.

Smith, J. and K. F. Wallis (2009). A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics 71(3), 331–355.

Soll, J. B. and R. P. Larrick (2009, May). Strategies for revising judgment: how (and how well) people use others' opinions. Journal of Experimental Psychology Learning Memory Cognition 35(3), 780–805.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. Granger, and A. Timmermann (Eds.), Handbook of Economic Forecasting, Vol 1. Elsevier, Amsterdam.

Winkler, R. L. and R. T. Clemen (1992). Sensitivity of weights in combining forecasts. Operations Research 40(3), 609–614.

Appendix A: Proofs and supplementary results

Proof of Lemma 1: It is sufficient to show that

$$\begin{aligned}\frac{\partial F}{\partial a_j} &= \lim_{h\to 0}\frac{F(\mathbf{a}+\mathbf{h}_i)-F(\mathbf{a})}{h} \\ &= \int_{X_{\mathbf{a}}^{+}} u_j g(\nu_0,\mathbf{u})dv - \int_{X_{\mathbf{a}}^{-}} u_j g(\nu_0,\mathbf{u}^\top)dv.\end{aligned} \tag{A.1}$$

Partitioning each X^+ and X^- sets into mutually exclusive sets, such that

$$\begin{aligned}X_{\mathbf{a}+\mathbf{h}_i}^{+} &= \left[X_{\mathbf{a}+\mathbf{h}_i}^{+}\cap X_{\mathbf{a}}^{+}\right]\cup\left[X_{\mathbf{a}+\mathbf{h}_i}^{+}\cap\left(X_{\mathbf{a}}^{-}\cup X_{\mathbf{a}}^{0}\right)\right] \\ X_{\mathbf{a}+\mathbf{h}_i}^{-} &= \left[X_{\mathbf{a}+\mathbf{h}_i}^{-}\cap X_{\mathbf{a}}^{-}\right]\cup\left[X_{\mathbf{a}+\mathbf{h}_i}^{-}\cap\left(X_{\mathbf{a}}^{+}\cup X_{\mathbf{a}}^{0}\right)\right] \\ X_{\mathbf{a}}^{+} &= \left[X_{\mathbf{a}+\mathbf{h}_i}^{+}\cap X_{\mathbf{a}}^{+}\right]\cup\left[\left(X_{\mathbf{a}+\mathbf{h}_i}^{-}\cup X_{\mathbf{a}+\mathbf{h}_i}^{0}\right)\cap X_{\mathbf{a}}^{+}\right] \\ X_{\mathbf{a}}^{-} &= \left[X_{\mathbf{a}+\mathbf{h}_i}^{-}\cap X_{\mathbf{a}}^{-}\right]\cup\left[\left(X_{\mathbf{a}+\mathbf{h}_i}^{+}\cup X_{\mathbf{a}+\mathbf{h}_i}^{0}\right)\cap X_{\mathbf{a}}^{-}\right]\end{aligned}$$

then equation (A.1) can be rewritten as

$$\frac{\partial F}{\partial a_j} = \lim_{h\to 0}\frac{1}{h}\left(A_1 + A_2 - A_3 + A_4\right) \tag{A.2}$$

where

$$A_1 = \int\limits_{X_{\mathbf{a}+\mathbf{h}_i}^{+}\cap X_{\mathbf{a}}^{+}} u_i h G(dv) - \int\limits_{X_{\mathbf{a}+\mathbf{h}_i}^{-}\cap X_{\mathbf{a}}^{-}} u_i h G(dv) \tag{A.3}$$

$$A_2 = \int\limits_{X_{\mathbf{a}+\mathbf{h}_i}^{+}\cap X_{\mathbf{a}}^{-}} \left(\nu_0 + \mathbf{u}^\top(\mathbf{a}+\mathbf{h}_i)\right) G(dv) + \int\limits_{X_{\mathbf{a}+\mathbf{h}_i}^{+}\cap X_{\mathbf{a}}^{-}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv) \tag{A.4}$$

$$A_3 = \int\limits_{X_{\mathbf{a}+\mathbf{h}_i}^{-}\cap X_{\mathbf{a}}^{+}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a} + u_i\mathbf{h}\right) G(dv) + \int\limits_{X_{\mathbf{a}+\mathbf{h}_i}^{-}\cap X_{\mathbf{a}}^{+}} \left(\nu + \mathbf{u}^\top\mathbf{a}\right) G(dv) \tag{A.5}$$

$$A_4 = \int\limits_{X_{\mathbf{a}+\mathbf{h}_i}^{+}\cap X_{\mathbf{a}}^{0}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a} + u_i\mathbf{h}\right) G(dv) - \int\limits_{X_{\mathbf{a}+\mathbf{h}_i}^{-}\cap X_{\mathbf{a}}^{0}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a} + u_i\mathbf{h}\right) G(dv) \tag{A.6}$$

$$- \int\limits_{X_{\mathbf{a}}^{+}\cap X_{\mathbf{a}}^{0}+\mathbf{h}_i} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv) + \int\limits_{X_{\mathbf{a}}^{-}\cap X_{\mathbf{a}+\mathbf{h}_i}^{0}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv) \tag{A.7}$$

For A_1,

$$\lim_{h\to 0}\frac{1}{h}A_1 = \int\limits_{X^+_{\mathbf{a}+\mathbf{h}_i}\cap X^+_{\mathbf{a}}} u_i G(dv) - \int\limits_{X^-_{\mathbf{a}+\mathbf{h}_i}\cap X^-_{\mathbf{a}}} u_i G(dv) \to \int\limits_{X^+_{\mathbf{a}}} u_i G(dv) - \int\limits_{X^-_{\mathbf{a}}} u_i G(dv)$$

For A_2,

$$\begin{aligned}\lim_{h\to 0}\frac{1}{h}A_2 &= \lim_{h\to 0}\frac{1}{h}\left\{\int\limits_{X^+_{\mathbf{a}+\mathbf{h}_i}\cap X^-_{\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top(\mathbf{a}+\mathbf{h}_i)\right) G(dv) + \int\limits_{X^+_{\mathbf{a}+\mathbf{h}_i}\cap X^-_{\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv)\right\}\\
&= \lim_{h\to 0}\frac{2}{h}\int\limits_{X^+_{\mathbf{a}+\mathbf{h}_i}\cap X_{-\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv) + \lim_{h\to 0}\int\limits_{X^+_{\mathbf{a}+\mathbf{h}_i}\cap X^-_{\mathbf{a}}} u_i G(dv)\\
&= \lim_{h\to 0}\frac{2}{h}\int\limits_{X^+_{\mathbf{a}+\mathbf{h}_i}\cap X_{-\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv).\end{aligned}$$

Similarly for A_3,

$$\begin{aligned}\lim_{h\to 0}\frac{1}{h}A_3 &= \lim_{h\to 0}\frac{1}{h}\left\{\int\limits_{X^-_{\mathbf{a}+\mathbf{h}_i}\cap X^+_{\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a} + u_i\mathbf{h}\right) G(dv) + \int\limits_{X^-_{\mathbf{a}+\mathbf{h}_i}\cap X^+_{\mathbf{a}}} \left(\nu + \mathbf{u}^\top\mathbf{a}\right) G(dv)\right\}\\
&= \lim_{h\to 0}\frac{2}{h}\int\limits_{X^-_{\mathbf{a}+\mathbf{h}_i}\cap X^+_{\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv) + \lim_{h\to 0}\int\limits_{X^-_{\mathbf{a}+\mathbf{h}_i}\cap X^+_{\mathbf{a}}} u_i G(dv)\\
&= \lim_{h\to 0}\frac{2}{h}\int\limits_{X^-_{\mathbf{a}+\mathbf{h}_i}\cap X^+_{\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv).\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{h\to 0}\frac{1}{h}A_2 - A_3 &= \lim_{h\to 0}\frac{2}{h}\left\{\int\limits_{X^+_{\mathbf{a}+\mathbf{h}_i}\cap X^-_{\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv) - \int\limits_{X^-_{\mathbf{a}+\mathbf{h}_i}\cap X^+_{\mathbf{a}}} \left(\nu_0 + \mathbf{u}^\top\mathbf{a}\right) G(dv)\right\}\\
&\triangleq \Delta_i\left(\mathbf{a}\right).\end{aligned}$$

Using an axiom from the extended real line, specifically $\infty.0 = 0$, it is then clear that

$$\Delta_i\left(\mathbf{a}\right) = 0 \qquad \forall\mathbf{a}.$$

For A_4,

$$\begin{aligned}\lim_{h\to 0}\frac{A_4}{h} =&2\left\{\int\limits_{X^{+}_{\mathbf{a}+\mathbf{h}_i}\cap X^{0}_{\mathbf{a}}} u_i G(dv) + \int\limits_{X^{+}_{\mathbf{a}}\cap X^{0}_{\mathbf{a}+\mathbf{h}_i}} u_i G(dv)\right\}\\ =&0.\end{aligned}$$

This completes the proof. ■

Proof of Theorem 1: The Lagrangian function for the optimization problem as defined in equation (8):

$$L\left(\mathbf{a},\lambda\right) = \int\limits_{X^{+}_{\mathbf{a}}}\left(\nu_0+\mathbf{u}^{\top}\mathbf{a}\right)G(dv) - \int\limits_{X^{-}_{\mathbf{a}}}\left(\nu_0+\mathbf{u}^{\top}\mathbf{a}\right)G(dv) + \lambda\left(\mathbf{1}^{\top}\mathbf{a}-1\right).$$

Using Lemma 1, the first-order conditions are:

$$\left.\frac{\partial L}{\partial \mathbf{a}}\right|_{\mathbf{a}=\mathbf{a}^*,\lambda=\lambda^*} =\boldsymbol{\omega}'(\mathbf{a}^*)+\lambda^*\mathbf{1}^{\top}=0 \tag{A.8}$$

$$\left.\frac{\partial L}{\partial \lambda}\right|_{\mathbf{a}=\mathbf{a}^*,\lambda=\lambda^*} =\mathbf{1}^{\top}\mathbf{a}-1=0. \tag{A.9}$$

Rearranging equation (A.8) gives the result. This completes the proof. ■

The following lemmas are useful for proving Proposition 1.

Lemma 2. *Let $g : X \to Y$ and $h : Y \to Z$, where h and g are twice differentiable convex functions with $X \subset \mathbb{R}^k$ and $Y, Z \subset \mathbb{R}$. Consider the following optimization problems:*

$$\begin{aligned}&\textit{minimize} && g(\mathbf{a})\\ &\textit{subject to} && \mathbf{1}^{\top}\mathbf{a}=1,\end{aligned} \tag{A.10}$$

and

$$\begin{aligned}&\textit{minimize} && h(g(\mathbf{a}))\\ &\textit{subject to} && \mathbf{1}^{\top}\mathbf{a}=1,\end{aligned} \tag{A.11}$$

where $\mathbf{a}^*_g$ and $\mathbf{a}^*_h$ are the solutions to problems (A.10) and (A.11), respectively, then

$$\mathbf{a}^*_g = \mathbf{a}^*_h$$

Moreover, let λ_g^* and λ_h^* be the associated Lagrange multipliers for Problems A.10 and A.11, respectively; then,

$$\lambda_h^* = h'\left[g(\mathbf{a}_g^*)\right]\lambda_g^*.$$

Proof of Lemma 2: The first-order necessary conditions for Problem A.10 are

$$\nabla g(\mathbf{a}_g^*) = \lambda_g^* \mathbf{1} \tag{A.12}$$

$$\mathbf{1}^\top \mathbf{a}_g^* = 1 \tag{A.13}$$

Note that equation (A.12) yields exactly $k-1$ unique equations, specifically, $\nabla_i g(\mathbf{a}_g^*) = \nabla_k g(\mathbf{a}_g^*)$ for all $i = 1, \ldots, k-1$. Along with equation (A.13), these yield exactly k equations to identify $\mathbf{a}_g^*$. Given $\mathbf{a}_g^*$, λ_g^* can be obtained by evaluating the gradient vector $\nabla g(\mathbf{a})$ at $\mathbf{a}_g^*$.

Now, consider the first order necessary conditions for Problem A.11:

$$h'(g^*)\nabla g(\mathbf{a}_h^*) = \lambda_h^* \mathbf{1} \tag{A.14}$$

$$\mathbf{1}^\top \mathbf{a}_h^* = 1 \tag{A.15}$$

where $g^* = g(\mathbf{a}_h^*) \neq 0$ and since h is a scalar function, equation (A.14) can be rewritten as

$$\nabla g(\mathbf{a}_h^*) = \frac{\lambda_h^*}{h'(g^*)}\mathbf{1}.$$

Along with equation (A.15), these equations yield the same system of simultaneous equations as the first-order conditions for Problem (A.10) for $\mathbf{a}_g^*$ and therefore $\mathbf{a}_h^* = \mathbf{a}_g^*$. The relation between the two Lagrange multipliers follows directly from the conditions above. This completes the proof. ■

Lemma 3. *Let $h(\mathbf{x}) : \mathbb{R}^K \to \mathbb{R}$ and $f(\mathbf{x}) : \mathbb{R}^K \to \mathbb{R}$ be C^1 functions and consider the following:*

$$\mathbf{a}_h^* = \arg\min_{\mathbf{a}} h(\mathbf{a}) + \lambda_h \left(1 - \mathbf{1}^\top \mathbf{a}\right) \tag{A.16}$$

$$\mathbf{a}_f^* = \arg\min_{\mathbf{a}} f(\mathbf{a}) + \lambda_f \left(1 - \mathbf{1}^\top \mathbf{a}\right) \tag{A.17}$$

where λ_h and λ_f are scalars with $\mathbf{a}_h^*$ and $\mathbf{a}_f^*$ being the unique solutions to equations (A.16) and (A.17), respectively. If there exists a $g(\mathbf{a}) : \mathbb{R}^K \to \mathbb{R}$ and a $\mathbf{p}(\mathbf{a}) : \mathbb{R}^K \to \mathbb{R}^K$ such that

$$\frac{\partial h}{\partial \mathbf{a}} = g(\mathbf{a})\frac{\partial f}{\partial \mathbf{a}} + \mathbf{p}(\mathbf{a}) \tag{A.18}$$

with $g(\mathbf{a}_h^*) = c \neq 0$ *and* $\mathbf{p}(\mathbf{a}_h^*) = c_1 \mathbf{1}_K$ *then* $\mathbf{a}_h^* = \mathbf{a}_f^*$.

Proof of Lemma 3: Note that

$$\left.\frac{\partial h}{\partial \mathbf{a}}\right|_{\mathbf{a}_h^*} = \lambda_h^* \mathbf{1}, \tag{A.19}$$

and under the condition of the theorem, specifically equation (A.18), it follows that

$$\begin{aligned} c\left.\frac{\partial f}{\partial \mathbf{a}}\right|_{\mathbf{a}_h^*} + c_1 \mathbf{1} =& \lambda_h^* \mathbf{1} \\ \left.\frac{\partial f}{\partial \mathbf{a}}\right|_{\mathbf{a}_h^*} =& c^{-1}\left(\lambda_h^* - c_1\right)\mathbf{1}. \end{aligned}$$

This means that $\hat{\mathbf{a}}_h$ satisfies the first-order conditions of the optimization problem (A.17) and under the condition of uniqueness, $\mathbf{a}_h^* = \mathbf{a}_f^*$ and $\lambda_h^* = c\lambda_f^* + c_1$. This completes the proof. ■

Lemma 4. *Let the functions* $h(\mathbf{a})$ *satisfy the conditions as stated in Lemma 3. Let* $\{h_T(\mathbf{a})\}$ *be a sequence of* C^1 *functions that converge to* $h(\mathbf{a})$ *for some point* $\mathbf{a}_0$ *on some closed interval* $[a, b]$ *with the property that* $\frac{\partial h_T}{\partial \mathbf{a}}$ *converges uniformly on* $\prod_{i=1}^{K}[a_i, b_i]$. *Define*

$$\hat{\mathbf{a}}_h^* = \arg\min_{\mathbf{a}} h_T(\mathbf{a}) + \lambda_h\left(1 - \mathbf{1}^\top \mathbf{a}\right) \tag{A.20}$$

then $\hat{\mathbf{a}}_h^* - \mathbf{a}_h^* = o_p(1)$.

Proof of Lemma 4: Note that

$$\frac{\partial h_T}{\partial \mathbf{a}} - \frac{\partial h}{\partial \mathbf{a}} = o_p(1)$$

by Theorem 7.17 in Rudin (1976) and hence $\hat{\mathbf{a}}_h - \mathbf{a}_h^* = o_p(1)$, and the result follows from Lemma 3. This completes the proof. ■

Proof of Proposition 1 : Define $g(\mathbf{a}) = T^{-1}\sum_{t=1}^{T}|\nu_t|$, $h(g) = Tg^2(\mathbf{a})$ and $f(\mathbf{a}) = T^{-1}\sum_{t=1}^{T}\nu_t^2$ with

$$\begin{aligned} \text{minimize} \quad & g(\mathbf{a}) \\ \text{subject to} \quad & \mathbf{1}^\top \mathbf{a} = 1, \end{aligned} \tag{A.21}$$

$$\begin{aligned} \text{minimize} \quad & h(\mathbf{a}) \\ \text{subject to} \quad & \mathbf{1}^\top \mathbf{a} = 1, \end{aligned} \tag{A.22}$$

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{a}) \\ \text{subject to} \quad & \mathbf{1}^\top \mathbf{a} = 1, \end{aligned} \tag{A.23}$$

where $\hat{\mathbf{a}}^*_g$, $\hat{\mathbf{a}}^*_h$ and $\hat{\mathbf{a}}^*_{\text{MSE}}$ are the optimal solutions to (A.21) – (A.23). Note that in problems (A.21) and (A.23), the MAE and MSE loss functions are minimized, respectively.

Under the assumptions of Lemma 2, $\hat{\mathbf{a}}^*_g = \hat{\mathbf{a}}^*_h$ for all $T > 0$. Now, rewrite $h(g)$ as

$$\begin{aligned} h(\mathbf{a}) =& T^{-1}\left[\sum_{t=1}^{T} \nu_t^2 + \sum_{t=1}^{T} \sum_{\tau=1,\tau\neq t}^{T} |\nu_t||\nu_\tau|\right] \\ \nabla h(\mathbf{a}) =& T^{-1}\left[2\sum_{t=1}^{T} \nu_t \mathbf{u}_t + \sum_{\tau=1}^{T}\sum_{\nu_t>0} \mathbf{u}_t|\nu_\tau| - \sum_{\tau=1}^{T}\sum_{\nu_t<0} \mathbf{u}_t|\nu_\tau| + \sum_{t=1}^{T}\sum_{\nu_\tau>0} \mathbf{u}_\tau|\nu_t| - \sum_{t=1}^{T}\sum_{\nu_\tau<0} \mathbf{u}_\tau|\nu_t|\right]. \end{aligned}$$

By the continuous mapping theorem and the law of large numbers, $\hat{\mathbf{a}}^*_g = \mathbf{a}^*_{\text{MAE}} + o_p(1)$ where $\mathbf{a}^*_{\text{MAE}}$ is the solution to the following optimization problem:

$$\begin{aligned} \underset{\mathbf{a}}{\text{minimize}} \quad & \mathbb{E}|\nu_t| \\ \text{subject to} \quad & \mathbf{1}^\top \mathbf{a} = 1. \end{aligned}$$

This implies that $\mathbf{a}^*_{\text{MAE}}$ must also satisfy the first-order condition as defined in equation (11).

As T becomes sufficiently large

$$\begin{aligned} \nabla h(\mathbf{a}) =& 2\mathbb{E}\mathbf{u}_t\nu_t + \mathbb{E}\left(\mathbf{u}_t|\nu_\tau||\nu_t > 0\right) - \mathbb{E}\left(\mathbf{u}_t|\nu_\tau||\nu_t < 0\right) \\ & + \mathbb{E}\left(\mathbf{u}_\tau|\nu_t||\nu_\tau > 0\right) - \mathbb{E}\left(\mathbf{u}_\tau|\nu_t||\nu_\tau < 0\right) + o_p(1) \\ =& 2\mathbb{E}\mathbf{u}_t\nu_t + \left[\mathbb{E}\left(\mathbf{u}_t|\nu_t > 0\right)\Pr(\nu_t > 0) - \mathbb{E}\left(\mathbf{u}_t|\nu_t < 0\right)\Pr(\nu_t < 0)\right]\mathbb{E}|\nu_\tau| \\ & + \left[\mathbb{E}\left(\mathbf{u}_\tau|\nu_\tau > 0\right)\Pr(\nu_\tau > 0) - \mathbb{E}\left(\mathbf{u}_\tau|\nu_\tau < 0\right)\Pr(\nu_\tau < 0)\right]\mathbb{E}|\nu_t| + o_p(1) \end{aligned}$$

The last line follows from the independent properties of $\mathbf{u}_t$ and ν_τ. Equation (11) implies that

$$\begin{aligned} \nabla h(\mathbf{a}^*_{\text{MAE}}) =& 2\mathbb{E}\mathbf{u}_t\nu_t + o_p(1) \\ =& \nabla f(\mathbf{a}^*_{\text{MAE}}), \end{aligned}$$

for sufficiently large T. The result then follows directly from Lemmas 3 and 4. This completes the proof. ■