Revisiting Forecast Combination Puzzle

An Empirical Study

XIEFEI (SAPPHIRE) LI

Department of Econometrics and Business Statistics
xlii0145@student.monash.edu

Supervisor: David T. Frazier

MONASH University

Forecast combination - point and density

Combining multiple forecasts can dramatically improve the accuracy of the forecast (Bates and Granger, 1969).

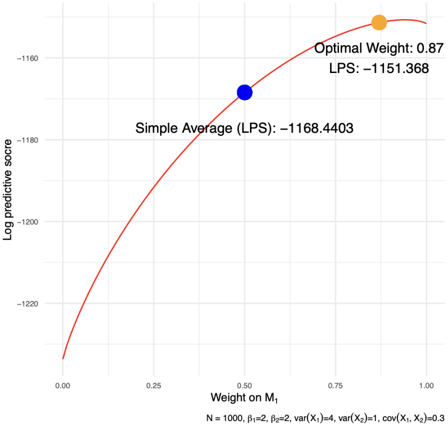
Point Forecast Combination

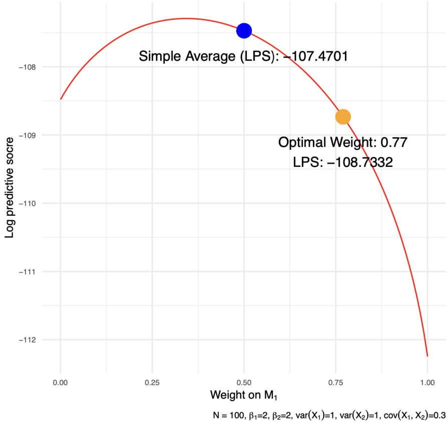
$$\hat{y}_t(\omega) = \omega\ \hat{y}_{1t} + (1-\omega)\ \hat{y}_{2t}$$

Density Forecast Combination

$$\hat{f}_\omega(y_t) = \omega\ \hat{f}_1(y_t) + (1-\omega)\hat{f}_2(y_t)$$

Forecast combination puzzle





- Complicated Weighting Schemes
- Simple Averaging / Equal Weights

When should we expect the puzzle

In the linear regression context, the optimal weight $\hat{\omega}_{opt}$ has a closed-form expression when using the Mean Squared Error (MSE) weighting scheme.

$$\omega_\star = \frac{1}{2} \Rightarrow \alpha_1' \Sigma_{11} \alpha_1 = \alpha_2' \Sigma_{22} \alpha_2$$

- Coefficients α_1, α_2
- Variances of regressors Σ_{11}, Σ_{22}
- In-sample performance

Preliminary Conjecture

Consider a two-model combination.

		M_2	
		Good	Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$

Table 1: Initial conjecture on the presence of forecast combination puzzle

The in-sample fit of two models may indicate the presence of the puzzle.

The puzzle will be in evidence when both models are good or both are bad.

Preliminary Conjecture

Consider a two-model combination.

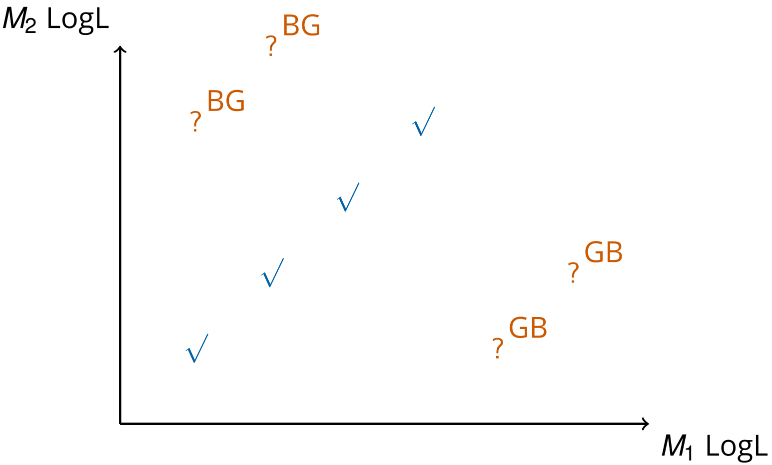
		M_2	
		Good	Bad
M_1	Good	√	?
	Bad	?	√

Table 1: Initial conjecture on the presence of forecast combination puzzle

The in-sample fit of two models may indicate the presence of the puzzle.

The puzzle will be in evidence when both models are good or both are bad.

Preliminary Conjecture



Explanations of the puzzle in literature

Uncertainty in Weight Estimation

The simple averaging does not require any estimation (Stock and Watson, 1998, Stock and Watson, 2004, and Smith and Wallis, 2009).

Trade-off between Bias and Variance

The equally weighted combination is unbiased and its variance has only one component (Elliott, 2011 and Claeskens et al., 2016).

A Recent (General) Explanation

Estimation Uncertainty on Forecast Performance

Asymptotically, the bias and sampling variability mainly come from the estimation of the models used to produce the constituent model forecasts (Zischke et al., 2022 and Frazier et al., 2023).

These explanations all implicitly assume that **the puzzle will be in evidence** when combining forecasts, regardless of the choice of constituent models or the weighing scheme.

Research Gap

Forecast combination has attracted wide attention and contributions in the literature, both theoretical and applied (Clemen, 1989 and Timmermann, 2006).

Researchers have examined a variety of combination methods for both point and density forecasts over the past 50 years, see Wang et al., 2022 for a modern literature review.

No attention appears to have been given to **the cross-sectional setting**.

Research objectives

1. To **substantiate the presence** of the combination puzzle in the time series setting with empirical data.

Research objectives

1. To **substantiate the presence** of the combination puzzle in the time series setting with empirical data.
2. To systematically investigate the determinants behind, and evidence for, the **forecast combination puzzle in the cross-sectional setting** using simulated data.

Research objectives

1. To **substantiate the presence** of the combination puzzle in the time series setting with empirical data.
2. To systematically investigate the determinants behind, and evidence for, the **forecast combination puzzle in the cross-sectional setting** using simulated data.
3. To **validate our preliminary conjecture** with empirical evidence.

Forecast Combination

We focus on the combination of forecasts from non-nested models for a given dataset, which is commonly performed in two stages:

1. **producing** separate point or probabilistic **forecasts** for the next time point using observed data and constituent models, and

Forecast Combination

We focus on the combination of forecasts from non-nested models for a given dataset, which is commonly performed in two stages:

1. **producing** separate point or probabilistic **forecasts** for the next time point using observed data and constituent models, and
2. **combining forecasts** based on a given accuracy criteria.

First Stage - Parameters Estimation

The unknown parameters of each model, θ, are estimated by maximizing the log likelihood function over the in-sample period (R):

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max} \sum_{t=1}^{R} log\ f(y_t | \mathcal{F}_{t-1}, \theta) \tag{1}$$

$\mathcal{F}_{t-1}$ = all information available at time $(t-1)$.

- Point forecasts
- Density forecasts

Point: Linear Combination

Two constituent points y_{1t} and y_{2t} are aggregated linearly:

$$y_t(\omega) = \omega \, y_{1t} + (1 - \omega) \, y_{2t} \tag{2}$$

where $\omega \in [0, 1]$ is the non-negative weight allocated to the point expressed based on the first model.

Density: Linear Pools

A linear combination of two densities, $f(y_t)$, is constructed with two constituent densities $f_1(y_t)$ and $f_2(y_t)$:

$$f_\omega(y_t) = \omega\, f_1(y_t) + (1-\omega)\, f_2(y_t) \tag{3}$$

where $\omega \in [0,1]$ is the weight allocated to the first density.

The sum of two weights is implied to be 1, which is necessary and sufficient for the combination to be a density function [Geweke and Amisano, 2011].

Accuracy Criteria

The optimal weight assigned to the first point/density is estimated by satisfying one of the accuracy criteria over the in-sample period (R).

- Mean Squared Error (Smith and Wallis, 2009)

$$MSE = \frac{1}{R}\sum_{t=1}^{R}(y_t - \hat{y}_t)^2 \tag{4}$$

Accuracy Criteria

The optimal weight assigned to the first point/density is estimated by satisfying one of the accuracy criteria over the in-sample period (R).

- Mean Squared Error (Smith and Wallis, 2009)

$$MSE = \frac{1}{R}\sum_{t=1}^{R}(y_t - \hat{y}_t)^2 \tag{4}$$

- Log Score Function (Geweke and Amisano, 2011)

$$LS = \sum_{t=1}^{R} log\ \hat{f}(y_t|\mathcal{F}_{t-1}, \hat{\theta}) \tag{5}$$

Second Stage - Weight Estimation

- Mean Squared Error

$$\hat{\omega}_{\text{opt}} = \underset{\omega \in [0,1]}{\arg\min} \frac{1}{R} \sum_{t=1}^{R} \left\{ y_t - [\omega\, \hat{y}_{1t} + (1-\omega)\, \hat{y}_{2t}] \right\}^2 \tag{6}$$

Second Stage - Weight Estimation

- Mean Squared Error

$$\hat{\omega}_{\text{opt}} = \underset{\omega \in [0,1]}{\arg\min} \frac{1}{R} \sum_{t=1}^{R} \left\{ y_t - \left[\omega \, \hat{y}_{1t} + (1-\omega) \, \hat{y}_{2t} \right] \right\}^2 \tag{6}$$

- Log Score Function

$$\hat{\omega}_{\text{opt}} = \underset{\omega \in [0,1]}{\arg\max} \sum_{t=1}^{R} log \left[\omega \, \hat{f}_1(y_t) + (1-\omega) \, \hat{f}_2(y_t) \right] \tag{7}$$

$$\hat{f}_1(y_t) \equiv \hat{f}_1(y_t | \mathcal{F}_{t-1}, \hat{\theta}_1)$$
$$\hat{f}_2(y_t) \equiv \hat{f}_2(y_t | \mathcal{F}_{t-1}, \hat{\theta}_2)$$

Point Forecast Combination

The mean squared forecast error (MSFE) over the out-of-sample period (P) is:

$$MSFE = \frac{1}{P} \sum_{t=R+1}^{T} \left\{ y_t - \left[\hat{\omega}_{\text{opt}} \, \hat{y}_{1t} + (1 - \hat{\omega}_{\text{opt}}) \, \hat{y}_{2t}\right] \right\}^2 . \tag{8}$$

Density Forecast Combination

The log predictive score (LPS) over the out-of-sample period (P) is:

$$LPS = \sum_{t=R+1}^{T} log\Big[\hat{\omega}_{\text{opt}}\ \hat{f}_1(y_t) + (1 - \hat{\omega}_{\text{opt}})\ \hat{f}_2(y_t)\Big]. \tag{9}$$

$$\hat{f}_1(y_t) \equiv \hat{f}_1(y_t|\mathcal{F}_{t-1}, \hat{\theta}_1)$$
$$\hat{f}_2(y_t) \equiv \hat{f}_2(y_t|\mathcal{F}_{t-1}, \hat{\theta}_2)$$

Example - Standard and Poor's (S&P) 500 index

Daily S&P 500 index from Federal Reserve Economic Data (FRED, 2023)

- February 11, 2013 - February 10, 2023
- $T = 2519$
- $R = 1511$ (60%)
- $P = 1008$

We focus on three common classes of linear time series models:

- Autoregressive integrated moving average (ARIMA)
- Exponential smoothing (ETS)
- Linear regression model (LR) with ARIMA errors

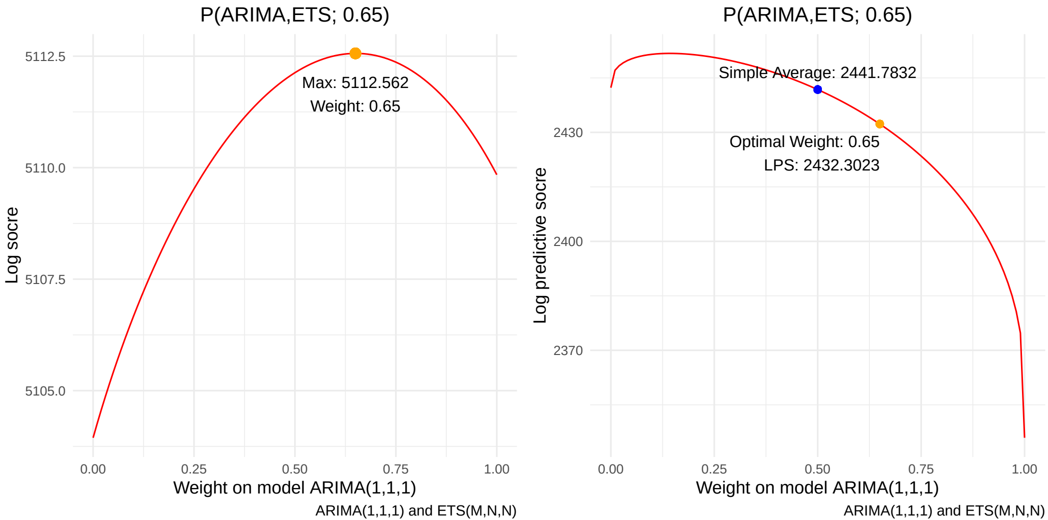


Figure 1: Log score of S&P 500 index predictive densities in P(ARIMA, ETS; 0.65).

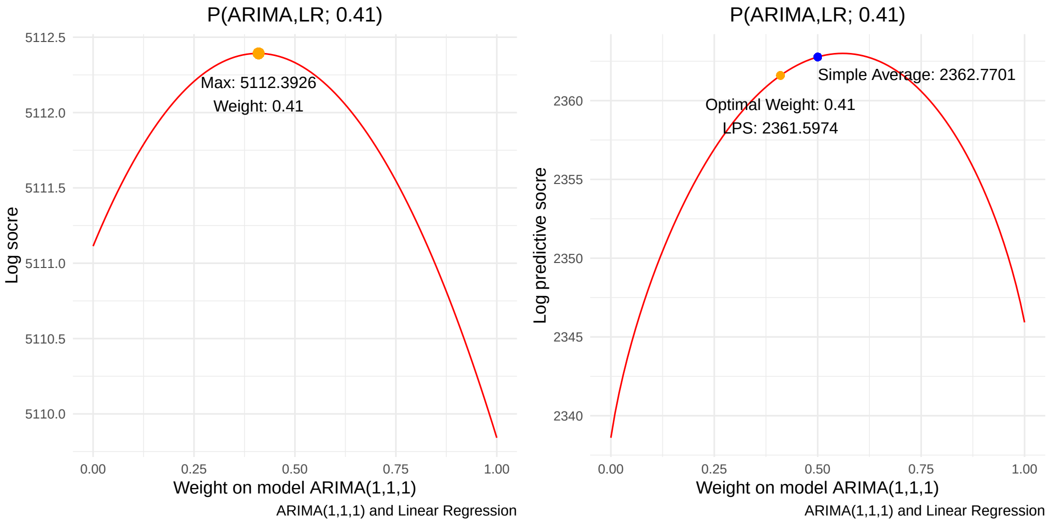


Figure 2: Log score of S&P 500 index predictive densities in P(ARIMA, LR; 0.41).

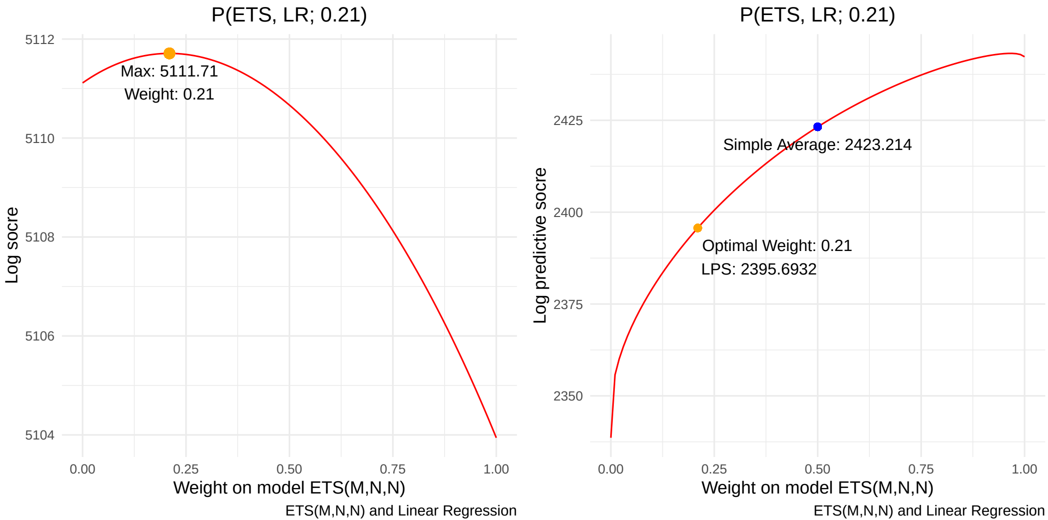


Figure 3: Log score of S&P 500 index predictive densities in P(ETS, LR; 0.21).

In-sample Fit Comparison

	P(ARIMA,ETS; 0.65)
1^{st} Model LogL	5113.694
2^{nd} Model LogL	1725.137
Difference	3388.556
Puzzle	Yes

	P(ARIMA,ETS)
Type	(G,B)
Puzzle	Yes

		M_2	
		Good	Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$

In-sample Fit Comparison

	P(ARIMA,ETS; 0.65)	P(ARIMA,LR; 0.41)
1^{st} Model LogL	5113.694	5113.694
2^{nd} Model LogL	1725.137	5116.014
Difference	3388.556	2.320
Puzzle	Yes	Yes

	P(ARIMA,ETS)	P(ARIMA,LR)
Type	(G,B)	(G,G)
Puzzle	Yes	Yes

		M_2 Good	M_2 Bad
M_1	Good	√	?
	Bad	?	√

In-sample Fit Comparison

	P(ARIMA,ETS; 0.65)	P(ARIMA,LR; 0.41)	P(ETS,LR; 0.21)
1^{st} Model LogL	5113.694	5113.694	1725.137
2^{nd} Model LogL	1725.137	5116.014	5116.014
Difference	3388.556	2.320	3390.876
Puzzle	Yes	Yes	Yes

	P(ARIMA,ETS)	P(ARIMA,LR)	P(ETS,LR)
Type	(G,B)	(G,G)	(B,G)
Puzzle	Yes	Yes	Yes

		M_2 Good	M_2 Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$

Conjecture Revision

Under a mild definition of the forecast combination puzzle, the empirical evidence suggest that the puzzle is in evidence in all the cases.

$$
\begin{array}{llcc}
 & & \multicolumn{2}{c}{M_2} \\
 & & \text{Good} & \text{Bad} \\
M_1 & \text{Good} & \surd & \surd \\
 & \text{Bad} & \surd & \surd
\end{array}
$$

However, there are too few examples to draw conclusions.

We may encounter situations where the optimal forecast combination is more accurate than the simple averaging.

Model Setup - True DGP

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$$

- $N = 10000$
- $E[X_{1i}] = E[X_{2i}] = 0$, $Var(X_{1i}) = Var(X_{2i}) = 1$
- $Cov(X_{1i}, X_{2i}) = 0.3$ exogenous and weakly correlated regressors
- $\beta = (\beta_1, \beta_2)' = (2, 2)'$, $\sigma_\epsilon^2 = 4$
- All classical assumptions

- Obtain y, x_{1i} and x_{2i}

Model Setup - Forecasting Models

The constituent models in matrix form are proposed as

$$M_1 : \boldsymbol{y} = \boldsymbol{x}_1\alpha_1 + \boldsymbol{u}_1, \quad \boldsymbol{u}_1 \overset{i.i.d}{\sim} N(\boldsymbol{0}, \sigma_1^2)$$
$$M_2 : \boldsymbol{y} = \boldsymbol{x}_2\alpha_2 + \boldsymbol{u}_2, \quad \boldsymbol{u}_2 \overset{i.i.d}{\sim} N(\boldsymbol{0}, \sigma_2^2).$$

- $\boldsymbol{y} \in \mathbb{R}^{\mathrm{R}}$ is a $(\mathrm{R} \times 1)$ vector, same for $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{u}_1$ and $\boldsymbol{u}_2$
- Obtain $\hat{\boldsymbol{y}}_1$ from M_1 and $\hat{\boldsymbol{y}}_2$ from M_2
- Aggregate them linearly $\hat{\boldsymbol{y}}(\omega) = \hat{\boldsymbol{y}}_1\omega + \hat{\boldsymbol{y}}_2(1-\omega)$

Optimal Weight Derivation - Minimization

The optimal weight is obtained over the in-sample period (R).

$$\hat{\omega}_{\text{opt}} = \underset{\omega \in [0,1]}{\arg\min} \frac{1}{R} \left[\boldsymbol{y} - (\hat{\boldsymbol{y}}_1 \omega + \hat{\boldsymbol{y}}_2 (1 - \omega)) \right]' \left[\boldsymbol{y} - (\hat{\boldsymbol{y}}_1 \omega + \hat{\boldsymbol{y}}_2 (1 - \omega)) \right]$$

Optimal Weight Derivation - Minimization

The optimal weight is obtained over the in-sample period (R).

$$\hat{\omega}_{\text{opt}} = \underset{\omega\in[0,1]}{\arg\min}\ \frac{1}{R}\left[\boldsymbol{y} - \left(\hat{\boldsymbol{y}}_1\omega + \hat{\boldsymbol{y}}_2(1-\omega)\right)\right]'\left[\boldsymbol{y} - \left(\hat{\boldsymbol{y}}_1\omega + \hat{\boldsymbol{y}}_2(1-\omega)\right)\right]$$

The first-order condition needs to be satisfied.

$$-\frac{2}{R}(\boldsymbol{x}_1\hat{\alpha}_1 - \boldsymbol{x}_2\hat{\alpha}_2)'(\boldsymbol{y} - (\boldsymbol{x}_1\hat{\alpha}_1 - \boldsymbol{x}_2\hat{\alpha}_2)\hat{\omega}_{opt} - \boldsymbol{x}_2\hat{\alpha}_2) = 0$$

Optimal Weight Derivation - Findings

$$\hat{\omega}_{opt} = \frac{(\boldsymbol{x}_1\hat{\alpha}_1 - \boldsymbol{x}_2\hat{\alpha}_2)'y - (\boldsymbol{x}_1\hat{\alpha}_1 - \boldsymbol{x}_2\hat{\alpha}_2)'x_2\hat{\alpha}_2}{\hat{\alpha}_1'\boldsymbol{x}_1'\boldsymbol{x}_1\hat{\alpha}_1 - 2\hat{\alpha}_1'\boldsymbol{x}_1'\boldsymbol{x}_2\hat{\alpha}_2 + \hat{\alpha}_2'\boldsymbol{x}_2'\boldsymbol{x}_2\hat{\alpha}_2}$$
$$= \frac{\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_1)\hat{\alpha}_1 - \hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_2)\hat{\alpha}_2}{\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_1)\hat{\alpha}_1 - 2\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_2)\hat{\alpha}_2 + \hat{\alpha}_2'\mathrm{cov}_R(\boldsymbol{x}_2, \boldsymbol{x}_2)\hat{\alpha}_2}$$

The estimated optimal weight $\hat{\omega}_{opt}$ is affected by

1. the magnitude and sign of estimated coefficients in constituent models

Optimal Weight Derivation - Findings

$$\hat{\omega}_{opt} = \frac{(\boldsymbol{x}_1\hat{\alpha}_1 - \boldsymbol{x}_2\hat{\alpha}_2)'y - (\boldsymbol{x}_1\hat{\alpha}_1 - \boldsymbol{x}_2\hat{\alpha}_2)'x_2\hat{\alpha}_2}{\hat{\alpha}_1'\boldsymbol{x}_1'\boldsymbol{x}_1\hat{\alpha}_1 - 2\hat{\alpha}_1'\boldsymbol{x}_1'\boldsymbol{x}_2\hat{\alpha}_2 + \hat{\alpha}_2'\boldsymbol{x}_2'\boldsymbol{x}_2\hat{\alpha}_2}$$
$$= \frac{\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_1)\hat{\alpha}_1 - \hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_2)\hat{\alpha}_2}{\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_1)\hat{\alpha}_1 - 2\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_2)\hat{\alpha}_2 + \hat{\alpha}_2'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_1)\hat{\alpha}_2}$$

The estimated optimal weight $\hat{\omega}_{opt}$ is affected by

1. the magnitude and sign of estimated coefficients in constituent models
2. sample variances of regressors

Optimal Weight Derivation - Findings

$$\hat{\omega}_{opt} = \frac{(\boldsymbol{x}_1\hat{\alpha}_1 - \boldsymbol{x}_2\hat{\alpha}_2)'y - (\boldsymbol{x}_1\hat{\alpha}_1 - \boldsymbol{x}_2\hat{\alpha}_2)'x_2\hat{\alpha}_2}{\hat{\alpha}_1'\boldsymbol{x}_1'\boldsymbol{x}_1\hat{\alpha}_1 - 2\hat{\alpha}_1'\boldsymbol{x}_1'\boldsymbol{x}_2\hat{\alpha}_2 + \hat{\alpha}_2'\boldsymbol{x}_2'\boldsymbol{x}_2\hat{\alpha}_2}$$

$$= \frac{\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_1)\hat{\alpha}_1 - \hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_2)\hat{\alpha}_2}{\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_1)\hat{\alpha}_1 - 2\hat{\alpha}_1'\mathrm{cov}_R(\boldsymbol{x}_1, \boldsymbol{x}_2)\hat{\alpha}_2 + \hat{\alpha}_2'\mathrm{cov}_R(\boldsymbol{x}_2, \boldsymbol{x}_2)\hat{\alpha}_2}$$

The estimated optimal weight $\hat{\omega}_{opt}$ is affected by

1. the magnitude and sign of estimated coefficients in constituent models
2. sample variances of regressors
3. sample covariance of regressors

Optimal Weight Derivation - Limiting Result

$$\hat{\omega}_{opt} \xrightarrow{p} \omega_{\star} = \frac{\alpha_1' \Sigma_{11} \alpha_1 - \alpha_1' \Sigma_{12} \alpha_2}{\alpha_1' \Sigma_{11} \alpha_1 - 2\alpha_1' \Sigma_{12} \alpha_2 + \alpha_2' \Sigma_{22} \alpha_2}$$

Σ_{mn} is the population covariance between regressors x_m and x_n.

For $\omega_{\star} = \frac{1}{2}$, it must be that

$$\alpha_1' \Sigma_{11} \alpha_1 = \alpha_2' \Sigma_{22} \alpha_2.$$

When this final equality is nearly satisfied will inevitably lead the optimal weight to be around a half.

The relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ is derived in Appendix 13.

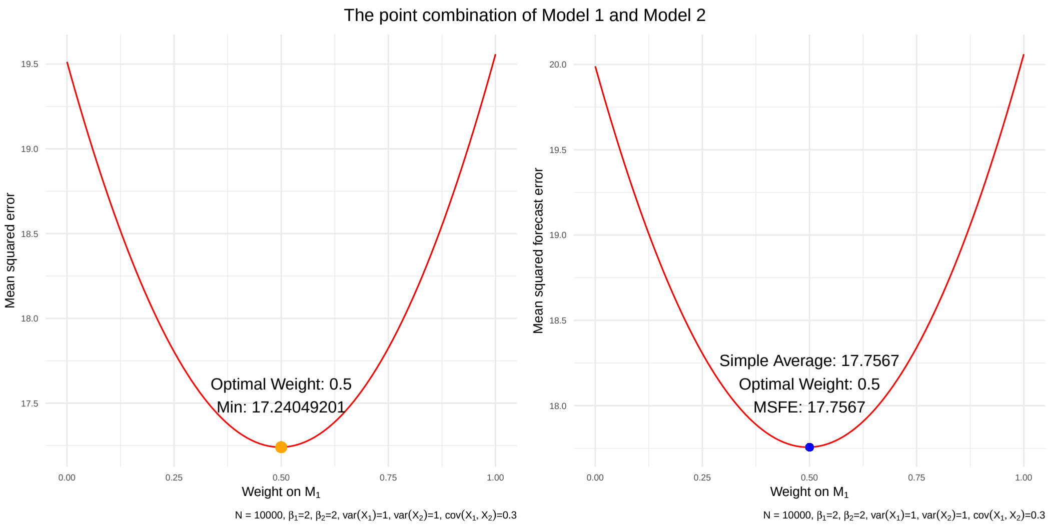


Figure 4: $\hat{\omega}_{opt}$ = 0.5, N = 10000, β_1=2, β_2=2, $var(X_1)$=1, $var(X_2)$=1, $cov(X_1, X_2)$=0.3

Density Simulations

Applying the learning to density combinations

- Log scoring rules
- No closed-form expression for $\hat{\omega}_{opt}$
- Applicability of findings

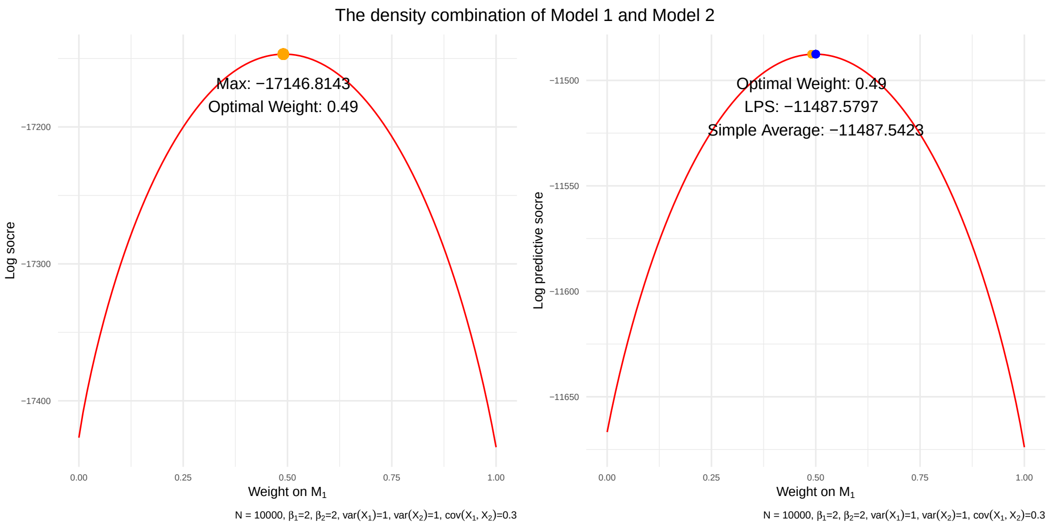


Figure 5: $\hat{\omega}_{opt}$ = 0.49, N = 10000, β_1=2, β_2=2, $var(X_1)$=1, $var(X_2)$=1, $cov(X_1, X_2)$=0.3

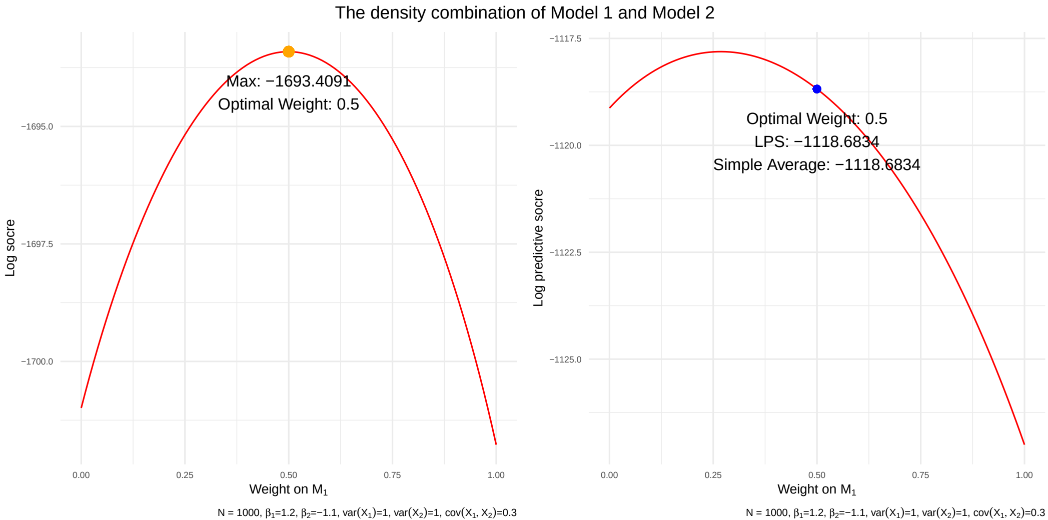


Figure 6: $\hat{\omega}_{opt}$ = 0.5, N = 1000, β_1=1.2, β_2=-1.1, $var(X_1)$=1, $var(X_2)$=1, $cov(X_1, X_2)$=0.3

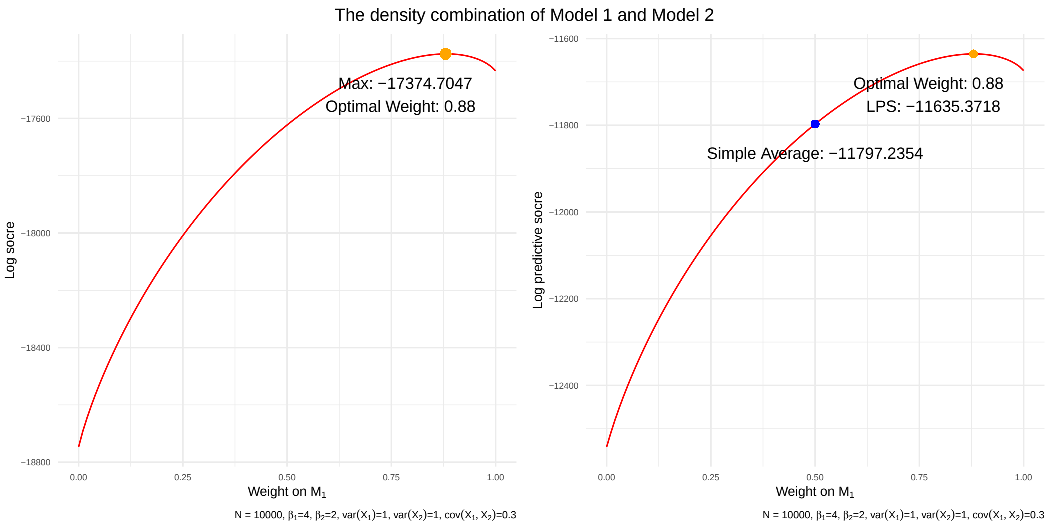


Figure 7: $\hat{\omega}_{opt}$ = 0.88, N = 10000, β_1=4, β_2=2, $var(X_1)$=1, $var(X_2)$=1, $cov(X_1, X_2)$=0.3

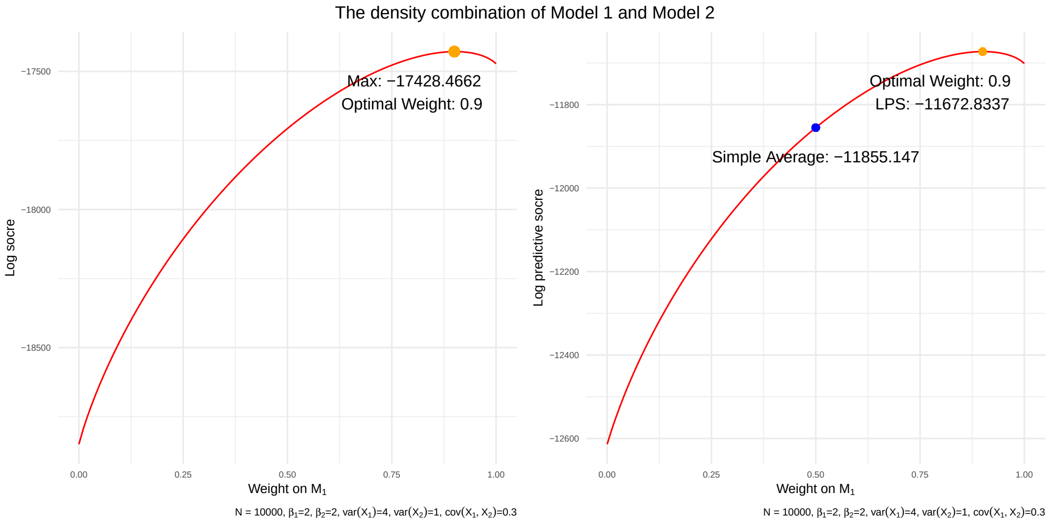


Figure 8: $\hat{\omega}_{opt}$ = 0.9, N = 10000, β_1=2, β_2=2, $var(X_1)$=4, $var(X_2)$=1, $cov(X_1, X_2)$=0.3

Density Simulation Findings

Similar to the MSE scheme, the estimated ω_{opt} is affected by

- the magnitude and sign of β,
- sample variances of regressors, and
- the sample size.

Surprisingly, under the log scoring rule, $\hat{\omega}_{opt}$ has a non-linear relationship with the proposed models.

Conjecture Revision

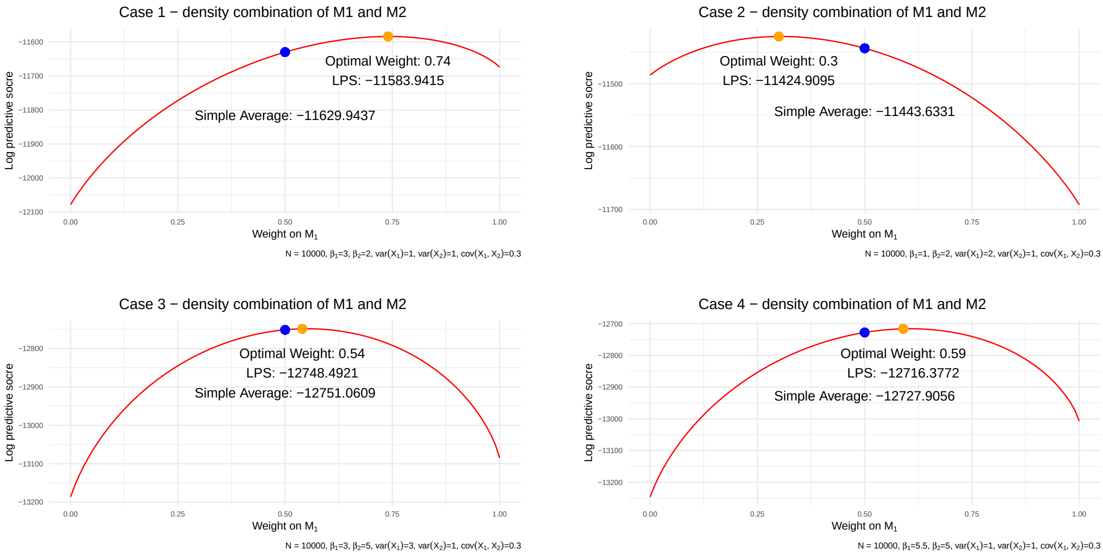
We can now further validate our preliminary conjecture on the presence of forecast combination puzzle.

		M_2 Good	M_2 Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$

Table 2: Initial conjecture

		M_2 Good	M_2 Bad
M_1	Good	$\surd$	$\surd$
	Bad	$\surd$	$\surd$

Table 3: Updated conjecture



Density forecast combination accuracy evaluation for the optimal combination and the simple averaging in different cases.

In-sample Fit Comparison

	Case 1
R^2 of M_1	0.393
R^2 of M_2	0.256
Difference	0.138
Type	(G,B)
Puzzle	No

		M_2	
		Good	Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$

In-sample Fit Comparison

	Case 1	Case 2
R^2 of M_1	0.393	0.141
R^2 of M_2	0.256	0.224
Difference	0.138	0.083
Type	(G,B)	(B,G)
Puzzle	No	No

		M_2 Good	M_2 Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$

In-sample Fit Comparison

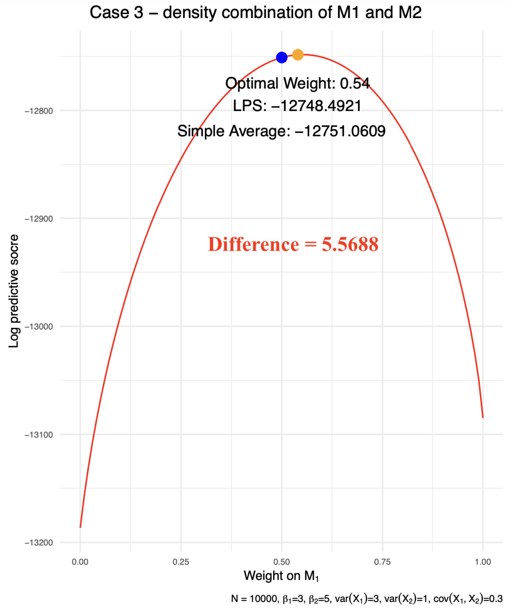
	Case 1	Case 2	Case 3
R^2 of M_1	0.393	0.141	0.476
R^2 of M_2	0.256	0.224	0.452
Difference	0.138	0.083	0.024
Type	(G,B)	(B,G)	(B,B)
Puzzle	No	No	Yes

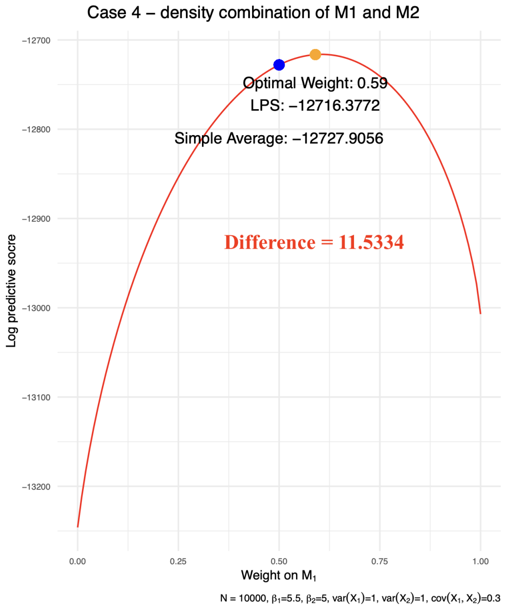
		M_2 Good	M_2 Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$

In-sample Fit Comparison

	Case 1	Case 2	Case 3	Case 4
R^2 of M_1	0.393	0.141	0.476	0.558
R^2 of M_2	0.256	0.224	0.452	0.504
Difference	0.138	0.083	0.024	0.053
Type	(G,B)	(B,G)	(B,B)	(G,B)
Puzzle	No	No	Yes	Yes

		M_2 Good	M_2 Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$


Case 3 – density combination of M1 and M2
Optimal Weight: 0.54
LPS: −12748.4921
Simple Average: −12751.0609
Difference = 5.5688
Log predictive socre
−12800
−12900
−13000
−13100
−13200
0.00
0.25
0.50
0.75
1.00
Weight on M_1
N = 10000, β_1=3, β_2=5, var(X_1)=3, var(X_2)=1, cov(X_1, X_2)=0.3


Case 4 – density combination of M1 and M2
Optimal Weight: 0.59
LPS: −12716.3772
Simple Average: −12727.9056
Difference = 11.5334
Log predictive socre
−12700
−12800
−12900
−13000
−13100
−13200
0.00
0.25
0.50
0.75
1.00
Weight on M_1
N = 10000, β_1=5.5, β_2=5, var(X_1)=1, var(X_2)=1, cov(X_1, X_2)=0.3

Accuracy Difference

- Formal testing?

  Diebold-Mariano Test (Diebold, 2015) is not appropriate.

Accuracy Difference

- Formal testing?

  Diebold-Mariano Test (Diebold, 2015) is not appropriate.

- An arbitrary choice?

  The magnitude of the log predictive score is closely related to the sample size and the true (unknown) value of the variance for the error in the actual DGP. Case-by-case basis.

Accuracy Difference

- Formal testing?

  Diebold-Mariano Test (Diebold, 2015) is not appropriate.

- An arbitrary choice?

  The magnitude of the log predictive score is closely related to the sample size and the true (unknown) value of the variance for the error in the actual DGP. Case-by-case basis.

Heuristic: If the absolute difference of the in-sample R^2 between two constituent models is less than 0.05, then we are in (G,G) or (B,B) cases.

Finalised Conjecture

		M_2	
		Good	Bad
M_1	Good	$\surd$	?
	Bad	?	$\surd$

Under a mild definition of the forecast combination puzzle,

- the puzzle is in evidence when constituent models fit the in-sample data both good or both bad, whereas
- the presence of the puzzle is uncertain when in-sample performance of two constituent models *differs a lot*.

Conclusion

1. Forecast combinations can deliver **improved accuracy** over single models.

Conclusion

1. Forecast combinations can deliver **improved accuracy** over single models.
2. The puzzle can be found in both pure **time series and cross-sectional** settings.

Conclusion

1. Forecast combinations can deliver **improved accuracy** over single models.
2. The puzzle can be found in both pure **time series and cross-sectional** settings.
3. The presence of the forecast combination puzzle is highly correlated with the **in-sample performance** of constituent models.

Conclusion

1. Forecast combinations can deliver **improved accuracy** over single models.
2. The puzzle can be found in both pure **time series and cross-sectional** settings.
3. The presence of the forecast combination puzzle is highly correlated with the **in-sample performance** of constituent models.
4. The optimal weight interacts with the **true data generating process** and is therefore related to the true DGP.

Limitations

1. Only two constituent models are considered

Limitations

1. Only two constituent models are considered
2. Restricted model assumptions

Limitations

1. Only two constituent models are considered
2. Restricted model assumptions
3. Simple model structures

Limitations

1. Only two constituent models are considered
2. Restricted model assumptions
3. Simple model structures
4. A mild definition of the forecast combination puzzle

Limitations

1. Only two constituent models are considered
2. Restricted model assumptions
3. Simple model structures
4. A mild definition of the forecast combination puzzle
5. Hard to determine the significance of the accuracy difference between optimal combination and simple averaging

References I

ABS. (2023). *Labour force, australia, detailed*. Retrieved March 28, 2023, from https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia-detailed/latest-release

Bates, J. M., & Granger, C. W. (1969).The combination of forecasts. *Journal of the operational research society*, *20*(4), 451–468. https://doi.org/https://doi.org/10.1057/jors.1969.103

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016).The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762.

Clemen, R. T. (1989).Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, *5*(4), 559–583.

Diebold, F. X. (2015).Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, *33*(1), 1–1.

References II

- Elliott, G. (2011).Averaging and the optimal combination of forecasts. *University of California, San Diego*.
- Frazier, D. T., Covey, R., Martin, G. M., & Poskitt, D. (2023).Solving the forecast combination puzzle. *arXiv preprint arXiv:2308.05263*.
- FRED. (2023). *S&p500*. Retrieved February 12, 2023, from https://fred.stlouisfed.org/series/SP500#0
- Geweke, J., & Amisano, G. (2011).Optimal prediction pools. *Journal of Econometrics*, *164*(1), 130–141. https://doi.org/https://doi.org/10.1016/j.jeconom.2011.02.017
- Smith, J., & Wallis, K. F. (2009).A simple explanation of the forecast combination puzzle. *Oxford bulletin of economics and statistics*, *71*(3), 331–355.
- Stock, J. H., & Watson, M. W. (1998). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series.

References III

- Stock, J. H., & Watson, M. W. (2004).Combination forecasts of output growth in a seven-country data set. *Journal of forecasting*, *23*(6), 405–430. https://doi.org/https://doi.org/10.1002/for.928
- Timmermann, A. (2006).Forecast combinations. *Handbook of economic forecasting*, *1*, 135–196.
- Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2022).Forecast combinations: An over 50-year review. *International Journal of Forecasting*. https://doi.org/https://doi.org/10.48550/arXiv.2205.04216
- Zischke, R., Martin, G. M., Frazier, D. T., & Poskitt, D. S. (2022).The impact of sampling variability on estimated combinations of distributional forecasts. *arXiv preprint arXiv:2206.02376*. https://doi.org/https://doi.org/10.48550/arXiv.2206.02376

Thank You!

Questions?

Example 1 (Nonstationary) - Model Specification I

- ARIMA(1,1,1) model

$$log(y_t) = c + log(y_{t-1}) + \phi_1 \left[log(y_{t-1}) - log(y_{t-2})\right] + \epsilon_t + \theta_1 \epsilon_{t-1}$$

- ETS(M,N,N) model

$$\begin{aligned} y_t &= \ell_{t-1}(1 + \epsilon_t) \\ \ell_t &= \ell_{t-1}(1 + \alpha\epsilon_t) \end{aligned}$$

Example 1 (Nonstationary) - Model Specification II

- A linear regression model of the natural logarithm of the S&P 500 index and ARIMA(1,0,0) errors.

$$log(y_t) = \beta_0 + \beta_1 t + u_t$$
$$u_t = \phi_1 u_{t-1} + \epsilon_t$$

The ϵ_t in each model is assumed to be independent and normally distributed with a zero mean and a constant variance.

Example 1 (Stationary) - Model Specification

- ARMA(1,1) model with an intercept of the natural logarithm of S&P 500 returns.

$$log(y_t) - log(y_{t-1}) = c + \phi_1 \left[log(y_{t-1}) - log(y_{t-2}) \right] + \epsilon_t + \theta_1 \epsilon_{t-1}$$

- A classical linear regression model of the natural logarithm of the S&P 500 returns and ARMA(1,1) errors.

$$\begin{aligned} log(y_t) &= \beta_0 + u_t \\ u_t &= \phi_1 u_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1} \end{aligned}$$

Example 1 - S&P 500 log returns

Same dataset but modelling the S&P 500 log returns

	P(ARMA,LR;0.69)
1^{st} model Log Likelihood	5109.8071
2^{nd} model Log Likelihood	5109.8054
Puzzle	Yes

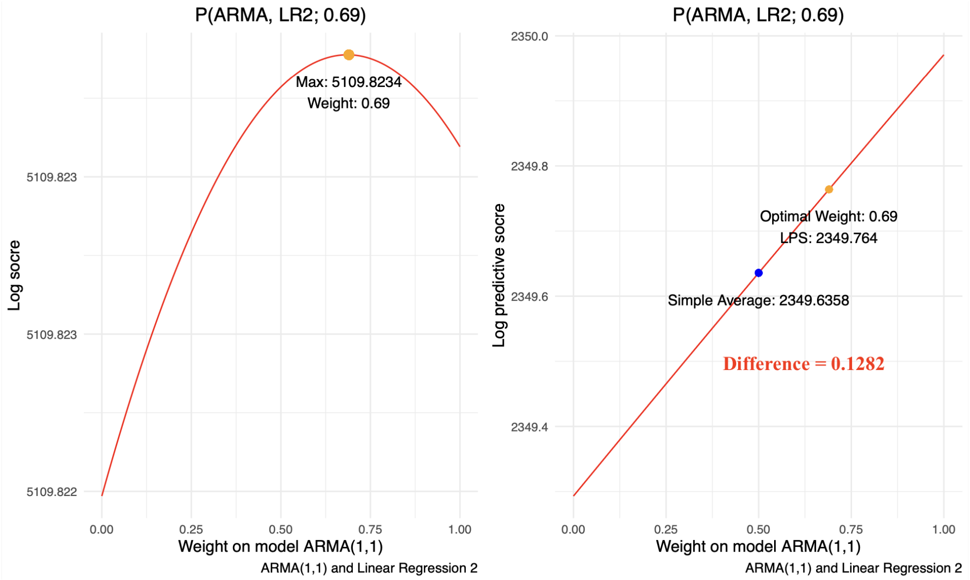


Figure 9: Log score of S&P 500 log returns predictive densities in two-model pools.

Example 2 - Seasonal Number of Unemployment

Quarterly total number of unemployed individuals (in thousands) retrieved from the Australia Bureau of Statistics (ABS, 2023)

- 1985 Q1 - 2023 Q1
- $T = 2519$
- $R = 1511$ (60%)
- $P = 1008$

We now consider the following linear time series models:

- Seasonal autoregressive integrated moving average (SARIMA)
- Exponential smoothing (ETS)

Example 2 - Well-specified Models

- ARIMA(2,0,2)(0,1,1)[4] model with an intercept of the natural logarithm of unemployed individuals.

$$
\begin{aligned}
log(y_t) = c + log(y_{t-4}) + \phi_1 \left[log(y_{t-1}) - log(y_{t-5})\right] \\
+ \phi_2 \left[log(y_{t-2}) - log(y_{t-6})\right] + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} \\
+ \Theta_1 \epsilon_{t-4} + \theta_1 \Theta_1 \epsilon_{t-5} + \theta_2 \Theta_1 \epsilon_{t-6}
\end{aligned}
$$

- ETS(A,A,A) model of the natural logarithm of unemployed individuals.

$$
\begin{aligned}
log(y_t) &= \ell_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t \\
\ell_t &= \ell_{t-1} + b_{t-1} + \alpha \epsilon_t \\
b_t &= b_{t-1} + \beta \epsilon_t \\
s_t &= s_{t-m} + \gamma \epsilon_t
\end{aligned}
$$

Example 2 - Poorly-specified Models

- ARIMA(2,1,0) model with an intercept of the natural logarithm of unemployed individuals.

$$log(y_t) = c + log(y_{t-1}) + \phi_1 \left[log(y_{t-1}) - log(y_{t-2})\right] + \phi_2 \left[log(y_{t-2}) - log(y_{t-3})\right] + \epsilon_t$$

- ETS(A,A,N) model of the natural logarithm of unemployed individuals.

$$\begin{aligned} log(y_t) &= \ell_{t-1} + b_{t-1} + \epsilon_t \\ \ell_t &= \ell_{t-1} + b_{t-1} + \alpha\epsilon_t \\ b_t &= b_{t-1} + \beta\epsilon_t \end{aligned}$$

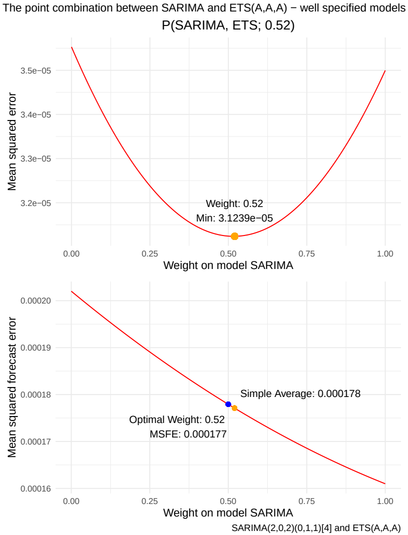


Figure 10: MSFE of predictive unemployment in two well-specified model pools.

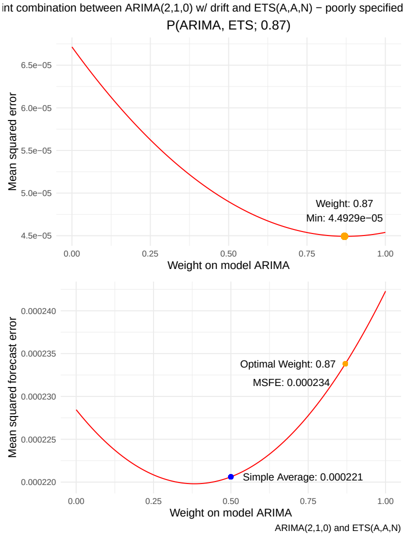


Figure 11: MSFE of predictive unemployment in two poorly-specified model pools.

Example 2 - In-sample Fit Comparison

	P(SARIMA,ETS;0.52)	P(ARIMA,ETS;0.87)
1^{st} Model LogL	321.4497	322.1642
2^{nd} Model LogL	260.9102	231.9507
Difference	60.5395	90.2135
Puzzle	Yes	Yes

Optimal Weight Derivation (Detail) - Model

Models can be written in matrix forms

$$y = x_1\beta_1 + x_2\beta_2 + \epsilon$$

$$M_1 : y = x_1\alpha_1 + u_1$$

$$M_2 : y = x_2\alpha_2 + u_2$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \; x_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{bmatrix}, \; x_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{N2} \end{bmatrix}, \; \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}.$$

Optimal Weight Derivation (Detail) - Parameter Estimation

Applying the OLS estimation over the in-sample period (R).

$$\begin{aligned}
\hat{\alpha}_1 &= (x_1'x_1)^{-1}x_1'y \\
&= (x_1'x_1)^{-1}x_1'(x_1\beta_1 + x_2\beta_2 + \epsilon) \\
&= \beta_1 + (x_1'x_1)^{-1}x_1'x_2\beta_2 \\
&= \beta_1 + \mathrm{var}_R(x_1)^{-1}\mathrm{cov}_R(x_1, x_2)\beta_2
\end{aligned}$$

$$\begin{aligned}
\hat{\alpha}_2 &= (x_2'x_2)^{-1}x_2'y \\
&= (x_2'x_2)^{-1}x_2'(x_1\beta_1 + x_2\beta_2 + \epsilon) \\
&= \beta_2 + (x_2'x_2)^{-1}x_2'x_1\beta_1 \\
&= \beta_2 + \mathrm{var}_R(x_2)^{-1}\mathrm{cov}_R(x_2, x_1)\beta_1
\end{aligned}$$

Optimal Weight Derivation (Detail) - MSE

$$
\begin{aligned}
\hat{y} &= \hat{y}_1\omega + \hat{y}_2(1-\omega) \\
&= x_1\hat{\alpha}_1\omega + \ x_2\hat{\alpha}_2(1-\omega) \\
&= x_1\hat{\alpha}_1\omega - x_2\hat{\alpha}_2\omega + x_2\hat{\alpha}_2 \\
&= (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\omega + x_2\hat{\alpha}_2
\end{aligned}
$$

$$
\begin{aligned}
\hat{\omega}_{\text{opt}} &= \underset{\omega\in[0,1]}{\arg\min}\ \frac{1}{R}(y-\hat{y}_\omega)'(y-\hat{y}_\omega) \\
&= \underset{\omega\in[0,1]}{\arg\min}\ \frac{1}{R}\left[y-(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\omega - x_2\hat{\alpha}_2\right]'\left[y-(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\omega - x_2\hat{\alpha}_2\right]
\end{aligned}
$$

Optimal Weight Derivation (Detail) - Optimal Weight

Solve the First-order condition

$$-\frac{2}{R}(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(y - (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\hat{\omega}_{opt} - x_2\hat{\alpha}_2) = 0.$$

$$\begin{aligned}
(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)\hat{\omega}_{opt} &= (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(y - x_2\hat{\alpha}_2) \\
\hat{\omega}_{opt} &= \frac{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(y - x_2\hat{\alpha}_2)}{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)} \\
\hat{\omega}_{opt} &= \frac{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(y - x_2\hat{\alpha}_2)}{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)} \\
\hat{\omega}_{opt} &= \frac{(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'y - (x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'x_2\hat{\alpha}_2}{\hat{\alpha}_1'x_1'x_1\hat{\alpha}_1 - 2\hat{\alpha}_1'x_1'x_2\hat{\alpha}_2 + \hat{\alpha}_2'x_2'x_2\hat{\alpha}_2}
\end{aligned}$$

Optimal Weight Derivation (Detail) - Meaningful Expression

$$
\begin{aligned}
\hat{\omega}_{opt} &= \frac{R^{-1}(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'y - R^{-1}(x_1\hat{\alpha}_1 - x_2\hat{\alpha}_2)'x_2\hat{\alpha}_2}{\hat{\alpha}_1'\frac{x_1'x_1}{R}\hat{\alpha}_1 - 2\hat{\alpha}_1'\frac{x_1'x_2}{R}\hat{\alpha}_2 + \hat{\alpha}_2'\frac{x_2'x_2}{R}\hat{\alpha}_2} \\
&= \frac{\hat{\alpha}_1'\mathrm{cov}_R(x_1, y) - \hat{\alpha}_2'\mathrm{cov}_R(x_2, y) - \hat{\alpha}_1'\mathrm{cov}_R(x_1, x_2)\hat{\alpha}_2 + \hat{\alpha}_2'\mathrm{cov}_R(x_2, x_2)\hat{\alpha}_2}{\hat{\alpha}_1'\mathrm{cov}_R(x_1, x_1)\hat{\alpha}_1 - 2\hat{\alpha}_1'\mathrm{cov}_R(x_1, x_2)\hat{\alpha}_2 + \hat{\alpha}_2'\mathrm{cov}_R(x_2, x_2)\hat{\alpha}_2} \\
&= \frac{\hat{\alpha}_1'\mathrm{cov}_R(x_1, x_1)\hat{\alpha}_1 - \hat{\alpha}_1'\mathrm{cov}_R(x_1, x_2)\hat{\alpha}_2}{\hat{\alpha}_1'\mathrm{cov}_R(x_1, x_1)\hat{\alpha}_1 - 2\hat{\alpha}_1'\mathrm{cov}_R(x_1, x_2)\hat{\alpha}_2 + \hat{\alpha}_2'\mathrm{cov}_R(x_2, x_2)\hat{\alpha}_2}
\end{aligned}
$$

Optimal Weight Derivation (Detail) - Limit Result

$$\omega_\star = \frac{\alpha_1'\Sigma_{11}\alpha_1 - \alpha_1'\Sigma_{12}\alpha_2}{\alpha_1'\Sigma_{11}\alpha_1 - 2\alpha_1'\Sigma_{12}\alpha_2 + \alpha_2'\Sigma_{22}\alpha_2}$$

For $\omega_\star = \frac{1}{2}$ it must be that

$$\begin{aligned}
\frac{1}{2} &= \frac{\alpha_1'\Sigma_{11}\alpha_1 - \alpha_1'\Sigma_{12}\alpha_2}{\alpha_1'\Sigma_{11}\alpha_1 - 2\alpha_1'\Sigma_{12}\alpha_2 + \alpha_2'\Sigma_{22}\alpha_2} \\
\alpha_1'\Sigma_{11}\alpha_1 - 2\alpha_1'\Sigma_{12}\alpha_2 + \alpha_2'\Sigma_{22}\alpha_2 &= 2\left(\alpha_1'\Sigma_{11}\alpha_1 - \alpha_1'\Sigma_{12}\alpha_2\right) \\
\alpha_1'\Sigma_{11}\alpha_1 + \alpha_2'\Sigma_{22}\alpha_2 &= 2\alpha_1'\Sigma_{11}\alpha_1 \\
\alpha_1'\Sigma_{11}\alpha_1 &= \alpha_2'\Sigma_{22}\alpha_2.
\end{aligned}$$

Optimal Weight Derivation (Detail) - F-statistic

Define the sum squared of errors (SSE) for the true model is $SSE_{full} = (y - x_1\hat{\beta}_1 - x_2\hat{\beta}_2)'(y - x_1\hat{\beta}_1 - x_2\hat{\beta}_2)$.

The unbiased estimator of the true model variance is $s^2 = \frac{SSE_{full}}{R-2}$.

The optimal weight can also be constructed by the F-statistics of M_1 and M_2.

$$\hat{\omega}_{opt} = \frac{F_{\alpha_1} - R\,\hat{\alpha}_1'\text{cov}_R(x_1, x_2)\hat{\alpha}_2/s^2}{F_{\alpha_1} + F_{\alpha_2} - 2R\,\hat{\alpha}_1'\text{cov}_R(x_1, x_2)\hat{\alpha}_2/s^2}.$$

Optimal Weight Derivation (Detail) - F-statistic

The F-statistic follows a F-distribution with degrees of freedom (1,R-2) under H_0, which is defined as

$$F_{\alpha_1} = R\, s^{-2}\, \hat{\alpha}_1' \text{cov}_R(x_1, x_1)\hat{\alpha}_1.$$

Similarly, we have

$$F_{\alpha_2} = R\, s^{-2}\, \hat{\alpha}_2' \text{cov}_R(x_2, x_2)\hat{\alpha}_2 \sim F_{1,R-2} \text{ under } \mathrm{H}_0.$$