When Should Forecasts Be Equally Weighted? Signal Recovery and Estimation Costs

Abstract

Whether estimated combination weights improve on equal weights depends not only on the signal in forecast differences, but also on how much of that signal the forecasting procedure recovers and at what cost. We express the expected mean squared forecast error (MSFE) gain in terms of signal recovery and estimation costs, giving a boundary for when a specified forecasting procedure improves on equal weighting. We complement this theoretical boundary with a decision rule based on past out-of-sample loss differences to choose between equal weights and the estimated weights. Under stationarity and weak dependence, with estimation windows fixed, the probability of choosing the lower-MSFE method tends to one as past evaluations accumulate, provided the two population MSFEs differ. Simulations and macroeconomic applications illustrate the comparison and the choices of equal and estimated weights.

Keywords: forecast combination, equal weights, signal recovery, estimation error, out-of-sample evaluation

1. Introduction

When is it worth estimating forecast combination weights rather than taking a simple average? Unequal weights can exploit differences in predictive information, but a forecasting procedure may recover only part of that information and incur estimation error. This trade-off underlies the forecast combination puzzle (Bates and Granger, 1969; Smith and Wallis, 2009; Claeskens et al., 2016). Whether estimation is worthwhile therefore depends on both the available signal and the procedure used to recover it.

Lee and Lee (2026) provide a direct starting point. Their factor-adjusted combinations treat the simple average as a common component and use component forecasts' deviations from that average to estimate an adjustment. Weight estimators such as Ridge, Lasso, adaptive Lasso, and boosting can recover this information to different degrees and with different estimation costs. We study when a specified forecasting procedure reduces mean squared forecast error (MSFE) relative to equal weights (EW). The decision is whether to use that procedure or retain EW for a fixed target and forecast horizon. The component forecasts are inputs to the weighting rule, which includes tuning and final weight estimation. We call the resulting forecast the candidate combination.

We use the quadratic bias–variance decomposition to express the gain over EW in terms of signal recovery and estimation costs. The resulting signal-recovery–cost boundary includes dependence between estimated weights and subsequent forecast errors. We then calculate how penalty selection and component estimation affect this comparison. In a Gaussian cross-validation example, treating the selected penalty as fixed gives the wrong sign of the gain after refitting. In a model with estimated component forecasts, we derive finite-window component moments and establish conditions for zero or positive signal.

We estimate the net gain from past out-of-sample squared-error differences. We use the candidate when its average loss is lower and otherwise retain EW. We study the probability of choosing the lower-MSFE procedure and the forecast risk of this choice. Under stationarity and weak dependence, with estimation windows fixed, the expected squared forecast error of the selected forecast converges to the smaller of the two MSFEs as more past forecast evaluations become available. When the two methods have similar expected MSFEs, we characterize how uncertainty in the estimated gain affects the probability of selecting the candidate.

For broader reviews of forecast combination, see Timmermann (2006) and Wang et al. (2023). More closely related to our question, several papers study why estimated combination weights may fail to improve on equal weighting. Blanc and Setzer (2016) derive decision

boundaries between the simple average and estimated optimal weights based on training-sample size and changes in forecast-error variances and correlations. Blanc and Setzer (2020) develop a bias–variance analysis of estimated weights and derive optimal shrinkage toward equal weighting. Regularization toward EW remains an important benchmark (Diebold and Pauly, 1990; Diebold and Shin, 2019). Elliott and Liao (2026) show that the population gains available from optimal weighting can be small. Claeskens et al. (2016) show how randomness in estimated weights changes the bias and variance of the combined forecast. Our focus is on a specified forecasting procedure: we characterize how much of the available signal it recovers and the estimation costs it incurs, including those from estimated component forecasts and data-selected tuning.

Past forecast performance has been used to adapt or combine forecasting procedures (Yang, 2004). Qian et al. (2022) give cumulative squared-error risk bounds for an adaptive combination of original and combined forecasts. Inoue and Kilian (2006) compare model selection based on out-of-sample prediction errors with information criteria, while Kourentzes et al. (2019) study forecast selection and combination through the choice of forecast pools. Here the choice is between EW and a specified estimated-weight combination. Frazier et al. (2023) show how two-step estimation of forecast combinations can affect the size and power of forecast-comparison tests. We use fixed estimation windows as the number of past out-of-sample evaluations increases. Our loss-difference comparison follows Diebold and Mariano (1995) and the fixed-window perspective of Giacomini and White (2006). We use self-normalization for uncertainty reporting following Shao (2010b) and Li et al. (2025).

The simulations quantify the signal-recovery–cost decomposition and evaluate the decision rule in 400 independent replications. With more past out-of-sample evaluations, the decision more often agrees with the sign of the expected gain. In the macroeconomic application, the rule makes one pre-evaluation choice for each of five variables at three forecast horizons. The rule matches the method with lower subsequent MSFE in 11 of 15 cases, compared with eight matches for always using EW and seven for always using Ridge. Of

these 11 matches, six retain EW and five use Ridge.

Section 2 develops the signal-recovery–cost decomposition and boundary. Section 3 studies tuning and estimated component forecasts. Section 4 develops the decision rule and its forecast risk. Sections 5 and 6 report the simulations and the macroeconomic application. All proofs and additional results are in the supplement.

2. Signal recovery and estimation costs

2.1. Forecasts and the risk comparison

Fix a target variable and forecast horizon. Let y_t be the outcome and let $\boldsymbol{f}_t = (f_{1,t}, \ldots, f_{N,t})'$ collect $N \geq 2$ component forecasts of y_t. With $\mathbf{1}$ denoting the N-vector of ones, write

$$\bar{f}_t = \frac{1}{N}\mathbf{1}'\boldsymbol{f}_t, \qquad \boldsymbol{\beta}_0 = \frac{1}{N}\mathbf{1}$$

for the equal-weight forecast and its weights. Define its forecast error $u_t = y_t - \bar{f}_t$ and the deviations from the simple average $\boldsymbol{d}_t = \boldsymbol{f}_t - \bar{f}_t\mathbf{1}$.

For weights $\boldsymbol{\beta}$ satisfying $\mathbf{1}'\boldsymbol{\beta} = 1$, let $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$. Since $\mathbf{1}'\boldsymbol{\delta} = 0$,

$$\boldsymbol{\beta}'\boldsymbol{f}_t = \bar{f}_t + \boldsymbol{\delta}'\boldsymbol{f}_t = \bar{f}_t + \boldsymbol{\delta}'\boldsymbol{d}_t.$$

Thus unequal weighting adjusts the simple average using differences among the component forecasts. Negative weights are allowed.

For fixed weights, define the population MSFE by

$$R(\boldsymbol{\beta}) = \mathbb{E}\Big[(y - \boldsymbol{\beta}'\boldsymbol{f})^2\Big].$$

Assume that y and $\boldsymbol{f}$ have finite second moments and that the minimizer below is unique:

$$\boldsymbol{\beta}^* = \underset{\mathbf{1}'\boldsymbol{\beta}=1}{\arg\min}\, R(\boldsymbol{\beta}), \qquad R_0 = R(\boldsymbol{\beta}_0), \qquad S = R_0 - R(\boldsymbol{\beta}^*). \tag{1}$$

The available signal S is the reduction in MSFE relative to EW attainable with the population-optimal weights.

A weighting rule π, including any tuning step, uses a training sample $\mathcal{D}$ to estimate weights $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_\pi(\mathcal{D})$ satisfying $\mathbf{1}'\widehat{\boldsymbol{\beta}} = 1$. Let $\overline{\boldsymbol{\beta}} = \mathbb{E}_\mathcal{D}\widehat{\boldsymbol{\beta}}$ be the mean estimated weights, where $\mathbb{E}_\mathcal{D}$ averages over training samples. Assume $\mathbb{E}_\mathcal{D}\|\widehat{\boldsymbol{\beta}}\|^2 < \infty$, and define

$$B = R(\overline{\boldsymbol{\beta}}) - R(\boldsymbol{\beta}^*), \qquad V = \mathbb{E}_\mathcal{D} R(\widehat{\boldsymbol{\beta}}) - R(\overline{\boldsymbol{\beta}}). \tag{2}$$

Here B is the MSFE cost of bias in the estimated weights, and V is the cost of weight-estimation variance.

If the forecasts and outcomes used for evaluation are independent of the training data, the expected gain over EW is

$$G^I = R_0 - \mathbb{E}_\mathcal{D} R(\widehat{\boldsymbol{\beta}}).$$

The outer expectation averages the risk of the fitted weights over training samples. For subsequent observations from the same time series, the expected gain is

$$G^C = \mathbb{E}\Big[(y - \bar{f})^2 - (y - \widehat{\boldsymbol{\beta}}'\boldsymbol{f})^2\Big],$$

with finite expected squared errors. The superscripts I and C denote independent and consecutive evaluation, respectively. Define

$$\Delta = G^C - G^I. \tag{3}$$

The two gains use the same training and evaluation distributions and differ only in their dependence. Thus Δ measures the effect of this dependence on the gain. Dependence between estimated weights and forecast errors is studied by Claeskens et al. (2016). By the definition of V,

$$G^C = \Big[R_0 - R(\overline{\boldsymbol{\beta}})\Big] - V + \Delta.$$

The term in brackets is $S - B$, the gain from the mean estimated weights. For a fixed evaluation block, the gain is averaged over its forecast dates.

To express this gain in terms of signal recovery, write the optimal and mean estimated adjustments to EW as

$$g^* = (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)'\boldsymbol{d}, \qquad \bar{g} = (\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\boldsymbol{d}.$$

The corresponding forecast errors are $u - g^*$ and $u - \bar{g}$. The first-order condition gives $\mathbb{E}[(u - g^*)g^*] = 0$, so

$$S = \mathbb{E}\Big[u^2 - (u - g^*)^2\Big] = \mathbb{E}[(g^*)^2].$$

For $S > 0$, define

$$a = \frac{\mathbb{E}[\bar{g}g^*]}{S}, \qquad D^{\perp} = \mathbb{E}\Big[(\bar{g} - ag^*)^2\Big]. \tag{4}$$

The multiple ag^* best approximates $\bar{g}$ in mean square, and $D^{\perp}$ is the remaining mean squared difference. When $D^{\perp} = 0$, the mean weights equal the EW weights for $a = 0$ and the population-optimal weights for $a = 1$.

Theorem 1 (Signal recovery and estimation costs)**.** *Under the assumptions above,*

$$G^I = S - B - V, \qquad G^C = S - B - V + \Delta. \tag{5}$$

If $S > 0$, then

$$B = (1 - a)^2 S + D^{\perp} \tag{6}$$

and hence

$$G^C = a(2 - a)S - D^{\perp} - V + \Delta. \tag{7}$$

The identity $G^I = S - B - V$ has the usual bias–variance form (Blanc and Setzer, 2020). For $S > 0$, the first-order condition for the optimal weights gives $\mathbb{E}[(u - g^*)\bar{g}] = 0$. By the definition of a, $\mathbb{E}[u\bar{g}] = \mathbb{E}[g^*\bar{g}] = aS$. Also, (4) gives $\mathbb{E}[\bar{g}^2] = a^2 S + D^{\perp}$. Expanding the

squared errors therefore gives

$$R_0 - R(\overline{\boldsymbol{\beta}}) = 2\mathbb{E}[u\bar{g}] - \mathbb{E}[\bar{g}^2] = a(2-a)S - D^{\perp}.$$

Subtracting V and adding Δ yields (7). Thus the candidate has lower MSFE than EW on subsequent observations if and only if

$$a(2-a)S > D^{\perp} + V - \Delta. \tag{8}$$

The coefficient a can lie outside $[0,1]$, and Δ can have either sign. When $S = 0$, equation (5) applies without defining a.

2.2. Recovery under group symmetry

Consider two nonempty groups of component forecasts. Suppose that swapping the labels of any two forecasts within either group leaves the distributions of the training and evaluation data unchanged. Assume also that, under the same relabeling, the weighting rule swaps the corresponding estimated weights without changing the combined forecast.

Corollary 1 (Two exchangeable groups)**.** *Under the assumptions of Theorem 1 and the symmetry conditions above, suppose $S > 0$ and no combination intercept is fitted. Then $\bar{g} = ag^*$ and $D^{\perp} = 0$, so*

$$G^C > 0 \quad \Longleftrightarrow \quad a(2-a)S > V - \Delta. \tag{9}$$

Symmetry makes both the optimal weights and the mean estimated weights equal within each group. Since the weights sum to one, the total weight assigned to one group determines the total assigned to the other. Their adjustments to EW are therefore proportional. The quantities a, V, and Δ may change with signal strength.

For example, consider four forecasts in groups $\{1, 2\}$ and $\{3, 4\}$, and suppose that

$$\boldsymbol{\beta}^* = (0.35, 0.35, 0.15, 0.15)', \qquad \overline{\boldsymbol{\beta}} = (0.30, 0.30, 0.20, 0.20)'.$$

EW assigns 0.25 to each forecast. The mean weights move halfway from EW to the optimal weights, so $a = 1/2$ and $D^{\perp} = 0$. Their MSFE gain over EW is $3S/4$, and the fitted combination has gain $G^C = 3S/4 - V + \Delta$.

3. Shrinkage and estimated component forecasts

The following examples show how shrinkage and component estimation affect signal recovery and estimation costs in Section 2.

3.1. Fixed and data-dependent shrinkage

Shrinking the estimated adjustment toward zero moves the combined forecast toward EW (Blanc and Setzer, 2020). Throughout this subsection, the evaluation data are independent of the training data, and the gain is G^I.

For this example, use $p = N - 1$ linear combinations $\widetilde{d}_j$ of the forecast differences to represent every adjustment to EW, normalized so that

$$\mathbb{E}[\widetilde{d}_j^2] = 1, \qquad \mathbb{E}[\widetilde{d}_i \widetilde{d}_j] = 0 \quad (i \neq j).$$

Write the optimal adjustment as $g^* = \sum_{j=1}^{p} \theta_j \widetilde{d}_j$. With estimated coefficients $\widehat{\theta}_j$, the combined forecast is $\bar{f} + \sum_{j=1}^{p} \widehat{\theta}_j \widetilde{d}_j$, and its error is

$$u - \sum_{j=1}^{p} \widehat{\theta}_j \widetilde{d}_j = (u - g^*) + \sum_{j=1}^{p} (\theta_j - \widehat{\theta}_j) \widetilde{d}_j.$$

The first term is the error of the optimal combination. The normalization uses population moments. Optimality gives $\mathbb{E}[(u - g^*)\widetilde{d}_j] = 0$ for each j. Together with independent

evaluation and the normalization above, this yields

$$\mathbb{E}\left[\left(u - \sum_{j=1}^{p} \widehat{\theta}_j \widetilde{d}_j\right)^2\right] = R(\boldsymbol{\beta}^*) + \sum_{j=1}^{p} \mathbb{E}\Big[(\widehat{\theta}_j - \theta_j)^2\Big].$$

Thus the normalization gives

$$S = \mathbb{E}[(g^*)^2] = \sum_{j=1}^{p} \theta_j^2, \qquad G^I = S - \sum_{j=1}^{p} \mathbb{E}\Big[(\widehat{\theta}_j - \theta_j)^2\Big].$$

The coefficient errors therefore measure the additional MSFE from estimating the adjustment.

Let n be the estimation sample size and Z_j the estimate of θ_j before shrinkage. Suppose these estimates are independent, with $Z_j \sim N(\theta_j, \sigma^2/n)$ and $\sigma^2 > 0$. For fixed shrinkage factors η_j, use $\widehat{\theta}_j = \eta_j Z_j$. Since

$$\eta_j Z_j - \theta_j = (\eta_j - 1)\theta_j + \eta_j(Z_j - \theta_j),$$

the cross term vanishes in expectation because $\mathbb{E}[Z_j - \theta_j] = 0$. Consequently,

$$\mathbb{E}\Big[(\eta_j Z_j - \theta_j)^2\Big] = (1 - \eta_j)^2\theta_j^2 + \eta_j^2 \frac{\sigma^2}{n},$$

and substitution into the gain gives

$$G^I = \sum_{j=1}^{p} \eta_j(2 - \eta_j)\theta_j^2 - \frac{\sigma^2}{n} \sum_{j=1}^{p} \eta_j^2. \tag{10}$$

With a common factor $\eta \in (0, 1]$, this becomes $G^I = \eta(2-\eta)S - \eta^2 p\sigma^2/n$. Since $\eta(2-\eta) > 0$,

$$G^I > 0 \quad \Longleftrightarrow \quad S > \frac{\eta}{2 - \eta} \frac{p\sigma^2}{n}. \tag{11}$$

Blanc and Setzer (2020) obtain the same fixed-shrinkage risk structure for estimated opti-

mal weights, with a variance term determined by their estimator. For $S > 0$, this is the comparison in Section 2 with $a = \eta$, $D^{\perp} = \Delta = 0$, and $V = \eta^2 p\sigma^2/n$. For fixed p, η, and σ^2, a larger estimation sample lowers the required signal. With different shrinkage factors, the distribution of the signal across coefficients also matters. For $p = 2$, take $(\eta_1, \eta_2) = (.9, .05)$, $S = .05$, and $\sigma^2/n = .01$. Putting all the signal in the first coefficient gives a gain of .041375; putting it in the second gives $-.00325$.

When shrinkage is selected from the data, the selected factor and the coefficient estimate need not be independent. With one adjustment coefficient, let Z be an estimate before shrinkage satisfying $\mathbb{E}Z = \theta$ and $\operatorname{Var} Z = \tau^2$. Select $\widehat{\eta}$ from the training data and use $\widehat{\theta} = \widehat{\eta}Z$. The expected gain is

$$G^I = \theta^2 - \mathbb{E}\Big[(\widehat{\eta}Z - \theta)^2\Big].$$

Assuming the required moments are finite, expansion gives

$$G^I = \mathbb{E}\Big[(2\widehat{\eta} - \widehat{\eta}^2)\theta^2 - \tau^2\widehat{\eta}^2\Big] + J, \tag{12}$$

where

$$J = 2\theta \operatorname{Cov}(\widehat{\eta}, Z) - \operatorname{Cov}(\widehat{\eta}^2, Z^2).$$

The term J accounts for dependence between the selected factor $\widehat{\eta}$ and the estimate Z. Ignoring this dependence amounts to omitting J. Related effects of tuning are studied by Tibshirani and Rosset (2019).

To illustrate selection followed by re-estimation, divide a sample of even size n into two equal blocks for estimation and validation (Arlot and Celisse, 2010). Let Z_1, Z_2 be independent $N(\theta, 2\sigma^2/n)$ estimates from the two blocks. Choose $\eta \in \{0, .25, .5, 1\}$ to minimize $(Z_2 - \eta Z_1)^2$, then refit using $Z = (Z_1 + Z_2)/2$ and the coefficient $\widehat{\theta} = \widehat{\eta}Z$. Table 1 reports the expected gain and the value obtained when J is omitted, both calculated by Gaussian

integration. At $n\theta^2/\sigma^2 = .85$, ignoring selection predicts an improvement over EW, whereas the forecast has a negative expected gain. At $n\theta^2/\sigma^2 = 4$, ignoring selection understates the gain.

Table 1: Expected gains with data-selected shrinkage

$n\theta^2/\sigma^2$	Expected gain	Ignoring selection
0.00	−0.547	−0.322
0.85	−0.0404	0.0172
1.00	0.0587	0.0949
4.00	2.42	2.30

Notes: Entries are scaled by n/σ^2. Both columns use the same selected shrinkage factor and the full-sample estimate $Z = (Z_1 + Z_2)/2$. "Ignoring selection" omits J in (12), with $\tau^2 = \sigma^2/n$.

3.2. Estimated component forecasts

We now study the gain when the component forecasts are estimated from finite samples. Let F_t be a common factor, $\boldsymbol{x}_t$ an N-vector of predictors, and y_t the outcome. Consider the stationary Gaussian model

$$\begin{aligned} F_t &= \rho F_{t-1} + \xi_t, & \xi_t &\sim N(0, \sigma_F^2(1-\rho^2)), \\ \boldsymbol{x}_t &= F_t \mathbf{1} + \boldsymbol{e}_t, & \boldsymbol{e}_t &\sim N_N(0, \delta \boldsymbol{I}_N), \\ y_t &= F_t + \boldsymbol{\theta}' \boldsymbol{e}_t + \nu_t, & \nu_t &\sim N(0, v). \end{aligned} \tag{13}$$

Here σ_F^2 is the stationary factor variance, ρ its autoregressive coefficient, δ the variance of each predictor-specific innovation, and v the variance of the outcome noise. In this model, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)'$ gives the coefficients on the predictor-specific innovations; $\boldsymbol{I}_N$ is the identity matrix. Assume $\sigma_F^2, \delta, v > 0$ and $|\rho| < 1$. The innovations ξ_t, $\boldsymbol{e}_t$, and ν_t are mutually independent and independent across dates. The factor starts in its stationary distribution.

The predictors $\boldsymbol{x}_t$ are observed before y_t. Each component regresses y on an intercept and one predictor using the previous $m \geq 2$ observations, then predicts y_t from $x_{j,t}$. Estimation

uses OLS, except that a slope is set to zero when the centered predictor sum of squares is at most 10^{-12}; the intercept is then the sample mean of y.

Using n past forecast–outcome pairs, the combination regresses the EW errors on the forecast differences by ridge regression (Hoerl and Kennard, 1970), without a combination intercept. For a finite penalty $\lambda \geq 0$, a training fold of n_f observations uses the objective

$$n_f^{-1}\|\boldsymbol{u} - \boldsymbol{D}_s\boldsymbol{c}\|^2 + \lambda\|\boldsymbol{c}\|^2.$$

Here $\boldsymbol{u}$ contains the EW errors and $\boldsymbol{c}$ contains the adjustment coefficients. Each column of $\boldsymbol{D}_s$ is a forecast difference divided by its training standard deviation, without subtracting its mean. Five chronological cross-validation folds use expanding training samples to select the penalty from a fixed grid that includes zero penalty and EW. The selected penalty is used to refit on all n observations.[1]

Write S_m for the signal available with component estimation window m, and $G^I_{m,n}, G^C_{m,n}$ for this Ridge combination's gains under independent and subsequent time-series evaluation, respectively.

Proposition 1 (Signal and gains with estimated components)**.** *In* (13)*, the population-optimal weights subject to summing to one are unique. If* $\boldsymbol{\theta} = \theta_0\mathbf{1}$ *for a scalar* θ_0*, then* $S_m = 0$ *and* $G^I_{m,n} < 0$ *whenever the fitted adjustment is nonzero with positive probability.*

For $N = 2$*,* $\rho = 0$*, and* $m > 3$*, define* $P_j = (\sigma_F^2 + \delta\theta_j)^2/(\sigma_F^2 + \delta)$*,* $j = 1, 2$*. Then*

$$\mathbb{E}\{u(f_1 - f_2)\} = \frac{\kappa_m}{2}(P_1 - P_2), \qquad \kappa_m > 0, \tag{14}$$

where κ_m *depends on* m*, the predictor variance, and the component slope threshold. Thus* $P_1 \neq P_2$ *implies* $S_m > 0$*.*

For $N = 2$*,* $\rho = 0$*,* $m = 120$*,* $\sigma_F^2 = 25$*,* $\delta = v = 1$*, and* $\boldsymbol{\theta} = (1, 0)'$*, both* $G^I_{m,n}$ *and* $G^C_{m,n}$

[1]The penalty grid and numerical rules are specified in Supplementary Section S3.1.

are positive for all sufficiently large n divisible by six. For each such n, $G^I_{m,n} = 0$ at some $\boldsymbol{\theta} = (\vartheta, 0)'$ with $0 < \vartheta < 1$.

With the component estimation window fixed, this Ridge combination can have positive or negative expected gain under independent evaluation.

4. Choosing between equal and estimated weights

4.1. Past forecast errors and the decision rule

Fix a target, a forecast horizon $h \geq 1$, and a forecasting procedure. An evaluation starting at date t covers H forecasts for outcomes $y_{t+h}, \ldots, y_{t+h+H-1}$. Let $\bar{f}_{t,j}$ and $f_{\pi,t,j}$ be the EW and candidate forecasts for $y_{t+h+j-1}$, with coefficients updated as specified by the procedure. Define the average loss difference

$$L_t = \frac{1}{H} \sum_{j=1}^{H} \Big\{ (y_{t+h+j-1} - \bar{f}_{t,j})^2 - (y_{t+h+j-1} - f_{\pi,t,j})^2 \Big\}. \tag{15}$$

A positive L_t means that the candidate has lower average squared forecast error over these outcomes.

Each evaluation uses a fixed span of w observations through its start date for the initial fit. At date T, outcomes are observed through T, so completed evaluations start at $w, \ldots, T - h - H + 1$. Assuming at least one is available, their number is $M = T - w - h - H + 2$. Applying the same procedure at each start date gives

$$\widehat{G}_M = \frac{1}{M} \sum_{t=w}^{T-h-H+1} L_t. \tag{16}$$

This is the sample mean of out-of-sample loss differences (Diebold and Mariano, 1995), with estimation based on fixed windows (Giacomini and White, 2006).

Proposition 2 (Estimation of the expected gain)**.** *Suppose the data process is strictly stationary, each L_t is computed by the same measurable function of a fixed-length data window, and $\mathbb{E}|L_0| < \infty$. Then $\mathbb{E}\widehat{G}_M = G^C$. If (L_t) is ergodic, $\widehat{G}_M \to G^C$ almost surely.*

The mean estimates the gain in (5), including the effects of weight estimation. The evaluations can share observations, so their loss differences can be dependent. With $\gamma_L(k) = \mathrm{Cov}(L_0, L_k)$ and finite second moments,

$$\mathrm{Var}(\widehat{G}_M) = \frac{1}{M}\left\{\gamma_L(0) + 2\sum_{k=1}^{M-1}(1 - k/M)\gamma_L(k)\right\}. \tag{17}$$

Write $G = G^C$ throughout this section. Let $A = 1$ select the candidate and $A = 0$ select EW. Choosing the forecast with lower past MSFE gives

$$\widehat{A}_M = \mathbf{1}\{\widehat{G}_M > 0\}, \qquad f^A = \bar{f} + \widehat{A}_M(f_\pi - \bar{f}), \tag{18}$$

where $\mathbf{1}\{\cdot\}$ is the event indicator and $\bar{f}, f_\pi$ are the EW and candidate forecasts for the date being forecast. A tie retains EW. The selected procedure continues its specified coefficient updates. This choice compares estimated risks (Hirano and Porter, 2009).

Proposition 3 (Selection error and forecast risk)**.** *Let* $A^* = \mathbf{1}\{G > 0\}$ *and* $G_+ = \max(G, 0)$. *For any estimate* $\widehat{G}$ *and the decision* $\widehat{A} = \mathbf{1}\{\widehat{G} > 0\}$,

$$G_+ - \widehat{A}G = |G|\mathbf{1}\{\widehat{A} \neq A^*\} \leq |\widehat{G} - G|. \tag{19}$$

For any binary decision A *and integrable future loss difference* L_T^+ *with* $\mathbb{E}L_T^+ = G$,

$$\mathcal{R}(A) := G_+ - \mathbb{E}[AL_T^+] = |G|\mathbb{P}(A \neq A^*) - \mathrm{Cov}(A, L_T^+). \tag{20}$$

The quantity G_+ is the gain from always using the procedure with lower MSFE. A wrong choice loses the difference $|G|$ between the two MSFEs. For a data-dependent decision, the future gain is AL_T^+, whose expectation satisfies

$$\mathbb{E}[AL_T^+] = G\mathbb{E}A + \mathrm{Cov}(A, L_T^+).$$

Thus $\mathcal{R}(A)$ is the selected forecast's expected squared error minus that of always using the better procedure. The covariance matters because the decision and future errors can depend on overlapping or serially related observations. It can have either sign, so $\mathcal{R}(A)$ can also be negative.

4.2. Forecast risk and decisions near the boundary

Let $\boldsymbol{X}_t$ collect the observations used by the forecasting procedure: outcomes and supplied component forecasts, or outcomes and predictors when the component forecasts are estimated. Write

$$L_t = \Psi(\boldsymbol{X}_{t-w+1}, \ldots, \boldsymbol{X}_{t+h+H-1}),$$

where Ψ is the same measurable function at each date. Keep w, h, H, and the forecasting procedure fixed as $M \to \infty$. Suppose $(\boldsymbol{X}_t)$ is strictly stationary and strongly mixing, with mixing coefficient $\alpha_X(k) \to 0$. This requires dependence between distant groups of observations to vanish. Assume $\mathbb{E}|L_0|^{1+\epsilon} < \infty$ for some $\epsilon > 0$ and the central limit theorem

$$\sqrt{M}(\widehat{G}_M - G) \Rightarrow N(0, \sigma_L^2), \qquad 0 < \sigma_L^2 < \infty. \tag{21}$$

Here σ_L^2 is the long-run variance. Only the number of past evaluations increases; the estimation windows remain fixed.

Theorem 2 (Forecast risk of the decision rule)**.** *Under the assumptions above, let L_T^+ be the average loss difference for the evaluation starting at T, so $\mathbb{E}L_T^+ = G$. Then*

$$\mathbb{E}[\widehat{A}_M L_T^+] \longrightarrow G_+. \tag{22}$$

The probability of selecting the candidate tends to zero for $G < 0$, to one for $G > 0$, and to one half for $G = 0$.

When the two mean MSFEs differ, the rule chooses the lower-MSFE procedure with probability tending to one. Its expected squared forecast error also converges to the smaller

of the two MSFEs. For one decision followed by a fixed number of monthly forecasts, weights may be re-estimated each month. The result applies if historical and future forecasts use the same procedure and each has mean gain G.

When the two MSFEs are close, uncertainty in the estimated gain affects the choice. To describe this case, allow the data distribution to vary with M so that its mean gain G_M approaches zero. The forecasting procedure and estimation windows remain fixed. Let Φ be the standard normal distribution function.

Corollary 2 (Selection when MSFEs are close)**.** *Suppose, for a positive long-run standard deviation ω,*

$$\frac{\sqrt{M}(\widehat{G}_M - G_M)}{\omega} \Rightarrow N(0,1), \qquad \frac{\sqrt{M}G_M}{\omega} \to \delta_0, \qquad 0 < \omega < \infty.$$

Then $\mathbb{P}(\widehat{A}_M = 1) \to \Phi(\delta_0)$ and

$$\frac{\sqrt{M}}{\omega}\Big\{(G_M)_+ - G_M \mathbb{E}\widehat{A}_M\Big\} \longrightarrow |\delta_0|\Phi(-|\delta_0|), \tag{23}$$

where $(G_M)_+ = \max(G_M, 0)$ and δ_0 is the limiting mean gain measured in standard-error units.

The ratio $\sqrt{M}G_M/\omega$ compares the mean gain with the asymptotic standard deviation of its estimate. For a family indexed by signal S, with other model features fixed, suppose $G(S_*) = 0$, $G'(S_*) \neq 0$, and $\sqrt{M}(S_M - S_*) \to c$. Then $\delta_0 = G'(S_*)c/\omega$. Selection therefore depends on the distance from the boundary, the slope of the gain, and estimation uncertainty.

The term $(G_M)_+ - G_M\mathbb{E}\widehat{A}_M$ equals $|G_M|$ times the probability of choosing the higher-MSFE procedure. The same normalized limit for future forecast risk also requires

$$\frac{\sqrt{M}}{\omega}\operatorname{Cov}(\widehat{A}_M, L_M^+) \to 0,$$

where L_M^+ is the future loss difference under the corresponding data distribution, with $\mathbb{E}L_M^+ =$

G_M.

We report a self-normalized (SN) confidence interval for G (Shao, 2010b,a). The corresponding statistic for testing $G = 0$ is the self-normalized Diebold–Mariano statistic of Li et al. (2025). Self-normalization uses cumulative centered loss differences in place of a direct long-run variance estimate. Its asymptotic validity follows under a functional central limit theorem. We also report a heteroskedasticity and autocorrelation consistent (HAC) interval. The simulations also compare SN and HAC rules that select the candidate only when $\widehat{G}_M$ exceeds a positive threshold based on its uncertainty.[2]

5. Monte Carlo experiments

5.1. Signal recovery and estimation costs

We first estimate the risk components in Section 2, then assess the decision rule using separate replications. Both experiments use (13) and the component estimation and Ridge combination in Section 3. The estimation windows are $m = 120$ and $n = 480$ ($w = m + n = 600$), and each evaluation covers $H = 120$ subsequent outcomes with $h = 1$ in the notation of Section 4. Combination weights remain fixed during each evaluation, while the component models are re-estimated at each date. The factor starts in its stationary distribution.

With two forecasts, the adjustment uses a single forecast difference. The $N = 119$ designs study a larger forecast pool, matching the number of components in Section 6. For $N = 2$, set $\sigma_F^2 = 25$, $\rho = 0$, $\delta = v = 1$, and $\boldsymbol{\theta} = (\vartheta, 0)'$ with $\vartheta \in \{0, .25, .5, 1\}$. For $N = 119$, set $\sigma_F^2 = 25/(1 - .7^2)$, $\rho = .7$, and $\delta = v = 1$. s entries of $\boldsymbol{\theta}$ equal $\sqrt{r/(1-r)}/\sqrt{s}$, with all other entries set to zero, for

$$(s, r) \in \{(3, 0), (3, .025), (3, .08), (8, 0), (8, .048), (8, .08)\}.$$

The parameter r sets the squared loading magnitude: $\|\boldsymbol{\theta}\|^2 = r/(1 - r)$. The two $r = 0$

[2]The SN interval and comparison thresholds are given in Supplementary Section S5.

configurations describe the same zero-signal process and use separate sets of replications.

We use 2,048 replications for each configuration. Each fitted combination is evaluated on subsequent observations and on an independently generated series from the same model. These replications provide reference estimates of the risk components and expected gains.

Table 2: Signal recovery and expected MSFE gains

Design	Signal S_m	Recovery a	Gain from mean weights $a(2-a)S_m$	Variance cost V	Dependence Δ	Net gain G^C
$N = 2,\ \vartheta = .25$	0.0310	0.832	0.0301	0.00390	0.00007	0.0266
$N = 2,\ \vartheta = .5$	0.130	0.908	0.129	0.00350	−0.00504	0.120
$N = 2,\ \vartheta = 1$	0.502	0.962	0.501	0.00319	−0.00207	0.495
$s = 3,\ r = .025$	0.0266	0.058	0.00299	0.00181	0.00019	0.00141
$s = 3,\ r = .08$	0.0903	0.129	0.0218	0.00665	0.00068	0.0157
$s = 8,\ r = .048$	0.0494	0.084	0.00795	0.00329	0.00005	0.00488
$s = 8,\ r = .08$	0.0864	0.125	0.0202	0.00608	0.00120	0.0153

Notes: Entries are Monte Carlo estimates from 2,048 replications per configuration. The lower four rows use $N = 119$. The table shows positive-signal configurations; a is undefined when $S_m = 0$. Direct gain estimates and gains reconstructed from the estimated risk components differ by at most 0.00041 across all ten configurations.

With 119 forecasts, the mean estimated weights recover about 11–24% of the available MSFE reduction in the four positive-signal designs (Table 2). Group symmetry gives $D^{\perp} = 0$, so this gain is $a(2-a)S_m$. For $s = 3$ and $r = .025$, the available signal is estimated as .02661, but the mean adjustment yields a gain of only .00299, before a variance cost of .00181. At $r = .08$, the gain from the mean adjustment rises to .02184, compared with a variance cost of .00665. The estimated recovery coefficient a ranges from .058 to .129. The reference estimates of G^C are positive in these four designs and negative in the zero-signal designs.

5.2. Performance of the decision rule

We evaluate the decision rule in 400 replications independent of the reference experiment: 100 for $N = 2$, $\vartheta = .25$, and 100 for each of the $N = 119$, $s = 3$, $r \in \{0, .025, .08\}$ designs. The rule and designs were fixed before these replications were generated.

Each replication contains 4,800 past observations and 120 subsequent outcomes. We compute $\widehat{G}_M$ using either the final 1,200 past observations, giving $M = 481$ evaluations, or all 4,800 observations, giving $M = 4{,}081$. Both decisions use the same candidate fitted from the final $m + n = 600$ observations and are evaluated on the same future outcomes. Only the history used to choose between EW and Ridge changes; the estimation windows remain fixed.

Benchmarks are EW, always using Ridge, fixed-shrinkage combinations $\bar{f} + \eta(f_\pi - \bar{f})$ with $\eta = .25, .50, .75$, and the SN and HAC rules described in Section 4. For each design, all 100 replications enter every comparison. We refer to $r = .025$ and $r = .08$ as the weak- and stronger-signal designs, respectively. Monte Carlo standard errors (MCSEs) of differences between rules are calculated from paired differences within replications (Morris et al., 2019).

More past evaluations make the estimated gain more precise. With the longer history, the sampling standard deviation of $\widehat{G}_M$ falls to .29–.38 of its short-history value. Mean estimates differ from the reference gains by at most 2.23 MCSEs, allowing for Monte Carlo error in both estimates. The nominal 95% SN intervals contain the separately estimated reference gain in 73–80 of 100 replications with the shorter history and 88–99 with the longer history.

Table 3: Forecast choices and mean MSFE gains

Design	Past observations	Ridge selections	Rule	SN	Ridge	Rule–Ridge difference	MCSE
$N = 2,\ \vartheta = .25$	1,200	95/100	0.0206	0.0127	0.0214	−0.000846	0.000883
$N = 2,\ \vartheta = .25$	4,800	100/100	0.0214	0.0214	0.0214	0	0
$N = 119,\ r = 0$	1,200	22/100	−0.000191	0	−0.00138	0.00119	0.000669
$N = 119,\ r = 0$	4,800	13/100	−0.000288	0	−0.00138	0.00109	0.000685
$N = 119,\ r = .025$	1,200	64/100	0.00121	0.000170	0.000843	0.000362	0.000423
$N = 119,\ r = .025$	4,800	87/100	0.000716	0.000384	0.000843	−0.000127	0.000134
$N = 119,\ r = .08$	1,200	97/100	0.0179	0.0120	0.0183	−0.000369	0.000313
$N = 119,\ r = .08$	4,800	100/100	0.0183	0.0178	0.0183	0	0

Notes: "Rule" denotes (18), and "Ridge selections" counts how often it chooses Ridge. The Rule, SN, and Ridge columns report mean MSFE reductions relative to EW over the subsequent $H = 120$ outcomes. Gains are in squared-error units, with EW equal to zero. Ridge denotes always using Ridge. The final two columns report paired Rule–Ridge differences and their MCSEs. All $N = 119$ designs have $s = 3$.

With the longer history, the rule retains EW in 87 of the 100 zero-signal replications and selects Ridge in 87 of the 100 weak-signal replications. It selects Ridge in every two-forecast and stronger-signal replication, matching the gains from always using Ridge. The largest mean gains over the SN rule occur with the shorter history in these last two designs. The weak-signal differences from SN are less precisely estimated.

Fixed shrinkage also yields useful forecasts. With $\eta = .75$, its mean gain is .001069 in the weak-signal design and .015171 in the stronger-signal design, compared with .000716 and .018272 for the long-history rule.

At zero signal, always using Ridge has a mean gain of $-.001377$ (MCSE .000692). With either history, the decision rule reduces this loss but still has higher mean MSFE than EW. SN retains EW throughout, while HAC selects Ridge once with the shorter history. For the longer history, the sampling standard deviation of $\widehat{G}_M$ is .000631, close to the magnitude of the reference gain, .000627.

More frequent agreement with the reference ranking does not raise mean gains in every design. Using the longer rather than shorter history changes the mean gain by $-.000097$ (MCSE .000202) at zero signal and by $-.000488$ (MCSE .000437) at weak signal. Both differences are imprecisely estimated. As (20) shows, forecast performance depends on which replications change their decision and on the losses in those replications, not only on the number of correct choices.

6. Macroeconomic applications

6.1. Data and forecasting procedure

We use the FRED-MD data in Lee and Lee (2026) for five targets: real personal income (RPI), the consumer price index (CPIAUCSL), the personal consumption expenditures price index (PCEPI), industrial production (INDPRO), and the unemployment rate (UNRATE). FRED-MD is described by McCracken and Ng (2016). The raw series span January 1960–December 2019. We use first log differences for RPI and INDPRO, second log differences for

CPIAUCSL and PCEPI, and first differences for UNRATE. The price-index targets therefore measure changes in monthly inflation. Dropping the first two months leaves 718 observations beginning in March 1960, with 119 retained predictors for each target.

At forecast origin o, component j predicts y_{o+h} by

$$f_{j,o+h|o} = \widehat{\alpha}_{j,o} + \widehat{b}_{j,o} x_{j,o}, \qquad h = 1, 2, 3.$$

The intercept and slope are estimated by OLS using predictor–outcome pairs whose outcomes are observed by o. Following the authors' program, the component windows contain 117, 116, and 115 pairs at the three horizons. The components are re-estimated monthly, and EW is their simple average.

The candidate is the factor-adjusted Ridge combination of Lee and Lee (2026), estimated from 480 past forecast–outcome pairs. Forecast differences are divided by their training root mean squares without subtracting their means. Ten contiguous cross-validation blocks each select a penalty from 100 values between 10^{-5} and 10^{6}; the mean selected penalty is mapped to the nearest grid value for the final fit. The penalty is normalized by the number of component forecasts, as in their program. Writing $\widehat{\boldsymbol{c}}$ for the fitted coefficients on the unscaled forecast differences, the combination intercept is $\widehat{\alpha} = \bar{u} - \overline{\boldsymbol{d}}'\widehat{\boldsymbol{c}}$, where $\bar{u}$ and $\overline{\boldsymbol{d}}$ are the training-sample means. The candidate is $\bar{f} + \widehat{\alpha} + \boldsymbol{d}'\widehat{\boldsymbol{c}}$.[3]

At every horizon, the component and combination estimation windows end at forecast origin o. All training outcomes are therefore observed when the forecast is made.

For each target and horizon, the rule makes one choice before forecasting January 2015. The decision dates are December, November, and October 2014 for horizons one, two, and three, using $M = 60, 57, 54$ past loss differences, respectively. Each historical loss difference is for one forecast, so $H = 1$ in Section 4. A positive mean gain selects Ridge; otherwise the rule retains EW. The choice remains fixed for the 60 target months from January 2015 to

[3]Supplementary Section S1.3 extends the risk decomposition to combinations with an intercept.

December 2019, while coefficients continue to update monthly.

The data and evaluation period were examined during development, so this is a retrospective comparison. It uses one data vintage, the authors' full-period predictor completeness filter, and the convention that observations dated o are available at month-end.

6.2. Forecast choices and subsequent MSFE

Evidence on forecast combination differs across targets. Stock and Watson (2004) find that simple combinations work well for output growth, while Genre et al. (2013) find more gains over equal weighting for inflation than for GDP growth or unemployment. We assess each target and horizon separately.

Let $\widehat{R}_{\mathrm{EW},k,h}$ and $\widehat{R}_{\pi,k,h}$ be the two MSFEs over the evaluation period for target k and horizon h. With $A_{k,h}$ denoting the choice made before evaluation, the selected forecast has MSFE

$$\widehat{R}_{A,k,h} = (1 - A_{k,h})\widehat{R}_{\mathrm{EW},k,h} + A_{k,h}\widehat{R}_{\pi,k,h}. \tag{24}$$

Table 4 separates agreement with the subsequent ranking from the size of the MSFE differences. Panel A counts choices matching the lower MSFE of EW and Ridge. Panel B reports every choice and the MSFE ratios relative to EW.

The rule matches the lower subsequent MSFE in 11 of 15 cases, compared with eight for always using EW and seven for always using Ridge. It correctly retains EW in six of the eight cases where EW is better and selects Ridge in five of the seven cases where Ridge is better.

The largest captured gains are for the one-month CPI and PCE targets, where Ridge reduces MSFE relative to EW by 12.23% and 14.87%. The historical CPI gain is close to zero despite its large subsequent improvement. Correct choices of EW occur for RPI and industrial production at horizons one and two, and for both price-index targets at horizon three.

The four mismatches involve errors in both directions. Retaining EW for RPI and indus-

trial production at horizon three gives MSFEs 2.90% and 2.11% higher than using Ridge. Selecting Ridge for unemployment at horizons one and two raises MSFE by 2.02% and 2.42% relative to EW. These losses show why agreement with the later ranking and the magnitude of forecast errors are distinct measures of performance.

With the fitted slopes and penalties unchanged, removing the intercept gives one-month CPI and PCE MSFE ratios of .8755 and .8499 relative to EW. The corresponding ratios with the intercept are .8777 and .8513 (Supplementary Table S7).

Table 4: Forecast choices and subsequent MSFE

Panel A: Choices matching the lower MSFE

Lower MSFE	Cases	Rule	Always EW	Always Ridge
EW	8	6	8	0
Ridge	7	5	0	7
Total	15	**11**	8	7

Panel B: Individual choices and MSFE ratios

Target	h	Choice	Lower MSFE	Match	Ridge/EW	Rule/EW
RPI	1	EW	EW	Yes	1.0098	1.0000
RPI	2	EW	EW	Yes	1.0040	1.0000
RPI	3	EW	Ridge	No	0.9718	1.0000
CPIAUCSL	1	Ridge	Ridge	Yes	0.8777	0.8777
CPIAUCSL	2	Ridge	Ridge	Yes	0.9988	0.9988
CPIAUCSL	3	EW	EW	Yes	1.0011	1.0000
PCEPI	1	Ridge	Ridge	Yes	0.8513	0.8513
PCEPI	2	Ridge	Ridge	Yes	0.9938	0.9938
PCEPI	3	EW	EW	Yes	1.0010	1.0000
INDPRO	1	EW	EW	Yes	1.0037	1.0000
INDPRO	2	EW	EW	Yes	1.0140	1.0000
INDPRO	3	EW	Ridge	No	0.9793	1.0000
UNRATE	1	Ridge	EW	No	1.0202	1.0202
UNRATE	2	Ridge	EW	No	1.0242	1.0242
UNRATE	3	Ridge	Ridge	Yes	0.9812	0.9812

Notes: Choices use only losses observed before the first January 2015 target forecast. Lower MSFE refers to the EW–Ridge comparison over January 2015–December 2019; there are no ties. "Match" indicates agreement with that ranking. Both ratios use EW as the denominator. All 15 cases share one evaluation period.

7. Conclusion

The decision to retain EW or use estimated weights depends on the forecasting procedure as well as the signal in forecast differences. The risk comparison separates the available improvement from the part recovered after estimation. Penalty selection and component estimation can change that comparison.

Past out-of-sample losses provide evidence for this choice. Under the stated conditions, the choice converges to the lower-MSFE procedure when the two mean MSFEs differ, while the selected forecast's expected squared error approaches the smaller of the two MSFEs. The estimation windows remain fixed as the number of past evaluations increases.

The macroeconomic results show useful choices of EW for some targets and horizons and of Ridge for others. A decision rule gains by retaining an accurate simple average or adopting a useful adjustment. Its forecasting value depends on the size of the improvements it captures and the losses from mistaken choices.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, ChatGPT was used to assist with language polishing. The authors reviewed and edited the content as needed and take full responsibility for the content of the submitted article.

References

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468.

Blanc, S. M. and Setzer, T. (2016). When to choose the simple average in forecast combination. *Journal of Business Research*, 69(10):3951–3962.

Blanc, S. M. and Setzer, T. (2020). Bias–variance trade-off and shrinkage of weights in forecast combination. *Management Science*, 66(12):5720–5737.

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

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Diebold, F. X. and Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, 6(4):503–508.

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

Elliott, G. and Liao, J. (2026). Combining forecasts—on why averaging beats optimal linear weights. *Journal of Business & Economic Statistics*. Advance online publication.

Frazier, D. T., Covey, R., Martin, G. M., and Poskitt, D. (2023). Solving the forecast combination puzzle. *arXiv preprint arXiv:2308.05263*.

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

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Inoue, A. and Kilian, L. (2006). On the selection of forecasting models. *Journal of Econo-*

metrics, 130(2):273–306.

Kourentzes, N., Barrow, D., and Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, 209:226–235.

Lee, S. and Lee, T.-H. (2026). Improving the simple average combined forecast via factor-adjusted regularization. *International Journal of Forecasting*. In press.

Li, H., Zhang, N., and Zhou, J. (2025). A new self-normalized forecast comparison test. *Economics Letters*, 256:112646.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Qian, W., Rolling, C. A., Cheng, G., and Yang, Y. (2022). Combining forecasts for universally optimal performance. *International Journal of Forecasting*, 38(1):193–208.

Shao, X. (2010a). Corrigendum: A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):695–696.

Shao, X. (2010b). A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):343–366.

Smith, J. and 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. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Tibshirani, R. J. and Rosset, S. (2019). Excess optimism: How biased is the apparent error of an estimator tuned by SURE? *Journal of the American Statistical Association*,

114(526):697–712.

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, 1:135–196.

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547.

Yang, Y. (2004). Combining forecasting procedures: Some theoretical results. *Econometric Theory*, 20(1):176–222.