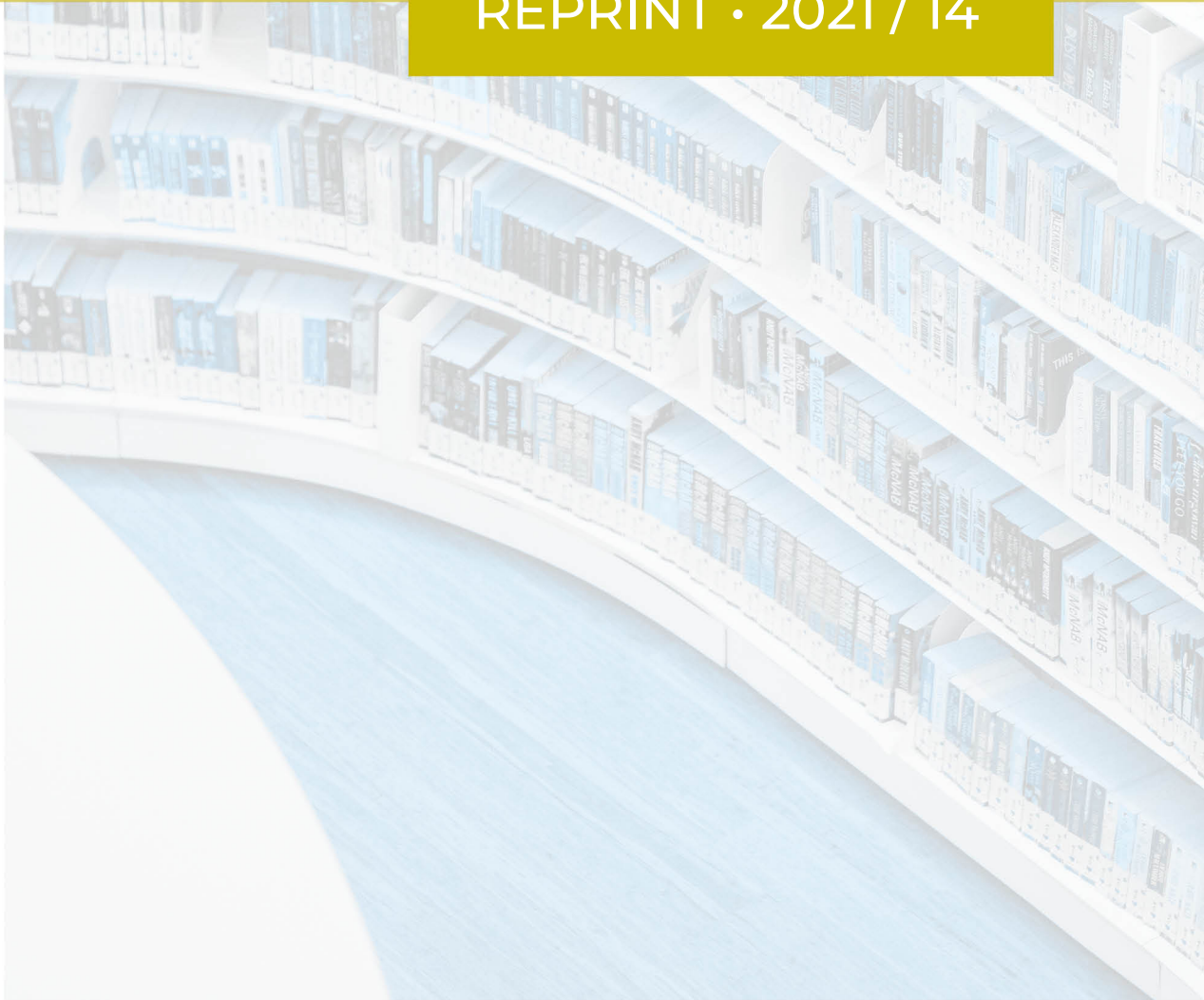


OPTIMAL AND ROBUST COMBINATION OF FORECASTS VIA CONSTRAINED OPTIMIZATION AND SHRINKAGE

Francesco Roccuzzella, Paolo Gambetti,
Frédéric Vrms

REPRINT · 2021 / 14



LFIN

Voie du Roman Pays 34, L1.03.01 B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: lidam-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/lfin/publications.html>



Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Optimal and robust combination of forecasts via constrained optimization and shrinkage

Francesco Roccazzella^{*}, Paolo Gambetti, Frédéric Vrins

LIDAM - LFIN, UCLouvain, Louvain-la-Neuve, B-1348, Belgium

ARTICLE INFO

Keywords:

Forecast combination
Model selection
Machine learning
Shrinkage
Robust methods

ABSTRACT

We introduce various methods that combine forecasts using constrained optimization with penalty. A non-negativity constraint is imposed on the weights, and several penalties are considered, taking the form of a divergence from a reference combination scheme. In contrast with most of the existing approaches, our framework performs forecast selection and combination in one step, allowing for *potentially* sparse combining schemes. Moreover, by exploiting the analogy between forecasts combination and portfolio optimization, we provide the analytical expression of the optimal penalty strength when penalizing with the L2-divergence from the equally-weighted scheme. An extensive simulation study and two empirical applications allow us to investigate the impact of the divergence function, the reference scheme, and the non-negativity constraint on the predictive performance. Our results suggest that the proposed models outperform those considered in previous studies.

© 2021 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

1. Introduction

Often, m different models can be used to produce a set of forecasts $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_m)^T$ of a given random variable y . In such cases, two strategies can be considered to obtain an aggregated forecast $\hat{y}$: the best prediction can be chosen (i.e., $\hat{y} = \hat{y}_{i^*}$, where model $i^* \in \{1, \dots, m\}$ is selected according to some criterion), or the forecasts can be combined into a single prediction (i.e., $\hat{y} = \Phi(\hat{\mathbf{y}})$, for some function $\Phi: \mathbb{R}^m \rightarrow \mathbb{R}$).¹

The first strategy usually provides disappointing results, and it is now widely recognized that combining models is preferable (Atiya, 2020; Granger & Ramanathan,

1984; Taylor, 2020). Most combination strategies include two stages. The first stage consists of removing poorly performing models via trimming based on rules of thumb (e.g., retaining a fixed percentage of models as in Aiolfi & Favero, 2005), formal tests (Amendola et al., 2020; Samuels & Sekkel, 2017), measures of interest (Granger & Jeon, 2004), or machine learning methods (Diebold & Shin, 2019) so that only k models out of m candidates are kept. Then, the selected candidates are combined using a combination scheme, usually in a linear way (i.e., $\Phi(\hat{\mathbf{y}}) = \mathbf{w}^T \hat{\mathbf{y}}$ where $\mathbf{w}$ is an m -by-1 vector of combining weights).

When the individual forecasts are unbiased, the most standard approach consists of minimizing a loss function featuring the aggregated forecast error under the constraint that the weights sum to 1.² For instance, the minimum-variance combination scheme $\mathbf{w}^*$ is obtained

^{*} Correspondence to: Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium.

E-mail addresses: francesco.roccazzella@uclouvain.be (F. Roccazzella), paolo.gambetti@uclouvain.be (P. Gambetti), frederic.vrins@uclouvain.be (F. Vrins).

¹ Throughout the paper, we use the terms “forecast”, “prediction” and “model” as synonyms. We denote matrices as bold capital letters, while vectors and scalars are represented by bold lowercase letters and lowercase letters, respectively.

² We have assumed individual forecasts to be unbiased throughout the paper, but this assumption can be easily relaxed. If we reformulate the optimal forecast combination problem in the form of a linear regression problem the bias can be easily corrected by adding the intercept to the model as discussed in Timmermann (2006).

by minimizing $\mathbb{E}\|y - w^T \hat{y}\|_2^2$ under the constraint $w^T \mathbf{1} = 1$ or, equivalently (Granger & Ramanathan, 1984):

$$w^* := \underset{w \in \mathcal{W}}{\operatorname{argmin}} w^T S w \quad \text{where} \quad \mathcal{W} := \{w \in \mathbb{R}^m \mid \mathbf{1}^T w = 1\}, \quad (1)$$

where S is the population covariance matrix of models' prediction error, $v_i := y - \hat{y}_i$. The solution to (1) is

$$w^* = \Psi(S) \quad \text{where} \quad \Psi(M) := [\mathbf{1}^T M^{-1} \mathbf{1}]^{-1} M^{-1} \mathbf{1}.$$

In practice, S is typically unknown and must be estimated from the data. Given a sample of n observations, the quality of the optimal scheme depends to a large extent on the estimator of S . One possibility is to consider the sample counterpart of S , i.e., $\hat{S} := \frac{1}{n} V^T V$, where V is the $n \times m$ matrix of realized prediction errors. However, despite being consistent, $\hat{S}$ breaks down in high dimensional frameworks ($n \ll m$), and can be imprecise due to sample size limitations or the presence of considerable background noise. Hence, the estimated w^* will be unstable and the corresponding predicting strategy will perform poorly out of sample (Smith & Wallis, 2009).

Previous literature attempted to tackle this issue using two distinct approaches. The first approach is to rely on regularization (or shrinkage) techniques such as constrained optimization with penalty (COP), with shrinkage towards equal weights (Diebold & Pauly, 1990; Diebold & Shin, 2019).³ The second route is to impose some constraints on the combining weights, such as restricting them to be non-negative (Conflitti et al., 2015). In this respect, it is worth noting that this field shares many similarities with that of portfolio optimization in finance. The popular investment strategy called the *minimum-variance portfolio* (Markowitz, 1952) precisely corresponds to (1) and its performance suffers from the same shortcomings regarding the estimator $\hat{S}$ (DeMiguel et al., 2007; Michaud, 1989).⁴ Shrinkage techniques on the weights (DeMiguel et al., 2009) or no-shortsales constraints corresponding to the non-negativity restriction (Jagannathan & Ma, 2003) are similarly applied in that domain. Interestingly, however, the portfolio optimization literature also brings further solutions. For instance, it has been proposed to *combine* the weights resulting from advanced (i.e., asymptotically unbiased but potentially high-variance) and naive (i.e., biased but low-variance, such as the equally weighted) schemes (Tu & Zhou, 2011). Another proposed method is instead to alleviate the noise impacting $\hat{S}$ by using the *shrinkage of the covariance matrix*, a convex combination of the sample covariance matrix $\hat{S}$ and a highly structured covariance matrix $\bar{S}$ (Ledoit & Wolf, 2004b).⁵ These methods have never been

explored in a forecast combination context. In particular, existing combination schemes still strongly rely on the noisy estimator $\hat{S}$ of S (Claeskens et al., 2016).

More generally, the literature on forecasts combination still suffers from several limitations. First, (i) by featuring a model selection step, existing methods implicitly assume the best forecasting strategy to be a *sparse* combination of the candidate forecasts and/or they still rely on the *a priori* identification of the forecasts to combine. The decision regarding how to trim the set of the model is generally arbitrary and can deteriorate the quality of the combination (Winkler & Makridakis, 1983).⁶ Second, (ii) COP schemes and the non-negativity constraint have been explored independently so far; it is unclear under what conditions combining the two might be beneficial. Third, (iii) existing COP methods feature L1 or L2 penalty to shrink the solution towards a reference scheme but other divergence measures (possibly tailored to deal with non-negative weights) remain unexplored. Finally, (iv) methods that shrink the solution towards a reference scheme often rely on cross-validation to estimate the penalty strength. However, cross-validation can be harmful in a small sample, and its properties are not trivial when there is auto-correlation in forecast errors.

In this paper, we address the above gaps as follows. Regarding (i), we design a robust single-step framework that allows the obtaining of *potentially sparse* combining weights. This contrasts with the standard two-step procedures which *a priori* assume the combination of forecasts to be sparse (Aiolfi & Favero, 2005; Diebold & Shin, 2019; Samuels & Sekkel, 2017). Regarding (ii), we propose for the first time to combine the non-negativity constraint with COP methods leading to COP⁺ schemes. This restriction is shown to blend forecast selection and weights estimation in a single step and helps mitigating estimation uncertainty (in line with the observations made in the literature on portfolio theory). We also present and discuss the scenarios where dropping the non-negativity restriction can be beneficial. We find that this occurs only under specific circumstances: when a candidate's prediction errors are strongly positively correlated to those of the other forecasts and when, at the same time, its predictive ability is also significantly inferior. Moreover, we contribute to (iii) by introducing new divergence measures in COP (namely, elastic net and cross-entropy) and study their performance. We find promising results for cross-entropy (only valid when working under the non-negativity constraint), which consistently outperforms the standard L1 penalization. Eventually, (iv)

³ Diebold and Shin (2019) extend the standard COP method using a two-step procedure: having selected a subset of forecasts using LASSO, the weights of the selected forecasts are estimated by penalizing the L1 or L2 norm of the difference between the estimated and equal weights.

⁴ In that context, the vector of weights w represents the proportion of wealth assigned to each asset, S represents the population asset returns' covariance matrix and $\hat{S}$ is its sample estimator.

⁵ There are deep connections between these approaches, especially when minimum variance is the objective function. For instance, $\bar{S} = I$

amounts to penalize the Euclidean distance between the weights and the equally-weighted scheme (DeMiguel et al., 2009). Similarly, setting an upper limit on the L2-norm corresponds to add a penalty to prevent the weights to be too far from the equally-weighted scheme (DeMiguel et al., 2009). Alternatively, imposing a constraint on the L1-norm on the weights amounts to add a no short-sales constraint (DeMiguel et al., 2009). This latter also corresponds to reducing the large elements of the sample covariance matrix (Jagannathan & Ma, 2003).

⁶ For instance, forecasts tend to be highly correlated, and statistical procedures such as LASSO often fail to correctly select the best subset of forecasts (Zhao & Yu, 2006).

we use the correspondence between the shrinkage estimators of the sample covariance matrix (Ledoit & Wolf, 2004b) and the L2-norm penalty (DeMiguel et al., 2009) to obtain an analytical expression for the optimal penalty strength, thereby making the economy of a potentially harmful cross-validation step in this case. In an extensive simulation study and two empirical applications, we find that these “explicit COP models” outperform cross-validation based COP methods, the simple average forecast, trimming methods, and other standard benchmarks from previous studies. We document the flexibility and robustness of the single-step combinations with explicit penalty strength, methods called *optimal and robust combinations of forecasts* hereafter.

The remainder of this paper is structured as follows. Section 2 describes the proposed optimal and robust combination of forecasts. Section 3 includes the simulation study. Section 4 is dedicated to empirical applications. Section 5 concludes the paper.

2. The one-step optimal and robust combination of forecasts

We now study the features of the proposed optimal and robust combination of forecasts.

2.1. The optimal combination of forecasts with non-negative weights

The proposed methods aim to blend forecast selection and weights estimation in a single step. Let us briefly explain how and why constraining the weights to be non-negative and to sum to 1 implicitly selects which forecasts to combine. Intuitively, if a candidate forecast j has a higher covariance with other strategies than other candidates, the optimization procedure will reduce w_j accordingly and eventually set it to 0 if the marginal contribution of $\hat{y}_j$ to the variance of the aggregate prediction error is always larger than the other forecasts' marginal contributions. Furthermore, imposing non-negativity and portfolio constraints also reduces the estimation error by inducing a shrinkage-like effect on the sample covariance matrix (Jagannathan & Ma, 2003).

We consider the following constrained minimization problem (the covariance matrix is always assumed to be positive definite in population in the sequel):

$$\mathbf{w}^* := \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}^+} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} \quad \text{where} \quad \mathcal{W}^+ := \{\mathbf{w} \in \mathcal{W} \mid \min \mathbf{w} \geq 0\}. \quad (2)$$

This expression is similar to (1) except that the optimization set is now $\mathcal{W}^+$ (the set of non-negative weights summing to one) and that the (unknown) covariance matrix $\mathbf{S}$ has been replaced by its sample counterpart, $\boldsymbol{\Sigma}$. The optimal solution is known analytically and can be retrieved by imposing the Karush-Kuhn-Tucker (KKT) conditions (Jagannathan & Ma, 2003):

Proposition 1. *The solution to the optimization problem (2) is given by*

$$\mathbf{w}^* := \Psi^+(\boldsymbol{\Sigma}) \quad \text{where} \quad \Psi^+(\mathbf{M}) := \frac{\mathbf{M}^{-1} \boldsymbol{\eta}(\mathbf{M})}{\boldsymbol{\eta}^T(\mathbf{M}) \mathbf{M}^{-1} \boldsymbol{\eta}(\mathbf{M})} \quad (3)$$

and $\boldsymbol{\eta}(\mathbf{M})$ is a conformable column vector whose j th element is defined as follows:

$$\eta_j(\mathbf{M}) := \begin{cases} 0 & \text{if } \left[\frac{\mathbf{M}^{-1} \mathbf{1}}{\mathbf{1}^T \mathbf{M}^{-1} \mathbf{1}} \right]_j \leq 0 \\ 1 & \text{otherwise} \end{cases}$$

Conflitti et al. (2015) notice that Proposition 1 is a special case of LASSO regression since the non-negativity restriction and the sum to one constraint imply that the weight vector belongs to the unit simplex, and therefore weights have a unitary L1-norm.

2.2. Constrained optimization with penalty (COP) on the weighting scheme

We introduce the COP combination scheme as the solution to an optimization problem on $\mathcal{W}^+$. Similarly to (2), we directly consider the sample covariance matrix of prediction errors in the objective function. The minimum-variance objective function is associated with a penalty term, whose purpose is to shrink the solution towards a predetermined reference scheme $\bar{\mathbf{w}} \in \mathcal{W}^+$ using some divergence measure. To study the impact of the non-negativity constraint, we define COP^+ and COP as follows.

Definition 1 (Constrained Optimization with Penalty - COP). The COP weighting scheme with non-negativity constraint, COP^+ , is defined as

$$\mathbf{w}_\delta^* := \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}^+} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} + \delta d(\mathbf{w}, \bar{\mathbf{w}}), \quad (4)$$

where $\delta \geq 0$ is the *penalty strength* and $d(\cdot, \cdot) := \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}^+$, $(\mathbf{u}, \mathbf{v}) \mapsto d(\mathbf{u}, \mathbf{v})$ is a *divergence measure* between two vectors.

COP takes the same form as (4) but where $\mathcal{W}^+$ is replaced by $\mathcal{W}$.

The coefficient δ controls the impact of the penalty, whereas $d(\mathbf{w}, \bar{\mathbf{w}})$ quantifies the “distance” between $\mathbf{w}$ and the reference scheme $\bar{\mathbf{w}}$. The divergences that will be considered in our simulation study and the empirical applications are:

- elastic net (EN): $d(\mathbf{u}, \mathbf{v}) = a \|\mathbf{u} - \mathbf{v}\|_1 + (1 - a) \|\mathbf{u} - \mathbf{v}\|_2^2$ where $a \in [0, 1]$,
- cross-entropy (CE): $d(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^m u_j \ln \left(\frac{u_j}{v_j} \right)$ where $\mathbf{u}, \mathbf{v} \in \mathcal{W}^+$.

Notice that the L1 and L2 distances are special cases of EN with $a = 1$ and $a = 0$, respectively, and do not require to impose non-negativity, in contrast with CE.

2.3. The shrinkage direction of combining weights

Potentially, any vector satisfying $\bar{\mathbf{w}} \in \mathcal{W}^+$ can be taken as $\bar{\mathbf{w}}$ in (4); for example, $\bar{\mathbf{w}}$ can represent a combination strategy that the forecaster believes to be suitable for a particular predictive problem or can express some prior belief about the exclusion of a model from the

set of candidate forecasts. Nevertheless, in practice, we often prefer a fixed reference point to alternatives that require the estimation of individual losses, as the latter can be subject to estimation error. In particular, the simple average $\bar{\mathbf{w}}^E := \mathbf{1}/m$ seems a natural reference scheme. Assigning equal weights often performs quite well for both finite samples (Smith & Wallis, 2009) and the population (Aruoba et al., 2012). Moreover, the equally-weighted scheme is rather intuitive: shrinking the weights towards $\bar{\mathbf{w}}^E$ amounts to penalize for the L2-norm of the weights, since the constraint $\|\mathbf{w} - \bar{\mathbf{w}}^E\|_2^2$ coincides with $\|\mathbf{w}\|_2^2$ on $\mathcal{W}$. Finally, this choice can also offer some computational advantages, as it often simplifies the optimization problem. For example, the COP estimators of the combining weights are known in closed-form (see Appendix A for a simple proof):

Proposition 2. *In the special case where $\bar{\mathbf{w}} = \bar{\mathbf{w}}^E$ and $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2^2$, the COP⁺ and COP weighting schemes in Definition 1 are respectively given by:*

$$\mathbf{w}_\delta^* = \Psi^+(\Sigma_\delta) \text{ or } \Psi(\Sigma_\delta) \quad \text{where } \Sigma_\delta := \Sigma + \delta \mathbf{I}, \quad (5)$$

with $\mathbf{I}$ the m -by- m identity matrix.

Nevertheless, shrinking the combining weights towards equality can significantly damage the predictive performance if the individual prediction error variances truly differ among candidate forecasts. In this case, considering the inverse-loss weighted average as a reference, i.e., $\bar{\mathbf{w}} = \bar{\mathbf{w}}^L$ where $\bar{\mathbf{w}}^L := [\sigma_i^2 \sum_{i=1}^m 1/\sigma_i^2]^{-1}$, $\sigma_i^2 := \Sigma_{i,i}$, is a more appealing alternative to equal weights. However, the former is data-driven and, therefore, is subjected to estimation error. There is no clear way to balance misspecification and estimation error in the reference scheme. Hence, in the empirical section, we test whether restricting ourselves to equal weights offers a significant disadvantage compared to the inverse-loss weighted average or whether the misspecification introduced in the former is balanced by an increased sensitivity to estimation error in the latter.

2.4. Robust combination methods and the shrinkage estimator of the covariance matrix

In the previous section, we derived a forecast combination scheme that makes use of constrained optimization with penalty to estimate the optimal set of weights, but other methods can be adopted. In particular, we can consider shrinkage estimators of Σ . This approach has been developed in a series of papers by Ledoit and Wolf (2003, 2004a, 2004b, 2012, 2017), who introduced linear and nonlinear shrinkage techniques to reduce the instability of the estimated covariance matrix in a portfolio optimization context. As recalled in the introduction, this consists of combining the sample covariance matrix (which can be easily computed and is asymptotically unbiased but prone to estimation error) with a highly structured estimator (that contains relatively little estimation error but tends to be misspecified and biased). Transposing this

technique to our forecast combination problem leads to the following combination scheme:

Definition 2 (Constrained Optimization with Shrinkage (COS) of the Covariance Matrix). The COS weighting scheme with non-negativity constraint, COS⁺, is defined as

$$\mathbf{w}_\lambda^* := \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}^+} \mathbf{w}^T \Sigma_\lambda \mathbf{w} \quad \text{where } \Sigma_\lambda := (1 - \lambda) \Sigma + \lambda \bar{\Sigma}, \quad (6)$$

$\lambda \in [0, 1]$ is the shrinkage intensity and $\bar{\Sigma}$ is the reference covariance matrix. COS takes the same form as (6) but where $\mathcal{W}^+$ is replaced by $\mathcal{W}$.

This problem takes the same form as (2), hence, the solution to COS⁺ and COS are simply $\mathbf{w}_\lambda^* = \Psi^+(\Sigma_\lambda)$ and $\Psi(\Sigma_\lambda)$, respectively.

As explained above, shrinkage towards equal weights is natural and standard. Moreover, Timmermann (2006) shows that this combining scheme is optimal when the covariance matrix of the prediction errors satisfies $\Sigma_{i,i} = \sigma^2$ and $\Sigma_{i,j \neq i} = \rho \sigma^2$. Thus, shrinking towards equal weights can indirectly be achieved by shrinking the sample covariance matrix towards a special reference matrix, namely, the covariance matrix that would be implied if equal weights indeed represented the optimal combination strategy. Without loss of generality (see Appendix B),

$$\bar{\Sigma}^E := \bar{\sigma}^2 \mathbf{I}, \quad \text{where } \bar{\sigma}^2 := \frac{1}{m} \sum_{i=1}^m \sigma_i^2. \quad (7)$$

We can extend this framework to consider other reference matrices $\bar{\Sigma}$ that are closely related to alternative reference weighting schemes. For example, we can indirectly obtain the inverse-loss weighted average as reference weights by shrinking the sample covariance matrix towards a specific reference $\bar{\Sigma}^L$. This requires choosing

$$\bar{\Sigma}_{i,i}^L := \sigma_i^2 \quad \text{and} \quad \bar{\Sigma}_{i,j}^L := 0, \quad \forall i \neq j. \quad (8)$$

2.5. A closer look at COP and COS methods

Despite following different approaches, the optimal weighting schemes in COP and COS are closely related. In particular, in the field of portfolio optimization, DeMiguel et al. (2009) show that there is a correspondence between COS and the portfolio that solves the minimum-variance problem, subject to the additional constraint that the norm of the portfolio-weights vector is smaller than a given threshold. A similar relation holds in the context of constrained optimization with penalty.

Proposition 3. *The COS model (6) with reference covariance matrix $\bar{\Sigma} = \bar{\Sigma}^E$ given in (7) corresponds to the COP model (4) with $\bar{\mathbf{w}} = \bar{\mathbf{w}}^E$, $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2^2$ and $\delta = \frac{\lambda}{1-\lambda} \bar{\sigma}^2$.*

The proof of Proposition 3 is trivial. Provided that $\lambda > 0$,⁷ Σ_λ can be rescaled by $(1 - \lambda)$ to obtain $\tilde{\Sigma}_\lambda := \frac{\Sigma_\lambda}{1 - \lambda}$

⁷ The $\lambda = 0$ case corresponds to $\delta = 0$, i.e., no shrinkage.

$= \Sigma + \frac{\lambda}{1-\lambda} \bar{\sigma}^2 \mathbf{I}$. Considering $\tilde{\Sigma}_\lambda$ instead of Σ_λ does not impact the COS weighting scheme 2 in the sense that $\mathbf{w}_\lambda = \Psi(\Sigma_\lambda) = \Psi(\tilde{\Sigma}_\lambda)$; similar expressions hold for COS^+ . The equivalence between COP and COS schemes simply results from the fact that in this specific case, $\tilde{\Sigma}_\lambda$ takes the same form as Σ_δ in (5) for the considered expression of δ . The equivalence holds exactly $\forall \lambda \in [0, 1)$. The equivalence also holds in the limiting case since both weighting schemes converge to equal weights, i.e.,

$$\delta \xrightarrow{\lambda \rightarrow 1} +\infty \quad \text{and} \quad \mathbf{w}_\delta \xrightarrow{\lambda \rightarrow 1} \bar{\mathbf{w}}^E = \mathbf{w}_\lambda. \quad (9)$$

Proposition 3 has an important implication for the choice of the penalty strength δ . In fact, expressing $\delta = g(\lambda)$ extends the use of an estimated shrinkage intensity in the COS with reference $\bar{\Sigma}^E$ to the COP (with L2 divergence and with equal weights as reference scheme), and vice versa.

2.6. Estimation of penalty strength

The estimate of $\mathbf{w}_\delta^*$ in **Definition 1** depends on the chosen divergence measure $d(\cdot, \cdot)$, the reference scheme $\bar{\mathbf{w}}$, and the penalty strength δ . One option for selecting δ is to rely on standard k -fold cross-validation, but more advanced methods such as the forward-looking cross-validation technique used in **Diebold and Shin (2019)**, could be considered as well. Cross-validation can mitigate the consequences of a subjective choice of $d(\cdot, \cdot)$ and $\bar{\mathbf{w}}$. Once δ has been found, we simply solve (4) by imposing the KKT conditions.

However, cross-validation forces us to choose how many observations to retain for validation, which implicitly reduces the information set that can be used for the estimation of Σ , thereby potentially amplifying estimation error. When working with small samples, few additional observations can make the difference between singularity and more stable estimates of Σ . In this regard, **Varoquaux (2018)** shows that k -fold cross-validation can strongly underestimate the variance of the prediction errors in small samples, causing forecasters to “over-shrink” towards the reference weighting scheme. Hence, a closed-form expression for the penalty strength would be highly desirable.

We provide this contribution by relying on the COP-COS equivalence mentioned in **Proposition 3**. Following **(Ledoit & Wolf, 2004b)**, we estimate the optimal shrinkage intensity λ^* by minimizing the expected value of the Frobenius norm of the difference between the shrinkage estimator of the covariance matrix and the true covariance matrix of prediction errors $\mathbf{S}$, i.e.,

$$\lambda^* := \underset{\lambda \in [0, 1]}{\operatorname{argmin}} \mathbb{E} \left\| (1-\lambda) \Sigma + \lambda \bar{\Sigma} - \mathbf{S} \right\|_2^2. \quad (10)$$

with solution **(Ledoit & Wolf, 2003)**:

$$\lambda^* = \frac{\pi - k}{\gamma}, \quad (11)$$

$$\pi = \sum_{i=1}^m \sum_{j=1}^m \operatorname{Var}(\Sigma_{ij}),$$

$$k = \sum_{i=1}^m \sum_{j=1}^m \operatorname{Cov}(\Sigma_{ij}, \bar{\Sigma}_{ij}),$$

$$\gamma = \sum_{i=1}^m \sum_{j=1}^m \operatorname{Var}(\Sigma_{ij}) + (\bar{\Sigma}_{ij} - \Sigma_{ij})^2.$$

By denoting the asymptotic variance and the asymptotic covariance as AsyVar and AsyCov , respectively⁸ and using the assumptions of i.i.d. data and finite fourth moments, is a consistent estimator of λ^* **(Ledoit & Wolf, 2003)**:

$$\hat{\lambda}^* = \frac{1}{n} \frac{\hat{\pi} - \hat{k}}{\hat{\gamma}}, \quad (12)$$

$$\hat{\pi} = \sum_{i=1}^m \sum_{j=1}^m \operatorname{AsyVar}(\sqrt{n} \Sigma_{ij}),$$

$$\hat{k} = \sum_{i=1}^m \sum_{j=1}^m \operatorname{AsyCov}(\sqrt{n} \Sigma_{ij}, \sqrt{n} \bar{\Sigma}_{ij}),$$

$$\hat{\gamma} = \sum_{i=1}^m \sum_{j=1}^m (\bar{\Sigma}_{ij} - \Sigma_{ij})^2.$$

Clearly, the optimal shrinkage intensity λ^* depends on the reference $\bar{\Sigma}$ that we have selected. Indeed, while the $\hat{\pi}$ and $\hat{k}$ measure respectively the instability of the elements of the sample covariance matrix⁹ and the covariances between the elements of the reference and the sample covariance matrices,¹⁰ $\hat{\gamma}$ measures the degree of misspecification of $\bar{\Sigma}$. For example, if we impose the reference matrix in (7), $\bar{\Sigma} = \bar{\Sigma}^E$, then a completely inappropriate choice for the reference $\bar{\sigma}$ can easily neutralize the COS by coaxing $\lambda^* = 0$. However, by setting $\bar{\sigma}$ as in (7), we avoid running into this problem since they assume *a priori* that the considered predictive strategies have the same performance (i.e., the same variance of the prediction error) and that possible differences are due to chance. We note $\hat{\delta}^* := \frac{\lambda^*}{1-\lambda^*} \bar{\sigma}^2$ the penalty strength in **Proposition 3** associated with $\lambda = \hat{\lambda}^*$.¹¹

2.7. Negative or non-negative weights

Forecast selection and the reduction in the estimation error come at the cost of misspecification if the constraint $\mathbf{w} \in \mathcal{W}^+$ is violated in population. Imposing non-negative weights in portfolio selection may not look

⁸ We refer to Appendix D for the explicit formula to compute $\hat{\lambda}^*$. The consistent estimators of $\hat{\pi}$, $\hat{k}$ and $\hat{\gamma}$ are derived using the linear shrinkage estimator of the covariance matrix described in **Ledoit and Wolf (2003, 2004a, 2004b)** and adapting the reference matrix and the notation to our framework.

⁹ The sum of the asymptotic variances of the elements of Σ scaled by $\sqrt{n}$.

¹⁰ The sum of asymptotic covariances of the entries of reference matrix with the entries of the sample covariance matrix scaled by $\sqrt{n}$. Notice that restricting the elements of Σ to be constant and/or independent of Σ implies that $\hat{k} = 0$. This is the case when we shrink towards equal weights.

¹¹ By contrast, it does not depend on the optimization problem underlying the combination scheme nor on constraints that could be imposed on the latter.

like a good idea as, perhaps, short and long positions could be the appropriate solution to the portfolio optimization problem at hand (Green & Hollifield, 1992). Nevertheless, as underlined in Section 1, imposing a non-negativity constraint on the weights might still be beneficial because it rules out extreme solutions that essentially result from estimation errors: $\mathbf{w} \in \mathcal{W}^+$ implies $0 \leq w_i \leq 1$ and, when dealing with the minimum-variance objective function, coincides with an $L1$ -constraint on the weights; see, e.g., DeMiguel et al. (2009) and Jagannathan and Ma (2003).¹²

In forecasting, negative weights would arise in very specific circumstances (Bunn, 1985; Smith & Wallis, 2009). In particular, this depends on the difference in the variance of prediction errors between models, on the correlation across candidate forecasts, and the number of models with superior performance.

For example, consider a set of 20 forecasts where a subset of them significantly under-performs the others. Table 1 displays the minimum difference in RMSFE necessary to attribute an optimal weight of -0.05 , -0.1 , -0.2 and -0.5 to a candidate forecast for a fixed level of correlation across prediction errors.¹³ The presence of negative weights seems to be beneficial the more a candidate's prediction errors are positively correlated to the ones of the other forecasts. Nonetheless, this requires inferior (and sometimes considerably worse) performance compared to the best forecasts.

Nevertheless, negative weights generally suggest the existence of a noise forecast, i.e., a prediction that individually does not help to predict the target, but whose inclusion in the combination of forecasts is only justified by the objective of minimizing the *in-sample* variance of prediction errors. It is well known that the inclusion of such forecasts can negatively impact the stability of the combining weights (Bunn, 1981; Dickinson, 1973). Imposing the non-negativity constraint acknowledges this fact directly in the estimation step: this differs from the use of trimming methods in a pre-screening phase. We further discuss the properties of negative and non-negative weights for COP schemes in the simulation study (Section 3) and the empirical applications (Section 4).

3. Monte Carlo simulations

We study the behavior of the combinations of forecasts based on constrained optimization with penalty in a controlled environment. Specifically, we analyze their sensitivity to various specifications of the variance-covariance matrix of prediction errors and to the choice of the penalty

function $d(\mathbf{w}, \bar{\mathbf{w}})$. We also examine the impact on the performance of the non-negative restriction $\mathbf{w} \in \mathcal{W}^+$ compared to $\mathbf{w} \in \mathcal{W}$. Table 2 summarizes the combination strategies under consideration.

In the following sections, we consider several COP methods (Definition 1), namely, COP1, COP2, COPEN, and COPE⁺, which are associated with $L1$, $L2$, elastic net, and cross-entropy divergences, respectively. The considered reference scheme $\bar{\mathbf{w}}$ (equal weights or inverse loss-weights) is indicated by the subscript E or IL. Minimization is performed over $\mathcal{W}$ except for the models with the non-negative restriction, where it is performed on $\mathcal{W}^+$. We indicate these latter with a ⁺ superscript.¹⁴ For example, COP2_IL⁺ refers to the COP scheme with $L2$ penalty, reference scheme $\bar{\mathbf{w}}^{IL}$ and $\mathbf{w} \in \mathcal{W}^+$, whereas COP2_IL is similar except that $\mathbf{w} \in \mathcal{W}$. Note that in these approaches, the penalty strength δ is chosen using cross-validation.

We also consider two COS methods (Definition 2) that rely on the analytical expression for the shrinkage intensity $\hat{\lambda}^*$, given in (7). For instance, COS_IL⁺ and COS_IL use $\hat{\lambda} = \hat{\lambda}^*$ and shrink towards the reference covariance matrix $\bar{\Sigma}^{IL}$, with and without non-negativity constraint, respectively. When considering $\bar{\Sigma}^E$, Proposition 3 suggests that COS_E is equivalent to COP2_E provided that $\delta = \delta^*$. These equivalent COP-COS models are hence referred to as COPS_E⁺ and COPS_E.

3.1. Simulation designs

We study a set of $m = 20$ candidate forecasts of a target variable $y \sim \mathcal{N}(0, \sigma_y^2)$, where a subset of k forecasts exhibits superior performance. As before, we denote by $\hat{y}_i$ the i th forecast and by $v_i = y - \hat{y}_i$ its prediction error. The vectors of forecasts and prediction errors are noted $\hat{\mathbf{y}}$ and $\mathbf{v}$, respectively. We consider two scenarios for $\mathbf{S}$.

In scenario (a) we form forecasts using $\hat{y}_i = y + v_i$ where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{S})$ is independent of y . To model the fact that k models exhibit superior performance, we postulate that their RMSFEs are, on average, 50% lower than the last $m - k$ ones.¹⁵ Hence, we set $S_{i,i} = \alpha^2$ for $i \in \{1, \dots, k\}$, $S_{i,i} = (2\alpha)^2$ for $i \in \{k+1, \dots, m\}$ and $S_{i,j \neq i} = r\sqrt{S_{i,i}S_{j,j}}$. Notice that, with a 50% difference in performance, the best forecasting strategy can feature negative weights, especially when the correlation across prediction errors is considerable. Hence, we use this scenario to study whether the non-negative restriction significantly penalizes COP schemes.

In scenario (b), we deal with the case where $m - k$ forecasts are “pure noise”. Intuitively, an optimal combination strategy is expected to identify the k superior models while assigning negligible weights to the $m - k$ uninformative forecasts. In this setup, we model the i th forecast as $\hat{y}_i = y + \epsilon_i$ for $i \in \{1, \dots, k\}$ and $\hat{y}_i = \epsilon_i$ for $i \in \{k+1, \dots, m\}$ where $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i^2] = \alpha^2$ and

¹² Note that some authors restrict the weights to belong to some interval $[a, b]$, where the bounds are computed from the data. For example, Behr et al. (2013) show that when the investment universe is large enough, one typically gets $0 \approx a < b \leq 1$.

¹³ Throughout the paper and in this simulation exercise, we assume the individual forecasts to be unbiased. In the case of biased forecasts, the bias can be corrected by reformulating the optimal forecast combination problem in the form of a linear regression problem and adding the intercept to the model as discussed in Timmermann (2006) and indicated in footnote 2. Alternatively, the minimum-variance objective function can be adjusted to account for a bias-variance tradeoff.

¹⁴ Recall that non-negativity is always required when using COPE.

¹⁵ We also considered scenarios in which the k superior models provide RMSFEs that are on average either equal or 10%, 30% and 70% lower than the inferior candidates. The results are comparable and available upon request.

Table 1

Differences in performances (% RMSFE) between the k superior forecasts and $m - n$ inferior forecasts needed to achieve negative weights of different magnitude. These differences in performances are computed for distinct levels of correlation among the candidate forecast errors and for a different number k of superior models.

ρ	Minimum negative weight -5%					ρ	Minimum negative weight -10%				
	k = 1	k = 2	k = 5	k = 10	k = 15		k = 1	k = 2	k = 5	k = 10	k = 15
0.01	>100%	>100%	>100%	>100%	>100%	0.01	>100%	>100%	>100%	>100%	>100%
0.1	>100%	>100%	>100%	>100%	>100%	0.1	>100%	>100%	>100%	>100%	>100%
0.25	>100%	>100%	>100%	>100%	>100%	0.25	>100%	>100%	>100%	>100%	>100%
0.5	>100%	>100%	>100%	27%	16%	0.5	>100%	>100%	>100%	>100%	34%
0.75	>100%	>100%	16%	8%	5%	0.75	>100%	>100%	>100%	13%	8%
0.9	33%	13%	5%	3%	2%	0.9	>100%	>100%	8%	4%	3%

ρ	Minimum negative weight -20%					ρ	Minimum negative weight -50%				
	k = 1	k = 2	k = 5	k = 10	k = 15		k = 1	k = 2	k = 5	k = 10	k = 15
0.01	>100%	>100%	>100%	>100%	>100%	0.01	>100%	>100%	>100%	>100%	>100%
0.1	>100%	>100%	>100%	>100%	>100%	0.1	>100%	>100%	>100%	>100%	>100%
0.25	>100%	>100%	>100%	>100%	>100%	0.25	>100%	>100%	>100%	>100%	>100%
0.5	>100%	>100%	>100%	>100%	>100%	0.5	>100%	>100%	>100%	>100%	>100%
0.75	>100%	>100%	>100%	>100%	16%	0.75	>100%	>100%	>100%	>100%	>100%
0.9	>100%	>100%	>100%	7%	5%	0.9	>100%	>100%	>100%	>100%	>100%

Table 2

List of the considered predictive algorithms and forecast combination methods. Minimization is performed over $\mathcal{W}$ except for the models with the non-negative restriction, where it is performed on $\mathcal{W}^+$. We indicate these latter with a $^+$ superscript, e.g., CO^+ .

Combination methods	Acronym	Weights
Standard methods		
Simple average forecast	SA	$\bar{w}^E$
Loss-based weighted average forecast	IL	$\bar{w}^{IL}$
Constrained optimization	$\text{CO}^{(+)}$	$\psi^{(+)}(\Sigma)$
partially-egalitarian LASSO	DS_egal_lasso	w^{DS}
Trimmed averages		
Trimmed simple average forecast	T-SA- p	$\mathbf{1}/\text{card}(\mathcal{M}_p)$ if $i \in \mathcal{M}_p$, 0 otherwise.
Trimmed simple average forecast via cross-validation	T-SA-CV	$\mathbf{1}/\text{card}(\mathcal{M}_p)$ if $i \in \mathcal{M}_p$, 0 otherwise.
Trimmed-MCS simple average forecast	T-MCS- α	$\mathbf{1}/\text{card}(\mathcal{M}_\alpha)$ if $i \in \mathcal{M}_\alpha$, 0 otherwise.
Constrained optimization with shrinkage direction equal weights		
COP with L1 penalty	$\text{COP1_E}^{(+)}$	$\arg\min w^T \Sigma w + \delta \ w - \bar{w}^E\ _1$
COP with L2 penalty	$\text{COP2_E}^{(+)}$	$\arg\min w^T \Sigma w + \delta \ w - \bar{w}^E\ _2^2$
COP with EN penalty	$\text{COPEN_E}^{(+)}$	$\arg\min w^T \Sigma w + \delta \left[a \ w - \bar{w}^E\ _1 + (1-a) \ w - \bar{w}^E\ _2^2 \right]$
COP with EN penalty and shrinkage towards 0	COPEN	$\arg\min w^T \Sigma w + \delta \left[a \ w\ _1 + (1-a) \ w\ _2^2 \right]$
COP with cross-entropy	COPE_E^+	$\arg\min w^T \Sigma w + \delta \sum_{j=1}^m w_j \ln \left(\frac{w_j}{1/m} \right)$
COS with reference $\bar{\Sigma}^E \equiv \text{COP with L2 penalty and } \hat{\delta}^*$	$\text{COPS_E}^{(+)}$	$\psi^{(+)}(\Sigma_{\hat{\delta}^*}^E) = \psi^{(+)}(\Sigma_{\hat{\lambda}^*}^E)$
Constrained optimization with shrinkage direction inverse loss-weights		
COP with L1 penalty	$\text{COP1_IL}^{(+)}$	$\arg\min w^T \Sigma w + \delta \ w - \bar{w}^{IL}\ _1$
COP with L2 penalty	$\text{COP2_IL}^{(+)}$	$\arg\min w^T \Sigma w + \delta \ w - \bar{w}^{IL}\ _2^2$
COP with EN penalty	$\text{COPEN_IL}^{(+)}$	$\arg\min w^T \Sigma w + \delta \left[a \ w - \bar{w}^{IL}\ _1 + (1-a) \ w - \bar{w}^{IL}\ _2^2 \right]$
COP with cross-entropy	COPE_IL^+	$\arg\min w^T \Sigma w + \delta \sum_{j=1}^m w_j \ln \left(w_j / \bar{w}_j^{IL} \right)$
COS with reference $\bar{\Sigma}^{IL}$ and $\hat{\lambda}^*$	$\text{COS_IL}^{(+)}$	$\psi^{(+)}(\Sigma_{\hat{\lambda}^*}^{IL})$
We denote with the superscript + the CO, COP1, COP2 and COPEN schemes with non-negative restriction, $w \in \mathcal{W}^+$		

Notes. We denote the cardinality of the set $\mathcal{A}$ as $\text{card}(\mathcal{A})$. For example, $\mathcal{M}_p$ is the subset of forecasts containing $m - p$ candidates, such that $\text{card}(\mathcal{M}_p) = m - p$. We denote the trimmed simple average forecast as T-SA- p once the worst p models have been excluded. We consider the possibility of estimating the number of models to discard via cross validation in T-SA-CV. $\mathcal{M}_\alpha$ is the subset of forecasts that are included in the superior set of models with confidence level α (Hansen et al., 2011). We denote the simple average forecast derived from the candidate forecasts included in $\mathcal{M}_\alpha$ as T-MCS- α . Hence, T-MCS- α represents the simple average forecast that includes only the candidates having at least a probability α of belonging to the superior set of models. For the elastic net divergence measure, we set $a = 0.8$. Following (Diebold & Shin, 2019), the partially-egalitarian LASSO consists of selecting the subset of forecasts to combine in the first step, and shrinking the combining weights of the selected candidates (w^{DS}) towards equal weights (using L1 divergence measure). We estimate the penalties via 5-fold cross validation both in the first and second step.

$\mathbb{E}[\epsilon_i \epsilon_{j \neq i}] = r\alpha^2$. In this case, the population covariance matrix of prediction errors takes the form

$$S_{i,i} = \alpha^2 + \sigma_y^2 \mathbb{1}_{\{i>k\}}, \quad S_{i,j \neq i} = r\alpha^2 + \sigma_y^2 \mathbb{1}_{\{\min(i,j)>k\}}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Notice that σ_y^2 and α control the signal to noise ratio (SNR) in this simulation exercise. We set $\sigma_y^2 = \alpha = 1$, which is representative of a scenario where the SNR = 1.

In both scenarios (a) and (b), we consider four *designs*, based on the number of superior forecasts k . In *design 1*, only the first two candidate forecasts have superior predictive performances ($k = 2$) for *scenario (a)* or must be identified for *scenario (b)*; *design 2*, *design 3* and *design 4* deal with $k = 5$, $k = 10$ and $k = 15$, respectively.

We compare the considered combination schemes (Table 2) with the optimal forecasting strategy w^* . In particular, we compute the relative distance between the root mean square forecast error (RMSFE) of the strategy under consideration and that of the benchmark w^* .¹⁶ Specifically, for the i th combination scheme, we compute the performance criterion

$$\mu_i = \frac{\text{RMSFE}_i}{\text{RMSFE}_{w^*}} - 1. \quad (13)$$

We simulate $n = 1030$ observations (30 for training and 1000 for testing) for 500 replications. The difference in size between the training and testing sets should not be surprising. While having 1000 observations in the testing sample provides a rather accurate measurement of predictive performance, having 30 observations in the training sample is representative of a realistic scenario where, relative to the number of candidate forecasts, few observations are retained for estimating combining weights, and most of the data are used to train the individual candidate forecasts.¹⁷

3.2. Simulation results

We now discuss how combination methods perform across the considered scenarios. Since results are comparable across designs, Fig. 1 only includes the results of *design 1* and *design 4* while we refer to Appendix E for *design 2* and *design 3*. For each design, the left-hand side (LHS) panel refers to *scenario (a)* (superior models yield RMSFEs that are 50% lower than the inferior candidates) and the right-hand side panel (RHS) to *scenario (b)* (the best forecasting strategy is sparse). We refer to Appendix F for an additional proof of the consistency of COS methods in discarding the pure noise forecasts.

We sort the combination schemes (y-axis) based on their average performances across the four correlation levels, which are indicated using colors (yellow $r = 0$, orange $r = .25$, red $r = .5$ and black $r = .9$). The closer the combination strategy is to the origin of the axes, the closer its average performance to the optimal forecasting strategy w^* .

Fig. 1 points out that the performances depend on the correlation level r of prediction errors and on whether the best forecasting strategy is sparse or not. Indeed, as anticipated in Section 2.7, negative weights are likely to

arise when $r > .90$, which penalizes the strategies that restrict combining weights to be non-negative.

In *scenario (a)* (LHS panel), we confirm this fact: non-negative COP methods outperform their negative counterparts when $r \leq .50$, but underperform the latter when $r = .90$. Such an impact can be so considerable to flip over the overall ranking. For example, in *design 1*, COPS_E⁺ and COS_IL⁺ perform the best when $r \leq 0.5$, but rank only fourth and sixth overall due to the poor performance in high correlation case. This can also hold for unpenalized methods: in *design 3*, while CO (i.e., the combination that only minimizes the in sample variance of prediction errors) generally performs poorly, the benefits of dropping the non-negative constraint when $r = .90$ is so considerable to counteract the superior performance of CO⁺ (same as CO but with non-negativity constraint) in the other cases. We can generalize this finding to COP methods but not to COS: the non-negativity constraint does not penalize methods that rely on a closed-form expression for the shrinkage intensity. Both in *design 1* and *design 4*, restricting the weights to be non-negative offers an advantage: COPS_E⁺ and COS_IL⁺ outplay their counterparts COPS_E and COS_IL.

The non-negative restriction on the combining weights offers a significant advantage when w^* is sparse. In *scenario (b)*, (RHS panel) COP methods with non-negative weights always outperform their negative counterparts, both when the penalty strength is estimated via cross-validation and via the analytical formula. In this case, the correlation level is no longer the main driver of the variability of the performance across combination strategies. Only T-MCSs remain particularly sensitive to the correlation among prediction errors.

Fig. 1 also underlines the generally poor performance of trimming methods: both T-MCSs and T-SAs rank among the worst strategies together with IL and SA. Furthermore, estimating the number of forecasts to discard using cross-validation (T-SA-CV) does not offer any significant advantage. Only T-SA-5 performs remarkably well in the *design 4* of *scenario (b)* but the reason is clear: conditional on identifying the right candidates, T-SA-5 coincides with the best forecasting strategy w^* for that design.

The partially-egalitarian LASSO procedure (DS_egal_lasso) is competitive with COP methods only when $r = .90$ and when w^* is not sparse. We also confirm the *forecast combination puzzle*: CO performs poorly, especially when $r \leq .50$, where it performs worse than the simple average forecast (SA), IL, and T-SAs.

This simulation study sheds the light on four aspects of great practical relevance for the use of combination strategies in forecasting.

First, all COP and COS methods perform remarkably well, especially compared to trimming-based combination strategies in all considered designs. In particular, it is difficult to identify a unique trimming method that is optimal under various specifications of S . In general, however, forecast selection is inappropriate when correlations among the prediction errors of candidates are particularly low ($r = 0$). Conversely, COP methods limit the impact of subjective choices on the number of models to discard.

¹⁶ Note that while w^* is implied by S , the considered forecast combination methods only rely on the sample counterpart of S , i.e., $\hat{S} := \frac{1}{n} V^T V$.

¹⁷ T-MCS methods that use bootstrapping to estimate the MCS rely on 1000 replications of the training set. Methods that use a validation fold to estimate the penalty strength retain 5 observations from the training set and rely on the remaining 25 observations for estimating the covariance matrix of prediction errors.

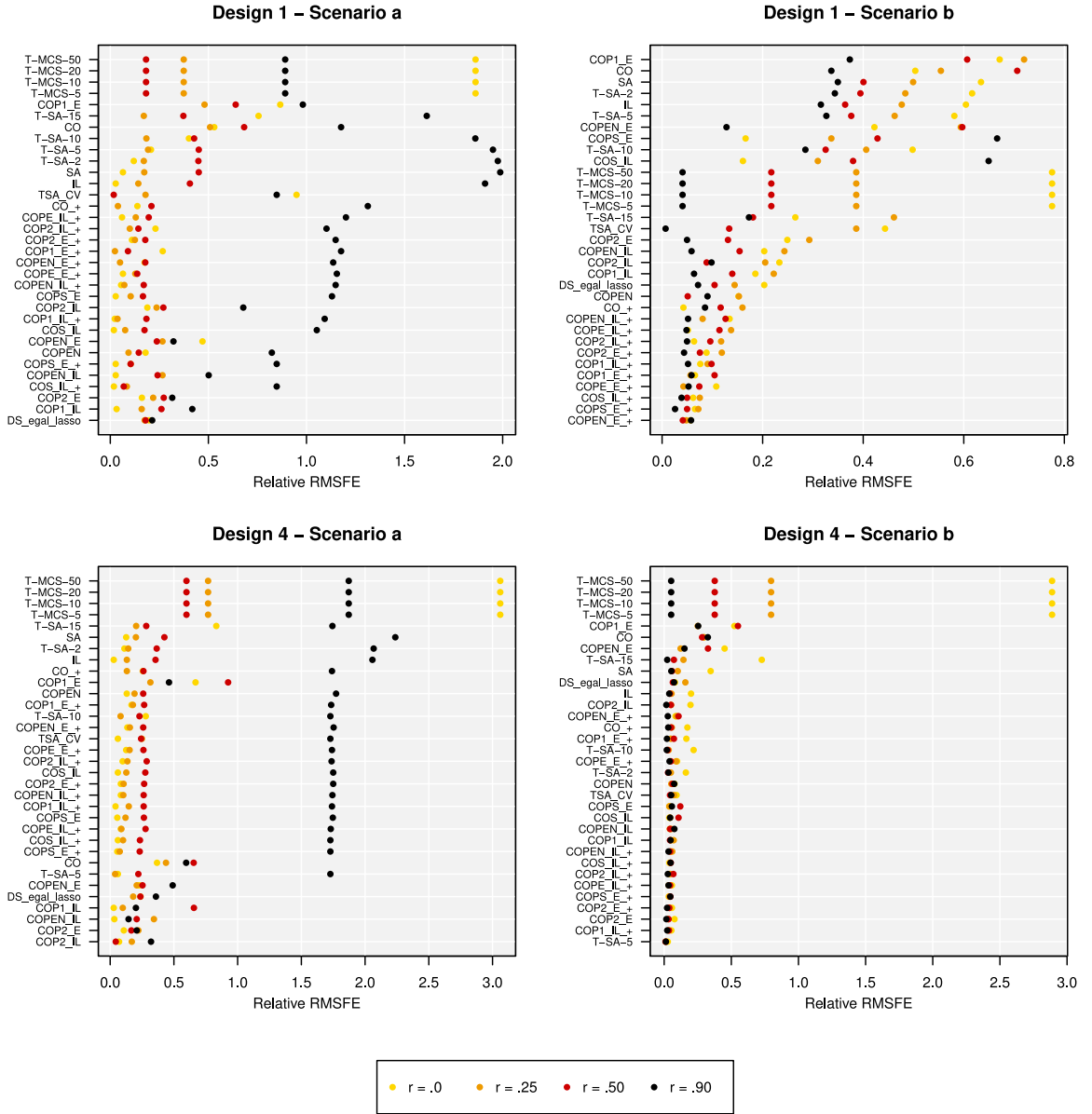


Fig. 1. Relative RMSFE with respect to the true optimal forecast combination w^* . In the left-hand side panel, superior models yield RMSFEs that are 50% lower than the inferior candidates (*scenario a*) and in the right-hand side panel the best forecasting strategy is sparse (*scenario b*). Combination schemes are ordered according to their average performances across correlation levels. In *design 1*, the first two candidates have superior predictive performance; in *design 4* the first fifteen candidates have superior performance. Right panels show the outstanding performance of the non-negativity constraint in scenario (b). The same holds true for scenario (a), but only for moderate correlation levels.

Second, the choice of the penalty function (L1, L2, EN, and cross-entropy) does not remarkably affect predictive performance when $w \in \mathcal{W}^+$. Conversely, if w is unrestricted, the L1 divergence measure may negatively affect performance depending on the shrinkage direction ($\bar{w}_E$ vs $\bar{w}_L$) and the DGP under consideration.

Third, the non-negative restriction penalizes performance only when forecast errors are strongly correlated across candidate forecasts. Otherwise, in the other cases,

it significantly improves predictive ability. We do not find any relevant difference between using equal weights or inverse loss-weights as reference scheme unless L1 is involved.

Fourth, $COPS_E^+$ and COS_IL^+ yield the minimum model risk relative to the considered combination strategies. Compared to $COP1$, $COP2$, $COPEN$, and $COPE^+$, the analytical expressions of the shrinkage intensity in $COPS_E$ and COS_IL allows us to make full use of the training

set to estimate the covariance of prediction errors. This is a crucial advantage for more precisely estimating the combining weights.

4. Empirical study

In the empirical study, we consider two prediction problems taken from economics. In these applications, we train the individual algorithms using 70% of the data, build forecasts combinations on 10% and keep the remaining 20% for testing.¹⁸ In the following sections, we present the evaluation criteria used to compare the considered predictive strategies and the predictive algorithms that are used to form the candidate forecasts.

4.1. Evaluation criteria

We must define some evaluation criteria for comparing the predictive strategies. For this purpose, we consider the RMSFE¹⁹ and two statistical procedures: the Diebold–Mariano test (hereafter DM, Diebold & Mariano, 1995) and the model confidence set (hereafter MCS, Hansen et al., 2011).

The DM tests whether one model is superior to another for a given confidence level. Formally, let the loss difference between forecast i and j for observation t be $d_{i,j}^t = (v_i^t)^2 - (v_j^t)^2$. The DM tests whether $H_0 : DM = \frac{\mathbb{E}[d_{i,j}]}{\sigma[d_{i,j}]} = 0$, with $DM \sim \mathcal{N}(0, 1)$. However, the original test can be oversized in small samples and not robust to the heavy-tailed distributions of the forecast errors. To tackle these problems, we consider a scaled DM statistic $\sim$ standard t-student(df), as proposed by Harvey et al. (1997).

The MCS tests whether a subset of methods enters jointly in the superior set of models by repeatedly testing the null hypothesis of equal predictive performance with significance level α . Let $\mathcal{M}_0$ be the set of all forecasting models (both individual candidates and forecasts combinations), and let $\mathcal{M}^*$ be the superior set of models. Formally, the MCS tests $H_0 : \mathbb{E}[d_{i,j}] = 0, \forall i, j$. If the null hypothesis is rejected, then the procedure eliminates the model with the greatest relative loss from the set $\mathcal{M}$. This procedure is sequentially repeated until the null hypothesis is not rejected at the chosen probability level α . We compute the MCS p-values via bootstrapping (10000 replications) and using the Oxford MFE Toolbox publicly available at <https://www.kevinshppard.com/code/matlab/mfe-toolbox/>.

¹⁸ This choice implies that only 21 observations are retained for constructing forecast combinations when predicting the Chicago Fed National Activity Index, and 51 observations are retained for predicting Boston housing prices. The results are robust to a large estimation window, and these additional analyses are available upon request.

¹⁹ We transform the target variable to have unitary variance given the information available in the training set. Therefore, RMSFE and RMSPE (RMS-Percentage-FE) display comparable results unless the variance of the y variable significantly differs in the testing set. We present both the RMSFE and RMSPE in Appendix G.

4.2. Individual predictive algorithms

We consider individual forecasts from 13 standard predictive algorithms that are publicly available on software such as MATLAB. Table 3 lists the predictive algorithms used to form candidate forecasts.

We estimate the hyperparameters of all the considered models via 10-fold cross-validation. Ensemble methods such as gradient boosted regression tree (GBRT), random forest (RF) and feed-forward neural network (NN) methods consist of 1000, 1000 and 5 base learners, respectively. This choice is an appealing compromise between computational time and out-of-sample results. Furthermore, datasets are standardized (unitary variance and zero-mean), and predictors are cross-sectionally normalized to be within $[-1, 1]$. These procedures are standard in the field of machine learning; see Gu et al. (2020) for a financial case. Combination methods that rely on validation to estimate the penalty strength use 5-fold cross-validation.

4.3. Empirical study I: Boston housing prices

In the first application, Boston housing prices are predicted using 14 attributes (e.g., per capita crime rate, average number of rooms per dwelling, or % lower status of the population). This dataset consists of 507 observations, and it is freely available at <http://lib.stat.cmu.edu/datasets/boston>. Although it was originally studied to provide quantitative estimates of the willingness to pay for air quality (Harrison & Rubinfeld, 1978), it is commonly used in the literature to benchmark predictive algorithms (Ben-Porat & Tennenholtz, 2017; Breiman & Friedman, 1985; Quinlan, 1993).

Fig. 2 presents RMSFE for different combination strategies classified in two-step combinations, cross-validation-based COP methods and one step combining strategies. In this latter group, we also include the performance of the *ex-post* best candidate forecasting. The vertical dotted line marks the threshold to beat before joining the superior set of models at 10% significance level. We refer the DM statistics to Appendix G.

In Fig. 2, we notice that despite using standard algorithms, our results are competitive with those of other predictive exercises on the same dataset (average RMSFE ≈ 36 ,²⁰ see <https://www.kaggle.com/c/boston-housing/leaderboard>). Nevertheless, the benefit of combining forecasts, especially via COP and COS methods, remains clear. Indeed, COPS_E, COS_IL (both with and without the non-negative weights), COP2_E⁺, COP2_IL⁺ and COPE_IL⁺ rank as the best forecasting strategies. COP2_IL, COPE_IL⁺ and the two-step partially-egalitarian LASSO of Diebold and Shin (2019) stand on the edge of the superior set of models.

²⁰ In our application the data are standardized; therefore, to compare our results to the leaderboard, we compute the average of the mean square forecast error of the 20 best models in the leaderboard, which gives ≈ 10.88 . Then, we compute $\sqrt{10.88/84.58} = 0.3586 \approx 0.36$, where 84.58 corresponds roughly to the variance of the target variable in the data set.

Table 3

List of the considered predictive algorithms.

Predictive model	Abbreviation	MATLAB function
Generalized linear models		
Ordinary least squares	OLS	fitlm
LASSO	L1	lasso
Ridge	L2	lasso
Elastic net	EN	lasso
Linear SVM with L1 regularization	svml1	fitrlinear
Linear SVM with L2 regularization	svml2	fitrlinear
Partial least squares	PLS	plsregress
Regression trees		
Gradient boosted trees	GBRT	fitrensemble
Random forests	RF	fitrensemble
Feed-forward neural networks (# of neurons per layer)		
1 Hidden layer (8)	NN1	fitnet
1 Hidden layer (16)	NN2	fitnet
2 Hidden layers (8 - 4)	NN3	fitnet
2 Hidden layers (16 - 8)	NN4	fitnet

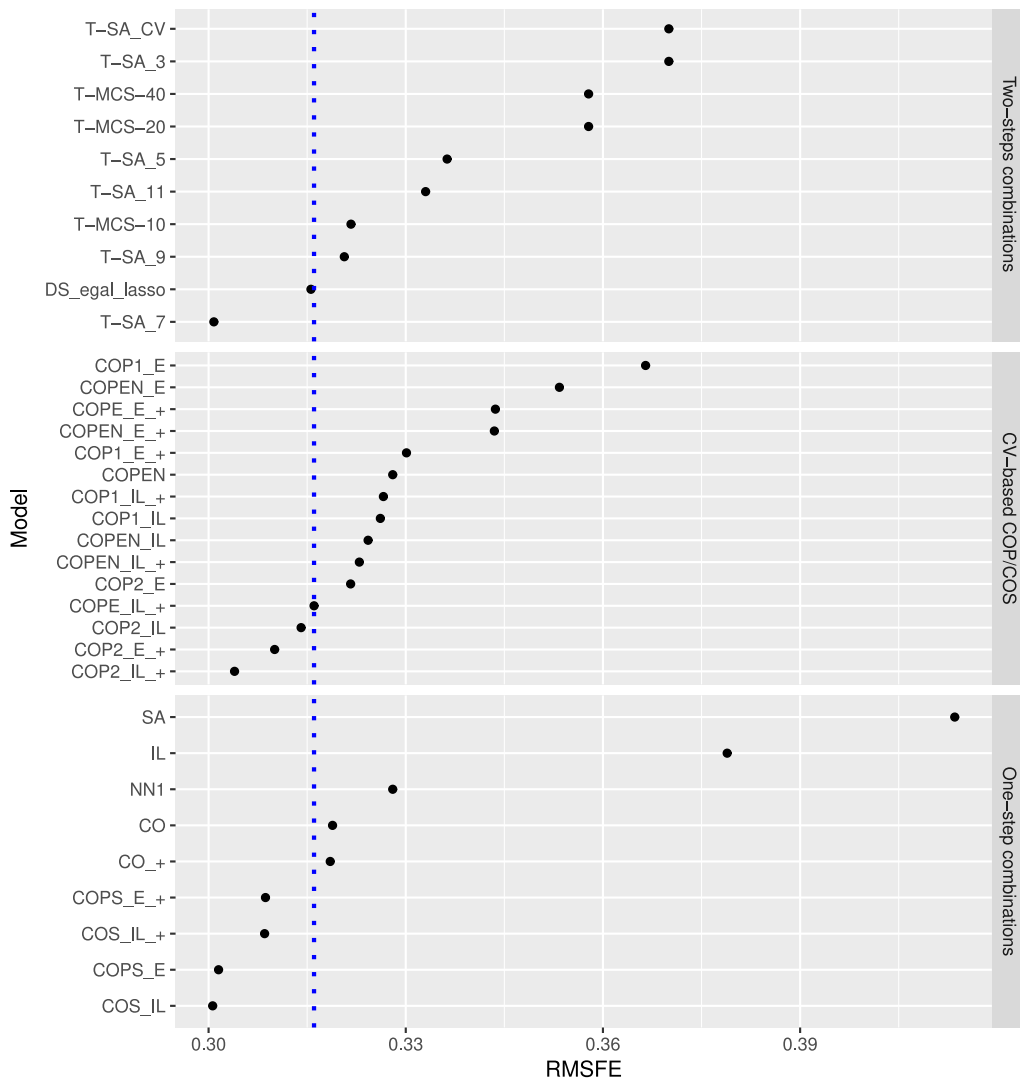


Fig. 2. Predicting Boston housing prices, out of sample results. Performance of the forecasts combination schemes grouped in two-step combinations, cross-validation based COP methods, and one-step combining strategies. The vertical dotted line marks the threshold to overcome to join the superior set of models at 10% significance level.

Table 4

Predicting Boston housing prices, combination weights.

	OLS	L1	L2	svmL1	svmL2	EN	PLS	RF	GBRT	NN1	NN2	NN3	NN4
DS_egal_lasso	.	.	.	.	.	.	.	.25	.25	.25	.25	.	.
CO	.90	.77	(.223)	(.23)	(.23)	.02	1.00	.40	.15	.27	.27	.15	(.23)
COPEN	.08	.09	.	.	.	.	.27	.18	.16	.09	.08	.12	(.06)
COP1_E	.08	.08	.05	(.06)	.08	.08	.08	.08	.22	.09	.08	.08	.08
COP2_E	.08	.08	.08	(.05)	(.05)	(.05)	.08	.08	.20	.20	.20	.08	.08
COPEN_E	.38	.71	(1.31)	(.54)	(.29)	.03	1.	.42	.17	.10	.15	.18	.01
COPS_E	.14	.15	(.04)	(.27)	(.26)	(.05)	.35	.25	.20	.16	.16	.19	.02
COP1_IL	.03	.03	.03	.03	.03	.03	.03	.09	.09	.15	.15	.15	.15
COP2_IL	.09	.09	(.04)	(.04)	(.04)	(.04)	.09	.09	.21	.21	.21	.09	.09
COPEN_IL	.05	.08	.02	.05	(.36)	(.30)	(.02)	.50	.26	.19	.14	.15	.13
COS_IL	.12	.13	(.03)	(.22)	(.21)	(.05)	.27	.23	.20	.17	.20	.22	(.04)
CO ⁺	.03	.	.	.04	.	.	.03	.18	.36	.14	.	.11	.11
COP1_E ⁺	.07	.07	.	.	.	.07	.	.08	.20	.14	.14	.08	.14
COP2_E ⁺	.08	.	.	.	.	.	.	.14	.21	.27	.15	.15	.
COPE_E ⁺	.06	.06	.06	.03	.03	.03	.03	.09	.21	.15	.09	.09	.09
COPEN_E ⁺	.04	.04	.04	.04	.04	.04	.04	.10	.15	.13	.13	.10	.10
COPS_E ⁺	.02	.03	.	.	.	.	.02	.15	.25	.31	.	.11	.11
COP1_IL ⁺	.05	.05	.05	.02	.02	.02	.02	.08	.14	.14	.14	.14	.14
COP2_IL ⁺	.04	.	.	.	.	.	.	.13	.19	.19	.19	.19	.07
COPE_IL ⁺	.03	.03	.03	.03	.03	.03	.03	.09	.12	.15	.15	.15	.15
COPEN_IL ⁺	.02	.02	.02	.02	.02	.02	.02	.14	.14	.14	.14	.14	.14
COS_IL ⁺ ($r = 0$)	.02	.03	.	.	.	.	.02	.16	.25	.31	.	.11	.11

Notes. We denote with · candidates that have been discarded and with (·) negative weights. Weights do sum to 1 before rounding, differences are due to the approximation to two decimal digits to enhance readability.

Among COP and COS schemes, the L2 penalty offers the best predictive performance, while L1 performs the worst. This result does not depend on the reference combination strategy. Thus, shrinking towards equal weights or inverse loss-weights does not impact performance. However, this has one exception: COPE_E⁺ performs significantly worse than COPE_IL⁺.

Concerning T-SAs, the difference in predictive performance strongly depends on how many candidate forecasts we discard in the trimming phase. The findings of Aiolfi and Favero (2005) and Granger and Jeon (2004) are confirmed: among trimmed simple average strategies, those discarding most of the candidate forecasts perform best. Nevertheless, we cannot identify a clear pattern regarding the optimal number of models to discard. For example, while T-SA-7 ranks among the best performers and it is included with at least 10% in the superior set of models, T-SA-3,5,9 and 11 perform significantly worse and have less than 10% probability of being included. T-MCSs perform poorly and strongly depend on the choice of the significance level in the model confidence set. Overall, the choice of the number of candidate forecasts to consider in T-SAs and T-MCSs is subjective. Conversely, COP methods do not require any preliminary decision about whether to trim and by how much: they only rely on the estimated covariance matrix of prediction errors to perform selection and combination in a single step.

Table 4 reports the combining weights and confirms that the weights are constrained to be non-negative for forecast selection. Yet, the use of different divergence measures for COP implies different optimal combination schemes. For example, the “sharp” edges of the L1 penalty tends to force some of the weights to be exactly equal to 1/k. Conversely, L2 better tolerates small deviations from the reference and increases weights variability, which in this case results in better predictive performance.

We also notice that, among COP methods that rely on cross-validation, only COP1_E⁺, COP2_E⁺ and COP2_IL⁺ provide sparse solutions: the non-negative restriction plays a crucial role in this. Also notice, that while COPS_E⁺ simply discards the candidates that display negative weights in COPS_E, we cannot make the same analogy for cross-validation-based methods. In the latter case, the performance in the validation fold depends on the estimated weights that depend on the non-negativity restriction. Conversely, COPS_E⁺, COPS_E and COS_IL⁺, COS_IL have respectively identical penalty strengths, since the estimation of $\hat{\lambda}^*$ is independent of the estimation of the weights.

However, although the weights can differ between COP methods, they largely agree on the relevance of the candidate forecasts: linear models are either discarded or play a marginal role in the aggregate forecast. In contrast, NNs, GBRT, and RF are always included and are the main drivers of the aggregate forecast. We also confirm the particular behavior of the DS_egal_lasso: it selects a small subset of the candidates and it sets their weights to equality.

4.4. Empirical study II: Chicago fed national activity index

The second application aims to predict the Chicago Fed National Activity Index (CFNAI), a monthly index measuring the current economic activity in the United States. The index, publicly available at <https://fred.stlouisfed.org/series/CFNAI>, is a weighted average of 85 monthly indicators of national economic activity (e.g., production and income, employment, personal consumption, housing, sales, and inventories). The CFNAI is constructed to have an average value of zero and a standard deviation of one. Zero values indicate that the US economy is expanding at its historical trend rate of growth, negative

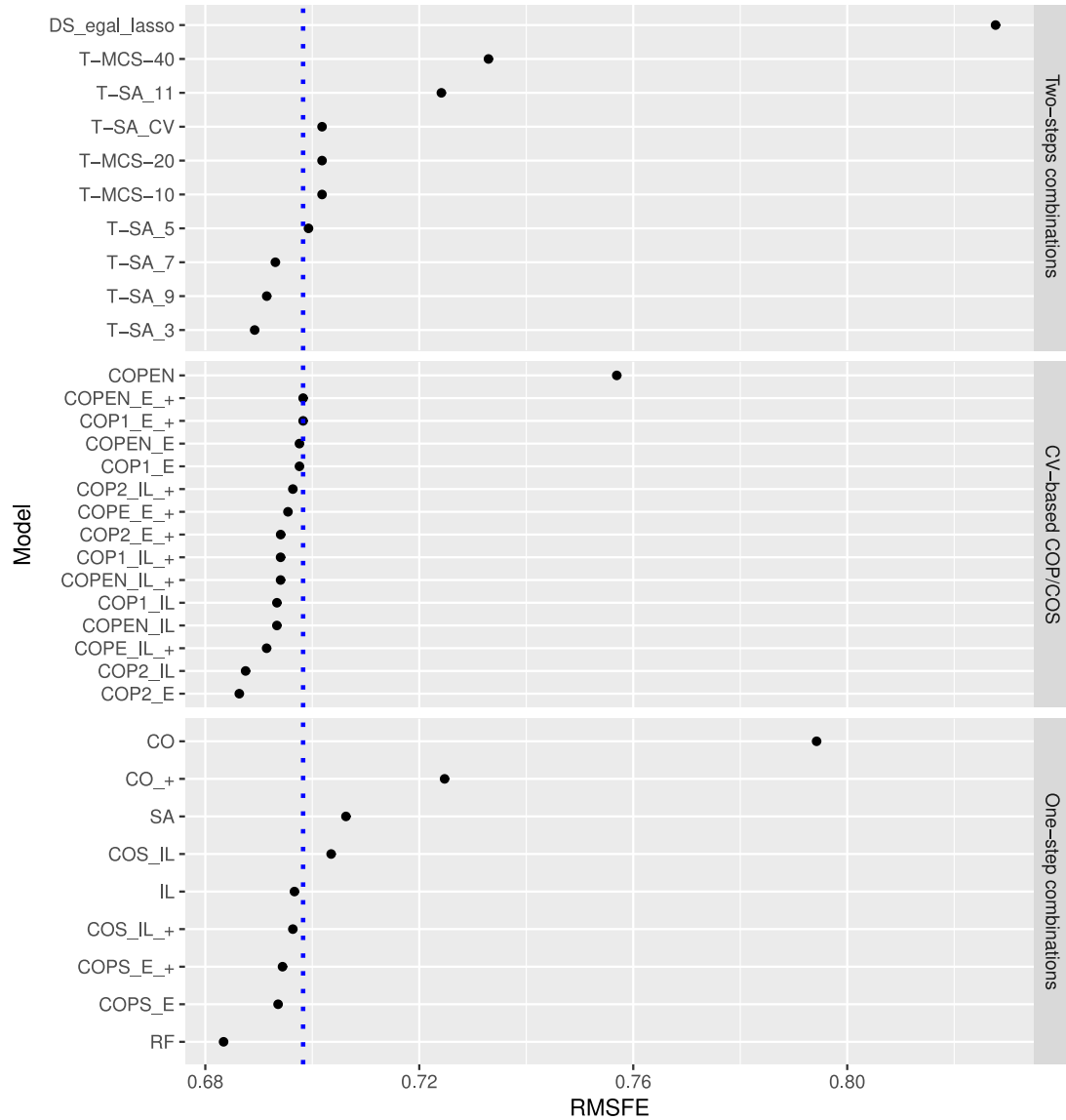


Fig. 3. Predicting Chicago Fed National Activity Index (CFNAI), out of sample results. Performance of the forecasts combination schemes grouped across two-step combinations, cross-validation based COP methods and one step combining strategies. The vertical dotted line marks the threshold to overcome to join the superior set of models at 10% significance level.

values indicate below-average growth, and positive values indicate above-average growth.

We predict 1-step ahead values of CFNAI using a large dataset of backward-looking and forward-looking variables. We retrieve the backward-looking variables from the monthly dataset of the Federal Reserve Bank of St. Louis (FRED-MD) and transform the original series according to the work of [McCracken and Ng \(2016\)](#). We also include measures of macroeconomic, financial and real uncertainty from [Jurado et al., 2015](#)) and [Ludvigson et al. \(2015\)](#). We retrieve the forward-looking variables from the black Chip Financial Forecasts (BCFF) database. This database relies on a survey of professional forecasters that includes the expected term structure of U.S. interest rates, expected real GDP, price level (GDP price deflator and

Consumer Price Index) and the expected level of the major currency index. BCFF also includes the average expected yield for home mortgages, and the yield of Aaa- and Bbb-rated corporate bonds. We use the first three principal components to summarize the information of the term structure of expected yields. We further compute the standard deviation and the conditional asymmetry ([Bowley, 1920](#)) to capture the dispersion and the skewness of the empirical distribution of forecasts, respectively. We count 125 predictive variables. Our sample starts in January 2001 and stops in May 2018 (209 monthly observations). We train the individual models using 146 observations, form a combination of forecasts using 21 observations, and keep the remaining 42 observations for testing. Thus, COPS_E and COS_IL have $21 - 13 = 8$

Table 5

Predicting Chicago Fed National Activity Index (CFNAI), combination weights.

	OLS	L1	L2	svmL1	svmL2	EN	PLS	RF	GBRT	NN1	NN2	NN3	NN4
DS_egal_lasso	.	.	.	.	.	.	.	.	.	.	.	1.	.
CO	(.09)	(.93)	.01	.99	1.	(.46)	.55	.10	(.18)	(.67)	(.02)	1.	(.31)
COPEN	.	.	.	1.	.	.	.	.	.	.	.	.	.
COP1_E	.05	.08	.08	.11	.08	.08	.08	.08	.08	.08	.08	.08	.08
COP2_E	(.03)	.03	.04	.14	.12	.04	.04	.16	.06	.06	.05	.23	.06
COPEN_E	.05	.08	.08	.11	.08	.08	.08	.08	.08	.08	.08	.08	.08
COPS_E	(.08)	(.09)	(.08)	.34	.48	(.20)	.02	.49	(.17)	(.17)	(.17)	.90	(.26)
COP1_IL	.01	.08	.08	.10	.10	.08	.08	.08	.08	.08	.08	.08	.07
COP2_IL	(.04)	.03	.04	.15	.13	.04	.04	.16	.05	.06	.06	.23	.05
COPEN_IL	.01	.08	.08	.10	.10	.08	.08	.08	.08	.08	.08	.08	.07
COS_IL	(.08)	(.10)	(.11)	.38	.60	(.24)	.01	.49	(.19)	(.20)	(.20)	.92	(.28)
CO ⁺	.	.	.	.54	.33	.	.	.	.	.13	.	.	.
COP1_E ⁺	.05	.08	.08	.11	.08	.08	.08	.08	.08	.08	.08	.08	.08
COP1_IL ⁺	.01	.08	.08	.10	.10	.08	.08	.08	.08	.08	.08	.08	.07
COP2_E ⁺	.	.	.	.27	.26	.	.	.24	.	.10	.	.13	.
COP2_IL ⁺	.	.	.	.31	.27	.	.	.21	.	.09	.	.13	.
COPE_E ⁺	.05	.08	.08	.11	.08	.08	.08	.08	.08	.08	.08	.08	.08
COPE_IL ⁺	.01	.08	.08	.10	.10	.08	.08	.08	.08	.08	.08	.08	.07
COPEN_E ⁺	.03	.07	.07	.09	.09	.08	.07	.10	.08	.07	.07	.10	.07
COPEN_IL ⁺	.	.07	.08	.11	.10	.08	.07	.10	.07	.08	.08	.11	.07
COPS_E ⁺	.	.	.	.26	.24	.	.	.25	.	.08	.05	.13	.
COS_IL ⁺ (r = 0)	.	.	.	.30	.27	.	.	.21	.	.09	.	.13	.

Notes. We denote with . candidates that have been discarded and with (.) negative weights. Weights do sum to 1 before rounding, differences are due to the approximation to two decimal digits to enhance readability.

degrees of freedom, while methods that rely on 5-fold cross-validation have only $17 - 13 = 4$ degrees of freedom for computing Σ .

Note that in this predictive exercise, the time dimension can affect the predictive ability of the individual models and, consequently, combination methods. Therefore, as a robustness check, we also consider an expanding window to estimate the individual models, to form the combination of forecasts, and to form evaluation statistics. This means that we estimate the individual models and the combining weights 42 times. The training set, made of the data available up to time t , is partitioned in two folds. The *combination fold* is composed of 21 observations selected randomly from the training set and is used to estimate the combining weights. The other fold is used to train the individual models. Results are robust to this second specification and we refer to Appendix H for the results that build on the expanding window specification.

Fig. 3 presents the obtained RMSFEs; the Diebold–Mariano statistics are provided in Appendix G. The results are in line with the previous findings.

In Fig. 3, we remark that the COP and COS methods significantly outperform our main benchmarks, i.e., SA, CO⁺ (Conflitti et al., 2015) and DS_egal_lasso (Diebold & Shin, 2019). The use of a single step method with penalization is what marks the difference between being excluded (SA, CO, CO⁺, and DS_egal_lasso) and being included in the superior set of models (COP and COS schemes). On the one hand, despite imposing the non-negative constraint on the weights improves the performance (CO⁺ outperforms CO), this restriction alone is not sufficient to be included in the superior set of models (CO⁺ is excluded). On the other hand, the partially-egalitarian LASSO (DS_egal_lasso) actually collapses to NN3 (see Table 5), which however is not the best candidate forecast in the testing set: this underlines how the failure to identify the model(s) in the first step can have

relevant consequences on performance. We can draw similar conclusions on T-SAs and T-MCSs.

T-SAs are competitive with the best forecasting strategies, but it remains difficult to determine the *right* number of models to discard: while TSA_3_9, and TSA_7 are included in the superior set of models, TSA_5 and TSA_11 are excluded. As stressed earlier, a wrong choice for this number can be very harmful: TSA_11 ranks among the worst-performing strategies overall. T-MCS methods are always excluded from the superior set of models. Conversely, COS⁺ schemes are always included in the superior set of models and, at 10% significance level, their performance is not statistically different from the best *ex-post* forecasting strategy, i.e., RF.

Table 5 gives an overview of the estimated weights. We find evidence that COP methods can achieve sparsity; i.e., they combine only a subset of the available candidate forecasts. Furthermore, as in the previous application, COP and COS methods and largely agree on which models to select and on the weight to assign to each of them. Interestingly, we notice that COS_IL, COP2_IL, and COPS_E, COP2_E respectively estimate almost identical combining weights. Furthermore, the non-negativity restriction does not play any role in COP1_E⁺, COP1_IL⁺, and COPEN_IL⁺ because the corresponding weights are positive even when the non-negativity constraint is relaxed.

5. Conclusions

Constrained optimization with penalty and shrinkage techniques have been successfully used in finance, and they are at the heart of asset allocation strategies. In this paper, we point out that these methods also provide an appealing route for forming forecast combinations. In particular, we design a robust single-step framework that allows the obtaining of *potentially sparse* combining

weights. We show that these methods, relying on constrained optimization with penalty (COP) or with shrinkage of the covariance matrix (COS) and the non-negativity constraint on the combining weights, limit the subjectivity in the trimming phase because they do not require any predetermined selection rule or predetermined combination strategy; they rely on the estimated covariance matrix of prediction errors.

We also present and discuss the scenarios where dropping the non-negativity restriction can be beneficial. We find that this occurs only under specific circumstances: when a candidate's prediction errors are strongly positively correlated to those of the other forecasts and when, at the same time, its predictive ability is also significantly inferior. Moreover, we introduce new divergence measures in COP (namely, elastic net and cross-entropy) and study their performance. We find promising results for cross-entropy (only valid when working under the non-negativity constraint), which consistently outperforms the standard L1 penalization. Eventually, we use the correspondence between the shrinkage estimators of the sample covariance matrix (Ledoit & Wolf, 2004b) and the L2-norm penalty (DeMiguel et al., 2009) to obtain an analytical expression for the optimal penalty strength, thereby making the economy of a potentially harmful cross-validation step in this case.

In an extensive simulation study and two empirical applications, we document the flexibility and robustness of the single-step combinations with non-negative weights and explicit penalty strength (or shrinkage intensity). We find that these models outperform cross-validation based COP methods, the simple average forecast, trimming methods and other standard benchmarks from previous studies.

Acknowledgments

Francesco Roccazzella gratefully acknowledges financial support from the Economics School of Louvain. Paolo Gambetti's research is funded by the Fonds de la Recherche Scientifique (F.S.R.-FNRS) under Grant ASP FC23545. The work of Frédéric Vrins is supported by the Fonds de la Recherche Scientifique - (F.S.R.-FNRS) grant CDR J.0037.18 and by the grant ARC 18-23/089 of the Belgian Federal Science Policy Office.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Proof of Proposition 2

Specializing the optimization problem to the Euclidean distance between $\mathbf{w}$ and the equally-weighted scheme yields (1),

$$\mathbf{w}_\delta^* := \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \mathbf{w}^T \Sigma \mathbf{w} + \delta \|\mathbf{w} - \bar{\mathbf{w}}^E\|_2^2.$$

The penalty term is developed by making use of the fact that $\mathbf{w} \in \mathcal{W}^+ \subset \mathcal{W}$:

$$\|\mathbf{w} - \mathbf{1}/m\|_2^2 = \mathbf{w}^T \mathbf{w} - 3/m.$$

By discarding the constant and grouping the quadratic terms,

$$\mathbf{w}_\delta^* := \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}^+} \mathbf{w}^T (\Sigma + \delta \mathbf{I}) \mathbf{w},$$

which takes a similar form as (2). The proof is concluded by applying Proposition 1, replacing Σ by $\Sigma + \delta \mathbf{I}$ (observe that because $\Sigma + \delta \mathbf{I}$ is positive definite, which is always assumed to be true, Σ and $\delta \geq 0$ are positive definite). Notice that this result holds for a weighting scheme $\mathbf{w} \in \mathcal{W}$.

Appendix B. Proof of equivalence: A general case

The general form of the reference covariance matrix associated with the equally weighted scheme can be written as $\bar{\Sigma} = \bar{\sigma}^2((1 - \rho)\mathbf{I} + \rho\mathbf{E})$, where $\mathbf{E}$ is a conformable matrix of ones. It is easy to see that the optimization problem (6) associated with the specific reference matrix above collapses to that associated with $\rho = 0$ under the $\mathbf{w} \in \mathcal{W}^+$ constraint. Indeed, for every $\lambda \in [0, 1]$, the objective function can be decomposed as

$$\begin{aligned} \mathbf{w}^T \Sigma_\lambda \mathbf{w} &= \mathbf{w}^T \left(\Sigma + \frac{\lambda}{1 - \lambda} (1 - \rho) \bar{\sigma}^2 \mathbf{I} \right) \mathbf{w} \\ &\quad + \frac{\lambda}{1 - \lambda} \rho \bar{\sigma}^2 \mathbf{w}^T \mathbf{E} \mathbf{w}. \end{aligned}$$

Due to the restriction $\mathbf{w} \in \mathcal{W}^+$, the last term collapses to the constant $\lambda \rho \bar{\sigma}^2 / (1 - \lambda)$, which can be discarded. Hence, the cost function is that of a COS scheme with $\bar{\Sigma} = \bar{\sigma}^2 \mathbf{I}$ and shrinkage intensity $\frac{\lambda}{1 - \lambda} (1 - \rho)$. Thus, we can set $\rho = 0$ without loss of generality.

Appendix C. A bayesian reformulation of the cop problem

We motivated the use of COP and COS schemes with non-negativity constraint to improve the stability of the combining weights when the sample covariance matrix is poorly estimated or when the number of candidate forecasts is greater than the number of available observations. Nevertheless, we can also motivate the use of shrinkage from a Bayesian perspective. On the one hand, it is a way to incorporate prior information into the estimation of combining weights as suggested by Diebold and Pauly (1990). On the other hand, a Bayesian reformulation of COP helps the forecaster to quantify the uncertainty around the estimated weights and incorporate the doubts about the non-negative restriction on the combining weights.

In this section, we reformulate the COP problem with L1 and L2 divergence measure in a Bayesian framework. We restrict ourselves to consider the reference scheme $\bar{\mathbf{w}}^E$ since Bayesian prior should not depend on the specific sample under analysis. Following Granger and Ramanathan (1984), we recast the optimal forecast combination as a linear regression problem. Denoting by $\mathbf{Y} := [\mathbf{y}, \dots, \mathbf{y}]$, the $n \times m$ matrix containing m copies of $\mathbf{y}$, and

by $\mathbf{V}$ the $n \times m$ matrix containing the realized forecast errors, we can reformulate the CO regression problem as

$$\begin{aligned} \mathbf{w}^* &:= \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} [\mathbf{y} - \hat{\mathbf{Y}}\mathbf{w}]^T [\mathbf{y} - \hat{\mathbf{Y}}\mathbf{w}] \\ &= \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} [\mathbf{y} - (\mathbf{Y} + \mathbf{V})\mathbf{w}]^T [\mathbf{y} - (\mathbf{Y} + \mathbf{V})\mathbf{w}]. \end{aligned}$$

Recalling that $\mathbf{1}^T \mathbf{w} = 1$, we notice that $\mathbf{Y}\mathbf{w} = \mathbf{y}$. Therefore,

$$\mathbf{w}^* := \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} [\mathbf{V}\mathbf{w}]^T [\mathbf{V}\mathbf{w}] = \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}.$$

Finally, we can retrieve COP2 by adding the appropriate penalty term. In Proposition 2, we show that when shrinkage towards equal weights, the solution to COP2_E consists of shrinking diagonal elements of $\boldsymbol{\Sigma}$ by δ . In this case, the posterior mode coincides with the Ridge estimate of $\mathbf{w}$ when reformulating COP2 in the form of a linear regression problem (see for example Hastie et al., 2008). In particular, by using the Gaussian likelihood, the Normal-inverse-Wishart distribution as prior and appropriately rescaling the weights to sum to 1, we obtain the Bayesian reformulation of COP2_E (see Geweke, 2005, for an introduction to Bayesian regression models). Similarly, we can retrieve COP1_E by imposing a Laplace prior (centered on the reference weighting scheme) on the weights. COP1_E can be easily implemented by modifying the Gibbs sampler proposed by Park and Casella (2008) to center the Laplace prior around $\bar{\mathbf{w}}^E$.

Notice that we neither directly impose a truncated Gaussian (COP2_E) or a truncated Laplace (COP1_E) prior on the weights nor truncate the posterior distribution in $[0, 1]$. Indeed, despite we can quantify the uncertainty on the estimated weights also when restricting weights to be non-negative, dropping this assumption permits us to also incorporate the doubts about the non-negative restriction itself. While the underlying hypothesis is that negative weights signal noise forecasts that should be discarded as discussed in Section 2.1, if the posterior indicates that a weight is significantly negative, the Bayesian framework would have helped us identify a case where this restriction is violated in population. The mode, but most importantly the dispersion of the posterior distribution of the weights, should guide the forecaster on whether to include or discard a subset of forecasts. However, dropping the non-negative restriction comes at one cost: COP1_E and COP2_E cannot perform forecast selection and therefore handle high dimensional combination problems.

Appendix D. The analytical formula for the optimal shrinkage intensity

Following Ledoit and Wolf (2003, 2004a, 2004b), we present the analytical formula to estimate the optimal shrinkage intensity for COS methods. This formula for λ^* makes the linear shrinkage estimator of $\boldsymbol{\Sigma}$ asymptotically optimal, i.e., it asymptotically minimizes the Frobenius Norm (quadratic loss function) of the difference between the true and the shrinkage estimator $\hat{\boldsymbol{\Sigma}}_\lambda$ as the number of observations and the number of variables go to infinity together. As indicated in (12), the optimal shrinkage

intensity is

$$\hat{\lambda}^* = \frac{1}{n} \frac{\hat{\pi} - \hat{k}}{\hat{\gamma}},$$

$$\hat{\pi} = \sum_{i=1}^m \sum_{j=1}^m \operatorname{AsyVar}(\sqrt{n} \Sigma_{ij}),$$

$$\hat{k} = \sum_{i=1}^m \sum_{j=1}^m \operatorname{AsyCov}(\sqrt{n} \Sigma_{ij}, \sqrt{n} \bar{\Sigma}_{ij}),$$

$$\hat{\gamma} = \sum_{i=1}^m \sum_{j=1}^m (\bar{\Sigma}_{ij} - \Sigma_{ij})^2.$$

While $\hat{\gamma}$ can be easily computed from (12), the formulas for $\hat{\pi}$ and $\hat{k}$ are less straightforward. Following the work of Ledoit and Wolf, we provide the corresponding formulas below.

Denoting the z^{th} realized forecast error of the i^{th} model and i^{th} model's sample average forecast error with $V_{z,i}$ and $V_{\cdot,i}$, respectively, we start by considering $\hat{\pi}$, i.e., a consistent estimator of π (Lemma 1 p. 612 Ledoit & Wolf, 2003)

$$\hat{\pi} = \sum_{i=1}^m \sum_{j=1}^m \hat{\pi}_{ij}, \quad (\text{D.1})$$

$$\hat{\pi}_{ij} = \frac{1}{n} \sum_{z=1}^n \left[\left(V_{zi} - \frac{1}{n} V_{\cdot i} \right) \left(V_{zj} - \frac{1}{n} V_{\cdot j} \right) - \Sigma_{ij} \right]^2. \quad (\text{D.2})$$

Notice that π and its consistent estimator $\hat{\pi}$ do not depend on the reference matrix $\bar{\boldsymbol{\Sigma}}$. Conversely, consider $\hat{k}$, i.e a consistent estimator of k (Lemma 2 p. 612-613 Ledoit & Wolf, 2003).

$$\hat{k} = \sum_{i=1}^m \sum_{j=1}^m \hat{k}_{ij}.$$

By construction, $\hat{k}$ depends on the reference matrix $\bar{\boldsymbol{\Sigma}}$. When $\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\Sigma}}^E = \bar{\sigma}^2 \mathbf{I}$, the covariance each element of the sample covariance matrix and the identity matrix scaled by $\bar{\sigma}^2$ is constantly equal to zero. Therefore, in the case of COS_E⁽⁺⁾, the consistent estimator of the optimal shrinkage intensity λ^* , is given by $\hat{\lambda}^* = \frac{\hat{\pi}}{\hat{\gamma}}$. This corresponds to the optimal shrinkage intensity derived in Theorem 2.1 in Ledoit and Wolf (2004a).

Conversely, in the case of COS_IL⁽⁺⁾, i.e., when $\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\Sigma}}^{\text{IL}}$, the reference matrix depends on the sample estimator of the covariance matrix of prediction errors. While the off-diagonal elements of $\bar{\boldsymbol{\Sigma}}^{\text{IL}}$ are restricted to be zero, i.e., $\bar{\Sigma}_{ij, i \neq j}^{\text{IL}} = \Sigma$, the elements on the diagonal of $\bar{\boldsymbol{\Sigma}}^{\text{IL}}$ are the sample estimates of the variances of prediction errors, i.e., $\bar{\Sigma}_{ii}^{\text{IL}} = \Sigma_{i,i}$. In this case, we obtain

$$\hat{k}_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ \hat{\pi}_{ii} & \text{otherwise} \end{cases}.$$

This result is a particular case of optimal shrinkage intensity in the case the constant correlation reference matrix presented in Appendix B in Ledoit and Wolf (2004b).

Appendix E. Simulation study: Results for Design 2 and Design 3

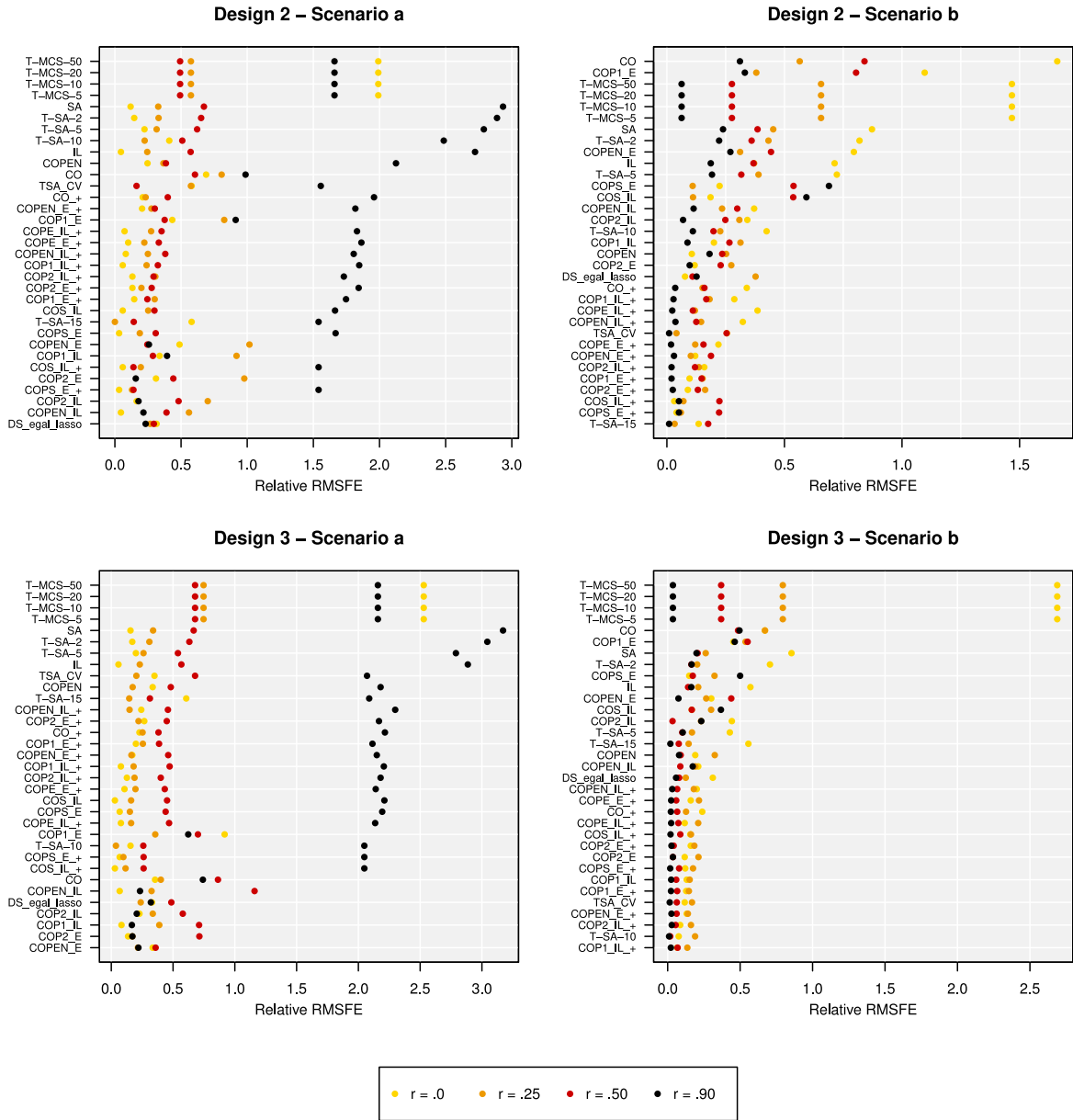


Fig. E.1. Relative RMSFE with respect to the true optimal forecast combination w^* . Performance of the forecasts combination schemes for various specifications of the covariance matrix Σ_v of the forecasts errors. Combination schemes are ordered according to their average performances across correlation levels. In each design, the candidates that have superior predictive performance yield RMSFEs that are 50% lower than the remaining ones. In *design 2*, the first five candidates have superior predictive performance; in *design 3* the first ten candidates have superior performance.

Appendix F. Forecast selection using CO+ and COS(+) methods

Table F.1

L2 norm of weights assigned to noise forecasts.

	r = 0	r= 0.25	r=0.5	r=0.9
T = 30	$\alpha_{noise} = 10$			
CO+	0.009	0.007	0.007	0.007
COPS_E	0.021	0.004	0.001	0.002
COS_IL	0.001	0.001	0.001	0.002
COPS_E+	0.021	0.001	0.000	0.000
COS_IL+	0.001	0.000	0.000	0.000
T = 30	$\alpha_{noise} = 1$			
CO+	0.017	0.017	0.018	0.015
COPS_E	0.020	0.031	0.046	0.083
COS_IL	0.018	0.027	0.040	0.068
COPS_E+	0.015	0.012	0.012	0.014
COS_IL+	0.013	0.011	0.011	0.013
T = 100	$\alpha_{noise} = 10$			
CO+	0.008	0.007	0.007	0.007
COPS_E	0.009	0.000	0.000	0.000
COS_IL	0.001	0.000	0.000	0.000
COPS_E+	0.009	0.000	0.000	0.000
COS_IL+	0.001	0.000	0.000	0.000
T = 100	$\alpha_{noise} = 1$			
CO+	0.015	0.015	0.015	0.011
COPS_E	0.015	0.024	0.033	0.065
COS_IL	0.014	0.023	0.032	0.061
COPS_E+	0.010	0.010	0.010	0.006
COS_IL+	0.010	0.010	0.009	0.006

Notes. L2 norm of the estimated weights assigned to noise forecasts. We model the i -th forecast as $\hat{y}_i = y + \epsilon_i$ for $i \in \{1, \dots, 10\}$ and $\hat{y}_i = \alpha_{noise} \epsilon_i$ for $i \in \{11, \dots, 20\}$ where $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i^2] = \alpha^2 = 1$ and $\mathbb{E}[\epsilon_i \epsilon_j] = r \alpha_i \alpha_j$. Despite the reference weighting schemes $\bar{w}^{IL}$ and $\bar{w}^F$ assign non-zero weights to noise forecasts, an aspect that penalizes COS' selection ability, COS methods consistently discard noise forecasts. The reason is that as the sample size increases, we also more precisely estimate Σ : the shrinkage intensity λ^* will asymptotically decay, providing the consistency result of the linear shrinkage estimator of the covariance matrix of Ledoit and Wolf (2003). Thus, the L2 norm of the weights assigned to noise forecasts approaches zero as the sample size increases. This holds for any level of correlation.

Appendix G. Performance of all the considered models in the empirical studies

Table G.1

Forecasting Boston housing prices, out-of-sample performance.

$i \backslash j$:	Diebold-Mariano		RMSFE	RMSPFE
	COPS-E ⁺	COS-IL ⁺		
OLS	⋄	⋄	.5639	.5713
L1	⋄	⋄	.5638	.5712
L2	⋄	⋄	.5687	.5761
svmLASSO	⋄	⋄	.5963	.6041
svmRIDGE	⋄	⋄	.5947	.6025
EN	⋄	⋄	.5952	.6030
PLS	⋄	⋄	.5845	.5922
RF	⋄	⋄	.4567	.4627
GBRT	⋄	⋄	.3763	.3812
NN1*	=	=	.3280	.3323
NN2	⋄	⋄	.3578	.3625
NN3	=	=	.3320	.3364
NN4	=	=	.3304	.3348
SA	⋄	⋄	.4135	.4183

(continued on next page)

Table G.1 (continued).

T-SA-3	⋄	⋄	.3700	.3748
T-SA-5	⋄	⋄	.3363	.3411
T-SA-7	=	=	.3008	.3056
T-SA-9	=	=	.3206	.3254
T-SA-11	<	<	.3330	.3378
T-SA-CV	⋄	⋄	.3700	.4048
T-MCS-60%	=	=	.3217	.3264
T-MCS-80%	⋄	⋄	.3578	.3626
T-MCS-90%	⋄	⋄	.3578	.3626
IL	⋄	⋄	.3789	.3837
DS_egal_lasso	=	=	.3156	.3203
CO	=	=	.3188	.3236
COPEN	⋄	⋄	.3280	.3328
CO ⁺	=	=	.3185	.3223
COP1_E ⁺	⋄	⋄	.3301	.3349
COP2_E ⁺	=	=	.3100	.3148
COPE_E ⁺	<	<	.3437	.3484
COPEN_E ⁺	⋄	⋄	.3435	.3482
COP1_IL ⁺	⋄	⋄	.3266	.3314
COP2_IL ⁺	=	=	.3040	.3087
COPE_IL ⁺	=	=	.3160	.3208
COPEN_IL ⁺	⋄	⋄	.3230	.3277
COP1_E	⋄	⋄	.3665	.3713
COP2_E	⋄	⋄	.3216	.3263
COPEN_E	<	<	.3533	.3581
COP1_IL	⋄	⋄	.3261	.3309
COP2_IL	=	=	.3140	.3088
COPEN_IL	<	<	.3243	.3290
COPS_E	=	=	.3015	.3063
COS_IL	=	=	.3006	.3054
COPS_E ⁺	.	=	.3087	.3134
COS_IL ⁺	=	.	.3086	.3133

Notes. The symbols <, ⋄ and ⋄ indicate that model j (column index) surpasses model i (row index) at confidence levels of 90%, 95% and 99%, respectively. Failure to reject the null hypothesis of equal predictive performance at a 10% significance level is indicated with ".".

Table G.2

Predicting Chicago Fed National Activity Index (CFNAI), out-of-sample performance.

$i \backslash j$:	Diebold-Mariano		RMSFE	RMSPFE
	COPS-E ⁺	COS-IL ⁺		
OLS	⋄	⋄	1.280	1.321
L1	=	=	.700	.723
L2	=	=	.698	.721
svmLASSO	=	=	.757	.782
svmRIDGE	=	=	.733	.757
EN	=	=	.686	.708
PLS	=	=	.718	.741
RF*	=	=	.683	.706
GBRT	=	=	.740	.764
NN1	⋄	⋄	.822	.849
NN2	⋄	⋄	.821	.848
NN3	⋄	⋄	.829	.856
NN4	⋄	⋄	.856	.884
SA	=	=	.706	.729
T-SA-3	=	=	.689	.712
T-SA-5	=	=	.699	.722
T-SA-7	=	=	.693	.716
T-SA-9	=	=	.691	.714
T-SA-11	⋄	⋄	.724	.748
T-SA-CV	=	=	.702	.725
T-MCS-60%	=	=	.702	.725
T-MCS-80%	=	=	.702	.725
T-MCS-90%	=	=	.733	.757
IL	=	=	.697	.719

(continued on next page)

Table G.2 (continued).

DS_egal_lasso	<<	<	.828	.855
CO	<	<	.794	.820
COPEN	=	=	.757	.782
CO ⁺	<	<	.725	.748
COP1_E ⁺	=	=	.698	.721
COP2_E ⁺	=	=	.694	.717
COPE_E ⁺	=	=	.698	.721
COPEN_E ⁺	=	=	.695	.718
COP1_IL ⁺	=	=	.694	.717
COP2_IL ⁺	=	=	.696	.719
COPE_IL ⁺	=	=	.694	.717
COPEN_IL ⁺	=	=	.691	.714
COP1_E	=	=	.698	.720
COP2_E	=	=	.686	.709
COPEN_E	=	=	.698	.720
COP1_IL	=	=	.693	.716
COP2_IL	=	=	.688	.710
COPEN_IL	=	=	.693	.716
COPS_E	=	=	.694	.716
COS_IL	=	=	.704	.726
COPS_E ⁺	.	=	.694	.717
COS_IL ⁺	=	.	.696	.719

Notes. The symbols <, << and <<< indicate that model j (column index) surpasses model i (row index) at confidence levels of 90%, 95% and 99%, respectively. Failure to reject the null hypothesis of equal predictive performance at a 10% significance level is indicated with “=”.

Appendix H. (Pseudo) out-of-sample time-series exercise in Chicago Fed National Activity Index (CFNAI)

In the prediction exercise described in Section 4.4, we randomly split the sample into 3 independent subsamples: training, testing, and a fold of 21 observations to estimate the combinations of forecasts. Here, we test the robustness of the proposed methods to time-series frameworks, we re-conducted the forecasting exercise in CFNAI using an expanding window to estimate the individual models and to form the combinations of forecasts.

We start our prediction exercise in December 2014. The training phase uses data available up to time t and then we extend the data by using an expanding window. The training set, made of the data available up to time t , is partitioned in two folds. The combination fold is composed of 21 observations selected randomly from the training set, and is used to estimate the combining weights. The other fold is used to train the individual models. Notice that the first iteration of this expanding window setup and the exercise described in Section 4.4 count the same number of observations available for training (146 for the estimation of the candidate forecasts and 21 to form forecast combinations). However, in the time series exercise, the size of the training set increases by one observation at each iteration, expanding the available information to estimate the baseline forecasts (e.g., in the last iteration, we count 187 observations for the estimation of the candidate forecasts and 21 to form the forecast combinations).

Because of the time-series dimension, we compute the Diebold-Mariano statistics by considering Newey-West standard errors (Newey & West, 1987) where we select the bandwidth of the Bartlett kernel using an AR(1) process. In the MCS test, we instead consider a block-bootstrap procedure with a window of four observations

Table H.1

Predicting Chicago Fed National Activity Index (CFNAI), (pseudo) out-of-sample performance in time-series.

$i \backslash j$:	Diebold-Mariano		RMSFE	MCS
	COPS-E ⁺	COS-IL ⁺		
OLS	<<<	<<<	0.928	
L1	=	=	0.512	
L2	=	=	0.512	
svmLASSO	<<	<<	0.524	
svmRIDGE	<<<	<<<	0.556	
EN	<	<	0.517	
PLS	<	<	0.535	
RF	=	=	0.477	*
GBRT	=	=	0.513	
NN1	<	<	0.534	
NN2	<	<	0.531	
NN3	<<	<<	0.572	
NN4	<	<	0.549	
SA	=	=	0.481	*
TSA_3	=	=	0.476	*
TSA_5	=	=	0.486	
TSA_7	=	=	0.488	
TSA_9	=	=	0.490	
TSA_11	=	=	0.504	
T-MCS-60%	=	=	0.490	*
T-MCS-80%	=	=	0.494	
T-MCS-90%	=	=	0.525	
IL	=	=	0.478	*
DS_egal_lasso	<<<	<<<	0.557	
CO	<<<	<<<	0.657	
COPEN	=	=	0.461	*
CO ⁺	<<<	<<<	0.505	
COP1_E ⁺	=	=	0.463	*
COP2_E ⁺	=	=	0.496	
COPEN_E ⁺	=	=	0.467	*
COPE_E ⁺	=	=	0.460	*
COP1_IL ⁺	=	=	0.465	*
COP2_IL ⁺	=	=	0.500	
COPEN_IL ⁺	=	=	0.471	*
COPE_IL ⁺	=	=	0.465	*
COP1_E	=	=	0.467	*
COP2_E	=	=	0.498	
COPEN_E	=	=	0.494	
COP1_IL	=	=	0.501	
COP2_IL	=	=	0.495	
COPEN_IL	=	=	0.485	*
COPS_E	<	<	0.527	
COS_IL	<	<	0.540	
COPS_E ⁺	.	=	0.484	*
COS_IL ⁺	=	.	0.483	*

Notes. The symbols <, << and <<< indicate that model j (column index) surpasses model i (row index) at confidence levels of 90%, 95% and 99%, respectively. Failure to reject the null hypothesis of equal predictive performance at a 10% significance level is indicated with “=”.

(four months). In the MCS test, because of the time-series dimension, we consider a block-bootstrap procedure with a window of four observations.

References

- Aiolfi, M., & Favero, C. A. (2005). Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting*, 24(4), 233–254.
- Amendola, A., Braione, M., Candila, V., & Storti, G. (2020). A model confidence set approach to the combination of multivariate volatility forecasts. *International Journal of Forecasting*, 36(3), 873–891.

- Aruoba, S. B., Diebold, F. X., Nalewaik, J., Schorfheide, F., & Song, D. (2012). Improving U.S. GDP measurement: A forecast combination perspective. In *Recent advances and future directions in causality, prediction, and specification analysis* (pp. 1–25). Springer New York.
- Atiyya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, 36(1), 197–200.
- Behr, P., Guettler, A., & Miebs, F. (2013). On portfolio optimization: Imposing the right constraints. *Journal of Banking & Finance*, 37(4), 1232–1242.
- Ben-Porat, O., & Tennenholtz, M. (2017). Best response regression. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 1498–1507). Red Hook, NY, USA: Curran Associates Inc.
- Bowley, A. L. (1920). *Elements of statistics*. New York: Charles Scribner's Sons.
- Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391), 580–598.
- Bunn, D. W. (1981). Two methodologies for the linear combination of forecasts. *Journal of the Operational Research Society*, 32(3), 213–222.
- Bunn, D. W. (1985). Statistical efficiency in the linear combination of forecasts. *International Journal of Forecasting*, 1(2), 151–163.
- 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.
- Conflitti, C., Mol, C. D., & Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, 31(4), 1096–1103.
- DeMiguel, V., Garlappi, L., Nogales, F. J., & Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5), 798–812.
- DeMiguel, V., Garlappi, L., & Uppal, R. (2007). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5), 1915–1953.
- Dickinson, J. (1973). Some statistical results in the combination of forecasts. *Journal of the Operational Research Society*, 24(2), 253–260.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Diebold, F. X., & Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, 6(4), 503–508.
- Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting*, 35(4), 1679–1691.
- Geweke, J. (2005). Elements of Bayesian inference. In *Wiley series in probability and statistics* (pp. 21–71). John Wiley & Sons, Inc.
- Granger, C. W., & Jeon, Y. (2004). Thick modeling. *Economic Modelling*, 21(2), 323–343.
- Granger, C. W. J., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3(2), 197–204.
- Green, R. C., & Hollifield, B. (1992). When will mean-variance efficient portfolios be well diversified? *The Journal of Finance*, 47(5), 1785–1809.
- Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5).
- Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 81–102.
- Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291.
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). Linear methods for regression. In *The elements of statistical learning* (pp. 43–99). Springer New York.
- Jagannathan, R., & Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4), 1651–1683.
- Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105(3), 1177–1216.
- Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5), 603–621.
- Ledoit, O., & Wolf, M. (2004a). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2), 365–411.
- Ledoit, O., & Wolf, M. (2004b). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4), 110–119.
- Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2), 1024–1060.
- Ledoit, O., & Wolf, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies*, 30(12), 4349–4388.
- Ludvigson, S., Ma, S., & Ng, S. (2015). *Uncertainty and business cycles: Exogenous impulse or endogenous response?*. NBER working papers 21803, National Bureau of Economic Research, Inc.
- Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77–91.
- McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4), 574–589.
- Michaud, R. (1989). The markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45, 31–42.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703.
- Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681–686.
- Quinlan, J. R. (1993). Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on international conference on machine learning* (pp. 236–243). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..
- Samuels, J. D., & Sekkel, R. M. (2017). Model confidence sets and forecast combination. *International Journal of Forecasting*, 33(1), 48–60.
- Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3), 331–355.
- Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, 36(2), 428–441.
- Timmermann, A. (2006). Chapter 4 forecast combinations. In *Handbook of economic forecasting* (pp. 135–196). Elsevier.
- Tu, J., & Zhou, G. (2011). Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, 99, 204–215.
- Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, 180, 68–77.
- Winkler, R. L., & Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)*, 146(2), 150.
- Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7, 2541–2563.