Forecast Combination with Random Subspaces: Accuracy Gains and Interpretability

Boris Kozyrev*

March 2, 2026

Abstract

This paper studies Random Subspace Regressions (RSM) as a forecast-combination strategy. We develop a finite-sample framework that makes explicit the trade-off between using more forecasts (lower combination variance) and estimating more weights (higher estimation error). Monte Carlo experiments, where individual forecasts are themselves estimated, show that RSM typically delivers lower RMSFE than equal weighting and restricted OLS. In an empirical application using real-time FRED-MD vintages to nowcast U.S. GDP growth, RSM significantly improves the forecast combination performance relative to the mean with the largest gains during the COVID-19 recession and compares favorably to Lasso, Ridge, and Random Forest benchmarks. Finally, we link RSM to Shapley value decomposition, yielding a transparent decomposition of the combined forecast into contributions of individual forecasts even in high-dimensional settings.

Keywords: Forecasting, Forecast combination, Random Subset, Forecast Combination Puzzle.

JEL: C53, C52, C32.

*Halle Institute for Economic Research (IWH), boris.kozyrev@iwh-halle.de

1 Introduction

Nowcasting, as well as short-horizon forecasting of economic activity, is crucial for decision-makers such as firms and governments. Over the last couple of decades, the amount of data used for (improving) macroeconomic forecasting has been increasing sharply. This means researchers may generate hundreds of different forecasts for a given macroeconomic indicator, for example, GDP. These individual predictions may be derived from diverse methods or data sources. A natural question arises: how to combine and/or select from numerous forecasts in order to produce a final prediction?

One of the most straightforward techniques used by practitioners and researchers is to assign equal weight to each prediction. This simple aggregation method proves to be a solid benchmark and often outperforms more complex combination approaches. In the literature, this fact is known as a "forecast combination puzzle". However, despite its straightforward implementation and often empirical robustness, in many cases, the rule of applying equal weight to each individual forecast can be harmful. The two potential and interconnected problems are (i) under- or overestimating the true weight of a forecast, which in turn leads to a second challenge (ii) a wrongly estimated relative contribution of this forecast to the final prediction, since it crucially depends on the weight.

First, apart from empirical robustness, the theoretical advantage of using the ad hoc equal-weighting is that it avoids estimating weights and therefore avoids estimation error. The latter can be substantial if the number of forecasts is rather high (especially compared to the time dimension). Still, in many practically-relevant cases, taking the simple average is rather restrictive since it does not take into account any prior historical performance of predictions. Intuitively, by unconditionally taking the mean of a set of forecasts, one is likely to misspecify the optimal weights: too much weight to weak forecasters and too little to strong ones. This can be especially important during turbulent times such as the COVID-19 crisis, when some of the forecasts (for instance, if they are built using the leading indicators) respond substantially differently to the shock.

Another concern related to misspecified forecast combination weights is interpretability. In many applications, the estimated weights can not only be used to form the combined forecast, but also to quantify how much each individual forecast contributes to the final prediction. As a result, the interpretability of the combined forecast depends directly on the quality of the estimated weights. This matters for decision-makers, who often require empirical results that are transparent and easy to communicate. In this paper, interpretability means decomposing the combined forecast into additive contributions that sum to the model's forecast deviation from a baseline (e.g., a historical-mean benchmark).

Economists and policymakers care not only about predictive accuracy, but also about understanding the forces driving the forecasted outcomes. Forecasting (and forecast combination) should help motivate economic and public policy decisions rather than deliver intractable, though sometimes ac-

curate, predictions (Burgess et al., 2013). Even the "best" single model can occasionally produce large errors. In principle, understanding a model's mechanics makes it easier to diagnose unusual predictions in light of the inputs and the underlying economic environment.

These considerations have motivated a growing literature on model interpretability and feature-attribution methods. The Shapley value decomposition (SVD) is among the most widely used tools for allocating such contributions across the input. However, exact Shapley values computation oftentimes becomes computationally expensive in high dimensions and non linear settings, which forces researchers to either focus on a small set or groups of indicators or to rely on model-specific algorithms for which shortcuts are known such as linear models or Random Forest (RF). RSM benefits from the shortucuts for linear models, making attribution feasible and meaningful.

In this article, we propose a framework that connects two seemingly independent fields: forecast combination and model interpretability. We build on the Random Subspace Regression method (RSM) Boot and Nibbering (2019), originally proposed for forecasting rather than forecast combination. A specific advantage of RSM is that this model makes it feasible to remain within a linear framework even when the number of predictors (here, individual forecasts) is large. Hence, it also allows for obtaining the contributions for each variable directly via developed shortcuts. Essentially, RSM estimates many low-dimensional regressions on randomly selected subsets of forecasts and then aggregates the resulting subspace predictions. Thus, it replaces a single high-dimensional (and potentially infeasible) regression with a collection of smaller, related estimation problems, therefore reducing weight-estimation error while preserving or even improving forecast performance.

First, we study the theoretical properties of RSM for forecast combination. We derived the conditions under which RSM achieves a lower Mean Squared Error (MSE) than both the optimal linear combination under a sum-to-one restriction (restricted OLS) and the equal weighting benchmark. The results are in line with, for example, Elliott and Liao (2025) in the sense that in the exogenous case – when forecasts are treated as given – in order to beat the equal weighting scheme, RSM requires overcoming the estimation penalty. Simultaneously, RSM can mitigate this penalty relative to estimating the full set of optimal weights. Also, the general expressions as well as closed-form ones (in some specific cases) are obtained to illustrate the subset size k^*, which minimizes MSE under the RSM framework even if in a specific case (such as compound symmetry) RSM cannot outperform equal weighting because the latter is itself optimal. The analysis implies, that k^* depends strongly but not exclusively on the sample size T. In particular, we show that $k^* \propto \sqrt{T-1}$ and often takes a form $k^* = c\sqrt{T-1}$, where c is some constant which can depend on several other model-specific values, for instance, the pairwise correlation among forecasts.

Second, we conduct the Monte Carlo simulations to evaluate the RSM forecast combination performance compared to the same benchmarks (equal weighting and restricted OLS "optimal" weights). Unlike the theoretical analysis, in this exercise, we consider a different setting. Namely, when the

forecasts are not exogenous, but rather generated from each variable included in the data-generated process (DGP) of the dependent variable. This design – univariately constructing covariate-based predictions – is standard in the forecast combination literature when a high-dimensional setting is considered (among others, see Chen and Maung (2023), Samuels and Sekkel (2017)). Moreover, this specification helps us to understand how the theoretical results derived for the exogenous case translate to cases when the forecasts themselves are estimated.

The Monte Carlo simulations results suggest that RSM yields the lower root mean squared forecasting error (RMSFE) than both benchmarks when the subset dimension is chosen near the k^* across all DGPs considered. However, the size of the gains depends on various factors, including the time dimension T. Consistent with the provided theoretical results, equal weighting is often more resilient than the optimal weights in the case when the number of forecasts is large, because the latter suffers from the weight-estimation error.

Typically, RMSFE, as a function of k, takes a U-shaped form highlighting a non-linear relationship. This finding indicates the importance of choosing the right k: if the subset size is too high or too low, the relative performance of RSM can deteriorate. However, the value of k^* which minimizes RMFSE across all scenarios tends to lay in the close proximity to $k = \sqrt{T}$ rounded to the nearest integer. In most cases, there are no additional gains of estimating true k^* via cross validation. The ad hoc rule of choosing $k = \sqrt{T}$ yield the similar to (sometimes exact) minimum RMSFE value. Thus, this choice of k is an empirically robust rule of thumb for the endogenous simulation and empirical settings considered. Therefore, we treat $k = \sqrt{T}$ as a practically relevant starting point for applying RSM for forecast combination in the endogenous case. Alternatively, the cross validation, if desired, can be implemented around the proposed value.

Additionally, we explore a synergy between RSM and the SVD. Because RSM implies a linear structure at the subspace level, we can decompose the RSM prediction into forecast contributions using standard linear-model shortcuts, while still accommodating high-dimensional sets of forecasts. Thus, under this framework, it is possible to circumvent the challenges of using many variables while dealing with the SVD. We make sure that the main SVD advantage is fulfilled: the difference between the final prediction and a baseline (e.g., the historical mean) is fairly distributed among the individual forecasts. That is, all contributions sum up exactly to the total deviation holding up to an intercept term.

In our empirical application, we use real-time vintages of the FRED-md dataset, provided by the Federal Reserve Bank of St.Louis, to nowcast US GDP using a high-dimensional set of predictions. For the main analysis, the optimal subset size is set to $k = \sqrt{T}$ rounded to the nearest integer, motivated by the Monte Carlo simulation results. We also examine robustness to alternative values of k and found only marginal gains of using k^*, found by cross-validation. The results indicate that RSM outperforms the equal weighting, delivering lower RMSFE across the specifications. In the full sample, the relative

to the mean RMSE for RSM is 0.67 and 0.88, if the COVID-19 quarters are excluded. This strengthens the intuition that during crises, the RSM produces a more accurate forecast combination due to its flexibility. Additionally, we also compare RSM to other high-dimensional benchmarks such as LASSO, Ridge regression, and Random Forests, with RSM achieving the lowest RMSE among the competing approaches. These empirical evidence supports the practical usefulness of the rule-of-thumb choice of $k = \sqrt{T}$, even though estimating k^* can sometimes yield further improvements.

Finally, we empirically investigate the difference between the weights and corresponding Shapley Values obtained by the equal weighting and RSM, respectively. We observe that equal weighting tends to underestimate the global weight of some important predictions (for example, industrial production), which, in turn, leads to underestimating the relative contribution of this forecast, especially during the COVID-19 crisis. The tight link between weights and contributions also potentially helps interpret negative weights, which are often puzzling in practice.

Literature review and contributions. This paper relates to several strands of literature. First, this paper contributes to the literature of forecast combination by providing a flexible and easy-to-implement regression-based estimation of forecast combination weights using random subspaces. So far, RSM and its various modifications is used mostly to produce forecasts based on the actual data (Boot and Nibbering (2019), Wichitaksorn et al. (2023), Kotchoni et al. (2019), Pick and Carpay (2022), Kant et al. (2025), Meligkotsidou et al. (2021), Dinh et al. (2024)) with little being known about its properties in the forecast combination framework. One closely related contribution is Chen and Maung (2023), who, in the spirit of the complete subset regression idea of Elliott et al. (2015), fix a subset size k, estimate many k-variate forecasting regressions, and then average the resulting subset-model forecasts. In their empirical implementation, k is kept small (e.g., $k \in 1, 2, 3, 5$) largely because the number of candidate subsets grows exponentially with k, making larger values computationally burdensome. By contrast, our random-subspace approach treats subsetting as an integral part of the forecast-combination estimator. We repeatedly estimate constrained combination weights within randomly drawn subspaces and use a theory-guided estimation-error trade-off to motivate the choice of k. Our framework yields practical guidance for selecting k (typically higher that proposed values used in Chen and Maung (2023)). Consistent with this perspective, Chen and Maung (2023) report that forecasting performance improves as k increases over the small grid they consider, suggesting that further gains may arise.

Second, we contribute to the growing literature of high-dimensional model/variable selection and dimension reduction in the forecast combination settings. While dealing with a large number of variables, researchers often assume sparsity. Sparse models explicitly imply that the least relevant variables are discarded. For example, LASSO, as proposed in Tibshirani (1996), forces coefficients for many predictors to equal zero, which can be too restrictive. RSM avoids assuming sparsity – an assumption often imposed in the high-dimensional settings. Moreover, Giannone et al. (2017) argue,

based on Bayesian forecasting exercise, that posterior predictive distributions are characterized by a combination of various different models rather than being concentrated on a single sparse model or a single dense model. This intuition can transfer to the frequentist forecast combination settings suggesting that a well-designed model averaging technique can outperform any sparse model, as shown for the forecasting case in Kotchoni et al. (2019).

Furthermore, RSM allows for computing each predictor's contribution, linear attribution scheme inspired by Shapley Value Decomposition.

The rest of the paper is organized as follows: Section 2 provides the current state of the forecast combination research. In Section 3, we describe the RSM method and how it can be used to combine forecasts. Section 4 considers the theoretical framework. Section 5 explains the Monte Carlo framework and simulation results. Section 6 introduces Shapley Values and establishes their correspondence to RSM. Section 7 contains an empirical application of the RSM method. Finally, Section 8 concludes and provides an outlook for future research.

2 Forecast Combination

Many researchers from various disciplines have analyzed a wide range of forecast combination methods. In many practical settings, there are multiple forecasts of the same variable obtained either by different methods or by using different sources of information (Timmermann, 2006). A common premise is that combining different individual predictions can help to improve accuracy. The aggregation strategy enables the incorporation of much more available information, which indeed may be more fruitful than trying to seek a "best" model. For a review of the extensive literature on forecast combinations, see Wang et al. (2023). In this section, we describe methods dealing with point forecast combinations based on different linear combinations. We focus on this specific class of forecast aggregation because, as can be seen further, it is possible to merge the latter with RSM. To conclude, a forecast combination puzzle is discussed.

2.1 Linear Combinations of Individual Forecasts

One of the most straightforward strategies to combine various forecasts is to assign weights to each individual prediction: the more precise the forecast, the greater the weight. There is a plethora of different weighting schemes. Thus, it sounds feasible to construct a linear combination of the individual forecasts and treat the OLS coefficients as weights.

More formally, suppose an N-dimensional vector of h-step ahead forecasts $\hat{\boldsymbol{y}}_{T+h|T}$ is given. Additionally, an n-dimensional vector of past observations of the targeted variable $\boldsymbol{y_T}$ is observed. Under this framework, regression-based weights can be defined as follows (Granger and Ramanathan, 1984):

$$y_{T+h} = w_{T+h|T,0} + \boldsymbol{w'_{T+h|T}}\hat{\boldsymbol{y}}_{T+h|T} + \epsilon_{T+h}, \tag{1}$$

where $\hat{\boldsymbol{y}}_{T+h|T}$ are past individual forecasts.

Some restrictions may still be imposed on $\boldsymbol{w}'_{\boldsymbol{T+h|T}}$. For instance, a constant term can be omitted, and/or the weights can be adjusted so that their sum adds up to one. In the latter case, we obtain the so-called "optimal" weights Bates and Granger (1969):

$$y_{T+h} = \boldsymbol{w}'_{\boldsymbol{T+h|T}}\hat{\boldsymbol{y}}_{T+h|T} + \epsilon_{T+h}, \quad \sum \boldsymbol{w}'_{\boldsymbol{T+h|T}} = 1 \tag{2}$$

The use of regression-based weights for forecast combinations has a long history. Both advantages and limitations of this approach have already been examined in the last century. One of the critiques is shown in De Menezes et al. (2000). While exploiting the unrestricted regression, one should address potential multicollinearity and serial correlation in forecast errors. We assume that both of these challenges are mitigated when combining this approach with RSM, as discussed later.

Upon establishing the regression-based weighting method, another concern arises. What if the total number of $\hat{\boldsymbol{y}}_{T+h|T}$ is exceptionally large? In this case, a simple OLS estimation becomes impractical, particularly when the number of forecasts (N) exceeds the number of observations in the training set (n). It makes sense to use some variable selection methods, such as Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), to end up only with a handful of individual forecasts.

Typically, when dealing with high-dimensional data (in the form of $N > n$), LASSO or other penalized regressions have been extensively used for dimension reduction in many areas, including economics. However, sometimes it is necessary to use as much data as possible without excluding variables by assuming non-sparsity. RSM can manage this level of selection 'strictness' by utilizing all the available data.

2.2 Forecast Combination Puzzle

Although various forecast combination methods have been proposed over the last couple of decades, in practice, the simple average (assigning equal weights to each of the forecasts) tends to perform similarly to (or even better than) more sophisticated aggregation approaches. This fact perplexed researchers, who had been trying to find a prolific answer to that paradox. Stock and Watson (2004) were among those who studied this phenomenon. In that paper, the term "Forecast combination puzzle" was introduced.

Indeed, one should expect some gains in terms of forecast accuracy while using more rigorous tools for forecast combination rather than averaging. Moreover, theoretical results for some approaches were formally derived, proving this initial claim. However, often a simple mean overshadows other weighting schemes. Again, in Wang et al. (2023), a great summary of why the "Forecast combination puzzle" holds empirically is presented. To summarize, a possible explanation is that the weights are misspecified or wrongly computed due to, for instance, structural changes or limited data availability. Alternatively, benefits from using a sophisticated combination may be relatively small, thus estimation

per se overwhelms the potential gains, especially if the weights are treated as random and not fixed Claeskens et al. (2016).

However, if individual forecasts come from different sources or models, the variance can be high, in particular due to some outliers. If the number of outliers is high and they are not equally distributed around the mean, then averaging over this forecasting set may not be efficient. Although taking a median can arise as a potential method to produce a final prediction, as used in Stock and Watson (1999), it does not indicate how varied the individual predictions are in the set. Moreover, individual forecasts, even without outliers, can be distributed in such a way that taking a mean can be almost uninformative. For example, in Figure 1 and in Figure 2, the histograms for the expected change in prices during the next year based on the Surveys of Consumers, provided by the University of Michigan, for November 2024 and July 2025 are presented.

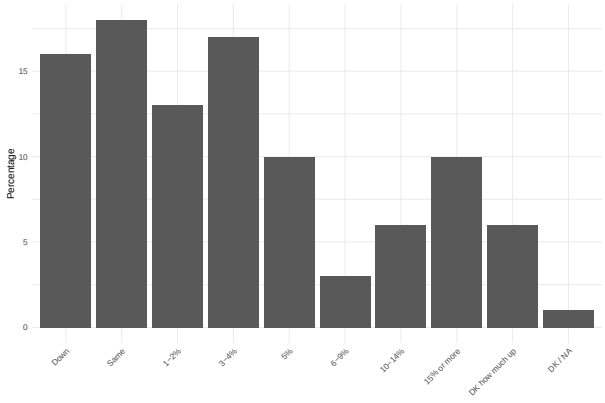


Figure 1: Expected Change In Prices During the Next Year (November 2024)

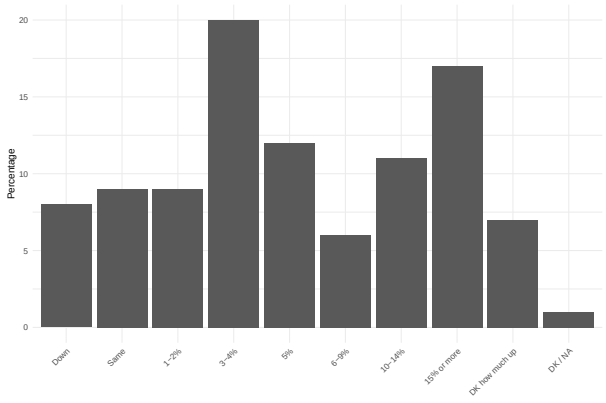


Figure 2: Expected Change In Prices During the Next Year (July 2025)

Because the Survey of Consumers extracts inflation expectations in pre-defined categories rather than as continuous numerical responses, any "mean expectation" is not directly observed but constructed by assigning representative values to bins. This construction is particularly consequential in the presence of open-ended categories such as "15% or more" and "Down", where the set of admissible values is wide, and the bin midpoint is undefined. Both histograms exhibit fat right tails, so the mean (as reported in the survey) can be disproportionately influenced by a minority of extreme responses and may therefore provide a poor summary of the "typical" respondent's beliefs.

For these reasons, aggregation methods that are less tail-sensitive are often more informative than a single mean. Using RSM for forecast combination allows for relying on the mean even in the presence of outliers and/or if there are fat tails in the initial set. This is achieved by obtaining a new set of forecasts – based on the initial one – which is supposed to be distributed in a way that taking a mean makes sense and provides good empirical performance.

3 Random Subspace

Initially, RSM methods appeared in the field of machine learning without much application in macroeconomic forecasting. Ho (1998) introduced this approach to train tree models, later Bay (1998) extended the same logic to the nearest neighbour classifiers. Wichitaksorn et al. (2023) applied RSM to feature selection in a logistic regression framework. Recently, random or complete[1] subspace methods have become a powerful tool used in macroeconomics (Elliott et al. (2013),Kotchoni et al. (2019), Pick and Carpay (2022)) and finance (Meligkotsidou et al. (2021)). Some further examples of papers utilizing subspace methods can be seen in Boot and Nibbering (2020).

Random subset regression is a randomized reduction method, according to which a small number of indicators are (randomly) selected to estimate a lower-dimensional problem. This procedure is repeated a sufficient number of times. Thus, one ends up with many approximations to the initial – sometimes infeasible to calculate directly – model. Finally, the forecasts from each low-dimensional model are combined (averaged). This allows for reducing forecast variance while utilizing all available information.

More formally, suppose that the following data-generating process (DGP) is defined as follows:

$$y_{t+1} = x_t'\beta + \epsilon_{t+1}, \tag{3}$$

for $t = 1, \cdots, T$, x_t' is a vector of predictors in $\mathbb{R}^N$, $\epsilon_{t+1} \sim iid(0, \sigma^2)$.

Initially, this model has been proposed for using observed time series to predict the target. Instead, we use this approach for forecast combination. To draw a parallel with previous sections and align the notation, in Equation 3, x_t' corresponds to the N-dimensional set of individual forecasts, whereas the n-dimensional vector y_{t+1} is the final prediction of the independent variable.

As discussed previously, when the number of estimated coefficients is large, it is either impossible to run OLS or there is a high variance of the latter, which can significantly decrease forecast accuracy. To overcome both potential challenges, following Boot and Nibbering (2019), it is possible to make a projection of N-dimensional vector x_t onto a k-dimensional subspace using a permutation matrix, denoted $R_i \in \mathbb{R}^{N \times k}$, such that:

$$\tilde{x}_t' = x_t' R_i$$

A permutation matrix R_i is responsible for randomly selecting (without replacement) a subset of k predictors out of all N covariates. To illustrate this idea, imagine a simple example where there are $N = 5$ variables and $k = 3$ of those are drawn. One possible R_i can look like as follows (Boot and

[1] Complete subspace methods imply that for a given set of predictors, forecasts from all possible linear regressions are combined.

Nibbering, 2020):

$$R_i = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

It should be noted that there are different methods to generate R_i. Instead of selecting a subset of covariates, weighted averages can be used to construct a different set of predictions. Weights are chosen randomly from a normal distribution. Moreover, the proposed specification of RSM implies that the selection is carried out using a uniform distribution (the probability that a given variable is selected is equal across the covariate set). It is also feasible to use so-called importance sampling probabilities or leverage scores (Mahoney et al., 2011)[2]. The rest of the paper uses coordinate subsampling via a 0–1 selection matrix with uniform inclusion probabilities.

Assigning an equal probability to each of the covariates can be partially explained by the following logic. Suppose there is a concern that, for a given model, randomly selected covariates might be less relevant than the rest of the pool. At the same time, most (if not all) of the time series used for macroeconomic predictions are correlated. That fact hints at two possible conclusions. First, a "less" relevant covariate is likely correlated with one that possesses more predictive power. Moreover, repeating this procedure – drawing variables randomly and averaging results over each iteration – makes the model less prone to 'uninformative' variable selection.

Conditional on a drawn R_i, a k-dimensional vector of OLS coefficients $\hat{\beta}^i$ can be obtained, as well as a single aggregated forecast value of $\hat{y}_{t+1}$. We are interested in both statistics. First, by averaging different $\hat{y}_{t+1}$, we claim that the RSM method could be beneficial for forecast aggregation, providing better predictions compared to more standard metrics. Later, we demonstrate the forecast performance of this method compared to benchmarks.

4 Theoretical Model

Following Elliott and Liao (2025), we consider the setup:

$$\begin{pmatrix} y_{t+1} \\ f^m_{t,t+1} \end{pmatrix} \sim \left[\begin{pmatrix} \mu_{t,t+1} \\ \mu_{t,t+1} \end{pmatrix}, \begin{pmatrix} \sigma^2_\epsilon & 0 \\ 0 & \tilde{\Sigma}_m \end{pmatrix} \right]$$

,

where y_{t+1} is the $T \times 1$ target variable with conditional mean $\mu_{t,t+1}$, $f^m_{t,t+1}$ – the $m \times T$-dimensional matrix of unbiased forecasts to combine, σ^2_ϵ – the irreducible noise, $\tilde{\Sigma}_m$ – the $m \times m$ variance-covariance matrix of the forecasts.

[2]It is worth mentioning that leverage scores are commonly used while randomly selecting rows rather than columns.

We consider a fixed forecast horizon (without loss of generality, $h = 1$), and treat the set of m candidate forecasts as exogenous draws from an unbiased forecast distribution conditional on information at time t. $\tilde{\Sigma}_m$ summarizes cross-forecast dispersion: its diagonal elements capture the marginal forecast-error variance (no forecasters know the true model) and its off-diagonals capture how strongly forecast errors co-move.

Additionally, the vector of forecast errors and the variance-covariance matrix of the latter are required to be defined. The m-dimensional vector of forecasts errors is denoted as $e^m_{t,t+1} = y_{t+1}\iota_m - f^m_{t,t+1}$, where ι_m is the m-dimensional vector of ones. Correspondingly, the variance-covariance matrix of $e^m_{t,t+1}$, which is assumed to be constant over time, can be described as follows:

$$\tilde{\Omega}_m = \sigma^2_\epsilon \iota_m \iota'_m + \tilde{\Sigma}_m$$

This setup allows for constructing the Mean Squared Error (MSE) loss function of any combined forecast if the latter is obtained by a weighted average (linear combination) of given forecasts under the assumption that the weights sum up to one. That is, if one produces a point forecast $\hat{y}_{t+1} = \omega'_m f^m$ such that $\sum \omega'_m = 1$, the desired loss function has the following form:

$$L = \underbrace{(1 - \omega'_m \iota_m)^2 \mu^2_{t,t+1}}_{Bias} + \underbrace{\sigma^2_\epsilon + \omega'_m \tilde{\Sigma} \omega_m}_{Variance} \tag{4}$$

However, when combining unbiased forecasts under the constraint that their corresponding weights add up to one, the final prediction will be unbiased. Thus, the first (bias) term in the Equation 4 is zero, and the comparison across such different methods comes down to minimizing the variance estimation part, which is equal to $\sigma^2_\epsilon + \omega' \tilde{\Sigma} \omega$.

The loss function depends on two entries: the vector of weights ω_m and the variance-covariance matrix of individual forecasts $\tilde{\Sigma}$. For both equal weighting and optimal weights, it is possible to derive the closed-form expression for the loss function. When averaging, each weight is the same for each forecast. Therefore, the vector of weights, in this case, is equal to $w^{ave} = m^{-1}\iota_m$ with the final loss $L^{ave} = \sigma^2_\epsilon + m^{-2}\iota'_m \tilde{\Sigma} \iota_m$. The weights for the optimal combination weights can be denoted as $w^{opt} = (\iota'_m \tilde{\Sigma}^{-1} \iota_m)^{-1} \iota'_m \tilde{\Sigma}^{-1}$ which yields the following loss function $L^{opt} = \sigma^2_\epsilon + (\iota'_m \tilde{\Sigma}^{-1} \iota_m)^{-1}$.

Finally, we impose two additional assumptions, following Elliott and Liao (2025). The first one deals with the structure of the matrix $\tilde{\Sigma}_m$. More specifically, we assume that noise in each of the individual forecasts around the conditional mean is the same, meaning we can express $\tilde{\Sigma}_m = \sigma^2_f \Sigma$, where σ^2_f scales the variance of forecast innovations around the conditional mean, Σ_m is a correlation matrix with ones on the main diagonal and pairwise correlation elsewhere.

Second, a specific estimation error form of the loss L^{opt} is assumed. Specifically, the L^{opt} value is inflated by the factor $(1 + \frac{m-1}{T})$ due to the estimation error arising from running restricted OLS. This multiplicative term is a reduced-form way to capture finite-sample estimation noise from estimating

$m-1$ free weights under the sum-to-one constraint. In this case, the loss for the optimal combination weights could be approximated by the following expression:

$$L^{opt} \approx (\sigma_\varepsilon^2 + Q^{opt}) \left(1 + \frac{m-1}{T}\right), \tag{5}$$

where $Q^{opt} = (\iota_m' \tilde{\Sigma}_m^{-1} \iota_m)^{-1}$.

This expression of the loss function would be used later to compare optimal weights to the weights obtained by RSM.

4.1 Random Subspace Method for Forecast Combination

We start from the simple example when one random subset S of size $k \in \{1...m\}$ is chosen. Based on the Equation 5, we can define the loss for using the subset S by replacing m by k:

$$L^{S_k} \approx (\sigma_\varepsilon^2 + Q^S(k)) \left(1 + \frac{k-1}{T}\right), \tag{6}$$

where $Q^S = (\iota_k' \tilde{\Sigma}_k^{-1} \iota_k)^{-1}$ is the k-dimensional counterpart of Q^{opt}, $\tilde{\Sigma}_k$ — the $k \times k$ variance-covariance matrix of the forecasts in the subset S.

The RSM procedure relies not on a single draw, however, on B draws. First, we show that the penalty for a single subset regression characterizes the estimation-error penalty of the bagged predictor, which averages across many fits. Thus, the effective degrees of freedom for the bagged estimator scale linearly with k, which allows us to use $(1 + \frac{k-1}{T})$. Suppose, a single random subset S_b with k variables is chosen. In this case, we have:

$$\hat{y}^{(S_b)} = H_{S_b} y,$$

where H_{S_b} is is the matrix of the estimated OLS coefficients for the subspace regression.

Trace of the matrix under optimal weighting is the number of free parameters, thus $tr(H_S) = k-1$. Averaging over B random subspaces yields:

$$\hat{y}^{RSM} = \frac{1}{B} \sum_{i=1}^{B} y^{(S_b)} = \left(\frac{1}{B} \sum_{i=1}^{B} H_{S_b}\right) y = \hat{H} y$$

The bagged estimator is also linear in y with the final matrix $\hat{H}$. Then, the effective degrees of freedom is as follows:

$$df(\hat{y}^{RSM}) = tr(\hat{H}) = \frac{1}{B} \sum_{i=1}^{B} tr(H_{S_b}) = k - 1$$

Since each subset regression has $tr(H_{S_b}) = k-1$. Hence, the penalty for the bagged predictor coincides with the penalty for a single subspace regression. Later, we use this common df-based approximation in reduced-form finite-sample penalties.

There is an additional potential advantage of iterating over different samples. In one realization of k-dimensional subset, the forecasts chosen at random can generate $Q^S(k) \gg Q^{opt}$, implying that increase in loss will not be attenuated by the variance reduction, Moreover, denote $\bar{\omega} = \frac{1}{B}\sum_{b=1}^{B} \hat{w}^{(b)}$ the average weight RSM weights based on B draws. Notice, that $\phi(w) = \omega'\tilde{\Sigma}\omega$ is a convex function because the matrix $\tilde{\Sigma}$ is positive semi-definite. Thus, by applying Jensen's inequality:

$$\phi(\bar{\omega}) = \phi(\frac{1}{B}\sum_{b=1}^{B} \hat{w}^{(b)}) \leq \frac{1}{B}\sum_{b=1}^{B} \phi(\hat{w}^{(b)}) \tag{7}$$

The Equation 7 shows that inflating each draw's loss and then averaging (right-hand side) is not smaller than the loss of using $\bar{\omega}$ (left-hand side) for the same k. Translating this back to the loss function:

$$\mathbb{E}(L^{RSM}) \leq \mathbb{E}_{S_k}\left[\left(\sigma_\varepsilon^2 + Q^S(k)\right)\left(1 + \frac{k-1}{T}\right)\right]$$

For later results, we would use the upper-bound:

$$\mathbb{E}(L^{RSM}) \approx \left(\sigma_\varepsilon^2 + Q^{RSM}(k)\right)\left(1 + \frac{k-1}{T}\right), \tag{8}$$

where $Q^{RSM} = \mathbb{E}_{S_k}(\iota_k'\tilde{\Sigma}_k^{-1}\iota_k)^{-1}$ is the average of optimal weights across S subsets of dimension k.

First, notice that the inflation factor (as in the case where $B = 1$) decreases compared to the optimal weights scenario in Equation 5. This happens because RSM reduces the number of parameters to estimate (from $m-1$ to $k-1$), thus making the estimation error smaller. However, this reduction comes at the cost of increasing the loss value itself as the optimal weights using all m individual forecasts is always strictly better than taking any subsets of them. We can express the latter fact in two interchangeable ways:

- $Q^{RSM}(k) \geq Q^{opt}$ with equality holding when $k = m$ or when additional forecasts are redundant.
- $\dfrac{\mathrm{d}Q^{RSM}(k)}{\mathrm{d}k}$ is non-increasing in k and achieves its minimum at $k = m$ or when additional forecasts are redundant.

In other words, the optimal weight using all m forecasts is known to be the best forecast combination method unless some forecasts are not useful for the combination; therefore, usually $Q^{RSM}(k) > Q^{opt} \quad \forall k \neq m$. However, due to the estimation error using the finite or even small sample data, it can produce suboptimal results. Alternatively, by increasing the subset dimension k, the $Q^{RSM}(k)$ decreases and collapses with Q^{opt} when $k = m$. This provides the intuition for the second inequality: the first derivative is non-positive with respect to k.

Additionally, we assume that the second order derivative of $Q^{RSM}(k)$ is positive $\left(\dfrac{\mathrm{d}^2Q^{RSM}(k)}{\mathrm{d}k^2}\right) > 0$. This assumption helps us derive the optimal k, which minimizes the total loss. Therefore, we

consider Q^{RSM} to be a decreasing and convex function: it keeps decreasing but flattens out with bigger k. Intuitively, when instead of m available forecasts one selects $m-1$, the resulting values of $Q^{RSM}(m)$ and $Q^{RSM}(m-1)$ should be much closer to each other then in the case, when k is small.

4.2 RSM and Optimal Weights Loss Comparison

After establishing the loss function for RSM, it is possible to compare it to the optimal weight counterpart to study under which conditions RSM can potentially produce less loss. The desired inequality:

$$\underbrace{(\sigma_\varepsilon^2 + Q^{opt})\left(1 + \frac{m-1}{T}\right)}_{L^{opt}} \geq \underbrace{(\sigma_\varepsilon^2 + Q^{RSM}(k))\left(1 + \frac{k-1}{T}\right)}_{L^{RSM}(k)} \tag{9}$$

holds when:

$$\frac{m-k}{T+k-1}\sigma_\varepsilon^2 + \underbrace{\frac{T+m-1}{T+k-1}}_{\geq 1} Q^{opt} \geq Q^{RSM}(k)$$

Notice that if all the m forecasts are used within the RSM procedure, one would obtain that $Q^{RSM}(m) = Q^{opt}$, which is fulfilled by default.

4.3 RSM and Equal Weights Loss Comparison

Recall, for averaging, we get $Q^{ave} = n^{-2}\iota_n'\tilde{\Sigma}\iota_n$. Additionally, unlike the optimal weighting and RSM, there is no penalty term for estimating weights. By comparing loss functions, we can derive the condition under which RSM would be superior to the mean forecast:

$$\underbrace{(\sigma_\varepsilon^2 + Q^{ave})}_{L^{ave}} > \underbrace{(\sigma_\varepsilon^2 + Q^{RSM}(k))\left(1 + \frac{k-1}{T}\right)}_{L^{RSM}} \tag{10}$$

or, alternatively:

$$Q^{ave} > \underbrace{\frac{T}{T+k-1}}_{\leq 1} Q^{ave} - \frac{k-1}{T+k-1}\sigma_\varepsilon^2 > Q^{RSM}(k)$$

The resulting condition further shows the well-known result: not only should RSM be better than the mean, but it also needs to overcome the estimation penalty, which does not exist for the equal weighting framework. Later, we check if RSM can improve forecast combination performance via Monte Carlo simulations and an empirical study in comparison to the equal weighting and optimal weights (if conditions, as imposed in Equation 9 and Equation 10 are satisfied).

4.4 Optimal Subset Size k^*

Formally, it is possible to minimize the loss function with respect to the number of variables to be selected k:

$$(\sigma_\varepsilon^2 + Q^{RSM}(k)) \left(1 + \frac{k-1}{T}\right) \to min,$$

which yields:

$$\frac{\mathrm{d}Q^{RSM}(k)}{\mathrm{d}k} = -\frac{Q^{RSM}(k) + \sigma_\epsilon^2}{T + k - 1}$$

The proof can be found in Appendix A.

This condition cannot further be simplified in the general case without assuming restrictions on $Q^{RSM}(k)$. However, it is still possible to interpret this condition as follows: at the optimal k^* the relative gain from increasing k (reducing $Q^{RSM}(k)+\sigma_\varepsilon^2$) exactly matches the relative cost from increasing k via the penalty term. This interpretation follows the motivation of using RSM by decreasing the number of variables to combine from m to k; we increase the loss with the hope of reducing the estimation error.

4.5 Special cases of k^*

There are cases, when we can explicitly derive the closed-form k^*. This can be achieved either by imposing a structure on the variance-covariance matrix $\tilde{\Sigma}$ or by assuming a specific functional form of $Q^{RSM}(k)$.

Compound Symmetry[3] First, we can show that if all forecasts are equicorrelated (all pairs in a set share an identical correlation coefficient $0 \leq \rho < 1$), the optimal subset size k^* equals to:

$$k^* \approx \sqrt{\frac{(T-1)(1-\rho)}{\rho + \frac{\sigma_\epsilon^2}{\sigma_f^2}}}, \qquad k^* = \text{nearest integer in } [2,\, m] \tag{11}$$

The proof can be found in Appendix B.

The size of k^* depends on three quantity: the number of observations $(T-1)$, the correlation coefficient (ρ), and the ratio $\frac{\sigma_\epsilon^2}{\sigma_f^2}$. With a bigger sample size, one should use a higher subset dimension k^*. Notice that if T is large enough, according the Equation 11, k^* could be bigger m, that is why the corner solution is presented. This also proves the intuition that if there is a sufficient amount of observations, the optimal weight would be superior to RSM. The pairwise correlation ρ also plays a significant role. The higher ρ, the smaller the subset size is supposed to be. This result indicates that when the forecasts are highly correlated, the smaller subset is enough to capture the dynamics. However, if ρ is close to perfect correlation (say, $\rho = 0.99$), the optimal subset size k^* can shrink close to 0, thus the corner solution of $k^* = 2$ is implemented. The ratio $\frac{\sigma_\epsilon^2}{\sigma_f^2}$ indicate when the target is

[3]We acknowledge the fact that under compound symmetry, averaging all the forecasts is the optimal strategy, yielding the lowest loss. Thus, RSM cannot outperform equal weighting. However, it provides an intuition for the results in the further sections.

inherently unpredictable, combining many imperfect forecasts offers limited gains. Hence, k^* decreases with respect to it.

Independent Forecasts

On the contrary, when the forecasts are fully independent ($\rho = 0$), the optimal subset dimension is as follows[4]:

$$k^* \approx \sqrt{\frac{(T-1)\sigma_f^2}{\sigma_\epsilon^2}} \tag{12}$$

Function form of $Q^{RSM}(k)$

Alternatively, we can assume a functional form of Q^{RSM}. The compound-symmetry benchmark already implies a parsimonious decay in k: $Q^{RSM}(k) = \big(\rho + (1-\rho)/k\big)\sigma_f^2$, i.e. a constant component plus a term of order $1/k$, as shown in Appendix B.

More generally, this structure is consistent with the bias–variance trade-off that arises when prediction is formed after restricting attention to a k-dimensional subspace. In particular, in the dimension-reduction regression literature, Slawski (2018) considers least squares based on a reduced design $XR \in \mathbb{R}^{T\times k}$ and shows that the excess prediction risk decomposes as

$$E(R) = \underbrace{\|(I - P_{XR})Xw^*\|_2^2/T}_{\text{approximation (bias)}} + \underbrace{\sigma^2 \operatorname{rank}(XR)/T}_{\text{estimation (variance)}},$$

so that under the common simplification $\operatorname{rank}(XR) = k$ the variance term is linear in k, whereas the bias is decreasing at the order of $O(\frac{1}{k})$.

Motivated by these findings, we adopt the reduced-form specification

$$Q^{RSM}(k) = q_\infty + \frac{b}{k^\alpha}, \qquad q_\infty > 0,\ b > 0,\ \alpha \geq 0, \tag{13}$$

where $\alpha = 1$ corresponds to the canonical $1/k$ rate (including compound symmetry), while $\alpha > 1$ allows for faster decay when the forecast pool is effectively low-dimensional.

Thus, if we assume a more general form of $Q^{RSM}(k) = q_\infty + \frac{b}{k^\alpha}$, for $q_\infty > 0, b > 0, \alpha \geq 1$, we can solve the optimization problem:

$$(\sigma_\varepsilon^2 + \underbrace{q_\infty + \frac{b}{k^\alpha}}_{Q^{RSM}(k)})\left(1 + \frac{k-1}{T}\right) \to min,$$

which yields the following first order condition for k:

$$(\sigma_\varepsilon^2 + q_\infty)k^{\alpha+1} + (1-\alpha)bk = \alpha b(T-1) \tag{14}$$

The proof can be found in Appendix C.

Generally, a nonlinear first order condition for k appears, which if all the constants are known can be solved numerically. If $\alpha = 1$ (as in the case of compound symmetry), $Q^{RSM}(k) = q_\infty + \frac{b}{k}$. Inserting $\alpha = 1$ into Equation 14, the rule for optimal k^* collapses to:

[4]The same logic with high T also holds in this case: if T is high enough, using all m forecasts is preferable.

$$k^* \approx \sqrt{\frac{b(T-1)}{\sigma_\varepsilon^2 + q_\infty}}$$

5 Monte Carlo Simulations

5.1 Monte Carlo Design

We evaluate the forecast-combination performance of RS relative to the equal-weighted mean via Monte Carlo experiments. The simulation design follows Samuels and Sekkel (2017) and Alvarez et al. (2012). In all experiments, the number of predictors is fixed at $N = 50$. We run $P = 500$ replications and consider the time series length $T \in \{100, 500\}$.

The data generating process (DGP) for independent variables x_t is specified as a dynamic factor structure:

$$x_t = F_t\Lambda + \epsilon_t, \tag{15}$$

where Λ is a column vector of N ones. F_t is defined as an AR(1) process:

$$F_t = \phi_f F_{t-1} + \epsilon_t^f \qquad \epsilon_t^f \overset{iid}{\sim} N(0,1)$$

We selected $\phi_f = \{0.3, 0.7\}$ to account for different persistence of F_t. Additionally, the error term in Equation 15 also follows the AR(1) process:

$$\epsilon_t = \phi_\eta \epsilon_{t-1} + \epsilon_t^\eta \qquad \epsilon_t^\eta \overset{iid}{\sim} N(0, \Sigma),$$

where the variance-covariance Σ is given by Toeplitz matrix of the form $\Sigma_{ij} = \rho^{|i-j|}$ or alternatively:

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho_s & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{pmatrix}.$$

The chosen form of Σ allows us to introduce correlation into the idiosyncratic components across different predictors. We consider $\rho = \{0, 0.8\}$. Similar to ϕ_f, we fix $\phi_\eta = \{0.3, 0.7\}$

Finally, the DGP for the dependent variable y_{t+1} is defined as a linear combination of predictors $\{x_{it}\}_{i=1}^N$:

$$y_{t+1} = \beta' x_t + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \overset{iid}{\sim} N(0,1)$$

The literature suggests multiple specifications for β (Samuels and Sekkel, 2017). In this paper, we restrict our attention to a case where the sparsity is introduced. To be more specific, in the simulations

$\beta = c[1_{1\times 10}, 0_{1\times(N-10)}]$ is assumed. Under this assumption, only the first ten variables – without loss of generality – have equal importance, whereas the rest of the predictors have zero loading in the DGP for y_{t+1}. Such sparsity is standard in macroeconomic applications, especially in forecasting.

Following Inoue and Kilian (2008), the scaling constant c is chosen to match target population R^2 values of $\{0.1, 0.3, 0.5\}$. This parameterization spans low-, medium-, and high-signal environments for the dependent variable.

Based on each x_{it}, an out-of-sample prediction for y_{t+1} is constructed with the help of a simple bivariate OLS regression:

$$y_{t+1} = \kappa_i x_{it} + \epsilon_{it}$$

Thus, the forecast pool consists of N univariate (bivariate-in-estimation) forecasts, each using only one predictor x_{it}, re-estimated with an expanding window at each origin. Later, these forecasts are combined via the RSM, the equal-weighted mean, or the optimal weighting and evaluated against the realized y_{t+1}. The RSM subset size is selected by cross-validation: we consider $k = \{2 \dots N\}$, where if $n = k$, the estimation collapses to full OLS. Conditional on the selected k, we execute the RSM 1,000 times and average the resulting predictions.

The evaluation sample is $T_{eval} = 20$ periods, and κ_i is estimated using an expanding window. We measure forecasting accuracy using Relative Root Mean Squared Forecast Error (rRMSFE), defined as follows:[5]

$$rRMSFE_{t+h} = \frac{\sqrt{\sum_{i=1}^{N}(y_{t+h} - \hat{y}_{t+h|t}^{RSM})^2}}{\sqrt{\sum_{i=1}^{N}(y_{t+h} - \hat{y}_{t+h|t}^{mean})^2}} = \frac{RMSFE_{t+h}^{RSM}}{RMSFE_{t+h}^{mean}}, \tag{16}$$

This metric facilitates comparison of the forecast-combination performance of the two approaches. A value below one indicates that RSM delivers more accurate forecasts than the standard equal-weighting scheme.

5.2 Monte Carlo Results

Figure 3 reports relative forecasting performance for $T = 100$, $\rho = 0.8$, $\phi_f = 0.3$, $\phi_\eta = 0.7$ first highlighting the role of R^2 in a small-sample setting. The vertical line corresponds to the $k = \sqrt{T}$ rounded to the nearest integer.

[5]Here, $h = 1$, as we focus on one-step-ahead forecasts in these simulations.

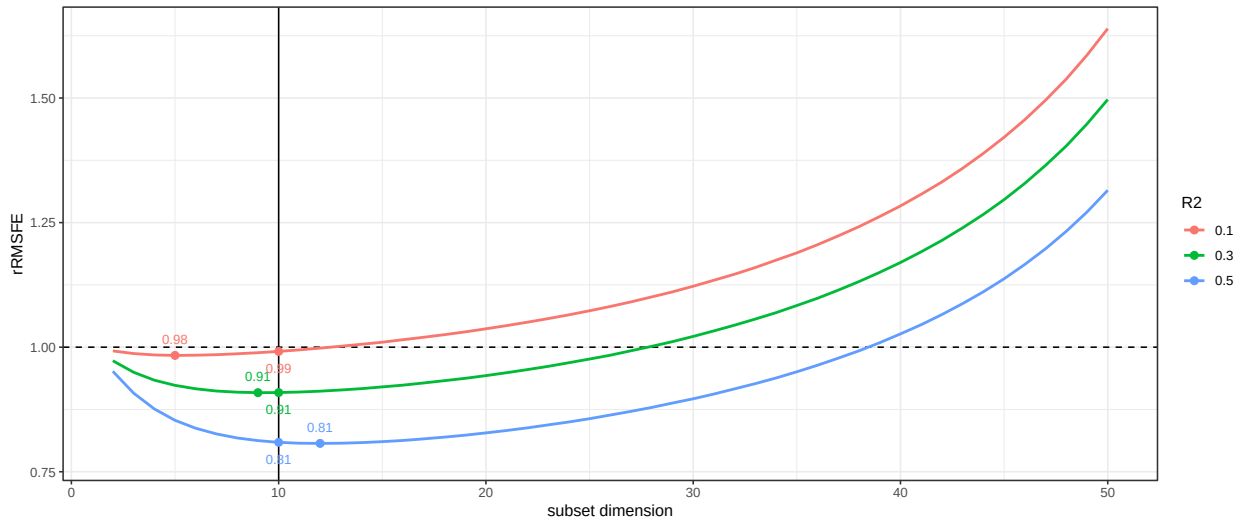


Figure 3: Relative forecasting performance ($T = 100$, $\rho = 0.8$, $\phi_f = 0.3$, $\phi_\eta = 0.7$)

The curves are (approximately) parallel, indicating that RSM's performance improves monotonically with R^2. This is unsurprising since higher values of R^2 imply that the predictors explain more of the target's variation. Accordingly, combining forecasts can improve accuracy.

It is worth noticing that for $R^2 \in \{0.3, 0.5\}$ the RSM dominates the equal-weighted mean. Their minimum values lie below one. By contrast, when $R^2 = 0.1$, gains are negligible: with a weak signal, individual forecasts have little predictive content, and their RSM aggregate is (almost) no better than the simple mean. Notice that by setting $k = \sqrt{T}$ the loss of gains is negligible compared to the true optimal subset size k^*. Those results are verified through all the scenarios, indicating that $\alpha \approx 1$ in the Equation 13. The rest of the plots depicting the relative forecasting performance can be found in Appendix D.

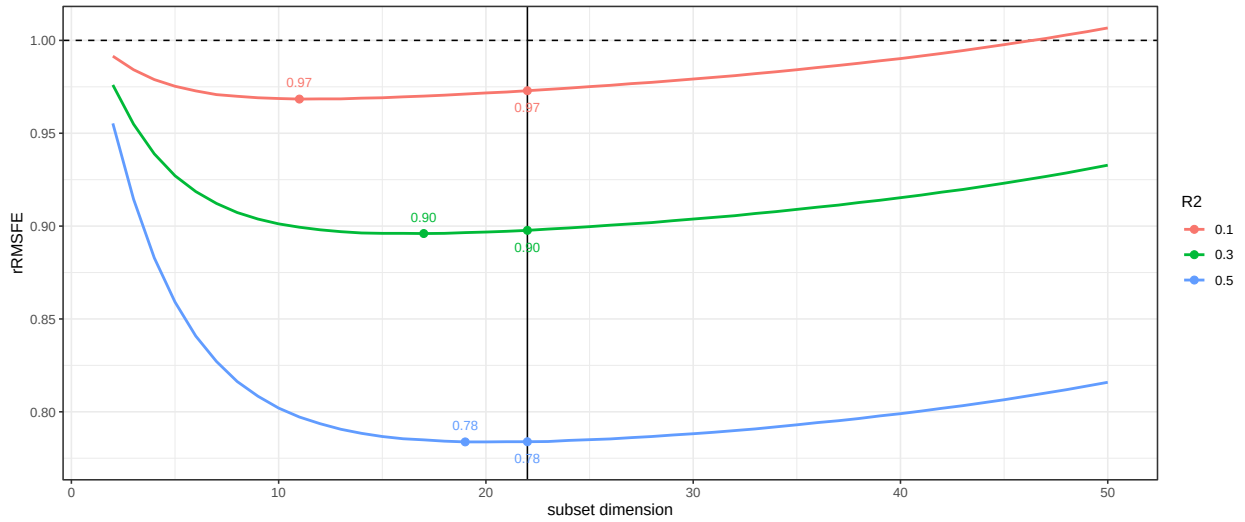


Figure 4: Relative forecasting performance ($T = 500$, $\rho = 0.8$, $\phi_f = 0.3$, $\phi_\eta = 0.7$)

Figure 4 presents relative forecasting performance for $T = 500$, $\rho = 0.8$, $\phi_f = 0.7$, $\phi_\eta = 0.3$. In

other words, the effect of more data availability can be analyzed since all other parameters (ρ, ϕ_f, ϕ_η remain unchanged. The larger sample size yields substantial gains from RSM over the equal-weighted mean, particularly at moderate and high R^2 values. When $R^2 = 0.5$, with other values fixed, the RSM can improve forecast combination results by almost 22%.

While tuning the optimal (or near-optimal) RSM subset size can be desirable for $T = 500$ and $R^2 \in \{0.3, 0.5\}$, a broad range of subset dimensions enhances forecasting accuracy (including $k = \sqrt{T}$). Across simulations, the optimal subset size either coincides with the rule-of-thumb choice or is nearby. Deviations from k^* yield only negligible changes in forecast error relative to the exact optimum.

The Monte Carlo setup also allows us to evaluate $Q^{RSM}(k)$. Figure 5 depicts the latter for the case when $T = 500$, $\rho = 0.8$, $\phi_f = 0.3$, $\phi_\eta = 0.7$. For all other scenarios, $Q^{RSM}(k)$ is also depicted, and the corresponding plots could be found in Appendix D.1.

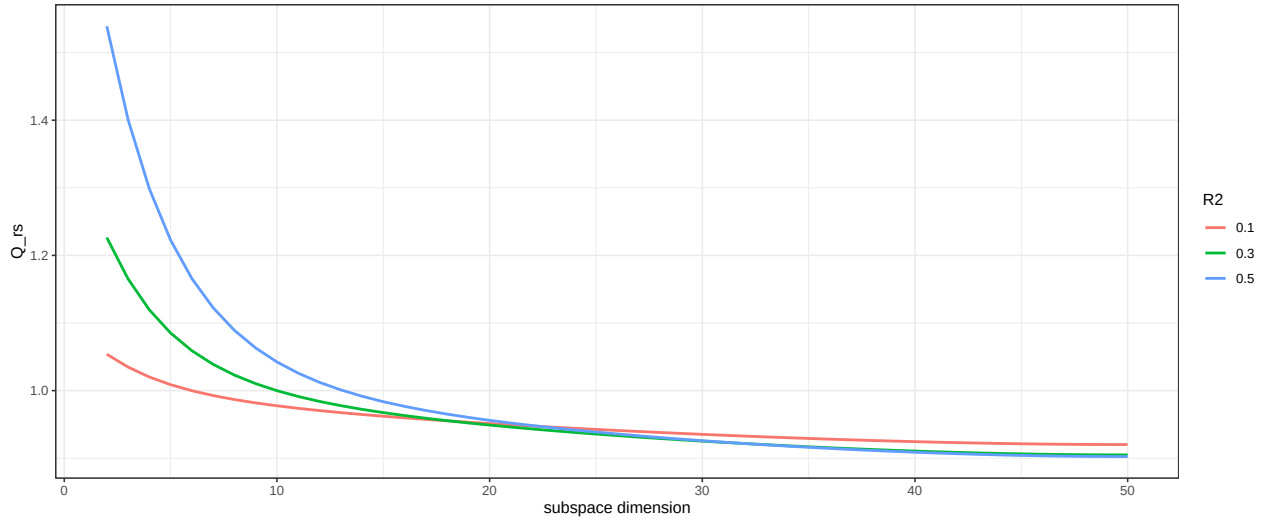


Figure 5: Q^{RSM} ($T = 500$, $\rho = 0.8$, $\phi_f = 0.3$, $\phi_\eta = 0.7$)

As expected, the lowest value of $Q^{RSM}(k)$ is attained at $k = 50$, i.e., when all available predictions are included in the combination (and this holds in all other cases as well). However, the performance improvement relative to smaller subsets becomes very small once k reaches moderate values. Overall, this result confirms that combining only a subset of k forecasts typically incurs little additional loss compared with using all N forecasts, while potentially reducing weight-estimation error.

6 Shapley Value Decomposition

One may explicitly pose a question of how important a given variable is to produce a predicted value of a target series. Interestingly, a possible answer to this question comes from a different field of study, namely game theory.

Shapley et al. (1953) introduced a solution to the problem of fairly dividing a joint pay-off across all individual players in a cooperative game. For almost six decades, this concept has had nothing to

do with forecasting literature. This connection remained unexplored until Strumbelj and Kononenko (2010) established a connection between players and variables used for forecasting, as well as the pay-off and a final prediction.

Following Buckmann et al. (2022), it is possible to define a contribution of variable k in observation x_i and model f:

$$\phi_k^S(x_i; f) = \sum_{x' \subseteq \mathcal{C}(x) \setminus \{k\}} \frac{|x'|!(|N| - |x'| - 1)!}{|N|!} [f(x_i|x' \cup \{k\}) - f(x_i|x')],$$

where $\mathcal{C}(x) \setminus \{k\}$ corresponds to the set of all possible combinations of variables excluding k, $|x'|$ is the number of variables included in the combination.

Similar to the initial idea, the SVD represents a weighted sum of marginal contributions of a given variable (player) to all possible combinations excluding this variable. Note that we se the SVD for forecast combination. It means that a given variable, in this case, is an individual forecast. Thus, the contribution of an individual forecast to the final prediction is obtained.

Buckmann et al. (2022) described obstacles while estimating Equation 6. There are two major concerns. As stated earlier, the number of possible combinations with l variables is 2^l, thus it grows exponentially. With sufficiently large l, evaluation of all possible combinations seems infeasible. Additionally, estimating the term $f(x_i|x')$ is usually challenging (unless feature independence is assumed). Both of these problems, in principle, can be dodged if one opts for tree-based (see Lundberg et al. (2020)) or linear frameworks. In the following subsection, we describe the SVD in a linear setting and establish a bridge between the SVD and RSM.

6.1 SVD and RSM

In this subsection, we introduce the connection between RSM and the SVD. In a general forecast combination framework:

$$y_{T+h} = w_{T+h|T,0} + \boldsymbol{w'_{T+h|T}} \hat{\boldsymbol{y}}_{T+h|T} + \epsilon_{T+h}, \tag{17}$$

If optimal weights are used, then $w_{T+h|T,0} = 0$ and $\sum \boldsymbol{w'_{T+h|T}} = 1$. However, as shown further, both restrictions do not influence the SVD decomposition. In other words, practically, it does not matter whether to exploit the restricted or unrestricted OLS combined with RSM.

The contribution ϕ_j of a j-th forecast $\hat{y}_{T+h|T,j}$ on the total prediction y_{T+h} is the difference between the feature effect and the average effect. More formally, one could express it as follows:

$$\phi_j = w'_{T+h|T,j} \hat{y}_{T+h|T,j} - E(w'_{T+h|T,j} X_j) = w'_{T+h|T,j} [\hat{y}_{T+h|T,j} - E(\hat{y}_{T+h|T,j})]$$

Let us combine this decomposition with RSM. It is straightforward to observe that in order to obtain ϕ_j, one only misses the OLS coefficient (restricted or unrestricted) $w'_{T+h|T,j}$. However, in a high-dimensional case, directly obtaining β_{x_j} may be complicated. RSM can assist in addressing this issue.

To illustrate the idea, suppose a variable x_j was selected 50 times within RSM. This implies that one has 50 different estimations of $w'_{T+h|T,j}$, each based on different subsamples. However, a term $[\hat{y}_{T+h|T,j} - E(\hat{y}_{T+h|T,j})]$ remains the same across all the iterations. Thus, it is possible to obtain an averaged contribution of ϕ_j by averaging across different realizations of estimated $w'_{T+h|T,j}$, as depicted below:

$$\bar{\phi_j} = \hat{w'}_{T+h|T,j}[\hat{y}_{T+h|T,j} - E(\hat{y}_{T+h|T,j})], \tag{18}$$

where $\hat{w'}_{T+h|T,j}$ is the average of (restricted or unrestricted) OLS coefficients corresponding to different draws.

Recall that under a standard linear framework, the SVD possesses a valuable characteristic – the difference between the final prediction and the historical average prediction is fairly distributed among the covariates (individual forecasts). Maintaining this property is crucial when combining the SVD with RSM, as it is not inherently fulfilled by the combination by default.

The explanation lies in the following observation: $\hat{w'}_{T+h|T,j}$ and the final prediction are averaged across different numbers. That is, $\hat{\beta}_{x_j}$ is divided by the n_{x_j} – the number of times a variable has been selected, meanwhile the final prediction is divided by B, where B is the number of RSM iterations.

More formally, define inclusion probability and conditional mean weight:

$$p_j := \Pr(j \in S), \qquad m_j := \mathbb{E}[\,\beta_j \mid j \in S\,].$$

Additionally, define the weights: zero-bagged weight (including 0 when the forecast is not selected):

$$\bar{w}_j^{\mathrm{zb}} := \mathbb{E}[\mathbf{1}\{j \in S\}\, w_j] \;=\; p_j\, m_j, \qquad \Rightarrow \quad \hat{y}_{t+1}^{RSM} = \sum_{i=1}^{N} \bar{w}_j^{\mathrm{zb}} \hat{y}_{j,t+1}$$

and the conditional weight (only when the forecast is selected in a subsample S) which is equal to the conditional mean:

$$\bar{w}_j^{\mathrm{nz}} := \mathbb{E}[w_j \mid j \in S] \;=\; m_j.$$

Thus, the average Shapley value for a j-th forecast uses the zero-bagged weight, that is:

$$\bar{\phi_j} = \bar{w}_j^{\mathrm{zb}}[\hat{y}_{T+h|T,j} - E(\hat{y}_{T+h|T,j})] \tag{19}$$

However, the sum of the contributions defined in Equation 19 explains the difference between the model prediction and its average prediction. To explain this difference in terms of the dependent variable ($\hat{y}_{T+h} - \bar{y}_t$), an additional intercept denoted ϕ_0 should be added such that:

$$\hat{y}_{T+h} - \bar{y}_t = \phi_0 + \sum_j \phi_j, \tag{20}$$

where $\phi_0 = \sum_j \bar{w}_j^{zb} E[\hat{y}_{T+h|T,j}] - \bar{y}_t$ is the deviation between the model's average prediction and expected model prediction of the dependent variable, ϕ_j are the average Shapley values as defined

in Equation 19. Notice that adding the constant term ϕ_0 does not influence the estimation of ϕ_j. Instead, it just ensures that the sum of ϕ_j explains the difference only in terms of the target variable y_t. Intuitively, the intercept term ϕ_0 reconciles two baselines: the model's average fitted value (implied by averaged subset-specific combinations) and the unconditional mean of the realized target series.

This notation also helps to derive the key rule for comparing global $\bar{w}_j^{\text{zb}}$ and conditional $\bar{w}_j^{\text{nz}}$ weights to the equal weighting. That is, if $p_j = k/N$ is a uniform probability, then there is a correspondence between zero-bagged, no-zeros, and mean as follows:

$$\bar{w}_j^{\text{zb}} = \frac{k}{N} m_j \quad \Longrightarrow \quad \boxed{\bar{w}_j^{\text{zb}} > \tfrac{1}{N} \iff m_j > \tfrac{1}{k}} \tag{21}$$

This means that the zero-bagged weight is higher than the mean weight if the conditional weight is higher that the reciprocal of the subset dimension k. In other words, given that the variable is in the subset S, its expected weight should be higher than the mean weight in the subsample for the global weight to be higher than the mean.

7 Empirical Application

7.1 Data and Methods

We employ the FRED-MD dataset to nowcast GDP growth rates. Real-time vintages of the last months of each quarter (March, June, September, December) are used to produce a forecast. This vintage selection ensures the fullest data availability before actually observing the outcome. The number of time series in the dataset varies over time, usually rounding up to 120. The optimal subset space k^* is set as proposed in the previous section, that is k^* as the nearest integer to $\sqrt{T}$. The validation of the subset choice k^* is presented in Appendix E.

We produce out-of-sample forecasts from Q3 1999 through Q1 2025, yielding 102 quarters for evaluation[6]. As in the Monte Carlo setup, we run separate bivariate OLS regressions —one for each predictor (aggregated to the quarterly frequency) — to obtain a unique forecast from each series. Thus, for a given quarter, we have as many forecasts as there are variables in the dataset. Current-quarter ragged-edge observations from the real time vintages (also aggregated to the quarterly frequency) are used to generate the out-of-sample nowcast from each predictor. At each vintage, the nowcast uses the latest available monthly releases within the quarter for each series; missing within-quarter observations are handled by the same aggregation rule used for all predictors to ensure comparability across vintages. To produce bivariate forecasts after the COVID-19-crisis (Q2 2020 and Q3 2020) as well as RSM and all the benchmarks forecasts, those two quarters were eliminated to avoid using outliers, which may severely affect the estimation. More specifically, for vintages after Q3 2020, we drop Q2 2020 and Q3 2020 from the estimation sample of every bivariate forecasting regression and

[6]Due to missing values in some of the indicators, Q4 2003 is discarded.

from the RSM weight estimation step, while still evaluating forecasts on the full out-of-sample period unless explicitly stated otherwise. For the baseline scenario, we fix $k = \sqrt{T}$ rounded to the nearest integer.

Other benchmarks for high-dimensional forecast combination are estimated to compare the performance. Following Chen and Maung (2023), we use Random Forest, Lasso, and Ridge.

7.2 Forecast Combination Evaluation

Table 1 reports GDP nowcast combination results comparing RSM with the benchmarks for the full sample. RSM substantially outperforms the equal weighting, Ridge, and Random Forest in terms of the RMSFE. Lasso, on the other hand, produces similar results, thus being the closest competitor. However, the time span used for forecast evaluation includes the COVID-19 crisis, which can influence the nowcasting performance of all models.

Model	MAFE	RMSFE
Mean	0.54	1.10
RSM	0.45**	0.74*
Lasso	0.51	0.76*
Ridge	0.50*	1.00
Random Forest	0.55	1.24

Table 1: GDP Forecast Evaluation: Full Sample

Table 2 provides the same metrics without the two COVID-19 quarters (Q2 and Q3 2020). Excluding those quarters noticeably improves the forecast combination performance across the models. Still, RSM shows the best forecast combination performance (both in terms of MAFE and RMSFE). However, unlike the case with full data, both ridge and Random Forest appear to be more accurate than Lasso, the performance of which is the same as the equal weighting. This indicates that Lasso performed substantially better during the COVID-19 crisis than other competitors (except the RSM).

Model	MAFE	RMSFE
Mean	0.42	0.60
RSM	0.39**	0.53**
Lasso	0.45	0.60
Ridge	0.39**	0.55**
Random Forest	0.41**	0.58*

Table 2: GDP Forecast Evaluation: Excluding COVID-19

Recall, RSM takes an initial set of N forecasts as input and uses them to produce another set of forecasts of size B, where B is the number of iterations (in our case, $B = 1000$). Therefore, there is supposed to be a change in the distributions of forecasts, which helps RSM to predict the target better than the mean. However, this might happen only when both equal weighting and RSM produce substantially different results.

First, Figure 6 and Figure 7 depict both distributions for Q3 2004. This quarter is selected due to the similar prediction of the GDP growth rate for both methods[7]. Even though the RSM distribution is closer to the bell-shaped form, both are centered around the true realization without outliers. That is why taking the mean of both sets produces accurate estimations.

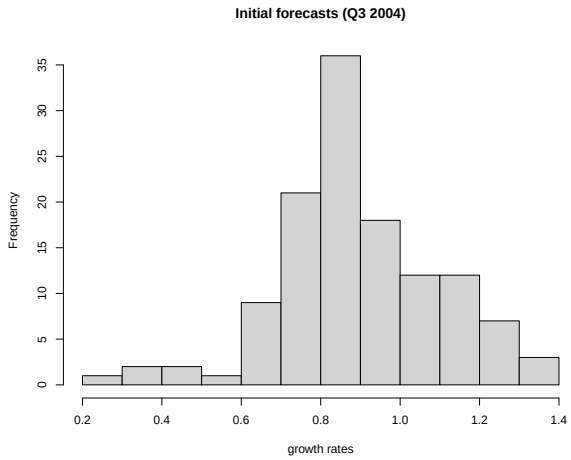


Figure 6: Initial set of GDP forecasts (Q3 2004)

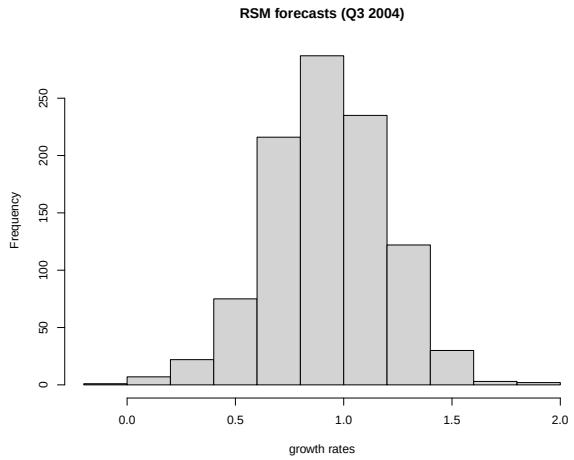


Figure 7: RS set of GDP forecasts (Q3 2004)

In crisis periods, RSM delivers significantly stronger out-of-sample performance (Figure 8 and Figure 9). Whereas many individual forecasts remain clustered near zero and fail to adjust, RSM adapts by shifting the combined forecast toward a more plausible mean level. Thus, RSM is particularly valuable when the target temporarily deviates from its typical behavior.

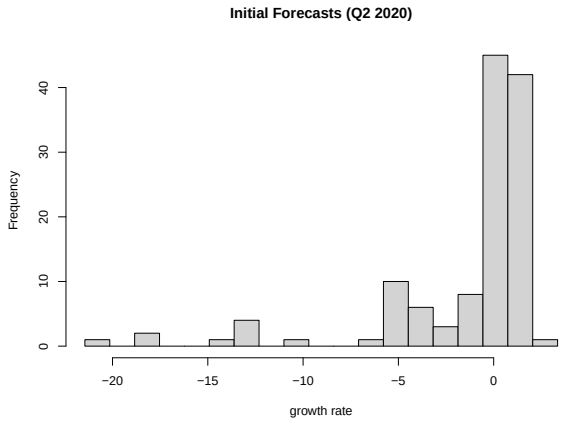


Figure 8: Initial set of GDP forecasts (Q2 2020)

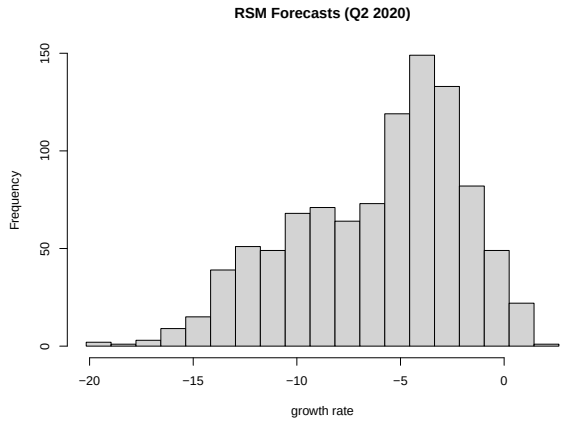


Figure 9: RS set of GDP forecasts (Q2 2020)

7

7.3 RSM Weights

RSM allows us to study the weight dynamics over time for any given variable and how it differs from the equal weighting scheme. Moreover, it is possible to study the global weight $\bar{w}_j^{zb}$ and its conditional counterpart m_j as defined per Equation 21. In the main analysis, we stick to the global weight $\bar{w}_j^{zb}$ used to produce the final forecast. In Appendix F, a more detailed discussion of the correspondence between the global and conditional weights is presented.

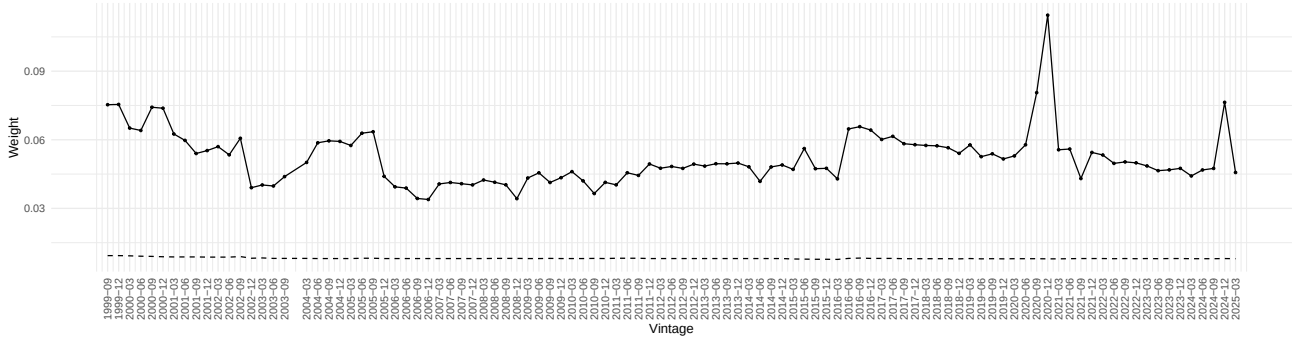


Figure 10: Industrial Production Weight over Time

We choose Industrial Production (FRED-md mnemonics INDPRO) to analyze the time-varying dynamics. This variable is chosen as one of the most important leading indicators used to nowcast/forecast GDP growth rates. IP is a standard real-activity indicator that tends to co-move with GDP and is widely used in short-horizon output tracking exercises. As shown in Figure 10, the solid line is always above the dotted one, meaning that Industrial Production always has a larger in absolute value than equal weight in the aggregated prediction. This confirms that forecasts based solely on IP are more informative and have stronger forecasting predictability than an average one, thus is supposed to have a higher weight in the optimal combination. Moreover, by implying the equal weighting, not only does one lose the flexibility, but also it influences the total relative contribution via the SVD as shown in Equation 18. Underestimating (as in this case) the weight of Industrial production leads to the corresponding underestimation of its contribution to the GDP growth rate which may negatively impact public policy decisions.

7.3.1 Negative Weights

In the forecast combination literature, a common assumption is that the weights should be non-negative and sum up to one in order to preserve the weight interpretability. The optimal weight, as defined in Equation 2, on the other hand, does not rely on the additional constraint of weights being non-negative. In other words, the optimal weights obtained via linear combination without additional assumptions can be negative. Moreover, if a large number of variables is included in the regression with the non-negativity constraints, it can lead to the case when most of the forecasts have a zero

weight, which can underestimate their actual weight.

The negative weights can cause ambiguity in terms of interpretability. However, as shown in Radchenko et al. (2023), negative weights can arise (and it should be expected) in the case when forecasts to combine are highly correlated because of the multicollinearity. Since we do not impose a restriction of global weights to be non-negative, some of them in the final combination happened to be negative. For example, let us take the S&P 500 index (FRED-md mnemonics S_P_500), which is considered an important financial variable to predict the GDP growth rate Koop et al. (2021). In Figure 11 (upper plot), the global weight $\bar{w}^{zb}$ of the latter is depicted.

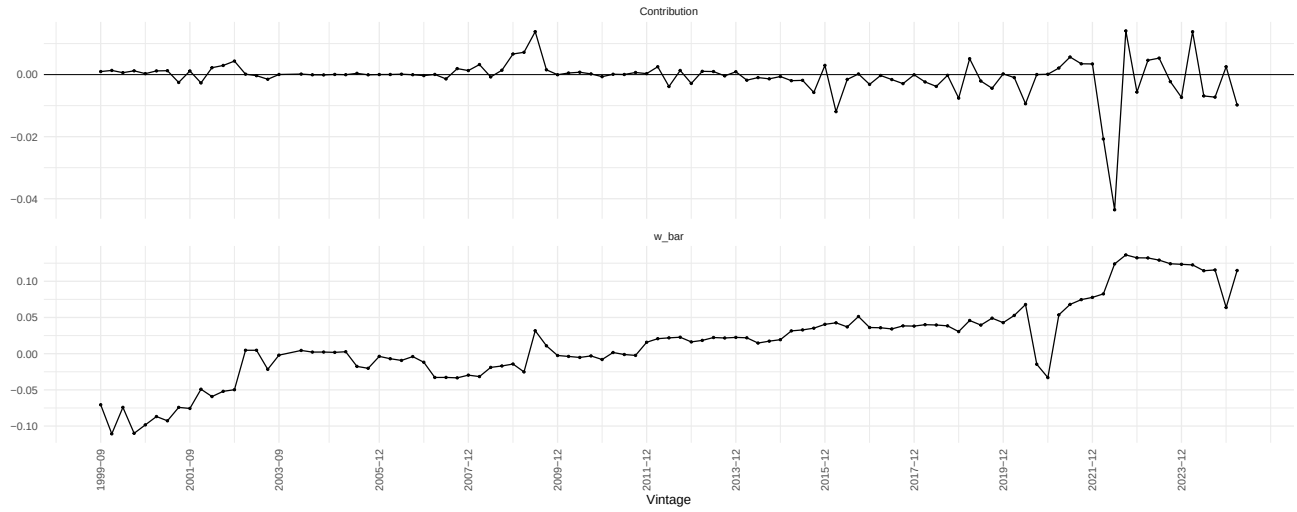


Figure 11: S&P 500: RSM global Weight and its Contribution

As can be seen from the plot, there are periods at the beginning of the out-of-sample exercise, when the RSM weights are indeed negative. It is worth mentioning that it happens quite often that some of the weights for different forecasts are less than 0 under the RSM forecast aggregation procedure. Again, according to Radchenko et al. (2023), it is expected to happen, since some of the employed forecasts in most subsets can be highly collinear, thus forcing the global weight to be negative.

We propose two alternative ways of interpreting negative weights within the RSM procedure. First, according to the Equation 17, when the simple OLS is estimated including a constant to capture a potential bias (if all the forecasts are unbiased, this constant is essentially 0), the obtained estimations are not interpreted as weights per se. Rather, they depict the OLS coefficients (the marginal change), and the focus is on the forecast combination performance instead of the interpretability of weights.

Second, we treat the weights as the input to compute the Shapley value of a given variable. This is, to the best of our knowledge, is an understudied property of the negative weights. The negative weights still allow for decomposing the contribution in such a way that the difference between the actual prediction and the historical average is fully explained. In other words, instead of focusing on the weight per se, we propose to focus on the Shapley value decomposition, where the (negative) weight serves as a definer. Because the combination assigns a negative weight to this forecast, its

deviations from its own historical mean move the combined forecast in the opposite direction. A below-average forecast produces a positive contribution because the model is effectively subtracting that negative deviation.

In Figure 11 (lower plot), the corresponding attributions of the S&P 500 to the GDP growth rate are presented. Despite the negative weights at the start, its contribution was slightly positive, meaning that the deviation between the prediction and the historical average was negative. However, both the weight and the discrepancy being very close to 0 (even though both negative) result in the virtual absence of the contribution.

7.4 Shapley Value Decomposition

Once the linear decomposition values $\bar{\phi}_j$ are obtained for each variable, it is possible to plot them to observe the relative contributions. We choose the same quarters as in the Section 7.2, that is, Q3 2004 and Q2 2020.

In Figure 12 the ten highest (in absolute values) relative contributions are depicted for Q3 2004 as well as the sum of all other contributions, denoted "Other". This group aggregates the remaining forecasts' Shapley values (all contributions outside the top-10 in absolute value) so that the displayed bars still sum to the total explained deviation. As discussed in Section 7.2, in the given quarter, both the equal weighting and the RSM forecast combination methods provide similar results (0.89 and 0.94, respectively), essentially precisely predicting the outcome. Moreover, the actual growth rate of GDP in Q3 2004 (0.95) and the historical average of the latter (0.83) are very close to each other. Hence, the difference to be explained by the individual forecast is virtually small and $\phi_0 < 0.0007$. With the number of forecasts being high (124), a relative contribution of a given variable is expected to be close to 0.

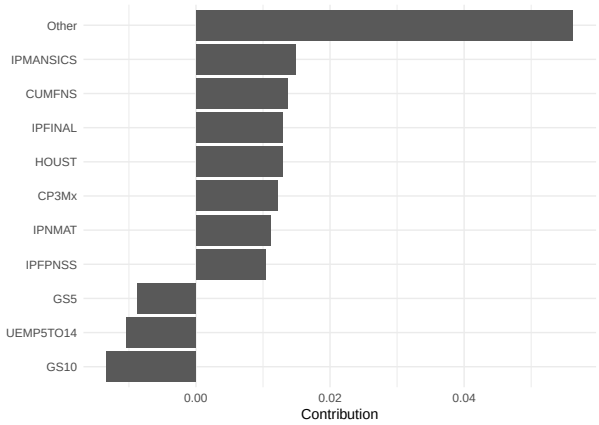


Figure 12: Shapley Value Decomposition (Q3 2004)

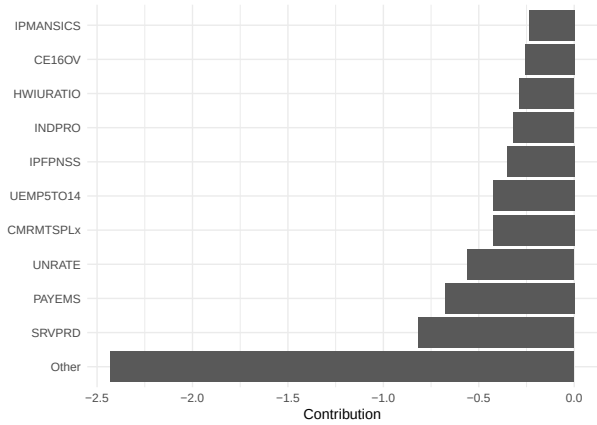


Figure 13: Shapley Value Decomposition (Q2 2020)

A more illustrative example is the same decomposition for Q2 2020, presented in Figure 13. Recall,

RSM significantly outperforms the mean combiner for a given quarter in terms of forecast precision. More specifically, RSM predicts the GDP growth rate to be -6.1, whereas the mean combination yields -1.3, hence severely underestimates the true realization (-7.9). The estimated $\phi_0 = 0.032$.

Since the discrepancy between the RSM prediction and its historical average is substantial, the magnitude of each variable's Shapley value is higher than in normal times. This decomposition also allows for interpreting the results descriptively. The highest contributions have labor market variables alongside production.

7.4.1 RSM and Equal Weight Contributions

Not only does one oftentimes lose the forecasting precision by using the mean, but also the corresponding average Shapley values ϕ_j can be underestimated, which can negatively impact public policy decisions. As in Section 7.3, the Industrial production and S&P index are chosen to depict potential difference while estimating relative contributions using RSM weights and the equal weighting.

In Figure 14, average Shapley Values for Industrial production obtained by RSM and the equal weighting are depicted. Since, as discussed before, the individual contribution is usually around 0 almost by construction. Still, the difference between the lines is noticeable. Especially, it is clear in both the COVID-19 recession and the following rebound (Q2 2020 and Q3 2020, respectively). In both quarters, Industrial production has a much higher average Shapley Values compare to the mean.

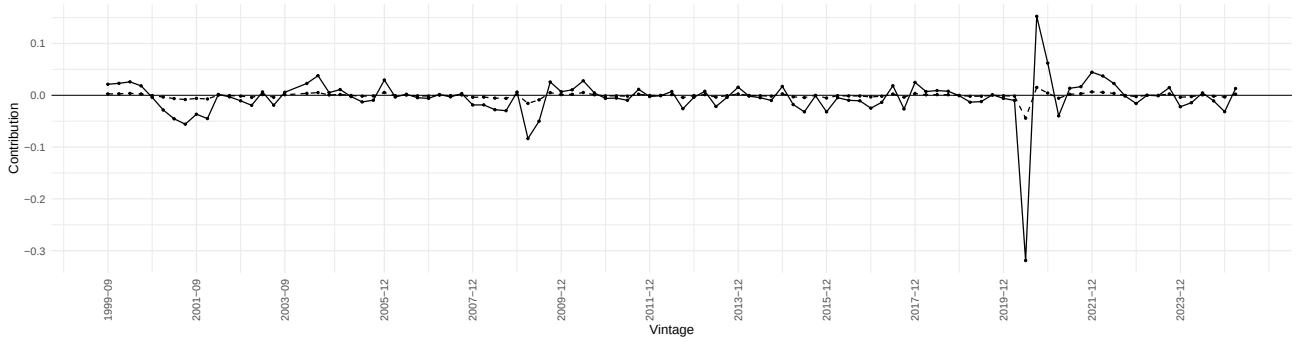


Figure 14: Industrial Production Contribution over Time

In Figure 15, average Shapley values obtained for S&P index by RSM and the equal weighting are depicted.

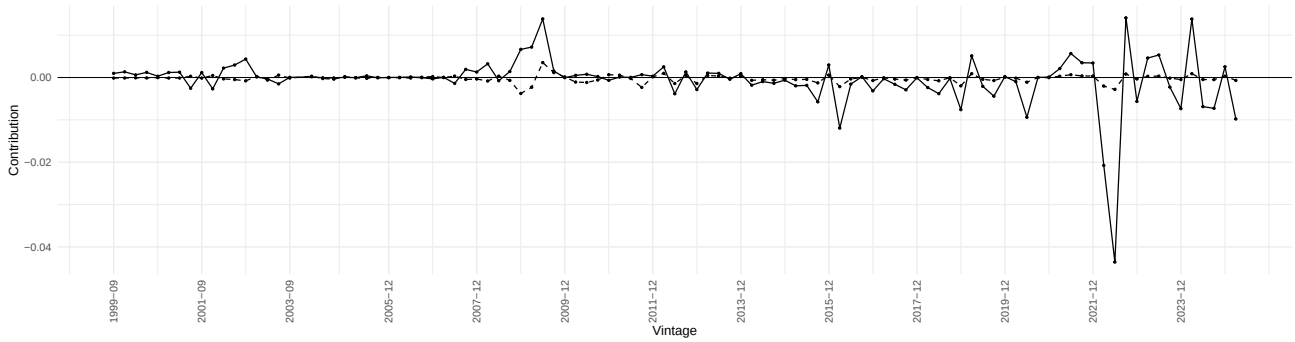


Figure 15: S& P Contribution over Time

8 Conclusion

This paper studies forecast aggregation in high-dimensional environments using RSM and its connection to the Shapley Value decomposition. The central idea is that instead of estimating a single large forecast-combination regression (often infeasible or noisy when the number of forecasts is large), we repeatedly draw small subsets of forecasts, estimate low-dimensional OLS combinations, and then average the resulting predictions. This strategy retains the information from large forecast pools while reducing estimation noise, which is a key friction behind the forecast combination puzzle and the weak empirical performance of many "optimal" weighting schemes.

Theoretical results show that there is room for RSM to improve forecast combination performance compared to the optimal weight and equal-weight frameworks in the finite (small) sample cases. Unsurprisingly, it is harder to beat the equal weighting because the latter does not imply the estimation error. Also, the performance of RSM crucially depends on the size of the subsets k. The ways of selecting k in a general case and in some special cases are discussed. In line with the simulation evidence, a convenient and robust rule of thumb is to choose the subset size close to $k \approx \sqrt{T}$ rounded to the nearest integer, which tends to be near-optimal across designs and sample sizes.

The empirical application uses real-time vintages of the FRED-MD data (provided by the Federal Reserve Bank of St. Louis) to nowcast U.S. GDP growth, constructing a large set of univariate forecasts and then combining them. RSM consistently improves on the equal-weight mean and performs favorably relative to high-dimensional benchmarks such as Lasso, Ridge, and Random Forest. The gains are particularly pronounced in periods of unusual dynamics (notably the COVID-19 recession), when many individual predictors react unevenly and the simple mean can be pulled toward systematic under- or over-reaction.

Finally, the paper connects forecast combination with interpretability by showing how RSM can be paired with a Shapley value decomposition. RSM weights allow the model's prediction to be easily decomposed into contributions of individual forecasts. This provides an interpretable attribution even in high dimensions and is crucial for policymakers.

References

Rocio Alvarez, Maximo Camacho, and Gabriel Perez-Quiros. Finite sample performance of small versus large scale dynamic factor models. 2012.

John M Bates and Clive WJ Granger. The combination of forecasts. *Journal of the operational research society*, 20(4):451–468, 1969.

Stephen D Bay. Combining nearest neighbor classifiers through multiple feature subsets. In *ICML*, volume 98, pages 37–45. Citeseer, 1998.

Tom Boot and Didier Nibbering. Forecasting using random subspace methods. *Journal of Econometrics*, 209(2):391–406, 2019.

Tom Boot and Didier Nibbering. Subspace methods. *Macroeconomic Forecasting in the Era of Big Data: Theory and Practice*, pages 267–291, 2020.

Marcus Buckmann, Andreas Joseph, and Helena Robertson. An interpretable machine learning workflow with an application to economic forecasting. Technical report, Bank of England, 2022.

Stephen Burgess, Emilio Fernandez-Corugedo, Charlotta Groth, Richard Harrison, Francesca Monti, Konstantinos Theodoridis, Matt Waldron, et al. *The Bank of England's forecasting platform: COMPASS, MAPS, EASE and the suite of models*. SSRN, 2013.

Bin Chen and Kenwin Maung. Time-varying forecast combination for high-dimensional data. *Journal of Econometrics*, 237(2):105418, 2023.

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

Lilian M De Menezes, Derek W Bunn, and James W Taylor. Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, 120(1):190–204, 2000.

Viet Hoang Dinh, Didier Nibbering, and Benjamin Wong. Random subspace local projections. *Review of Economics and Statistics*, pages 1–33, 2024.

Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006.

Graham Elliott and Jie Liao. Combining forecasts-on why averaging beats optimal linear weights. 2025.

Graham Elliott, Antonio Gargano, and Allan Timmermann. Complete subset regressions. *Journal of Econometrics*, 177(2):357–373, 2013.

Graham Elliott, Antonio Gargano, and Allan Timmermann. Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54:86–110, 2015.

Domenico Giannone, Michele Lenza, and G Primiceri. Macroeconomic prediction with big data: The illusion of sparsity. *echnical report, Federal Reserve Bank of New York*, 97, 2017.

Clive WJ Granger and Ramu Ramanathan. Improved methods of combining forecasts. *Journal of forecasting*, 3(2):197–204, 1984.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

Atsushi Inoue and Lutz Kilian. How useful is bagging in forecasting economic time series? a case study of us consumer price inflation. *Journal of the American Statistical Association*, 103(482): 511–522, 2008.

Dennis Kant, Andreas Pick, and Jasper de Winter. Nowcasting gdp using machine learning methods. *AStA Advances in Statistical Analysis*, 109(1):1–24, 2025.

Gary Koop, Stuart McIntyre, James Mitchell, and Aubrey Poon. Nowcasting 'true'monthly us gdp during the pandemic. *National Institute Economic Review*, 256:44–70, 2021.

Rachidi Kotchoni, Maxime Leroux, and Dalibor Stevanovic. Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics*, 34(7):1050–1072, 2019.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Michael W McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

Marcelo C Medeiros, Gabriel FR Vasconcelos, Álvaro Veiga, and Eduardo Zilberman. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119, 2021.

Loukia Meligkotsidou, Ekaterini Panopoulou, Ioannis D Vrontos, and Spyridon D Vrontos. Out-of-sample equity premium prediction: A complete subset quantile regression approach. *The European Journal of Finance*, 27(1-2):110–135, 2021.

Andreas Pick and Matthijs Carpay. Multi-step forecasting with large vector autoregressions. In *Essays in Honor of M. Hashem Pesaran: Prediction and Macro Modeling*, pages 73–98. Emerald Publishing Limited, 2022.

Peter Radchenko, Andrey L Vasnev, and Wendun Wang. Too similar to combine? on negative weights in forecast combination. *International Journal of Forecasting*, 39(1):18–38, 2023.

Jon D Samuels and Rodrigo M Sekkel. Model confidence sets and forecast combination. *International Journal of Forecasting*, 33(1):48–60, 2017.

Lloyd S Shapley et al. A value for n-person games. 1953.

Martin Slawski. On principal components regression, random projections, and column subsampling. 2018.

James H Stock and Mark W Watson. Forecasting inflation. *Journal of monetary economics*, 44(2): 293–335, 1999.

James H Stock and Mark W Watson. Combination forecasts of output growth in a seven-country data set. *Journal of forecasting*, 23(6):405–430, 2004.

Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Allan Timmermann. Forecast combinations. *Handbook of economic forecasting*, 1:135–196, 2006.

Xiaoqian Wang, Rob J Hyndman, Feng Li, and Yanfei Kang. Forecast combinations: an over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547, 2023.

Nuttanan Wichitaksorn, Yingyue Kang, and Faqiang Zhang. Random feature selection using random subspace logistic regression. *Expert Systems with Applications*, 217:119535, 2023.

Appendices

A General Solution to the Minimization Problem

We would like to minimize the expected loss with respect to k which has the following form:

$$L^{RSM} = (\sigma_\varepsilon^2 + Q^{RSM}(k)) \left(1 + \frac{k-1}{T}\right) \rightarrow min,$$

Notice that the constant factor $\frac{1}{T}$ does not influence the solution. Thus, we can rewrite the minimization function as follows:

$$L^{RSM} = (\sigma_\varepsilon^2 + Q^{RSM}(k)) (T + k - 1) \rightarrow min,$$

Differentiation with respect to k yields:

$$\frac{\mathrm{d}L^{RSM}}{\mathrm{d}k} = \sigma_\varepsilon^2 + T\frac{\mathrm{d}Q^{RSM}}{\mathrm{d}k} + Q^{RSM} + k\frac{\mathrm{d}Q^{RSM}}{\mathrm{d}k} - \frac{\mathrm{d}Q^{RSM}}{\mathrm{d}k} = 0$$

After re-arranging terms, we obtain the following condition:

$$\frac{\mathrm{d}Q^{RSM}(k)}{\mathrm{d}k^*} = -\frac{Q^{RSM}(k^*) + \sigma_\epsilon^2}{T + k^* - 1}$$

B Special Case: Compound Symmetry

Assume that for any subset of size k:

$$\tilde{\Sigma}_S = \sigma_f^2((1-\rho)I_k + \rho\iota_k\iota_k') \quad 0 \leq \rho < 1$$

Using the Sherman-Morrison formula, it is possible to define $\tilde{\Sigma}_S^{-1}$ as follows:

$$\tilde{\Sigma}_S^{-1} = \frac{1}{\sigma_f^2}\frac{1}{1-\rho}\left(I - \frac{\rho}{1+(k-1)\rho}\iota_k\iota_k'\right)$$

$$\iota_k'\tilde{\Sigma}_S^{-1}\iota_k = \frac{1}{\sigma_f^2}\frac{1}{1-\rho}\left(\iota_k'I_k\iota_k - \frac{\rho}{1+(k-1)\rho}\iota_k'(\iota_k\iota_k')\iota_k\right)$$

Notice that $\iota_k'I_k\iota_k = k$ and $\iota_k'(\iota_k\iota_k')\iota_k = k^2$. This leads to the following simplification:

$$\iota_k'\tilde{\Sigma}_S^{-1}\iota_k = \frac{1}{\sigma_f^2}\frac{1}{1-\rho}\left(k - \frac{\rho k^2}{1+(k-1)\rho}\right) = \frac{1}{\sigma_f^2}\left(\frac{k}{1+(k-1)\rho}\right)$$

Thus, the inverse of the expression yields:

$$Q^{RSM}(k) = \left(\iota_k'\tilde{\Sigma}_S^{-1}\iota_k\right)^{-1} = \left(\frac{1-\rho}{k} + \rho\right)\sigma_f^2$$

Then, minimization problem has a following form:

$$L^{RSM} = \left(\sigma_\varepsilon^2 + \left(\frac{1-\rho}{k} + \rho\right)\sigma_f^2\right)\left(1 + \frac{k-1}{T}\right) \to min,$$

Differentiation with respect to k yields:

$$\frac{\mathrm{d}L^{RSM}}{\mathrm{d}k} = \sigma_\varepsilon^2 + \rho\sigma_f^2 - \frac{(T-1)(1-\rho)\sigma_f^2}{k^2} = 0$$

Thus, k^* which minimizes the loss is described as follows:

$$k^* = \sqrt{\frac{(T-1)(1-\rho)}{\rho + \frac{\sigma_\varepsilon^2}{\sigma_f^2}}}$$

C Special Case: Functional form of $Q^{RSM}(k)$

The minimization problem has a following form:

$$L^{RSM} = (\sigma_\varepsilon^2 + \underbrace{q_\infty + \frac{b}{k^\alpha}}_{Q^{RSM}(k)}) \left(1 + \frac{k-1}{T}\right) \to min,$$

where $q_\infty > 0, b > 0, \alpha \geq 1$.

Since the factor $\frac{1}{T}$ does not influence the solution, we can rewrite the problem as follows:

$$L^{RSM} = (\sigma_\varepsilon^2 + q_\infty + \frac{b}{k^\alpha})\,(T + k - 1) \to min$$

Define $u(k) = \left(\sigma_\varepsilon^2 + q_\infty + \frac{b}{k^\alpha}\right)$ and $g(k) = \left(1 + \frac{k-1}{T}\right)$. The minimization problem can be rewritten as $L^{RSM} = u(k)g(k)$. Using the product rule, the first order condition is:

$$\frac{\mathrm{d}L^{RSM}}{\mathrm{d}k} = \frac{\mathrm{d}u(k)}{\mathrm{d}k} g(k) + \frac{\mathrm{d}g(k)}{\mathrm{d}k} u(k) = 0$$

Substituting $\frac{\mathrm{d}u(k)}{\mathrm{d}k} = -b\alpha k^{-(\alpha+1)}$ and $\frac{\mathrm{d}g(k)}{\mathrm{d}k} = 1$, we obtain:

$$\frac{\mathrm{d}L^{RSM}}{\mathrm{d}k} = -b\alpha k^{-(\alpha+1)}(T + k - 1) + (\sigma_\varepsilon^2 + q_\infty + \frac{b}{k^\alpha}) = 0$$

Multiplying both sides by $k^{\alpha+1}$ yields:

$$(\sigma_\varepsilon^2 + q_\infty)k^{\alpha+1} + bk - \alpha b(T + k - 1) = 0$$

After rearranging terms, we obtain:

$$(\sigma_\varepsilon^2 + q_\infty)k^{\alpha+1} + (1 - \alpha)bk = \alpha b(T - 1) \tag{22}$$

The optimal subset size k^* depends critically on the decay rate α in the functional form $Q_{RSM}(k) = q_\infty + \frac{b}{k^\alpha}$. Here we provide solutions for three cases: $\alpha = 2$ (fast decay), $\alpha = 1$ (moderate decay, baseline case), and $\alpha = 1/2$ (slow decay).

Case 1: $\alpha = 1$ (Moderate Decay)

This is the baseline case derived in the main text. Substituting $\alpha = 1$ into Equation 22:

$$(\sigma_\varepsilon^2 + q_\infty)k^2 = b(T - 1)$$

Solving for k yields:

$$\tilde{k}^* = \sqrt{\frac{b(T-1)}{\sigma_\varepsilon^2 + q_\infty}}$$

Under compound symmetry with equal pairwise correlation ρ, we have $q_\infty = \rho\sigma_f^2$ and $b = (1-\rho)\sigma_f^2$ (see Appendix B), yielding:

$$\tilde{k}^* = \sqrt{\frac{(T-1)(1-\rho)}{\rho + \frac{\sigma_\varepsilon^2}{\sigma_f^2}}}$$

Case 2: $\alpha = 2$ **(Fast Decay)**

Substituting $\alpha = 2$ into Equation 22:

$$(\sigma_\varepsilon^2 + q_\infty)k^3 - bk = 2b(T-1)$$

After rearranging the terms, we obtain:

$$k^3 - \frac{b}{\sigma_\varepsilon^2 + q_\infty}k - \frac{2b(T-1)}{\sigma_\varepsilon^2 + q_\infty} = 0$$

This is a depressed cubic equation of the form $k^3 + pk + q = 0$ where:

$$p = -\frac{b}{\sigma_\varepsilon^2 + q_\infty}$$
$$q = -\frac{2b(T-1)}{\sigma_\varepsilon^2 + q_\infty}$$

Using the Cardano's formula, the real solution is as follows:

$$\tilde{k}^* = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}$$

Substituting back:

$$\tilde{k}^* = \sqrt[3]{\frac{b(T-1)}{\sigma_\varepsilon^2 + q_\infty} + \sqrt{\Delta}} + \sqrt[3]{\frac{b(T-1)}{\sigma_\varepsilon^2 + q_\infty} - \sqrt{\Delta}}$$

where the discriminant is:

$$\Delta = \left(\frac{b(T-1)}{\sigma_\varepsilon^2 + q_\infty}\right)^2 - \frac{b^3}{27(\sigma_\varepsilon^2 + q_\infty)^3}$$

For large T, the first term dominates and we obtain the asymptotic approximation:

$$\tilde{k}^* \approx \left(\frac{2b(T-1)}{\sigma_\varepsilon^2 + q_\infty}\right)^{1/3}$$

Case 3: $\alpha = 1/2$ **(Slow Decay)**

Substituting $\alpha = 1/2$ into Equation 22:

$$(\sigma_\varepsilon^2 + q_\infty)k^{3/2} + \frac{1}{2}bk = \frac{1}{2}b(T-1)$$

After rearranging, we get the following equation:

$$(\sigma_\varepsilon^2 + q_\infty)k^{3/2} = \frac{b}{2}\left[(T-1) - k\right]$$

For large T where $k \ll T$, we can approximate:

$$(\sigma_\varepsilon^2 + q_\infty)k^{3/2} \approx \frac{b(T-1)}{2}$$

Thus, solving for k yields:

$$\tilde{k}^* = \left[\frac{b(T-1)}{2(\sigma_\varepsilon^2 + q_\infty)}\right]^{2/3}$$

D Additional Discussion of Monte Carlo Simulations Results

Here, all the plots for both relative RMSFE as well as $Q^{RSM}(k)$ across the scenarios are presented.

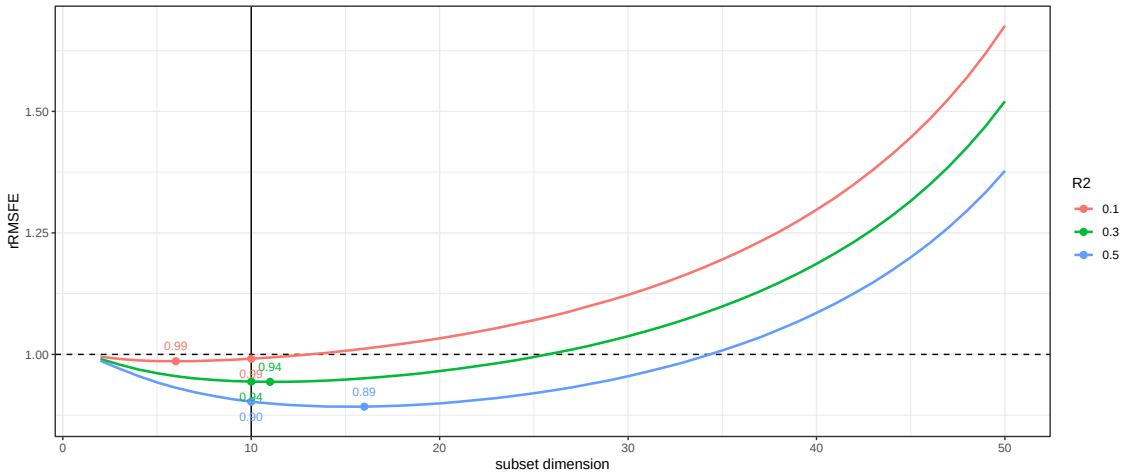


Figure 16: Relative forecasting performance ($T = 100$, $\rho = 0$, $\phi_f = 0.3$, $\phi_\eta = 0.7$)

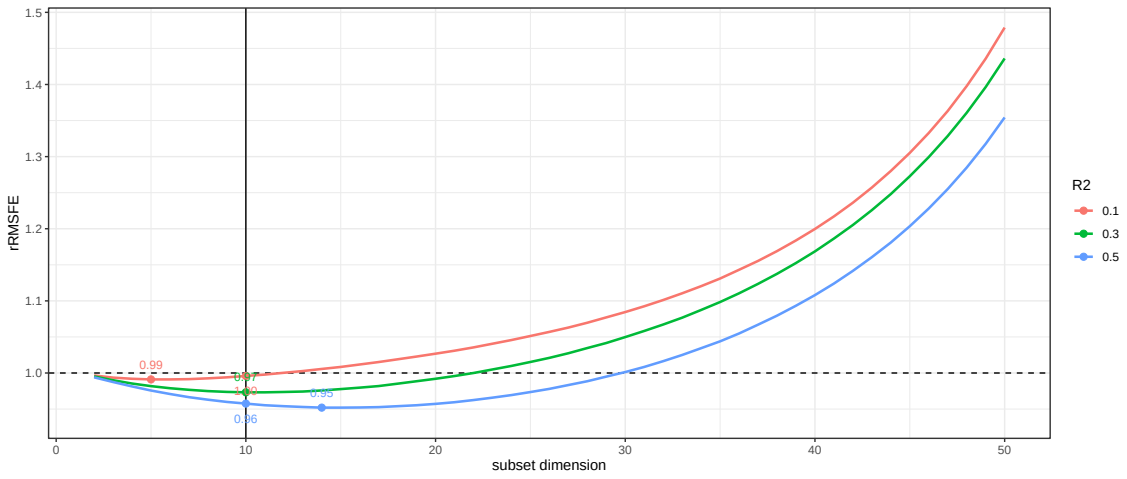


Figure 17: Relative forecasting performance ($T = 100$, $\rho = 0$, $\phi_f = 0.7$, $\phi_\eta = 0.3$)

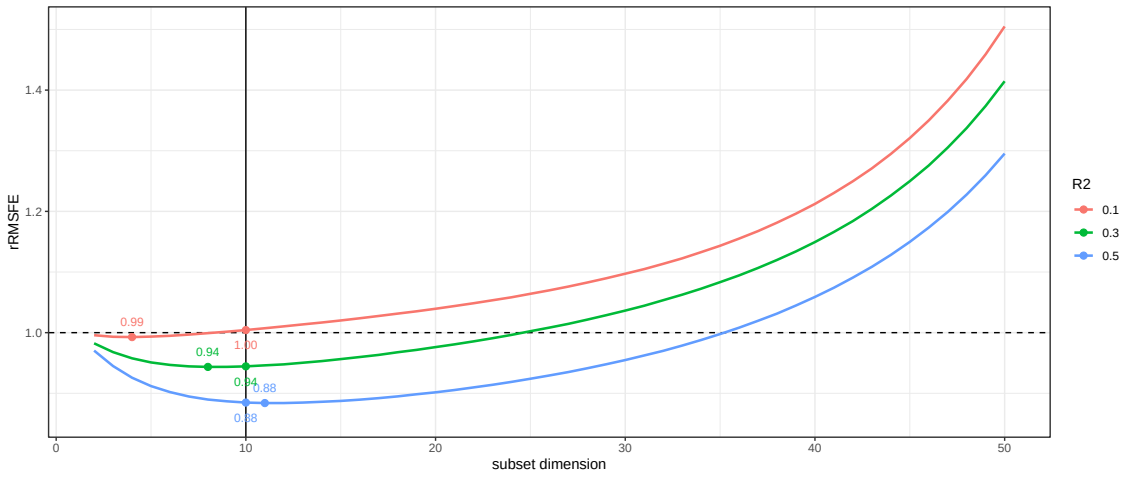


Figure 18: Relative forecasting performance ($T = 100$, $\rho = 0.8$, $\phi_f = 0.7$, $\phi_\eta = 0.3$)

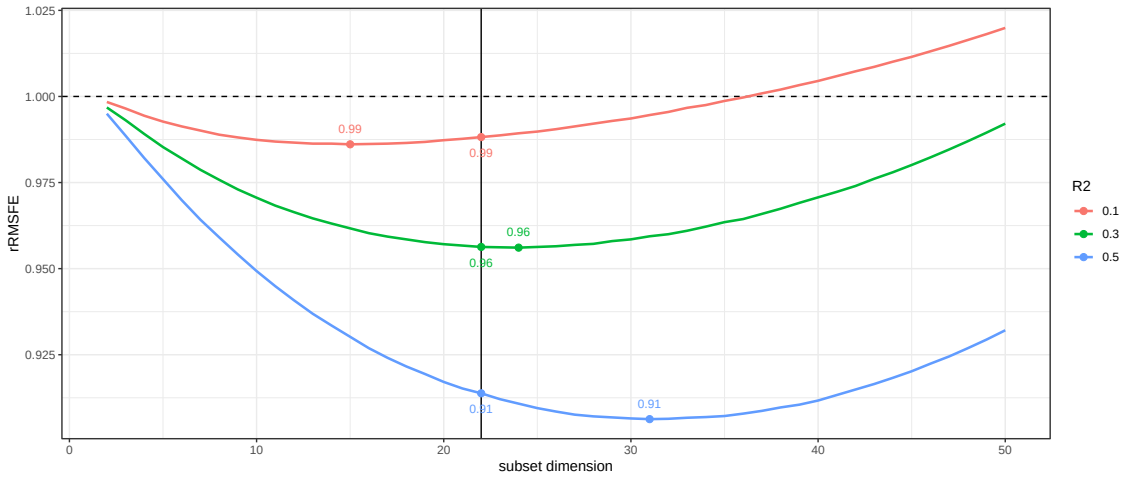


Figure 19: Relative forecasting performance ($T = 500$, $\rho = 0$, $\phi_f = 0.3$, $\phi_\eta = 0.7$)

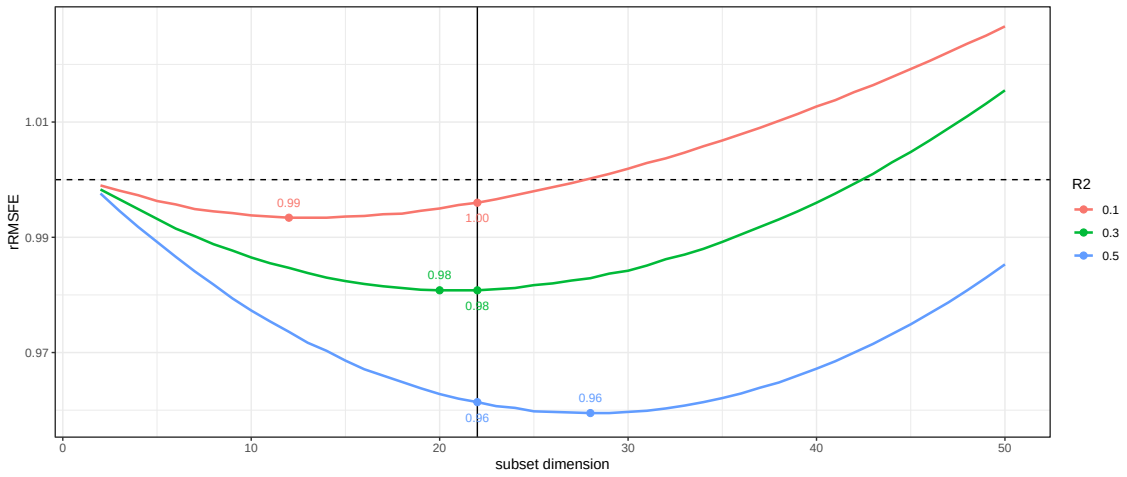


Figure 20: Relative forecasting performance ($T = 500$, $\rho = 0$, $\phi_f = 0.7$, $\phi_\eta = 0.3$)

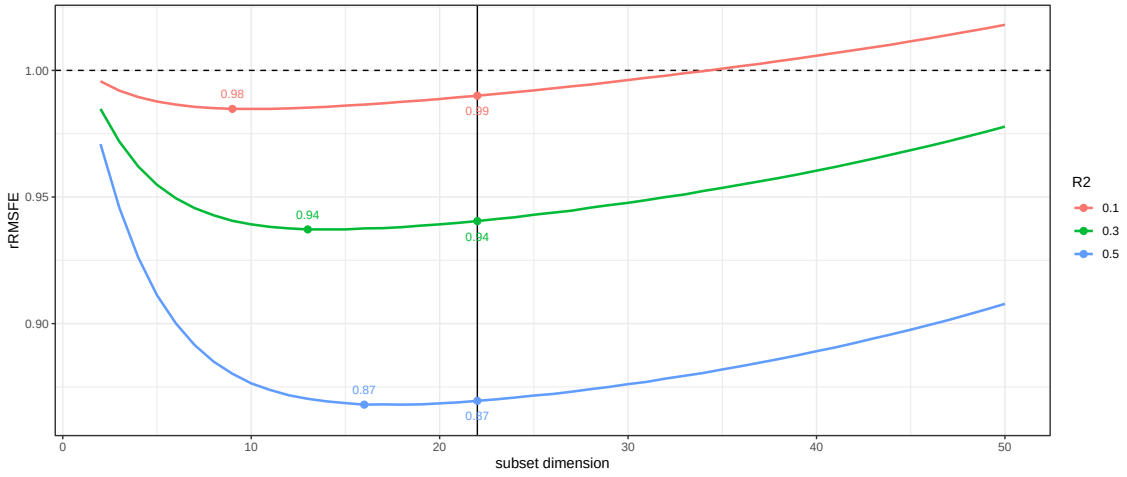


Figure 21: Relative forecasting performance ($T = 500$, $\rho = 0.8$, $\phi_f = 0.7$, $\phi_\eta = 0.3$)

D.1 $Q^{RSM}(k)$ for Different Scenarios

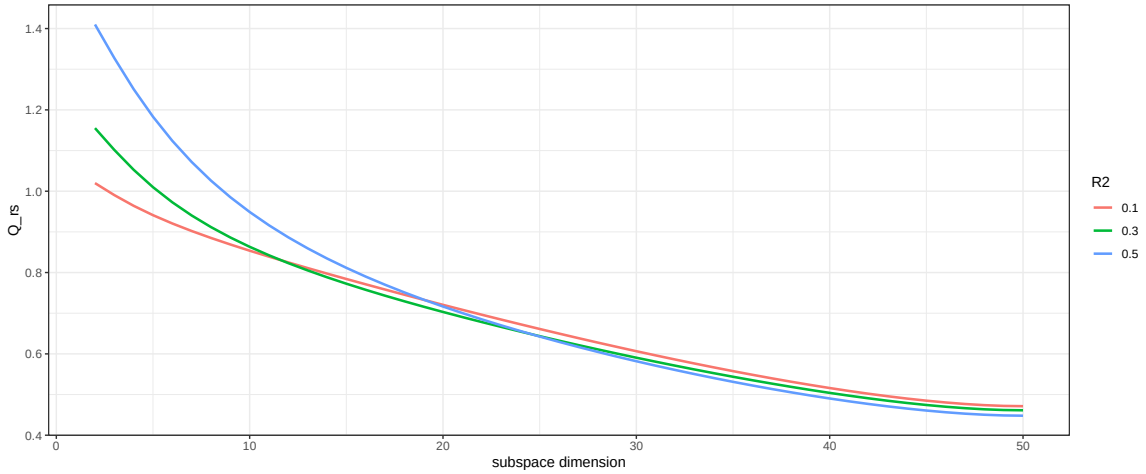


Figure 22: Q^{RSM} $(T = 100,\ \rho = 0,\ \phi_f = 0.3,\ \phi_\eta = 0.7)$

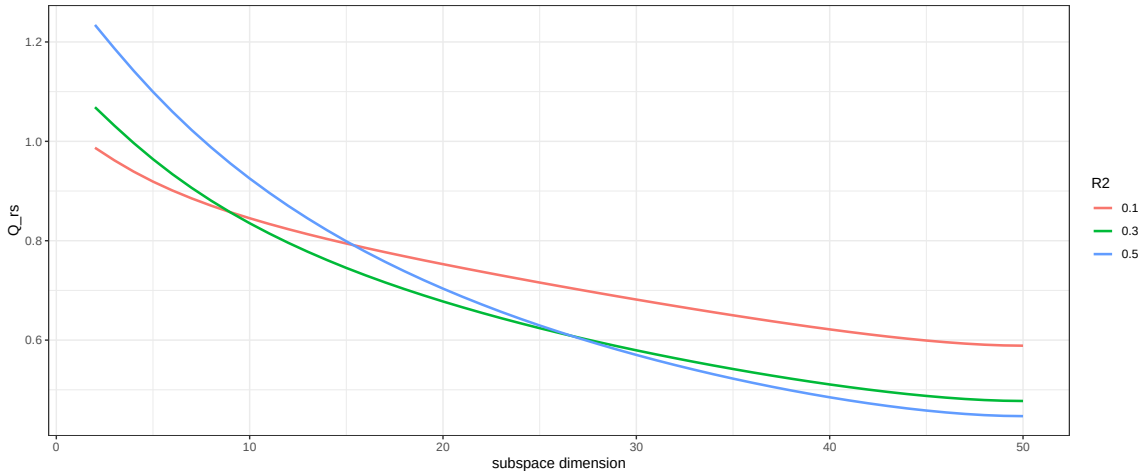


Figure 23: Q^{RSM} $(T = 100,\ \rho = 0,\ \phi_f = 0.7,\ \phi_\eta = 0.3)$

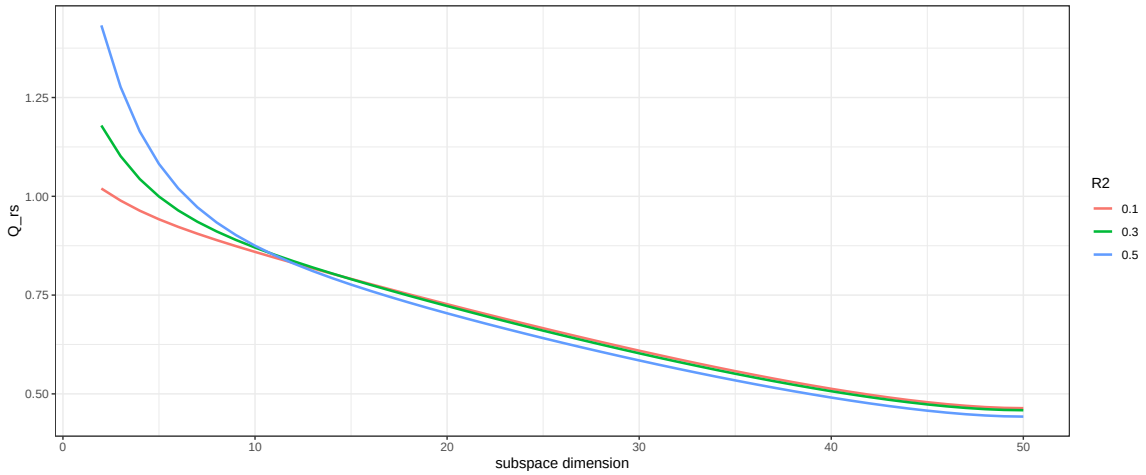


Figure 24: Q^{RSM} $(T = 100,\ \rho = 0.8,\ \phi_f = 0.3,\ \phi_\eta = 0.7)$

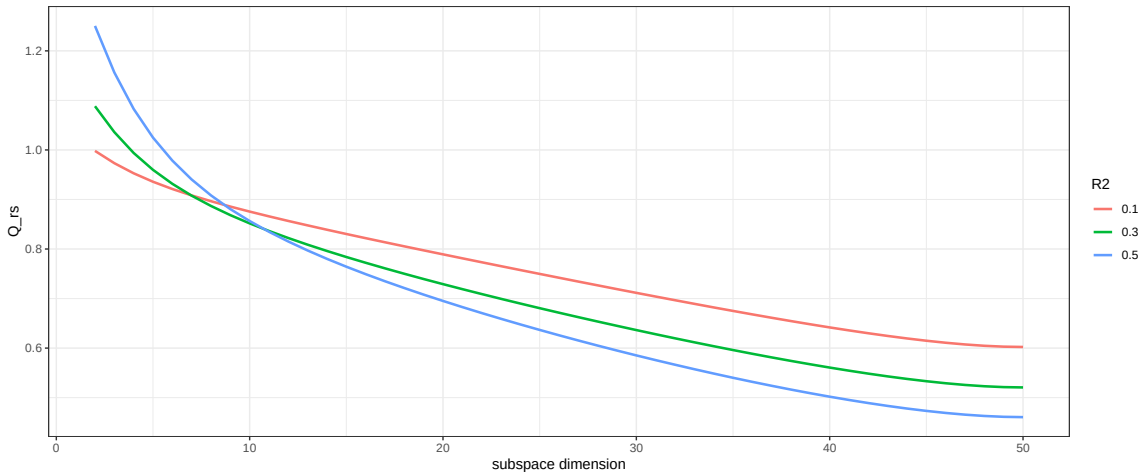


Figure 25: Q^{RSM} $(T = 100,\ \rho = 0.8,\ \phi_f = 0.7,\ \phi_\eta = 0.3)$

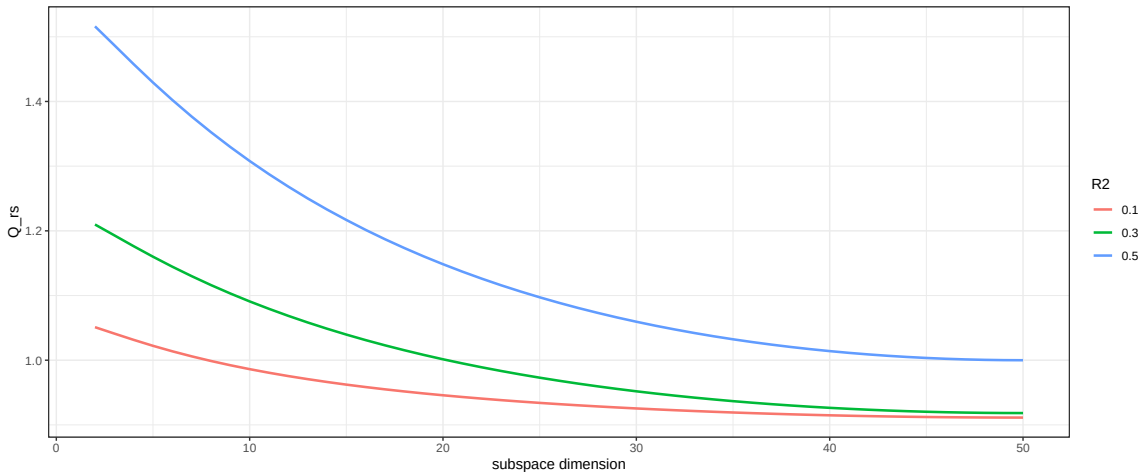


Figure 26: Q^{RSM} $(T = 500,\ \rho = 0,\ \phi_f = 0.3,\ \phi_\eta = 0.7)$

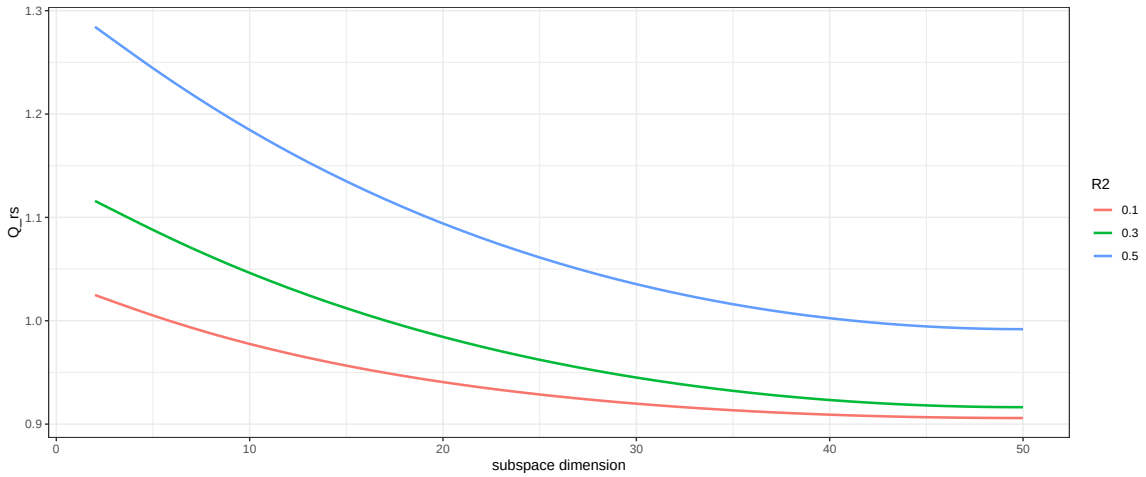


Figure 27: Q^{RSM} $(T = 500,\ \rho = 0,\ \phi_f = 0.7,\ \phi_\eta = 0.3)$

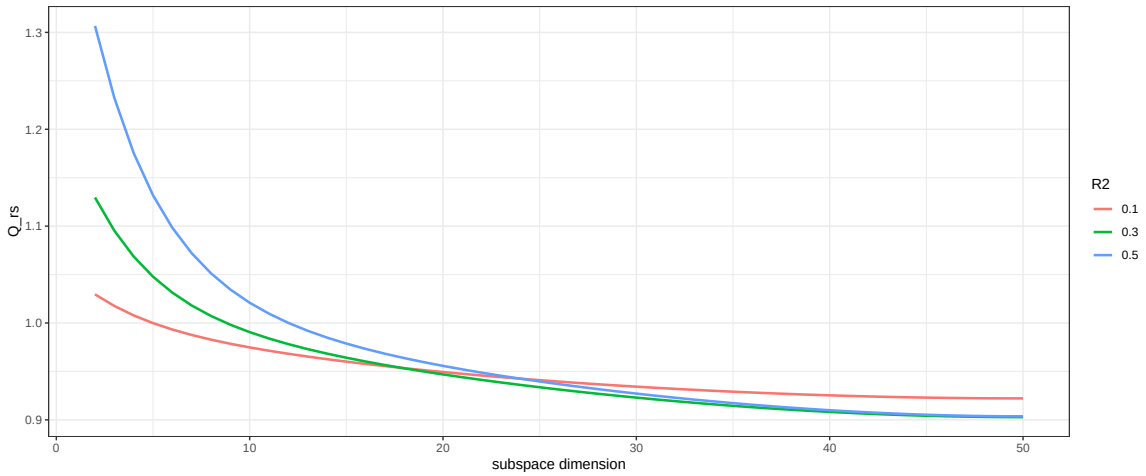


Figure 28: Q^{RSM} $(T = 500,\ \rho = 0.8,\ \phi_f = 0.7,\ \phi_\eta = 0.3)$

E Robustness Checks

E.0.1 RMSFE for Different Subset Dimensions k

In Figures 29 and 30, the RMSFE metrics for each subset $k \in [2, 55]$ are depicted with or without the COVID-19 quarters. Similar to Monte Carlo evidence, both curves have a U-shape, further illustrating that the optimal k^* lies somewhere in between. The exception is for $k = 37$, where a sudden jump is observed, which can be treated as an outlier. In our baseline scenario, we choose k^* as the nearest integer to the $\sqrt{T}$, which corresponds to $k^* = 9, 10, 11$. Again, the imposed rule-of-thumb to select k^* is in line with the empirical findings in Boot and Nibbering (2019), that k^* is somewhat stable over time.

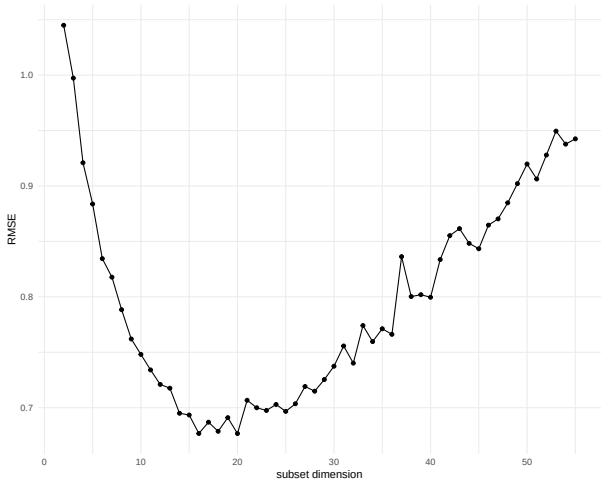


Figure 29: RMSFE full data

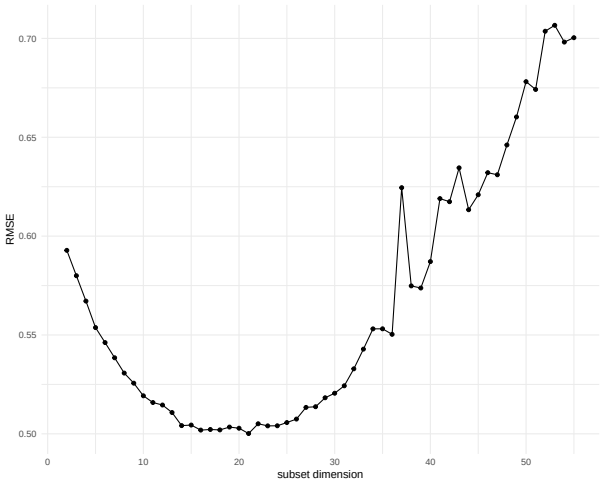


Figure 30: RMSFE without COVID-19

As can be seen in Table 1 and Table 2, the rule-of-thumb rule produces the RMSFE of 0.74 for the full sample and 0.53 for the COVID-19 restricted sample[8]. The corresponding RMSE minima are 0.68 for $k = 18$ and 0.5 for $\forall k \in [14, 26]$, respectively. In both cases, the optimal subset, which is characterized by the lower RMSFE, is higher that the ad-hoc k^* selection. Even though, the proposed k^* choice during the COVID-19 pandemic is able to produce more or less accurate results compared to the mean, the bigger subset dimension can improve the forecasting performance even further.

These findings show that the grid search over all possible k, implemented around $k = \sqrt{T}$, can improve forecasting performance, although sometimes only marginally (if COVID-19 is excluded).

[8]Recall, that RSM randomly selects variables. When implementing this exercise, a different seed is used, yielding 0.73 and 0.51, which are even closer to the corresponding minimum values

F Global and Conditional Weights

For illustrative purposes, we show the correspondence between the global weight $\bar{w}_j^{zb}$ and conditional weight m_j using All Employees: Retail Trade variable (FRED-md mnemonics USTRADE). This variable is chosen (without loss of generality) to depict the difference in the evolution of the global weight compared to the equal weighting.

Recall that both weights are interconnected via the following expression (Equation 21 in the main text):

$$\bar{w}_j^{\text{zb}} = \frac{k}{N} m_j \quad \Longrightarrow \quad \boxed{\bar{w}_j^{\text{zb}} > \tfrac{1}{N} \iff m_j > \tfrac{1}{k}}$$

Therefore, when comparing to the mean, one can use both weights (global and conditional), compared to different values (the equal weight $\frac{1}{N}$ and the subspace equal weight $\frac{1}{k}$, respectively). Both of these comparisons should provide the same results, and the plots should look similar to each other. To be more specific, the global weight should be scaled to $\frac{k}{N}$, which in our case is approximately equal to 0.1, since $k \in [9, 11]$ and $N \approx 120$ for each vintage.

In Figure 31 and Figure 32, the time-varying USTRADE weights are presented.

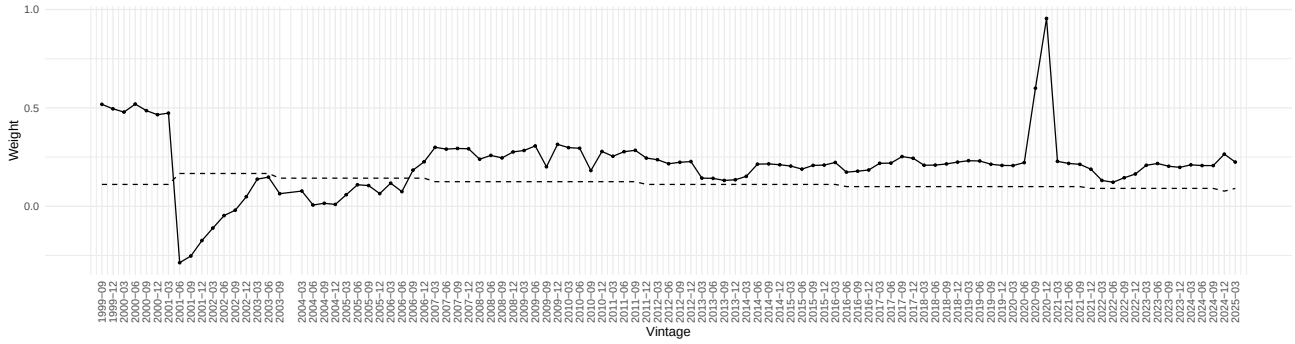


Figure 31: USTRADE: Global Weight against Equal Weight

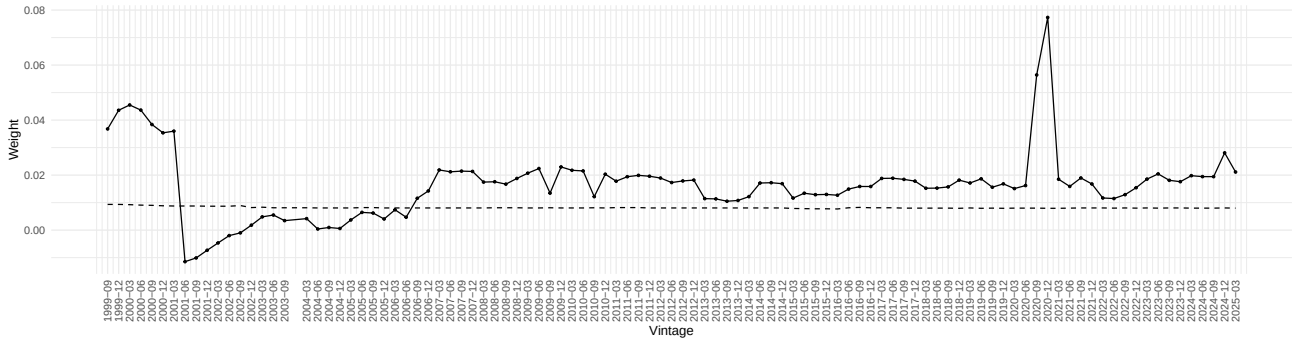


Figure 32: USTRADE: Conditional Weight against Subspace Equal Weight

First, as expected, the plots show the same dynamics. That is for all $\bar{w}_j^{zb} > \frac{1}{N}$ we can observe that $m_j > \frac{1}{k}$. Moreover, in our case, both N and k are somewhat stable over time. That happens

because the number of time series in the FRED-md does not vary substantially with rare occasions of dropouts. Whereas the established k^* changes at the speed $\sqrt{(T)}$, which is also slow. Thus, both dotted lines in Figure 31 and 32 appear to be straight, even though some slight jumps occurred. If we choose k^* differently (say via cross-validation) then the subset dimension can significantly vary over time (the dotted line in Figure 32, but the established correspondence between $\bar{w}_j^{\hat{z}^b}$ and m_j would still hold.