MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://monash.edu/business/ebs/research/publications

Free Lunch Multivariate Forecasting: reducing forecast variance using linear combinations

Yangzhuoran Fin Yang, Rob J. Hyndman,
George Athanasopoulos, Anastasios Panagiotelis

January 2024

Working Paper ??/??

Free Lunch Multivariate Forecasting: reducing forecast variance using linear combinations

Yangzhuoran Fin Yang
Monash University
Melbourne, Australia
Email: Fin.Yang@monash.edu

Rob J. Hyndman
Monash University
Melbourne, Australia

George Athanasopoulos
Monash University
Melbourne, Australia

Anastasios Panagiotelis
University of Sydney
Sydney, Australia

19 January 2024

Free Lunch Multivariate Forecasting: reducing forecast variance using linear combinations

Abstract

Abstact to be written.

1 Introduction

We introduce a new method for improving the accuracy of any multivariate time series method, often substantially. This is done without introducing any new data, or any new information. Thus, we call it a "free lunch" method: it is a simple addition to any existing multivariate forecasting method that can improve its accuracy.

The method is based on the idea that the forecasts of linear combinations of the series should be consistent with the forecasts of the series themselves. For example, suppose we have two observed series $z_{t,1}$ and $z_{t,2}$, and we also construct the combination $c_t = 2z_{t,1} - z_{t,1}$, then the forecasts of c_t should satisfy the same linear constraint: $\hat{c}_t = 2\hat{z}_{t,1} - \hat{z}_{t,2}$. If they do not, then we can improve the forecasts of all three series by adjusting them to be consistent. While the forecasts of z_t may be of no interest in themselves, they can be used to improve the forecasts of x_t and y_t.

The idea can be extended to any number of linear combinations, and can be applied to any number of multivariate series. It does not depend on the forecasting method being used, and works well even if all series are forecast using univariate models. In fact, when used with univariate models, this allows cross-correlations between the series to be implicitly captured in the forecasts.

We call these linear combinations of the observed time series "components", and we call the original forecasts of the observed series "base forecasts". Our free-lunch method is to adjust the base forecasts to be consistent with the forecasts of the components by projecting all forecasts onto the space where the linear constraints are satisfied. We show (theoretically and empirically) that this method leads to significant reduction in the forecast variance, without introducing any bias.

Our free-lunch method has close connections to forecast reconciliation in the hierarchical forecasting literature. See Athanasopoulos et al. (2023) for a recent review of the area. In particular, the projection formulation is inspired by the minimum trace (MinT, Wickramasuriya, Athanasopoulos &

Hyndman 2019) solution of the forecast reconciliation problem. Forecast reconciliation is a method to modify forecasts using projection so that they conform to a specific hierarchical, grouped or temporal structure. Notably, Wickramasuriya, Athanasopoulos & Hyndman (2019), Athanasopoulos et al. (2017), and Di Fonzo & Girolimetto (2023) have shown that forecast reconciliation can reduce forecast variance theoretically and empirically, in cross-sectional settings, temporal settings, and in cross-temporal settings. Panagiotelis et al. (2021) have provided insight into the geometric interpretation of the projection used in forecast reconciliation. However, forecast reconciliation cannot be directly applied in a general multivariate time series unless the series satisfy some linear constraints such as a hierarchical structure. In contrast, our method can be applied to any multivariate time series. It can also be used in conjunction with forecast reconciliation to further reduce forecast variance.

The idea bears some similarity to bootstrap aggregation or "bagging" (Breiman 1996; Bergmeir, Hyndman & Benítez 2016), where the final prediction is produced from an ensemble of predictions made on bootstrapped data. Bagging can reduce prediction variance without increasing bias (Hastie, Tibshirani & Friedman 2003), by mitigating model uncertainty (Petropoulos, Hyndman & Bergmeir 2018), and it does so without introducing any new data (just bootstrapped versions of the existing data). Our method also reduces forecast variance without introducing new data, but using linear combinations of the existing data, rather than bootstrapped versions of the data. A second difference is that bagging is model dependent: it is a procedure applied to enhance the models that produce the forecasts, where the same models are fitted repeatedly. Our method is model independent: it linearly transforms a set of forecasts, regardless of which models they come from. As a result, the two methods can be used in conjunction: the projections can be applied to forecasts produced by a bagged predictor.

Another approach to improve forecast accuracy is forecast combination. Point forecast combinations usually involve combining multiple forecasts of the same series from different models. See Wang et al. (2023) for a recent comprehensive review. Our free-lunch method differs from forecast combination in (a) the forecasts we combine, and (b) in how we combine them. First, rather than combine multiple forecasts of a single series, we combine single forecasts of many different linear combinations of all observed series. Second, our combinations are obtained via projections, and so the final forecasts of a particular series are linear combinations of all series in the collection, including the observed series and all constructed components. Our approach can be used in conjunction with standard forecast combination, as the base forecasts can be obtained from any combination of forecasts.

In a broad sense, forecast reconciliation, bagging, and our proposed free-lunch method can all be viewed as forms of forecast combination with different objects to be combined. Petropoulos & Spiliotis (2021) overview a group of methods utilising combination techniques which they call "the wisdom of data" including bagging, theta method (Assimakopoulos & Nikolopoulos 2000), temporal aggregation (Kourentzes, Petropoulos & Trapero 2014; Athanasopoulos et al. 2017), forecasting with sub-seasonal series (FOSS, Li, Petropoulos & Kang 2022) and forecast combination with multiple starting points (Disney & Petropoulos 2015). These differ in transformations, series to forecast, forecasting models, and combination weights. Similarly, our free-lunch method aims to exploit information in the data, with a focus on the shared information that can be captured by linear combinations of the series.

Finally, the use of components bears a resemblance to Dynamic Factor Models (DFMs), specifically those used in a forecasting setting where the components (factors) are estimated using Principal Component Analysis (PCA) (Stock & Watson 2002b,a, 2012), and their extension in the machine learning literature (De Stefani et al. 2019). DFMs assume that the multivariate time series possesses common components and the dynamics of the observed series are governed by the dynamics of these unobserved components, typically assumed to follow a Vector AutoRegressive (VAR) model. In contrast, our method makes no assumptions on the parametric form of the dynamics. In fact, the forecasts from a DFM can be further improved by applying the free-lunch method to the DFM forecasts. This is demonstrated in Section 3 and Section 4, where we also show how the performance of projected forecasts from univariate ARIMA models is comparable to base DFM forecasts, with the help of components.

The rest of the paper is structured as follows. In Section 2, we propose the free lunch forecast projection method, and highlight its theoretical properties and associated estimation methods. In Section 3, we present a simulation example demonstrating its performance and discuss the implications for sources of uncertainty. Section 4 examines the performance of free-lunch forecast projection in two empirical applications: forecasting Australian domestic tourism and forecasting macroeconomic variables in the FRED-MD data set. Section 5 concludes with some thoughts on future research directions.

2 Free-Lunch Forecast Projection

2.1 Definitions and properties

We use $\boldsymbol{I}_n$ to denote the $n \times n$ identity matrix, and $\boldsymbol{O}_{n\times k}$ to denote the $n \times k$ zero matrix. Define the selection matrix $\boldsymbol{J}_{n,k} = \begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{O}_{n\times k}\end{bmatrix}$, so that $\boldsymbol{J}_{n,k}\boldsymbol{A}$ picks out the first n rows of a matrix $\boldsymbol{A}$.

Let $\boldsymbol{z}_t \in \mathbf{R}^m$ be a vector of m observed time series at time t. Let $\boldsymbol{y}_t = \begin{bmatrix}\boldsymbol{z}_t', \boldsymbol{c}_t'\end{bmatrix}'$ be the collection of series $\boldsymbol{z}_t$ and components $\boldsymbol{c}_t$, where $\boldsymbol{c}_t = \boldsymbol{\Phi}\boldsymbol{z}_t \in \mathbf{R}^p$, and let $\hat{\boldsymbol{y}}_{t+h}$ denote the h-step-ahead base forecast

of $\boldsymbol{y}_t$. The forecast variance covariance matrix is $\text{Var}(\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) = \boldsymbol{W}_h$. We project the base forecasts onto the space where the constraints are imposed:

$$\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{M}\hat{\boldsymbol{y}}_{t+h} \tag{1}$$

with projection matrix

$$\boldsymbol{M} = \boldsymbol{I}_{m+p} - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}, \tag{2}$$

where $\boldsymbol{C} = \begin{bmatrix} -\boldsymbol{\Phi} & \boldsymbol{I}_p \end{bmatrix}$ defines the $p \times (m+p)$ constraint matrix such that $\boldsymbol{C}\boldsymbol{y}_t = \boldsymbol{c}_t - \boldsymbol{\Phi}\boldsymbol{z}_t = \boldsymbol{0}$ for any t. Let $\hat{\boldsymbol{z}}_{t+h}$ and $\tilde{\boldsymbol{z}}_{t+h}$ denote the first p elements of $\hat{\boldsymbol{y}}_{t+h}$ and $\tilde{\boldsymbol{y}}_{t+h}$, comprising the base and projected forecasts of $\boldsymbol{z}_t$ respectively. Similarly, let $\hat{\boldsymbol{c}}_{t+h}$ and $\tilde{\boldsymbol{c}}_{t+h}$ denote the last p elements of $\hat{\boldsymbol{y}}_{t+h}$ and $\tilde{\boldsymbol{y}}_{t+h}$, comprising the base and projected forecasts of $\boldsymbol{c}_t$ respectively. Then the projected forecast of $\boldsymbol{z}_t$ can be found by

$$\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{J}\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{J}\boldsymbol{M}\hat{\boldsymbol{y}}_{t+h}, \tag{3}$$

where $\boldsymbol{J} = \boldsymbol{J}_{m,p}$.

We are ready to present a few immediate results, with proofs provided in the Appendix.

Lemma 2.1. *The projected forecast $\tilde{\boldsymbol{y}}_{t+h}$ satisfies the constraint*

$$\boldsymbol{C}\tilde{\boldsymbol{y}}_{t+h} = \tilde{\boldsymbol{c}}_{t+h} - \boldsymbol{\Phi}\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{0}.$$

Lemma 2.2. *The mapping matrix $\boldsymbol{M}$ projected a vector onto the space where the constraint $\boldsymbol{C}$ is satisfied. For $\boldsymbol{y}_{t+h}$ that already satisfies the constraint, the projection does not change its value:*

$$\boldsymbol{M}\boldsymbol{y}_{t+h} = \boldsymbol{y}_{t+h}.$$

Lemma 2.3. *If the base forecasts are unbiased such that*

$$\text{E}(\hat{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \text{E}(\boldsymbol{y}_{t+h}|\mathscr{I}_t),$$

then the projected forecasts are also unbiased:

$$\text{E}(\tilde{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \text{E}(\boldsymbol{y}_{t+h}|\mathscr{I}_t).$$

The only requirement for the current theory of forecast projection to work is the given base forecasts need to be unbiased. This is not a very strict requirement as unbiasedness can be achieved by most

of the common forecasting models with an intercept. Even if there are transformations applied to the data set before the models are fitted, bias correction can be applied as suggested by Panagiotelis et al. (2021). Note this is not a requirement on model specification: we do not assume the model producing the base forecast is correctly specified like in the DFM literature (e.g. Stock & Watson (2002a)). In fact, the power of forecast projection manifests when the models are misspecified, as discussed in Section 3.

Lemma 2.4. *The forecast variance covariance matrix of the component-constrained projected h-step-ahead forecasts $\tilde{\boldsymbol{y}}_{t+h}$ is $\boldsymbol{M}\boldsymbol{W}_h$, i.e.*

$$\text{Var}(\tilde{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) = \boldsymbol{M}\boldsymbol{W}_h\boldsymbol{M}' = \boldsymbol{M}\boldsymbol{W}_h,$$

and the forecast variance covariance matrix of the projected h-step-ahead forecasts $\tilde{\boldsymbol{z}}_{t+h}$ is

$$\text{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) = \boldsymbol{J}\boldsymbol{M}\boldsymbol{W}_h\boldsymbol{J}'.$$

Lemma 2.4 is a well known results (e.g. Di Fonzo & Girolimetto (2023)) but was rarely focused on.

Theorem 2.1 (Positive Semi-Definiteness of Variance Reduction)**.** *The difference between the forecast variance covariance matrix of the base forecast and the projected forecast*

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) - \text{Var}(\tilde{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) &= \boldsymbol{W}_h - \boldsymbol{M}\boldsymbol{W}_h \\ &= \boldsymbol{W}_h - (\boldsymbol{I} - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C})\boldsymbol{W}_h \\ &= \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\end{aligned}$$

is positive semi-definite. The difference between the forecast variance of $\hat{\boldsymbol{z}}_{t+h}$ and $\tilde{\boldsymbol{z}}_{t+h}$

$$\text{Var}(\hat{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) - \text{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) = \boldsymbol{J}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{J}'$$

is therefore positive semi-definite.

Theorem 2.1 is why forecast projection works. The trace of $\boldsymbol{J}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{J}'$ is the sum of forecast variances that can be reduced by forecast projection. Because the matrix is positive semi-definite, such trace is nonnegative. It means we can reduce the forecast variance by simply forecasting the components - the artificially constructed linear combination of the original data, and mapping the forecasts using matrix $\boldsymbol{M}$. For the improvement to be zero, the trace needs to be zero,

and because the matrix is positive semi-definite, this implies that the entire $\boldsymbol{J}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{J}'$ is a zero matrix, which rarely happens in practice. See Theorem 2.3 for more discussions.

We give a simple example to show how the variance reduction works.

Example 2.1 (Identity W_h)**.** Let

$$\boldsymbol{W}_h = \boldsymbol{I}_{m+p},$$

that is, let $\boldsymbol{y}_t$ consist of m original series and p components whose forecasts are all uncorrelated with each other with variance 1, then

$$\begin{aligned}\operatorname{Var}(\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) - \operatorname{Var}(\tilde{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) &= \boldsymbol{C}'(\boldsymbol{C}\boldsymbol{C}')^{-1}\boldsymbol{C} \\ &= \begin{bmatrix} -\boldsymbol{\Phi}' \\ \boldsymbol{I}_p \end{bmatrix} (\boldsymbol{\Phi}\boldsymbol{\Phi}' + \boldsymbol{I})^{-1} \begin{bmatrix} -\boldsymbol{\Phi} & \boldsymbol{I}_p \end{bmatrix},\end{aligned}$$

where

$$\boldsymbol{C} = \begin{bmatrix} -\underset{p\times m}{\boldsymbol{\Phi}} & \underset{p\times p}{\boldsymbol{I}_p} \end{bmatrix}.$$

Let $\boldsymbol{\Phi}$ consists of unit vectors that are orthogonal to each other, for example, those obtained from Principal Component Analysis (PCA, see Jolliffe 2002, among others). That is,

$$\boldsymbol{\Phi}\boldsymbol{\Phi}' = \boldsymbol{I}_p \text{ when } p \le m$$

and

$$\boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{I}_m \text{ when } p = m,$$

then

$$\begin{aligned}\operatorname{Var}(\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) - \operatorname{Var}(\tilde{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) &= \frac{1}{2}\begin{bmatrix} \boldsymbol{\Phi}'\boldsymbol{\Phi} & -\boldsymbol{\Phi}' \\ -\boldsymbol{\Phi} & \boldsymbol{I}_p \end{bmatrix} \text{ when } p \le m. \\ &= \frac{1}{2}\begin{bmatrix} \boldsymbol{I}_m & -\boldsymbol{\Phi}' \\ -\boldsymbol{\Phi} & \boldsymbol{I}_p \end{bmatrix} \text{ when } p = m.\end{aligned}$$

We only focus on the forecast variance of the original series, which is $\operatorname{tr}(\operatorname{Var}(\hat{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) - \operatorname{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h})) = \frac{1}{2}\operatorname{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi})$. When $p < m$, since $\boldsymbol{\Phi}'\boldsymbol{\Phi}$ is idempotent (i.e. $(\boldsymbol{\Phi}'\boldsymbol{\Phi})(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = \boldsymbol{\Phi}'\boldsymbol{\Phi}\boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{\Phi}'\boldsymbol{\Phi}$), we have $\operatorname{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = \operatorname{rank}(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = p$. The reduction on the forecast variance of the original series is $p/2$. When $p = m$, we have $\operatorname{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = \operatorname{tr}(\boldsymbol{I}_m) = m$, and the reduction is $m/2$. If we have two series ($m = 2$) to begin with, using 1 component ($p = 1$) in the mapping will reduce the sum of forecast variance by

0.5, and using 2 components ($p = m = 2$) will reduce the sum of forecast variance by 1, that is a 50% reduction when the original sum of forecast variance is only 2.

If we keep increasing the number of components, the result in Theorem 2.1 still holds, but $\boldsymbol{\Phi}$ can no longer contain orthogonal vectors, and the example here becomes intractable. This is an artificial example in the sense that the forecast variance $\boldsymbol{W}_h$ can hardly be identity in practice, as the forecasts of a linear combination of two series are likely to be correlated with the forecasts of these series. Nonetheless, we can see how the forecast variance can be reduced as a result of uncorrelatedness between the forecast of components and the series, representing the new information brought by the components.

The variance reduction becomes larger as we increase the number of component p from 1 to 2 in Example 2.1. In fact, this is not a coincidence but rather a much desired property of forecast projection, as shown in the next section.

2.2 Monotonicity

In the results that follow, we break the base forecast variance covariance matrix into smaller blocks:

$$\boldsymbol{W}_h = \begin{bmatrix} \boldsymbol{W}_{z,h} & \boldsymbol{W}_{zc,h} \\ \boldsymbol{W}_{cz,h} & \boldsymbol{W}_{c,h} \end{bmatrix},$$

where $\boldsymbol{W}_{z,h}$ is the forecast variance covariance matrix of $\hat{\boldsymbol{z}}_{t+h}$, $\boldsymbol{W}_{c,h}$ is the forecast variance covariance matrix of $\hat{\boldsymbol{c}}_{t+h}$, and $\boldsymbol{W}_{zc,h}$ ($\boldsymbol{W}_{cz,h}$) consists of covariance between elements of $\hat{\boldsymbol{z}}_{t+h}$ and $\hat{\boldsymbol{c}}_{t+h}$ ($\hat{\boldsymbol{c}}_{t+h}$ and $\hat{\boldsymbol{z}}_{t+h}$).

Theorem 2.2 (Monotonicity)**.** *The sum of reduced forecast variance*

$$\operatorname{tr}(\operatorname{Var}(\hat{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) - \operatorname{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h})) = \operatorname{tr}(\boldsymbol{J}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{J}') \tag{4}$$

is non-decreasing as p increases.[1]

Theorem 2.2 is the key result that justifies the usefulness of forecast projection by providing a practical way to increase its power. It means we can keep increasing the number of components to reduce forecast variance, even when the number of components exceeds the number of original series, assuming we know the true base forecast variance covariance of all the series and components. It requires $\boldsymbol{C}$ to be $\begin{bmatrix} -\boldsymbol{\Phi} & \boldsymbol{I}_p \end{bmatrix}$ or $\begin{bmatrix} -\boldsymbol{\Phi} & \boldsymbol{L} \end{bmatrix}$ where $\boldsymbol{L}$ is a lower triangular matrix (the upper right corner needs to be all 0s). This implies that the components can also be constructed from existing

[1]We thank Daniele Girolimetto for contributing to the initial proof (of the second statement in the proof) in his unpublished work.

components, not only from the original series. This has little significance since a linear combination of components (which are linear combinations themselves) of the original series, is just a linear combination of the original series.

Extending the proof of Theorem 2.2, we can outline the condition for the reduced variance to be positive. That is, if the new component satisfies these relationships with the previous (projected) forecasts, the projected forecast variance with the new component is smaller than the base forecast variance or the projected variance with only the previous components. Denote $\boldsymbol{\phi}_i$ as the row vector containing the weights associated with the ith component, so that with p components, the weights matrix is $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_1' & \boldsymbol{\phi}_2' & \dots & \boldsymbol{\phi}_p' \end{bmatrix}'$.

Theorem 2.3 (Positive Variance Reduction Condition)**.** *For the first component to have a guaranteed reduction of forecast variance (for the reduced variance matrix in Theorem 2.1 to have positive trace), the following condition must be satisfied:*

$$\boldsymbol{\phi}_1 \boldsymbol{W}_{z,h} \neq \boldsymbol{W}_{c_1 z,h}, \tag{5}$$

where $\boldsymbol{W}_{c_1 z,h}$ denotes the base forecast covariance of the first components and the original series. For the ith component to have a positive reduction on forecast variance of the original series, the following condition is satisfied:

$$\boldsymbol{\phi}_i \boldsymbol{W}_{\tilde{z},h}^{(i-1)} \neq \boldsymbol{W}_{\tilde{c}_i \tilde{z},h}^{(i-1)}, \tag{6}$$

where $\boldsymbol{W}_{\tilde{z},h}^{(i-1)}$ denotes the projected forecast variance of the original series using the first $i-1$ components, and $\boldsymbol{W}_{\tilde{c}_i \tilde{z},h}^{(i-1)}$ denotes the covariance between the projected forecast of the original series using the first $i-1$ components and the projected forecast of the ith component using the first $i-1$ components. See the proof in the Appendix for an algebraic definition.

If we look at each element of the terms in Equation 5, we can interpret the condition in the following way: for at least one series, the forecast covariance between the new component and this series is not exactly a linear combination, defined by $\boldsymbol{\phi}_1$, of the forecast variance of this series and the forecast covariance between this series and all other series. For the ith component in Equation 6, the interpretation is similar, where the linear combination is defined by $\boldsymbol{\phi}_i$ and the forecast variance is replaced by the projected forecast variance and covariance associated with the first $i-1$ components. This interpretation constitutes a measure of forecast information. For the new component to be beneficial, the information brought by this new component, measured as the covariance, cannot be a combination of already existing information.

Theorem 2.3 can potentially provide insights into the selection of component weights and forecast models to satisfy the conditions. We leave this issue in a later article, as practically the conditions in Theorem 2.3 are either almost always satisfied if the weights are simulated randomly on a continuous scale, or the loss associated with the rare occasions where the conditions are not satisfied is neglectable compared to the estimation error imposed by the limited sample size as the number of components increases, as discussed in Section 3 and Section 4.

2.3 Alternative Interpretations

Equation 1 can be seen as a solution to the optimisation problem:

$$\begin{aligned}\underset{\check{\boldsymbol{y}}_{T+h}}{\arg\min}(\hat{\boldsymbol{y}}_{T+h}-\check{\boldsymbol{y}}_{T+h})'\boldsymbol{W}_h^{-1}(\hat{\boldsymbol{y}}_{T+h}-\check{\boldsymbol{y}}_{T+h})\\ \text{s.t. } \boldsymbol{C}\check{\boldsymbol{y}}_{T+h}=0.\end{aligned}$$

In this case, the projection can be interpreted as finding the set of forecast that is closest (on the transformed space) to the base forecast that satisfies the linear constraints imposed by the components.

Moreover, this is equivalent to the optimisation problem:

$$\begin{aligned}\underset{\check{\boldsymbol{y}}_{T+h}}{\arg\min}(\hat{\boldsymbol{y}}_{T+h}-\check{\boldsymbol{y}}_{T+h})'\boldsymbol{W}_h^{-1}(\hat{\boldsymbol{y}}_{T+h}-\check{\boldsymbol{y}}_{T+h})\\ \text{s.t. } \boldsymbol{\Phi}\check{\boldsymbol{z}}_{T+h}=\check{\boldsymbol{c}}_{T+h},\end{aligned}$$

where $\check{\boldsymbol{c}}_{T+h}$ is the vector of the last p elements of $\check{\boldsymbol{y}}_{T+h}$, corresponding to the forecast of the components as part of the solution. This equivalence is discussed in Wickramasuriya, Athanasopoulos & Hyndman (2019), where the authors find the solution by minimising the sum of forecast variance of all series (See Ando & Narita (2022) for a corrected proof). The result is the MinT solution

$$\tilde{\boldsymbol{y}}_{t+h}=\boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h}, \tag{7}$$

where $\boldsymbol{S}=\begin{bmatrix}\boldsymbol{I}_m\\ \boldsymbol{\Phi}\end{bmatrix}$ contains the constraints in a different order from $\boldsymbol{C}$, so we also have $\boldsymbol{y}_t=\boldsymbol{S}\boldsymbol{z}_t$, and

$$\boldsymbol{G}=(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}. \tag{8}$$

In Equation 7, $\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h}$ can be viewed as mapping all the series to a selected few. In the forecast reconciliation context, this is mapping series at all levels to the bottom level series; in our multivariate forecasting context, this is mapping all series including the given number of components, to the

original series. The structure matrix $\boldsymbol{S}$ in the forecast reconciliation context is to map the forecast at the bottom level to the entire hierarchical structure; in our context, this is not necessary, as our focus is on the forecast of the original series only, regardless of the forecasting performance on the components. Thus, our solution in the multivariate forecast context in Equation 3 is equivalent to finding the $\boldsymbol{G}$ in

$$\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{G}\hat{\boldsymbol{y}}_{t+h}. \tag{9}$$

Recognising Equation 1 is equivalent to Equation 7, the solution in Equation 8 is the solution that minimises the sum of variance of original series and all the components. Here we show in Theorem 2.4 that Equation 8 is also the solution to minimise each individual variance (and the sum) of the original series only, which is optimal in Equation 9. This can be viewed as a special case of Theorem 3.3 of Panagiotelis et al. (2021), or as illustrated by Ando & Narita (2022), but applied in a different context from forecast reconciliation. The early work we can find that noted this interpretation in a non-forecasting context can go back as far as Luenberger (1969, p. 85). We establish a few basic results leading to the optimality of this solution first, also to check that Lemma 2.1-2.4 hold under this alternative representation. It may not be immediately obvious that the constraints imposed by the components are satisfied by looking at Equation 9 only, but this can be seen from Equation 7 and Lemma 2.1 can be easily checked without referencing to a specific $\boldsymbol{G}$.

Lemma 2.5. *The projected forecast in Equation 9 satisfies the constraint such that*

$$\boldsymbol{C}\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{C}\boldsymbol{S}\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{0}.$$

Lemma 2.6. *For $\boldsymbol{y}_{t+h}$ that already satisfies the constraint, the projection does not change its value:*

$$\boldsymbol{G}\boldsymbol{y}_{t+h} = \boldsymbol{z}_{t+h}.$$

Lemma 2.7. *If the base forecasts are unbiased such that*

$$\mathrm{E}(\hat{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{y}_{t+h}|\mathscr{I}_t),$$

then the mapping in Equation 9 are also unbiased:

$$\mathrm{E}(\tilde{\boldsymbol{z}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{z}_{t+h}|\mathscr{I}_t),$$

provided

$$\boldsymbol{GS} = \boldsymbol{I}.$$

Lemma 2.8. *The forecast variance covariance matrix of the mapped forecasts from Equation 9 is given by*

$$\text{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) = \boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}'.$$

Putting them together, we have the following theorem.

Theorem 2.4 (Minimum Variance Unbiased Projected Forecast)**.** *The solution to*

$$\begin{aligned} &\arg\min_{\boldsymbol{G}} \boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}' \\ &\quad s.t.\ \boldsymbol{G}\boldsymbol{S} = \boldsymbol{I} \end{aligned} \tag{10}$$

is Equation 8.

Recognising $\boldsymbol{G}$ is of dimension $m \times (m+p)$, this problem can be effectively split into independent subproblems such that

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{g}_1 & \boldsymbol{g}_2 & \cdots & \boldsymbol{g}_n \end{bmatrix}',$$

where $\boldsymbol{g}_i$ is the solution to the subproblem of the ith series

$$\begin{aligned} &\arg\min_{\boldsymbol{g}_i} \boldsymbol{g}_i'\boldsymbol{W}_h\boldsymbol{g}_i \\ &s.t.\ \boldsymbol{g}_i'\boldsymbol{s}_j = \delta_{ij}, \quad j = 1, 2, \ldots, m, \end{aligned} \tag{11}$$

where $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & j \neq j \end{cases}$ *is the Kronecker delta function.*

In other words, the forecast projection method gives optimal projected forecast for a given set of component, in the sense that the unbiased forecast of each series has minimum variance, which is a step further from the collective optimum where the sum of the variances is minimised.

2.4 Estimation of W_h

In practice, the base forecast variance $\boldsymbol{W}_h$ is unknown and needs to be estimated. Denote $\hat{\boldsymbol{e}}_{t,h} = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-h}$ as the h-step-ahead base forecast residual. The conventional forecast variance covariance matrix estimator

$$\widehat{\boldsymbol{W}_h} = \frac{1}{T-1}\sum_{i=1}^{T} \hat{\boldsymbol{e}}_{t,h}\hat{\boldsymbol{e}}_{t,h}',$$

albeit unbiased, is not considered a good approximation to the true forecast variance in a finite sample when $(m+p) \approx T$. It is even singular when $(m+p) > T$, which makes the quantities discussed in the previous sections impossible to calculate. For this reason, we adopt the variance shrinkage method by Schäfer & Strimmer (2005), which is treated as the MinT(Shrink) method by Wickramasuriya, Athanasopoulos & Hyndman (2019), and the covariance shrinkage method by Opgen-Rhein & Strimmer (2007). The estimated forecast variance matrix is guaranteed to be positive definite with few numerical problems. This estimator is denoted as $\widehat{\boldsymbol{W}}_h^{shr} = (\hat{w}_{ij,h}^{shr})_{1\leq i,j\leq m+p}$ whose elements are

$$\hat{w}_{ij,h}^{shr} = \hat{r}_{ij,h}^{shr}\sqrt{\hat{v}_{i,h}\hat{v}_{j,h}}$$

with

$$\hat{r}_{ij,h}^{shr} = (1-\hat{\lambda}_{cor})\hat{r}_{ij,h}$$

and

$$\hat{v}_{i,h} = \hat{\lambda}_{var}\hat{w}_{h,median} + (1-\hat{\lambda}_{var})\hat{w}_{i,h},$$

with $\hat{\lambda}_{cor}$ being the shrinkage intensity parameter for the correlation:

$$\hat{\lambda}_{cor} = \min(1, \frac{\sum_{i\neq j}\widehat{\text{var}}(\hat{r}_{ij,h})}{\sum_{i\neq j}\hat{r}_{ij,h}^2}),$$

and $\hat{\lambda}_{var}$ being the shrinkage intensity parameter for the variance:

$$\hat{\lambda}_{var} = \min(1, \frac{\sum_{i=1}^{m+p}\widehat{\text{var}}(\hat{w}_{i,h})}{\sum_{i=1}^{m+p}(\hat{w}_{i,h}-\hat{w}_{h,median})^2}),$$

where $\hat{r}_{ij,h}$ is the sample correlation of the h-step-ahead forecast error between the ith and the jth series (component) in $\boldsymbol{y}_t$, $\hat{w}_{i,h}$ is the h-step-ahead sample base forecast variance associated with the ith series (the ith diagonal element of $\widehat{\boldsymbol{W}}_h$) and $\hat{w}_{h,median}$ is the median of the h-step-ahead sample forecast variance of the series and components (the median of the diagonal elements of $\widehat{\boldsymbol{W}}_h$). The estimation of $\widehat{\boldsymbol{W}}_h^{shr}$ in the following sections are implemented using the package `corpcor` (Schafer et al. 2021) in R (R Core Team 2022).

Estimating $\widehat{\boldsymbol{W}}_h^{shr}$ for each forecast horizon h is desirable but computationally intensive. It involves the calculation of multi-step-ahead in-sample residuals of the forecast models, which is especially challenging for iterative forecasts. Because of this, in practice it is not unreasonable to assume the h-step forecast variance is proportional to the 1-step forecast variance by a constant η_h, as do

Wickramasuriya, Athanasopoulos & Hyndman (2019):

$$\widehat{\boldsymbol{W}}_h^{shr} = \eta_h \widehat{\boldsymbol{W}}_1^{shr}.$$

Under this assumption, when $\widehat{\boldsymbol{W}}_h^{shr}$ is used in Equation 2, the proportionality constant η_h cancels out regardless of the value of h. We can effectively use only the one-stop forecast variance in forecast projection, if only the point forecasts are concerned. We calculate $\widehat{\boldsymbol{W}}_h^{shr}$ for each h for the simulation example in Section 3, but assumes this proportionality for the application in Section 4.

3 Simulation

3.1 Benchmarks

In this section, we illustrate the performance of forecast projection in a simulation example. In each sample, we simulate $T = 400$ observations from a VAR(3) process with $m = 70$ variables. The coefficients of the VAR model are estimated from the first 70 series in the Australian tourism data set used in Section 4.1. The innovation in the VAR model is simulated from a multivariate normal distribution with an identity variance covariance matrix. The estimation and simulation are done using package `tsDyn` (Fabio Di Narzo, Aznarte & Stigler 2009). We simulate 220 such samples and the forecast is evaluated on each sample.

The first benchmark we use is the univariate ARIMA model. For each series, we fit an ARIMA model using the `auto.arima()` function from the `forecast` package (Hyndman et al. 2023). The function implements an automatic model selection procedure proposed by Hyndman & Khandakar (2008). The number of first differences is determined by repeated KPSS tests (Kwiatkowski et al. 1992) and the number of seasonal differences is determined by the seasonal strength computed from an STL decomposition (Cleveland et al. 1990). The algorithm then chooses different orders of the autoregressive (AR) and moving average (MA) parts by comparing AICc between the corresponding models in a stepwise fashion, up to a maximal order of 5. Univariate ARIMAs are also used to produce base forecasts of the components used in projection, regardless of the base model.

Another base model is the DFM following Stock & Watson (2002b):

$$\hat{y}_{T+h} = \hat{\alpha}_h + \sum_{j=1}^{n} \hat{\boldsymbol{\beta}}_{hj}' \hat{\boldsymbol{F}}_{T-j+1} + \sum_{j=1}^{s} \hat{\gamma}_{hj} y_{T-j+1},$$

where $\hat{\boldsymbol{F}}_t$ is the vector of k estimated factors, and $\hat{y}_t$ is the target series to forecast. The factors are estimated using PCA on demeaned and scaled data. The optimal model is selected for each series

based on the Bayesian information criterion (BIC) from models fitted using different combinations of meta-parameters in their corresponding range: $1 \leq k \leq 6$, $1 \leq n \leq 3$ and $1 \leq s \leq 3$. Note here DFM produces direct forecasts in the sense that a different model is fitted for each forecast horizon h, compared to indirect forecast or iterative forecast.

We use different weighting methods to construct the components. The types of components are listed below. For the ones randomly simulated from a distribution, we normalise them into unit vectors to maintain some level of consistency with $\boldsymbol{\phi}_i / \sqrt{\sum_j (\phi_{ij}^2)}$ where ϕ_{ij} is the jth value in the weight vector of the i components.

PCA+Norm The m principal components in PCA are taken first, implemented with the `prcomp` function in package `stats` (R Core Team 2022). Weights of the additional components are simulated from a standard normal distribution before normalised to unit vectors.

PCA+Unif The m principal components in PCA are taken first, and the weights of the additional components are simulated from a uniform distribution with minimum -1 and maximum 1 before normalised to unit vectors.

Norm The weights of components are simulated from a standard normal distribution before normalised to unit vectors.

Unif The weights of components are simulated from a uniform distribution with minimum -1 and maximum 1 before normalised to unit vectors.

Ortho+Norm A random orthonormal matrix is generated using package `pracma` (Borchers 2023) as the weights of the first m components. Weights of the additional components are simulated from a standard normal distribution before normalised to unit vectors.

We employ the Friedman test (Friedman 1937, 1939) along with post-hoc Nemenyi tests (Nemenyi 1963; See Hollander, Wolfe & Chicken 2013, for details) to compare forecast performance between different methods. The analysis involves the use of Multiple Comparisons with the Best (MCB) plot introduced by Koning et al. (2005) to visualise the comparison. The mean squared error (MSE) of each series over different samples is calculated, and the MSEs of all the series are treated as observations in the Nemenyi test. Our objective is to assess whether there are statistically significant differences between the projected and base forecasts. The average ranks are plotted in Figure 1 for forecast horizons 1, 6 and 12. The methods using forecast projection are named "{Model}-{Comp. Weights}-{No. Comp.}". The maximum number of components is chosen to be 300. The base models are named "{Model}-Base" and these points are marked with triangles. The shaded region is the interval of the best-performing model. Methods outside the shaded region are significantly worse than the best model. We also plot the specific MSE values by the number of components p in Figure 2.

Here we include the performance of the true data generating process (DGP) VAR model (VAR - GDP), the estimated VAR model with the correct specification (VAR - Est.), and their projections. We do not include methods involving uniform distribution and random orthonormal matrices as they are visually identical to the methods with normal distribution. The vertical black line indicates the number of series $m = 70$. We group our findings in the following categories.

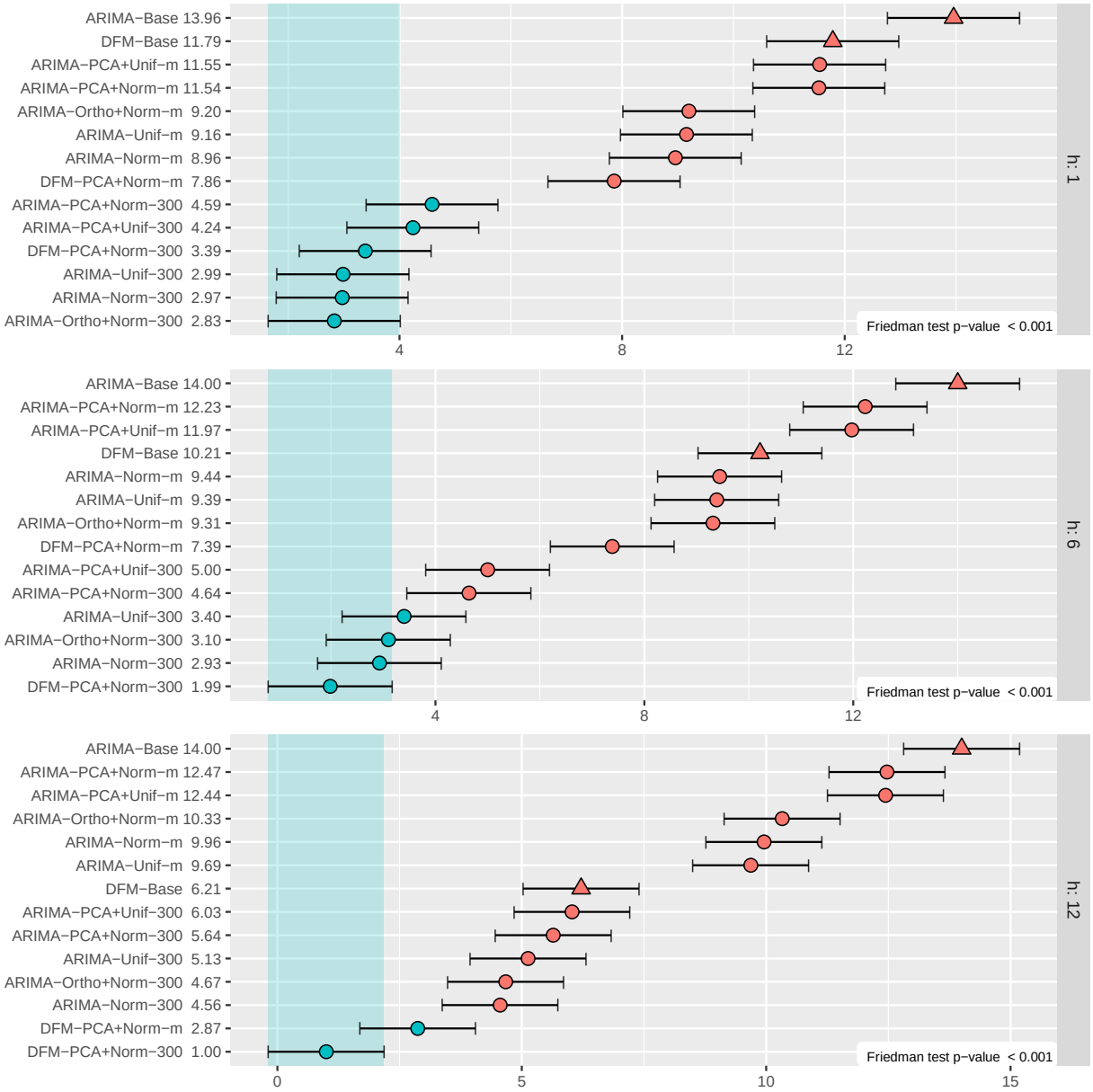

Figure 1: *Average ranks of* 1-, 6- *and* 12*-step-ahead MSE of different model and component specifications in the simulation. The methods using forecast projection are named as "{Model}-{Comp. Weights}-{No. Comp.}". The base models are named as "{Model}-Base" and these points are marked with triangles. The shaded region is the interval of the best performing model. Methods outside the shaded region are significantly worse than the best model.*

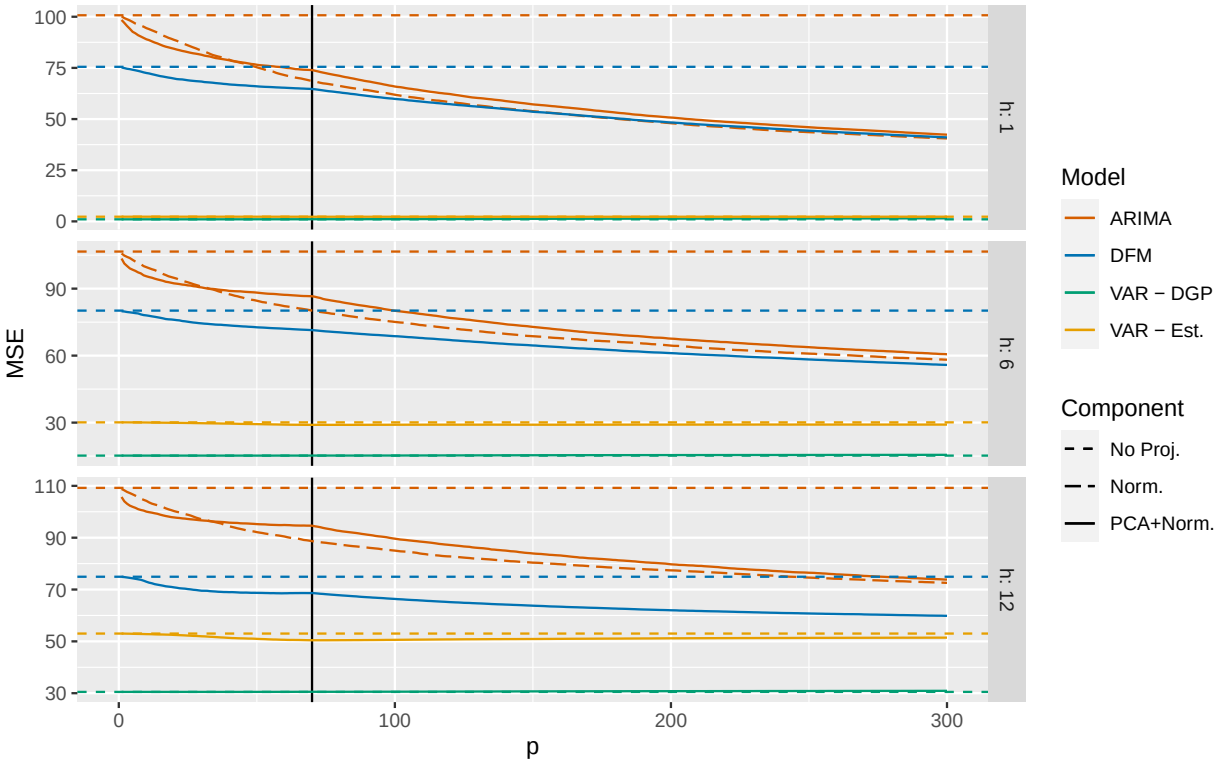

Figure 2: *MSE of different forecast models and component construction methods by the number of components p used in forecast projection in the simulation, for forecast horizons* 1*,* 6 *and* 12*. "VAR - DGP" indicates the performance of the true data generating VAR model. "VAR - Est." indicates the performance of the VAR model with the same structure as the true model and estimated parameter values. The vertical black line indicates the location of* $p = m$ *the number of series.*

3.2 Projection over base forecast

The first thing we note is the overall performance difference between the projected forecasts and the base forecast. In Figure 1, the average ranks of the projected forecasts are better than the corresponding base forecast at all forecast horizons, and the differences are all significant except the PCA-related ones with only $m = 70$ components for forecast horizon 6 and 12. Note here we need to compare forecasts with the same model: the projected forecast of ARIMA to base ARIMA and the projected forecast of DFM to base DFM. The number of components seems very important: the best-performing models are all with 300 components. Between the one-step-ahead forecasts, the methods with 300 components are not significantly different from each other, regardless of the forecast model or how we construct the components.

Indeed, from Figure 2 we can see the MSEs for model ARIMA and DFM keep decreasing from the base forecast, as the number of components increases. This confirms Theorem 2.2 that the more components we include, the more variance reduction we can achieve. This is only obvious in this ideal setting where we have 400 observations in each group while we only use at most 300 components in

the projection. This ample number of observations and the simple DGP can ease the challenge of estimation. This continued reduction in variance is not always achievable with real data, as we can see in Section 4, especially with FRED-MD in Section 4.2.

3.3 Base forecast model

If we compare ARIMA and DFM, under a VAR DGP, we expect DFM to pick up the correlation between series and not univariate ARIMA, so DFM should have better performance over ARIMA. This is indeed the case. Looking at the base forecasts in Figure 1, base DFM is significantly better than base ARIMA, except for $h = 1$ where it is close to significant. This is also very obvious in Figure 2. The horizontal line representing the MSE of base DFM is always far below the horizontal line for base ARIMA.

With the help of forecast projection, a simple model like ARIMA can achieve comparable performance to more sophisticated models like DFM. In Figure 1, all projected ARIMA forecasts except the ones having m PCs are significantly better than base DFM at $h = 1$, and all projected ARIMA with 300 components are significantly better than base DFM at $h = 6$. In Figure 2, The valid and long-dashed lines of projected ARIMA with corresponding component construction go down monotonically as the number of components p increases and reaches the MSE of base DFM at some point: at or below m for $h = 1$, at or above m for $h = 6$, and around $p = 300$ for $h = 12$. This is because forecast projection utilises shared information between series by capturing them in the components, making up for the overlooked correlation in univariate ARIMA models.

Interestingly enough, at $h = 1$, while the MSE of projected DFM also goes down as p increases, the MSE of the projected ARIMA and the MSE of the projected DFM seems to converge to the same value as p reaches 300, no matter how the components are constructed. Note here the same forecasts of the components, coming from univariate ARIMA of these components, are used for both the projected ARIMA and the projected DFM. This implies that much information in the series is not captured by ARIMA or DFM, but is captured by the components. As the number of components becomes high enough, the information captured by the components overpowers the information captured by the base models, dominating the performance of the projected forecasts. Once again, this emphasises the importance of the components and forecast projection. In this extreme case, the simple model is as good as the more complicated model after projection, because the forecast model itself is not as valuable as forecast projection.

This observation is not as obvious as the forecast horizon increases. This is because while ARIMA produces forecasts iteratively, DFM is a direct forecast model. With this simple DGP, the performance of DFM can be well maintained with larger h since a different model is fitted for each h. This can be

seen as the MSE of the base DFM does not change much with different h, but the MSE of base ARIMA keeps increasing as h increases.

3.4 Component constuction

The construction of components is obviously important in forecast projection, but might not be as important as expected in this simulation example. In Figure 1, the main difference that can be observed exists between using a combination of PCA and random weights, and purely using random weights. The distribution that generates the random weights is less relevant: in Figure 1, with the same number of components and the same base ARIMA model, the MSEs are not significantly different, regardless of whether the weights are simulated from a normal distribution (Norm) or a uniform distribution (Unif), or a combination of random orthonormal weights and random normal weights (Ortho+Norm). The same conclusion can be found when PCA is used. As long as PCA is used, the performance is not different whether the additional components are simulated from a normal distribution (PCA+Norm) or a uniform distribution (PCA+Unif).

Because the distribution is less important, in Figure 2, we only look at the inclusion of PCA with distribution set to normal. When p is smaller, the MSE drops faster when PCA is used, but the speed decreases as p increases. The performance of forecasts without PCA reaches and exceeds the performance with PCA before the number of components reaches m, and stays in the lead thereafter, although the gap seems to diminish with large p. This difference of PCA comes from the variances of principal components being maximised and ranked from largest to smallest, not from the orthogonality of the components, because the performance of using random orthonormal weight matrix is the same as using only random normal weights as discussed before. This might suggest the use of simple random weights if one is going to include a lot of components in the projection and to use PCA only when the number of components is small, but as we can see in Section 4, this is not the case with real data, and PCA is the preferred components even when the number of components is large.

Different constructions of components remain an important aspect of forecast projection. One important future direction would be to find alternative and optimal components, as we do not limit the structure of the weight matrix in Section 2. This should be studied together with the selection of the forecast model since both the weight matrix $\mathbf{\Phi}$ and the base forecast variance $\boldsymbol{W}_h$ are operatable and affect the projection simultaneously in Equation 2. This is likely to be an extension of the forecast combination literature, focusing on the properties of the base forecast, and the diversity and robustness of the forecast model and components. Examples of studies on this issue in the forecast combination literature include Batchelor & Dua (1995), Kang et al. (2022) and Lichtendahl & Winkler (2020).

3.5 Sources of uncertainty

At the bottom of each panel in Figure 2, the best forecasts come from the true DGP VAR model (the dashed green line that is partially covered by the solution green line). The forecast projection on the true model does not improve its forecast (the solid green line) as expected, as the uncertainty comes from the intrinsic error in the DGP that cannot be reduced. The second best forecast is from the estimated VAR model, as the uncertainty, apart from the intrinsic error, only comes from estimation error, not model misspecification error like ARIMA and DFM, which are both misspecified in this simulation example. The gap between the estimated VAR and the true VAR becomes bigger for a longer forecast horizon, because VAR produces iterative forecasts, and estimation error accumulates as h becomes larger.

Forecast projection shows little, if any, improvement over the estimated VAR. This means forecast projection cannot reduce estimation error or the parameter uncertainty described in Petropoulos, Hyndman & Bergmeir (2018). On the other hand, it shows significant improvement over misspecified base models. This implies that the uncertainty it can reduce is mainly the model misspecification error, or the model uncertainty in Petropoulos, Hyndman & Bergmeir (2018), similar to how bagging reduces variance. The data uncertainty in Petropoulos, Hyndman & Bergmeir (2018) is not examined and is less translatable to forecast projection, but it is partially the uncertainty associated with how the components are constructed as discussed in Section 3.4.

4 Empirical applications

Here we apply forecast projection to two real data examples and draw most of the same conclusions in simulation, with a few key differences.

4.1 Australian domestic tourism

The Australian Tourism Data Set compiled from the National Visitor Survey by Tourism Research Australia contains the total number of nights spent by Australians away from home, which we will refer to as visitor nights in what follows. The visitor nights are recorded monthly for each of the $m = 77$ regions, covering the period from January 1998 to December 2019. To measure the performance of forecast projection, we conduct time series cross-validation. The first $T = 84$ observations are kept as the training sample of the first evaluation, and the following 12 periods are taken as the test set on which the error is calculated. We repeat the evaluation for the rest of the data, with each training sample including one more observation than the previous one, and each test set shifting one period to the future.

The base forecasts of both the series and the components are produced by univariate ETS models selected and fitted using the `ets()` function in the `forecast` package (Hyndman et al. 2023; Hyndman & Khandakar 2008). In an ETS model, different term treats different patterns of a time series: the trend term treats the direction of the long-term tendency, the seasonality term treats the periodically recurring pattern with a fixed periodicity, and the error term measures the uncertainty. There does not need to be a trend term or seasonality term. If a trend term exists, we allow it to be additive or addictive damped. If a seasonality term exists, it can be addictive or multiplicative. The error can be additive or multiplicative. Excluding models with numerical instabilities, we choose the model with the smallest AICc among the 15 models.

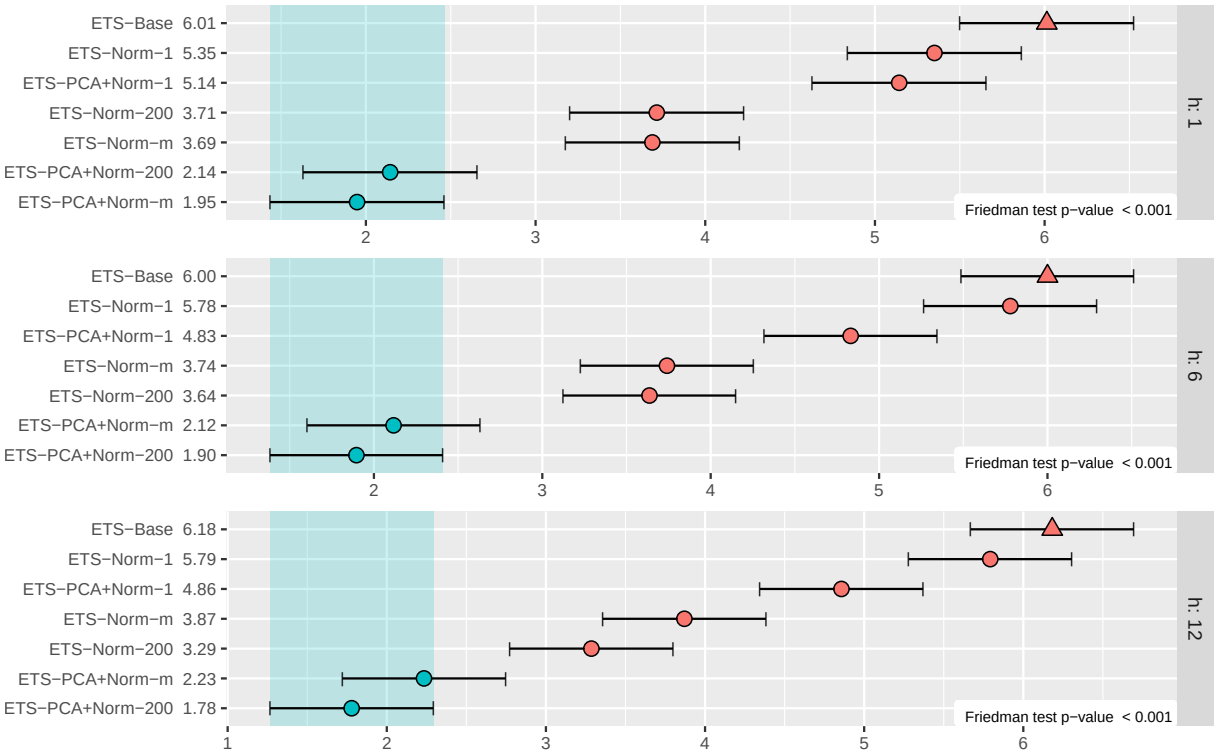

Figure 3: *Average ranks of* 1*-,* 6*- and* 12*-step-ahead cross-validation MSE of different model and component specifications on the visitor nights data. The methods using forecast projection are named as "{Model}-{Comp. Weights}-{No. Comp.}". The base models are named as "{Model}-Base" and these points are marked with triangles. The shaded region is the interval of the best performing model. Methods outside the shaded region are significantly worse than the best model.*

The MCB plot and the MSE plot can be found in Figure 3 and Figure 4. Most of the findings are consistent with Section 3. From Figure 3, the base forecast is always ranked last. Even projections with only one component are significantly better than the base forecast for $h = 6$ and 12.

We highlight two differences. Firstly, from Figure 4, the MSE of projection does not always go down as the number of components increases, especially for $h = 1$. This can also be seen from Figure 3, where the two best methods are not significantly different, even though they have very different

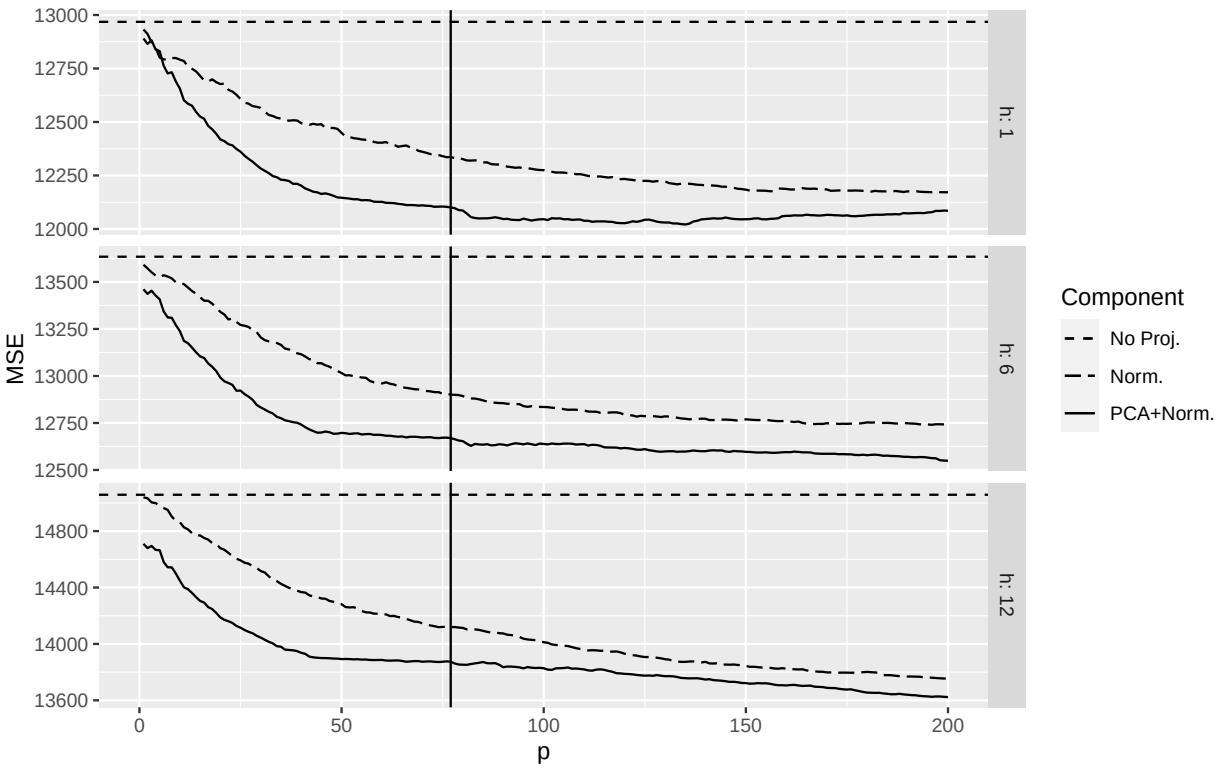

Figure 4: *MSE of different component construction methods by the number of components p used in forecast projection with ETS models on the visitor nights data, for forecast horizons* 1*,* 6 *and* 12*. The vertical black line indicates the location of $p = m$ the number of series.*

numbers of components $m = 77$ and $m = 200$. Intuitively, choosing the number of components is a tradeoff between the increasing estimation error as the dimension of forecast variance $\boldsymbol{W}_h$ increases, and the additional benefit brought by the information embedded in the new components, depending on the complexity of the DGP. For the visitor nights data set, the benefit of components above the estimation error diminishes after the number reaches about $m = 77$.

Secondly, unlike Section 3, using PCA as components is significantly better than simply using random normal weights, for the same large enough number of components (Figure 3). This is also clear from Figure 4, where the reduction of variance from using PCA is always in the lead, even after the m PCAs are exhausted and random normal weighted components are added. As discussed in Section 3.4, the rationale awaits future research, but we propose two potential explanations for this superior performance, related to the variance maximisation and ranking of PCA: 1. Optimality: By maximizing and ranking the variance of PCs from largest to smallest, we ensure that the projection utilizes components containing significantly more information (as measured by variance) compared to randomly weighted components; 2. Diversity: Actively seeking PCs with the highest variance results in the incorporation of a more diverse set of components into the projection.

4.2 FRED-MD

The FRED-MD (McCracken & Ng 2016) data set is a popular monthly data set for macroeconomic variables. It shares similar properties with the Stock & Watson (2002a) (and others) data. We download and transform the data set using the `fbi` (Chen, Ng & Bai 2023) package. The period we use for this exercise is from January 1959 to September 2023, containing 777 observations. Following McCracken & Ng (2016), we replace observations that deviate from the sample median by more than 10 interquartile ranges, which are recognised as outliers, with missing values. We then drop any series with more than 5% observations missing. This left us with $m = 122$ series. We fill in the missing values using the expectation-maximization (EM) algorithm described in Stock & Watson (2002b) with 8 factors. The number 8 is identified by McCracken & Ng (2016), albeit with a different time span. As the theory shows a reduction in forecast variance, we want to use MSE as the error measure, instead of other scaled or percentage error measures. To reliably calculate MSE over series with different scales, we demean the series to have mean 0 and scale the series to have variance 1. The MSEs are calculated on this standardised scale without back-transformation.

Similar to the visitor nights data, we evaluate the performance of forecasts using time series cross-validation. Starting with 300 observations in the first training set and the following 12 observations as the test set, we repeat the evaluation for the rest of the data with the size of the training set increasing 1 in each turn. The base models are the univariate ARIMA model and the DFM model, as described in Section 3. The ranges of the meta-parameters in the DFM models are $1 \leq k \leq 8$ (since 8 factors are identified and used to fill in the missing values), $1 \leq n \leq 3$ and $0 \leq s \leq 6$. The MCB plot and the MSE plot can be found in Figure 5 and Figure 6.

The first thing we note is the performance of base ARIMA exceeds that of the base DFM. This difference is significant at $h = 6$ (Figure 5). In terms of forecast projection, the best models at all forecast horizons are still forecast projections with PCA, and they are significantly better than the base models at $h = 1$ and 6. The fact that the projection with PCA is better than random weights, which can be seen from both the rankings in Figure 5 and the MSE in Figure 6, reaffirms our finding about the difference between PCA and random weights in Section 4.1.

In Figure 6, the forecast projection seems to be worse than the base models at the beginning, when the number of components is small, but improves with larger p and outperforms the base models gradually. As naturally the tradeoff between the benefit of new components and the difficulty of estimation of a large dimension still happens here, it seems to be more extreme: the MSEs start to increase as p increases, once p becomes larger than m. The projected forecast even worsens to the same level as the base forecast for ARIMA at $h = 6$ and DFM at $h = 12$ when $p = 300$. The

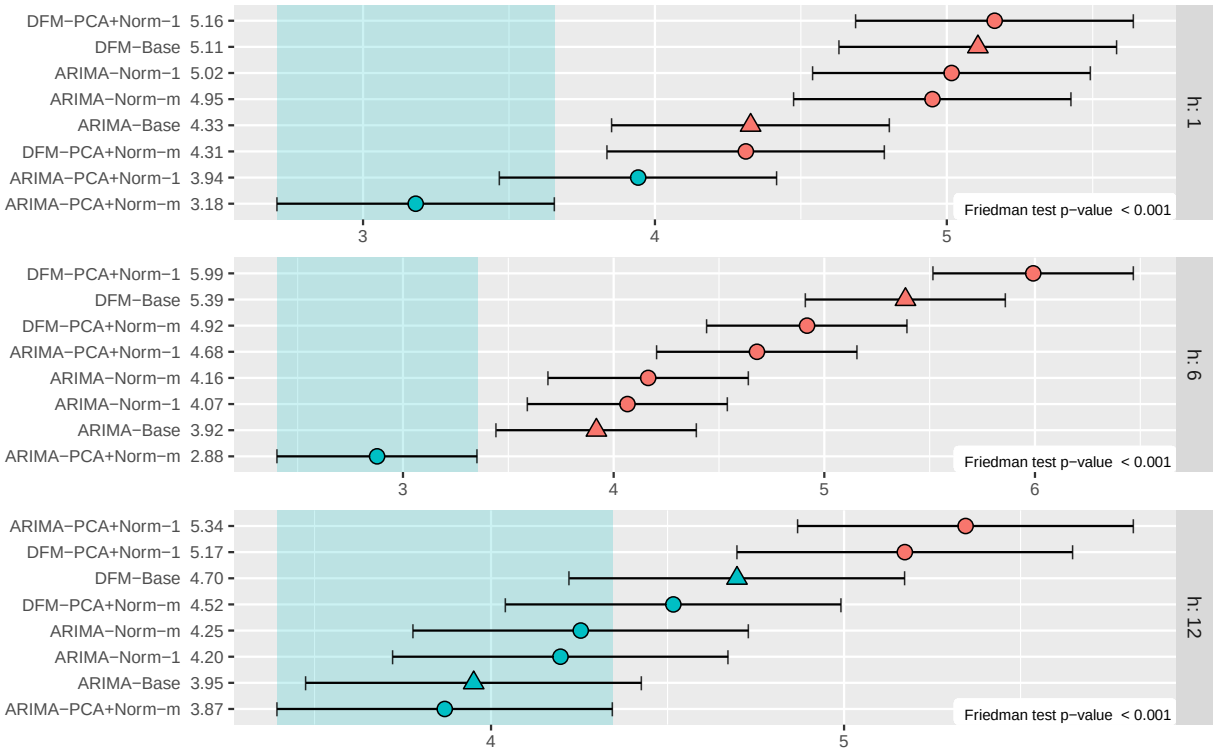

Figure 5: *Average ranks of* 1*-,* 6*- and* 12*-step-ahead cross-validation MSE of different model and component specifications on the FRED-MD data. The methods using forecast projection are named as "{Model}-{Comp. Weights}-{No. Comp.}". The base models are named as "{Model}-Base" and these points are marked with triangles. The shaded region is the interval of the best performing model. Methods outside the shaded region are significantly worse than the best model.*

turning point seems to be at m, but as we have seen in Section 3 and Section 4.1, m is not the clear cut-off point. Where the performance of forecast projection turns should be jointly determined by the number of series m, the sample size T, the component construction method, and the DGP. In the case of FRED-MD, it signals the importance of PCA, as m is the point that the component changes from PCA to random normal weighted linear combinations, implying PCA can exploit the information in the data while random weights cannot. This is more obvious for $h = 1$ and $h = 6$, as PCA works when $p < m$, but random normal weights do not seem to work from the beginning.

5 Conclusion

In this article, we propose a new approach called forecast projection that can reduce forecast variance on top of given forecasts without requiring any additional information, by projecting the forecasts of the series and the forecasts of their linear combinations in a specific way. We prove in theory, that the forecast projection can reduce forecast variance with the help of the components, and it can be reduced monotonically by including more components, assuming we know the base forecast variance matrix. Furthermore, we show that for a given number of components, within the class of linear

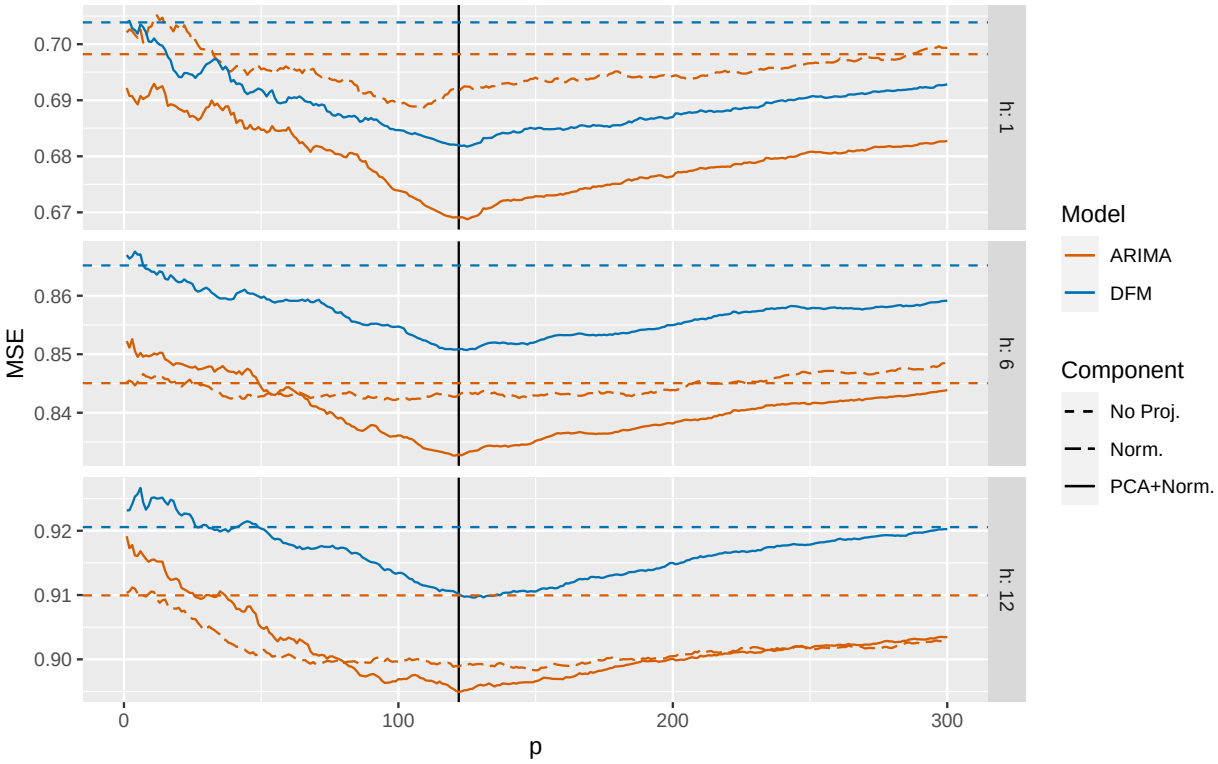

Figure 6: *MSE of different forecast models and component construction methods by the number of components p used in forecast projection on the FRED-MD data, for forecast horizons* 1*,* 6 *and* 12*. The vertical black line indicates the location of $p = m$ the number of series.*

projection, the proposed method is the best in the sense that it minimises the forecast variance of the series. To handle the difficulty of estimating the forecast variance matrix, we suggest using a shrinkage estimator, which shrinks variances toward their median and covariances toward zero.

We illustrate the proposed forecast projection outperforms base forecasts significantly and confirm the theoretical results in simulation and two empirical applications on the Australian domestic tourism data set and the FRED-MD data set. We find using PCA to construct components can achieve satisfactory variance reduction, and leave the issue of finding alternative optimal components to a later article. We recognise, in certain ideal cases, that the usage of forecast projection is even more important than the choice of base forecast model, emphasising the relative importance of forecast projection. We discuss that the source of the variance reduction is the reduction of model misspecification error and the forecast projection has little impact on estimation error.

References

Ando, S & F Narita (2022). An alternative proof of minimum trace reconciliation. *SSRN Electronic Journal*.

Assimakopoulos, V & K Nikolopoulos (2000). The theta model: a decomposition approach to forecasting. *International journal of forecasting* **16** (4), 521–530.

Athanasopoulos, G, RJ Hyndman, N Kourentzes & A Panagiotelis (2023). Forecast reconciliation: A review. *International journal of forecasting*.

Athanasopoulos, G, RJ Hyndman, N Kourentzes & F Petropoulos (2017). Forecasting with temporal hierarchies. *European journal of operational research* **262** (1), 60–74.

Batchelor, R & P Dua (1995). Forecaster diversity and the benefits of combining forecasts. *Management science* **41** (1), 68–75.

Bergmeir, C, RJ Hyndman & JM Benítez (2016). Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *International journal of forecasting* **32** (2), 303–312.

Borchers, HW (2023). *pracma: Practical Numerical Math Functions*. R package version 2.4.4. https://CRAN.R-project.org/package=pracma.

Breiman, L (1996). Bagging predictors. *Machine learning* **24** (2), 123–140.

Chen, Y(, S Ng & J Bai (2023). *fbi: Factor-Based Imputation and FRED-MD/QD Data Set*. R package version 0.7.0. https://github.com/cykbennie/fbi.

Cleveland, RB, WS Cleveland, JE McRae & I Terpenning (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of official statistics* **6** (1), 3–73.

De Stefani, J, YA Le Borgne, O Caelen, D Hattab & G Bontempi (2019). Batch and incremental dynamic factor machine learning for multivariate and multi-step-ahead forecasting. *International journal of data science and analytics* **7** (4), 311–329.

Di Fonzo, T & D Girolimetto (2023). Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives. *International journal of forecasting* **39** (1), 39–57.

Disney, SM & F Petropoulos (2015). Forecast combinations using multiple starting points. In: Logistics and Operations Management Section Annual Conference (Cardiff, Jan. 9, 2015).

Fabio Di Narzo, A, JL Aznarte & M Stigler (2009). *tsDyn: Time series analysis based on dynamical systems theory*. R package version 0.7. https://CRAN.R-project.org/package=tsDyn.

Friedman, M (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association* **32** (200), 675–701.

Friedman, M (1939). A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*.

Hastie, T, R Tibshirani & J Friedman (2003). *The elements of statistical learning: Data mining, inference, and prediction*. 1st ed. Springer series in statistics. New York, NY: Springer. 536 pp.

Hollander, M, DA Wolfe & E Chicken (2013). *Nonparametric Statistical Methods*. John Wiley & Sons. 848 pp.

Hyndman, R, G Athanasopoulos, C Bergmeir, G Caceres, L Chhay, M O'Hara-Wild, F Petropoulos, S Razbash, E Wang & F Yasmeen (2023). *forecast: Forecasting functions for time series and linear models*. R package version 8.21.1. https://pkg.robjhyndman.com/forecast/.

Hyndman, RJ & Y Khandakar (2008). Automatic time series forecasting: The forecast package for R. *Journal of statistical software* **27** (3), 1–22.

Jolliffe, IT (2002). *Principal Component Analysis*. Springer, New York, NY.

Kang, Y, W Cao, F Petropoulos & F Li (2022). Forecast with forecasts: Diversity matters. *European journal of operational research* **301** (1), 180–190.

Koning, AJ, PH Franses, M Hibon & HO Stekler (2005). The M3 competition: Statistical tests of the results. *International journal of forecasting* **21** (3), 397–409.

Kourentzes, N, F Petropoulos & JR Trapero (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International journal of forecasting* **30** (2), 291–302.

Kwiatkowski, D, PCB Phillips, P Schmidt & Y Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics* **54** (1-3), 159–178.

Li, X, F Petropoulos & Y Kang (2022). Improving forecasting by subsampling seasonal time series. *International journal of production research* **61** (3), 1–17.

Lichtendahl Jr, KC & RL Winkler (2020). Why do some combinations perform better than others? *International journal of forecasting* **36** (1), 142–149.

Luenberger, DG (1969). *Optimization by vector space methods*. Nashville, TN: John Wiley & Sons. 344 pp.

McCracken, MW & S Ng (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of business & economic statistics: a publication of the American Statistical Association* **34** (4), 574–589.

Nemenyi, PB (1963). "Distribution-free multiple comparisons". PhD thesis. Princeton University.

Opgen-Rhein, R & K Strimmer (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical applications in genetics and molecular biology* **6** (1).

Panagiotelis, A, G Athanasopoulos, P Gamakumara & RJ Hyndman (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International journal of forecasting* **37** (1), 343–359.

Petropoulos, F, RJ Hyndman & C Bergmeir (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European journal of operational research* **268** (2), 545–554.

Petropoulos, F & E Spiliotis (2021). The wisdom of the data: Getting the most out of univariate time series forecasting. *Forecasting* **3** (3), 478–497.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Schafer, J, R Opgen-Rhein, V Zuber, M Ahdesmaki, APD Silva & K Strimmer. (2021). *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*. R package version 1.6.10. https://CRAN.R-project.org/package=corpcor.

Schäfer, J & K Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology* **4** (1).

Stock, JH & MW Watson (2002a). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association* **97** (460), 1167–1179.

Stock, JH & MW Watson (2002b). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of business & economic statistics: a publication of the American Statistical Association* **20** (2), 147–162.

Stock, JH & MW Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of business & economic statistics: a publication of the American Statistical Association* **30** (4), 481–493.

Wang, X, RJ Hyndman, F Li & Y Kang (2023). Forecast combinations: An over 50-year review. *International journal of forecasting* **39** (4), 1518–1547.

Wickramasuriya, SL, G Athanasopoulos & RJ Hyndman (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association* **114** (526), 804–819.

A Proofs for Section 2 (Free-Lunch Forecast Projection)

Proof of Lemma 2.1.

$$\begin{aligned}\boldsymbol{C}\tilde{\boldsymbol{y}}_{t+h} &= \boldsymbol{C}\boldsymbol{M}\hat{\boldsymbol{y}}_{t+h} \\ &= \boldsymbol{C}(\boldsymbol{I}_{n+q} - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C})\hat{\boldsymbol{y}}_{t+h} \\ &= \boldsymbol{C}\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\hat{\boldsymbol{y}}_{t+h} \\ &= \boldsymbol{0}\end{aligned}$$

□

Proof of Lemma 2.2.

$$\boldsymbol{M}\boldsymbol{y}_{t+h} = (\boldsymbol{I}_{m+p} - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C})\boldsymbol{y}_{t+h} = \boldsymbol{y}_{t+h}$$

since $\boldsymbol{C}\boldsymbol{y}_{t+h} = \boldsymbol{0}$. □

Proof of Lemma 2.3.

$$\mathrm{E}(\tilde{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{M}\hat{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \boldsymbol{M}\,\mathrm{E}(\hat{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \boldsymbol{M}\,\mathrm{E}(\boldsymbol{y}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{M}\boldsymbol{y}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{y}_{t+h}|\mathscr{I}_t)$$

□

Proof of Lemma 2.4.

$$\mathrm{Var}(\tilde{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) = \mathrm{Var}(\boldsymbol{M}\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{M}\boldsymbol{y}_{t+h}) = \boldsymbol{M}\,\mathrm{Var}(\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h})\boldsymbol{M}' = \boldsymbol{M}\boldsymbol{W}_h\boldsymbol{M}'.$$

If we simplify it further, we have

$$\begin{aligned}\boldsymbol{M}\boldsymbol{W}_h\boldsymbol{M}' &= (\boldsymbol{I} - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C})\boldsymbol{W}_h(\boldsymbol{I} - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C})' \\ &= \boldsymbol{W}_h - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h \\ &\quad + \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h \\ &= \boldsymbol{W}_h - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h \\ &= \boldsymbol{M}\boldsymbol{W}_h.\end{aligned}$$

To get $\mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h})$, we just need to recognise that it is the first $m \times m$ leading principal submatrix of $\mathrm{Var}(\tilde{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h})$. □

Proof of Theorem 2.1. Trivially, $\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h$ and $\boldsymbol{J}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{J}'$ are positive semi-definite. Note that $\text{Var}(\hat{\boldsymbol{z}}_{t+h}-\boldsymbol{y}_{t+h})-\text{Var}(\tilde{\boldsymbol{z}}_{t+h}-\boldsymbol{y}_{t+h})$ is the leading principal submatrix of $\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h$, and the leading principal submatrix of a positive semi-definite matrix is positive semi-definite. □

Proof of Theorem 2.2. Suppose now that we want to include q more components $\boldsymbol{c}_t^* = \boldsymbol{\Phi}^*\boldsymbol{z}_t$ in the reconciliation. We define $\boldsymbol{y}_t^* = \begin{bmatrix}\boldsymbol{y}_t\\ \boldsymbol{c}_t^*\end{bmatrix}$, the constraint matrix

$$\boldsymbol{C}^* = \begin{bmatrix}\boldsymbol{C} & \underset{p\times q}{\boldsymbol{0}}\\ \underset{q\times m}{-\boldsymbol{\Phi}^*} & \underset{q\times p}{\boldsymbol{0}} & \boldsymbol{I}_q\end{bmatrix} = \begin{bmatrix}\underset{p\times m}{-\boldsymbol{\Phi}} & \boldsymbol{I}_p & \underset{p\times q}{\boldsymbol{0}}\\ \underset{q\times m}{-\boldsymbol{\Phi}^*} & \underset{q\times p}{\boldsymbol{0}} & \boldsymbol{I}_q,\end{bmatrix} = \begin{bmatrix}\overline{\boldsymbol{C}}\\ \underline{\boldsymbol{C}}\end{bmatrix} \tag{12}$$

where $\overline{\boldsymbol{C}}$ contains the first p rows of $\boldsymbol{C}^*$ and $\underline{\boldsymbol{C}}$ contains the remaining q rows of $\boldsymbol{C}^*$, the forecast variance covariance matrix

$$\text{Var}(\hat{\boldsymbol{y}}_{t+h}^* - \boldsymbol{y}_{t+h}^*) = \boldsymbol{W}_h^* = \begin{bmatrix}\boldsymbol{W}_h & \boldsymbol{W}_{yc,h}^*\\ \boldsymbol{W}_{cy,h}^* & \boldsymbol{W}_{c,h}^*\end{bmatrix}.$$

where $\hat{\boldsymbol{y}}_{t+h}^*$ is the h-step-ahead base forecasts of $\boldsymbol{y}_t^*$:

$$\hat{\boldsymbol{y}}_{t+h}^* = \begin{bmatrix}\hat{\boldsymbol{y}}_{t+h}\\ \hat{\boldsymbol{c}}_{t+h}^*\end{bmatrix},$$

and the corresponding

$$\boldsymbol{M}^* = \boldsymbol{I} - \boldsymbol{W}_h^*\boldsymbol{C}^{*\prime}(\boldsymbol{C}^*\boldsymbol{W}_h^*\boldsymbol{C}^{*\prime})^{-1}\boldsymbol{C}^*.$$

Proving Theorem 2.2 requires proving the following two items.

1. Including additional components in the mapping without including corresponding component constraints is equivalent to not including these additional components at all.
2. For a fixed set of components to be included in the mapping, adding constraints will reduce forecast variance.

We start by proving the first statement. Consider the case where we include the additional series $\boldsymbol{c}_t^*$ without using the additional constraint $\boldsymbol{\Phi}^*$. Defining $\boldsymbol{M}^+$ only with $\overline{\boldsymbol{C}}$:

$$\boldsymbol{M}^+ = \boldsymbol{I}_{m+p+q} - \boldsymbol{W}_h^*\overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}_h^*\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}}, \tag{13}$$

we have $\tilde{\boldsymbol{y}}^{+}_{t+h} = \boldsymbol{M}^{+}\hat{\boldsymbol{y}}^{*}_{t+h}$. Furthermore, we have

$$\boldsymbol{W}^{*}_{h}\overline{\boldsymbol{C}}' = \begin{bmatrix} \boldsymbol{W}_{h} & \boldsymbol{W}^{*}_{yc,h} \\ \boldsymbol{W}^{*}_{cy,h} & \boldsymbol{W}^{*}_{c} \end{bmatrix} \begin{bmatrix} \boldsymbol{C}' \\ \underset{q\times p}{\boldsymbol{0}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{W}_{h}\boldsymbol{C}' \\ \boldsymbol{W}^{*}_{cy,h}\boldsymbol{C}' \end{bmatrix}$$

and

$$\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_{h}\overline{\boldsymbol{C}}' = \begin{bmatrix} \boldsymbol{C} & \underset{p\times q}{\boldsymbol{0}} \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_{h}\boldsymbol{C}' \\ \boldsymbol{W}^{*}_{cy,h}\boldsymbol{C}' \end{bmatrix} = \boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}',$$

which gives

$$\begin{aligned} \boldsymbol{M}^{+} &= \boldsymbol{I}_{m+p+q} - \begin{bmatrix} \boldsymbol{W}_{h}\boldsymbol{C}' \\ \boldsymbol{W}^{*}_{cy,h}\boldsymbol{C}' \end{bmatrix} (\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1} \begin{bmatrix} \boldsymbol{C} & \underset{p\times q}{\boldsymbol{0}} \end{bmatrix} \\ &= \boldsymbol{I}_{m+p+q} - \begin{bmatrix} \boldsymbol{W}_{h}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1}\boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{W}^{*}_{cy,h}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1}\boldsymbol{C} & \boldsymbol{0} \end{bmatrix}, \end{aligned}$$

and

$$\begin{aligned} \tilde{\boldsymbol{y}}^{+}_{t+h} &= \boldsymbol{M}^{+}\hat{\boldsymbol{y}}^{*}_{t+h} \\ &= \left(\boldsymbol{I}_{m+p+q} - \begin{bmatrix} \boldsymbol{W}_{h}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1}\boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{W}^{*}_{cy,h}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1}\boldsymbol{C} & \boldsymbol{0} \end{bmatrix} \right) \begin{bmatrix} \hat{\boldsymbol{y}}_{t+h} \\ \hat{\boldsymbol{c}}^{*}_{t+h} \end{bmatrix} \\ &= \begin{bmatrix} (\boldsymbol{I}_{n+p} - \boldsymbol{W}_{h}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1}\boldsymbol{C})\hat{\boldsymbol{y}}_{t+h} \\ \hat{\boldsymbol{c}}^{*}_{t+h} - \boldsymbol{W}^{*}_{cy,h}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1}\boldsymbol{C}\hat{\boldsymbol{y}}_{t+h} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{M}\hat{\boldsymbol{y}}_{t+h} \\ \hat{\boldsymbol{c}}^{*}_{t+h} - \boldsymbol{W}^{*}_{cy,h}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1}\boldsymbol{C}\hat{\boldsymbol{y}}_{t+h} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\boldsymbol{y}}_{t+h} \\ \hat{\boldsymbol{c}}^{*}_{t+h} - \boldsymbol{W}^{*}_{cy,h}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_{h}\boldsymbol{C}')^{-1}\boldsymbol{C}\hat{\boldsymbol{y}}_{t+h} \end{bmatrix}. \end{aligned}$$

If we only consider the forecast performance relevant to $\boldsymbol{z}_{t+h}$, and define $\boldsymbol{J}^{*} = \boldsymbol{J}_{m,p+q} = \begin{bmatrix} \boldsymbol{I}_{m} & \boldsymbol{0}_{m\times(p+q)} \end{bmatrix}$, we have

$$\tilde{\boldsymbol{z}}^{+}_{t+h} = \boldsymbol{J}^{*}\tilde{\boldsymbol{y}}^{+}_{t+h} = \boldsymbol{J}\tilde{\boldsymbol{y}}_{t+h} = \tilde{\boldsymbol{z}}_{t+h}.$$

This means adding additional components without imposing the corresponding constraints will yield the same projected forecasts as if these additional components are not added, which implies that the

forecast variance stays the same:

$$\text{Var}(\tilde{\boldsymbol{z}}^{+}_{t+h} - \boldsymbol{z}_{t+h}) = \text{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) = \boldsymbol{J}\boldsymbol{M}\boldsymbol{W}_h\boldsymbol{J}'. \tag{14}$$

This finishes the proof of the first statement. Now we move on to proving the second statement. We have the forecast variance matrices

$$\begin{aligned} &\text{Var}(\tilde{\boldsymbol{y}}^{+}_{t+h} - \boldsymbol{y}^{*}_{t+h}) = \boldsymbol{M}^{+}\boldsymbol{W}^{*}_h = (\boldsymbol{I}_{m+p+q} - \boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}})\boldsymbol{W}^{*}_h \\ \text{and} \quad &\text{Var}(\tilde{\boldsymbol{y}}^{*}_{t+h} - \boldsymbol{y}^{*}_{t+h}) = \boldsymbol{M}^{*}\boldsymbol{W}^{*}_h = (\boldsymbol{I}_{m+p+q} - \boldsymbol{W}^{*}_h\boldsymbol{C}^{*\prime}(\boldsymbol{C}^{*}\boldsymbol{W}^{*}_h\boldsymbol{C}^{*\prime})^{-1}\boldsymbol{C}^{*})\boldsymbol{W}^{*}_h. \end{aligned}$$

Taking the difference, we have

$$\begin{aligned} \text{Var}(\tilde{\boldsymbol{y}}^{+}_{t+h} - \boldsymbol{y}^{*}_{t+h}) - \text{Var}(\tilde{\boldsymbol{y}}^{*}_{t+h} - \boldsymbol{y}^{*}_{t+h}) &= (\boldsymbol{W}^{*}_h\boldsymbol{C}^{*\prime}(\boldsymbol{C}^{*}\boldsymbol{W}^{*}_h\boldsymbol{C}^{*\prime})^{-1}\boldsymbol{C}^{*} - \boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}})\boldsymbol{W}^{*}_h \\ &= \boldsymbol{W}^{*}_h(\boldsymbol{C}^{*\prime}(\boldsymbol{C}^{*}\boldsymbol{W}^{*}_h\boldsymbol{C}^{*\prime})^{-1}\boldsymbol{C}^{*} - \overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}})\boldsymbol{W}^{*}_h. \end{aligned}$$

Using block matrix inversion, we have

$$\begin{aligned} \boldsymbol{C}^{*\prime}(\boldsymbol{C}^{*}\boldsymbol{W}^{*}_h\boldsymbol{C}^{*\prime})^{-1}\boldsymbol{C}^{*} &= \begin{bmatrix}\overline{\boldsymbol{C}}' & \underline{\boldsymbol{C}}'\end{bmatrix}\begin{bmatrix}\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}' & \overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}' \\ \underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}' & \underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}'\end{bmatrix}^{-1}\begin{bmatrix}\overline{\boldsymbol{C}} \\ \underline{\boldsymbol{C}}\end{bmatrix} \\ &= \begin{bmatrix}\overline{\boldsymbol{C}}' & \underline{\boldsymbol{C}}'\end{bmatrix}\begin{bmatrix}a & b \\ c & d\end{bmatrix}\begin{bmatrix}\overline{\boldsymbol{C}} \\ \underline{\boldsymbol{C}}\end{bmatrix} \\ &= \overline{\boldsymbol{C}}'a\overline{\boldsymbol{C}} + \overline{\boldsymbol{C}}'b\underline{\boldsymbol{C}} + \underline{\boldsymbol{C}}'c\overline{\boldsymbol{C}} + \underline{\boldsymbol{C}}'d\underline{\boldsymbol{C}}, \end{aligned}$$

where

$$\begin{aligned} a &= (\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1} + (\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}' \\ &\qquad (\underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}' - \underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}')^{-1}\underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1} \\ &= (\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1} + (\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}'(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}')^{-1}\underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}, \\ b &= -(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}'(\underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}' - \underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}')^{-1} \\ &= -(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}'(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}')^{-1}, \\ c &= -(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}')^{-1}\underline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}'(\overline{\boldsymbol{C}}\boldsymbol{W}^{*}_h\overline{\boldsymbol{C}}')^{-1}, \\ d &= (\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}^{*}_h\underline{\boldsymbol{C}}')^{-1}. \end{aligned}$$

Thus,

$$\begin{aligned}
\boldsymbol{C}^{*\prime}(\boldsymbol{C}^{*}\boldsymbol{W}_h^{*}\boldsymbol{C}^{*\prime})^{-1}\boldsymbol{C}^{*} &= \overline{\boldsymbol{C}}^{\prime}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}} \\
&\quad + \overline{\boldsymbol{C}}^{\prime}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}} \\
&\quad - \overline{\boldsymbol{C}}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}} \\
&\quad - \underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}}^{\prime}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}} \\
&\quad + \underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}} \\
&= \overline{\boldsymbol{C}}^{\prime}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}} \\
&\quad - \overline{\boldsymbol{C}}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}}\boldsymbol{M}^{+} \\
&\quad + \underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}}\boldsymbol{M}^{+} \\
&= \overline{\boldsymbol{C}}^{\prime}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}} + \boldsymbol{M}^{+\prime}\underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}}\boldsymbol{M}^{+}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathrm{Var}(\tilde{\boldsymbol{y}}_{t+h}^{+} - \boldsymbol{y}_{t+h}^{*}) - \mathrm{Var}(\tilde{\boldsymbol{y}}_{t+h}^{*} - \boldsymbol{y}_{t+h}^{*}) \\
&\quad = \boldsymbol{W}_h^{*}(\overline{\boldsymbol{C}}^{\prime}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}} + \boldsymbol{M}^{+\prime}\underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}}\boldsymbol{M}^{+} - \overline{\boldsymbol{C}}^{\prime}(\overline{\boldsymbol{C}}\boldsymbol{W}_h^{*}\overline{\boldsymbol{C}}^{\prime})^{-1}\overline{\boldsymbol{C}})\boldsymbol{W}_h^{*} \\
&\quad = \boldsymbol{W}_h^{*}(\boldsymbol{M}^{+\prime}\underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}}\boldsymbol{M}^{+})\boldsymbol{W}_h^{*}
\end{aligned}$$

is positive semi-definite. This concludes the proof of the second statement. Combining the results above, we have

$$\begin{aligned}
\mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) - \mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}^{*} - \boldsymbol{z}_{t+h}) &= \mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}^{+} - \boldsymbol{z}_{t+h}) - \mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}^{*} - \boldsymbol{z}_{t+h}) \\
&= \boldsymbol{J}^{*}\,\mathrm{Var}(\tilde{\boldsymbol{y}}_{t+h}^{+} - \boldsymbol{y}_{t+h}^{*})\boldsymbol{J}^{*\prime} - \boldsymbol{J}^{*}\,\mathrm{Var}(\tilde{\boldsymbol{y}}_{t+h}^{*} - \boldsymbol{y}_{t+h}^{*})\boldsymbol{J}^{*\prime} \qquad (15) \\
&= \boldsymbol{J}^{*}\boldsymbol{W}_h^{*}\boldsymbol{M}^{+\prime}\underline{\boldsymbol{C}}^{\prime}(\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\underline{\boldsymbol{C}}^{\prime})^{-1}\underline{\boldsymbol{C}}\boldsymbol{M}^{+}\boldsymbol{W}_h^{*}\boldsymbol{J}^{*\prime}
\end{aligned}$$

being positive semi-definite. Finally, we have

$$\begin{aligned}
&\mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) - \mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}^{*} - \boldsymbol{z}_{t+h})) - \mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) - \mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h})) \\
&\qquad = \mathrm{tr}(\mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) - \mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}^{*} - \boldsymbol{z}_{t+h}))
\end{aligned}$$

being the trace of a positive semi-definite matrix, which is non-negative. This means using a larger number of components in the mapping achieves a lower sum of forecast variance, giving Theorem 2.2.

□

Proof of Theorem 2.3. Denote $\boldsymbol{\psi}_i = \begin{bmatrix} -\boldsymbol{\phi}_i & \mathbf{0}_{1\times(i-1)} & 1 \end{bmatrix}$ and $\boldsymbol{W}_h^{(i)}$ to be the base forecast variance of the original series and the first i components. Starting with the first component, Equation 4 becomes

$$\operatorname{tr}(\boldsymbol{J}_{m,1}\boldsymbol{W}_h^{(1)}\boldsymbol{\psi}_1'(\boldsymbol{\psi}_1\boldsymbol{W}_h^{(1)}\boldsymbol{\psi}_1')^{-1}\boldsymbol{\psi}_1\boldsymbol{W}_h^{(1)}\boldsymbol{J}_{m,1}') = (\boldsymbol{\psi}_1\boldsymbol{W}_h^{(1)}\boldsymbol{\psi}_1')^{-1}\boldsymbol{\psi}_1\boldsymbol{W}_h^{(1)}\boldsymbol{J}_{m,1}'\boldsymbol{J}_{m,1}\boldsymbol{W}_h^{(1)}\boldsymbol{\psi}_1', \tag{16}$$

$$\text{where} \qquad \boldsymbol{\psi}_1\boldsymbol{W}_h^{(1)}\boldsymbol{J}_{m,1}' = \begin{bmatrix} -\boldsymbol{\phi}_1 & 1 \end{bmatrix}\begin{bmatrix} \boldsymbol{W}_{z,h} & \boldsymbol{W}_{zc_1,h} \\ \boldsymbol{W}_{c_1z,h} & \boldsymbol{W}_{c_1,h} \end{bmatrix}\begin{bmatrix} \boldsymbol{I}_m \\ 0 \end{bmatrix}$$
$$= -\boldsymbol{\phi}_1\boldsymbol{W}_{z,h} + \boldsymbol{W}_{c_1z,h}.$$

Equation 16 is obviously non-negative. For it to be larger than 0, we need $\boldsymbol{\psi}_1\boldsymbol{W}_h^{(1)}\boldsymbol{J}_{m,1}' \neq 0$, which gives $\boldsymbol{\phi}_1\boldsymbol{W}_{z,h} \neq \boldsymbol{W}_{c_1z,h}$.

When it comes to adding the ith component on top of the first $i-1$ components, we define

$$\overline{\boldsymbol{C}}_i = \begin{bmatrix} \boldsymbol{\psi}_1 & \mathbf{0}_{1\times i} \\ \boldsymbol{\psi}_2 & \mathbf{0}_{1\times(i-1)} \\ \vdots & \vdots \\ \boldsymbol{\psi}_i & 0 \end{bmatrix}$$

and

$$\boldsymbol{M}_i^+ = \boldsymbol{I}_{m+i} - \boldsymbol{W}_h^{(i)}\overline{\boldsymbol{C}}_{i-1}'(\overline{\boldsymbol{C}}_{i-1}\boldsymbol{W}_h^{(i)}\overline{\boldsymbol{C}}_{i-1}')^{-1}\overline{\boldsymbol{C}}_{i-1}$$

analogously to Equation 12 and Equation 13. Following Equation 15, the additional reduction of forecast variance when adding the ith component becomes

$$\boldsymbol{J}_{m,i}\boldsymbol{W}_h^{(i)}\boldsymbol{M}_i^{+\prime}\boldsymbol{\psi}_i'(\boldsymbol{\psi}_i\boldsymbol{M}_i^+\boldsymbol{W}_h^{(i)}\boldsymbol{\psi}_i')^{-1}\boldsymbol{\psi}_i\boldsymbol{M}_i^+\boldsymbol{W}_h^{(i)}\boldsymbol{J}_{m,i}' = (\boldsymbol{\psi}_i\boldsymbol{M}_i^+\boldsymbol{W}_h^{(i)}\boldsymbol{\psi}_i')^{-1}\boldsymbol{\psi}_i\boldsymbol{M}_i^+\boldsymbol{W}_h^{(i)}\boldsymbol{J}_{m,i}'\boldsymbol{J}_{m,i}\boldsymbol{W}_h^{(i)}\boldsymbol{M}_i^{+\prime}\boldsymbol{\psi}_i'.$$

Similar to before, we would want $\boldsymbol{\psi}_i\boldsymbol{M}_i^+\boldsymbol{W}_h^{(i)}\boldsymbol{J}_{m,i}' \neq \mathbf{0}$. Note that $\boldsymbol{\psi}_i$ concerns the first m rows and the last row of $\boldsymbol{M}_i^+\boldsymbol{W}_h^{(i)}$, and $\boldsymbol{J}_{m,i}'$ concerns the first m columns. Combined with the implication from Equation 14 that the $m\times m$ leading principal submatrix in equation $\boldsymbol{J}_{m,i}\boldsymbol{M}_i^+\boldsymbol{W}_h^{(i)}\boldsymbol{J}_{m,i}' = \boldsymbol{J}_{m,i-1}\boldsymbol{M}_{i-1}\boldsymbol{W}_h^{(i-1)}\boldsymbol{J}_{m,i-1}'$ is the same, we suppress the straightforward yet tiresome details, and obtain

$$\boldsymbol{\phi}_i\boldsymbol{W}_{\tilde{z},h}^{(i-1)} \neq \begin{bmatrix} \mathbf{0}_{1\times m+i-1} & 1 \end{bmatrix}\boldsymbol{M}_i^+\boldsymbol{W}_h^{(i)}\boldsymbol{J}_{m,i}',$$

where $\boldsymbol{W}_{\tilde{z},h}^{(i-1)} = \boldsymbol{J}_{m,i-1}\boldsymbol{M}_{i-1}\boldsymbol{W}_h^{(i-1)}\boldsymbol{J}_{m,i-1}'$ is the projected forecast variance of the original series using the first $i-1$ components, and the right hand side of the inequality is simply a one-row matrix

consisting of the first m elements in the last row of $\boldsymbol{M}_i^{+}\boldsymbol{W}_h^{(i)}$, which can be denoted as $\boldsymbol{W}_{\tilde{c}_i\tilde{\boldsymbol{z}},h}^{(i-1)}$ and interpreted as the covariance between the projected forecast of the original series using the first $i-1$ components, and the projected forecast of the ith component using the first $i-1$ components. □

Proof of Lemma 2.5.

$$\boldsymbol{C}\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{C}\boldsymbol{S}\tilde{\boldsymbol{z}}_{t+h} = \begin{bmatrix} -\boldsymbol{\Phi} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{\Phi} \end{bmatrix} \tilde{\boldsymbol{z}}_{t+h} = (-\boldsymbol{\Phi} + \boldsymbol{\Phi})\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{0}.$$

□

Proof of Lemma 2.6.

$$\boldsymbol{G}\boldsymbol{y}_{t+h} = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{y}_{t+h} = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S}\boldsymbol{z}_{t+h} = \boldsymbol{z}_{t+h}.$$

□

Proof of Lemma 2.7. If $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$, then

$$\mathrm{E}(\tilde{\boldsymbol{z}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \boldsymbol{G}\,\mathrm{E}(\hat{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \boldsymbol{G}\,\mathrm{E}(\boldsymbol{y}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{G}\boldsymbol{S}\boldsymbol{z}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{z}_{t+h}|\mathscr{I}_t).$$

□

Proof of Lemma 2.8. Let the base and projected forecast errors be given as

$$\begin{aligned}
&\hat{\boldsymbol{e}}_{z,t+h} = \boldsymbol{z}_{t+h} - \hat{\boldsymbol{z}}_{t+h},\\
&\hat{\boldsymbol{e}}_{y,t+h} = \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h},\\
&\tilde{\boldsymbol{e}}_{z,t+h} = \boldsymbol{z}_{t+h} - \tilde{\boldsymbol{z}}_{t+h},\\
\text{and}\quad &\tilde{\boldsymbol{e}}_{y,t+h} = \boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}\boldsymbol{z}_{t+h} - \boldsymbol{S}\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{S}\tilde{\boldsymbol{e}}_{z,t+h}.\\
\text{Then we have}\quad &\tilde{\boldsymbol{e}}_{y,t+h} = \hat{\boldsymbol{e}}_{y,t+h} + \hat{\boldsymbol{y}}_{t+h} - \tilde{\boldsymbol{y}}_{t+h}\\
&\qquad\qquad = \hat{\boldsymbol{e}}_{y,t+h} + \hat{\boldsymbol{y}}_{t+h} - \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h}\\
&\qquad\qquad = \hat{\boldsymbol{e}}_{y,t+h} + (\boldsymbol{I} - \boldsymbol{S}\boldsymbol{G})(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{e}}_{y,t+h})\\
\text{and}\quad &\qquad\qquad = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{e}}_{y,t+h} + (\boldsymbol{I} - \boldsymbol{S}\boldsymbol{G})\boldsymbol{S}\boldsymbol{z}_{t+h}\\
&\boldsymbol{S}\tilde{\boldsymbol{e}}_{z,t+h} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{e}}_{y,t+h},
\end{aligned}$$

where the last line comes from $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$. Left multiply $\boldsymbol{G}$ to both sides, we have

$$\boldsymbol{GS}\tilde{\boldsymbol{e}}_{z,t+h} = \boldsymbol{GSG}\hat{\boldsymbol{e}}_{y,t+h} \qquad \text{and} \qquad \tilde{\boldsymbol{e}}_{z,t+h} = \boldsymbol{G}\hat{\boldsymbol{e}}_{y,t+h},$$

and therefore

$$\text{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) = \text{Var}(\tilde{\boldsymbol{e}}_{z,t+h}) = \text{Var}(\boldsymbol{G}\hat{\boldsymbol{e}}_{y,t+h}) = \boldsymbol{G}\,\text{Var}(\hat{\boldsymbol{e}}_{y,t+h})\boldsymbol{G}' = \boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}'.$$

□

Proof of Theorem 2.4. This can be proved in a few different ways. We adopt the approach of Ando & Narita (2022) to obtain the solution to Equation 10, but the procedure from Luenberger (1969, p. 85) can also be used, where the problem is divided to Equation 11 and reconstructed to find the solution to Equation 10.

There exists a Lagrange multiplier $\boldsymbol{\Lambda}$ such that

$$L(\boldsymbol{G}) = \text{tr}(\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}') + \text{tr}(\boldsymbol{\Lambda}'(\boldsymbol{I} - \boldsymbol{GS}))$$

is stationary at an extremum $\boldsymbol{G}$ (Luenberger 1969, p. 243, Theorem 1). We set the Gateaux differential (Luenberger 1969, p. 171) to zero for any matrix $\boldsymbol{H}$:

$$\begin{aligned}
\lim_{\alpha\to 0}\frac{L(\boldsymbol{G}+\alpha\boldsymbol{H}) - L(\boldsymbol{G})}{\alpha} &= 0 \\
\text{tr}(\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{H}') + \text{tr}(\boldsymbol{H}\boldsymbol{W}_h\boldsymbol{G}') - \text{tr}(\boldsymbol{\Lambda}'(\boldsymbol{HS})) &= \text{tr}(2\boldsymbol{H}\boldsymbol{W}_h\boldsymbol{G}' - \boldsymbol{\Lambda}'\boldsymbol{HS}) \\
&= \text{tr}(\boldsymbol{H}(2\boldsymbol{W}_h\boldsymbol{G}' - \boldsymbol{S}\boldsymbol{\Lambda}')) \\
&= 0 \\
2\boldsymbol{W}_h\boldsymbol{G} &= \boldsymbol{S}\boldsymbol{\Lambda}' \\
\boldsymbol{G}' &= \frac{1}{2}\boldsymbol{W}_h^{-1}\boldsymbol{S}\boldsymbol{\Lambda}'.
\end{aligned}$$

Multiplying $\boldsymbol{S}'$ to the left of both sides. we have

$$\boldsymbol{S}'\boldsymbol{G}' = \boldsymbol{I} = \frac{1}{2}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S}\boldsymbol{\Lambda}' \qquad \text{and} \qquad \boldsymbol{\Lambda}' = 2(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}$$

because $\boldsymbol{GS} = \boldsymbol{I}$. Putting it back in, we have

$$\boldsymbol{G}' = \boldsymbol{W}_h^{-1}\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1} \qquad \text{and} \qquad \boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}.$$

□