Monash University

Monash Business School

ISSN 1440-771X

Department of Econometrics and Business Statistics

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

Forecast Linear Augmented Projection (FLAP): A free lunch to reduce forecast error variance

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

March 2024

Working Paper ??/??

Forecast Linear Augmented Projection (FLAP): A free lunch to reduce forecast error variance

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

20 March 2024

Forecast Linear Augmented Projection (FLAP): A free lunch to reduce forecast error variance

Abstract

A novel forecast linear augmented projection (FLAP) method that provably reduces the forecast error variance of any unbiased multivariate forecast without introducing bias, is introduced. The method first constructs new series as linear combinations of the original series. Forecasts are then generated for both the original and new series. Finally, the full vector of forecasts are projected onto a linear subspace where the constraints implied by the combination weights hold. It is proven that the trace of the forecast variance is non-increasing with the number of components, and mild conditions are established for which it is strictly decreasing. It is also shown that the proposed method achieves maximum forecast variance reduction among linear projections. The theoretical results are validated through simulations and two empirical applications based on Australian tourism and FRED-MD data. Notably, using FLAP with Principal Component Analysis (PCA) to construct the new series leads to substantial forecast error variance reduction.

1 Introduction

Multivariate forecasting arises in a number of disciplines including macroeconomics and finance; see Carriero, Galvao & Kapetanios (2019) and Tsay (2013) respectively, and references therein. We introduce a new post processing framework that (i) augments the data by constructing new series that are linear combinations of the original series, (ii) forecasts both the original and new series and (iii) recovers a new set of forecasts for the original series via projections. We refer to this method as Forecast Linear Augmented Projection (FLAP). We prove that the method reduces the forecast variance of the original series in a way that is agnostic both with respect to the weights of the linear combinations used at step (i) and with respect to the model used to generate forecasts at step (ii). The model is inspired by the forecast reconciliation literature (Athanasopoulos et al. 2023) whereby forecasts are adjusted to cohere with known linear constraints. In contrast to that literature, the FLAP method focuses on multivariate forecasting where such constraints are not present. Indeed, the method need not only be applied to forecasting problems, but multivariate predictions in general.

It may appear puzzling that forecast accuracy can be improved, not by introducing any new information, but by simply taking linear combinations of existing time series. To give an intuition into how this puzzle can be resolved, we consider a toy example of two series z_1 and z_2 that are of concern to the forecaster and two linear combinations or *components* of the series $c_1 = 0.5z_1 + 0.5z_2$ and $c_2 = 0.5z_1 - 0.5z_2$. Denote by $\hat{z}_1$, $\hat{z}_2$, $\hat{c}_1$, $\hat{c}_2$ any forecasts of these original series and components, which we collectively refer to as *base forecasts*. The base forecasts may be generated by univariate methods, multivariate methods, or even based on expert judgement. When considering z_1, there is both a *direct* forecast $\hat{z}_1$ and *indirect* forecast $\hat{c}_1 + \hat{c}_2$, similarly for z_2 the direct forecast is $\hat{z}_2$ and the indirect forecast $\hat{c}_1 - \hat{c}_2$.[1] With the exception of some pathological cases, the direct and indirect forecasts for the same variable will not, in general, be equal. Therefore forecast accuracy can be improved by combining direct and indirect forecasts, something implicitly achieved by the proposed FLAP method. The puzzle is thus resolved; while no new information is created at the data augmentation step, there is new information embedded into forecasts of the augmented series, which can be leveraged via model combination (see Wang et al. (2023) for a review of forecast combination). Something obscured by the simple toy example is the way our FLAP method differs from the usual forecast combination methods, in particular our combinations are potentially non-convex since they are obtained via projections, in a way that we now elaborate upon.

More formally and more generally, the FLAP method considers a vector of original series $\mathbf{z} \in \mathbb{R}^m$ and a vector of components $\mathbf{c} \in \mathbb{R}^p$. While $\left(\mathbf{z}', \mathbf{c}'\right)'$ is a $p+m$-vector, the construction of components as linear combinations of the original series implies that $\left(\mathbf{z}', \mathbf{c}'\right)'$ lies on a linear subspace of at most dimension m. The corresponding vector of forecasts $\left(\hat{\mathbf{z}}', \hat{\mathbf{c}}'\right)'$ will, in general, have support on $\mathbb{R}^{m+p}$. FLAP projects $\left(\hat{\mathbf{z}}', \hat{\mathbf{c}}'\right)'$ onto the m-dimensional linear subspace on which $\left(\mathbf{z}', \mathbf{c}'\right)'$ has support. The setup of this problem bears similarities to the well known problem of forecast reconciliation where Panagiotelis et al. (2021) provide similar geometric intuition, while Wickramasuriya, Athanasopoulos & Hyndman (2019), Athanasopoulos et al. (2017), and Di Fonzo & Girolimetto (2023) have all shown that reconciliation can reduce forecast variance theoretically and empirically. However, we note that these papers establish that reconciliation improves forecast accuracy for the hierarchy *as a whole*. In the general multivariate setting that we consider, this would imply improvements in forecast accuracy for $\mathbf{z}$ and $\mathbf{c}$ taken together. This poses a problem if improvements in forecast accuracy for $\mathbf{c}$ could be offset by a deterioration in forecast accuracy for $\mathbf{z}$, since the former are not of interest in and of themselves. A key insight we make in this paper in the hieararchical setting, is that reductions in forecast error variance accrue even for a subset of variables in the hierarchy. It is this contribution

[1] This argument, as well as the terminology direct and indirect forecasts, is inspired by Hollyman, Petropoulos & Tipping (2021) who discuss this in the forecast reconciliation setting.

that allows us to propose a method that goes beyond the case where time series adhere to linear constraints, and that instead applies to the more general setting.

While the theoretical results apply for any linear combinations of the original series, in practice we propose to augment the data with principal components. When doing so, the FLAP method bears a resemblance to Dynamic Factor Models (DFMs), specifically those common in macroeconomic forecasting (Stock & Watson 2002b,a, 2012), their extensions in the machine learning literature (De Stefani et al. 2019) as well as the factor augmented VAR (Bernanke, Boivin & Eliasz 2005). The factor models assume that the multivariate time series possesses common components and the dynamics of the observed series are governed by the dynamics of these unobserved factors, often assumed to follow some parametric model. In contrast, FLAP is a post-forecasting step, indeed forecasts can even be made using a DFM and then further improved by implementing the FLAP method, something we demonstrate in Section 3 and Section 4.

In the sense that forecast accuracy can be improved without any new information, FLAP has parallels with bootstrap aggregation or "bagging" (Breiman 1996; Bergmeir, Hyndman & Benítez 2016). Bagging can reduce prediction variance without increasing bias (Hastie, Tibshirani & Friedman 2003), by mitigating model uncertainty (Petropoulos, Hyndman & Bergmeir 2018), and does so without introducing any new data, but rather resampled versions of the existing data. Our FLAP method also reduces forecast variance without introducing new data, but using linear combinations of the existing data, rather than bootstrapping. The FLAP method (in addition to forecast reconciliation and bagging) can be viewed as contributing to the literature where forecasts are improved by combination and data augmentation methods. This includes the 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); a review of all these methods can be found in Petropoulos & Spiliotis (2021) who refer to them as using "the wisdom of data". Our FLAP method is distinct in that it aims to exploit information in the data with a focus on linear combinations of multivariate series.

The remainder of the paper is structured as follows. In Section 2, we propose the FLAP 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 FLAP 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. The methods introduced in this paper are implemented in the `flap` package (Yang 2024). This paper is fully reproducible with code and documentation provided at https://github.com/FinYang/paper-forecast-projection.

2 Forecast Linear Augmented Projection

2.1 Forecasted Linear Augmented Projections (FLAP) method

In the following, all vectors and matrices are denoted in bold font. 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.

Let $\boldsymbol{z}_t \in \mathbf{R}^m$ be a vector of m observed time series at time t, that are the series that we are interested in forecasting. The FLAP method involves three steps:

1. *Form components*. Set $\boldsymbol{c}_t = \boldsymbol{\Phi}\boldsymbol{z}_t \in \mathbf{R}^p$ as a vector of p linear combinations of $\boldsymbol{z}_t$ at time t, where $\boldsymbol{\Phi} \in \mathbf{R}^{p\times m}$. We call $\boldsymbol{c}_t$ the components of $\boldsymbol{z}_t$ and the component weights $\boldsymbol{\Phi}$ are known in the sense that they are chosen by the statistician using the FLAP method. Letting $\boldsymbol{y}_t = \left[\boldsymbol{z}_t', \boldsymbol{c}_t'\right]'$ be the concatenation of series $\boldsymbol{z}_t$ and components $\boldsymbol{c}_t$. We note that $\boldsymbol{y}_t$ will be constrained in the sense that $\boldsymbol{C}\boldsymbol{y}_t = \boldsymbol{c}_t - \boldsymbol{\Phi}\boldsymbol{z}_t = \boldsymbol{0}$ for any t where $\boldsymbol{C} = \begin{bmatrix} -\boldsymbol{\Phi} & \boldsymbol{I}_p \end{bmatrix}$ is referred to as the constraint matrix
2. *Produce forecasts* of $\boldsymbol{y}_t$. The h-step-ahead base forecast of $\boldsymbol{y}_t$ is denoted $\hat{\boldsymbol{y}}_{t+h}$. The method used to produce forecasts can be univariate or multivariate. In the more general setting where $\boldsymbol{y}_t$ are not time series but cross sectional data, any prediction method can be used. In general, the constraints that hold for $\boldsymbol{y}_t$ will not hold for $\hat{\boldsymbol{y}}_{t+h}$, i.e $\boldsymbol{C}\hat{\boldsymbol{y}}_{t+h} \neq \boldsymbol{0}$
3. *Project the base forecasts* onto the space where the constraints hold:

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

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 $\mathrm{Var}(\hat{\boldsymbol{y}}_{t+h} - \boldsymbol{y}_{t+h}) = \boldsymbol{W}_h$ is the forecast covariance matrix. For the proofs of this section, we will assume that $\boldsymbol{W}_h$ is known, in practice a plug-in estimate can be used that will be discussed in Section 2.5.

In practice we are not interested in forecasts of the full vector $\boldsymbol{y}_t$ but of $\boldsymbol{z}_t$. We now introduce some more notation to handle this issue. 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 $\hat{\boldsymbol{z}}_{t+h}$ and $\tilde{\boldsymbol{z}}_{t+h}$ denote the first m 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 now present some theoretical results regarding the FLAP method, with proofs provided in the Appendix. The main results are Theorem 2.1, which establishes that the FLAP method dominates the base forecasts, in the sense that the difference between the forecast error covariance matrices of $\hat{\boldsymbol{z}}_{t+h}$ and $\tilde{\boldsymbol{z}}_{t+h}$ is always positive definite. Theorem 2.2 establishes that the trace of the variance covariance of $\tilde{\boldsymbol{z}}_{t+h}$ is non-increasing in the number of components p. The conditions needed to make the trace strictly decreasing are discussed in Theorem 2.3. In Theorem 2.4 we prove that the projection in Equation 2 is optimal amongst the class of all projections. Finally, while the theoretical results imply that components could in principle continue to be added to improve forecasts, in practice, larger values of p will make estimation of the plug-in variance covariance matrix $\boldsymbol{W}_h$ unreliable. We explore this issue empirically in Section 3 and Section 4.

2.2 Positive Semi-Definiteness of Variance Reduction

We first provide some intermediate results.

Lemma 2.1. *The matrix $\boldsymbol{M}$ is a projection onto the space where the constraint $\boldsymbol{C}$ is satisfied.*

Based on the attractive properties of projections, we have the following corollaries.

Corollary 2.1.

1. *The projected forecast $\tilde{\boldsymbol{y}}_{t+h}$ satisfies the constraint $\boldsymbol{C}$.*
2. *For $\boldsymbol{y}_{t+h}$ that already satisfies the constraint, the projection does not change its value: $\boldsymbol{M}\boldsymbol{y}_{t+h} = \boldsymbol{y}_{t+h}$ (Rao 1974, Lemma 2.4).*
3. *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 projected forecasts are also unbiased: $\mathrm{E}(\tilde{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{y}_{t+h}|\mathscr{I}_t)$.*

Note that the unbiasedness of the base forecasts is true for most common forecasting models, and can otherwise be easily achieved with a simple bias correction 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 2.2. *The forecast covariance matrix of the component-constrained projected h-step-ahead forecasts* $\tilde{\boldsymbol{y}}_{t+h}$ *is* $\boldsymbol{M}\boldsymbol{W}_h$*, i.e.*

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

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

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

Lemma 2.2 is a well-known result in the forecast reconciliation literature (e.g., Di Fonzo & Girolimetto 2023).

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

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

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

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

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 the reduction in forecast variances, and is non-negative because the matrix is positive semi-definite. It means we can reduce the forecast variance by simply forecasting the components (the artificially constructed linear combinations 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. Suppose $\boldsymbol{y}_t$ comprises m original series and p components whose forecasts are all uncorrelated with each other and have variance 1. Then $\boldsymbol{W}_h = \boldsymbol{I}_{m+p}$, and

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

where

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

Let $\boldsymbol{\Phi}$ consist of orthogonal unit vectors, for example, those obtained from Principal Component Analysis (PCA, Jolliffe 2002). That is,

$$\boldsymbol{\Phi}\boldsymbol{\Phi}' = \boldsymbol{I}_p \text{ when } p \le m \qquad \text{and} \qquad \boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{I}_m \text{ when } p = m.$$

Then

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

We only focus on the forecast variance reduction for the original series, which is $\text{tr}(\text{Var}(\hat{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h}) - \text{Var}(\tilde{\boldsymbol{z}}_{t+h} - \boldsymbol{z}_{t+h})) = \frac{1}{2}\,\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi})$. When $p < m$, since $\boldsymbol{\Phi}'\boldsymbol{\Phi}$ is idempotent, we have $\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = \text{rank}(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = p$. The reduction in the forecast variance of the original series is $p/2$. When $p = m$, we have $\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = \text{tr}(\boldsymbol{I}_m) = m$, which means the reduction on the sum is $m/2$ and the reduction on each individual series is $1/2$. If we have two series ($m = 2$) to begin with, using $p = 1$ component in the mapping will reduce the sum of forecast variance by 0.5, and using $p = 2$ components will reduce the sum of forecast variance by 1; that is a 50% reduction as the original sum of forecast variances is 2.

If we keep increasing the number of components, the result in Theorem 2.1 still holds, although $\boldsymbol{\Phi}$ can no longer contain orthogonal vectors, and the example here becomes intractable. This is an artificial example because the forecast variance $\boldsymbol{W}_h$ can hardly be an 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 the component forecasts bringing new information about the original series.

The variance reduction becomes larger as we increase the number of components p from 1 to 2 in Example 2.1. This is not a coincidence but a desirable property of forecast projection, as shown in the next section.

2.3 Monotonicity

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

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

where $\boldsymbol{W}_{z,h}$ is the forecast covariance matrix of $\hat{\boldsymbol{z}}_{t+h}$, $\boldsymbol{W}_{c,h}$ is the forecast covariance matrix of $\hat{\boldsymbol{c}}_{t+h}$, and $\boldsymbol{W}_{zc,h}$ contains covariances between elements of $\hat{\boldsymbol{z}}_{t+h}$ and $\hat{\boldsymbol{c}}_{t+h}$.

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

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

is non-decreasing as p increases.

Theorem 2.2 is the key result that demonstrates the usefulness of forecast projection. It means that 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 base forecast 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 (with upper right corner all 0s). This implies that the components can also be constructed from existing components, not only from the original series. This has little significance since a linear combination of components 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. 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 $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_1' & \boldsymbol{\phi}_2' & \dots & \boldsymbol{\phi}_p' \end{bmatrix}'$. Let $\boldsymbol{W}_{\tilde{z},h}^{(i-1)}$ denote the covariance matrix of the projected forecasts of the original series based on the first $i-1$ components, $\boldsymbol{w}_{c_1 z,h}$ denote a vector of covariances of the first component and the base forecasts of the original series, and $\boldsymbol{w}_{\tilde{c}_i\tilde{z},h}^{(i-1)}$ denote a vector of covariances of the projected ith component, and the projected forecasts of the original series, based on the first $i-1$ components.

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:*

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

For the ith component to have a positive reduction on forecast variance of the original series, the following condition must be satisfied:

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

We can interpret the condition of Equation 5 in the following way: for the new component to be beneficial, the information brought by this new component, measured as the covariance, cannot be a linear 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 to 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.

2.4 Optimality of the projection

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

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

That is, the projection can be interpreted as finding the set of forecasts that are closest (on the transformed space) to the base forecasts, while satisfying the linear constraints imposed by the components.

Moreover, this is equivalent to the optimisation problem:

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

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 simpler 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, so that $\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{z}_t$, 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 of 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 components, to the space of the original series, leading to the solution

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

as equivalent to Equation 3. Recognising that Equation 7 is equivalent to Equation 1, it is the solution that minimises the sum of forecast variances of the original series and all the components. We go further in Theorem 2.4, and show that Equation 7 is also the solution to minimise each individual forecast variance of the original series, and their sum. 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 earliest work we can find that noted this interpretation in a non-forecasting context is 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, Corollary 2.1 and Lemma 2.2 hold under this alternative representation.

Lemma 2.3. *The matrix $\boldsymbol{SG}$ is a projection onto the space where the constraint $\boldsymbol{C}$ is satisfied, provided that $\boldsymbol{GS} = \boldsymbol{I}$.*

Corollary 2.2. *Provided that $\boldsymbol{GS} = \boldsymbol{I}$, the following results hold.*

1. *The projected forecast in Equation 9 satisfies the constraint $\boldsymbol{C}\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{C}\boldsymbol{S}\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{0}$.*
2. *For $\boldsymbol{y}_{t+h}$ that already satisfies the constraint, the mapping does not change its value: $\boldsymbol{G}\boldsymbol{y}_{t+h} = \boldsymbol{z}_{t+h}$.*
3. *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 projected forecasts in Equation 9 are also unbiased: $\mathrm{E}(\tilde{\boldsymbol{z}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{z}_{t+h}|\mathscr{I}_t)$.*

Lemma 2.4. *The covariance matrix of the projected forecasts from Equation 9 is given by*

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

Putting these together, we have the following theorem.

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

$$\underset{\boldsymbol{G}}{\arg\min}\ \mathrm{tr}(\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}') \qquad s.t.\ \boldsymbol{GS} = \boldsymbol{I} \tag{10}$$

is Equation 8. This problem can be effectively split into independent subproblems such that $\boldsymbol{G} = \begin{bmatrix} \boldsymbol{g}_1 & \boldsymbol{g}_2 & \dots & \boldsymbol{g}_m \end{bmatrix}'$*, where* $\boldsymbol{g}_i$ *is the solution to the subproblem of the ith series*

$$\underset{\boldsymbol{g}_i}{\arg\min}\ \boldsymbol{g}_i'\boldsymbol{W}_h\boldsymbol{g}_i \qquad s.t.\ \boldsymbol{g}_i'\boldsymbol{s}_j = \delta_{ij}, \quad j = 1, 2, \dots, m, \tag{11}$$

where $\boldsymbol{s}_j$ *is the jth column of* $\boldsymbol{S}$*, and* δ_{ij} *is the Kronecker delta function taking value 1 if* $i = j$ *and 0 otherwise.*

In other words, the forecast projection method gives optimal projected forecast for a given set of components, in the sense that the unbiased forecast of each series has minimum variance.

2.5 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 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 covariance shrinkage method of Schäfer & Strimmer (2005), which is the "MinT(Shrink)" method of Wickramasuriya, Athanasopoulos & Hyndman (2019), and the variance shrinkage method of Opgen-Rhein & Strimmer (2007). Then, the estimated forecast variance matrix is guaranteed to be positive definite with few numerical problems. This estimator is denoted as $\widehat{\boldsymbol{W}}_h^{shr} = (\hat{w}_{ij,h}^{shr})_{1 \le i,j \le m+p}$ with elements

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

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

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

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

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

$\hat{r}_{ij,h}$ the sample correlation of the h-step-ahead forecast error between the ith and the jth series (component) in $\boldsymbol{y}_t$, $\hat{w}_{i,h}$ the h-step-ahead sample base forecast variance associated with the ith series (the ith diagonal element of $\widehat{\boldsymbol{W}_h}$), and $\hat{w}_{h,median}$ 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 2023).

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-step forecast variance in forecast projection, if we only need point forecasts. We calculate $\widehat{\boldsymbol{W}}_h^{shr}$ for each h for the simulation example in Section 3, but assume this proportionality for the application in Section 4.

3 Simulation

3.1 Benchmarks

In this section, we illustrate the performance of the flap method 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 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 a 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; Hyndman & Athanasopoulos 2021). 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: $1 \le k \le 6$, $1 \le n \le 3$ and $1 \le s \le 3$. Note here DFM produces direct forecasts in the sense that a different model is fitted for each forecast horizon h, in contrast to indirect or iterative forecasts.

We use different weighting methods to construct the components. The types of components are listed below. For those randomly simulated, 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 ith component.

PCA+Norm The m principal components in PCA are taken first, implemented with the `prcomp` function in package `stats` (R Core Team 2023). Weights of the additional components are simulated from a standard normal distribution before being 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 being normalised to unit vectors.

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

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

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

We employ the Friedman test (Friedman 1937, 1939) along with post-hoc Nemenyi tests (Nemenyi 1963; Hollander, Wolfe & Chicken 2013) 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 flap 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 confidence 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 – DGP), the estimated VAR model with the correct specification (VAR – Est.), and their projections. We do not include those methods involving a uniform distribution and random orthonormal matrices, as they are visually identical to the methods with a normal distribution. The vertical black line indicates the number of series $m = 70$.

3.2 Projection over base forecast

The first thing we note is the overall performance difference between the projected forecasts and the base forecasts. 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 that 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 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 relatively large 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 will see in Section 4, especially with FRED-MD in Section 4.2.

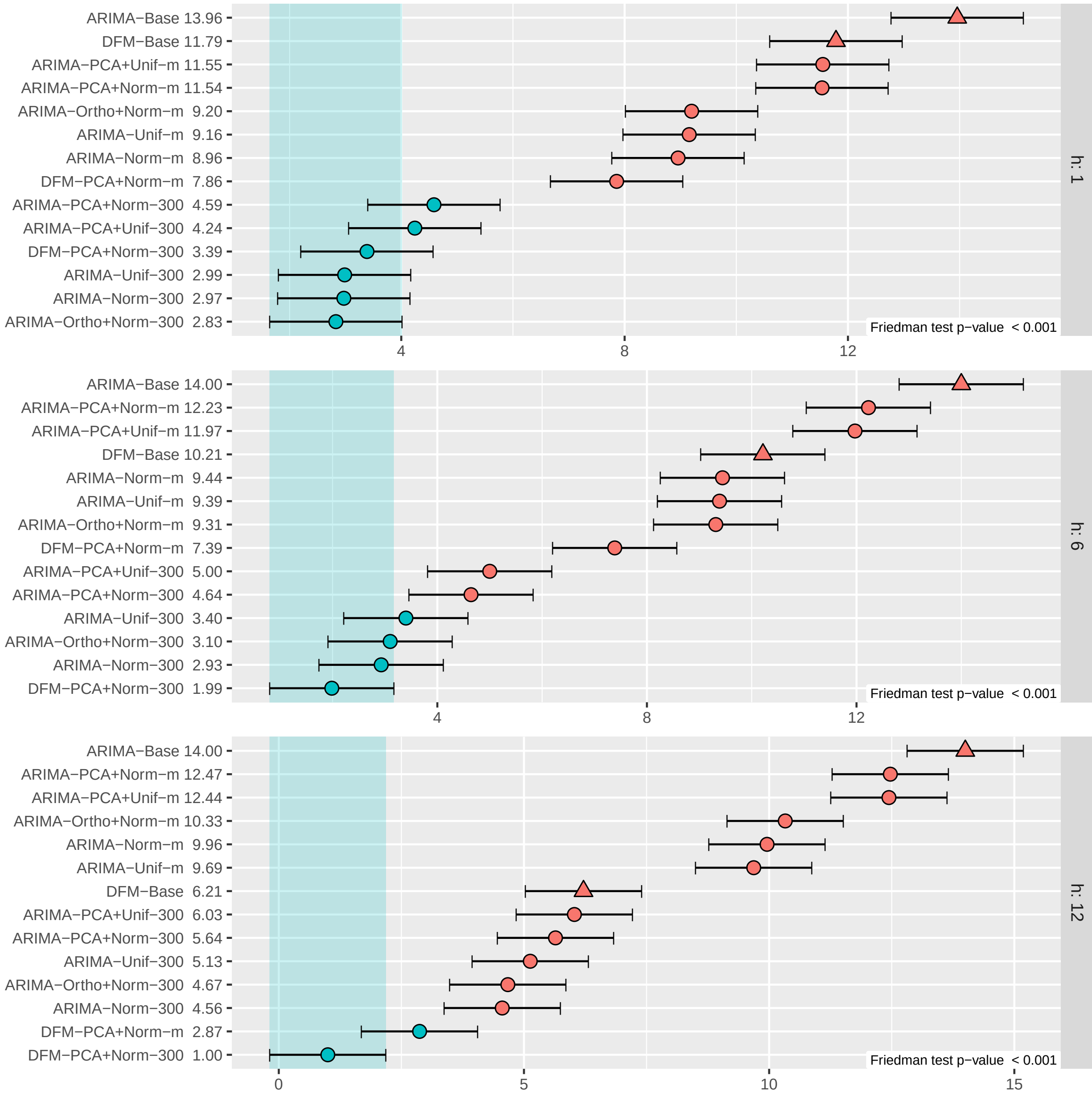

Figure 1: *Average ranks of 1-, 6- and 12-step-ahead MSE of different model and component specifications in the simulation. The methods using flap are named as "{Model}-{Comp. Weights}-{No. Comp.}". The base models are named as "{Model}-Base" and these points are marked with triangles. The shaded region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

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 but 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 obvious in Figure 2. The horizontal line representing the MSE of base DFM is always far below the horizontal line for base ARIMA.

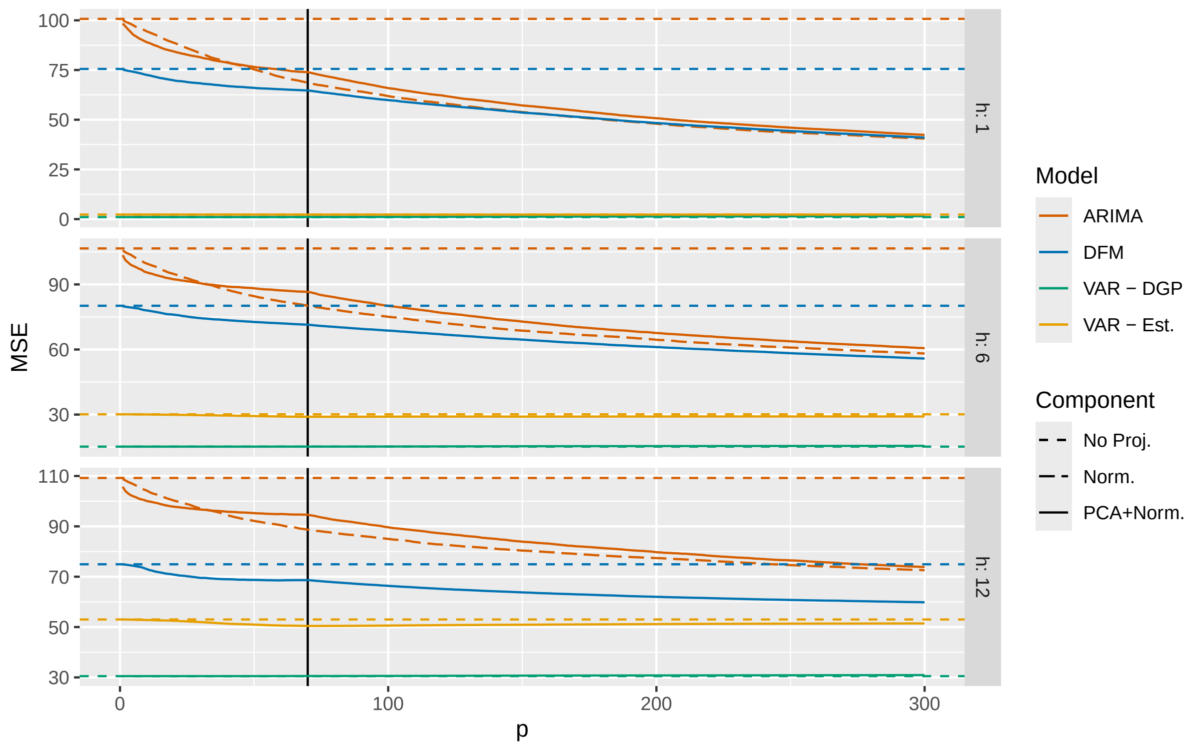

Figure 2: *MSE of different forecast models and component construction methods by the number of components p used in flap in the simulation, for forecast horizons* 1, 6 *and* 12*. "VAR – DGP" indicates the performance of the true data generating VAR model. "VAR – Est." indicates the performance of the 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.*

With the help of flap, 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 solid 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 cross-correlations 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 construction

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 is between using a combination of PCA and random weights, and purely using random weights. The distribution that generates the random weights has limited effect: 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 of limited importance, in Figure 2, we look only 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 will see in Section 4, this is not the case with real data, and PCA is the preferred approach to obtaining 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 $\boldsymbol{\Phi}$ and the base forecast variance $\boldsymbol{W}_h$ can 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 forecasts, 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, which is termed as the parameter uncertainty 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, referred to as the model uncertainty in Petropoulos, Hyndman & Bergmeir (2018). This is similar to bagging as bagging also reduces variance by controlling model uncertainty. The data uncertainty in Petropoulos, Hyndman & Bergmeir (2018) is not examined since the proposed method relies on one given data set, it is not clear how it translates to the forecast projection problem, but it relates to the uncertainty in how the components are constructed, as discussed in Section 3.4, and awaits future research.

4 Empirical applications

Here we apply the flap method to two real data examples and draw most of the same conclusions we did from the simulations, 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 each month away from home, which we will refer to as visitor nights in what follows. The monthly visitor nights are recorded 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 (Hyndman & Athanasopoulos 2021). 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 errors are 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, the trend term describes the direction of the long-term tendency, the seasonal term describes the periodically recurring pattern with a fixed periodicity, and the error term measures the uncertainty. The trend and seasonal terms are optional. If a trend term exists, we allow it to be additive or additive 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 possible models.

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

We highlight two differences. First, from Figure 4, the MSE of projection does not always go down as the number of components increases — see the $h = 1$ panel with $p > 150$. This can also be seen from Figure 3, where the two best methods are not significantly different, even though they have very different numbers of components ($p = 77$ and $p = 200$). 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 $p = m$.

Second, 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,

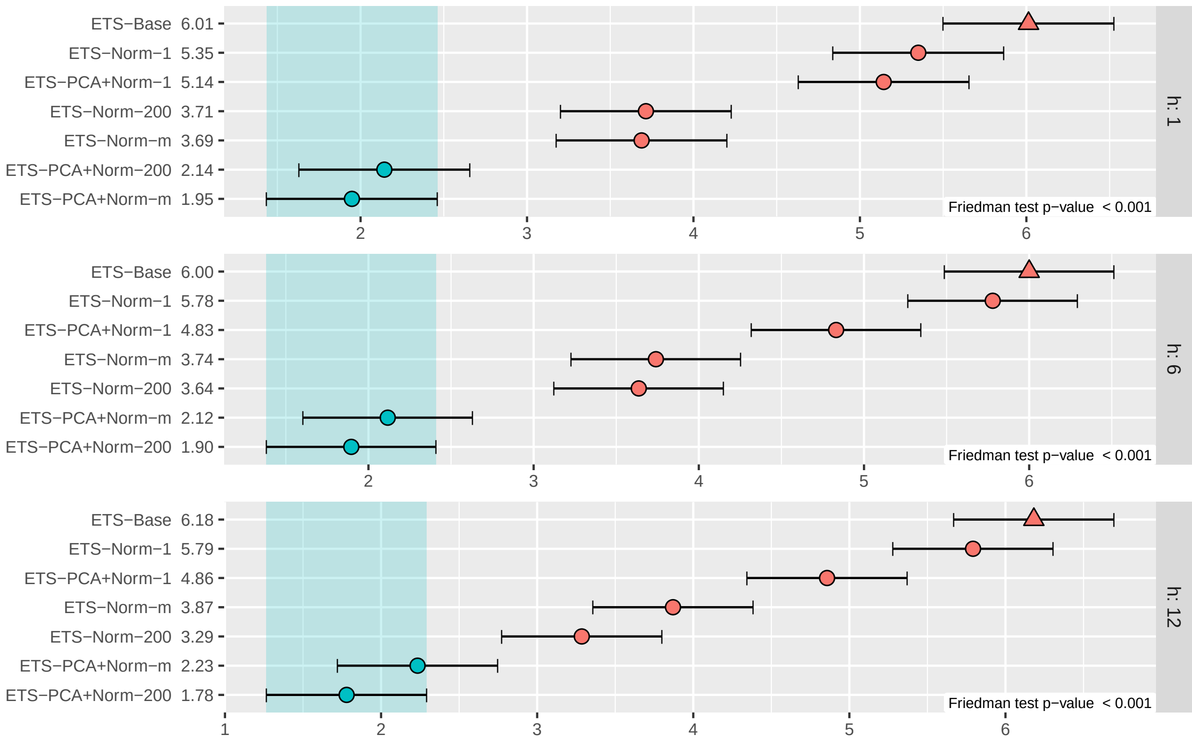

Figure 3: *Average ranks of* 1-, 6- *and* 12-*step-ahead cross-validation MSE of different model and component specifications on the visitor nights data. The methods using forecast projection are named as "{Model}-{Comp. Weights}-{No. Comp.}". The base models are named as "{Model}-Base" and these points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

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, and shares similar properties with the Stock & Watson (2002a) data. We downloaded 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

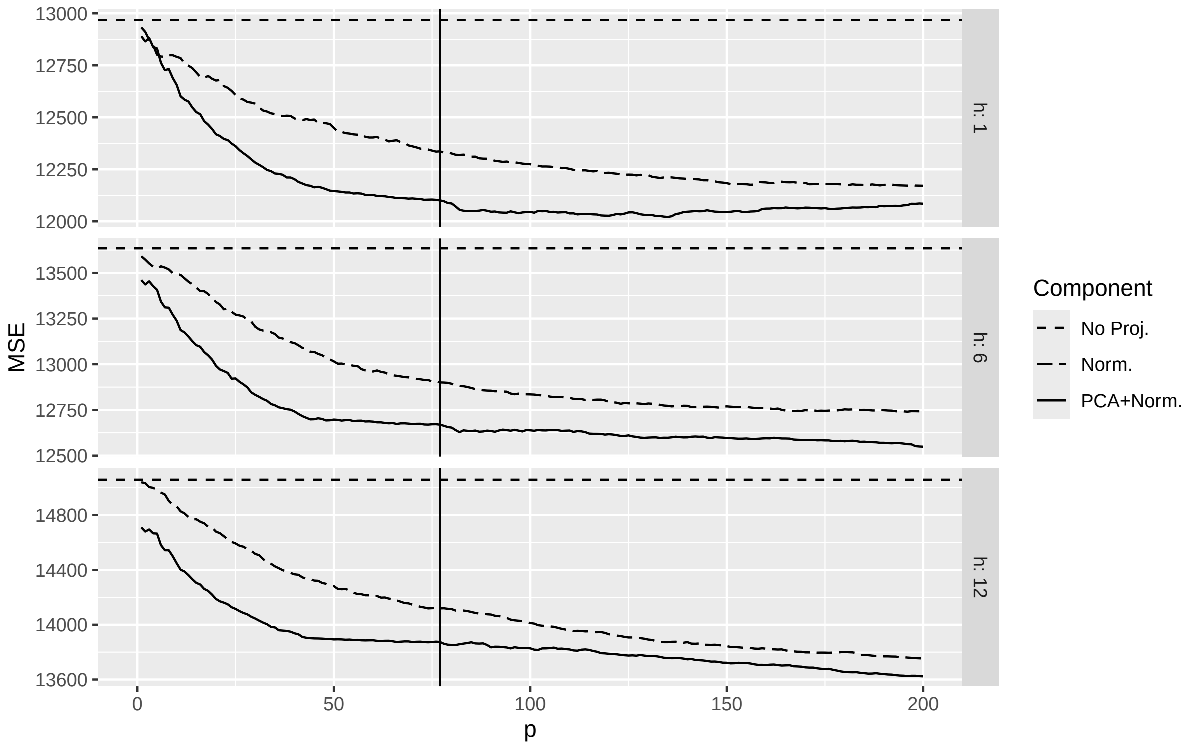

Figure 4: *MSE of different component construction methods by the number of components p used in forecast projection with ETS models on the visitor nights data, for forecast horizons* 1*,* 6 *and* 12*. The vertical black line indicates the location of $p = m$, the number of series.*

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 by 1 in each iteration. 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 are shown 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

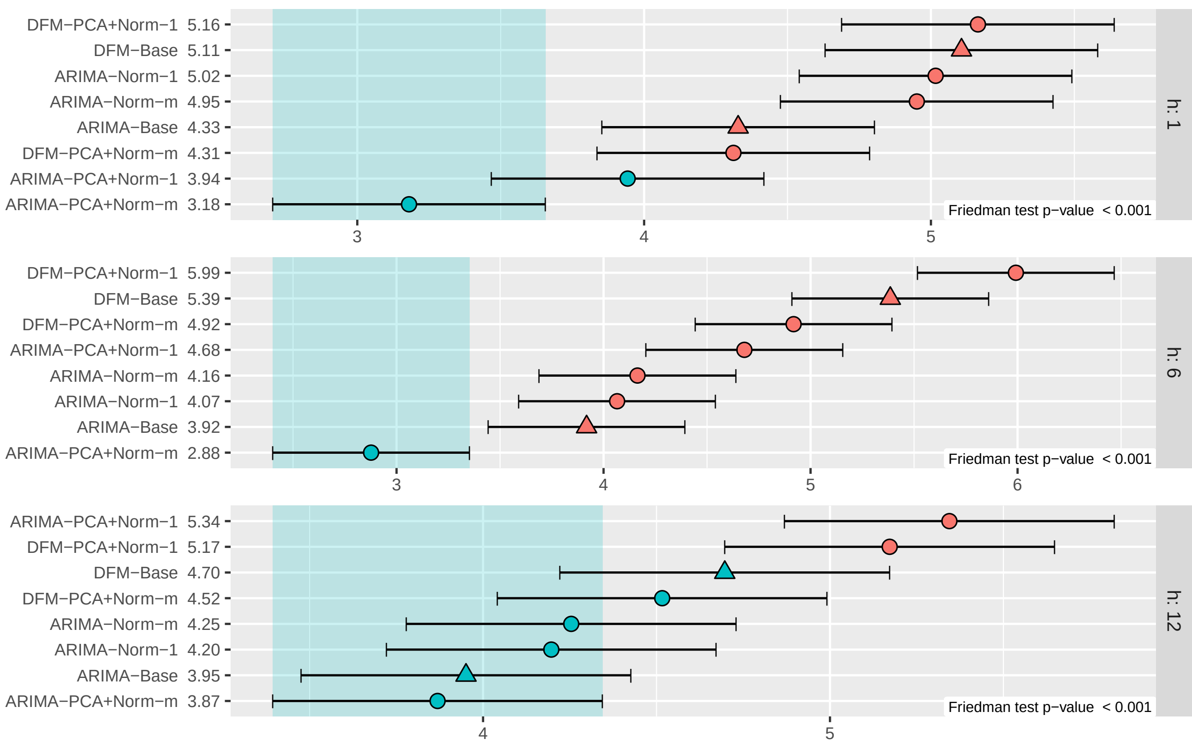

Figure 5: *Average ranks of* 1-, 6- *and* 12-*step-ahead cross-validation MSE of different model and component specifications on the FRED-MD data. The methods using forecast projection are named as "{Model}-{Comp. Weights}-{No. Comp.}". The base models are named as "{Model}-Base" and these points are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

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. The tradeoff between the benefit of new components and the difficulty of estimation of a large dimension is also seen here, but 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 turning point seems to be at m, but as we have seen in Section 3 and Section 4.1, m is not always a 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

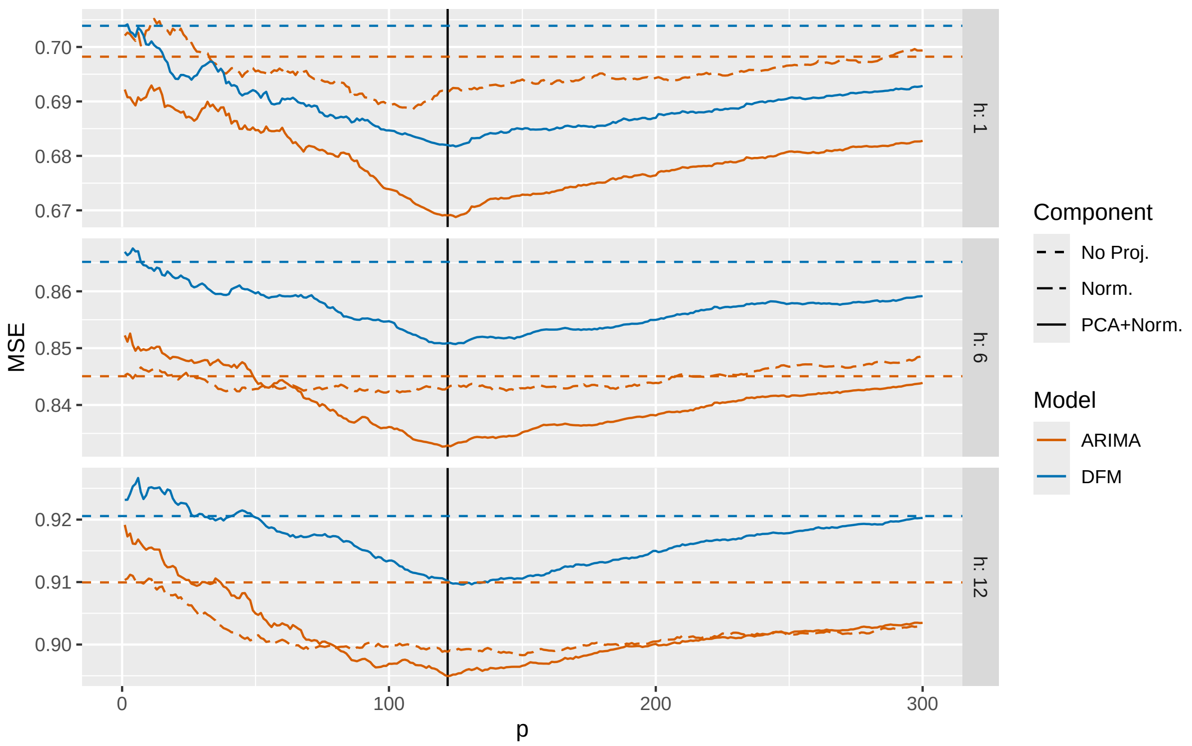

Figure 6: *MSE of different forecast models and component construction methods by the number of components p used in forecast projection on the FRED-MD data, for forecast horizons* 1, 6 *and* 12*. The vertical black line indicates the location of $p = m$ the number of series.*

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

The proposed forecast linear augmented projection (FLAP) method has been shown to be a simple but effective way to reduce forecast error variance of any multivariate forecasting problem. It simply involves augmenting the data with linear combinations, forecasting these and then projecting the augmented vector of forecasts. We have shown theoretically that FLAP will continue to reduce forecast error covariance as more components are added, assuming we that the forecast error covariance matrix is known. In practice, a plug in estimate of this covariance matrix can be used, and in both simulated and empirical data we demonstrate that a simple shrinkage estimator does indeed lead to improvements in forecast accuracy. Regarding the construction of components we find that PCA in practice achieves significant improvements in forecast accuracy. Another appealing property of FLAP is that the projection step can even compensate for a poor choice of base forecasting model. This is particularly attractive since it makes FLAP robust against model misspecification in the base forecasting step.

One outstanding issue is to find alternatives to PCA to select component weights. For example, Goerg (2013) proposed “forecastable components” that are optimal in the sense of minimising the forecast variance of the components, while Matteson & Tsay (2011) proposed “dynamic orthogonal components” that reduce a multivariate time series to a set of uncorrelated univariate time series. It would be interesting to explore whether these components (or other similar suggestions) can be used effectively in FLAP. Also, another route to improving FLAP my be found by optimizing Equation 10 over $\boldsymbol{\Phi}$ and $\boldsymbol{G}$ rather than just $\boldsymbol{G}$.

Finally, while FLAP is motivated by the forecast reconcililation literature, the focus is very much on multivariate time series with no constraints. However, it would be possible to use both forecast reconciliation and forecast projection together. This may be particularly useful when there are relationships between series that are not captured in the known hierarchical structure.

Acknowledgements

We thank Daniele Girolimetto for contributing to the initial proof of Theorem 2.2 in his unpublished work.

References

Ando, S & F Narita (2022). “An alternative proof of minimum trace reconciliation”. https://doi.org/10.2139/ssrn.4199143.

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.

Bernanke, BS, J Boivin & P Eliasz (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly journal of economics* **120**(1), 387–422.

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.

Carriero, A, AB Galvao & G Kapetanios (2019). A comprehensive evaluation of macroeconomic forecasting methods. *International Journal of Forecasting* **35**(4), 1226–1239.

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*.

Goerg, G (2013). Forecastable Component Analysis. In: *Proceedings of The 30th International Conference on Machine Learning*, pp.64–72. http://jmlr.org/proceedings/papers/v28/goerg13.pdf.

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.

Hollyman, R, F Petropoulos & ME Tipping (2021). Understanding forecast reconciliation. *European Journal of Operational Research* **294** (1), 149–160.

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 & G Athanasopoulos (2021). *Forecasting: principles and practice*. 3rd ed. Melbourne, Australia: OTexts. http://OTexts.org/fpp3.

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.

Matteson, DS & RS Tsay (2011). Dynamic Orthogonal Components for Multivariate Time Series. *Journal of the American Statistical Association* **106**(496), 1450–1463. http://dx.doi.org/10.1198/jasa.2011.tm10616.

McCracken, MW & S Ng (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics* **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 (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Rao, CR (1974). Projectors, generalized inverses and the Blue'S. *Journal of the Royal Statistical Society* **36** (3), 442–448.

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* **20** (2), 147–162.

Stock, JH & MW Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics* **30** (4), 481–493.

Tsay, RS (2013). *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons.

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.

Yang, YF (2024). *flap: Forecast Linear Augmented Projection*. R package version 0.1.0. https://CRAN.R-project.org/package=flap.

A Proofs for Section 2 (Forecast Linear Augmented Projection)

Proof of Lemma 2.1. We have

$$\begin{aligned}
\boldsymbol{M}\boldsymbol{M} &= \boldsymbol{I}_{m+p} - 2\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C} \\
&\quad + \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C} \\
&= \boldsymbol{I}_{m+p} - \boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C} \\
&= \boldsymbol{M},
\end{aligned}$$

so $\boldsymbol{M}$ is a projection matrix. For any $\boldsymbol{y}$ such that $\boldsymbol{M}\boldsymbol{y} = \boldsymbol{x}$ for some $\boldsymbol{x}$, we have

$$\boldsymbol{C}\boldsymbol{x} = \boldsymbol{C}\boldsymbol{M}\boldsymbol{y} = \boldsymbol{C}\boldsymbol{y} - \boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{y} = \boldsymbol{0}.$$

Thus, $\boldsymbol{M}$ projects any vector onto the space where the constraint $\boldsymbol{C}$ is satisfied.

□

Proof of Corollary 2.1. Items 1 and 2 are trivial application of Lemma 2.1. To prove 3, we have

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

□

Proof of Lemma 2.2.

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

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 projection. 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 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 $\tilde{\boldsymbol{y}}_{t+h}^+ = \boldsymbol{M}^+\hat{\boldsymbol{y}}_{t+h}^*$. Further, we obtain

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

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:

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

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

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

Taking the difference, we have

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

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,

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

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

Therefore,

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

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

$$
\begin{aligned}
\mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}-\boldsymbol{z}_{t+h})-\mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}^*-\boldsymbol{z}_{t+h})&=\mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}^+-\boldsymbol{z}_{t+h})-\mathrm{Var}(\tilde{\boldsymbol{z}}_{t+h}^*-\boldsymbol{z}_{t+h})\\
&=\boldsymbol{J}^*\,\mathrm{Var}(\tilde{\boldsymbol{y}}_{t+h}^+-\boldsymbol{y}_{t+h}^*)\boldsymbol{J}^{*\prime}-\boldsymbol{J}^*\,\mathrm{Var}(\tilde{\boldsymbol{y}}_{t+h}^*-\boldsymbol{y}_{t+h}^*)\boldsymbol{J}^{*\prime}\\
&=\boldsymbol{J}^*\boldsymbol{W}_h^*\boldsymbol{M}^{+\prime}\underline{\boldsymbol{C}}'(\underline{\boldsymbol{C}}\boldsymbol{M}^+\boldsymbol{W}_h^*\underline{\boldsymbol{C}}')^{-1}\underline{\boldsymbol{C}}\boldsymbol{M}^+\boldsymbol{W}_h^*\boldsymbol{J}^{*\prime}
\end{aligned}
\tag{15}
$$

being positive semi-definite. Finally, we have

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

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 $\boldsymbol{\psi}_i=\begin{bmatrix}-\boldsymbol{\phi}_i & \boldsymbol{0}_{1\times(i-1)} & 1\end{bmatrix}$ 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

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

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

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 $\boldsymbol{\phi}_1 \boldsymbol{W}_{z,h} \neq \boldsymbol{w}_{c_1 z,h}$.

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

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

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 $\boldsymbol{\psi}_i \boldsymbol{M}_i^+ \boldsymbol{W}_h^{(i)} \boldsymbol{J}'_{m,i} \neq \boldsymbol{0}$. 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

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

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 $\boldsymbol{w}_{\tilde{c}_i \tilde{z},h}^{(i-1)}$ 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 2.3. If $\boldsymbol{GS} = \boldsymbol{I}$, $\boldsymbol{SG}$ is a projection matrix: $\boldsymbol{SGSG} = \boldsymbol{SG}$.

For any $\boldsymbol{y}$ such that $\boldsymbol{SGy} = \boldsymbol{x}$ for some $\boldsymbol{x}$, we have $\boldsymbol{Cx} = \boldsymbol{CSGy} = \boldsymbol{0}$ because $\boldsymbol{CS} = \begin{bmatrix}-\boldsymbol{\Phi} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{\Phi}'\end{bmatrix}' = \boldsymbol{0}$. Similarly to $\boldsymbol{M}$, $\boldsymbol{SG}$ projects a vector to the same space where $\boldsymbol{C}$ is satisfied.

□

Proof of Corollary 2.2. Item 1 is an direct application of Lemma 2.3. From Lemma 2.3 and Lemma 2.4 in Rao (1974), we have

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

Left multiplying by $\boldsymbol{G}$ on both sides, we have $\boldsymbol{Gy}_{t+h} = \boldsymbol{z}_{t+h}$ and item 2 is proven. To prove Item 3, we have

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

□

Proof of Lemma 2.4. Let the base and projected forecast errors be given as

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

where the last line comes from $\boldsymbol{GS} = \boldsymbol{I}$. Left multiplying by $\boldsymbol{G}$ on 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

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

□

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

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

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 $\boldsymbol{H}$:

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

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 $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$. Putting it back in, we have

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

□