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

Abstract

We propose a novel forecast linear augmented projection (FLAP) method that can reduce the forecast error variance of any multivariate forecast. The method first constructs new component series which are linear combinations of the original series. Forecasts are then generated for both the original and component series. Finally, the full vector of forecasts is projected onto a linear subspace where the constraints implied by the combination weights hold. We show that the projection using the original forecast error covariance matrix will result in improved forecasts. Notably, the new forecast error variance of each series 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 error variance reduction among linear projection methods. We demonstrate our proposed method with an estimated covariance matrix using simulations and two empirical applications based on Australian tourism and FRED-MD data. In all cases, forecasts are improved. Notably, using FLAP with Principal Component Analysis (PCA) to construct the new series leads to substantial forecast error variance reduction.

Keywords: Forecast combinations; High-dimensional time series; Components; Forecast reconciliation.

1 Introduction

We will show that any multivariate forecasting method can be improved through a post-processing framework involving linear combinations of the original series. We call this FLAP: Forecast Linear Augmented Projection. This has immediate implications for multivariate forecasting in all disciplines including macroeconomics (Carriero et al. 2019) and finance (Tsay 2013).

Our FLAP post-processing framework: (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 prove that under projections based on the original forecast error covariance matrix, the method reduces the forecast error 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 y_1 and y_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 $\hat{y}_1$, $\hat{y}_2$, $\hat{c}_1$, $\hat{c}_2$ 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 y_1, there is both a *direct* forecast $\hat{y}_1$ and *indirect* forecast $\hat{c}_1 + \hat{c}_2$, similarly for y_2 the direct forecast is $\hat{y}_2$ 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.

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

More formally and more generally, the FLAP method considers a vector of original series $\boldsymbol{y} \in \mathbb{R}^m$ and a vector of components $\boldsymbol{c} \in \mathbb{R}^p$. While $(\boldsymbol{y}', \boldsymbol{c}')'$ is a $p+m$-vector, the construction of components as linear combinations of the original series implies that $(\boldsymbol{y}', \boldsymbol{c}')'$ lies on a linear subspace of at most dimension m. In general, the collection of all the corresponding vector of forecasts $(\hat{\boldsymbol{y}}', \hat{\boldsymbol{c}}')'$ spans the linear space $\mathbb{R}^{m+p}$. FLAP projects $(\hat{\boldsymbol{y}}', \hat{\boldsymbol{c}}')'$ onto the m-dimensional linear subspace in which $(\boldsymbol{y}', \boldsymbol{c}')'$ lives. 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 et al. (2019), Athanasopoulos et al. (2017), and Di Fonzo and Girolimetto (2023) have all shown that reconciliation can reduce forecast error 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 $\boldsymbol{y}$ and $\boldsymbol{c}$ taken together. This poses a problem if improvements in forecast accuracy for $\boldsymbol{c}$ could be offset by a deterioration in forecast accuracy for $\boldsymbol{y}$, since the former are not of interest in and of themselves. A key insight we make in this paper, is that reductions in forecast error variance accrue even for single series in the bottom level of the hierarchy. It is this contribution 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 (PCs). When doing so, the FLAP method bears a resemblance to Dynamic Factor Models (DFMs), specifically those common in macroeconomic forecasting (Stock and Watson 2002a; b; 2012), their extensions in the machine learning literature (De Stefani et al. 2019), as well as the factor augmented VAR (Bernanke et al. 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" (Bergmeir et al. 2016; Breiman 1996). Bagging can reduce prediction variance without increasing bias (Hastie et al. 2003), by mitigating model uncertainty (Petropoulos et al. 2018), and does so without introducing any new data, but rather resampled versions of the existing data. Our FLAP method also reduces forecast error variance without introducing new data, but using linear combinations of the existing data, rather than bootstrapping. The FLAP method improves forecast accuracy through forecast combination and data augmentation. Other methods that also do this include the theta method (Assimakopoulos and Nikolopoulos 2000), temporal aggregation (Athanasopoulos et al. 2017; Kourentzes et al. 2014), forecasting with sub-seasonal series (FOSS, Li et al. 2022) and forecast combination with multiple starting points (Disney and Petropoulos 2015); a review of all these methods can be found in Petropoulos and 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 for R (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 (FLAP)

2.1 Method and theoretical properties

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 $\boldsymbol{y}_t \in \boldsymbol{R}^m$ be a vector of m observed time series we are interested in forecasting. The FLAP method involves three steps:

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

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

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 $\text{Var}(\boldsymbol{z}_{t+h} - \hat{\boldsymbol{z}}_{t+h}) = \boldsymbol{W}_h$ is the forecast error 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 interested in forecasts of $\boldsymbol{y}_t$ and not the full vector $\boldsymbol{z}_t$. We now introduce some 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 $\boldsymbol{J}_{n,k}\boldsymbol{A}$ selects the first n rows of a matrix $\boldsymbol{A}$. Let $\hat{\boldsymbol{y}}_{t+h}$ and $\tilde{\boldsymbol{y}}_{t+h}$ denote the first m elements of $\hat{\boldsymbol{z}}_{t+h}$ and $\tilde{\boldsymbol{z}}_{t+h}$, comprising the base and projected forecasts of $\boldsymbol{y}_t$ respectively. Similarly, let $\hat{\boldsymbol{c}}_{t+h}$ and $\tilde{\boldsymbol{c}}_{t+h}$ denote the last p elements of $\hat{\boldsymbol{z}}_{t+h}$ and $\tilde{\boldsymbol{z}}_{t+h}$, comprising the base and projected forecasts of $\boldsymbol{c}_t$ respectively. Then the projected forecast of $\boldsymbol{y}_t$ can be found by

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

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

We now present some theoretical results regarding the FLAP method, with proofs provided in the Appendix. Theorem 2.1 establishes that the forecasts produced by the FLAP method, $\tilde{\boldsymbol{y}}_{t+h}$, dominate the base forecasts, $\hat{\boldsymbol{y}}_{t+h}$, in the sense that the difference between their forecast error variances is always positive semi-definite. Theorem 2.2 establishes that each of the forecast error variances of $\tilde{\boldsymbol{y}}_{t+h}$ is non-increasing with 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 achieves the minimum forecast error variance 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 covariance matrix $\boldsymbol{W}_h$ unreliable. We explore this issue in a simulation setting and empirically in Section 3 and Section 4, respectively.

2.2 Positive semi-definiteness of error variance reduction

We first provide some intermediate results.

Lemma 2.1. *Matrix $\boldsymbol{M}$ is a projection onto the space where the constraint $\boldsymbol{C}\boldsymbol{z}_t = \boldsymbol{0}$ is satisfied.*

Proof in Appendix A.

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

Corollary 2.1.

1. *The projected forecast $\tilde{\boldsymbol{z}}_{t+h}$ satisfies the constraint $\boldsymbol{C}\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{0}$.*
2. *For $\boldsymbol{z}_{t+h}$ that already satisfies the constraint, the projection does not change its value, i.e., $\boldsymbol{M}\boldsymbol{z}_{t+h} = \boldsymbol{z}_{t+h}$ (Rao 1974, Lemma 2.4).*
3. *If the base forecasts are unbiased such that $\mathrm{E}(\hat{\boldsymbol{z}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{z}_{t+h}|\mathscr{I}_t)$, then the projected forecasts are also unbiased, i.e., $\mathrm{E}(\tilde{\boldsymbol{z}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{z}_{t+h}|\mathscr{I}_t)$.*

Proof in Appendix A.

The assumption of unbiasedness of the base forecasts is not unreasonable in practice, and where it does not hold, a bias correction method can be applied. 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 and Watson 2002b). In fact, the power of FLAP manifests when the models are misspecified, as discussed in Section 3.

Lemma 2.2. *The forecast error covariance matrix of the component-augmented projected h-step-ahead forecasts* $\tilde{\boldsymbol{z}}_{t+h}$ *is*

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

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

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

Proof in Appendix A.

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

Theorem 2.1 (Positive Semi-Definiteness of Error Variance Reduction)**.** *The difference between the forecast error variances of the base and projected component-augmented forecasts,*

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

is positive semi-definite. The difference between the forecast error variances of the base and projected forecasts of the original series,

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

is therefore also positive semi-definite.

Proof in Appendix A.

Theorem 2.1 is why FLAP works. Note when $\boldsymbol{J}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{J}'$ is semi-definite, each of its diagonal elements, the reduction in forecast error variance for each series, is non-negative. It implies that the forecast error variance can be reduced 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 must be zero. This implies that $\boldsymbol{J}\boldsymbol{W}_h\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{C}')^{-1}\boldsymbol{C}\boldsymbol{W}_h\boldsymbol{J}' = \boldsymbol{O}_{m\times m}$, as this is a positive semi-definite matrix, something rarely observed in practice. See Theorem 2.3 for more discussion.

The following example illustrates the mechanism of the reduction in the forecast error variance.

Example 2.1. Suppose $\boldsymbol{z}_t$ comprises m original series $\boldsymbol{y}_t$ and p components $\boldsymbol{c}_t$. Let $\hat{\boldsymbol{z}}_{t+h}$ and $\tilde{\boldsymbol{z}}_{t+h}$ be h-step-ahead base and projected forecasts of $\boldsymbol{z}_t$. Assume that their corresponding forecast errors are uncorrelated with unit variance such that $\boldsymbol{W}_h = \boldsymbol{I}_{m+p}$. Then

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

where

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

Let $\boldsymbol{\Phi}$ consist of $p \leq m$ orthogonal unit vectors, for example, those obtained from Principal Component Analysis (PCA, Jolliffe 2002). In this case $\boldsymbol{\Phi}\boldsymbol{\Phi}' = \boldsymbol{I}_p$ and

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

Focusing on the forecast error variance reduction of the forecasts of the original series $\boldsymbol{y}_t$, i.e., $\text{tr}(\text{Var}(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h}) - \text{Var}(\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h})) = \frac{1}{2}\,\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi})$.

- When $p < m$, since $\boldsymbol{\Phi}'\boldsymbol{\Phi}$ is idempotent, $\text{tr}(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = \text{rank}(\boldsymbol{\Phi}'\boldsymbol{\Phi}) = p$. Hence, focusing on the original series the reduction in the total forecast error variance of the FLAP forecasts relative to the base forecasts, is $p/2$.
- When $p = m$, where all principal components are used, $\boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{I}_m$. This implies a total reduction in the error variance of $m/2$, and that for each of the m individual series the error variance is halved.

If we keep increasing the number of components beyond m, 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 as the forecast error variance $\boldsymbol{W}_h$ can hardly be an identity matrix in practice. It is likely that the forecast error of a linear combination of series will be correlated to the forecast error of forecasting these series directly. Nonetheless, the aim of the example is to demonstrate how the forecast error variance can be reduced as a result of the component forecasts bringing new information about the original series. The forecast error variance reduction becomes larger as we increase the number of components p. This is not a coincidence but a desirable property of FLAP, as shown in the next section.

2.3 Monotonicity

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

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

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

Theorem 2.2 (Monotonicity)**.** *The forecast error variance reductions for each series, i.e., the diagonal elements in the matrix*

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

are non-decreasing as p increases. In particular, the sum of forecast error variance reductions

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

is non-decreasing as p increases.

Proof in Appendix A.

Theorem 2.2 is the key result that demonstrates the usefulness of FLAP. It means that we can keep increasing the number of components to reduce forecast error variance, even when the number of components exceeds the number of original series. 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. 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. This of course assumes that the forecast error covariances are known, something we explore in Section 2.5 and Section 3.

Extending the proof of Theorem 2.2, we can state the condition required for the reduction sum of forecast error variances to be strictly 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 $\boldsymbol{W}_{\tilde{y},h}^{(i-1)}$ denote the error covariance matrix of the projected forecasts of the original series based on the first $i-1$ components, $\boldsymbol{w}_{c_1\hat{y},h}$ denote a vector of covariances of the first component and the base forecasts of the original series, and $\boldsymbol{w}_{c_i\tilde{y},h}^{(i-1)}$ 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 Forecast Error Variance Reduction Condition)**.** *For the first component to achieve a guaranteed reduction of forecast error variance (for the matrix in Theorem 2.1 to have strictly positive trace),*

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

For the ith component to have a positive reduction on forecast error variance of the original series,

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

Proof in Appendix A.

The condition of Equation 5 demonstrates that for a new component to be beneficial, the information introduced by this new component, reflected in the error 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. Heuristically, we want the components to be introducing new information, and this is more likely when they are uncorrelated, which motivates out choice of principal components which are uncorrelated by constructions. Optimal choices of weights we leave to future research, but note that practically speaking, 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 negligible 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,

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

If we consider the transformed space where all the vectors are first transformed via pre-multiplying by $\boldsymbol{W}_h^{-1/2}$, where $\boldsymbol{W}_h^{-1}=(\boldsymbol{W}_h^{-1/2})'\boldsymbol{W}_h^{-1/2}$, then this optimisation problem can be interpreted as finding the set of forecasts that are closest to the base forecasts on the transformed space, while satisfying the linear constraints imposed by the components.

An alternative way of characterising the same constraints is

$$\boldsymbol{\Phi}\check{\boldsymbol{y}}_{T+h}=\check{\boldsymbol{c}}_{T+h},$$

where $\check{\boldsymbol{c}}_{T+h}$ is the vector of the last p elements of $\check{\boldsymbol{z}}_{T+h}$, corresponding to the forecast of the components as part of the solution. This equivalence is discussed in Wickramasuriya et al. (2019), where the authors find the solution by minimising the sum of forecast error variance of all series (See Ando and Narita (2024) for an alternative proof). The result is

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

where $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{I}_m \\ \boldsymbol{\Phi} \end{bmatrix}$ contains the constraints, so that $\boldsymbol{z}_t = \boldsymbol{S}\boldsymbol{y}_t$, and

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

In Equation 7, $\boldsymbol{G}\hat{\boldsymbol{z}}_{t+h}$ can be viewed as mapping all of the series to a selected few. In the forecast reconciliation context, this is a mapping of all series to the "bottom level". In our multivariate forecasting context, this is a mapping of all series including the components, to the space of the original series. This leads to the solution

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

as equivalent to Equation 3. Recognising that Equation 7 is equivalent to Equation 1, it is the solution that minimises the sum of forecast error 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 error variance of the original series, and their sum. This can be viewed as a special case of Theorem 3.3 in Panagiotelis et al. (2021) applied in a forecast projection context, by taking $\boldsymbol{W} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$ in Panagiotelis et al. (2021, Thm. 3.3), as illustrated by Ando and Narita (2024). 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 $\boldsymbol{S}\boldsymbol{G}$ is a projection onto the space where the constraint $\boldsymbol{C}\boldsymbol{z}_t = \boldsymbol{0}$ is satisfied, provided that $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$.*

Proof in Appendix A.

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

1. *The projected forecast in Equation 9 satisfies the constraint $\boldsymbol{C}\tilde{\boldsymbol{z}}_{t+h} = \boldsymbol{C}\boldsymbol{S}\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{0}$.*
2. *For $\boldsymbol{z}_{t+h}$ that already satisfies the constraint, the mapping does not change its value, i.e., $\boldsymbol{G}\boldsymbol{z}_{t+h} = \boldsymbol{y}_{t+h}$.*

3. *If the base forecasts are unbiased such that* $\mathrm{E}(\hat{\boldsymbol{z}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{z}_{t+h}|\mathscr{I}_t)$*, then the projected forecasts in Equation 9 are also unbiased:* $\mathrm{E}(\tilde{\boldsymbol{y}}_{t+h}|\mathscr{I}_t) = \mathrm{E}(\boldsymbol{y}_{t+h}|\mathscr{I}_t)$*.*

Proof in Appendix A.

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

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

Proof in Appendix A.

We are now ready to present the following theorem.

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

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

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.*

Proof in Appendix A.

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

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

albeit unbiased, is not considered a good approximation to the true forecast error variance in a finite sample when $(m+p) \approx T-h$. It is even singular when $(m+p) > T-h$, which makes the quantities discussed in the previous sections impossible to calculate. For this reason, while other shrinkage estimators can be considered, we adopt the covariance shrinkage method of Schäfer and Strimmer (2005) and the variance shrinkage method of Opgen-Rhein and Strimmer (2007). Then, the estimated forecast error variance matrix is guaranteed to be positive definite with no numerical problems when computing their inverse. This estimator is denoted as $\widehat{\boldsymbol{W}}_h^{shr} = (\hat{w}_{ij,h}^{shr})_{1\le i,j\le m+p}$ with the element in row i and column j

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

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

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

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 $\boldsymbol{z}_t$, $\hat{w}_{i,h}$ the h-step-ahead sample base forecast error 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 error 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 error variance is proportional to the 1-step forecast error variance by a constant η_h, as do Wickramasuriya et al. (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 error 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

In this section, we illustrate the performance of FLAP in a simulation setting. We generate time series of length $T = 400$ from a $m = 70$ variable VAR(3) data generating process (DGP). The coefficients for the VAR DGP are estimated from the first 70 series in the Australian tourism data set used in Section 4.1. The innovations are simulated from a multivariate normal distribution with an identity covariance matrix. The estimation and simulation is performed using the `tsDyn` package (Fabio Di Narzo et al. 2009).

For each simulated sample we generate $h = 1$ to 12-step-ahead base forecasts from benchmark models. We implement FLAP with a varying number of components and component construction methods. This process is repeated 220 times. The improvements of FLAP forecasts over base forecasts is assessed, and the statistical significance of these improvements is evaluated.

3.1 Generating base and FLAP forecasts

We generate base forecasts from two benchmarks. The first benchmark is the univariate ARIMA model. For each series, we generate base forecasts from an ARIMA model using the `auto.arima()` function from the `forecast` package (Hyndman et al. 2023) with the default settings.

The second benchmark is the dynamic factor model (DFM). Following Stock and Watson (2002a), base forecasts

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

where $\hat{\boldsymbol{F}}_t$ is the vector of k estimated factors, and $\hat{y}_t$ is the target series to forecast. The factors are estimated using PCA on demeaned and scaled data. The optimal model is selected for each series based on the Bayesian information criterion (BIC) from models fitted using different combinations of meta-parameters in their corresponding range: $1 \leq k \leq 6$, $1 \leq n \leq 3$ and $1 \leq s \leq 3$. We note that the 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 generated by the ARIMA models.

We use several sets of weights to construct linear components for the FLAP method. The types of components are listed in Table 1. Principal components from PCA are established using the `prcomp()` function in the `stats` package (R Core Team 2023). We generate random components using weights simulated from a standard normal (Norm) distribution and a uniform (Unif) distribution with range $(-1, 1)$. We normalise the weights of all randomly generated components into unit vectors to ensure numerically stable computation, 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. The total number of components p is selected to be either $m = 70$ or 300.

Table 1: *Component construction methods for FLAP*

Component	Description
PCA	Principal components from PCA.
Norm	The weights of all components are simulated from a standard normal distribution.
Unif	The weights of all components are simulated from a uniform distribution
PCA+Norm	m principal components from PCA are complemented with random components whose weights are simulated from a standard normal distribution.
PCA+Unif	m principal components from PCA are complemented with random components whose weights are simulated from a uniform distribution.
Ortho	A random orthonormal matrix is generated using package `pracma` (Borchers 2023) as the weight matrix.

Component	Description
Ortho+Norm	A random orthonormal matrix is generated using package `pracma`, forming the weights of the first m components. Weights of the additional components are simulated from a standard normal distribution.

3.2 Forecast evaluation

We employ the Friedman test (Friedman 1937, 1939) along with post-hoc Nemenyi tests (Hollander et al. 2013; Nemenyi 1963) using the `tsutils` package (Kourentzes 2023) 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 MSE of each series over different samples is calculated, and the MSEs of all the series are treated as observations in the Nemenyi test. The average ranks are plotted in Figure 1 for forecast horizons 1, 6 and 12. The methods using FLAP are labelled "Model – Component Weights – Number of Components". The benchmarks are labelled "Model – Benchmark". These points are marked with triangles. The shaded region is the confidence interval of the best-performing method. Methods outside the shaded region are significantly worse than the best model.

In Figure 2, we plot the out-of-sample MSE values as a function of the number of components p included in the FLAP method. Here we also include the performance of the DGP VAR model (VAR – DGP), and an estimated VAR model with the correct specification (VAR – Est.), as well as their corresponding FLAP generated forecasts. We have not presented the FLAP forecasts using components generated from a uniform distribution or random orthonormal matrices, as these are visually identical to the methods with random weights generated from a standard normal distribution. The vertical black line indicates when $p = m = 70$, the number of original series.

Evaluating base forecasts

Having simulated data from a VAR DGP, we expect the DFM to capture the correlations between series, something not possible with univariate ARIMA models. Hence, we expect the DFM benchmark to perform better than the ARIMA benchmark. This is indeed the case. Figure 1 shows that

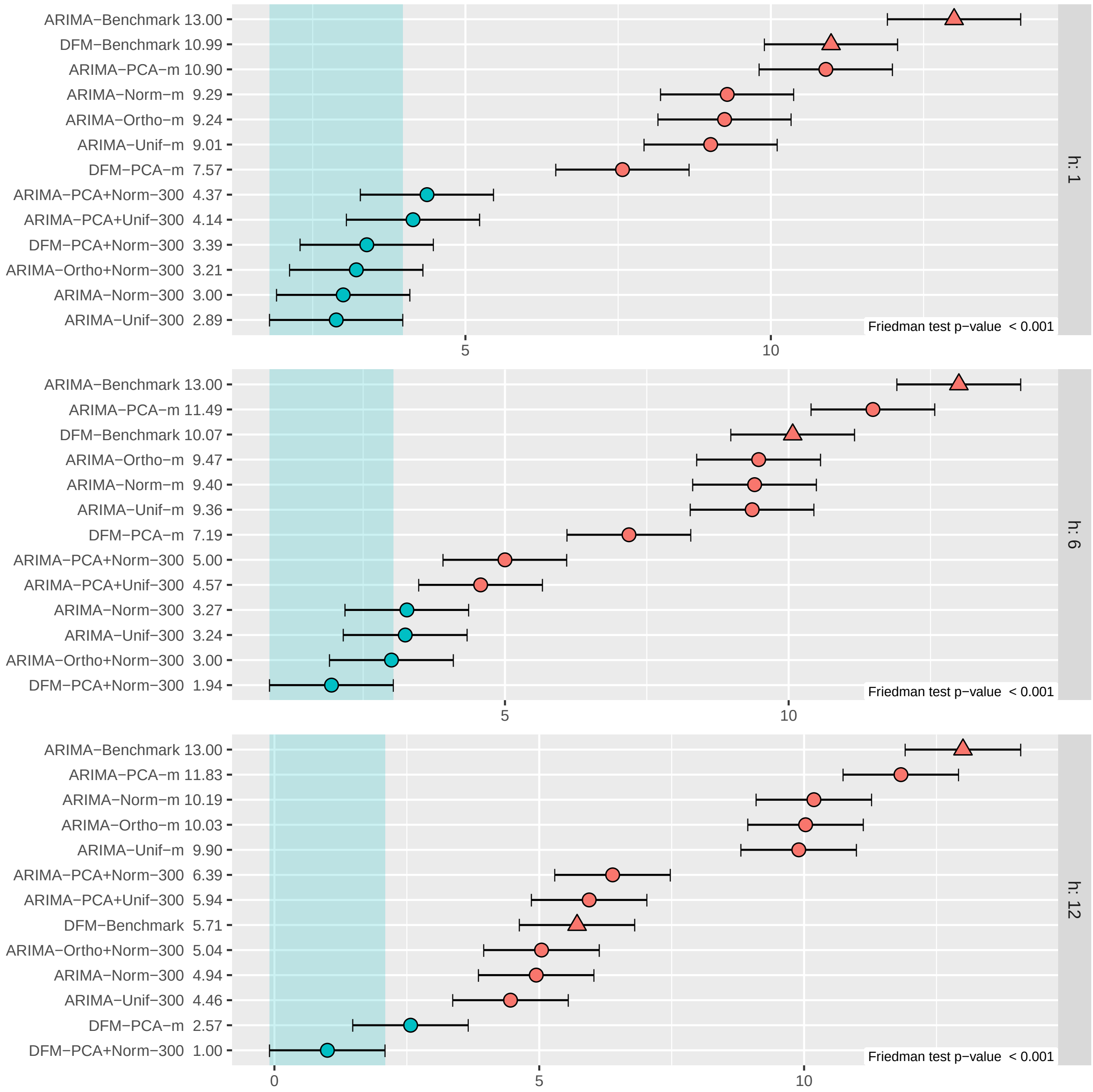

Figure 1: *Average ranks of* 1-, 6- *and* 12-*step-ahead MSE of different model and component specifications in the simulation. The methods using FLAP are labelled as "Model – Component Weights – Number of Components". The two benchmark models generating base forecasts are labelled "Model – Benchmark". Their MSEs are marked with triangles. The shaded region is the confidence interval of the best performing method. Methods outside the shaded region are significantly worse than the best method.*

the base DFM forecasts are significantly better than base ARIMA forecasts, with the exception for $h = 1$, where the difference in rank is statistically insignificant but only marginally.

This is also observed in Figure 2. The solid horizontal line representing the MSE of the base DFM forecasts is always much lower than the horizontal solid line of the base ARIMA forecasts.

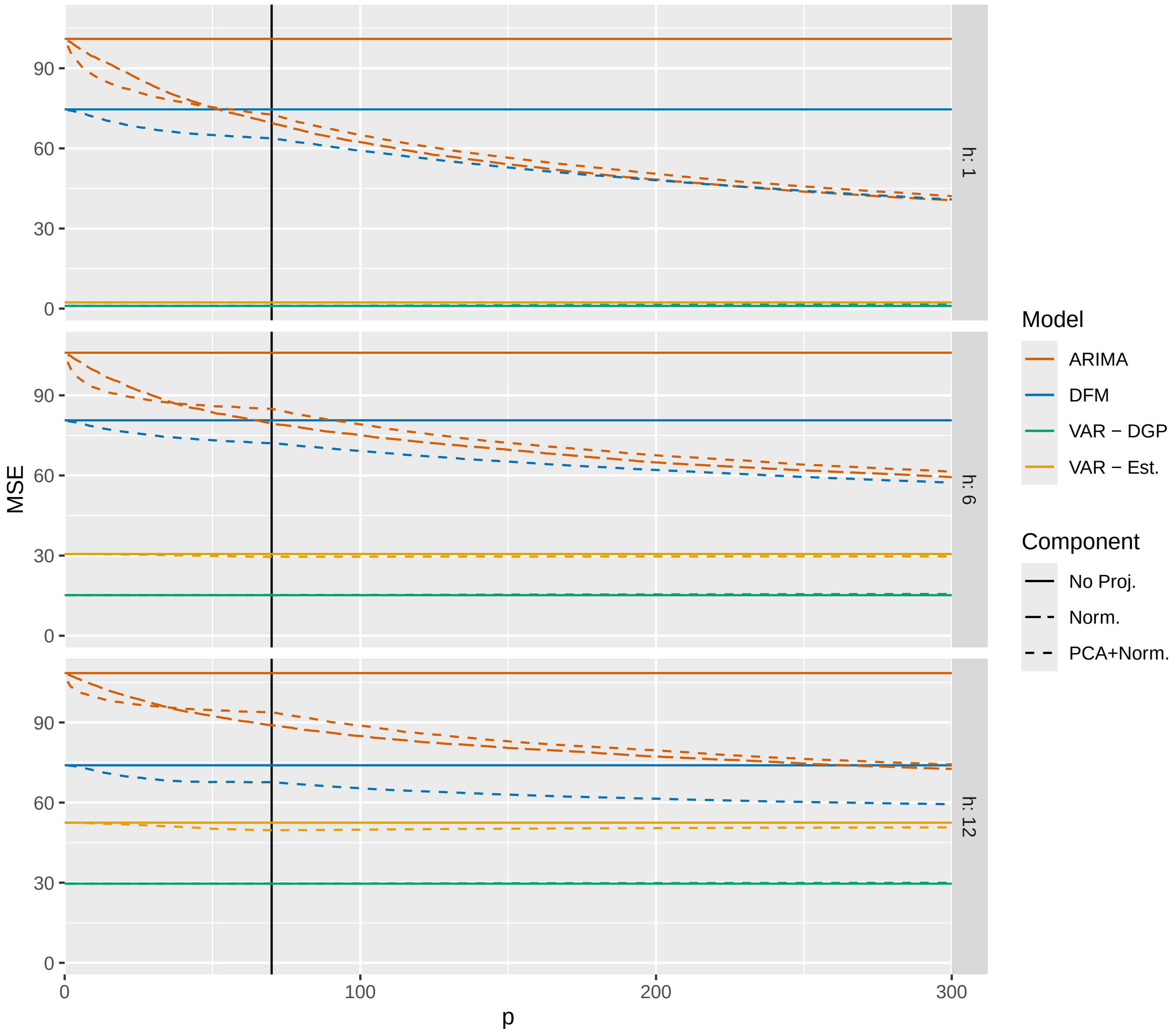

Figure 2: *Out-of-sample MSE for base and FLAP forecasts as the number of components p increases, for forecast horizons 1, 6 and 12. "VAR – DGP" indicates the performance of the true data generating VAR model. "VAR – Est." indicates the performance of the VAR model with the same structure as the true model but with estimated parameters. The solid horizontal lines show the MSE for the base forecasts (with no projection) while the dashed lines show the MSEs of the FLAP forecasts. The vertical black line indicates the location of $p = m = 70$, the number of series.*

Evaluating FLAP forecasts

The most important observation from Figure 1 are the statistically significant improvements of the FLAP forecasts over their corresponding benchmark base forecasts. Note that we are referring here to forecasts corresponding to the same model, i.e., FLAP ARIMA forecasts, compared to base ARIMA forecasts and FLAP DFM forecasts, compared to base DFM forecasts. The average ranks of the FLAP forecasts are better than the corresponding base forecasts for all forecast horizons, and the differences are all statistically significant. The only exception is for ARIMA base forecasts with

FLAP using only the $p = m = 70$ principal components, where the difference in rank is statistically insignificant.

Hence, the number of components seems to be important. The best-performing models all comprise the maximum $p = 300$ components. For $h = 1$-step-ahead forecasts, the differences between the methods with $p = 300$ components are not significantly significant, regardless of the model that generated the base forecasts or the method used to construct the components.

Indeed, Figure 2 shows that the MSEs for the FLAP ARIMA and DFM forecasts, represented by the dashed lines, decrease monotonically relative to the base forecasts, as the number of components increases. This supports Theorem 2.2, demonstrating that increasing the number of components in FLAP reduces the forecast error variance. Of course this is only feasible in this ideal setting where each time series has 400 observations and we use no more than 300 components in FLAP. The relatively large number of observations coupled with the relatively simple VAR DGP, has potentially eased the challenge of estimation. This continuous reduction in forecast error variance is not always achievable with real data, as demonstrated in the empirical applications in Section 4, especially with FRED-MD dataset in Section 4.2.

Worth noting from Figure 2 is the performance of the FLAP ARIMA forecasts relative to DFM forecasts as the number of components p increases. The MSE of the FLAP ARIMA forecasts becomes lower than the DFM base forecasts for $p < 70$ for $h = 1$, for $70 < p < 100$ for $h = 6$ and for a larger p for $h = 12$. Hence, FLAP is able to enhance univariate ARIMA forecasts with shared information between series by capturing and projecting these cross-correlations with the components.

Interestingly, for $h = 1$, the MSEs of FLAP ARIMA and FLAP DFM forecasts converge to the same value as p reaches 300, no matter how the components are constructed. For the longer forecast horizons, the MSEs again converge; they do not reach the same value for $p = 300$, although they are closer for $h = 6$ than $h = 12$. The reason for this is not the contribution of FLAP, but the starting point of the two sets of base forecasts. As the forecast horizon increases from $h = 1$ to 6 and 12 the ARIMA base forecasts MSE increases. In contrast the MSEs for DFM base forecasts remains at a relatively constant level. Thus, it seems that the direct forecasts from the DFM model, particularly fitting a different model for each forecast horizon h, is advantageous for the longer forecast horizons, in this setting of forecasting data generated from a simple VAR DGP. This is in contrast to the iterative nature of ARIMA generated forecasts.

Note that the same forecasts of the components, generated from univariate ARIMA models, are used for both the FLAP ARIMA and the FLAP DFM methods. It is the increase in the number of the components, that leads to the dominant performance of the FLAP forecasts. Hence, there is valuable information in the series that is not captured by either the ARIMA model or the DFM, but is captured by the components. Once again, this emphasises the importance of the components and FLAP. In this extreme case, a simple model after FLAP, performs as well as a more complicated model, while the forecast model itself is not as important as FLAP.

3.3 Forming components for FLAP

The construction of components is obviously a primary element in the proposed FLAP method. However, the simulation results show that the method of generating the components may not be as important as one may have expected.

The most important feature in Figure 1, as discussed in the previous section, is the number of components used in this simulation setting. The results clearly show that the FLAP forecasts using the maximum considered $p = 300$ components outrank the FLAP forecasts using $p = m = 70$ components, no matter how the components are generated. The only exception seems to be for $h = 12$ and for the DFM base forecasts when using all $p = m = 70$ PCA components or complementing these with randomly generated components from normal weights. Our conjecture, as also stated above, is that the key difference here lies in the quality of the direct DFM base forecasts for the longer forecast horizons, rather than in the component generating mechanism itself. Furthermore, between the methods using the maximum considered $p = 300$, there seems to be no evidence that the component generating mechanism makes any difference to the ranks, and especially there seems to be no difference between the distribution used to generate the random components.

Hence, in Figure 2, we present only the results for principal components and randomly generated components from a standard normal distribution. The results from using components generated from a uniform distribution or random orthonormal matrices are omitted, as these are visually identical to the latter.

Focusing on the ARIMA base forecasts it seems that for a relatively small p, the MSE decreases at a faster rate when principal components are used compared to random components. However, FLAP with random components seems to achieve a lower MSE than PCA as p increases. The gap between the two diminishes for large p. We conjecture that the initial favourable performance of PCA stems from the construction of the orthogonal components aimed at capturing the maximum variance in the data. These components are then ranked from largest to smallest variance, which likely contributes to the method's effectiveness for smaller p compared to random components. We comment more on PCA when analysing the results in Section 4.

The results show that the performance of FLAP using orthonormal weights is almost identical to FLAP using weights simulated from a normal distribution. This leads us to infer that it is not the orthogonality of the principal components that leads to the strong performance of FLAP using PCA. It may also suggest using simple random weights if one is prepared to include a relatively large number of components in FLAP; and using PCA when the intention is to use a small number of components. However, as we will see in Section 4, this is not the case with real data. In this case, PCA seems to be the preferred approach even when the number of components is large.

3.4 Sources of uncertainty

The MSEs of the true VAR data generating process, VAR-DGP, and the correctly specified but estimated VAR, VAR-Est, are plotted at the bottom of each panel in Figure 2. As expected, this confirms that they are the best performing models. As the forecast horizon increases so does the MSE difference between these, with estimation error for the VAR-Est accumulating as forecasts are iteratively generated for the longer forecast horizons.

For both these models, FLAP cannot improve on the base forecasts. More importantly, in the case of VAR-Est, FLAP does not seem to be able to reduce the estimation error or the parameter uncertainty. On the other hand, FLAP shows significant improvements over base forecasts generated from misspecified models, such as the two benchmarks. This implies that the uncertainty which FLAP can account for and reduce, is mainly due to model misspecification. It seems that FLAP operates similarly to bagging, as bagging also reduces variance by controlling model uncertainty

(Petropoulos et al. 2018). The uncertainty in how the components are constructed, unique in FLAP problems, is discussed in Section 3.3 and awaits future research.

4 Empirical applications

In this section, we apply the FLAP method in an empirical setting using two diverse datasets. The analysis confirms many of the results and conclusions drawn from the simulation setting, but it also identifies a few key differences.

4.1 Australian domestic tourism

The Australian Tourism Data Set compiled from the National Visitor Survey by Tourism Research Australia contains the total number of nights spent by Australians away from home. We refer to these as visitor nights. The monthly visitor nights are recorded for $m = 77$ regions around Australia, covering the period January 1998 to December 2019. To evaluate the forecast performance, we conduct am expanding window time series cross-validation (Hyndman and Athanasopoulos 2021). More specifically, the first $T = 84$ observations are used as the first training set and the following 12 months are used as the test set for evaluation. We repeat the evaluation for the rest of the data, by expanding each training sample by one observation at a time. This generates 169 forecasts for evaluation, for each of the forecast horizons, $h = 1$ to 12. Base forecasts for each series and for the components are generated using univariate ETS models selected and fitted using the `ets()` function in the `forecast` package (Hyndman et al. 2023; Hyndman and Khandakar 2008). See Hyndman and Athanasopoulos (2018,chap. 7) for more details.

MCB and MSE plots for $h = 1$, 6 and 12 (as described in Section 3.2) are presented in Figure 3 and Figure 4, respectively. The FLAP forecasts in general outperform the base forecasts, which is consistent with the findings in Section 3. In Figure 3, the base forecasts are always ranked last. Interestingly, FLAP forecasts improve over base forecasts, even when FLAP is using either one or at most two principal components. These improvements are statistically significant across all forecast horizons.

We highlight two findings that diverge from those observed in Section 3.

First, in contrast to the results from Section 3, the method implemented to construct components matters. Using principal components in FLAP improves base forecasts more compared to using random components with normal weights. Figure 3 shows that the best ranked forecasts are the FLAP forecasts using either the $p = m = 77$ principal components, or complementing these with random components so that $p = 200$. These improve on FLAP forecasts using only $p = 200$ random components and the improvements are statistically significant.

This is also clearly reflected in Figure 4, where the reduction in terms of MSE from using PCA is always lower, even after the $p = m$ principal components are exhausted and random components with weights generated from a normal distribution are added.

PCA aims to find components along the direction where the data varies the most. The first principal

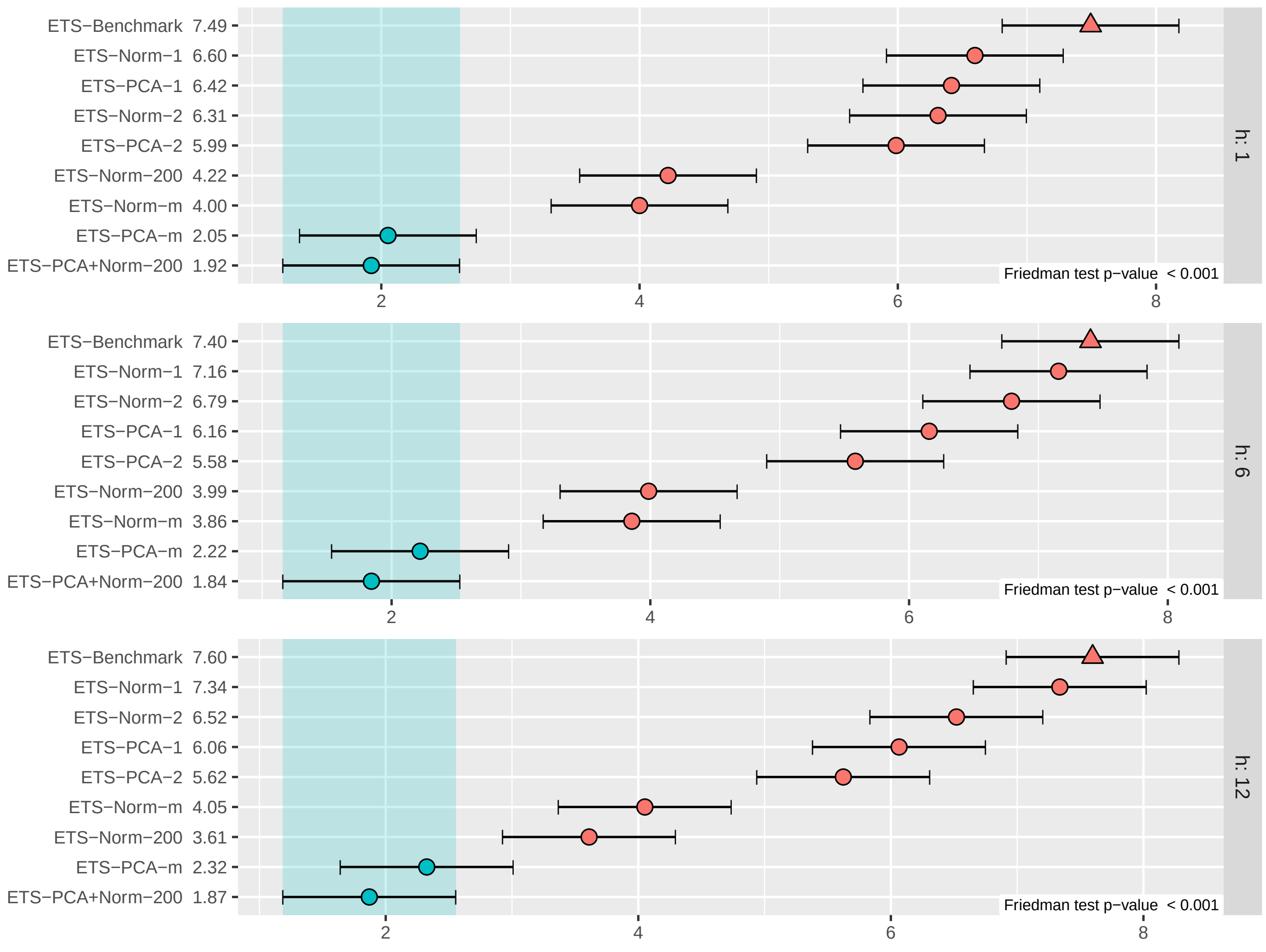

Figure 3: *Average ranks of* 1-, 6- *and* 12-*step-ahead cross-validation MSE of different model and component specifications on the visitor nights data. The methods using forecast projection are named as "Model – Component Weights – Number of Components". The base models are named as "Model – Benchmark" 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.*

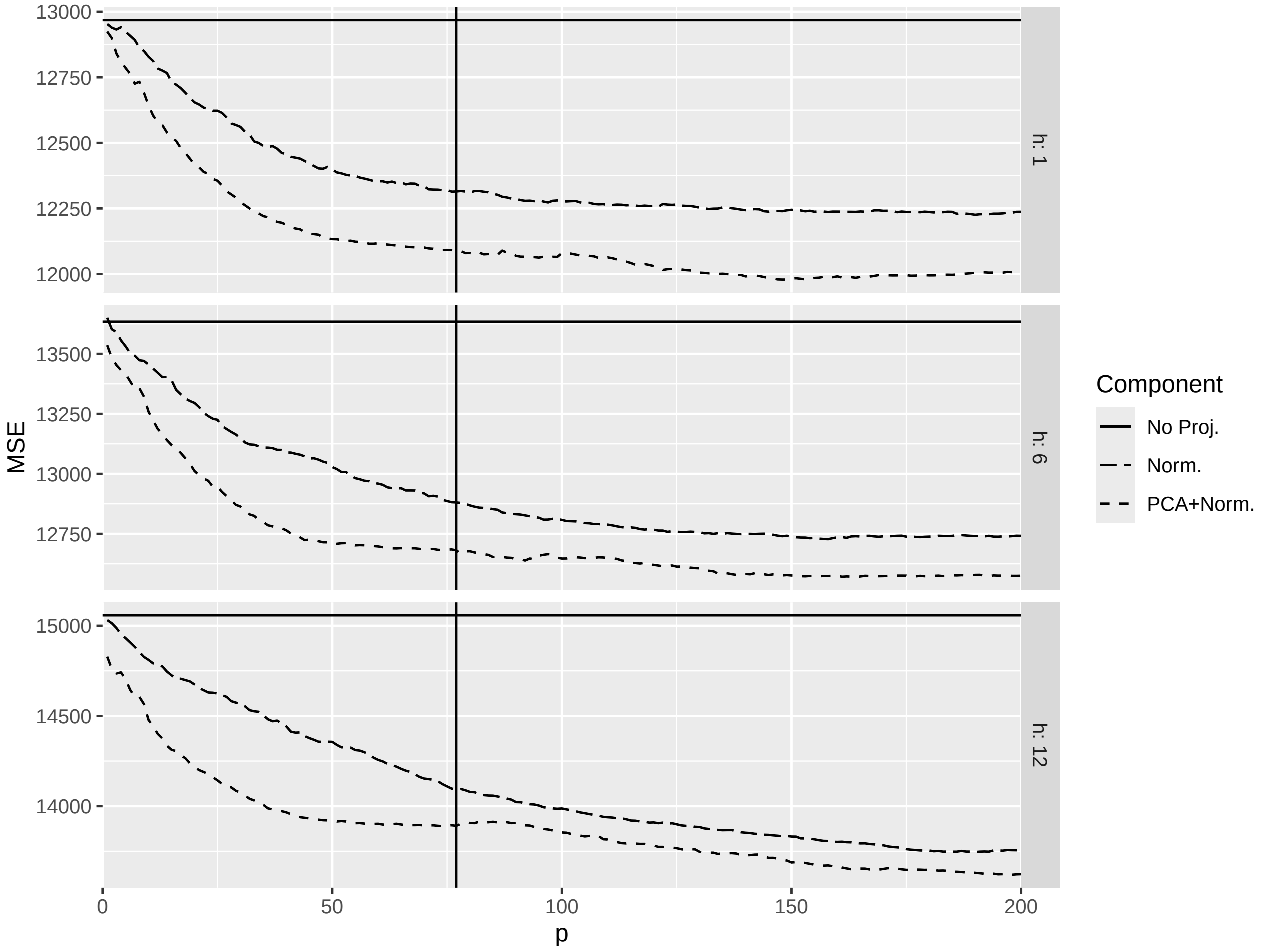

Figure 4: *Out-of-sample MSE of base and FLAP forecasts as the number of components p increases, for forecast horizons* 1, 6 *and* 12*, using the visitor nights data. The solid horizontal lines show the MSE for the base forecasts while the dashed lines show the MSEs of the FLAP forecasts. The vertical black line indicates the location of* $p = m$*, the number of series.*

components accounts for the maximum variance in the data. Subsequent principal components are orthogonal to the previous ones and capture the maximum remaining variance. Based on this variance maximisation and ranking of PCA, we propose two potential explanations for its superior performance.

1. Optimality: By maximising and ranking the variance of PCs from largest to smallest, we ensure that the projection utilises components containing significantly more information (as measured by variance) compared to randomly weighted components.
2. Diversity: Actively seeking principal components with the highest variance, results in the incorporation of a more diverse set of components into the projection.

Second, consistent with the results in Section 3, the MSEs of FLAP forecasts shown in Figure 4 decrease as the number of components increases. However, this decreases ceases, especially for

$h = 1$, approximately for $p > 150$. This is also reflected in the results presented in Figure 3, where the difference in ranking of the two best methods is not statistically significant, even though they use very different numbers of components ($p = m = 77$ and $p = 200$).

Choosing the number of components is a trade-off between the increasing estimation error as the dimension of forecast error variance $\boldsymbol{W}_h$ increases, and the additional benefit brought by the information embedded in the new components. Of course this all depends on the complexity of the DGP. For the visitor nights data, the benefit of components over and above the estimation error, diminishes after the number of components reaches $p = m = 77$.

4.2 FRED-MD

The FRED-MD (McCracken and Ng 2016) is a popular monthly data set of macroeconomic variables, and shares similar properties with the Stock and Watson (2002b) data. We downloaded and transformed the data using the `fbi` package (Chen et al. 2023). The period we use for this exercise is from January 1959 to September 2023, containing 777 observations. Following McCracken and Ng (2016), we replace observations that deviate from the sample median by more than 10 interquartile ranges (which are recognized as outliers), with missing values. We then drop any series with more than 5% observations missing. This results in $m = 122$ series. We fill in the missing values using the expectation-maximization (EM) algorithm described in Stock and Watson (2002a) with 8 factors. The number 8 is identified by McCracken and Ng (2016), albeit with a different time span.

In order to relate to the theoretical forecast error variance reduction, we 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 and scale them to have variance 1. The MSEs are calculated on this standardized scale without back-transformation.

We evaluate the performance of forecasts using time series cross-validation. We start with 300 observations in the first training set and the following 12 observations as the test set. We repeat the exercise by expanding the training set increasing by 1 observation at a time. This provides us with 466 forecasts for each of the forecast horizons from $h = 1$ to 12 for evaluation.

We generate base forecasts from ARIMA models using `auto.arima()` function from the `forecast` package (Hyndman et al. 2023) with the default settings, and DFMs. The ranges of the meta-parameters in the DFMs 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 $0 \leq s \leq 6$. For more details see Section 3.1. The MCB and MSE plots for $h = 1$, 6 and 12 are presented in Figure 5 and Figure 6, respectively.

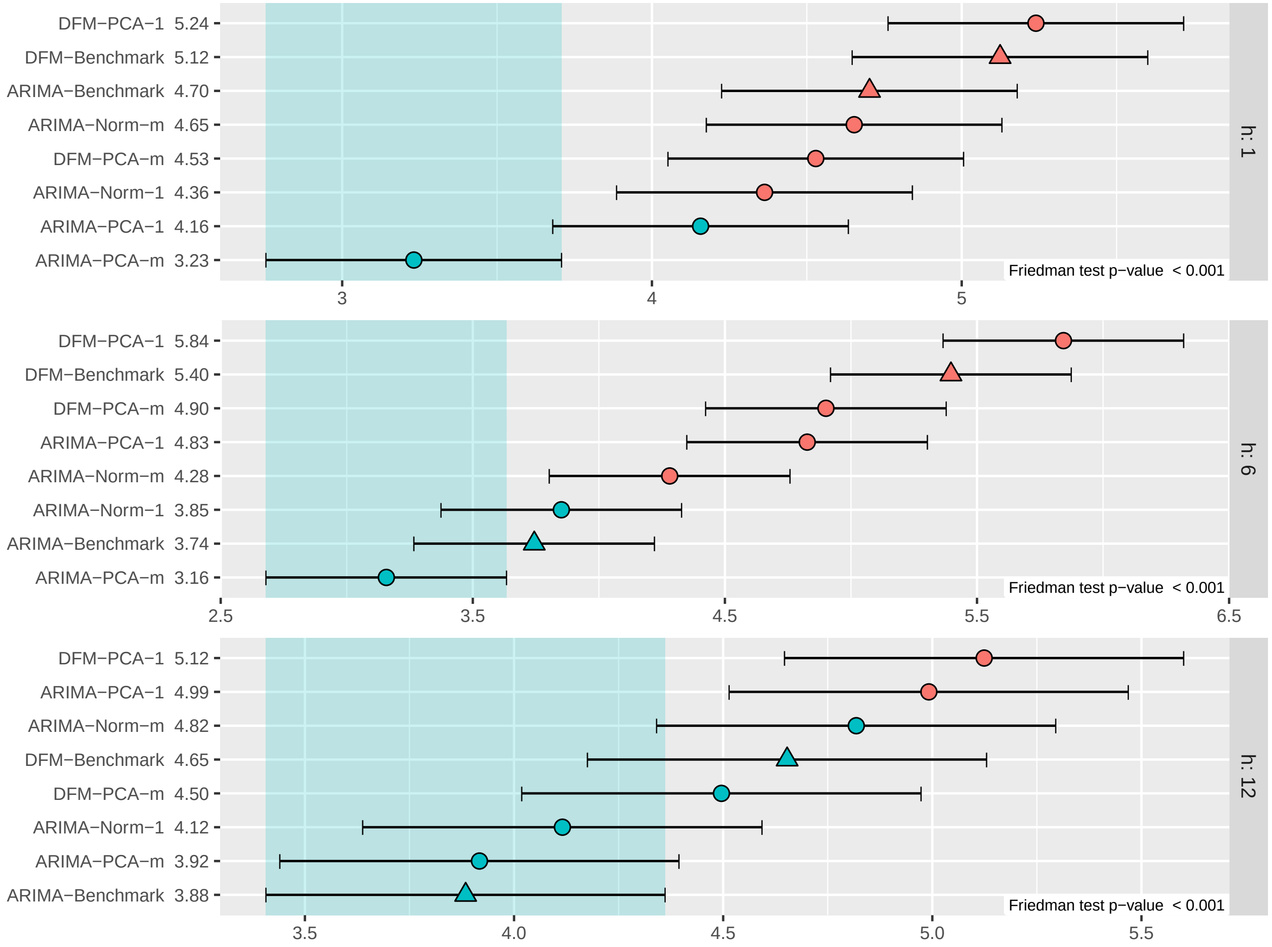

Figure 5: *Average ranks of* 1-, 6- *and* 12-*step-ahead cross-validation MSE of different model and component specifications on the FRED-MD data. The methods using forecast projection are named as "Model – Component Weights – Number of Components". The base models are named as "Model – Benchmark" and 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 results presented in Figure 5 show that the ARIMA base forecasts rank better that the DFM base forecasts. This difference is statistically significant for $h = 6$. Likewise, ARIMA base forecasts return a lower MSE compared to DFM as shown in Figure 6 across all forecast horizons.

The best ranked forecasts across all forecast horizons are once again FLAP forecasts generated using principal components. The improvements over the ARIMA base forecasts are significantly significant for $h = 1$ and 6. Furthermore, FLAP forecasts using PCA rank better than FLAP forecasts

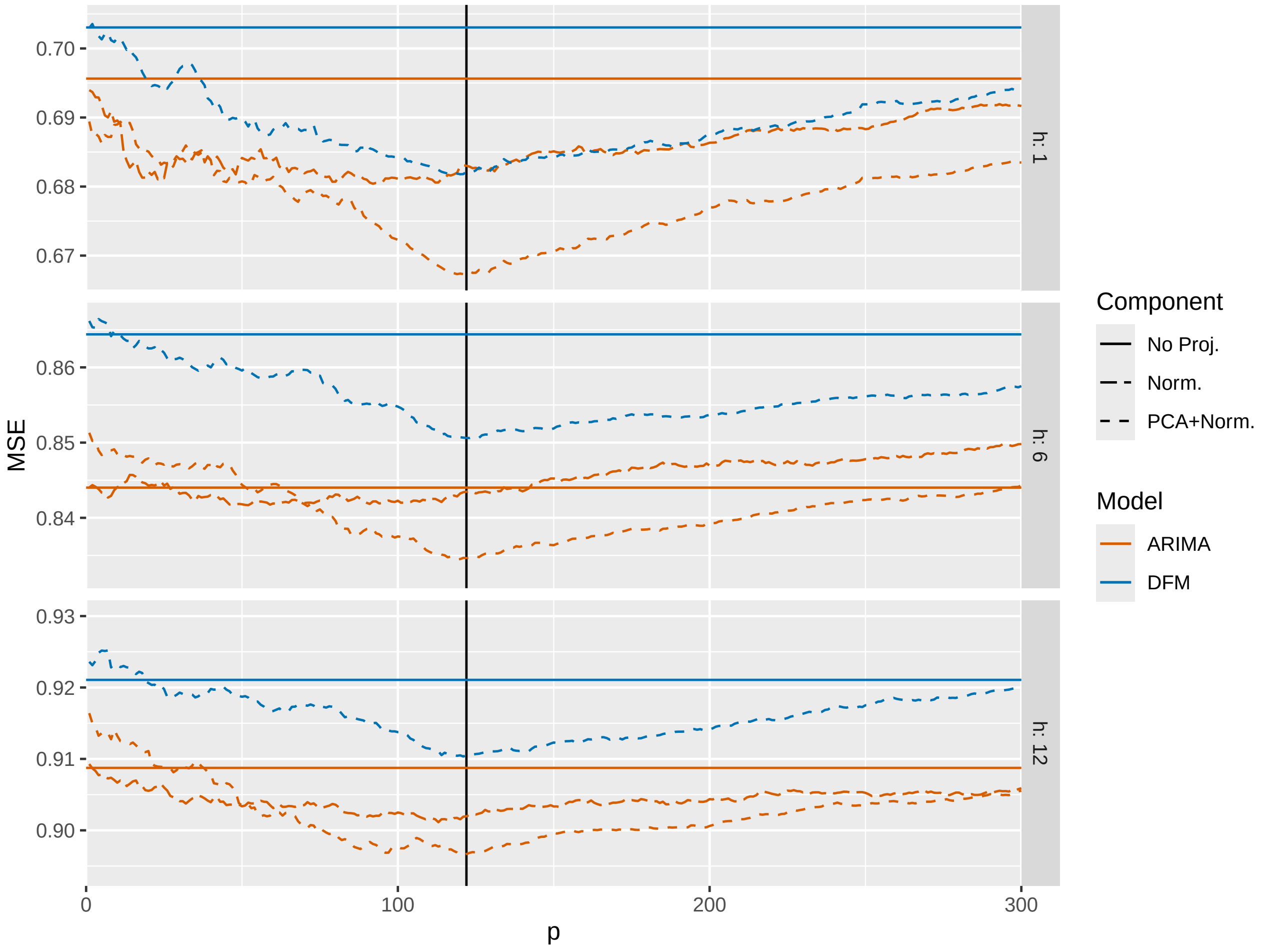

Figure 6: *Out-of-sample MSE for base and FLAP forecasts as the number of components p increases, for forecast horizons 1, 6 and 12, using the FRED-MD data. The solid horizontal lines show the MSE for the base forecasts while the dashed lines show the MSEs of the FLAP forecasts. The vertical black line indicates the location of $p = m$ the number of series.*

using components with random weights, and these differences are statistically significant. This is also observed in terms of MSE in Figure 6, reaffirming our findings based on the tourism data set discussed in Section 4.1. Figure 6 shows that both FLAP ARIMA and FLAP DFM forecasts, seem to be worse than the base forecasts when using a small number of components p for $h = 6$ and 12. However, this improves as p increases and FLAP forecasts outperform the base forecasts gradually. The trade-off between the benefit of including more components and the loss in degrees of freedom due to estimation in a large dimension is once again observed. However in this case it seems to be more extreme, compared to the tourism example, with MSEs starting to increase for $p > m$.

As we have seen in Section 3 and Section 4.1, m is not always a clear cut-off point in determining the number of components to use. The performance of FLAP is jointly determined by the number of series m, the sample size T, the component construction method, and the DGP. The example of FRED-MD shows the importance of PCA, as m is the point that the component changes from PCA to random normal weighted linear combinations, implying that PCA can exploit the information in the data while random weights cannot. This is more obvious for $h = 1$ and $h = 6$, as PCA works when $p < m$, but random normal weights do not seem to work from the beginning.

5 Conclusion

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 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 error variance of the components, while Matteson and 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. Another route to improving FLAP may be found by optimizing Equation 10 over $\boldsymbol{\Phi}$ and $\boldsymbol{G}$ rather than just $\boldsymbol{G}$.

Component construction should be studied together with the selection of the forecast model since both the weight matrix $\boldsymbol{\Phi}$ and the base forecast error variance $\boldsymbol{W}_h$ can affect the projection simultaneously in Equation 2. This is likely to be related to forecast combination methods, 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 and Dua (1995), Kang et al. (2022) and Lichtendahl and Winkler (2020).

Finally, while FLAP is motivated by the forecast reconciliation literature, our focus here 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.

Supplementary materials

The online supplementary materials contain the appendices for the article.

References

Ando, S., and Narita, F. (2024), "An Alternative Proof of Minimum Trace Reconciliation," *Forecasting*, 6, 456–461. https://doi.org/10.3390/forecast6020025.

Assimakopoulos, V., and Nikolopoulos, K. (2000), "The Theta Model: A Decomposition Approach to Forecasting," *International Journal of Forecasting*, 16, 521–530. https://doi.org/10.1016/S0169-2070(00)00066-2.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., and Panagiotelis, A. (2023), "Forecast Reconciliation: A Review," *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2023.10.010.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., and Petropoulos, F. (2017), "Forecasting with Temporal Hierarchies," *European Journal of Operational Research*, 262, 60–74. https://doi.org/10.1016/j.ejor.2017.02.046.

Batchelor, R., and Dua, P. (1995), "Forecaster Diversity and the Benefits of Combining Forecasts," *Management Science*, 41, 68–75. https://doi.org/10.1287/mnsc.41.1.68.

Bergmeir, C., Hyndman, R. J., and Benítez, J. M. (2016), "Bagging Exponential Smoothing Methods Using STL Decomposition and Box–Cox Transformation," *International Journal of Forecasting*, 32, 303–312. https://doi.org/10.1016/j.ijforecast.2015.07.002.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *The Quarterly Journal of Economics*, 120, 387–422.

Borchers, H. W. (2023), *pracma: Practical Numerical Math Functions*. https://doi.org/10.32614/CRAN.package.pracma.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. https://doi.org/10.1007/bf00058655.

Carriero, A., Galvao, A. B., and Kapetanios, G. (2019), "A Comprehensive Evaluation of Macroeconomic Forecasting Methods," *International Journal of Forecasting*, 35, 1226–1239.

Chen, Y., Ng, S., and Bai, J. (2023), *fbi: Factor-Based Imputation and FRED-MD/QD Data Set*.

De Stefani, J., Le Borgne, Y.-A., Caelen, O., Hattab, D., and Bontempi, G. (2019), "Batch and Incremental Dynamic Factor Machine Learning for Multivariate and Multi-Step-Ahead Forecasting," *International Journal of Data Science and Analytics*, 7, 311–329. https://doi.org/10.1007/s41060-018-0150-x.

Di Fonzo, T., and Girolimetto, D. (2023), "Cross-Temporal Forecast Reconciliation: Optimal Combination Method and Heuristic Alternatives," *International Journal of Forecasting*, 39, 39–57. https://doi.org/10.1016/j.ijforecast.2021.08.004.

Disney, S. M., and Petropoulos, F. (2015), "Forecast Combinations Using Multiple Starting Points."

Fabio Di Narzo, A., Aznarte, J. L., and Stigler, M. (2009), *tsDyn: Time Series Analysis Based on Dynamical Systems Theory*. https://doi.org/10.32614/CRAN.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, 675–701. https://doi.org/10.1080/01621459.1937.10503522.

Friedman, M. (1939), “A Correction,” *Journal of the American Statistical Association*, 34, 109–109. https://doi.org/10.1080/01621459.1939.10502372.

Goerg, G. (2013), “Forecastable Component Analysis,” in *Proceedings of the 30th International Conference on Machine Learning*, 64–72.

Hastie, T., Tibshirani, R., and Friedman, J. (2003), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics, New York, NY: Springer. https://doi.org/10.1007/978-0-387-21606-5.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013), *Nonparametric Statistical Methods*, John Wiley & Sons.

Hollyman, R., Petropoulos, F., and Tipping, M. E. (2021), “Understanding Forecast Reconciliation,” *European Journal of Operational Research*, 294, 149–160. https://doi.org/10.1016/j.ejor.2021.01.017.

Hyndman, R. J., and Athanasopoulos, G. (2018), *Forecasting: Principles and Practice*, Melbourne, Australia: OTexts.

Hyndman, R. J., and Athanasopoulos, G. (2021), *Forecasting: Principles and Practice*, Melbourne, Australia: OTexts.

Hyndman, R. J., and Khandakar, Y. (2008), “Automatic Time Series Forecasting: The Forecast Package for R,” *Journal of Statistical Software*, 27, 1–22. https://doi.org/10.18637/jss.v027.i03.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O’Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2023), *forecast: Forecasting Functions for Time Series and Linear Models*. https://doi.org/10.32614/CRAN.package.forecast.

Jolliffe, I. T. (2002), *Principal Component Analysis*, Springer, New York, NY. https://doi.org/10.1007/b98835.

Kang, Y., Cao, W., Petropoulos, F., and Li, F. (2022), “Forecast with Forecasts: Diversity Matters,” *European Journal of Operational Research*, 301, 180–190. https://doi.org/10.1016/j.ejor.2021.10.024.

Koning, A. J., Franses, P. H., Hibon, M., and Stekler, H. O. (2005), "The M3 Competition: Statistical Tests of the Results," *International Journal of Forecasting*, 21, 397–409. https://doi.org/10.1016/j.ijforecast.2004.10.003.

Kourentzes, N. (2023), *tsutils: Time Series Exploration, Modelling and Forecasting*. https://doi.org/10.32614/CRAN.package.tsutils.

Kourentzes, N., Petropoulos, F., and Trapero, J. R. (2014), "Improving Forecasting by Estimating Time Series Structural Components Across Multiple Frequencies," *International Journal of Forecasting*, 30, 291–302. https://doi.org/10.1016/j.ijforecast.2013.09.006.

Li, X., Petropoulos, F., and Kang, Y. (2022), "Improving Forecasting by Subsampling Seasonal Time Series," *International Journal of Production Research*, 61, 1–17. https://doi.org/10.1080/00207543.2021.2022800.

Lichtendahl, K. C., Jr, and Winkler, R. L. (2020), "Why Do Some Combinations Perform Better Than Others?" *International Journal of Forecasting*, 36, 142–149. https://doi.org/10.1016/j.ijforecast.2019.03.027.

Luenberger, D. G. (1969), *Optimization by Vector Space Methods*, Nashville, TN: John Wiley & Sons.

Matteson, D. S., and Tsay, R. S. (2011), "Dynamic Orthogonal Components for Multivariate Time Series," *Journal of the American Statistical Association*, 106, 1450–1463. https://doi.org/10.1198/jasa.2011.tm10616.

McCracken, M. W., and Ng, S. (2016), "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34, 574–589. https://doi.org/10.1080/07350015.2015.1086655.

Nemenyi, P. B. (1963), "Distribution-Free Multiple Comparisons."

Opgen-Rhein, R., and Strimmer, K. (2007), "Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach," *Statistical Applications in Genetics and Molecular Biology*, 6. https://doi.org/10.2202/1544-6115.1252.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., and Hyndman, R. J. (2021), "Forecast

Reconciliation: A Geometric View with New Insights on Bias Correction," *International Journal of Forecasting*, 37, 343–359. https://doi.org/10.1016/j.ijforecast.2020.06.004.

Petropoulos, F., Hyndman, R. J., and Bergmeir, C. (2018), "Exploring the Sources of Uncertainty: Why Does Bagging for Time Series Forecasting Work?" *European Journal of Operational Research*, 268, 545–554. https://doi.org/10.1016/j.ejor.2018.01.045.

Petropoulos, F., and Spiliotis, E. (2021), "The Wisdom of the Data: Getting the Most Out of Univariate Time Series Forecasting," *Forecasting*, 3, 478–497. https://doi.org/10.3390/forecast3030029.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. (1974), "Projectors, Generalized Inverses and the Blue'S," *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 36, 442–448. https://doi.org/10.1111/j.2517-6161.1974.tb01019.x.

Schafer, J., Opgen-Rhein, R., Zuber, V., Ahdesmaki, M., Silva, A. P. D., and Strimmer., K. (2021), *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*. https://doi.org/10.32614/CRAN.package.corpcor.

Schäfer, J., and Strimmer, K. (2005), "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics," *Statistical Applications in Genetics and Molecular Biology*, 4. https://doi.org/10.2202/1544-6115.1175.

Stock, J. H., and Watson, M. W. (2002a), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162. https://doi.org/10.1198/073500102317351921.

Stock, J. H., and Watson, M. W. (2002b), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. https://doi.org/10.1198/016214502388618960.

Stock, J. H., and Watson, M. W. (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business & Economic Statistics*, 30, 481–493. https://doi.org/10.1080/07350015.2012.715956.

Tsay, R. S. (2013), *Multivariate Time Series Analysis: With R and Financial Applications*, John Wiley & Sons.

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2023), "Forecast Combinations: An over 50-Year Review," *International Journal of Forecasting*, 39, 1518–1547. https://doi.org/10.1016/j.ijforecast.2022.11.005.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2019), "Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization," *Journal of the American Statistical Association*, 114, 804–819. https://doi.org/10.1080/01621459.2018.1448825.

Yang, Y. F. (2024), *flap: Forecast Linear Augmented Projection*. https://doi.org/10.32614/CRAN.package.flap.

A Appendix

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 $\boldsymbol{z}$ such that $\boldsymbol{M}\boldsymbol{z} = \boldsymbol{y}$ for some $\boldsymbol{y}$, we have

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

Thus, $\boldsymbol{M}$ projects any vector onto the space where the constraint $\boldsymbol{C}\boldsymbol{y} = 0$ is satisfied.

Proof of Corollary 2.1

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

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

Proof of Lemma 2.2

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

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

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 $\mathrm{Var}(\boldsymbol{y}_{t+h}-\hat{\boldsymbol{y}}_{t+h})-\mathrm{Var}(\boldsymbol{y}_{t+h}-\tilde{\boldsymbol{y}}_{t+h})$ 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 $\boldsymbol{c}_t^*=\boldsymbol{\Phi}^*\boldsymbol{y}_t$ in the projection. We define $\boldsymbol{z}_t^*=\begin{bmatrix}\boldsymbol{z}_t\\ \boldsymbol{c}_t^*\end{bmatrix}$, the constraint matrix

$$\boldsymbol{C}^*=\begin{bmatrix}\multicolumn{2}{c}{\boldsymbol{C}} & \underset{p\times q}{\boldsymbol{0}}\\ -\underset{q\times m}{\boldsymbol{\Phi}^*} & \underset{q\times p}{\boldsymbol{0}} & \boldsymbol{I}_q\end{bmatrix}=\begin{bmatrix}-\underset{p\times m}{\boldsymbol{\Phi}} & \boldsymbol{I}_p & \underset{p\times q}{\boldsymbol{0}}\\ -\underset{q\times m}{\boldsymbol{\Phi}^*} & \underset{q\times p}{\boldsymbol{0}} & \boldsymbol{I}_q,\end{bmatrix}=\begin{bmatrix}\overline{\boldsymbol{C}}\\ \underline{\boldsymbol{C}}\end{bmatrix} \tag{12}$$

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 error variance matrix

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

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

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

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

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

and

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

which gives

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

and

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

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

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

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 error variance stays the same:

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

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

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

Taking the difference, we have

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

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}} \\
&\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,

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

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

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

being positive semi-definite. Finally, we have

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

being a positive semi-definite matrix where the diagonal terms are non-negative, whose trace, therefore, is non-negative. This means using a larger number of components in the mapping achieves lower or equal forecast error variances. In other words, the reduction in forecast error variance of each series is non-decreasing as more components are added, 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 $\boldsymbol{W}^{(i)}_h$ to be the base forecast error variance of the original series and the first i components. Starting with the first component, Equation 4 becomes

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

$$\text{where}\qquad \boldsymbol{\psi}_1\boldsymbol{W}^{(1)}_h\boldsymbol{J}'_{m,1} = \begin{bmatrix}-\boldsymbol{\phi}_1 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{W}_{y,h} & \boldsymbol{w}'_{c_1y,h}\\ \boldsymbol{w}_{c_1y,h} & W_{c_1,h}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_m\\ 0\end{bmatrix}$$

$$= -\boldsymbol{\phi}_1\boldsymbol{W}_{y,h}+\boldsymbol{w}_{c_1y,h}.$$

Equation 16 is obviously non-negative. For it to be larger than 0, we need $\boldsymbol{\psi}_1\boldsymbol{W}^{(1)}_h\boldsymbol{J}'_{m,1}\neq 0$, which gives $\boldsymbol{\phi}_1\boldsymbol{W}_{y,h}\neq\boldsymbol{w}_{c_1y,h}$.

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 error 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

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

where $\boldsymbol{W}_{\tilde{y},h}^{(i-1)} = \boldsymbol{J}_{m,i-1} \boldsymbol{M}_{i-1} \boldsymbol{W}_h^{(i-1)} \boldsymbol{J}_{m,i-1}'$ is the projected forecast error 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 $\boldsymbol{w}_{\tilde{c}_i \tilde{y},h}^{(i-1)}$ 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 $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{I}$, $\boldsymbol{S}\boldsymbol{G}$ is a projection matrix: $\boldsymbol{S}\boldsymbol{G}\boldsymbol{S}\boldsymbol{G} = \boldsymbol{S}\boldsymbol{G}$.

For any $\boldsymbol{z}$ such that $\boldsymbol{S}\boldsymbol{G}\boldsymbol{z} = \boldsymbol{y}$ for some $\boldsymbol{y}$, we have $\boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\boldsymbol{S}\boldsymbol{G}\boldsymbol{z} = \boldsymbol{0}$ because $\boldsymbol{C}\boldsymbol{S} = \begin{bmatrix}-\boldsymbol{\Phi} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{\Phi}'\end{bmatrix}' = \boldsymbol{0}$. Similarly to $\boldsymbol{M}$, $\boldsymbol{S}\boldsymbol{G}$ 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

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

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

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

Proof of Lemma 2.4

Let the base and projected forecast errors be given as

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

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

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

and therefore

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

Proof of Theorem 2.4

This can be proved in a few different ways. We adopt the approach of Ando and Narita (2024) 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

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

is stationary at an extremum $\boldsymbol{G}$ (Luenberger 1969,p. 243, Theorem 1). We find the numerator of the Gateaux differential (Luenberger 1969,p. 171)

$$\lim_{\alpha\to 0}\frac{L(\boldsymbol{G}+\alpha\boldsymbol{H}) - L(\boldsymbol{G})}{\alpha}$$

to be

$$\begin{aligned}
L(\boldsymbol{G}+\alpha\boldsymbol{H}) - L(\boldsymbol{G}) =& \operatorname{tr}((\boldsymbol{G}+\alpha\boldsymbol{H})\boldsymbol{W}_h(\boldsymbol{G}+\alpha\boldsymbol{H})') + \operatorname{tr}(\boldsymbol{\Lambda}'(\boldsymbol{I} - (\boldsymbol{G}+\alpha\boldsymbol{H})\boldsymbol{S})) - \operatorname{tr}(\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}') - \operatorname{tr}(\boldsymbol{\Lambda}'(\boldsymbol{I} - \boldsymbol{G}\boldsymbol{S}))\\
=& \operatorname{tr}(\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}') + \operatorname{tr}(\alpha^2\boldsymbol{H}\boldsymbol{W}_h\boldsymbol{H}') + \operatorname{tr}(\alpha\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{H}') + \operatorname{tr}(\alpha\boldsymbol{H}\boldsymbol{W}_h\boldsymbol{G}') + \operatorname{tr}(\boldsymbol{\Lambda}'(\boldsymbol{I} - \boldsymbol{G}\boldsymbol{S})) - \operatorname{tr}(\alpha\boldsymbol{\Lambda}'\boldsymbol{H}\boldsymbol{S})\\
&- \operatorname{tr}(\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}') - \operatorname{tr}(\boldsymbol{\Lambda}'(\boldsymbol{I} - \boldsymbol{G}\boldsymbol{S}))\\
=& \alpha^2\operatorname{tr}(\boldsymbol{H}\boldsymbol{W}_h\boldsymbol{H}') + \alpha\operatorname{tr}(\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{H}') + \alpha\operatorname{tr}(\boldsymbol{H}\boldsymbol{W}_h\boldsymbol{G}') - \alpha\operatorname{tr}(\boldsymbol{\Lambda}'\boldsymbol{H}\boldsymbol{S}).
\end{aligned}$$

Thus, the Gateaux differential becomes

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

which we set to zero with a value of $\boldsymbol{G}^*$

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

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

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

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

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

To see how Equation 10 can be split into separate problems, recognize

$$\text{tr}(\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}')=\sum_{i=1}^{m}\boldsymbol{g}_i'\boldsymbol{W}_h\boldsymbol{g}_i,$$

where $\boldsymbol{W}_h$ is a positive definite variance covariance matrix, which makes each $\boldsymbol{g}_i'\boldsymbol{W}_h\boldsymbol{g}_i$ positive. Additionally, the element in the ith row and jth column of $\boldsymbol{G}\boldsymbol{S}$ is $\boldsymbol{g}_i'\boldsymbol{s}_j$, and the element in the ith row and jth column of an identity matrix is δ_{ij}. Therefore, the problem in Equation 10 is m separate problems.