Optimal Multi-step VAR Forecasting Averaging

Jen-Che Liao*and Wen-Jen Tsay†

Institute of Economics, Academia Sinica

June 2017

Abstract

This paper proposes the frequentist multiple-equation least squares averaging approaches for multi-step forecasting with vector autoregressive (VAR) models. The proposed VAR forecasting averaging methods are based on the multivariate Mallows model averaging (MMMA) and multivariate leave-h-out cross-validation averaging (MCVA$_h$) criteria (with h denoting the forecast horizon), which are valid for iterative and direct multi-step forecasting averaging, respectively. Under the framework of stationary VAR processes of infinite order, we provide theoretical justifications by establishing asymptotic unbiasedness and asymptotic optimality of the proposed forecasting averaging approaches. Specifically, MMMA exhibits asymptotic optimality for one-step ahead forecast averaging, whereas for direct multi-step forecasting averaging the asymptotically optimal combination weights are determined separately for each forecast horizon based on the MCVA$_h$ procedure. The finite-sample behaviour of the proposed averaging procedures under misspecification is investigated via simulation experiments. An empirical application to a three-variable monetary VAR, based on the U.S. data, is also provided to present our methodology.

Keywords: *Asymptotic optimality, Forecast combination/averaging, Iterative and direct multi-step forecasting, Vector autoregressions*
JEL Classification: C13, C32, C53

*Correspondence to: Liao, Institute of Economics, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan. Tel.: +886-2-2782-2791; Fax: +886-2-2785-3946. E-mail: jcliao@econ.sinica.edu.tw.

†Tsay: wtsay@econ.sinica.edu.tw. This paper was previously circulated under the title "Multivariate Least Squares Forecasting Averaging by Vector Autoregressive Models."

1 Introduction

Vector autoregressive (VAR) models are one of the most prevalent ways to analyze multivariate time series in the econometric and statistical literature. As a technique to characterize the joint dynamic behavior of economic variables, the VAR model has gained widespread use in theoretical and applied macroeconomic and financial economic research since being introduced by Sims (1980), with primary applications to forecasting and policy analysis. A key practical question of using VAR models is the number of lagged terms to be introduced to the VAR model.[1] This kind of model uncertainty may considerably impact the performance of a VAR-based estimation, inference, forecasting, and other analysis. This paper addresses the issue of VAR model specification via a frequentist model averaging approach under iterated and direct multi-step forecasting frameworks.

To examine the issue of model specification, a great deal of attention has been paid to model selection and model averaging in the statistics and econometrics literature. Model selection and model averaging are appealing, because they result in a lower mean squared error (MSE) by trading off bias and variance, which is a standard problem with this big strand in the literature. In the context of VAR lag length determination, two basic selection strategies have been used. The first approach uses the sequential likelihood ratio tests suggested by Tiao and Box (1981). The second one selects the VAR lag order based on information criteria. The three commonly used selection criteria are the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan-Quinn (HQ). Other selection criteria that have been studied in the literature include final prediction error and some variants of AIC and BIC that are designed to correct for overfitting in VAR models.[2]

As a more general approach versus model selection, model averaging methods are introduced in order to reduce variability in model selection and thus increase estimation accuracy. In fact, the application of model averaging techniques has largely focused on either single-equation forecasting procedures or multivariate forecasting based on Bayesian model averaging (e.g., Andersson and Karlsson (2007) and Clark and McCracken (2010)). For the former, Hansen (2008), among others, proposes a least-squares forecast averaging method based on Mallows model averaging for stationary time series observations. Cheng and Hansen (2015) consider forecast averaging with factor-augmented regression models. Zhang, Wan, and Zou (2013) and Cheng, Ing, and Yu (2015) respectively suggest a jackknife averaging approach and an autocorrelation-robust averaging method under the time series framework. Gao, Zhang, Wang, and Zou (2016) propose a leave-subject-out model averaging procedure for longitudinal data models and time series models with heteroskedastic errors.

Under a similar VAR setting to ours, Hansen (2016) introduces the Stein combination shrinkage for VARs in which unrestricted least squares estimates are shrunk toward multiple

[1] As pointed out by Elliott and Timmermann (2016), what makes VARs a popular forecasting tool is due to their relative simplicity, whereby only the choices of the variables to be forecast and lag length of variables need to be made for the forecaster to construct forecasts.

[2] Interested readers may refer to McQuarrie and Tsai (1998, Chapter 5) for a detailed discussion.

restricted least squares estimates. The major difference between Hansen (2016) and the present paper is that Hansen (2016) focuses on an estimation error for the model parameters and that the parameter of interest is a non-linear function of VAR coefficients such as iterated multi-step forecasts and multi-step impulse responses, whereas we are concerned with iterative as well as direct multi-step VAR forecasting averaging. Moreover, in contrast to Hansen (2016), assuming that the underlying VAR model has a finite and known order, we consider VAR processes of unknown and possibly infinite order. On the other hand, Hansen (2016) establishes an approximating MSE for which the VAR coefficients are assumed to be local to the restrictions. This local misspecification is not needed in our analysis.

We propose two novel multi-step VAR forecast averaging procedures based on the MMMA and MCVA_h criteria, both of which are well designed for prediction. To our knowledge, these two criteria have not yet been introduced nor investigated either theoretically or empirically in the problem of multi-step VAR forecasting averaging. Specifically, this paper offers several contributions in the following aspects. First, we propose an easy-to-implement multivariate forecast combination procedure based on the MMMA criterion that extends the frequentist forecast/model averaging to the setting of multivariate response variables. In the single-equation forecasting as a special case, our MMMA procedure reduces to Hansen's (2008) Mallows averaging. The implementation involves an ordinary least squares (OLS) estimation and solving for quadratic programming problems, where the latter can be easily solved through existing statistical programming software such as Gauss, Matlab, and R. The proposed MMMA method is designed for one-step forecast averaging, from which the averaging multi-step forecasts can be obtained by the iterative strategy.

A second contribution is that we further extend the VAR forecasting averaging to the direct forecasting framework, where the problem of serial correlation in forecast errors arises due to overlaps in the data when a forecast horizon of more than a single period is considered. To address this issue, we propose a new direct multi-step VAR forecasting averaging method based on the idea of leave-h-out cross-validation. Moreover, it is worth emphasizing that the main distinction of our multivariate averaging criteria with the single-equation version (e.g., Hansen (2008)) lies in the use of the inverse of the estimated forecast error covariance matrix. This is motivated with the aim to scale each response variable to be of equal importance and to incorporate potential correlations across equations in the VAR system, thereby likely improving forecast accuracy. This view is evidenced in our simulation experiments in Section 6.

Theoretical and empirical investigations of iterative and direct multi-step forecasting with time series models based on a fixed lag order or lag selection have been widely studied in the statistics and econometrics literature, e.g. Kunitomo and Yamamoto (1985), Bhansali (1996, 1997, 1999), Ing (2003), Chen, Yang, and Hafner (2004), Schorfheide (2005), Chevillon and Hendry (2005), Marcellino, Stock, and Watson (2006), Chevillon (2007), Pesaran, Pick, and Timmermann (2011), among others. However, to our knowledge, no efforts have been made for iterative and direct VAR forecasting averaging problems (in a non-Bayesian sense). The present paper offers a third contribution by filling this gap in the literature. On the one hand,

the proposed multi-step VAR forecasting averaging methods are theoretically examined. Our theoretical justifications hinge on the properties of asymptotic unbiasedness and asymptotic optimality of the proposed multivariate averaging criteria. Our theory shows that the classic observation, whereby the MMMA and MCVA$_h$ criteria are unbiased estimators of the MSE, still holds for forecast averaging in a multivariate time series framework. Furthermore, while there are considerable views about the optimality theory for model selection, there is very little theory under the time series context, even for single-equation model averaging problems. Most of the theory for model averaging is for the cross-section case, e.g., Hansen (2007), Wan, Zhang, and Zou (2010), Hansen and Racine (2012), and Liu, Okui, and Yoshimura (2016). For the dependent data, Zhang, Wan, and Zou (2013) generalize Hansen and Racine's (2012) jackknife averaging criterion to two time series cases: serially correlated errors and lagged dependent variables. Cheng, Ing, and Yu (2015) propose a autocorrelation-robust Mallows criterion under time series errors. Our theory extends these existing asymptotic optimality results to a setting of multi-step VAR forecasting averaging. In particular, our optimality results show that MMMA is asymptotically efficient for one-step ahead forecast averaging, whereas for multi-step ahead forecast averaging the asymptotically optimal combination weights are determined separately for each forecast horizon by the direct method based on the MCVA$_h$ criterion.

From an empirical perspective, this paper sheds new light on the relative merits of iterative versus direct methods in the context of VAR forecasting averaging. We provide numerical evidence via simulation experiments and an empirical application to a prototypical monetary VAR model for three U.S. macroeconomic time series of GDP, the GDP deflator, and the federal funds rate. Specifically, our numerical results reveal that the iterative MMMA tends to be preferable when the candidate model set contains VAR models with sufficiently long lags, when the candidate models are not highly misspecified, and when forecasting the GDP and federal funds rate series; conversely, the direct MCVA$_h$ exhibits substantial advantages when the model misspecification is severe, and when forecasting the GDP deflator. On the other hand, the direct MCVA$_h$ deteriorates as the forecast horizon lengthens under correct model specification or mild misspecification. Generally speaking, as the forecast horizon and maximum lag order increase, the robustness of the direct MCVA$_h$ tends to be outweighted by its efficiency loss.

The rest of the paper is organized as follows. Section 2 sets out the framework of multivariate time series forecasting with VAR models and discusses the determination of the VAR lag order. To deal with uncertainty arising from the VAR order selection, Section 3 suggests an iterative multi-step forecast averaging procedure based on the MMMA criterion. Section 4 further proposes a MCVA$_h$ procedure to address the serial correlation problem that arises under the direct multi-step forecasting scheme. Built on asymptotic unbiasedness and asymptotic optimality, Section 5 provides theoretical validity of our methods. Sections 6 and 7 respectively present the numerical performance of our methodology via finite-sample simulation experiments and an empirical application to a three-variable monetary VAR based on the U.S. data constructed by Stock and Watson (2009). We conclude the paper in Section 8.

Mathematical proofs of the theorems are collected in Appendix.

2 Forecasting problems by fitting VAR(p) models

Let the stationary K-dimensional multivariate time series $\{\mathbf{y}_t\}$ be a vector-valued moving average (MA) process:

$$\mathbf{y}_t = \sum_{j=0}^{\infty} \mathbf{\Phi}_j \boldsymbol{\varepsilon}_{t-j}, \tag{2.1}$$

where $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{Kt})'$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{Kt})'$, $t = 0, \pm 1, \pm 2, \ldots$, is a sequence of i.i.d. random vector with $E(\mathbf{y}_t) = 0$ and $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') = \mathbf{\Sigma}$, and $\mathbf{\Phi}_j$ are MA coefficients with $\mathbf{\Phi}_0$ set to the $K \times K$ identity matrix, denoted by $\mathbf{I}_K$. The intercept term has been dropped by assuming without loss of generality that the mean $E(\mathbf{y}_t)$ is already subtracted out.

Under the assumptions that $\sum_{j=0}^{\infty} \|\mathbf{\Phi}_j\| < \infty$ and $\det(\mathbf{\Phi}(z)) \neq 0$ for $|z| \leq 1$, where $\|\mathbf{\Phi}_j\| = \sqrt{\mathrm{tr}(\mathbf{\Phi}_j' \mathbf{\Phi}_j)}$, $\mathbf{\Phi}(z) = \sum_{j=0}^{\infty} \mathbf{\Phi}_j z^j$, and $\det(\mathbf{A})$ and $\mathrm{tr}(\mathbf{A})$ denote the determinant and trace of a matrix $\mathbf{A}$, respectively, (2.1) can be expressed as an infinite-order vector autoregression process, i.e:

$$\mathbf{y}_t = \sum_{i=1}^{\infty} \boldsymbol{\pi}_i \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t, \tag{2.2}$$

where $\boldsymbol{\pi}_i$'s are VAR coefficient matrices satisfying $\sum_{i=1}^{\infty} \|\boldsymbol{\pi}_i\| < \infty$. We note that any stationary invertible finite-order ARMA(p, q) models are included as a special case of model (2.2).

For the purpose of forecasting, let $\mathbf{y}_{t+h}$ be the future value of $\mathbf{y}$ at time $t + h$. It is known that the *minimum MSE predictor* for the h-step ahead forecast of $\mathbf{y}_{t+h}$ at origin t is the conditional expectation $E(\mathbf{y}_{t+h}|\mathcal{F}_t) \equiv \mathbf{y}^*_{t+h|t}$, where $\mathcal{F}_t = \sigma(\mathbf{y}_s : s \leq t)$ denotes the σ-algebra built from the past of the process $\{\mathbf{y}_s\}_{s \leq t}$, representing the information up to time t. We then consider the linear h-step ahead forecast of $\mathbf{y}_{t+h}$ by employing an approximating K-dimensional vector autoregressive (VAR) model of the finite-order p fitted to a realization $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$ of length T. Specifically, let $\mathbf{y}_{t+h|t}(p)$ denote the *minimum MSE linear predictor* of $\mathbf{y}_{t+h}$ based on $\mathcal{F}_t$:

$$\mathbf{y}_{t+h|t}(p) = \boldsymbol{\pi}_1(p)\mathbf{y}_{t+h-1|t} + \boldsymbol{\pi}_2(p)\mathbf{y}_{t+h-2|t} + \cdots + \boldsymbol{\pi}_p(p)\mathbf{y}_{t+h-p|t}, \tag{2.3}$$

where $\mathbf{y}_{t+j|t} = \mathbf{y}_{t+j}$ for $j \leq 0$, and $\boldsymbol{\pi}_i(p)$, $i = 1, \ldots, p$, are $K \times K$ autoregressive coefficient matrices for a VAR(p) model.

In matrix notation, for the one-step ahead forecast (i.e., $h = 1$), we have:

$$\mathbf{Y} = \mathbf{Z}(p)\mathbf{\Pi}(p) + \boldsymbol{\varepsilon}(p), \tag{2.4}$$

where $\mathbf{Y} = (\mathbf{Y}_1\, \mathbf{Y}_2\, \cdots\, \mathbf{Y}_K)$ is the $(T-p)\times K$ matrix with $\mathbf{Y}_k = (y_{k,p+1}, \ldots, y_{kT})'$ being the $(T-p)\times 1$ vector of observations on the k-th equation of the VAR(p) system, $\mathbf{Z}(p)$ is the $(T-p)\times m$ matrix with $m = Kp$ and the $(t-p+1)$-th row given by $\mathbf{z}_t(p)' = (\mathbf{y}_t', \ldots, \mathbf{y}_{t-p+1}')$ for $t = p, \ldots, T-1$, $\mathbf{\Pi}(p)' = (\boldsymbol{\pi}_1(p), \ldots, \boldsymbol{\pi}_p(p))$ is the $K\times m$ coefficient matrix, and $\boldsymbol{\varepsilon}(p)$ is the $(T-p)\times K$ matrix with k-th column being $\boldsymbol{\varepsilon}_k(p) = (\varepsilon_{k,p+1}, \ldots, \varepsilon_{kT})'$.

We use the OLS method to estimate the VAR(p) model. It is known that, as first shown by Zellner (1962), the OLS and generalized least squares (GLS) methods produce the same estimates when applied to a VAR(p) model as every equation in a VAR(p) model contains the same set of right-hand-side variables $\mathbf{Z}(p)$. Specifically, the OLS estimator $\widehat{\mathbf{\Pi}}(p) = (\widehat{\boldsymbol{\pi}}_1(p), \ldots, \widehat{\boldsymbol{\pi}}_p(p))'$ is given by:

$$\widehat{\mathbf{\Pi}}(p) = (\mathbf{Z}(p)'\mathbf{Z}(p))^{-1}\mathbf{Z}(p)'\mathbf{Y}.$$

Using estimated parameters $\widehat{\boldsymbol{\pi}}_i(p)$ in a fitted (one-step) VAR(p) model with unknown future values replaced with their own forecasts, the h-step ahead predictor of $\mathbf{y}_{t+h}$ at the origin t can be iteratively computed as follows:

$$\widehat{\mathbf{y}}^I_{t+h|t}(p) = \sum_{i=1}^{p} \widehat{\boldsymbol{\pi}}_i(p)\widehat{\mathbf{y}}^I_{t+h-i|t}(p), \tag{2.5}$$

where $\widehat{\mathbf{y}}^I_{t+j|t}(p) = \mathbf{y}_{t+j}$ if $j \le 0$, and the superscript "I" indicates indirect multi-step forecasts. The direct h-step ahead predictor of $\mathbf{y}_{t+h}$ based on a fitted h-step VAR(p) model will be discussed in Section 4.

In practice one must determine the lag length p to proceed with VAR forecasting.[3] Two approaches to select the lag order p have been studied in the literature. The first approach uses the sequential likelihood ratio tests suggested by Tiao and Box (1981). The second approach selects the VAR order based on information criteria. We briefly discuss the information criterion method for one-step ahead forecasting. We specifically let:

$$\widehat{\mathbf{\Sigma}}(p) = \frac{1}{T-\bar{p}}\sum_{t=\bar{p}}^{T-1} \widehat{\boldsymbol{\varepsilon}}_{t+1}(p)\widehat{\boldsymbol{\varepsilon}}_{t+1}(p)' \tag{2.6}$$

be the residual covariance matrix without the adjustment for degrees of freedom from a VAR(p) model, where $\widehat{\boldsymbol{\varepsilon}}_{t+1}(p)'$, $t = p, \ldots, T-1$, are the $1\times K$ row vectors of the OLS residual matrix $\widehat{\boldsymbol{\varepsilon}}(p) = \mathbf{Y} - \mathbf{Z}(p)\widehat{\mathbf{\Pi}}(p)$.

Based on $\widehat{\mathbf{\Sigma}}(p)$, the three commonly used selection criteria for the VAR model are AIC,

[3]For the VAR(p) model, we follow common practice that all lags are included up to the lag order p, i.e., no gaps in the lags are allowed. One may instead want to allow for more flexible lag structures in the considered models based on, for example, economic-theoretic considerations (e.g., Cho and Moreno (2006)) or statistical evidence, in which case GLS is preferable to OLS for efficiency gains in the estimation. The GLS-based VAR model averaging is beyond the scope of the present paper and is left to future research.

BIC, and Hannan-Quinn (HQ) under the following form:

$$\text{AIC}(p) = \ln(\det(\widehat{\mathbf{\Sigma}}(p))) + 2pK^2/T, \tag{2.7}$$

$$\text{BIC}(p) = \ln(\det(\widehat{\mathbf{\Sigma}}(p))) + pK^2 \ln T/T, \tag{2.8}$$

$$\text{HQ}(p) = \ln(\det(\widehat{\mathbf{\Sigma}}(p))) + 2pK^2 \ln\ln T/T, \tag{2.9}$$

where pK^2 is the total model-dependent number of coefficients of lagged variables in the VAR(p) model. As a standard, the second components on the right-hand side of equations (2.7)-(2.9) are introduced to penalize large models.

3 Iterative multi-step VAR forecasting averaging

This section proposes a new MMMA criterion for one-step ahead VAR forecast averaging based on a set of VAR candidate models fitted to the single period horizon, i.e., $h = 1$. The averaging multi-step forecasts are then obtained by iterating forward for multiple periods. We defer the discussion of direct multi-step VAR forecasting averaging using the idea of leave-h-out cross-validation to Section 4.

Consider the following multivariate Mallows criterion for VAR model selection:

$$C_T(p) = (T - \bar{p}) \cdot \text{tr}\left(\widetilde{\mathbf{\Sigma}}(\bar{p})^{-1}\widehat{\mathbf{\Sigma}}(p)\right) + 2pK^2, \tag{3.1}$$

where $\widehat{\mathbf{\Sigma}}(p)$ is given by (2.6) and:

$$\widetilde{\mathbf{\Sigma}}(\bar{p}) = \frac{1}{T - \bar{p} - \bar{m}} \sum_{t=\bar{p}}^{T-1} \widehat{\boldsymbol{\varepsilon}}_{t+1}(\bar{p})\widehat{\boldsymbol{\varepsilon}}_{t+1}(\bar{p})' \tag{3.2}$$

is a bias-corrected residual covariance matrix from the largest model VAR($\bar{p}$) with $\bar{m} = K\bar{p}$. The multivariate Mallows selection criterion (3.1) has been employed by, e.g. Fujikoshi and Satoh (1997), who consider the (modified) AIC and the Mallows criterion for selecting multivariate linear regression models.

We note here that the sample is set to be of equal size across different candidate models for valid comparison. To be explicit, we fix the effective sample with $T - \bar{p}$ observations $(\mathbf{y}_{t+1}, \mathbf{z}_t(p))$, $t = \bar{p}, \ldots, T-1$, and then estimate all VAR(p) models and compute $\widehat{\mathbf{\Sigma}}(p)$ using the same $T - \bar{p}$ observations. Using this effective sample, $\mathbf{Z}(p)$ becomes a $(T - \bar{p}) \times m$ matrix with the $(t - \bar{p} + 1)$-th row given by $\mathbf{z}_t(p)' = (\mathbf{y}_t', \ldots, \mathbf{y}_{t-p+1}')$ for $t = \bar{p}, \ldots, T-1$.

As will be shown in Section 5.2, $C_T(p)$ is an approximately unbiased estimator of the trace of the MSFE matrix for a candidate model VAR(p). This means that the Mallows selection criterion $C_T(p)$ (3.1) is a sensible empirical criterion. The VAR forecasting selection problem is to choose the VAR lag order p among the candidate models $\{\text{VAR}(p)\}_{p=1}^{\bar{p}}$ for which the value of $C_T(p)$ is minimized.

We now turn to the method of VAR forecast averaging based on the Mallows criterion, which is the first main focus of the present paper. To begin with, let $\mathbf{w} = (w(1), \ldots, w(\bar{p}))'$ be the weight vector associated with candidate models, and $\widehat{\mathbf{\Pi}}(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\overline{\mathbf{\Pi}}(p)$ is the weighted VAR coefficient matrix, where $\overline{\mathbf{\Pi}}(p)$ is a $\bar{m} \times K$ matrix satisfying that for the (i,j)-th element $\overline{\mathbf{\Pi}}_{ij}(p) = \widehat{\mathbf{\Pi}}_{ij}(p)$ for $1 \leq i \leq Kp$ and $1 \leq j \leq K$, and $\overline{\mathbf{\Pi}}_{ij}(p) = 0$ elsewhere, i.e., $\overline{\mathbf{\Pi}}(p)' = \left(\widehat{\mathbf{\Pi}}(p)' \ \mathbf{0}_{K \times K(\bar{p}-p)}\right)$, where $\mathbf{0}_{r \times s}$ is a $r \times s$ zero matrix. Note that the combination residuals $\widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w})$ can be expressed as:

$$\begin{aligned}
\widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w}) &= \mathbf{y}_{t+1} - \widehat{\mathbf{\Pi}}(\mathbf{w})'\mathbf{z}_t(\bar{p}) \\
&= \mathbf{y}_{t+1} - \sum_{p=1}^{\bar{p}} w(p)\overline{\mathbf{\Pi}}(p)'\mathbf{z}_t(\bar{p}) \\
&= \sum_{p=1}^{\bar{p}} w(p)\left(\mathbf{y}_{t+1} - \widehat{\mathbf{\Pi}}(p)'\mathbf{z}_t(p)\right) \\
&= \sum_{p=1}^{\bar{p}} w(p)\widehat{\boldsymbol{\varepsilon}}_{t+1}(p), \qquad (3.3)
\end{aligned}$$

where for the third equality we assume $\sum_{p=1}^{\bar{p}} w(p) = 1$. To simplify the exposition, by an abuse of notation, we use the same function as $\widehat{\boldsymbol{\varepsilon}}_{t+1}(p)$ to denote $\widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w})$ in (3.3), where the latter is a function of the weight vector $\mathbf{w}$. This type of abuse of notation appears in several places throughout the paper when it comes to a weighted version of a given function f of the form $f(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)f(p)$.

Based on (3.1) and (3.3), the proposed multivariate Mallow criterion for model averaging takes the following form:

$$\begin{aligned}
C_T(\mathbf{w}) &= (T - \bar{p}) \cdot \mathrm{tr}\left(\widetilde{\mathbf{\Sigma}}(\bar{p})^{-1}\widehat{\mathbf{\Sigma}}(\mathbf{w})\right) + 2\sum_{p=1}^{\bar{p}} w(p)pK^2 \\
&= (T - \bar{p}) \cdot \mathrm{tr}\left(\widetilde{\mathbf{\Sigma}}(\bar{p})^{-1}\widehat{\mathbf{\Sigma}}(\mathbf{w})\right) + 2K^2\mathbf{p}'\mathbf{w}, \qquad (3.4)
\end{aligned}$$

where

$$\widehat{\mathbf{\Sigma}}(\mathbf{w}) = \frac{1}{T - \bar{p}} \sum_{t=\bar{p}}^{T-1} \widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w})\widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w})',$$

$\mathbf{p} = (1, \ldots, \bar{p})'$, and we write $\sum_{p=1}^{\bar{p}} w(p)p = \mathbf{p}'\mathbf{w}$.

Consider the first term of the right-hand side of (3.4):

$$
\begin{aligned}
(T-\bar{p})\cdot \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\widehat{\boldsymbol{\Sigma}}(\mathbf{w})\right) &= \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\sum_{t=\bar{p}}^{T-1}\widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w})\widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w})'\right) \\
&= \sum_{t=\bar{p}}^{T-1}\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\left(\sum_{p=1}^{\bar{p}}w(p)\widehat{\boldsymbol{\varepsilon}}_{t+1}(p)\right)\left(\sum_{p=1}^{\bar{p}}w(p)\widehat{\boldsymbol{\varepsilon}}_{t+1}(p)\right)'\right) \\
&= \sum_{t=\bar{p}}^{T-1}\sum_{i=1}^{\bar{p}}\sum_{j=1}^{\bar{p}}\widetilde{\varepsilon}_{t+1,ij}w(i)w(j) \\
&= \mathbf{w}'\widehat{\mathbf{S}}\mathbf{w}, \qquad (3.5)
\end{aligned}
$$

where $\widehat{\mathbf{S}}$ is a $\bar{p}\times\bar{p}$ matrix whose (i,j)-th element is $\widehat{S}_{ij}=\sum_{t=\bar{p}}^{T-1}\widetilde{\varepsilon}_{t+1,ij}$ with:

$$
\widetilde{\varepsilon}_{t+1,ij}=\sum_{k=1}^{K}\sum_{\ell=1}^{K}\widetilde{\sigma}_{k\ell}\widehat{\varepsilon}_{k,t+1}(i)\widehat{\varepsilon}_{\ell,t+1}(j),
$$

and $\widetilde{\sigma}_{k\ell}$ is the (k,ℓ)-th component of the matrix $\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}$.

For the fourth equality in (3.5) and an alternative expression of $\widehat{\mathbf{S}}$, please see Appendix A1. As also shown in Appendix A1, in the univariate case (i.e., $K=1$) the multivariate Mallows averaging criterion (3.4) reduces to the single-equation Mallows criterion as considered by Hansen (2007, 2008).

Equation (3.5) shows that the term $(T-\bar{p})\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\widehat{\boldsymbol{\Sigma}}(\mathbf{w})\right)$ has a quadratic form, leading the $C_T(\mathbf{w})$ criterion to be linear-quadratic in $\mathbf{w}$:

$$
C_T(\mathbf{w})=\mathbf{w}'\widehat{\mathbf{S}}\mathbf{w}+2K^2\mathbf{p}'\mathbf{w}. \qquad (3.6)
$$

The Mallows weight vector $\widehat{\mathbf{w}}$ is defined by:

$$
\widehat{\mathbf{w}}=\arg\min_{\mathbf{w}\in\mathcal{H}_T}C_T(\mathbf{w}), \qquad (3.7)
$$

where $\mathcal{H}_T$ is a unit simplex of $\mathbb{R}^{\bar{p}_T}$ that allows the weights to be continuous - namely:

$$
\mathcal{H}_T=\left\{\mathbf{w}\in[0,1]^{\bar{p}_T}:\sum_{p=1}^{\bar{p}_T}w(p)=1\right\}.
$$

The quadratic programming problem in (3.7) can be solved via several widely used statistical programming software, such as the Guass function "qprog", Matlab function "quadprog", and R package "quadprog".

Note that for the asymptotic analysis in Section 5, the maximum lag order $\bar{p}_T$ is allowed to increase to infinity with the sample size T. The rate at which $\bar{p}_T$ grows with T will be

specified later. The averaging iterative h-step ahead forecast at origin t based on Mallows weights $\widehat{\mathbf{w}}$ is obtained by:

$$\widehat{\mathbf{y}}^{I}_{t+h|t}(\widehat{\mathbf{w}}) = \sum_{p=1}^{\bar{p}} \widehat{w}(p)\widehat{\mathbf{y}}^{I}_{t+h|t}(p), \tag{3.8}$$

where $\widehat{\mathbf{y}}^{I}_{t+h|t}(p)$ is given by (2.5).

4 Direct multi-step VAR forecasting averaging

This section proposes a VAR forecasting averaging procedure based on the MCVA$_h$ criterion for direct multi-step forecasting. We first note that the VAR(∞) model in (2.2) can be recursively represented as a h-step forecasting model:

$$\mathbf{y}_{t+h} = \sum_{i=1}^{\infty} \boldsymbol{\psi}_{hi}\mathbf{y}_{t-i+1} + \boldsymbol{\epsilon}_{t+h} \equiv \boldsymbol{\mu}^{h}_{t} + \boldsymbol{\epsilon}_{t+h}, \tag{4.1}$$

where $\boldsymbol{\mu}^{h}_{t} = (\mu^{h}_{1t}, \mu^{h}_{2t}, \ldots, \mu^{h}_{Kt})'$, $\boldsymbol{\psi}_{hi}$ is the VAR coefficient matrix in the linear least-squared predictor based on regressing $\mathbf{y}_{t+h}$ on the infinite past $\{\mathbf{y}_j\}_{j\leq t}$, and $\boldsymbol{\epsilon}_{t+h}$ is the associated h-step forecast error with $E(\boldsymbol{\epsilon}_{t+h}\boldsymbol{\epsilon}'_{t+h}) = \boldsymbol{\Sigma}_h$ and is combined with (2.1) to be expressed as $\boldsymbol{\epsilon}_{t+h} = \sum_{i=0}^{h-1} \boldsymbol{\Phi}_i\boldsymbol{\varepsilon}_{t+h-i}$, which is known to follow a moving average process of order $h-1$.

The direct h-step forecast for K-dimensional time series can be generated from the following h-step ahead VAR(p) forecasting model:

$$\mathbf{y}_{t+h} = \boldsymbol{\psi}_{h1}(p)\mathbf{y}_t + \boldsymbol{\psi}_{h2}(p)\mathbf{y}_{t-1} + \cdots + \boldsymbol{\psi}_{hp}(p)\mathbf{y}_{t-p+1} + \boldsymbol{\epsilon}_{t+h}(p), \tag{4.2}$$

where the dependent variable $\mathbf{y}_{t+h}$ is the h-step ahead value being forecasted and $\boldsymbol{\epsilon}_{t+h}(p) = \sum_{i=p+1}^{\infty} \boldsymbol{\psi}_{hi}\mathbf{y}_{t-i+1} + \boldsymbol{\epsilon}_{t+h}$. The subscript h in (4.2) reflects the fact that, in contrast to Section 3 where the iterated forecasts are made using a one-step ahead VAR(p) model and then iterated forward, a separate VAR(p) model is fitted here for each forecast horizon h.

Let $\boldsymbol{\mu}_h = (\boldsymbol{\mu}^{h}_{\bar{p}}, \ldots, \boldsymbol{\mu}^{h}_{T-h})'$. In matrix notation, for the full effective sample $\{\mathbf{y}_{t+h}, \mathbf{z}_t(p)\}_{t=\bar{p}}^{T-h}$, we can write $\mathbf{Y}_h = \boldsymbol{\mu}_h + \mathbf{e}_h$ and $\mathbf{Y}_h = \mathbf{Z}_h(p)\boldsymbol{\Psi}_h(p) + \mathbf{e}_h(p)$, where $\mathbf{Y}_h = (\mathbf{Y}^{h}_1\, \mathbf{Y}^{h}_2 \ldots \mathbf{Y}^{h}_K)$ is the $(T-\bar{p}-h+1)\times K$ matrix with $\mathbf{Y}^{h}_k = (y_{k,\bar{p}+h}, \ldots, y_{kT})'$, $\mathbf{e}_h = (\boldsymbol{\epsilon}_{\bar{p}+h}, \ldots, \boldsymbol{\epsilon}_T)'$, $\mathbf{Z}_h(p)$ is the $(T-\bar{p}-h+1)\times m$ matrix with the $(t-\bar{p}+1)$-th row given by $\mathbf{z}_t(p)' = (\mathbf{y}'_t, \ldots, \mathbf{y}'_{t-p+1})$ for $t = \bar{p}, \ldots, T-h$, $\boldsymbol{\Psi}_h(p)' = (\boldsymbol{\psi}_{h1}(p), \ldots, \boldsymbol{\psi}_{hp}(p))$ is a $K\times m$ coefficient matrix, and $\mathbf{e}_h(p) = (\boldsymbol{\epsilon}_{\bar{p}+h}(p), \ldots, \boldsymbol{\epsilon}_T(p))'$. The full-sample OLS coefficient estimate $\widehat{\boldsymbol{\Psi}}_h(p)$ of $\boldsymbol{\Psi}_h(p)$ is given by $\widehat{\boldsymbol{\Psi}}_h(p) = (\mathbf{Z}_h(p)'\mathbf{Z}_h(p))^{-1}\mathbf{Z}_h(p)'\mathbf{Y}_h$, and the resulting residual matrix is $\widehat{\mathbf{e}}_h(p) = \mathbf{Y}_h - \mathbf{Z}_h(p)\widehat{\boldsymbol{\Psi}}_h(p)$. The direct h-step ahead forecast using the fitted h-step VAR(p) model is then formed by $\widehat{\mathbf{y}}_{t+h|t}(p) = \widehat{\boldsymbol{\Psi}}_h(p)'\mathbf{z}_t(p)$.

We now introduce more notation for the leave-h-out OLS estimation for the construction of the MCVA$_h$ criterion. For a particular observation t ($t = \bar{p}, \ldots, T-h$) and forecast

horizon h, we denote by $\underline{\ell}_{ht} = \max(\bar{p}, t-(h-1))$ and by $\bar{\ell}_{ht} = \min(t+(h-1), T-h)$ the left- and right-end points of the observation window that is deleted, respectively, and hence $\ell_{ht} = \bar{\ell}_{ht} - \underline{\ell}_{ht} + 1$ is the number of observations deleted. Note that $\ell_{ht} = 2h-1$ for $\bar{p}+h-1 \leq t \leq T-2h+1$. We also denote by $\ell_h = \sum_{t=\bar{p}}^{T-h} \ell_{ht}$ the total number of observations deleted. Taking $h=2$ for example, for the first ($t=\bar{p}$) and second ($t=\bar{p}+1$) observations in the effective sample, their corresponding deleted observation windows have size $\ell_{ht}=2$ (from $\underline{\ell}_{ht}=\bar{p}$ to $\bar{\ell}_{ht}=\bar{p}+1$) and $\ell_{ht}=3$ (from $\underline{\ell}_{ht}=\bar{p}$ to $\bar{\ell}_{ht}=\bar{p}+2$), respectively.

For series k, denote by $\widetilde{\epsilon}_{k,t+h}(p)$ the OLS residual from the regression of $y_{k,t+h}$ on $\mathbf{z}_t(p)$ with ℓ_{ht} observations $\{y_{k,j+h}, \mathbf{z}_j(p)\}_{j=\underline{\ell}_{ht}}^{\bar{\ell}_{ht}}$ deleted. Let $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p) = (\widetilde{\epsilon}_{1,t+h}(p), \widetilde{\epsilon}_{2,t+h}(p), \ldots, \widetilde{\epsilon}_{K,t+h}(p))'$, which is obtained by $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p) = \mathbf{y}_{t+h} - \widetilde{\boldsymbol{\Psi}}_{h,t}(p)'\mathbf{z}_t(p)$, where:

$$\widetilde{\boldsymbol{\Psi}}_{h,t}(p) = (\widetilde{\mathbf{Z}}_{h,t}(p)'\widetilde{\mathbf{Z}}_{h,t}(p))^{-1}\widetilde{\mathbf{Z}}_{h,t}(p)'\widetilde{\mathbf{Y}}_{h,t}$$

is the $m \times K$ matrix of the leave-h-out OLS estimates of VAR(p) coefficients for observation t, and $\widetilde{\mathbf{Z}}_{h,t}(p)$ and $\widetilde{\mathbf{Y}}_{h,t}$ are the resulting data matrices with ℓ_{ht} observations removed from $\mathbf{Z}_h(p)$ and $\mathbf{Y}_h$, respectively. The direct h-step ahead forecast of $\mathbf{y}_{t+h}$ at origin t based on the h-step VAR(p) model fitted by the leave-h-out approach is formed by:

$$\widetilde{\mathbf{y}}_{t+h|t}^{D}(p) = \widetilde{\boldsymbol{\Psi}}_{h,t}(p)'\mathbf{z}_t(p). \tag{4.3}$$

Similar to (3.3), let $\widetilde{\boldsymbol{\epsilon}}_{t+h}(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)$ denote the weighted average of the leave-h-out residuals. The MCVA$_h$ criterion for the direct h-step forecast combination is proposed as follows:

$$\begin{aligned}
CV_{T,h}(\mathbf{w}) &= (T-\bar{p}-h+1)\cdot \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\widetilde{\boldsymbol{\Sigma}}_h(\mathbf{w})\right) \\
&= \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\sum_{t=\bar{p}}^{T-h}\widetilde{\boldsymbol{\epsilon}}_{t+h}(\mathbf{w})\widetilde{\boldsymbol{\epsilon}}_{t+h}(\mathbf{w})'\right) \\
&= \sum_{t=\bar{p}}^{T-h}\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\left(\sum_{p=1}^{\bar{p}}w(p)\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)\right)\left(\sum_{p=1}^{\bar{p}}w(p)\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)\right)'\right) \\
&= \sum_{t=\bar{p}}^{T-h}\sum_{i=1}^{\bar{p}}\sum_{j=1}^{\bar{p}}\widetilde{\epsilon}_{t+h,ij}w(i)w(j) \\
&= \mathbf{w}'\widetilde{\mathbf{S}}_h\mathbf{w},
\end{aligned} \tag{4.4}$$

where

$$\widetilde{\boldsymbol{\Sigma}}_h(\bar{p}) = \frac{1}{T-\bar{p}-h-\bar{m}+1}\sum_{t=\bar{p}}^{T-h}\widetilde{\boldsymbol{\epsilon}}_{t+h}(\bar{p})\widetilde{\boldsymbol{\epsilon}}_{t+h}(\bar{p})', \tag{4.5}$$

and $\widetilde{\mathbf{S}}_h$ is a $\bar{p} \times \bar{p}$ matrix with the (i,j)-th element $\widetilde{S}_{hij} = \sum_{t=\bar{p}}^{T-h} \widetilde{\epsilon}_{t+h,ij}$,

$$\widetilde{\epsilon}_{t+h,ij} = \sum_{k=1}^{K} \sum_{\ell=1}^{K} \widetilde{\sigma}^{h}_{k\ell} \widetilde{\epsilon}_{k,t+h}(i) \widetilde{\epsilon}_{\ell,t+h}(j),$$

and $\widetilde{\sigma}^{h}_{k\ell}$ is the (k,ℓ)-th element of $\widetilde{\mathbf{\Sigma}}_h(\bar{p})^{-1}$.

The estimated MCVA$_h$ weight vector for direct VAR forecast averaging is defined by:

$$\widehat{\mathbf{w}}_{cv,h} = \arg \min_{\mathbf{w} \in \mathcal{H}_T} CV_{T,h}(\mathbf{w}), \tag{4.6}$$

where $\mathcal{H}_T$ is defined as before. The estimated weight vector $\widehat{\mathbf{w}}_{cv,h}$ is indexed by h to reflect the fact that $\widehat{\mathbf{w}}_{cv,h}$ is selected anew for each h by minimizing the MCVA$_h$ criterion. Similar to (3.7), (4.6) takes the form of quadratic programming problems and therefore can be solved by using the same statistical programming packages/functions as those mentioned in Section 3, without the need to specify the linear component of the criterion.

The resulting averaging direct h-step ahead forecast at origin t based on leave-h-out cross-validation is produced by:

$$\widetilde{\mathbf{y}}^{D}_{t+h|t}(\widehat{\mathbf{w}}_{cv,h}) = \sum_{p=1}^{\bar{p}} \widehat{w}_{cv,h}(p) \widetilde{\mathbf{y}}^{D}_{t+h|t}(p), \tag{4.7}$$

where $\widehat{\mathbf{w}}_{cv,h} = (\widehat{w}_{cv,h}(1), \ldots, \widehat{w}_{cv,h}(\bar{p}))'$ and $\widetilde{\mathbf{y}}^{D}_{t+h|t}(p)$ is given by (4.3).

Efficient computation of $CV_{T,h}(\mathbf{w})$. Computing the $CV_{T,h}(\mathbf{w})$ criterion is known to be computationally expensive; specifically, its computation is on the order of T^2, i.e. $O(T^2)$. This is because by using the standard approach, we need to compute $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p) = \mathbf{y}_{t+h} - \widetilde{\mathbf{\Psi}}_{h,t}(p)'\mathbf{z}_t(p)$, and $\widetilde{\mathbf{\Psi}}_{h,t}(p)$ is of order $O(T)$. We now discuss about how to efficiently compute the leave-h-out residual vector $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)$. First note that $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)$ is the $\min(t-\bar{p}+1, h)$-th row of the $\ell_{ht} \times K$ removed leave-h-out residual matrix, denoted by $\widetilde{\mathbf{e}}_{t:h}(p)$.

A computationally convenient formula for $\widetilde{\mathbf{e}}_{t:h}(p)$ can be derived as follows:

$$\widetilde{\mathbf{e}}_{t:h}(p) = \mathbf{Y}_{t:h} - \mathbf{Z}_{t:h}(p)\widetilde{\mathbf{\Psi}}_{h,t}(p) = \left(\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p)\right)^{-1} \widehat{\mathbf{e}}_{t:h}(p), \tag{4.8}$$

where $\mathbf{Y}_{t:h}$ and $\mathbf{Z}_{t:h}(p)$ are $\ell_{ht} \times K$ and $\ell_{ht} \times m$ block matrices for the removed observations: $\underline{\ell}_{ht}, \ldots, t, \ldots, \bar{\ell}_{ht}$ in $\mathbf{Y}$ and $\mathbf{Z}(p)$, respectively, $\mathbf{P}_{t:h}(p) = \mathbf{Z}_{t:h}(p)(\mathbf{Z}_h(p)'\mathbf{Z}_h(p))^{-1}\mathbf{Z}_{t:h}(p)'$, and $\widehat{\mathbf{e}}_{t:h}(p)$ is the $\ell_{ht} \times K$ block matrix of $\widehat{\mathbf{e}}_h(p)$ for the removed observations. The second equality

in (4.8) follows from using the following formula for $\widetilde{\mathbf{\Psi}}_{h,t}(p)$:

$$\begin{aligned}\widetilde{\mathbf{\Psi}}_{h,t}(p) &= \left(\mathbf{Z}_h(p)'\mathbf{Z}_h(p) - \mathbf{Z}_{t:h}(p)'\mathbf{Z}_{t:h}(p)\right)^{-1}\left(\mathbf{Z}_h(p)'\mathbf{Y}_h - \mathbf{Z}_{t:h}(p)'\mathbf{Y}_{t:h}\right) \\ &= \widehat{\mathbf{\Psi}}_h(p) - \left(\mathbf{Z}_h(p)'\mathbf{Z}_h(p)\right)^{-1}\mathbf{Z}_{t:h}(p)'\Big(\mathbf{Y}_{t:h} - \left(\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p)\right)^{-1}\mathbf{Z}_{t:h}(p)\left(\mathbf{Z}_h(p)'\mathbf{Z}_h(p)\right)^{-1}\mathbf{Z}_h(p)'\mathbf{Y}_h \\ &\quad + \left(\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p)\right)^{-1}\mathbf{P}_{t:h}(p)\mathbf{Y}_{t:h}\Big). \qquad (4.9)\end{aligned}$$

Formulae (4.8) and (4.9) are derived by directly applying the arguments in Racine (1997) and Hansen (2010) to our VAR setting. The detailed derivations are omitted here and are available upon request from the authors. Using (4.8), it is not necessary to actually fit $T-\bar{p}-h+1$ separate models when computing the $CV_{T,h}(\mathbf{w})$ criterion and as a result, the computation of the $CV_{T,h}(\mathbf{w})$ criterion is of order $O(T)$ instead of $O(T^2)$.

Let $\mathbf{P}_h(p) = \mathbf{Z}_h(p)(\mathbf{Z}_h(p)'\mathbf{Z}_h(p))^{-1}\mathbf{Z}_h(p)'$ be the regular $(T-\bar{p}-h+1)\times(T-\bar{p}-h+1)$ projection matrix to the subspace spanned by the columns of $\mathbf{Z}_h(p)$. We next wish to examine the relationship between $\mathbf{P}_h(p)$ and its leave-h-out version, denoted by $\widetilde{\mathbf{P}}_h(p)$. This relationship will be used in several places in the proof for the asymptotic optimality of our MCVA$_h$ procedure, as will be shown in Section A5 in Appendix. We first need to develop some notation.

Denote by $\mathbf{S}_{t:h}$ the $\ell_{ht}\times(T-\bar{p}-h+1)$ selection matrix with a $\ell_{ht}\times\ell_{ht}$ block matrix equal to $\mathbf{I}_{\ell_{ht}}$ and 0 elsewhere - namely, for a particular t, matrix $\mathbf{S}_{t:h}$ is used to extract the block matrix corresponding to ℓ_{ht} removed observations. For example, $\widehat{\mathbf{e}}_{t:h}(p)$ in (4.8) can be taken from $\widehat{\mathbf{e}}_h(p)$ by using $\widehat{\mathbf{e}}_{t:h}(p) = \mathbf{S}_{t:h}\widehat{\mathbf{e}}_h(p)$. We also denote by e_{ht} the $\ell_{ht}\times 1$ selection vector with 1 in its $\min(t-\bar{p}+1,h)$-th element and 0 elsewhere. To be more explicit, $\mathbf{S}_{t:h} = \left(\mathbf{0}_{\ell_{ht}\times(\underline{\ell}_{ht}-\bar{p})}\ \mathbf{I}_{\ell_{ht}}\ \mathbf{0}_{\ell_{ht}\times(T-h-\bar{\ell}_{ht})}\right)$ if $\underline{\ell}_{ht}-\bar{p}>0$; $\mathbf{S}_{t:h} = \left(\mathbf{I}_{\ell_{ht}}\ \mathbf{0}_{\ell_{ht}\times(T-h-\bar{\ell}_{ht})}\right)$ if $\underline{\ell}_{ht}-\bar{p}=0$; and $e_{ht} = (\mathbf{0}_{1\times(\min(t-\bar{p}+1,h)-1)}, 1, \mathbf{0}_{1\times(\ell_{ht}-\min(t-\bar{p}+1,h))})'$.

Using the selection matrix $\mathbf{S}_{t:h}$, $\widetilde{\mathbf{e}}_{t:h}(p)$ in (4.8) can be equivalently rewritten as: $\mathbf{S}_{t:h}(\mathbf{Y}_h - \widetilde{\mathbf{P}}_h(p)\mathbf{Y}_h) = (\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p))^{-1}\mathbf{S}_{t:h}(\mathbf{Y}_h - \mathbf{P}_h(p)\mathbf{Y}_h)$. Cancelling out $\mathbf{Y}_h$ on both sides of the above equation and then rearranging yield $\mathbf{S}_{t:h}\widetilde{\mathbf{P}}_h(p) = (\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p))^{-1}\mathbf{S}_{t:h}(\mathbf{P}_h(p) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{S}_{t:h}$. Denote $\widetilde{\mathbf{P}}_{t:h}(p) = \mathbf{S}_{t:h}\widetilde{\mathbf{P}}_h(p)$ and $\mathbf{D}_{t:h}(p) = (\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p))^{-1}\mathbf{S}_{t:h}$. Applying the selection vector e_{ht} to $\widetilde{\mathbf{P}}_{t:h}(p)$ gives the $(t-\bar{p}+1)$-th row of the leave-h-out projection matrix $\widetilde{\mathbf{P}}_h(p)$, i.e.:

$$e_{ht}'\widetilde{\mathbf{P}}_{t:h}(p) = e_{ht}'(\mathbf{D}_{t:h}(p)(\mathbf{P}_h(p) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{S}_{t:h}). \qquad (4.10)$$

Lastly, stacking (4.10) vertically over $t=\bar{p},\ldots,T-h$ results in $\widetilde{\mathbf{P}}_h(p)$, as stated in Lemma 1 below.

For the presentation of Lemma 1, we denote by $\mathbf{E}_h$ the $(T-\bar{p}-h+1)\times\ell_h$ matrix with the $(t-\bar{p}+1)$-th row that is formed by e_{ht}' as its $\sum_{i=\bar{p}-1}^{t-1}(\ell_{hi}+1),\ldots,\sum_{i=\bar{p}-1}^{t}\ell_{hi}$ column row subvector and 0 elsewhere, and with $\ell_{h,\bar{p}-1}$ set to 0. We also denote by $\mathbf{D}_h(p)$ and $\mathbf{S}_h$ the $\ell_h\times(T-\bar{p}-h+1)$ matrices vertically stacking $(\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p))^{-1}\mathbf{S}_{t:h}$ and $\mathbf{S}_{t:h}$, respectively.

Lemma 1. *The leave-h-out estimates $\widetilde{\boldsymbol{\mu}}_h(p)$ of $\boldsymbol{\mu}_h$ based on the fitted h-step VAR(p) model*

can be represented by $\widetilde{\boldsymbol{\mu}}_h(p) = \widetilde{\mathbf{P}}_h(p)\mathbf{Y}_h$, *where* $\widetilde{\mathbf{P}}_h(p)$ *is related to* $\mathbf{P}_h(p)$ *as follows:*

$$\widetilde{\mathbf{P}}_h(p) = \mathbf{E}_h \left(\mathbf{D}_h(p)(\mathbf{P}_h(p) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{S}_h\right). \tag{4.11}$$

Alternatively, (4.11) *can also be expressed as:*

$$\widetilde{\mathbf{P}}_h(p) = \widetilde{\mathbf{D}}_h(p)(\mathbf{P}_h(p) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{I}_{T-\bar{p}-h+1}, \tag{4.12}$$

where we use the fact that $\mathbf{E}_h\mathbf{S}_h = \mathbf{I}_{T-\bar{p}-h+1}$ *and denote* $\widetilde{\mathbf{D}}_h(p) = \mathbf{E}_h\mathbf{D}_h(p)$.

Lemma 1 generalizes to $h > 1$ for the projection matrix based on leave-h-out cross-validation. To see this, in an important special case when $h = 1$ (corresponding to leave-one-out or Jackknife cross-validation), let $q_{ij}(p)$ denote the (i, j)-th element of the one-step projection matrix, denoted by $\mathbf{P}(p)$. In this particular case, we have $\ell_{ht} = 1$ for all t, $\ell_h = T - \bar{p}$, and the matrices $\mathbf{E}_h$, $\mathbf{D}_h(p)$, and $\mathbf{S}_h$ in (4.11) become $\mathbf{I}_{T-\bar{p}}$, the diagonal matrix $\mathbf{D}(p)$ of dimension $(T - \bar{p})$ with the i-th diagonal element equal to $(1 - q_{ii}(p))^{-1}$, and $\mathbf{I}_{T-\bar{p}}$, respectively. As a consequence, (4.11) reduces to equation (1.4) of Li (1987): $\widetilde{\mathbf{P}}(p) = \mathbf{D}(p)(\mathbf{P}(p) - \mathbf{I}_{T-\bar{p}}) + \mathbf{I}_{T-\bar{p}}$.

5 Asymptotic theory

This section provides theoretical justifications of our methods. In Sections 5.1 and 5.2, we discuss the relation of the proposed MMMA and MCVA$_h$ criteria to MSE and MSFE in multivariate settings and establish the property that the proposed averaging criteria are approximately unbiased estimators of the underlying risk functions, including in-sample MSE and out-of-sample MSFE. Section 5.3 provides the conditions under which the proposed procedures are asymptotically optimal, in the sense that the averaged squared error evaluated at the empirical weights is asymptotically equivalent to the infeasible optimum.

5.1 MSFE of multi-step forecast averaging

Given the h-step ahead forecast $\widehat{\mathbf{y}}_{T+h|T}(p) = \widehat{\boldsymbol{\Psi}}_h(p)'\mathbf{z}_T(p)$ produced by a fitted direct h-step VAR(p) model using the full effective sample, the associated forecast error is:

$$\begin{aligned}\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h|T}(p) &= (\mathbf{y}_{T+h} - \mathbf{y}^*_{T+h|T}) + (\mathbf{y}^*_{T+h|T} - \widehat{\mathbf{y}}_{T+h|T}(p)) \\ &= \boldsymbol{\epsilon}_{T+h} + (\mathbf{y}^*_{T+h|T} - \widehat{\mathbf{y}}_{T+h|T}(p)).\end{aligned} \tag{5.1}$$

Here, recall that $\mathbf{y}^*_{T+h|T} \equiv E(\mathbf{y}_{T+h}|\mathcal{F}_T)$ is the optimal MSE predictor of $\mathbf{y}_{T+h}$.

In this paper we consider the trace of the standardized MSFE matrix as a scalar measure

of the risk matrix, or more explicitly:

$$\begin{aligned}
\mathrm{MSFE}_h(p) &= E\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h|T}(p))(\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h|T}(p))'\right)\right) \\
&= E\left(\mathrm{tr}(\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\epsilon}_{T+h}\boldsymbol{\epsilon}_{T+h}')\right) + E\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_T^h - \widehat{\boldsymbol{\mu}}_T^h(p))(\boldsymbol{\mu}_T^h - \widehat{\boldsymbol{\mu}}_T^h(p))'\right)\right) \\
&\simeq K + E\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_t^h - \widehat{\boldsymbol{\mu}}_t^h(p))(\boldsymbol{\mu}_t^h - \widehat{\boldsymbol{\mu}}_t^h(p))'\right)\right) \\
&= K + (T - \bar{p} - h + 1)E\left(\mathrm{tr}\left(\frac{1}{T - \bar{p} - h + 1}\boldsymbol{\Sigma}_h^{-1}\sum_{t=\bar{p}}^{T-h}(\boldsymbol{\mu}_t^h - \widehat{\boldsymbol{\mu}}_t^h(p))(\boldsymbol{\mu}_t^h - \widehat{\boldsymbol{\mu}}_t^h(p))'\right)\right) \\
&= K + (T - \bar{p} - h + 1)E\left(\mathrm{tr}\left(\frac{1}{T - \bar{p} - h + 1}\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h(p))'(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h(p))\right)\right) \\
&\equiv K + (T - \bar{p} - h + 1)E(L_{T,h}(p)), \qquad (5.2)
\end{aligned}$$

where $\widehat{\boldsymbol{\mu}}_h(p) = (\widehat{\boldsymbol{\mu}}_{\bar{p}}^h(p), \ldots, \widehat{\boldsymbol{\mu}}_{T-h}^h(p))' = \mathbf{P}_h(p)\mathbf{Y}_h$ is the matrix of fitted values of $\mathbf{Y}_h$ with $\widehat{\boldsymbol{\mu}}_t^h(p) = (\widehat{\mu}_{1t}^h(p), \widehat{\mu}_{2t}^h(p), \ldots, \widehat{\mu}_{Kt}^h(p))'$, the approximation follows from the assumed stationarity of $\mathbf{y}_t$, and $E(\mathrm{tr}(\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\epsilon}_{T+h}\boldsymbol{\epsilon}_{T+h}')) = \mathrm{tr}(\boldsymbol{\Sigma}_h^{-1}E(\boldsymbol{\epsilon}_{T+h}\boldsymbol{\epsilon}_{T+h}')) = \mathrm{tr}(\mathbf{I}_K) = K$. $L_{T,h}(p)$ in (5.2) is defined as the trace of the standardized in-sample average squared error - namely:

$$\begin{aligned}
L_{T,h}(p) &= \frac{1}{(T - \bar{p} - h + 1)}\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h(p))'(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h(p))\right) \\
&= \frac{1}{(T - \bar{p} - h + 1)}\sum_{t=\bar{p}}^{T-h}\sum_{\ell=1}^{K}\sum_{k=1}^{K}(\mu_{kt}^h - \widehat{\mu}_{kt}^h(p))\sigma_{k\ell}^h(\mu_{\ell t}^h - \widehat{\mu}_{\ell t}^h(p)), \qquad (5.3)
\end{aligned}$$

where $\sigma_{k\ell}^h$ is the (k,ℓ)-th entry of $\boldsymbol{\Sigma}_h^{-1}$, and $E(L_{T,h}(p))$ is the associated in-sample risk.

Equation (5.2) implies that $\mathrm{MSFE}_h(p)$ equals $E(L_{T,h}(p))$ up to an additive constant K as well as a multiplicative constant $(T - \bar{p} - h + 1)$. This relates the out-of-sample MSFE to in-sample MSE in the framework of VAR forecast selection. We then regard the selected model $\mathrm{VAR}(p^*)$ for which the risk measured by $\mathrm{MSFE}_h(p)$ in (5.2) is minimized over $p = 1, \ldots, \bar{p}$ as the best model.

It is worth emphasizing that $\mathrm{MSFE}_h(p)$ defined in (5.2) is standardized by the true error covariance matrix $\boldsymbol{\Sigma}_h$. This standardized version of the risk function has been used by, for example, Fujikoshi and Satoh (1997) and Yanagihara and Satoh (2010), in multivariate regression settings. Weighted risk by $\boldsymbol{\Sigma}_h^{-1}$ is in contrast to the single-equation model selection/averaging problem. The main motivation of the introduction of $\boldsymbol{\Sigma}_h^{-1}$ to the MSFE matrix is two-fold. First, weighted by $\boldsymbol{\Sigma}_h^{-1}$ makes the MSFE criterion scale-independent, so that the individual MSFE for each response variable is scaled to be of equal importance. Second, it incorporates the potential interrelationships among forecast errors and thus may make better use of the information the data contain, thereby likely improving forecast accuracy. This weighted MSFE is consistent with the use of the inverse of the estimated forecast error covariance matrices: $\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}$ in (3.4) and $\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}$ in (4.4).

For any forecast combination $\mathbf{w}$, the h-step ahead forecast combination is given by:

$$\widehat{\mathbf{y}}_{T+h|T}(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\widehat{\mathbf{y}}_{T+h|T}(p).$$

Denote $\widehat{\boldsymbol{\mu}}_h(\mathbf{w}) = (\widehat{\boldsymbol{\mu}}^h_{\bar{p}}(\mathbf{w}), \ldots, \widehat{\boldsymbol{\mu}}^h_{T-h}(\mathbf{w}))'$ with $\widehat{\boldsymbol{\mu}}^h_t(\mathbf{w}) = (\widehat{\mu}^h_{1t}(\mathbf{w}), \ldots, \widehat{\mu}^h_{Kt}(\mathbf{w}))'$ and $\widehat{\boldsymbol{\mu}}^h_{kt}(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\widehat{\boldsymbol{\mu}}^h_{kt}(p)$. Using the same arguments as those in (5.2), we denote by $\text{MSFE}_h(\mathbf{w})$ the trace of the associated standardized MSFE matrix of the h-step ahead forecast combination $\widehat{\mathbf{y}}_{T+h|T}(\mathbf{w})$:

$$\begin{aligned}
\text{MSFE}_h(\mathbf{w}) &= E\left(\text{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h|T}(\mathbf{w}))(\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h|T}(\mathbf{w}))'\right)\right) \\
&\simeq K + E\left(\text{tr}\left(\frac{1}{T-\bar{p}-h+1}\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h(\mathbf{w}))'(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h(\mathbf{w}))\right)\right) \\
&\equiv K + (T-\bar{p}-h+1)E(L_{T,h}(\mathbf{w})), \qquad (5.4)
\end{aligned}$$

where once again the approximation follows from the assumed stationarity of $\mathbf{y}_t$, and $L_{T,h}(\mathbf{w})$ is the in-sample average squared error from h-step ahead forecast combination:

$$\begin{aligned}
L_{T,h}(\mathbf{w}) &= \frac{1}{T-\bar{p}-h+1}\text{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h(\mathbf{w}))'(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h(\mathbf{w}))\right) \\
&= \frac{1}{(T-\bar{p}-h+1)}\sum_{t=\bar{p}}^{T-h}\sum_{\ell=1}^{K}\sum_{k=1}^{K}(\mu^h_{kt} - \widehat{\mu}^h_{kt}(\mathbf{w}))\sigma^h_{k\ell}(\mu^h_{\ell t} - \widehat{\mu}^h_{\ell t}(\mathbf{w})), \qquad (5.5)
\end{aligned}$$

and $E(L_{T,h}(\mathbf{w}))$ is the associated expected in-sample squared error.

5.2 Asymptotic unbiasedness

Let $L_T(\mathbf{w})$ be the in-sample average squared error from one-step ahead forecast averaging, as defined by (5.5) when $h = 1$. First, we wish to show the property that the $C_T(\mathbf{w})$ criterion is an asymptotically unbiased estimator of $L_T(\mathbf{w})$. For this case of $h = 1$, we simply remove the superscript/subscript h in corresponding notations defined before by denoting $\mathbf{P}(p) = \mathbf{Z}(p)(\mathbf{Z}(p)'\mathbf{Z}(p))^{-1}\mathbf{Z}(p)'$, $\widehat{\boldsymbol{\mu}}(p) = (\widehat{\boldsymbol{\mu}}_{\bar{p}}(p), \ldots, \widehat{\boldsymbol{\mu}}_{T-1}(p))' = \mathbf{P}(p)\mathbf{Y}$ with $\widehat{\boldsymbol{\mu}}_t(p) = (\widehat{\mu}_{1t}(p), \ldots, \widehat{\mu}_{Kt}(p))'$, $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\widehat{\boldsymbol{\mu}}(p)$, and $\text{MSFE}(\mathbf{w}) \equiv \text{MSFE}_h(\mathbf{w})$ defined in (5.4) when $h = 1$.

In matrix notation, let $\widehat{\mathbf{e}}(\mathbf{w}) = \mathbf{Y} - \widehat{\boldsymbol{\mu}}(\mathbf{w})$. We first calculate the first term on the right

side of $C_T(\mathbf{w})$ in (3.4), which can be rewritten as:

$$
\begin{aligned}
(T-\bar{p})\cdot \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\widehat{\boldsymbol{\Sigma}}(\mathbf{w})\right) &= \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\sum_{t=\bar{p}+1}^{T}\widehat{\boldsymbol{\varepsilon}}_t(\mathbf{w})\widehat{\boldsymbol{\varepsilon}}_t(\mathbf{w})'\right) \\
&= \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\widehat{\mathbf{e}}(\mathbf{w})'\widehat{\mathbf{e}}(\mathbf{w})\right) \\
&= \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\mathbf{Y}-\widehat{\boldsymbol{\mu}}(\mathbf{w}))'(\mathbf{Y}-\widehat{\boldsymbol{\mu}}(\mathbf{w}))\right). \qquad (5.6)
\end{aligned}
$$

Let $\mathbf{P}(\mathbf{w})=\sum_{p=1}^{\bar{p}} w(p)\mathbf{P}(p)$ be the weighted average of the projection matrices $\mathbf{P}(p)$, $p=1,\ldots,\bar{p}$. Denote $\mathbf{e}=(\boldsymbol{\varepsilon}_{\bar{p}+1},\ldots,\boldsymbol{\varepsilon}_T)'$. Using $\mathbf{Y}=\boldsymbol{\mu}+\mathbf{e}$ and $\mathbf{Y}-\widehat{\boldsymbol{\mu}}(p)=(\mathbf{I}_{T-\bar{p}}-\mathbf{P}(p))\mathbf{Y}$ with $\mathbf{I}_{T-\bar{p}}$ denoting the identity matrix of dimension $T-\bar{p}$, (5.6) can be decomposed into:

$$
\begin{aligned}
\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\mathbf{Y}-\widehat{\boldsymbol{\mu}}(\mathbf{w}))'(\mathbf{Y}-\widehat{\boldsymbol{\mu}}(\mathbf{w}))\right) &= \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w}))'(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w}))\right) \\
&\quad + \operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\mathbf{e}'\mathbf{e}) + 2\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\mathbf{e}'(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w})))\right), \qquad (5.7)
\end{aligned}
$$

where the first two terms on the right-hand side of (5.7) correspond to the in-sample squared error and error covariance, respectively. The $\operatorname{tr}(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\mathbf{e}'\mathbf{e})$ term does not depend on the candidate model, with the expectation being $E(\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{e}))=\operatorname{tr}(\boldsymbol{\Sigma}^{-1}E(\mathbf{e}'\mathbf{e}))=(T-\bar{p})K$. Furthermore, since $\widehat{\boldsymbol{\mu}}(\mathbf{w})=\mathbf{P}(\mathbf{w})(\boldsymbol{\mu}+\mathbf{e})$ and thus $\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w})=(\mathbf{I}_{T-\bar{p}}-\mathbf{P}(\mathbf{w}))\boldsymbol{\mu}-\mathbf{P}(\mathbf{w})\mathbf{e}$, the last term on the right-hand side of (5.7) can be further written as:

$$
\begin{aligned}
2\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\mathbf{e}'((\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w})))\right) &= 2\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\mathbf{e}'(\mathbf{I}_{T-\bar{p}}-\mathbf{P}(\mathbf{w}))\boldsymbol{\mu}-\mathbf{P}(\mathbf{w})\mathbf{e}))\right) \\
&= 2\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\mathbf{e}'(\mathbf{I}_{T-\bar{p}}-\mathbf{P}(\mathbf{w}))\boldsymbol{\mu}))\right) - 2\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{e})\right). \qquad (5.8)
\end{aligned}
$$

We now just have to investigate the asymptotic behavior of the two terms on the right-hand side of (5.8), which is the main focus of the proof of Theorem 1.

Built on the asymptotic result derived by Lewis and Reinsel (1985) which allows the VAR lag order to increase with the sample size, Theorem 1 below formally summarizes the arguments presenting the asymptotic unbiasedness of $C_T(\mathbf{w})$. To establish Theorem 1, we make the following assumptions.

Assumption 1. *(a) The multivariate time series* $\mathbf{y}_t=(y_{1t},y_{2t},\ldots,y_{Kt})'$ *satisfies* (2.1)-(2.2) *and conditions therein, where the error term vector* $\boldsymbol{\varepsilon}_t=(\varepsilon_{1t},\varepsilon_{2t},\ldots,\varepsilon_{Kt})'$ *is a* K*-dimensional i.i.d. white noise process, satisfying* $E(\boldsymbol{\varepsilon}_t|\mathcal{F}_t)=0$ *with a non-singular variance-covariance matrix* $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')=\boldsymbol{\Sigma}$.

(b) Moreover, $\boldsymbol{\varepsilon}_t$ *is assumed to have a finite fourth moment in the sense that for some finite constant* C, $E|\varepsilon_{ut}\varepsilon_{vt}\varepsilon_{wt}\varepsilon_{xt}|\leq C$ *for* $u,v,w,x=1,\ldots,K$ *and all* t.

(c) The VAR maximum length-order $\bar{p}$ depends on the sample size T such that as $T \to \infty$, $\bar{p} = \bar{p}_T \to \infty$, $\bar{p} = o\left(T^{1/3}\right)$.

Assumption 1 collects the standard assumptions in multivariate regression to ensure that the suitable law of large numbers and the central limit theorem are both satisfied. Assumption 1(c) is standard for the multivariate OLS estimator of fitting a finite-order VAR(p) model to potentially infinite-order processes, e.g., Lewis and Reinsel (1985). The condition $\bar{p} = \bar{p}_T = o\left(T^{1/3}\right)$ imposes an upper bound on the rate at which the maximum lag order $\bar{p}$ goes to infinity.

Theorem 1. *Suppose that Assumption 1 holds. For the maximum lag order $\bar{p} = \bar{p}_T$ and the fixed weight vector $\mathbf{w}$, as $T \to \infty$, the proposed MMMA criterion $C_T(\mathbf{w})$ given by (3.4) can be expressed as:*

$$C_T(\mathbf{w}) = (T - \bar{p})L_T(\mathbf{w}) + (T - \bar{p})K + r_{1T}(\mathbf{w}) + r_{2T}(\mathbf{w}) + 2K^2\mathbf{p}'\mathbf{w}, \tag{5.9}$$

where $L_T(\mathbf{w})$ is the in-sample average squared error defined by (5.5) when $h = 1$, and:

$$r_{1T}(\mathbf{w}) = 2vec(\boldsymbol{\mu}')' \left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(\mathbf{w}))' \otimes \mathbf{I}_K\right) \left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right) vec(\mathbf{e}'),$$

and

$$r_{2T}(\mathbf{w}) = -2vec(\mathbf{e}')'(\mathbf{P}(\mathbf{w})' \otimes \mathbf{I}_K) \left(\mathbf{I}_{T-p} \otimes \boldsymbol{\Sigma}^{-1}\right) vec(\mathbf{e}'),$$

satisfying $E(r_{1T}(\mathbf{w})) = 0$ and $E(r_{2T}(\mathbf{w})) = -2K^2\mathbf{p}'\mathbf{w}$ as $T \to \infty$.

Remark 1 Theorem 1 is equivalent to stating that $E(C_T(\mathbf{w})) = (T - \bar{p})E(L_T(\mathbf{w})) + (T - \bar{p})K$ as $T \to \infty$. As a special case of VAR forecast averaging, it is immediate to see that for VAR forecast selection, $E(C_T(p)) = (T - \bar{p})E(L_T(p)) + (T - \bar{p})K$ for $1 \leq p \leq \bar{p}$ as $T \to \infty$, where $L_T(p)$ is defined as (5.3) when $h = 1$, establishing asymptotic unbiasedness of $C_T(p)$.

Remark 2 As expressed in (3.4), the MMMA criterion $C_T(\mathbf{w})$ consists of two components: the trace of the standardized weighted residual covariance matrix and the penalty term. Theorem 1 implies that the leading term of $C_T(\mathbf{w})$ is a downward biased estimator of the expected loss $E(L_T(\mathbf{w}))$. This downward bias arises as we use the observations $\mathbf{Y}$ in replace of the unknown conditional mean $\boldsymbol{\mu}$ in $E(L_T(\mathbf{w}))$, while $\widehat{\boldsymbol{\mu}}(\mathbf{w})$ is estimated based on $\mathbf{Y}$. From Theorem 1 it is clear that the source of the downward bias is the term $r_{2T}(\mathbf{w})$ with its expected value shown to be $-2K^2\mathbf{p}'\mathbf{w}$, or the negative of the penalty term.

We next turn to establishing asymptotic unbiasedness of the $CV_{T,h}(\mathbf{w})$ criterion, as stated in Theorem 2. Let $\widetilde{\boldsymbol{\mu}}_h(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\widetilde{\boldsymbol{\mu}}_h(p)$, where $\widetilde{\boldsymbol{\mu}}_h(p) = \widetilde{\mathbf{P}}_h(p)\mathbf{Y}_h$ as given in Lemma 1. We also denote $\widetilde{\mathbf{P}}_h(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\widetilde{\mathbf{P}}_h(p)$. Define $\widetilde{L}_{T,h}(\mathbf{w})$ as the in-sample average squared

errors of the averaging h-step ahead forecast produced from the MCVA$_h$ procedure, i.e.:

$$\begin{aligned}\widetilde{L}_{T,h}(\mathbf{w}) &= \frac{1}{T-\bar{p}-h+1}\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))'(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))\right) \\ &= \frac{1}{(T-\bar{p}-h+1)}\sum_{t=\bar{p}}^{T-h}\sum_{\ell=1}^{K}\sum_{k=1}^{K}(\mu_{kt}^h - \widetilde{\mu}_{kt}^h(\mathbf{w}))\sigma_{k\ell}^h(\mu_{\ell t}^h - \widetilde{\mu}_{\ell t}^h(\mathbf{w})), \end{aligned} \tag{5.10}$$

and $\widetilde{V}_{T,h}(\mathbf{w}) = E(\widetilde{L}_{T,h}(\mathbf{w}))$ is the expected in-sample squared error of the averaging h-step ahead forecast based on leave-h-out cross-validation.

Theorem 2. *Suppose that Assumption 1 holds. For the maximum lag order $\bar{p} = \bar{p}_T$ and the fixed weight vector $\mathbf{w}$, as $T \to \infty$, the proposed MCVA$_h$ criterion, $CV_{T,h}(\mathbf{w})$ given by (4.4), can be expressed as:*

$$CV_{T,h}(\mathbf{w}) = (T-\bar{p}-h+1)\widetilde{L}_{T,h}(\mathbf{w}) + (T-\bar{p}-h+1)K + \widetilde{r}_{1Th}(\mathbf{w}) + \widetilde{r}_{2Th}(\mathbf{w}), \tag{5.11}$$

where $\widetilde{r}_{1Th}(\mathbf{w}) = 2vec(\boldsymbol{\mu}_h')'\left((\mathbf{I}_{T-\bar{p}-h+1} - \widetilde{\mathbf{P}}_h(\mathbf{w}))' \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}^{-1}\right) vec(\mathbf{e}_h')$ *and* $\widetilde{r}_{2Th}(\mathbf{w}) = -2vec(\mathbf{e}_h')'(\widetilde{\mathbf{P}}_h(\mathbf{w})' \otimes \mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right) vec(\mathbf{e}_h')$, *satisfying* $E(\widetilde{r}_{1Th}(\mathbf{w})) = 0$ *as* $T \to \infty$ *and* $E(\widetilde{r}_{2Th}(\mathbf{w})) = 0$.

5.3 Asymptotic optimality

This section shows that our MMMA and MCVA$_h$ procedures proposed in Sections 3 and 4, respectively, are asymptotically optimal, in the sense that asymptotically our procedures with the estimated combination weights perform as well as the infeasible procedures with the optimal weights. To begin with, the MMMA procedure is said to be asymptotically optimal with respect to the criterion $L_T(\mathbf{w})$ if:

$$\text{(OPT 1)}: \frac{L_T(\widehat{\mathbf{w}})}{\inf\limits_{\mathbf{w}\in\mathcal{H}_T} L_T(\mathbf{w})} \xrightarrow{p} 1 \quad \text{as } T \to \infty \tag{5.12}$$

is satisfied, where $\widehat{\mathbf{w}}$ is the estimated Mallows weight vector obtained from (3.7).

Define $C_T^*(\mathbf{w}) = C_T(\mathbf{w})/(T-\bar{p})$. To establish (5.12), the key is to show:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{C_T^*(\mathbf{w}) - L_T(\mathbf{w})}{L_T(\mathbf{w})}\right| \xrightarrow{p} 0. \tag{5.13}$$

Let $\overline{\mathbf{Z}} \equiv \mathbf{Z}(\bar{p})$ be the $(T-\bar{p}) \times K\bar{p}$ regressor matrix using the maximum lag order $\bar{p}$, and $\overline{\mathbf{P}} = \overline{\mathbf{Z}}\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\overline{\mathbf{Z}}'$ is the associated projection matrix. Denote $\mathbf{A}(\mathbf{w}) = \mathbf{I}_{T-\bar{p}} - \mathbf{P}(\mathbf{w})$ and

define:

$$V_T(\mathbf{w}) = E(L_T(\mathbf{w})) = \frac{1}{T-\bar{p}} E\left(\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}))'(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}))\right)\right)$$
$$= \frac{1}{T-\bar{p}} \mathrm{tr}\left(\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{A}(\mathbf{w})\right) + E\left(\mathrm{tr}\left(\mathbf{P}(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}(\mathbf{w})\right)\right), \tag{5.14}$$

and $\xi_T^* = \inf_{\mathbf{w}\in\mathcal{H}_T}(T-\bar{p})V_T(\mathbf{w})$. It is implicitly assumed that $\xi_T^* \to \infty$ as $T \to \infty$ since in our VAR framework, there is non-zero approximation error for all candidate models VAR(p) of finite order.

If the estimation loss $L_T(\mathbf{w})$ and resulting estimation risk $V_T(\mathbf{w})$ are shown to be asymptotically equivalent to each other, i.e.:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{L_T(\mathbf{w})}{V_T(\mathbf{w})} - 1\right| \xrightarrow{p} 0, \tag{5.15}$$

then the goal (5.13) to prove asymptotic optimality becomes:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{C_T^*(\mathbf{w}) - L_T(\mathbf{w})}{V_T(\mathbf{w})}\right| \xrightarrow{p} 0. \tag{5.16}$$

We make the following assumptions to prove (5.15) and (5.16), and the optimality result is stated in Theorem 3 below.

Assumption 2. *(a) As $T \to \infty$, $\bar{p}{\xi_T^*}^{-1} = o_p(1)$, and $\bar{p}{\xi_T^*}^{-2} vec(\boldsymbol{\mu}')'\left(\overline{\mathbf{P}} \otimes \mathbf{I}_K\right) vec(\boldsymbol{\mu}') = o_p(1)$.*

(b) Let S denote the set of all real K-vectors $\boldsymbol{\alpha}$ of Euclidean length one, and $P(E)$ denotes the probability of the event E. The innovation vector $\boldsymbol{\varepsilon}_t$ is uniformly Lipschitz over all directions, in the sense that there exist positive constants M, δ, and ρ such that for all u, v satisfying $0 < u - v \le \delta$, $\sup_{\boldsymbol{\alpha}\in S} P(v < \boldsymbol{\alpha}'\boldsymbol{\varepsilon}_t < u) \le M(u-v)^\rho$ holds for all t.

(c) Denote $\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) = (T-\bar{p})^{-1}\sum_{t=\bar{p}}^{T-1}\mathbf{z}_t(\bar{p})\mathbf{z}_t(\bar{p})' = (T-\bar{p})^{-1}\overline{\mathbf{Z}}'\overline{\mathbf{Z}}$. Assume that $E\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\|_1 = O(\bar{p}^{2+\theta})$ for all large T and any $\theta > 0$, where $\|\mathbf{A}\|_1^2 = \lambda_{max}(\mathbf{A}'\mathbf{A})$ is the maximum eigenvalue of the matrix $\mathbf{A}'\mathbf{A}$ and $\|\mathbf{A}\|_1^2 = \lambda_{max}^2(\mathbf{A})$ if the matrix $\mathbf{A}$ is symmetric.

(d) $\bar{p}^{6+\delta_1} = O(T)$ for some $\delta_1 > 0$.

(e) $\bar{p}^{2+\delta_1} = O(T)$ for some $\delta_1 > 0$ and $\sup_{-\infty<t<\infty} E|\varepsilon_{k_1 t}\cdots\varepsilon_{k_s t}| < \infty$ for $s = 1, 2, \ldots$ and $k_1, \ldots, k_s = 1, \ldots, K$.

Assumption 2(a) ${\bar{p}\xi_T^*}^{-1} = o_p(1)$ places a restriction on the growth rate of the maximum lag order $\bar{p}$, i.e., $\bar{p}$ must diverge slower than $\xi_T^* \to \infty$. On the other hand, the requirement ${\bar{p}\xi_T^*}^{-2}\mathrm{vec}(\boldsymbol{\mu}')'\left(\overline{\mathbf{P}} \otimes \mathbf{I}_K\right)\mathrm{vec}(\boldsymbol{\mu}') = o_p(1)$ can sometimes be viewed as a weaker condition than the convergence condition (21) of Zhang, Wan, and Zou (2013) and the similar condition (8)

of Wan, Zhang, and Zou (2010). Please see detailed discussions therein. Assumption 2(b) is directly from Findley and Wei (2002), which is a multivariate generalization of Condition (K.2) of Ing and Wei (2003) and Condition (C.3) of Zhang, Wan, and Zou (2013). This so-called uniform Lipschitz condition on the distributions of the independent process of $\boldsymbol{\varepsilon}_t$ is required to obtain the moment bound of the inverse regressor matrix, as shown in Theorem 4.1 of Findley and Wei (2002).

As discussed in Findley and Wei (2002), a rich class of distributions has the uniform Lipschitz property. Similar to Equation (2.16) of Ing and Wei (2003), Assumption 2(c) places an upper bound that goes to infinity as $\bar{p} = \bar{p}_T$ increases to infinity. As shown in Lemma 2 in Appendix, this condition plays a key role in further improving the upper bound on $\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\|$. Moreover, the conditions in Assumption 2(d) and (e) are the same as the conditions in Theorem 2 of Ing and Wei (2003) and Condition (C.4) of Zhang, Wan, and Zou (2013). These two sets of assumptions provide alternative restrictions on and a tradeoff between the growth rate of $\bar{p}$ and the existence of moments of $\boldsymbol{\varepsilon}_t$.

Theorem 3. *If either Assumptions 1*(a)-(b) *and 2*(a)-(d) *or Assumptions 1*(a) *and 2*(a)-(c) *and* (e) *are satisfied, then our MMMA procedure is asymptotically optimal in the sense that the optimality condition* (5.12) *holds.*

Remark 3 Theorem 3 extends the existing asymptotic optimality results for model averaging to the multivariate Mallows criterion in the context of VAR forecasting averaging, as discussed in the Introduction section.

Remark 4 Theorem 3 shows that $L_T(\mathbf{w})$ and hence MSFE($\mathbf{w}$) can be uniformly approximated by $C_T^*(\mathbf{w})$ (and thus by $C_T(\mathbf{w})$), implying that from a forecasting point of view, the estimated weight obtained from $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{H}_T} C_T(\mathbf{w})$ can be viewed as the optimal weight for the one-step ahead forecast combination. However, this asymptotic optimality, in general, does not carry over to multi-step ahead forecasting (i.e., $h > 1$) by the iterative MMMA. To address this limitation, our next focus is on exploring the possibility of the h-step ahead generalization of asymptotic optimality for multi-step VAR forecasting averaging, where the combination weights are selected for each h by the direct method based on our MCVA$_h$ procedure.

Similar to (5.12), the asymptotic optimality condition for the MCVA$_h$ procedure with respect to the criterion $L_{T,h}(\mathbf{w})$ (5.5) is given by:

$$\text{(OPT 2):}\ \frac{L_{T,h}(\widehat{\mathbf{w}}_{cv,h})}{\inf\limits_{\mathbf{w}\in\mathcal{H}_T} L_{T,h}(\mathbf{w})} \xrightarrow{p} 1 \quad \text{as } T\to\infty, \tag{5.17}$$

where $\widehat{\mathbf{w}}_{cv,h}$ is the estimated weight vector obtained from (4.6).

Let $V_{T,h}(\mathbf{w}) = E(L_{T,h}(\mathbf{w}))$ and $\widetilde{V}_{T,h}(\mathbf{w}) = E(\widetilde{L}_{T,h}(\mathbf{w}))$ be the associated estimation risk of the averaging h-step ahead forecasts formed by full-sample and leave-h-out estima-

tors, respectively. We also define $\xi^*_{T,h} = \inf_{\mathbf{w}\in\mathcal{H}_T}(T-\bar{p}-h+1)V_{T,h}(\mathbf{w})$ and $CV^*_{T,h}(\mathbf{w}) = CV_{T,h}(\mathbf{w})/(T-\bar{p}-h+1)$, where $CV_{T,h}(\mathbf{w})$ is given by (4.4).

Analogous to (5.13), (5.15), and (5.16), the asymptotic optimality conditions we wish to show are:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{CV^*_{T,h}(\mathbf{w})-\widetilde{L}_{T,h}(\mathbf{w})}{\widetilde{V}_{T,h}(\mathbf{w})}\right| \xrightarrow{p} 0 \quad \text{and} \quad \sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{L}_{T,h}(\mathbf{w})}{\widetilde{V}_{T,h}(\mathbf{w})}-1\right| \xrightarrow{p} 0, \tag{5.18}$$

establishing:

$$\frac{\widetilde{L}_{T,h}(\widehat{\mathbf{w}}_{cv,h})}{\inf_{\mathbf{w}\in\mathcal{H}_T}\widetilde{L}_{T,h}(\mathbf{w})} \xrightarrow{p} 1 \quad \text{as } T\to\infty, \tag{5.19}$$

which is the asymptotic optimality of $\widehat{\mathbf{w}}_{cv,h}$ with respect to the criterion $\widetilde{L}_{T,h}(\mathbf{w})$. Lastly, combining (5.19) with:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{L}_{T,h}(\mathbf{w})}{L_{T,h}(\mathbf{w})}-1\right| \xrightarrow{p} 0 \tag{5.20}$$

yields (5.17), as desired. To establish (5.17), we make the following conditions.

Assumption 3. *(a) As $T\to\infty$, $\bar{p}_T\xi^{*-1}_{T,h} = o_p(1)$ and $\bar{p}_T\xi^{*-2}_{T,h}vec(\boldsymbol{\mu}'_h)'\left(\overline{\mathbf{P}}_h\otimes\mathbf{I}_K\right)vec(\boldsymbol{\mu}'_h) = o_p(1)$, where $\overline{\mathbf{P}}_h = \overline{\mathbf{Z}}_h(\overline{\mathbf{Z}}'_h\overline{\mathbf{Z}}_h)^{-1}\overline{\mathbf{Z}}'_h$ and $\overline{\mathbf{Z}}_h = \mathbf{Z}_h(\bar{p})$.*

(b) Let $q^*_h = \max_{1\le p\le\bar{p}}\max_{\bar{p}\le t\le T-h}\max_{\underline{\ell}_{ht}-\bar{p}+1\le j\le\bar{\ell}_{ht}-\bar{p}+1}\left(\mathbf{P}_h(p)\right)_{t-\bar{p}+1,j}$, where $(\mathbf{A})_{ij}$ denotes the (i,j)-th element of matrix $\mathbf{A}$. Assume that q^*_h satisfies $q^*_h\bar{p}_T^{-1}T\to 0$ alomst surely as $T\to\infty$.

Assumption 3(a) is analogous to Assumption 2(a) for the MMMA case. Denote $\widehat{\mathbf{\Gamma}}_{T,h}(\bar{p}) = (T-\bar{p}-h+1)^{-1}\sum_{t=\bar{p}}^{T-h}\mathbf{z}_t(\bar{p})\mathbf{z}_t(\bar{p})' = (T-\bar{p}-h+1)^{-1}\overline{\mathbf{Z}}'_h\overline{\mathbf{Z}}_h$. Under the condition $E\|\widehat{\mathbf{\Gamma}}^{-1}_{T,h}(\bar{p})\|_1 = O(\bar{p}^{2+\theta})$ for all large T and any $\theta>0$, which is implied by Assumptions 1, 2(b), and 3, and using similar arguments to those employed to prove (A.21) and (A.22), it is not hard to establish the leave-h-out version of (A.21) and (A.22), i.e., $E\|\widehat{\mathbf{\Gamma}}^{-1}_{T,h}(\bar{p})\|_1 = O(1)$ and for every h, $E(\mathrm{tr}(\mathbf{e}'_h\overline{\mathbf{Z}}_h\overline{\mathbf{Z}}'_h\mathbf{e}_h))/(T-\bar{p}+h-1) = O(\bar{p}_T)$. On the other hand, combining with the fact that $\boldsymbol{\varepsilon}_{h,t+h} = \sum_{i=0}^{h-1}\mathbf{\Phi}_i\boldsymbol{\varepsilon}_{t+h-i}$, it can be shown that the uniform Lipschitz condition for the disturbance $\boldsymbol{\varepsilon}_t$ imposed in Assumption 2(b) implies that the h-step forecast error $\boldsymbol{\varepsilon}_{h,t}$ is also uniformly Lipschitz in the sense of Assumption 2(b). Assumption 3(b) is the leave-h-out generalization of the conditions that are commonly used in the literature on asymptotic optimality of leave-one-out cross-validation, e.g., Li (1987), Andrews (1991), Hansen and Racine (2012), and Zhang, Wan, and Zou (2013). This assumption requires that for a particular t, the contributions of ℓ_{ht} omitted observations, $\{\mathbf{y}_{j+h},\mathbf{z}_j(p)\}_{j=\underline{\ell}_{ht}}^{\bar{\ell}_{ht}}$, to the fitted value of $\mathbf{y}_{t+h}$ are asymptotically negligible for all candidate models.

Theorem 4 below states that the direct h-step combination weights $\widehat{\mathbf{w}}_{cv,h}$ determined by minimizing the criterion $CV_{T,h}(\mathbf{w})$ are asymptotically efficient for all fixed $h \geq 1$.

Theorem 4. *Suppose that either Assumptions 1(a)-(b), 2(b)-(d), and 3(a)-(b); or Assumptions 1(a), and 2(b),(c),(e), and 3(a)-(b) are satisfied; for all fixed $h \geq 1$, the proposed direct MCVA$_h$ procedure based on the criterion $CV_{T,h}(\mathbf{w})$ is then asymptotically optimal in the sense that the optimality condition (5.17) holds.*

Remark 5 From a theoretical perspective, Theorem 4 extends the forecast/model averaging optimality results based on leave-one-out cross-validation (e.g., Hansen and Racine (2012) and Zhang, Wan, and Zou (2013)) to multivariate time series forecasting for the forecast horizon $h > 1$. On the other hand, it also addresses the limitation of Theorem 3 where the MMMA weight estimator $\widehat{\mathbf{w}}$ is shown to be asymptotically efficient for only one-step ahead forecast averaging. Namely, for each forecast horizon $h \geq 1$, the asymptotically optimality for multi-step VAR forecast averaging can still be achieved by selecting $\widehat{\mathbf{w}}_{cv,h}$ from the direct MCVA$_h$ procedure.

6 Simulation

This section illustrates our iterative and direct multi-step VAR forecast averaging methods via simulation experiments under three data generating processes (DGPs). In particular, we focus on the finite-sample forecast performance of our methods under correct specification and misspecification of forecasting models to shed some light on the relative merits of our iterative and direct forecast averaging methods. We also compare our methods with other competing methods and examine whether or not our multivariate averaging procedures have gains in forecasting relative to the existing single-equation averaging approach.

6.1 Simulation design

The first DGP is directly from Lewis and Reinsel (1985), who consider the bivariate ARMA(1,1) model of the form:

$$\text{DGP 1:} \quad \mathbf{y}_t - \mathbf{\Phi}\mathbf{y}_{t-1} = \boldsymbol{\varepsilon}_t - \boldsymbol{\theta}\boldsymbol{\varepsilon}_{t-1}$$

with:

$$\mathbf{\Phi} = \begin{bmatrix} 1.2 & -0.5 \\ 0.6 & 0.3 \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} -0.6 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 1.00 & 0.50 \\ 0.50 & 1.25 \end{bmatrix}.$$

The second DGP is a medium-scaled VAR(5) process of seven dimensions considered in Hansen (2016):

$$\text{DGP 2:} \quad \mathbf{y}_t - \sum_{i=1}^{5} \mathbf{\Phi}_i \mathbf{y}_{t-i} = \boldsymbol{\varepsilon}_t,$$

where the coefficient matrices are $\mathbf{\Phi}_1 = (a+b)\mathbf{I}_7 + c\mathbf{1}_7$, $\mathbf{\Phi}_2 = -(ab+d)\mathbf{I}_7 - (a+b)c\mathbf{1}_7$, $\mathbf{\Phi}_3 = (a+b)d\mathbf{I}_7 + (ab+d)c\mathbf{1}_7$, $\mathbf{\Phi}_4 = -abd\mathbf{I}_7 - (a+b)cd\mathbf{1}_7$, $\mathbf{\Phi}_5 = abcd\mathbf{1}_7$ with $(a,b,c,d) = (0.5, 0.3, 0.1, 0.3)$, $\mathbf{\Sigma}$ is a diagonal matrix with diagonal elements 0.027^2, and $\mathbf{I}_7$ and $\mathbf{1}_7$ are the 7×7 identity matrix and 7×7 matrix of ones, respectively. Under this design, we are interested in examining the effect of model specification, in the sense for whether or not the true DGP is contained as one of the candidate models, on forecast accuracy of our averaging methods.

To further investigate the iterative and direct multi-step forecast performances based on our VAR averaging methods under misspecification, where under DGP 3 the data are generated from the drifting bivariate ARMA(1,10) process, as previously considered by Schorfheide (2005):

$$\text{DGP 3:} \quad \mathbf{y}_t - \mathbf{\Phi}_1 \mathbf{y}_{t-1} = \boldsymbol{\varepsilon}_t + \frac{\alpha}{\sqrt{T}} \sum_{i=1}^{10} \boldsymbol{\theta}_i \boldsymbol{\varepsilon}_{t-i},$$

where

$$\mathbf{\Phi}_1 = \begin{bmatrix} 0.754 & 0.146 \\ 0.254 & 0.646 \end{bmatrix}, \quad \boldsymbol{\theta}_1 = \begin{bmatrix} 0.87 & 0.69 \\ -1.37 & -0.03 \end{bmatrix}, \quad \boldsymbol{\theta}_2 = \begin{bmatrix} -0.05 & 0.85 \\ -0.81 & 0.14 \end{bmatrix}, \quad \boldsymbol{\theta}_3 = \begin{bmatrix} 0.30 & 0.30 \\ 0.27 & -0.10 \end{bmatrix},$$
$$\boldsymbol{\theta}_4 = \begin{bmatrix} 0.11 & -0.10 \\ -0.20 & -0.12 \end{bmatrix}, \quad \boldsymbol{\theta}_5 = \begin{bmatrix} 0.24 & -0.17 \\ -0.19 & 0.33 \end{bmatrix}, \quad \boldsymbol{\theta}_6 = \begin{bmatrix} -0.24 & -0.18 \\ -0.15 & -0.29 \end{bmatrix}, \quad \boldsymbol{\theta}_7 = \begin{bmatrix} 0.08 & 0.15 \\ -0.17 & 0.13 \end{bmatrix},$$
$$\boldsymbol{\theta}_8 = \begin{bmatrix} 0.01 & -0.05 \\ -0.14 & 0.06 \end{bmatrix}, \quad \boldsymbol{\theta}_9 = \begin{bmatrix} -0.50 & -0.12 \\ -0.21 & 0.03 \end{bmatrix}, \quad \boldsymbol{\theta}_{10} = \begin{bmatrix} 0.15 & -0.03 \\ 0.24 & 0.01 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 1.00 & 0.80 \\ 0.80 & 4.00 \end{bmatrix}.$$

We follow Schorfheide (2005) by setting $\alpha = 0, 2, 5, 10$ to allow for different degrees of (local) misspecification. It is noted that when $\alpha = 0$ in DGP 3, DGP reduces to a pure VAR process of order one.

We set the maximum lag order $\bar{p} = 3, 4, \ldots, 15$ (for DGPs 1 and 3) and $\bar{p} = 3, 4, \ldots, 8$ (for DGP 2), due to consideration of the degrees of freedom. The sample sizes $T = 100$ and $T = 200$ are considered. We examine h-step-ahead forecast errors up to $h = 12$ in our simulation experiment. The computation in simulations and the empirical application in the next section is carried out with R programming.

We compare the forecasting performance of our forecasting combination approach based on VAR model averaging with those of existing VAR lag selection/averaging methods, including the AIC, BIC, HQ, Smoothed AIC (SAIC), smoothed BIC (SBIC), and equal-weight (EQ) approaches. We also incorporate OLS using the fixed lag $\bar{p}$ as a benchmark. The lag selection using AIC, BIC, and HQ criteria is discussed in Section 2. The SAIC weights are specified to be proportional to $\exp(-\text{AIC}(p)/2)$, where $\text{AIC}(p)$ is the AIC score for candidate model p, i.e., $w_{\text{AIC}}(p) = \exp(-\text{AIC}(p)/2) / \sum_{j=1}^{\bar{p}} \exp(-\text{AIC}(j)/2)$. The SBIC weight specification, as a simple approximation to the posterior probability of the candidate model being correct, is given in a similar form to $w_{\text{AIC}}(p)$ with BIC scores in place of AIC scores. The EQ weights are simply the uniform weight given to each candidate model. We also compare

forecast performance with the VAR Stein combination shrinkage estimator (Stein) proposed by Hansen (2016).

For the above forecast selection and averaging methods, we use the iterative approach for multi-step forecasting. Specifically, under the iterative strategy and employing the estimated VAR coefficients, the h-step-ahead forecasts of the future values at times $T+1, \ldots, T+h$ are formed iteratively and compared to their actual values. On the other hand, we conduct direct multi-step forecast averaging based on the proposed MCVA_h criterion. Given the selected and combined iterative and direct h-step ahead forecasts, we then compute and report the average of their weighted MSFE values, i.e., using the inverse of $\widetilde{\mathbf{\Sigma}}_h(\bar{p})$ given in (4.5) as weights, across $R = 2,500$ random samples from the DGP under investigation:

$$\widehat{\text{MSFE}}_h(\bar{p}; M) = \frac{1}{R}\sum_{r=1}^{R}\left[\text{tr}\left(\widetilde{\mathbf{\Sigma}}_h^{(r)}(\bar{p})^{-1}\left(\mathbf{y}_{T+h}^{(r)} - \widehat{\mathbf{y}}_{T+h|T}^{(r)}(\bar{p}; M)\right)\left(\mathbf{y}_{T+h}^{(r)} - \widehat{\mathbf{y}}_{T+h|T}^{(r)}(\bar{p}; M)\right)'\right)\right], \tag{6.1}$$

where $\widehat{\mathbf{y}}_{T+h|T}^{(r)}(\bar{p}; M)$ is the h-step ahead forecast computed by the iterative or direct VAR forecast selection/averaging method M using the maximum lag order $\bar{p}$, and the superscript "(r)" indicates the r-th simulation repetition. We note that the forecast performance (measured by $\widehat{\text{MSFE}}_h(\bar{p}; M)$ of the competing methods depends on the pre-specified $\bar{p}$. To address this uncertainty, we also compute and report "maximum regret" as a secondary and supporting measure of forecast performance, as discussed in Section 6.2.1.

6.2 Simulation results

6.2.1 Multi-step forecasting performance under misspecification

Figures A1-A2 present the iterative and direct multi-step forecast performance (measured by the relative MSFE to OLS(I)) by VAR model fitting for the data generated under three DGPs considered, where "D" and "I" in parentheses refer to direct and iterative multi-step forecasts, respectively. Several findings from our simulation results are summarized as follows.

For DGP 1 (i.e., bivariate ARMA(1,1)), the panels in the first two rows of Figure A1 present relative MSFEs at forecast horizons up to $h = 12$, which can be viewed as an extension of Table 1 of Lewis and Reinsel (1985) by adding data-driven selection and averaging methods for the lag order determination. To save space, only the results using the maximum lag length $\bar{p} = 3, 5, 10$, and 15 are reported. The first finding is that the relative MSFEs of MCVA_h(D) are seen to be generally greater than those of MMMA(I) except for $h = 1$, and the former deteriorates as the forecast horizon h lengthens and $\bar{p}$ increases. For example, when $T = 100$ and $\bar{p} = 10$, MMMA(I) improves upon MCVA_h by 4.6%, 6.7%, and 8.5% for $h = 4, 8$, and 12, respectively. These figures are 6.1%, 9.0%, and 11.8% when $\bar{p}$ increases to 15. This may be expected due to the fact that fewer observations are available for es-

timation at longer forecast horizons, making inefficiency of the direct multi-step forecast methods more prominent. The quantitatively similar pattern can also be seen when OLS(D) and OLS(I) are compared. Second, when restricting the attention to the iterative multi-step methods, AIC(I), Stein(I), and HQ(I) perform notably worse than the other methods: AIC(I) is dominated by SAIC(I), and MMMA(I) always outperforms Stein(I) and HQ(I). Further improvement of MMMA(I) upon Stein(I) can be seen as $\bar{p}$ increases. Third, BIC(I) and SBIC(I) seem sensitive to $\bar{p}$: SBIC(I) performs better than BIC(I) when $\bar{p} = 3$ and 5, and the reverse can be seen when $\bar{p} = 10$ and 15. Fourth, MMMA(I) are comparable to SAIC(I), SBIC(I), and EQ(I) when $\bar{p}$ is small, whereas the outperformance of MMMA(I) is noticeable when $\bar{p}$ is sufficiently large, say $\bar{p} \geq 10$. For instance, when $T = 100$ and $\bar{p} = 15$, MMMA(I) improves upon SAIC(I), SBIC(I), and EQ(I) by respectively 3.8%, 1.6%, and 3.7% for $h = 1$; 7.4%, 5.2%, and 7.1% for $h = 4$; 5.7%, 4.1%, and 5.5% for $h = 8$; and 4.2%, 2.9%, and 4.0% for $h = 12$. In sum, under DGP 1 where misspecification is not so severe that DGP could be well approximated by finite-order VARs, MMMA(I) is superior to MCVA$_h$(D), particularly at longer lead times. In addition, overall MMMA(I) presents better performance than other competing iterative multi-step forecasting methods in most cases. The advantage of MMMA(I) is even more prominent when sufficient long VAR candidates are fitted.

Under the pure VAR(5) process of dimension 7 (DGP 2), the relative MSFEs are shown in the panels of the last two rows of Figure A1. We only report the forecast performance based on $\bar{p} = 3, 5$, and 8, corresponding to the cases of under-order, correct-order, and over-order fitting with respect to the largest candidate model. We find that, similar to DGP 1, overall MMMA(I) performs well in most of the cases considered here, and the relative performance of MMMA(I) improves as $\bar{p}$ increases. A few exceptions can be seen, such as Stein(I) slightly performs better than MMMA(I) for $h \geq 8$ when $T = 100$ and $\bar{p} = 3$, but the outperformance of Stein(I) over MMMA(I) shrinks when either the sample size or maximum lag order increases. For example, the improvement of Stein(I) upon MMMA(I) shrinks to $h = 11$ and 12 in the case of $T = 200$ and $\bar{p} = 3$. We also note that MMMA(I) is inferior to MCVA$_h$(D) only when $h = 1$. Moreover, for $\bar{p} = 3$ where all candidate models are under-specified, BIC(I) outperforms AIC(I) in most cases, and BIC(I) is seen to clearly uniformly dominate AIC(I) when $T = 100$ and $\bar{p} = 5$ and 8. This is consistent with the well-known property that BIC is consistent in model selection, in the sense of choosing the true model with probability approaching one. On the other hand, in the cases of correct specification ($\bar{p} = 5$) and over specification ($\bar{p} = 8$) where in both cases the true DGP is contained in the set of candidate models, MMMA(I), BIC(I), and SBIC(I) appear to outperform other methods and are comparable to each other. Among these best three, MMMA(I), followed by BIC(I), tends to dominate for $h \leq 8$, and MMMA(I) and SBIC(I) show very similar performances for $h > 8$.

Turning to DGP 3, we restrict our attention to the comparison of three new averaging methods: MMMA(I), MCVA$_h$(D), and Stein(I) under correct specification and misspecification of forecasting models. It can be seen from Figure A2 that in the absence of mis-

specification (i.e., $\alpha = 0$), the iterative multi-step methods generally outperform the direct multi-step methods, which is a similar finding to that in DGP 2. For example, the relative MSFEs of OLS(D) are greater than those of OLS(I), and OLS(D) deteriorates as h increases. This is expected, because, as discussed before, there is no misspecification bias and thus the bias advantage of the direct multi-step forecast methods does not appear to outweigh their variance disadvantage. Moreover, in the absence of model misspecification, it is clear that MMMA(I) performs best among all the iterative and direct methods considered. On the other hand, MCVA$_h$(D) is dominated by Stein(I) when $\bar{p} = 3$ and as the forecast horizon lengthens. However, a greater improvement of MCVA$_h$(D) and MMMA(I) upon Stein(I), OLS(I), and OLS(D) can be seen as $\bar{p}$ increases. For example, in the case of $\bar{p} = 15$, MMMA(I) and MCVA$_h$(D) are superior to Stein(I), OLS(I), and OLS(D) uniformly over all forecast horizons. The above findings apply to the case when $\alpha = 2$, i.e., the misspecification is mild.

Under DGP 3 with $\alpha \geq 5$, the outperformance of the iterative multi-step forecast selection/averaging methods may not necessarily hold. In such cases, the misspecification is large and the quality of approximating the generated processes depends crucially on the pre-specified maximum lag order $\bar{p}$. When the approximating VAR models using small $\bar{p}$, say $\bar{p} = 3$, are fitted, MCVA$_h$(D) and OLS(D) substantially dominate the iterative counterparts for $h = 7 \sim 10$ ($\alpha = 5$) and for $h = 5 \sim 11$ ($\alpha = 10$), while for other forecast horizons direct and iterative methods are comparable to each other. Taking $\alpha = 10$, the relative MSFEs of MCVA$_h$(D) are smaller by as much as 30.1% and 31.2% than those of Stein(I) and MMMA(I), respectively. As $\bar{p}$ increases to 10, MCVA$_h$(D) outperforms Stein(I) in all horizons and is superior to MMMA(I) except for $h = 11$ and 12. This appears to be consistent with a previous finding (e.g., Bhansali (1997, 1999)) that in the presence of model misspecification, under-parameterization may benefit the direct methods. We also find that the forecast performance of MMMA(I) relative to other methods tends to improve as sufficiently long VARs in the candidate model set are fitted, i.e., when $\bar{p}$ is large enough. For example, as $\bar{p}$ further grows large from 10 to 15, while the dominance of MCVA$_h$(D) over Stein(I) remains, MMMA(I) is reversely and slightly preferable to MCVA$_h$(D) except for $h = 1$. This indicates that as forecast horizons and autoregressions lengthen, the robustness of MCVA$_h$(D) is likely to be outweighted by the efficiency of MMMA(I).

As our simulation results reveal, the ranking of the competing methods based on MSFEs may vary with the pre-specified maximum lag order $\bar{p}$. To address uncertainty arising from the choice of $\bar{p}$, Figure A3 presents the normalized maximum regret based on MSFEs over different values of $\bar{p}$ under DGPs 1-3. This maximum regret criterion allows a unique ranking across maximum lag orders. The regret of the different forecast selection/averaging methods is defined as the gap between their MSFEs for a given $\bar{p}$ and the best possible MSFE across all methods (collected in the set $\mathcal{M}$) under consideration for that $\bar{p}$, namely:

$$\widehat{\mathrm{R}}_h(\bar{p}; M) = \widehat{\mathrm{MSFE}}_h(\bar{p}; M) - \min_{M \in \mathcal{M}} \widehat{\mathrm{MSFE}}_h(\bar{p}; M). \tag{6.2}$$

Given $\widehat{\mathrm{R}}_h(\bar{p}; M)$ in (6.2), the maximum regret, which is the worst-case regret, is then taken over all $\bar{p}$'s and then normalized by the maximum regret of OLS(I) - here a value of normalized maximum regret smaller than 1 implies that the method considered is superior to OLS(I). The results in Figure A2 reveal a clearer dominance of MMMA(I) (MCVA$_h$(D)) over other competing methods under no or mild (large) model misspecification when uncertainty from the choice of $\bar{p}$ is taken into account.

6.2.2 Multivariate vs. single-equation forecast averaging

As mentioned before, the VAR model is estimated efficiently using the equation-by-equation OLS estimator due to the fact that each equation of the VAR model has the same set of regressors. As a matter of fact, it seems natural to think about conducting forecast averaging for each variable separately for VAR model averaging, leading the forecast combination weights to be specific to the forecast variable. We consider applying the single-equation Mallows model averaging (SMMA) method to the VAR system in an equation-by-equation manner, in the same spirit as Hansen (2008) under a single multivariate regression setting (i.e., the Autoregressive-Distributed Lag model). Intuitively, SMMA, when compared to our MMMA, may provide some flexibility in the sense that it allows combination weights to be different across VAR equations, while the single-equation approach does not take into account the information about the correlation between response variables. Such information may be beneficial in improving forecast accuracy. To compare these two forecast averaging approaches, we additionally consider DGP 1 with different values of the covariance component, denoted by σ_{12}, in $\mathbf{\Sigma}$. We set $\sigma_{12} = 0.5, 0.6, 0.7, 0.8, 0.9$, and 1.0.

Table A1 reports the normalized maximum regret of SMMA(I) under DGPs 1-3, where the maximum regret of MMMA(I) is normalized to 1, and what "I" in parentheses refers to is the same as before. For the cases when the maximum regret of MMMA(I) is zero (meaning that MMMA(I) is the best (uniformly in $\bar{p}$) among all competing methods), we simply put "> 1.00" in Table A1 to indicate that SMMA(I) is inferior to MMMA(I). We find from Table A1 that overall MMMA(I) has better forecast performance than SMMA(I). In particular, SMMA(I) has a larger maximum regret than MMMA(I) in many cases under DGPs 2 and 3. The forecast improvement of MMMA(I) over SMMA(I) is even more prominent under DGP 2. Under DGP 1 for $T = 100$, SMMA(I) performs better than MMMA(I) in longer forecast horizons (i.e., $h > 3$) when $\sigma_{12} = 0.5$ and 0.7. MMMA(I), however, tends to dominate SMMA(I) as σ_{12} increases to 0.9 and 1.0, i.e., the correlation across equations is larger. Specifically, under DGP 1 when $T = 100$ and $\sigma_{12} = 0.9$ and 1.0, MMMA(I) shows uniform outperformance over SMMA(I) for all forecast horizons. On the other hand, when the sample size increases to $T = 200$, it is seen that MMMA(I) improves over SMMA(I) in 8 out of 12 forecast horizons in the case of $\sigma_{12} = 0.9$, while the uniform dominance of MMMA(I) in forecast horizons remains in the case of $\sigma_{12} = 1.0$.

These simulation findings, to some extent, confirm our conjecture. In other words, instead of conducting single-equation forecast averaging, incorporating the correlation between

multivariate response variables through our multivariate criterion for VAR forecast averaging may help improve the forecast performance.

7 Empirical illustration

A common interest among economists is analyzing the relationship among economic data series. This partly explains the popularity of the VAR model advanced by Sims (1980) in theoretical studies and empirical applications. For empirical illustration, this section applies our iterative and direct multi-step VAR forecasting averaging methods to forecast the U.S. macroeconomic time series.

Our empirical example uses the quarterly U.S. dataset constructed by Stock and Watson (2009). Following Giannone, Lenza, and Primiceri (2015), we consider a small-scale three-variable VAR that is a prototypical monetary VAR consisting of three endogenous variables: GDP (Y), the GDP deflator (P), and the federal funds rate (FF). In this empirical application, the variables Y and P are transformed by log differencing, while the FF series enters the model in a first-differencing form.

The dataset contains the quarterly observations ranging from 1959:Q1 to 2008:Q4. We use $T = 100$ observations for estimation. We perform the forecast exercise as follows. Using the first $T = 100$ observations ($t = 1, \ldots, 100$ from 1959:Q2-1984:Q1), VAR coefficients are estimated and forecasts are computed by using the iterative or direct methods for all the horizons up to $h = 12$ quarters ahead. We then use the rolling forecast scheme for forecast updates, i.e., we update the estimation sample by using observations $t = 2, \ldots, 101$ in the second iteration.[4] The VAR coefficients are then re-estimated using the updated sample, and a new set of h-period ahead forecasts is produced. This forecasting procedure is repeated until the sample is exhausted. The first h-step ahead forecast is for time 1984:Q2+$h-1$ for $h = 1, \ldots, 12$. The last forecast at horizon h is for time 2006:Q1+$h-1$, based on the estimation sample 1981:Q1 to 2005:Q4. This produces 88 point forecasts for each pair of 3 variables and 12 forecast horizons. The alternative VAR lag selection and averaging methods to be compared are the same as those considered in the simulation section.

Out-of-sample forecast performance is evaluated using the averages of sample MSFEs over the full forecasting evaluation period. Specifically, the sample MSFE for the h-step ahead forecast of each of the three variables $i =$Y, P, and FF using data available up to time t for estimation is:

$$\widehat{\text{MSFE}}_h^i(\bar{p}; M) = \frac{1}{t_1 - h - t_0 + 1} \sum_{t=t_0}^{t_1-h} \left(\widehat{i}_{t+h|t}(\bar{p}; M) - i_{t+h} \right)^2, \tag{7.1}$$

where t_0 and t_1 are set to 1984:Q1 and 2005:Q4, respectively, and $\widehat{i}_{t+h|t}(\bar{p}; M)$ is the h-step

[4]In our empirical example, we realize that potential structural breaks may be present in the form of parameters changing over time for instance. As a consequence, we employ the rolling estimation scheme as a simple way to give more weight to the most recent observations.

ahead forecast of variable i computed by iterative or direct VAR forecast selection/averaging method M with the maximum lag length $\bar{p}$. We also compute an aggregate version of the sample weighted MSFEs (by $\widetilde{\mathbf{\Sigma}}_h(\bar{p})$) based on (5.2) and (5.4)) for the whole VAR system as:

$$\widehat{\text{MSFE}}_h^A(\bar{p}; M) = \frac{1}{t_1 - h - t_0 + 1} \sum_{t=t_0}^{t_1-h} \text{tr}\left(\widetilde{\mathbf{\Sigma}}_{ht}(\bar{p})^{-1} \left(\mathbf{y}_{t+h} - \widehat{\mathbf{y}}_{t+h|t}(\bar{p}; M) \right) \left(\mathbf{y}_{t+h} - \widehat{\mathbf{y}}_{t+h|t}(\bar{p}; M) \right)' \right), \tag{7.2}$$

where $\mathbf{y}_{t+h} = (\text{Y}_{t+h}, \text{P}_{t+h}, \text{FF}_{t+h})'$, $\widehat{\mathbf{y}}_{t+h|t}(\bar{p}; M) = (\widehat{\text{Y}}_{t+h|t}(\bar{p}; M), \widehat{\text{P}}_{t+h|t}(\bar{p}; M), \widehat{\text{FF}}_{t+h|t}(\bar{p}; M))'$, $\widetilde{\mathbf{\Sigma}}_{ht}(\bar{p})$ is the residual covariance matrix $\widetilde{\mathbf{\Sigma}}_h(\bar{p})$ estimated using the rolling window up to time t, and the superscript "A" refers to the aggregate of MSFEs for the VAR system.

Figure A4 summarizes the relative MSFEs of h-step ahead point forecasts of the individual Y, P, and FF series and those for the VAR system of our MMMA, MCVA$_h$ and other competing methods,[5] all relative to OLS(I). The individual and aggregated MSFEs are computed from formulae (7.1) and (7.2), respectively. We also report the resulting maximum regret normalized by OLS(I), present only the results for $\bar{p} = 5, 10$, and 15 for brevity, and discuss several findings that emerge from Figure A4 as follows.

We note overall that MMMA(I) and MCVA$_h$(D) perform reasonably well, particularly when incorporating VARs that fit long $\bar{p}$ lags into the candidate models. More specifically, when $\bar{p} = 5$ (the first-row panels in Figure A4), MMMA(I) and MCVA$_h$(D) are preferred to Stein(I) in forecasting Y at most horizons and in forecasting FF at horizons $h \leq 5$; on the other hand, Stein(I) makes a substantial improvement upon MMMA(I) and MCVA$_h$(D) for the P series under all horizons. As $\bar{p}$ increases to 10 and 15 (the second- and third-row panels, respectively, in Figure A4), however, the performances of both MMMA(I) and MCVA$_h$(D) improve and are better than Stein(I) under most horizons, while the advantage of Stein(I) in forecasting P can be seen at horizons $h \geq 7$ when $\bar{p} = 10$. On the other hand, it can be seen from Figure A4 that BIC(I), HQ(I), and EQ(I) also perform well in many cases. In particular, BIC(I) is preferred to AIC(I) uniformly across all horizons and all variables when $\bar{p} = 15$. AIC(I) does a great job predicting the P series when $\bar{p} = 5$ and $\bar{p} = 10$, but on the contrary BIC(I) has poor performance in forecasting the P series when $\bar{p}$ is set to be short to modest, say $\bar{p} \leq 10$. Moreover, HQ(I) is particularly good at forecasting Y at long horizons, say $h > 9$. EQ(I) performs quite well in forecasting P, particularly at short horizons, say $h \leq 6$, and when $\bar{p}$ is sufficiently long, say $\bar{p} > 5$. We also notice that the iterative forecasts using the fixed lag order $\bar{p}$, i.e., OLS(I), nearly uniformly dominate their direct counterpart OLS(D) across all horizons, all maximum lag orders, and all variables, except for the cases of the long-horizon and short-lagged P forecasts. Moreover, also as expected, OLS(D) gets markedly worse as the lag length increases, which is in line with the previous finding that the robustness of the long-lagged direct forecast tends to be outweighted by its efficiency

[5] We do not report the results for SAIC(I) and SBIC(I), because their performances vary dramatically in our application.

loss.

We next shift the focus on the comparison between the proposed iterative MMMA(I) and direct MCVA$_h$(D) methods. First of all, it is often the case that MMMA(I) tends to have smaller relative MSFEs than MCVA$_h$(D) in forecasting Y, particularly when $\bar{p} \leq 10$, while MMMA(I) and MCVA$_h$(D) have similar performances for the Y series when $\bar{p} > 10$. For the P series, MCVA$_h$(D) tends to improve upon MMMA(I) based on low-order candidate VARs, particularly at longer horizons, with the improvements ranging from 4.6% ($h = 1$) to 22.9% ($h = 4$) when $\bar{p} = 5$ for instance.

It can also be seen that the advantage of MCVA$_h$(D) over MMMA(I) in forecasting P becomes less prominent when averaging forecasts from higher-order VAR candidates. For example, when $\bar{p} = 10$ and $\bar{p} = 15$ are specified, the respective improvements of MCVA$_h$(D) in forecasting P are about 3.5% ($h = 1$) to 16.5% ($h = 4$) and 1.7% ($h = 2$) to 11.9% ($h = 12$). This finding is consistent with Marcellino, Stock, and Watson (2006), where the authors pointed out that for the series measuring wages, prices, and money, there could be a large moving average root or long lags in the optimal linear predictor. Moreover, when forecasting the FF series, MMMA(I) is more desirable than MCVA$_h$(D) by a substantial margin at most horizons, while MMMA(I) and MCVA$_h$(D) perform similarly only for short ($h = 1$) and long ($h = 12$) horizons. For example, MMMA(I) improves upon MCVA$_h$(D) by 23.8% (at $h = 4$) and by 18.2% (at $h = 8$) when $\bar{p} = 10$. If the attention is restricted to the aggregated MSFEs (i.e., the panels labelled A in Figure A4), then MMMA(I) and MCVA$_h$(D) are competitive to each other at short to modest horizons, say $h \leq 6$, while MCVA$_h$(D) tends to dominate MMMA(I) at longer horizons, e.g., MCVA$_h$(D) improves upon MMMA(I) by 10.1% ($\bar{p} = 5$), 11.5% ($\bar{p} = 10$), and 6.3% ($\bar{p} = 15$) at horizon $h = 12$.

As far as the normalized maximum regret is concerned, it is clear to see that MMMA(I) performs quite well in forecasting Y and FF series, while MCVA$_h$(D) has prominent forecast advantages for the P series. In terms of the aggregated MSFEs for the VAR system, MMMA(I) and MCVA$_h$(D) are competitive with each other at short to modest horizons, say $h < 6$, and MCVA$_h$(D) tends to dominate MMMA(I) at longer horizons, $h \geq 6$. Moreover, both MMMA(I) and MCVA$_h$(D) appear to be competitive with EQ(I) and HQ(I) at around horizons 4∼8 for which EQ(I) and HQ(I) have the lowest and the second lowest normalized maximum regrets, respectively.

8 Conclusion

This paper has employed a frequentist model averaging approach based on the MMMA and MCVA$_h$ criteria for combinations of multi-step forecasts with VAR models. The former is designed for iterative multi-step VAR forecast averaging, while the latter aims to deal with the issue of serial correlation that is due to overlapping data under the direct multi-step forecasting framework. The proposed methods are straightforward to implement, because our procedures are based on least squares estimation and quadratic programming for ob-

taining the combination weights. We have also shown that our approaches are theoretically grounded by the properties of asymptotic unbiasedness and asymptotic optimality. We have further investigated the numerical performances of our methods and have compared them to other competing methods in a Monte Carlo simulation and an empirical application to U.S. macroeconomic variables, illustrating the usefulness of our methods as econometric tools for multi-step VAR forecast combinations.

Several directions built on the present paper are worth exploring for future research. For example, as datasets with large cross-sectional dimension have drawn growing attention in theoretical and applied econometrics, dimension reduction techniques, such as factors or index variables, should be introduced into the framework of VAR forecasting averaging. It would also be interesting to extend our methodology for future research into non-stationary processes.

References

Andersson, M. K., and S. Karlsson (2007): "Bayesian Forecast Combination for VAR Models," Working Papers 2007:13, Orebro University, School of Business.

Andrews, D. W. K. (1991): "Asymptotic Optimality of Generalized CL, Cross-validation, and Generalized Cross-validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377.

Berk, K. N. (1974): "Consistent Autoregressive Spectral Estimates," *The Annals of Statistics*, 2(3), 489–502.

Bhansali, R. J. (1996): "Asymptotically Efficient Autoregressive Model Selection for Multistep Prediction," *Annals of the Institute of Statistical Mathematics*, 48(3), 577–602.

——— (1997): "Direct Autoregressive Predictors for Multistep Prediction: Order Selection and Performance Relative to the Plug in Predictors," *Statistica Sinica*, 7(2), 425–449.

——— (1999): "Parameter Estimation and Model Selection for Multistep Prediction of Time Series: A Review," in *Asymptotics, Nonparametrics and Time Series*, ed. by S. Gosh, pp. 201–225. Marcel Dekker.

Chen, R., L. Yang, and C. Hafner (2004): "Nonparametric Multistep-Ahead Prediction in Time Series Analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(3), 669–686.

Cheng, T.-C. F., C.-K. Ing, and S.-H. Yu (2015): "Toward Optimal Model Averaging in Regression Models with Time Series Errors," *Journal of Econometrics*, 189(2), 321–334, Frontiers in Time Series and Financial Econometrics.

Cheng, X., and B. E. Hansen (2015): "Forecasting with Factor-augmented Regression: A Frequentist Model Averaging Approach," *Journal of Econometrics*, 186(2), 280–293, High Dimensional Problems in Econometrics.

Chevillon, G. (2007): "Direct Multi-step Estimation and Forecasting," *Journal of Economic Surveys*, 21(4), 746–785.

Chevillon, G., and D. F. Hendry (2005): "Non-parametric Direct Multi-step Estimation for Forecasting Economic Processes," *International Journal of Forecasting*, 21(2), 201–218.

Cho, S., and A. Moreno (2006): "A Small-Sample Study of the New-Keynesian Macro Model," *Journal of Money, Credit and Banking*, 38(6), 1461–1481.

Clark, T. E., and McCracken (2010): "Averaging Forecasts from VARs with Uncertain Instabilities," *Journal of Applied Econometrics*, 25, 5–29.

Elliott, G., and A. Timmermann (2016): *Economic Forecasting*. Princeton University Press.

Findley, D. F., and C.-Z. Wei (2002): "AIC, Overfitting Principles, and the Boundedness of Moments of Inverse Matrices for Vector Autotregressions and Related Models," *Journal of Multivariate Analysis*, 83(2), 415–450.

Fujikoshi, Y., and K. Satoh (1997): "Modified AIC and C_p in Multivariate Linear Regression," *Biometrica*, 84(3), 707–716.

Gao, Y., X. Zhang, S. Wang, and G. Zou (2016): "Model Averaging Based on Leave-subject-out Cross-validation," *Journal of Econometrics*, 192(1), 139–151.

Giannone, D., M. Lenza, and G. Primiceri (2015): "Prior Selection For Vector Autoregressions," *The Review of Economics and Statsitics*, 97(2), 436–451.

Hansen, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75(4), 1175–1189.

——— (2008): "Least-squares Forecast Averaging," *Journal of Econometrics*, 146(2), 342–350, Honoring the research contributions of Charles R. Nelson.

Hansen, B. E. (2010): "Multi-step Forecast Model Selection," Working paper, University of Wisconsin.

——— (2016): "Stein Combination Shrinkage for Vector Autoergressions," Discussion paper, University of Wisconsin.

Hansen, B. E., and J. S. Racine (2012): "Jackknife Model Averaging," *Journal of Econometrics*, 167(1), 38–46.

Ing, C.-K. (2003): "Multistep Prediction in Autoregressive Processes," *Econometric Theory*, 19(2), 254–279.

Ing, C.-K., and C.-Z. Wei (2003): "On Same-realization Prediction in an Infinite-order Autoregressive Process," *Journal of Multivariate Analysis*, 85(1), 130–155.

Kunitomo, N., and T. Yamamoto (1985): "Properties of Predictors in Misspecified Autoregressive Time Series Models," *Journal of the American Statistical Association*, 80(392), 941–950.

Lewis, R., and G. Reinsel (1985): "Prediction of Multivariate Time Series by Autoregressive Model Fitting," *Journal of Multivariate Analysis*, 16(3), 393–411.

Li, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15(3), 958–975.

Liu, Q., R. Okui, and A. Yoshimura (2016): "Generalized Least Squares Model Averaging," *Econometric Reviews*, 35(8-10), 1692–1752.

Lütkepohl, H. (2005): *New Introduction to Multiple Time Series Analysis*. Springer.

Marcellino, M., J. Stock, and M. Watson (2006): "A Comparison of Direct and Iterated Multistep AR MMethod for Forecasting Macroeconomic Time Series," *Journal of Econometrics*, 135, 499–526.

McQuarrie, A. D., and C.-L. Tsai (1998): *Regression and Time Series Model Selection*. World Scientific Publishing Co. Pte. Ltd.

Pesaran, M. H., A. Pick, and A. Timmermann (2011): "Variable Selection, Estimation and Inference for Multi-period Forecasting Problems," *Journal of Econometrics*, 164(1), 173–187, Annals Issue on Forecasting.

Racine, J. (1997): "Feasible Cross-Validation Model Selection for General Stationary Processes," *Journal of Applied Econometrics*, 12(2), 169–179.

Schorfheide, F. (2005): "VAR Forecasting under Misspecification," *Journal of Econometrics*, 128(1), 99–136.

Sims, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48(1), 1–48.

Stock, J., and M. Watson (2009): *Forecasting in Dynamic Factor Models Subject to Structural Instability*pp. 1–57. Oxford University Press.

Tiao, G. C., and G. E. P. Box (1981): "Modeling Multiple Times Series with Applications," *Journal of the American Statistical Association*, 76(376), 802–816.

Wan, A. T., X. Zhang, and G. Zou (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156(2), 277–283.

Yanagihara, H., and K. Satoh (2010): "An Unbiased Criterion for Multivariate Ridge Regression," *Journal of Multivariate Analysis*, 101(5), 1226–1238.

Zellner, A. (1962): "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57(298), 348–368.

Zhang, X., A. T. Wan, and G. Zou (2013): "Model Averaging by Jackknife Criterion in Models with Dependent Data," *Journal of Econometrics*, 174(2), 82–94.

A Appendix

A1 Derivations of equation (3.5)

$$
\begin{aligned}
&(T-\bar{p})\cdot \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\widehat{\boldsymbol{\Sigma}}(\mathbf{w})\right) = \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\sum_{t=\bar{p}}^{T-1}\widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w})\widehat{\boldsymbol{\varepsilon}}_{t+1}(\mathbf{w})'\right)\\
&=\sum_{t=\bar{p}}^{T-1}\operatorname{tr}\left\{\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\left(\sum_{p=1}^{\bar{p}}w(p)\widehat{\boldsymbol{\varepsilon}}_{t+1}(p)\right)\left(\sum_{p=1}^{\bar{p}}w(p)\widehat{\boldsymbol{\varepsilon}}_{t+1}(p)\right)'\right\}\\
&=\sum_{t=\bar{p}}^{T-1}\operatorname{tr}\left\{\underbrace{\begin{bmatrix}\widetilde{\sigma}_{11} & \widetilde{\sigma}_{12} & \cdots & \widetilde{\sigma}_{1K}\\ \widetilde{\sigma}_{21} & \widetilde{\sigma}_{22} & \cdots & \widetilde{\sigma}_{2K}\\ \vdots & \vdots & \ddots & \vdots\\ \widetilde{\sigma}_{K1} & \cdots & \cdots & \widetilde{\sigma}_{KK}\end{bmatrix}}_{\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}}\left(w(1)\begin{bmatrix}\widehat{\varepsilon}_{1,t+1}(1)\\ \widehat{\varepsilon}_{2,t+1}(1)\\ \vdots\\ \widehat{\varepsilon}_{K,t+1}(1)\end{bmatrix}+\cdots+w(\bar{p})\begin{bmatrix}\widehat{\varepsilon}_{1,t+1}(\bar{p})\\ \widehat{\varepsilon}_{2,t+1}(\bar{p})\\ \vdots\\ \widehat{\varepsilon}_{K,t+1}(\bar{p})\end{bmatrix}\right)\right.\\
&\quad\left.\left(w(1)\begin{bmatrix}\widehat{\varepsilon}_{1,t+1}(1)\\ \widehat{\varepsilon}_{2,t+1}(1)\\ \vdots\\ \widehat{\varepsilon}_{K,t+1}(1)\end{bmatrix}'+\cdots+w(\bar{p})\begin{bmatrix}\widehat{\varepsilon}_{1,t+1}(\bar{p})\\ \widehat{\varepsilon}_{2,t+1}(\bar{p})\\ \vdots\\ \widehat{\varepsilon}_{K,t+1}(\bar{p})\end{bmatrix}'\right)\right\}\\
&=\sum_{t=\bar{p}}^{T-1}\sum_{i=1}^{\bar{p}}\sum_{j=1}^{\bar{p}}w(i)w(j)\left\{\sum_{k=1}^{K}\sum_{\ell=1}^{K}\widetilde{\sigma}_{k\ell}\widehat{\varepsilon}_{k,t+1}(i)\widehat{\varepsilon}_{\ell,t+1}(j)\right\}\\
&\equiv\sum_{t=\bar{p}}^{T-1}\sum_{i=1}^{\bar{p}}\sum_{j=1}^{\bar{p}}w(i)w(j)\widetilde{\varepsilon}_{t+1,ij}\\
&=\mathbf{w}'\widehat{\mathbf{S}}\mathbf{w}.
\end{aligned}
$$

Let $\widetilde{\mathbf{R}}$ be the squared root of $\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}$, i.e., $\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}=\widetilde{\mathbf{R}}\widetilde{\mathbf{R}}$. The matrix $\widehat{\mathbf{S}}$ can be alternatively expressed in a compact form as:

$$
\widehat{\mathbf{S}}=\sum_{t=\bar{p}}^{T-1}\widetilde{\boldsymbol{\varepsilon}}_{t+1}\widetilde{\boldsymbol{\varepsilon}}_{t+1}', \tag{A.1}
$$

where

$$
\widetilde{\boldsymbol{\varepsilon}}_{t+1}=\begin{bmatrix}\widehat{\varepsilon}_{1,t+1}(1) & \widehat{\varepsilon}_{2,t+1}(1) & \cdots & \widehat{\varepsilon}_{K,t+1}(1)\\ \widehat{\varepsilon}_{1,t+1}(2) & \widehat{\varepsilon}_{2,t+1}(2) & \cdots & \widehat{\varepsilon}_{K,t+1}(2)\\ \vdots & \vdots & \ddots & \vdots\\ \widehat{\varepsilon}_{1,t+1}(\bar{p}) & \widehat{\varepsilon}_{2,t+1}(\bar{p}) & \cdots & \widehat{\varepsilon}_{K,t+1}(\bar{p})\end{bmatrix}\widetilde{\mathbf{R}}\equiv\begin{bmatrix}\widetilde{\varepsilon}_{1,t+1}(1) & \widetilde{\varepsilon}_{2,t+1}(1) & \cdots & \widetilde{\varepsilon}_{K,t+1}(1)\\ \widetilde{\varepsilon}_{1,t+1}(2) & \widetilde{\varepsilon}_{2,t+1}(2) & \cdots & \widetilde{\varepsilon}_{K,t+1}(2)\\ \vdots & \vdots & \ddots & \vdots\\ \widetilde{\varepsilon}_{1,t+1}(\bar{p}) & \widetilde{\varepsilon}_{2,t+1}(\bar{p}) & \cdots & \widetilde{\varepsilon}_{K,t+1}(\bar{p})\end{bmatrix}
$$

is a $\bar{p}\times K$ matrix.

Moreover, equation (3.5) can also be conveniently expressed as:

$$\mathbf{w}'\widehat{\mathbf{S}}\mathbf{w} = \mathbf{w}'\left(\sum_{t=\bar{p}}^{T-1}\widetilde{\boldsymbol{\varepsilon}}_{t+1}\widetilde{\boldsymbol{\varepsilon}}_{t+1}'\right)\mathbf{w} = \mathbf{w}'\bar{\boldsymbol{\varepsilon}}'\bar{\boldsymbol{\varepsilon}}\mathbf{w}, \tag{A.2}$$

where

$$\bar{\boldsymbol{\varepsilon}}' = \begin{bmatrix} \widetilde{\varepsilon}_{1,\bar{p}+1}(1) & \dots & \widetilde{\varepsilon}_{1T}(1) & \widetilde{\varepsilon}_{2,\bar{p}+1}(1) & \dots & \widetilde{\varepsilon}_{2T}(1) & \cdots & \widetilde{\varepsilon}_{K,\bar{p}+1}(1) & \dots & \widetilde{\varepsilon}_{KT}(1) \\ \widetilde{\varepsilon}_{1,\bar{p}+1}(2) & \dots & \widetilde{\varepsilon}_{1T}(2) & \widetilde{\varepsilon}_{2,\bar{p}+1}(2) & \dots & \widetilde{\varepsilon}_{2T}(2) & \cdots & \widetilde{\varepsilon}_{K,\bar{p}+1}(2) & \dots & \widetilde{\varepsilon}_{KT}(2) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \widetilde{\varepsilon}_{1,\bar{p}+1}(\bar{p}) & \dots & \widetilde{\varepsilon}_{1T}(\bar{p}) & \widetilde{\varepsilon}_{2,\bar{p}+1}(\bar{p}) & \dots & \widetilde{\varepsilon}_{2T}(\bar{p}) & \cdots & \widetilde{\varepsilon}_{K,\bar{p}+1}(\bar{p}) & \dots & \widetilde{\varepsilon}_{KT}(\bar{p}) \end{bmatrix} \tag{A.3}$$

is a $\bar{p} \times K(T-\bar{p})$ matrix.

In a special case of $K = 1$ (i.e., the univariate $AR(p)$), it is obvious to see that $(T-\bar{p}) \cdot \text{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\widehat{\boldsymbol{\Sigma}}(\mathbf{w})\right)$ reduces to:

$$\sum_{t=\bar{p}}^{T-1}\sum_{i=1}^{\bar{p}}\sum_{j=1}^{\bar{p}} w(i)w(j)\widetilde{\sigma}^{-2}\widehat{\varepsilon}_{t+1}(i)\widehat{\varepsilon}_{t+1}(j) = \widetilde{\sigma}^{-2}\mathbf{w}'\bar{\boldsymbol{\varepsilon}}'\bar{\boldsymbol{\varepsilon}}\mathbf{w},$$

where $\widehat{\varepsilon}_{t+1}(p)$ for $t = \bar{p}, \dots, T-1$ and $p = 1, \dots, \bar{p}$ are OLS residuals and $\widetilde{\sigma}^2$ is the estimated variance from the largest model, i.e.:

$$\widetilde{\sigma}^2 = \frac{1}{T-\bar{p}}\sum_{t=\bar{p}}^{T-1}\widehat{\varepsilon}_{t+1}(\bar{p})^2,$$

and $\bar{\boldsymbol{\varepsilon}}$ defined in (A.3) reduces to:

$$\bar{\boldsymbol{\varepsilon}} = \begin{bmatrix} \widehat{\varepsilon}_{\bar{p}+1}(1) & \widehat{\varepsilon}_{\bar{p}+1}(2) & \cdots & \widehat{\varepsilon}_{\bar{p}+1}(\bar{p}) \\ \widehat{\varepsilon}_2(1) & \widehat{\varepsilon}_2(2) & \cdots & \widehat{\varepsilon}_2(\bar{p}) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\varepsilon}_T(1) & \widehat{\varepsilon}_T(2) & \cdots & \widehat{\varepsilon}_T(\bar{p}) \end{bmatrix},$$

which is a $(T-\bar{p}) \times \bar{p}$ matrix. The Mallows averaging criterion becomes:

$$C_T(\mathbf{w}) = \widetilde{\sigma}^{-2}\mathbf{w}'\bar{\boldsymbol{\varepsilon}}'\bar{\boldsymbol{\varepsilon}}\mathbf{w} + 2\mathbf{p}'\mathbf{w}. \tag{A.4}$$

Since the constant $\widetilde{\sigma}^{-2}$ plays no practical role in model selection/averaging criterion, multiplying (A.4) by $\widetilde{\sigma}^2$ gives another equivalent expression of (A.4):

$$C_T(\mathbf{w}) = \mathbf{w}'\bar{\boldsymbol{\varepsilon}}'\bar{\boldsymbol{\varepsilon}}\mathbf{w} + 2\widetilde{\sigma}^2\mathbf{p}'\mathbf{w}, \tag{A.5}$$

which equals equation (13) in Hansen (2007, p.1180) or equation (16) in Hansen (2008, p.344).

A2 Proof of Theorem 1

For each candidate VAR(p) model, recall $\mathbf{P}(p) = \mathbf{Z}(p)(\mathbf{Z}(p)'\mathbf{Z}(p))^{-1}\mathbf{Z}(p)'$ and $\mathbf{P}(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\mathbf{P}(p)$. We write and expand the sum of squared residuals as:

$$\begin{aligned}
&\operatorname{tr}\left((\mathbf{Y} - \widehat{\boldsymbol{\mu}}(\mathbf{w}))\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}(\mathbf{w}))'\right) \\
&\quad = \operatorname{tr}\left((\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}) + \mathbf{e})\,\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\,((\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}) + \mathbf{e})'\right) \\
&\quad = \operatorname{vec}((\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}) + \mathbf{e})')'\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}((\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}) + \mathbf{e})') \\
&\quad = \operatorname{vec}((\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}((\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}))') \\
&\qquad + \operatorname{vec}(\mathbf{e}')'\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}') \\
&\qquad + 2\operatorname{vec}((\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}'), \qquad \text{(A.6)}
\end{aligned}$$

where vec and $\otimes$ denote a column stacking operator and kronecker product, respectively, and for the second equality we use the property that for conformable matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, $\operatorname{tr}(\mathbf{ABC}) = \operatorname{vec}(\mathbf{A}')'(\mathbf{I} \otimes \mathbf{B})\operatorname{vec}(\mathbf{C})$. The first two terms on the right-hand side of equation (A.6) correspond to the in-sample squared error and error covariance, respectively, and the latter term does not depend on the candidate model.

We next examine the third term on the right-hand side of equation (A.6). Rewriting $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \mathbf{P}(\mathbf{w})(\boldsymbol{\mu} + \mathbf{e})$ and thus $\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}) = (\mathbf{I}_{T-\bar{p}} - \mathbf{P}(\mathbf{w}))\boldsymbol{\mu} - \mathbf{P}(\mathbf{w})\mathbf{e}$, we have:

$$\begin{aligned}
&2\operatorname{vec}((\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}') \\
&\quad = 2\operatorname{vec}(((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(\mathbf{w}))\boldsymbol{\mu})' - (\mathbf{P}(\mathbf{w})\mathbf{e})')'\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}') \\
&\quad = 2\operatorname{vec}((\boldsymbol{\mu}'(\mathbf{I}_{T-\bar{p}} - \mathbf{P}(\mathbf{w}))'\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}') \\
&\qquad - 2\operatorname{vec}((\mathbf{e}'\mathbf{P}(\mathbf{w})')'\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}') \\
&\quad = 2\operatorname{vec}(\boldsymbol{\mu}')'\left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(\mathbf{w}))' \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}') \\
&\qquad - 2\operatorname{vec}(\mathbf{e}')'(\mathbf{P}(\mathbf{w})' \otimes \mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}') \\
&\quad \equiv r_{1T}(\mathbf{w}) + r_{2T}(\mathbf{w}), \qquad \text{(A.7)}
\end{aligned}$$

where the third equality follows from the property that $\operatorname{vec}(\mathbf{AB}) = (\mathbf{B}' \otimes \mathbf{I})\operatorname{vec}(\mathbf{A})$ for conformable matrices $\mathbf{A}$ and $\mathbf{B}$.

We first examine the term $r_{1T}(\mathbf{w})$. Note that $p = p_T$ is assumed to increase with the sample size T. For each candidate model VAR(p), $p = 1, \ldots, \bar{p}$, we define:

$$\begin{aligned}
\xi_{1T}(p) &\equiv \frac{1}{\sqrt{T-\bar{p}}}\operatorname{vec}(\boldsymbol{\mu}')'\left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(p)) \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}') \\
&= \frac{1}{\sqrt{T-\bar{p}}}\operatorname{vec}(\boldsymbol{\mu}')'\left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(p)) \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\operatorname{vec}(\mathbf{e}'), \qquad \text{(A.8)}
\end{aligned}$$

where $\xi_{1T}(p)$ satisfies $\xi_{1T}(p)/\Gamma_1(p)^{1/2} \xrightarrow{d} N(0,1)$ with $\Gamma_1(p) = \operatorname{plim} \boldsymbol{\nu}(p)'\left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}\right)\boldsymbol{\nu}(p)/(T-$

$\bar{p})$, $\boldsymbol{\nu}(p)' \equiv \text{vec}(\boldsymbol{\mu}')' \left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(p)) \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1} \right)$, and $\xrightarrow{d}$ denotes convergence in distribution. This large sample result for (A.8) is an application of Proposition 15.1 of Lütkepohl (2005, p.533). Equation (A.8) says that $\xi_{1T}(p)/\Gamma_1(p)^{1/2}$ is a standard normal random variable. Since the term $r_{1T}(\mathbf{w})$ in (A.7) can be expressed as $2\sum_{p=1}^{\bar{p}} w(p)\xi_{1T}(p)$, $r_{1T}(\mathbf{w})$ is a weighted sum of mean-zero normal random variables, implying $E(r_{1T}(\mathbf{w})) = 0$.

We next move to evaluate the term $r_{2T}(\mathbf{w})$. We note that for each VAR(p) candidate model, $\mathbf{P}(p) = \mathbf{Z}(p) \left(\mathbf{Z}(p)'\mathbf{Z}(p) \right)^{-1} \mathbf{Z}(p)'$ and thus:

$$\begin{aligned}
\text{vec}(\mathbf{e}'\mathbf{P}(p)) &= \text{vec} \left(\mathbf{e}'\mathbf{Z}(p) \left(\mathbf{Z}(p)'\mathbf{Z}(p) \right)^{-1} \mathbf{Z}(p)' \right) \\
&= \left(\mathbf{Z}(p) \left(\mathbf{Z}(p)'\mathbf{Z}(p) \right)^{-1} \otimes \mathbf{I}_K \right) \text{vec}(\mathbf{e}'\mathbf{Z}(p)) \\
&= \left(\mathbf{Z}(p) \otimes \mathbf{I}_K \right) \left(\left(\mathbf{Z}(p)'\mathbf{Z}(p) \right)^{-1} \otimes \mathbf{I}_K \right) \text{vec}(\mathbf{e}'\mathbf{Z}(p)).
\end{aligned}$$

As a result, we write:

$$\begin{aligned}
&- 2\text{vec}((\mathbf{e}'\mathbf{P}(p))' \left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1} \right) \text{vec}(\mathbf{e}') \\
&= -2\text{vec}(\mathbf{e}'\mathbf{Z}(p))' \left(\left(\mathbf{Z}(p)'\mathbf{Z}(p) \right)^{-1} \otimes \mathbf{I}_K \right) \left(\mathbf{Z}(p)' \otimes \mathbf{I}_K \right) \left(\mathbf{I}_{Kp} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1} \right) \text{vec}(\mathbf{e}') \\
&= -2\text{vec}(\mathbf{e}'\mathbf{Z}(p))' \left(\left(\mathbf{Z}(p)'\mathbf{Z}(p) \right)^{-1} \otimes \mathbf{I}_K \right) \left(\mathbf{I}_{Kp} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1} \right) \left(\mathbf{Z}(p)' \otimes \mathbf{I}_K \right) \text{vec}(\mathbf{e}') \\
&= -2\text{vec}(\mathbf{e}'\mathbf{Z}(p))' \left(\left(\mathbf{Z}(p)'\mathbf{Z}(p) \right)^{-1} \otimes \mathbf{I}_K \right) \left(\mathbf{I}_{Kp} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1} \right) \text{vec}(\mathbf{e}'\mathbf{Z}(p)). \qquad \text{(A.9)}
\end{aligned}$$

To evaluate (A.9), we first note that:

$$\begin{aligned}
\left(\left(\mathbf{Z}(p)'\mathbf{Z}(p) \right)^{-1} \otimes \mathbf{I}_K \right) \text{vec} \left(\mathbf{e}'\mathbf{Z}(p) \right) &= \left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}} \right)^{-1} \otimes \mathbf{I}_K \right) \text{vec} \left((T-\bar{p})^{-1} \sum_{t=\bar{p}+1}^{T} \boldsymbol{\varepsilon}_t \mathbf{z}_t' \right) \\
&= \text{vec} \left((T-\bar{p})^{-1} \left(\sum_{t=\bar{p}+1}^{T} \boldsymbol{\varepsilon}_t \mathbf{z}_t(p)' \right) \left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}} \right)^{-1} \right), \qquad \text{(A.10)}
\end{aligned}$$

where we recall that $\mathbf{z}_t(p)' = (\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}')$, $t = \bar{p}+1, \ldots, T$, is the $(t-\bar{p})$-th row of $\mathbf{Z}(p)$, and $\mathbf{e}' = (\boldsymbol{\varepsilon}_{\bar{p}+1}, \ldots, \boldsymbol{\varepsilon}_T)$. Denote $\boldsymbol{\Gamma}_2(p) = \text{plim}\, \mathbf{Z}(p)'\mathbf{Z}(p)/(T-\bar{p})$, and let $\ell(p)$ be a sequence of $K^2p \times 1$ vectors such that $0 < c_1 \leq \ell(p)'\ell(p) \leq c_2 < \infty$ for positive constants c_1 and c_2. We thus define:

$$s_T = (T-\bar{p})^{1/2} \ell(p)' \text{vec} \left((T-\bar{p})^{-1} \left(\sum_{t=\bar{p}+1}^{T} \boldsymbol{\varepsilon}_t \mathbf{z}_t(p)' \right) \boldsymbol{\Gamma}_2(p)^{-1} \right),$$

and $v_T^2 = \text{Var}(s_T) = \ell(p)' \left(\boldsymbol{\Gamma}_2(p)^{-1} \otimes \boldsymbol{\Sigma} \right) \ell(p)$. Under Assumption 1, Theorem 3 of Lewis and Reinsel (1985) shows that as $T \to \infty$:

$$s_T / v_T \xrightarrow{d} N(0,1). \qquad \text{(A.11)}$$

Putting together the arguments in (A.10) and (A.11), it is obvious to derive the following

limiting distribution result that allows the lag order p to increase with the sample size:

$$\frac{\left(\ell(p)'\text{vec}(\mathbf{e}'\mathbf{Z}(p))\right)/\sqrt{T-\bar{p}}}{\left(\ell(p)'\left(\mathbf{\Gamma}_2(p)^{-1}\otimes\mathbf{\Sigma}\right)\ell(p)\right)^{1/2}} \equiv \ell(p)'\boldsymbol{\phi}_T(p) \xrightarrow{d} N(0,1), \tag{A.12}$$

where

$$\boldsymbol{\phi}_T(p) \equiv \frac{\text{vec}(\mathbf{e}'\mathbf{Z}(p))/\sqrt{T-\bar{p}}}{\left(\mathbf{\Gamma}_2(p)^{-1}\otimes\mathbf{\Sigma}\right)^{1/2}}.$$

By the Cramér-Wold theorem, $\boldsymbol{\phi}_T(p)$ then converges in distribution to a K^2p-dimensional vector of multivariate standard normal random variables. Denote $\xi_{2T}(p) \equiv \boldsymbol{\phi}_T(p)'\boldsymbol{\phi}_T(p)$. We thus have $\xi_{2T}(p) \xrightarrow{d} \chi^2(K(p))$, where $\chi^2(K(p))$ is a chi-squared distribution with degrees of freedom $K(p) = K^2p$.

We now turn back to (A.9). Using the asymptotic normality results discussed above, the consistency of $\widetilde{\mathbf{\Sigma}}(\bar{p})$ (since the maximum lag order $\bar{p}$ increases with sample size), and the fact that (A.9) (ignoring the constant -2 for a moment) is a quadratic form in multivariate normal random variables, (A.9) is asymptotically equivalent to $\xi_{2T}(p)$, i.e.:

$$\begin{aligned}&\frac{1}{\sqrt{T-\bar{p}}}\text{vec}(\mathbf{e}'\mathbf{Z}(p))'\left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}}\right)\otimes\widetilde{\mathbf{\Sigma}}(\bar{p})\right)^{-1}\frac{1}{\sqrt{T-\bar{p}}}\text{vec}(\mathbf{e}'\mathbf{Z}(p))\\ &\quad-\xi_{2T}(p) = o_p(1),\end{aligned} \tag{A.13}$$

where we use

$$\begin{aligned}\left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}}\right)^{-1}\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{Kp}\otimes\widetilde{\mathbf{\Sigma}}(\bar{p})^{-1}\right) &= \left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}}\right)^{-1}\otimes\widetilde{\mathbf{\Sigma}}(\bar{p})^{-1}\right)\\ &= \left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}}\right)\otimes\widetilde{\mathbf{\Sigma}}(\bar{p})\right)^{-1}.\end{aligned} \tag{A.14}$$

Denote $\xi_{2T}(\mathbf{w}) \equiv \sum_{p=1}^{\bar{p}} w(p)\xi_{2T}(p)$. It then follows that $\xi_{2T}(\mathbf{w})$ converges in distribution to a weighted sum of chi-squared random variables $\xi_{2T}(p)$, $p = 1,\ldots,\bar{p}$, with the mean $E(\xi_{2T}(\mathbf{w})) = \sum_{p=1}^{\bar{p}} w(p)E(\xi_{2T}(p)) = \sum_{p=1}^{\bar{p}} w(p)K(p)$. This implies:

$$E(r_{2T}(\mathbf{w})) = -2\sum_{p=1}^{\bar{p}} w(p)K^2p = -2K^2\mathbf{p}'\mathbf{w}. \tag{A.15}$$

Combining (A.6), (A.7), (A.8), and (A.15), we conclude that $E(C_T(\mathbf{w})) = (T-\bar{p})E(L_T(\mathbf{w}))$, completing the proof.

A3 Proof of Theorem 2

Similar to (A.6), we write and expand the sum of squared leave-h-out cross-validation residuals as:

$$\begin{aligned}
&\mathrm{tr}\left((\mathbf{Y}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}(\mathbf{Y}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))'\right) \\
&\quad = \mathrm{tr}\left((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}) + \mathbf{e}_h)\,\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\,((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}) + \mathbf{e}_h)'\right) \\
&\quad = \mathrm{vec}((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right)\mathrm{vec}((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))') \\
&\qquad + \mathrm{vec}(\mathbf{e}_h')'\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right)\mathrm{vec}(\mathbf{e}_h') \\
&\qquad + 2\mathrm{vec}((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right)\mathrm{vec}(\mathbf{e}_h'), \qquad \text{(A.16)}
\end{aligned}$$

where the first term on the right-hand side of (A.16) corresponds the leave-h-out in-sample squared error and the second term does not involve $\mathbf{w}$. Writing $\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}) = \boldsymbol{\mu}_h - \widetilde{\mathbf{P}}_h(\mathbf{w})(\boldsymbol{\mu}_h + \mathbf{e}_h) = (\mathbf{I}_{T-\bar{p}-h+1} - \widetilde{\mathbf{P}}_h(\mathbf{w}))\boldsymbol{\mu}_h - \widetilde{\mathbf{P}}_h(\mathbf{w})\mathbf{e}_h$, we further decompose the third term on the right-hand side of (A.16) into:

$$\begin{aligned}
&2\mathrm{vec}((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right)\mathrm{vec}(\mathbf{e}_h') \\
&\quad = 2\mathrm{vec}(\boldsymbol{\mu}_h{}')'\left((\mathbf{I}_{T-\bar{p}-h+1} - \widetilde{\mathbf{P}}_h(\mathbf{w}))' \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right)\mathrm{vec}(\mathbf{e}_h') \\
&\qquad - 2\mathrm{vec}(\mathbf{e}_h')'(\widetilde{\mathbf{P}}_h(\mathbf{w})' \otimes \mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right)\mathrm{vec}(\mathbf{e}_h') \\
&\quad \equiv \widetilde{r}_{1Th}(\mathbf{w}) + \widetilde{r}_{2Th}(\mathbf{w}). \qquad \text{(A.17)}
\end{aligned}$$

We now examine the $\widetilde{r}_{1Th}(\mathbf{w})$ and $\widetilde{r}_{2Th}(\mathbf{w})$ terms as follows. Using the similar arguments to those used in the proof of Theorem 1, specifically (A.7) and (A.8), with $\boldsymbol{\mu}_h$, $\widetilde{\boldsymbol{\mu}}_h$, $\widetilde{\mathbf{P}}_h(\mathbf{w})$, $\mathbf{I}_{T-\bar{p}-h+1}$, $\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})$, and $\mathbf{e}_h$ in place of $\boldsymbol{\mu}$, $\widehat{\boldsymbol{\mu}}$, $\mathbf{P}(\mathbf{w})$, $\mathbf{I}_{T-\bar{p}}$, $\widetilde{\boldsymbol{\Sigma}}(\bar{p})$, and $\mathbf{e}$, respectively, it can be shown that, similar to the $r_{1T}(\mathbf{w})$ term in (A.7), $\widetilde{r}_{1Th}(\mathbf{w})$ is a weighted sum of mean-zero normal random variables, i.e., $E(\widetilde{r}_{1Th}(\mathbf{w})) = 0$ as $T \to \infty$.

Turning to the $\widetilde{r}_{2Th}(\mathbf{w})$ term, we have $E(\widetilde{r}_{2Th}(\mathbf{w})) = E(\mathrm{tr}(\widetilde{\mathbf{P}}_h(\mathbf{w})\mathbf{e}_h\mathbf{e}_h')) = \mathrm{tr}(\widetilde{\mathbf{P}}_h(\mathbf{w})E(\mathbf{e}_h\mathbf{e}_h')) = 0$ since for any given $h \geq 1$ and a particular t, the $(t - \bar{p} + 1)$-th row of $\widetilde{\mathbf{P}}_h(\mathbf{w})$ has ℓ_{ht} zero elements (corresponding to $\underline{\ell}_{ht} - \bar{p} + 1, \ldots, \bar{\ell}_{ht} - \bar{p} + 1$ columns) and non-zero elements elsewhere. Conversely, the matrix $E(\mathbf{e}_h\mathbf{e}_h')$ has an exactly opposite non-zero/zero structure to $\widetilde{\mathbf{P}}_h(\mathbf{w})$, and, as a result, the element-wise multiplication of the same rows of $\widetilde{\mathbf{P}}_h(\mathbf{w})$ and $E(\mathbf{e}_h\mathbf{e}_h')$ is always zero and $E(\mathbf{e}_h\mathbf{e}_h')$ is symmetric. This completes the proof.

A4 Proof of Theorem 3

In the sequel we use C to denote a generic positive constant that is independent of the sample size and may be different in different places. Specifically, we begin with the

observation that:

$$C_T^*(\mathbf{w}) = L_T(\mathbf{w}) + K + \frac{2}{T-\bar{p}}\text{vec}(\boldsymbol{\mu}')'\left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(\mathbf{w}))' \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}') \\ - \frac{2}{T-\bar{p}}\text{vec}(\mathbf{e}')'(\mathbf{P}(\mathbf{w})' \otimes \mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}') + \frac{2K^2\mathbf{p}'\mathbf{w}}{T-\bar{p}}. \tag{A.18}$$

Based on (A.18), to prove (5.16) we need to verify the following two uniform convergence results of the form:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\text{vec}(\boldsymbol{\mu}')'\left(\mathbf{P}(\mathbf{w}) \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}')\right| / V_T(\mathbf{w}) = o_p(1), \tag{A.19}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\text{vec}(\mathbf{e}')'(\mathbf{P}(\mathbf{w})' \otimes \mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}') - K^2\mathbf{p}'\mathbf{w}\right| / V_T(\mathbf{w}) = o_p(1). \tag{A.20}$$

To verify (A.19) and (A.20), we show the following statements:

$$T\lambda_{\max}\left(\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\right) = O_p(1), \tag{A.21}$$

$$T^{-1}\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}' \otimes \mathbf{I}_K\right)\text{vec}(\mathbf{e}') = O_p(1), \tag{A.22}$$

where $\lambda_{\max}(\mathbf{A})$ denotes the maximum eigenvalue of a matrix $\mathbf{A}$.

We now take (A.21). Denote $\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) = (T-\bar{p})^{-1}\sum_{t=\bar{p}+1}^{T}\mathbf{z}_t(\bar{p})\mathbf{z}_t(\bar{p})' = (T-\bar{p})^{-1}\overline{\mathbf{Z}}'\overline{\mathbf{Z}}$ and $\boldsymbol{\Gamma}(\bar{p}) = E(\mathbf{z}_{T+1}(\bar{p})\mathbf{z}_{T+1}(\bar{p})')$ is a $\bar{p}K \times \bar{p}K$ matrix whose (i,j)-th $(K\times K)$ block of elements is $\boldsymbol{\Gamma}_{i-j}$, $i,j = 1,\ldots,\bar{p}$ with $\boldsymbol{\Gamma}_j = E(\mathbf{y}_t\mathbf{y}_{t+j}')$. We also denote $\|\mathbf{A}\|_1^2 = \lambda_{\max}(\mathbf{A}'\mathbf{A})$ as the maximum eigenvalue of the matrix $\mathbf{A}'\mathbf{A}$ and $\|\mathbf{A}\|_1^2 = \lambda_{\max}^2(\mathbf{A})$ if the matrix $\mathbf{A}$ is symmetric. The following lemma places the moment bound on $\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\|_1$.

Lemma 2. *Suppose that either (1) Assumptions 1(a)-(b) and 2(b)-(d) or (2) Assumptions 1(a) and 2(b)-(c) and (e) are satisfied. Thus, $E\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\|_1 = O(1)$ for sufficiently large T.*

Proof. The proof is similar to that of Theorem 2 of Ing and Wei (2003) in the context of univariate autoregressions. To begin with, according to Lewis and Reinsel (1985, p.397), we have:

$$E\left(\left\|\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right\|_1^2\right) \le E\left(\left\|\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right\|^2\right) \le C\frac{\bar{p}^2K^2}{T-\bar{p}} = C\frac{\bar{p}^2}{T-\bar{p}}, \tag{A.23}$$

where the first inequality holds by $\|\mathbf{A}\|_1^2 = \lambda_{\max}^2(\mathbf{A}) \le \sum_{\ell=1}^{m}\lambda_\ell^2(\mathbf{A}) = \|\mathbf{A}\|^2$ for a $m\times m$ symmetric matrix $\mathbf{A}$ with eigenvalues λ_ℓ, $\ell = 1,\ldots,m$.

We next observe that:

$$\left\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p}) - \boldsymbol{\Gamma}^{-1}(\bar{p})\right\|_1 = \left\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\left(\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right)\boldsymbol{\Gamma}^{-1}(\bar{p})\right\|_1 \\ \le \left\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\right\|_1\left\|\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right\|_1\left\|\boldsymbol{\Gamma}^{-1}(\bar{p})\right\|_1, \tag{A.24}$$

almost surely for large T, and hence we can write for sufficiently large T and any $\theta > 0$:

$$\begin{aligned} E\left(\left\|\widehat{\mathbf{\Gamma}}_T^{-1}(\bar{p}) - \mathbf{\Gamma}^{-1}(\bar{p})\right\|_1\right) &\leq \bar{p}^{2+\theta} E\left(\left\|\widehat{\mathbf{\Gamma}}_T(\bar{p}) - \mathbf{\Gamma}(\bar{p})\right\|_1\right) C \\ &\leq C\frac{\bar{p}}{(T-\bar{p})^{1/2}}\bar{p}^{2+\theta} \\ &= C\frac{\bar{p}^{3+\theta}}{(T-\bar{p})^{1/2}} \\ &\leq C\left(\frac{\bar{p}^{6+\delta_1}}{T-\bar{p}}\right)^{1/2}, \end{aligned} \tag{A.25}$$

where the first inequality uses Assumption 2(c). As in the univariate case (see, e.g., Berk (1974, p.491)), $\|\mathbf{\Gamma}^{-1}(p)\|_1$ is uniformly bounded by a positive constant for all $1 \leq p \leq \bar{p}$, as stated in Lewis and Reinsel (1985, p.397), the second inequality follows from (A.23), and the last inequality follows by setting $2\theta \leq \delta_1$.

Based on (A.25), it is not hard to see $E\|\widehat{\mathbf{\Gamma}}_T^{-1}(\bar{p})\|_1 \leq C$, provided that $\bar{p}^{6+\delta_1} = O(T)$ and $E\|\mathbf{\Gamma}^{-1}(\bar{p})\|_1 \leq C$. This completes the proof of the lemma under the first set of assumptions.

Using the similar arguments to those employed in Ing and Wei (2003, p.140), one can show that the statement of the lemma still holds under the second set of assumptions. The proof is omitted for brevity. □

Using Lemma 2, we obtain:

$$E\left\|\widehat{\mathbf{\Gamma}}_T^{-1}(\bar{p})\right\|_1 \equiv E\left\|\left(\frac{\overline{\mathbf{Z}}'\overline{\mathbf{Z}}}{T-\bar{p}}\right)^{-1}\right\|_1 = E\left[\lambda_{\max}\left(\left(\frac{\overline{\mathbf{Z}}'\overline{\mathbf{Z}}}{T-\bar{p}}\right)^{-1}\right)\right] < \infty, \tag{A.26}$$

where the second equality in (A.26) follows since the matrix $\overline{\mathbf{Z}}'\overline{\mathbf{Z}}$ is symmetric. Thus, (A.21) follows from combining (A.26) and Markov's inequality. Next, note that $\text{vec}(\mathbf{e}')'(\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}' \otimes \mathbf{I}_K)\text{vec}(\mathbf{e}') = \text{tr}(\mathbf{e}'\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\mathbf{e})$. We show (A.22) in the following lemma.

Lemma 3. *Under assumptions that the second moment of $\boldsymbol{\varepsilon}_t$ exists for all t and that $E\left(|y_{i,t-\ell}y_{j,t-\ell}|\right) \leq C$ for all $i, j = 1, \ldots, K$, t and ℓ, we have:*

$$E\left\|\frac{1}{\sqrt{T-\bar{p}}}\sum_{t=\bar{p}+1}^{T}\boldsymbol{\varepsilon}_t\mathbf{z}_t(\bar{p})'\right\|^2 \equiv \frac{1}{T-\bar{p}}E\left[tr(\mathbf{e}'\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\mathbf{e})\right] = O(\bar{p}_T).$$

Proof. Recall that $\mathbf{z}_t(p)' = (\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}')$ and observe that:

$$E\left\|\frac{1}{\sqrt{T-\bar{p}}}\sum_{t=\bar{p}+1}^{T}\boldsymbol{\varepsilon}_t\mathbf{z}_t(\bar{p})'\right\|^2 \leq \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{\ell=1}^{\bar{p}}E\left[(T-\bar{p})^{-1}\left|\sum_{t=\bar{p}+1}^{T}\varepsilon_{it}y_{j,t-\ell}\right|^2\right]. \tag{A.27}$$

Since $E(|\varepsilon_{it}\varepsilon_{jt}|) \leq C$ for $i, j = 1, \ldots, K$ and all t, the summand $E\left[(T-\bar{p})^{-1}\left|\sum_{t=\bar{p}+1}^{T}\varepsilon_{it}y_{j,t-\ell}\right|^2\right]$

in (A.27) is bounded by:

$$CE\left((T-\bar{p})^{-1}\sum_{t=\bar{p}+1}^{T}|y_{j,t-\ell}y_{j',t-\ell}|\right)=C(T-\bar{p})^{-1}\sum_{t=\bar{p}+1}^{T}E\left(|y_{j,t-\ell}y_{j',t-\ell}|\right)=O(1). \quad \text{(A.28)}$$

Lastly, the lemma follows from combining (A.27), (A.28), and the condition that $E\left(|y_{j,t-\ell}y_{j',t-\ell}|\right)\leq C$ for all j, j', t and ℓ, where the last condition follows from the assumption of $\sum_{j=0}^{\infty}\|\mathbf{\Phi}_j\|<\infty$. This yields the desired result.

□

Equipped with (A.21) and (A.22), we first verify (A.19) by writing:

$$\begin{aligned}
&\xi_T^{*-1}\left|\text{vec}(\boldsymbol{\mu}')'\left(\mathbf{P}(\mathbf{w})\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}')\right| \\
&=\xi_T^{*-1}\left|\text{vec}(\boldsymbol{\mu}')'\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\left(\mathbf{P}(\mathbf{w})\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}')\right| \\
&\leq\xi_T^{*-1}\left\{\text{vec}(\boldsymbol{\mu}')'\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\text{vec}(\boldsymbol{\mu}')\text{vec}(\mathbf{e}')'\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\left(\mathbf{P}(\mathbf{w})\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}')\right\}^{1/2} \\
&\leq\xi_T^{*-1}\left\{\text{vec}(\boldsymbol{\mu}')'\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\text{vec}(\boldsymbol{\mu}')\text{vec}(\mathbf{e}')'\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}')\right\}^{1/2} \\
&=\left\{\underbrace{\left[\bar{p}\xi_T^{*-2}\text{vec}(\boldsymbol{\mu}')'\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\text{vec}(\boldsymbol{\mu}')\right]}_{=o_p(1)\text{ by Assumption 2(a)}}\left[\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}')\right]\right\}^{1/2} \\
&=o_p(1), \quad \text{(A.29)}
\end{aligned}$$

where the first inequality follows from the Schwarz inequality, and the last equality holds since:

$$\begin{aligned}
&\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}}\otimes\mathbf{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}') \\
&\leq C\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\text{vec}(\mathbf{e}') \\
&\leq\underbrace{C\,T\lambda_{\max}\left(\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\right)}_{=O_p(1)\text{ by (A.21)}}\underbrace{T^{-1}\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\otimes\mathbf{I}_K\right)\text{vec}(\mathbf{e}')}_{=O_p(1)\text{ by (A.22)}} \\
&=O_p(1). \quad \text{(A.30)}
\end{aligned}$$

The second inequality in (A.30) follows from the fact that $\text{vec}(\mathbf{e}')'\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\text{vec}(\mathbf{e}')=\text{tr}\left(\mathbf{e}'\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{e}\right)=\text{tr}\left((\overline{\mathbf{Z}}'\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{e}\mathbf{e}'\overline{\mathbf{Z}}\right)$ and the trace inequality, whereby setting $\mathbf{A}=(\overline{\mathbf{Z}}'\overline{\mathbf{Z}})^{-1}$ and $\mathbf{B}=\mathbf{e}'\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\mathbf{e}$, we have $\text{tr}(\mathbf{AB})\leq\lambda_{\max}(\mathbf{A})\text{tr}(\mathbf{B})$ for squared matrices $\mathbf{A}$ and $\mathbf{B}$ with $\mathbf{A}$ being symmetric and $\mathbf{B}\geq 0$.

We next move to (A.20). Using Assumption 2(a) and a similar argument to show (A.30),

we have:

$$\begin{aligned}
&\xi_T^{*\,-1}\left|\mathrm{vec}(\mathbf{e}')'(\mathbf{P}(\mathbf{w})'\otimes\mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathrm{vec}(\mathbf{e}')-K^2\mathbf{p}'\mathbf{w}\right| \\
&\leq \xi_T^{*\,-1}\mathrm{vec}(\mathbf{e}')'(\overline{\mathbf{P}}\otimes\mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathrm{vec}(\mathbf{e}')+\underbrace{\xi_T^{*\,-1}K^2\bar{p}}_{\substack{=o_p(1)\\ \text{by Assumption 2}(a)}} \\
&\leq \underbrace{\xi_T^{*\,-1}\bar{p}}_{\substack{=o_p(1)\\ \text{by Assumption 2}(a)}}\underbrace{T\lambda_{\max}\left(\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\right)}_{=O_p(1)\text{ by (A.21)}}\underbrace{T^{-1}\bar{p}^{-1}\mathrm{vec}(\mathbf{e}')'\left(\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\otimes\mathbf{I}_K\right)\mathrm{vec}(\mathbf{e}')}_{=O_p(1)\text{ by (A.22)}}+o_p(1) \\
&= o_p(1).
\end{aligned} \tag{A.31}$$

The last thing to show is (5.15). The argument is essentially the same as the above. Recall that $\widehat{\boldsymbol{\mu}}(\mathbf{w})=\mathbf{P}(\mathbf{w})\mathbf{Y}$ and $\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w})=\mathbf{A}(\mathbf{w})\boldsymbol{\mu}-\mathbf{P}(\mathbf{w})\mathbf{e}$. We first calculate:

$$\begin{aligned}
L_T(\mathbf{w}) &= \frac{1}{T-\bar{p}}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w})\right)'\left(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w})\right)\right) \\
&= \frac{1}{T-\bar{p}}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{A}(\mathbf{w})\mathbf{A}(\mathbf{w})\boldsymbol{\mu}\right)-2\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{A}(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\right)+\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\right)\right],
\end{aligned} \tag{A.32}$$

and thus:

$$\begin{aligned}
L_T(\mathbf{w})-V_T(\mathbf{w}) = &-\frac{2}{T-\bar{p}}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{A}(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\right) \\
&+\frac{1}{T-\bar{p}}\left\{\mathrm{tr}\left(\mathbf{P}(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}(\mathbf{w})\right)-E\left[\mathrm{tr}\left(\mathbf{P}(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}(\mathbf{w})\right)\right]\right\}.
\end{aligned} \tag{A.33}$$

Take the first term on the right-hand side of (A.33). We calculate:

$$\begin{aligned}
\xi_T^{*\,-1}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{A}(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\right) &\leq \xi_T^{*\,-1}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{P}(\mathbf{w})\mathbf{e}\right)+\xi_T^{*\,-1}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\mathrm{tr}\left(\boldsymbol{\mu}'\mathbf{P}(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\right) \\
&\leq \xi_T^{*\,-1}\lambda_{\max}\left(\boldsymbol{\Sigma}^{-1}\right)\mathrm{tr}\left(\boldsymbol{\mu}'\mathbf{P}(\mathbf{w})\mathbf{e}\right)+\xi_T^{*\,-1}\lambda_{\max}\left(\boldsymbol{\Sigma}^{-1}\right)\lambda_{\max}\left(\mathbf{P}(\mathbf{w})\right)\mathrm{tr}\left(\boldsymbol{\mu}'\mathbf{P}(\mathbf{w})\mathbf{e}\right) \\
&\leq C\xi_T^{*\,-1}\mathrm{tr}\left(\boldsymbol{\mu}'\mathbf{P}(\mathbf{w})\mathbf{e}\right) \\
&= C\xi_T^{*\,-1}\mathrm{tr}\left(\boldsymbol{\mu}'\overline{\mathbf{P}}\mathbf{P}(\mathbf{w})\mathbf{e}\right) \\
&= C\xi_T^{*\,-1}\mathrm{vec}(\overline{\mathbf{P}}\boldsymbol{\mu}')'\mathrm{vec}(\mathbf{P}(\mathbf{w})\mathbf{e}) \\
&\leq C\xi_T^{*\,-1}\left[\mathrm{vec}(\overline{\mathbf{P}}\boldsymbol{\mu}')'\mathrm{vec}(\overline{\mathbf{P}}\boldsymbol{\mu}')\right]^{1/2}\left[\mathrm{vec}(\mathbf{P}(\mathbf{w})\mathbf{e})'\mathrm{vec}(\mathbf{P}(\mathbf{w})\mathbf{e})\right]^{1/2} \\
&= C\underbrace{\left[\bar{p}\xi_T^{*\,-2}\mathrm{tr}\left(\boldsymbol{\mu}\overline{\mathbf{P}}\boldsymbol{\mu}'\right)\right]^{1/2}}_{=o_p(1)\text{ by Assumption 2}(a)}\underbrace{\left[\bar{p}^{-1}\mathrm{tr}\left(\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\right)\right]^{1/2}}_{=O_p(1)} \\
&= o_p(1),
\end{aligned} \tag{A.34}$$

where the first and second inequalities follow from the trace inequality, and the fourth inequality follows from the Schwarz inequality. Using (A.21) and (A.22), the second part on

the right-hand side of the third equality in (A.34) holds since:

$$\begin{aligned}
\left(\bar{p}^{-1}\text{tr}\left(\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\right)\right)^{1/2} &\leq \left(\bar{p}^{-1}\lambda_{\max}\left(\mathbf{P}(\mathbf{w})\right)\text{tr}\left(\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{e}\right)\right)^{1/2} \\
&= C\left(\bar{p}^{-1}\text{tr}\left(\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{e}\right)\right)^{1/2} \\
&\leq C\left(\bar{p}^{-1}\text{tr}\left(\mathbf{e}'\overline{\mathbf{P}}\mathbf{e}\right)\right)^{1/2} \\
&\leq \left(\underbrace{T\lambda_{\max}\left(\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\right)}_{=O_p(1)\text{ by (A.21)}}\underbrace{T^{-1}\bar{p}^{-1}\text{tr}\left(\mathbf{e}'\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\mathbf{e}\right)}_{=O_p(1)\text{ by (A.22)}}\right)^{1/2} \\
&= O_p(1). \qquad \text{(A.35)}
\end{aligned}$$

We next take the second term on the right-hand side of (A.33). Using (A.35) and Assumption 2(a), we have:

$$\begin{aligned}
&{\xi_T^*}^{-1}\left\{\text{tr}\left(\mathbf{P}(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}(\mathbf{w})\right) - E\left[\text{tr}\left(\mathbf{P}(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}(\mathbf{w})\right)\right]\right\} \\
&\quad \leq C{\xi_T^*}^{-1}\underbrace{\text{tr}\left(\mathbf{e}'\mathbf{P}(\mathbf{w})\mathbf{P}(\mathbf{w})\mathbf{e}\right)}_{=O_p(\bar{p})\text{ by (A.35)}} + C{\xi_T^*}^{-1}\underbrace{\text{tr}\left(\mathbf{P}(\mathbf{w})E(\mathbf{e}\mathbf{e}')\mathbf{P}(\mathbf{w})\right)}_{=O(\bar{p})} \\
&\quad = O_p\left({\xi_T^*}^{-1}\bar{p}\right) + O\left({\xi_T^*}^{-1}\bar{p}\right) \\
&\quad = o_p(1), \qquad \text{(A.36)}
\end{aligned}$$

where the last equality follows from Assumption 2(a).

A5 Proof of Theorem 4

As discussed in (5.18)-(5.20) in the text, to prove (5.17), it suffices to show (5.19) and (5.20), where (5.19) is implied by (5.18). First take (5.18). Based on the following decomposition of $CV_{T,h}^*(\mathbf{w})$:

$$\begin{aligned}
CV_{T,h}^*(\mathbf{w}) &\equiv CV_{T,h}(\mathbf{w})/(T-\bar{p}-h+1) \\
&= \widetilde{L}_{T,h}(\mathbf{w}) + K \\
&\quad + \frac{2}{T-\bar{p}-h+1}\text{vec}(\boldsymbol{\mu}_h')'\left(\left(\mathbf{I}_{T-\bar{p}-h+1} - \widetilde{\mathbf{P}}_h(\mathbf{w})\right)' \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\text{vec}(\mathbf{e}_h') \\
&\quad - \frac{2}{T-\bar{p}-h+1}\text{vec}(\mathbf{e}_h')'\left(\widetilde{\mathbf{P}}_h(\mathbf{w})' \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\text{vec}(\mathbf{e}_h') \\
&\quad + \frac{2}{T-\bar{p}-h+1}\text{vec}\left(\left(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w})\right)'\right)'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\text{vec}(\mathbf{e}_h'), \qquad \text{(A.37)}
\end{aligned}$$

to establish the first condition in (5.18), it is sufficient to show the following uniform convergence results:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\mathrm{vec}(\boldsymbol{\mu}_h')'\left(\widetilde{\mathbf{P}}_h(\mathbf{w})'\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})=o_p(1), \tag{A.38}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\mathrm{vec}(\mathbf{e}_h')'(\widetilde{\mathbf{P}}_h(\mathbf{w})'\otimes\mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})=o_p(1), \tag{A.39}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\mathrm{vec}((\boldsymbol{\mu}_h-\widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})=o_p(1), \tag{A.40}$$

where we have replaced $\widetilde{V}_{T,h}(\mathbf{w})$ with $V_{T,h}(\mathbf{w})$ in the denominator of (A.38)-(A.40) under the condition: $\sup_{\mathbf{w}\in\mathcal{H}_T}|\widetilde{V}_{T,h}(\mathbf{w})/V_{T,h}(\mathbf{w})-1|\xrightarrow{p}0$, which will be established in (A.45) below.

Under Assumption 3(a), (A.38) and (A.39) can be shown to hold by using similar arguments to those in (A.29), (A.30), and (A.31) under the conditions (A.21) and (A.22) with ξ_T^*, $\overline{\mathbf{Z}}$, $\mathbf{e}$, $\boldsymbol{\mu}$, $\overline{\mathbf{P}}$, $\mathbf{I}_{T-\bar{p}}$, and $\boldsymbol{\Sigma}^{-1}$ replaced by $\xi_{T,h}^*$, $\overline{\mathbf{Z}}_h$, $\mathbf{e}_h$, $\boldsymbol{\mu}_h$, $\overline{\mathbf{P}}_h$, $\mathbf{I}_{T-\bar{p}-h+1}$, and $\boldsymbol{\Sigma}_h^{-1}$, respectively. Next turn to (A.40). Using $\widetilde{\boldsymbol{\mu}}_h(\mathbf{w})=\widetilde{\mathbf{P}}_h(\mathbf{w})(\boldsymbol{\mu}_h+\mathbf{e}_h)$ and ignoring the term that does not involve $\mathbf{w}$, we only need to show:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\mathrm{vec}((\boldsymbol{\mu}_h'\widetilde{\mathbf{P}}_h(\mathbf{w})')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})=o_p(1), \tag{A.41}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\mathrm{vec}((\mathbf{e}_h'\widetilde{\mathbf{P}}_h(\mathbf{w})')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})=o_p(1). \tag{A.42}$$

Take (A.41). Using Lemma 1, we rewrite $\widetilde{\mathbf{P}}_h(p)=\widetilde{\mathbf{D}}_h(p)(\mathbf{P}_h(p)-\mathbf{I}_{T-\bar{p}-h+1})+\mathbf{I}_{T-\bar{p}-h+1}$ as $\widetilde{\mathbf{P}}_h(p)=\mathbf{P}_h(p)+\mathbf{T}_h(p)-\mathbf{Q}_h(p)$, where $\mathbf{Q}_h(p)=\widetilde{\mathbf{D}}_h(p)-\mathbf{I}_{T-\bar{p}-h+1}$ and $\mathbf{T}_h(p)=\mathbf{Q}_h(p)\mathbf{P}_h(p)$. As a result, we have $\widetilde{\mathbf{P}}_h(\mathbf{w})=\sum_{p=1}^{\bar{p}}w(p)\widetilde{\mathbf{P}}_h(p)=\mathbf{P}_h(\mathbf{w})+\mathbf{T}_h(\mathbf{w})-\mathbf{Q}_h(\mathbf{w})$, where $\mathbf{T}_h(\mathbf{w})=\sum_{p=1}^{\bar{p}}w(p)\mathbf{T}_h(p)$ and $\mathbf{Q}_h(\mathbf{w})$ is defined analogously. Using this, we rewrite the left-hand side of (A.41) as:

$$\begin{aligned}
&\left|\mathrm{vec}((\boldsymbol{\mu}_h'\widetilde{\mathbf{P}}_h(\mathbf{w})')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})\\
&\le\left|\mathrm{vec}((\boldsymbol{\mu}_h'\mathbf{P}_h(\mathbf{w})')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})\\
&\quad+\left|\mathrm{vec}((\boldsymbol{\mu}_h'\mathbf{T}_h(\mathbf{w})')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})\\
&\quad+\left|\mathrm{vec}((\boldsymbol{\mu}_h'\mathbf{Q}_h(\mathbf{w})')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\mathrm{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})\\
&\le o_p(1)+C\xi_{T,h}^{*-1}q_h^*\mathrm{tr}(\boldsymbol{\mu}_h\mathbf{e}_h')\\
&=o_p(1)+C(\bar{p}\xi_{T,h}^{*-1})(\bar{p}^{-1}Tq_h^*)(T^{-1}\mathrm{tr}(\boldsymbol{\mu}_h\mathbf{e}_h'))\\
&=o_p(1),
\end{aligned} \tag{A.43}$$

where the second inequality in (A.43) follows from using the identical arguments to those in (A.29) with ξ_T^*, $\boldsymbol{\mu}$, $\mathbf{P}(\mathbf{w})$, $\overline{\mathbf{P}}$, $\mathbf{I}_{T-\bar{p}}$, $\boldsymbol{\Sigma}$, and $\mathbf{e}$ replaced by $\xi_{T,h}^*$, $\boldsymbol{\mu}_h$, $\mathbf{P}_h(\mathbf{w})$, $\overline{\mathbf{P}}_h$, $\mathbf{I}_{T-\bar{p}-h+1}$, $\boldsymbol{\Sigma}_h$, and $\mathbf{e}_h$, respectively, and under Assumption 3(a); the last equality follows from Assumption 3 and from $\mathrm{tr}(\boldsymbol{\mu}_h\mathbf{e}_h')=\sum_{t=\bar{p}}^{T-h}\sum_{k=1}^{K}\mu_{kt}^h\epsilon_{k,t+h}=O_p(T)$ under the conditions $E(|\varepsilon_{it}\varepsilon_{jt}|)=O(1)$

for $i, j = 1, \ldots, K$ and $E(|y_{j,t-\ell} y_{j',t-\ell}|) = O(1)$ for all j, j', t, and ℓ.

Turning to (A.42), once again using $\widetilde{\mathbf{P}}_h(p) = \mathbf{P}_h(p) + \mathbf{T}_h(p) - \mathbf{Q}_h(p)$, we can write:

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_T} \left|\text{vec}((\mathbf{e}_h'\widetilde{\mathbf{P}}_h(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \text{vec}(\mathbf{e}_h')\right| / V_{T,h}(\mathbf{w}) \\
&\quad \leq \sup_{\mathbf{w}\in\mathcal{H}_T} \left|\text{vec}((\mathbf{e}_h'\mathbf{P}_h(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \text{vec}(\mathbf{e}_h')\right| / V_{T,h}(\mathbf{w}) \\
&\qquad + \sup_{\mathbf{w}\in\mathcal{H}_T} \left|\text{vec}((\mathbf{e}_h'(\mathbf{T}_h(\mathbf{w}) - \mathbf{Q}_h(\mathbf{w}))')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \text{vec}(\mathbf{e}_h')\right| / V_{T,h}(\mathbf{w}) \\
&\quad \leq o_p(1) + \xi_{T,h}^{*-1} \sup_{\mathbf{w}\in\mathcal{H}_T} \sum_{p=1}^{\bar{p}} w(p) \left|\text{vec}((\mathbf{e}_h'(\mathbf{T}_h(p) - \mathbf{Q}_h(p))')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \text{vec}(\mathbf{e}_h')\right| \\
&\quad \leq o_p(1) + C\xi_{T,h}^{*-1} q_h^* \text{tr}(\mathbf{e}_h\mathbf{e}_h') = o_p(1) + C(\bar{p}\xi_{T,h}^{*-1})(\bar{p}^{-1} T q_h^*)(T^{-1}\text{tr}(\mathbf{e}_h\mathbf{e}_h')) \\
&= o_p(1),
\end{aligned} \tag{A.44}
$$

where the second inequality follows from the arguments showing (A.30) with $\mathbf{e}$, $\overline{\mathbf{P}}$, $\mathbf{I}_{T-\bar{p}}$, and $\boldsymbol{\Sigma}$ replaced by $\mathbf{e}_h$, $\overline{\mathbf{P}}_h$, $\mathbf{I}_{T-\bar{p}-h+1}$, and $\boldsymbol{\Sigma}_h$, respectively, under suitable conditions stated in Assumption 3, and the last equality is satisfied by Assumption 3 and by the fact that $\text{tr}(\mathbf{e}_h\mathbf{e}_h') = O_p(T)$ under once again $E(|\varepsilon_{it}\varepsilon_{jt}|) = O(1)$ for $i, j = 1, \ldots, K$.

It now remains to establish $\sup_{\mathbf{w}\in\mathcal{H}_T} |\widetilde{L}_{T,h}(\mathbf{w})/L_{T,h}(\mathbf{w}) - 1| \xrightarrow{p} 0$ as $T \to \infty$. To prove this, we first show:

$$
\sup_{\mathbf{w}\in\mathcal{H}_T} \left|\frac{\widetilde{V}_{T,h}(\mathbf{w})}{V_{T,h}(\mathbf{w})} - 1\right| \to 0 \tag{A.45}
$$

almost surely as $T \to \infty$. Define $\widetilde{\mathbf{A}}_h(\mathbf{w}) = \mathbf{I}_{T-\bar{p}-h+1} - \widetilde{\mathbf{P}}_h(\mathbf{w})$. Using $\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}) = \widetilde{\mathbf{A}}_h(\mathbf{w})\boldsymbol{\mu}_h - \widetilde{\mathbf{P}}_h(\mathbf{w})\mathbf{e}_h$, the leave-$h$-out risk $\widetilde{V}_{T,h}(\mathbf{w})$ for averaging h-step forecasts is given by:

$$
\begin{aligned}
\widetilde{V}_{T,h}(\mathbf{w}) &= E(\widetilde{L}_{T,h}(\mathbf{w})) \\
&= \frac{1}{T-\bar{p}-h+1} E\left[\text{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))'(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h(\mathbf{w}))\right)\right] \\
&= \frac{1}{T-\bar{p}-h+1} \text{tr}\left(\widetilde{\mathbf{A}}_h(\mathbf{w})\boldsymbol{\mu}_h\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h'\widetilde{\mathbf{A}}_h(\mathbf{w})'\right) + E\left[\text{tr}\left(\widetilde{\mathbf{P}}_h(\mathbf{w})\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h(\mathbf{w})'\right)\right].
\end{aligned} \tag{A.46}
$$

Based on (A.46), it is seen that $V_{T,h}(\mathbf{w})$ is equal to $\widetilde{V}_{T,h}(\mathbf{w})$ with $\widetilde{\mathbf{A}}_h(\mathbf{w})$ and $\widetilde{\mathbf{P}}_h(\mathbf{w})$ replaced by $\mathbf{A}_h(\mathbf{w})$ and $\mathbf{P}_h(\mathbf{w})$, respectively. As a consequence, it is sufficient to establish that for any pair of candidate models i and j, the following conditions hold:

$$
\text{tr}\left(\widetilde{\mathbf{A}}_h(i)\boldsymbol{\mu}_h\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h'\widetilde{\mathbf{A}}_h(j)'\right) = \text{tr}\left(\mathbf{A}_h(i)\boldsymbol{\mu}_h\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h'\mathbf{A}_h(j)'\right)(1+o(1)), \tag{A.47}
$$

$$
E\left[\text{tr}\left(\widetilde{\mathbf{P}}_h(i)\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h(j)'\right)\right] = E\left[\text{tr}\left(\mathbf{P}_h(i)\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\mathbf{P}_h(j)'\right)\right](1+o(1)), \tag{A.48}
$$

where the $o(1)$ terms are uniform in $1 \leq i, j \leq \bar{p}$. Using $\widetilde{\mathbf{A}}_h(i) = \mathbf{I}_{T-\bar{p}-h+1} - \widetilde{\mathbf{P}}_h(i) = \mathbf{A}_h(i) - \mathbf{T}_h(i) + \mathbf{Q}_h(i) = \mathbf{A}_h(i) + \mathbf{Q}_h(i)\mathbf{A}_h(i)$, it can be shown that $\widetilde{\mathbf{A}}_h(i) = \mathbf{A}_h(i)(1+o(1))$ since $\mathbf{Q}_h(i) = o(1)$ under Assumption 3(b). This establishes (A.47). Next take (A.48). Using

$\widetilde{\mathbf{P}}_h(p) = \widetilde{\mathbf{D}}_h(p)(\mathbf{P}_h(p) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{I}_{T-\bar{p}-h+1}$ implied by (4.11) in Lemma 1, we have:

$$\begin{aligned}
E\left[\mathrm{tr}\left(\widetilde{\mathbf{P}}_h(i)\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h(j)'\right)\right] &= \mathrm{tr}\left(\widetilde{\mathbf{P}}_h(j)'(\widetilde{\mathbf{D}}_h(i)(\mathbf{P}_h(i) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{I}_{T-\bar{p}-h+1})E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) \\
&= \mathrm{tr}\left(\widetilde{\mathbf{P}}_h(j)'\widetilde{\mathbf{D}}_h(i)\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) - \mathrm{tr}\left(\widetilde{\mathbf{P}}_h(j)'\widetilde{\mathbf{D}}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) \\
&\quad + \mathrm{tr}\left(\widetilde{\mathbf{P}}_h(j)'E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) \\
&= \mathrm{tr}\left(\widetilde{\mathbf{P}}_h(j)'\widetilde{\mathbf{D}}_h(i)\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right)(1 + o(1)) \\
&= \Big[\mathrm{tr}\left((\mathbf{P}_h(j) - \mathbf{I}_{T-\bar{p}-h+1})\widetilde{\mathbf{D}}_h(j)'\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) \\
&\quad + \mathrm{tr}\left(\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right)\Big](1 + o(1)) \\
&= \mathrm{tr}\left(\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\mathbf{P}_h(j)\right)(1 + o(1)), \qquad \text{(A.49)}
\end{aligned}$$

where the third equality follows from $\mathrm{tr}\left(\widetilde{\mathbf{P}}_h(j)'\widetilde{\mathbf{D}}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) = \mathrm{tr}\left(\widetilde{\mathbf{P}}_h(j)'E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right)(1+o(1))$ under Assumption 3(b) and from $\mathrm{tr}\left(\widetilde{\mathbf{P}}_h(j)'E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) = 0$ by the diagonal elements of $\widetilde{\mathbf{P}}_h(j)'E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')$ being zero (using the similar arguments to those used in showing $E(\widetilde{r}_{2hT}(\mathbf{w})) = 0$ in (A.17)), and the last equality holds by Assumption 3(b) again. This establishes (A.48) and thus, combined with (A.47), yields (A.45).

Second, it is straightforward to show $\sup_{\mathbf{w}\in\mathcal{H}_T}|L_{T,h}(\mathbf{w})/V_{T,h}(\mathbf{w}) - 1| \xrightarrow{p} 0$ as $T \to \infty$ by following the identical arguments to those in (A.32)-(A.36) with $L_T(\mathbf{w})$, $V_T(\mathbf{w})$, $\boldsymbol{\mu}$, $\widehat{\boldsymbol{\mu}}(\mathbf{w})$, $\boldsymbol{\Sigma}$, $\mathbf{A}(\mathbf{w})$, $\mathbf{P}(\mathbf{w})$, $\overline{\mathbf{P}}$, $\overline{\mathbf{Z}}$, $\mathbf{e}$, and ξ_T^* replaced by $L_{T,h}(\mathbf{w})$, $V_{T,h}(\mathbf{w})$, $\boldsymbol{\mu}_h$, $\widehat{\boldsymbol{\mu}}_h(\mathbf{w})$, $\boldsymbol{\Sigma}_h$, $\mathbf{A}_h(\mathbf{w})$, $\mathbf{P}_h(\mathbf{w})$, $\overline{\mathbf{P}}_h$, $\overline{\mathbf{Z}}_h$, $\mathbf{e}_h$, and $\xi_{T,h}^*$, respectively. Next, to show $\sup_{\mathbf{w}\in\mathcal{H}_T}|\widetilde{L}_{T,h}(\mathbf{w})/\widetilde{V}_{T,h}(\mathbf{w}) - 1| \xrightarrow{p} 0$ as $T \to \infty$, we first write:

$$\begin{aligned}
\widetilde{L}_{T,h}(\mathbf{w}) - \widetilde{V}_{T,h}(\mathbf{w}) = &-\frac{2}{T-\bar{p}-h+1}\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h'\widetilde{\mathbf{A}}_h(\mathbf{w})'\widetilde{\mathbf{P}}_h(\mathbf{w})\mathbf{e}_h\right) \\
&+ \frac{1}{T-\bar{p}-h+1}\left\{\mathrm{tr}\left(\widetilde{\mathbf{P}}_h(\mathbf{w})\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h(\mathbf{w})'\right) - E\left[\mathrm{tr}\left(\widetilde{\mathbf{P}}_h(\mathbf{w})\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h(\mathbf{w})'\right)\right]\right\}. \qquad \text{(A.50)}
\end{aligned}$$

Using (A.45) and similar arguments to those for proving (A.47)-(A.48), it is not hard to show $(\widetilde{L}_{T,h}(\mathbf{w}) - \widetilde{V}_{T,h}(\mathbf{w}))/\widetilde{V}_{T,h}(\mathbf{w}) = (L_{T,h}(\mathbf{w}) - V_{T,h}(\mathbf{w}))/V_{T,h}(\mathbf{w})(1+o(1))$, establishing the second condition in (5.18): $\sup_{\mathbf{w}\in\mathcal{H}_T}|\widetilde{L}_{T,h}(\mathbf{w})/\widetilde{V}_{T,h}(\mathbf{w}) - 1| \xrightarrow{p} 0$ as $T \to \infty$. Combining these above conditions implies

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{L}_{T,h}(\mathbf{w})}{L_{T,h}(\mathbf{w})} - 1\right| \le \sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{L}_{T,h}(\mathbf{w})}{\widetilde{V}_{T,h}(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{V}_{T,h}(\mathbf{w})}{V_{T,h}(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{V_{T,h}(\mathbf{w})}{L_{T,h}(\mathbf{w})}\right| - 1 \xrightarrow{p} 0 \qquad \text{(A.51)}$$

as $T \to \infty$, establishing (5.20). Putting together (A.38)-(A.40), (A.45), and (A.51) completes the proof.

A6 Tables and Figures

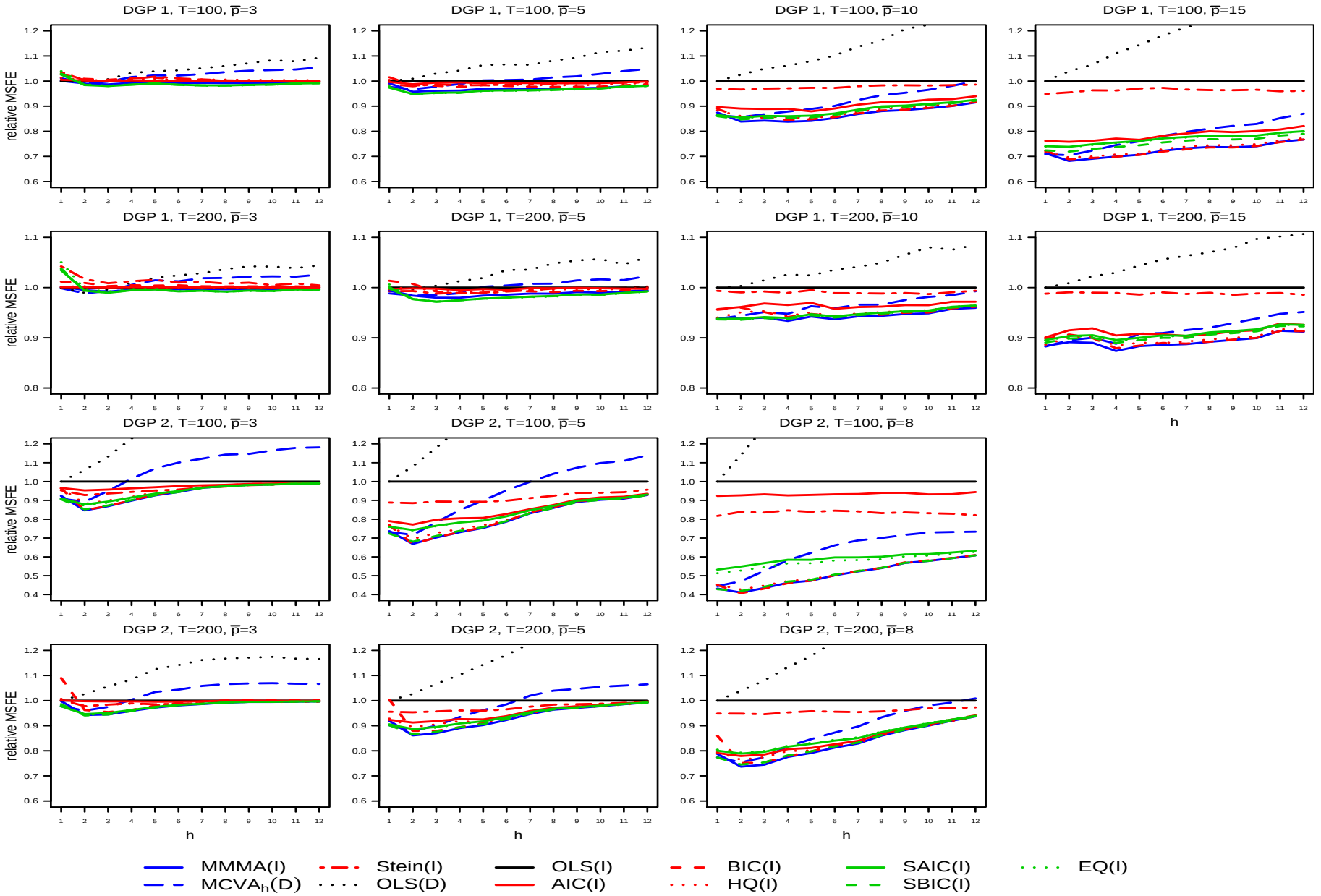


Figure A1: Multi-step forecast performance for DGPs 1-2

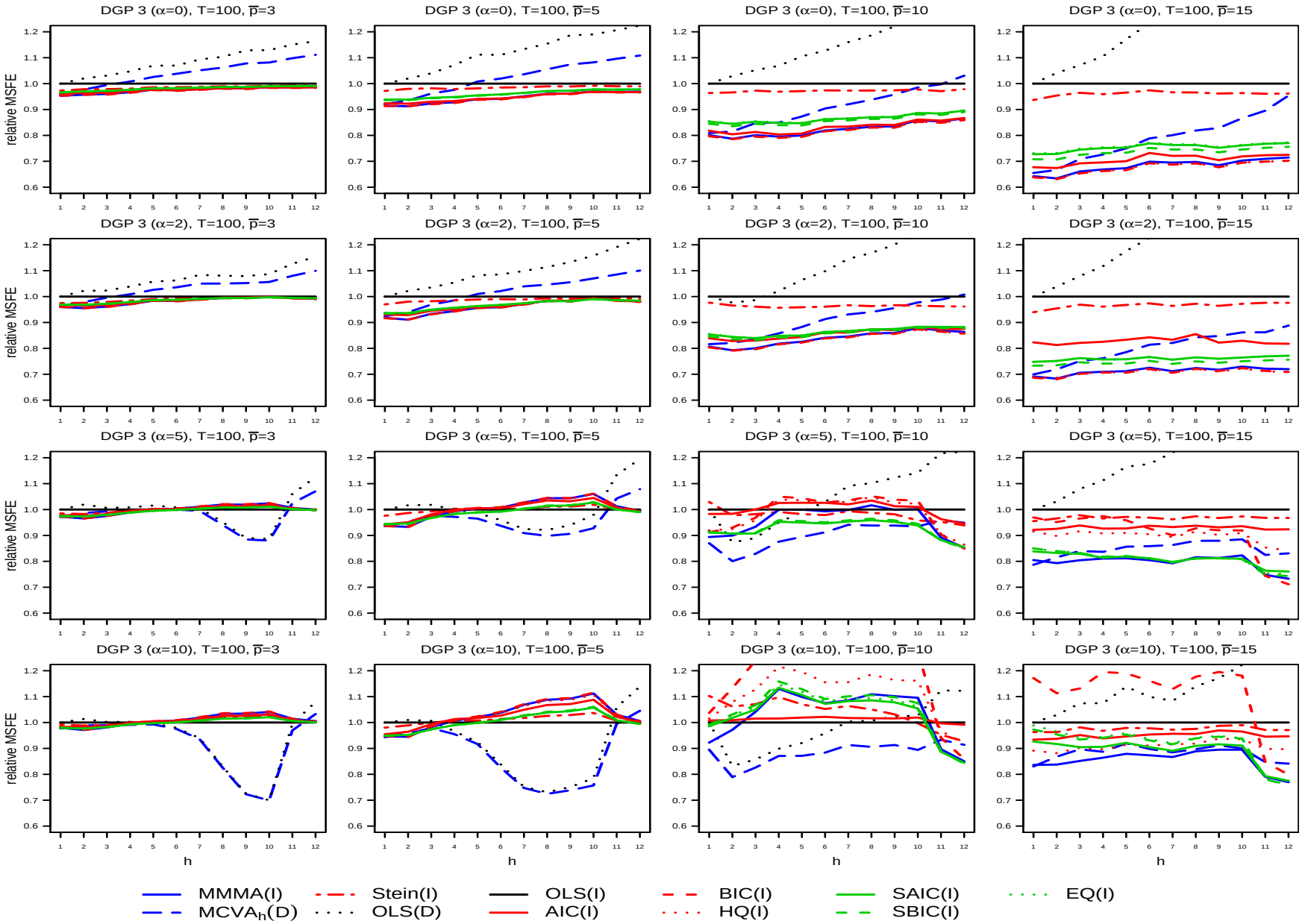


Figure A2: Multi-step forecast performance for DGP 3 ($T = 100$): bivariate drifting ARMA(1,10)

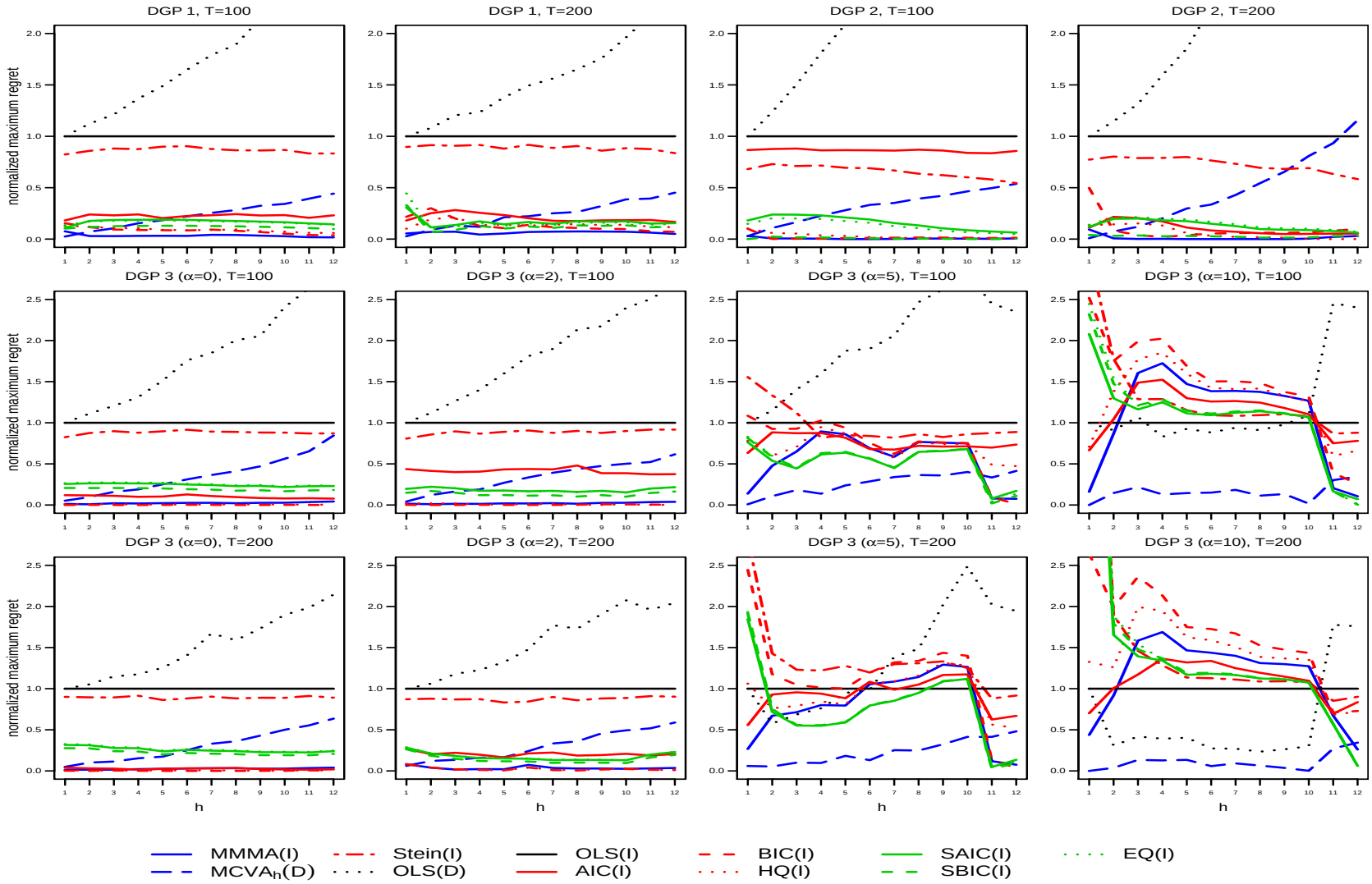
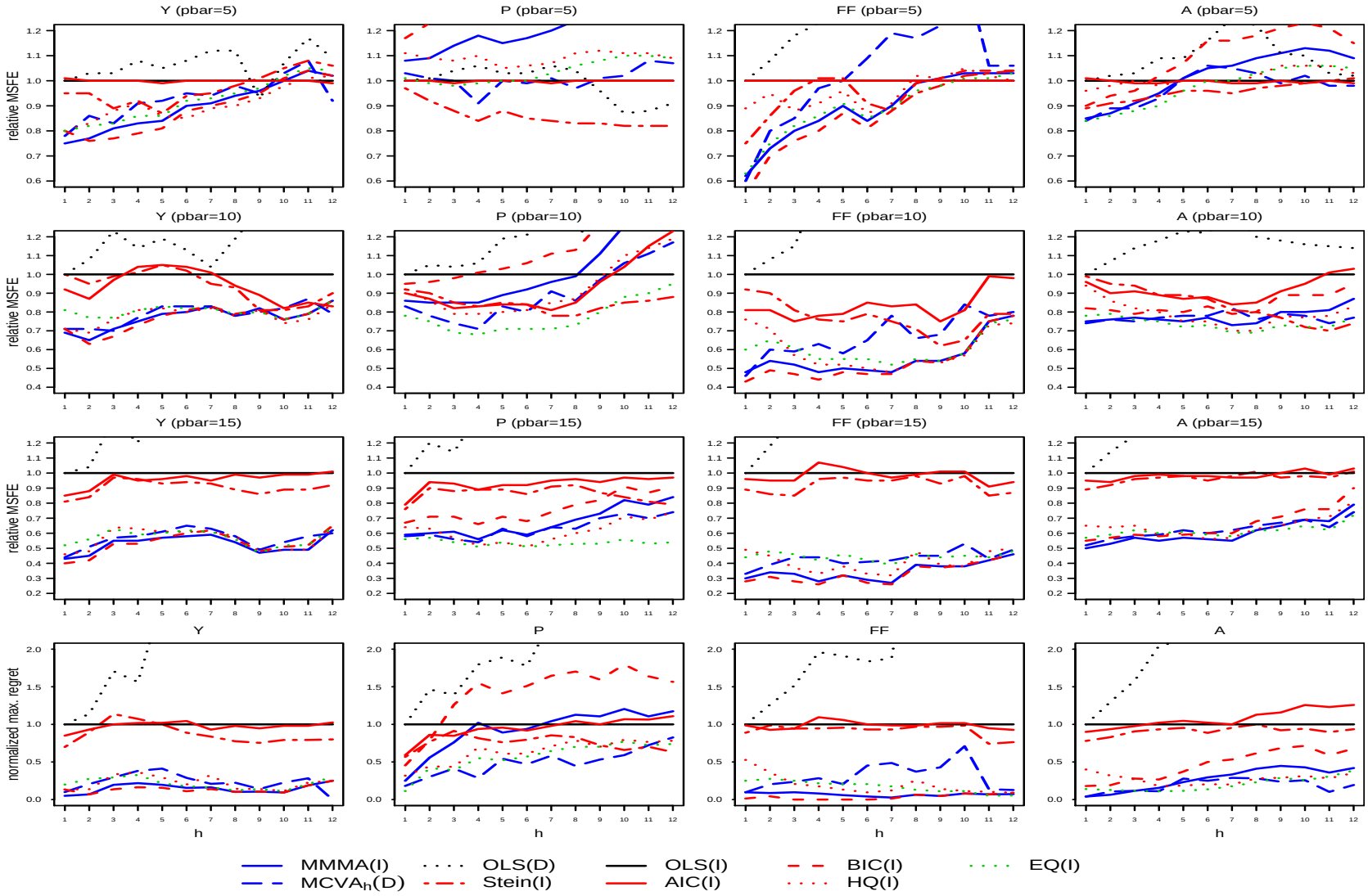


Figure A3: Normalized maximum regret under DGPs 1-3

Table A1: Forecast performance (measured by normalized maximum regret) of MMMA(I) and SMMA(I) under DGPs 1-3

		SMMA(I)								
		DGP 1				DGP 2	DGP 3			
h	MMMA(I)	$\sigma_{12} = .5$	$\sigma_{12} = .7$	$\sigma_{12} = .9$	$\sigma_{12} = 1$		$\alpha = 0$	$\alpha = 2$	$\alpha = 5$	$\alpha = 10$
					$T = 100$					
1	1.00	3.55	3.58	2.75	3.01	3.15	0.90	1.27	3.65	2.34
2	1.00	2.04	1.95	6.36	17.18	0.66	0.92	1.12	1.27	1.22
3	1.00	1.93	1.57	5.02	13.80	1.06	0.96	0.97	1.05	1.08
4	1.00	0.87	0.84	2.03	6.18	6.30	0.97	1.08	1.03	1.08
5	1.00	0.77	0.47	2.09	5.18	> 1.00	0.98	1.18	1.04	1.11
6	1.00	0.55	0.27	1.49	3.36	29.23	0.95	1.19	1.03	1.04
7	1.00	0.32	0.17	1.51	2.80	507.68	1.02	1.06	1.01	1.07
8	1.00	0.26	0.16	1.36	2.24	10.25	0.95	0.95	1.03	1.06
9	1.00	0.33	0.25	1.00	2.27	6.87	0.97	0.99	1.01	1.04
10	1.00	0.33	0.33	1.19	1.78	4.23	1.00	1.00	1.01	1.05
11	1.00	0.27	0.17	1.02	2.47	3.45	1.02	1.10	1.11	1.04
12	1.00	0.25	0.16	1.13	2.37	2.83	1.01	1.04	0.60	0.84
					$T = 200$					
1	1.00	2.75	2.78	2.63	2.08	4.85	1.10	1.04	3.40	2.79
2	1.00	1.71	1.40	2.23	17.67	8.76	0.91	1.30	1.13	1.35
3	1.00	1.34	1.22	1.73	13.36	8.11	0.79	1.12	1.12	1.26
4	1.00	1.24	1.58	6.48	5.64	>1.00	0.95	1.03	1.12	1.10
5	1.00	1.45	1.36	5.40	5.34	>1.00	0.94	1.18	1.02	1.09
6	1.00	1.23	0.91	2.61	13.62	>1.00	1.00	1.12	1.14	1.08
7	1.00	0.73	0.54	1.31	2.96	57.55	1.03	1.08	1.13	1.14
8	1.00	0.27	0.24	0.71	4.45	9.73	0.98	1.07	1.02	1.14
9	1.00	0.40	0.21	0.83	3.47	6.78	1.09	1.04	1.01	1.09
10	1.00	0.46	0.40	1.03	2.53	4.91	1.02	0.95	1.01	1.09
11	1.00	0.46	0.26	0.53	2.58	4.54	1.01	0.93	1.20	1.06
12	1.00	0.49	0.35	0.55	4.48	4.04	1.03	0.87	0.81	1.05

Note: (1) MMMA and SMMA refer to multivariate and single-equation Mallows model average, respectively; (2) The maximum regret for the sample size considered is taken over pre-specified maximum lag orders: $\bar{p} = 3, 4, \ldots, 15$ for DGPs 1 and 3 and $\bar{p} = 3, 4, \ldots, 8$ for DGP 2; (3) This corresponds to the case where the maximum regret of MMMA(I) is zero. We simply put "> 1.00" to indicate that the performance is inferior to MMMA(I).



Note: (1) The MSFEs for Y (GDP), P (the GDP deflator), and FF (the federal funds rate) are computed from (7.1) in the text, while the MSFEs for A are the aggregated weighted MSFEs computed from (7.2) in the text. All is relative to OLS(I); (2) Normalized maximum regret is taken over 13 pre-specified maximum lag orders: $\bar{p} = 3, \ldots, 15$, with OLS(I) normalized to unity; (3) "I" and "D" in parentheses refer to iterative and direct multi-step forecasts, respectively.

Figure A4: Empirical results: forecast performance (measured by relative MSFEs and normalized maximum regret) of MMMA(I), MCVA$_h$(D), and competing methods based on three-variable VARs ($\bar{p} = 5, 10, 15$ and $T = 100$)