Econometric Theory, **36**, 2020, 1099–1126.
doi:10.1017/S0266466619000434

OPTIMAL MULTISTEP VAR FORECAST AVERAGING*

JEN-CHE LIAO
Fu Jen Catholic University
WEN-JEN TSAY
Academia Sinica

This article proposes frequentist multiple-equation least-squares averaging approaches for multistep forecasting with vector autoregressive (VAR) models. The proposed VAR forecast 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 multistep forecast 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 forecast averaging approaches. Specifically, MMMA exhibits asymptotic optimality for one-step-ahead forecast averaging, whereas for direct multistep forecast averaging, the asymptotically optimal combination weights are determined separately for each forecast horizon based on the MCVA$_h$ procedure. To present our methodology, we investigate the finite-sample behavior of the proposed averaging procedures under model misspecification via simulation experiments.

1. INTRODUCTION

As a technique to characterize the joint dynamic behavior of economic variables, the vector autoregressive (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 into the VAR analysis.[1] This kind of model

*This article was previously circulated under the title "Multivariate Least Squares Forecasting Averaging by Vector Autoregressive Models." We thank the Editor Peter Phillips, the Co-Editor Robert Taylor, and two anonymous referees for their constructive comments on earlier versions of this article. We appreciate helpful comments and suggestions from Le-Yu Chen, Yi-Ting Chen, Graham Elliott, Bruce Hansen, Chu-An Liu, Chor-Yiu (CY) Sin, and participants of the 2016 Cross-Strait Dialogue III, the 2016 Taiwan Economics Research workshop, IAAE 2017, SETA 2019, CMES 2019, and econometrics seminars at several universities. We appreciate research support from Institute of Economics at Academia Sinica. We assume responsibility for any errors in the article. Address correspondence to Jen-Che Liao, Department of Economics, Fu Jen Catholic University, 510 Zhongzheng Road, Xinzhuang, New Taipei City 24205, Taiwan; e-mail: jccepd@gmail.com.

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

uncertainty arising from the choice of lag length may considerably impact the performance of any VAR-based estimation, inference, forecasting, and other analysis. This article addresses the issue of VAR model specification via a frequentist model averaging approach under iterated and direct multistep 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 such a large strand of the literature. As a more general approach versus model selection, model averaging methods are introduced in order to lower variability in model selection and thus increase estimation accuracy. In fact, the application of model averaging techniques has focused largely 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) suggest a jackknife averaging approach and an autocorrelation-robust averaging method under the time series framework, respectively. Gao et al. (2016) propose a leave-subject-out model averaging procedure for longitudinal data models and time series models with heteroskedastic errors. Under a known finite-order VAR with coefficients assumed to be local to the restrictions, Hansen (2016b) introduces the Stein combination shrinkage for VARs in which unrestricted least-squares estimates are shrunk toward multiple restricted least-squares estimates.

The main contribution of the article lies in the focus on both iterated and direct multistep-ahead forecast averaging problems. We propose two multistep VAR forecast averaging procedures 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). These two criteria have not yet been introduced or investigated either theoretically or empirically in multistep VAR forecasting settings. Our VAR forecast averaging methods allow for multiple response variables and include iterative and direct forecasting schemes; and their properties of asymptotic optimality are investigated in multistep forecasting settings.

This article offers several contributions to the literature. 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 time series 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.

The proposed MMMA method is designed for one-step forecast averaging, from which one can obtain the averaging multistep forecasts via the iterative strategy.

A second contribution is that we further extend the VAR forecast averaging to the direct forecasting framework, where serial correlations in h-step errors arise due to overlaps in the data when a forecast horizon of more than a single period (i.e., $h > 1$) is considered. To address this issue, we propose a new direct multistep VAR forecast averaging method based on the idea of leave-h-out cross-validation. Moreover, the main distinction of our two 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 employment is motivated by the aim to scale each response variable to have equal importance and to incorporate potential correlations across equations in the VAR system, thereby likely improving forecast accuracy.

Theoretical and empirical investigations of iterative and direct multistep forecasting with time series models based on a fixed lag or lag selection have been widely studied in statistics and econometrics, for example, 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), and Pesaran, Pick, and Timmermann (2011), among others. However, to the best of our knowledge, no efforts have been made for iterative and direct VAR forecast averaging problems (in a non-Bayesian sense). This article offers theoretical and empirical contributions by filling this gap in the literature.

Our main theoretical justifications hinge, on the one hand, upon a demonstration of the asymptotic optimality of the proposed VAR averaging methods for multistep forecasting. For one-step-ahead forecasting problems, Shibata (1980, 1981), and Shibata (1983) establishes the asymptotic optimality of the Akaike information criterion (AIC) and its variants, and while there are considerable views about the optimality theory for model selection under various circumstances (such as finite- versus infinite-dimensional models, cross-section versus time-series models, independent- versus same-realization predictions, homoskedastic versus heteroskedastic errors, and one- versus multistep forecasting),[2] there is very little theory in the time series context, even for single-equation model averaging problems. Most of the theory for model averaging applies to the cross-sectional case, for example, Hansen (2007), Wan, Zhang, and Zou (2010), Hansen and Racine (2012), and Liu, Okui, and Yoshimura (2016). For dependent data, Zhang et al. (2013) generalize Hansen and Racine's (2012) jackknife averaging criterion to two time series cases: serially correlated errors and lagged dependent variables. Cheng et al. (2015) propose an autocorrelation-robust Mallows criterion under time series errors.

[2]See Shao (1997) and Leeb and Pötscher (2009) for an excellent review of the pointwise asymptotic optimality (or loss-efficiency) results in the model selection literature.

Our optimality theory extends these existing asymptotic optimality results in the (frequentist) model averaging literature to a setting of multistep VAR forecast averaging, including iterative and direct forecasting strategies. In particular, our optimality results show that the proposed MMMA and MCVA_h are asymptotically efficient for *one-step-ahead* forecast averaging. As pointed out by an anonymous referee, for *one-step-ahead* forecasting, our proposed approaches, which generalize single-equation Mallows selection/averaging to multiple-equation model averaging, lie in a broad class of forecast selection (e.g., AIC and its variants [Shibata, 1980, Li, 1987, and Ing and Wei, 2005, among others]) and forecast averaging (e.g., Mallows averaging [Hansen, 2008]) methods that share the similar asymptotic optimality property.[3]

A different picture emerges when it comes to *multistep-ahead* forecasting, which is the main focus of this article. Unlike the case of *one-step-ahead* forecasting, our proposed MCVA_h method is shown to be theoretically preferred over MMMA as well as better than the other usual selection methods mentioned above for multistep forecasting under an infinite-order VAR setting, due to its asymptotic optimality for each forecast horizon $h > 1$. Namely, for *multistep-ahead* forecast averaging, the asymptotically optimal combination weights are determined separately for each forecast horizon by the direct method based on our proposed MCVA_h criterion. Our results appear to be the first demonstration that investigates the validity of MMMA and MCVA_h in multistep VAR forecast averaging settings.

From an empirical perspective, we illustrate the proposed methods through Monte Carlo simulations and highlight the importance of the misspecification bias when comparing iterative versus direct VAR forecast averaging methods. Specifically, our simulation results reveal that the iterative MMMA tends to be preferable when the candidate model set contains VAR models with sufficiently long lags and when the candidate models are not highly misspecified; conversely, the direct MCVA_h exhibits substantial advantages when model misspecification is severe. 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 outweighed by its efficiency loss.

This article is related to the large and growing literature on forecast combination. The development of forecast combination dates back to the seminal works of Reid (1968, 1969), and Bates and Granger (1969). Since then, forecast combination methods have been investigated in numerous studies, for example, Granger (1989), Clemen (1989), Diebold and Lopez (1996), Hendry and Clements (2004), Timmermann (2006), and Stock and Watson (2006), among others. More recent studies, including Hansen (2008), Zhang et al. (2013), Cheng and Hansen (2015), and

[3] In this view, asymptotic optimality is a weak property, and it would thus call for establishing a more meaningful optimality property that is able to further distinguish these similarly optimal methods for one-step-ahead forecasting. We leave this important issue for future research.

Gao et al. (2016), extend the forecast combination literature by developing frequentist model averaging methods for univariate time series forecast combinations.

The rest of the article is organized as follows. Section 2 sets out the framework of multivariate time series forecasting with VAR models. Section 3 suggests an iterative multistep forecast averaging procedure based on the MMMA criterion. Section 4 further proposes an MCVA$_h$ procedure to address the serial correlation problem that arises under the direct multistep forecasting scheme. Built upon asymptotic unbiasedness and asymptotic optimality, Section 5 provides the theoretical validity of our methods. Sections 6 presents the numerical performance of our methodology via finite-sample simulation experiments. We conclude the article in Section 7. Mathematical proofs of the theorems, additional simulation results, and an empirical application to a three-variable monetary VAR based on U.S. data can be found online, in the Supplementary Material section of this article.

2. FORECASTING PROBLEMS BY FITTING VAR(p) MODELS

Consider a stationary K-dimensional moving average (MA) process $\{\mathbf{y}_t\}$,

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

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 vectors 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 in (2.1) by assuming without loss of generality that the mean $E(\mathbf{y}_t)$ is already subtracted out.

We assume 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. Thus, (2.1) can be expressed as an infinite-order vector autoregression process, that is,

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

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 special cases of (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.

Fixing the finite lag-order p, we consider the linear h-step ahead forecast of $\mathbf{y}_{t+h}$ by employing an approximating K-dimensional 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.

To construct the one-step-ahead forecast (i.e., $h = 1$), in matrix notation we write

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

where $\mathbf{Y} = (\mathbf{Y}_1\, \mathbf{Y}_2 \ldots \mathbf{Y}_K)$ 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 kth 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 $t = p, \ldots, T-1$, $\boldsymbol{\Pi}(p)' = (\boldsymbol{\pi}_1(p), \ldots, \boldsymbol{\pi}_p(p))$ is the $K \times m$ coefficient matrix, and $\boldsymbol{\varepsilon}(p)$ is the $(T-p) \times K$ matrix with the kth column being $(\varepsilon_{k,p+1}(p), \ldots, \varepsilon_{kT}(p))'$.

We use the OLS method to estimate the VAR(p) model. It is known, as first shown by Zellner (1962), that the OLS and generalized least-squares 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 $\widehat{\boldsymbol{\Pi}}(p) = (\widehat{\boldsymbol{\pi}}_1(p), \ldots, \widehat{\boldsymbol{\pi}}_p(p))'$ is given by

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

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 $j \leq 0$, and the superscript "I" indicates indirect multistep 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. For a VAR(p) model, we follow the common standard that all lags are included up to the lag order p, that is, no gaps in the lags are allowed. 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. Let $\bar{p}$ denote the maximum lag order. We specifically let

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

be the residual covariance matrix without the adjustment for degrees of freedom from a VAR(p) model, where $\hat{\boldsymbol{\varepsilon}}_{t+1}(p)'$, $t = \bar{p}, \ldots, T-1$, are the $1 \times K$ row vectors of the OLS residual matrix $\hat{\boldsymbol{\varepsilon}}(p) = \mathbf{Y} - \mathbf{Z}(p)\widehat{\boldsymbol{\Pi}}(p)$. Based on $\widehat{\boldsymbol{\Sigma}}(p)$, the three commonly used criteria for VAR lag selection are AIC, Bayesian information criterion (BIC), and Hannan–Quinn (HQ). Other selection criteria include final prediction error and some variants of AIC and BIC that are designed to correct for overfitting in VAR models.[4]

Instead of using lag selection methods, this article considers combining forecasts generated from candidate VAR models using different lags. In general, given a prespecified $\bar{p}$,[5] constructing a combined multistep forecast involves a sequence of choices and several potential alternatives that may exist for these choices:[6] (i) estimate each of the candidate VAR models by some estimation method and produce the resulting forecasts based on the fitted models; (ii) seek a combination scheme that aggregates individual VAR forecasts;[7] (iii) estimate the combination weights by minimizing some criterion for $h \geq 1$; and (iv) combine the multistep forecasts using the estimated weights. For (i), both of our proposed multistep VAR forecast averaging methods are based on OLS estimation. The main focus of the article lies in (iii) and (iv). Specifically, our proposed MMMA (MCVA_h) approach estimates the combination weights by minimizing a multivariate Mallows (leave-h-out cross-validation) averaging criterion and constructs the combined multistep forecasts in an iterative (a direct) manner. These two averaging methods are discussed in detail in Sections 3 and 4, with their theoretical justifications provided in Section 5.

It is known that the number of VAR parameters increases quadratically with the number of variables. As a consequence, instead of the unrestricted OLS estimation in a frequentist setting considered here for (i), under high-dimensional settings, it might be beneficial to consider Bayesian VARs (established in the seminal papers of Litterman, 1986 and Doan, Litterman, and Sims, 1984) that use shrinkage priors on the model parameters, as suggested by an anonymous referee. Existing related literature also covers automated data-based model determination methods (Phillips, 1995, 1996) that use optimized information criteria that are specifically suited to forecast model selection (e.g., the determination of lag length or cointegrating rank in a VAR, or hyperparameters in Bayesian VAR settings). Moreover, the Bayesian VAR offers a natural way to combine with Bayesian model

[4]Interested readers may refer to Lütkepohl (2005) and McQuarrie and Tsai (1998, Ch. 5) for detailed discussions.

[5]The choice of $\bar{p}$ is a common issue encountered in classical VAR lag determination methods, such as sequential testing procedures or model selection criteria. To keep our presentation focused, our analysis simply proceeds with a prespecified $\bar{p}$. To take into account the uncertainty from selecting $\bar{p}$, we examine in the simulation experiments the sensitivity of the forecast performance of the proposed methods to the choice of $\bar{p}$, using the normalized maximum regret based on mean squared forecast errors (MSFEs) over different values of $\bar{p}$; see Section 6 for details.

[6]We thank an anonymous referee for drawing our attention to this point.

[7]This article focuses on the family of *linear* forecast combinations, which are more commonly used in practice. Other possibilities for (ii) are nonlinear and time-varying combination schemes; for which we refer the interested reader to Timmermann (2006, Sect. 4).

averaging, as an alternative for (iii), to deal with model uncertainty. Another class of the shrinkage type of estimators (i.e., penalized least-squares methods such as Lasso and its variants) also makes it possible to address the dimensionality issue. These potential directions, especially various comparisons of these alternative methods against our VAR forecast averaging approaches, are certainly worth investigating in future research.

3. ITERATIVE MULTISTEP VAR FORECAST 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, that is, $h = 1$. The averaging multistep forecasts are then obtained by iterating forward for multiple periods.

Consider the following multivariate Mallows criterion for VAR model selection:

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

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

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

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 Sparks, Coutsourides, and Troskie (1983) and Fujikoshi and Satoh (1997) for selecting multivariate regression models. The weighted sum of squared residuals, as in the first component of (3.1), has been recently employed in several studies (e.g., Lee and Liu, 2012 and Basu and Michailidis, 2015) on shrinkage estimation of sparse large VAR models to incorporate information on possible correlations among variables. The weighted criterion considered here is based on the underlying weighted loss or risk function; see Section 5 for a related discussion.

For valid comparison, the sample is set to be of equal size across different candidate models. 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 $t = \bar{p}, \ldots, T-1$.

We now turn to the method of VAR forecast averaging based on the Mallows criterion. To begin with, let $\mathbf{w} = (w(1), \ldots, w(\bar{p}))'$ be the weight vector associated with candidate models, and $\widehat{\mathbf{\Pi}}^*(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p) \overline{\mathbf{\Pi}}(p)$ be 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 $1 \le i \le Kp$ and $1 \le j \le K$, and $\overline{\mathbf{\Pi}}_{ij}(p) = 0$ elsewhere, that is, $\overline{\mathbf{\Pi}}(p)' = \left(\widehat{\mathbf{\Pi}}(p)' \, \mathbf{0}_{K \times K(\bar{p}-p)}\right)$, where $\mathbf{0}_{r \times s}$ is a $r \times s$ zero matrix.

The combination residuals $\hat{\boldsymbol{\varepsilon}}^*_{t+1}(\mathbf{w})$ can be expressed as

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

where for the third equality, we assume $\sum_{p=1}^{\bar{p}} w(p) = 1$.

Weight estimation. Armed with $\hat{\boldsymbol{\varepsilon}}^*_{t+1}(\mathbf{w})$, $t = \bar{p}, \ldots, T-1$, given in (3.3), the proposed MMMA criterion takes the following form:

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

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

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

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

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

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:

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

and $\widetilde{\sigma}_{k\ell}$ is the (k, ℓ)th element of $\widetilde{\mathbf{\Sigma}}(\bar{p})^{-1}$. For the derivation of the last equality in (3.5) and, in particular, the more detailed construction of the matrix $\widehat{\mathbf{S}}$, see Appendix A in the online supplementary material available at Cambridge Journals Online (journals.cambridge.org/ect). As also shown there, in the univariate case (i.e., $K = 1$) the MMMA criterion (3.4) reduces to a Hansen's (2007) single-equation Mallows averaging criterion.

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

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

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

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

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

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

As in single-equation model averaging problems, the quadratic programming problem in (3.7) can be solved via several types of widely used statistical programming software, such as the Guass function "qprog," Matlab function "quadprog," and R package "quadprog."

Lastly, 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 MULTISTEP VAR FORECAST AVERAGING

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 $\{\mathbf{y}_j\}_{j\le t}$, and $\boldsymbol{\epsilon}_{t+h}$ is the associated h-step error with $E(\boldsymbol{\epsilon}_{t+h}\boldsymbol{\epsilon}'_{t+h}) = \mathbf{\Sigma}_h$ and can be expressed as $\boldsymbol{\epsilon}_{t+h} = \sum_{i=0}^{h-1} \mathbf{\Phi}_i\boldsymbol{\varepsilon}_{t+h-i}$, which is known to follow a MA 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 the 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 residual matrix is $\widehat{\mathbf{e}}_h(p) = \mathbf{Y}_h - \mathbf{Z}_h(p)\widehat{\boldsymbol{\Psi}}_h(p)$. The resulting direct h-step-ahead forecast is then formed by

$$\widehat{\mathbf{y}}^D_{t+h|t}(p) = \widehat{\boldsymbol{\Psi}}_h(p)'\mathbf{z}_t(p), \tag{4.3}$$

where the superscript D stands for the direct method.

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 $\bar{p} + h - 1 \le t \le T - 2h + 1$. 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.

Let $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p) = (\widetilde{\epsilon}_{1,t+h}(p), \widetilde{\epsilon}_{2,t+h}(p), \ldots, \widetilde{\epsilon}_{K,t+h}(p))'$, where $\widetilde{\epsilon}_{k,t+h}(p)$ is 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. Specifically, $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)$ is then obtained by

$$\widetilde{\boldsymbol{\epsilon}}_{t+h}(p) = \mathbf{y}_{t+h} - \widetilde{\boldsymbol{\Psi}}_{h,t}(p)'\mathbf{z}_t(p), \tag{4.4}$$

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.

Weight estimation. 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 residual matrices. The proposed MCVA$_h$ criterion for the weight estimation that is used to combine direct h-step forecasts is given by

$$\begin{aligned} CV_{T,h}(\mathbf{w}) &= (T-\bar{p}-h+1)\cdot \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\widetilde{\boldsymbol{\Sigma}}^*_h(\mathbf{w})\right) \\ &= \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\sum_{t=\bar{p}}^{T-h}\widetilde{\boldsymbol{\epsilon}}^*_{t+h}(\mathbf{w})\widetilde{\boldsymbol{\epsilon}}^*_{t+h}(\mathbf{w})'\right) \\ &= \sum_{t=\bar{p}}^{T-h}\operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\left(\sum_{p=1}^{\bar{p}} w(p)\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)\right)\left(\sum_{p=1}^{\bar{p}} w(p)\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)\right)'\right) \\ &= \sum_{t=\bar{p}}^{T-h}\sum_{i=1}^{\bar{p}}\sum_{j=1}^{\bar{p}}\widetilde{\epsilon}_{t+h,ij}\,w(i)w(j) \\ &= \mathbf{w}'\widetilde{\mathbf{S}}_h\mathbf{w}, \end{aligned} \tag{4.5}$$

where

$$\widetilde{\boldsymbol{\Sigma}}_h(\bar{p}) = \frac{1}{T-\bar{p}-h-\bar{m}+1}\sum_{t=\bar{p}}^{T-h}\widetilde{\boldsymbol{\epsilon}}_{t+h}(\bar{p})\widetilde{\boldsymbol{\epsilon}}_{t+h}(\bar{p})', \tag{4.6}$$

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 $\widetilde{\sigma}^h_{k\ell}$ being the (k,ℓ)th element of $\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}$.

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

$$\widehat{\mathbf{w}}_{cv,h} = \arg\min_{\mathbf{w}\in\mathcal{H}_T} CV_{T,h}(\mathbf{w}), \tag{4.7}$$

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), (4.7) also takes the form of quadratic programming problems, 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

$$\widehat{\mathbf{y}}^{D*}_{t+h|t}(\widehat{\mathbf{w}}_{cv,h}) = \sum_{p=1}^{\bar{p}}\widehat{w}_{cv,h}(p)\widehat{\mathbf{y}}^{D}_{t+h|t}(p), \tag{4.8}$$

where $\widehat{\mathbf{w}}_{cv,h} = (\widehat{w}_{cv,h}(1),\ldots,\widehat{w}_{cv,h}(\bar{p}))'$ and $\widehat{\mathbf{y}}^{D}_{t+h|t}(p)$ 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. In Section B in the online supplementary material, we discuss how to efficiently compute the leave-h-out residual vector $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)$.

5. ASYMPTOTIC THEORY

This section provides theoretical justifications of our methods proposed in Sections 3 and 4, including the relation of the proposed iterative and direct VAR forecast averaging criteria to MSE and MSFE in multistep forecast settings and the asymptotic unbiasedness and asymptotic optimality of the proposed averaging procedures.

5.1. MSFE of MultiStep Forecast Averaging

In this subsection, we simply denote by $\widehat{\mathbf{y}}_{T+h|T}(p) = \widehat{\mathbf{y}}^D_{T+h|T}(p)$ the h-step ahead forecast at the origin T produced by a fitted direct h-step VAR(p) model using the full effective sample. For any forecast combination $\mathbf{w}$, the h-step-ahead forecast combination is given by

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

with the associated forecast error

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

Let $\widehat{\boldsymbol{\mu}}_h(p) = (\widehat{\boldsymbol{\mu}}^h_p(p), \ldots, \widehat{\boldsymbol{\mu}}^h_{T-h}(p))'$ denote the matrix of fitted values of $\mathbf{Y}_h$ based on a fitted VAR(p) with $\widehat{\boldsymbol{\mu}}^h_t(p) = (\widehat{\mu}^h_{1t}(p), \widehat{\mu}^h_{2t}(p), \ldots, \widehat{\mu}^h_{Kt}(p))'$, and analogously, $\widehat{\boldsymbol{\mu}}^*_h(\mathbf{w}) = (\widehat{\boldsymbol{\mu}}^{h*}_{\bar{p}}(\mathbf{w}), \ldots, \widehat{\boldsymbol{\mu}}^{h*}_{T-h}(\mathbf{w}))'$ with $\widehat{\boldsymbol{\mu}}^{h*}_t(\mathbf{w}) = (\widehat{\boldsymbol{\mu}}^{h*}_{1t}(\mathbf{w}), \ldots, \widehat{\boldsymbol{\mu}}^{h*}_{Kt}(\mathbf{w}))'$ and $\widehat{\boldsymbol{\mu}}^{h*}_{kt}(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\widehat{\boldsymbol{\mu}}^h_{kt}(p)$. Define

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

where the approximation follows from the assumed stationarity of $\mathbf{y}_t$ and $E(\mathrm{tr}(\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\epsilon}_{T+h}\boldsymbol{\epsilon}'_{T+h})) = \mathrm{tr}(\boldsymbol{\Sigma}_h^{-1}E(\boldsymbol{\epsilon}_{T+h}\boldsymbol{\epsilon}'_{T+h})) = \mathrm{tr}(\mathbf{I}_K) = K$ and $L_{T,h}(\mathbf{w})$ is defined as the in-sample average squared error from the h-step ahead forecast

combination

$$L_{T,h}(\mathbf{w}) = \frac{1}{T-\bar{p}-h+1}\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h^*(\mathbf{w}))'(\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h^*(\mathbf{w}))\right)$$
$$= \frac{1}{(T-\bar{p}-h+1)}\sum_{t=\bar{p}}^{T-h}\sum_{\ell=1}^{K}\sum_{k=1}^{K}(\mu_{kt}^h - \widehat{\mu}_{kt}^{h*}(\mathbf{w}))\sigma_{k\ell}^h(\mu_{\ell t}^h - \widehat{\mu}_{\ell t}^{h*}(\mathbf{w})), \quad \textbf{(5.3)}$$

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

It is worth emphasizing that $\mathrm{MSFE}_h(\mathbf{w})$ defined in (5.2) is weighted by the inverted true error covariance matrix $\boldsymbol{\Sigma}_h^{-1}$, which contrasts with the single-equation model selection/averaging problem.[8] The use of the weighted MSFEs here has two motivations. 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 contains, thereby likely improving forecast accuracy.

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 (5.3) when $h=1$. 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 $\mathrm{MSFE}(\mathbf{w})$ $\equiv \mathrm{MSFE}_h(\mathbf{w})$ defined in (5.2) when $h=1$. To establish the property that the $C_T(\mathbf{w})$ criterion is an asymptotically unbiased estimator of $L_T(\mathbf{w})$, 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) and (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 nonsingular variance–covariance matrix $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \boldsymbol{\Sigma}$.

(b) $\boldsymbol{\varepsilon}_t$ is assumed to have a finite fourth moment in the sense that for some finite constant C, $E|\varepsilon_{ut}\varepsilon_{vt}\varepsilon_{wt}\varepsilon_{xt}| \le C$ 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

[8] The $\mathrm{MSFE}_h(\mathbf{w})$ defined in (5.2) can be viewed as a type of the Mahalanobis distance that accounts for the covariance structure (Varmuza and Filzmoser, 2009, p. 46). The weighted loss or risk functions have been considered in the literature, for example, Andrews (1991, Sect. 6) and Hansen (2016a, Sect. 2.2); Fujikoshi and Satoh (1997) and Yanagihara and Satoh (2010) in multivariate regression settings. Moreover, in the context of VAR forecasting using Lasso, Hsu, Hung, and Chang (2008), Ren and Zhang (2010), and Ren, Xiao, and Zhang (2013) use the normalized MSFE to evaluate forecast accuracy in their simulation work and empirical examples.

satisfied. Assumption 1(c) is standard for the multivariate OLS estimator of fitting a finite-order VAR(p) model to potentially infinite-order processes, for example, 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 below formally summarizes the arguments presenting the asymptotic unbiasedness of $C_T(\mathbf{w})$.

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

$$C_T(\mathbf{w}) = (T-\bar{p})L_T(\mathbf{w}) + (T-\bar{p})K + r_{1T}(\mathbf{w}) + r_{2T}(\mathbf{w}) + 2K^2\mathbf{p}'\mathbf{w}, \tag{5.4}$$

where $L_T(\mathbf{w})$ is the in-sample average squared error defined by (5.3) 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. Similar to the univariate case, Theorem 1 implies that the leading term of $C_T(\mathbf{w})$ is a downward-biased estimator of the expected loss $E(L_T(\mathbf{w}))$, and this downward bias arises as we use $\mathbf{Y}$ to replace the unknown $\boldsymbol{\mu}$ in $E(L_T(\mathbf{w}))$, while $\widehat{\boldsymbol{\mu}}(\mathbf{w})$ is also estimated based on $\mathbf{Y}$. The source of the downward bias is $r_{2T}(\mathbf{w})$, and built on Lewis and Reinsel (1985), its expected value can be shown to be $-2K^2\mathbf{p}'\mathbf{w}$, or the negative of the penalty term.

We next turn to the $CV_{T,h}(\mathbf{w})$ criterion. 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 (presented in the online supplementary material). 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, that is,

$$\begin{aligned}\widetilde{L}_{T,h}(\mathbf{w}) &= \frac{1}{T-\bar{p}-h+1}\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h^*(\mathbf{w}))'(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h^*(\mathbf{w}))\right) \\ &= \frac{1}{(T-\bar{p}-h+1)}\sum_{t=\bar{p}}^{T-h}\sum_{\ell=1}^{K}\sum_{k=1}^{K}(\mu_{kt}^h - \widetilde{\mu}_{kt}^{h*}(\mathbf{w}))\sigma_{k\ell}^h(\mu_{\ell t}^h - \widetilde{\mu}_{\ell t}^{h*}(\mathbf{w})), \end{aligned} \tag{5.5}$$

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 $\mathbf{w}$, the proposed MCVA$_h$ criterion $CV_{T,h}(\mathbf{w})$ given by (4.5) can be expressed as*

$$CV_{T,h}(\mathbf{w}) = (T-\bar{p}-h+1)\widetilde{L}_{T,h}(\mathbf{w}) + (T-\bar{p}-h+1)K + \widetilde{r}_{1Th}(\mathrm{w}) + \widetilde{r}_{2Th}(\mathbf{w}), \tag{5.6}$$

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

Remark 2. The asymptotic unbiasedness of the MCVA_h criterion in Theorem 2 essentially builds on the argument that $E(\widetilde{r}_{2Th}(\mathbf{w})) = E(\mathrm{tr}(\widetilde{\mathbf{P}}_h^*(\mathbf{w})\mathbf{e}_h\mathbf{e}_h')) = 0$, which is based on the observations that the matrix $E(\mathbf{e}_h\mathbf{e}_h')$ has an exactly opposite nonzero/zero structure to the matrix $\widetilde{\mathbf{P}}_h^*(\mathbf{w})$; $E(\mathbf{e}_h\mathbf{e}_h')$ is symmetric, 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.

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

$$\text{(OPT 1)}: \frac{L_T(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{H}_T} L_T(\mathbf{w})} \xrightarrow{p} 1 \quad \text{as } T \to \infty \tag{5.7}$$

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 (5.7), the key is to show

$$\sup_{\mathbf{w}\in H_T}\left|\frac{C_T^*(\mathbf{w}) - L_T(\mathbf{w})}{L_T(\mathbf{w})}\right| \xrightarrow{p} 0. \tag{5.8}$$

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 $\bar{p}$; $\overline{\mathbf{P}} = \overline{\mathbf{Z}}\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\overline{\mathbf{Z}}'$ is the associated projection matrix. Denote $\mathbf{A}^*(\mathbf{w}) = \mathbf{I}_{T-\bar{p}} - \mathbf{P}^*(\mathbf{w})$ and define

$$\begin{aligned} V_T(\mathbf{w}) = E(L_T(\mathbf{w})) &= \frac{1}{T-\bar{p}} E\left(\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}^*(\mathbf{w}))'(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}^*(\mathbf{w}))\right)\right) \\ &= \frac{1}{T-\bar{p}}\mathrm{tr}\left(\mathbf{A}^*(\mathbf{w})\boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{A}^*(\mathbf{w})\right) \\ &\quad + E\left(\mathrm{tr}\left(\mathbf{P}^*(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}^*(\mathbf{w})\right)\right), \end{aligned} \tag{5.9}$$

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 nonzero approximation error for all candidate models VAR(p) of finite order.

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

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{L_T(\mathbf{w})}{V_T(\mathbf{w})}-1\right|\xrightarrow{p}0, \tag{5.10}$$

then the goal (5.8) to prove asymptotic optimality becomes

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{C_T^*(\mathbf{w})-L_T(\mathbf{w})}{V_T(\mathbf{w})}\right|\xrightarrow{p}0. \tag{5.11}$$

We make the following assumptions to prove (5.10) and (5.11), 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 $\bar{p}\xi_T^{*-2}\mathrm{vec}(\boldsymbol{\mu}')'\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\mathrm{vec}(\boldsymbol{\mu}')=o_p(1)$.

(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 $\sup_{-\infty<t<\infty}E|\varepsilon_{k_1t}\cdots\varepsilon_{k_st}|<\infty$ 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 $\bar{p}$, that is, $\bar{p}$ 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 et al. (2013) and the similar condition (8) of Wan et al. (2010); 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 et al. (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 Thm. 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 Ing and Wei (2003, eqn. (2.16)), 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 the Appendix, this condition plays a key role

in further improving the upper bound on $\|\widehat{\mathbf{\Gamma}}_T^{-1}(\bar{p})\|$. Moreover, the conditions in Assumption 2(d) and (e) are the same as the conditions in Thm. 2 of Ing and Wei (2003) and condition (C.4) of Zhang et al. (2013). These two sets of assumptions provide alternative restrictions on and a trade-off between the growth rate of $\bar{p}$ and the existence of moments of $\boldsymbol{\varepsilon}_t$.

THEOREM 3. *If either Assumptions 1(a) and (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 (5.7) holds.*

Remark 3. Theorem 3 extends the existing asymptotic optimality results for model averaging to the multivariate Mallows criterion in the context of VAR forecast averaging. This result 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.

Remark 4. It is essential to note that in general, the asymptotic optimality of the MMMA for one-step-ahead forecast averaging, as stated in Theorem 3, does not carry over to multistep-ahead forecasting (i.e., $h > 1$) in either a direct or an iterative scheme. For the former, one straightforward aspect to see the asymptotic nonoptimality of the direct MMMA for $h > 1$ is through its invalidity under serial correlations. Specifically, it is not hard to show that the asymptotic unbiasedness (and, hence, asymptotic optimality) of the h-step version of the MMMA criterion $C_T(\mathbf{w})$ in (3.4) does not hold under the direct method based on h-step VAR models, where the serial correlation problem arises naturally; for detailed discussions, see Section D in the online supplementary material. For the iterative MMMA, we note that a formal investigation of the nonoptimality of the MMMA when used iteratively for multistep forecasting requires extending Bhansali's (1996) analysis to the general forecast averaging; see Remark 7 for further discussions.

To address the aforementioned limitation of Theorem 3, our next focus is on exploring the possibility of the h-step-ahead generalization of asymptotic optimality for multistep VAR forecast averaging, where the combination weights are selected for each h by our direct MCVA$_h$ procedure.

To begin with, similar to (5.7), the asymptotic optimality condition for the MCVA$_h$ procedure with respect to the criterion $L_{T,h}(\mathbf{w})$ (5.3) is given by

$$\text{(OPT 2):}\ \frac{L_{T,h}(\widehat{\mathbf{w}}_{cv,h})}{\inf_{\mathbf{w}\in\mathcal{H}_T} L_{T,h}(\mathbf{w})} \xrightarrow{p} 1 \quad \text{as } T\to\infty, \tag{5.12}$$

where $\widehat{\mathbf{w}}_{cv,h}$ is the estimated weight vector obtained from (4.7).

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 estimators, 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 (4.5).

Analogous to (5.8), (5.10), and (5.11), the asymptotic optimality conditions we wish to show are

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{CV^*_{T,h}(\mathbf{w})-\widetilde{L}_{T,h}(\mathbf{w})}{\widetilde{V}_{T,h}(\mathbf{w})}\right| \xrightarrow{p} 0 \quad \text{and} \quad \sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{L}_{T,h}(\mathbf{w})}{\widetilde{V}_{T,h}(\mathbf{w})}-1\right| \xrightarrow{p} 0, \tag{5.13}$$

establishing

$$\frac{\widetilde{L}_{T,h}(\widehat{\mathbf{w}}_{cv,h})}{\inf_{\mathbf{w}\in\mathcal{H}_T}\widetilde{L}_{T,h}(\mathbf{w})} \xrightarrow{p} 1 \quad \text{as } T\to\infty, \tag{5.14}$$

which is the asymptotic optimality of $\widehat{\mathbf{w}}_{cv,h}$ with respect to the criterion $\widetilde{L}_{T,h}(\mathbf{w})$. Lastly, combining (5.14) with

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{L}_{T,h}(\mathbf{w})}{L_{T,h}(\mathbf{w})}-1\right| \xrightarrow{p} 0 \tag{5.15}$$

yields (5.12), as desired. To establish (5.12), we make the following conditions.

Assumption 3. (a) As $T\to\infty$, $\bar{p}_T\xi^{*-1}_{T,h} = o_p(1)$, and $\bar{p}_T\xi^{*-2}_{T,h}\text{vec}(\boldsymbol{\mu}_h')'\left(\overline{\mathbf{P}}_h\otimes\mathbf{I}_K\right)\text{vec}(\boldsymbol{\mu}_h') = o_p(1)$, 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 $q^*_h = \max_{1\le p\le\bar{p}}\max_{\bar{p}\le t\le T-h}\max_{\underline{\ell}_{ht}-\bar{p}+1\le j\le\bar{\ell}_{ht}-\bar{p}+1}(\mathbf{P}_h(p))_{t-\bar{p}+1,j}$, where $(\mathbf{A})_{ij}$ denotes the (i,j)th element of matrix $\mathbf{A}$. Assume that q^*_h satisfies $q^*_h\bar{p}_T^{-1}T\to 0$ almost surely as $T\to\infty$.

Assumption 3(a) is analogous to Assumption 2(a) for the MMMA case. Denote $\widehat{\boldsymbol{\Gamma}}_{T,h}(\bar{p}) = (T-\bar{p}-h+1)^{-1}\sum_{t=\bar{p}}^{T-h}\mathbf{z}_t(\bar{p})\mathbf{z}_t(\bar{p})' = (T-\bar{p}-h+1)^{-1}\overline{\mathbf{Z}}_h'\overline{\mathbf{Z}}_h$. Under the condition $E\|\widehat{\boldsymbol{\Gamma}}^{-1}_{T,h}(\bar{p})\|_1 = O(\bar{p}^{2+\theta})$ 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 (C.15) and (C.16) in the online supplementary material, it is not difficult to establish the leave-h-out version of (C.15) and (C.16), that is, $E\|\widehat{\boldsymbol{\Gamma}}^{-1}_{T,h}(\bar{p})\|_1 = O(1)$ and for every h, $E(\text{tr}(\mathbf{e}_h'\overline{\mathbf{Z}}_h\overline{\mathbf{Z}}_h'\mathbf{e}_h))/(T-\bar{p}+h-1) = O(\bar{p}_T)$. On the other hand, combined with the fact that $\boldsymbol{\varepsilon}_{h,t+h} = \sum_{i=0}^{h-1}\boldsymbol{\Phi}_i\boldsymbol{\varepsilon}_{t+h-i}$, 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 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, for example, Li (1987), Andrews (1991), Hansen and Racine (2012), and Zhang et al. (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 weight estimator $\widehat{\mathbf{w}}_{cv,h}$ 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) and (b), 2(b)–(d), and 3(a) and (b); or Assumptions 1(a), and 2(b),(c),(e), and 3(a) and (b) are satisfied; for all fixed $h \geq 1$, the proposed direct MCVA$_h$ procedure based on the criterion $CV_{T,h}(\boldsymbol{w})$ is then asymptotically optimal in the sense that the optimality condition (5.12) 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 et al., 2013) to VAR forecasting for forecast horizons $h > 1$. 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 multistep VAR forecast averaging can still be achieved by selecting $\widehat{\mathbf{w}}_{cv,h}$ from the direct MCVA$_h$ procedure.

Remark 6. It is worth emphasizing that our asymptotic optimality established in Shibata's sense is somewhat weak and that it would therefore be desirable to call for more meaningful optimality properties. In particular, the notion of Shibata's asymptotic optimality is only a pointwise property and does not hold uniformly in the parameter space, which may yield inefficient small-sample performance for asymptotically optimal selection procedures, as demonstrated in Kabaila (2002); see also Leeb and Pötscher (2009, Sect. 3) for a related discussion.

Remark 7. It is worth emphasizing that further investigation is still needed to formally show the nonoptimality of the MMMA when used iteratively for multistep forecasting for $h > 1$. Here, we briefly describe the heuristics as follows. Analogous to MSFE$_h(\mathbf{w})$ defined in (5.2), we can define the weighted MSFE for the iterative averaging method as

$$\text{MSFE}_h^I(\mathbf{w}) = E\left(\text{tr}(\boldsymbol{\Sigma}_h^{-1}(\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h|T}^{I*}(\mathbf{w}))(\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h|T}^{I*}(\mathbf{w}))')\right), \tag{5.16}$$

where $\widehat{\mathbf{y}}_{T+h|T}^{I*}(\mathbf{w}))$ is given by (3.8) at forecast origin T. It is then useful to establish that for each $h \geq 1$, the second-order MSFE, $\text{MSFE}_h^I(\mathbf{w}) - K$, of the iterative h-step averaging forecast $\widehat{\mathbf{y}}_{T+h|T}^{I*}(\mathbf{w}))$ can be *uniformly* approximated in $\mathbf{w}$ by

$$L_{T,h}^I(\mathbf{w}) = \|\boldsymbol{\Psi}_h^*(\mathbf{w}) - \boldsymbol{\Psi}_h\|_{\mathbf{R}}^2 + \mathbf{p}'\mathbf{w}K/T, \tag{5.17}$$

where $\|\mathbf{A}\|_{\mathbf{B}} = \|\mathbf{A}'\mathbf{B}\mathbf{A}\|^{1/2}$ for a matrix $\mathbf{A}$ and a positive definite matrix $\mathbf{B}$, $\boldsymbol{\Psi}_h^*(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\bar{\boldsymbol{\Psi}}_h(p)$, $\bar{\boldsymbol{\Psi}}_h(p) = (\boldsymbol{\Psi}_h(p)' \quad \mathbf{0}_{K\times K(\bar{p}-p)})'$, $\boldsymbol{\Psi}_h = (\boldsymbol{\psi}_{h1}\boldsymbol{\psi}_{h2}\ldots)'$, and $\mathbf{R} = [\boldsymbol{\Gamma}(u-v)]$ for $u, v = 1, 2, \ldots$ with $\boldsymbol{\Gamma}(u) = E(\mathbf{y}_t\mathbf{y}'_{t+u})$. In (5.17), $\bar{\boldsymbol{\Psi}}_h(p)$ is understood as an infinite-dimensional matrix with elements $\boldsymbol{\Psi}_h(p)'$,

$p = 1, 2, \ldots$. Lastly, we wish to show that for $h \geq 1$ and the MMMA weight estimate $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{H}_T} C_T(\mathbf{w})$:

$$\lim_{T\to\infty} \frac{\mathrm{MSFE}_h^I(\widehat{\mathbf{w}}) - K}{\min_{\mathbf{w}\in\mathcal{H}_T} L_{T,h}^I(\mathbf{w})} \geq 1, \tag{5.18}$$

with the equality holding when $h = 1$. Equation (5.18) implies that if the MMMA weight estimate $\widehat{\mathbf{w}}$ is used iteratively for multistep forecasting, then the resulting second-order MSFE for $h > 1$ is ultimately greater than the lower bound $\min_{\mathbf{w}\in\mathcal{H}_T} L_{T,h}^I(\mathbf{w})$. In fact, we conjecture that the nonattainability of the lower bound for the iterative method, as previously claimed by Bhansali (1996), should carry over to the averaging setting under an infinite-order VAR process.

To establish (5.17) and (5.18), it is necessary to extend Lewis and Reinsel (1985), Bhansali (1996), and Ing and Wei (2003, 2005) to the general averaging case for VAR multistep forecasting. We do not pursue such an extension in the present paper and leave it for future research.

6. SIMULATION

This section presents the finite-sample forecast performance of our proposed approaches under correct specification and misspecification of forecasting models to shed some light on the relative merits of our iterative and direct forecast averaging methods.

6.1. Simulation Design

The data-generating process (DGP) we consider is the drifting bivariate ARMA(1,10) process, as previously considered by Schorfheide (2005):

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

where

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

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

We set the maximum lag order $\bar{p} = 3, 4, \ldots, 15$. The sample sizes $T = 100$, 200, and 500 are considered. We examine h-step-ahead forecast errors up to $h = 12$ in our simulation experiment. The number of simulation repetitions is $R = 2,500$. The computation in simulations and the empirical application in the online supplementary material 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 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, that is, $w_{\text{AIC}}(p) = \exp(-\text{AIC}(p)/2)/\sum_{j=1}^{\bar{p}} \exp(-\text{AIC}(j)/2)$. The SBIC weight specification 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 (2016b).

Given the selected and combined iterative and direct h-step ahead forecasts, we then compute and report the average of their weighted MSFE values, that is, using the inverse of $\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})$ given in (4.6) as weights, across $R = 2,500$ random samples from the DGP under investigation

$$\widehat{\text{MSFE}}_h(\bar{p}; M) = \frac{1}{2500}\sum_{r=1}^{2500}\Big[\text{tr}\Big(\widetilde{\boldsymbol{\Sigma}}_h^{(r)}(\bar{p})^{-1}\Big(\mathbf{y}_{T+h}^{(r)} - \widehat{\mathbf{y}}_{T+h|T}^{(r)}(\bar{p}; M)\Big) \times \Big(\mathbf{y}_{T+h}^{(r)} - \widehat{\mathbf{y}}_{T+h|T}^{(r)}(\bar{p}; M)\Big)'\Big)\Big],$$

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 based on the maximum lag order $\bar{p}$, and the superscript "(r)" indicates the rth simulation repetition.

6.2. Simulation Results

Figure 1 presents the iterative and direct multistep VAR forecast performance (measured by the relative MSFE to OLS(I)), where "D" and "I" in parentheses refer to direct and iterative multistep forecasts, respectively. To save space, we report here the results for $T = 100$ only; for the cases of $T = 200$ and $T = 500$, see Figures A2 and A3, respectively, in the online supplementary material. Several findings from our simulation results are summarized as follows.[9]

It can be seen from Figure 1 that in the absence of misspecification (i.e., $\alpha = 0$), the iterative multistep methods generally outperform the direct multistep methods. For example, the relative MSFEs of OLS(D) are greater than those of OLS(I), and OLS(D) deteriorates as h increases. This result is expected, as there

[9] More simulation results and discussions are provided in Section E in the online supplementary material.

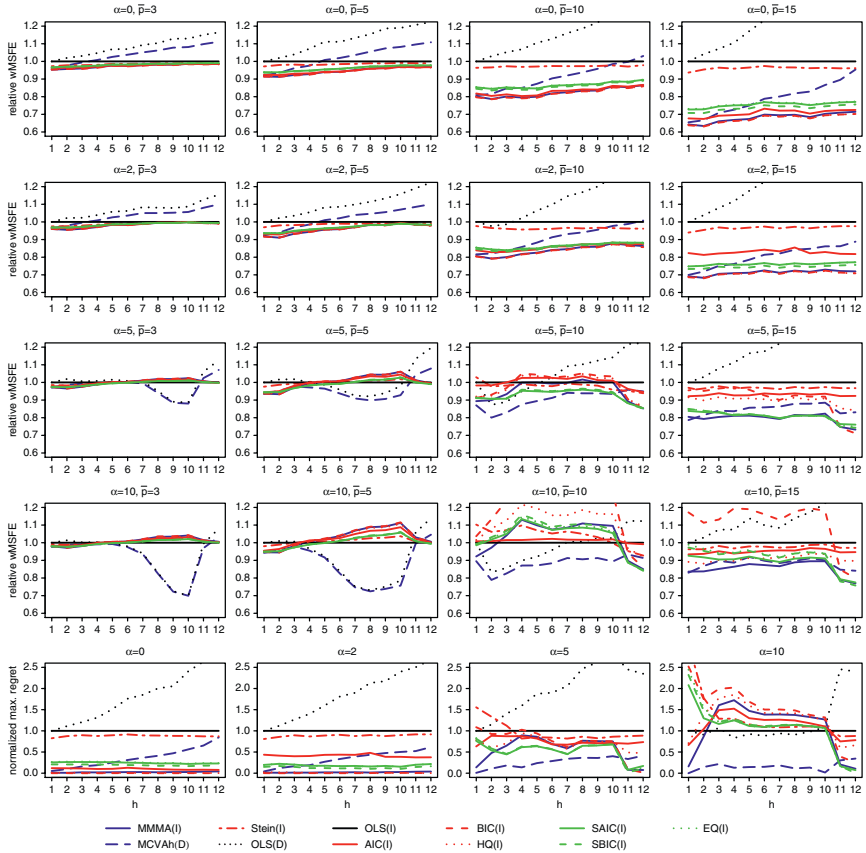

FIGURE 1. Multistep forecast performance under bivariate drifting ARMA(1,10): $T = 100$.

is no misspecification bias and the bias advantage of the direct multistep forecast methods thus does not appear to outweigh their variance disadvantage. Moreover, in the absence of model misspecification, it is clear that among all the iterative and direct methods considered, MMMA(I) performs best. 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 $\alpha = 2$, that is, the misspecification is mild.

Under the DGP with $\alpha \geq 5$, the outperformance of the iterative multistep 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 prespecified 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 finding 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, that is, when $\bar{p}$ is large enough. For example, as $\bar{p}$ further grows 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 finding indicates that as forecast horizons and autoregressions lengthen, the robustness of MCVA$_h$(D) is likely to be outweighed by the efficiency of MMMA(I).

We next investigate the effect of errors introduced by estimation of the combination weights on the forecast performance by comparing our averaging methods with the simple (equal) averaging (denoted by EQ(I)).[10] We find from Figure 1 (for $T = 100$) that in our simulation settings MMMA(I) is comparable to EQ(I) and both of them dominate MCVA$_h$(D) when no ($\alpha = 0$) or mild ($\alpha = 2$) misspecification is present and the number of candidate models is small ($\bar{p} \leq 5$). The dominance of MMMA(I) and EQ(I) over MCVA$_h$(D) vanishes as the sample size increases (Figures A2 ($T = 200$) and A3 ($T = 500$) in the online supplementary material). It can also be seen that MMMA(I) tends to outperform EQ(I) as $\bar{p}$ increases, that is, the number of candidate models becomes large. This is expected because when misspecification is not severe, unlike EQ(I) imposing equal weights, MMMA(I) places zero or small weights on large (over-specified) VAR models that lead to poor performance (Figures A4–A6 in the online supplementary material). On the other hand, MCVA$_h$(D) appears to dominate MMMA(I) and EQ(I) in most cases when $\alpha = 5$ and 10. Similar to Cheng and Hansen's (2015) simulation results, we also find overall that relative to our averaging methods, EQ(I) is somewhat sensitive to $\bar{p}$, particularly for $h = 1$ and $\alpha = 5$ and 10 (Table A1 in the online supplementary material). We note that these findings may not apply to more complicated DGPs that can generate large estimation errors of estimated weights.

We also compare the unweighted MSFEs (i.e., a simple sum of the MSFEs associated with individual response variables) for the entire VAR system (which are not reported here to save space and are available upon request). From there, we find that in most cases, the relative forecast performances of our methods when compared with other competitors are qualitatively similar to those based on the weighted MSFEs discussed above.

[10] We thank an anonymous referee for drawing our attention to this important issue.

As our simulation results reveal, the ranking of the competing methods based on MSFEs may vary with the prespecified maximum lag order $\bar{p}$. To address uncertainty arising from the choice of $\bar{p}$, the last row of Figure 1 presents the normalized maximum regret based on MSFEs over different values of $\bar{p}$. 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,

$$\widehat{\mathrm{R}}_h(\bar{p}; M) = \widehat{\mathrm{MSFE}}_h(\bar{p}; M) - \min_{M \in \mathcal{M}} \widehat{\mathrm{MSFE}}_h(\bar{p}; M). \tag{6.1}$$

Given $\widehat{\mathrm{R}}_h(\bar{p}; M)$, 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 the last row of Figure 1 reveal a clearer dominance of MMMA(I) (MCVA_h(D)) over other competing methods under no or mild (large) model misspecification when the uncertainty from the choice of $\bar{p}$ is taken into account.

7. CONCLUSION

This article has employed a frequentist multiple-equation model averaging approach based on the MMMA and MCVA_h criteria for combinations of multistep forecasts with VAR models. The former criterion is designed for iterative multistep VAR forecast averaging, while the latter aims to deal with the issue of the serial correlation that is due to overlapping data under the direct multistep forecasting framework. The proposed methods are straightforward to implement because our procedures are based on least-squares estimation and quadratic programming to obtain 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 (reported in the online supplementary material), illustrating the usefulness of our methods as econometric tools for multistep VAR forecast combinations.

Several directions built on this article are worth exploring in future research. For example, it would be useful to introduce dimension reduction techniques, such as shrinkage or factors, into the framework of VAR forecast averaging. It would also be interesting to extend our methodology to nonstationary processes.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0266466619000434 . We provide proofs of the theoretical results, addi-

tional simulation results, 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.

REFERENCES

Andersson, M.K. & 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.

Basu, S. & G. Michailidis (2015) Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics* 43, 1535–1567.

Bates, J.M. & C.W.J. Granger (1969) The combination of forecasts. *Journal of the Operational Research Society* 20, 451–468.

Bhansali, R.J. (1996) Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics* 48, 577–602.

Bhansali, R.J. (1997) Direct autoregressive predictors for multistep prediction: Order selection and performance relative to the plug in predictors. *Statistica Sinica* 7, 425–449.

Bhansali, R.J. (1999) Parameter estimation and model selection for multistep prediction of time series: A review. In S. Gosh (ed.). *Asymptotics, Nonparametrics and Time Series*, pp. 201–225. Marcel Dekker.

Chen, R., L. Yang, & C. Hafner (2004) Nonparametric multistep-ahead prediction in time series analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 66, 669–686.

Cheng, T.-C.F., C.-K. Ing, & S.-H. Yu (2015) Toward optimal model averaging in regression models with time series errors. *Journal of Econometrics* 189, 321–334.

Cheng, X. & B.E. Hansen (2015) Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics* 186, 280–293.

Chevillon, G. (2007) Direct multi-step estimation and forecasting. *Journal of Economic Surveys* 21, 746–785.

Chevillon, G. & D.F. Hendry (2005) Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting* 21, 201–218.

Clark, T.E., & McCracken (2010) Averaging forecasts from VARs with uncertain instabilities. *Journal of Applied Econometrics* 25, 5–29.

Clemen, R.T. (1989) Combining forecast: A review and annotated bibliography. *International Journal of Forecasting* 5, 559–581.

Diebold, F.X., & J.A. Lopez (1996) Forecast evaluation and combination. In *Statistical Methods in Finance. Handbook of Statistics*, vol. 14, pp. 241–268. Elsevier.

Doan, T., R. Litterman, & C. Sims (1984) Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews* 3, 1–100.

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

Findley, D.F. & 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, 415–450.

Fujikoshi, Y. & K. Satoh (1997) Modified AIC and C_p in multivariate linear regression. *Biometrica* 84, 707–716.

Gao, Y., X. Zhang, S. Wang, & G. Zou (2016) Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics* 192, 139–151.

Granger, C. (1989) Combining forecasts twenty years later. *Journal of Forecast* 8, 167–173.

Hansen, B.E. (2007) Least squares model averaging. *Econometrica* 75, 1175–1189.

Hansen, B.E. (2008) Least-squares forecast averaging. *Journal of Econometrics* 146, 342–350.

Hansen, B.E. (2016a) Efficient shrinkage in parametric models. *Journal of Econometrics* 190, 115–132.

Hansen, B.E. (2016b) Stein Combination Shrinkage for Vector Autoergressions, Discussion paper. University of Wisconsin.
Hansen, B.E. & J.S. Racine (2012) Jackknife model averaging. *Journal of Econometrics* 167, 38–46.
Hendry, D. & M. Clements (2004) Pooling of forecasts. *Econometric Journal* 7, 1–31.
Hsu, N.-J., H.-L. Hung, & Y.-M. Chang (2008) Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis* 52, 3645–3657.
Ing, C.-K. (2003) Multistep prediction in autoregressive processes. *Econometric Theory* 19, 254–279.
Ing, C.-K. & C.-Z. Wei (2003) On same-realization prediction in an infinite-order autoregressive process. *Journal of Multivariate Analysis* 85, 130–155.
Ing, C.-K. & C.-Z. Wei (2005) Order selection for same-realization predictions in autoregressive processes. *Annals of Statistics* 33, 2423–2474.
Kabaila, P. (2002) On variable selection in linear regression. *Econometric Theory* 18, 913–925.
Kunitomo, N. & T. Yamamoto (1985) Properties of predictors in misspecified autoregressive time series models. *Journal of the American Statistical Association* 80, 941–950.
Lee, W. & Y. Liu (2012) Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis* 111, 241–255.
Leeb, H. & B.M. Pötscher (2009) Model selection. In T.G. Andersen, R.A. Davis, J.-P. Kreib, and T. Mikosch (eds.). *The Handbook of Financial Time Series*, pp. 889–925. Springer.
Lewis, R. & G. Reinsel (1985) Prediction of multivariate time Series by autoregressive model fitting. *Journal of Multivariate Analysis* 16, 393–411.
Li, K.-C. (1987) Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* 15, 958–975.
Litterman, R.B. (1986) Forecasting with Bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics* 4, 25–38.
Liu, Q., R. Okui, & A. Yoshimura (2016) Generalized least squares model averaging. *Econometric Reviews* 35, 1692–1752.
Lütkepohl, H. (2005) *New Introduction to Multiple Time Series Analysis*. Springer.
Marcellino, M., J. Stock, & M. Watson (2006) A comparison of direct and iterated multistep AR method for forecasting macroeconomic time series. *Journal of Econometrics* 135, 499–526.
McQuarrie, A.D. & C.-L. Tsai (1998) *Regression and Time Series Model Selection*. World Scientific Publishing Co. Pte. Ltd.
Pesaran, M.H., A. Pick, & A. Timmermann (2011) Variable selection, estimation and inference for multi-period forecasting problems. *Journal of Econometrics* 164, 173–187.
Phillips, P.C.B. (1995) Automated forecasts of Asia-Pacific economic activity. *Asia-Pacific Economic Review* 1, 92–102.
Phillips, P.C.B. (1996) Econometric model determination. *Econometrica* 64, 763–812.
Reid, D.J. (1968) Combining three estimates of gross domestic product. *Economica* 35, 431–444.
Reid, D.J. (1969) A Comparative Study of Time Series Prediction Techniques on Economic Data, Ph.d. Thesis. University of Nottingham, Nottingham, UK.
Ren, Y., Z. Xiao, & X. Zhang (2013) Two-step adaptive model selection for vector autoregressive processes. *Journal of Multivariate Analysis* 116, 349–364.
Ren, Y. & X. Zhang (2010) Subset selection for vector autoregressive processes via adaptive Lasso. *Statistics and Probability Letters* 80, 1705–1712.
Schorfheide, F. (2005) VAR forecasting under misspecification. *Journal of Econometrics* 128, 99–136.
Shao, J. (1997) An asymptotic theory for linear model selection. *Statistica Sinica* 7, 221–264.
Shibata, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8, 147–164.
Shibata, R. (1981) An optimal selection of regression variables. *Biometrika* 68, 45–54.
Shibata, R. (1983) Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics* 35, 415–423.
Sims, C.A. (1980) Macroeconomics and reality. *Econometrica* 48, 1–48.

Sparks, R., D. Coutsourides, & L. Troskie (1983) The multivariate C_p. *Communications in Statistics—Theory and Methods* 12, 1775–1793.

Stock, J. & M. Watson (2006) Forecast with many predictors. In *Handbook of Economic Forecasting*, pp. 515–554. Elsevier Press.

Tiao, G.C. & G.E.P. Box (1981) Modeling multiple times series with applications. *Journal of the American Statistical Association* 76, 802–816.

Timmermann, A. (2006) Forecast combinations. In *Handbook of Economic Forecasting*, pp. 135–196. Elsevier Press.

Varmuza, K. & P. Filzmoser (2009) *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press.

Wan, A.T., X. Zhang, & G. Zou (2010) Least squares model averaging by Mallows criterion. *Journal of Econometrics* 156, 277–283.

Yanagihara, H. & K. Satoh (2010) An unbiased criterion for multivariate ridge regression. *Journal of Multivariate Analysis* 101, 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, 348–368.

Zhang, X., A.T. Wan, & G. Zou (2013) Model averaging by Jackknife criterion in models with dependent data. *Journal of Econometrics* 174, 82–94.