Online Supplementary Material on "Optimal Multi-step VAR Forecast Averaging"

Jen-Che Liao and Wen-Jen Tsay

A Derivations of equation (3.5)

We first write:

$$
\begin{aligned}
&(T-\bar{p})\cdot \operatorname{tr}\left(\widetilde{\mathbf{\Sigma}}(\bar{p})^{-1}\widehat{\mathbf{\Sigma}}^{*}(\mathbf{w})\right) = \operatorname{tr}\left(\widetilde{\mathbf{\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{\mathbf{\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{\mathbf{\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.\\
&\qquad\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

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

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 $(T - \bar{p}) \cdot \operatorname{tr}\left(\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\widehat{\boldsymbol{\Sigma}}^*(\mathbf{w})\right)$ 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 $t = \bar{p}, \ldots, T-1$ and $p = 1, \ldots, \bar{p}$ 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).

B Efficient Computation of $CV_{T,h}(\mathbf{w})$

First note that $\widetilde{\boldsymbol{\epsilon}}_{t+h}(p)$ defined in (4.4) in the main text 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:

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

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 (B.1) follows from using the following formula for $\widetilde{\boldsymbol{\Psi}}_{h,t}(p)$:

$$\begin{aligned}
\widetilde{\boldsymbol{\Psi}}_{h,t}(p) &= \left(\mathbf{Z}_h(p)'\mathbf{Z}_h(p) - \mathbf{Z}_{t:h}(p)'\mathbf{Z}_{t:h}(p)\right)^{-1}\left(\mathbf{Z}_h(p)'\mathbf{Y}_h - \mathbf{Z}_{t:h}(p)'\mathbf{Y}_{t:h}\right) \\
&= \widehat{\boldsymbol{\Psi}}_h(p) - \left(\mathbf{Z}_h(p)'\mathbf{Z}_h(p)\right)^{-1}\mathbf{Z}_{t:h}(p)'\Big(\mathbf{Y}_{t:h} - \left(\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p)\right)^{-1}\mathbf{Z}_{t:h}(p)\left(\mathbf{Z}_h(p)'\mathbf{Z}_h(p)\right)^{-1}\mathbf{Z}_h(p)'\mathbf{Y}_h \\
&\quad + \left(\mathbf{I}_{\ell_{ht}} - \mathbf{P}_{t:h}(p)\right)^{-1}\mathbf{P}_{t:h}(p)\mathbf{Y}_{t:h}\Big). \qquad \text{(B.2)}
\end{aligned}$$

Formulae (B.1) and (B.2) 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 (B.1), 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 C.4 in the 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 (B.1) 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 (B.1) 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.:

$$e'_{ht}\widetilde{\mathbf{P}}_{t:h}(p) = e'_{ht}(\mathbf{D}_{t:h}(p)(\mathbf{P}_h(p) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{S}_{t:h}). \tag{B.3}$$

Lastly, stacking (B.3) 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. *(Shortcut formula) 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:*

$$\widetilde{\mathbf{P}}_h(p) = \mathbf{E}_h\left(\mathbf{D}_h(p)(\mathbf{P}_h(p) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{S}_h\right). \tag{B.4}$$

Alternatively, (B.4) *can also be expressed as:*

$$\widetilde{\mathbf{P}}_h(p) = \widetilde{\mathbf{D}}_h(p)(\mathbf{P}_h(p) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{I}_{T-\bar{p}-h+1}, \tag{B.5}$$

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 (B.4) 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, (B.4) reduces to equation (1.4) of Li (1987): $\widetilde{\mathbf{P}}(p) = \mathbf{D}(p)(\mathbf{P}(p) - \mathbf{I}_{T-\bar{p}}) + \mathbf{I}_{T-\bar{p}}$.

C Mathematical proof

C.1 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:

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

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 $\mathbf{C}$, $\mathrm{tr}(\mathbf{ABC}) = \mathrm{vec}(\mathbf{A}')'(\mathbf{I} \otimes \mathbf{B})\mathrm{vec}(\mathbf{C})$. The first two terms on the right-hand side of equation (C.1) 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 (C.1). 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:

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

where the third equality follows from the property that $\mathrm{vec}(\mathbf{AB}) = (\mathbf{B}' \otimes \mathbf{I})\mathrm{vec}(\mathbf{A})$ 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$(p), p = 1, \dots, \bar{p}$, we define:

$$\begin{aligned}\xi_{1T}(p) &\equiv \frac{1}{\sqrt{T-\bar{p}}}\mathrm{vec}(\boldsymbol{\mu}')' \left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(p)) \otimes \mathbf{I}_K\right) \left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right) \mathrm{vec}(\mathbf{e}') \\ &= \frac{1}{\sqrt{T-\bar{p}}}\mathrm{vec}(\boldsymbol{\mu}')' \left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(p)) \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right) \mathrm{vec}(\mathbf{e}'), \qquad \text{(C.3)}\end{aligned}$$

where $\xi_{1T}(p)$ satisfies $\xi_{1T}(p)/\Gamma_1(p)^{1/2} \xrightarrow{d} N(0,1)$ with $\Gamma_1(p) = \mathrm{plim}\ \boldsymbol{\nu}(p)' \left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}\right) \boldsymbol{\nu}(p)/(T-\bar{p})$, $\boldsymbol{\nu}(p)' \equiv \mathrm{vec}(\boldsymbol{\mu}')' \left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}(p)) \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)$, and $\xrightarrow{d}$ denotes convergence in distribution. This large sample result for (C.3) is an application of Proposition 15.1 of Lütkepohl (2005, p.533). Equation (C.3) says that $\xi_{1T}(p)/\Gamma_1(p)^{1/2}$ is a standard normal random variable. Since the term $r_{1T}(\mathbf{w})$ in (C.2) 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:

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

As a result, we write:

$$\begin{aligned}&-2\mathrm{vec}((\mathbf{e}'\mathbf{P}(p))' \left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\mathrm{vec}(\mathbf{e}') \\ &= -2\mathrm{vec}(\mathbf{e}'\mathbf{Z}(p))' \left(\left(\mathbf{Z}(p)'\mathbf{Z}(p)\right)^{-1} \otimes \mathbf{I}_K\right)(\mathbf{Z}(p)' \otimes \mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)\mathrm{vec}(\mathbf{e}') \\ &= -2\mathrm{vec}(\mathbf{e}'\mathbf{Z}(p))' \left(\left(\mathbf{Z}(p)'\mathbf{Z}(p)\right)^{-1} \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{Kp} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right)(\mathbf{Z}(p)' \otimes \mathbf{I}_K)\mathrm{vec}(\mathbf{e}') \\ &= -2\mathrm{vec}(\mathbf{e}'\mathbf{Z}(p))' \left((\mathbf{Z}(p)'\mathbf{Z}(p)) \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})\right)^{-1}\mathrm{vec}(\mathbf{e}'\mathbf{Z}(p)), \qquad \text{(C.4)}\end{aligned}$$

where we use $\left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}}\right)^{-1} \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{Kp} \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}\right) = \left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}}\right) \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p})\right)^{-1}$.

To evaluate (C.4), consider:

$$
\begin{aligned}
&\left(\left(\mathbf{Z}(p)'\mathbf{Z}(p)\right)^{-1} \otimes \boldsymbol{\Sigma}^{1/2}\right) \mathrm{vec}\left(\mathbf{e}'\mathbf{Z}(p)\right) \\
&\quad = (T-\bar{p})^{1/2} \left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}} \right)^{-1/2} \otimes \boldsymbol{\Sigma}^{1/2} \right) \mathrm{vec} \left((T-\bar{p})^{-1} \sum_{t=\bar{p}+1}^{T} \boldsymbol{\varepsilon}_t \mathbf{z}_t' \right) \\
&\quad = (T-\bar{p})^{1/2} \mathrm{vec} \left((T-\bar{p})^{-1} \boldsymbol{\Sigma}^{1/2} \left(\sum_{t=\bar{p}+1}^{T} \boldsymbol{\varepsilon}_t \mathbf{z}_t(p)' \right) \left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}} \right)^{-1/2} \right), \qquad \text{(C.5)}
\end{aligned}
$$

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 $\mathbf{e}' = (\boldsymbol{\varepsilon}_{\bar{p}+1}, \ldots, \boldsymbol{\varepsilon}_T)$, and for the second equality we use the fact that $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$ for conformable matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. Denote $\boldsymbol{\Gamma}_2(p) = \mathrm{plim}\, \mathbf{Z}(p)'\mathbf{Z}(p)/(T-\bar{p})$, and let $\ell(p)$ be a sequence of $K^2 p \times 1$ 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:

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

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

$$
s_T / v_T \xrightarrow{d} N(0,1). \qquad \text{(C.6)}
$$

Putting together the arguments in (C.5) and (C.6), it is obvious to derive the following limiting distribution result that allows the lag order p to increase with the sample size:

$$
\frac{\left(\ell(p)' \mathrm{vec}(\boldsymbol{\Sigma}^{-1/2} \mathbf{e}' \mathbf{Z}(p) \boldsymbol{\Gamma}_2(p)^{-1/2}) \right) / \sqrt{T-\bar{p}}}{\left(\ell(p)'\ell(p) \right)^{1/2}} \equiv \ell(p)' \boldsymbol{\phi}_T(p) \xrightarrow{d} N(0,1), \qquad \text{(C.7)}
$$

where

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

By the Cramér-Wold theorem, for any $p \in \{1, 2, \ldots, \bar{p}\}$, $\boldsymbol{\phi}_T(p)$ then converges in distribution to a $K^2 p$-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 $K(p) = K^2 p$.

We now turn back to (C.4). Using the asymptotic normality results discussed above, the consistency of $\widetilde{\boldsymbol{\Sigma}}(\bar{p})$ (since the maximum lag order $\bar{p} = \bar{p}_T$ increases with sample size), and the fact that (C.4) (ignoring the constant -2 for a moment) is a quadratic form in multivariate normal random variables, (C.4) is asymptotically equivalent to $\xi_{2T}(p)$, i.e., for each $p \in \{1, 2, \ldots, \bar{p}\}$:

$$\frac{1}{\sqrt{T-\bar{p}}}\operatorname{vec}(\mathbf{e}'\mathbf{Z}(p))' \left(\left(\frac{\mathbf{Z}(p)'\mathbf{Z}(p)}{T-\bar{p}} \right) \otimes \widetilde{\boldsymbol{\Sigma}}(\bar{p}) \right)^{-1} \frac{1}{\sqrt{T-\bar{p}}}\operatorname{vec}(\mathbf{e}'\mathbf{Z}(p)) - \xi_{2T}(p) = o_p(1). \tag{C.8}$$

Denote $\xi^*_{2T}(\mathbf{w}) \equiv \sum_{p=1}^{\bar{p}} w(p)\xi_{2T}(p)$. It then follows that $E(\xi^*_{2T}(\mathbf{w})) = \sum_{p=1}^{\bar{p}} w(p)E(\xi_{2T}(p)) = \sum_{p=1}^{\bar{p}} w(p)K(p)$. This further implies:

$$E(r_{2T}(\mathbf{w})) = -2\sum_{p=1}^{\bar{p}} w(p)K^2 p = -2K^2\mathbf{p}'\mathbf{w}. \tag{C.9}$$

Combining (C.1), (C.2), (C.3), and (C.9), we conclude that $E(C_T(\mathbf{w})) = (T-\bar{p})E(L_T(\mathbf{w}))$, completing the proof.

C.2 Proof of Theorem 2

Similar to (C.1), we write and expand the sum of squared leave-h-out cross-validation residuals as:

$$\begin{aligned}
&\operatorname{tr}\left((\mathbf{Y}_h - \widetilde{\boldsymbol{\mu}}^*_h(\mathbf{w}))\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}(\mathbf{Y}_h - \widetilde{\boldsymbol{\mu}}^*_h(\mathbf{w}))' \right) \\
&\quad = \operatorname{tr}\left((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}^*_h(\mathbf{w}) + \mathbf{e}_h)\, \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1} \left((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}^*_h(\mathbf{w}) + \mathbf{e}_h \right)' \right) \\
&\quad = \operatorname{vec}((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}^*_h(\mathbf{w}))')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1} \right) \operatorname{vec}((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}^*_h(\mathbf{w}))') \\
&\qquad + \operatorname{vec}(\mathbf{e}'_h)' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1} \right) \operatorname{vec}(\mathbf{e}'_h) \\
&\qquad + 2\operatorname{vec}((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}^*_h(\mathbf{w}))')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1} \right) \operatorname{vec}(\mathbf{e}'_h),
\end{aligned} \tag{C.10}$$

where the first term on the right-hand side of (C.10) 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 (C.10) into:

$$\begin{aligned}
&2\text{vec}((\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h^*(\mathbf{w}))')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right) \text{vec}(\mathbf{e}_h') \\
&\quad = 2\text{vec}({\boldsymbol{\mu}_h}')' \left((\mathbf{I}_{T-\bar{p}-h+1} - \widetilde{\mathbf{P}}_h^*(\mathbf{w}))' \otimes \mathbf{I}_K\right) \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right) \text{vec}(\mathbf{e}_h') \\
&\qquad - 2\text{vec}(\mathbf{e}_h')'(\widetilde{\mathbf{P}}_h^*(\mathbf{w})' \otimes \mathbf{I}_K) \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}\right) \text{vec}(\mathbf{e}_h') \\
&\quad \equiv \widetilde{r}_{1Th}(\mathbf{w}) + \widetilde{r}_{2Th}(\mathbf{w}). \qquad \text{(C.11)}
\end{aligned}$$

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 (C.2) and (C.3), 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 (C.2), $\widetilde{r}_{1Th}(\mathbf{w})$ 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 $E(\widetilde{r}_{2Th}(\mathbf{w})) = E(\text{tr}(\widetilde{\mathbf{P}}_h^*(\mathbf{w})\mathbf{e}_h\mathbf{e}_h')) = \text{tr}(\widetilde{\mathbf{P}}_h^*(\mathbf{w})E(\mathbf{e}_h\mathbf{e}_h')) = 0$ 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.

C.3 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:

$$\begin{aligned}
C_T^*(\mathbf{w}) &= L_T(\mathbf{w}) + K + \frac{2}{T-\bar{p}}\text{vec}(\boldsymbol{\mu}')' \left((\mathbf{I}_{T-\bar{p}} - \mathbf{P}^*(\mathbf{w}))' \otimes \mathbf{I}_K\right) \left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right) \text{vec}(\mathbf{e}') \\
&\quad - \frac{2}{T-\bar{p}}\text{vec}(\mathbf{e}')'(\mathbf{P}^*(\mathbf{w})' \otimes \mathbf{I}_K) \left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right) \text{vec}(\mathbf{e}') + \frac{2K^2\mathbf{p}'\mathbf{w}}{T-\bar{p}}. \qquad \text{(C.12)}
\end{aligned}$$

Based on (C.12), to prove (5.11) we need to verify the following two uniform convergence results of the form:

$$\sup_{\mathbf{w}\in\mathcal{H}_T} \left|\mathrm{vec}(\boldsymbol{\mu}')' \left(\mathbf{P}^*(\mathbf{w}) \otimes \mathbf{I}_K\right) \left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right) \mathrm{vec}(\mathbf{e}')\right| / V_T(\mathbf{w}) = o_p(1), \tag{C.13}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_T} \left|\mathrm{vec}(\mathbf{e}')'(\mathbf{P}^*(\mathbf{w})' \otimes \mathbf{I}_K) \left(\mathbf{I}_{T-\bar{p}} \otimes \boldsymbol{\Sigma}^{-1}\right) \mathrm{vec}(\mathbf{e}') - K^2\mathbf{p}'\mathbf{w}\right| / V_T(\mathbf{w}) = o_p(1). \tag{C.14}$$

To verify (C.13) and (C.14), we show the following statements:

$$T\lambda_{\max}\left(\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\right) = O_p(1), \tag{C.15}$$

$$T^{-1}\bar{p}^{-1}\mathrm{vec}(\mathbf{e}')' \left(\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}' \otimes \mathbf{I}_K\right) \mathrm{vec}(\mathbf{e}') = O_p(1), \tag{C.16}$$

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

We now take (C.15). 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:

$$E\left(\left\|\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right\|_1^2\right) \le E\left(\left\|\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right\|^2\right) \le C\frac{\bar{p}^2K^2}{T-\bar{p}} = C\frac{\bar{p}^2}{T-\bar{p}}, \tag{C.17}$$

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:

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p}) - \boldsymbol{\Gamma}^{-1}(\bar{p})\right\|_1 &= \left\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\left(\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right)\boldsymbol{\Gamma}^{-1}(\bar{p})\right\|_1 \\ &\le \left\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\right\|_1\left\|\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right\|_1\left\|\boldsymbol{\Gamma}^{-1}(\bar{p})\right\|_1,\end{aligned} \tag{C.18}$$

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

$$\begin{aligned} E\left(\left\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p}) - \boldsymbol{\Gamma}^{-1}(\bar{p})\right\|_1\right) &\leq \bar{p}^{2+\theta} E\left(\left\|\widehat{\boldsymbol{\Gamma}}_T(\bar{p}) - \boldsymbol{\Gamma}(\bar{p})\right\|_1\right) C \\ &\leq C\frac{\bar{p}}{(T-\bar{p})^{1/2}}\bar{p}^{2+\theta} \\ &= C\frac{\bar{p}^{3+\theta}}{(T-\bar{p})^{1/2}} \\ &\leq C\left(\frac{\bar{p}^{6+\delta_1}}{T-\bar{p}}\right)^{1/2}, \end{aligned} \tag{C.19}$$

where the first inequality uses Assumption 2(c). As in the univariate case (see, e.g., Berk (1974, p.491)), $\|\boldsymbol{\Gamma}^{-1}(p)\|_1$ 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 (C.17), and the last inequality follows by setting $2\theta \leq \delta_1$.

Based on (C.19), it is not hard to see $E\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\|_1 \leq C$, provided that $\bar{p}^{6+\delta_1} = O(T)$ and $E\|\boldsymbol{\Gamma}^{-1}(\bar{p})\|_1 \leq C$. 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:

$$E\left\|\widehat{\boldsymbol{\Gamma}}_T^{-1}(\bar{p})\right\|_1 \equiv E\left\|\left(\frac{\overline{\mathbf{Z}}'\overline{\mathbf{Z}}}{T-\bar{p}}\right)^{-1}\right\|_1 = E\left[\lambda_{\max}\left(\left(\frac{\overline{\mathbf{Z}}'\overline{\mathbf{Z}}}{T-\bar{p}}\right)^{-1}\right)\right] < \infty, \tag{C.20}$$

where the second equality in (C.20) follows since the matrix $\overline{\mathbf{Z}}'\overline{\mathbf{Z}}$ is symmetric. Thus, (C.15) follows from combining (C.20) 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 (C.16) 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:

$$E\left\| \frac{1}{\sqrt{T-\bar{p}}} \sum_{t=\bar{p}+1}^{T} \boldsymbol{\varepsilon}_t \mathbf{z}_t(\bar{p})' \right\|^2 \leq \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{\ell=1}^{\bar{p}} E\left[(T-\bar{p})^{-1} \left| \sum_{t=\bar{p}+1}^{T} \varepsilon_{it} y_{j,t-\ell} \right|^2 \right]. \tag{C.21}$$

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 (C.21) is bounded by:

$$CE\left((T-\bar{p})^{-1} \sum_{t=\bar{p}+1}^{T} |y_{j,t-\ell} y_{j',t-\ell}| \right) = C(T-\bar{p})^{-1} \sum_{t=\bar{p}+1}^{T} E\left(|y_{j,t-\ell} y_{j',t-\ell}| \right) = O(1). \tag{C.22}$$

Lastly, the lemma follows from combining (C.21), (C.22), 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 (C.15) and (C.16), we first verify (C.13) by writing:

$$\begin{aligned}
&\xi_T^{*\,-1} \left| \mathrm{vec}(\boldsymbol{\mu}')' \left(\mathbf{P}^*(\mathbf{w}) \otimes \mathbf{I}_K \right) \left(\mathbf{I}_{T-\bar{p}} \otimes \mathbf{\Sigma}^{-1} \right) \mathrm{vec}(\mathbf{e}') \right| \\
&= \xi_T^{*\,-1} \left| \mathrm{vec}(\boldsymbol{\mu}')' \left(\overline{\mathbf{P}} \otimes \mathbf{I}_K \right) \left(\mathbf{P}^*(\mathbf{w}) \otimes \mathbf{I}_K \right) \left(\mathbf{I}_{T-\bar{p}} \otimes \mathbf{\Sigma}^{-1} \right) \mathrm{vec}(\mathbf{e}') \right| \\
&\leq \xi_T^{*\,-1} \left\{ \mathrm{vec}(\boldsymbol{\mu}')' \left(\overline{\mathbf{P}} \otimes \mathbf{I}_K \right) \left(\overline{\mathbf{P}} \otimes \mathbf{I}_K \right) \mathrm{vec}(\boldsymbol{\mu}') \mathrm{vec}(\mathbf{e}')' \left(\mathbf{I}_{T-\bar{p}} \otimes \mathbf{\Sigma}^{-1} \right) \left(\mathbf{P}^*(\mathbf{w}) \otimes \mathbf{I}_K \right) \left(\mathbf{I}_{T-\bar{p}} \otimes \mathbf{\Sigma}^{-1} \right) \mathrm{vec}(\mathbf{e}') \right\}^{1/2} \\
&\leq \xi_T^{*\,-1} \left\{ \mathrm{vec}(\boldsymbol{\mu}')' \left(\overline{\mathbf{P}} \otimes \mathbf{I}_K \right) \mathrm{vec}(\boldsymbol{\mu}') \mathrm{vec}(\mathbf{e}')' \left(\mathbf{I}_{T-\bar{p}} \otimes \mathbf{\Sigma}^{-1} \right) \left(\overline{\mathbf{P}} \otimes \mathbf{I}_K \right) \left(\mathbf{I}_{T-\bar{p}} \otimes \mathbf{\Sigma}^{-1} \right) \mathrm{vec}(\mathbf{e}') \right\}^{1/2} \\
&= \Bigg\{ \underbrace{\left[\bar{p} \xi_T^{*\,-2} \mathrm{vec}(\boldsymbol{\mu}')' \left(\overline{\mathbf{P}} \otimes \mathbf{I}_K \right) \mathrm{vec}(\boldsymbol{\mu}') \right]}_{= o_p(1) \text{ by Assumption 2(a)}} \left[\bar{p}^{-1} \mathrm{vec}(\mathbf{e}')' \left(\mathbf{I}_{T-\bar{p}} \otimes \mathbf{\Sigma}^{-1} \right) \left(\overline{\mathbf{P}} \otimes \mathbf{I}_K \right) \left(\mathbf{I}_{T-\bar{p}} \otimes \mathbf{\Sigma}^{-1} \right) \mathrm{vec}(\mathbf{e}') \right] \Bigg\}^{1/2} \\
&= o_p(1),
\end{aligned} \tag{C.23}$$

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

$$\begin{aligned}
&\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\mathbf{I}_{T-\bar{p}}\otimes\boldsymbol{\Sigma}^{-1}\right)\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}}\otimes\boldsymbol{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}')\\
&\le C\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\overline{\mathbf{P}}\otimes\mathbf{I}_K\right)\text{vec}(\mathbf{e}')\\
&\le C\underbrace{T\lambda_{\max}\left(\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\right)}_{=O_p(1)\text{ by (C.15)}}\underbrace{T^{-1}\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\otimes\mathbf{I}_K\right)\text{vec}(\mathbf{e}')}_{=O_p(1)\text{ by (C.16)}}\\
&=O_p(1). \qquad\qquad (C.24)
\end{aligned}$$

The second inequality in (C.24) 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 $\text{tr}(\mathbf{AB}) \le \lambda_{\max}(\mathbf{A})\text{tr}(\mathbf{B})$ for squared matrices $\mathbf{A}$ and $\mathbf{B}$ with $\mathbf{A}$ being symmetric and $\mathbf{B}\ge 0$.

We next move to (C.14). Using Assumption 2(a) and a similar argument to show (C.24), we have:

$$\begin{aligned}
&\xi_T^{*\,-1}\left|\text{vec}(\mathbf{e}')'(\mathbf{P}^*(\mathbf{w})'\otimes\mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}}\otimes\boldsymbol{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}') - K^2\mathbf{p}'\mathbf{w}\right|\\
&\le \xi_T^{*\,-1}\text{vec}(\mathbf{e}')'(\overline{\mathbf{P}}\otimes\mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}}\otimes\boldsymbol{\Sigma}^{-1}\right)\text{vec}(\mathbf{e}') + \underbrace{\xi_T^{*\,-1}K^2\bar{p}}_{\substack{=o_p(1)\\ \text{by Assumption 2}(a)}}\\
&\le \underbrace{\xi_T^{*\,-1}\bar{p}}_{\substack{=o_p(1)\\ \text{by Assumption 2}(a)}}\underbrace{T\lambda_{\max}\left(\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\right)}_{=O_p(1)\text{ by (C.15)}}\underbrace{T^{-1}\bar{p}^{-1}\text{vec}(\mathbf{e}')'\left(\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\otimes\mathbf{I}_K\right)\text{vec}(\mathbf{e}')}_{=O_p(1)\text{ by (C.16)}}+o_p(1)\\
&=o_p(1). \qquad\qquad (C.25)
\end{aligned}$$

The last thing to show is (5.10). 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:

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

(C.26)

and thus:

$$
\begin{aligned}
L_T(\mathbf{w}) - V_T(\mathbf{w}) = &-\frac{2}{T-\bar{p}} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{A}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\mathbf{e}\right) \\
&+ \frac{1}{T-\bar{p}} \left\{\mathrm{tr}\left(\mathbf{P}^*(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}^*(\mathbf{w})\right) - E\left[\mathrm{tr}\left(\mathbf{P}^*(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}^*(\mathbf{w})\right)\right]\right\}.
\end{aligned}
\tag{C.27}
$$

Take the first term on the right-hand side of (C.27). We calculate:

$$
\begin{aligned}
{\xi_T^*}^{-1}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{A}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\mathbf{e}\right) &\leq {\xi_T^*}^{-1}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}'\mathbf{P}^*(\mathbf{w})\mathbf{e}\right) + {\xi_T^*}^{-1}\lambda_{\max}(\boldsymbol{\Sigma}^{-1})\mathrm{tr}\left(\boldsymbol{\mu}'\mathbf{P}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\mathbf{e}\right) \\
&\leq {\xi_T^*}^{-1}\lambda_{\max}\left(\boldsymbol{\Sigma}^{-1}\right)\mathrm{tr}\left(\boldsymbol{\mu}'\mathbf{P}^*(\mathbf{w})\mathbf{e}\right) \\
&\quad + {\xi_T^*}^{-1}\lambda_{\max}\left(\boldsymbol{\Sigma}^{-1}\right)\lambda_{\max}\left(\mathbf{P}^*(\mathbf{w})\right)\mathrm{tr}\left(\boldsymbol{\mu}'\mathbf{P}^*(\mathbf{w})\mathbf{e}\right) \\
&\leq C{\xi_T^*}^{-1}\mathrm{tr}\left(\boldsymbol{\mu}'\mathbf{P}^*(\mathbf{w})\mathbf{e}\right) \\
&= C{\xi_T^*}^{-1}\mathrm{tr}\left(\boldsymbol{\mu}'\overline{\mathbf{P}}\mathbf{P}^*(\mathbf{w})\mathbf{e}\right) \\
&= C{\xi_T^*}^{-1}\mathrm{vec}(\overline{\mathbf{P}}\boldsymbol{\mu}')'\mathrm{vec}(\mathbf{P}^*(\mathbf{w})\mathbf{e}) \\
&\leq C{\xi_T^*}^{-1}\left[\mathrm{vec}(\overline{\mathbf{P}}\boldsymbol{\mu}')'\mathrm{vec}(\overline{\mathbf{P}}\boldsymbol{\mu}')\right]^{1/2}\left[\mathrm{vec}(\mathbf{P}^*(\mathbf{w})\mathbf{e})'\mathrm{vec}(\mathbf{P}^*(\mathbf{w})\mathbf{e})\right]^{1/2} \\
&= C\underbrace{\left[\bar{p}{\xi_T^*}^{-2}\mathrm{tr}\left(\boldsymbol{\mu}\overline{\mathbf{P}}\boldsymbol{\mu}'\right)\right]^{1/2}}_{=o_p(1)\text{ by Assumption 2(}a\text{)}}\underbrace{\left[\bar{p}^{-1}\mathrm{tr}\left(\mathbf{e}'\mathbf{P}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\mathbf{e}\right)\right]^{1/2}}_{=O_p(1)} \\
&= o_p(1),
\end{aligned}
\tag{C.28}
$$

where the first and second inequalities follow from the trace inequality, and the fourth inequality follows from the Schwarz inequality. Using (C.15) and (C.16), the second part on the right-hand side of the third equality in (C.28) holds since:

$$
\begin{aligned}
\left(\bar{p}^{-1}\mathrm{tr}\left(\mathbf{e}'\mathbf{P}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\mathbf{e}\right)\right)^{1/2} &\leq \left(\bar{p}^{-1}\lambda_{\max}\left(\mathbf{P}^*(\mathbf{w})\right)\mathrm{tr}\left(\mathbf{e}'\mathbf{P}^*(\mathbf{w})\mathbf{e}\right)\right)^{1/2} \\
&= C\left(\bar{p}^{-1}\mathrm{tr}\left(\mathbf{e}'\mathbf{P}^*(\mathbf{w})\mathbf{e}\right)\right)^{1/2} \\
&\leq C\left(\bar{p}^{-1}\mathrm{tr}\left(\mathbf{e}'\overline{\mathbf{P}}\mathbf{e}\right)\right)^{1/2} \\
&\leq \left(\underbrace{T\lambda_{\max}\left(\left(\overline{\mathbf{Z}}'\overline{\mathbf{Z}}\right)^{-1}\right)}_{=O_p(1)\text{ by (C.15)}}\underbrace{T^{-1}\bar{p}^{-1}\mathrm{tr}\left(\mathbf{e}'\overline{\mathbf{Z}}\,\overline{\mathbf{Z}}'\mathbf{e}\right)}_{=O_p(1)\text{ by (C.16)}}\right)^{1/2} \\
&= O_p(1).
\end{aligned}
\tag{C.29}
$$

We next take the second term on the right-hand side of (C.27). Using (C.29) and Assumption 2(a), we have:

$$\begin{aligned}
&{\xi_T^*}^{-1}\left\{\operatorname{tr}\left(\mathbf{P}^*(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}^*(\mathbf{w})\right)-E\left[\operatorname{tr}\left(\mathbf{P}^*(\mathbf{w})\mathbf{e}\boldsymbol{\Sigma}^{-1}\mathbf{e}'\mathbf{P}^*(\mathbf{w})\right)\right]\right\}\\
&\quad\leq C{\xi_T^*}^{-1}\underbrace{\operatorname{tr}\left(\mathbf{e}'\mathbf{P}^*(\mathbf{w})\mathbf{P}^*(\mathbf{w})\mathbf{e}\right)}_{=O_p(\bar{p})\text{ by (C.29)}}+C{\xi_T^*}^{-1}O(\bar{p})\\
&\quad=O_p\left({\xi_T^*}^{-1}\bar{p}\right)+O\left({\xi_T^*}^{-1}\bar{p}\right)\\
&\quad=o_p(1),
\end{aligned}\tag{C.30}$$

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

C.4 Proof of Theorem 4

As discussed in (5.13)-(5.15) in the text, to prove (5.12), it suffices to show (5.14) and (5.15), where (5.14) is implied by (5.13). First take (5.13). Based on the following decomposition of $CV_{T,h}^*(\mathbf{w})$:

$$\begin{aligned}
CV_{T,h}^*(\mathbf{w})&\equiv CV_{T,h}(\mathbf{w})/(T-\bar{p}-h+1)\\
&=\widetilde{L}_{T,h}(\mathbf{w})+K\\
&\quad+\frac{2}{T-\bar{p}-h+1}\operatorname{vec}({\boldsymbol{\mu}_h}')'\left((\mathbf{I}_{T-\bar{p}-h+1}-\widetilde{\mathbf{P}}_h^*(\mathbf{w}))'\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\operatorname{vec}(\mathbf{e}_h')\\
&\quad-\frac{2}{T-\bar{p}-h+1}\operatorname{vec}(\mathbf{e}_h')'(\widetilde{\mathbf{P}}_h^*(\mathbf{w})'\otimes\mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\operatorname{vec}(\mathbf{e}_h')\\
&\quad+\frac{2}{T-\bar{p}-h+1}\operatorname{vec}((\boldsymbol{\mu}_h-\widetilde{\boldsymbol{\mu}}_h^*(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\operatorname{vec}(\mathbf{e}_h'),
\end{aligned}\tag{C.31}$$

to establish the first condition in (5.13), it is sufficient to show the following uniform convergence results:

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\operatorname{vec}(\boldsymbol{\mu}_h')'\left(\widetilde{\mathbf{P}}_h^*(\mathbf{w})'\otimes\mathbf{I}_K\right)\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\operatorname{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})=o_p(1),\tag{C.32}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\operatorname{vec}(\mathbf{e}_h')'(\widetilde{\mathbf{P}}_h^*(\mathbf{w})'\otimes\mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\operatorname{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})=o_p(1),\tag{C.33}$$

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\operatorname{vec}((\boldsymbol{\mu}_h-\widetilde{\boldsymbol{\mu}}_h^*(\mathbf{w}))')'\left(\mathbf{I}_{T-\bar{p}-h+1}\otimes\boldsymbol{\Sigma}_h^{-1}\right)\operatorname{vec}(\mathbf{e}_h')\right|/V_{T,h}(\mathbf{w})=o_p(1),\tag{C.34}$$

where we have replaced $\widetilde{V}_{T,h}(\mathbf{w})$ with $V_{T,h}(\mathbf{w})$ in the denominator of (C.32)-(C.34) 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 (C.39) below.

Under Assumption 3(a), (C.32) and (C.33) can be shown to hold by using similar arguments to those in (C.23), (C.24), and (C.25) under the conditions (C.15) and (C.16) 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 (C.34). 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:

$$\sup_{\mathbf{w}\in\mathcal{H}_T} \left|\operatorname{vec}((\boldsymbol{\mu}_h'\widetilde{\mathbf{P}}_h^*(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \operatorname{vec}(\mathbf{e}_h')\right| /V_{T,h}(\mathbf{w}) = o_p(1), \tag{C.35}$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}_T} \left|\operatorname{vec}((\mathbf{e}_h'\widetilde{\mathbf{P}}_h^*(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \operatorname{vec}(\mathbf{e}_h')\right| /V_{T,h}(\mathbf{w}) = o_p(1). \tag{C.36}$$

Take (C.35). 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 (C.35) as:

$$\begin{aligned}
&\left|\operatorname{vec}((\boldsymbol{\mu}_h'\widetilde{\mathbf{P}}_h^*(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \operatorname{vec}(\mathbf{e}_h')\right| /V_{T,h}(\mathbf{w}) \\
&\leq \left|\operatorname{vec}((\boldsymbol{\mu}_h'\mathbf{P}_h^*(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \operatorname{vec}(\mathbf{e}_h')\right| /V_{T,h}(\mathbf{w}) \\
&\quad + \left|\operatorname{vec}((\boldsymbol{\mu}_h'\mathbf{T}_h^*(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \operatorname{vec}(\mathbf{e}_h')\right| /V_{T,h}(\mathbf{w}) \\
&\quad + \left|\operatorname{vec}((\boldsymbol{\mu}_h'\mathbf{Q}_h^*(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \operatorname{vec}(\mathbf{e}_h')\right| /V_{T,h}(\mathbf{w}) \\
&\leq o_p(1) + C\xi_{T,h}^{*-1} q_h^* \operatorname{tr}(\boldsymbol{\mu}_h\mathbf{e}_h') \\
&= o_p(1) + C(\bar{p}\xi_{T,h}^{*-1})(\bar{p}^{-1}Tq_h^*)(T^{-1}\operatorname{tr}(\boldsymbol{\mu}_h\mathbf{e}_h')) \\
&= o_p(1),
\end{aligned} \tag{C.37}$$

where the second inequality in (C.37) follows from using the identical arguments to those in (C.23) 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 (C.36), once again using $\widetilde{\mathbf{P}}_h(p) = \mathbf{P}_h(p) + \mathbf{T}_h(p) - \mathbf{Q}_h(p)$, we can write:

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_T} \left| \mathrm{vec}((\mathbf{e}_h' \widetilde{\mathbf{P}}_h^*(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \mathrm{vec}(\mathbf{e}_h') \right| / V_{T,h}(\mathbf{w}) \\
&\quad \leq \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \mathrm{vec}((\mathbf{e}_h' \mathbf{P}_h^*(\mathbf{w})')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \mathrm{vec}(\mathbf{e}_h') \right| / V_{T,h}(\mathbf{w}) \\
&\qquad + \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \mathrm{vec}((\mathbf{e}_h' (\mathbf{T}_h^*(\mathbf{w}) - \mathbf{Q}_h^*(\mathbf{w}))')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \mathrm{vec}(\mathbf{e}_h') \right| / V_{T,h}(\mathbf{w}) \\
&\quad \leq o_p(1) + \xi_{T,h}^{*-1} \sup_{\mathbf{w}\in\mathcal{H}_T} \sum_{p=1}^{\bar{p}} w(p) \left| \mathrm{vec}((\mathbf{e}_h' (\mathbf{T}_h(p) - \mathbf{Q}_h(p))')' \left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \boldsymbol{\Sigma}_h^{-1}\right) \mathrm{vec}(\mathbf{e}_h') \right| \\
&\quad \leq o_p(1) + C\xi_{T,h}^{*-1} q_h^* \mathrm{tr}(\mathbf{e}_h \mathbf{e}_h') = o_p(1) + C(\bar{p}\xi_{T,h}^{*-1})(\bar{p}^{-1} T q_h^*)(T^{-1}\mathrm{tr}(\mathbf{e}_h \mathbf{e}_h')) \\
&= o_p(1), \qquad \text{(C.38)}
\end{aligned}
$$

where the second inequality follows from the arguments showing (C.24) 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 $\mathrm{tr}(\mathbf{e}_h \mathbf{e}_h') = O_p(T)$ 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:

$$
\sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{\widetilde{V}_{T,h}(\mathbf{w})}{V_{T,h}(\mathbf{w})} - 1 \right| \to 0 \qquad \text{(C.39)}
$$

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:

$$
\begin{aligned}
\widetilde{V}_{T,h}(\mathbf{w}) &= E(\widetilde{L}_{T,h}(\mathbf{w})) \\
&= \frac{1}{T-\bar{p}-h+1} E\left[\mathrm{tr}\left(\boldsymbol{\Sigma}_h^{-1}(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h^*(\mathbf{w}))'(\boldsymbol{\mu}_h - \widetilde{\boldsymbol{\mu}}_h^*(\mathbf{w}))\right)\right] \\
&= \frac{1}{T-\bar{p}-h+1} \mathrm{tr}\left(\widetilde{\mathbf{A}}_h^*(\mathbf{w})\boldsymbol{\mu}_h \boldsymbol{\Sigma}_h^{-1} \boldsymbol{\mu}_h' \widetilde{\mathbf{A}}_h^*(\mathbf{w})'\right) + E\left[\mathrm{tr}\left(\widetilde{\mathbf{P}}_h^*(\mathbf{w})\mathbf{e}_h \boldsymbol{\Sigma}_h^{-1} \mathbf{e}_h' \widetilde{\mathbf{P}}_h^*(\mathbf{w})'\right)\right].
\end{aligned}
\qquad \text{(C.40)}
$$

Based on (C.40), 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:

$$\operatorname{tr}\left(\widetilde{\mathbf{A}}_h(i)\boldsymbol{\mu}_h\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h'\widetilde{\mathbf{A}}_h(j)'\right) = \operatorname{tr}\left(\mathbf{A}_h(i)\boldsymbol{\mu}_h\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h'\mathbf{A}_h(j)'\right)(1+o(1)), \tag{C.41}$$

$$E\left[\operatorname{tr}\left(\widetilde{\mathbf{P}}_h(i)\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h(j)'\right)\right] = E\left[\operatorname{tr}\left(\mathbf{P}_h(i)\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\mathbf{P}_h(j)'\right)\right](1+o(1)), \tag{C.42}$$

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 (C.41). Next take (C.42). 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 (B.4) in Lemma 1, we have:

$$\begin{aligned}
E\left[\operatorname{tr}\left(\widetilde{\mathbf{P}}_h(i)\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h(j)'\right)\right] &= \operatorname{tr}\left(\widetilde{\mathbf{P}}_h(j)'(\widetilde{\mathbf{D}}_h(i)(\mathbf{P}_h(i) - \mathbf{I}_{T-\bar{p}-h+1}) + \mathbf{I}_{T-\bar{p}-h+1})E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) \\
&= \operatorname{tr}\left(\widetilde{\mathbf{P}}_h(j)'\widetilde{\mathbf{D}}_h(i)\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) - \operatorname{tr}\left(\widetilde{\mathbf{P}}_h(j)'\widetilde{\mathbf{D}}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) \\
&\quad + \operatorname{tr}\left(\widetilde{\mathbf{P}}_h(j)'E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) \\
&= \operatorname{tr}\left(\widetilde{\mathbf{P}}_h(j)'\widetilde{\mathbf{D}}_h(i)\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right)(1+o(1)) \\
&= \Big[\operatorname{tr}\left((\mathbf{P}_h(j) - \mathbf{I}_{T-\bar{p}-h+1})\widetilde{\mathbf{D}}_h(j)'\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) \\
&\quad + \operatorname{tr}\left(\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right)\Big](1+o(1)) \\
&= \operatorname{tr}\left(\mathbf{P}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\mathbf{P}_h(j)\right)(1+o(1)),
\end{aligned} \tag{C.43}$$

where the third equality follows from $\operatorname{tr}\left(\widetilde{\mathbf{P}}_h(j)'\widetilde{\mathbf{D}}_h(i)E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) = \operatorname{tr}\left(\widetilde{\mathbf{P}}_h(j)'E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right)(1+o(1))$ under Assumption 3(b) and from $\operatorname{tr}\left(\widetilde{\mathbf{P}}_h(j)'E(\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h')\right) = 0$ 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 (C.11)), and the last equality holds by Assumption 3(b) again. This establishes (C.42) and thus, combined with (C.41), yields (C.39).

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 (C.26)-(C.30) 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:

$$\widetilde{L}_{T,h}(\mathbf{w}) - \widetilde{V}_{T,h}(\mathbf{w}) = -\frac{2}{T-\bar{p}-h+1}\operatorname{tr}\left(\boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h'\widetilde{\mathbf{A}}_h^*(\mathbf{w})'\widetilde{\mathbf{P}}_h^*(\mathbf{w})\mathbf{e}_h\right) \\ + \frac{1}{T-\bar{p}-h+1}\left\{\operatorname{tr}\left(\widetilde{\mathbf{P}}_h^*(\mathbf{w})\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h^*(\mathbf{w})'\right) - E\left[\operatorname{tr}\left(\widetilde{\mathbf{P}}_h^*(\mathbf{w})\mathbf{e}_h\boldsymbol{\Sigma}_h^{-1}\mathbf{e}_h'\widetilde{\mathbf{P}}_h^*(\mathbf{w})'\right)\right]\right\}. \tag{C.44}$$

Using (C.39) and similar arguments to those for proving (C.41)-(C.42), 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 (5.13): $\sup_{\mathbf{w}\in\mathcal{H}_T} |\widetilde{L}_{T,h}(\mathbf{w})/\widetilde{V}_{T,h}(\mathbf{w}) - 1| \xrightarrow{p} 0$ as $T \to \infty$. Combining these above conditions implies

$$\sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{L}_{T,h}(\mathbf{w})}{L_{T,h}(\mathbf{w})} - 1\right| \leq \sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{L}_{T,h}(\mathbf{w})}{\widetilde{V}_{T,h}(\mathbf{w})}\right| \sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{\widetilde{V}_{T,h}(\mathbf{w})}{V_{T,h}(\mathbf{w})}\right| \sup_{\mathbf{w}\in\mathcal{H}_T}\left|\frac{V_{T,h}(\mathbf{w})}{L_{T,h}(\mathbf{w})}\right| - 1 \xrightarrow{p} 0 \tag{C.45}$$

as $T \to \infty$, establishing (5.15). Putting together (C.32)-(C.34), (C.39), and (C.45) completes the proof.

D Non-optimality of MMMA under serial correlation in the direct multi-step forecasting scheme

Seeing the asymptotic non-optimality of our MMMA for $h > 1$ can be done through its invalidity (in terms of asymptotic biasedness) under serial correlation. To begin with, we first recall that when $h > 1$, the serial correlation problem arises due to the fact that the h-step error $\boldsymbol{\epsilon}_{t+h}$ in (4.1) follows a moving average process of order $h-1$. Let $\boldsymbol{\Omega}_h = E(\operatorname{vec}(\mathbf{e}_h)\operatorname{vec}(\mathbf{e}_h)')$, where $\mathbf{e}_h = (\boldsymbol{\epsilon}_{\bar{p}+h}, \ldots, \boldsymbol{\epsilon}_T)'$. Analogous to the MMMA criterion $C_T(\mathbf{w})$ defined in (3.4) in the manuscript, we consider the following h-step version of the MMMA criterion:

$$C_{T,h}(\mathbf{w}) = (T-\bar{p}-h+1)\cdot\operatorname{tr}\left(\widetilde{\widetilde{\boldsymbol{\Sigma}}}_h(\bar{p})^{-1}\widehat{\boldsymbol{\Sigma}}_h^*(\mathbf{w})\right) + 2K^2\mathbf{p}'\mathbf{w}, \tag{3.4'}$$

where

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

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

Using similar decomposition arguments to those used in (C.1)-(C.4) in the online supplementary material (pages A5-A7), we define the h-step counterparts of $r_{2T}(\mathbf{w})$ (defined in (C.2) and (C.4)) and rewrite them respectively as:

$$r_{2Th}(\mathbf{w}) = -2\text{vec}(\mathbf{e}_h')'(\mathbf{P}_h^*(\mathbf{w})' \otimes \mathbf{I}_K)\left(\mathbf{I}_{T-\bar{p}-h+1} \otimes \widetilde{\widetilde{\boldsymbol{\Sigma}}}_h(\bar{p})^{-1}\right)\text{vec}(\mathbf{e}_h')$$

and

$$-2\text{vec}(\mathbf{e}_h'\mathbf{Z}_h(p))'\left((\mathbf{Z}_h(p)'\mathbf{Z}_h(p))^{-1} \otimes \mathbf{I}_K\right)\left(\mathbf{I}_{Kp} \otimes \widetilde{\widetilde{\boldsymbol{\Sigma}}}_h(\bar{p})^{-1}\right)(\mathbf{Z}_h(p)' \otimes \mathbf{I}_K)\,\text{vec}(\mathbf{e}_h'), \quad \text{(C.4')}$$

where $\mathbf{P}_h^*(\mathbf{w}) = \sum_{p=1}^{\bar{p}} w(p)\mathbf{P}_h(p)$ with $\mathbf{P}_h(p) = \mathbf{Z}_h(p)(\mathbf{Z}_h(p)'\mathbf{Z}_h(p))^{-1}\mathbf{Z}_h'(p)$, and $\mathbf{Z}_h(p)$ is defined in the text. To examine (C.4'), it is not difficult to show under serial correlation that the asymptotic variance of s_{Th}, defined as

$$\begin{aligned} s_{Th} &= (T-\bar{p}-h+1)^{-1/2}\ell(p)'\text{vec}\left(\boldsymbol{\Sigma}_h^{-1/2}\mathbf{e}_h'\mathbf{Z}_h(p)\boldsymbol{\Gamma}_{2h}(p)^{-1/2}\right) \\ &\equiv \ell(p)'\boldsymbol{\phi}_{Th}(p), \end{aligned} \quad \text{(D.1)}$$

is given by:

$$v_{Th}^2 = \text{Var}(s_{Th}) = \ell(p)'(\boldsymbol{\Gamma}_{2h}(p)^{-1/2} \otimes \boldsymbol{\Sigma}_h^{-1/2})\boldsymbol{\Lambda}_h(p)(\boldsymbol{\Gamma}_{2h}(p)^{-1/2} \otimes \boldsymbol{\Sigma}_h^{-1/2})\ell(p), \quad \text{(D.2)}$$

where $\ell(p)$ is 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, $\boldsymbol{\Gamma}_{2h}(p) = \text{plim}\left(\mathbf{Z}_h(p)'\mathbf{Z}_h(p)\right)/(T-\bar{p}-h+1)$, and:

$$\boldsymbol{\Lambda}_h(p) = \text{plim}\left[(T-\bar{p}-h+1)^{-1}\left(\mathbf{Z}_h(p)' \otimes \mathbf{I}_K\right)\boldsymbol{\Omega}_h\left(\mathbf{Z}_h(p) \otimes \mathbf{I}_K\right)\right]. \quad \text{(D.3)}$$

Equations (D.2) and (D.3) together imply that for multi-step forecasting, $\boldsymbol{\phi}_{Th}(p)$ does not converge in distribution to a K^2p-dimensional vector of multivariate *standard* normal random variables; i.e., $(\boldsymbol{\Gamma}_{2h}(p)^{-1/2} \otimes \boldsymbol{\Sigma}_h^{-1/2})\boldsymbol{\Lambda}_h(p)(\boldsymbol{\Gamma}_{2h}(p)^{-1/2} \otimes \boldsymbol{\Sigma}_h^{-1/2})$ in (D.2) is not equal to $\mathbf{I}_{k^2p}$ in the presence of serial correlation. As a consequence, the term (C.4') (ignoring the constant -2) does not take a quadratic form in multivariate standard normal random variables, meaning that in general, $E(r_{2Th}(\mathbf{w})) \neq -2K^2\mathbf{p}'\mathbf{w}$. This finding reveals the asymptotic non-unbiasedness and, hence, non-optimality of $C_{T,h}(\mathbf{w})$ under serial correlation, which invalidates the use of the $C_{T,h}(\mathbf{w})$ criterion for direct multi-step forecast averaging.

E Additional simulation results

E.1 Sensitivity analysis to the maximum lag order $\bar{p}$

Through the simulation experiments, this subsection examines the variability in the trace of the weighted MSFE as the value of $\bar{p}$ varies. Specifically, we report in Table A1 and Figure A1 (for $T = 100$) the sample variance of weighted MSFEs computed from the competing methods over the considered pre-specified maximum lag orders $\bar{p} = 3, 4, \ldots, 15$. The simulation results reveal that relative to other competing methods, our MMMA(I) and MCVA$_h$(D) methods are not very sensitive to the choice of $\bar{p}$ in most cases.

E.2 Estimation effects of $\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}$ ($\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}$) on the forecast performance of MMMA (MCVA$_h$)

As an anonymous referee points out, due to the use of the transformation matrices $\widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}$ and $\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}$ in our averaging criteria (3.4) and (4.5), respectively, another source of estimation error may be introduced into our forecast averaging methods. To empirically examine this issue, we compare the forecast accuracy of the MMMA and MCVA$_h$ approaches using $\mathbf{W} = \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}$ and $\boldsymbol{\Sigma}_h(\bar{p})^{-1}$ for weighting the associated sum of squared residual matrices, where these two transformation matrices correspond to the feasibly standardized and infeasibly standardized versions of our VAR forecast averaging approaches. Specifically, the MMMA and MCVA$_h$ criteria using the general matrix $\mathbf{W}$ for transformation are of the following form:

$$GC_T(\mathbf{w};\mathbf{W}) = (T-\bar{p})\cdot \operatorname{tr}\left(\mathbf{W}\widehat{\boldsymbol{\Sigma}}(\mathbf{w})\right) + 2K^2\mathbf{p}'\mathbf{w}, \tag{E.1}$$

$$GCV_{T,h}(\mathbf{w};\mathbf{W}) = (T-\bar{p}-h+1)\cdot \operatorname{tr}\left(\mathbf{W}\widetilde{\boldsymbol{\Sigma}}_h(\mathbf{w})\right). \tag{E.2}$$

For a fair comparison, we compare the forecast performance based on the sum of the actual h-step MSFEs:

$$\widehat{\mathrm{MSFE}}_h(\mathbf{W}) == \frac{1}{2500}\sum_{r=1}^{2500}\left[\operatorname{tr}\left(\left(\mathbf{y}_{T+h}^{(r)} - \widehat{\mathbf{y}}_{T+h|T}^{(r)}(\mathbf{W})\right)\left(\mathbf{y}_{T+h}^{(r)} - \widehat{\mathbf{y}}_{T+h|T}^{(r)}(\mathbf{W})\right)'\right)\right],$$

where $\widehat{\mathbf{y}}_{T+h|T}^{(r)}(\mathbf{W})$ is the combined h-step ahead forecast, with the combination weights computed by minimizing $GC_T(\mathbf{w};\mathbf{W})$ for iterative forecast averaging or $GCV_{T,h}(\mathbf{w};\mathbf{W})$ for direct forecast averaging, and the superscript "(r)" indicates the r-th simulation repetition.

We consider the error covariance matrix given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.00 & \rho \\ \rho & 4.00 \end{bmatrix},$$

where the covariance parameter ρ measures the degree of correlation between the two response variables in our DGP. We set $\rho = 0.8, 1.2,$ and 1.8.

Table A2 reports the forecast performance of the MMMA and MCVA$_h$ methods using the considered transformation matrices $\mathbf{W}$ for $\bar{p} = 5, 10, 15$, where we normalize $\widehat{\mathrm{MSFE}}_h(\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}) = 1$. It appears from Table A2 that the estimation error of $\widetilde{\boldsymbol{\Sigma}}_h(\bar{p})$ does not considerably impact the forecast performance of our MMMA and MACV$_h$ methods; specifically, the performance of MMMA using $\mathbf{W} = \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}$ is nearly identical to that of the infeasible MMMA using $\mathbf{W} = \boldsymbol{\Sigma}(\bar{p})^{-1}$ in almost all sample sizes, forecast horizons, covariance ρ's, and maximum lag orders considered. On the other hand, the performance of the feasible MCVA$_h$ using $\mathbf{W} = \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}$ is slightly worse than the infeasible MCVA$_h$ using $\mathbf{W} = \boldsymbol{\Sigma}_h^{-1}$ under the small sample size of $T = 100$. We also find that, as expected, the inferior performance of the feasible MCVA$_h$ using $\mathbf{W} = \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}$ relative to its infeasible version using $\mathbf{W} = \boldsymbol{\Sigma}_h^{-1}$ becomes prominent as the covariance ρ increases and that increasing the sample size usually leads to improvements in the feasible MCVA$_h$, with its performance nearly identical to that of the infeasible MCVA$_h$ in most cases when $T = 200$ and $T = 500$.

E.3 Two more DGPs

This section additionally considers two DGPs. The first DGP is directly from Lewis and Reinsel (1985), who consider the bivariate ARMA(1,1) model of the form:

$$\text{DGP A1:}\quad \mathbf{y}_t - \boldsymbol{\Phi}\mathbf{y}_{t-1} = \boldsymbol{\varepsilon}_t - \boldsymbol{\theta}\boldsymbol{\varepsilon}_{t-1}$$

with:

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

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

$$\text{DGP A2:}\quad \mathbf{y}_t - \sum_{i=1}^{5} \boldsymbol{\Phi}_i \mathbf{y}_{t-i} = \boldsymbol{\varepsilon}_t,$$

where the coefficient matrices are $\boldsymbol{\Phi}_1 = (a+b)\mathbf{I}_7 + c\mathbf{1}_7$, $\boldsymbol{\Phi}_2 = -(ab+d)\mathbf{I}_7 - (a+b)c\mathbf{1}_7$, $\boldsymbol{\Phi}_3 = (a+b)d\mathbf{I}_7 + (ab+d)c\mathbf{1}_7$, $\boldsymbol{\Phi}_4 = -abd\mathbf{I}_7 - (a+b)cd\mathbf{1}_7$, $\boldsymbol{\Phi}_5 = abcd\mathbf{1}_7$ with $(a,b,c,d) = (0.5, 0.3, 0.1, 0.3)$, $\boldsymbol{\Sigma}$ 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 of whether or not the true DGP is contained as one of the candidate models, on forecast accuracy of our averaging methods.

The maximum lag order for DGP A2 is set to $\bar{p} = 3, 4, \ldots, 8$ due to consideration of the degree of freedom. The simulation results under DGPs A1 and A2 for $T = 100$ and $T = 200$ (displayed in Figure A7) are summarized as follows.

For DGP A1 (i.e., bivariate ARMA(1,1)), the panels in the first two rows of Figure A7 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 estimation 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 A1 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 A2), the relative MSFEs are shown in the panels of the last two rows of Figure A7. 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 A1, 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 $h > 8$.

F Empirical illustration

A common interest among economists is analyzing the relationship among economic data series, which 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 forecast 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 employ the recursive forecast scheme (using expanding estimation windows) for forecast updates. 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 1959:Q2 to 2005:Q4. This procedure 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

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

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 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 $\widetilde{\mathbf{\Sigma}}_h(\bar{p})^{-1}$ based on (5.2) for the whole VAR system as

$$\widehat{\mathrm{MSFE}}_h^A(\bar{p}; M) = \frac{1}{t_1 - h - t_0 + 1} \sum_{t=t_0}^{t_1 - h} \mathrm{tr} \left(\widetilde{\mathbf{\Sigma}}_{ht}(\bar{p})^{-1} \left(\mathbf{y}_{t+h} - \widehat{\mathbf{y}}_{t+h|t}(\bar{p}; M) \right) \left(\mathbf{y}_{t+h} - \widehat{\mathbf{y}}_{t+h|t}(\bar{p}; M) \right)' \right) , \tag{F.2}$$

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

Figure A8 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,[1] all relative to OLS(I). The individual and aggregated MSFEs are computed from formulae (F.1) and (F.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 A8 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 A8), MMMA(I) is preferred to MCVA$_h$(D) and

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

Stein(I) in forecasting Y at most horizons and in forecasting FF at horizons $h \leq 7$; 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 A8), however, the performances of both MMMA(I) and MCVA$_h$(D) improve and are better than Stein(I) under many horizons, while the advantage of Stein(I) in forecasting P can be seen at horizon $h \geq 7$ when $\bar{p} = 10$. On the other hand, it can be seen from Figure A8 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 of predicting the P series when $\bar{p} = 5$, but on the contrary, BIC(I) has good performance in forecasting the P series when $\bar{p}$ is set to be modest to long, say, $\bar{p} \geq 10$. Moreover, HQ(I) is particularly good at forecasting Y at long horizons. EQ(I) performs quite well in forecasting P 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. 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 outweighed by its efficiency loss.

We next discuss the comparison between the proposed iterative MMMA(I) and direct MCVA$_h$(D) methods. First of all, MMMA(I) often tends to have smaller relative MSFEs than MCVA$_h$(D) in forecasting Y, particularly when $\bar{p} \leq 9$, while MCVA$_h$(D) appears to dominate MMMA(I) for the Y series when $\bar{p} > 9$.

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 2.0% ($h = 1$) to 7.1% ($h = 4$) when $\bar{p} = 5$, for instance. 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 $\bar{p} = 10$ is specified, the improvements of MCVA$_h$(D) in forecasting P are approximately 0.1% $\sim$ 6.0%. 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 ($h = 11, 12$) horizons. For example, MMMA(I) improves upon MCVA$_h$(D) by 9.0% $\sim$ 38.6% ($h = 2 \sim 10$) when $\bar{p} = 10$. If the attention is restricted to the aggregated MSFEs (i.e., the panels labeled A in Figure A8), then MMMA(I) and MCVA$_h$(D) are competitive to each other at short to modest horizons, say, $h \leq 4$ and $\bar{p} = 10$, while MCVA$_h$(D) tends to dominate MMMA(I) at longer horizons for $\bar{p} = 5$ and $\bar{p} = 10$.

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 normalized maximum regret for the VAR system (displayed at the bottom-right corner in Figure A8), MMMA(I) improves upon MCVA$_h$(D) for $h = 4 \sim 10$, and MCVA$_h$(D) tends to dominate MMMA(I) at short and long horizons.

Table A1: Relative variability in weighted MSFEs over different $\bar{p}$'s

h	MMMA(I)	MCVA$_h$(D)	OLS(D)	Stein(I)	OLS(I)	AIC(I)	BIC(I)	HQ(I)	SAIC(I)	SBIC(I)	EQ(I)
$\alpha = 0$					$T = 100$						
1	0.018	0.022	1.000	0.861	1.000	0.033	0.018	0.018	0.093	0.065	0.099
4	0.032	0.074	1.469	0.902	1.000	0.058	0.028	0.029	0.141	0.105	0.148
8	0.067	0.195	2.348	0.846	1.000	0.097	0.062	0.063	0.170	0.138	0.177
12	0.128	0.373	3.832	0.853	1.000	0.149	0.117	0.118	0.218	0.190	0.223
average	0.071	0.200	2.408	0.854	1.000	0.095	0.065	0.066	0.171	0.139	0.177
$\alpha = 2$											
1	0.113	0.091	1.000	0.902	1.000	0.427	0.108	0.109	0.194	0.169	0.196
4	0.088	0.136	1.594	0.892	1.000	0.369	0.081	0.082	0.164	0.136	0.165
8	0.075	0.217	3.049	0.856	1.000	0.378	0.066	0.066	0.131	0.106	0.130
12	0.112	0.325	4.008	0.868	1.000	0.316	0.100	0.101	0.195	0.168	0.195
average	0.090	0.212	2.657	0.863	1.000	0.348	0.081	0.083	0.160	0.133	0.160
$\alpha = 5$											
1	0.600	0.394	1.000	1.449	1.000	0.963	2.865	1.220	1.148	1.336	1.327
4	0.628	0.427	1.607	0.980	1.000	0.918	1.275	0.919	0.563	0.592	0.584
8	0.233	0.721	4.691	0.721	1.000	0.535	0.503	0.437	0.160	0.158	0.157
12	0.102	0.241	3.474	0.837	1.000	0.775	0.084	0.426	0.147	0.120	0.131
average	0.364	0.469	3.047	0.890	1.000	0.755	0.884	0.666	0.380	0.399	0.398
$\alpha = 10$											
1	0.659	0.548	1.000	2.471	1.000	0.976	5.424	0.943	2.832	3.633	4.004
4	1.684	0.461	1.093	1.354	1.000	1.407	2.858	1.904	1.313	1.460	1.442
8	1.543	1.355	4.446	0.701	1.000	1.091	1.887	1.394	0.465	0.478	0.473
12	0.109	0.219	3.121	0.804	1.000	0.732	0.240	0.589	0.117	0.092	0.094
average	1.167	0.688	2.532	1.124	1.000	1.087	2.282	1.326	0.943	1.072	1.095
$\alpha = 0$					$T = 200$						
1	0.044	0.050	1.000	0.897	1.000	0.048	0.040	0.043	0.154	0.131	0.160
4	0.162	0.180	1.177	0.937	1.000	0.169	0.166	0.164	0.242	0.223	0.247
8	0.122	0.175	1.940	0.911	1.000	0.124	0.127	0.125	0.190	0.174	0.194
12	0.116	0.252	3.282	0.836	1.000	0.116	0.106	0.108	0.193	0.180	0.197
average	0.129	0.173	1.701	0.904	1.000	0.129	0.127	0.127	0.205	0.189	0.210
$\alpha = 2$											
1	0.293	0.238	1.000	1.049	1.000	0.387	0.292	0.298	0.372	0.357	0.375
4	0.270	0.305	1.269	1.092	1.000	0.411	0.266	0.273	0.345	0.331	0.347
8	0.181	0.266	2.078	0.928	1.000	0.225	0.182	0.185	0.230	0.216	0.232
12	0.084	0.201	2.729	0.858	1.000	0.117	0.076	0.077	0.160	0.144	0.162
average	0.194	0.260	1.794	0.973	1.000	0.264	0.193	0.197	0.267	0.254	0.270
$\alpha = 5$											
1	0.991	0.646	1.000	3.131	1.000	0.941	7.362	1.994	3.070	3.391	3.353
4	0.935	0.351	0.806	1.395	1.000	1.044	2.115	1.064	0.894	0.934	0.922
8	2.307	0.811	4.136	1.123	1.000	1.653	1.397	1.343	0.917	0.871	0.864
12	0.116	0.201	2.223	0.834	1.000	0.708	0.102	0.393	0.149	0.138	0.145
average	1.095	0.431	1.776	1.468	1.000	1.068	2.345	1.161	1.038	1.090	1.076
$\alpha = 10$											
1	0.834	0.778	1.000	5.321	1.000	0.861	1.624	0.964	9.151	10.961	12.611
4	2.522	0.245	0.429	1.730	1.000	1.583	4.698	3.487	1.898	2.051	2.067
8	2.763	0.174	0.389	1.122	1.000	1.858	3.155	3.094	0.791	0.676	0.582
12	0.136	0.216	1.530	0.741	1.000	0.708	0.400	0.458	0.118	0.116	0.117
average	2.357	0.249	0.529	1.519	1.000	1.622	3.418	2.870	1.589	1.681	1.727

Table A1: Relative variability in weighted MSFEs over different $\bar{p}$'s (cont'd)

h	MMMA(I)	MCVA$_h$(D)	OLS(D)	Stein(I)	OLS(I)	AIC(I)	BIC(I)	HQ(I)	SAIC(I)	SBIC(I)	EQ(I)
$\alpha = 0$						$T = 500$					
1	0.438	0.450	1.000	0.999	1.000	0.478	0.433	0.437	0.565	0.556	0.567
4	0.728	0.775	0.985	0.922	1.000	0.746	0.710	0.714	0.720	0.721	0.720
8	0.839	0.843	1.259	0.978	1.000	0.813	0.821	0.818	0.832	0.829	0.832
12	0.389	0.530	2.098	0.845	1.000	0.343	0.354	0.354	0.450	0.442	0.452
average	0.590	0.615	1.204	0.947	1.000	0.592	0.584	0.586	0.633	0.628	0.634
$\alpha = 2$											
1	0.688	0.591	1.000	1.208	1.000	0.784	0.682	0.685	0.781	0.777	0.783
4	0.621	0.674	1.081	0.948	1.000	0.662	0.614	0.616	0.614	0.613	0.614
8	0.774	0.673	1.527	0.915	1.000	0.700	0.776	0.770	0.690	0.690	0.690
12	0.337	0.514	2.829	0.809	1.000	0.311	0.326	0.328	0.391	0.386	0.392
average	0.638	0.597	1.411	0.947	1.000	0.640	0.641	0.641	0.627	0.625	0.628
$\alpha = 5$											
1	0.975	0.870	1.000	2.876	1.000	0.905	5.640	1.376	2.894	3.020	3.010
4	1.068	0.425	0.609	1.164	1.000	1.106	1.532	1.025	0.754	0.763	0.758
8	1.293	0.934	1.908	0.812	1.000	1.165	1.408	1.088	0.746	0.747	0.745
12	0.193	0.330	1.722	0.683	1.000	0.477	0.178	0.332	0.237	0.231	0.236
average	1.119	0.535	1.113	1.244	1.000	1.120	1.990	1.048	1.013	1.033	1.029
$\alpha = 10$											
1	1.032	0.953	1.000	10.507	1.000	0.964	1.192	1.081	13.678	14.891	16.817
4	1.645	0.090	0.122	1.363	1.000	1.096	3.412	2.482	1.136	1.158	1.148
8	1.946	0.202	0.274	0.994	1.000	1.349	2.964	2.424	0.442	0.396	0.340
12	0.264	0.350	1.239	0.819	1.000	0.766	0.375	0.469	0.242	0.238	0.240
average	1.798	0.135	0.216	1.463	1.000	1.242	3.090	2.420	1.210	1.233	1.258

Notes: (1) The DGP is a drifting bivariate ARMA(1,10) with the parameter α measuring the degree of local misspecification; see Section 6 in the manuscript for details; (2) Entries are the sample variances of weighted MSFEs computed from a specific method using 13 pre-specified maximum lag orders: $\bar{p} = 3, \ldots, 15$ with OLS(I) normalized to unity. "average" refers to the averages of the sample variances over forecast horizons $h = 1, 2, \ldots, 12$; (3) "I" and "D" in parentheses refer to iterative and direct multi-step forecasts, respectively.

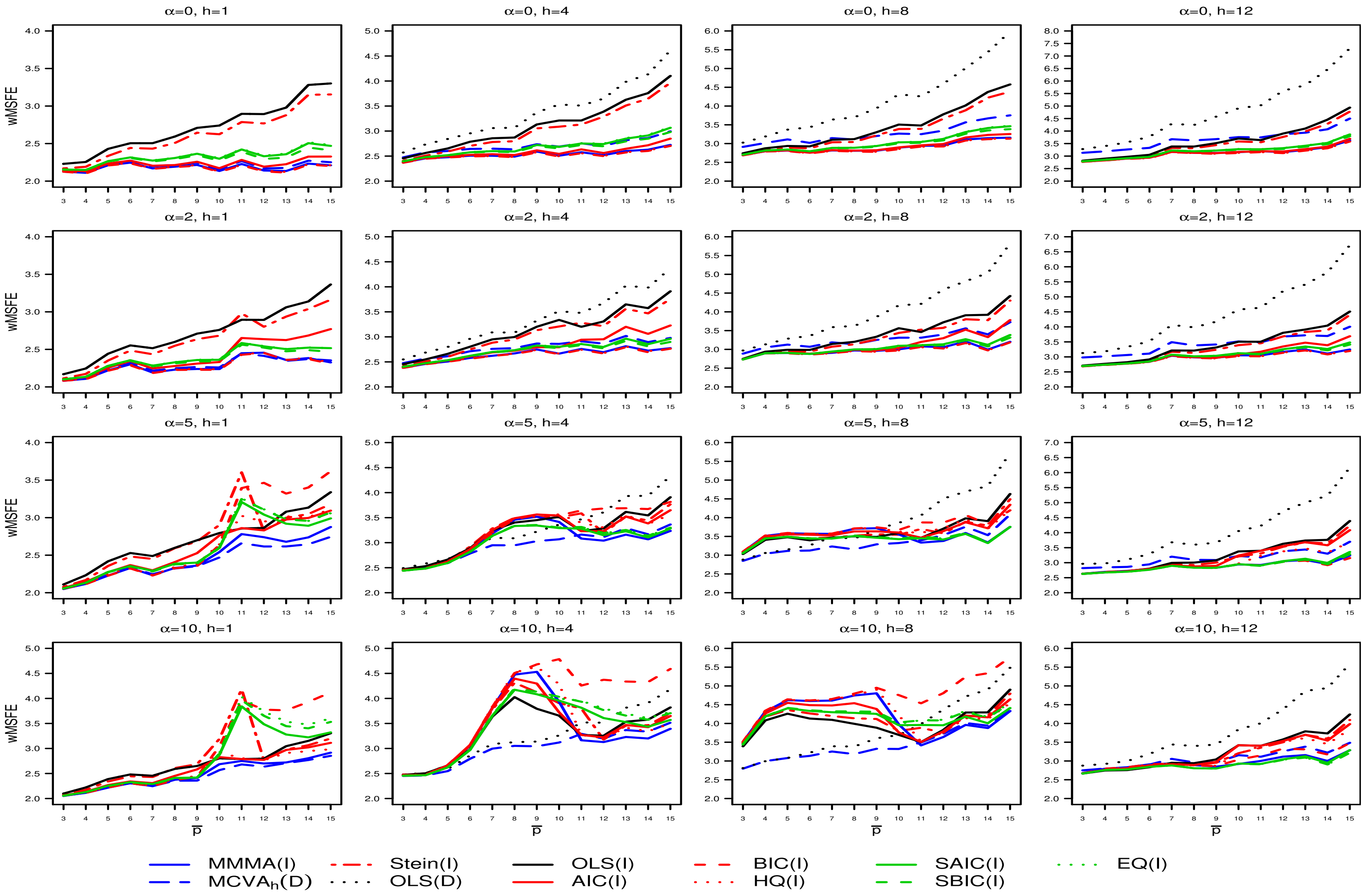


Figure A1: Sensitivity of weighted MSFEs of competing methods to the choice of $\bar{p}$ for $h = 1, 4, 8, 12$ ($T = 100$)

Table A2: Estimation effects on the relative multi-step forecast performance of MMMA and MCVA$_h$ ($\bar{p} = 5, 10, 15$)

	$\mathbf{w} = \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}$		$\mathbf{w} = \boldsymbol{\Sigma}_h^{-1}$					
	Reference		$T = 100$		$T = 200$		$T = 500$	
h	MCVA$_h$	MMMA	MCVA$_h$	MMMA	MCVA$_h$	MMMA	MCVAh	MMMA
$\rho = 0.8$				$\bar{p} = 5$				
1	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
8	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
12	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000
$\rho = 1.2$								
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	0.997	1.000	1.000	1.000	1.000	1.000
8	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
12	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
$\rho = 1.8$								
1	1.000	1.000	0.996	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	0.992	0.998	0.999	1.000	1.000	1.000
8	1.000	1.000	0.984	0.998	0.999	1.000	1.000	1.000
12	1.000	1.000	0.978	1.000	0.997	1.000	1.000	1.000
$\rho = 0.8$				$\bar{p} = 10$				
1	1.000	1.000	0.997	0.999	1.000	1.000	1.000	1.000
4	1.000	1.000	0.996	0.999	0.999	1.000	1.000	1.000
8	1.000	1.000	0.991	1.000	0.999	1.000	1.000	1.000
12	1.000	1.000	0.989	1.000	0.999	1.000	1.000	1.000
$\rho = 1.2$								
1	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	0.995	1.000	1.000	1.000	1.000	1.000
8	1.000	1.000	0.990	1.000	0.999	1.000	1.000	1.000
12	1.000	1.000	0.990	1.000	0.998	1.000	1.000	1.000
$\rho = 1.8$								
1	1.000	1.000	0.999	0.999	1.000	1.000	1.000	1.000
4	1.000	1.000	0.998	1.000	0.999	1.000	1.000	1.000
8	1.000	1.000	0.991	1.000	1.000	1.000	1.000	1.000
12	1.000	1.000	0.987	1.000	0.999	1.000	1.000	1.000
$\rho = 0.8$				$\bar{p} = 15$				
1	1.000	1.000	0.997	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	0.988	0.999	0.999	1.000	1.000	1.000
8	1.000	1.000	0.983	1.000	0.999	1.000	0.999	1.000
12	1.000	1.000	0.977	1.000	0.997	1.000	1.000	1.000
$\rho = 1.2$								
1	1.000	1.000	0.997	0.999	1.000	1.000	1.000	1.000
4	1.000	1.000	0.991	1.001	1.000	1.000	1.000	1.000
8	1.000	1.000	0.979	1.000	0.999	1.000	1.000	1.000
12	1.000	1.000	0.981	1.000	0.998	1.000	1.000	1.000
$\rho = 1.8$								
1	1.000	1.000	0.996	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	0.992	0.998	0.999	1.000	1.000	1.000
8	1.000	1.000	0.984	0.998	0.999	1.000	1.000	1.000
12	1.000	1.000	0.978	1.000	0.997	1.000	1.000	1.000

Note: Entries less than one indicate superior performance relative to the proposed MMMA and MCVA$_h$ using $\mathbf{W} = \widetilde{\boldsymbol{\Sigma}}(\bar{p})^{-1}$ and $\mathbf{W} = \widetilde{\boldsymbol{\Sigma}}_h(\bar{p})^{-1}$, respectively.

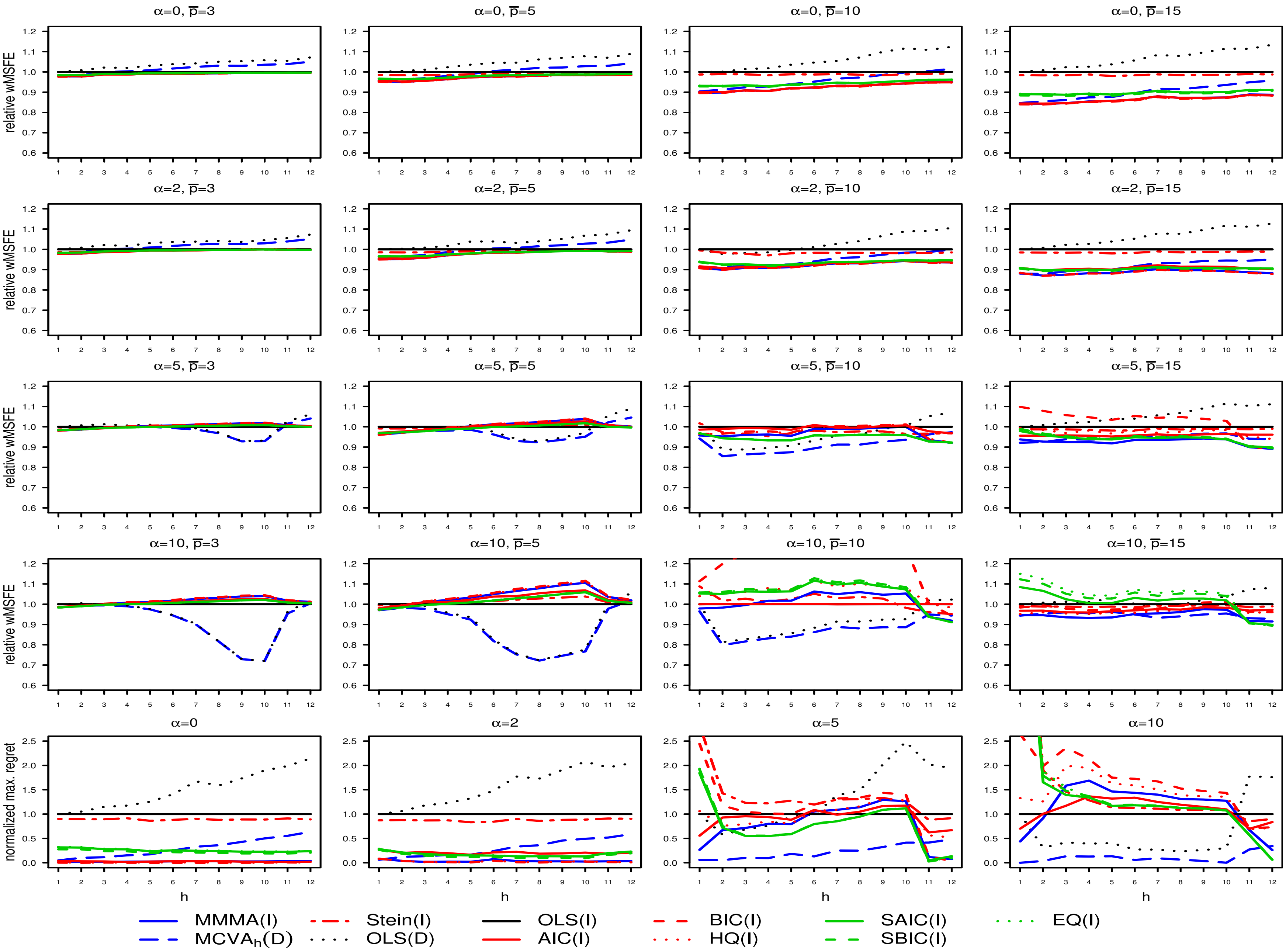


Figure A2: Multi-step forecast performance under bivariate drifting ARMA(1,10): $T = 200$

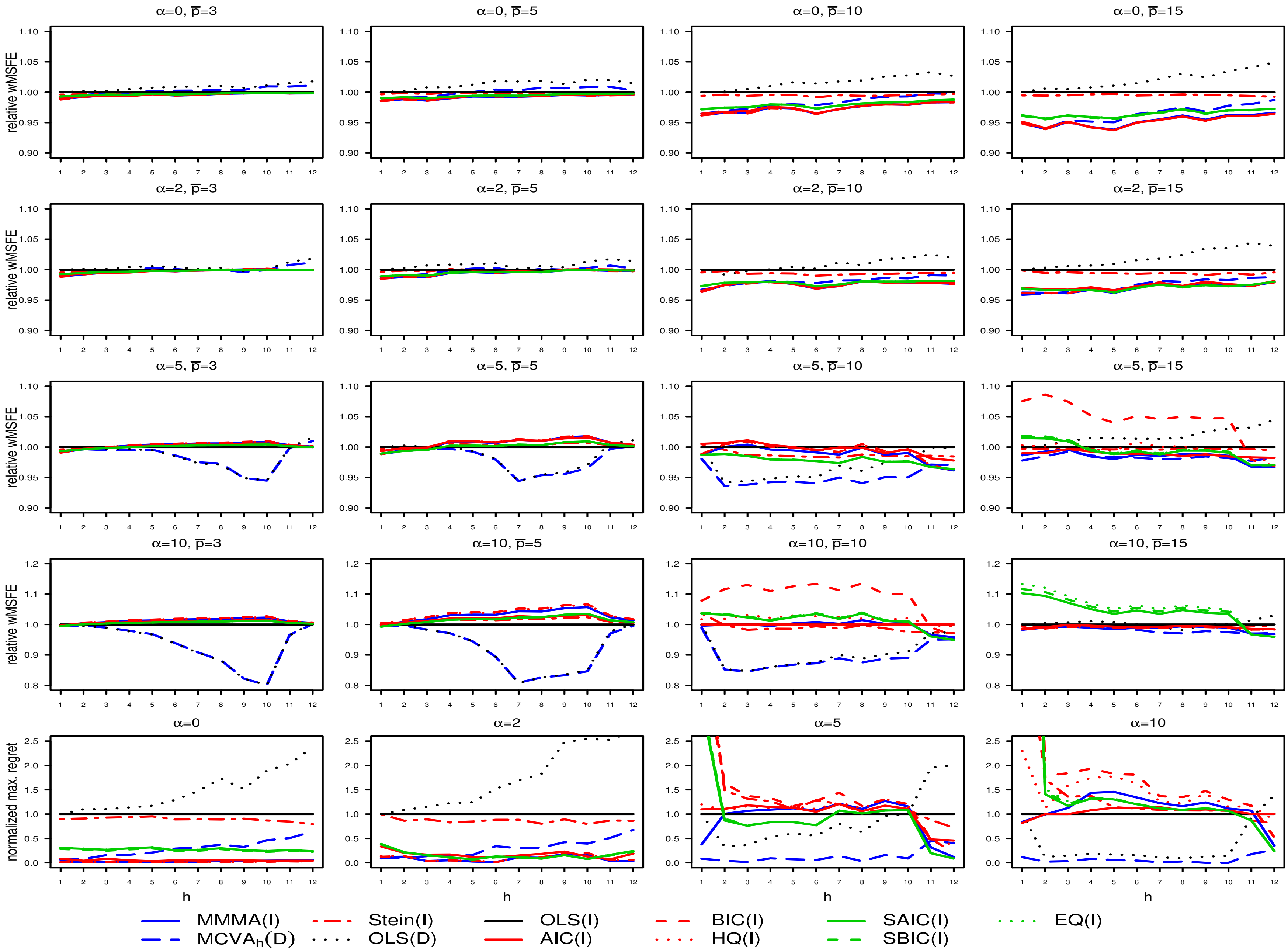
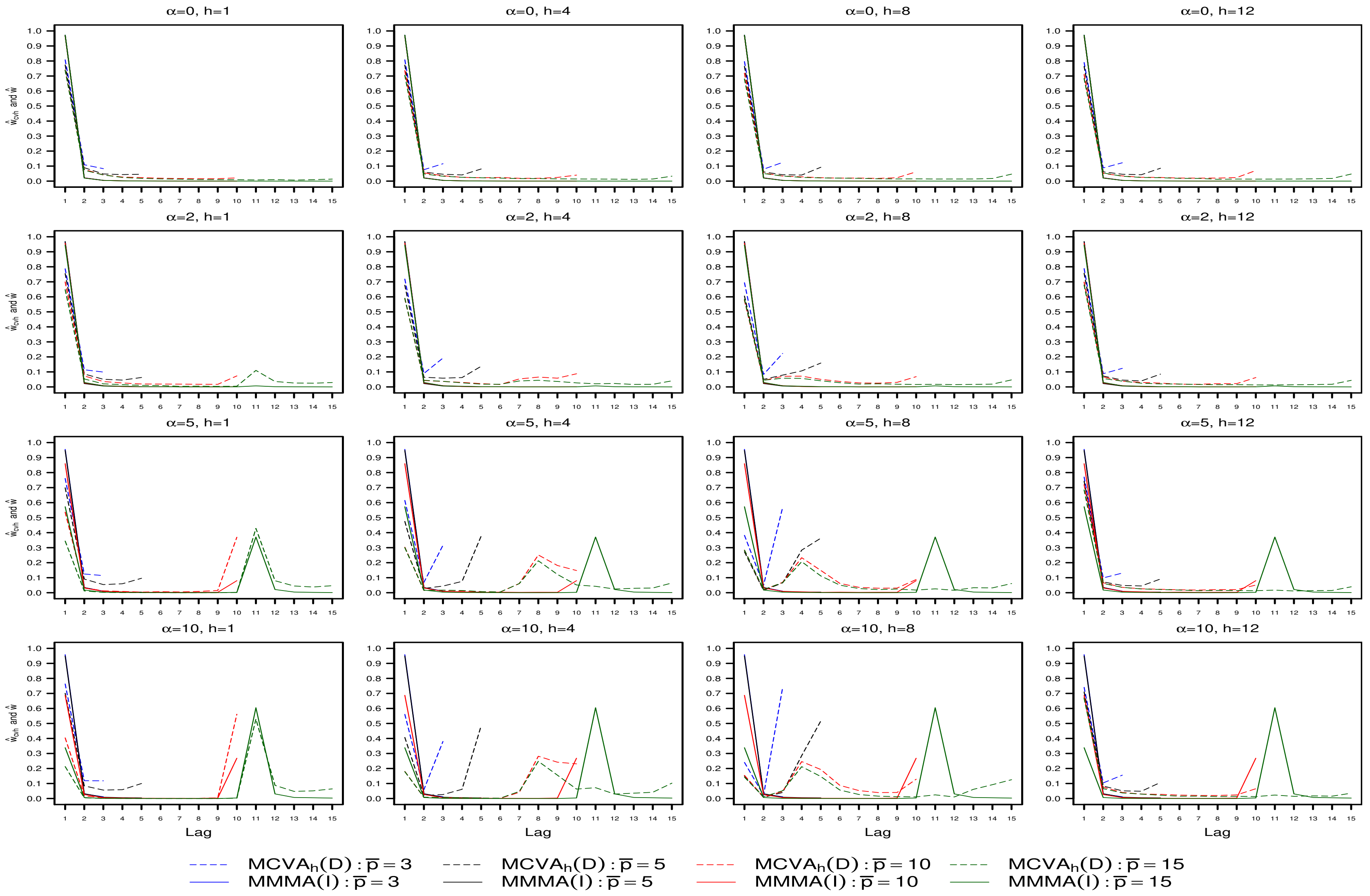
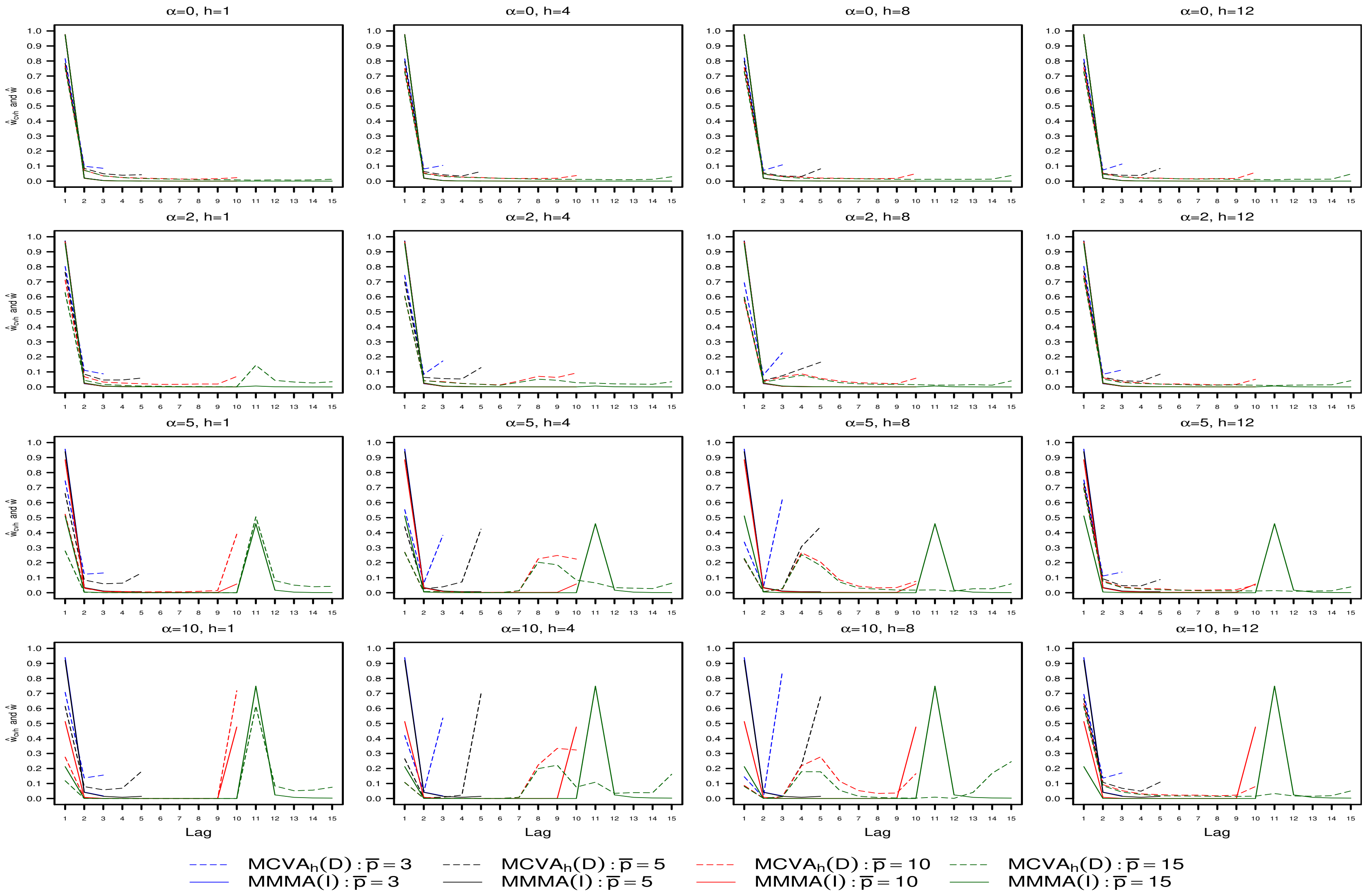


Figure A3: Multi-step forecast performance under bivariate drifting ARMA(1,10): $T = 500$



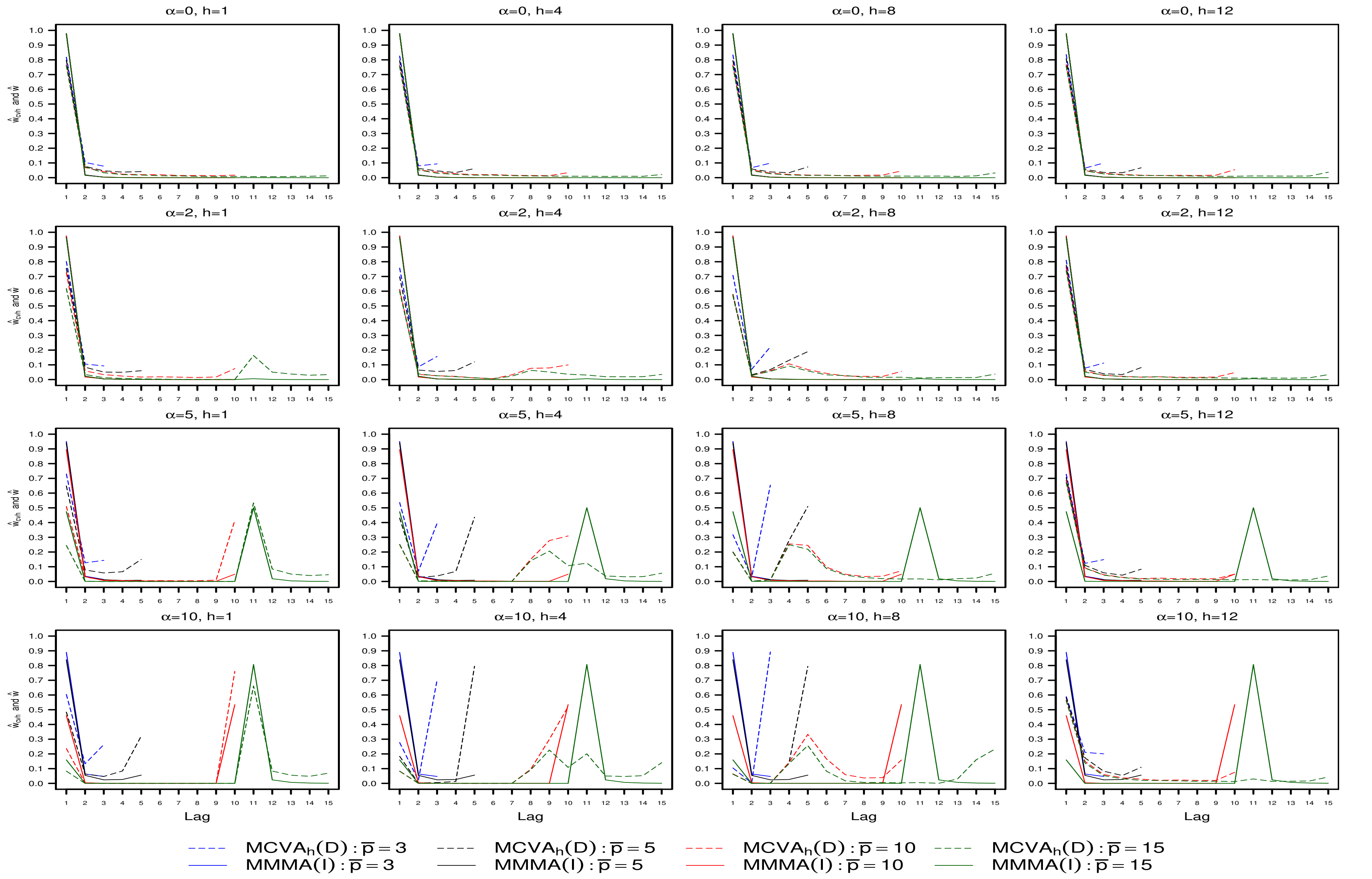
Note: For each of the values of α considered, MMMA(I) uses the same weights obtained from one-step-ahead forecast averaging across forecast horizons.

Figure A4: Weight estimates obtained from MCVA$_h$(D) and MMMA(I) based on $\bar{p} = 3, 5, 10, 15$ ($T = 100$)



Note: For each of the values of α considered, MMMA(I) uses the same weights obtained from one-step-ahead forecast averaging across forecast horizons.

Figure A5: Weight estimates obtained from MCVA$_h$(D) and MMMA(I) based on $\bar{p} = 3, 5, 10, 15$ ($T = 200$)



Note: For each of the values of α considered, MMMA(I) uses the same weights obtained from one-step-ahead forecast averaging across forecast horizons.

Figure A6: Weight estimates obtained from MCVA$_h$(D) and MMMA(I) based on $\bar{p} = 3, 5, 10, 15$ ($T = 500$)

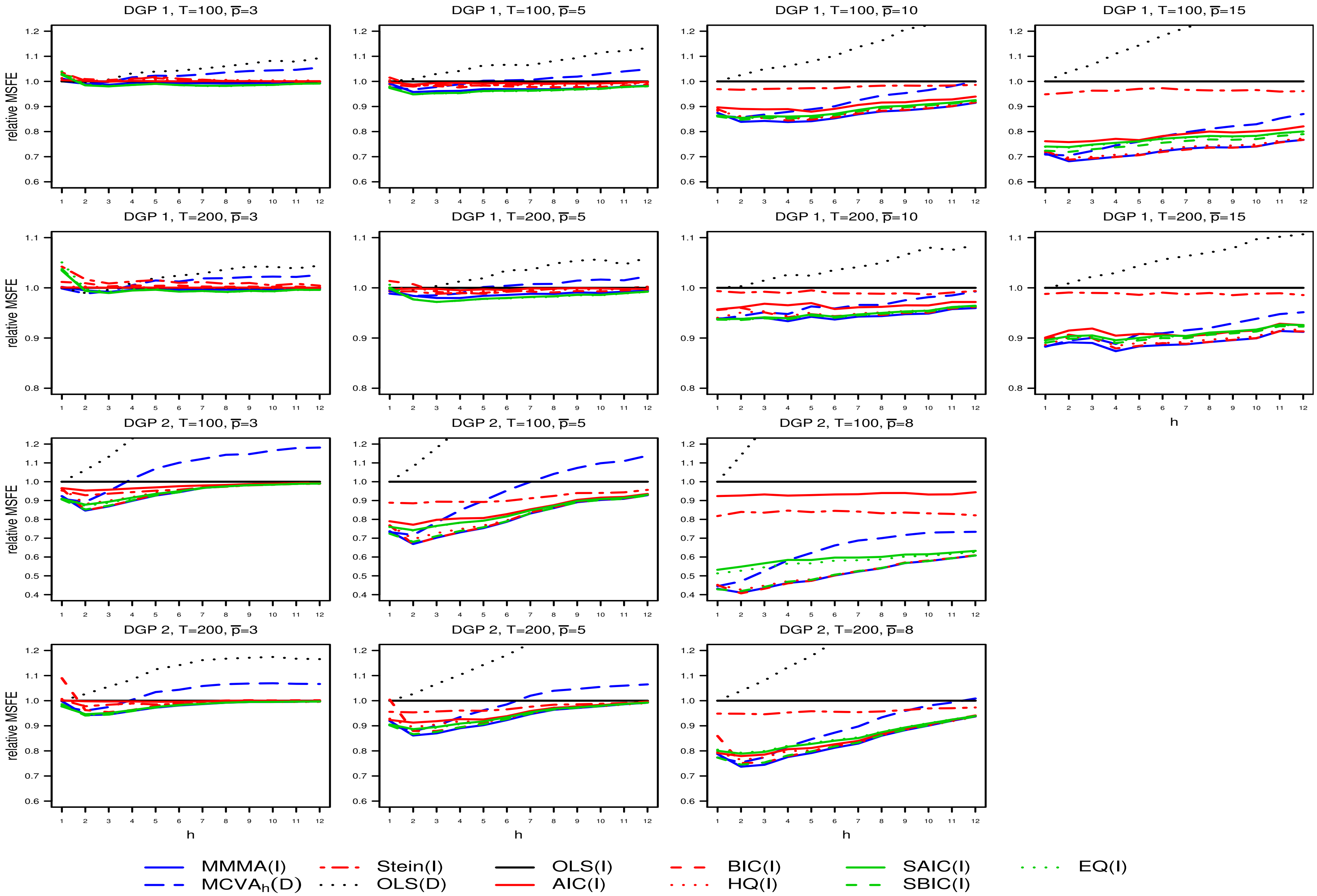
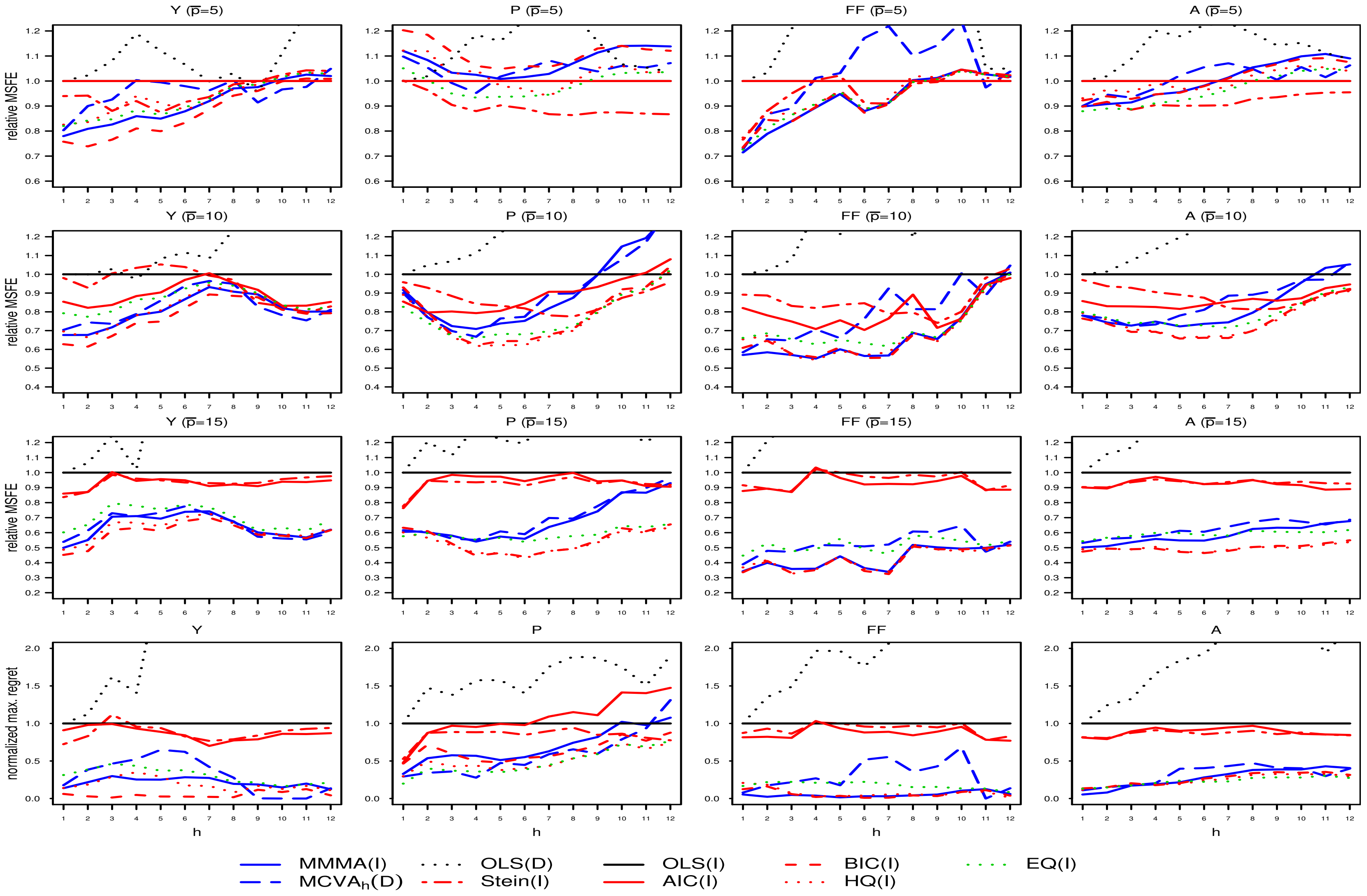


Figure A7: Multi-step forecast performance for DGPs A1-A2



Note: (1) The MSFEs for Y (GDP), P (the GDP deflator), and FF (the federal funds rate) are computed from (F.1), while the MSFEs for A are the aggregated weighted MSFEs computed from (F.2). 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 A8: 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 (with $\bar{p} = 5, 10, 15$)

References

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

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.

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.

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," *Annals of Statistics*, 15(3), 958–975.

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.

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

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," in *The Methodology and Practice of Econometrics: Festschrift in Honor of D.F. Hendry*, ed. by N. Shephard, and J. Castle, pp. 1–57. Oxford University Press.