Econometric Reviews

ISSN: 0747-4938 (Print) 1532-4168 (Online) Journal homepage: https://www.tandfonline.com/loi/lecr20

Focused information criterion for locally misspecified vector autoregressive models

Jan Lohmeyer, Franz Palm, Hanno Reuvers & Jean-Pierre Urbain

To cite this article: Jan Lohmeyer, Franz Palm, Hanno Reuvers & Jean-Pierre Urbain (2019) Focused information criterion for locally misspecified vector autoregressive models, Econometric Reviews, 38:7, 763-792, DOI: 10.1080/07474938.2017.1409410

To link to this article: https://doi.org/10.1080/07474938.2017.1409410

ECONOMETRIC REVIEWS
2019, VOL. 38, NO. 7, 763–792
https://doi.org/10.1080/07474938.2017.1409410

Taylor & Francis
Taylor & Francis Group

Focused information criterion for locally misspecified vector autoregressive models

Jan Lohmeyer, Franz Palm, Hanno Reuvers, and Jean-Pierre Urbain*

Department of Quantitative Economics, Maastricht University SBE, Maastricht, The Netherlands

ABSTRACT

This paper investigates the focused information criterion and plug-in average for vector autoregressive models with local-to-zero misspecification. These methods have the advantage of focusing on a quantity of interest rather than aiming at overall model fit. Any (sufficiently regular) function of the parameters can be used as a quantity of interest. We determine the asymptotic properties and elaborate on the role of the locally misspecified parameters. In particular, we show that the inability to consistently estimate locally misspecified parameters translates into suboptimal selection and averaging. We apply this framework to impulse response analysis. A Monte Carlo simulation study supports our claims.

KEYWORDS

Focused information criteria; frequentist model averaging; impulse responses; local misspecification; model selection; model uncertainty; vector autoregressive models

JEL CLASSIFICATION

C1; C51; C53; C54

1. Introduction

The motivation for this paper stems from Hansen (2005). The author considers a Gausssian ARMA(1,1) model approximated by AR(k) models with $k \in \{0, 1, \dots, k_{max}\}$ and is interested in the impulse responses. Table 1 of Hansen (2005) shows that the MSE-minimizing AR order depends strongly on parameter values and impulse response horizon. An extreme case is the specification $y_t = 0.5y_{t-1} + \epsilon_t - 0.9\epsilon_{t-1}$. The MSE-minimizing autoregressive orders equal 0 and 10 for the impulse responses at horizon 2 and 6, respectively. Ivanov and Kilian (2005) report a similar issue in a VAR setting. They simulate VAR processes similar to those often found in empirical work, and rank different model selection criteria (AIC, BIC, HQ, and serial correlation tests) based on the MSE of the estimated impulse responses implied by the selected model. A uniformly best criterion was not found. This might be expected since information criteria like AIC and BIC aim at global model fit and do not take into account the quantity of interest (such as the impulse response at a particular horizon). The Focused Information Criteria introduced by Claeskens and Hjort (2003) does take into account the interest of the researcher. Hansen (2005) acknowledged the opportunities for the FIC for the estimation of impulse responses when he remarked based on simulation outcomes: “The message from Tables 2 and 3 is that the FIC is an intriguing challenger to existing model selection methods and deserves attention and scrutiny.” A theoretical justification of these simulation results was not provided.

We develop a theoretical framework starting from a vector autoregression where part of the coefficients are local-to-zero, i.e., declining to zero at a rate of $T^{-1/2}$ with T denoting sample size. This case

CONTACT Franz Palm f.palm@maastrichtuniversity.nl Department of Quantitative Economics, Maastricht University SBE, P.O. Box 616, 6200 MD Maastricht, The Netherlands.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lecr.

Supplemental data for this article can be accessed on the publisher's website.

*To our deepest regret our friend and co-author Jean-Pierre Urbain passed away shortly after the submission of this paper.

is fundamentally different from a static setup because dynamic properties are also varying with sample size.[1] Building on ideas from Claeskens and Hjort (2003), Claeskens and Hjort (2008) and Liu (2015), we propose an estimator that can be used for both model selection and model averaging. This estimator is fairly general as it only requires the parameters of interest to be a sufficiently smooth transformation of the model's parameters. The results are subsequently applied to the specific case of impulse responses. A slight generalization of a theorem by Liu (2015) enables us to not only construct confidence intervals for a specific horizon, but to also construct confidence bands for multiple horizons. In addition, we provide an in depth discussion on the role of the local-to-zero parameters. These parameters cannot be estimated consistently, and we show that as a consequence the FIC and plug-in averages do not fully minimize the asymptotic mean squared error.

Our paper is related to the literature on model selection/averaging and the literature on impulse response analysis. We now discuss both. One of the earliest references on frequentist model averaging is the paper by Bates and Granger (1969). The literature on model averaging that is unrelated to forecasting is of a more recent origin. One literature branch on frequentist model averaging started with the paper by Hansen (2007). This paper shows that weight selection by minimization of Mallow's criterion will asymptotically lead to the lowest possible squared error among a class of estimators. His regression setup with homoskedastic errors was generalized to regression forecasts in Hansen (2008), and was modified by Hansen and Racine (2012) to allow for heteroskedastic errors. A time series application to stationary autoregressions of infinite order is Zhang et al. (2013). Zhang and Liu (2017) report results on the distribution of Mallow's and Jackknife-based model averaging weights in linear regressions with irrelevant variables.

The second branch of literature on frequentist model averaging evolves around locally misspecified models. The FIC was proposed by Claeskens and Hjort (2003) and extended by Claeskens and Hjort (2008). The underlying idea has been applied to various settings. We will report a nonexhaustive list. Liu (2015) considered the linear regression setup and derived asymptotically valid confidence intervals. Two additions to the treatment effects literature are Lu (2015) and Kitagawa and Muris (2016). DiTraglia (2016) provides results for generalized method of moments estimation. Liu and Kuo (2016) consider predictive regressions.

Finally, we briefly discuss the literature on impulse response (IR) analysis in autoregressive models. A comprehensive discussion on IR analysis can be found in Section 3.7 of Lütkepohl (2005). Both Lütkepohl (1990) and Benkwitz et al. (2000) have reported that the coverage of the impulse response confidence intervals can be low since the convergence rate of the estimators to their asymptotic distribution is nonconstant over the whole parameter space. Another important topic for impulse response analysis is the construction of joint confidence bands. A naive Cartesian product of the individual confidence intervals leads to severe undercoverage, whereas confidence bands based on the Bonferroni inequality have good coverage but are at the same time excessively wide. Often considered alternatives are bootstrap methods, e.g., Kilian (2001), Lütkepohl et al. (2015), and Bruder and Wolf (2017).

The remainder of this paper is organized as follows. Section 2 presents the model framework, the estimation procedure, and the asymptotic properties of: (1) the parameter estimates, (2) the feasible FIC, and (3) the elements of the weighting matrix. A discussion and illustration of the consequences of the inconsistent estimation of the local-to-zero parameter follows. Our theoretical findings are subsequently supported by various Monte Carlo simulations in Section 3. Section 4 concludes, and the mathematical proofs are presented in the Appendix.

In terms of notation, we follow Abadir and Magnus (2002) as closely as possible; in particular $\xrightarrow{d}$ and $\xrightarrow{p}$ signify convergence in distribution and convergence in probability, respectively. The stochastic and the strict stochastic order relations are denoted by $O_p(\cdot)$ and $o_p(\cdot)$. Vectors are printed in bold and

[1] Dynamic models under local-to-zero misspecification were discussed in Claeskens et al. (2007) and Rohan and Ramanathan (2011). Both papers first derive the asymptotic results in a setting without local misspecification and subsequently introduce the misspecification (see p. 363 of Claeskens et al. (2007) and Equation (8) on p. 221 of Rohan and Ramanathan (2011)). The theoretical implications of this two step procedure are not completely clear.

denote column vectors by default. $\mathbf{0}_j$ is the column vector of length j consisting of zeros only. We permit small deviations from Abadir and Magnus (2002) to keep our notation in line with the notation of Liu (2015).

2. Theory

2.1. Model framework

Let the K-dimensional multiple time series $\{\{\boldsymbol{y}_{T,t}\}_{t=-\infty}^{\infty}\}_{T=1}^{\infty}$ constitute a vector triangular array generated by the vector autoregressive (VAR) processes

$$\boldsymbol{y}_{T,t} = \boldsymbol{B}_1\boldsymbol{y}_{T,t-1} + \ldots + \boldsymbol{B}_{p_1}\boldsymbol{y}_{T,t-p_1} + \frac{\boldsymbol{\Delta}_1}{\sqrt{T}}\boldsymbol{y}_{T,t-p_1-1} + \ldots + \frac{\boldsymbol{\Delta}_{p_2}}{\sqrt{T}}\boldsymbol{y}_{T,t-p_1-p_2} + \boldsymbol{u}_t, \tag{2.1}$$

where $\boldsymbol{B}_i$ $(i \in \{1, 2, \ldots, p_1\})$ and $\boldsymbol{\Delta}_i$ $(i \in \{1, 2, \ldots, p_2\})$ are $(K \times K)$ coefficient matrices. Equation (2.1) differs in one important aspect from the usual VAR specifications, namely some of the coefficient matrices are premultiplied by $T^{-1/2}$ with T denoting sample size. This local-to-zero misspecification causes different dynamics for every T. Mathematically, this decay rate will prove to be crucial for the development of the asymptotic theory because it prevents the omitted variable bias from diverging with increasing sample size. Intuitively, we could think of this model specification as expressing a degree of uncertainty concerning the true lag order. The VAR process includes $p := p_1 + p_2$ lags for finite T, yet asymptotically a VAR(p_1) remains. This can be interpreted as exploring a shrinking neighborhood of the VAR(p_1) model.

2.2. Parameter estimation and asymptotics

To simplify notation, we collect all the parameters in the matrices $\boldsymbol{B} = [\boldsymbol{B}_1 \;\; \boldsymbol{B}_2 \;\; \cdots \;\; \boldsymbol{B}_{p_1}]$, $\boldsymbol{C}_T = [\boldsymbol{\Delta}_1 \;\; \boldsymbol{\Delta}_2 \;\; \cdots \;\; \boldsymbol{\Delta}_{p_2}]/\sqrt{T} = \boldsymbol{\Delta}/\sqrt{T}$, and define $\boldsymbol{\Theta}_T = [\boldsymbol{B} \; \boldsymbol{C}_T]$. Similarly to Lütkepohl (2005), we also stack the observations over time to obtain,

$$\begin{array}{rcll}
\boldsymbol{Y}_T & := & (\boldsymbol{y}_{T,1}, \boldsymbol{y}_{T,2}, \ldots, \boldsymbol{y}_{T,T}) & (K \times T), \\
\boldsymbol{z}_{T,t} & := & \begin{bmatrix} \boldsymbol{y}_{T,t} \\ \boldsymbol{y}_{T,t-1} \\ \vdots \\ \boldsymbol{y}_{T,t-p+1} \end{bmatrix} & (Kp \times 1), \\
\boldsymbol{Z}_T & := & (\boldsymbol{z}_{T,0}, \boldsymbol{z}_{T,1}, \ldots, \boldsymbol{z}_{T,T-1}) & (Kp \times T), \\
\boldsymbol{U} & := & (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_T) & (K \times T).
\end{array} \tag{2.2}$$

The model can now be expressed as $\boldsymbol{Y}_T = \boldsymbol{\Theta}_T\boldsymbol{Z}_T + \boldsymbol{U}$. A variety of approximating models can be considered but we will restrict our attention to models that use the same lag order in every equation (see Remark 1 for further details). Using the same lag order in the cross-section is common practice and will decrease the notational burden. Selection matrices are used to relate all estimators to the estimator using p lags. That is, for some integer m, such that $p_1 \leq m \leq p$,

$$\begin{array}{rcll}
\boldsymbol{L} & := & \boldsymbol{L}^{(1)} \otimes \boldsymbol{I}_K, \text{ with } \boldsymbol{L}^{(1)} = \begin{bmatrix} \boldsymbol{I}_{p_1} \\ \mathbf{O}_{p_2 \times p_1} \end{bmatrix} & (Kp \times Kp_1), \\
\boldsymbol{S}_0 & := & \boldsymbol{S}_0^{(1)} \otimes \boldsymbol{I}_K, \text{ with } \boldsymbol{S}_0^{(1)} = \begin{bmatrix} \mathbf{O}_{p_1 \times p_2} \\ \boldsymbol{I}_{p_2} \end{bmatrix} & (Kp \times Kp_2), \\
\boldsymbol{S}_m & := & \boldsymbol{S}_m^{(1)} \otimes \boldsymbol{I}_K, \text{ with } \boldsymbol{S}_m^{(1)} = \begin{bmatrix} \boldsymbol{I}_m \\ \mathbf{O}_{(p-m)\times m} \end{bmatrix} & (Kp \times Km), \\
\boldsymbol{\Pi}_m' & := & \boldsymbol{\Pi}_m'^{(1)} \otimes \boldsymbol{I}_K, \text{ with } \boldsymbol{\Pi}_m'^{(1)} = \begin{bmatrix} \boldsymbol{I}_{m-p_1} \\ \mathbf{O}_{(p-m)\times(m-p_1)} \end{bmatrix} & (Kp_2 \times K(m - p_1)).
\end{array} \tag{2.3}$$

The Kronecker products with $\boldsymbol{I}_K$ are a direct consequence of estimating all equations with the same lag order. The regressor matrix for the estimation of a VAR(m) model satisfies $\boldsymbol{Z}_{T,m} = \boldsymbol{S}'_m\boldsymbol{Z}_T$. The implied OLS estimator is the $(K \times Km)$ matrix $\hat{\boldsymbol{\Theta}}_{T,m}$ given by

$$\hat{\boldsymbol{\Theta}}_{T,m} = \boldsymbol{\Theta}_{T,m} + \boldsymbol{C}_T(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m)\boldsymbol{S}'_0\boldsymbol{Z}_T\boldsymbol{Z}'_{T,m}(\boldsymbol{Z}_{T,m}\boldsymbol{Z}'_{T,m})^{-1} + \boldsymbol{U}\boldsymbol{Z}'_{T,m}(\boldsymbol{Z}_{T,m}\boldsymbol{Z}'_{T,m})^{-1}. \tag{2.4}$$

Some rearranging and rescaling produces,

$$\sqrt{T}\left(\hat{\boldsymbol{\Theta}}_{T,m} - \boldsymbol{\Theta}_{T,m}\right) = \underbrace{\sqrt{T}\boldsymbol{C}_T}_{\boldsymbol{\Delta}}\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{S}'_0\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\left[\boldsymbol{S}'_m\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\right]^{-1}$$
$$+\left(\frac{1}{\sqrt{T}}\boldsymbol{U}\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\left[\boldsymbol{S}'_m\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\right]^{-1}, \tag{2.5}$$

and it can be seen that the $T^{1/2}$-consistency of the estimator precisely matches the decay rate of $T^{-1/2}$ in the elements of the parameter matrix $\boldsymbol{C}_T$. As a final step we apply the vec operator to transform the parameter matrices into a single parameter vector,

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,m} - \boldsymbol{\theta}_{T,m}\right) = \left(\left[\boldsymbol{S}'_m\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}'_m\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_0\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\otimes\boldsymbol{I}_K\right)\boldsymbol{\delta}$$
$$+\left(\left[\boldsymbol{S}'_m\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}'_m\otimes\boldsymbol{I}_K\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\text{vec}\left(\boldsymbol{u}_t\boldsymbol{z}'_{T,t-1}\right), \tag{2.6}$$

where $\hat{\boldsymbol{\theta}}_{T,m} = \text{vec}(\hat{\boldsymbol{\Theta}}_{T,m})$, $\boldsymbol{\theta}_{T,m} = \text{vec}(\boldsymbol{\Theta}_{T,m})$, and $\boldsymbol{\delta} = \text{vec}(\boldsymbol{\Delta})$. Equation (2.6) depends on: (1) various selection matrices, (2) the random matrix $\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1}$, and (3) the random vector $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\text{vec}\big(\boldsymbol{u}_t\boldsymbol{z}'_{T,t-1}\big)$. The latter two rescaled sums are typically encountered in laws of large numbers and central limit theorems, respectively. The following three assumptions guarantee that such theorems are applicable.

Assumption 1. *The sequence $\{\boldsymbol{u}_t\}$ of random K-vectors is an independent and identically distributed sequence with mean vector zero, a positive definite covariance matrix* $\text{E}(\boldsymbol{u}_t\boldsymbol{u}'_t) = \boldsymbol{\Sigma}$, *and there exists a* $c > 0$ *such that* $\text{E}|u_{it}u_{jt}u_{kt}u_{mt}| < c < \infty$ *for* $i, j, k, m = 1, 2, \ldots K$.

Assumption 2. $\det(\boldsymbol{B}_T(z)) = \det\left(\boldsymbol{I}_K z^p - \boldsymbol{B}_1 z^{p-1} - \ldots - \boldsymbol{B}_{p_1} z^{p_2} - \frac{\boldsymbol{\Delta}_1}{\sqrt{T}} z^{p_2-1} - \ldots - \frac{\boldsymbol{\Delta}_{p_2}}{\sqrt{T}}\right) \neq 0$ *for all* $|z| \geq 1$ *and* $\forall T \in \mathbb{N}$.

Assumption 3. $\det(\boldsymbol{B}_\infty(z)) = \det\left(\boldsymbol{I}_K z^{p_1} - \boldsymbol{B}_1 z^{p_1-1} - \ldots - \boldsymbol{B}_{p_1}\right) \neq 0$ *for all* $|z| \geq 1$.

Assumption 1 provides moment bounds and independence between the innovation $\boldsymbol{u}_t$ and its past. This latter property is exploited to apply limit theorems for martingale differences.[2] Assumptions 2 and 3 require the vector autoregressive process to be stationary for every finite T and also in the absence of local misspecification. The asymptotic properties of the OLS estimators are stated in Theorem 1.

[2] The requirement of i.i.d. innovations can be relaxed to the assumption that $\{\boldsymbol{u}_t\}$ is a martingale difference sequence. Formally, let $\mathcal{F}_t = \sigma(\boldsymbol{u}_s, -\infty < s \leq t)$ denote the sigma field generated by the innovations up to and including time t. Our results remain valid if the conditions $\text{E}(\boldsymbol{u}_t) = \boldsymbol{0}$ and $\text{E}(\boldsymbol{u}_t\boldsymbol{u}'_t) = \boldsymbol{\Sigma}$ are replaced by $\text{E}\left(\boldsymbol{u}_t|\mathcal{F}_{t-1}\right) = \boldsymbol{0}$ and $\text{E}\left(\boldsymbol{u}_t\boldsymbol{u}'_t|\mathcal{F}_{t-1}\right) = \boldsymbol{\Sigma}$, respectively.

Theorem 1 (Asymptotic Normality of the Least Squares Estimator). *Let Assumptions 1–3 hold. Then*

(a) *In the limit $T \to \infty$, we have for any $m \in \mathcal{M} = \{p_1, p_1 + 1, \dots, p\}$,*

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,m} - \boldsymbol{\theta}_{T,m}\right) \xrightarrow{d} \boldsymbol{A}_m \boldsymbol{\delta} + \left(\left[\boldsymbol{S}_m' \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \boldsymbol{S}_m' \otimes \boldsymbol{I}_K\right) \boldsymbol{R} \sim \mathrm{N}\left(\boldsymbol{A}_m \boldsymbol{\delta}, \left[\boldsymbol{S}_m' \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \otimes \boldsymbol{\Sigma}\right),$$

with $\boldsymbol{\Omega} = \mathrm{plim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_{T,t-1} \boldsymbol{z}_{T,t-1}'$, $\boldsymbol{A}_m = \left[\boldsymbol{S}_m' \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \boldsymbol{S}_m' \boldsymbol{\Omega} \boldsymbol{S}_0 \left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}_m' \boldsymbol{\Pi}_m\right) \otimes \boldsymbol{I}_K$ and $\boldsymbol{R} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}\right)$.

(b) *Let $\hat{\boldsymbol{u}}_{T,t}^{m}$ denote the OLS residuals from the estimation of a VAR(m). Consider $\hat{\boldsymbol{\Sigma}}_u^m = \frac{1}{T} \sum_{t=1}^{T} \hat{\boldsymbol{u}}_{T,t}^{m} \hat{\boldsymbol{u}}_{T,t}^{m\prime}$ as an estimator for $\boldsymbol{\Sigma}_u$. The result in part (a) can be strengthened to joint asymptotic normality with the covariance matrix estimator $\hat{\boldsymbol{\Sigma}}_u^m$,*

$$\begin{bmatrix} \sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,m} - \boldsymbol{\theta}_{T,m}\right) \\ \sqrt{T}\mathrm{vech}\left(\hat{\boldsymbol{\Sigma}}_u^m - \boldsymbol{\Sigma}\right) \end{bmatrix} \xrightarrow{d} \mathrm{N}\left(\begin{bmatrix} \boldsymbol{A}_m \\ \boldsymbol{O} \end{bmatrix} \boldsymbol{\delta}, \begin{bmatrix} \left[\boldsymbol{S}_m' \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \otimes \boldsymbol{\Sigma} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Xi}_{22} \end{bmatrix}\right).$$

The matrix $\boldsymbol{\Xi}_{22}$ is specified in the Appendix.

(c) *The estimator convergence as discussed in parts (a) and (b) of this theorem is also a joint convergence across different $m \in \mathcal{M}$. That is, for $\{i_1, i_2, \dots, i_M\} \in \mathcal{M}$, any $m \in \mathcal{M}$, and $i_1 < i_2 < \dots < i_M$, we have*

$$\begin{bmatrix} \sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,i_1} - \boldsymbol{\theta}_{i_1}\right) \\ \sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,i_2} - \boldsymbol{\theta}_{i_2}\right) \\ \vdots \\ \sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,i_M} - \boldsymbol{\theta}_{i_M}\right) \\ \sqrt{T}\mathrm{vech}\left(\hat{\boldsymbol{\Sigma}}_u^m - \boldsymbol{\Sigma}\right) \end{bmatrix} \xrightarrow{d} \mathrm{N}\left(\begin{bmatrix} \boldsymbol{A}_{i_1} \\ \boldsymbol{A}_{i_2} \\ \vdots \\ \boldsymbol{A}_{i_M} \\ \boldsymbol{O} \end{bmatrix} \boldsymbol{\delta}, \begin{bmatrix} \boldsymbol{V}_{i_1 i_1} & \boldsymbol{V}_{i_1 i_2} & \dots & \boldsymbol{V}_{i_1 i_M} & \boldsymbol{O} \\ \boldsymbol{V}_{i_2 i_1} & \boldsymbol{V}_{i_2 i_2} & \dots & \boldsymbol{V}_{i_2 i_M} & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{V}_{i_M i_1} & \boldsymbol{V}_{i_M i_2} & \dots & \boldsymbol{V}_{i_M i_M} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \dots & \boldsymbol{O} & \boldsymbol{\Xi}_{22} \end{bmatrix}\right).$$

The matrices $\boldsymbol{V}_{jk}$ are given by $\boldsymbol{V}_{jk} = \left[\boldsymbol{S}_j' \boldsymbol{\Omega} \boldsymbol{S}_j\right]^{-1} \boldsymbol{S}_j' \boldsymbol{\Omega} \boldsymbol{S}_k \left[\boldsymbol{S}_k' \boldsymbol{\Omega} \boldsymbol{S}_k\right]^{-1} \otimes \boldsymbol{\Sigma}$. It suffices to consider a single estimator for $\boldsymbol{\Sigma}$ because all the estimators are asymptotically equivalent.

The matrix $\boldsymbol{\Omega}$ deserves further attention. It is defined as the probability limit of the Gram matrix $\frac{1}{T}\boldsymbol{Z}_T \boldsymbol{Z}_T'$. The proof of Theorem 1 reveals that this probability limit equals $\mathrm{E}(\boldsymbol{z}_{\infty,t} \boldsymbol{z}_{\infty,t}')$ where $\boldsymbol{z}_{\infty,t}$ is defined as in Equation (2.2) but being generated by a VAR without local misspecification. We illustrate this remark with the AR process defined by $y_{T,t} = \alpha y_{T,t-1} + \frac{\delta_1}{\sqrt{T}} y_{T,t-2} + \frac{\delta_2}{\sqrt{T}} y_{T,t-3} + u_t$, that is an AR model with $p_1 = 1, p_2 = 2$ and $p = 3$: in that case $\boldsymbol{\Omega} = \frac{\sigma^2}{1-\alpha^2} \begin{bmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{bmatrix}$.

Remark 1. The consequences of the local misspecification framework are visible in Theorem 1. Standard asymptotics will fail if relevant parameters are left out since omitted variable bias will dominate asymptotically.[3] The local-to-zero rate of $T^{-1/2}$ balances this diverging behavior such that a finite asymptotic bias remains. This reasoning applies to all models that contain all the fixed parameters (i.e., the lag order should be no less than p_1) and leave out arbitrary parameters that are local-to-zero.

Remark 2. Assumption 2 is rather strict because it requires stationarity for all T in the natural numbers. Is it even possible for any parameter combination to satisfy this assumption? We can answer this question

[3] Let us consider the data generating process $y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + u_t$. Suppose that we estimate an AR(1) model. The OLS parameter estimator of the first lag coefficient, say $\hat{\alpha}$, satisfies $\sqrt{T}(\hat{\alpha} - \alpha_1) = \sqrt{T}\alpha_2 \frac{\frac{1}{T}\sum_{t=1}^{T} y_{t-1} y_{t-2}}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2} + \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_{t-1} u_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2}$. The first term on the RHS diverges at large T for $\alpha_2 \neq 0$. The divergence rate is $\sqrt{T}$.

in the affirmative for the univariate case but the result does not generalize easily to the multivariate case. For the univariate case, we define the lag polynomial $\beta_T(L)$ by

$$\beta_T(L)y_{T,t} = \left(1 - \beta_1 L - \ldots - \beta_{p_1} L^{p_1} - \frac{\delta_1}{\sqrt{T}} L^{p_1+1} - \ldots - \frac{\delta_{p_2}}{\sqrt{T}} L^p\right) y_{T,t} = u_t. \tag{2.7}$$

Fujiwara (1916) has shown that the largest modulus root of a polynomial $a(z) = a_0 z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n$ is bounded above by $2 \max \left\{ |a_1/a_0|, |a_2/a_0|^{1/2}, \ldots, |a_n/a_0|^{1/n} \right\}$. The largest modulus root of the lag polynomial β_T is thus bounded by

$$2 \max \left\{ |\beta_1|, |\beta_2|^{\frac{1}{2}}, \ldots, |\beta_{p_1}|^{\frac{1}{p_1}}, |\frac{\delta_1}{\sqrt{T}}|^{\frac{1}{p_1+1}}, \ldots, |\frac{\delta_{p_2}}{\sqrt{T}}|^{\frac{1}{p}} \right\}. \tag{2.8}$$

We deduce from Equation (2.8) that $2 \max \left\{ |\beta_1|, |\beta_2|^{\frac{1}{2}}, \ldots, |\beta_{p_1}|^{\frac{1}{p_1}}, |\delta_1|^{\frac{1}{p_1+1}}, \ldots, |\delta_{p_2}|^{\frac{1}{p}} \right\} < 1$ guarantees stationarity for all T.[4]

2.3. Quantities of interest

The focused information criterion (FIC) introduced by Claeskens and Hjort (2003) focusses on a quantity of interest rather than general model fit. Quantities of interest could be a single parameter, several parameters, or parameter transformations. Natural quantities of interest in the current dynamical setting are the impulse responses. In general, let $\boldsymbol{\mu} : \mathbb{R}^{K^2 p + K(K+1)/2} \to \mathbb{R}^l$ define the mapping from the model parameters to the l-dimensional focus quantity. The first $K^2 p$ arguments of the function $\boldsymbol{\mu}$ are reserved for the conditional mean parameters, whereas the last $K(K+1)/2$ arguments refer to the parameters in $\boldsymbol{\Sigma}$. As such we define $\boldsymbol{\sigma} = \text{vech}(\boldsymbol{\Sigma})$ and $\widehat{\boldsymbol{\sigma}} = \text{vech}(\widehat{\boldsymbol{\Sigma}})$, and write $\mu(\boldsymbol{\theta}, \boldsymbol{\sigma})$.[5] We additionally assume that evaluating the quantity of interest at $\boldsymbol{\mu}((\boldsymbol{\theta}_{T,m}, \mathbf{0}_{K^2(p-m)}), \boldsymbol{\sigma})$ provides an estimate for the quantity of interest in the model with m lags. The Auxiliary Result in the Appendix shows that this is true for the impulse responses. The next theorem follows from Theorem 1 and the multivariate first-order delta method.

Theorem 2 (Asymptotic Normality of the Quantities of Interest). *Let* $\boldsymbol{\mu} : \mathbb{R}^{K^2 p + K(K+1)/2} \to \mathbb{R}^l$ *have a continuous first derivative at all points* $(\boldsymbol{\theta}_m, \mathbf{0}_{K^2(p-m)}, \boldsymbol{\sigma})$, *with* $m \in \mathcal{M}$. *Let* $\boldsymbol{\theta}_\infty$ *denote the parameters obtained by taking* $\boldsymbol{\theta}_{T,p}$ *but setting* $\boldsymbol{\Delta}_1 = \boldsymbol{\Delta}_2 = \ldots = \boldsymbol{\Delta}_{p_2} = \mathbf{O}$, *and define the Jacobian matrices* $\boldsymbol{D}_\theta = \partial \boldsymbol{\mu}(\boldsymbol{\theta}_\infty, \boldsymbol{\sigma})/\partial \boldsymbol{\theta}'$ *and* $\boldsymbol{D}_\sigma = \partial \boldsymbol{\mu}(\boldsymbol{\theta}_\infty, \boldsymbol{\sigma})/\partial \boldsymbol{\sigma}'$. *For* $\boldsymbol{D}_\theta$ *and* $\boldsymbol{D}_\sigma$ *not having zero rows, under Assumptions 1–3, and as* $T \to \infty$,

$$\sqrt{T}\Big(\boldsymbol{\mu}((\hat{\boldsymbol{\theta}}_{T,m}, \mathbf{0}_{K^2(p-m)}), \hat{\boldsymbol{\sigma}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_{T,p}, \boldsymbol{\sigma})\Big) \xrightarrow{d} \mathrm{N}\Big(\boldsymbol{D}_\theta \boldsymbol{C}_m \boldsymbol{\delta}, \boldsymbol{D}_\theta \boldsymbol{P}_m (\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}) \boldsymbol{P}_m \boldsymbol{D}_\theta' + \boldsymbol{D}_\sigma \boldsymbol{\Xi}_{22} \boldsymbol{D}_\sigma'\Big),$$

with $\boldsymbol{P}_m = \boldsymbol{S}_m \left[\boldsymbol{S}_m' \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \boldsymbol{S}_m' \otimes \boldsymbol{I}_K$ *and*

$$\boldsymbol{C}_m = \left(\boldsymbol{S}_m \left[\boldsymbol{S}_m' \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \boldsymbol{S}_m' \boldsymbol{\Omega} - \boldsymbol{I}_{Kp}\right) \boldsymbol{S}_0 \left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}_m' \boldsymbol{\Pi}_m\right) \otimes \boldsymbol{I}_K.$$

We define the impulse response at horizon h as the hth coefficient matrix of the MA(∞) representation $\boldsymbol{y}_t = \sum_{i=0}^{\infty} \boldsymbol{\Phi}_i \boldsymbol{u}_{t-i}$ with $\boldsymbol{\Phi}_0 = \boldsymbol{I}_K$, hence $\boldsymbol{\mu} : \mathbb{R}^{K^2 p + K(K+1)/2} \to \mathbb{R}^{K^2}$. Theorem 2 can be applied if the Jacobian matrices $\boldsymbol{D}_\theta$ and $\boldsymbol{D}_\sigma$ are known. Lütkepohl (1990) lists these Jacobian matrices for the impulse responses, the orthogonalized impulse responses, the accumulated responses, the total accumulated

[4] This condition is a sufficient but by no means a necessary condition. For $p_1 = 1$ and $p_2 = 1$, the model $y_{T,t} = 0.7 y_{T,t-1} + \frac{0.75}{\sqrt{T}} y_{T,t-2} + u_t$ is stationary for all T but the parameters violate the requirement based on Fujiwara's bound.

[5] Theorem 1 showed that all the $\widehat{\boldsymbol{\Sigma}}_u^m$ are asymptotically equivalent. We omit the superscript m from now on.

responses, and the forecast error variance decomposition. The specific case of the (orthogonalized) impulse responses is highlighted in the following Corollary.

Corollary (An Application to Impulse Responses). *Let $\boldsymbol{A}_\infty$ denote the $(Kp \times Kp)$ companion matrix related to the process in which the misspecification coefficients have been set to zero, $\boldsymbol{J} = [\boldsymbol{I}_K\ \mathbf{O}\ \cdots\ \mathbf{O}]$ a matrix of dimensions $(K \times Kp)$ and $\boldsymbol{\Sigma}_u = \boldsymbol{P}\boldsymbol{P}'$. Then, under the assumptions of Theorem 2,*

(a) *The asymptotic distribution of the estimated impulse response at horizon i, $\hat{\boldsymbol{\Phi}}_i$, follows*

$$\sqrt{T}\left(\mathrm{vec}(\hat{\boldsymbol{\Phi}}_i) - \mathrm{vec}(\boldsymbol{\Phi}_i)\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{G}_i\boldsymbol{C}_m\boldsymbol{\delta}, \boldsymbol{G}_i\boldsymbol{P}_m\left(\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma}\right)\boldsymbol{P}_m\boldsymbol{G}_i'\right),$$

where $\boldsymbol{G}_i = \partial\mathrm{vec}\left(\boldsymbol{\Phi}_i\right)/\partial\boldsymbol{\theta} = \sum_{j=0}^{i-1}\boldsymbol{J}\left(\boldsymbol{A}_\infty\right)^{i-1-j}\otimes\boldsymbol{\Phi}_j$.

(b) *The asymptotic distribution of the estimated orthogonalized impulse response at horizon i, $\hat{\boldsymbol{\Psi}}_i$, follows*

$$\sqrt{T}\left(\mathrm{vec}(\hat{\boldsymbol{\Psi}}_i) - \mathrm{vec}(\boldsymbol{\Psi}_i)\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{F}_i\boldsymbol{C}_m\boldsymbol{\delta}, \boldsymbol{F}_i\boldsymbol{P}_m\left(\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma}\right)\boldsymbol{P}_m\boldsymbol{F}_i' + \bar{\boldsymbol{F}}_i\boldsymbol{\Xi}_{22}\bar{\boldsymbol{F}}_i'\right),$$

where $\boldsymbol{F}_0 = \mathbf{O}$ and $\boldsymbol{F}_i = \left(\boldsymbol{P}'\otimes\boldsymbol{I}_K\right)\boldsymbol{G}_i$ for $i > 0$. For all i we have $\bar{\boldsymbol{F}}_i = \left(\boldsymbol{I}_K\otimes\boldsymbol{\Phi}_i\right)\boldsymbol{H}$ with $\boldsymbol{H} = \partial\mathrm{vec}(\boldsymbol{P})/\partial\boldsymbol{\sigma}' = \boldsymbol{L}_K'[\boldsymbol{L}_K\left(\boldsymbol{I}_{K^2}+\boldsymbol{K}_{KK}\right)\left(\boldsymbol{P}\otimes\boldsymbol{I}_K\right)\boldsymbol{L}_K']^{-1}$ (see Lütkepohl (1990) for the definitions of $\boldsymbol{L}_K$ and $\boldsymbol{K}_{KK}$).

Remark 3. The first-order delta method is invalid if either $\boldsymbol{D}_\theta$ or $\boldsymbol{D}_\sigma$ has zero rows. It is well-documented in the literature that $\boldsymbol{D}_\theta$ can have zero rows for specific parameter combinations when impulse responses are considered. We refer to Lütkepohl (1990) and Benkwitz et al. (2000) for details.

2.3.1. Model selection: The focused information criterion (FIC)

The intuition behind the FIC of Claeskens and Hjort (2003) is most easily understood for a univariate quantity of interest, so we temporarily assume $l = 1$. The generalization to multiple quantities is covered in Remark 4. Theorem 2 implies that the asymptotic mean squared error (AMSE) of the focus quantity $\mu\left((\hat{\boldsymbol{\theta}}_{T,m}, \mathbf{0}_{K^2(p-m)}), \hat{\boldsymbol{\sigma}}\right)$ is

$$\mathrm{AMSE}\left(\mu\left((\hat{\boldsymbol{\theta}}_{T,m}, \mathbf{0}_{K^2(p-m)}), \hat{\boldsymbol{\sigma}}\right)\right) = \boldsymbol{D}_\theta\left[\boldsymbol{C}_m\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{C}_m' + \boldsymbol{P}_m(\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma})\boldsymbol{P}_m\right]\boldsymbol{D}_\theta' + \boldsymbol{D}_\sigma\boldsymbol{\Xi}_{22}\boldsymbol{D}_\sigma'. \tag{2.9}$$

There are three contributions to the AMSE: (1) the term $\boldsymbol{D}_\theta\boldsymbol{C}_m\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{C}_m'\boldsymbol{D}_\theta'$ is an asymptotic squared bias originating from the exclusion of local-to-zero parameters, (2) the asymptotic variance contribution $\boldsymbol{D}_\theta\boldsymbol{P}_m(\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma})\boldsymbol{P}_m\boldsymbol{D}_\theta'$, and (3) the contribution $\boldsymbol{D}_\sigma\boldsymbol{\Xi}_{22}\boldsymbol{D}_\sigma'$ which does not depend on the lag order m. Overall we face a bias-variance tradeoff when having to decide on m.

The FIC is an estimate of the AMSE. The quantities $\hat{\boldsymbol{\theta}}_{T,p}$ and $\hat{\boldsymbol{\Omega}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_{T,t}\boldsymbol{z}_{T,t}'$ provide consistent estimates for $\boldsymbol{\theta}_T$ and $\boldsymbol{\Omega}$, respectively. In view of the continuous mapping theorem, $\hat{\boldsymbol{P}}_m = \boldsymbol{S}_m\left[\boldsymbol{S}_m'\hat{\boldsymbol{\Omega}}\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}_m'\otimes\boldsymbol{I}_K$ and

$$\hat{\boldsymbol{C}}_m = \left(\boldsymbol{S}_m\left[\boldsymbol{S}_m'\hat{\boldsymbol{\Omega}}\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}_m'\hat{\boldsymbol{\Omega}} - \boldsymbol{I}_{Kp}\right)\boldsymbol{S}_0\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\otimes\boldsymbol{I}_K, \tag{2.10}$$

are consistent estimators as well. *A consistent estimator for $\boldsymbol{\delta}$ is not available due to the adopted misspecification framework*. We follow the existing literature (see Claeskens and Hjort (2003), Liu (2015), and Charkhi et al. (2016) among others) and use $\hat{\boldsymbol{\delta}} = \sqrt{T}\mathrm{vec}(\hat{\boldsymbol{\Theta}}_{T,p}\boldsymbol{S}_0)$ which satisfies[6]

$$\hat{\boldsymbol{\delta}} \xrightarrow{d} \boldsymbol{R}_\delta = \boldsymbol{\delta} + (\boldsymbol{S}_0'\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{I}_K)\boldsymbol{R} \sim \mathrm{N}(\boldsymbol{\delta}, \boldsymbol{S}_0'\boldsymbol{\Omega}^{-1}\boldsymbol{S}_0\otimes\boldsymbol{\Sigma}). \tag{2.11}$$

[6] $\hat{\boldsymbol{\delta}} = \sqrt{T}\,\mathrm{vec}(\hat{\boldsymbol{\Theta}}_{T,p}\boldsymbol{S}_0)$ is the sample equivalent of $\boldsymbol{\delta} = \mathrm{vec}(\boldsymbol{\Delta}) = \sqrt{T}\mathrm{vec}(\boldsymbol{C}_T) = \sqrt{T}\mathrm{vec}(\boldsymbol{\Theta}_{T,p}\boldsymbol{S}_0)$.

Asymptotically, we have $\mathrm{E}(\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}') = \boldsymbol{\delta}\boldsymbol{\delta}' + \boldsymbol{S}_0'\boldsymbol{\Omega}^{-1}\boldsymbol{S}_0 \otimes \boldsymbol{\Sigma}$. Using the asymptotically unbiased estimate $\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}' - \boldsymbol{S}_0'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{S}_0 \otimes \hat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\delta}\boldsymbol{\delta}'$, the FIC for the approximating model with m lags is defined as

$$\widehat{FIC}_m = \boldsymbol{D}_\theta \left[\hat{\boldsymbol{C}}_m \left(\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}' - \boldsymbol{S}_0'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{S}_0 \otimes \hat{\boldsymbol{\Sigma}}\right)\hat{\boldsymbol{C}}_m' + \hat{\boldsymbol{P}}_m(\hat{\boldsymbol{\Omega}} \otimes \hat{\boldsymbol{\Sigma}})\hat{\boldsymbol{P}}_m\right]\boldsymbol{D}_\theta' + \boldsymbol{D}_\sigma \hat{\boldsymbol{\Xi}}_{22}\boldsymbol{D}_\sigma'. \tag{2.12}$$

This estimate of the AMSE can be computed for every model, and the model with the lowest $\widehat{FIC}_m$ is selected. We elaborate on implications of inconsistent estimation of $\boldsymbol{\delta}$ in Section 2.4.

Remark 4. The same procedure can be followed when $l > 1$, but the AMSE becomes an $(l \times l)$ matrix. The trace or determinant are meaningful ways to describe this AMSE matrix by a see scalar Charkhi et al. (2016).[7] The trace is computationally convenient because the overall FIC will be the sum of the individual univariate FIC contributions.

2.3.2. Model averaging: Plug-in averaging

Liu (2015) proposed a model averaging approach along the lines of the FIC. It was named plug-in averaging. We again depart from the case $l = 1$, see Remark 5 for the generalization. Part (b) from Theorem 1 implies that linear combinations of the VAR parameter estimates are also asymptotically normally distributed. Interpret the coefficients in the linear combination as weights, i.e., define $\boldsymbol{w} = (w_{p_1}, w_{p_1+1}, \ldots, w_p)$ with $\boldsymbol{w} \in \mathcal{H} = \left\{\boldsymbol{w} \in [0,1]^{p_2+1} : \sum_{m=p_1}^p w_m = 1\right\}$.[8]

Theorem 3 details the asymptotic distribution of the following weighted estimator,

$$\bar{\mu}(\boldsymbol{w}) = \sum_{m=p_1}^{p} w_m \mu\big((\hat{\boldsymbol{\theta}}_{T,m}, \boldsymbol{0}_{K^2(p-m)}), \hat{\boldsymbol{\sigma}}\big). \tag{2.13}$$

Theorem 3 (Asymptotic Normality of the Plug-In Estimator). *Under the Assumptions of Theorem 2, we have for $T \to \infty$,*

$$\begin{aligned}\sqrt{T}\Big(\bar{\mu}(\boldsymbol{w}) - \mu(\boldsymbol{\theta}_T, \boldsymbol{\sigma})\Big) &\xrightarrow{d} \boldsymbol{D}_\theta \sum_{m=p_1}^{p} w_m \boldsymbol{C}_m \boldsymbol{\delta} + \boldsymbol{D}_\theta \sum_{m=p_1}^{p} w_m \boldsymbol{P}_m \boldsymbol{R} + \boldsymbol{D}_\sigma \boldsymbol{S} \\ &\sim \mathrm{N}\left(\boldsymbol{D}_\theta \sum_{m=p_1}^{p} w_m \boldsymbol{C}_m \boldsymbol{\delta}, \boldsymbol{V} + \boldsymbol{D}_\sigma \boldsymbol{\Xi}_{22}\boldsymbol{D}_\sigma'\right),\end{aligned}$$

with $\boldsymbol{V} = \sum_{m=p_1}^{p}\sum_{l=p_1}^{p} w_m w_l \boldsymbol{D}_\theta \boldsymbol{P}_m (\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}) \boldsymbol{P}_l \boldsymbol{D}_\theta'$.

As for the FIC, we compute the AMSE and find[9]

$$\mathrm{AMSE}(\bar{\mu}(\boldsymbol{w})) = \sum_{m=p_1}^{p}\sum_{l=p_1}^{p} w_m \boldsymbol{D}_\theta \Big(\boldsymbol{C}_m \boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{C}_l' + \boldsymbol{P}_m (\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}) \boldsymbol{P}_l\Big)\boldsymbol{D}_\theta' w_l = \boldsymbol{w}'\boldsymbol{\Psi}\boldsymbol{w}, \tag{2.14}$$

with the $((p_2 + 1) \times (p_2 + 1))$ matrix $\boldsymbol{\Psi}$ having the (m, l)th element $\boldsymbol{\Psi}_{m,l} = \boldsymbol{D}_\theta\Big(\boldsymbol{C}_m\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{C}_l' + \boldsymbol{P}_m (\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}) \boldsymbol{P}_l\Big)\boldsymbol{D}_\theta'$. The optimal weight vector that minimizes the AMSE is given by $\boldsymbol{w}^0 = \arg\min_{\boldsymbol{w}\in\mathcal{H}} \boldsymbol{w}'\boldsymbol{\Psi}\boldsymbol{w}$. But $\boldsymbol{w}^0$ depends on population quantities, so using the same estimates for population

[7] Any mapping from the AMSE matrix to a scalar can be used, e.g., matrix norms could be used as well.
[8] We assume that we average over all the models in $\mathcal{M} = \{p_1, p_1 + 1, \ldots, p\}$.
[9] The contribution $\boldsymbol{D}_\sigma \boldsymbol{\Xi}_{22}\boldsymbol{D}_\sigma'$ does not depend on m, and is therefore inconsequential for the analysis. This term will be omitted from the AMSE to allow for an easier presentation.

quantities as in Section 2.3.1, we compute feasible weights as

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{H}} \boldsymbol{w}'\hat{\boldsymbol{\Psi}}^{j}\boldsymbol{w}, \qquad j \in \{\text{Biased, Bias cor}\}, \tag{2.15}$$

with

$$\begin{aligned} \hat{\boldsymbol{\Psi}}^{\text{Biased}}_{m,l} &= \boldsymbol{D}_\theta \left[\hat{\boldsymbol{C}}_m \hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{C}}_l' + \hat{\boldsymbol{P}}_m\left(\hat{\boldsymbol{\Omega}} \otimes \hat{\Sigma}\right)\hat{\boldsymbol{P}}_l\right]\boldsymbol{D}_\theta', \\ \hat{\boldsymbol{\Psi}}^{\text{Bias cor}}_{m,l} &= \boldsymbol{D}_\theta \left[\hat{\boldsymbol{C}}_m\left(\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}' - \boldsymbol{S}_0'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{S}_0 \otimes \hat{\boldsymbol{\Sigma}}\right)\hat{\boldsymbol{C}}_l' + \hat{\boldsymbol{P}}_m\left(\hat{\boldsymbol{\Omega}} \otimes \hat{\Sigma}\right)\hat{\boldsymbol{P}}_l\right]\boldsymbol{D}_\theta'. \end{aligned} \tag{2.16}$$

The sole difference between the matrix elements in Equation (2.16) is an asymptotic bias correction for $\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}'$. For both versions, Equation (2.15) is a quadratic programming problem with linear constraints. Solvers are readily available (for example "quadprog" in Matlab, or "qprog" in Gauss). The estimator for $\boldsymbol{\delta}$ remains inconsistent and we again refer to Section 2.4 for a discussion of the implications.

Remark 5. As in Remark 4, we will obtain an $(l \times l)$ AMSE matrix for multiple quantities of interest. This matrix has to be summarized by a scalar. The trace has again computational benefits because the objective function will take the form $\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{H}} \boldsymbol{w}'\left(\sum_i \hat{\boldsymbol{\Psi}}^j_i\right)\boldsymbol{w}$ with $\hat{\boldsymbol{\Psi}}^j_i$ the matrix corresponding to the ith focus quantity.

Remark 6. Two remarks related to Charkhi et al. (2016) are in place.

1. The weight vector $\hat{\boldsymbol{w}}$ is only uniquely determined when $\hat{\boldsymbol{\Psi}}^j$ is positive definite. As such, the bias subtraction may lead to nonunique weights.
2. Charkhi et al. (2016) consider a weighting scheme in which the weights sum to one but are not necessarily positive. Simulation results have shown that it is advisable to keep the positivity constraint in our autoregressive setup because weights can otherwise become large in magnitude and rather unstable.

Remark 7. Autoregressive models of infinite order have been considered by Berk (1974) and Lewis and Reinsel (1985) among others. It is an intriguing question whether the current framework can be extended to VAR(∞) models.[10] We argue that the main difficulty is the estimation of the infinitely many local-to-zero parameters. Let us consider the univariate model $y_{T,t} = \alpha y_{T,t-1} + \sum_{j=1}^{\infty}\left(\frac{\delta_j}{\sqrt{T}}\right) y_{T,t-1-j} + u_t$ as an illustration. We conjecture[11] that the asymptotic distribution of the approximating AR(1) model follows $\sqrt{T}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathrm{N}\left(\sum_{j=1}^{\infty}\delta_j\alpha^j, 1-\alpha^2\right)$. The bias contribution to the AMSE now depends on infinitely many δ_j. Their estimation would require the lag order of the largest approximating model to grow with sample size. Our proof of Theorem 1 does not easily allow for such an extension since we currently rely on the finite dimension of the companion matrix. A full exploration of this topic is left for further research.

There is one final result that forces us to look at the case $l > 1$. Practitioners are usually interested in the impulse responses for several horizons. Using a separate weight vector for every horizon may: (1) create impulse responses that vary irregularly from one horizon to the next due to strong changes in the weights, and (2) result in confidence intervals that do not take into account the dependence between the horizons. Theorem 4 extends the result of Liu (2015) to obtain asymptotically valid confidence bands for several horizons.

[10] We thank an anonymous referee for bringing this topic to our attention.

[11] We can be more precise concerning our assumptions. Theorem 1 has shown that the asymptotic results are governed by the process with the local-to-zero parameters being set equal to zero. We assume that this remains true when there are infinitely many local-to-zero parameters.

Theorem 4 (Joint Confidence Bands). *Under the Assumptions of Theorem 2, if* $w_m(\hat{\boldsymbol{\delta}}) \xrightarrow{d} w_m(\boldsymbol{R}_\delta)$, *and if* $\boldsymbol{D}_\theta\left(\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma}\right)\boldsymbol{D}_\theta' + \boldsymbol{D}_\sigma\boldsymbol{\Xi}_{22}\boldsymbol{D}_\sigma \succ 0$, *then*

$$\left(\sqrt{T}\left(\bar{\boldsymbol{\mu}}(\hat{\boldsymbol{w}})-\boldsymbol{\mu}(\boldsymbol{\theta}_T,\boldsymbol{\sigma})\right)-\boldsymbol{D}_\theta\sum_{m=p_1}^{p}\hat{w}_m\hat{\boldsymbol{C}}_m\hat{\boldsymbol{\delta}}\right)'\left(\boldsymbol{D}_\theta\left(\hat{\boldsymbol{\Omega}}\otimes\hat{\boldsymbol{\Sigma}}\right)\boldsymbol{D}_\theta'+\boldsymbol{D}_\sigma\hat{\boldsymbol{\Xi}}_{22}\boldsymbol{D}_\sigma\right)^{-1}$$

$$\left(\sqrt{T}\left(\bar{\boldsymbol{\mu}}(\hat{\boldsymbol{w}})-\boldsymbol{\mu}(\boldsymbol{\theta}_T,\boldsymbol{\sigma})\right)-\boldsymbol{D}_\theta\sum_{m=p_1}^{p}\hat{w}_m\hat{\boldsymbol{C}}_m\hat{\boldsymbol{\delta}}\right)\leq\chi^2_{l,1-\alpha},$$

is an asymptotically correct $(1-\alpha)$ *confidence band, where* $\chi^2_{l,1-\alpha}$ *denotes the* $(1-\alpha)$ *quantile of a chi-squared distributed random variable with l degrees of freedom.*

Remark 8. There is one practical concern which has not been addressed, namely the choices for p_1 and p.[12] p_1 will turn out to be unimportant. To see this, we consider the expression for $\widehat{FIC}_m$ (a similar reasoning applies to the plug-in weights). The terms $\boldsymbol{D}_\theta\hat{\boldsymbol{P}}_m(\hat{\boldsymbol{\Omega}}\otimes\hat{\boldsymbol{\Sigma}})\hat{\boldsymbol{P}}_m\boldsymbol{D}_\theta'$ and $\boldsymbol{D}_\sigma\hat{\boldsymbol{\Xi}}_{22}\boldsymbol{D}_\sigma'$ in Equation (2.12) do not depend on p_1, so it remains to inspect the contribution $\boldsymbol{D}_\theta\hat{\boldsymbol{C}}_m\left(\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}'-\boldsymbol{S}_0'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{S}_0\otimes\hat{\boldsymbol{\Sigma}}\right)\hat{\boldsymbol{C}}_m'\boldsymbol{D}_\theta'$. Using $\hat{\boldsymbol{\delta}}=\sqrt{T}(\boldsymbol{S}_0'\otimes\boldsymbol{I}_K)\hat{\boldsymbol{\theta}}_{T,p}$ we can rewrite this contribution as

$$\boldsymbol{D}_\theta\left(\hat{\boldsymbol{C}_m}(\boldsymbol{S}_0'\otimes\boldsymbol{I}_K)\right)\left[(\sqrt{T}\hat{\boldsymbol{\theta}}_{T,p})(\sqrt{T}\hat{\boldsymbol{\theta}}_{T,p})'-\hat{\boldsymbol{\Omega}}^{-1}\otimes\hat{\boldsymbol{\Sigma}}\right]\left(\hat{\boldsymbol{C}}_m(\boldsymbol{S}_0'\otimes\boldsymbol{I}_K)\right)'\boldsymbol{D}_\theta'. \tag{2.17}$$

By definition of $\hat{\boldsymbol{C}}_m$, we have

$$\begin{aligned}\hat{\boldsymbol{C}}_m(\boldsymbol{S}_0'\otimes\boldsymbol{I}_K)&=\left(\boldsymbol{S}_m\left[\boldsymbol{S}_m'\hat{\boldsymbol{\Omega}}\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}_m'\hat{\boldsymbol{\Omega}}-\boldsymbol{I}_{Kp}\right)\left[\boldsymbol{S}_0\left(\boldsymbol{I}_{Kp_2}-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{S}_0'\right]\otimes\boldsymbol{I}_K\\&=\left(\boldsymbol{S}_m\left[\boldsymbol{S}_m'\hat{\boldsymbol{\Omega}}\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}_m'\hat{\boldsymbol{\Omega}}-\boldsymbol{I}_{Kp}\right)\underbrace{\left(\begin{bmatrix}\mathbf{O}&\mathbf{O}\\\mathbf{O}&\boldsymbol{I}_{K(p-m)}\end{bmatrix}\otimes\boldsymbol{I}_K\right)}_{\mathcal{B}},\end{aligned} \tag{2.18}$$

thereby showing that actually none of the contributions to $\widehat{FIC}_m$ depends on p_1. However, the zero pattern of the matrix $\mathcal{B}$ in Equation (2.18) will only cause a nondiverging value for $\widehat{FIC}_m$ if models are chosen such that $m\in\mathcal{M}=\{p_1,p_1+1,\ldots,p\}$. This supports the claim in Remark 1. The lag order of the full model, that is p, might be chosen by AIC or set equal to an a priori selected p_{max}.

2.4. Effects of inconsistently estimating delta

Equation (2.11) showed that $\hat{\boldsymbol{\delta}}$ converges to a normally distributed random vector centered around $\boldsymbol{\delta}$. How does this influence the selection and averaging procedures? Clearly, $\widehat{FIC}_m$, $\hat{\Psi}^{\text{Biased}}_{m,l}$, and $\hat{\Psi}^{\text{Bias cor}}_{m,l}$ will not converge in probability to the AMSE they are intended to estimate. The limiting distributions of these quantities are highlighted in Theorems 5 and 6. The plug-in results are stated for $\hat{\Psi}^{\text{Bias cor}}_{m,l}$, but a simple omission of the bias correction term would give the corresponding findings for $\hat{\Psi}^{\text{Biased}}_{m,l}$.

Theorem 5 ([The Asymptotic Behavior of $\widehat{FIC}_m$).] *Under the Assumptions of Theorem 2, we have for* $m\in\mathcal{M}\backslash p$,

$$\begin{aligned}\widehat{FIC}_m&\xrightarrow{d}\boldsymbol{D}_\theta\left[\boldsymbol{C}_m\left(\boldsymbol{R}_\delta\boldsymbol{R}_\delta'-(\boldsymbol{S}_0'\boldsymbol{\Omega}\boldsymbol{S}_0\otimes\boldsymbol{\Sigma})\right)\boldsymbol{C}_m'+\boldsymbol{P}_m(\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma})\boldsymbol{P}_m\right]\boldsymbol{D}_\theta'+\boldsymbol{D}_\sigma\boldsymbol{\Xi}_{22}\boldsymbol{D}_\sigma'\\&:=\ FIC_m^\infty\sim a_m\chi^2_{noncentral}\left(1,\left(\boldsymbol{D}_\theta\boldsymbol{C}_m\boldsymbol{\delta}\right)^2/a_m\right)-a_m+\boldsymbol{D}_\theta\boldsymbol{P}_m(\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma})\boldsymbol{P}_m\boldsymbol{D}_\theta'+\boldsymbol{D}_\sigma\boldsymbol{\Xi}_{22}\boldsymbol{D}_\sigma',\end{aligned}$$

[12] We thank an anonymous referee who rightfully conjectured that the choice of p_1 does not have an influence on the numerical outcome for $\widehat{FIC}_m$.

where $a_m = \boldsymbol{D}_\theta \boldsymbol{C}_m(\boldsymbol{S}_0'\boldsymbol{\Omega}^{-1}\boldsymbol{S}_0 \otimes \boldsymbol{\Sigma})\boldsymbol{C}_m'\boldsymbol{D}_\theta'$, *and* $\chi^2_{noncentral}(\nu, \lambda)$ *denotes a noncentral chi-squared distributed random variable with* ν *degrees of freedom and noncentrality parameter* λ. *It can be shown that* $\mathrm{E}\left(FIC_m^\infty\right) = AMSE\left(\mu(\hat{\boldsymbol{\theta}}_{T,m}, \boldsymbol{0}_{K^2(p-m)}, \hat{\boldsymbol{\sigma}})\right)$ *and* $\mathrm{var}(FIC_m^\infty) = 2a_m(a_m + 2(\boldsymbol{D}_\theta \boldsymbol{C}_m \boldsymbol{\delta})^2)$. *For the full model,* $m = p$, *we have*

$$\widehat{FIC}_p \xrightarrow{p} FIC_p^\infty = \boldsymbol{D}_\theta(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Sigma})\boldsymbol{D}_\theta' = AMSE\left(\mu(\hat{\boldsymbol{\theta}}_{T,p}, \hat{\boldsymbol{\sigma}})\right).$$

Theorem 6 (The Asymptotic Behavior of $\hat{\Psi}^{\text{Biased}}_{m,l}$ and $\hat{\Psi}^{\text{Bias cor}}_{m,l}$). *Under the Assumptions of Theorem 2, we have for* $m, l \in \mathcal{M}\backslash p$,

$$\begin{aligned}\hat{\Psi}^{Bias\ cor}_{m,l} &\xrightarrow{d} \boldsymbol{R}_\delta' \boldsymbol{C}_m' \boldsymbol{D}_\theta' \boldsymbol{D}_\theta \boldsymbol{C}_l \boldsymbol{R}_\delta + \boldsymbol{D}_\theta \boldsymbol{P}_m(\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma})\boldsymbol{P}_l \boldsymbol{D}_\theta' - \boldsymbol{D}_\theta \boldsymbol{C}_m(\boldsymbol{S}_0'\boldsymbol{\Omega}\boldsymbol{S}_0 \otimes \boldsymbol{\Sigma})\boldsymbol{C}_m'\boldsymbol{D}_\theta' \\ &:= \boldsymbol{\Psi}^{Bias\ cor,\infty}_{m,l}.\end{aligned}$$

Two cases can be distinguished:

(a) *If* $m = l$, *then* $\boldsymbol{\Psi}^{Bias\ cor,\infty}_{m,m} \sim a_m \chi^2_{noncentral}\left(1, \left(\boldsymbol{D}_\theta \boldsymbol{C}_m \boldsymbol{\delta}\right)^2 / a_m\right) - a_m + \boldsymbol{D}_\theta \boldsymbol{P}_m(\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma})\boldsymbol{P}_m \boldsymbol{D}_\theta'$.

(b) *Define* $\mathcal{A} = \boldsymbol{S}_0'\boldsymbol{\Omega}^{-1}\boldsymbol{S}_0 \otimes \boldsymbol{\Sigma}$, *and consider the eigenvalue decomposition* $\frac{1}{2}\mathcal{A}^{1/2}(\boldsymbol{C}_m'\boldsymbol{D}_\theta\boldsymbol{D}_\theta'\boldsymbol{C}_l + \boldsymbol{C}_l'\boldsymbol{D}_\theta\boldsymbol{D}_\theta'\boldsymbol{C}_m)\mathcal{A}^{1/2} = \sum_{i=1}^2 \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i'$, *where* λ_i *denotes the eigenvalue corresponding to the eigenvector* $\boldsymbol{v}_i$. *If* $m \neq l$, *then*

$$\begin{aligned}\boldsymbol{\Psi}^{Bias\ cor,\infty}_{m,l} \sim \sum_{i=1}^{2} \lambda_i \chi^2_{noncentral}\left(1, \left(\boldsymbol{v}_i'\mathcal{A}^{-1/2}\boldsymbol{\delta}\right)^2\right) + \boldsymbol{D}_\theta \boldsymbol{P}_m(\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma})\boldsymbol{P}_l \boldsymbol{D}_\theta' \\ - \boldsymbol{D}_\theta \boldsymbol{C}_m(\boldsymbol{S}_0'\boldsymbol{\Omega}\boldsymbol{S}_0 \otimes \boldsymbol{\Sigma})\boldsymbol{C}_m'\boldsymbol{D}_\theta'.\end{aligned}$$

If $m = p$ *and/or* $l = p$, *then* $\hat{\Psi}^{Bias\ cor}_{m,l} \xrightarrow{p} \boldsymbol{\Psi}_{m,l}$.

Theorems 5 and 6 stated the limiting distribution of the FIC and the matrix elements that enter the weighting scheme. Based on the random limits of these quantities, we might expect that our methods will not truly minimize the AMSE among either model choices or model weights. We proceed with a small illustration to stress the difference between knowing $\boldsymbol{\delta}$ and having an estimator $\hat{\boldsymbol{\delta}}$ that converges in distribution only.

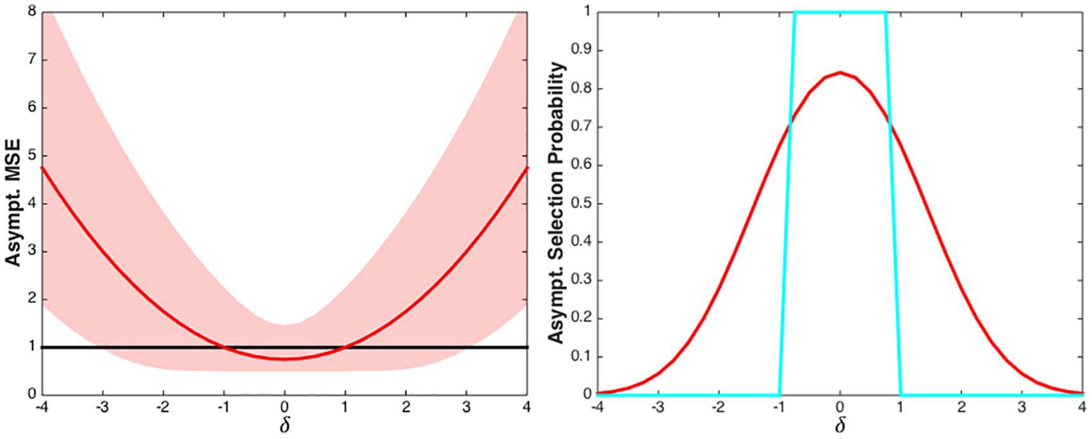
An illustration

Consider a simplified DGP, $y_{T,t} = 0.5y_{T,t-1} + \frac{\delta}{\sqrt{T}}y_{T,t-2} + u_t$, with $\mathrm{var}(u_t) = 1$, and a focus on the impulse response at horizon 1 (i.e., $\boldsymbol{D}_\theta = (1, 0)$ and $\boldsymbol{D}_\sigma = 0$). The model set is $\mathcal{M} = \{1, 2\}$. This simplified setting makes the asymptotic behavior of the FIC and plug-in weights analytically tractable. Figure 1(a) depicts FIC_1^∞ and FIC_2^∞ as a function of δ. Note that FIC_1^∞ converges in distribution and has a nonzero probability to give an outcome below FIC_2^∞. This asymptotic selection probability of the model with $m = 1$ can be calculated analytically using Theorem 5. Figure 1(b) shows that the FIC does not select the model with the smallest AMSE with probability one.

Our simplified model can also be used to examine the effect of $\hat{\delta}$ on the plug-in weights. We focus on the weights in the absence of bias correction.[13] The (2 × 2) limiting matrix $\boldsymbol{\Psi}^\infty$ is

$$\boldsymbol{\Psi}^\infty = \begin{bmatrix} a_1\chi^2_{noncentral}\left(1, (\boldsymbol{D}_\theta \boldsymbol{C}_1 \boldsymbol{\delta})^2/a_1\right) + \sigma^2 \boldsymbol{D}_\theta \boldsymbol{P}_1 \boldsymbol{D}_\theta' & \sigma^2 \boldsymbol{D}_\theta \boldsymbol{P}_1 \boldsymbol{D}_\theta' \\ \sigma^2 \boldsymbol{D}_\theta \boldsymbol{P}_1 \boldsymbol{D}_\theta' & \sigma^2 \boldsymbol{D}_\theta \boldsymbol{\Omega}^{-1} \boldsymbol{D}_\theta' \end{bmatrix}, \tag{2.19}$$

[13] There is a finite probability for the matrix $\boldsymbol{\Phi}^\infty$ to have a negative eigenvalue when the bias correction is applied. This severely complicates the derivations, so we exclude this case from our analysis. For the $\boldsymbol{\Psi}^\infty$ matrix without bias correction we will have $\boldsymbol{\Psi}^\infty \succ 0$ if $\boldsymbol{D}_\theta'(\boldsymbol{\Omega}^{-1} - \boldsymbol{P}_1)\boldsymbol{D}_\theta > 0$. The latter requirement is equivalent to $\boldsymbol{D}_\theta' \left[\begin{smallmatrix} -\omega_{12} \\ \omega_{11} \end{smallmatrix}\right] \neq 0$.



(a) (b)

Figure 1. (a) The asymptotic MSE of the models with one and two lags (red and black line, respectively). The area between the 5 and 95% quantiles of FIC_1^{∞} is shaded in red. $\widehat{FIC}_2$ converges in probability to the values of the black line. (b) The asymptotic selection probabilities of the model with $m = 1$. The infeasible estimator takes a binary decision based on whether the red or black line in graph (a) is lowest. Model selection based on the focused information criterion results in a smoothed asymptotic selection probability because $\widehat{FIC}_1$ converges in distribution.

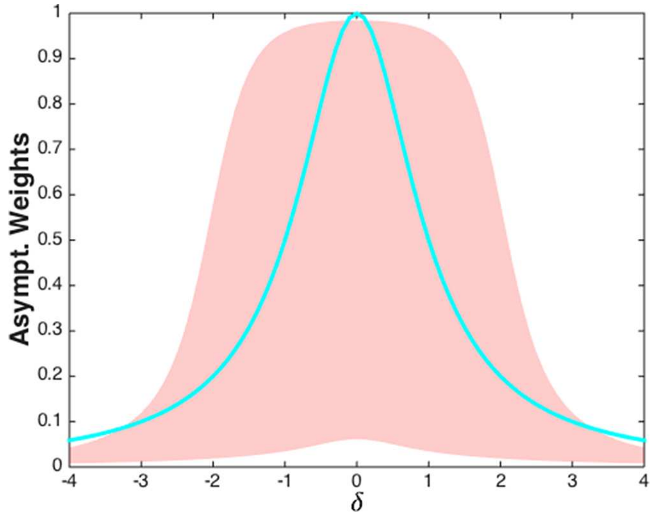


Figure 2. The 5 and 95% quantiles of the asymptotic distribution of the weights as a shaded red area, see Equation (2.20). The solid cyan lines are the asymptotically optimal weights which can only be obtained if either δ is known (infeasible) or a consistent estimator for δ is available.

where $a_1 = \sigma^2 \boldsymbol{D}_\theta \boldsymbol{C}_1 \boldsymbol{S}_0' \boldsymbol{\Omega} \boldsymbol{S}_0 \boldsymbol{C}_1' \boldsymbol{D}_\theta'$ (see Theorem 6). Let w^* denote the asymptotically optimal plug-in weight for the model with $m = 1$. We have

$$\Pr\left(w^* \leq x\right) = \Pr\left(\chi^2_{noncentral}\left(1, (\boldsymbol{D}_\theta \boldsymbol{C}_1 \boldsymbol{\delta})^2/a_1\right) \geq \frac{\sigma^2 \boldsymbol{D}_\theta (\boldsymbol{\Omega}^{-1} - \boldsymbol{P}_1)\boldsymbol{D}_\theta' \,[1-x]}{a_1 x}\right). \tag{2.20}$$

Figure 2 shows the area between the 5 and 95% quantiles of w^* together with the optimal weight for known delta. We see that the asymptotic distribution of w^* is located closer to zero than the optimal infeasible weights. This is unsurprising because the lack of bias-correction causes (on average) an overestimation of the AMSE of the model with $m = 1$.

Remark 9. The exposition in this section was based on a simplified model. We concluded that the absence of a consistent estimator for δ translates into suboptimal model selection and suboptimal model averaging. Also in more elaborate models, the FIC and the elements of the weighting matrix $\boldsymbol{\Psi}$ will converge to random variables (except for $m = p$). It is key to realize that the AMSE with estimated $\boldsymbol{\delta}$ will not coincide with the AMSE that can be attained if $\boldsymbol{\delta}$ was either known or consistently estimated. We conjecture that these considerations are equally relevant outside an autoregressive framework, e.g., in the regression framework discussed in Liu (2015) and the likelihood framework of Charkhi et al. (2016).

3. Simulations

This simulation section consists of three parts.[14] In the first part we will verify our derivations for the simplified DGP, and see how the suboptimal selection/averaging affects the finite sample MSE. This section is followed by a study of the impulse responses for different horizons in a univariate and multivariate setting. All our graphs are made as a function of the scalar δ. This scalar measures the amount of misspecification and the closeness to unit root.[15] Any missing starting values in the autoregressive recursion were replaced by zeros, and the first 100 data points were omitted as a presample. All results are based on 100,000 Monte Carlo replications.

The performance of the various methods was assessed using the empirical mean squared error. For model selection the featured methods are:

1. The Akaike information criterion ("**AIC**") and Bayesian information criterion ("**BIC**"), e.g., Section 4.3 of Lütkepohl (2005) and the original papers by Akaike (1998) and Schwarz et al. (1978).
2. The "**FIC**" from Equation (2.12) with estimated $\boldsymbol{\delta}$.
3. An infeasible version of the FIC abbreviated as "**Infeas.**" This information criterion is based on population quantities, especially $\boldsymbol{\delta}$ is known.

For the model averaging setup we consider:

1. "**sAIC**" and "**sBIC**" as smoothed counterparts of the AIC and BIC, see Burnham and Anderson (2002). To illustrate, let $AIC(m)$ denote the AIC for model $m \in \mathcal{M}$. The smoothed AIC assigns a weight of $\exp\left(-\frac{1}{2}AIC(m)\right) / \sum_{m\in\mathcal{M}} \exp\left(-\frac{1}{2}AIC(m)\right)$ to model m.
2. Three plug-in averages are reported. "**Plug-in**" and "**Plug-in Corr.**" are computed from Equations (2.15) and (2.16), where only the second average uses the bias correction on $\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}'$. The plug-in average based on known $\boldsymbol{\delta}$ is denoted "**Infeas.**"
3. The "**Jackknife**" model averaging procedure detailed in Hansen and Racine (2012) and Zhang et al. (2013).
4. The Stein combination shrinkage method used in the simulation section of Hansen (2016) is abbreviated "**SteinH.**" This shrinkage method combines VAR(1) through VAR(p) models as well as univariate AR(1) through AR(p) models. Our DGP contains considerable interaction between the cross-sectional units so we also consider a shrinkage method abbreviated "**Stein**" which only combines the VAR(1) through VAR(p).

[14] A selection of simulation results is reported here, extensive results can be found in the Supplementary Material.

[15] Previous studies (e.g., Hansen (2007), Hansen (2008), Hansen and Racine (2012), Liu and Okui (2013), Zhang et al. (2013) and Liu (2015)) show the performance as a function of the population R^2. This representation is inconvenient in our dynamic setup because it is unclear when we are approaching the boundary of the stationarity region.

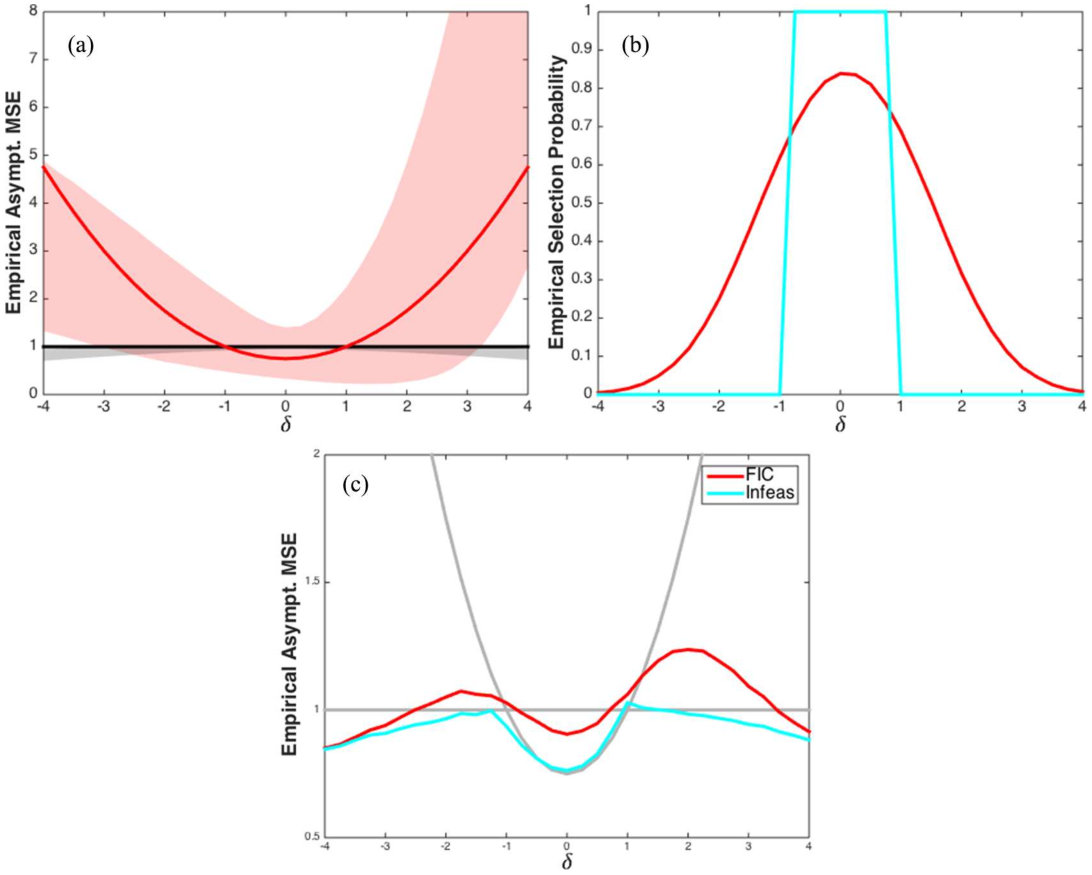


Figure 3. (a) The empirical asymptotic MSE of the models with one and two lags (red and black line, respectively). The area between the 5 and 95% empirical quantiles of $\widehat{FIC}_1$ and $\widehat{FIC}_2$ are shaded in red and grey. (b) The empirical selection probabilities of the FIC. (c) The AMSE of the models with $m = 1$ and $m = 2$ together with the empirical MSE of the feasible FIC (red) and infeasible FIC (cyan). This figure was obtained for $T = 100$ and should be compared with the asymptotic results in Figure 1.

3.1. Simplified DGP

Figures 3 and 4 provide the finite sample confirmation of the intuition we gained from the simplified DGP.[16] The wide spread in the empirical distribution of $\widehat{FIC}_1$ shown in Figure 3(a) results indeed in a smoothed instead of binary selection between the models (Figure 3(b)). The performance of the feasible FIC is therefore worse than that of the infeasible FIC that assumes δ to be known. At high $|\delta|$, we see that the probability to select the wrong model is small. The feasible FIC, therefore, performs similarly to its infeasible counterpart for large amounts of misspecification only.

In Figure 4, the three panels display results on the plug-in averages. The quantiles of the weight distribution without bias correction should be compared to those in Equation (2.20) and Figure 2. The results match. Figure 4(b) shows the quantiles of the weight distribution with bias correction. As expected, this distribution is shifted toward higher weights because the upward bias of the AMSE of the model with $m = 1$ is removed. We can see in 4(c) that the plug-in averages do not perform as well as the infeasible estimator. Unreported simulation results at a sample size of $T = 1000$ confirm that this effect does not disappear with sample size. The inconsistent estimation of δ again causes the feasible weights to differ from optimal weights.

[16] In this section we have rescaled the empirical MSE by the sample size to make it comparable to the asymptotic results of Figures 1 and 2, hence the label empirical asymptotic MSE.

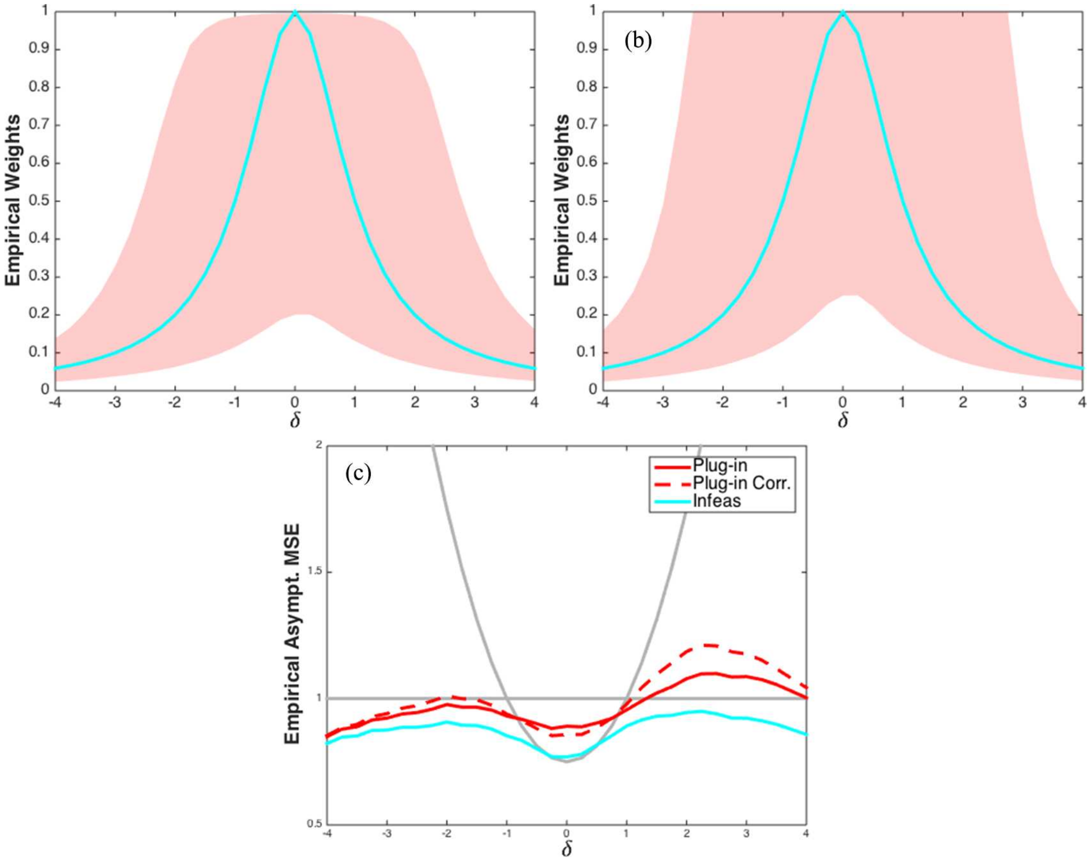


Figure 4. The 5 and 95% empirical quantiles of the weights distribution without bias correction (a) and with bias correction (b). The infeasible weights are displayed in cyan. The empirical MSE of plug-in methods is shown in (c). The sample size is $T = 100$.

3.2. Simulation results for an autoregressive model

Further simulations are based on the following model[17]:

$$y_{T,t} = 0.5y_{T,t-1} + \frac{\delta}{\sqrt{T}}y_{T,t-2} + \frac{\delta}{2\sqrt{T}}y_{T,t-3} + u_t, \qquad u_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1). \tag{3.1}$$

The second and third lag are local-to-zero implying that $\mathcal{M} = \{1,2,3\}$. The coefficients in front of the misspecified lags decline linearly as in Liu (2015), where δ governs the amount of misspecification.The largest modulus eigenvalue of the companion matrix is about 0.3 at $\delta = -0.2$ and increases monotonously to approximately 0.9 at the boundaries of the interval $[-4, 2]$.

Remark 10. The AMSEs of the impulse response at horizon 1 are the same for $m = 2$ and $m = 3$. The plug-in weights are not unique, also see Remark 6.

3.2.1. MSE comparison

The empirical MSE of the various selection methods are shown in Figure 5 for the impulse responses at horizon 1, 3, and 5. Due to the strong penalty on model complexity, the BIC performs well for small

[17] We show in the Supplementary Material that the simplified model with $p_1 = p_2 = 1$, i.e., $y_{T,t} = \alpha y_{T,t-1} + \frac{\delta}{\sqrt{T}}y_{T,t-2} + u_t$, is special because the gradient vector has no influence on model selection and plug-in averaging. We extend the model with an additional lag to see the influence of the impulse response horizon.

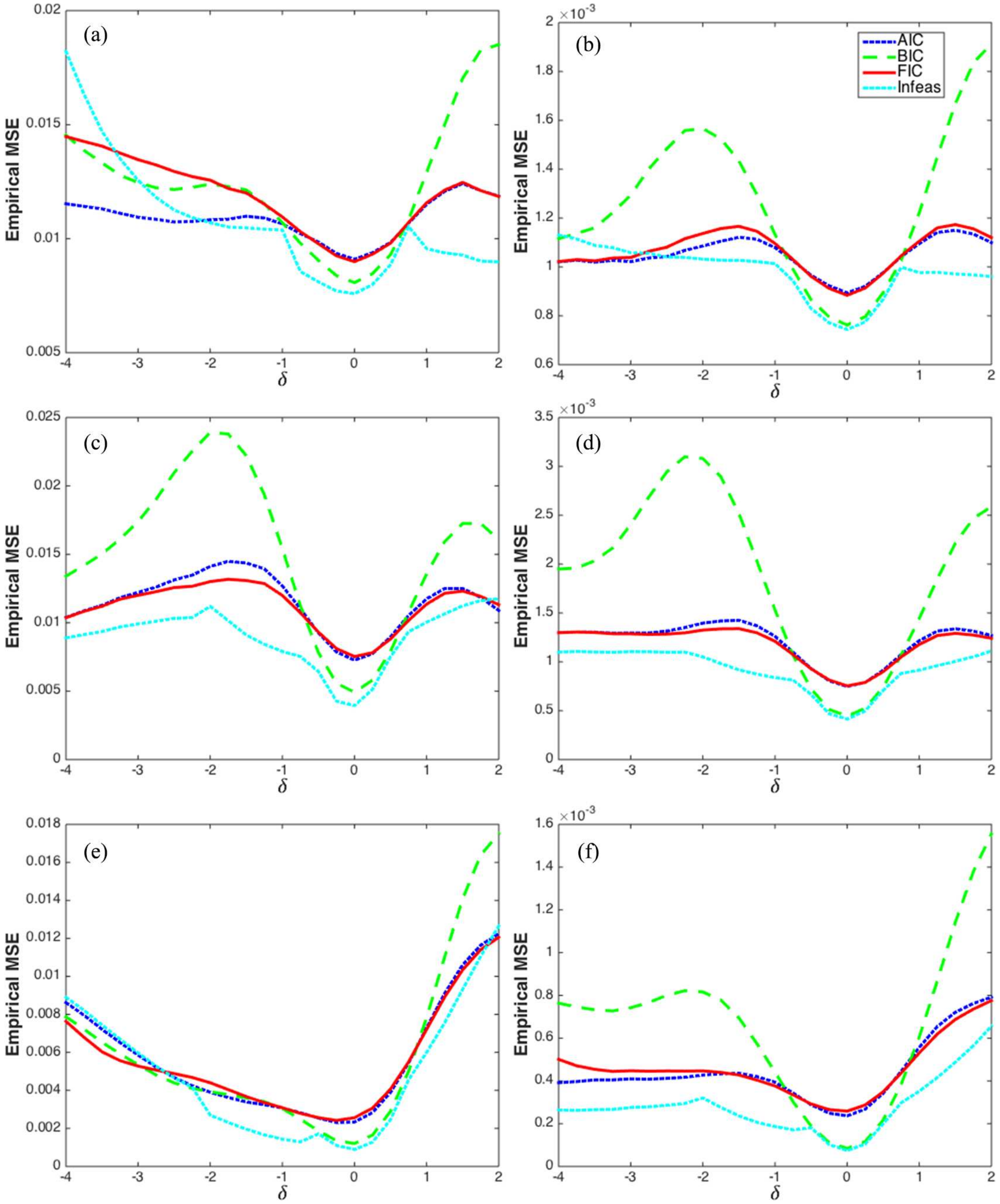


Figure 5. The empirical MSE for model selection. The DGP is $y_{T,t} = 0.5y_{T,t-1} + \frac{\delta}{\sqrt{T}}y_{T,t-2} + \frac{\delta}{2\sqrt{T}}y_{T,t-3} + u_t$. (a) $h = 1$, $T = 100$, (b) $h = 1$, $T = 1000$, (c) $h = 3$, $T = 100$, (d) $h = 3$, $T = 1000$, (e) $h = 5$, $T = 100$, and (f) $h = 5$, $T = 1000$.

amounts of misspecification, but its performance quickly deteriorates as $|\delta|$ increases. The performance of the AIC and the feasible version of the FIC are comparable for large areas of the parameter space, with neither of these methods being preferred to the other. The infeasible FIC is very frequently the preferred method.

Model averaging results are reported in Figure 6. The behavior of the smoothed BIC procedure is similar to that of its selection counterpart, i.e., it only performs well for small δ. The same remark applies to the plug-in average with bias correction. The Jackknife, smoothed AIC, and the plug-in average without bias correction are close competitors, where the plug-in average is a better candidate for

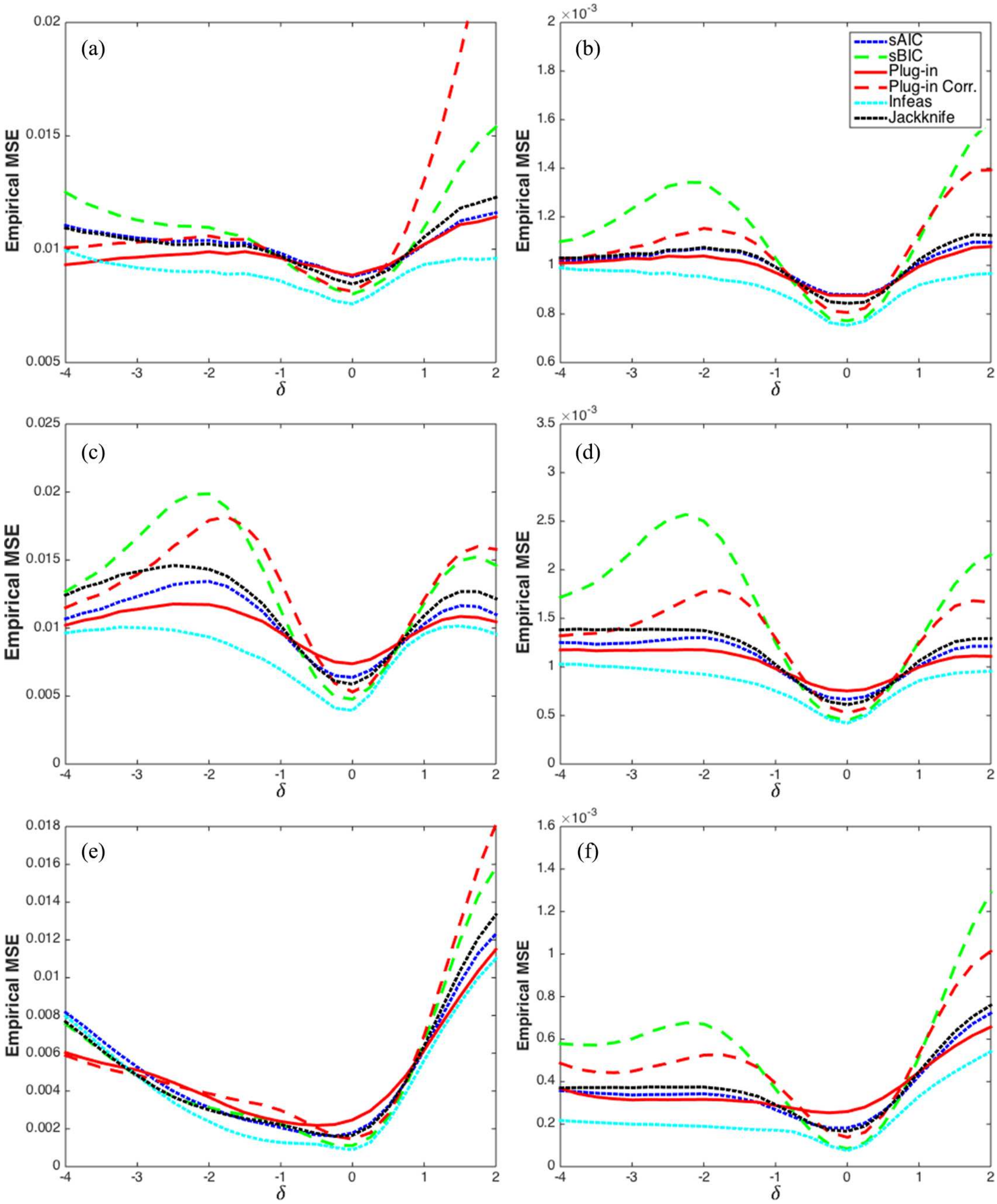
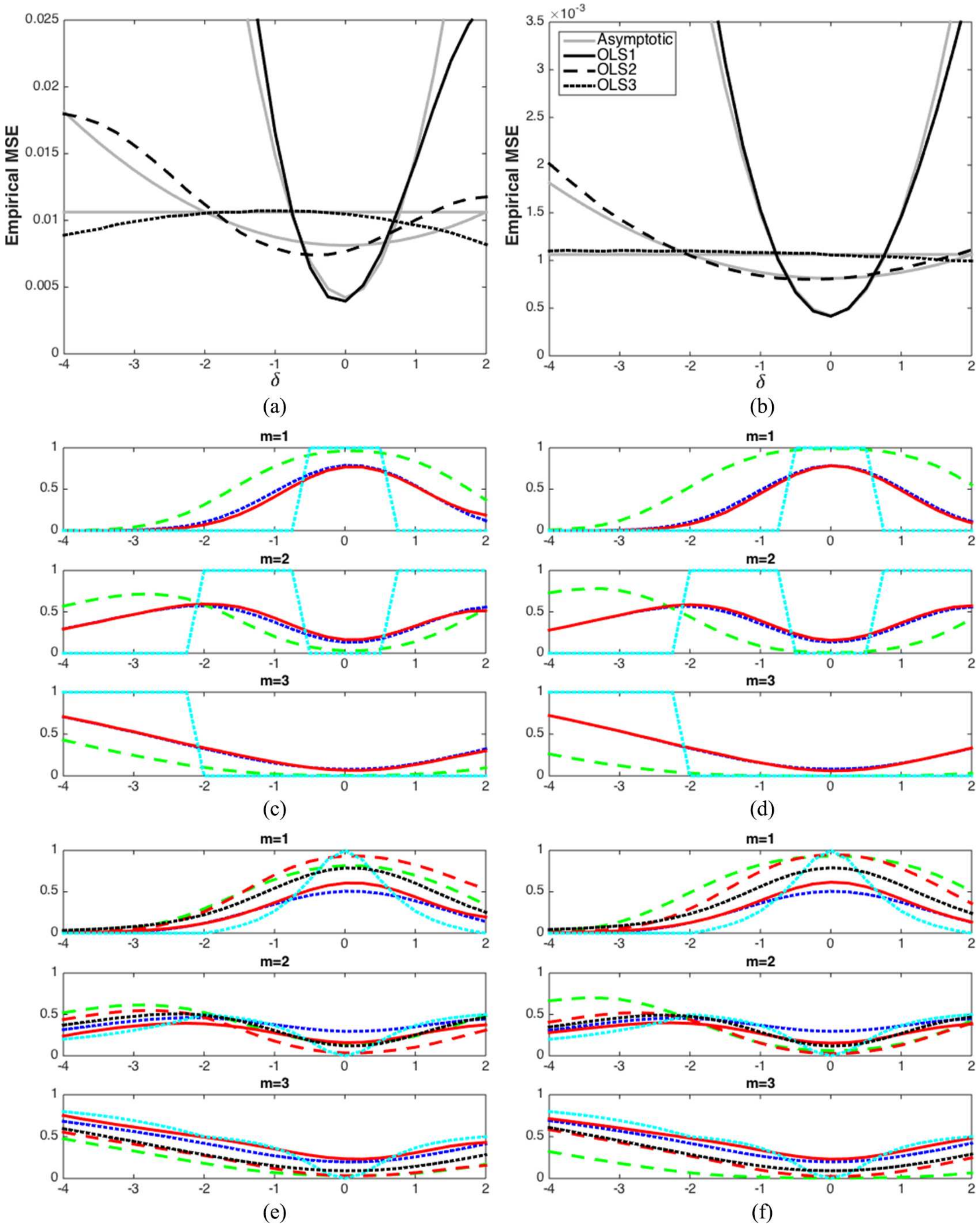


Figure 6. The empirical MSE for model averaging. The DGP is $y_{T,t} = 0.5y_{T,t-1} + \frac{\delta}{\sqrt{T}}y_{T,t-2} + \frac{\delta}{2\sqrt{T}}y_{T,t-3} + u_t$. (a) $h = 1, T = 100$, (b) $h = 1, T = 1000$, (c) $h = 3, T = 100$, (d) $h = 3, T = 1000$, (e) $h = 5, T = 100$, and (f) $h = 5, T = 1000$.



(a) (b) (c) (d) (e) (f)

Figure 7. (a) and (b) The empirical MSE of the OLS estimator of the model with 1 lag (OLS1), 2 lags (OLS2) and the full model with 3 lags (OLS3). Gray lines show the asymptotic MSE approximations as provided by the delta method. (c) and (d) The empirical selection probabilities (see Figure 5 for the appropriate legend). (e) and (f) The empirical distribution of the weights (see Figure 6 for the appropriate legend). The DGP is $y_{T,t} = 0.5y_{T,t-1} + \frac{\delta}{\sqrt{T}}y_{T,t-2} + \frac{\delta}{2\sqrt{T}}y_{T,t-3} + u_t$ for all graphs. (a) $h = 3$, $T = 100$, (b) $h = 3$, $T = 1000$, (c) $h = 3$, $T = 100$, (d) $h = 3$, $T = 1000$, (e) $h = 3$, $T = 100$, and (f) $h = 3$, $T = 1000$.

Table 1. The empirical coverage of 90% confidence intervals for the horizons: two, three and six. For several choices of δ and T in the DGP $y_{T,t} = 0.5y_{T,t-1} + \frac{\delta}{\sqrt{T}}y_{T,t-2} + \frac{\delta}{2\sqrt{T}}y_{T,t-3} + u_t$.

	$h = 2$				$h = 3$				$h = 6$			
δ	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 100$	$T = 250$	$T = 500$	$T = 1000$
−4.0	75.76	81.70	83.04	84.08	76.54	81.13	81.99	82.31	75.85	58.81	42.44	2.04
−3.0	80.06	82.92	83.72	84.40	79.82	82.08	82.42	82.42	67.85	47.84	6.19	11.01
−2.0	82.33	83.79	$\boxed{84.16}$	84.76	81.92	82.64	82.54	82.46	55.30	12.54	$\boxed{14.52}$	23.33
−1.0	83.59	84.44	84.50	84.99	82.23	82.38	82.13	82.07	24.76	23.12	27.72	31.39
0.0	83.92	84.59	84.61	85.00	81.01	81.52	81.39	81.43	53.66	46.20	44.24	43.45
1.0	83.20	84.33	84.51	84.94	78.11	79.67	80.13	80.58	68.96	62.81	57.59	53.38
2.0	80.60	83.49	84.20	84.82	71.70	76.94	78.61	79.59	76.69	73.19	67.54	61.70

large $|\delta|$. The performance of the plug-in average with known δ is best. It even performs uniformly the best at the larger sample size of $T = 1000$.

What causes the superior performance of the infeasible estimators? Our simulation findings can be understood from the intuition that was gained from the simplified DGP. Panel (a) and (b) from Figure 7 show the empirical MSE of the three models, $m \in \{1, 2, 3\}$, together with the AMSE of these models. The asymptotic approximation is close for $T = 100$ and improves further at $T = 1000$. The selection probabilities in panels (c) and (d) reveal how the infeasible estimator takes a binary decision with the lag length increasing with δ. For the simplified DGP we have seen how the convergence in distribution of $\hat{\delta}$ causes smeared out selection probabilities instead of the binary decision. This effect is also observed in the graphs, *even at the large sample size of* $T = 1000$. The panels (e) and (f) tell the same story for the plug-in weights.

We also perform simulation where we focus on several impulse responses simultaneously, see Remarks 4 and 5. The trace is used to map the AMSE matrix to a scalar. The simulation outcomes are qualitatively similar to our results for the impulse responses at a single horizon. Further details can be found in the Supplementary Material.

3.2.2. Confidence intervals

Confidence intervals/bands can be calculated based on Theorem 4. Simulation results are provided in Tables 1 and 2. The desired nominal coverage level was 90%. Table 1 shows that the empirical coverage of the individual confidence intervals is consistently too low. At horizon 2 and 3 this under-coverage is least severe and decreases with sample size. The coverage of the confidence level for horizon 6 varies strongly across δ and can be very low. It is well-established in the literature (e.g., Kilian (1998) and Kilian

Table 2. The empirical coverage of the 90% confidence regions (see Theorem 4) as a function of sample size T and misspecification parameter δ.

	$h = \{2,3\}$				$h = \{2,6\}$			
δ	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 100$	$T = 250$	$T = 500$	$T = 1000$
−4.0	89.71	89.94	90.01	90.08	87.14	86.47	71.25	9.87
−3.0	89.59	90.03	90.11	90.11	84.15	73.27	41.74	56.54
−2.0	89.64	90.05	90.07	90.02	74.91	56.85	68.57	93.01
−1.0	89.59	90.01	90.01	90.02	76.13	89.96	90.41	90.21
0.0	89.41	89.86	89.89	89.93	83.80	86.91	88.31	89.16
1.0	89.17	89.74	89.77	89.87	85.99	87.91	88.56	89.13
2.0	88.54	89.75	89.79	89.91	87.29	88.83	89.16	89.46

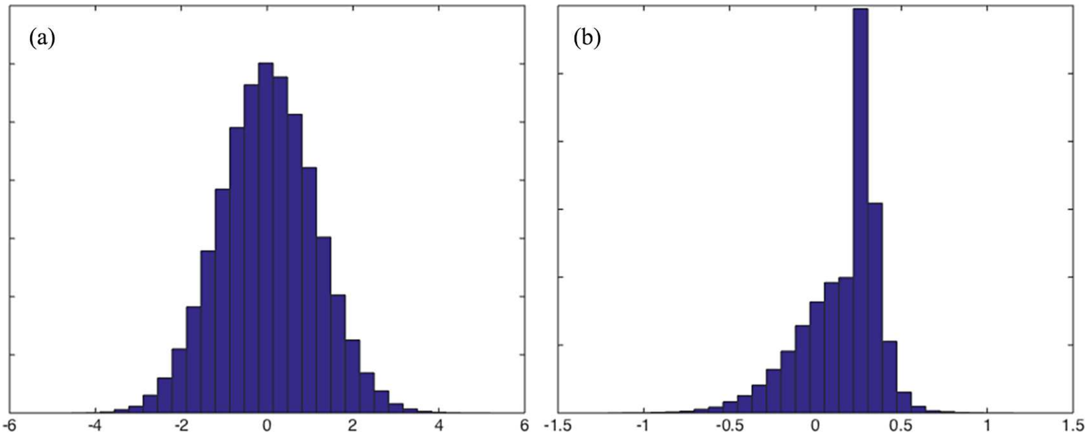


Figure 8. The confidence intervals are based on the asymptotic normality of $\sqrt{T}\left(\bar{\boldsymbol{\mu}}(\hat{\boldsymbol{w}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_T, \boldsymbol{\sigma})\right) - \boldsymbol{D}_\theta \sum_{m=p_1}^{p_1+p_2} \hat{w}_m \hat{\boldsymbol{C}}_m \hat{\boldsymbol{\delta}}$ (see Theorem 4). The displayed histograms are constructed for $y_{T,t} = 0.5y_{T,t-1} + \frac{\delta}{\sqrt{T}} y_{T,t-2} + \frac{\delta}{2\sqrt{T}} y_{T,t-3} + u_t$ with $\delta = -2$ and $T = 500$, i.e., the boxed entries in Table 1. The number of Monte Carlo replications is 100, 000. (a) $h = 2, T = 500$ and (b) $h = 6, T = 500$.

(2001)) that inference on impulse responses at higher horizons is inherently more difficult because of the increased nonlinearity in the parameters. This nonlinearity causes the delta method approximation to perform poorly. Figure 8 shows the histograms of $\sqrt{T}\Big(\bar{\boldsymbol{\mu}}(\hat{\boldsymbol{w}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_T, \boldsymbol{\sigma})\Big) - \boldsymbol{D}_\theta \sum_{m=p_1}^{p_1+p_2} \hat{w}_m \hat{\boldsymbol{C}}_m \hat{\boldsymbol{\delta}}$ for the impulse responses at horizons 2 and 6 (corresponding to the boxed numbers in Table 1). Note that the confidence intervals/bands defined in Theorem 4 are based on the asymptotic normality of this expression. The sometimes severe under-coverage at horizon 6 should therefore not come as a surprise. This poor asymptotic approximation at horizon 6 also influences the empirical coverage of the confidence bands as can be seen in Table 2.

3.3. Simulation results for a vector autoregressive model

Our simulation results are based on a bivariate VAR with DGP

$$\begin{aligned} \boldsymbol{y}_{T,t} &= \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0.5 \end{pmatrix} \boldsymbol{y}_{T,t-1} + \frac{\delta}{\sqrt{T}} \begin{pmatrix} 1 & 0 \\ 0.5 & 1 \end{pmatrix} \boldsymbol{y}_{T,t-2} + \frac{\delta}{2\sqrt{T}} \begin{pmatrix} 1 & 0 \\ 0.5 & 1 \end{pmatrix} \boldsymbol{y}_{T,t-3} + \boldsymbol{u}_t, \\ \boldsymbol{u}_t &\overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_u), \text{ where } \boldsymbol{\Sigma}_u = \begin{pmatrix} 1 & 0.17 \\ 0.17 & 0.33 \end{pmatrix}, \end{aligned} \tag{3.2}$$

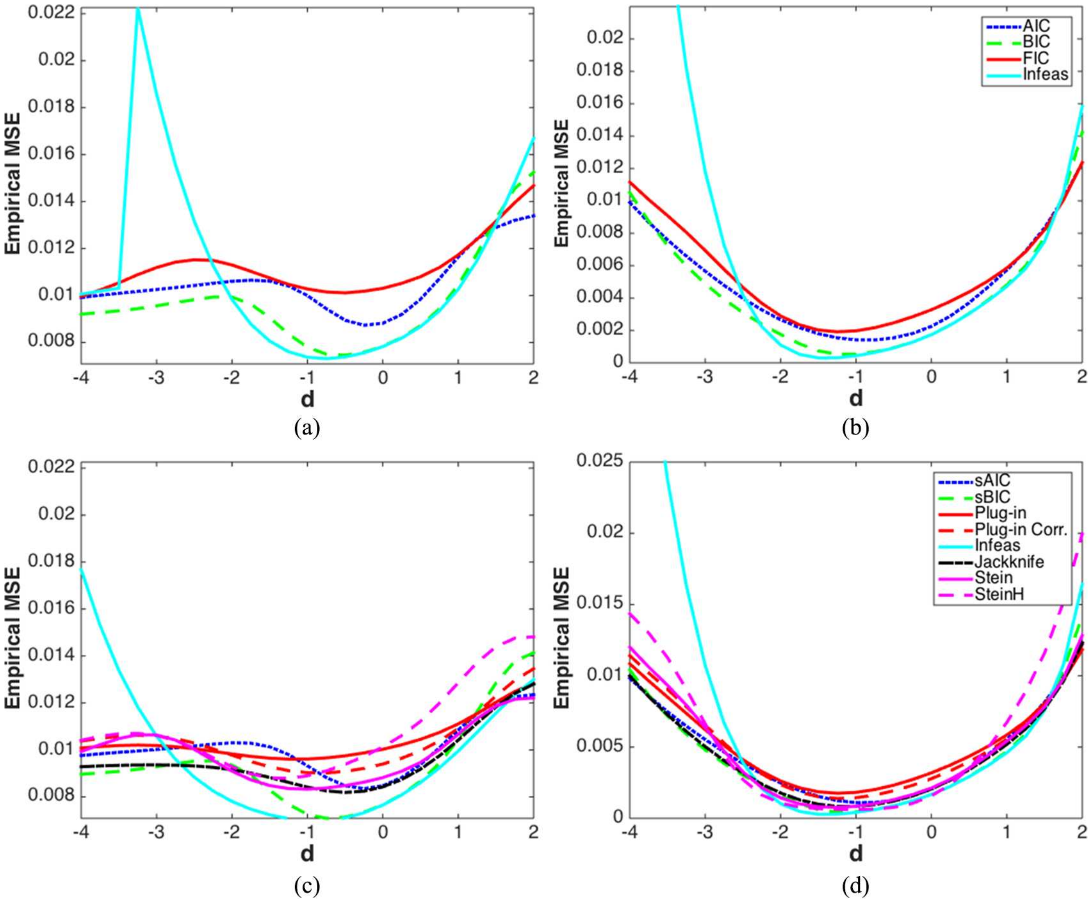


Figure 9. The empirical MSE of the impulse response estimator for several selection and averaging methods. We have displayed the results for the response of variable 1 to a structural shock in the variable 1 for horizons 2 and 6. The DGP is given in Equation (3.2). The sample size is $T = 100$. (a) $h = 2$, model selection, (b) $h = 6$, model selection, (c) $h = 2$, model averaging, and (d) $h = 6$, model averaging.

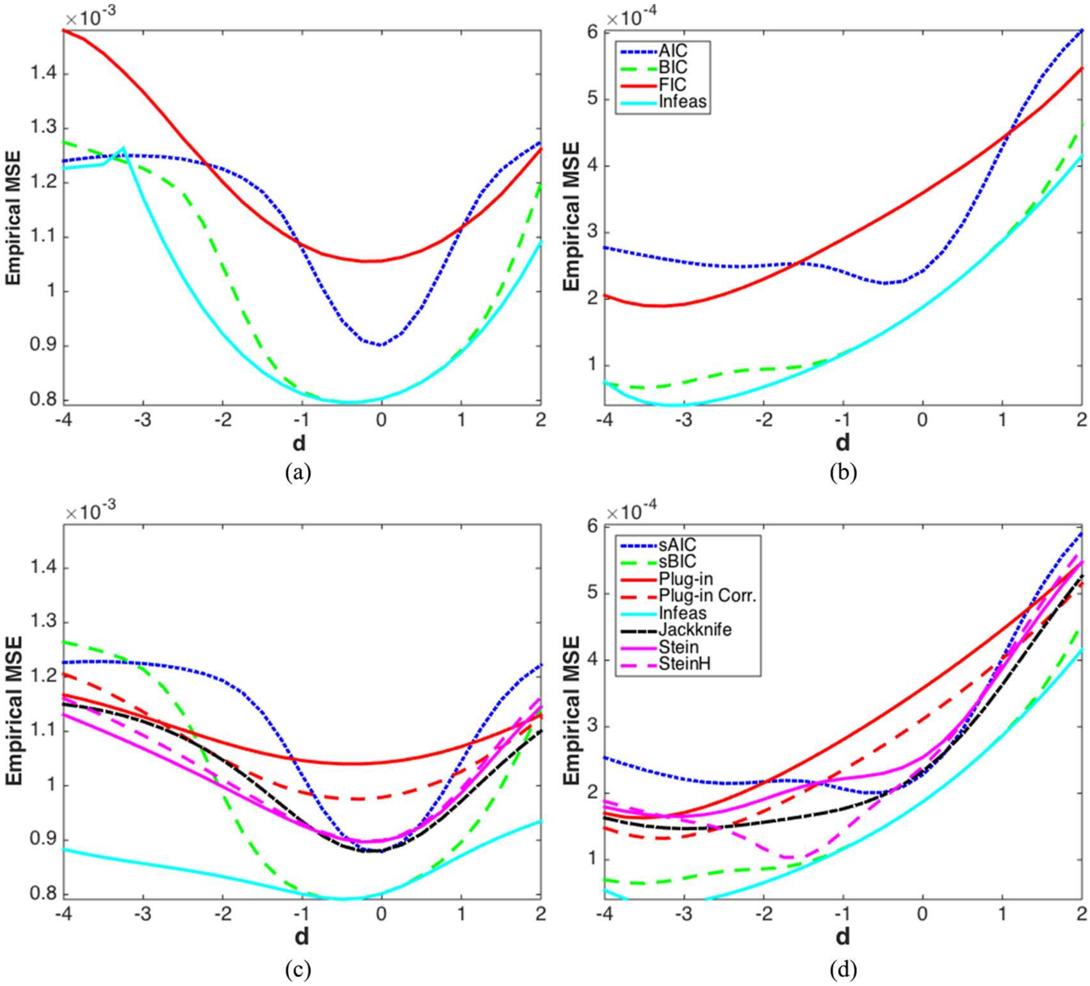


Figure 10. Identical to Figure 9, but for $T = 1000$. (a) $h = 2$, model selection, (b) $h = 6$, model selection, (c) $h = 2$, model averaging, and (d) $h = 6$, model averaging.

which is similar to the VAR used in Lütkepohl et al. (2015) for impulse response analyses. This process has the same roots as the univariate process of Equation (3.1) but with double multiplicity. The parameter δ governs the degree of misspecification. For brevity, we only report MSE results of the response of variable 1 to a structural shock in variable 1. Figures 9, and 10 show the results for horizons 2 and 6.[18] Similarly to the univariate results, none of the methods performs uniformly best. Only the infeasible methods get close to dominating all other methods for the large sample size of $T = 1000$. The ragged spike for "**Infeas**" in Figure 9(a) is caused by an abrupt binary decision to switch between models with different lag lengths. Finally, it is interesting to note that the Stein shrinkage methods perform well in comparison to the plug-in averaging procedure.

4. Conclusion

In this paper, we studied the issue of model selection and model averaging for multivariate autoregressive processes in a locally drifting asymptotic framework. Within this drifting framework we derived the

[18] The simulation results for all four impulse responses and horizons 1–6 can be found in the Supplementary Material.

asymptotic normality of the least squares estimators. The multivariate delta method subsequently ensured that this asymptotic normality carries over to sufficiently smooth parameter transformations, e.g., impulse responses. The focused information criterion and plug-in averaging estimator were defined as the minimizers of the estimated asymptotic mean squared error of the focus parameter estimator.

We highlighted the role of the misspecification parameter $\boldsymbol{\delta}$. Both Liu (2015) and DiTraglia (2016) mentioned that the feasible FIC remains random in the limit. We provided the explicit expressions for the limiting distribution of the FIC values and the elements of the weighting matrices, and illustrated that the feasible estimators do not truly minimize the asymptotic mean squared error. This latter result might encourage further research into different ways to deal with the misspecification parameter. There are to the best of our knowledge two alternatives reported in the literature. The recent paper by Kitagawa and Muris (2016) adopts a mixed frequentist and Bayesian framework to alleviate the estimation of $\boldsymbol{\delta}$ in their study of model averaging in semiparametric estimation of treatment effects. Hansen (2016) similarly adopts a local-to-zero framework but minimizes a risk quantity that does not require the direct estimation of $\boldsymbol{\delta}$.

Our simulation study of univariate and multivariate autoregressive processes underlined the previous paragraph because the infeasible estimator (the estimator knowing $\boldsymbol{\delta}$) frequently dominated the other methods. The latter was especially the case at the larger sample size of $T = 1000$. There was no clearly preferred method for *feasible* model selection/averaging.

A possible extension of this work is an application to forecasting. Such an extension would complement: (1) the predictive static regression setup discussed in Liu and Kuo (2016), and (2) the prediction focused model selection of autoregressive models in Claeskens et al. (2007). Forecasts for autoregressive models often start from the assumption that estimation and prediction are applied to two independent processes with the same stochastic structure. The link to this current paper is that (under this independence assumption) the asymptotic covariance matrix of the forecast is a continuous transformation of the autoregressive parameters, see Section 3.5 of Lütkepohl (2005) for further details.

Appendix

Mathematical proofs

Proof of Theorem 1. As in Liu (2015) we first relate the parameters of the VAR(m) models to the parameters of the VAR(p). With the aid of the selection matrices we have

$$\begin{aligned}
\hat{\boldsymbol{\Theta}}_{T,m} &= \boldsymbol{Y}_T\boldsymbol{Z}'_{T,m}(\boldsymbol{Z}_{T,m}\boldsymbol{Z}'_{T,m})^{-1} = (\boldsymbol{B}\boldsymbol{L}'\boldsymbol{Z}_T + \boldsymbol{C}_T\boldsymbol{S}'_0\boldsymbol{Z}_T + \boldsymbol{U})\boldsymbol{Z}'_{T,m}(\boldsymbol{Z}_{T,m}\boldsymbol{Z}'_{T,m})^{-1}\\
&= ([\boldsymbol{B}\ \ \boldsymbol{C}_T]\boldsymbol{S}_m\boldsymbol{S}'_m\boldsymbol{Z}_T + \boldsymbol{C}_T(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m)\boldsymbol{S}'_0\boldsymbol{Z}_T + \boldsymbol{U})\boldsymbol{Z}'_{T,m}(\boldsymbol{Z}_{T,m}\boldsymbol{Z}'_{T,m})^{-1}\\
&= (\boldsymbol{\Theta}_{T,m}\boldsymbol{Z}_{T,m} + \boldsymbol{C}_T(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m)\boldsymbol{S}'_0\boldsymbol{Z}_T + \boldsymbol{U})\boldsymbol{Z}'_{T,m}(\boldsymbol{Z}_{T,m}\boldsymbol{Z}'_{T,m})^{-1}\\
&= \boldsymbol{\Theta}_{T,m} + \boldsymbol{C}_T(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m)\boldsymbol{S}'_0\boldsymbol{Z}_T\boldsymbol{Z}'_{T,m}(\boldsymbol{Z}_{T,m}\boldsymbol{Z}'_{T,m})^{-1} + \boldsymbol{U}\boldsymbol{Z}'_{T,m}(\boldsymbol{Z}_{T,m}\boldsymbol{Z}'_{T,m})^{-1}. \qquad \text{(A.1)}
\end{aligned}$$

We rearrange terms to obtain the starting point of our analysis. Note especially how the scaling by $T^{1/2}$ cancels against the $T^{-1/2}$ decay rate of the elements in the matrix $\boldsymbol{C}_T$. See Remark 1 for a discussion. An expression in terms of the fixed parameter matrix $\boldsymbol{\Delta}$ remains:

$$\begin{aligned}
\sqrt{T}\left(\hat{\boldsymbol{\Theta}}_{T,m} - \boldsymbol{\Theta}_{T,m}\right) = {}& \underbrace{\sqrt{T}\boldsymbol{C}_T}_{\boldsymbol{\Delta}}\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{S}'_0\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\left[\boldsymbol{S}'_m\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\right]^{-1}\\
&+\left(\frac{1}{\sqrt{T}}\boldsymbol{U}\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\left[\boldsymbol{S}'_m\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T\right)\boldsymbol{S}_m\right]^{-1}. \qquad \text{(A.2)}
\end{aligned}$$

If we define the vector of parameter estimates as $\hat{\boldsymbol{\theta}}_{T,m} = \operatorname{vec}(\hat{\boldsymbol{\Theta}}_{T,m})$ and the true parameter vector $\boldsymbol{\theta}_{T,m} = \operatorname{vec}(\boldsymbol{\Theta}_{T,m})$, then the properties of the vec operator provide

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,m} - \boldsymbol{\theta}_{T,m}\right) = \left(\left[\boldsymbol{S}_m'\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}_T'\right)\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}_m'\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}_T'\right)\boldsymbol{S}_0\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right) \otimes \boldsymbol{I}_K\right)\boldsymbol{\delta}$$

$$+ \left(\left[\boldsymbol{S}_m'\left(\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}_T'\right)\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}_m' \otimes \boldsymbol{I}_K\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\operatorname{vec}\left(\boldsymbol{u}_t\boldsymbol{z}_{T,t-1}'\right), \qquad \text{(A.3)}$$

where $\boldsymbol{\delta} = \operatorname{vec}(\boldsymbol{\Delta})$. We will prove both $\operatorname{plim}_{T\to\infty}\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}_T' = \operatorname{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_{T,t-1}\boldsymbol{z}_{T,t-1}' = \boldsymbol{\Omega}$ and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\operatorname{vec}(\boldsymbol{u}_t\boldsymbol{z}_{T,t-1}') \xrightarrow{d} \mathrm{N}\,(\boldsymbol{0}, \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma})$. These two results prove part (a) of Theorem 1, because the continuous mapping theorem implies that $\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,m} - \boldsymbol{\theta}_{T,m}) = \boldsymbol{A}_m\boldsymbol{\delta} + ([\boldsymbol{S}_m'\boldsymbol{\Omega}\boldsymbol{S}_m]^{-1}\boldsymbol{S}_m' \otimes \boldsymbol{I}_K)\boldsymbol{R} + o_p(1)$.

We start with the proof of $\operatorname{plim}_{T\to\infty}\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}_T' = \boldsymbol{\Omega}$. The process $\{\boldsymbol{y}_{T,t}\}_{t=-\infty}^{\infty}$ is stationary and ergodic for every fixed T in view of Assumptions 1 and 2 (e.g., Theorem 3 on p. 204 of Hannan (1970)). Define the companion matrix $\boldsymbol{A}_T$ and innovation vector $\boldsymbol{E}_t$ such that $\boldsymbol{z}_{T,t} = \boldsymbol{A}_T\boldsymbol{z}_{T,t-1} + \boldsymbol{E}_t$, i.e.,

$$\boldsymbol{A}_T := \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_2 & \cdots & \boldsymbol{B}_{p_1-1} & \boldsymbol{B}_{p_1} & \frac{\boldsymbol{C}_1}{\sqrt{T}} & \frac{\boldsymbol{C}_2}{\sqrt{T}} & \cdots & \frac{\boldsymbol{C}_{p_2-1}}{\sqrt{T}} & \frac{\boldsymbol{C}_{p_2}}{\sqrt{T}} \\ \boldsymbol{I}_K & \mathbf{O} & \cdots & \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \cdots & \mathbf{O} & \mathbf{O} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \cdots & \boldsymbol{I}_K & \mathbf{O} \end{bmatrix} \qquad (Kp \times Kp), \qquad \text{(A.4)}$$

$$\boldsymbol{E}_t := \left(\boldsymbol{u}_t', \boldsymbol{0}', \ldots, \boldsymbol{0}'\right)' \qquad (Kp \times 1).$$

From this extended VAR(1) form we conclude that $\boldsymbol{z}_{T,t}\boldsymbol{z}_{T,t}' = \boldsymbol{A}_T\boldsymbol{z}_{T,t-1}\boldsymbol{z}_{T,t-1}'\boldsymbol{A}_T' + \boldsymbol{A}\boldsymbol{z}_{T,t-1}\boldsymbol{E}_t' + \boldsymbol{E}_{t-1}\boldsymbol{z}_{T,t-1}'\boldsymbol{A}_T' + \boldsymbol{E}_t\boldsymbol{E}_t'$. Stationarity implies

$$\lim_{T\to\infty}\left(\boldsymbol{I}_{K^2p^2} - \boldsymbol{A}_T \otimes \boldsymbol{A}_T\right)\operatorname{plim}\operatorname{vec}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_{T,t-1}\boldsymbol{z}_{T,t-1}'\right)$$

$$= \operatorname{plim}\operatorname{vec}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}_T\boldsymbol{z}_{T,t-1}\boldsymbol{E}_t'\right) + \operatorname{plim}\operatorname{vec}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{E}_t\boldsymbol{z}_{T,t-1}'\boldsymbol{A}_T'\right) + \operatorname{plim}\operatorname{vec}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{E}_t\boldsymbol{E}_t'\right). \qquad \text{(A.5)}$$

The nonrandom matrix $\boldsymbol{A}_T$ will converge to the matrix $\boldsymbol{A}_\infty$ for large T. $\boldsymbol{A}_\infty$ is thus obtained from $\boldsymbol{A}_T$ by replacing the ratios $\boldsymbol{C}_i/\sqrt{T}$ with zero matrices for $i \in \{1, 2, \ldots, p_2\}$. Note that the eigenvalues of $\boldsymbol{A}_\infty$ coincide with the roots of the matrix polynomial $\boldsymbol{B}_\infty(z)$ augmented with Kp_2 additional zero eigenvalues. Assumption 3 guarantees that the matrix $\boldsymbol{I}_{p^2} - \boldsymbol{A}_\infty \otimes \boldsymbol{A}_\infty$ is invertible.

Subsequently we consider the RHS of Equation (A.5). Let $y_{T,t-j,k}$ and $u_{t,k}$ denote the kth component of $\boldsymbol{y}_{T,t-j}$ and $\boldsymbol{u}_t$, respectively. If we can show that $\frac{1}{T}\sum_{t=1}^{T} y_{T,t-j,k}u_{t,l} \xrightarrow{p} 0$ for all $j \in \{1, 2, \ldots, p\}$ and $k, l \in \{1, 2, \ldots, K\}$, then the first two terms in the RHS of Equation (A.5) are $o_p(1)$. To prove this, we define the array $X_{T,t}^{jkl} = y_{T,t-j,k}u_{t,l}/T$ and the norming $c_T = 1/T$. $X_{T,t}^{jkl}$ is a martingale difference (m.d.) array with respect to the filtration $\mathcal{F}_t = \sigma(\boldsymbol{u}_s, -\infty < s \leq t)$ and $\mathrm{E}|X_{T,t}^{jkl}/c_T|^4 = \mathrm{E}|y_{T,t-j,k}u_{t,l}|^4$ is finite in view of Assumption 1. Result 12.10 from Davidson (1994) implies that $|X_{T,t}^{jkl}|^2$ is uniformly integrable and Result 19.7 from the same reference gives $\frac{1}{T}\sum_{t=1}^{T} y_{T,t-j,k}u_t \overset{L_2}{\to} 0$. The result for the first two terms follows. The third term in the RHS of Equation (A.5) is a sample mean of an i.i.d. sequence. Khinchine's

Theorem gives the probability limit. Combining all the results, we conclude that

$$\operatorname{vec}(\boldsymbol{\Omega}) := \operatorname{plim}_{T\to\infty}\operatorname{vec}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1}\right) = \left(\boldsymbol{I}_{K^2p^2} - \boldsymbol{A}_\infty \otimes \boldsymbol{A}_\infty\right)^{-1}\operatorname{vec}\left(\boldsymbol{\Sigma}^*\right) + o_p(1), \quad \text{(A.6)}$$

where $\boldsymbol{\Sigma}^* = \boldsymbol{e}\boldsymbol{\Sigma}\boldsymbol{e}'$ and $\boldsymbol{e}$ is the $(Kp \times K)$ matrix composed of the first K column of I_{Kp}. This shows that $\operatorname{plim}_{T\to\infty}\frac{1}{T}\boldsymbol{Z}_T\boldsymbol{Z}'_T = \boldsymbol{\Omega}$. Ergodicity for every T also provides the result $(\boldsymbol{I}_{K^2p^2} - \boldsymbol{A}_T \otimes \boldsymbol{A}_T)\operatorname{vec}(\mathrm{E}(\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1})) = \operatorname{vec}(\boldsymbol{\Sigma})$ and hence $\boldsymbol{\Omega} = \lim_{T\to\infty}\mathrm{E}(\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1})$ because $\boldsymbol{A}_T \to \boldsymbol{A}_\infty$.

We rely on the Cramer-Wold theorem (e.g., Result 25.5 from Davidson (1994)) to prove the convergence of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\operatorname{vec}(\boldsymbol{u}_t\boldsymbol{z}'_{T,t-1})$ to $\boldsymbol{R} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega}\otimes\boldsymbol{\Sigma})$. Let $\boldsymbol{\xi}$ denote a fixed $(K^2p \times 1)$ vector. $X^*_{T,t} = \boldsymbol{\xi}'\operatorname{vec}(\boldsymbol{u}_t\boldsymbol{z}'_{T,t-1})$ is a m.d. array with respect to $\mathcal{F}_t$. We note that $\sigma^2_{Tt} = \mathrm{E}(X^{*2}_{T,t}|\mathcal{F}_{t-1}) = \mathrm{E}(\boldsymbol{\xi}'\operatorname{vec}(\boldsymbol{u}_T\boldsymbol{z}'_{T,t-1})\operatorname{vec}(\boldsymbol{u}_T\boldsymbol{z}'_{T,t-1})'\boldsymbol{\xi}|\mathcal{F}_{t-1}) = \mathrm{E}(\boldsymbol{\xi}'((\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1})\otimes(\boldsymbol{u}_t\boldsymbol{u}_t))\boldsymbol{\xi}|\mathcal{F}_{t-1}) = \boldsymbol{\xi}'((\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1})\otimes \boldsymbol{\Sigma}\ )\boldsymbol{\xi}$ and that $X^{*2}_{T,t}$ is square integrable by Assumption 1. Moreover, from $s^2_T = \sum_{t=1}^{T}\mathrm{E}(X^{*2}_{T,t}) = \boldsymbol{\xi}'(\mathrm{E}(\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1})\otimes\boldsymbol{\Sigma})\boldsymbol{\xi}$ we have

$$\sup_T \frac{T}{s^2_T} = \sup_T \frac{1}{\boldsymbol{\xi}'\left(\mathrm{E}\left(\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1}\right)\otimes\boldsymbol{\Sigma}\right)\boldsymbol{\xi}} < \infty. \quad \text{(A.7)}$$

Equation (A.7) holds because the quadratic form in the denominator cannot be zero as both $\boldsymbol{\Sigma}$ and $\mathrm{E}(\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1})$ are positive definite matrices. The positive definiteness of $\boldsymbol{\Sigma}$ is part of Assumption 4. For finite T, $\mathrm{E}(\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1})$ cannot be positive semidefinite because this would imply the existence of a $\boldsymbol{\kappa} \neq \boldsymbol{0}$ such that $\boldsymbol{\kappa}'\mathrm{E}(\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1})\boldsymbol{\kappa} = \mathrm{E}(\boldsymbol{\kappa}'\boldsymbol{z}_{T,t-1})^2 = 0$ and at least one component of $z_{T,t-1}$ is zero for all t. Also it cannot approach a positive semidefinite matrix due to convergence to $\boldsymbol{\Omega}$. A generalization of Result 24.4 from Davidson (1994) to martingale difference arrays shows that

$$\frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\xi}'\operatorname{vec}\left(\boldsymbol{u}_t\boldsymbol{z}'_{T,t-1}\right)}{\sqrt{\boldsymbol{\xi}'\left(\mathrm{E}\left(\boldsymbol{z}_{T,t-1}\boldsymbol{z}'_{T,t-1}\right)\otimes\boldsymbol{\Sigma}\right)\boldsymbol{\xi}}} \xrightarrow{d} \mathrm{N}(0,1). \quad \text{(A.8)}$$

The expression under the square root is asymptotically equivalent to $\boldsymbol{\xi}'(\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma})\boldsymbol{\xi}$. The second result, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\operatorname{vec}(\boldsymbol{u}_t\boldsymbol{z}'_{T,t-1}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega}\otimes\boldsymbol{\Sigma})$, follows because $\boldsymbol{\xi}$ is arbitrary. The proof of part (a) is complete.

Part (b) of Theorem 1 is a joint convergence result with the estimator for the covariance matrix. For any model with $m \in \{p_1, p_1+1, \ldots, p\}$ lags we define the residual matrix (residuals are stacked columnwise) by $\hat{\boldsymbol{U}}^m_{T,t} = \boldsymbol{Y}_T - \hat{\boldsymbol{B}}_{T,m}\boldsymbol{Z}_{T,m}$. The estimated covariance matrix based on the residuals from the model with m lags satisfies

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}^m_u &= \frac{1}{T}\left(\boldsymbol{Y}_T - \hat{\boldsymbol{\Theta}}_{T,m}\boldsymbol{Z}_{T,m}\right)\left(\boldsymbol{Y}_T - \hat{\boldsymbol{\Theta}}_{T,m}\boldsymbol{Z}_{T,m}\right)' \\
&= \frac{1}{T}\left[\left(\boldsymbol{\Theta}_{T,m} - \hat{\boldsymbol{\Theta}}_{T,m}\right)\boldsymbol{S}'_m\boldsymbol{Z}_T + \boldsymbol{C}_T\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{S}'_0\boldsymbol{Z}_T + \boldsymbol{U}\right] \\
&\qquad \left[\left(\boldsymbol{\Theta}_{T,m} - \hat{\boldsymbol{\Theta}}_{T,m}\right)\boldsymbol{S}'_m\boldsymbol{Z}_T + \boldsymbol{C}_T\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{S}'_0\boldsymbol{Z}_T + \boldsymbol{U}\right]' \\
&= \left(\boldsymbol{\Theta}_{T,m} - \hat{\boldsymbol{\Theta}}_{T,m}\right)\boldsymbol{S}'_m\left(\frac{\boldsymbol{Z}_T\boldsymbol{Z}'_T}{T}\right)\boldsymbol{S}_m\left(\boldsymbol{\Theta}_{T,m} - \hat{\boldsymbol{\Theta}}_{T,m}\right)' \\
&\quad + \boldsymbol{C}_T\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)\boldsymbol{S}'_0\left(\frac{\boldsymbol{Z}_T\boldsymbol{Z}'_T}{T}\right)\boldsymbol{S}_0\left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m\boldsymbol{\Pi}_m\right)'\boldsymbol{C}'_T
\end{aligned}$$

$$
\begin{aligned}
&+\left(\boldsymbol{\Theta}_{T,m}-\hat{\boldsymbol{\Theta}}_{T,m}\right)\boldsymbol{S}_m'\left(\frac{\boldsymbol{Z}_T\boldsymbol{Z}_T'}{T}\right)\boldsymbol{S}_0\left(\boldsymbol{I}_{Kp_2}-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)'\boldsymbol{C}_T'+\left(\boldsymbol{\Theta}_{T,m}-\hat{\boldsymbol{\Theta}}_{T,m}\right)\boldsymbol{S}_m'\left(\frac{\boldsymbol{Z}_T\boldsymbol{U}'}{T}\right)\\
&+\boldsymbol{C}_T\left(\boldsymbol{I}_{Kp_2}-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{S}_0'\left(\frac{\boldsymbol{Z}_T\boldsymbol{Z}_T'}{T}\right)\boldsymbol{S}_m\left(\boldsymbol{\Theta}_{T,m}-\hat{\boldsymbol{\Theta}}_{T,m}\right)'+\boldsymbol{C}_T\left(\boldsymbol{I}_{Kp_2}-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)\boldsymbol{S}_0'\left(\frac{\boldsymbol{Z}_T\boldsymbol{U}'}{T}\right)\\
&+\left(\frac{\boldsymbol{U}\boldsymbol{Z}_T'}{T}\right)\boldsymbol{S}_m\left(\boldsymbol{\Theta}_{T,m}-\hat{\boldsymbol{\Theta}}_{T,m}\right)+\left(\frac{\boldsymbol{U}\boldsymbol{Z}_T}{T}\right)\boldsymbol{S}_0\left(\boldsymbol{I}_{Kp_2}-\boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m\right)'\boldsymbol{C}_T'+\frac{1}{T}\boldsymbol{U}\boldsymbol{U}'. \qquad \text{(A.9)}
\end{aligned}
$$

The stochastic orders of the various terms in Equation (A.9) are known from previous results. We have $\boldsymbol{\Theta}_{T,m}-\hat{\boldsymbol{\Theta}}_{T,m}=O_p(T^{-1/2})$, $\boldsymbol{Z}_T\boldsymbol{Z}_T'/T\xrightarrow{p}\boldsymbol{\Omega}$, $\boldsymbol{C}_T=O(T^{-1/2})$ and $\boldsymbol{Z}_T\boldsymbol{U}'/T=O_p(T^{-1/2})$ by Equation (A.8). We conclude that $\hat{\boldsymbol{\Sigma}}_u^m=\frac{1}{T}\boldsymbol{U}\boldsymbol{U}'+o_P(1)$. Every covariance estimator (every in the sense of for all $m\in\mathcal{M}$) has therefore the same asymptotic distribution as the covariance estimator based on the true innovations.

Joint asymptotic normality of the parameter estimates and the covariance estimator can be obtained along the lines of the proof of Proposition 11.2 of Hamilton (1994). That is, we define

$$
\boldsymbol{\lambda}_t=\operatorname{vech}\begin{bmatrix} u_{1t}^2-\sigma_{11} & u_{1t}u_{2t}-\sigma_{12} & \dots & u_{1t}u_{Kt}-\sigma_{1K}\\ u_{2t}u_{1t}-\sigma_{21} & u_{2t}^2-\sigma_{22} & \dots & u_{2t}u_{Kt}-\sigma_{2K}\\ \vdots & \vdots & \ddots & \vdots\\ u_{Kt}u_{1t}-\sigma_{K1} & u_{Kt}u_{2t}-\sigma_{K2} & \dots & u_{Kt}^2-\sigma_{KK}\end{bmatrix}. \qquad \text{(A.10)}
$$

The sequence $\{\boldsymbol{\lambda}_t\}$ is i.i.d. and thus also a martingale difference sequence. One can apply the Cramer-Wold Theorem to the extended martingale difference vector $(\operatorname{vec}(\boldsymbol{u}_t\boldsymbol{z}_{T,t-1}')',\boldsymbol{\lambda}_t')'$ to show

$$
\begin{bmatrix}(1\sqrt{T})\sum_{t=1}^T\operatorname{vec}(\boldsymbol{u}_t\boldsymbol{z}_{T,t-1}')\\ (1\sqrt{T})\sum_{t=1}^T\boldsymbol{\lambda}_t\end{bmatrix}\xrightarrow{d}\mathrm{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Xi}_{11} & \boldsymbol{\Xi}_{12}\\ \boldsymbol{\Xi}_{21} & \boldsymbol{\Xi}_{22}\end{bmatrix}\right). \qquad \text{(A.11)}
$$

We already know that $\boldsymbol{\Xi}_{11}=\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma}$. The elements in the covariance matrix $\boldsymbol{\Xi}_{12}$ take the form $\lim_{T\to\infty}\mathrm{E}(u_{k_1t}y_{T,t-j,k_2}(u_{k_3t}u_{k_4t}-\sigma_{k_3k_4}))$. They are zero because $\lim_{T\to\infty}\mathrm{E}(y_{T,t-j,k_2})=0$. Finally, $\boldsymbol{\Xi}_{22}=\mathrm{E}\left(\boldsymbol{\lambda}_t\boldsymbol{\lambda}_t'\right)$. The typical elements are $\mathrm{E}((u_{it}u_{jt}-\sigma_{ij})(u_{lt}u_{mt}-\sigma_{lm}))$.

Define two independent random vectors: $\boldsymbol{R}\sim\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma}\right)$ and $\boldsymbol{S}\sim\mathrm{N}(\mathbf{0},\boldsymbol{\Xi}_{22})$. We will proof the claim in part (c) of Theorem 1 for the case of three different models indexed by $m_1,m_2,m_3\in\mathcal{M}$. The proof is immediate, since

$$
\begin{bmatrix}\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,m_1}-\boldsymbol{\theta}_{T,m_1}\right)\\ \sqrt{T}\left(\hat{\boldsymbol{\theta}}_{T,m_2}-\boldsymbol{\theta}_{T,m_2}\right)\\ \sqrt{T}\operatorname{vech}\left(\hat{\boldsymbol{\Sigma}}_u^{m_3}-\boldsymbol{\Sigma}_u\right)\end{bmatrix}\xrightarrow{d}\begin{bmatrix}\boldsymbol{A}_{m_1}\\ \boldsymbol{A}_{m_2}\\ \mathbf{O}\end{bmatrix}\delta+\begin{bmatrix}\left(\left[\boldsymbol{S}_{m_1}'\boldsymbol{\Omega}\boldsymbol{S}_{m_1}\right]^{-1}\boldsymbol{S}_{m_1}'\otimes\boldsymbol{I}_K\right) & \mathbf{O}\\ \left(\left[\boldsymbol{S}_{m_2}'\boldsymbol{\Omega}\boldsymbol{S}_{m_2}\right]^{-1}\boldsymbol{S}_{m_2}'\otimes\boldsymbol{I}_K\right) & \mathbf{O}\\ \mathbf{O} & \boldsymbol{I}_{K(K+1)/2}\end{bmatrix}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{S}\end{bmatrix}. \qquad \text{(A.12)}
$$

Proof of the Auxiliary Result. The proof uses mathematical induction so let us compare the impulse responses of the VAR(p) and VAR($p+1$) models. For $p=0$ we are comparing a white noise model with a VAR(1) with coefficient matrix $\boldsymbol{A}$. The impulse responses at horizon h (the case $h=1$ is trivial so we focus on $h>1$) for these models are $\mathbf{O}_{K\times K}$ and $\boldsymbol{A}^h$, respectively. The base case $p=0$ holds.

We start the inductive step by defining the companion matrix of the VAR($p+1$),

$$
\boldsymbol{F}_{(p+1)}=\begin{bmatrix}\boldsymbol{A}_1 & \boldsymbol{A}_2 & \dots & \boldsymbol{A}_p & \boldsymbol{A}_{p+1}\\ \boldsymbol{I}_K & \mathbf{O} & \dots & \mathbf{O} & \mathbf{O}\\ \mathbf{O} & \boldsymbol{I}_K & \dots & \mathbf{O} & \mathbf{O}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \mathbf{O} & \mathbf{O} & \dots & \boldsymbol{I}_K & \mathbf{O}\end{bmatrix}. \qquad \text{(A.13)}
$$

This companion matrix is $(K(p+1)\times K(p+1))$. For an arbitrary matrix of this size, say $\boldsymbol{A}$, we will introduce the notation $[\boldsymbol{A}]^{ij}$ to denote its (i,j)th block of dimension $(K\times K)$. In this notation, the impulse response at horizon h for the VAR$(p+1)$ is simply $[(\boldsymbol{F}_{p+1})^h]^{11}$. Setting $\boldsymbol{A}_{p+1} = \mathbf{O}$ provides K zero columns, and hence

$$\begin{aligned}[(\boldsymbol{F}_{p+1})^h]^{11} &= \sum_{j_1=1}^{p+1}\cdots\sum_{j_{h-1}=1}^{p+1}[\boldsymbol{F}_{p+1}]^{1j_1}[\boldsymbol{F}_{p+1}]^{j_1j_2}\cdots[\boldsymbol{F}_{p+1}]^{j_{h-1}1}\\ &= \sum_{j_1=1}^{p}[\boldsymbol{F}_{p+1}]^{1j_1}\left(\sum_{j_2=1}^{p+1}\cdots\sum_{j_{h-1}=1}^{p+1}[\boldsymbol{F}_{p+1}]^{j_1j_2}\cdots[\boldsymbol{F}_{p+1}]^{j_{h-1}1}\right)\\ &= \ldots = \sum_{j_1=1}^{p}\cdots\sum_{j_{h-1}=1}^{p}[\boldsymbol{F}_{p+1}]^{1j_1}[\boldsymbol{F}_{p+1}]^{j_1j_2}\cdots[\boldsymbol{F}_{p+1}]^{j_{h-1}1} = \left[(\boldsymbol{F}_p)^h\right]^{11}. \qquad \text{(A.14)}\end{aligned}$$

where $\boldsymbol{F}_p$ is the companion matrix related to the VAR(p). The inductive step is also complete.

Proof of Theorem 2. We rewrite the expression in Theorem 2 as

$$\begin{aligned}&\sqrt{T}\left(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{T,m},\mathbf{0}_{K^2(p-m)},\hat{\boldsymbol{\sigma}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_{T,p},\boldsymbol{\sigma})\right)\\ &\quad = \sqrt{T}\left(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{T,m},\mathbf{0}_{K^2(p-m)},\hat{\boldsymbol{\sigma}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_{T,m},\mathbf{0}_{K^2(p-m)},\boldsymbol{\sigma})\right)\\ &\qquad - \sqrt{T}\left(\boldsymbol{\mu}(\boldsymbol{\theta}_{T,p},\boldsymbol{\sigma}) - \boldsymbol{\mu}(\boldsymbol{\theta}_{T,m},\mathbf{0}_{K^2(p-m)},\boldsymbol{\sigma})\right). \qquad \text{(A.15)}\end{aligned}$$

The first term in the RHS of Equation (A.15) contains a parameter transformation of the estimated parameters. The first-order delta method can be applied to this expression because Theorem 2 explicitly assumes the nonvanishing derivatives at the necessary points. The second term is nonrandom. It is the difference of two terms which only differ in locally misspecified coefficients which have been set to zero. A Taylor expansion can handle this second contribution. The result from the delta method is

$$\begin{aligned}&\sqrt{T}\left(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{T,m},\mathbf{0}_{K^2(p-m)},\hat{\boldsymbol{\sigma}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_{T,m},\mathbf{0}_{K^2(p-m)},\boldsymbol{\sigma})\right)\\ &\quad \xrightarrow{d} \left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta}_\infty,\boldsymbol{\sigma})}{\partial\boldsymbol{\theta}'}(\boldsymbol{S}_m\otimes\boldsymbol{I}_K)\right)\left(\boldsymbol{A}_m\boldsymbol{\delta} + \left(\left[\boldsymbol{S}_m'\boldsymbol{\Omega}\boldsymbol{S}_m\right]^{-1}\boldsymbol{S}_m'\otimes\boldsymbol{I}_K\right)\boldsymbol{R}\right) + \left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta}_\infty,\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}'}\right)\mathbf{S}, \qquad \text{(A.16)}\end{aligned}$$

where $\boldsymbol{\theta}_\infty$ denotes the parameter vector $\boldsymbol{\theta}_{T,p}$ but with all the misspecification parameters $\boldsymbol{C}_T$ set to zero. The result of the Taylor expansion is

$$\begin{aligned}&\sqrt{T}\left(\boldsymbol{\mu}(\boldsymbol{\theta}_{T,p},\boldsymbol{\sigma}) - \boldsymbol{\mu}(\boldsymbol{\theta}_{T,m},\mathbf{0}_{K^2(p-m)},\boldsymbol{\sigma})\right)\\ &= \sqrt{T}\boldsymbol{\mu}(\boldsymbol{\theta}_{T,m},\mathbf{0}_{K^2(p-m)},\boldsymbol{\sigma}) + \left(\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta}_\infty,\boldsymbol{\sigma})}{\partial\boldsymbol{\theta}'}(\boldsymbol{S}_0\otimes\boldsymbol{I}_K)\right)\left((\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m)\otimes\boldsymbol{I}_K\right)\boldsymbol{\delta} + O(T^{-1/2})\\ &\quad - \sqrt{T}\boldsymbol{\mu}(\boldsymbol{\theta}_{T,m},\mathbf{0}_{K^2(p-m)},\boldsymbol{\sigma})\\ &= \frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta}_\infty,\boldsymbol{\sigma})}{\partial\boldsymbol{\theta}'}\left(\boldsymbol{S}_0(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}_m'\boldsymbol{\Pi}_m)\otimes\boldsymbol{I}_K\right)\boldsymbol{\delta} + O(T^{-1/2}). \qquad \text{(A.17)}\end{aligned}$$

The notation can be made a little lighter using the definitions in Theorem 2: $\boldsymbol{D}_\theta = \partial\boldsymbol{\mu}(\boldsymbol{\theta}_\infty, \boldsymbol{\sigma})/\partial\boldsymbol{\theta}'$ and $\boldsymbol{D}_\sigma = \partial\boldsymbol{\mu}(\boldsymbol{\theta}_\infty, \boldsymbol{\sigma})/\partial\boldsymbol{\sigma}'$. Equations (A.15), (A.16) and (A.17) combine to

$$\begin{aligned}
&\sqrt{T}\Big(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{T,m}, \mathbf{0}_{p-m}, \hat{\boldsymbol{\sigma}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_{T,p}, \boldsymbol{\sigma})\Big) \\
&\quad \xrightarrow{d} \boldsymbol{D}_\theta \left[\left(\boldsymbol{S}_m \otimes \boldsymbol{I}_K\right) \boldsymbol{A}_m - \boldsymbol{S}_0 \left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m \boldsymbol{\Pi}_m\right) \otimes \boldsymbol{I}_K \right] \boldsymbol{\delta} \\
&\qquad + \boldsymbol{D}_\theta \left(\boldsymbol{S}_m \left[\boldsymbol{S}'_m \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \boldsymbol{S}'_m \otimes \boldsymbol{I}_K \right) \boldsymbol{R} + \boldsymbol{D}_\sigma \boldsymbol{S} \\
&\quad = \boldsymbol{D}_\theta \left[\left(\boldsymbol{S}_m \left[\boldsymbol{S}'_m \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \boldsymbol{S}'_m \boldsymbol{\Omega} - \boldsymbol{I}_{Kp} \right) \boldsymbol{S}_0 \left(\boldsymbol{I}_{Kp_2} - \boldsymbol{\Pi}'_m \boldsymbol{\Pi}_m\right) \otimes \boldsymbol{I}_K \right] \boldsymbol{\delta} \\
&\qquad + \boldsymbol{D}_\theta \left(\boldsymbol{S}_m \left[\boldsymbol{S}'_m \boldsymbol{\Omega} \boldsymbol{S}_m\right]^{-1} \boldsymbol{S}'_m \otimes \boldsymbol{I}_K \right) \boldsymbol{R} + \boldsymbol{D}_\sigma \boldsymbol{S}, \qquad \text{(A.18)}
\end{aligned}$$

where the final line follows from the definition of $\boldsymbol{A}_m$. With the given definitions of $\boldsymbol{C}_m$ and $\boldsymbol{P}_m$ we indeed recover the result stated in Theorem 2,

$$\begin{aligned}
\sqrt{T}\Big(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}_{T,m}, \mathbf{0}_{K^2(p-m)}, \hat{\boldsymbol{\sigma}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_{T,p}, \boldsymbol{\sigma})\Big) &\xrightarrow{d} \boldsymbol{D}_\theta \boldsymbol{C}_m \boldsymbol{\delta} + \boldsymbol{D}_\theta \boldsymbol{P}_m \boldsymbol{R} + \boldsymbol{D}_\sigma \boldsymbol{S} \\
&\sim \mathrm{N}\Big(\boldsymbol{D}_\theta \boldsymbol{C}_m \boldsymbol{\delta}, \boldsymbol{D}_\theta \boldsymbol{P}_m \left(\Omega \otimes \boldsymbol{\Sigma}\right) \boldsymbol{P}_m \boldsymbol{D}'_\theta + \boldsymbol{D}_\sigma \boldsymbol{\Xi}_{22} \boldsymbol{D}'_\sigma\Big). \\
&\qquad \text{(A.19)}
\end{aligned}$$

Proof of Theorem 3. By Equation (A.19),

$$\begin{aligned}
\sqrt{T}\Big(\bar{\mu}(\boldsymbol{w}) - \mu(\boldsymbol{\theta}_{T,p}, \boldsymbol{\sigma})\Big) &= \sum_{m=p_1}^{p} w_m \left[\sqrt{T}\Big(\mu(\hat{\boldsymbol{\theta}}_{T,m}, \mathbf{0}_{K^2(p-m)}, \hat{\boldsymbol{\sigma}}) - \mu(\boldsymbol{\theta}_{T,p}, \boldsymbol{\sigma})\Big) \right] \\
&\xrightarrow{d} \boldsymbol{D}_\theta \sum_{m=p_1}^{p} w_m \boldsymbol{C}_m \boldsymbol{\delta} + \boldsymbol{D}_\theta \sum_{m=p_1}^{p} w_m \boldsymbol{P}_m \boldsymbol{R} + \boldsymbol{D}_\sigma \boldsymbol{S}. \qquad \text{(A.20)}
\end{aligned}$$

The calculation of the mean vector and the asymptotic covariance matrix is straightforward.

Proof of Theorem 4. A valid confidence interval for a scalar focus was derived in Theorem 6 of Liu (2015). We follow the same reasoning. Similar to Equation (A.20), we have

$$\sqrt{T}\Big(\bar{\boldsymbol{\mu}}(\hat{\boldsymbol{w}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_T, \boldsymbol{\sigma})\Big) \xrightarrow{d} \boldsymbol{D}_\theta \sum_{m=p_1}^{p} w_m(\boldsymbol{R}_\delta) \boldsymbol{C}_m \boldsymbol{R}_\delta + \boldsymbol{D}_\theta \left(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{I}_K\right) \boldsymbol{R} + \boldsymbol{D}_\sigma \boldsymbol{S}. \qquad \text{(A.21)}$$

Next, by the convergence of $\hat{\boldsymbol{\delta}}$ to $\boldsymbol{R}_\delta$,

$$\sqrt{T}\Big(\bar{\boldsymbol{\mu}}(\hat{\boldsymbol{w}}) - \boldsymbol{\mu}(\boldsymbol{\theta}_T, \boldsymbol{\sigma})\Big) - \boldsymbol{D}_\theta \sum_{m=p_1}^{p} \hat{w}_m \hat{\boldsymbol{C}}_m \hat{\boldsymbol{\delta}} \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{D}_\theta (\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Sigma}) \boldsymbol{D}'_\theta + \boldsymbol{D}_\sigma \boldsymbol{\Xi}_{22} \boldsymbol{D}'_\sigma\right). \qquad \text{(A.22)}$$

The confidence region is constructed from the standardized quadratic form with population quantities replaced by their consistent estimates.

Proof of Theorem 5. Consider $m \neq p$. We define $\boldsymbol{\alpha}_m = \boldsymbol{C}'_m \boldsymbol{D}'_\theta$, $\mathcal{A} = \boldsymbol{S}'_0 \boldsymbol{\Omega}^{-1} \boldsymbol{S}_0 \otimes \boldsymbol{\Sigma}$, and introduce a standard normally distributed random vector $\boldsymbol{Z}_{K^2 p_2} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I}_{K^2 p_2})$ and random variable $Z \sim \mathrm{N}(0, 1)$.

Now $\boldsymbol{R}_\delta = \boldsymbol{\delta} + \mathcal{A}^{1/2}\boldsymbol{Z}_{K^2p_2}$, and

$$\boldsymbol{R}'_\delta \boldsymbol{C}'_m \boldsymbol{D}'_\theta \boldsymbol{D}_\theta \boldsymbol{C}_m \boldsymbol{R}_\delta = \left(\boldsymbol{\alpha}'_m(\boldsymbol{\delta} + \mathcal{A}^{1/2}\boldsymbol{Z}_{K^2p_2})\right)^2 = \|\boldsymbol{\alpha}'_m\mathcal{A}^{1/2}\|^2 \left(\frac{\boldsymbol{\alpha}'_m\mathcal{A}^{1/2}\boldsymbol{Z}_{K^2p_2}}{\|\boldsymbol{\alpha}'_m\mathcal{A}^{1/2}\|} + \frac{\boldsymbol{\alpha}'_m\boldsymbol{\delta}}{\|\boldsymbol{\alpha}'_m\mathcal{A}^{1/2}\|}\right)^2$$

$$= (\boldsymbol{\alpha}'_m\mathcal{A}\boldsymbol{\alpha}_m)\left(Z + \boldsymbol{\alpha}'_m\boldsymbol{\delta}/\sqrt{\boldsymbol{\alpha}'_m\mathcal{A}\boldsymbol{\alpha}_m}\right)^2, \qquad \text{(A.23)}$$

where $\boldsymbol{\alpha}'_m\mathcal{A}\boldsymbol{\alpha}_m = a_m$. The squared expression has the stated noncentral chi-squared distribution (see Chapter 29 of Johnson et al. (1994) for details and moments). Finally, all quantities in $\widehat{FIC}_p = \boldsymbol{D}_\theta(\hat{\boldsymbol{\Omega}}^{-1} \otimes \hat{\boldsymbol{\Sigma}})\boldsymbol{D}'_\theta + \boldsymbol{D}_\sigma\,\Xi_{22}\boldsymbol{D}'_\sigma$ are estimated consistently.

Proof of Theorem 6. If either $m = p$ and/or $l = p$, then there is no bias contribution and the matrix elements converge in probability. Now consider $m, l \neq p$, then

(a) For $m = l$, the proof is identical to the proof of Theorem 5, hence omitted.

(b) Start by noting that $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}'(\frac{\boldsymbol{A}+\boldsymbol{A}'}{2})\boldsymbol{x}$, this gives

$$\boldsymbol{R}'_\delta\boldsymbol{\alpha}_m\boldsymbol{\alpha}'_l\boldsymbol{R}_\delta = \left(\mathcal{A}^{-1/2}\boldsymbol{\delta} + \boldsymbol{Z}_{K^2p_2}\right)'\left[\mathcal{A}^{1/2}\left(\frac{\boldsymbol{\alpha}_m\boldsymbol{\alpha}'_l + \boldsymbol{\alpha}_l\boldsymbol{\alpha}'_m}{2}\right)\mathcal{A}^{1/2}\right]\left(\mathcal{A}^{-1/2}\boldsymbol{\delta} + \boldsymbol{Z}_{K^2p_2}\right) \qquad \text{(A.24)}$$

We subsequently use the transformation stated in Imhof (1961). The matrix in square brackets is symmetric and has a rank of at most two. The eigenvalue decomposition mentioned in Theorem 6 applies, and therefore

$$\boldsymbol{R}'_\delta\boldsymbol{\alpha}_m\boldsymbol{\alpha}'_l\boldsymbol{R}_\delta = \sum_{i=1}^{2}\lambda_i\left(\boldsymbol{v}'_i\mathcal{A}^{-1/2}\boldsymbol{\delta} + Z_i\right)^2 \sim \sum_{i=1}^{2}\lambda_i\chi^2_{noncentral}\left(1, \left(\boldsymbol{v}'_i\mathcal{A}^{-1/2}\boldsymbol{\delta}\right)^2\right), \qquad \text{(A.25)}$$

where the independence of the Z_i follows from orthonormality of the eigenvectors.

Funding

This research was financially supported by The Netherlands Organisation for Scientific Research (NWO) [Grant number 407-11-042].

References

Abadir, K., Magnus, J. (2002). Notation in econometrics: A proposal for a standard. *The Econometrics Journal* 5:76–90.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In: *Selected Papers of Hirotugu Akaike*. New York: Springer-Verlag, pp. 199–213.

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

Benkwitz, A., Neumann, M., Lütkepohl, H. (2000). Problems related to confidence intervals for impulse responses of autoregressive processes. *Econometric Reviews* 19:69–103.

Berk, K. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics* 2:489–502.

Bruder, S. Wolf, M. (2017). Balanced bootstrap joint confidence bands for structural impulse response functions. *Working Paper*.

Burnham, K. P. Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.

Charkhi, A., Claeskens, G., Hansen, B. (2016). Minimum mean squared error model averaging in likelihood models. *Statistica Sinica* 26:809–840.

Claeskens, G., Croux, C., van Kerckhoven, J. (2007). Prediction-focused model selection for autoregressive models. *Australian & New Zealand Journal of Statistics* 49:359–379.

Claeskens, G., Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association* 98:900–916.

Claeskens, G., Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford: Oxford University Press.

DiTraglia, F. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics* 195:187–208.

Fujiwara, M. (1916). Ueber die obere Schranke des absoluten Betrages der Wurzeln einer algebraischen Gleichung. *Tohoku Mathematica Journal* 10:161–171.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.

Hannan, E. (1970). *Multiple Time Series*. New York: John Wiley & Sons.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory* 21:60–68.

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

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

Hansen, B. E. (2016). Stein combination shrinkage for vector autoregressions. *Working Paper*.

Hansen, B. E., Racine, J. (2012). Jackknife model averaging. *Journal of Econometrics* 167:38–46.

Imhof, J.-P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* 48:419–426.

Ivanov, V., Kilian, L. (2005). A practitioner's guide to lag order selection for VAR impulse response analysis. *Studies in Nonlinear Dynamics & Econometrics* 9:1–36.

Johnson, N., Kotz, S., Balakrishnan, S. (1994). *Continuous Univariate Distributions, Vols. 1–2*. New York: John Wiley & Sons.

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics* 80:218–230.

Kilian, L. (2001). Impulse response analysis in vector autoregressions with unknown lag order. *Journal of Forecasting* 20:161–179.

Kitagawa, T., Muris, C. (2016). Model averaging in semiparametric estimation of treatment effects. *Journal of Econometrics* 193:271–289.

Lewis, R., Reinsel, G. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* 16:393–411.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics* 186:142–159.

Liu, C.-A., Kuo, B.-S. (2016). Model averaging in predictive regressions. *Econometrics Journal* 19:203–231.

Liu, Q., Okui, R. (2013). Heteroscedasticity-robust Cp model averaging. *The Econometrics Journal* 16:463–472.

Lu, X. (2015). A covariate selection criterion for estimation of treatment effects. *Journal of Business & Economic Statistics* 33:506–522.

Lütkepohl, H. (1990). Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *The Review of Economics and Statistics* 72:116–125.

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

Lütkepohl, H., Staszewska-Bystrova, A., Winker, P. (2015). Comparison of methods for constructing joint confidence bands for impulse response functions. *International Journal of Forecasting* 31:782–798.

Rohan, N., Ramanathan, T. V. (2011). Order selection in arma models using the focused information criterion. *Australian & New Zealand Journal of Statistics* 53:217–231.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6:461–464.

Zhang, X., Liu, C.-A. (2017). Inference after model averaging in linear regression models. *Working Paper*.

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