Order Selection in Vector Autoregression by Mean Square Information Criterion

Michael Hellstern[1] and Ali Shojaie[1]

[1]Department of Biostatistics, University of Washington

Abstract

Vector autoregressive (VAR) processes are ubiquitously used in economics, finance, and biology. Order selection is an essential step in fitting VAR models. While many order selection methods exist, all come with weaknesses. Order selection by minimizing AIC is a popular approach but is known to consistently overestimate the true order for processes of small dimension. On the other hand, methods based on BIC or the Hannan-Quinn (HQ) criteria are shown to require large sample sizes in order to accurately estimate the order for larger-dimensional processes. We propose the mean square information criterion (MIC) based on the observation that the expected squared error loss is flat once the fitted order reaches or exceeds the true order. MIC is shown to consistently estimate the order of the process under relatively mild conditions. Our simulation results show that MIC offers better performance relative to AIC, BIC, and HQ under misspecification. This advantage is corroborated when forecasting COVID-19 outcomes in New York City. Order selection by MIC is implemented in the `micvar` `R` package available on CRAN.

1 Introduction

In vector autoregressive (VAR) models, each variable is modeled as a linear function of the multivariate time series over prior lags. VARs were first introduced in macroeconometrics by Sims (1980) and have since become standard for macroeconomic forecasting. They have also become essential tools in a range of other fields, including in biomedical applications; in neuroscience to analyze functional connectivity in the brain (Seth et al., 2015) and in epidemiology to predict COVID-19 cases (Kitaoka and Takahashi, 2023). A fundamental problem in fitting VAR models is how to choose the lag order: using too few lags may result in underfitting, while too many lags can lead to overfitting, both decreasing the accuracy of forecasts. Incorrect selection of the lag order can also impact the selection of the relevant variables in the VAR model, resulting in ambiguous interpretations, especially when the goal is to infer Granger causal effects (Shojaie and Fox, 2022). Unfortunately, the lag order is typically unknown and must be chosen either by prior knowledge or in a data-dependent way.

Perhaps the most popular VAR order selection method is to choose the order that minimizes an information theoretic criterion, commonly the Akaike's Information Criterion (AIC) (Akaike, 1973, 1974). Minimizing the AIC selects the model with the lowest negative log likelihood plus a penalty term on the number of independently adjustable parameters. Despite its popularity, AIC has several drawbacks for use in VAR models. The first stems from AIC's inherent reliance on the likelihood. In the context of VAR models, this often amounts to assuming a Gaussian likelihood, which results in simplifying the log likelihood term to the log determinant of the prediction error matrix. When the errors are not Gaussian, this simplification of the likelihood is no longer valid. The second is that AIC may not provide a consistent estimate of the VAR order (Lütkepohl, 2005, Corollary 4.2.1). Although this limitation improves as the dimension of the process increases and is negligible for VAR models of dimension greater than 5 (Paulsen and Tjøstheim, 1985), it nonetheless limits the applicability of AIC. This lack of consistency is highlighted in the simulation results in the left panel of Figure 1. The plot shows that in the univariate case, AIC reaches its peak accuracy of 0.6 at around $n = 2000$ and does not improve substantially as n increases to 5000. These results are in contrast to those presented in the right panel of Figure 1 for a 10 dimensional VAR model: in this case, AIC has nearly perfect accuracy for $n \geq 2000$. Shortly after AIC was introduced, additional VAR order selection criteria were proposed, namely the Bayesian information criterion, (BIC, Schwarz, 1978) and the Hannan-Quinn (HQ) criterion (Hannan and Quinn, 1979; Quinn, 1980). Similar to AIC, the BIC criterion relies on the likelihood, which in the case of Gaussian errors amounts to the log determinant

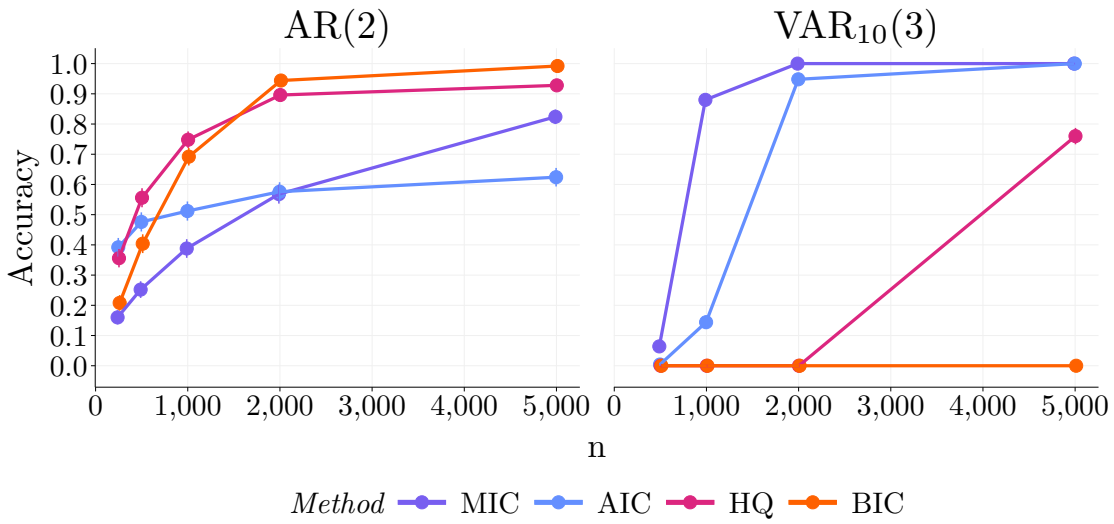

Figure 1: Accuracy of various order selection methods in detecting the order of simulated $\mathrm{AR}(2)$ (*left*) and $\mathrm{VAR}_{10}(3)$ (*right*) processes. Vertical lines represent standard errors. Accuracy is measured as the proportion of simulations where the correct order was chosen. See Section 5 for more details.

of the prediction error, plus a penalty on the number of model parameters. The HQ criterion is not likelihood-based but it too relies on the log determinant of the prediction error plus a penalty. Although all three criteria use the log determinant of the prediction error, they differ in the penalty used. Denoting by n the number of observations, AIC uses a penalty of $(2/n)(\#\text{parameters})$, while HQ and BIC use penalties of $((2\log\log n)/n)(\#\text{parameters})$ and $((\log n)/n)(\#\text{parameters})$, respectively. This change of penalty is essential: if the underlying process is a stationary and stable VAR with standard white noise, it can be shown that both HQ and BIC consistently estimate the true order of the process (Lütkepohl, 2005, Corollary 4.2.2). However, while HQ and BIC are consistent, simulation results in Figure 1 show they perform poorly for small sample sizes when the dimension of the process increases to a moderate size.

To address the above limitations of existing order selection methods, we propose a new method, mean squared information criterion (MIC). MIC is likelihood-free, consistent, and performs well in a variety of simulation settings. Our criterion leverages the novel observation that the expected squared error loss is constant when the fitted VAR order is at least as large as the true model order. We establish the consistency of MIC under mild assumptions and show that, compared with AIC, HQ, and BIC, it performs well in a variety of simulation settings and forecasting on real data.

The rest of the paper is structured as follows. In Section 2, we motivate our new information criterion and show that the expected squared error loss is constant when the fitted order is at least as large as the true model order. We extend this observation to the multivariate case and present theoretical results in Section 3, introduce our estimator in Section 4 and compare its performance to AIC, HQ and BIC using simulated data in Section 5. We apply our method to a financial application and COVID-19 forecasting in Section 6 and end with a discussion in Section 7.

Notation: We use uppercase letters X, Y, Z to denote random variables. We will use subscript t to denote the time component as in Z_t. When Z or Z_t are random vectors, the components can be accessed by Z_j, $Z_{t,j}$, respectively. Observed values of random variables are written in lowercase as in z_t. Additionally, observed vectors and matrices are denoted using bold lowercase and uppercase letters as in $\mathbf{z}$ and $\mathbf{Z}$, respectively. Hats or tildes will be used to specify sample estimates of population quantities, e.g. $\hat{\Gamma}_0$ and $\tilde{\Gamma}_0$ represent different sample estimates of Γ_0.

2 A new idea

We begin with the univariate case. Let Z_t be a stationary univariate time series. Without loss of generality, we assume Z_t is a mean zero process, as in practice, we can subtract the sample mean from the data. We denote the h^{th} autocovariance

of Z_t as $\gamma_h = \mathbb{E}(Z_t Z_{t-h})$ and denote $Y_t = Z_t, X_p = \begin{bmatrix} Z_{t-1} \dots Z_{t-p} \end{bmatrix}$. We wish to study the behavior of the expected squared error loss as a function of different orders p up to a prespecified maximum order, $p_{\max}$. Specifically, we study

$$L_{\mathrm{AR}}(p, \beta) := \mathbb{E}\left[(Y_t - X_p \beta)^2\right].$$

In this case, β is a nuisance parameter. For fixed p, $L_{\mathrm{AR}}(p, \beta)$ is the usual least squares problem and we can solve for β to get $\beta_p^* = \mathbb{E}\left(X_p^T X_p\right)^{-1} \mathbb{E}\left(X_p^T Y_t\right)$. Plugging β_p^* back in to the expected loss and simplifying, we get a loss that only depends on p, for which we use the shorthand notation $L_{\mathrm{AR}}(p)$:

$$L_{\mathrm{AR}}(p) = \mathbb{E}\left(Y_t^2\right) - \mathbb{E}\left(X_p^T Y_t\right)^T \mathbb{E}\left(X_p^T X_p\right)^{-1} \mathbb{E}\left(X_p^T Y_t\right).$$

By stationarity of Z_t, this simplifies to

$$L_{\mathrm{AR}}(p) = \gamma_0 - \begin{bmatrix} \gamma_1 & \dots & \gamma_p \end{bmatrix} \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \dots & \\ \gamma_1 & \gamma_0 & \gamma_1 & \dots & \\ \gamma_2 & \gamma_1 & \gamma_0 & \dots & \\ \vdots & \vdots & \vdots & \ddots & \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \dots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{bmatrix}.$$

Note that up to this point we have only assumed Z_t is stationary and have made no assumptions on the structure of the data generating process. Now suppose Z_t is not only stationary but is also a stable $\mathrm{AR}(p_0)$ process, that is,

$$Z_t = a_1 Z_{t-1} + \dots + a_{p_0} Z_{t-p_0} + \epsilon_t,$$

where $\mathbb{E}(\epsilon_t) = 0$, $\mathbb{E}\left(\epsilon_t^2\right) = \sigma_\epsilon^2$ and $\mathbb{E}(\epsilon_s \epsilon_t) = 0$ for $s \neq t$. Then, γ_h has a known form and we can calculate the expected squared error loss for each $p = 0, \dots, p_{\max}$ (Lütkepohl, 2005, pp. 26-30). For $p = 0$, there are no prior lags as predictors so we get that $L_{\mathrm{AR}}(0) = \gamma_0$.

We now observe that the expected squared error loss should be flat after the true order p_0. This is because the first p_0 lags should contain all the information needed for prediction. Furthermore, the error of the process is white noise and thus unpredictable. Therefore, the lowest achievable expected squared error should be the variance of the error. In Figure 2, we compute the expected squared error loss for several AR processes and see that our intuition holds. It can be seen that the expected squared error loss is indeed flat once the fitted order is at least as large as the true order p_0 and the eventual value of the expected loss is the variance of the error, σ_ϵ^2. The behavior of the loss observed in Figure 2 can be proven mathematically and follows immediately from the multivariate case proved in Theorem 1 in the next section. Given this behavior, if we add an appropriately sized penalty to the fitted order p and consider orders large enough ($p_{\max} > p_0$), we should be able to design a correct order selection procedure in the sense that when the process is AR(p_0) we will recover the true order p_0. Specifically, the true order can be found as

$$p_0 = \operatorname*{arg\,min}_{p \in \{0, \dots, p_{\max}\}} L_{\mathrm{AR}}(p) + \lambda p.$$

If λ is too large, we would underestimate the order. However, we must also have $\lambda > 0$ to avoid multiple solutions and an undefined parameter. In practice, we rarely deal with univariate time series so in the next section we extend these concepts to the multivariate case.

3 Extension to the multivariate case

In this section, we extend the concepts of Section 2 to the multivariate case and show a similar behavior of the expected squared error loss. Suppose $Z_t = \begin{bmatrix} Z_{t,1}, \dots, Z_{t,k} \end{bmatrix}^T$ is a k-dimensional column vector. We assume that Z_t is a stable process, as formalized in Assumption 1 (stability) in Section A. The h^{th} autocovariance matrix is defined as $\Gamma_h = \mathbb{E}\left(Z_t Z_{t-h}^T\right) \in \mathbb{R}^{k \times k}$. Note that Γ_h is not symmetric in general, but $\Gamma_h = \Gamma_{-h}^T$. Similar to the univariate case, we define $Y_t = Z_t, X_p = \begin{bmatrix} Z_{t-1}^T \dots Z_{t-p}^T \end{bmatrix}^T$, where $Y_t \in \mathbb{R}^{k \times 1}$ and $X_p \in \mathbb{R}^{kp \times 1}$. We study the expected squared error loss at different values of p,

$$L_{\mathrm{VAR}}(p, A_p) := \mathbb{E}\left[(Y_t - A_p X_p)^T (Y_t - A_p X_p)\right].$$

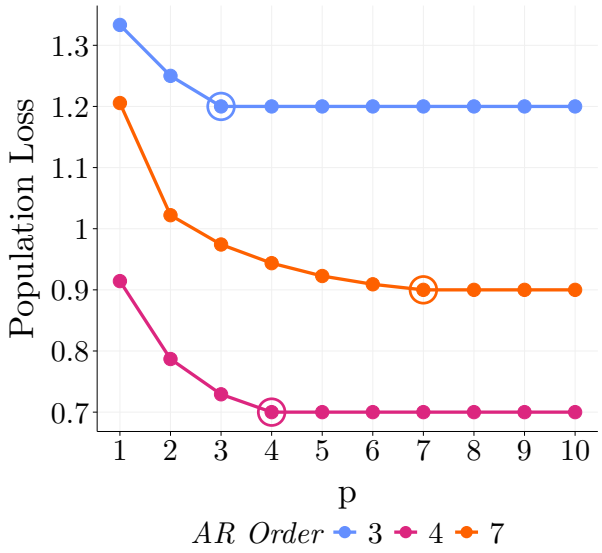

Figure 2: Population loss for AR(3), AR(4), and AR(7) processes with $\sigma_\epsilon^2 = 1.2, 0.7, 0.9$ respectively. As the fitted order increases to the true order the expected loss decreases monotonically to σ_ϵ^2. The true order for each line is denoted by a hollow circle.

We proceed by profiling out A_p. For fixed p, we have $A_p^* = \mathbb{E}\left(Y_t X_p^T\right) \mathbb{E}\left(X_p X_p^T\right)^{-1}$. Similar to the univariate case, we plug A_p^* back into the expected loss which we denote as $L_{\mathrm{VAR}}(p)$ after simplification. It is shown in Section B that

$$L_{\mathrm{VAR}}(p) = \mathrm{Tr}\left(\Gamma_0\right) - \mathrm{Tr}\left(\begin{bmatrix}\Gamma_1 & \dots & \Gamma_p\end{bmatrix} \begin{bmatrix}\Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0\end{bmatrix}^{-1} \begin{bmatrix}\Gamma_1^T \\ \vdots \\ \Gamma_p^T\end{bmatrix}\right). \tag{1}$$

For $p = 0$ there are no prior lags as predictors so we get that $L_{\mathrm{VAR}}(0) = \mathrm{Tr}\left(\Gamma_0\right)$. Similar to the univariate case, if Z_t is a stable VAR(p_0) process, the multivariate loss decreases until the fitted order is equal to the true order at which point the population loss is constant with a value of $\mathrm{Tr}(\Sigma_\epsilon)$. This behavior of the loss is formally stated in Theorem 1. The assumptions are stated in Appendix A

Theorem 1 (Flat Loss)**.** *Suppose Z_t is a* VAR(p_0) *process with standard white noise. That is, $Z_t = \sum_{i=1}^{p_0} A_i Z_{t-i} + \epsilon_t$ where $\mathbb{E}\left(\epsilon_t\right) = 0, \mathbb{E}\left(\epsilon_t \epsilon_t^T\right) = \Sigma_\epsilon$ and $\mathbb{E}\left(\epsilon_t \epsilon_s^T\right) = 0$ for $t \neq s$. If Assumption 1 (stability), Assumption 2 (invertibility), and Assumption 3 (irreducibility) hold, then*

$$\begin{cases} L_{\mathrm{VAR}}(p) < L_{\mathrm{VAR}}(p-1) & \text{if } p \leq p_0 \\ L_{\mathrm{VAR}}(p) = \mathrm{Tr}(\Sigma_\epsilon) & \text{if } p \geq p_0\,. \end{cases}$$

Assumptions 1 to 3 are mild and standard assumptions which hold in many applications. Further discussion on the assumptions is provided in Section A. For $L_{\mathrm{VAR}}(p_0)$, Theorem 1 implies both $L_{\mathrm{VAR}}(p_0) < L_{\mathrm{VAR}}(p_0 - 1)$ and $L_{\mathrm{VAR}}(p_0) = \mathrm{Tr}(\Sigma_\epsilon)$. If we penalize the fitted order p by an appropriate amount, and use $p_{\max} > p_0$, we would obtain a procedure that recovers the true order, p_0. This is formally stated in Corollary 1.

Corollary 1. *Let $M = \min\left(L_{\mathrm{VAR}}(p_0 - 1) - L_{\mathrm{VAR}}(p_0), \left[L_{\mathrm{VAR}}(p_0 - 2) - L_{\mathrm{VAR}}(p_0)\right]/2, \dots, \left[L_{\mathrm{VAR}}(0) - L_{\mathrm{VAR}}(p_0)\right]/p_0\right)$. Then, for a VAR($p_0$) process that satisfies the conditions of Theorem 1 and for $\lambda \in (0, M)$ and $p_{\max} > p_0$ we have*

$$p_0 = \underset{p \in \{0, \dots, p_{\max}\}}{\arg\min} \; L_{\mathrm{VAR}}(p) + \lambda p\,.$$

To estimate p_0, we need to estimate $L_{\text{VAR}}(p) + \lambda p$ for each p.While we use the form $L_{\text{VAR}}(p)$ with the autocovariance matrices to establish the theoretical results, $L_{\text{VAR}}(p)$ can be expressed as the expected squared error loss: $L_{\text{VAR}}(p) = \mathbb{E}\left[\left(Y_t - A_p^* X_p\right)^T \left(Y_t - A_p^* X_p\right)\right]$. Therefore, a natural estimator of $L_{\text{VAR}}(p)$ is the sample squared error loss, as defined formally below.

Let $\{\mathbf{z}_t\}_{t=1}^n$ denote the observed k-dimensional time series of length n. To estimate the squared error loss based on the least squares estimate of A_p^* at a fitted order p, we denote $\mathbf{Y}_p = \begin{bmatrix}\mathbf{z}_{1+p}, \ldots, \mathbf{z}_n\end{bmatrix} \in \mathbb{R}^{k \times (n-p)}$, $\mathbf{x}_{t,p} = \begin{bmatrix}\mathbf{z}_{t-1}^T & \ldots & \mathbf{z}_{t-p}^T\end{bmatrix}^T \in \mathbb{R}^{kp \times 1}$ and $\mathbf{X}_p = \begin{bmatrix}\mathbf{x}_{1+p,p} & \ldots & \mathbf{x}_{n,p}\end{bmatrix} \in \mathbb{R}^{kp \times (n-p)}$. The number of data points in $\mathbf{X}_p$ and $\mathbf{Y}_p$ depends on the fitted order p as fitting a $\text{VAR}(p)$ model requires the p prior data points as covariates. As a result, we only have $n - p$ usable data points for a $\text{VAR}(p)$ model. With these definitions, an estimate of the squared error loss at a fitted order p using the least squares estimates of A_p^* is given by

$$\begin{aligned}\hat{L}_{\text{VAR}}(p) &= \frac{1}{n-p} \operatorname{Tr}\left(\left(\mathbf{Y}_p - \hat{A}_p \mathbf{X}_p\right)^T \left(\mathbf{Y}_p - \hat{A}_p \mathbf{X}_p\right)\right) \\ &= \frac{1}{n-p} \operatorname{Tr}\left(\left(\mathbf{Y}_p - \hat{A}_p \mathbf{X}_p\right) \left(\mathbf{Y}_p - \hat{A}_p \mathbf{X}_p\right)^T\right),\end{aligned} \tag{2}$$

where $\hat{A}_p = \mathbf{Y}_p \mathbf{X}_p^T \left(\mathbf{X}_p \mathbf{X}_p^T\right)^{-1}$ and we used $\operatorname{Tr}(A^T B) = \operatorname{Tr}(A B^T)$.

With (2) and using $\lambda \in (0, M)$, we can define the estimated order based on the mean-squared information criterion (MIC), p_{MIC}^*, as

$$p_{\text{MIC}}^* = \underset{p \in \{0, \ldots, p_{\max}\}}{\arg\min} \; \hat{L}_{\text{VAR}}(p) + \lambda p\,. \tag{3}$$

Note that to use this estimator in practice, we need to specify the tuning parameter λ. We defer this to Eq. (4) below. To solve the minimization problem in Eq. (3), we can compute $\hat{L}_{\text{VAR}}(p)$ for each $p = 0, \ldots, p_{\max}$. Theorem 2 establishes the consistency of $\hat{L}_{\text{VAR}}(p)$.

Theorem 2 (Consistency of Loss)**.** *Under the assumptions of Theorem 1, we have that*

$$\left|\hat{L}_{\text{VAR}}(p) - L_{\text{VAR}}(p)\right| = o_P(n^{-1/2+\delta})\,,$$

for all $\delta > 0$.

As shown in the proof of Theorem 2 (see Section C.3), the consistency of $\hat{L}_{\text{VAR}}(p)$ relies on the consistency of the sample autocovariances and the rate of convergence of $\hat{L}_{\text{VAR}}(p)$ is the same as the rate of convergence of the sample autocovariance. One benefit of the proposed information criterion is that it does not rely on the form of the likelihood. In fact, Theorem 2 will hold as long as the sample autocovariance is consistent, with the rate of convergence matching that of the sample autocovariance. An immediate consequence of Theorem 2 is consistency of p_{MIC}^*.

Corollary 2 (Consistency of order estimate)**.** *Under the assumptions of Theorem 1, we have that*

$$p_{\text{MIC}}^* \to_p p_0,$$

where $\to_p$ *denotes convergence in probability.*

While our theoretical analyses and implementation estimate $L_{\text{VAR}}(p)$ using the least squares estimate of A_p^*, it is possible to use other estimators, such as the Yule-Walker estimate of A_p^*, and prove similar results. However, we chose to use the least squares estimate due to better small sample performance and lower bias (Lütkepohl, 2005; Tjøstheim and Paulsen, 1983). It is worth noting that, as pointed out in Lütkepohl (2005, pp. 86), the least squares and Yule-Walker estimators of A_p^* are asymptotically equivalent for stable processes, which are the focus of this work.

4 MIC estimator of VAR order

Our theoretical analyses in Section 3 assume $\lambda \in (0, M)$ is known. However, λ is unknown in practice and needs to be selected. The flat loss concept from Theorem 1 tells us that once the fitted order exceeds the true order, the loss should be constant. In practice, there is sampling variability so the loss will never completely stabilize. We generate an estimate of this variability using a "self-tuning" approach.

In our self-tuning approach, we fit models from lag order $p_{\max}+1$ to $2p_{\max}$ and take the absolute value of the mean of the difference between each loss and the subsequent loss to estimate the amount of variability, or noise, in the loss. While it is possible to use orders larger than $2p_{\max}$, the trade-off is fitting larger order models consumes more prior data points as covariates and reduces sample size. We find that $2p_{\max}$ works well in practice. Specifically, our self-tuning approach computes

$$\mathrm{MD} = \left|\mathrm{mean}\left(\hat{L}_{\mathrm{VAR}}(p_{\max}) - \hat{L}_{\mathrm{VAR}}(p_{\max}+1), \ldots, \hat{L}_{\mathrm{VAR}}(2p_{\max}-1) - \hat{L}_{\mathrm{VAR}}(2p_{\max})\right)\right| .$$

We then scale MD by $\sqrt{n/(k^2\log(n))}$ to get

$$\lambda_{\mathrm{ST}} = \mathrm{MD}\sqrt{\frac{n}{k^2\log(n)}} .$$

With this choice of λ, our estimator is defined as

$$\hat{p}_{\mathrm{MIC}} = \underset{p\in\{0,\ldots,p_{\max}\}}{\arg\min}\ \hat{L}_{\mathrm{VAR}}(p) + \lambda_{\mathrm{ST}} p . \tag{4}$$

We next discuss the choice of MD as well as the scaling $\sqrt{n/(k^2\log(n))}$.

Due to the flat loss property, each $\hat{L}_{\mathrm{VAR}}(p_{\max}+i) - \hat{L}_{\mathrm{VAR}}(p_{\max}+i+1)$ should represent an estimate of how the sample loss changes when we have exceeded the true order and increase the fitted order by 1. We average over $p_{\max}$ of these to reduce the variance in this estimate. When computing the mean, subsequent differences cancel and this quantity can be simplified as

$$\mathrm{MD} = \left|\frac{\hat{L}_{\mathrm{VAR}}(p_{\max}) - \hat{L}_{\mathrm{VAR}}(2p_{\max})}{p_{\max}}\right| .$$

Thus, MD can be computed efficiently as it only requires fitting one additional regression of order $2p_{\max}$. It is also worth noting that $\hat{L}_{\mathrm{VAR}}(p_{\max}), \ldots, \hat{L}_{\mathrm{VAR}}(2p_{\max})$ converge to the same asymptotic distribution and are asymptotically perfectly correlated. In this instance, the correlation is beneficial as it further reduces the variance of our estimate since for two random variables, X, Y, $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y)$.

To understand why it is necessary to scale MD, consider the case where we instead use $\lambda_{\mathrm{ST}} = \mathrm{MD}$. To simplify notation and provide a more concrete setting, consider $p_{\max} = 10$. With these, the score for order $2p_{\max} := 20$ based on Eq. (4) becomes

$$\begin{aligned}\hat{L}_{\mathrm{VAR}}(20) + \frac{\hat{L}_{\mathrm{VAR}}(10) - \hat{L}_{\mathrm{VAR}}(20)}{10}20 &= \frac{20}{10}\hat{L}_{\mathrm{VAR}}(10) - \frac{10}{10}\hat{L}_{\mathrm{VAR}}(20) \\ &= \hat{L}_{\mathrm{VAR}}(10) + \frac{\hat{L}_{\mathrm{VAR}}(10) - \hat{L}_{\mathrm{VAR}}(20)}{10}10 ,\end{aligned} \tag{5}$$

where we have assumed that $\hat{L}_{\mathrm{VAR}}(10) > \hat{L}_{\mathrm{VAR}}(20)$ so we can ignore the absolute value in MD. While it is possible that $\hat{L}_{\mathrm{VAR}}(10) < \hat{L}_{\mathrm{VAR}}(20)$ due to the different datasets used in fitting each model—e.g. $\hat{L}_{\mathrm{VAR}}(10)$ is estimated using $n-10$ observations while $\hat{L}_{\mathrm{VAR}}(20)$ uses $n-20$ observations—this is unlikely.

The last equation on the right hand side of Eq. (5) is exactly the value of the penalized loss for order $p_{\max} = 10$. In other words, setting $\lambda_{\mathrm{ST}} = \mathrm{MD}$ treats models $p_{\max}$ and $2p_{\max}$ as equally viable. However, these models are not equally viable and we want to enforce a belief that higher orders are less likely. Thus, we need to choose a penalty that is larger than MD. One way to do this is to scale MD by a factor greater than 1. From Theorem 2, we know that $\hat{L}_{\mathrm{VAR}}(p) = L_{\mathrm{VAR}}(p) + o_P(n^{-1/2+\delta}) \quad \forall\, \delta > 0$ so scaling by $n^{1/2}$ is too fast. In practice, we find that $\sqrt{n/\log(n)}$ works well. Lastly, we scale by $1/k$ since each additional order fitted requires estimating k more parameters than the prior order (see the proof of Theorem 1).

4.1 Relationship between MIC and other criteria

We now compare our method, MIC, to AIC, BIC, and HQ. For ease of comparison, we write the estimated error covariance matrix as $\hat{\Sigma}_p$.

$$\hat{\Sigma}_p = \frac{1}{n-p}\left(\mathbf{Y}_p - \hat{A}_p\mathbf{X}_p\right)\left(\mathbf{Y}_p - \hat{A}_p\mathbf{X}_p\right)^T .$$

Note that $\mathrm{Tr}(\hat{\Sigma}_p) = \hat{L}_{\mathrm{VAR}}(p)$. Denoting the natural logarithm as $\log$, the criteria considered can be written as follows

$$\mathrm{MIC}(p) = \mathrm{Tr}(\hat{\Sigma}_p) + \mathrm{MD}\sqrt{\frac{n}{k^2\log(n)}}p \qquad \mathrm{AIC}(p) = \log\left|\hat{\Sigma}_p\right| + \frac{2}{n}k^2 p\,;$$

$$\mathrm{BIC}(p) = \log\left|\hat{\Sigma}_p\right| + \frac{\log n}{n}k^2 p \qquad \mathrm{HQ}(p) = \log\left|\hat{\Sigma}_p\right| + \frac{2\log\log n}{n}k^2 p\,.$$

The above formulations show that all criteria rely on the same estimate of the error matrix $\hat{\Sigma}_p$ and differ only in how they use it—$\mathrm{Tr}()$ or $\log|\cdot|$—and penalize that information. The advantage of MIC is that it is consistent and likelihood-free. Our simulation results also show that MIC works well empirically.

4.2 Alternative choices of λ

We also considered two other methods to select the order based on the flat loss concept discussed in Section 3. The first method, which we refer to as MIC-sp, uses MIC with a penalty λ_{sp} that is chosen by splitting the data into train and test sets. Due to the time-dependent nature of our data, we use the first 70% of observations as training and the remainder as test data. Models from $p_{\max}$ to $2p_{\max}$ are fit on the training data to estimate $\hat{A}_p$ and their prediction errors computed on the test data are denoted as e.g. $\hat{\Sigma}_{\mathrm{test},p_{\max}}$. We set

$$\lambda_{\mathrm{sp}} = \mathrm{mean}\left(\left|\mathrm{Tr}(\hat{\Sigma}_{\mathrm{test},p_{\max}}) - \mathrm{Tr}(\hat{\Sigma}_{\mathrm{test},p_{\max}+1})\right|, \ldots, \left|\mathrm{Tr}(\hat{\Sigma}_{\mathrm{test},2p_{\max}-1}) - \mathrm{Tr}(\hat{\Sigma}_{\mathrm{test},2p_{\max}})\right|\right)\,.$$

Due to the flat loss property, each of these differences should be 0 and any sample variability should be captured in λ_{sp}.

We further consider a procedure, which we denote as MIC-mt. We again use a 70-30 train-test split and fit VAR models of order $0, \ldots, p_{\max}$ on the train dataset. Similarly, we compute the errors of each fitted model on the test data. MIC-mt then chooses the order $0, \ldots, p_{\max}$ that minimizes the test error. Simulations comparing all three methods in the case of a diagonal covariance matrix with Gaussian errors are shown in Figure S7. The results show that MIC, which indicates the MIC method with self-tuned λ, performs the best across a variety of sample sizes and dimensions. It is only consistently outperformed by MIC-sp in the AR(2) case. Results for other simulations in Section 5 are not shown but are qualitatively similar. Thus, we proceed with our proposed self-tuning approach for selecting λ.

5 Simulations

5.1 Order selection accuracy

In this section, we compare the accuracy of MIC, AIC, BIC, and HQ order selection methods using simulated data. In general, we will use $\mathrm{VAR}_k(p)$ to denote the dimension k and order p of the process. We will also use $\mathrm{U}(a, b)$ to denote a Uniform distribution with support (a, b). We consider VAR models with 4 different dimensions and 3 different error structures. The first is an autoregressive process of order 2, AR(2), with parameters (0.3, 0.1). The second is a $\mathrm{VAR}_2(2)$ process. The entries of the first lag coefficient matrix are 25% sparse and randomly drawn from either a $\mathrm{U}(0.1, 0.3)$ or a $\mathrm{U}(-0.3, -0.1)$ each with 50% probability. The entries of the second lag coefficient matrix are 50% sparse and randomly drawn from either a $\mathrm{U}(0.07, 0.2)$ or a $\mathrm{U}(-0.2, -0.07)$ each with 50% probability. The third simulation setting is a $\mathrm{VAR}_5(3)$. All lag coefficient matrices have 60% sparsity. In the first lag coefficient matrix, the non-zero entries are drawn from a $\mathrm{U}(0.1, 0.3)$ or a $\mathrm{U}(-0.3, -0.1)$ each with equal probability. The second lag coefficient matrix uses a $\mathrm{U}(0.1, 0.2)$ or a $\mathrm{U}(-0.2, -0.1)$ while the third uses a $\mathrm{U}(0.05, 0.1)$ or a $\mathrm{U}(-0.1, -0.05)$. The fourth simulation setting is a $\mathrm{VAR}_{10}(3)$ process where the first lag coefficient matrix has 40% sparsity and the non-zero entries are drawn from a $\mathrm{U}(0.1, 0.3)$ or a $\mathrm{U}(-0.3, -0.1)$. The second lag coefficient matrix has 80% sparsity, but the remaining entries are drawn from a $\mathrm{U}(-0.2, 0.2)$. The final lag coefficient matrix has 80% sparsity with remaining entries drawn from a $\mathrm{U}(-0.1, 0.1)$. All coefficient matrices are generated once and the same matrices are used throughout the simulations. Stability for each setting is verified using the method of Lütkepohl (2005, pp. 14 - 17).

All datasets are simulated using three error structures. The first is mean-zero Gaussian errors with an identity covariance matrix while the second uses a randomly generated covariance matrix. The covariance matrix is generated by generating a $k \times k$ matrix with entries drawn from a $\mathrm{U}(-3, 3)$. The matrix is then symmetrized by left multiplying by its transpose. We enforce a maximum condition number of 100 for each matrix by consecutively adding 0.001 to diagonal elements until the condition number is met. After the covariance matrix is reconditioned, it is scaled to

have unit variances. The third error structure is a Gaussian mixture model with 5 components. Each component is Gaussian where the mean vector is generated from a $\mathrm{U}(-5, 5)$. For each k, we then subtract the mean across all 5 components so that the mean of the component means is 0. The covariances are $k \times k$ matrices with entries drawn from a $\mathrm{U}(-3, 3)$. The matrices are subsequently symmetrized, reconditioned, and rescaled as explained above. We simulate $n = 250, 500, 1000, 2000, 5000$ observations for each setting except for $\mathrm{VAR}_{10}(3)$ where $n = 250$ is excluded as the number of parameters exceeds the number of data points.

Lastly, we consider a $\mathrm{VAR}_3(2)$ process that switches between one of two regimes. Both regimes share the same lag coefficient matrices and error covariances but differ in their means. Since both regimes are order 2, the true order is 2. The regime mean vectors are generated from $\mathrm{U}(-0.5, 0.5)$ and the regime switches every 10% of observations. For example, if $n = 5000$, there are nine regime swtiches at $n = 500, 1000, 1500, \ldots, 4500$. Due to the time-dependent switching mean, this process is not stationary and we present these results to study how the methods perform under misspecification. For this process, the first lag coefficient matrix is 30% sparse with entries randomly drawn from a $\mathrm{U}(0.1, 0.3)$ or $\mathrm{U}(-0.3, -0.1)$ while the second is 60% sparse with entries drawn from a $\mathrm{U}(0.1, 0.2)$ or $\mathrm{U}(-0.2, -0.1)$. The errors are generated from a Gaussian distribution with mean zero and randomly generated covariance matrix as above. Prior to analyzing the data, each of the three process components is centered so the overall mean of the components is 0.

We compute each criteria, MIC, AIC, BIC, and HQ, for $p = 0, \ldots, 10 := p_{\max}$. The estimated orders for each criteria are those that achieve the minimum value. That is,

$$\hat{p}_{\mathrm{CRITERION}} = \underset{p \in \{0, \ldots, p_{\max}\}}{\arg\min} \ \mathrm{CRITERION}(p)\,.$$

For each error structure and VAR model, we simulate $B = 250$ data sets and compute the proportion of times the correct order is estimated. That is, we use

$$\text{Accuracy} = \frac{1}{B} \sum_{b=1}^{B} I(\hat{p}_{\mathrm{CRITERION}} = p_0)\,.$$

The simulation results are summarized in Figures 3 to 6. When errors are diagonal Gaussian and the dimension is large, MIC outperforms AIC, HQ, and BIC, as shown in Figure 3. This makes sense as in this setting AIC, HQ, and BIC all use the entire error matrix, including the off-diagonals, through the determinant. When the true errors are diagonal, the estimated off-diagonals can contain incorrect information. On the other hand, MIC only uses the diagonals and so discards the potentially misleading off-diagonal information. For diagonal errors with small sample sizes and small dimension, HQ, and BIC perform slightly better, but MIC is still competitive. As previously noted, AIC is not consistent for the $\mathrm{AR}(2)$ and $\mathrm{VAR}_2(2)$ processes. Results for non-diagonal Gaussian errors in Figure 4 show much better performance for AIC, BIC, and HQ. MIC still appears to consistently estimate the order as sample size increases, but the small sample performance is now worse relative to the other methods. This makes sense as there is useful information contained in the off-diagonals of the error matrix that AIC, BIC, and HQ can leverage while MIC does not. It is also worth noting that the performance of MIC relative to other methods improves as the dimension of the process increases. We see similar trends for Gaussian mixture errors in Figure 5. Note that performance is sometimes better and sometimes worse than the corresponding results in Figure 4. This is likely due to variation in the simulated error covariances and means.

For the misspecification simulation—the $\mathrm{VAR}_3(2)$ switching process in Figure 6—the performance of AIC, BIC, and HQ deteriorates as sample size increases. While the performance of AIC deteriorates fastest, all methods except MIC have an accuracy of 0 for $n = 5000$. Conversely, the accuracy of MIC increases a n increases suggesting it consistently estimates the true order. To investigate this further, we explore the performance of each method with much larger sample sizes in Figure S8 and find that the results are consistent with Figure 6. That is, AIC, BIC, and HQ all have 0 accuracy for $n > 20000$. In contrast, MIC shows robust performance and perfectly estimates the order for all sample sizes in Figure S8.

While MIC is outperformed by AIC, BIC, and HQ when the true error matrix has off-diagonal elements, it offers better performance when the true error matrix is diagonal and appears to be more robust to misspecification as shown by the performance in the $\mathrm{VAR}_3(2)$ switching setting. Thus, MIC offers a viable alternative to AIC/BIC/HQ, as in many practical applications errors are unlikely to be Gaussian. In Section 6, we investigate the performance of MIC compared to AIC/BIC/HQ by comparing forecast performance on two datasets.

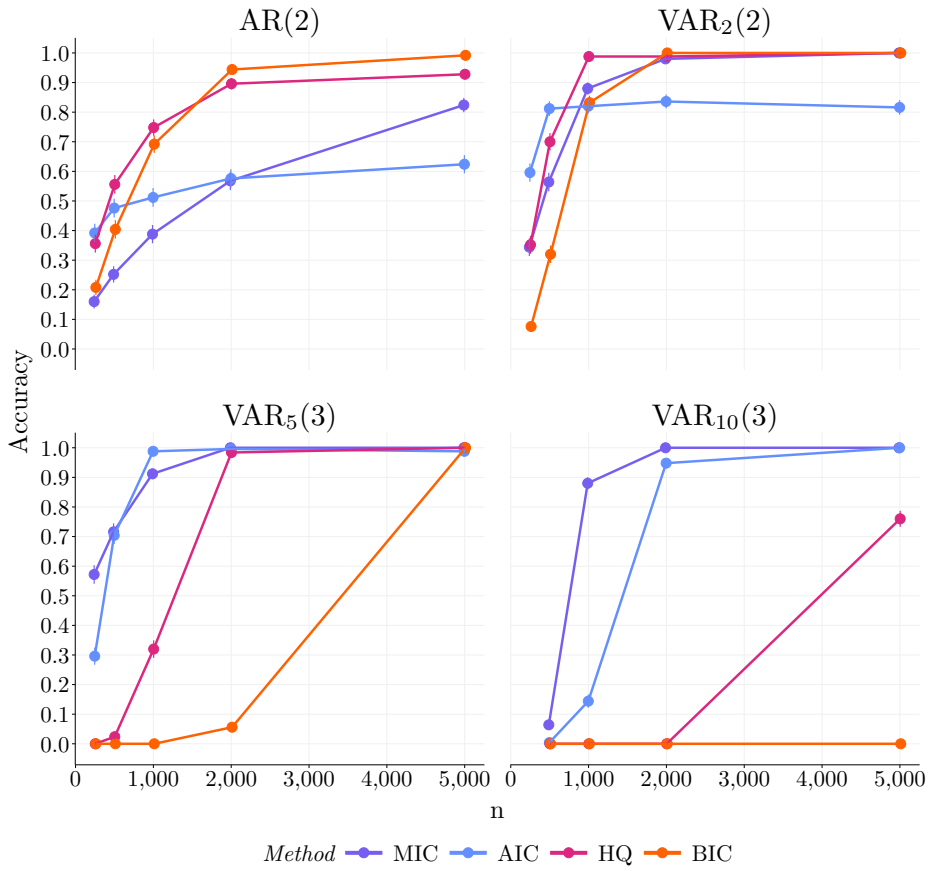

Figure 3: **Diagonal Gaussian errors**. Simulation results for accuracy of specific order selection method and simulation setting with diagonal Gaussian errors. Vertical lines indicate standard errors.

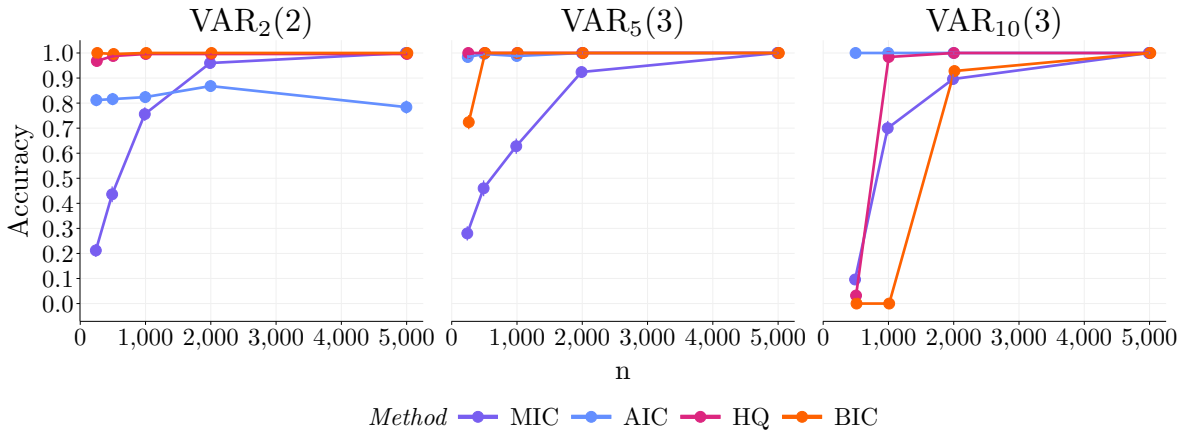

Figure 4: **Non-diagonal Gaussian errors**. Simulation results for accuracy of specific order selection method and simulation setting with non-diagonal Gaussian errors. Vertical lines indicate standard errors.

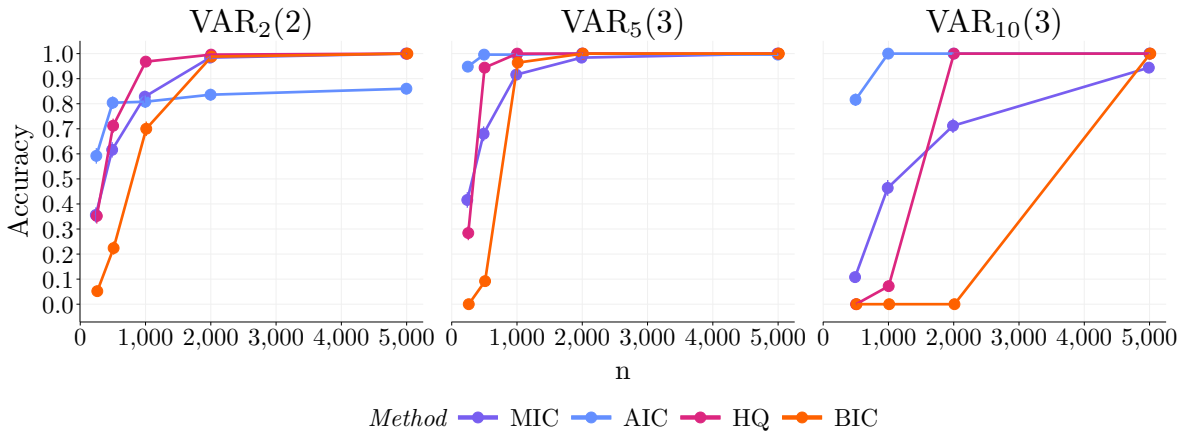

Figure 5: **Gaussian mixture errors**. Simulation results for accuracy of specific order selection method and simulation setting with Gaussian mixture errors. Vertical lines indicate standard errors.

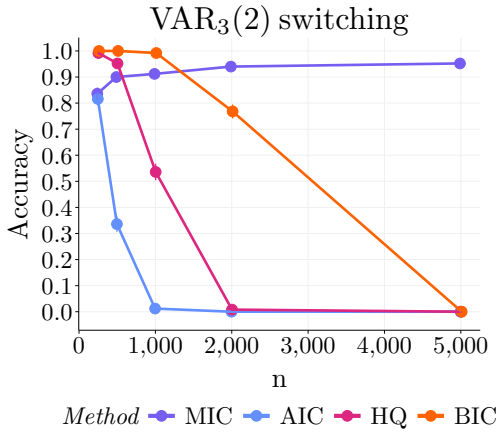

Figure 6: **VAR$_3$(2) switching**. Simulation results for accuracy of specific order selection method and simulation setting. Vertical lines indicate standard errors.

5.2 Over and under selection probability

As previously mentioned, it is known that AIC does not provide a consistent estimate of the VAR order when the dimension is small. To investigate this, we compare the likelihood of over and under selection of the model order for each of the order selection methods using simulations. Since MIC uses an estimated λ, we evaluate its theoretical properties with an oracle λ value, which we denote as MIC-oracle. Specifically, we choose $\lambda_{\text{oracle}} = M/2$ where M is defined in Corollary 1. The results for diagonal Gaussian errors are shown in Figures S9 and S10 while the non-diagonal gaussian errors are shown in Figure S11. Overall, we see that MIC-oracle tends to select an order that is larger than the true order (over selection) while the alternative methods AIC, BIC, and HQ, tend to select an order that is smaller than the true order (under selection). Figure S11 shows that, when the dimension of the process is large and errors are non-diagonal Gaussian, MIC-oracle suffers from worse over selection relative to the under selection from AIC, BIC, and HQ.

6 Applications

In this section, we apply our MIC method to two different forecasting problems and compare it to AIC, BIC, and HQ. The first problem is financial forecasting, while the second is forecasting COVID-19 outcomes. In both problems, we follow Nicholson et al. (2020) by comparing the weighted mean squared forecast error (wMSFE). In both problems, the first 80% of observations are used to estimate the order for MIC, AIC, BIC, and HQ, while the last 20% are used for evaluating the forecast accuracy. Formally, if n represents the total number of observations, then the first $T_1 = \lfloor 0.8n \rfloor$ data points are used to estimate the order for each method and the remaining observations are used for testing. We use a rolling window of size T_1 to perform one-step ahead forecasts. That is, if t indexes the observation we are forecasting, then we use observations $t - T_1, \ldots, t - 1$ as the rolling window. For each rolling window, we standardize each variable in the series by subtracting the mean and dividing by the standard deviation. The observation we are forecasting is also standardized using the mean/SD from the rolling window. We then fit a VAR model corresponding to the orders chosen by MIC, AIC, BIC, and HQ to this standardized rolling window and predict observation t. The wMSFE for method m is computed over all series and forecast time points as

$$\text{wMSFE}(m) = \frac{1}{k(n - T_1)} \sum_{i=1}^{k} \sum_{t=T_1+1}^{n} \left(\frac{y_{i,t} - \hat{y}_{i,t}^{m}}{\hat{\sigma}_i} \right)^2,$$

where $\hat{\sigma}_i$ is the standard deviation of the variable i computed over the forecast observations.

6.1 Daily realized stock variances

We compare the forecast performance of VAR models with order selected by MIC, AIC, BIC, and HQ using data from the Oxford-Man Institute of Quantitative Finance obtained from an older version of the `mfGARCH R` package, https://github.com/onnokleen/mfGARCH, from Conrad and Kleen (2020). Specifically, we analyze 5-minute return daily realized variances for up to 17 stocks from January 3, 2000 to June 27, 2018 ($n = 4,847$). We perform two analyses: one using the same $k = 16$ stocks as in Nicholson et al. (2020) as well as $k = 7$ stocks from Son et al. (2023). Many stocks from Nicholson et al. (2020) and Son et al. (2023) overlap and there are only 17 total unique stocks. Due to high levels of missingness (34%) we exclude OMXSPI, an index of the Stockholm Stock Exchange, from our analysis based on the Son et al. (2023) stocks. A full list of the stocks analyzed is given in Section E. All data are log-transformed to make them stationary. As we are not specifically interested in high-dimensional applications we estimate the order for each order selection method using a $p_{\max} = 10$, equivalent to two trading weeks.

Forecast results for both the $k = 16$ and $k = 7$ analysis are displayed in Table 1. Overall, we see that the methods considered give very similar forecast accuracy despite a large range of orders. For example, in Table 1(a), the order varies from a low of 2 chosen by BIC to a high of 9 chosen by MIC.

6.2 COVID-19 in New York City

We next compare the performance of order selection methods in forecasting COVID-19 outcomes in New York City. Daily data on deaths, cases, and hospitalizations in New York City due to COVID-19 are available starting February 29, 2020 at City of New York's website. We analyze data from February 29, 2020 to July 8, 2024 ($n = 1,592$). All data are

		AIC	BIC	HQ	MIC
$k = 16$	Order	7	2	4	9
	Forecast Error	0.485	0.504	0.487	0.491
$k = 7$	Order	8	4	6	6
	Forecast Error	0.493	0.497	0.492	0.492

Table 1: Comparison of order selection methods based on weighted Mean Squared Forecast Error for daily realized stock variances for $k = 16$ stocks and $k = 7$ stocks. Order selected by each method is also included.

first differenced to make them stationary and we use $p_{\max} = 30$. As a check, we run an Augmented Dickey-Full test (ADF, Said and Dickey, 1984) and a Kwiatkowski-Phillips-Schmidt-Shin (KPSS, Kwiatkowski et al., 1992) test for each series after differencing. The null hypothesis of the ADF test is that a unit root is present in the time series while the null hypothesis of the KPSS test is that the series is trend-stationary. All series pass the ADF test with $p < 0.01$ and the KPSS test with $p > 0.1$.

Forecast results are displayed in Table 2. We see that AIC, BIC, and HQ all fit models using around a month's worth of prior data points ($p = 30$) to forecast the next day. However, these models are substantially worse than the model fitted using MIC order selection, which only uses around a week of prior data points ($p = 8$). In fact, the forecast accuracy of the model fitted using MIC is around 30% better than the accruacy of those fit by AIC/BIC/HQ.

	AIC	BIC	HQ	MIC
Order	30	24	30	8
Forecast Error	1.334	1.301	1.334	1.036

Table 2: Comparison of order selection methods based on weighted Mean Squared Forecast Error for COVID-19 outcomes. Order selected by each method is also included.

7 Discussion

In this paper, we proposed the mean square information criterion (MIC), a new approach for estimating the order of VAR processes. MIC is based on a key new observation: the flatness of the expected squared error loss after the fitted order exceeds the true order. We show, under relatively mild assumptions, that the true order can be estimated consistently by minimizing the MIC. Specifically, consistency of MIC only requires consistent estimates of the autocovariances.

Our proposed method, MIC, was compared to three other order selection criteria, AIC, BIC, and HQ, based on the proportion of simulations in which the correct order was correctly estimated. Simulation settings ranged from univariate to 10-dimensional VAR models with between 250 to 5,000 observations. Simulations included both Gaussian and Gaussian mixture error structures, as well as a regime switching $\mathrm{VAR}_3(2)$ model. While outperformed in Gaussian errors and small sample sizes, relative to the other criteria, MIC showed the best performance when the process had regime changes. As errors are unlikely ever Gaussian, these results suggest that MIC can be very useful in practice. This is confirmed in our data applications where order selection via MIC achieved comparable forecast accuracy for daily realized stock variance and substantially better accuracy in forecasting COVID-19 outcomes in NYC when compared to order selected via AIC, BIC, or HQ.

An interesting direction for future work is to extend the proposed method to high-dimensional settings. In high dimensions, we can substitute the least squares estimate by e.g. ridge or LASSO estimators. Alternatively, estimates of the autocovariances in high dimensions may be used and plugged directly into the loss. It would then be interesting to compare the resulting estimator to recently proposed regularization-based approaches (Shojaie and Michailidis, 2010; Shojaie et al., 2012; Nicholson et al., 2017).

References

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–81, 1973.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723, 1974.

Christian Conrad and Onno Kleen. Two are better than one: volatility forecasting using multiplicative component garch-midas models. *Journal of Applied Econometrics*, 35(1):19–45, 2020.

Jean Gallier et al. The schur complement and symmetric positive semidefinite (and definite) matrices (2019). *URL https://www. cis. upenn. edu/jean/schur-comp. pdf*, 2020.

Edward J Hannan and Barry G Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B*, 41(2):190–195, 1979.

Takayoshi Kitaoka and Harutaka Takahashi. Improved prediction of new covid-19 cases using a simple vector autoregressive model: evidence from seven new york state counties. *Biology Methods and Protocols*, 8(1):bpac035, 2023.

Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1-3):159–178, 1992.

Tzon-Tzer Lu and Sheng-Hua Shiou. Inverses of $2\times$ 2 block matrices. *Computers & Mathematics with Applications*, 43(1-2):119–129, 2002.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

William B Nicholson, David S Matteson, and Jacob Bien. VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, 33(3):627–651, 2017.

William B Nicholson, Ines Wilms, Jacob Bien, and David S Matteson. High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, 21(166):1–52, 2020.

Jostein Paulsen and Dag Tjøstheim. On the estimation of residual variance and order in autoregressive time series. *Journal of the Royal Statistical Society: Series B*, 47(2):216–228, 1985.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7 (15):510, 2008.

Barry G Quinn. Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society: Series B*, 42(2):182–185, 1980.

Said E Said and David A Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 1984.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

Anil K Seth, Adam B Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9(1):289–319, 2022.

Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

Ali Shojaie, Sumanta Basu, and George Michailidis. Adaptive thresholding for reconstructing regulatory networks from time-course gene expression data. *Statistics in Biosciences*, 4:66–83, 2012.

Christopher A Sims. Macroeconomics and reality. *Econometrica*, pages 1–48, 1980.

Bumho Son, Yunyoung Lee, Seongwan Park, and Jaewook Lee. Forecasting global stock market volatility: The impact of volatility spillover index in spatial-temporal graph-based model. *Journal of Forecasting*, 42(7):1539–1559, 2023.

Dag Tjøstheim and Jostein Paulsen. Bias of some commonly-used time series estimates. *Biometrika*, 70(2):389–399, 1983.

Appendix

Table of Contents

A Assumptions

In this section, we list the assumptions used in our analyses and provide a brief discussion of each. All these assumptions are mild.

Assumption 1. *$Z_t \in \mathbb{R}^{k\times 1}$ is a stable mean zero process. That is,* $\det(I_k - A_1 z - \cdots - A_{p_0} z^{p_0}) \neq 0$ *for* $|z| \leq 1$.

This stability condition is standard and is identical to that used in Lütkepohl (2005, Eq. (2.1.12)). Note that stability implies stationarity (Lütkepohl, 2005, Proposition 2.1) and this assumption is required to replace second order expectations with autocovariances. That is, Assumption 1 is required to have $\mathbb{E}\left(Z_t Z_{t-h}^T\right) = \Gamma_h$. It is also required to use the Yule-Walker equations.

Assumption 2.

$$\mathbb{E}\left(X_{p_{\max}} X_{p_{\max}}^T\right) = \begin{bmatrix} \Gamma_0 & \Gamma_1 & \dots & \Gamma_{p_{\max}-1} \\ \Gamma_1^T & \Gamma_0 & \dots & \Gamma_{p_{\max}-2} \\ \vdots & & & \vdots \\ \Gamma_{p_{\max}-1}^T & \Gamma_{p_{\max}-2}^T & \dots & \Gamma_0 \end{bmatrix},$$

is invertible.

Note that due to the quadratic form of $\mathbb{E}\left(X_{p_{\max}} X_{p_{\max}}^T\right)$, this matrix is symmetric and positive semidefinite. By assuming invertibility, we ensure this matrix is also positive definite. We also only need to make this assumption for $p_{\max}$ and we will have positive definiteness for all $\mathbb{E}\left(X_{p_{\max}} X_{p_{\max}}^T\right)$ for $i = 1, \dots p_{\max}$ since a matrix is positive definite if and only if all its principle minors are positive (see Lütkepohl (2005) Appendix A.8.3). The ki^{th} principle minor of $\mathbb{E}\left(X_{p_{\max}} X_{p_{\max}}^T\right)$ is $\det\left(\mathbb{E}\left(X_i X_i^T\right)\right)$. Again, $\mathbb{E}\left(X_i X_i^T\right)$ is symmetric and positive semi-definite so ensuring a positive determinant ensures strictly positive eigenvalues and thus positive definiteness. In all simulation settings this assumption has been met.

Assumption 3. *We assume that for a VAR(p_0) process when $p < p_0$,*

$$\begin{bmatrix} \Gamma_1 & \dots & \Gamma_{p-1} \end{bmatrix} \begin{bmatrix} \Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \Gamma_1^T \\ \vdots \\ \Gamma_p^T \end{bmatrix} - \Gamma_p \neq 0.$$

This assumption essentially states that we need at least p_0 lags to generate the pth autocovariance Γ_p. This is a mild assumption. For example, it is implicitly made in Lütkepohl (2005, Eq. (2.1.37)) where for a VAR(p_0) process, the autocovariance matrix is determined by the prior p_0 lags.

B Simplifying multivariate loss

In this section, we show that the profiled loss can be written as

$$\mathbb{E}\left[\left(Y_t - A_p^* X_p\right)^T \left(Y_t - A_p^* X_p\right)\right] = \operatorname{Tr}\left(\Gamma_0\right) - \operatorname{Tr}\left(\begin{bmatrix} \Gamma_1 & \dots & \Gamma_p \end{bmatrix} \begin{bmatrix} \Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \Gamma_1^T \\ \vdots \\ \Gamma_p^T \end{bmatrix}\right).$$

To write the profiled loss in this way, we will need to replace expectations with atuocovariances as in $\mathbb{E}\left(Z_t Z_{t-h}^T\right) = \Gamma_h$. Thus Assumption 1 (stability) is required.

B.1 Simplifying equalities

To begin, we first compute some expectations that will be needed to simplify the loss. We make note of several identities we will use. First, $\text{vec}(A)^T \text{vec}(B) = \text{Tr}(A^T B)$ and $\text{vec}(ABC) = (C^T \otimes A)\text{vec}(B)$. The j^{th} row and column of Γ_i are denoted as $\Gamma_{i,j\cdot}$ and $\Gamma_{i,\cdot j}$ respectively.

We also note that $\mathbb{E}\left(X_p^T \otimes X_p^T\right) \in \mathbb{R}^{1\times k^2p^2}$ and can be written as

$$\mathbb{E}\left(X_p^T \otimes X_p^T\right) = \mathbb{E}\left(\begin{bmatrix} Z_{t-1,1} & \dots & Z_{t-1,k} & Z_{t-2,1} & \dots & Z_{t-p,k}\end{bmatrix} \otimes \begin{bmatrix} Z_{t-1,1} & \dots & Z_{t-1,k} & Z_{t-2,1} & \dots & Z_{t-p,k}\end{bmatrix}\right)$$
$$= \mathbb{E}\begin{pmatrix} [\, Z_{t-1,1}^2 & \dots & Z_{t-1,1}Z_{t-1,k} & Z_{t-1,1}Z_{t-2,1} & \dots & Z_{t-1,1}Z_{t-2,k} & \dots & Z_{t-1,1}Z_{t-p,k} \\ Z_{t-1,2}Z_{t-1,1} & \dots & Z_{t-1,2}Z_{t-1,k} & Z_{t-1,2}Z_{t-2,1} & \dots & Z_{t-1,2}Z_{t-2,k} & \dots & Z_{t-1,2}Z_{t-p,k} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ Z_{t-1,k}Z_{t-1,1} & \dots & Z_{t-1,k}Z_{t-1,k} & Z_{t-1,k}Z_{t-2,1} & \dots & Z_{t-1,k}Z_{t-2,k} & \dots & Z_{t-1,k}Z_{t-p,k} \\ Z_{t-2,1}Z_{t-1,1} & \dots & Z_{t-2,1}Z_{t-1,k} & Z_{t-2,1}Z_{t-2,1} & \dots & Z_{t-2,1}Z_{t-2,k} & \dots & Z_{t-2,1}Z_{t-p,k} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots\,] \end{pmatrix}.$$

We have

$$\mathbb{E}\left(\begin{bmatrix} Z_{t-1,1}^2 & \dots & Z_{t-1,1}Z_{t-1,k}\end{bmatrix}^T\right) = \Gamma_{0,\cdot 1}.$$

Similarly,

$$\mathbb{E}\left(\begin{bmatrix} Z_{t-1,1}Z_{t-2,1} & \dots & Z_{t-1,1}Z_{t-2,k}\end{bmatrix}^T\right) = \Gamma_{-1,\cdot 1} = (\Gamma_1^T)_{\cdot 1}\,,$$
$$\mathbb{E}\left(\begin{bmatrix} Z_{t-1,1}Z_{t-p,1} & \dots & Z_{t-1,1}Z_{t-p,k}\end{bmatrix}^T\right) = \left(\Gamma_{p-1}^T\right)_{\cdot 1}\,.$$

We define

$$C := \begin{matrix} [\Gamma_{0,\cdot 1} & (\Gamma_1^T)_{\cdot 1} & (\Gamma_2^T)_{\cdot 1} & \dots & (\Gamma_{p-1}^T)_{\cdot 1} \\ \Gamma_{0,\cdot 2} & (\Gamma_1^T)_{\cdot 2} & (\Gamma_2^T)_{\cdot 2} & \dots & (\Gamma_{p-1}^T)_{\cdot 2} \\ \vdots & & & & \vdots \\ \Gamma_{0,\cdot k} & (\Gamma_1^T)_{\cdot k} & (\Gamma_2^T)_{\cdot k} & \dots & (\Gamma_{p-1}^T)_{\cdot k} \\ \Gamma_{1,\cdot 1} & \Gamma_{0,\cdot 1} & (\Gamma_1^T)_{\cdot 1} & \dots & (\Gamma_{p-1}^T)_{\cdot k} \\ \vdots & & & & \vdots \\ \Gamma_{p-1,\cdot k} & \Gamma_{p-2,\cdot k} & \Gamma_{p-3,\cdot k} & \dots & \Gamma_{0,\cdot k}]\,. \end{matrix}$$

Note that each element in C is $\in \mathbb{R}^{k\times 1}$. For example, both $\Gamma_{0,\cdot 1}$ and $(\Gamma_1)_{\cdot 1}^T$ are $\in \mathbb{R}^{k\times 1}$. Thus, $C \in \mathbb{R}^{k\times kp^2}$, and we can write

$$\mathbb{E}\left(X_p^T \otimes X_p^T\right) = \mathrm{vec}(C)^T$$
$$= \mathrm{vec}\left(\begin{bmatrix} \Gamma_0 & \Gamma_1 & \dots & \Gamma_{p-1} \\ \Gamma_1^T & \Gamma_0 & \dots & \Gamma_{p-2} \\ \vdots & & & \vdots \\ \Gamma_{p-1}^T & \Gamma_{p-2}^T & \dots & \Gamma_0 \end{bmatrix}\right)^T.$$

Using a similar idea, we have that

$$\mathbb{E}\left(X_p^T \otimes Y_t^T\right) = \mathbb{E}\left(\begin{bmatrix} Z_{t-1,1} & \dots & Z_{t-1,k} & Z_{t-2,1} & \dots & Z_{t-p,k}\end{bmatrix} \otimes \begin{bmatrix} Z_{t,1} & \dots & Z_{t,k}\end{bmatrix}\right)$$
$$= \mathrm{vec}\left(\begin{bmatrix} \Gamma_1 & \dots & \Gamma_p\end{bmatrix}\right)^T.$$

For space considerations, we do not write out the steps in detail. The method proceeds similarly to the above. Next,

$$\mathbb{E}\left(X_p X_p^T\right) = \mathbb{E}\left(\begin{bmatrix} Z_{t-1} \\ Z_{t-2} \\ \vdots \\ Z_{t-p}\end{bmatrix}\begin{bmatrix} Z_{t-1}^T & Z_{t-2}^T & \dots & Z_{t-p}^T\end{bmatrix}\right) = \mathbb{E}\left(\begin{bmatrix} Z_{t-1}Z_{t-1}^T & Z_{t-1}Z_{t-2}^T & \dots & Z_{t-1}Z_{t-p}^T \\ Z_{t-2}Z_{t-1}^T & Z_{t-2}Z_{t-2}^T & \dots & Z_{t-2}Z_{t-2}^T \\ \vdots & & & \vdots \\ Z_{t-p}Z_{t-1}^T & Z_{t-p}Z_{t-2}^T & \dots & Z_{t-p}Z_{t-p}^T \end{bmatrix}\right)$$
$$= \begin{bmatrix} \Gamma_0 & \Gamma_1 & \dots & \Gamma_{p-1} \\ \Gamma_1^T & \Gamma_0 & \dots & \Gamma_{p-2} \\ \vdots & & & \vdots \\ \Gamma_{p-1}^T & \Gamma_{p-2}^T & \dots & \Gamma_0 \end{bmatrix}.$$

Lastly,

$$\mathbb{E}\left(Y_t X_p^T\right) = \begin{bmatrix}\Gamma_1 & \dots & \Gamma_p\end{bmatrix}.$$

B.2 Simplifying the loss

We can now proceed in simplifying the population loss. First, note that

$$\begin{aligned}\mathbb{E}\left[\left(Y_t - A_p^* X_p\right)^T \left(Y_t - A_p^* X_p\right)\right] &= \mathbb{E}\left(Y_t^T Y_t\right) - 2\mathbb{E}\left(Y_t^T A_p^* X_p\right) + \mathbb{E}\left(X_p^T A_p^{*T} A_p^* X_p\right) \\ &= \operatorname{Tr}\left(\Gamma_0\right) - 2\mathbb{E}\left(X_p^T \otimes Y_t^T\right)\operatorname{vec}\left(A_p^*\right) + \mathbb{E}\left(X_p^T \otimes X_p^T\right)\operatorname{vec}\left(A_p^{*T} A_p^*\right) \\ &= \operatorname{Tr}\left(\Gamma_0\right) - 2\operatorname{Tr}\left(\begin{bmatrix}\Gamma_1^T \\ \vdots \\ \Gamma_p^T\end{bmatrix} A_p^*\right) + \operatorname{Tr}\left(\begin{bmatrix}\Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0\end{bmatrix} A_p^{*T} A_p^*\right).\end{aligned}$$

Now, $A_p^{*T} = \mathbb{E}\left(X_p X_p^T\right)^{-1,T} E\left(Y_t X_p^T\right)^T = \begin{bmatrix}\Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0\end{bmatrix}^{-1,T} \begin{bmatrix}\Gamma_1^T \\ \vdots \\ \Gamma_p^T\end{bmatrix}$. Thus, $\begin{bmatrix}\Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0\end{bmatrix}$ is symmetric and so is its inverse. Therefore, using $\operatorname{Tr}(A^T B) = \operatorname{Tr}(AB^T)$, we get

$$\begin{aligned}\mathbb{E}\left[\left(Y_t - A_p^* X_p\right)^T \left(Y_t - A_p^* X_p\right)\right] &= \operatorname{Tr}\left(\Gamma_0\right) - \operatorname{Tr}\left(\begin{bmatrix}\Gamma_1^T \\ \vdots \\ \Gamma_p^T\end{bmatrix} A_p^*\right) \\ &= \operatorname{Tr}\left(\Gamma_0\right) - \operatorname{Tr}\left(\begin{bmatrix}\Gamma_1 & \dots & \Gamma_p\end{bmatrix} \begin{bmatrix}\Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0\end{bmatrix}^{-1} \begin{bmatrix}\Gamma_1^T \\ \vdots \\ \Gamma_p^T\end{bmatrix}\right).\end{aligned}$$

C Proofs

In this section, we prove Theorems 1 and 2 and Corollary 1. Specifically, Section C.1 focuses on Theorem 1, Section C.2 proves Corollary 1, and Section C.3 proves Theorem 2.

C.1 Flat loss

In this section, we prove Theorem 1 which establishes that the loss decreases until the true order p_0 at which point it remains constant at $\operatorname{Tr}(\Sigma_\epsilon)$. The proof proceeds in several steps. We first relate the loss at fitted order p to the loss at fitted order $p-1$. In the second step, we consider the cases when $p > p_0$ and $p \leq p_0$ separately. Finally, we show that the loss at p_0 is equal to $\operatorname{Tr}(\Sigma_\epsilon)$.

Proof of Theorem 1. Recall that

$$L_{\mathrm{VAR}}(p) := \operatorname{Tr}\left(\Gamma_0\right) - \operatorname{Tr}\left(\begin{bmatrix}\Gamma_1 & \dots & \Gamma_p\end{bmatrix} \begin{bmatrix}\Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0\end{bmatrix}^{-1} \begin{bmatrix}\Gamma_1^T \\ \vdots \\ \Gamma_p^T\end{bmatrix}\right).$$

To express $L_{\mathrm{VAR}}(p)$ as a function of $L_{\mathrm{VAR}}(p-1)$, we partition the matrices as follows

$$\left[\begin{array}{ccc|c}\Gamma_1 & \dots & \Gamma_{p-1} & \Gamma_p\end{array}\right] := \begin{bmatrix}g & \Gamma_p\end{bmatrix},$$

$$\left[\begin{array}{ccc|c} \Gamma_0 & \dots & \Gamma_{p-2} & \Gamma_{p-1} \\ \Gamma_1^T & \dots & \Gamma_{p-3} & \Gamma_{p-2} \\ \vdots & & \vdots & \vdots \\ \Gamma_{p-2}^T & \dots & \Gamma_0 & \vdots \\ \hline \Gamma_{p-1}^T & \dots & \Gamma_1^T & \Gamma_0 \end{array}\right] := \begin{bmatrix} B & C^T \\ C & D \end{bmatrix}. \tag{6}$$

Now, using from the 2 x 2 block matrix inversion formula (Lu and Shiou, 2002),

$$\begin{bmatrix} \Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0 \end{bmatrix}^{-1} = \begin{bmatrix} B & C^T \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1} + B^{-1}C^T HCB^{-1} & -B^{-1}C^T H \\ -HCB^{-1} & H \end{bmatrix},$$

where $H = (D - CB^{-1}C^T)^{-1}$ is the inverse of the Schur-complement of block B. Carrying out the multiplication,

$$\begin{aligned} \begin{bmatrix} \Gamma_1 & \dots & \Gamma_p \end{bmatrix} \begin{bmatrix} \Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \Gamma_1^T \\ \vdots \\ \Gamma_p^T \end{bmatrix} &= \begin{bmatrix} g & \Gamma_p \end{bmatrix} \begin{bmatrix} B^{-1} + B^{-1}C^T HCB^{-1} & -B^{-1}C^T H \\ -HCB^{-1} & H \end{bmatrix} \begin{bmatrix} g^T \\ \Gamma_p^T \end{bmatrix} \\ &= gB^{-1}g^T + gB^{-1}C^T HCB^{-1}g^T - \Gamma_p HCB^{-1}g^T - gB^{-1}C^T H\Gamma_p^T + \Gamma_p H\Gamma_p^T. \end{aligned}$$

Thus,

$$\begin{aligned} L_{\mathrm{VAR}}(p) &= \mathrm{Tr}\left(\Gamma_0\right) - \mathrm{Tr}\left(gB^{-1}g^T + gB^{-1}C^T HCB^{-1}g^T - \Gamma_p HCB^{-1}g^T - gB^{-1}C^T H\Gamma_p^T + \Gamma_p H\Gamma_p^T\right) \\ &= L_{\mathrm{VAR}}(p-1) - \mathrm{Tr}\left(gB^{-1}C^T HCB^{-1}g^T - \Gamma_p HCB^{-1}g^T - gB^{-1}C^T H\Gamma_p^T + \Gamma_p H\Gamma_p^T\right) \\ &= L_{\mathrm{VAR}}(p-1) - \mathrm{Tr}\left(gB^{-1}C^T HCB^{-1}g^T\right) + \mathrm{Tr}\left(\Gamma_p HCB^{-1}g^T\right) + \mathrm{Tr}\left(gB^{-1}C^T H\Gamma_p^T\right) - \mathrm{Tr}\left(\Gamma_p H\Gamma_p^T\right) \\ &= L_{\mathrm{VAR}}(p-1) - \mathrm{Tr}\left(\left(gB^{-1}C^T - \Gamma_p\right) H \left(gB^{-1}C^T - \Gamma_p\right)^T\right). \end{aligned}$$

Note that this relationship holds in general. To get this result we only rely on Assumption 1 (stability) and Assumption 2 (invertibility).

Now we use that the true process is VAR(p_0) to study the loss when $p > p_0$. If $p > p_0$, from the Yule-Walker equations,

$$\begin{bmatrix} \Gamma_1 & \dots & \Gamma_{p-1} \end{bmatrix} = \begin{bmatrix} A_1 & \dots & A_{p_0} \end{bmatrix} \begin{bmatrix} \Gamma_0 & \Gamma_1 & \dots & \Gamma_{p-2} \\ \Gamma_{-1} & \Gamma_0 & \dots & \Gamma_{p-3} \\ \vdots & & & \vdots \\ \Gamma_{-(p_0-1)} & \Gamma_{-(p_0-2)} & \dots & \Gamma_{-(p_0-(p-1))} \end{bmatrix}.$$

We can extend this to

$$\begin{bmatrix} \Gamma_1 & \dots & \Gamma_{p-1} \end{bmatrix} = \begin{bmatrix} A_1 & \dots & A_{p_0} & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \Gamma_0 & \Gamma_1 & \dots & \Gamma_{p-2} \\ \Gamma_{-1} & \Gamma_0 & \dots & \Gamma_{p-3} \\ \vdots & & & \vdots \\ \Gamma_{2-p} & \Gamma_{3-p} & \dots & \Gamma_0 \end{bmatrix}.$$

Using that $\Gamma_{-i} = \Gamma_i^T$ and substituting in the definitions of B and g and under Assumption 2 (invertibility),

$$gB^{-1}C^T = \begin{bmatrix} A_1 & \dots & A_{p_0} & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \Gamma_{p-1} \\ \vdots \\ \Gamma_1 \end{bmatrix} = \Gamma_p.$$

Thus, for any $p > p_0$

$$L_{\mathrm{VAR}}(p) = L_{\mathrm{VAR}}(p-1).$$

Next, we study the loss when $p \leq p_0$. From Assumption 2 (invertibility), the matrix

$$\begin{bmatrix} \Gamma_0 & \dots & \Gamma_{p-1} \\ \vdots & & \vdots \\ \Gamma_{p-1}^T & \dots & \Gamma_0 \end{bmatrix}$$

is positive definite for all $p = 1, \dots, p_{\max}$. From Proposition 2.2 of Gallier et al. (2020), the Schur-complement of block B in Eq. (6) is also positive definite and hence so is the inverse, H. From Assumption 3 (irreducibility), when $p \leq p_0$,

$$gB^{-1}C^T - \Gamma_p := \Delta \neq 0\,.$$

Recall our equation relating $L_{\text{VAR}}(p)$ to $L_{\text{VAR}}(p-1)$:

$$L_{\text{VAR}}(p) = L_{\text{VAR}}(p-1) - \text{Tr}\left(\left(gB^{-1}C^T - \Gamma_p\right) H \left(gB^{-1}C^T - \Gamma_p\right)^T\right)\,.$$

Using δ_i^T to denote the i^{th} row of Δ,

$$\text{Tr}\left(\Delta H \Delta^T\right) = \sum_{i=1}^{k} \delta_i^T H \delta_i\,.$$

Since H is positive definite and $\delta_i \neq 0$ for at least one $i = 1, \dots, k$, $\text{Tr}\left(\Delta H \Delta^T\right) > 0$ and thus

$$L_{\text{VAR}}(p) < L_{\text{VAR}}(p-1)\,.$$

Lastly, we show that the loss flattens out at $\text{Tr}(\Sigma_\epsilon)$. Since $L_{\text{VAR}}(p) = L_{\text{VAR}}(p-1)$ when $p > p_0$, it suffices to show that $L_{\text{VAR}}(p_0) = \text{Tr}(\Sigma_\epsilon)$:

$$\begin{aligned} L_{\text{VAR}}(p_0) &= \text{Tr}\left(\Gamma_0\right) - \text{Tr}\left(\begin{bmatrix}\Gamma_1 & \dots & \Gamma_{p_0}\end{bmatrix} \begin{bmatrix} \Gamma_0 & \dots & \Gamma_{p_0-1} \\ \vdots & & \vdots \\ \Gamma_{p_0-1}^T & \dots & \Gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \Gamma_1^T \\ \vdots \\ \Gamma_{p_0}^T \end{bmatrix}\right) \\ &= \text{Tr}(\Gamma_0) - \text{Tr}\left(\sum_{i=1}^{p_0} A_i \Gamma_{-i}\right) \\ &= \text{Tr}\left(\Gamma_0 - (\Gamma_0 - \Sigma_\epsilon)\right) \\ &= \text{Tr}(\Sigma_\epsilon)\,, \end{aligned}$$

where in the second line we use $\Gamma_i^T = \Gamma_{-i}$, and in the third line we use the form of Γ_0 from Lütkepohl (2005, Eq. (2.1.36)).

□

C.2 Penalized loss recovers true order

In this section, we prove Corollary 1 which states that, for the correct choice of penalty, the true VAR order can be recovered by penalizing the population loss.

Proof of Corollary 1. We first show that for $p > p_0$, $L_{\text{VAR}}(p) + \lambda p > L_{\text{VAR}}(p_0) + \lambda p_0$. This follows immediately by noting that $L_{\text{VAR}}(p) - L_{\text{VAR}}(p_0) = 0$ from Theorem 1 and $\lambda p > \lambda p_0$ for $p > p_0$ as $\lambda > 0$. Next, we show that for $p < p_0$, $L_{\text{VAR}}(p) + \lambda p > L_{\text{VAR}}(p_0) + \lambda p_0$. Since $p = p_0 - i$ for some i, it suffices to show that $\lambda < \left(L_{\text{VAR}}(p_0 - i) - L_{\text{VAR}}(p_0)\right)/i$ for every $i = 1, \dots, p_0$. This holds by definition of λ in Corollary 1. □

C.3 Consistency of sample loss

Next, we prove Theorem 2 which establishes that $\hat{L}_{\text{VAR}}(p)$ converges to the population loss $L_{\text{VAR}}(p)$. This is proved using Lemma 1 and the convergence of $\hat{\Gamma}_i$ to Γ_i.

Lemma 1. *Consider the function*

$$f(G) = \mathrm{Tr}\left(\Gamma_0^0 - \begin{bmatrix} \Gamma_1^0 & \dots & \Gamma_p^0 \end{bmatrix} \begin{bmatrix} \Gamma_0^1 & \Gamma_1^1 & \Gamma_2^1 & \dots & \Gamma_{p-1}^1 \\ (\Gamma_1^1)^T & \Gamma_0^2 & \Gamma_1^2 & \dots & \Gamma_{p-2}^2 \\ (\Gamma_2^1)^T & (\Gamma_1^2)^T & \Gamma_0^3 & \dots & \Gamma_{p-3}^3 \\ \vdots & & & \dots & \vdots \\ (\Gamma_{p-1}^1)^T & (\Gamma_{p-2}^2)^T & (\Gamma_{p-3}^3)^T & \dots & \Gamma_0^p \end{bmatrix}^{-1} \begin{bmatrix} (\Gamma_1^0)^T \\ \vdots \\ (\Gamma_p^0)^T \end{bmatrix} \right),$$

where the elements of the input vector $G \in \mathbb{R}^{\frac{k^2(p+2)(p+1)}{2}}$ *are denoted as*

$$G = \begin{bmatrix} \vec{\Gamma}_0^0 & \vec{\Gamma}_1^0 & \dots & \vec{\Gamma}_p^0 & \vec{\Gamma}_0^1 & \dots & \vec{\Gamma}_0^p & \vec{\Gamma}_1^1 & \dots & \vec{\Gamma}_1^{p-1} & \vec{\Gamma}_2^1 & \dots & \vec{\Gamma}_2^{p-2} & \dots & \vec{\Gamma}_{p-1}^1 \end{bmatrix},$$

and $\vec{\Gamma} := \mathrm{vec}(\Gamma)^T$ *is used to define the row vector formed by stacking the columns of* Γ *and transposing. The subscripts represent the autocovariance lag of interest while the superscripts represent (possibly) different estimates of that autocovariance. That is,* Γ_h^i *represents the* i^{th} *estimate of the* h^{th} *autocovariance.*

If G *and* $\hat{G}$ *are defined such that the corresponding inverses in* $f(G)$, $f(\hat{G})$ *and* $f((1-\nu)G + \nu\hat{G})$ *exist for all* $\nu \in (0,1)$*, then* $f(G)$ *is Lipschitz continuous with respect to* G*. That is*

$$\left| f(\hat{G}) - f(G) \right| \leq L \left\| \hat{G} - G \right\|_1,$$

for some $L < \infty$*.*

The proof of Lemma 1 is deferred until after the proof of Theorem 2 for clarity of presentation. The proof of Theorem 2 proceeds by first showing that the least squares estimator of the loss can be expressed in terms of many different autocovariance estimators. Next, we show that the autocovariance estimators in the least squares estimator of the loss are asymptotically equivalent to the autocovariance estimators that use all available data. Finally, this fact is combined with Lemma 1 to establish consistency and the rate of the least squares estimator of the loss.

Proof of Theorem 2. Consider our least squares estimator of the loss,

$$\begin{aligned} \hat{L}_{\mathrm{VAR}}(p) &= \frac{1}{n-p} \mathrm{Tr}\left(\left(\mathbf{Y}_p - \hat{A}_p \mathbf{X}_p \right) \left(\mathbf{Y}_p - \hat{A}_p \mathbf{X}_p \right)^T \right) \\ &= \frac{1}{n-p} \mathrm{Tr}\left(\mathbf{Y}_p \mathbf{Y}_p^T - \mathbf{Y}_p \mathbf{X}_p^T \left(\mathbf{X}_p \mathbf{X}_p^T \right)^{-1} \mathbf{X}_p \mathbf{Y}_p^T \right) \\ &= \mathrm{Tr}\left(\frac{1}{n-p} \mathbf{Y}_p \mathbf{Y}_p^T - \left(\frac{1}{n-p} \mathbf{Y}_p \mathbf{X}_p^T \right) \left(\frac{1}{n-p} \mathbf{X}_p \mathbf{X}_p^T \right)^{-1} \left(\frac{1}{n-p} \mathbf{X}_p \mathbf{Y}_p^T \right) \right), \end{aligned}$$

where for de-meaned data,

$$\begin{aligned} \frac{1}{n-p} \mathbf{Y}_p \mathbf{Y}_p^T &= \frac{1}{n-p} \sum_{t=1+p}^{n} (\mathbf{z}_t - \overline{\mathbf{z}})(\mathbf{z}_t - \overline{\mathbf{z}})^T \\ \frac{1}{n-p} \mathbf{Y}_p \mathbf{X}_p^T &= \frac{1}{n-p} \sum_{t=1+p}^{n} (\mathbf{z}_t - \overline{\mathbf{z}}) (\mathbf{x}_{t,p} - \overline{\mathbf{z}})^T \\ &= \frac{1}{n-p} \begin{bmatrix} \sum_{t=1+p}^{n} (\mathbf{z}_t - \overline{\mathbf{z}}) (\mathbf{z}_{t-1} - \overline{\mathbf{z}})^T & \dots & \sum_{t=1+p}^{n} (\mathbf{z}_t - \overline{\mathbf{z}}) \left(\mathbf{z}_{t-p}^T - \overline{\mathbf{z}} \right) \end{bmatrix} \\ \frac{1}{n-p} \mathbf{X}_p \mathbf{X}_p^T &= \frac{1}{n-p} \sum_{t=1+p}^{n} \begin{bmatrix} (\mathbf{z}_{t-1} - \overline{\mathbf{z}}) \\ \vdots \\ (\mathbf{z}_{t-p} - \overline{\mathbf{z}}) \end{bmatrix} \begin{bmatrix} (\mathbf{z}_{t-1} - \overline{\mathbf{z}})^T & \dots & (\mathbf{z}_{t-p} - \overline{\mathbf{z}})^T \end{bmatrix} \\ &= \frac{1}{n-p} \sum_{t=1+p}^{n} \begin{bmatrix} (\mathbf{z}_{t-1} - \overline{\mathbf{z}}) (\mathbf{z}_{t-1} - \overline{\mathbf{z}})^T & (\mathbf{z}_{t-1} - \overline{\mathbf{z}}) (\mathbf{z}_{t-2} - \overline{\mathbf{z}})^T & \dots & (\mathbf{z}_{t-1} - \overline{\mathbf{z}}) (\mathbf{z}_{t-p} - \overline{\mathbf{z}})^T \\ & \vdots & & \vdots \\ (\mathbf{z}_{t-p} - \overline{\mathbf{z}}) (\mathbf{z}_{t-1} - \overline{\mathbf{z}})^T & (\mathbf{z}_{t-p} - \overline{\mathbf{z}}) (\mathbf{z}_{t-2} - \overline{\mathbf{z}})^T & \dots & (\mathbf{z}_{t-p} - \overline{\mathbf{z}}) (\mathbf{z}_{t-p} - \overline{\mathbf{z}})^T \end{bmatrix}. \end{aligned}$$

By examining $\frac{1}{n-p}\mathbf{X}_p\mathbf{X}_p^T$ we can see that slightly different estimates of the same autocovariance are used. For example, the diagonals of $\frac{1}{n-p}\mathbf{X}_p\mathbf{X}_p^T$ show that p different estimates of Γ_0 are used. Similarly $p-1$ different estimates of Γ_1 are used and $p-2$ different estimates of Γ_2 are used and so on. All the matrices used in estimating the least squares loss can be written as

$$\hat{\Gamma}_{l-m,p} = \frac{1}{n-p}\sum_{t=1+p}^{n}(\mathbf{z}_{t-m}-\overline{\mathbf{z}})(\mathbf{z}_{t-l}-\overline{\mathbf{z}})^T \quad \text{for } l \in \{0,1,\ldots,p\}, 0 \leq m \leq l\,.$$

These estimators do not use all available data to estimate the autocovariances. For example, to estimate the autocovariance $l-m=1$, it is possible to use observations $2,\ldots,n$. However, the least squares estimator uses only observations $p+1,\ldots,n$. We denote the autocovariance estimates that use all available data as

$$\tilde{\Gamma}_{l-m} = \frac{1}{n}\sum_{t=(l-m)+1}^{n}(\mathbf{z}_t-\overline{\mathbf{z}})(\mathbf{z}_{t-(l-m)}-\overline{\mathbf{z}})^T\,.$$

Let $\tilde{\gamma}_{l-m} = (\tilde{\Gamma}_{l-m})_{ij}$ and $\gamma_{l-m} = (\Gamma_{l-m})_{ij}$. Then, under the assumptions of Theorem 1, it is shown in Quinn (1980) that $n^{1/2-\delta}(\tilde{\gamma}_{l-m}-\gamma_{l-m})$ converges almost surely to 0 for all $\delta > 0$. Thus, $n^{1/2-\delta}(\tilde{\gamma}_{l-m}-\gamma_{l-m})$ also converges in probability to 0, which implies that $|\tilde{\gamma}_{l-m}-\gamma_{l-m}| = o_P(n^{-1/2+\delta})$. These rates are not immediately applicable to $\hat{\gamma}_{l-m,p}$ so we next establish that $|\hat{\gamma}_{l-m,p}-\gamma_{l-m}| = o_P(n^{-1/2+\delta})$ by simplifying $\hat{\Gamma}_{l-m,p}$. We start by re-indexing the sum in $\hat{\Gamma}_{l-m,p}$ with $j = t-m$. Then,

$$\hat{\Gamma}_{l-m,p} = \frac{1}{n-p}\sum_{j=p-m+1}^{n-m}(\mathbf{z}_j-\overline{\mathbf{z}})(\mathbf{z}_{j-(l-m)}-\overline{\mathbf{z}})^T\,.$$

With this formulation, it is easy to see that the difference between $\tilde{\Gamma}_{l-m}$ and $\hat{\Gamma}_{l-m,p}$ is that $\tilde{\Gamma}_{l-m}$ has the additional indices $t = \{(l-m)+1,\ldots,p-m\}$ and $t = \{n-m+1,\ldots,n\}$ in the sum and also uses a scaling of $1/n$ instead of $1/(n-p)$. Note that the number of the extra terms does not grow with n and instead only depends on l,m,p. Thus, these terms are $o(1/n)$. The explicit relationship between the two is given by

$$\tilde{\Gamma}_{l-m} = \frac{n-p}{n}\left(\hat{\Gamma}_{l-m,p} + \frac{1}{n-p}\left[\sum_{j=(l-m)+1}^{p-m}(\mathbf{z}_j-\overline{\mathbf{z}})(\mathbf{z}_{j-(l-m)}-\overline{\mathbf{z}})^T + \sum_{j=n-m+1}^{n}(\mathbf{z}_j-\overline{\mathbf{z}})(\mathbf{z}_{j-(l-m)}-\overline{\mathbf{z}})^T\right]\right). \tag{7}$$

Using Eq. (7), we can write

$$\begin{aligned}
|\hat{\gamma}_{l-m,p}-\gamma_{l-m}| &= \left|\frac{n}{n-p}\tilde{\gamma}_{l-m} - o(n^{-1}) - \gamma_{l-m}\right| \\
&= \left|\frac{n-p}{n-p}\tilde{\gamma}_{l-m} + \frac{p}{n-p}\tilde{\gamma}_{l-m} - \gamma_{l-m} - o(n^{-1})\right| \\
&\leq |\tilde{\gamma}_{l-m}-\gamma_{l-m}| + o_P(n^{-1}) + o(n^{-1}) \\
&= o_P(n^{-1/2+\delta})\,,
\end{aligned}$$

where in the second to last line we used the fact that $\tilde{\gamma}_{l-m}$ converges in probability so $\tilde{\gamma}_{l-m}p/(n-p) = o_P(n^{-1})$. Since Assumption 2 holds by the assumptions of Theorem 1 and the data is drawn from a continuous distribution, we can apply Lemma 1 to see that

$$\begin{aligned}
|\hat{L}_{\text{VAR}}(p) - L_{\text{VAR}}(p)| &\leq L\left\|\hat{G}-G^*\right\|_1 \\
|\hat{L}_{\text{VAR}}(p) - L_{\text{VAR}}(p)| &\leq L\frac{k^2(p+2)(p+1)}{2}o_P(n^{-1/2+\delta}) \\
|\hat{L}_{\text{VAR}}(p) - L_{\text{VAR}}(p)| &= o_P(n^{-1/2+\delta})\,.
\end{aligned}$$

Note that the $\hat{G}$ is the input formed by the relevant autocovariance matrices used in the least squares loss and G^* is the input consisting of the true autocovariance matrices as given in Eq. (8). □

Proof of Lemma 1. For ease of notation, let

$$V(G) := \begin{bmatrix} \Gamma_1^0 & \dots & \Gamma_p^0 \end{bmatrix},$$
$$T(G) := \begin{bmatrix} \Gamma_0^1 & \dots & \Gamma_{p-1}^1 \\ \vdots & & \vdots \\ (\Gamma_{p-1}^1)^T & \dots & \Gamma_0^p \end{bmatrix}.$$

With this, we have that,

$$f(G) = \operatorname{Tr}\left(\Gamma_0^0 - V(G)T(G)V(G)^T .\right).$$

To simplify notation further, we will suppress the dependence on G in the notation of $V(G), T(G)$ and use the shorthands $V := V(G)$ and $T := T(G)$. Let G_i be the i^{th} element of G. Using $\frac{\partial \operatorname{Tr}(h(G))}{\partial G_i} = \operatorname{Tr}\left(\frac{\partial h(G)}{\partial G_i}\right)$, we have that

$$\frac{\partial f(G)}{\partial G_i} = \operatorname{Tr}\left(\frac{\partial \Gamma_0^0}{\partial G_i} - \frac{\partial V}{\partial G_i}T^{-1}V^T - V\frac{\partial T^{-1}}{\partial G_i}V^T - VT^{-1}\frac{\partial V^T}{\partial G_i}\right).$$

From (59) of Petersen et al. (2008), $\frac{\partial T^{-1}}{\partial G_i} = -T^{-1}\frac{\partial T}{\partial G_i}T^{-1}$. Thus,

$$\frac{\partial f(G)}{\partial G_i} = \operatorname{Tr}\left(\frac{\partial \Gamma_0^0}{\partial G_i} - \frac{\partial V}{\partial G_i}T^{-1}V^T + VT^{-1}\frac{\partial T}{\partial G_i}T^{-1}V^T - VT^{-1}\frac{\partial V^T}{\partial G_i}\right).$$

Note that each of $\frac{\partial \Gamma_0^0}{\partial G_i}, \frac{\partial V}{\partial G_i}, \frac{\partial T}{\partial G_i}$ are matrices containing 1 in the entries of Γ_0^0, V, T where G_i is present and 0 otherwise.

While matrix multiplication and the trace are continuous functions in general, matrix inversion is a continuous function only over the set of invertible matrices. However since we have assumed that $G, \hat{G}$ are such that the inverses exist, we conclude that $f(G)$ is continuously differentiable. By the mean value theorem and Hölder's inequality

$$\left|f(\hat{G}) - f(G)\right| \leq \left\|\nabla f((1-\nu)G + \nu\hat{G})\right\|_\infty \left\|\hat{G} - G\right\|_1,$$

for some $\nu \in (0,1)$. Since $\frac{\partial f(G)}{\partial G_i}$ is continuous for each G_i, it is bounded on any closed interval including between G and $\hat{G}$. Thus $\left\|\nabla f((1-\nu)G + \nu\hat{G})\right\|_\infty$ is bounded and

$$\left|f(\hat{G}) - f(G)\right| \leq L\left\|\hat{G} - G\right\|_1,$$

for some $L < \infty$. □

Remark 1. As shown in the proof of Theorem 2, the least squares estimator, $\hat{A}_p$, uses multiple estimates of the autocovariances Γ_i. Thus, it is necessary to define G and $f(G)$ in Lemma 1 to allow for multiple estimators of autocovariances Γ_i. In practice, however, the estimates of each of the autocvoariances Γ_i used in $\hat{A}_p$ are nearly identical.

Remark 2. It is worth noting that if we define

$$G^* = \begin{bmatrix} \vec{\Gamma}_0 & \vec{\Gamma}_1 & \dots & \vec{\Gamma}_p & \vec{\Gamma}_0 & \dots & \vec{\Gamma}_0 & \vec{\Gamma}_1 & \dots & \vec{\Gamma}_1 & \vec{\Gamma}_2 & \dots & \vec{\Gamma}_2 & \dots & \vec{\Gamma}_{p-1} \end{bmatrix}, \tag{8}$$

where Γ_i are the i^{th} population autocovariances, then under Assumption 2, $f(G^*) = L_{\mathrm{VAR}}(p)$. That is, $f(G^*)$ is the expected squared error loss at fitted order p.

Remark 3. We require G and $\hat{G}$ to be such that the inverses in $f(G), f(\hat{G}), f((1-\nu)G + \nu\hat{G})$ exist for all $\nu \in (0,1)$, as the inverse function is only continuous over the set of invertible matrices. In our theoretical analysis of Theorem 2, this holds by Assumption 2 and the fact that $\hat{G}$ is generated from a continuous distribution so that the inverse in $f(\hat{G})$ exists with probability 1 and similarly the inverse in $f((1-\nu)G + \nu\hat{G})$ is also exists with probability 1 for all $\nu \in (0,1)$.

To better understand teh structure of $\frac{\partial \Gamma_0^0}{\partial G_i}, \frac{\partial V}{\partial G_i}, \frac{\partial T}{\partial G_i}$ mentioned in the proof of Lemma 1, we consider two examples.

Example 1. For the first example consider taking the partial derivative with respect to the first component, $G_1 = (\Gamma_0^0)_{1,1}$. That is, G_1 is the $(1,1)$ entry of the Γ_0^0 parameter. Then, the partial derivatives are

$$\frac{\partial \Gamma_0^0}{\partial G_1} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix},$$
$$\frac{\partial V}{\partial G_1} = 0\,,$$
$$\frac{\partial T}{\partial G_1} = 0\,.$$

Example 2. For the second example consider taking the partial derivative with respect to the $(k^2(2p+1)+1)$-th component in G. We have that $G_{k^2(2p+1)+1} = (\Gamma_1^1)_{1,1}$. The resulting partial derivatives are

$$\frac{\partial \Gamma_0^0}{\partial G_{k^2(2p+1)+1}} = 0\,,$$
$$\frac{\partial V}{\partial G_{k^2(2p+1)+1}} = 0\,,$$
$$\frac{\partial T}{\partial G_{k^2(2p+1)+1}} = \begin{bmatrix} & & & \\ & 0_{k\times k} & & \\ & & & \end{bmatrix} \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & & \dots & \\ 0 & 0 & \dots & 0 \end{bmatrix} \begin{matrix} \\ \dots \\ \end{matrix} \begin{matrix} \\ 0_{k\times k} \\ \end{matrix}$$
$$\begin{bmatrix} & & & \\ & 0_{k\times k} & & \\ & & & \end{bmatrix} \begin{matrix} \\ 0_{k\times k} \\ \end{matrix} \begin{matrix} \\ \dots \\ \end{matrix} \begin{matrix} \\ 0_{k\times k} \\ \end{matrix}$$
$$\begin{bmatrix} & & & \\ & \vdots & & \\ & & & \end{bmatrix} \begin{matrix} \\ \ddots \\ \end{matrix} \begin{matrix} \\ \vdots \\ \end{matrix}$$
$$\begin{bmatrix} & & & \\ & 0_{k\times k} & & \\ & & & \end{bmatrix} \begin{matrix} \\ 0_{k\times k} \\ \end{matrix} \begin{matrix} \\ \dots \\ \end{matrix} \begin{matrix} \\ 0_{k\times k} \\ \end{matrix}.$$

D Additional simulation results

In this section, we present additional simulation results. In Figure S7 we compare order selection by minimizing MIC with λ_{ST} (denoted as MIC) to the alternative procedures, MIC-sp and MIC-mt, from Section 4.2. Recall that MIC-sp uses MIC with a penalty λ_{sp} that is chosen by using a 70-30 train-test split of the data while MIC-mt also uses a 70-30 split but chooses the order $0, \dots, p_{\max}$ that minimizes the test error. Overall we see that MIC offers the best performance over all settings and sample sizes. In Figure S8 we study the performance of each order selection method for large sample sizes in the $\mathrm{VAR}_3(2)$ switching simulation setting. MIC is the only method that consistently estimates the true order for all sample sizes. Lastly, as discussed in Section 5.2, we compare the likelihood of over and under selection of the true model order for MIC using an oracle λ to each of the competing order selection methods. The results in Figures S9 to S11 show that MIC-oracle tends to over select the true order while the alternative methods tend to under select the true order. Moreover, as seen in Figure S11, the probability of over selection in MIC-oracle may be worse than the corresponding probability of under selection from AIC, BIC, and HQ when the dimension is large and the errors are non-diagonal Gaussian.

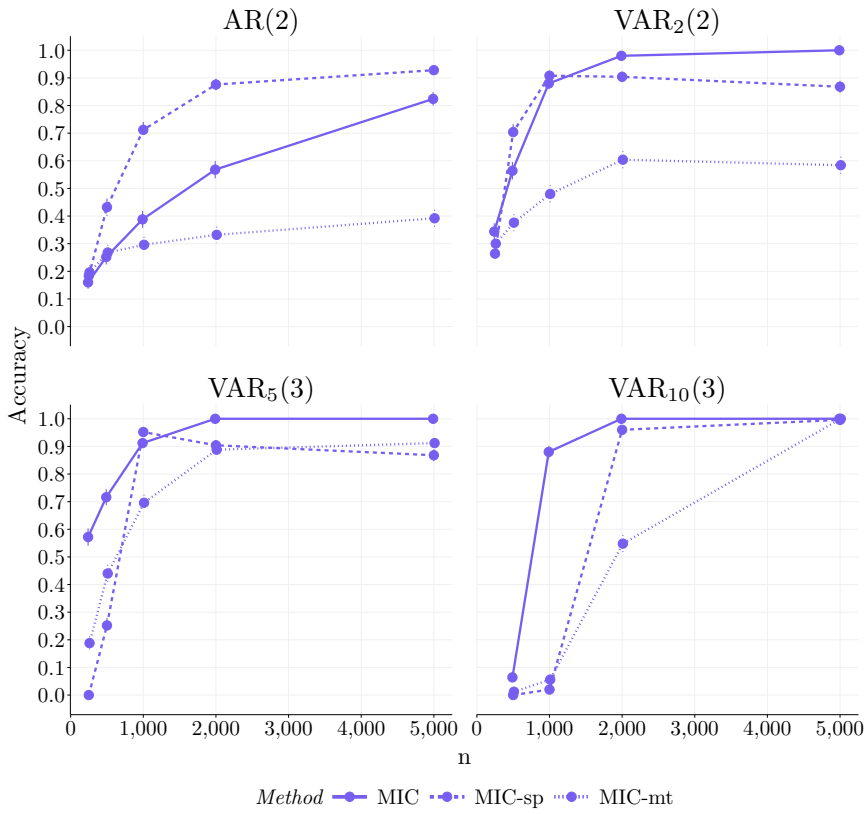

Figure S7: **Diagonal Gaussian errors**. Simulation results comparing accuracy of different MIC order selection methods. Vertical lines indicate standard errors.

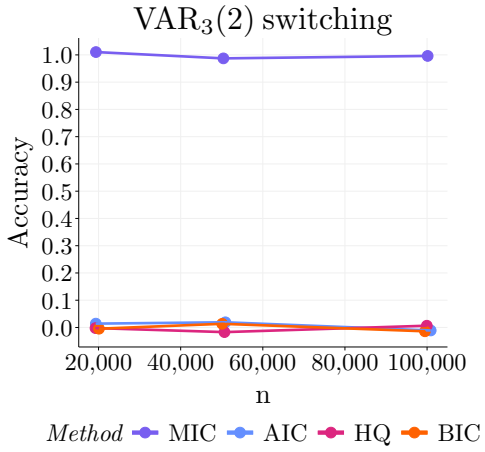

Figure S8: **VAR$_3$(2) switching**. Large sample size simulation results. Vertical lines indicate standard errors. Points have been jittered by 0.02 in the vertical and 1000 in the horizontal directions to improve readability.

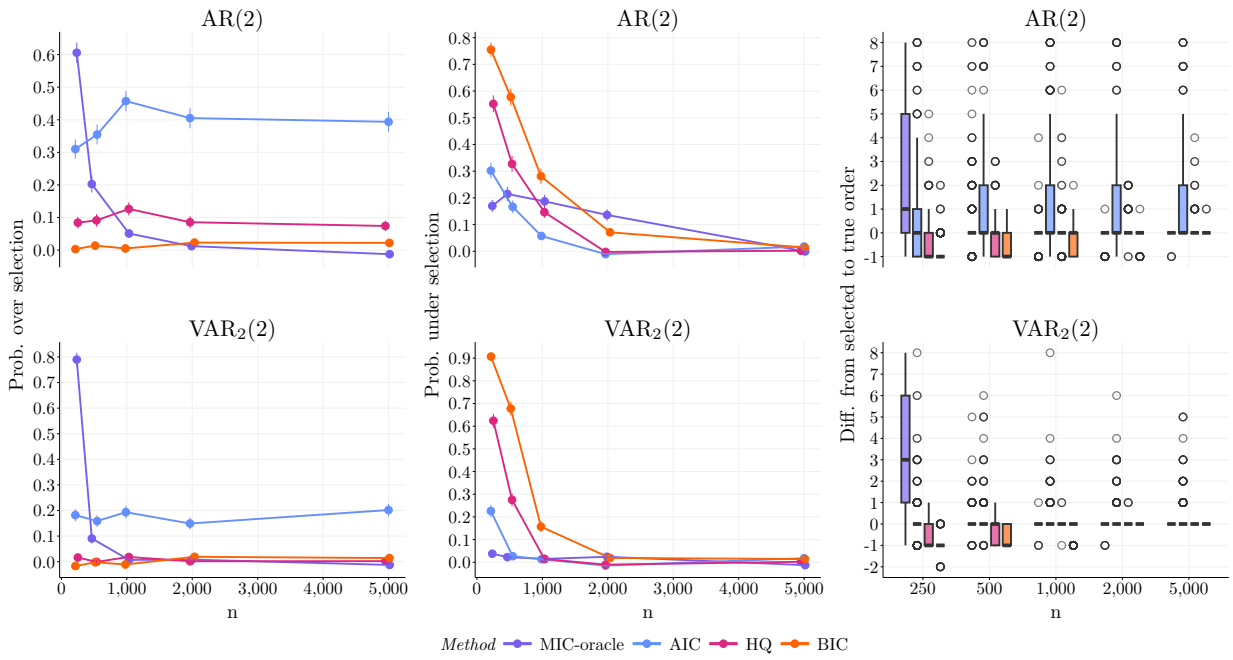

Figure S9: **Diagonal Gaussian errors**. Simulation results for over and under selection of order with diagonal Gaussian errors. Vertical lines indicate standard errors. Points in the over and under selection plots have been jittered by 0.02 in the vertical and 50 in the horizontal directions to improve readability.

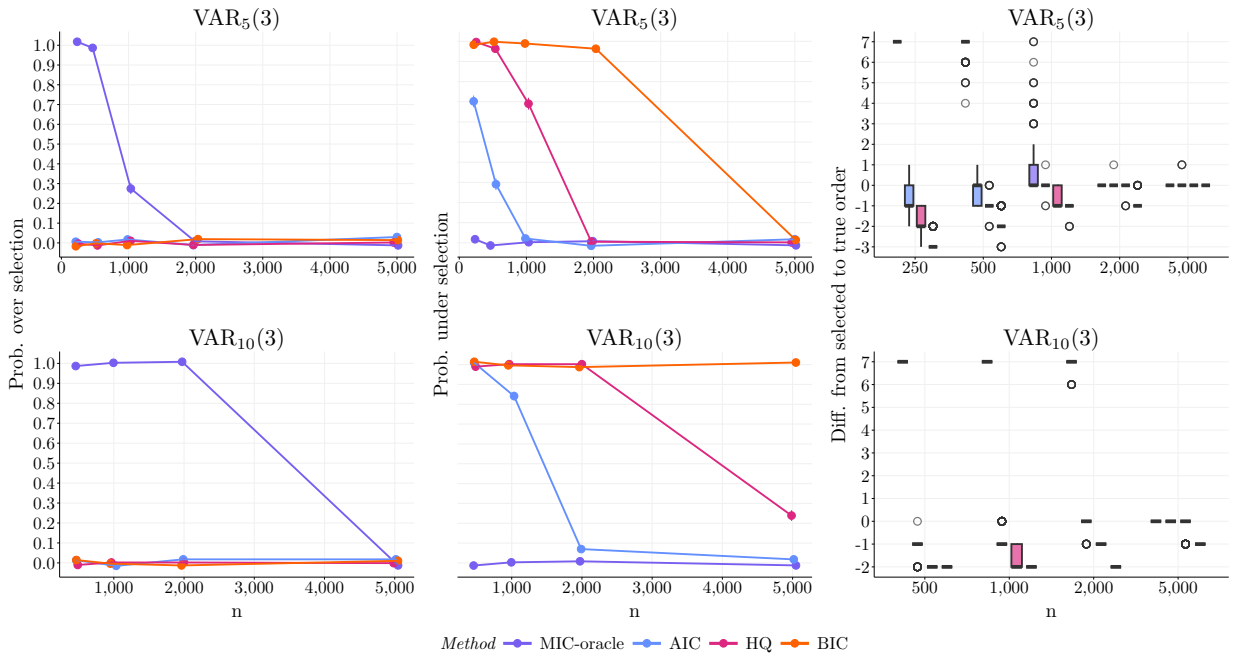

Figure S10: **Diagonal Gaussian errors**. Simulation results for over and under selection of order with diagonal Gaussian errors. Vertical lines indicate standard errors. Points in the over and under selection plots have been jittered by 0.02 in the vertical and 50 in the horizontal directions to improve readability.

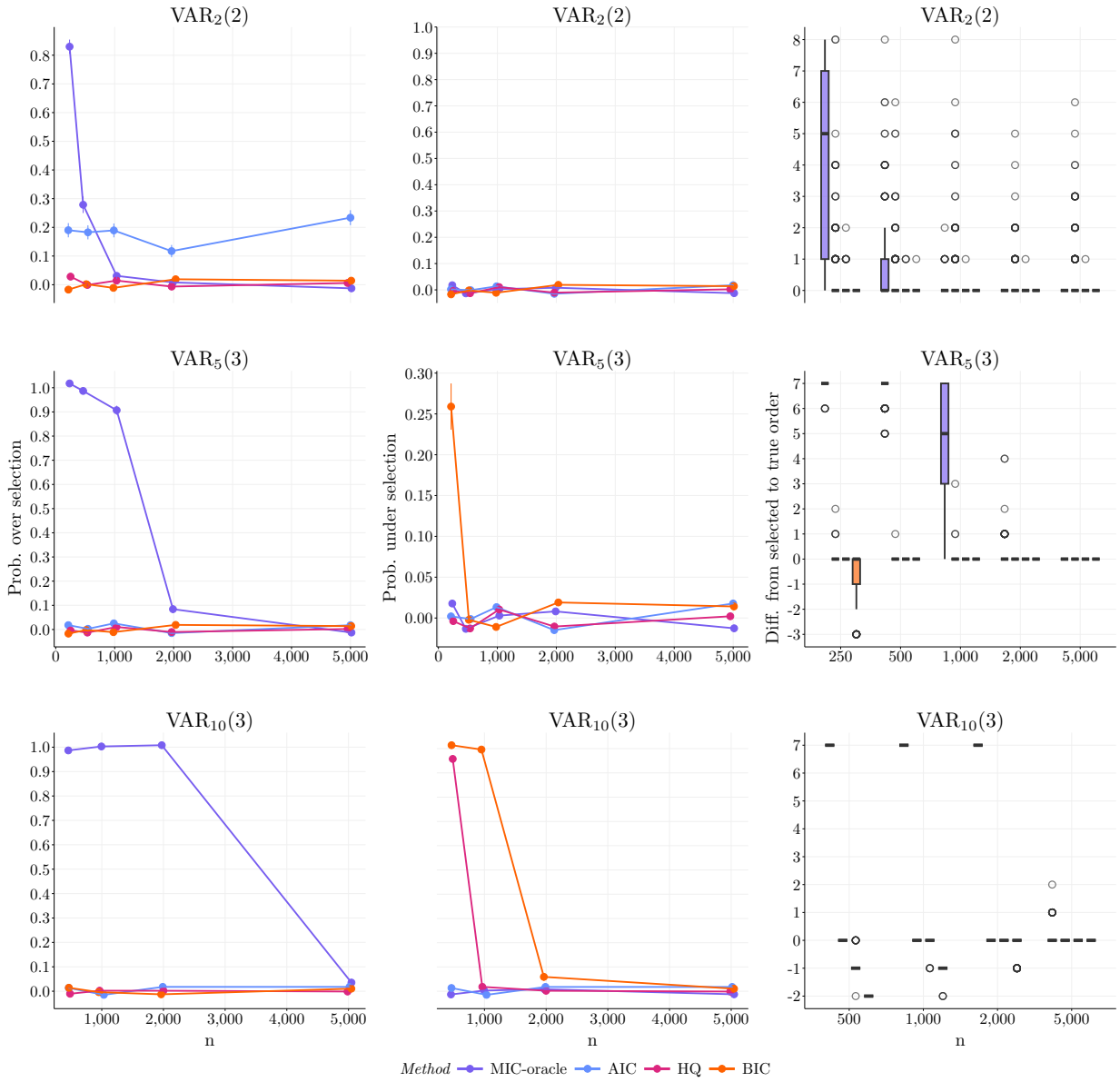

Figure S11: **Non-diagonal Gaussian errors**. Simulation results for over and under selection of order with non-diagonal Gaussian errors. Vertical lines indicate standard errors. Points in the over and under selection plots have been jittered by 0.02 in the vertical and 50 in the horizontal directions to improve readability.

E Daily realized stock variances

Table S3 shows the stocks and their corresponding descriptions that were included in the financial forecasting analysis from Section 6.1. A check mark in the $k = 16$ or $k = 7$ column indicates that the stock was included in the corresponding analysis shown in Table 1.

Stock	Description	$k = 16$	$k = 7$
AEX	Amsterdam Exchange Index	✓	
AORD	All Ordinaries Index	✓	
BFX	Belgium Bell 20 Index	✓	
BVSP	BOVESPA Index	✓	
DJI	Dow Jones Industrial Average	✓	
FCHI	Cotation Assist´ee en Continu Index	✓	✓
FTSE	Financial Times Stock Exchange Index 100	✓	✓
GDAXI	Deutscher Aktienindex	✓	✓
HSI	HANG SENG Index	✓	✓
IXIC	Nasdaq stock index	✓	
KS11	Korea Composite Stock Price Index	✓	✓
MXX	IPC Mexico	✓	
N225	Tokyo stock exchange index		✓
RUT	Russel 2000	✓	
SPX	Standard & Poor's 500 market index	✓	✓
SSMI	Swiss market index	✓	
STOXX50E	EURO STOXX 50	✓	

Table S3: Stocks analyzed in financial application