VAR Order Selection

Umidjon Abdullaev, Ulrich Gunter, Miaomiao Yan

Vector Autoregressive Models

January 16th 2008
Agenda

1. Introduction
Agenda

1. Introduction

2. A Sequence of Tests for Determining the VAR Order
   - The Impact of the Fitted VAR Order on the Forecast MSE
   - The Likelihood Ratio Test Statistic
   - A Testing Scheme for VAR Order Determination
Agenda

1. Introduction

2. A Sequence of Tests for Determining the VAR Order
   - The Impact of the Fitted VAR Order on the Forecast MSE
   - The Likelihood Ratio Test Statistic
   - A Testing Scheme for VAR Order Determination

3. Criteria for VAR Order Selection
   - Minimizing the Forecast MSE
   - Akaike’s Information Criterion (AIC)
   - Hannan-Quinn Criterion (HQ)
   - Schwarz Criterion (SC)
Introduction

A Sequence of Tests for Determining the VAR Order

The Impact of the Fitted VAR Order on the Forecast MSE

The Likelihood Ratio Test Statistic

A Testing Scheme for VAR Order Determination

Criteria for VAR Order Selection

Minimizing the Forecast MSE

Akaike’s Information Criterion (AIC)

Hannan-Quinn Criterion (HQ)

Schwarz Criterion (SC)

Comparison of Order Selection Criteria
Contrary to the simplifying assumption throughout chapters 1 to 3 in Lütkepohl’s (2005) textbook, the lag order of a vector autoregressive (VAR)(p) process is usually unknown.

However, one should be interested to know the true lag order because it can be shown that the (approximate) mean squared error (MSE) matrix of the one-step ahead predictor increases with lag order $p$. 
In other words: if the chosen – or the assumed – lag order is unnecessarily large, the forecast precision of the corresponding VAR($p$) model will be reduced. Therefore, it is useful to have certain procedures or criteria for choosing the adequate lag order $p$. 
In order to do so, there are several possibilities:

- One could try to visually inspect correlograms (as suggested by Box/Jenkins) similar to the univariate case to determine $p$. However, this method does not appear to be useful in the multivariate case.

- An alternative would be to run so-called diagnostic tests, which means that a lag is to be chosen such that the residuals of the process pass diagnostics. But this empirically turned out to be the least reliable approach.
A more reasonable suggestion is to run statistical tests with the hypothesis that a certain lag equals zero. Some examples for this kind of hypothesis tests would be the likelihood ratio statistic or approximations thereof such as the Wald F statistic or the LM statistic (see e.g. Wooldridge 2006).
• The last and maybe the most elaborated way of coping with the problem of choosing the right lag order is to use information criteria (IC). The common underlying principle of these IC is that each of them minimizes the MSE. Predictors that minimize the MSE are optimal predictors in case the loss of the forecasters can be approximated by squares.
In the following, we will skip the first and the second method mentioned above as they do not seem to be very accurate and concentrate on the third and the fourth one.

Afterwards, we will evaluate the discussed approaches and present a simple empirical example.
Consider the process $y_t$ is called a VAR($p$) order if

$$y_t = \nu + A_1y_{t-1} + \ldots + A_p y_{t-p} + u_t$$

where $A_p \neq 0$ and $A_i = 0$ for $i > p$. And $p$ is the VAR order.
From previous section, we know

$$\Sigma_{\hat{y}}(1) = \frac{T + Kp + 1}{T} \Sigma_u$$

Obviously, $\Sigma_{\hat{y}}(1)$ is an increasing function of the order of the model fitted to the data.

Empirically, forecast from lower order models are clearly superior to higher order models.
Assume $y_t$ is a stable Gaussian (normally distributed) VAR($p$) process, and its log-likelihood function is

$$
\ln \ell(\beta, \Sigma_u) = -\frac{KT}{2} \ln 2\pi - \frac{T}{2} \ln |\Sigma_u| - \frac{1}{2} [y - (Z' \otimes I_k)\beta]'(I_T \otimes \Sigma_u^{-1})[y - (Z' \otimes I_k)\beta]
$$

$$
\frac{\partial \ln \ell}{\partial \beta} = (Z \otimes \Sigma_u^{-1})y - (ZZ' \otimes \Sigma_u^{-1})\beta
$$

Equating to zero gives the unrestricted ML estimator

$$
\tilde{\beta} = ((ZZ')^{-1}Z \otimes I_K)y
$$
Suppose the restrictions on $\beta$ are

$$C\beta = c$$

where $C$ is a known $(N \times (K^2p + K))$ matrix of rank $N$ and $c$ is a known $(N \times 1)$ vector. The restricted ML estimator can be found by Lagrangian approach.

$$L(\beta, \gamma) = \ln \ell(\beta) + \gamma'(C\beta - c)$$

$$\frac{\partial L}{\partial \beta} = (Z \otimes \Sigma_u^{-1})\gamma - (ZZ' \otimes \Sigma_u^{-1})\beta + C'\gamma$$

$$\frac{\partial L}{\partial \gamma} = C\beta - c$$

$$\tilde{\beta}_r = \tilde{\beta} + [(ZZ')^{-1} \otimes \Sigma_u] C'[C((ZZ')^{-1} \otimes \Sigma_u)C']^{-1}(c - C\tilde{\beta})$$
The log-likelihood ratio statistic is

$$\lambda_{LR} = 2[\ln l(\tilde{\delta}, \tilde{\Sigma}_u) - \ln l(\tilde{\delta}_r, \tilde{\Sigma}_u^r)]$$

where

$$\tilde{\Sigma}_u = \frac{1}{T}(Y - \tilde{B}Z)(Y - \tilde{B}Z)'$$

$$\tilde{\Sigma}_u^r = \frac{1}{T}(Y - \tilde{B}^r Z)(Y - \tilde{B}^r Z)'$$
Proposition: Let $y_t$ be a stationary, stable VAR($p$) process with standard white noise $u_t$. (Here, $y_t$ is not necessarily Gaussian) $\tilde{\beta}$ and $\tilde{\beta}_r$ are the (quasi) ML and restricted (quasi) ML estimators. Then the log-likelihood ratio test statistic is

$$
\lambda_{LR} = 2[\ln \ell(\tilde{\delta}, \tilde{\Sigma}_u) - \ln \ell(\tilde{\delta}_r, \tilde{\Sigma}_u)]
\quad = T(\ln |\tilde{\Sigma}_u^r| - \ln |\tilde{\Sigma}_u|) \quad (1)
\quad = (\tilde{\beta}_r - \tilde{\beta})'(ZZ' \otimes (\Sigma_u^r)^{-1})(\tilde{\beta}_r - \tilde{\beta}) \quad (2)
\quad = (\tilde{\beta}_r - \tilde{\beta})'(ZZ' \otimes \Sigma_u^{-1})(\tilde{\beta}_r - \tilde{\beta}) + o_p(1) \quad (3)
\quad = (C\tilde{\beta} - c)'[C((ZZ')^{-1} \otimes \tilde{\Sigma}_u)C']^{-1}(C\tilde{\beta} - c) + o_p(1) \quad (4)
\quad = (C\tilde{\beta} - c)'[C((ZZ')^{-1} \otimes \tilde{\Sigma}_u^r)C']^{-1}(C\tilde{\beta} - c) + o_p(1) \quad (5)
$$

and

$$
\lambda_{LR} \xrightarrow{d} \chi^2(N)
$$
Proof: To get (1), note that the third item of the log-likelihood function,

\[
[y - (Z' \otimes I_k)\beta]'(I_T \otimes \Sigma_u^{-1})[y - (Z' \otimes I_k)\beta] \\
= tr[(Y - BZ)'\Sigma_u^{-1}(Y - BZ)] \\
= tr[\Sigma_u^{-1} (Y - BZ)'(Y - BZ)] \\
\text{For unbiased } B, \text{ is } T\Sigma_u
\]

\[
\ln \ell(\tilde{\beta}, \tilde{\Sigma}_u) = \text{constant} - \frac{T}{2} \ln |\tilde{\Sigma}_u| \\
\ln \ell(\tilde{\beta}^r, \tilde{\Sigma}_u^r) = \text{constant} - \frac{T}{2} \ln |\tilde{\Sigma}_u^r|
\]
To prove (2), by Taylor’s theorem (Appendix A.13, Proposition A.3),

\[
\ln \ell(\tilde{\beta}_r) = \ln \ell(\tilde{\beta}) + \frac{\partial \ln \ell(\tilde{\beta})}{\partial \tilde{\beta}_r'}(\tilde{\beta}_r - \tilde{\beta}) + \frac{1}{2}(\tilde{\beta}_r - \tilde{\beta})' \frac{\partial^2 \ln \ell(\tilde{\beta})}{\partial \tilde{\beta}_r \partial \tilde{\beta}'}(\tilde{\beta}_r - \tilde{\beta}) = 0
\]

\[
\lambda_{LR} = 2[\ln \ell(\tilde{\delta}, \tilde{\Sigma}_u) - \ln \ell(\tilde{\delta}_r, \tilde{\Sigma}'_u)] = -(\tilde{\beta}_r - \tilde{\beta})' \frac{\partial^2 \ln \ell(\tilde{\beta})}{\partial \tilde{\beta}_r \partial \tilde{\beta}'}(\tilde{\beta}_r - \tilde{\beta}) = -(\Sigma^{-1} Z Z') (\tilde{\beta}_r - \tilde{\beta})
\]
plim(\tilde{\Sigma}_u - \tilde{\Sigma}'_u) = 0 can be used to show (3).

To get (4) and (5), from the formula of \tilde{\beta}_r,

\[
\lambda_{LR} = (C\tilde{\beta} - c)'[C((ZZ')^{-1} \otimes \Sigma_u)C']^{-1} \\
\times C((ZZ')^{-1} \otimes \Sigma_u)((ZZ') \otimes \tilde{\Sigma}_u^{-1})((ZZ')^{-1} \otimes \Sigma_u)C' \\
\times [C((ZZ')^{-1} \otimes \Sigma_u)C']^{-1}(C\tilde{\beta} - c)
\]
The asymptotic distribution of $\lambda_{LR}$ follows

**Proposition C.15(5):** Suppose $\hat{\beta}$ is an estimator of the $(K \times 1)$ vector $\beta$ with $\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. If $\Sigma$ is nonsingular and $\text{plim}\hat{\Sigma} = \Sigma$, then

$$T(\hat{\beta} - \beta)'\hat{\Sigma}^{-1}(\hat{\beta} - \beta) \xrightarrow{d} \chi^2(K).$$

Here, $(ZZ'/T)^{-1} \otimes \hat{\Sigma}_r^u$ is a consistent estimator of $\Gamma^{-1} \otimes \Sigma_u$. ■

Alternatively, one may consider using the statistic

$$\frac{\lambda_{LR}}{N} \xrightarrow{d} F(N, T - Kp - 1)$$

in small samples.
Assuming that $M$ is known to be an upper bound for the VAR order, the following sequence of null and alternative hypotheses may be tested using LR tests:

\[ H^1_0 : A_M = 0 \quad \text{versus} \quad H^1_1 : A_M \neq 0 \]
\[ H^2_0 : A_{M-1} = 0 \quad \text{versus} \quad H^2_1 : A_{M-1} \neq 0 | A_M = 0 \]
\[ \vdots \]
\[ H^i_0 : A_{M-i+1} = 0 \quad \text{versus} \quad H^i_1 : A_{M-i+1} \neq 0 | A_M = \ldots = A_{M-i+2} = 0 \]
\[ \vdots \]
\[ H^M_0 : A_1 = 0 \quad \text{versus} \quad H^M_1 : A_1 \neq 0 | A_M = \ldots = A_2 = 0 \]

The likelihood ratio statistic for testing the $i$th null hypothesis is

\[ \lambda_{LR}(i) = T \left[ \ln |\tilde{\Sigma}_u(M - i)| - \ln |\tilde{\Sigma}_u(M - i + 1)| \right] \xrightarrow{d} \chi^2(K^2) \]
It is important to realize that the significance levels of individual tests must be distinguished from the Type I error of the whole procedure, because rejection of $H_0^i$ means $H_0^{i+1}, \ldots, H_0^M$ are automatically rejected. Denoting

$$D_i \ldots H_0^i$$ is rejected in the $j$th test when it is actually true.

$\epsilon_i$ ... the probability of a Type I error for the $i$th test.

$$\epsilon_i = Pr(D_1 \cup D_2 \cup \ldots \cup D_i)$$

$\gamma_j$ ... the significance level of the $j$th individual test.

$$\gamma_j = Pr(D_j)$$
It can be shown that for \( m \neq j \) and \( m, j \leq i \), \( \lambda_{LR}(m) \) and \( \lambda_{LR}(j) \) are asymptotically independent if \( H_0^{i+1}, \ldots, H_0^i \) are true. Hence, \( D_m \) and \( D_j \) are independent in large samples so that

\[
\epsilon_i = Pr(D_1 \cup D_2 \cup \ldots \cup D_{i-1}) + Pr(D_i) - Pr\{(D_1 \cup D_2 \cup \ldots \cup D_{i-1}) \cap D_i\}
\]

\[
= \epsilon_{i-1} + \gamma_i - \epsilon_{i-1}\gamma_i = \epsilon_{i-1} + \gamma_i(1 - \epsilon_{i-1}), \ i = 2, 3, \ldots, M
\]

By induction and \( \epsilon_1 = \gamma_1 \)

\[
\epsilon_i = 1 - (1 - \gamma_1)...(1 - \gamma_i), \ i = 2, 3, \ldots, M.
\]
For example, take $\gamma_i = .05$ then

$$\epsilon_1 = .05, \epsilon_2 = 1 - .95 \times .95 = .0975, \epsilon_3 = .142625.$$  

Hence, the actual rejection probability will become quite substantial if the sequence of null hypothesis to be tested is long.
Another testing scheme is also suggested in the literature

\[ H^1_0 : A_M = 0 \quad \text{versus} \quad H^1_1 : A_M \neq 0 \]

\[ \vdots \]

\[ H^i_0 : A_M = \ldots = A_{M-i+1} = 0 \quad \text{versus} \quad H^i_1 : A_M \neq 0 \text{or} \ldots \text{or} \ A_{M-i+1} \neq 0 \]

\[ \vdots \]

Here, \( H^i_0 \) is tested against the full VAR(M) model. Unfortunately, the LR statistics will not be independent so that the overall significance level is difficult to determine.
For prediction purposes, it may not be necessary to find the correct lag order $p$ of the data generation process (DGP) as done by running hypothesis tests because

- one does not automatically obtain a good model for prediction by using hypothesis testing,

- and hypothesis testing as such can have the backdraw that, with positive probability, one chooses the incorrect lag order even though the time-series length is large.
In order to maximize forecast precision, one should choose lag order $p$ such that the forecast MSE is minimized.

An approximation to the MSE matrix of the one-step ahead predictor is given by the formula:

$$
\Sigma_{\hat{y}}(1) = \frac{T + Km + 1}{T} \Sigma_u
$$

$T$ denotes the sample size, $K$ the dimension of the time series, $m$ the order of the VAR fitted to the data and $\Sigma_u$ the white-noise covariance matrix.
However, we do not know $\Sigma_u$.

$\Rightarrow$ We have to estimate it:

$$\hat{\Sigma}_u(m) = \frac{T}{T - Km - 1} \tilde{\Sigma}_u(m)$$

This is a least squares (LS) estimator of $\Sigma_u$ with degrees of freedom adjustment $T/(T - Km - 1)$, where $\tilde{\Sigma}_u(m)$ is the maximum likelihood (ML) estimator of $\Sigma_u$ obtained by fitting a VAR($m$) model to the data.
For a better interpretation and for obtaining a unique solution, the final criterion should be a scalar rather than a matrix.

⇒ We take the determinant and define the resulting criterion as **Final Prediction Error (FPE)**:

\[
FPE(m) = \det \left[ \frac{T + Km + 1}{T} - \frac{T}{T - Km - 1} \tilde{\Sigma}_u(m) \right]
\]

\[
= \left[ \frac{T + Km + 1}{T - Km - 1} \right]^K \det \tilde{\Sigma}_u(m)
\]
Based on the FPE criterion, we choose the estimation $\hat{p}(FPE)$ of the lag order $p$ such that

$$FPE[\hat{p}(FPE)] = \min \{ FPE(m) | m = 0, 1, \ldots, M \}$$

VAR models of order $m = 0, 1, \ldots, M$ are estimated and the corresponding FPE values are computed. Then, the order $m$ minimizing the FPE is chosen as estimate of the true underlying lag order $p$. 

Umidjon Abdullaev, Ulrich Gunter, Miaomiao Yan

VAR Order Selection
$FPE(m) = \left[ \frac{T + Km + 1}{T - Km - 1} \right]^K \text{det} \tilde{\Sigma}_u(m)$

The first term in the above equation increases if $m$ increases, whereas the second term decreases for increasing $m$.

Thus, when these two forces are balanced in equilibrium, one obtains the estimate of the lag order.
Akaike’s Information Criterion (AIC) is defined as follows:

\[
AIC(m) = \ln |\tilde{\Sigma}_u(m)| + \frac{2mK^2}{T}
\]

$|\tilde{\Sigma}_u(m)|$ is the same as $\det \tilde{\Sigma}_u(m)$ and decreasing in $m$, $mK^2$ is the number of freely estimated parameters in a VAR model. $2mK^2/T$ is the penalty term that increases in $m$ and converges to zero for $T \rightarrow \infty$. 
We choose the estimate $\hat{p}(\text{AIC})$ of $p$ such that $\text{AIC}(m)$ is minimized.

One could additionally show that $\text{FPE}(m)$ and $\text{AIC}(m)$ are similar for some constant $N$, although the AIC has been derived based on a different reasoning. Hence, both criteria have similar properties, for example the following...
Corollary: The estimators $\hat{p}(FPE)$ and $\hat{p}(AIC)$ are not consistent. However, this would be a desirable asymptotic property of an estimator.

An estimator $\hat{p}$ is **consistent** if:

$$\lim_{T \to \infty} Pr\{\hat{p} = p\} = 1$$

An estimator $\hat{p}$ is **strongly consistent** if:

$$Pr\{\lim \hat{p} = p\} = 1$$
A criterion

\[ Cr(m) = \ln |\tilde{\Sigma}_u(m)| + \frac{mc_T}{T} \]

with \( c_T \) denoting a non-decreasing sequence of real numbers which depends on the sample size \( T \) and maximum lag order \( M (M \geq p) \) is supposed to be minimized by choosing \( \hat{p} \).

Then \( \hat{p} \) is a consistent estimator if and only if \( c_T \to \infty \) and \( c_T/T \to 0 \) for \( T \to \infty \).
\( \hat{p} \) is strongly consistent if and only if the above holds and, in addition, \( c_T / 2 \ln \ln T > 1 \) for \( T \to \infty \).

If \( M > p \), \( \hat{p}(\text{AIC}) \) and due to similarity also \( \hat{p}(\text{FPE}) \) are not consistent estimators because in this case \( mc_T / T = 2mK^2 / T \) or \( c_T = 2K^2 \), which contradicts \( c_T \to \infty \) and \( c_T / T \to 0 \) for \( T \to \infty \).

Both estimators usually overestimate the true lag order with positive probability.
The **Hannan-Quinn Criterion (HQ)** is defined as follows:

\[
HQ(m) = \ln |\tilde{\Sigma}_u(m)| + \frac{2 \ln \ln T}{T} mK^2
\]

We choose the estimate \( \hat{p}(HQ) \) such that HQ\((m)\) is minimized.
Here $c_T = 2K^2 \ln \ln T$. $c_T \to \infty$ and $c_T/T \to 0$ for $T \to \infty$.

$\Rightarrow \hat{p}(HQ)$ is consistent.

$\hat{p}(HQ)$ is even strongly consistent for $K > 1$ because in this case $c_T/2 \ln \ln T = K^2 > 1$. 

The **Schwarz Criterion (SC)** is defined as follows:

\[
SC(m) = \ln |\tilde{\Sigma}_u(m)| + \frac{\ln T}{T} mK^2
\]

We choose the estimate \(\hat{p}(SC)\) such that \(SC(m)\) is minimized.
Here $c_T = K^2 \ln T$. $c_T \to \infty$ and $c_T / T \to 0$ for $T \to \infty$.

Moreover, $c_T / 2 \ln \ln T = K^2 \ln T / 2 \ln \ln T > 1$ for any $K$.

$\Rightarrow \hat{p}(SC)$ is always strongly consistent.
Small sample comparison of AIC, HQ, and SC

Proposition 4.3. Let $y_{-M+1}, \ldots, y_0, y_1, \ldots, y_T$ be any $K$-dimensional multiple time series and suppose that VAR($m$) models, $m = 0, 1, \ldots, M$, are fitted to $y_1, \ldots, y_T$. Then the following relations hold:

- $\hat{p}(SC) \leq \hat{p}(AIC)$ if $T \geq 8$
- $\hat{p}(SC) \leq \hat{p}(HQ)$ for all $T$
- $\hat{p}(HQ) \leq \hat{p}(AIC)$ if $T \geq 16$. 

Umidjon Abdullaev, Ulrich Gunter, Miaomiao Yan
A consequence of previous results is that

$$\lim_{T \to \infty} Pr\{\hat{p}(AIC) < p\} = 0 \text{ and}$$

$$\lim Pr\{\hat{p}(AIC) > p\} > 0$$

The same holds for FPE because it is asymptotically equivalent to AIC.

Large sample results sometimes are good approximations only if extreme sample sizes are available.
A sequential testing procedure may give the same order as a selection criterion if the significance levels are chosen accordingly.

For instance, AIC chooses an order smaller than the maximum order $M$ if $\text{AIC}(M-1) > \text{AIC}(M)$ or, equivalently, if

$$\lambda_{LR}(1) = T \left( \ln |\tilde{\Sigma}_u(M-1)| - \ln |\tilde{\Sigma}_u(M)| \right) < 2MK^2 - 2(M-1)K^2$$

For $K=2$, $2K^2 = 8 \approx \chi^2(4)_{.90}$. 
The sequential testing procedure will not lead to a consistent order estimator if the sequence of individual significance levels is held constant.

For $M > p$ and a fixed significance level $\gamma$, 

$$H_0 : A_M = 0$$

is rejected with probability $\gamma$.

This problem can be circumvented by letting the significance level go to zero as $T \to \infty$. 
Some simulation results

Data: quarterly, seasonally adjusted series for fixed investment, disposable income, consumption expenditures for West Germany in billions of DM, 1960Q1-1982Q4;