Studies in Nonlinear Dynamics & Econometrics

Volume 9, *Issue* 1 2005 *Article* 2

A Practitioner's Guide to Lag Order Selection For VAR Impulse Response Analysis

Ventzislav Ivanov[*] Lutz Kilian[†]

[*]University of Michigan,

[†]University of Michigan, lkilian@umich.edu

A Practitioner's Guide to Lag Order Selection For VAR Impulse Response Analysis*

Ventzislav Ivanov and Lutz Kilian

Abstract

It is common in empirical macroeconomics to fit vector autoregressive (VAR) models to construct estimates of impulse responses. An important preliminary step in impulse response analysis is the selection of the VAR lag order. In this paper, we compare the six lag-order selection criteria most commonly used in applied work. Our metric is the mean-squared error (MSE) of the implied pointwise impulse response estimates normalized relative to their MSE based on knowing the true lag order. Based on our simulation design we conclude that for monthly VAR models, the Akaike Information Criterion (AIC) tends to produce the most accurate structural and semi-structural impulse response estimates for realistic sample sizes. For quarterly VAR models, the Hannan-Quinn Criterion (HQC) appears to be the most accurate criterion with the exception of sample sizes smaller than 120, for which the Schwarz Information Criterion (SIC) is more accurate. For persistence profiles based on quarterly vector error correction models with known cointegrating vector, our results suggest that the SIC is the most accurate criterion for all realistic sample sizes.

*Acknowledgements: We thank the editor and two anonymous referees for helpful suggestions. We also have benefited from comments by Fabio Canova, John Geweke, Phil Howrey, Jan Kmenta, Hashem Pesaran, Paul Rilstone and Shinichi Sakata on earlier drafts of this paper. Correspondence to: Lutz Kilian, Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220. Phone: (734) 647-5612. Fax: (734) 764-2769. E-mail: lkilian@umich.edu.

1. INTRODUCTION

Impulse response analysis based on vector autoregressions (VARs) plays a central role in modern empirical macroeconomics (for reviews of this literature see Pesaran and Smith 1998; Christiano, Eichenbaum and Evans 1999). Many researchers study impulse responses in structural or semi-structural VAR models based on identifying assumptions about the short-run and long-run responses of the economy to individual structural shocks (e.g., Sims 1980; Bernanke 1986; Shapiro and Watson 1988; Blanchard and Quah 1989). Other researchers attempt to identify long-run equilibrium relationships in the data based on VAR models estimated in vector error correction (VEC) form. For these models, one can construct impulse responses that trace out the response of error correction terms to a one-time shock in the vector of disturbances. The latter type of impulse response is known as a persistence profile and, in many cases, can be interpreted as a measure of the speed of convergence toward equilibrium (e.g., Pesaran and Shin 1996; Kilian 1999).

It is well known that the dynamic properties of impulse responses may depend critically on the lag order of the VAR model fitted to the data. These differences can be large enough to affect the substantive interpretation of VAR impulse response estimates (see e.g., Hamilton and Herrera 2004; Kilian 2001). An important preliminary step in empirical studies is to select the order of the autoregression based on the same data used subsequently to construct the impulse response estimates. The most common strategy in empirical studies is to select the lag-order by some pre-specified criterion and to condition on this estimate in constructing the impulse response estimates. A number of such lag-order selection criteria are in use in the empirical literature, yet little is known about their implications for the accuracy of the implied impulse response estimates.

In this paper, we use Monte Carlo simulations to compare the six criteria most commonly used in applied work. Our metric is the mean-squared error (MSE) of a given VAR impulse response estimate obtained by using a lag-order selection criterion, normalized relative to the MSE of the same VAR impulse response estimate obtained after imposing the true lag order. The six criteria are the Schwarz Information Criterion (SIC), the Hannan-Quinn Criterion (HQC), the Akaike Information Criterion (AIC), the general-to-specific sequential Likelihood Ratio test (LR), a small-sample correction to that test (SLR) proposed by Sims (1980), and the specific-to-general sequential Portmanteau test. The latter test may be interpreted as a Lagrange Multiplier (LM) test of a given VAR model for zero coefficient restrictions at higher-order lags.[1]

[1] For a detailed review of five of the six procedures see Lütkepohl (1993). All criteria considered here are motivated by classical statistical theory for unrestricted vector autoregressions. We do not pursue Bayesian approaches to model selection, although we note that the SIC may be given a

Our objective is to provide recommendations about how to select the lag order in applied work if the primary purpose of estimating the vector autoregression is to construct accurate impulse response estimates. This is clearly not the only purpose vector autoregressions may be used for. For example, one might use the same VAR model for real-time forecasting. Given that the overwhelming majority of empirical VAR studies is concerned primarily (if not exclusively) with impulse response analysis from VAR models (several examples are listed below) and given that this study is intended to inform and - if necessary - correct current practice, we focus on the effect of lag order selection procedures on the accuracy of impulse response estimates.[2]

Much of the previous research on VAR lag order selection has focused on the ability of lag-order selection criteria to detect the true lag order (see, e.g., Nickelsburg 1985; Lütkepohl 1985). We depart from this practice because the lag order itself typically is of no economic interest. It matters only to the extent that it affects the accuracy of the implied impulse response estimators. In practice, there is no simple mapping from the distribution of lag order estimates to the finite-sample accuracy of impulse response estimators. Notably, underestimation of the VAR lag order may be beneficial for the MSE of the impulse response estimator if the reduction in estimation variance outweighs the misspecification bias. For this reason, in this paper, we will focus directly on the accuracy of the impulse response estimator and discuss results for the lag order estimates only in passing.

We break new ground along three dimensions. First, to the best of our knowledge this practically important question has not been analyzed before with the exception of some illustrative bivariate examples presented by Kilian (2001). In contrast, in this paper, we present simulation evidence for large-dimensional VAR models with many lags of the type routinely estimated by leading practitioners in the VAR literature. The VAR models considered include anywhere between two and seven variables.

Second, we study the six lag order selection criteria that are most widely used in the applied VAR literature: the LR, SLR and LM test and the AIC, HQC

Bayesian interpretation (see Schwarz 1978). The reader is referred to Sims and Zha (1998) for further discussion of the Bayesian approach to estimation and inference in VAR models.

[2] Few economists use unrestricted VAR models for forecasting. The consensus in the profession is that univariate ARMA models, Bayesian VAR models (which are implemented without the use of lag order selection criteria), and – more recently – dynamic factor models are the methods of choice for generating out-of-sample forecasts. In contrast, for the objective of studying the propagation mechanisms of the economy (in the form of impulse response functions) unrestricted VAR models remain the method of choice in the literature. These questions cannot be analyzed in a univariate framework or in a standard dynamic factor framework. This is not to say that it might not be of interest to study the effect of lag order selection on out-of-sample forecasts from unrestricted VARs, but this problem is clearly secondary from the point of view of practitioners who use VAR models.

and SIC.[3] Likelihood ratio tests have been used for example by Sims (1980), Blanchard (1989), Keating and Nye (1998), Bernanke and Mihov (1998a) and Hamilton and Herrera (2004); the LM test by Galí (1992), Söderlind and Vredin (1996), Rotemberg and Woodford (1996); and the AIC, SIC, and HQC have been used by Lütkepohl and Reimers (1992), Bernanke, Gertler and Watson (1997), and Bernanke and Mihov (1998b), among others.

Third, as noted by Lütkepohl (1993), a central concern in comparing the accuracy of lag-order selection criteria is the generality of the simulation results. We therefore employ a variety of data generating processes including monthly VAR models, quarterly VAR models and quarterly vector error correction (VEC) models, resulting in a total of 180 design points.

We briefly review the most common lag-order selection procedures in section 2. The simulation design and the motivation for our relative MSE criterion are discussed in section 3. The results are presented in section 4. We organize the discussion around a number of questions of interest to empirical researchers. The results are summarized in graphical response surfaces and tables for the three types of problems most common in applied work: structural and semi-structural impulse responses based on monthly or quarterly VAR models, and persistence profiles based on quarterly models in VEC form. Section 5 contains the concluding remarks.

2. A REVIEW OF LAG-ORDER SELECTION PROCEDURES

The most common strategy in empirical studies is to select the lag-order by some pre-specified criterion and to condition on this estimate in constructing the impulse response estimates. This strategy is sometimes criticized on the grounds that some researchers use more than one criterion to examine the robustness of the estimation results. Such a strategy does not seem to reflect common practice among many leading practitioners, however. For example, the empirical applications listed in the introduction tend to use just one criterion. Moreover, it is not clear how such a sensitivity analysis should be conducted and on what statistical basis. Finally - and most importantly - it is not clear how to proceed in the likely case that different criteria give different answers. For example, Lütkepohl (1990) provides an illustrative example in which he considers three alternative lag order selection criteria, yet decides to use only one of them when these criteria give conflicting results. Similarly, researchers who combine an initial lag order selection based on the AIC with an LM test for serial correlation in the error term of the model selected by the AIC, will in practice overrule the AIC choice when there is a contradiction. Thus, in the end, even researchers who

[3] We do not include the Posterior Information Criterion (PIC) of Phillips and Ploberger (1994), for example, because that criterion does not appear to have been used in the empirical VAR literature.

investigate the sensitivity of the estimation result tend to favor one criterion at the expense of the others.[4]

Although one could in principle report estimation results for a number of alternative lag orders, these results will tend to differ in practice and researchers will have to take a stand on the relative plausibility of their results. Ultimately researchers cannot avoid the trade-offs involved in the choice of lag order selection criteria. At a minimum, understanding the properties of each procedure will help researchers to make an informed decision. For these reasons, in this paper, we follow the common practice of relying on one criterion only in selecting the lag order.

We postulate that the true process is a *K*-dimensional autoregression of order p_0, which may be represented in VAR or in VEC form. Abstracting from deterministic regressors (such as seasonal dummies or intercepts), the first three lag-order selection criteria are:

$$SIC(p) = \ln\left|\bar{\Sigma}(p)\right| + \frac{\ln N}{N}(K^2 p)$$

$$HQC(p) = \ln\left|\bar{\Sigma}(p)\right| + \frac{2\ln\ln N}{N}(K^2 p)$$

$$AIC(p) = \ln\left|\bar{\Sigma}(p)\right| + \frac{2}{N}(K^2 p)$$

where N is the effective sample size and $\bar{\Sigma}$ is the quasi-maximum likelihood estimate of the innovation covariance matrix Σ (see Sin and White (1996) for further discussion of the theoretical rationale for these criteria). The lag order estimate $\hat{p}$ is chosen to minimize the value of the criterion function for $\{p : 1 \le p \le \bar{p}\}$ where $\bar{p} \ge p_0$ (see Quinn 1980; Paulsen and Tjøstheim 1985; Quinn 1988). It can be shown that $\hat{p}^{SIC} \le \hat{p}^{AIC}$ for $N \ge 8$, $\hat{p}^{SIC} \le \hat{p}^{HQC}$ for all N, and $\hat{p}^{HQC} \le \hat{p}^{AIC}$ for $N \ge 16$. As noted by Granger, King and White (1995), any one of these three information criteria may be interpreted as a sequence of LR tests with the critical value being implicitly determined by the penalty function. No one model is favored because it is chosen as the null hypothesis, and the order in which the criterion function is evaluated does not affect the lag order choice.

In contrast, the use of sequential LR and LM tests requires the explicit choice of a significance level. The general-to-specific LR test is implemented as

[4] Yet another possible strategy for model selection is to choose the lag order to ensure impulse response functions that look “sensible”. This strategy will not be pursued in this paper. It is not only at odds with statistical approaches to model selection, but there is no evidence that any of the studies by leading practitioners that we cited in section 1 employed this approach.

described by Lütkepohl (1993). We follow Lütkepohl (1985) in using the same nominal significance level of the LR test at each step in the sequential procedure, and we use the asymptotic $\chi^2(K^2)$ critical values. Note that the overall significance level of sequential tests will differ from the individual level. The LR test involves a sequence of tests of the form

$$LR(i) = N\left(\ln\left|\overline{\Sigma}(\overline{p}-i)\right| - \ln\left|\overline{\Sigma}(\overline{p}-i+1)\right|\right)$$

for $i = 1,...,\overline{p}-1$. If the null cannot be rejected, we repeat the test with $i = i+1$. The test sequence is terminated when we can reject the null hypothesis that $p_0 = p$ against $p_0 = p+1$ (or when $p = 1$). The resulting tests are denoted by LR1 (for the nominal 1% LR test) and LR5 (for the nominal 5% LR test). We also consider a small-sample correction of the LR test proposed by Sims (1980). This correction takes the form:

$$SLR(i) = (N-c)\left(\ln\left|\overline{\Sigma}(\overline{p}-i)\right| - \ln\left|\overline{\Sigma}(\overline{p}-i+1)\right|\right),$$

where $c = (\overline{p}-i+1)K$. The corresponding tests will be denoted by SLR1 and SLR5.

The Portmanteau test involves a sequence of tests of the null of no serial correlation in the residuals of the VAR(*p*) model against the alternative that at least one of the first *s* residual autocorrelations differs from zero. Hosking (1981) shows that this test can be interpreted as a special case of an LM test of the null of no serial correlation. Our Portmanteau (LM) test statistic is:

$$LM(p) = N^2 \sum_{i=1}^{s} (N-i)^{-1} tr\left(\hat{C}_i{}'\hat{C}_0^{-1}\hat{C}_i\hat{C}_0^{-1}\right),$$

where $\hat{C}_i = \sum_{t=i+1}^{N} \hat{u}_t \hat{u}_{t-i}{}' \Big/ N$ and $\hat{u}_t$ denotes the residual from a VAR(p) process. The LM test statistic is calculated for $p = 1,...,\overline{p}$, in ascending order. If the null of no residual correlation is rejected, we add one more lag to the VAR model and repeat the test. The test sequence is terminated when the null of no serial correlation cannot be rejected (or when $p = \overline{p}$). Provided that $s/N \to 0$ at a suitable rate as $N \to \infty$, the LM test has an asymptotic $\chi^2\left(K^2(s-p)\right)$-distribution where $s > p$ (see Lütkepohl 1993, pp. 150). We adopt the convention of setting *s* equal to the maximum of $N^{1/2}$ (rounded to the nearest integer) and

Table 1: Empirical Studies After Which the DGPs Are Modeled

Empirical Study	Dimension	Lag order	Model	Frequency	Variables (in Order)
Sims (1986)	6	4	VAR	Quarterly	Output, investment, price level, M1, unemployment rate, T-Bill rate
Rotemberg-Woodford (1996)	4	Not reported	VAR	Quarterly	Growth rate of nominal oil price, real price of oil, output growth, growth rate of real wage.
Christiano-Eichenbaum-Evans (1996)	7	4	VAR	Quarterly	Output, price level, commodity prices, FedFunds rate, nonborrowed reserves, total reserves, M1
Galí (1999)	2	4	VAR	Quarterly	Growth rate of labor productivity, growth rate of hours worked
Strongin (1995)	5	12	VAR	Monthly	Output, price level, total reserves, nonborrowed reserve ratio, FedFunds rate.
Eichenbaum-Evans (1995)	5	6	VAR	Monthly	Output, price level, nonborrowed reserves ratio, U.S.-UK. short-term interest rate differential and real exchange rate.
Bernanke-Gertler (1995)	4	12	VAR	Monthly	Output, price level, commodity prices, FedFunds rate.
Leeper (1997)	6	18*	VAR	Monthly	T-Bill rate, output, price level, T-Bond rate, total reserves, commodity prices.
Johansen-Juselius (1990)	4	2	VEC	Quarterly	Finnish data for real balances, real income, short-term interest rate, inflation rate.
Kilian (1999)	2	4	VEC	Quarterly	Percent change in U.S.-Canadian spot exchange rate, deviation of spot rate from monetary fundamental.
Pesaran-Shin-Smith (2000)	5	2	VEC	Quarterly	U.K. price level, ROW price level, U.K.-ROW exchange rate, U.K. T-Bill rate, short-term ROW interest rate.

NOTES: * with Bayesian prior on lag structure.

$\overline{p}+1$. This rule results in choices of s that are similar to values used in many empirical studies. The resulting tests are denoted by LM1 (for the nominal 1% LM test) and LM5 (for the nominal 5% LM test.[5]

3. SIMULATION DESIGN AND PERFORMANCE CRITERIA

We consider three classes of DGPs based on monthly and quarterly data sets drawn from empirical studies published by leading VAR practitioners. A list of these studies is provided in Table 1. We study structural and semi-structural impulse responses based on four quarterly VAR models (Sims 1986; Rotemberg and Woodford 1996; Christiano, Eichenbaum and Evans 1996; Galí 1999) and four monthly VAR models (Bernanke and Gertler 1995; Eichenbaum and Evans 1995; Strongin 1995; Leeper 1997). These impulse responses are obtained by imposing contemporaneous identifying assumptions, with the exception of the study by Galí (1999), which uses long-run identifying assumptions instead. For precise definitions of these impulse response estimators the reader is referred to Christiano, Eichenbaum and Evans (1999), Lütkepohl (1993) and Pesaran and Smith (1998).

We also analyze persistence profiles of three quarterly VEC models: a money market equilibrium relationship based on Johansen and Juselius (1990), an exchange rate arbitrage condition based on monetary fundamentals from Kilian (1999), and a VEC model by Pesaran, Shin and Smith (2000) that involves two equilibrium relationships: the uncovered interest parity condition and the purchasing power parity condition. The latter are treated separately in the analysis.[6] Note that persistence profiles (which may be viewed as generalized impulse response functions) differ from conventional impulse response functions and will in general have different statistical properties. For a precise definition of persistence profiles and their relationship to other impulse response estimators see

[5] With one exception, these lag order selection criteria are known to be robust to the presence of a unit root in the autoregressive lag order polynomial. Notably, Paulsen (1984) formally establishes the consistency of the SIC and HQC in the presence of a unit root. To the best of our knowledge, no formal analysis exists of the properties of the AIC in the unit root case. Watson (1994, p. 2860) shows that for $p \geq 1$ the LR and LM tests remain asymptotically valid even in the presence of a unit root.

[6] As Wickens (1996) shows, estimated multiple cointegrating vectors cannot be given an economic interpretation without additional a priori information. In estimating the VEC models, we therefore impose coefficient values on the cointegrating vectors that are consistent with economic theory, even when the original studies rely on estimated cointegrating vectors. Specifically, for the Johansen-Juselius model we impose that real money demand is homogeneous of degree 1 in income and that nominal interest rates and inflation rates are stationary. For the Pesaran-Shin-Smith model we impose that uncovered interest parity and purchasing power parity hold in the long run.

Pesaran and Smith (1998). There are no monthly VEC models in our simulation study, because we could not find any examples in the literature of persistence profile estimation on monthly data.

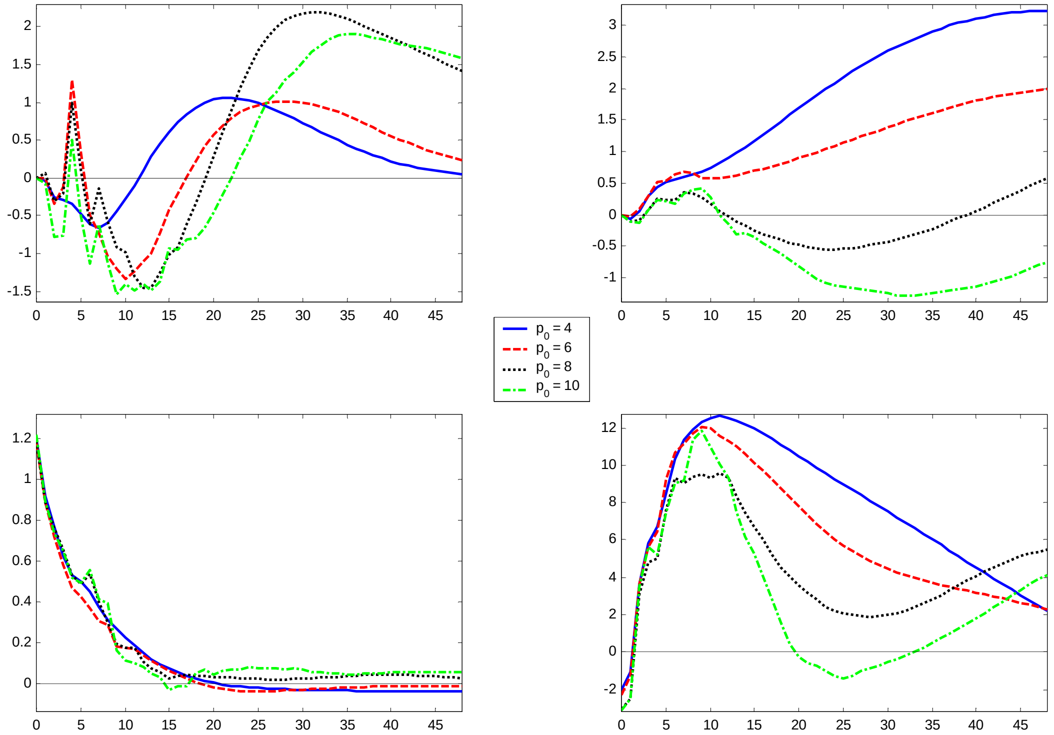
For the quarterly data, we postulate values of $p_0 \in \{2,4,6\}$ for each DGP and for the monthly data values of $p_0 \in \{4,6,8,10\}$.[7] The DGPs are constructed as follows: For each data set and value of p_0, we fit a VAR(p_0) model to the data used in the original empirical study. The resulting model estimate is subsequently treated as a DGP for the simulation study. Note that each such model will have different parameter values by construction. The hope is that the resulting DGPs will be more representative for empirical VAR studies than any ad hoc choice of parameter values would have been. In fitting the VAR models, we impose unit roots and cointegration constraints whenever the original studies did so, and we closely follow the original studies in including deterministic regressors (seasonal dummies, dummies, intercepts) and enforcing exogeneity constraints. The model innovations are postulated to be Gaussian white noise with the same innovation covariance matrix as the fitted VAR(p_0) model. For each DGP we consider several sample sizes $T = N + \bar{p}$. For quarterly data, $T \in \{80, 100, 120, 160, 200\}$ and for monthly data $T \in \{240, 300, 360, 480, 600\}$. Note that we only consider sample sizes that are relevant for empirical research using quarterly and monthly data. Altogether, our simulation study includes 180 different design points.[8] For quarterly data, we set $\bar{p}$ = 8 and for monthly data $\bar{p}$ = 12. This choice ensures that in all cases in the simulation study the true lag order is included in the set of lag orders considered. The choice of p_0 in many cases has nontrivial implications for the shape and persistence of the implied population impulse response functions. Figure 1 provides four examples for how responses to a given shock may change, as we vary p_0.

For each design point, we generate 5,000 independent draws of data of length T. Initial values are obtained by randomly drawing blocks of data of length p_0 with replacement from the original data set. For each draw $\{y_t\}_{t=1}^{T}$, we select the best-fitting VAR model based on each of the nine lag order selection criteria

[7] Few applied users would consider more than 12 lags in selecting the order of a VAR model for monthly post-war data and most studies end up using 5 or 6 lags. We did not consider p_0 = 12 for monthly data because such a design point may have favored the AIC, given its tendency to overfit asymptotically.

[8] For the DGP based on Pesaran et al. (2000), the point estimate for $p_0 = 6$ was explosive. We therefore eliminated this design point from the analysis.

Figure 1: Effect of p_0 on Selected Population Impulse Response Functions in Eichenbaum-Evans Model



(SIC, HQC, AIC, LR1, LR5, SLR1, SLR5, LM1, LM5).[9] In fitting the VAR models, we impose whatever unit root, exogeneity, or cointegration restrictions hold for the underlying DGP.[10] We are interested in the impulse response functions of the VAR system up to horizon *h*. For each of the seven VAR model estimates, we compute the $K^2(h+1)$ implied pointwise impulse response coefficient estimates using the same identifying assumptions as the original studies in Table 1. We restrict ourselves to horizons of up to four years for the

[9] We do not consider the possibility that $p = 0$ in the simulation study. This simplifying assumption involves little loss of generality, because most macroeconomic time series tend to be so persistent that a VAR(0) model is a priori implausible.

[10] In particular, for the VEC models the cointegrating vector is assumed to be known and is imposed before the lag order is selected. As shown in section 4, selecting the lag order prior to imposing the cointegrating vector would make little difference for our results.

quarterly and monthly VAR models and horizons of up to six years for the persistence profiles.[11] Finally, for each criterion, we calculate the mean squared deviation of each of the impulse response coefficient estimates from their true values, which will be referred to as the mean-squared error (MSE) of the estimate.

An obvious difficulty in comparing MSE results for different impulse response coefficients is that the size of the MSE will be sensitive to the size of the underlying impulse response coefficient. This makes it impossible to compare the MSE of any two impulse response estimates directly. We address this problem by expressing the MSEs relative to the corresponding MSEs based on knowing the true lag order. This normalization allows us to average results across horizons for the same impulse response function, across different impulse responses for the same horizon, and across DGPs. Specifically, the relative MSE for a given lag order selection method is given by:

$$\frac{\left[\hat{\theta}_{jk,i}(\hat{p} \mid p_0) - \theta_{jk,i}(p_0)\right]^2}{\left[\hat{\theta}_{jk,i}(p_0) - \theta_{jk,i}(p_0)\right]^2},$$

where the expression $\theta_{jk,i}$ denotes the population response of variable *j* to shock *k* at horizon *i*, and $\hat{\theta}_{jk,i}$ denotes the corresponding estimator. These quantities depend on the lag order of the underlying VAR model. Here p_0 denotes the true, but in practice unknown VAR lag order of the data generating process, and $\hat{p} \mid p_0$ is the lag order selected by a given method conditional on the data having been generated by a VAR(p_0) process.

Throughout the paper, averages of these MSE ratios are calculated as geometric means rather than arithmetic means.[12] In a preliminary investigation, we also computed results based on the ratios of the mean-absolute error of the impulse response estimates. These results tended to be qualitatively similar.

[11] In very small samples, there is some probability that the persistence profile estimate is explosive (in the sense of diverging toward infinity). We follow Pesaran and Shin (1996, pp. 141) in discarding these rare explosive draws.

[12] Arithmetic means of ratios may be misleading. Consider a sequence of MSEs for two methods A and B: MSE(A)= [2 1 4] and MSE(B)=[3 2 2]. Then the arithmetic mean of the sequence of pointwise ratios MSE(A)/MSE(B)=[2/3 1/2 4/2] is 1.06, suggesting that B is more accurate than A, yet the arithmetic mean of the pointwise reciprocals MSE(B)/MSE(A)=[3/2 2/1 2/4] is 1.33, suggesting that A is more accurate than B. We therefore compute geometric means by exponentiating the arithmetic mean of the log-differences of MSE(A) and MSE(B).

4. SIMULATION RESULTS

Given the large number of design points, the simulation results are summarized in graphical response surfaces and tables. We begin with the discussion of some general regularities. We organize the discussion around a number of questions of interest to empirical researchers.

Question 1: How does the overall ranking of the criteria depend on the sample size?

Tables 2 and 3 are based on overall averages of the relative MSEs for each class of DGPs. These averages are computed for the impulse responses of all variables with respect to all underlying shocks and at all horizons. The results in Tables 2 and 3 show that the choice of lag order selection criterion is practically important for impulse response analysis, and that there are important differences across alternative lag-order selection criteria.

Table 2 shows the average relative MSE for each criterion as a function of T only. The results in Table 2 are appropriate if a researcher cares equally about all horizons h and is completely unsure about the lag order p_0 of the underlying process. For impulse responses based on monthly VAR processes, we find that the AIC-based estimates are always at least as accurate as those based on other criteria. For T = 240 only the HQC is as accurate as the AIC, and for larger sample sizes the AIC dominates the other criteria across the board. In contrast, for impulse responses based on quarterly VAR processes, the AIC cannot be recommended. The SIC dominates the other criteria for sample sizes up to 120 quarters, whereas for all larger sample sizes the HQC is the most accurate criterion. Finally, for persistence profiles based on quarterly VEC processes, the SIC dominates the other criteria for all sample sizes considered. Note that the latter results are not directly comparable to the quarterly VAR results both because the statistic of interest differs and because the VEC models are estimated subject to the cointegration constraint, whereas the VAR models are estimated by unrestricted least-squares.

Since the LR, SLR and LM tests at the nominal 1% level are systematically more reliable than the corresponding nominal 5% tests, we report only the former. The sequential LR1 and LM1 tests tend to perform poorly for all three classes of models, especially for small sample sizes. The SLR1 test in all cases is more accurate than the LR1 test, often by a wide margin. It also is more accurate than the LM1 test in most cases, but even the SLR1 test is clearly dominated by other criteria for all sample sizes. On the basis of these results, and keeping in mind the purpose of this study, we recommend that applied users rely on the AIC for all monthly VAR models, the HQC for all quarterly VAR models with the exception of sample sizes up to 120 quarters, for which the SIC is preferred, and the SIC for all quarterly VEC models. The simulation evidence does not support the use of sequential LM, LR or SLR tests in applied work.

Table 2: Average MSE Ratio for Impulse Response Estimates by Criterion Relative to Model Based on True Lag Order

(a) Monthly VAR Models

T	AIC	HQC	SIC	LR1	SLR1	LM1
80	0.97	0.97	0.99	1.17	1.00	1.10
100	0.99	1.03	1.07	1.11	1.01	1.07
120	1.00	1.04	1.15	1.08	1.02	1.05
160	1.00	1.05	1.28	1.06	1.02	1.03
200	1.00	1.06	1.38	1.04	1.02	1.03

(b) Quarterly VAR Models

T	AIC	HQC	SIC	LR1	SLR1	LM1
80	1.76	1.18	0.79	2.17	1.14	1.36
100	1.28	0.93	0.89	1.74	1.09	1.20
120	1.10	0.97	0.96	1.60	1.09	1.17
160	1.06	1.00	1.10	1.43	1.10	1.15
200	1.06	1.01	1.17	1.34	1.11	1.14

(c) Quarterly VEC Models

T	AIC	HQC	SIC	LR1	SLR1	LM1
80	3.88	1.04	0.70	5.38	1.48	1.58
100	1.37	0.90	0.79	3.26	1.15	1.15
120	1.05	0.94	0.85	2.26	1.10	1.04
160	1.02	0.98	0.91	1.58	1.07	0.99
200	1.02	0.99	0.95	1.34	1.05	0.99

NOTES: Averages of ratios are calculated as geometric means.

Table 3: Selected Average MSE Ratios for Impulse Response Estimates

(a) Monthly VAR Models

T	AIC/HQC	AIC/SIC	AIC/LR1	AIC/SLR1	AIC/LM1
240	1.00	0.98	0.83	0.97	0.88
300	0.97	0.93	0.89	0.98	0.93
360	0.96	0.87	0.92	0.98	0.95
480	0.95	0.78	0.95	0.98	0.97
600	0.94	0.73	0.96	0.98	0.97

(b) Quarterly VAR Models

T	HQC/AIC	HQC/SIC	HQC/LR1	HQC/SLR1	HQC/LM1
80	0.67	1.50	0.55	1.04	0.87
100	0.72	1.05	0.53	0.85	0.78
120	0.88	1.00	0.60	0.88	0.82
160	0.94	0.91	0.70	0.91	0.87
200	0.96	0.86	0.76	0.92	0.89

(c) Quarterly VEC Models

T	SIC/AIC	SIC/HQC	SIC/LR1	SIC/SLR1	SIC/LM1
80	0.18	0.67	0.13	0.47	0.44
100	0.57	0.87	0.24	0.68	0.69
120	0.80	0.90	0.38	0.77	0.81
160	0.89	0.93	0.57	0.85	0.91
200	0.93	0.96	0.71	0.90	0.96

NOTES: Averages of ratios are calculated as geometric means.

Question 2: What are the costs of not knowing the true lag order?

So far we have focused on the ranking of the criteria as a function of the sample size. A closely related question is how quantitatively important the effects of lag order uncertainty are relative to knowing the true lag order. Table 2a documents the fact that the success of a lag order selection criterion in impulse response analysis is not directly related to its ability to estimate accurately the true lag order. For T = 240, for example, all three information criteria result in more accurate impulse response estimates than would have been obtained by imposing the true lag order. It can be shown that these criteria all tend to underestimate the true lag order for T = 240 in Table 2a, yet their MSE ratios are slightly below one.[13] The explanation is that in small samples the bias induced by using a lower lag order than p_0 is more than offset by the reduced variance of the impulse response estimator. Interestingly, the AIC-based and HQC-based impulse response estimates both are not only more accurate than those based on the true lag order, but they also are more accurate than the impulse response estimates based on the SIC. The reason for this outcome is that the SIC tends to underestimate severely the true lag order in small samples. Although slight underestimation is beneficial for the MSE of the impulse responses for T = 240, severe underestimation is a serious problem in this case.

To illustrate further this phenomenon, we also constructed the average MSE of the monthly VAR models for a grid of fixed lag orders. Table 4 shows the average MSE based on lag orders of $p_0 + i$ relative to that for models based on p_0. For expository purposes we set $p_0 = 6$ and let $i \in \{-5, -4, -3, -2, -1, 1, 2, 3\}$.

Table 4: Average MSE Ratio for Impulse Response Estimates by Lag Order Relative to Model Based on True Lag Order

Monthly VAR Models with Fixed Lag Order

T	$p_0 - 5$	$p_0 - 4$	$p_0 - 3$	$p_0 - 2$	$p_0 - 1$	$p_0 + 1$	$p_0 + 2$	$p_0 + 3$
240	1.00	0.95	0.94	0.96	0.97	1.08	1.16	1.24
300	1.10	1.04	0.99	0.99	0.99	1.07	1.14	1.22
360	1.19	1.11	1.04	1.03	1.00	1.07	1.14	1.21
480	1.37	1.25	1.11	1.09	1.02	1.07	1.13	1.20
600	1.54	1.38	1.19	1.14	1.04	1.07	1.13	1.20

NOTES: Averages of ratios are calculated as geometric means. True lag order fixed at $p_0 = 6$.

[13] To conserve space, we do not report tables for the distribution of lag-order estimates. The reason is that for each model, true lag order p_0 and sample size we would require a different table.

It is evident that the results are not symmetric about p_0. Whereas overfitting always raises the MSE relative to the true lag order, a moderate degree of underfitting may actually lower the MSE in small samples. These gains dissipate quickly as T increases. For larger T the losses from overfitting and underfitting become almost symmetric.

Table 4 also illustrates the increasing cost of a strong degree of underfitting as the sample size is increased. This fact helps to explain why the accuracy of the SIC in Table 2a (and to a lesser extent that of the HQC) actually deteriorates with increasing sample size. Whereas the AIC's MSE quickly approaches that for the true model in Table 2a, the MSEs of the SIC- and HQC-based impulse response estimators worsen, as the sample size is increased. Note that this happens despite the consistency of the SIC and HQC for the true lag order p_0.

This deterioration of the MSE is closely related to the fact that the estimator of the parameter vector of models obtained by consistent model selection need not converge uniformly to the true parameter vector (see Kabaila 1995; Pötscher 1995). Note that for small *T*, the SIC (and to a lesser extent the HQC) is strongly downward biased relative to p_0. In very small samples, the bias induced by the underestimation of p_0 is more than offset by a large variance reduction. In short, parsimony is beneficial for the MSE of the impulse response estimator. As the sample size increases, however, this variance effect becomes less and less important and the effect of misspecification bias becomes dominant. Although the degree of underestimation by the SIC (and by the HQC) diminishes as $T \to \infty$, as expected, in practice, it may cause a large increase in the relative MSE of the impulse response estimator. Only for sample sizes much larger than those considered here, this bias will vanish and the MSE ratios of the SIC will begin to improve. For example, it can be shown that the SIC ratio drops from 1.38 for $T = 600$ to 1.28 for $T = 1200$ and that of the HQC drops from 1.06 to 1.01. Asymptotically, both ratios will approach unity.

In contrast, the AIC lag order estimates are clustered increasingly close to the true lag order with increasing sample size. The probability that the AIC underestimates the true lag order shrinks toward zero, as the sample size increases. This finding is consistent with standard theoretical results. In addition, the probability of the AIC overestimating the true lag order shrinks almost to zero for T = 600, in line with theoretical results by Paulsen and Tjøstheim (1985), who showed that asymptotically the AIC will estimate the true lag order with probability 99.0% for $K = 4$, 99.8% for $K = 5$ and even higher probability for $K \geq 6$. This fact helps to explain the relative performance of the AIC, SIC and HQC for large *T*.

Table 2b shows the corresponding results for the quarterly VAR models. For small *T*, all criteria but the SIC are associated with dramatic losses in accuracy relative to the true lag order model, in some cases by a factor of more than two. For larger samples, the relative MSE of all criteria but the SIC improves in Table 2c. As in the case of the monthly VAR models, the worsening MSE ratio of the SIC for larger *T* reflects the strong downward bias of this lag order selection criterion in small samples. It can be shown that for our DGPs the SIC systematically underestimates the true lag order for all sample sizes. This small-sample bias appears to be beneficial for the accuracy of the impulse responses for small *T*, regardless of p_0, but it is a liability for larger *T*, especially when p_0 is large. In contrast, the comparatively high MSE of the AIC- (and to a lesser extent of the HQC-) based estimates for *T* = 80 reflects the fact that these criteria tend to overestimate severely the true lag order. For larger *T*, this tendency weakens. Whereas the AIC continues to overestimate the true lag order to some extent, the HQC estimates closely track p_0. This fact helps to explain the low MSE ratios of the HQC in Table 2b. We conclude that a high degree of parsimony is beneficial in very small samples, but that for larger sample sizes both overestimation and underestimation of the lag order have large costs in terms of the MSE of the impulse response estimates. The HQC minimizes these two risks.

Table 2c shows the corresponding results for the quarterly VEC model. Again in small samples - with the exception of the SIC - MSE ratios relative to the true model are high, reaching a factor of more than 5 for the LR1 tests, of about 1.5 for the SLR1 tests, of about 1.6 for the LM1 tests and of almost 4 for the AIC. For larger sample sizes, the relative accuracy of all criteria improves. The reasons for the relative ranking of the AIC, HQC and SIC for *T* = 80 are the same as for the quarterly VAR models. For such small sample sizes parsimony is beneficial and criteria that tend to overfit such as the AIC (and to a much lesser extent the HQC) perform poorly. Unlike in the quarterly VAR case, however, the SIC performs very well even for larger sample sizes. Although the SIC has a similar tendency to underfit – especially when p_0 is large – this tendency does not seem to affect the accuracy of the estimated persistence profiles.

Question 3: How large are the differences in accuracy across criteria?

We now turn to the closely related question of how quantitatively important the differences between alternative criteria are. Table 3a shows that the gains from choosing the best criterion can be substantial both for small and for large sample sizes. For example, for a monthly VAR model with *T* = 600, the AIC can be expected to reduce the MSE of the impulse responses by up to 27% relative to the

SIC.[14] For T = 240, the gains range from 0% relative to the HQC and 2% relative to the SIC to 17% relative to the LR1 test.

Table 3b shows the corresponding results for the quarterly VAR models. The relative performance of the three information criteria in Table 3b depends on the sample size. For T = 80, the SIC promises gains of up to 55% (33%) relative to the AIC (HQC). For T = 100, these gains diminish and for T = 120, the SIC and the HQC are virtually tied, both having MSEs about 12% lower than the AIC. For larger sample sizes, the ranking of the SIC and the HQC is reversed, and the HQC promises MSE reductions of up to 4% (14%) relative to the AIC (SIC).

Table 3c shows the corresponding results for the quarterly VEC model. The differences in accuracy are dramatic. The SIC tends to improve the accuracy of the persistence profiles by 82% (33%) relative to the AIC (HQC) for T = 80. Relative to the LR1 test the relative gains are even larger, reaching 87%. The relative gains diminish as the sample size increases, but they still amount to 7% (4%) relative to the AIC (HQC) for T = 200 and up to 29% for the other criteria.

Question 4: How sensitive are the results to the impulse response horizon?

An important question for applied users is to what extent the results in Table 3 hold for alternative horizons h. In applied research, we may care more about some horizons than about others. For example, a policy-maker may be primarily concerned about the responses at horizons of one year or less. A standard argument in forecasting is that the prediction mean-squared error may be reduced in small samples if the model is slightly underfit. Similar arguments apply to impulse response analysis. A natural conjecture therefore is that highly parsimonious lag order selection criteria such as the SIC may produce impulse response estimates that are more accurate at least at short horizons than the AIC for example. On the other hand, parsimonious lag order selection criteria may fail to capture complicated and non-smooth dynamics of impulse response functions, especially at longer horizons, as shown by Kilian (2001).

It is unclear a priori which of these effects will dominate in practice. We therefore disaggregate the MSE results in Table 3 by time horizon. These disaggregated results will be appropriate if we know T and the range of horizons we are interested in, but we have no idea whether p_0 is small or large. Given the large number of simulation results and the relatively poor performance of sequential LM and LR tests, we do not provide detailed results for each criterion, but focus on the three penalized likelihood criteria. Our main finding is that

[14] In this example, Table 2a shows that the MSE-ratio of the AIC is 1.00 and the MSE ratio of the SIC is 1.38. We obtain a reduction of (1.38-1.00)/1.38=27.54 percent relative to the SIC's MSE ratio of 1.38. This gain is represented as a ratio of 0.73 for AIC/SIC in Table 3a. The slight difference is due to the rounding of the results in Table 2a.

parsimony matters, but that the required degree of parsimony may differ greatly depending on the sample size and class of DGP.

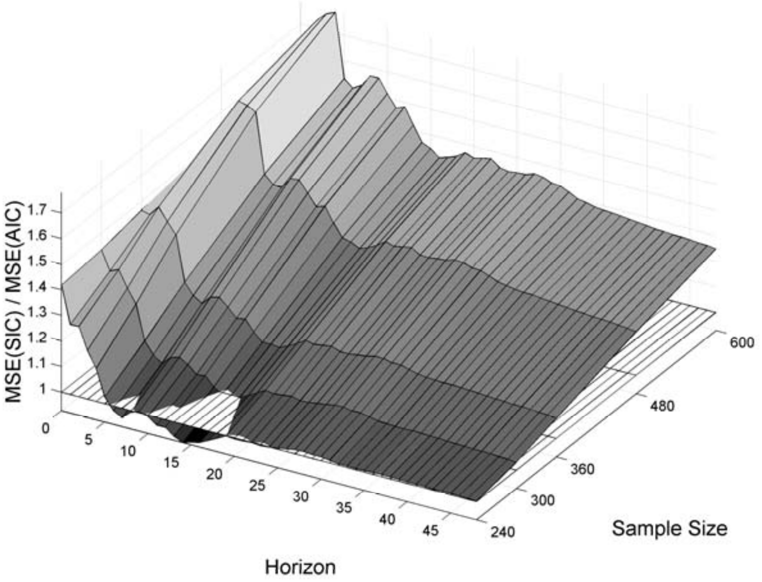
Figures 2-4 show selected response surfaces for the ratios MSE(SIC)/MSE(AIC), MSE(HQC)/MSE(AIC) and MSE(SIC)/ MSE(HQC) as a function of T and h. For the reader's benefit, we also impose a horizontal plane indicating MSE ratios of unity. Figure 2a shows a surface that rarely drops below unity, indicating that the AIC has smaller MSE than the SIC throughout with the exception of a small region for $T = 240$ and intermediate horizons. The most important gains of the AIC relative to the SIC occur at horizons of up to two years. Consistent with the examples in Kilian (2001), the response surfaces in Figure 2a are quite choppy for the first two years after the shock. At short horizons, the MSE of the SIC is up to 1.7 times as high as that of the AIC. For large h and small T, the surface drops back toward 1. There is a clear tendency for the relative accuracy of the AIC to increase with the sample size, however, and at longer horizons the MSE ratio may easily exceed 1.2 for T large.

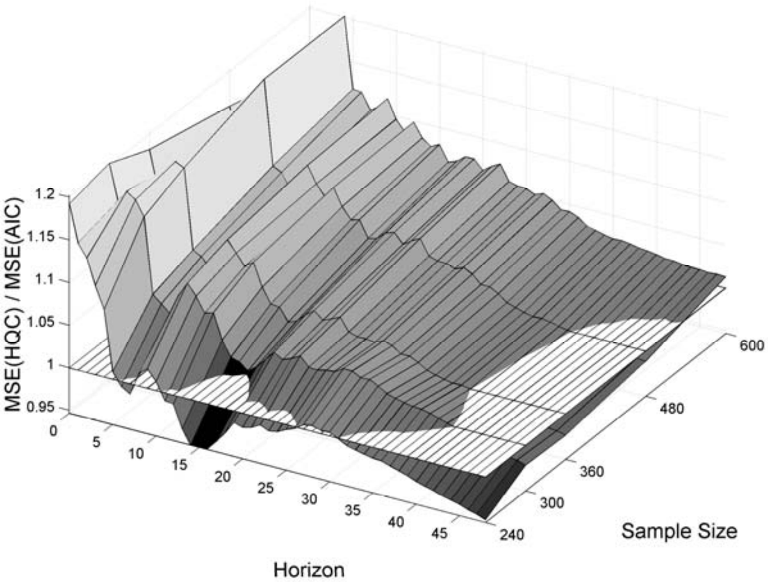
Figure 2b shows the average MSE of the AIC relative to the HQC. In addition to the same anomaly as in Figure 2a for small T and intermediate h, we observe a second region for which the surface drops below unity at horizons in excess of three years and T up to 480. This drop is most pronounced for $T = 240$. Even for $T = 240$, however, as Table 3a shows, the average performance of the AIC is still as good as that of the HQC. For larger T, the relative gains of the AIC for horizons shorter than three years easily compensate for the slight advantages of the HQC for horizons in excess of three years.

For quarterly VAR models, Figure 3a shows that the HQC in most cases is more accurate than the AIC. The MSE ratio may fall as low as 50% in small samples. Only for large sample sizes and short horizon, this pattern is reversed and the HQC actually has an MSE that exceeds that of the AIC by up to 10%. Figure 3b shows the corresponding MSE(SIC)/MSE(HQC) ratios. The MSE ratio is increasing in the sample size and decreasing in h, and reach a factor of almost 1.5 for large T and small h. Figure 3b suggests that, except for $T < 120$, the HQC is clearly the preferred criterion for impulse response analysis in quarterly VAR models for all but the longest horizons of interest. For $T < 120$ the SIC in turn is most accurate for all but the shortest horizons. Thus, the tradeoffs between alternative horizons are minimal. In both subplots, the response surfaces are much smoother than for the monthly models. The apparent reason is that the population impulse response functions for the quarterly VAR models tend to be much smoother than for the monthly VAR models.

For the quarterly VEC models in Figure 4a, the HQC uniformly dominates the AIC. The relative gains from using the HQC range from an MSE reduction of more than 80% for the smallest sample sizes to a few percent for $T = 200$. Figure 4b shows that the SIC dominates the HQC (and by implication the AIC) for most

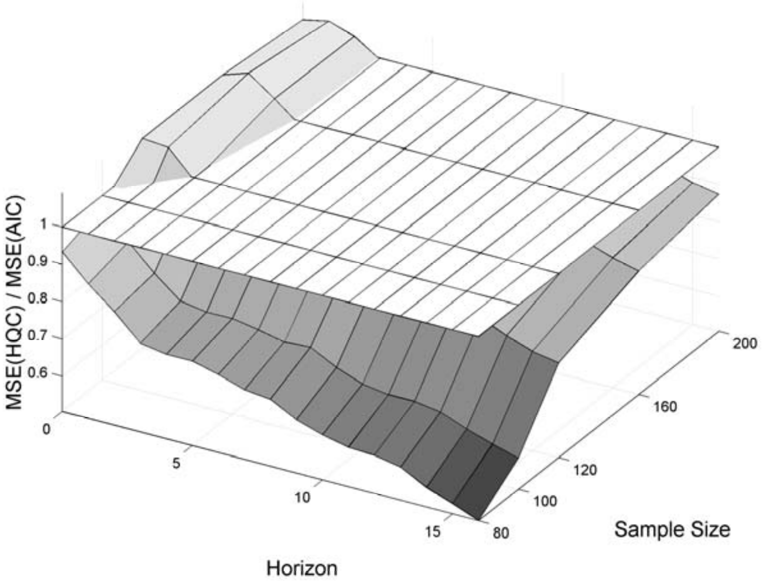
Figure 2
Relative MSEs by Horizon and Sample Size
Monthly VAR Models

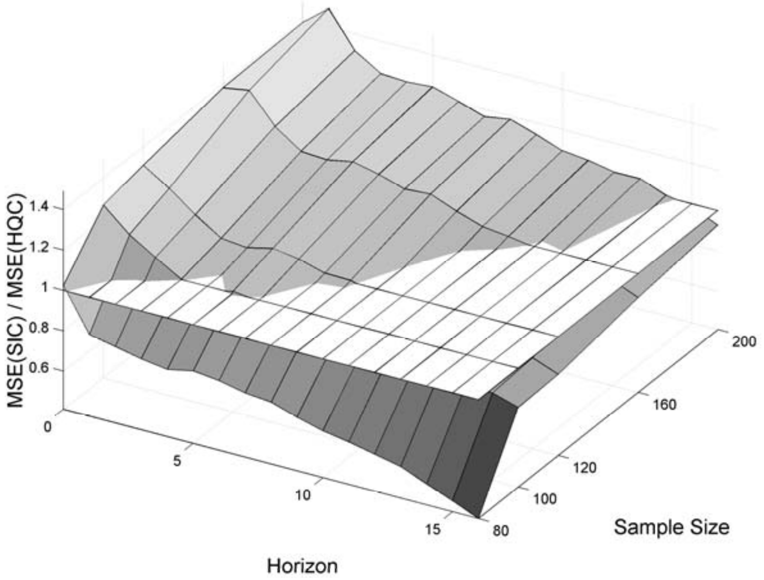




NOTES: Average MSE ratios across all impulse response estimates for all DGPs within model class.

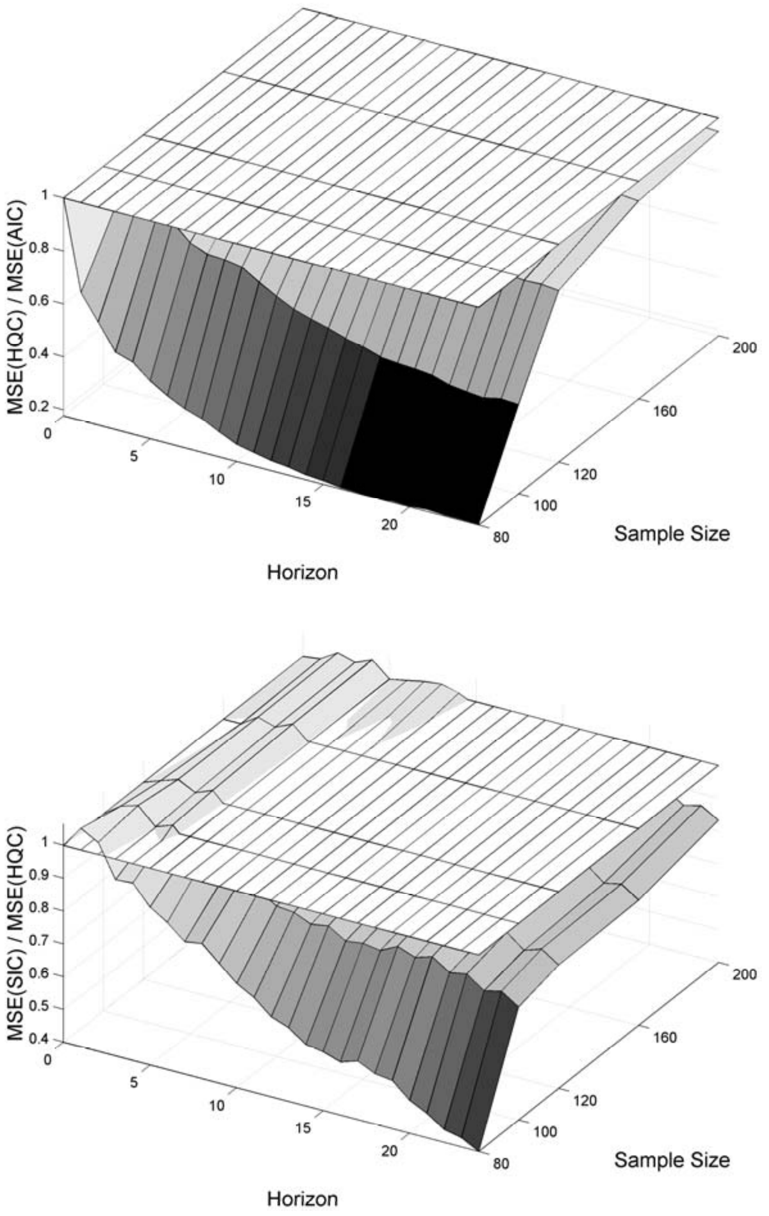
Figure 3
Relative MSEs by Horizon and Sample Size
Quarterly VAR Models





NOTES: See Figure 2.

Figure 4
Relative MSEs by Horizon and Sample Size
Quarterly VEC Models



NOTES: See Figure 2.

horizons, with the exception of the first few quarters. At longer horizons the gains from using the SIC may be as large as 60% relative to the HQC for T = 80, but they decline to about 15% for T = 200. In contrast, for short horizons, the worst loss from using the SIC is a less than 10% increase in the MSE for T = 80. Thus, overall, the SIC compares favorably to the HQC when used for the purpose of constructing persistence profiles based on quarterly VEC models. The relatively smooth response surfaces reflect the fact that the underlying population persistence profiles themselves are fairly smooth. There are no significant tradeoffs across horizons of interest.

Question 5: Why are the results different for the three types of DGPs?

Our results show that the distinction between different classes of models and types of impulse responses is practically important. For example, the practical recommendations for quarterly and for monthly VAR models differ systematically. There are two basic reasons for this result. One is the difference in sample sizes. Although one could attempt to control for the sample size by considering the same sample sizes for quarterly and for monthly VAR models such an exercise would be of little interest for applied work. For example, sample sizes of 240 to 600 amount to time spans of 60 to 150 years of quarterly data. Clearly, such large samples are unrealistic for quarterly data.

The other reason for the differences in results is that the dynamics embodied in quarterly and monthly VAR models differ (even controlling for the sample size). Monthly models are not simply quarterly models estimated on larger sample sizes. First, monthly and quarterly VAR models tend to use different data series for the same concept (e.g., GDP deflator vs. CPI for the price level, GDP vs. industrial production for output) and these series will in general behave differently. Second, one would generally expect a high-frequency model to require a different lag structure than a low-frequency model. This phenomenon is familiar to applied researchers. To be concrete, suppose an AR(1) model well-approximates annual inflation. This does not imply that the corresponding monthly model of inflation is an AR(1). Typically, the monthly data will be well-approximated by an AR model of low order, say an AR(4) and there is no compelling reason for the impulse responses of the two models to look the same nor for the performance of the lag-order selection criteria to be the same, even if the annual AR(1) model is estimated with as many observations as the monthly AR(4) model.

One way of disentangling the effects of the data frequency and of the data set on the relative accuracy of information criteria would be to aggregate the monthly data to quarterly frequency, while maintaining the same VAR model. This proposal is not as straightforward as it may seem because time aggregation destroys the validity of the identifying assumptions used in the monthly structural VAR models. Specifically, these models use exclusion restrictions that postulate no feedback from one variable to another within a month. Clearly, after time aggregation, there will be feedback within the quarter, so we can no longer use the same identification scheme. If we did, this would amount to changing the VAR model along with the data frequency.

We can, however, illustrate the role of time aggregation by conducting a similar exercise using reduced form (as opposed to structural) impulse responses. The basic idea is to repeat the analysis underlying Table 2 using reduced form impulse responses, with the aim of showing, first, that for the original monthly data the reduced form VAR results are qualitatively the same as the structural VAR results, and second, that time aggregation to the quarterly frequency can explain the differences in the rankings between Table 2a and Table 2b. For expository purposes we focus on the Eichenbaum-Evans data set. Table 5a and 5b below show that indeed there is little difference between the MSE rankings, whether we study structural impulse responses or reduced form impulse responses. In both case, for the Eichenbaum-Evans data set, the AIC works best.

Table 5c shows the results for reduced form impulse responses based on the quarterly data obtained by aggregating the Eichenbaum-Evans data set from monthly to quarterly frequency. For $T = 80$ the SIC implies the most accurate impulse responses; for larger sample sizes the HQC does. This is qualitatively the same pattern we found for quarterly VAR models in Table 2b and supports the conjecture that the differences in simulation results between monthly and quarterly VAR models are mainly a consequence of time aggregation.

This leaves the question of why there are differences between the practical recommendations for quarterly VAR models and for quarterly VEC models. We already stressed that persistence profiles in VEC models are fundamentally different statistics from conventional impulse response functions in unrestricted VAR models. Another difference is that our VEC models utilize additional structure in estimation in that we (appropriately) impose the known cointegrating vector prior to selecting the lag order of the VEC presentation of the system. A natural conjecture would be that if instead we selected the lag order on the unrestricted VAR presentation (as we did for the quarterly VAR models) and imposed the known cointegrating vector only in constructing the persistence profiles, our VEC model results might be more similar to those for the quarterly VAR models. Additional simulation analysis suggests that this is not the case.

For expository purposes, we focus on the Johansen-Juselius design. Table 6a serves as a reminder that for the Johansen-Juselius design – as shown in Table 2c for the average of all quarterly VEC models – the SIC is the most accurate model selection criterion regardless of sample size, when the lag order is selected after imposing the known cointegrating vector. Table 6b shows that the SIC is still

Table 5: Average MSE Ratio for Impulse Response Estimates by Criterion Relative to Model Based on True Lag Order

Eichenbaum-Evans Data Set

(a) Monthly VAR Model: Structural Impulse Responses

T	AIC	HQC	SIC	LR1	SLR1	LM1
240	1.00	1.26	1.35	1.22	1.02	1.17
300	1.00	1.29	1.48	1.14	1.03	1.11
360	1.00	1.15	1.63	1.10	1.03	1.08
480	1.00	1.02	1.94	1.07	1.02	1.05
600	1.00	1.01	2.13	1.05	1.02	1.04

(b) Monthly VAR Model: Reduced Form Impulse Responses

T	AIC	HQC	SIC	LR1	SLR1	LM1
240	1.00	1.12	1.17	1.24	1.03	1.18
300	1.00	1.19	1.30	1.16	1.03	1.11
360	1.00	1.11	1.43	1.12	1.03	1.08
480	1.00	1.01	1.71	1.07	1.03	1.05
600	1.00	1.01	1.91	1.06	1.02	1.04

(c) Time-Aggregated Quarterly VAR Model: Reduced Form Impulse Responses

T	AIC	HQC	SIC	LR1	SLR1	LM1
80	2.05	1.04	0.88	3.10	1.17	1.93
100	1.19	1.01	1.05	2.37	1.12	1.56
120	1.06	1.00	1.08	1.92	1.09	1.39
160	1.02	1.00	1.06	1.50	1.08	1.25
200	1.01	1.00	1.03	1.34	1.06	1.14

NOTES: Averages of ratios are calculated as geometric means.

the most accurate criterion for all sample sizes, even when the lag order is estimated on the unrestricted VAR presentation. In fact, the differences in the SIC's accuracy are extremely small, whether the lag order is determined based on the unrestricted VAR presentation or the restricted VEC presentation of the system.

Table 6: Average MSE Ratio for Impulse Response Estimates by Criterion Relative to Model Based on True Lag Order

Johansen-Juselius Data Set

(a) Known Cointegrating Vector, Lag Order Estimation on Restricted VAR

T	AIC	HQC	SIC
80	1.92	0.63	0.36
100	1.15	0.72	0.48
120	1.07	0.85	0.56
160	1.03	0.94	0.71
200	1.02	0.98	0.85

(b) Known Cointegrating Vector, Lag Order Estimation on Unrestricted VAR

T	AIC	HQC	SIC
80	2.25	0.64	0.36
100	1.18	0.73	0.48
120	1.08	0.85	0.56
160	1.04	0.94	0.71
200	1.02	0.98	0.85

NOTES: Averages of ratios are calculated as geometric means.

Question 6: How robust are the simulation results?

We now address the sensitivity of our main results to (a) the true lag order, (b) the number of variables in the VAR model, (c) the choice of the data sets.

(a) How sensitive are the main results to the true lag order?

One would expect that – all else equal - highly parsimonious lag order selection criteria such as the SIC or HQC will tend to be at a disadvantage for large p_0.

This conjecture suggests that for a given sample size the accuracy of the AIC-based impulse responses should improve relative to the HQC- and SIC-based estimates (and similarly for the HQC relative to the SIC), as p_0 increases. We find no support, however, for the notion that larger values of p_0 for the same sample size favor less parsimonious criteria such as the AIC. For monthly VAR processes, the AIC tends to be more accurate than the other criteria for all sample sizes almost regardless of the value of p_0. Figure 5 illustrates this point. It shows the average MSE ratios for alternative horizons *h* and values of p_0, given the sample size *T*. We focus on the accuracy of the AIC relative to the SIC. The first panel shows that, even for sample sizes as small as *T* = 240, the AIC compares favorably to the SIC. As *T* increases, the relative accuracy of the AIC further improves. For sample sizes in excess of *T* = 300 the AIC uniformly dominates the SIC for all values of *h* and p_0. The greatest gains are achieved for *T* = 600, as shown in the second panel.

Similarly, for quarterly VAR processes, there is no evidence that less parsimonious criteria are more accurate for large p_0. Finally, for quarterly VEC models, the SIC dominates the other criteria for all sample sizes, regardless of the value of p_0. We conclude that at least over the range of p_0 we considered, the value of p_0 is not an important determinant of the relative accuracy of alternative lag order selection criteria.

(b) How sensitive are the main results to the number of model variables?

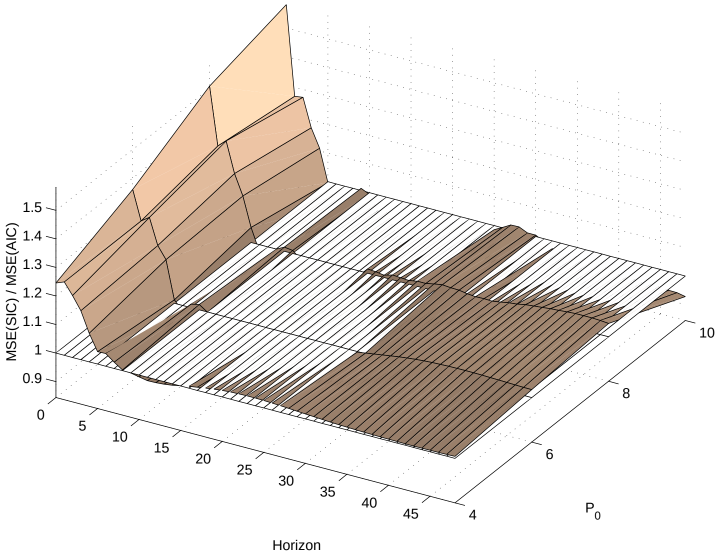
Another question of practical interest is how much our results are affected by *K*, the number of variables included in the VAR system. A simulation study that systematically analyzes this question for a given class of models, while controlling for *h*, *T*, and p_0, would be computationally prohibitive. Based on the available evidence, we nevertheless can report that there is no evidence that the average MSE ratios in our study are systematically affected by *K*.

(c) How sensitive are the main results to the choice of data sets?

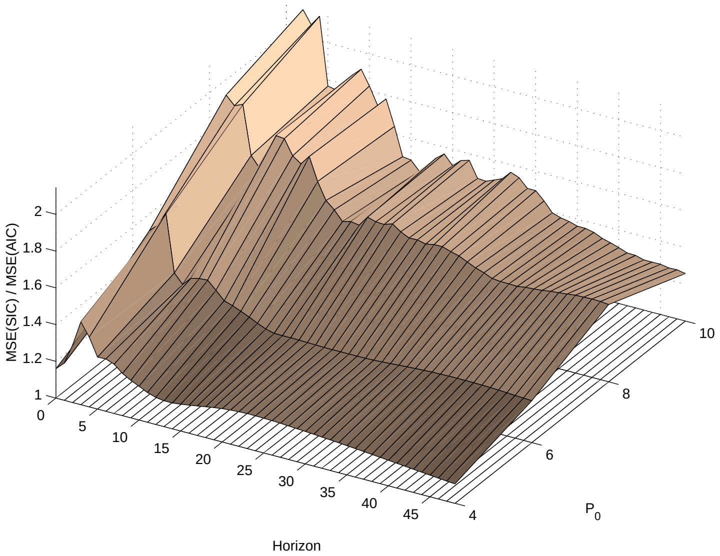
An obvious concern is how sensitive the main results are to the choice of data sets listed in Table 1. We address this question in part by using a comparatively large number of alternative specifications. We also conducted a sensitivity analysis fashioned after the idea of the delete-one jackknife. Specifically, we recalculated the results for each class of models after deleting one data set (and all associated DGPs) at a time. This procedure allowed us to construct a crude measure of the sensitivity of the results to the choice of data sets.

Figure 5
Relative MSEs by Horizon and True Lag Order
Monthly VAR Models

T = 240



T = 600



NOTES: See Figure 2.

For the monthly DGPs we find that the rankings in Table 2a are virtually unchanged after discarding one data set at a time. Only in one of four cases, the HQC and SIC appear more accurate for small sample sizes than the AIC. Specifically, when the Eichenbaum-Evans data set is discarded, the HQC (and to a lesser extent the SIC) are more accurate than the AIC for T = 240 and T = 300 by 0.07 and 0.06, respectively. On the other hand, after dropping the Strongin data set, for example, the MSE ratio of the SIC and HQC for T = 240 (300) increases by 0.04 (0.05) and 0.06 (0.06), respectively. On balance, our results appear to be representative. Moreover, other qualitative features (such as the tendency of the accuracy of SIC-based estimates to worsen drastically as the sample size is increased) are robust. Similarly, for the quarterly VAR DGPs we find that when some data sets are excluded, the ranking of the SIC and the HQC for T = 120 may alternate. Overall, the rankings in Table 2b are remarkably robust, however. For the quarterly VEC models the only difference in results is that after excluding the Johansen-Juselius data set, for T = 120 and T = 160 the ranking of the SIC and HQC changes, but the differences in accuracy are very small in all cases (0.02 and 0.01, respectively) and do not reflect important practical advantages of either criterion. In the other three cases, the ranking of the SIC is not affected. We conclude that we can be reasonably confident that our results and practical recommendations are not inadvertently driven by the choice of data sets.

As noted earlier, only the VAR models based on the Galí (1999) data set employ long-run identifying restrictions, whereas all other models rely on short-run exclusion restrictions for identification. Thus, it is of special interest to investigate whether there are any systematic differences in the results. Table 7a shows the average MSE ratios by criterion for all quarterly VAR models excluding the Galí data set; Table 7b shows the corresponding results only for the VAR models based on the Galí data set. There is some indication that parsimony is slightly more beneficial for the VAR models based on Galí (1999), but overall the conclusions are similar. In both cases we find that the SIC works best for small sample sizes, whereas for larger sample sizes the HQC results in more accurate impulse responses than the SIC.

5. CONCLUDING REMARKS

We compared the most commonly used lag-order selection criteria for VAR models in terms of the MSE of the implied impulse response estimates relative to the MSE based on knowing the true lag order. The criteria included in the study were the SIC, the HQC, the AIC, the sequential likelihood ratio (LR) test and its modification by Sims (referred to as SLR), and finally the sequential Portmanteau (or LM) test. The latter three criteria were evaluated at an individual nominal significance level of 1% and 5% each, resulting in a total of nine criteria. Our

Table 7: Average MSE Ratio for Impulse Response Estimates by Criterion Relative to Model Based on True Lag Order

(a) Quarterly VAR Models Excluding Galí (1999) Data Set

T	AIC	HQC	SIC	LR1	SLR1	LM1
80	2.06	1.34	0.81	2.60	1.18	1.65
100	1.34	0.95	0.92	1.94	1.08	1.34
120	1.08	0.99	1.00	1.73	1.07	1.26
160	1.01	1.01	1.14	1.49	1.07	1.17
200	1.01	1.01	1.23	1.35	1.07	1.12

(b) Quarterly VAR Models Based on Galí (1999) Data Set

T	AIC	HQC	SIC	LR1	SLR1	LM1
80	1.09	0.82	0.73	1.24	1.03	0.76
100	1.13	0.86	0.81	1.26	1.12	0.86
120	1.15	0.91	0.87	1.27	1.15	0.96
160	1.20	0.96	0.97	1.28	1.22	1.10
200	1.21	1.02	1.03	1.30	1.24	1.18

NOTES: Averages of ratios are calculated as geometric means.

simulation study involved a total of 180 design points and supported the following main conclusions. First, our results show that that the choice of lag-order selection criterion has quantitatively important implications for the accuracy of VAR impulse response estimates. In some cases, the MSE of the impulse response estimates increased more than five-fold relative to the estimates based on knowing the true lag order, and the MSEs of different criteria in some cases differed by a factor of eight. Second, we found that, contrary to what one might have conjectured, our practical recommendations do not appear to be very sensitive to the horizon of interest. The chief determinant of the relative accuracy of alternative criteria for a given class of models appears to be the sample size. Third, we concluded that no criterion dominates in all circumstances, but the results of the simulation study are nevertheless clear-cut and informative and allow several practical recommendations for applied researchers.

We showed that, in general, sequential LR and LM tests do not perform well, especially for small sample sizes. The SLR test performs better than the LR test in all cases and better than the LM test in most cases. All three sequential tests are in turn dominated by one or more of the information criteria. The relative performance of the other AIC, HQC and SIC differs across classes of models and types of impulse responses. We focused on three classes of models and impulse response functions common in applied work. First, for structural and semi-structural impulse responses in monthly VAR models, the AIC tends to produce the most accurate impulse response estimates for all realistic sample sizes. The average reduction in mean-squared error from using the AIC can be as high as 27% relative to the SIC and 6% relative to the HQC. Second, for structural and semi-structural impulse responses based on quarterly VAR models, the HQC appears to be most accurate except for sample sizes of fewer than 120 quarters, for which the SIC was found to improve accuracy by up to 33% relative to the HQC. Third, for persistence profiles based on quarterly VEC models, the SIC tends to be most accurate for all sample sizes we considered (with gains of up to 82% relative to the AIC and 33% relative to the HQC).

The fact that no one criterion works best for all classes of models should not be surprising. We discussed in detail the sources of these differences. It should be borne in mind that our simulation results – although based on an extensive study - are necessarily tentative and limited by the simulation design. We also note that the results may differ if the objective of estimating the VAR model is forecasting or the construction of variance decompositions, for example, or if the underlying process is of infinite order. Finally, we have postulated Gaussian VAR innovations throughout this paper. As noted for example by Kilian and Demiroglu (2000), there is considerable evidence of fat tails and skewness in the unconditional distribution of VAR residuals and of conditional heteroskedasticity in the error term. A study of the effect of such departures from normality on the accuracy of lag-order selection criteria is left as an extension for future work.

REFERENCES

Bernanke, B.S. (1986), ‘Alternative Explorations of the Money-Income Correlation’, *Carnegie-Rochester Conference on Public Policy*, **25**, 49-100.

Bernanke, B.S., and M. Gertler (1995), ‘Inside the Black Box: The Credit Channel of Monetary Policy Transmission’, *Journal of Economic Perspectives*, **9**, 27-48.

Bernanke, B.S., M. Gertler and M.W. Watson (1997), ‘Systematic Monetary Policy and the Effects of Oil Price Shocks’, *Brookings Papers on Economic Activity*, 91-142.

Bernanke, B.S., and I. Mihov (1998a), ‘Measuring Monetary Policy’, *Quarterly Journal of Economics*, **113**, 869-902.

Bernanke, B.S., and I. Mihov (1998b), ‘The Liquidity Effect and Long-run Neutrality’, *Carnegie-Rochester Conference Series on Public Policy*, **49**, 149-194.

Blanchard, O. (1989), ‘A Traditional Interpretation of Macroeconomic Fluctuations’, *American Economic Review*, **79**, 1146-1164.

Blanchard, O., and D. Quah (1989), ‘The Dynamic Effects of Aggregate Demand and Supply Disturbances’, *American Economic Review*, **79**, 655-673.

Christiano, L.J., Eichenbaum, M., and C.L. Evans (1996), ‘The Effects of Monetary Policy Shocks: Some Evidence from the Flow of Funds’, *Review of Economics and Statistics*, **78**, 16-34.

Christiano, L.J., Eichenbaum, M., and C.L. Evans (1999), ‘The Effects of Monetary Policy Shocks: What Have We Learned and to What End?’, forthcoming in: J.B. Taylor and M. Woodford (eds.), *Handbook of Macroeconomics. Vol. 1A*, Elsevier Science Publishers, Amsterdam.

Eichenbaum, M., and C.L. Evans (1995), ‘Some Empirical Evidence on the Effects of Shocks to Monetary Policy on Exchange Rates’, *Quarterly Journal of Economics*, **110**, 975-1010.

Galí, J. (1992), ‘How Well Does the IS-LM Model Fit Postwar U.S. Data?’, *Quarterly Journal of Economics*, **107**, 709-738.

Galí, J. (1999), ‘Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?’, *American Economic Review,* **89**, 249-271.

Granger, C.W.J., M.L. King, and H. White (1995), ‘Comments on Testing Economic Theories and the Use of Model Selection Criteria’, *Journal of Econometrics*, **67**, 173-187.

Hamilton, J.D., and A.M. Herrera (2004), “Oil Shocks and Aggregate Macroeconomic Behavior: The Role of Monetary Policy”, *Journal of Money, Credit, and Banking,* **36**, 265-286.

Hosking, J.R.M. (1981), ‘Lagrange-Multiplier Tests of Multivariate Time Series Models’, *Journal of the Royal Statistical Society (Series B)*, **43**, 219-230.

Johansen, S., and K. Juselius (1990), ‘Maximum Likelihood Estimation and Inference on Cointegration – With Applications to the Demand for Money’, *Oxford Bulletin of Economics and Statistics*, **52**, 169-210.

Kabaila, P.M. (1995), ‘The Effect of Model Selection on Confidence Regions and Prediction Regions’, *Econometric Theory*, **11**, 537-549.

Keating, J.W., and J.V. Nye (1998), ‘Permanent and Transitory Shocks to Real Output: Estimates from Nineteenth Century and Postwar Data’, *Journal of Money, Credit and Banking*, **30**, 231-251.

Kilian, L. (1999), ‘Exchange Rates and Monetary Fundamentals: What Do We Learn From Long-Horizon Regressions?’, *Journal of Applied Econometrics*, **14**, 491-510.

Kilian, L. (2001), ‘Impulse Response Analysis in Vector Autoregressions with Unknown Lag Order’, *Journal of Forecasting*, **20**, 161-179.

Kilian, L., and U. Demiroglu (2000), ‘Residual-Based Tests of Normality in Autoregressions: Asymptotic Theory and Simulation Evidence’, *Journal of Business and Economic Statistics*, **18**, 40-50.

Leeper, E.M. (1997), ‘Narrative and VAR Approaches to Monetary Policy: Common Identification Problems’, *Journal of Monetary Economics*, **40**, 641-657.

Lütkepohl, H. (1985), ‘Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process’, *Journal of Time Series Analysis*, **6**, 35-52.

Lütkepohl, H. (1990), ‘Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models’, *Review of Economics and Statistics*, 72, 116-125.

Lütkepohl, H. (1993), *An Introduction to Multiple Time Series Analysis*, 2nd ed., Springer-Verlag, New York.

Lütkepohl, H., and H.-E. Reimers (1992), ‘Impulse Response Analysis of Cointegrated Systems’, *Journal of Economic Dynamics and Control*, **16**, 53-78.

Nickelsburg, G. (1985), ‘Small-Sample Properties of Dimensionality Statistics for Fitting VAR Models to Aggregate Economic Data. A Monte Carlo Study’, *Journal of Econometrics*, **28**, 183-192.

Paulsen, J. (1984), ‘Order Determination of Multivariate Autoregressive Time Series with Unit Roots’, *Journal of Time Series Analysis*, **5**, 115-127.

Paulsen, J., and D. Tjøstheim (1985), ‘On the Estimation of Residual Variance and Order in Autoregressive Time Series’, *Journal of the Royal Statistical Society (Series B)*, **47**, 216-228.

Pesaran, M.H., and Y. Shin (1996), ‘Cointegration and Speed of Convergence to Equilibrium’, *Journal of Econometrics,* **71**, 117-143.

Pesaran, M.H., Y. Shin, and R.J. Smith (2000), ‘Structural Analysis of Vector Error Correction Models with Exogenous I(1) Variables’, *Journal of Econometrics,* **97**, 293-343.

Pesaran, M.H., and R.P. Smith (1998), ‘Structural Analysis of Cointegrating VARs’, *Journal of Economic Surveys,* **12**, 471-506.

Phillips, P.C.B., and W. Ploberger (1994), ‘Posterior Odds Testing for a Unit Root with Data-Based Model Selection’, *Econometric Theory*, 10, 774-808.

Pötscher, B.M. (1995), ‘Comment on “The Effect of Model Selection on Confidence Regions and Prediction Regions” by P. Kabaila’, *Econometric Theory*, **11**, 550-559.

Quinn, B.G. (1980), ‘Order Determination for a Multivariate Autoregression’, *Journal of the Royal Statistical Society (Series B)*, **42**, 182-185.

Quinn, B.G. (1988), ‘A Note on AIC Order Determination for Multivariate Autoregressions’, *Journal of Time Series Analysis*, **9**, 241-245.

Rotemberg, J.-J., and M. Woodford (1996), ‘Imperfect Competition and the Effects of Energy Price Increases on Economic Activity, *Journal of Money, Credit and Banking*, **28**, 550-577.

Schwarz, G. (1978), ‘Estimating the Dimension of a Model’, *Annals of Statistics,* 6, 461-464.

Shapiro, M., and M.W. Watson (1988), 'Sources of Business Cycle Fluctuations', in: *NBER Macroeconomics Annual*, **3**, Cambridge, MA: MIT Press, 111-148.

Sims, C.A. (1980), 'Macroeconomics and Reality', *Econometrica*, **48**, 1-48.

Sims, C.A. (1986), 'Are Forecasting Models Usable for Policy Analysis?', *Federal Reserve Bank of Minneapolis Quarterly Review*, **10**, 2-16.

Sims, C.A., and T. Zha (1998), 'Bayesian Methods for Dynamic Multivariate Models', *International Economic Review*, **39**, 949-968.

Sin, C.-Y., and H. White (1996), 'Information Criteria for Selecting Possibly Misspecified Parametric Models', *Journal of Econometrics*, **71**, 207-225.

Söderlind, , P., and A. Vredin (1996), 'Applied Cointegration Analysis in the Mirror of Macroeconomic Theory', *Journal of Applied Econometrics*, **11**, 363-381.

Strongin, S. (1995), 'The Identification of Monetary Policy Disturbances: Explaining the Liquidity Effect', *Journal of Monetary Economics*, **35**, 463-497.

Watson, M.W. (1994), 'Vector Autoregressions and Cointegration', in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics – Volume IV*, Elsevier Science, 2843-2915.

Wickens, M.R. (1996), 'Interpreting Cointegrating Vectors and Common Stochastic Trends', *Journal of Econometrics*, **74**, 255-271.