DISCUSSION PAPER SERIES

No. 2685

A PRACTITIONER'S GUIDE TO LAG-ORDER SELECTION FOR VECTOR AUTOREGRESSIONS

Ventzislav Ivanov and Lutz Kilian

INTERNATIONAL MACROECONOMICS

Centre for Economic Policy Research

www.cepr.org

An online version of this Paper can be found at www.cepr.org/pubs/dps/DP2685.asp

ISSN 0265-8003

A PRACTITIONER'S GUIDE TO LAG ORDER SELECTION FOR VECTOR AUTOREGRESSIONS

Ventzislav Ivanov, University of Michigan
Lutz Kilian, University of Michigan and CEPR

Discussion Paper No. 2685
January 2001

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 2685

January 2001

ABSTRACT

A Practitioner's Guide to Lag Order Selection for Vector Autoregressions*

An important preliminary step in impulse response analysis is to select the vector autoregressive (VAR) lag order from the data, yet little is known about the implications of alternative lag order selection criteria for the accuracy of the impulse response estimates. In this Paper, we compare the criteria most commonly used in applied work in terms of the mean-squared error of the implied impulse response estimates. We conclude that for monthly VAR models, the Akaike information criterion (AIC) produces 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 (VEC) models, the SIC is the most accurate criterion for all realistic sample sizes. Sequential Lagrange multiplier and likelihood ratio tests cannot be recommended.

JEL Classification: C32, C51, E37, E47
Keywords: impulse responses, lag orders, model selection, VAR, VEC

Lutz Kilian
Department of Economics
Lorch Hall, M-109
University of Michigan
Ann Arbor, MI 48109-1220
USA
Tel: (1 734) 764 2320
Fax: (1 734) 764 2769
Email: lkilian@umich.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=132598

Ventzislav Ivanov
Department of Economics
Lorch Hall, M-109
University of Michigan
Ann Arbor, MI 48109-1220
USA
Email: ventzi@umich.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid= 153163

*We thank Phil Howrey, Jan Kmenta, Hashem Pesaran, Paul Rilstone and Shinichi Sakata for comments on earlier drafts of this Paper.

Submitted 22 September 2000

NON-TECHNICAL SUMMARY

Impulse response analysis based on estimated vector autoregressive (VAR) models plays an important role in empirical macroeconomics both in assessing the effectiveness of economic policies and in establishing facts to be explained by economic theory. This Paper addresses a practical problem that all users of VAR models face in applied work: how to decide the lag order of the VAR model. Dealing with this problem is important because it may affect the substantive conclusions that are drawn from empirical VAR studies.

VAR models typically focus on a relatively small subset of variables of economic interest and attempt to explain each variable's movements in terms of the lagged values of all economic variables in that system. In the absence of plausible restrictions on the dynamics of the economic system, no restrictions are imposed on the lag structure other than an upper bound on the number of lagged values for each variable. This upper bound is commonly referred to as the lag order of the VAR model.

It is common to study the properties of estimated VAR models by plotting the implied impulse response estimates. Impulse responses trace out the responses of economic variables to unanticipated shocks to the economic system. For example, it is common to study the response of inflation and output to unanticipated changes in interest rates (or other variables typically regarded as being under the control of the monetary authorities). These impulse responses are called structural or semi-structural because they invariably require some short-run or long-run identifying assumptions on the effects of unanticipated changes in economic variables. Other forms of impulse responses (such as persistence profiles) measure the speed with which economic variables revert toward a long-run equilibrium relationship suggested by theory.

Regardless of the type of impulse response used, the shape, sign and magnitude of the impulse responses will be sensitive to the lag order of the estimated VAR system. Unfortunately, economic theory rarely provides guidance as to how many lags are needed to characterize the interaction of economic variables. For that reason, various statistical tools have been developed to help practitioners to select the lag order of the VAR model, given a set of variables of interest. Some tools such as the Akaike information criterion (AIC), the Schwarz criterion (SIC) and the Hannan-Quinn criterion (HQC) trade off a better fit of the model against a penalty term that depends on the number of parameters fitted. Other tools such as likelihood ratio tests and Lagrange multiplier tests involve a sequence of pairwise tests of larger against smaller models.

These criteria have been in use for many years, yet little is known about the implications of alternative lag order selection criteria for the accuracy of the

implied impulse response estimates. In this Paper, we compare the criteria most commonly used in applied work in terms of the mean-squared error of the impulse response estimates. We carefully distinguish several cases: monthly VAR models involving larger sample sizes, but also more lags; quarterly VAR models with fewer lags, but also smaller sample sizes; and finally quarterly vector error correction (VEC) models that impose additional restrictions on the time-series behaviour of economic variables. We conduct a simulation study that is based on a large number of models estimated by leading practitioners in the empirical literature, ranging from bivariate models to models with seven variables. Thus, our findings are of immediate interest to other VAR users in empirical macroeconomics.

In total, our study includes 180 different models. Based on this evidence, we conclude that for monthly VAR models, the AIC produces the most accurate structural and semi-structural impulse response estimates for realistic sample sizes. For quarterly VAR models, the HQC appears to be the most accurate criterion with the exception of sample sizes smaller than 120, for which the SIC is more accurate. For persistence profiles based on quarterly VEC models, the SIC is the most accurate criterion for all realistic sample sizes. In contrast, sequential Lagrange multiplier and likelihood ratio tests cannot be recommended.

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. An important preliminary step in empirical studies is to select the order of the autoregression based on the data. A number of alternative 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 compare the five criteria most commonly used in applied work in terms of the mean-squared error of the implied impulse response estimates. These 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) 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. For a detailed review of all five procedures see Lütkepohl (1993).[1]

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. 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 (2000).[2] 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 focus on the five lag order selection criteria that are most widely used in the applied VAR literature. For example, the LR test has been used by Blanchard (1989), Keating and Nye (1998) and Bernanke and Mihov (1998a); 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), 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 quarterly VAR models, monthly VAR models and quarterly vector error correction (VEC) models, resulting in a total of 180 design points. We also conduct a sensitivity analysis of our main results.

[1] All these criteria are motivated by classical statistical theory. We do not pursue Bayesian approaches to model selection. The reader is referred to Sims and Zha (1998).

[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 Nickelsburg 1985, Lütkepohl 1985). There is no simple mapping between this criterion and the accuracy of the impulse response estimates. Lütkepohl (1985, 1993) in addition presents some simulation evidence for VAR prediction mean squared errors, but these results have no direct implications for impulse response analysis and are based on a very small set of stylized bivariate and trivariate models.

We briefly review the five lag-order selection criteria in section 2. The simulation design is 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. We provide details for the three types of models most commonly used in applied work: monthly VAR models in levels, quarterly VAR models in levels and quarterly models in VEC form. Section 5 contains the concluding remarks.

2. REVIEW OF LAG-ORDER SELECTION CRITERIA

We postulate that the true process is an *N*-dimensional autoregression of order p_0, which may be represented in VAR or in VEC form. Abstracting from deterministic regressors, the first three lag-order selection criteria are:

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

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

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

where *T* is the effective sample size and $\overline{\Sigma}$ 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 $\{p : 1 \leq p \leq \overline{p}\}$ where $\overline{p} \geq p_0$ (see Quinn 1980; Paulsen and Tjøstheim 1985; Quinn 1988). It can be shown that $\hat{p}^{SIC} \leq \hat{p}^{AIC}$ for $T \geq 8$, $\hat{p}^{SIC} \leq \hat{p}^{HQC}$ for all T, and $\hat{p}^{HQC} \leq \hat{p}^{AIC}$ for $T \geq 16$. 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 described by Lütkepohl (1993). We follow Lütkepohl (1985) in fixing the nominal significance level of the LR test at 5% (1%) at each step in the sequential procedure and use the asymptotic $\chi^2(N^2)$ 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

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

for $i = 1,\ldots,\overline{p}-1$. 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 and LR5.

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. Hoskings (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:

$$LM(p) = T^2 \sum_{i=1}^{s} (T-i)^{-1} tr\left(\hat{C}_i{}'\hat{C}_0^{-1}\hat{C}_i\hat{C}_0^{-1}\right),$$

where $\hat{C}_i = \frac{1}{T}\sum_{t=i+1}^{T} \hat{u}_t\hat{u}_{t-i}{}'$ and $\hat{u}_t$ denotes the residual from a VAR(p) process. The LM test statistic is calculated for $p = 1,\ldots,\overline{p},$ in ascending order. The nominal significance level at each step is set to 5% (1%). 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 $s/T \to 0$ at a suitable rate as $T \to \infty$, the LM test has an asymptotic $\chi^2\left(N^2(s-p)\right)$-distribution where $s > p$. We adopt the convention of setting s equal to the maximum of $T^{1/2}$ (rounded to the nearest integer) and $\overline{p}+1$. This rule results in choices of s that are similar to values used in many empirical studies. The resulting tests are denoted by LM1 and LM5.[3]

3. SIMULATION DESIGN

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). 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

[3] For the processes considered in this study, the asymptotic validity of our criteria does not require the autoregressive roots of the process to be in the stationary region. Watson (1994, p. 2860) shows that for $p \geq 1$ the LR and LM tests remain valid even in the presence of a unit root. Paulsen (1984) establishes the asymptotic validity of the information criteria for unit root processes.

uncovered interest parity condition and the purchasing power parity condition. These are treated separately in the analysis.[4] For a precise definition of persistence profiles and their relationship to other impulse response estimators see Pesaran and Smith (1998).

For the quarterly data, we postulate values of $p_0 \in \{2, 4, 6\}$ for each DGP and for the monthly data values of $p_0 \in \{4, 6, 8, 10\}$. 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, intercepts, etc.) and enforcing exogeneity constraints. The model innovations are assumed to be Gaussian white noise with the same innovation variance as the fitted VAR(p_0) model. For each DGP we consider several sample sizes. For quarterly data, $T \in \{80, 100, 120, 160, 200\}$ and for monthly data $T \in \{240, 300, 360, 480, 600\}$. Altogether, our simulation study includes 180 different design points.[5] For quarterly data, we set $\bar{p} = 8$ and for monthly data $\bar{p} = 12$.

For each design point, we generate 5,000 independent draws of data of length *T*. For each draw, we fit seven VAR models based on the lag orders chosen by the seven alternative lag

---

[4] As Wickens (1996) shows, estimated cointegrating vectors cannot be given an economic interpretation without additional a priori information. We therefore impose cointegrating vectors suggested by theory in estimating the VEC models, even when the original studies do not.

[5] 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.

order selection criteria (SIC, HQC, AIC, LR1, LR5, LM1, LM5). Then we compute all implied 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 quarterly and monthly VAR models and horizons of up to six years for the persistence profiles.[6] Finally, for each criterion, we calculate the mean squared deviation (MSE) of each of the impulse response coefficient estimates from their true values.

4. SIMULATION RESULTS

Given the large number of design points, the simulation results are summarized in graphical response surfaces and tables. We organize the discussion around a number of questions of interest to empirical researchers. We begin with the discussion of some general regularities. Tables 2 and 3 demonstrate 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.

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

A key question of concern to empirical researchers is which of the seven criteria can be expected to produce the most accurate impulse response estimates. Tables 2a-c show the average MSE for each criterion as a function of T only. The MSEs have been expressed relative to the MSE based on knowing the true model and averaged across all DGPs. Throughout the paper, averages of ratios are calculated as geometric means rather than arithmetic means.[7]

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

[7] 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).

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.

Our results show that the distinction between different classes of models and types of impulse responses is practically important. 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 five 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. The sequential LR and LM tests tend to perform poorly for all three classes of models, especially for small sample sizes. The LM and LR tests at the nominal 1% level are more reliable than the corresponding nominal 5% tests. The LM1 test in all cases is the most accurate of the sequential tests, but even the LM1 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. We also recommend that sequential LM and LR tests not be used in applied work.

Question 2: What are the costs and benefits of not imposing 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 model. Table 2a documents 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. Additional simulation evidence (not included in the paper) shows 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. This example is interesting by itself because it is often thought that the AIC will necessarily tend to overestimate the true lag order. More importantly for our purposes, the resulting 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.

Another interesting feature of Table 2a is that the accuracy of the SIC (and to a lesser extent the HQC) actually deteriorates as the sample size is increased. Whereas the AIC's MSE quickly approaches that for the true model, the MSEs of the SIC- and HQC-based impulse response estimates actually worsen, as the sample size is increased. The apparent reason is that these criteria (especially the SIC) systematically and severely underestimate the true lag order for all sample sizes, even for fairly large sample sizes. In contrast, the AIC lag order estimates are clustered increasingly close to the true lag order with increasing sample size. In fact, the probability of the AIC overestimating the true lag order shrinks almost to zero for *T* = 600, consistent with the theoretical results by Paulsen and Tjøstheim (1985) for VAR models. This fact helps to explain the relative performance of these three criteria for large *T*.

Table 2b shows the corresponding results for the quarterly VAR models. All criteria but the SIC lead to dramatic losses in accuracy relative to the true model in small samples, in some cases by a factor of more than two. For larger samples, the MSE of all criteria but the SIC improves in Table 2c. As in the case of the monthly VAR models, the worsening MSE of the SIC for larger *T* reflects the strong downward bias of this lag order selection criterion in small samples. Additional simulation evidence (not included in the paper) shows 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 severely overestimate 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 huge, reaching a factor of more than 5 for the LR tests, of about 2 for the LM 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 a sample size of 600 observations, the AIC can be expected to reduce the MSE of the impulse responses by up to 27% relative to the SIC. For T = 240, the gains range from 0% relative to the HQC and 2 % relative to the SIC to 24% relative to the LR5 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 LR5 test the relative gains are even larger, reaching 88%. 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 41% for the other criteria.

Question 4: How robust are the main simulation results?

We now study the sensitivity of our main results in Tables 2 and 3 to (a) the true lag order, (b) the number of variables in the VAR model, (c) the choice of DGPs, and (d) to the horizon of the impulse response function. To conserve space, in this section we only display the most important simulation results and briefly summarize the other findings.

(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 dominates the other six criteria for all sample sizes almost regardless of the value of p_0. 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 six 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 *N*, 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 *N*.

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

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 DGPs.

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. There are, however, other processes that strongly favor the SIC and HQC, without affecting much the accuracy of the AIC. For example, after dropping the Strongin data set, the MSE 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 SIC's MSE drastically to worsen 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 any one set of DGPs.

(d) How sensitive are the main 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 than the AIC for example. Kilian (2000), on the other hand, stresses the inability of parsimonious lag order selection criteria to capture the nonlinear dynamics of impulse response functions especially at longer horizons. Based on some stylized examples, he suggests that it often will be far less harmful - especially at longer horizons – possibly to overfit the model than to risk underfitting the model. For that reason, one would expect criteria such as the AIC to produce more accurate impulse responses than the SIC or HQC.

This line of reasoning suggests that the relative accuracy of lag order selection criteria may depend on the horizon. To the extent that there is considerable variation in relative accuracy across horizons, the average results in Table 3 for a given sample size may be misleading. We therefore disaggregate the MSE results 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 parsimony matters, but that the required degree of parsimony may differ greatly depending on the sample size and class of DGPs. Although a certain degree of parsimony appears to be helpful for small sample sizes, especially at very long horizons, it tends to be harmful for short horizons regardless of sample size and for all horizons as the sample size increases.

Figures 1-3 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 1a 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 (2000), the response surfaces in Figure 1a 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 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 1b shows the average MSE of the AIC relative to the HQC. In addition to the same anomaly as in Figure 1a 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. We conclude that for most horizons of practical interest, the results in Table 3 will be representative.

For quarterly VAR models, Figure 2a 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 2b 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 2b 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 3a, 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 3b shows that the SIC dominates the HQC (and by implication the AIC) for most 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 well below a 10% increase in the MSE for T = 80. Thus, overall the SIC compares favorably to the HQC 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.

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. The criteria included in the study were the SIC, the HQC, the AIC, the sequential LR test and the sequential Portmanteau (or LM) test. The latter two criteria were evaluated at an individual nominal significance level of 1% and 5% each, resulting in a total of seven criteria. Our 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 even for large sample sizes. In some cases, the MSE of the impulse response estimates increased more than five-fold relative to the true model, and the MSEs of different criteria in some cases differed by a factor of eight. This finding suggests that it is important to report in empirical studies how the lag order of the VAR model was arrived at. 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, to the number of variables in the model, or to the lag order of the underlying true process. 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. These tests cannot be recommended for impulse response analysis. The relative performance of the other three criteria (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 produces 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 is recommended 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 appears 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. There are important differences across classes of models not just in terms of the relevant sample size, but also in the dimensions of the fitted models, in the parameter values of the underlying DGPs and – in some cases – in the method of estimation. For example, a monthly VAR model cannot be interpreted simply as a quarterly VAR model with a larger sample size. Although we could have analyzed the question of how further increases in sample size for a given class of models affect the relative performance of the various criteria, we chose to limit the analysis to sample sizes relevant for applied work.

It should be borne in mind that our simulation results – although based on an extensive study - are necessarily tentative and limited to impulse response analysis. We addressed the first concern by conducting a sensitivity analysis. The available evidence suggests that our simulation results are reasonably robust to the choice of DGPs. 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. We leave such extensions 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.

Hosking, J.R.M. (1981), 'Lagrange-Multiplier Tests of Multivariate Time Series Models', *Journal of 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.

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. (2000), 'Impulse Response Analysis in Vector Autoregressions with Unknown Lag Order', forthcoming: *Journal of Forecasting*.

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. (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', forthcoming: *Journal of Econometrics*.

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

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.

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.

Table 1: Empirical Studies After Which the DGPs Are Modeled

Empirical Study	Number Of Variables	Type Of Model	Data Frequency	Variables (in Order)
Sims (1986)	6	VAR	Q	Output, investment, price level, M1, unemployment rate, T-Bill rate.
Rotemberg-Woodford (1996)	4	VAR	Q	Growth rate of nominal oil price, real price of oil, output growth, growth rate of real wage.
Christiano-Eichenbaum-Evans (1996)	7	VAR	Q	Output, price level, commodity prices, FedFunds rate, nonborrowed reserves, total reserves, M1
Galí (1999)	2	VAR	Q	Growth rate of labor productivity, growth rate of hours worked.
Strongin (1995)	5	VAR	M	Output, price level, total reserves, nonborrowed reserve ratio, FedFunds rate.
Eichenbaum-Evans (1995)	5	VAR	M	Output, price level, nonborrowed reserves ratio, U.S.-UK. short-term interest rate differential and real exchange rate.
Bernanke-Gertler (1995)	4	VAR	M	Output, price level, commodity prices, FedFunds rate.
Leeper (1997)	6	VAR	M	T-Bill rate, output, price level, T-Bond rate, total reserves, commodity prices.
Johansen-Juselius (1990)	4	VEC	Q	Finnish data for real balances, real income, short-term interest rate, inflation rate.
Kilian (1999)	2	VEC	Q	Percent change in U.S.-U.K. spot exchange rate, deviation of spot rate from monetary fundamental.
Pesaran-Shin-Smith (2000)	5	VEC	Q	U.K. price level, ROW price level, U.K.-ROW exchange rate, U.K. T-Bill rate, short-term ROW interest rate.

NOTES: Q = quarterly data; M = monthly data

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

(a) Monthly VAR Models

T	AIC	HQC	SIC	LR5	LR1	LM5	LM1
240	0.97	0.97	0.99	1.28	1.17	1.20	1.10
300	0.99	1.03	1.07	1.22	1.11	1.15	1.07
360	1.00	1.04	1.15	1.18	1.08	1.11	1.05
480	1.00	1.05	1.28	1.14	1.06	1.08	1.03
600	1.00	1.06	1.38	1.12	1.04	1.06	1.03

(b) Quarterly VAR Models

T	AIC	HQC	SIC	LR5	LR1	LM5	LM1
80	1.76	1.18	0.79	2.42	2.17	1.72	1.36
100	1.28	0.93	0.89	1.98	1.74	1.45	1.20
120	1.10	0.97	0.96	1.84	1.60	1.38	1.17
160	1.06	1.00	1.10	1.68	1.43	1.32	1.15
200	1.06	1.01	1.17	1.59	1.34	1.27	1.14

(c) Quarterly VEC Models

T	AIC	HQC	SIC	LR5	LR1	LM5	LM1
80	3.88	1.04	0.70	5.87	5.38	2.43	1.58
100	1.37	0.90	0.79	3.64	3.26	1.57	1.15
120	1.05	0.94	0.85	2.61	2.26	1.28	1.04
160	1.02	0.98	0.91	1.90	1.58	1.11	0.99
200	1.02	0.99	0.95	1.60	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/LR5	AIC/LR1	AIC/LM5	AIC/LM1
240	1.00	0.98	0.76	0.83	0.81	0.88
300	0.97	0.93	0.81	0.89	0.87	0.93
360	0.96	0.87	0.84	0.92	0.90	0.95
480	0.95	0.78	0.88	0.95	0.93	0.97
600	0.94	0.73	0.89	0.96	0.94	0.97

(b) Quarterly VAR Models

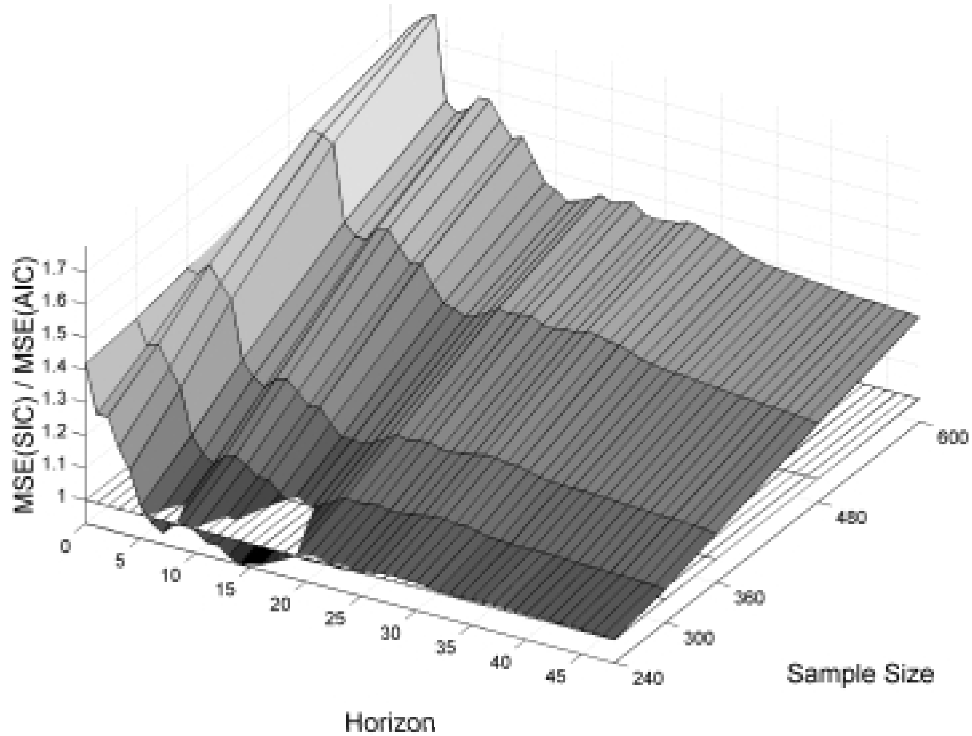
T	HQC/AIC	HQC/SIC	HQC/LR5	HQC/LR1	HQC/LM5	HQC/LM1
80	0.67	1.50	0.49	0.55	0.69	0.87
100	0.72	1.05	0.47	0.53	0.64	0.78
120	0.88	1.00	0.53	0.60	0.70	0.82
160	0.94	0.91	0.59	0.70	0.76	0.87
200	0.96	0.86	0.64	0.76	0.80	0.89

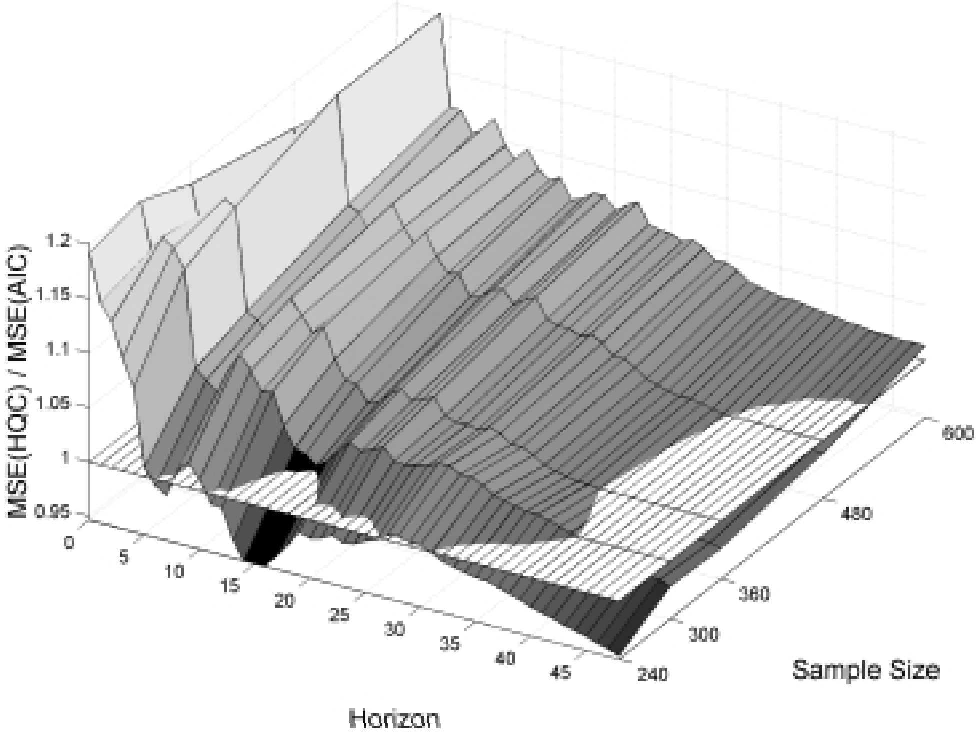
(c) Quarterly VEC Models

T	SIC/AIC	SIC/HQC	SIC/LR5	SIC/LR1	SIC/LM5	SIC/LM1
80	0.18	0.67	0.12	0.13	0.29	0.44
100	0.57	0.87	0.22	0.24	0.50	0.69
120	0.80	0.90	0.32	0.38	0.66	0.81
160	0.89	0.93	0.48	0.57	0.82	0.91
200	0.93	0.96	0.59	0.71	0.90	0.96

NOTES: Averages of ratios are calculated as geometric means.

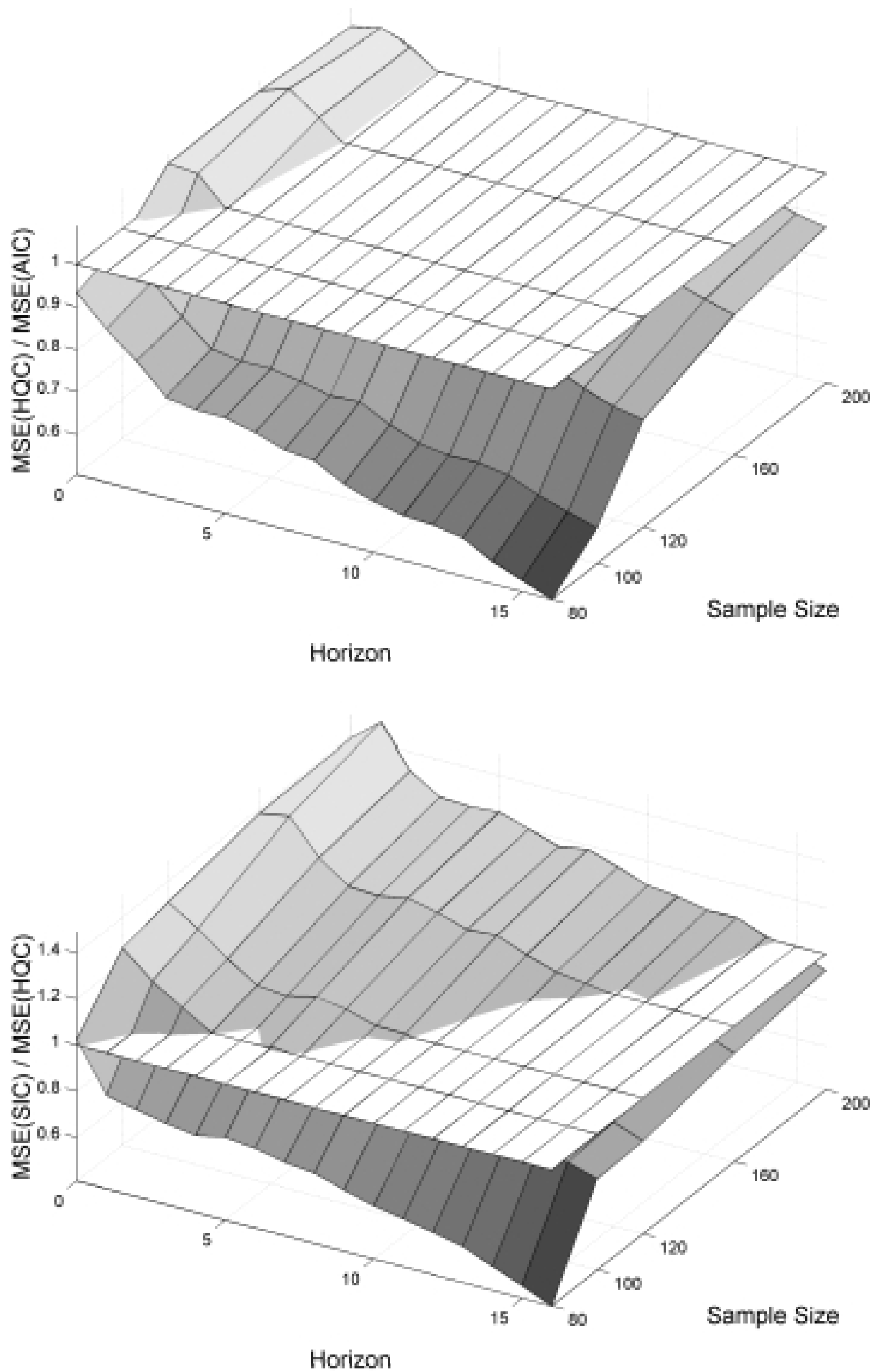
Figure 1
Relative MSEs by Horizon and Sample Size
Monthly VAR Models





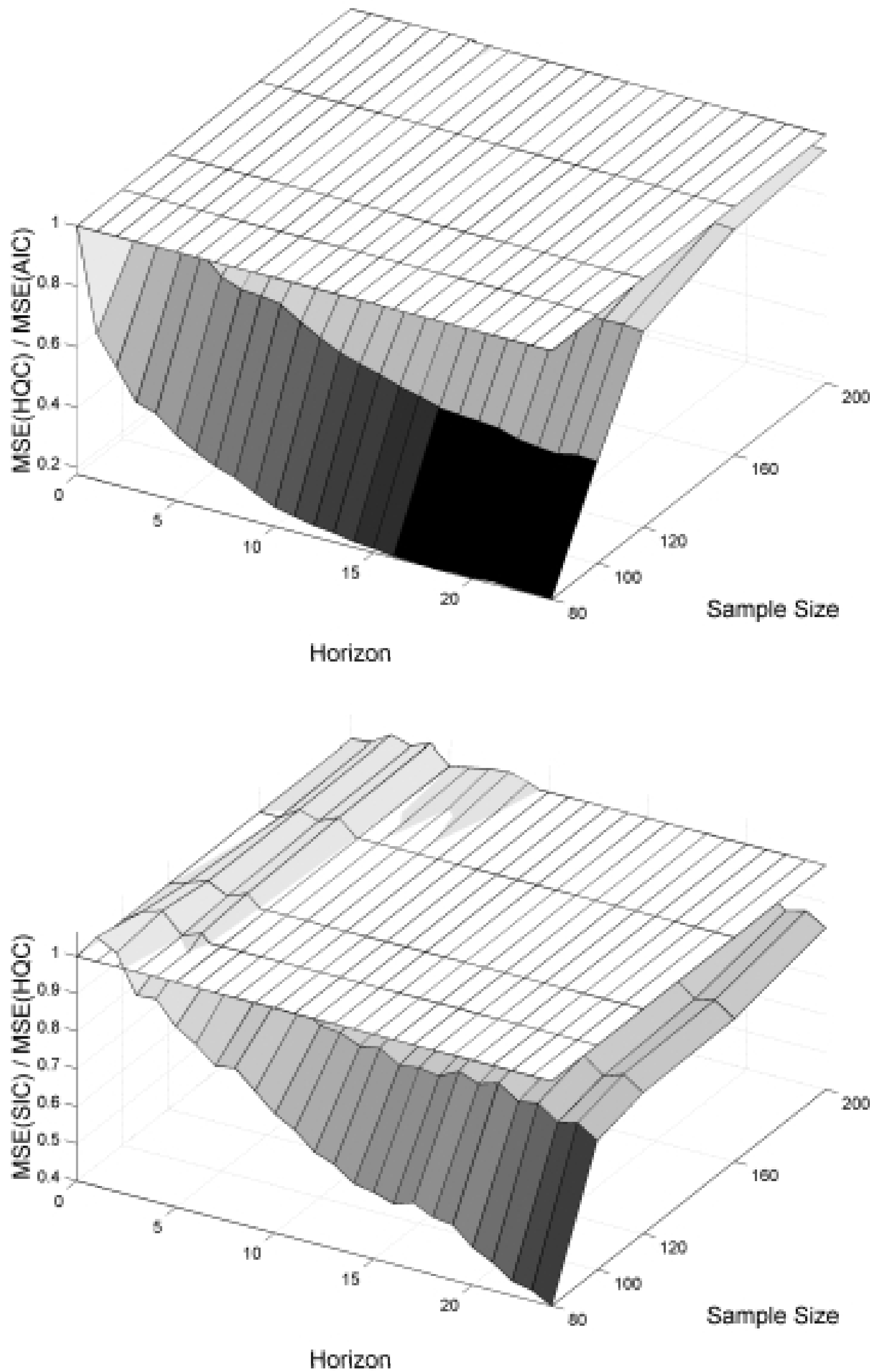
NOTES: Average MSE ratios across all impulse response estimates for all DGPs within model class.

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



NOTES: See Figure 1.

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



NOTES: See Figure 1.