Journal of Forecasting
J. Forecast. **20**, 161–179 (2001)

Impulse Response Analysis in Vector Autoregressions with Unknown Lag Order

LUTZ KILIAN*
University of Michigan, USA, and CEPR, UK

ABSTRACT

We show that the effects of overfitting and underfitting a vector autoregressive (VAR) model are strongly asymmetric for VAR summary statistics involving higher-order dynamics (such as impulse response functions, variance decompositions, or long-run forecasts). Underfit models often underestimate the true dynamics of the population process and may result in spuriously tight confidence intervals. These insights are important for applied work, regardless of how the lag order is determined. In addition, they provide a new perspective on the trade-offs between alternative lag order selection criteria. We provide evidence that, contrary to conventional wisdom, for many statistics of interest to VAR users the point and interval estimates based on the AIC compare favourably to those based on the more parsimonious Schwarz Information Criterion and Hannan–Quinn Criterion. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS VAR; lag order selection; model uncertainty; bootstrap

Much of what we know about macroeconomic dynamics is based on summary statistics calculated from estimates of vector-autoregressive (VAR) models. These dynamics crucially depend on the lag order choice, because the statistics of interest are functions of the order of the autoregressive lag polynomial. In this paper, it is argued that the effects of overfitting and underfitting the VAR model are strongly asymmetric for VAR summary statistics involving higher-order dynamics (such as impulse response functions, variance decompositions, or long-run forecasts). It is shown that underfit models tend to underestimate the true dynamics of the population process and may result in spuriously tight confidence intervals. These insights are important for applied work, regardless of how the lag order is determined from the data. In addition, they provide a new perspective on the trade-offs between lag order selection criteria.

Although it is common in applied work to determine the VAR lag order based on information-based lag order selection criteria (see Nishi, 1988; Granger, King, and White, 1995; Sin and White, 1996), to date very little is known about the implications of alternative lag order selection procedures for estimation and inference. This gap in the literature is surprising, given the great importance attached to the substantive conclusions from vector autoregressions. It is widely

* Correspondence to: Lutz Kilian, Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220, USA. E-mail: lkilian@umich.edu

believed that strongly consistent lag order selection criteria such as the Schwarz Information Criterion (SIC) and the Hannan–Quinn Criterion (HQC) are better suited for the analysis of finite-lag order VAR models than the less parsimonious Akaike Information Criterion (AIC). In contrast, for infinite-order autoregressions, the AIC is regarded as more appropriate. In this paper, the case is made that *even* in finite-order VAR models for many statistics of interest to VAR users the use of the AIC tends to result in more accurate point and interval estimates.

Previous studies of the lag order choice in finite-lag order VAR models tentatively concluded that the SIC performs best in small samples. That conclusion was based on simulation evidence about the distribution of the lag order estimates and about the short-run forecasting performance of the estimated VAR model (e.g. Nickelsburg, 1985; Lütkepohl, 1985, 1991). Evidence for other statistics of interest such as impulse response functions, variance decompositions, measures of predictability, or long-term forecasts apparently has not been presented. This paper takes the view that the latter statistics differ in important ways from short-run forecasts. For example, impulse response functions can be thought of as curves well approximated by higher-order polynomials. Underfitting the lag order amounts to approximating these curves by lower-order polynomials. Consequently, much of the curvature of the impulse response function is effectively erased, resulting in misleading estimates and inference. In contrast, overfitting only results in less precise estimates of the impulse response function. Thus, the effects of overfitting and underfitting the model are strongly asymmetric, especially at long time horizons, and the relative performance of the lag order selection criteria may differ substantially from the results reported for short-run forecasts. The fact that the costs associated with underfitting the model tend to be disproportionately larger suggests that less parsimonious lag order selection criteria such as the AIC may result in more accurate impulse response estimates compared to the highly parsimonious SIC and HQC.

A Monte Carlo study illustrates that, in the presence of higher-order dynamics in impulse response functions, the AIC indeed has better finite-sample properties than more parsimonious lag order selection criteria. The simulation study compares the performance of four well-known information-based lag order selection criteria based on: (1) the small-sample distribution of the lag order estimates; (2) the mean squared errors of the implied impulse response point estimates; and (3) the coverage accuracy and average length of the implied impulse response confidence intervals. The latter part of the paper builds on a recent study of Kilian (1998a) which compared various confidence intervals for VAR impulse responses under the assumption that the lag order is known, and concluded that bias-corrected non-parametric bootstrap intervals tend to be most accurate in small samples. A description of the algorithm is provided in the Appendix.

The remainder of the paper is organized as follows. The next section briefly reviews the lag order selection criteria used in the simulation study. The design of the Monte Carlo study is explained in the third section. The fourth section contains a summary of the simulation results. The fifth section relates the findings to the existing literature on lag order selection, and the sixth section contains an example of how the differences between the Akaike and the Schwarz Information Criterion may affect the substantive interpretation of macroeconomic VAR models. In the final section we summarize the results and outline several extensions.

LAG ORDER SELECTION CRITERIA FOR VECTOR AUTOREGRESSIONS

Consider a covariance stationary N-dimensional VAR process with finite lag order p_0 and iid disturbances u_t with vector mean zero and unknown positive definite covariance matrix Σ_u:

$$y_t = B_1 y_{t-1} + B_2 y_{t-2} + \ldots + B_p y_{t-p_0} + u_t \tag{1}$$

Let $\beta = vec(B_1, B_2, \ldots, B_{p_0})$ and $\sigma = vech(\Sigma_u)$, where *vec* denotes the column stacking operator and *vech* is the column stacking operator that stacks the elements on and below the diagonal only. The statistic of interest is the estimated response of variable k to a one-time impulse in variable l, i periods ago, denoted by $\hat{\theta}_{kl,i}(\hat{\beta}, \hat{\sigma}, \hat{p})$ where $\hat{p}$ is an estimator of p_0 (see Lütkepohl, 1991, for further discussion).

In practice, $\hat{p}$ must be determined from the data. It is common to use information-based lag order selection criteria for this purpose. We consider four such criteria, which differ by the severity of the penalty imposed for parameter profligacy and hence in the parsimony of the model selected: the Schwarz Information Criterion (SIC), the Hannan–Quinn Criterion (HQC), the Akaike Information Criterion (AIC), and the bias-corrected Akaike Information Criterion (AIC_{BC}) of Hurvich and Tsai (1993). The AIC_{BC} is a modification of the AIC designed to bridge the middle ground between the HQC and the AIC by reducing the AIC's tendency to overfit in small samples. For an N-dimensional VAR(p) process without deterministic components:

$$\begin{aligned}
SIC(p) &= \ln|\bar{\Sigma}_u(p)| + \frac{\ln T}{T}(N^2 p) \\
HQC(p) &= \ln|\bar{\Sigma}_u(p)| + \frac{2\ln\ln T}{T}(N^2 p) \\
AIC(p) &= \ln|\bar{\Sigma}_u(p)| + \frac{2}{T}(N^2 p) \\
AIC_{BC}(p) &= T(\ln|\bar{\Sigma}_u(p)| + N) + 2b\{N^2 p + N(N+1)/2\}
\end{aligned} \tag{2}$$

where T is the effective sample size, $\bar{\Sigma}_u$ the maximum likelihood estimate of Σ_u, and the bias-correction factor for the AIC_{BC} is $b = T/\{T-(pN+N+1)\}$. In each case, the lag order $\hat{p}$ is chosen to minimize the value of the criterion over a range of alternative lag orders p given by $\{p: 1 \leqslant p \leqslant \bar{p}\}$. It is assumed that the true lag order p_0 is contained in this set. Our ultimate goal is to use estimates $\hat{p}$ to construct point and interval estimates for $\hat{\theta}_{kl,i}(\hat{\beta}, \hat{\sigma}, \hat{p})$.

Only the SIC and HQC are strongly consistent for p_0 (see Quinn, 1980, p. 182), but all four criteria imply consistent estimates of $\theta_{kl,i}(\beta, \sigma, p_0)$. For further discussion see Shibata (1976, 1980), Hannan and Quinn (1979), Quinn (1980), Paulsen and Tjøstheim (1985), Shibata (1986), Quinn (1988), Pötscher (1991, 1995), and Kabaila (1995). Although the AIC will tend to overestimate p_0, the asymptotic probability that the AIC selects the true lag order is 0.88–0.89 for bivariate processes, about 0.96 for trivariate processes, 0.99 for dimension 4, and 0.998 for dimension 5 (see Paulsen and Tjøstheim, 1985). This means that the asymptotic probability of overestimating the lag order can be safely neglected in most multivariate applications.

An important drawback of strongly consistent lag order selection criteria is that they have a tendency to underestimate the true lag order in small samples. As a result, the implied parameter vector need not converge uniformly to the true parameter vector (see Kabaila, 1995; Pötscher, 1995). We conjecture that, in practice, this tendency to underfit may result in severe misspecification bias for VAR statistics like impulse responses, especially at higher horizons. The practical importance of this problem, however, is not obvious because biased estimates tend to have lower variance than unbiased estimates. In many instances a researcher may prefer a biased estimate, provided the MSE is reduced. It thus seems natural to compare the performance of

alternative lag-order selection criteria in terms of the MSE of the parameter estimator $\hat{\theta}_{kl,i}(\hat{\beta}, \hat{\sigma}, \hat{p})$. The Monte Carlo study below will examine the bias–variance tradeoff in greater detail.

SIMULATION DESIGN

The population process underlying the Monte Carlo study is designed to produce impulse response functions with shapes characteristic of impulse response functions encountered in applied work, notably the existence of higher-order dynamics in the impulse response functions. The data-generating process is a bivariate VAR(4) with coefficient matrices:

$$B_1 = \begin{bmatrix} 0.6362 & -0.0012 \\ 0.0190 & 0.5782 \end{bmatrix} \quad B_2 = \begin{bmatrix} -0.0168 & -0.0285 \\ 0.5211 & -0.3041 \end{bmatrix}$$
$$B_3 = \begin{bmatrix} 0.0273 & -0.0028 \\ 0.1568 & 0.2229 \end{bmatrix} \quad B_4 = \begin{bmatrix} 0.1517 & -0.0198 \\ -0.7600 & -0.3168 \end{bmatrix} \tag{3}$$

with a dominant root of 0.8894. The iid innovations are normally distributed with vector mean zero and variance–covariance matrix Σ_u:

$$\Sigma_u = \begin{bmatrix} 0.025 & 0.009 \\ 0.009 & 0.387 \end{bmatrix} \times 10^{-3} \tag{4}$$

This particular process was chosen for two reasons: (1) its impulse response functions are similar in shape to responses that might be encountered in larger systems such as the empirical example given later; (2) the process is persistent, but its dominant root is small enough to allow us to abstract from any complications that may arise in models with roots close to unity. In practice, interest often centres on VAR models estimated subject to cointegration constraints. We do not address the subject of cointegration in this paper because it is not central to our point. The purpose of the Monte Carlo study is to illustrate the potential quantitative importance of the bias–variance tradeoff in as simple a setting as possible. A comprehensive study of alternative models and estimators would more appropriately be the subject of a separate study.

Based on draws from this data-generating process, we compare the MSEs of the orthogonalized impulse response point estimates $\hat{\theta}_{kl,i}(\hat{\beta}, \hat{\sigma}, \hat{p})$ for each of the four lag-order selection criteria discussed earlier. We also evaluate the effective coverage accuracy and average length of the corresponding nominal 95% bootstrap confidence intervals. Effective coverage is defined as the relative frequency with which the confidence interval covers the true, but in practice unknown, impulse response value in repeated trials.

The intervals are obtained by conditioning on the estimated lag order, as though it were the true lag order. As a standard of comparison, the counterfactual interval that would be obtained, if the true lag order were known, is also included. By construction the true lag order of the process is $p_0 = 4$. In the Monte Carlo study, the lag order is estimated under the maintained assumption that $1 \leqslant p \leqslant 8$. This assumption is likely to favour parsimonious criteria like the SIC by preventing them from selecting $\hat{p} = 0$. The number of Monte Carlo trials is 400, implying a Monte Carlo standard error of 0.01 for the 95% interval, and the number of bootstrap replications

is 2000. The sample sizes are 80 and 160, which may be thought of as twenty and forty years of quarterly data.

SIMULATION RESULTS

Lag order estimates

It is well known that consistent lag order selection criteria are more likely to underestimate the true lag order than to overestimate it in finite samples. In contrast, inconsistent criteria such as the AIC tend to be more balanced about p_0 in small samples, with a tendency to overestimate p_0 slightly as the sample size grows. We begin by examining the accuracy of these claims for the data-generating process in equations (3) and (4).

Table I summarizes the finite-sample distribution of $\hat{p}$ for each lag order selection criterion. For sample size 80, the SIC underestimates the true lag order in 98% of the 400 trials. In fact, with probability 0.92 it picks a lag order of one. This strong downward bias of the SIC lag order estimator confirms that the SIC has much higher risk of underestimating the lag order than of overestimating it. Table I also shows that the SIC has almost zero probability of overestimating p_0. In contract, the AIC lag order distribution is roughly centred on the true value. It underestimates the lag order with probability 0.26 and overestimates it with probability 0.18. As the sample size increases to 160, the performance of both criteria improves, but the basic pattern is preserved. The SIC still picks a lag order of one with probability 0.61, whereas the AIC selects the true lag order with probability 0.83. It is noteworthy that the AIC underestimates the lag order in only three cases, and that the probability of overestimation further declines with the sample size, consistent with the asymptotic results of Paulsen and Tjøstheim (1985). The fact that in small samples the AIC lag order distribution tends to be more balanced about the true lag order than the SIC lag order estimates is consistent with findings by Nickelsburg (1985) and Lütkepohl (1985, 1991) for other data-generating processes (see also Shibata, 1983).

The evidence in Table I supports the view that the SIC may be extremely unbalanced even for

Table I. Percentage distribution of lag order estimates by criterion

	Lag orders							
Criterion	1	2	3	4	5	6	7	8
Sample size 80								
SIC	92.0	5.8	0	2.0	0.3	0	0	0
HQC	54.3	11.3	1.3	30.3	2.8	0.3	0	0
AIC_{BC}	32.8	15.3	2.0	45.0	4.3	0.8	0	0
AIC	14.5	9.3	1.8	56.5	11.3	2.8	2.3	1.8
Sample size 160								
SIC	60.8	11.5	0.3	27.5	0	0	0	0
HQC	7.8	4.3	0.3	84.5	3.3	0	0	0
AIC_{BC}	0.8	0.8	0.3	88.8	7.8	1.3	0.5	0
AIC	0.5	0	0.3	83.3	10.8	2.5	2.8	0

Source
Results based on 400 Monte Carlo trials and the data-generating process described in the fourth section. The lag orders are constrained to lie between 1 and 8. The true lag order is $p_0 = 4$.

sample size 160. It thus seems sensible to explore less parsimonious alternatives. Since the HQC is the least parsimonious, yet still consistent, model selection criterion, one would expect the HQC to be more balanced and less downward biased than the SIC. The Monte Carlo simulations confirm that view. Table I shows that for sample size 80 (160) the HQC underestimates p_0 with probability 0.67 (0.12) and overestimates p_0 with probability 0.03 (0.03). However, the performance of the HQC interval still falls far short of the AIC interval, especially for sample size 80. This suggests to consider alternative criteria such as the AIC_{BC} that bridge the middle ground between the HQC and the AIC. This criterion was introduced by Hurvich and Tsai (1993, p. 271) to reduce the tendency of the AIC to overfit in VAR models, especially in small samples. Table I shows that the AIC_{BC} indeed reduces the tendency of the AIC to overfit in small samples, but at the expense of underfitting more often. Given the earlier discussion, one would therefore conjecture that AIC_{BC} impulse response estimates ought to be more accurate at longer horizons than SIC or HQC estimates, but less accurate than AIC estimates.

Impulse response point estimates

We now turn to the question of how the accuracy of the impulse response point estimates is affected by the choice of lag order selection criterion. The accuracy of the point estimates $\hat{\theta}_{kl,i}(\hat{\beta}, \hat{\sigma}, \hat{p})$ is evaluated in terms of their mean-squared deviation from the population value of the impulse response at each time horizon. Figures 1 and 2 plot the results for sample size 80 and 160. To put the magnitude of the MSEs into perspective, the plots also include results under the counterfactual assumption that the true lag order is known. To assist in the interpretation of the results, the left column in each figure shows the underlying population impulse response functions to be estimated. The shapes of the true impulse response functions involve various degrees and patterns of higher-order dynamics. They represent four common patterns found in applied work. Figures 1 and 2 show that, regardless of the shape of the impulse response function, the AIC-based estimate overall has the lowest mean squared error, followed by the AIC_{BC}, HQC, and SIC, in that order. As the sample size increases, the accuracy of all estimates improves, but the relative accuracy of the SIC-based estimates deteriorates further. The ranking of the four criteria is exactly as conjectured, and the differences in accuracy can be substantial.

The magnitude of the MSE appears closely linked to the curvature of the population impulse response function. Note the tendency of the MSE of the SIC estimate (and to a lesser extent of the other estimates) to oscillate as the time horizon grows. This type of pattern is what one would expect from fitting a smoothly decaying low-order polynomial to the population impulse response functions in the left column. Such oversmoothing of the impulse response function would also account for the fact that in some cases the MSE of the SIC estimate temporarily drops below that of the estimates based on the AIC or the true lag order. Further indirect evidence that higher-order dynamics are driving the result is the fact that in many cases there is little difference between alternative lag order selection criteria at low time horizons. In two cases, for sample size 80, the SIC estimate has substantially lower MSE in the very short run. This evidence is consistent with earlier findings by Lütkepohl (1985, 1991) for short-run prediction MSEs in somewhat simpler VAR models. It is also consistent with the argument presented above that the adverse effects of oversmoothing will only become apparent at higher time horizons. However, the advantages of the SIC in the very short run disappear for moderate samples and in all cases are dwarfed by increases in the MSE at higher time horizons.

Figures 1 and 2 confirm Shibata's (1983) point that consistency for p_0 is not necessarily a desirable property if we are interested in a good estimator or predictor. Lower-order models

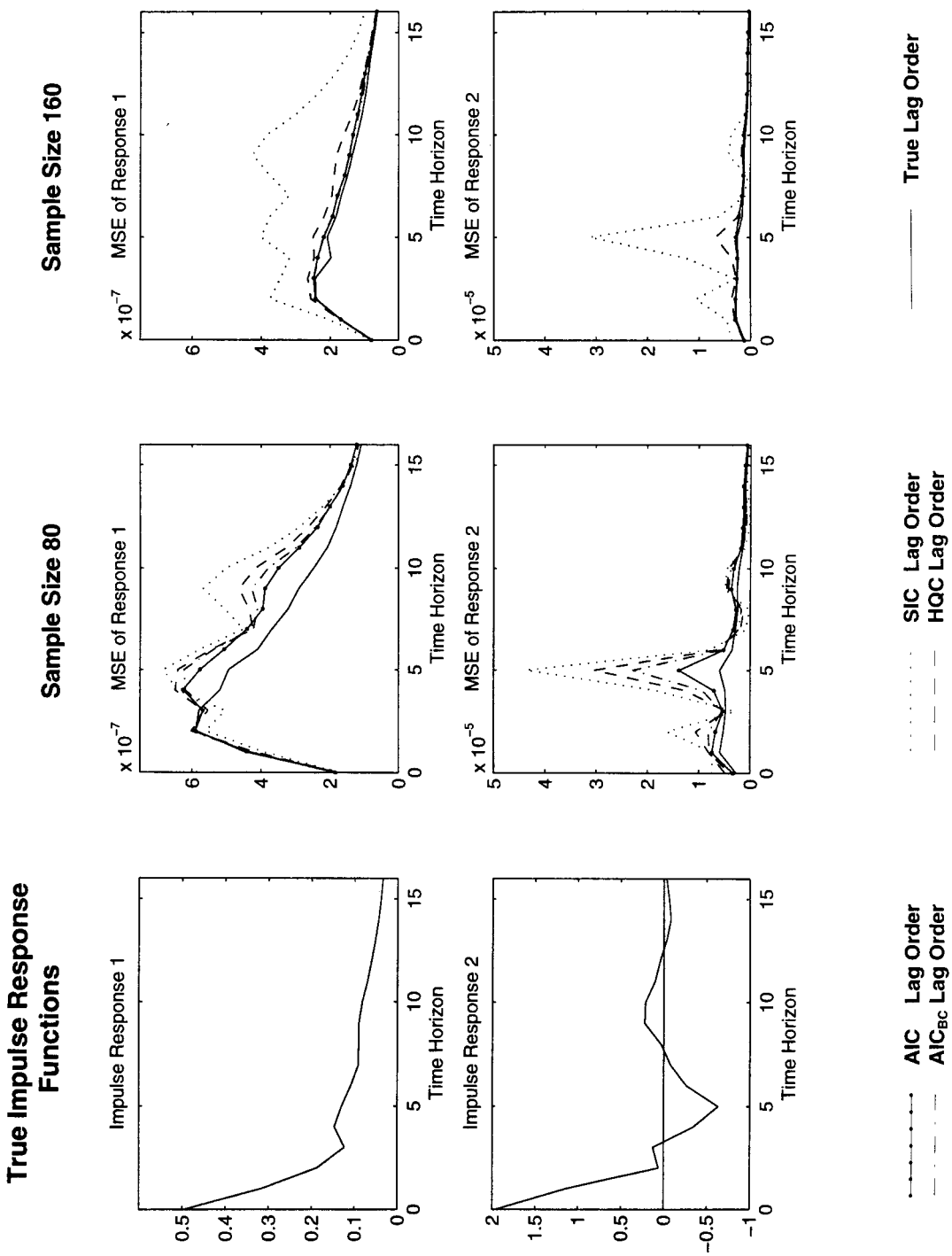


Figure 1. Mean-squared errors of impulse response estimates by lag order selection criterion

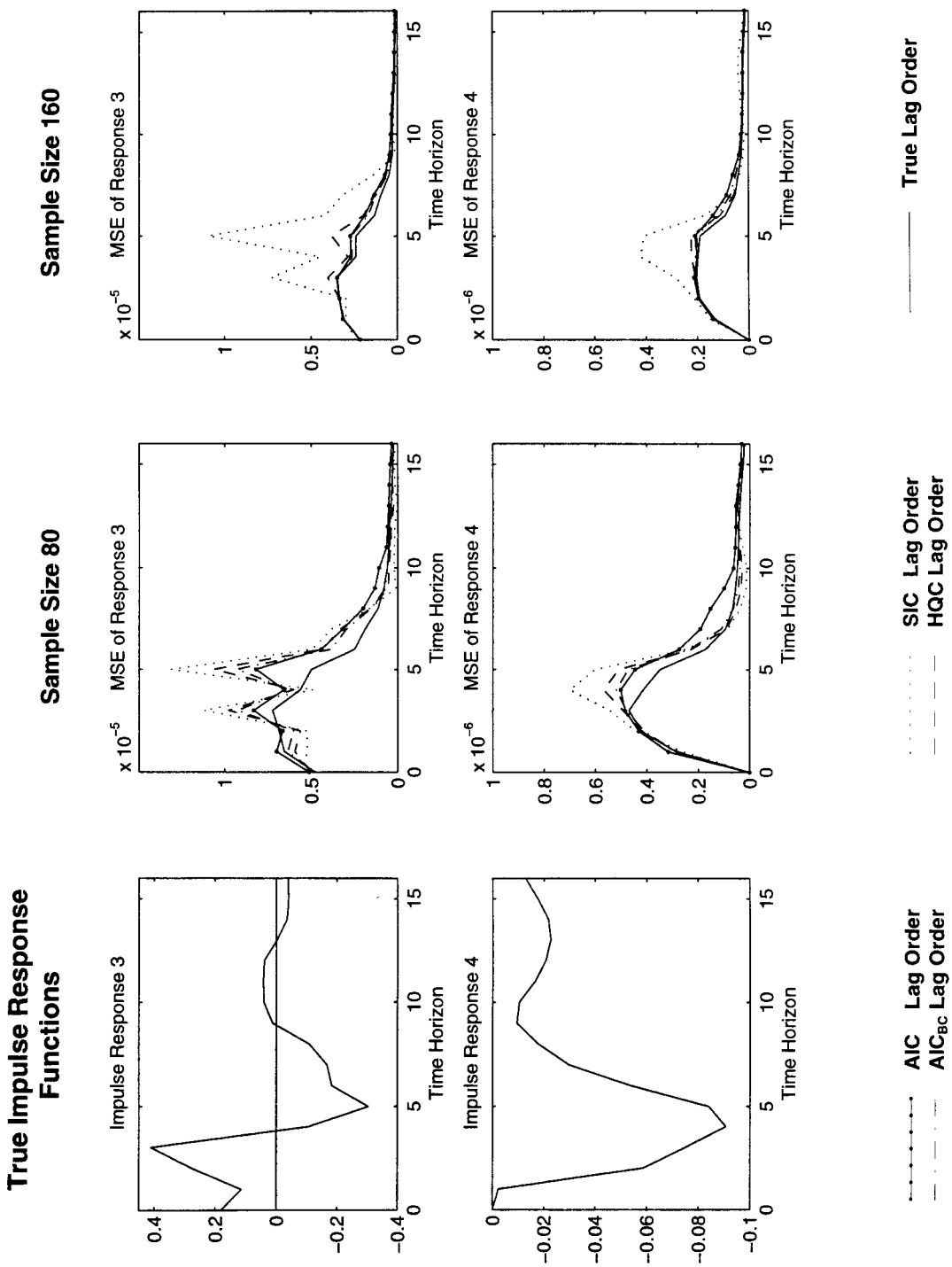


Figure 2. Mean-squared errors of impulse response estimates by lag order selection criterion

based on parsimonious, consistent model-selection criteria tend to do relatively well at very short time horizons, but often poorly at higher time horizons. In contrast, inconsistent, but more balanced model selection criteria tend to do well at all time horizons, even after accounting for the increase in variance that accompanies a reduction in misspecification bias.

Impulse response confidence intervals

It has become standard in the macroeconometric literature to interpret VAR impulse response estimates after accounting for sampling uncertainty. This subsection therefore examines the implications of the choice of model-selection criteria for confidence intervals, rather than point estimates. To conserve space, results for only two impulse response functions are presented. The results for the other impulse response functions are qualitatively similar.

Figures 3 and 4 plot the coverage rates and average length of the implied confidence intervals for impulse response functions 1 and 2. The left column in Figure 3 plots the true impulse response functions. The other two columns plot the effective coverage rates of the SIC, HQC, AIC_{BC}, and AIC intervals. As a reference point, the plots also include the coverage of the same interval under the counterfactual assumption that the true lag order is known. Each subplot shows the effective coverage rates of the nominal 95% intervals for a time horizon of 16 periods after the initial shock. A horizontal line at 0.95 would imply perfect coverage accuracy at all time horizons and has been imposed on the plots as a reference line. The true lag order interval fluctuates around this ideal value for impulse response 2, and comes very close for impulse response 1. This result is consistent with evidence in Kilian (1998a) that the bias-corrected bootstrap algorithm performs quite well, if the true lag order is known. However, in applied work the lag order is rarely known. A more realistic exercise recognizes that the lag order must be estimated from the data before fitting a VAR. The middle column in Figure 3 shows that for sample size 80 the effects of lag order uncertainty can be striking. Consider first the SIC and AIC intervals. For the exponentially decaying impulse response function in the upper panel the effective coverage drops by up to 0.47 if the lag order p_0 is estimated by the SIC and by up to 0.13 if p_0 is estimated by the AIC. For the cyclical impulse response function in the lower panel the effects are even more dramatic. For the SIC, the coverage of the nominal 95% interval becomes extremely erratic and may drop as low as 3%; for the AIC, coverage also becomes very unstable and for some time horizons falls to 71%. The peaks and troughs of the coverage plot reflect the oscillation in the underlying true impulse response function. The relative performance of the SIC and AIC is exactly as conjectured, and consistent with the evidence for the point estimates.

The right column of Figure 3 shows that if the sample size is raised to 160, the effective coverage of the AIC intervals is quite close to nominal coverage for the cyclical as well as the smoothly decaying impulse response function. In sharp contrast, the coverage of the SIC interval slightly deteriorates in the upper panel, and only somewhat improves in the lower panel. While its coverage rates become more stable for the cyclical impulse response function, the coverage of the SIC interval may still be as low as 27%, even for sample size 160. Estimating the lag order by the SIC rather than the AIC can reduce coverage by as much as 69 percentage points, despite the fairly large sample size. The intuitive explanation is that the SIC lag order estimate apparently converges more slowly than the impulse response estimates. The lower coverage of the SIC interval with higher sample size arises because the confidence interval converges conditional on $\hat{p} < p_0$, but not necessarily to the true value of the impulse response coefficient. This finding is

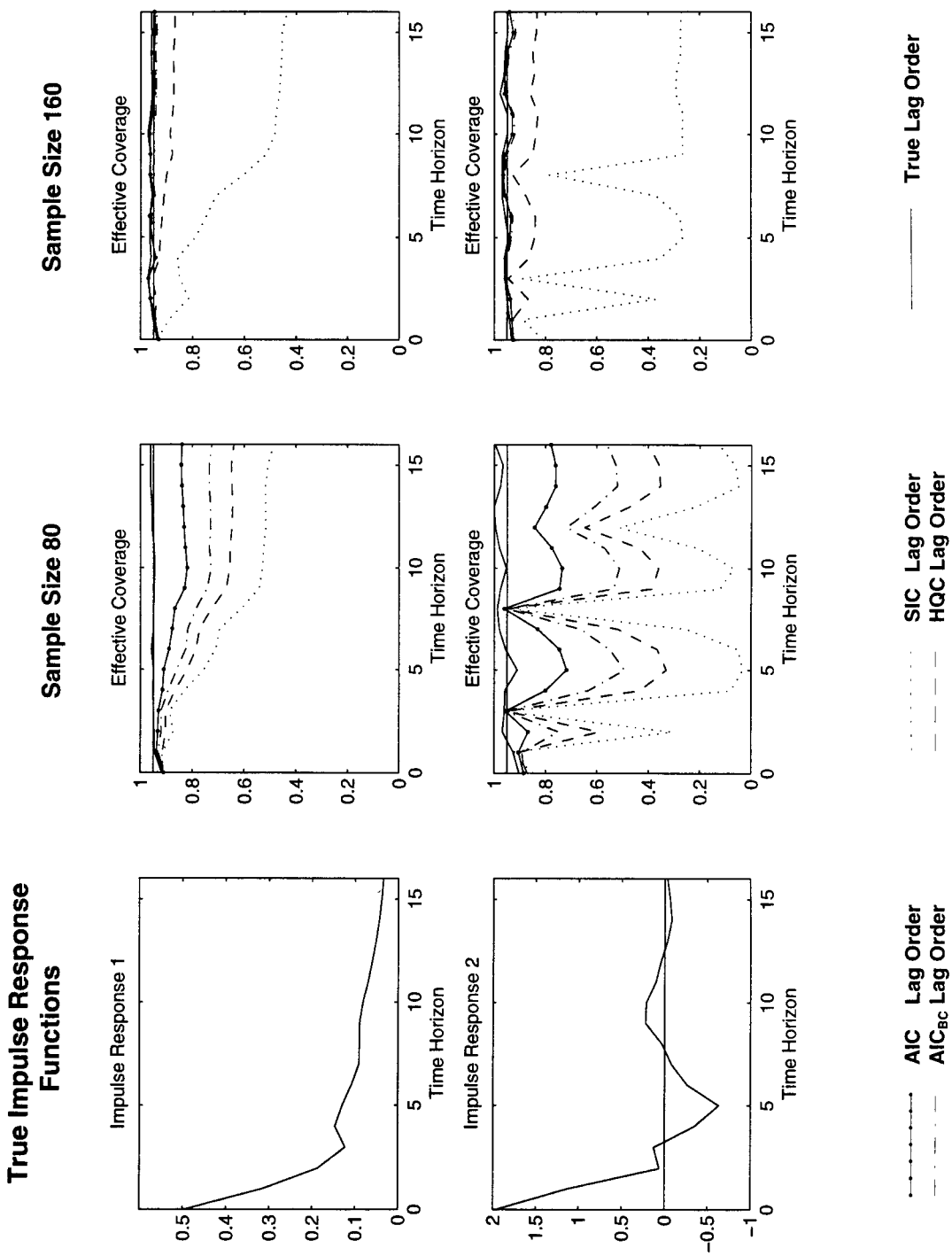


Figure 3. Effective coverage rates of 95% intervals by lag order selection criterion

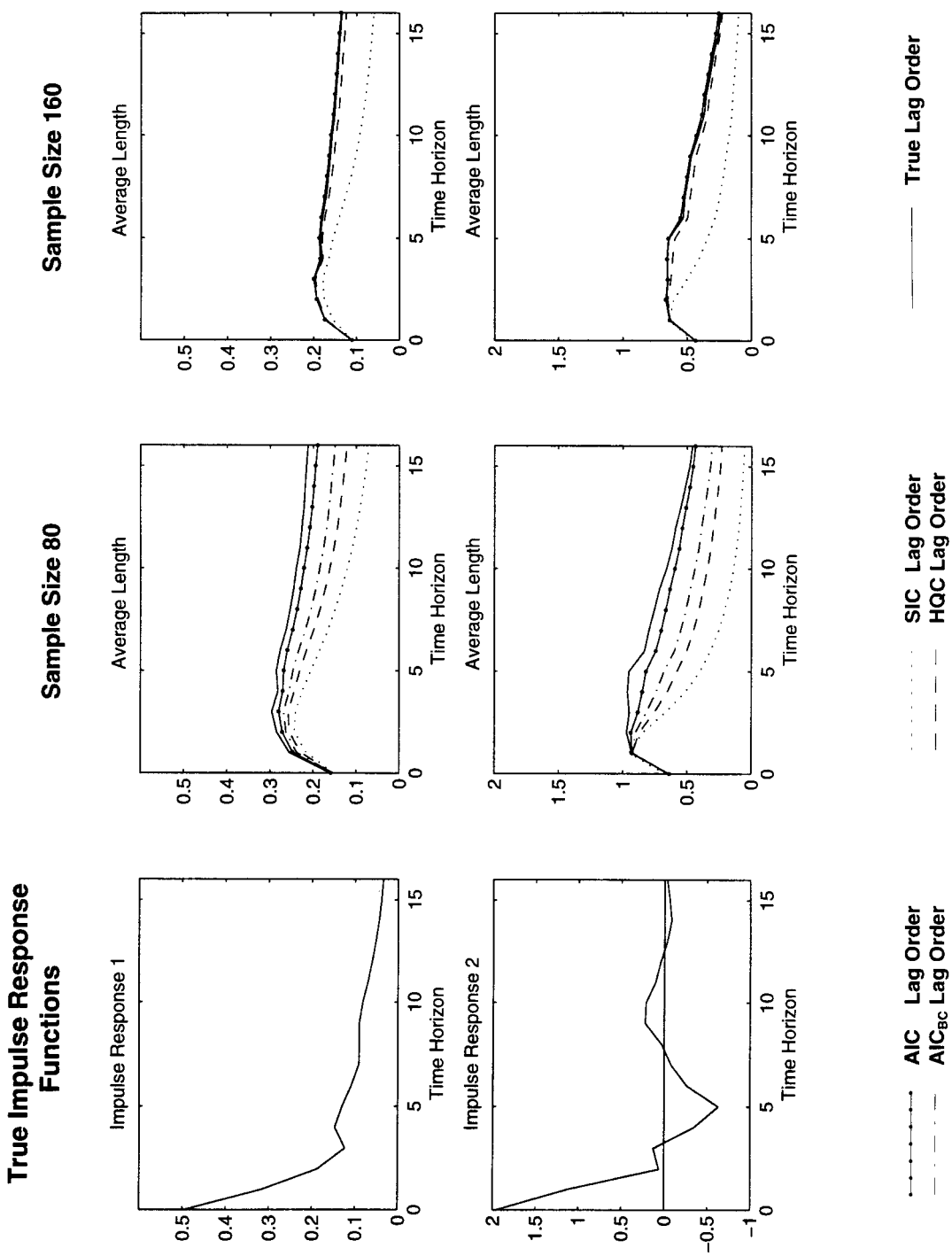


Figure 4. Average lengths of 95% intervals by lag order selection criterion

consistent with the point made by Kabaila (1995) and Pötscher (1995) that the coverage of the SIC-based interval may become arbitrarily small in finite samples.

It is also instructive to compare the average length of the intervals. For sample size 80, the AIC interval in Figure 4 is slightly shorter than the true lag order interval in the upper panel, and somewhat shorter in the lower panel. This evidence is roughly consistent with the coverage results in Figure 3. However, the SIC interval is much shorter than the true lag order interval. Considering its coverage deficiencies, this tendency is evidence that underfitting the lag order produces intervals that grossly understate the true extent of sampling uncertainty. For sample size 160, the basic pattern of the results is preserved. The SIC interval is still too short. In contrast, the AIC interval has about the same length as the true lag order interval in both panels. This pattern is consistent with the results in Table I. As expected, AIC intervals are only slightly less efficient than the counterfactual intervals based on p_0. Overestimation of p_0 evidently is not a serious problem.

In addition, similar Monte Carlo experiments were conducted for the HQC and the AIC_{BC}. Since the HQC lag order distribution in Table I is more balanced than the distribution of the SIC lag order estimate, one would expect the HQC interval to perform better than the SIC interval. Monte Carlo simulation confirms this conjecture. For sample size 80, the coverage rates of the HQC interval exceed the corresponding SIC rates by up to 0.15 for the smoothly decaying impulse response function, and by up to 0.35 for the oscillating impulse response function. For sample size 160, the HQC interval converges rapidly, and its coverage is at most 0.12 short of nominal coverage. The improvement in coverage is consistent with the evidence in Table I that the HQC underestimates the lag order much less frequently than the SIC. However, as expected, the performance of the HQC interval still falls far short of the AIC interval. As a final check consider the AIC_{BC} interval. Based on Table I, one would expect the AIC_{BC} interval to be less accurate in small samples than the AIC interval, but more accurate than the SIC and HQC intervals. Monte Carlo simulation again confirms that conjecture. The AIC_{BC} interval performs about as well as the AIC interval for sample size 160, but has lower coverage for sample size 80 by as much as 0.24. At the same time, it clearly dominates the SIC and HQC intervals. Interestingly, the fact that the AIC_{BC} succeeds in reducing the tendency of the AIC to overfit the model, as intended by its creators, rather than being a virtue, becomes a liability in the present context, as the AIC_{BC} is more likely to underfit the model for sample size 80 and to miss the higher-order dynamics of the impulse response function.

The relative performance of the four lag order selection criteria matches the results for the point estimates and is consistent with the view that the effect of misspecifying the lag order is strongly asymmetric, depending on whether the model is underfit or overfit. Since underfit models often imply distorted impulse responses, the coverage accuracy of the intervals is directly correlated with the probability of underestimating the lag order. In particular, for sample size 80, the AIC interval (with $pr(\hat{p} < p_0) = 0.26$) is most accurate, followed by the AIC_{BC} interval (0.50), the HQC interval (0.67), and the SIC interval (0.98). For sample size 160, the AIC interval (0.01) ranks first, followed by the AIC_{BC} interval (0.02), the HQC interval (0.12), and the SIC interval (0.73).

DISCUSSION

Previous Monte Carlo studies usually focused on the question of which lag order selection criterion is likely to select the true lag order most often (e.g. Nickelsburg, 1985). Results for the

lag order distribution may be of theoretical interest, but they are of limited interest for applied users interested in VAR statistics such as forecasts, impulse responses or variance decompositions. This paper argues that for applied work, a more useful criterion for comparing alternative lag order selection criteria is the MSE of the statistics of interest. Note that the relative frequency distribution of lag orders for a given process does not allow us to predict *a priori* which criterion will imply the smallest MSE for forecasts, impulse responses, and other statistics of interest and how these MSEs will vary with the forecast horizon. For example, it is quite possible that underestimation improves the MSE relative to the model based on the true lag order (see Lütkepohl, 1985). It is also possible that adding extra lags has little effect on the MSE of non-linear functions of VAR slope parameters such as impulse responses. Unlike Nickelsburg (1985), we therefore focus directly on the MSEs of the statistic of interest.

We illustrate our point in the context of impulse response analysis. The work most closely related to ours is Lütkepohl (1985, 1991) who, based on the MSEs of short-run forecasts, tentatively concluded that the SIC performs best in small samples. This paper shows that Lütkepohl's results for short-run forecasts do not necessarily extend to statistics involving higher-order dynamics such as impulse response functions. In fact, overly parsimonious models may completely miss the higher-order dynamics of the impulse response function and yield severely misleading confidence intervals. In practice, the reduction in variance from fitting more parsimonious models is outweighed by the increase in misspecification bias. As a result, applied researchers interested in policy analysis or multi-step-ahead forecasts based on small samples are better served by less parsimonious criteria, even if those criteria are not consistent estimators of the lag order. A similar point has been made by Härdle and Bowman (1988) in the context of non-parametric regression. They find that the bandwidth for non-linear and oscillating curves must be adapted to the local curvature to reduce the bias in curve estimation. In their regression model, excessive smoothing eliminates the higher-order dynamics of the underlying curve. In the VAR model, the lag order plays the role of the bandwidth parameter, and underestimating the lag order similarly eliminates the oscillation in the estimates of the underlying impulse response curve.

It seems worth emphasizing that the simulation evidence in this paper only applies to impulse response estimates (and presumably related quantities like variance decompositions or multi-step-ahead predictions). It does not imply that the AIC is a superior lag order selection criterion for all purposes. In particular, the simulation evidence in Lütkepohl (1985, 1991) for some simple *ad hoc* data-generating processes suggests that for short-run forecasts the SIC or the HQC may be preferable. The evidence in this paper for a data-generating process with much richer and perhaps more realistic dynamics (judged by the shape of the impulse response functions) is broadly consistent with Lütkepohl's results, but it also shows that the potential advantages of the SIC at very short time horizons tend to be dwarfed by severe distortions at time horizons beyond five periods.

Furthermore, the results in Table I suggest that the evidence in Lütkepohl (1985, 1991) of the greater accuracy of the SIC in selecting the true lag order may have been overstated. Note that the data-generating processes on which this conclusion has been based are mostly VAR(1) models (or VAR(2) models with many elements of the coefficient matrices set to zero or close to zero). This assumption greatly favours parsimonious lag order selection criteria like the SIC with a built-in bias in small samples toward selecting a very low lag order, regardless of the true lag order. This fact largely explains why the lag order estimates in Table I are so much more favourable to the AIC than previous results. To the extent that higher-order data-generating

processes like the one used in this paper are more realistic than previous models, the findings of this paper seem to be of greater relevance for applied work. The relatively good performance of the AIC in our example is no guarantee of success, however. There is a tendency for all criteria, including the AIC, to underfit the VAR model in *small samples*, especially for higher-order data-generating processes.

The emphasis in this paper on longer time horizons also raises the question of the appropriateness of model-selection criteria designed to minimize one-step-ahead prediction error variances. While criteria like the AIC seem to work reasonably well in practice, their motivation for impulse response analysis seems awkward and alternative designs for model-selection criteria may yield further improvements. For example, one could imagine re-estimating the model with different lag orders depending on the time horizon of interest or specifying an explicit loss function for a given time horizon.

It is interesting to contrast the use of information-based lag order selection criteria with the use of Bayesian priors for the lag structure. It is common in Bayesian VAR analysis to specify some rate of decay for the lag order weights with a finite upper bound. Leaving aside the question of where lag order priors come from in practice, this approach is clearly more flexible than the rigid priors implicit in information-based criteria such as the SIC. For example, the use of slowly decaying lag weights in Bayesian analysis allows consideration of higher lag orders and avoids sharp cut-off points. Thus, from a Bayesian point of view, lag order estimates based on information-based lag order selection criteria may simply reflect unreasonable priors about the lag order weights. They impose either too many or not enough restrictions on the lag structure. This does not mean that the results of this paper are of no use for Bayesian analysis. The basic tradoffs this paper has documented continue to apply in any VAR analysis, whether the lag order is selected *ad hoc*, based on model-selection criteria, or based on formal priors. For example, by choosing a rate of decay that is too fast, a Bayesian lag order prior may oversmooth the impulse response functions in much the same way that truncating the lag order in classical analysis would. The difference is only a matter of degree. The same applies to the selection of the upper bound for the lag structure. If that prior cut-off point is chosen too low, the MSE of the impulse response estimates will be adversely affected much as in the case of the SIC or HQC.

Finally, it appears that some applied VAR users simply rely on conventional lag order choices (say, 4 or 8 lags for quarterly data, 6 or 12 lags for monthly data) rather than explicitly estimating the lag order from the data. The results of this paper suggest that in that case it appears to be safer to err on the side of including extra lags rather than to truncate the lag order polynomial too early. We noted that even the AIC is likely to underestimate the lag order in small samples. If the AIC lag order estimate appears counterintuitively low, as is often the case in applied work, a researcher may be justified in considering even higher lag orders than suggested by the AIC.

APPLICATION: THE EFFECTS OF MONETARY POLICY

The Monte Carlo evidence presented above suggests that the choice of lag order selection criterion can have important effects on statistical accuracy. To establish that these statistical differences may be important enough to affect the economic interpretation of the estimated VAR model in real-life applications requires additional evidence. This section presents one such real-life application based on Eichenbaum (1992). Eichenbaum studies the effects of monetary policy on the US economy based on impulse responses for a variety of VAR models, including a four-variable

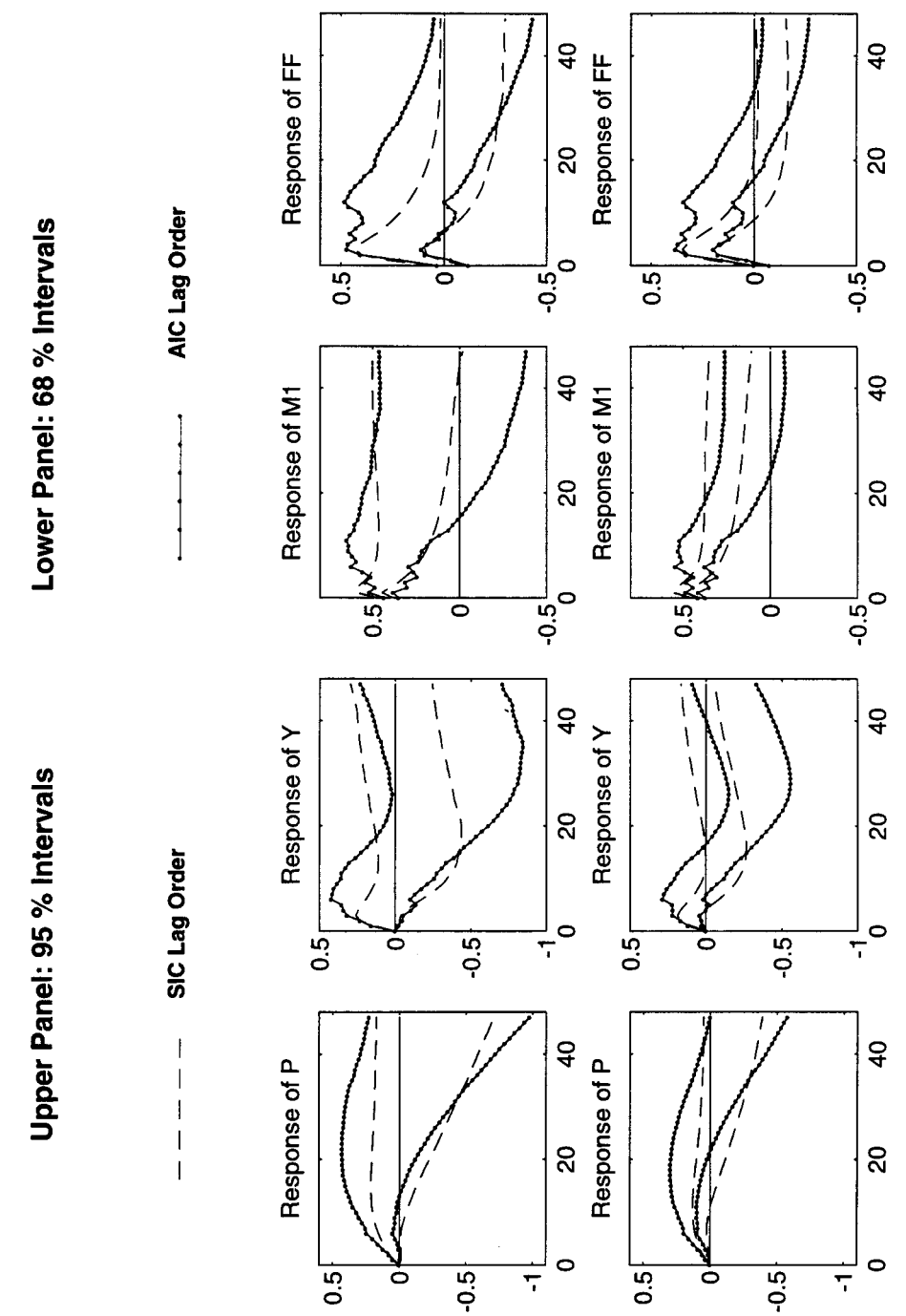
VAR with intercept based on monthly data for 1965:1–1990:1. The ordering of the variables is industrial production, consumer prices, M1, and the Federal Funds rate. All data but the interest rate are in logs. For both lag order selection criteria the upper and lower bounds on the lag order were specified as $1 \leqslant p \leqslant 12$. The SIC selects a lag order of two, whereas the AIC selects a lag order of eight.

Figure 5 plots the 95% confidence intervals (upper panel) and 68% confidence intervals (lower panel) for selected responses to an unanticipated rise in M1. To avoid notational clutter, only the confidence intervals based on the AIC and SIC are shown. As the sign of the impulse response estimate is primarily determined by the first-order autoregressive coefficients, one would expect the SIC and AIC intervals to typically move in the same direction. However, the SIC can be expected to ignore the higher-order dynamics of the impulse response. As a result, SIC intervals will be smoother and typically tighter than AIC intervals. These differences are important, because often analysts are interested in the magnitude of impulse responses in addition to their direction. For example, in the upper panel the response of M1 to a monetary expansion is highly persistent according to the SIC, but one cannot reject that it decays quickly according to the AIC. Similarly, the response of the price level is significantly positive under the AIC, but not significantly different from zero under the SIC. In general, AIC intervals are much wider, especially at higher time horizons, indicating substantially higher sampling uncertainty about the effects of monetary policy. In fact, the interval width somewhat obscures the differences between the SIC and AIC in the upper panel. The lower panel of Figure 5 therefore plots the corresponding 68% intervals for the same responses. Note that the response of M1 now is strongly significant and highly persistent according to the SIC, but still insignificant for time horizons of two years or higher according to the AIC. In contrast to the earlier findings, the response of output to a monetary expansion is significantly negative based on the AIC, but not based on the SIC. This finding is important, because the negative response of output was one of the key reasons which led Eichenbaum to reject this particular model. Other significant differences between the SIC and AIC arise for the responses of the price level and the Fed Funds rate. For example, the response of the price level to a monetary expansion remains significant for almost twice as long according to the AIC compared to the SIC. These results show that despite the increase in interval width, using the AIC rather than more parsimonious criteria does not render VAR impulse response analysis futile. Moreover, the AIC may convey a quite different picture of which responses are significant and at what time horizons.

The evidence in Figure 5 does not tell us which method is more reliable in this particular model. That question could only be answered by a Monte Carlo study (the computational cost of which is likely to be prohibitive). However, Figure 5 illustrates the great practical importance of finding reliable lag order selection criteria for inference based on VAR model estimates.

CONCLUSION

The paper provides evidence of the trade-offs between four commonly used information-based lag order selection criteria: the SIC, HQC, AIC_{BC}, and AIC. It is widely believed that strongly consistent lag order selection criteria such as the SIC and the Hannan–Quinn Criterion are better suited for VAR analysis than less parsimonious criteria such as the AIC, if the true lag order is finite. In contrast, the AIC is regarded as more reliable for infinite-order autoregressions. The



Source: VAR with intercept using monthly data for 1965:1-1990:1 for U.S. industrial production (Y), consumer prices (P), M1, and the Federal Funds rate (FF). All data but the interest rate are in logs. All data are from CitiBase. The example is based on Eichenbaum (1992).

Figure 5. Confidence intervals for selected responses to a monetary expansion

novel conclusion of this paper was that *even* in finite-order VAR models in many cases of practical interest the AIC is likely to be preferable to the SIC or HQC.

The paper emphasized the fact that the effects of model misspecification in small samples are strongly asymmetric depending on whether the model is under- or overfit, if interest centres on VAR statistics involving higher-order dynamics such as impulse response functions, variance decompositions, measures of predictability, or long-term forecasts. In terms of the MSE, one would therefore expect less parsimonious criteria like the AIC to result in more accurate estimates and inference than strongly consistent lag order selection criteria like the SIC and HQC that have a stronger tendency to underestimate the lag order in finite samples. A Monte Carlo study illustrated the potential quantitative importance of this point. It was found that the MSEs of impulse response estimates based on the AIC tend to be substantially lower than for estimates based on more parsimonious criteria. Similarly, impulse response confidence intervals based on the AIC lag order estimate tended to be by far the most accurate intervals. In contrast, the SIC estimates typically missed the higher-order dynamics in impulse response functions and often resulted in severely misleading and spuriously tight interval estimates, even for fairly large samples.

The aim of this paper has been to raise the awareness of applied researchers of the implicit trade-offs in the use of lag order selection criteria and to re-open the debate over model selection by demonstrating that parsimony is not necessarily a virtue. It was demonstrated that applied researchers need to give careful thought to the lag order choice because in many practical applications parsimony of the model may obscure the true dynamics. There are several directions for future research. First, the choice of model-selection criterion involves an implicit tradeoff between location (or bias) and scale (or variance) effects. This tradeoff may differ depending on the sample size, the forecast horizon, and the statistic of interest. Further research is needed to identify these tradeoffs in model selection for other statistics, models, and time horizons.

Second, an important extension would be to generalize the results to autoregressive models with possibly infinite lag order. The common assumption that the true model is contained among a set of finite order VAR models is clearly unrealistic. While the results in this paper only apply to finite lag order processes, in practice, one would not expect fundamental differences in results between higher-order VAR(p) and infinite order VAR models, as the former may be viewed as an approximation of the latter. Some preliminary work on lag order selection in infinite-order processes has been presented by Berkowitz, Birgean, and Kilian (1999).

A third extension would be to allow for the effects of lag order uncertainty in inference. This paper has documented that coverage rates of confidence intervals may drop drastically if the lag order is unknown and has to be estimated from the data. Part of the problem is that standard inference ignores the uncertainty associated with the lag order estimate in finite samples (see Pötscher, 1991). Kilian (1998b) examines a modified bootstrap algorithm which accounts for the fact that the lag order is determined based on the same data set used for the estimation of the VAR model. This modification can be shown to substantially enhance the coverage accuracy of bootstrap confidence intervals in many cases.

APPENDIX: THE NON-PARAMETRIC BIAS-CORRECTED BOOTSTRAP ALGORITHM OF KILIAN (1998a)

Let $\Psi = E(\hat{\beta}-\beta) = b(\hat{\beta})/T + O(T^{-3/2})$, where $\hat{\beta}$ denotes the OLS estimate of β (see Pope, 1990). Given the assumptions of the second section of this paper, under some regularity conditions, a

bias-corrected bootstrap confidence interval for this impulse response estimate may be constructed as follows:

- *Step 1a*: Determine the lag order p by an appropriate model selection criterion such as the AIC and fit a VAR($\hat{p}$) model to the data $\{y_t\}$. Estimate $\hat{\beta}$ and calculate the first-order bias $\hat{\Psi} = b(\hat{\beta})/T$ using the closed-form solution given in Pope (1990).
- *Step 1b*: The companion matrix is the autoregressive coefficient matrix obtained by expressing the N-dimensional VAR(p) process as an N p-dimensional VAR(1) process (see Lütkepohl, 1991, p. 11). Calculate the modulus of the largest root of the companion matrix associated with $\hat{\beta}$. Denote the modulus by $m(\hat{\beta})$. If $m(\hat{\beta}) \geqslant 1$, set $\tilde{\beta} = \hat{\beta}$ without any adjustments. If $m(\hat{\beta}) < 1$, construct the bias-corrected coefficient estimate $\tilde{\beta} = \hat{\beta} - \hat{\Psi}$. If $m(\tilde{\beta}) \geqslant 1$, let $\hat{\Psi}_1 = \hat{\Psi}$ and $\delta_1 = 1$ and define $\hat{\Psi}_{i+1} = \delta_i \hat{\Psi}_i$ and $\delta_{i+1} = \delta_i - 0.01$. Set $\tilde{\beta} = \tilde{\beta}_i$ after iterating on $\tilde{\beta}_i = \hat{\beta} - \hat{\Psi}_i$ for $i = 1, 2, \ldots$, until $m(\tilde{\beta}_i) < 1$. By changing the grid for δ, one can make $m(\tilde{\beta})$ arbitrarily close to unity. The purpose of this stationarity correction is to avoid pushing stationary impulse response estimates into the non-stationary region. The adjustment has no effect asymptotically and does not restrict the parameter space of the OLS estimator, since it does not shrink the OLS estimate $\hat{\beta}$ itself, but only its bias estimate.
- *Step 2a*: Using standard non-parametric resampling techniques, generate R bootstrap replications $\{y_t^*\}$ of $\{y_t\}$ based on the recursion:

$$y_t^* = \tilde{B}_1 y_{t-1}^* + \tilde{B}_2 y_{t-2}^* + \ldots + \tilde{B}_{\hat{p}} y_{t-\hat{p}}^* + u_t^*$$

  For each bootstrap replication $\{y_t^*\}$, fit a VAR($\hat{p}$) to $\{y_t^*\}$. Estimate $\hat{\beta}^*$ and $\hat{\sigma}_u^*$ and construct the first-order bias estimate $\hat{\Psi}^* = b(\hat{\beta}^*)/T$.
- *Step 2b*: Calculate the bias-corrected estimate $\tilde{\beta}^*$ from $\hat{\beta}^*$ and $\hat{\Psi}^*$ following the instructions in step 1b with the obvious changes in notation.
- *Step 3*: Read off the α and $1-\alpha$ percentile interval endpoints of the distribution of the bootstrap impulse response estimate $\theta_{kl,i}(\tilde{\beta}^*, \hat{\sigma}^*, \hat{p})$.

ACKNOWLEDGEMENTS

I thank the editor, the associate editor, Bob Barsky, Larry Christiano, Frank Diebold, Phil Howrey, Bob Stine, Chris Sims, Tao Zha and two anonymous referees for helpful discussions and comments on an earlier version of this paper.

REFERENCES

Berkowitz J, Birgean I, Kilian L. 1999. On the finite-sample accuracy of nonparametric resampling algorithms for economic time series. Forthcoming in *Advances in Econometrics: Applying Kernel and Nonparametric Estimation to Economic Topics*, Fomby TB, Hill RC (eds). **14**.

Eichenbaum M. 1992. Comments 'Interpreting the macroeconomic time series facts: the effects of monetary policy' by Christopher Sims. *European Economic Review* **36**: 1001–1011.

Granger CWJ, King ML, White H. 1995. Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics* **67**: 173–187.

Härdle W, Bowman A. 1988. Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *Journal of the American Statistical Association* **83**: 102–110.

Hannan EJ, Quinn BG. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* **41**: 190–195.
Hurvich CM, Tsai C-L. 1993. A corrected Akaike Information Criterion for vector autoregressive model selection. *Journal of Time Series Analysis* **14**: 271–279.
Kabaila PM. 1995. The effect of model selection on confidence regions and prediction regions. *Econometric Theory* **11**: 537–549.
Kilian L. 1998a. Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics* **80**: 218–230.
Kilian L. 1998b. Accounting for lag order uncertainty in autoregressions: the endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis* **19**: 531–548.
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. 1991. *Introduction to Multiple Time Series Analysis*. Springer-Verlag: New York.
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.
Nishi R. 1988. Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis* **27**: 392–403.
Paulsen J, Tjøstheim D. 1985. On the estimation of residual variance and order in autoregressive time series. *Journal of the Royal Statistical Society B* **47**: 216–228.
Pope AL. 1990. Biases of estimators in multivariate non-Gaussian autoregressions. *Journal of Time Series Analysis* **11**: 249–258.
Pötscher BM. 1991. Effects of model selection on inference. *Econometric Theory* **7**: 63–185.
Pötscher BM. 1995. Comment on 'The effect of model selection on confidence regions and prediction regions' by P. Kabaila. *Econometric Theory* **11**: 550–559.
Quinn BG. 1980. Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society B* **42**: 182–185.
Quinn BG. 1988. A note on AIC order determination for multivariate autoregressions. *Journal of Time Series Analysis* **9**: 241–245.
Shibata R. 1976. Selection of the order of an autoregressive model by Akaike's Information Criterion. *Biometrika* **63**: 117–126.
Shibata R. 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear regression. *Annals of Statistics* **8**: 147–164.
Shibata R. 1983. A theoretical view of the use of the AIC. In Time Series Analysis: Theory and Practice 4, Anderson OD (ed.). Elsevier/North-Holland: Amsterdam; 237–244.
Shibata R. 1986. Consistency of model selection and parameter estimation. In Essays in Time Series and Allied Processes, Gani J, Priestley MB (eds). Applied Probability Trust: Sheffield; 127–141.
Sin CY, White H. 1996. Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics* **71**: 207–223.

Author's biography:
Lutz Kilian is Assistant Professor of Economics at the University of Michigan and Research Affiliate at the Centre for Economic Policy Research, London, UK. He holds a PhD in economics from the University of Pennsylvania. His current research interests include empirical macroeconomics and bootstrap inference for time series. His articles have appeared in *Advances in Econometrics*, *Econometric Reviews*, *Journal of Applied Econometrics*, *Journal of Business and Economic Statistics*, *Journal of Money, Credit and Banking*, *Journal of Time Series Analysis*, and *Review of Economics and Statistics*.

Author's address: **Lutz Kilian**, Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220, USA.

