NAOMI
A New Quarterly Forecasting Model
Part I: Proposed Model Selection Strategy

Stephen Murchison

Department of Finance Working Paper
2001-19

The author thanks Patrick Georges, Alain Guay and Nicholas Moreau for very useful comments and suggestions. The author also thanks Robert Amano, Don Coletti , Robert Tetlow and the participants of the 2000 meetings of the Canadian Economics Association for helpful discussions.

Abstract

This is the first of three papers describing NAOMI[1], a new quarterly forecasting model developed in the Economic Analysis and Forecasting Division at the Department of Finance. NAOMI's intended purpose is twofold. First, it is capable of producing reliable, judgement-free macroeconomic forecasts on a timely basis. Second, it can accurately quantify the level of uncertainty associated with each forecast while ensuring this uncertainty is minimised. Jointly achieving these two objectives represents a formidable task and necessitates a somewhat unique approach to the model building exercise. This paper outlines the model building strategy used to create NAOMI. The proposed approach differs from existing methods in that its sole objective is to minimise forecast uncertainty. We demonstrate that the high level of forecast imprecision typically associated with multi-equation forecasting models such as vector autoregressions stems from the large number of free parameters embodied in such systems. Furthermore, it is shown that lag length selection procedures based on in-sample fit often produce sub-optimal forecasting models in small samples. This paper proposes a technique whereby the sample is divided into two distinct sub-samples, called the estimation and evaluation sets. We then suggest choosing the lag order that minimises the mean squared error over the evaluation sample conditional on the OLS parameter estimates over the estimation sample. Monte Carlo results indicate that when the number of free parameters is large relative to the sample size there are appreciable gains to employing the proposed strategy. While the focus of this paper is on VAR models, this technique is well suited to a wide range of single-equation models used by applied macroeconomists.

[1] NAOMI stands for North American Open-Economy Macroeconometric Intergrated Model.

Résumé

Voici le premier de trois documents qui décrivent MIOAN[2], un nouveau modèle de prévisions trimestrielles mis au point dans la Division de l'analyse et des prévisions économiques au Ministère des finances. MIOAN a un double but : d'abord, produire rapidement des prévisions macro-économiques sans jugements mais fiables et ensuite, quantifier avec précision l'incertitude que comporte chaque prévision tout en veillant à réduire le plus possible cette incertitude. L'atteinte simultanée de ces deux objectifs représente un défi de taille et nécessite une démarche particulière au chapitre de la conception de modèles. Le présent document énonce la stratégie de conception de modèles utilisée pour créer MIOAN. Elle diffère des méthodes actuelles, son seul objectif consistant à réduire l'incertitude des prévisions. Nous démontrons que le niveau élevé d'imprécision qui caractérise habituellement les modèles de prévisions à plusieurs équations, comme les autorégressions vectorielles, découle du grand nombre de paramètres libres que comportent ces systèmes. En outre, il est prouvé que les méthodes de sélection de la durée du retard intégrées à un échantillon débouchent sur des modèles de prévisions de second rang lorsqu'elles sont utilisées avec de petits échantillons. Dans le présent document, nous proposons une technique qui permet de scinder l'échantillon en deux sous-échantillons distincts appelés ensembles d'estimation et d'évaluation. Nous entendons ensuite choisir l'ordre de retard qui réduit le plus l'erreur quadratique moyenne à l'égard de l'échantillon d'évaluation, à partir des estimations des paramètres par MCO (moindres carrés ordinaires) dans l'échantillon d'estimation. Les simulations de Monte Carlo révèlent que lorsque le nombre de paramètres libres est important par rapport à la taille de l'échantillon, il est très avantageux d'appliquer la stratégie proposée. Bien que le présent document porte plus particulièrement sur les modèles d'autorégressions vectorielles, cette technique convient bien à une vaste gamme de modèles à équation unique utilisés en macro-économie appliquée.

[2] MIOAN désigne le modèle Macro-économique Intégré de l'économie Ouverte de l'Amérique du Nord.

1.0 Introduction

A key commitment of the Canadian Forecasting Group at the Department of Finance is to constantly assess the nature and degree of economic risks that the federal government needs to factor into its fiscal plans. Knowledge of both the most probable economic outcome and the likelihood of a specific range of outcomes is critical to the process fiscal policy design. One approach to evaluating the magnitude and sources of risk is through the use of an econometric forecasting model. A highly desirable property of such a model is forecast accuracy since this will reduce the uncertainty surrounding its predictions.[3] A second feature is ease of use, the size and degree of complexity of the model must be such that forecasts and their corresponding confidence bands can be computed quickly.

Experience has shown that given the relatively small samples typically available to economists, forecast accuracy is often closely related to the size of the model. As the number of estimated parameters grow, the degree of precision with which one is able to estimate each parameter is eroded. This problem manifests itself in poor out-of-sample forecasts and wide confidence bands. Consequently our objective of forecast precision requires the selection of a relatively parsimonious representation of only a few key macro variables. Such a model will also serve well the second stated objective, ease of use. However, the search for a highly parsimonious structure must be tempered by the knowledge that too small a model may omit important economic variables. In addition, such a model will tend limit the number of interesting questions that we can ask of it. In practice, striking an optimal balance between economic richness and parsimony is difficult.

This paper presents a procedure that is reasonably easy to implement and is well suited to the task of model selection when one is primarily interested in making accurate predictions. By explicitly incorporating out-of-sample forecast information into the lag length/variable selection process this procedure is demonstrated to perform well relative to other criteria. This performance advantage is most significant with realistic sized models and sample sizes. While this paper

[3] The size of the estimated confidence bands is directly related to the average size of historical forecast errors.

focuses on vector autoregressions (VAR) as the forecasting device, this procedure can easily be generalised to a much wider class of dynamic economic models.[4]

The second paper in this series applies this technique to the construction of the Canadian side of NAOMI, a new quarterly forecasting model developed in the Economic Forecasting and Division at Finance.[5] NAOMI is a small, fully-estimated model of the Canadian and US economies. In addition to producing a model-consistent forecast of key Canadian and US variables each quarter, NAOMI also provides a solid foundation for assessing both the magnitude and sources of uncertainty over the forecast horizon. For instance, the model is capable of providing objective answers to questions such as; what is the probability that nominal income will grow by at least 2.5% over the next year? or how confident can we be that long term interest rates will not rise over the course of the next 6 quarters? Furthermore, statements such as 'the main source of output growth uncertainty 1-2 years from now is interest rate uncertainty over the next 2 quarters' are now possible. This capacity can easily be extended to the assessment of budget balance risk. Specifically, NAOMI can generate fiscal prudence factors that take into consideration both shock and parameter uncertainty.

NAOMI can be loosely thought of as a restricted VAR model.[6] Specifically, we permit different lag lengths for each variable in each equation so as to reduce the number of free parameters. We also permit the exclusion of variables in certain equations. The lag length selection technique proposed in this paper is generalised to that of a model selection procedure for the purpose of building NAOMI. Variable inclusion is determined by a combination of economic theory and this procedure whereas lag length is determined entirely by the procedure.[7]

The paper is divided as follows; section 2.0 provides a brief summary of two key problems associated with using VARs as forecasting devices. Section 3.0 addresses the second problem in more detail and provides an example using Canadian data. Section 4.0 describes and compares

---

[4]We focus on VAR models mainly because of their symmetry. Since there is only one lag parameter, *k*, to choose it is easy to run and summarise the monte carlo experiments.

[5] NAOMI stands for North American Open-economy Macroeconometric Integrated model. The second and third papers describe the structure of the Canadian and US sides of the model, respectively.

[6] With the notable exception that the instrument of monetary policy in the model is explicitly forward looking.

[7] Economic theory rarely provides much guidance in selecting lag length

three popular lag length selection criteria currently employed in empirical research. Section 5.0 provides a formal introduction to the proposed methodology and outlines the Monte Carlo framework used to compare the criteria. Section 6.0 provides results for two artificial data generating processes (d.g.p.), a small and large VAR. Section 7.0 provides a brief summary of extensions and future applications of the technique while Section 8.0 concludes.

2.0 Unrestricted VARs and forecasting

One of the most popular approaches to the task of applied forecasting in recent years has been the vector autoregression model (VAR). In its unrestricted form a VAR may be thought of as a system of equations whereby each equation contains a pre-specified number of lags of each variable in the system. This approach has a number of desirable features including;

(a) The economist is not required to classify variables as endogenous or exogenous. Every variable in the VAR is taken as endogenous.

(b) There is no simultaneity to model. Contemporary relationships are 'hidden' within the residual covariance matrix. As such no 'incredible' identifying restrictions are necessary (Sims 1980).

(c) Because no simultaneity is modelled and because each equation contains exactly the same set of regressors, the system may estimated efficiently and quickly by OLS.

(d) Because each variable helps forecast the others, the system is self-contained. Unlike single equation methods, the model does not require a set of forecasts for all the right hand side variables from another source.

The VAR model is a very convenient tool for forecasting, the researcher is required only to choose the appropriate variables and lags to include. Unfortunately, practical experience with unrestricted VARs has been rather poor. The inability of such models to produce accurate forecasts with reasonable sized confidences bands stems from two fundamental problems. The first problem arises as a consequence of the reduced form nature of a VAR. Because proper account is not taken of such factors as agents' expectations or shifts in monetary/fiscal policy regime, VARs tend to suffer from temporal parameter instability. The second shortcoming stems from the large number of free parameters in the model relative to the typical sample size, i.e. the

overparameterization problem. The first issue can be addressed within a time varying parameter framework using the Kalman filter. This paper deals explicitly with the second problem.

3.0 Consistency, bias, efficiency and the process of model selection

When building any dynamic model, one is faced with the problem of selecting the relevant variables and the correct number of lags, k, to include. Since the number of lags embodied in the d.g.p is usually unknown, k is typically treated as a random variable and some form of (quasi) statistical criterion is employed in its selection. It is interesting to note that due to the consistency of OLS, the issue of lag length is relevant only in finite samples. With enough data one could always choose an arbitrarily high lag length and allow the data to set the irrelevant lags to zero. Unfortunately, the applied forecaster rarely possesses enough data to render such an approach feasible.

At this point it is useful to explore the consequences of choosing too many/few lags when the sample size is small. While the focus here is on VAR models, the results can easily be extended to a much larger class of multivariate structures. In order to develop these concepts it is first necessary to introduce the definition of mean squared error (mse). Consider the p-dimensional VAR(k) process given by;

$$(3.0.1) \quad \mathbf{y}_t = \sum_{i=1}^{k} \boldsymbol{\beta}_i \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t$$

with $\mathbf{y}_t = (y_{1,t}, y_{2,t}, \ldots, y_{p,t},)'$, $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, \ldots, \varepsilon_{p,t},)'$ and $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t') = \boldsymbol{\Sigma}$. For our purposes, it is convenient to rewrite the VAR(k) as a VAR(1) as follows;

$$(3.0.2) \quad \mathbf{Y}_t = \boldsymbol{\Theta} \mathbf{Y}_{t-1} + \mathbf{e}_t$$

with

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \boldsymbol{\beta}_3 & \cdots & \boldsymbol{\beta}_{k-1} & \boldsymbol{\beta}_k \\ \mathbf{I}_p & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_p & \mathbf{0} \end{bmatrix}; \quad \mathbf{Y}_t = \begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-k+1} \end{bmatrix}; \quad \mathbf{e}_t = \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}$$

The mse matrix for the optimal v-period ahead forecast is given as;

$$(3.0.3)\quad \boldsymbol{\Xi}^{v} \equiv \mathrm{E}\left[\left(\mathbf{Y}_{t+v} - \mathbf{Y}_{t+v|t}\right)\left(\mathbf{Y}_{t+v} - \mathbf{Y}_{t+v|t}\right)'\right] \text{ with } \mathbf{Y}_{t+v|t} = \boldsymbol{\Theta}^{v}\mathbf{Y}_{t} \text{ so}$$

$$(3.0.4)\quad \boldsymbol{\Xi}^{v} = \mathrm{E}\left[\left(\mathbf{Y}_{t+v} - \boldsymbol{\Theta}^{v}\mathbf{Y}_{t}\right)\left(\mathbf{Y}_{t+v} - \boldsymbol{\Theta}^{v}\mathbf{Y}_{t}\right)'\right] = \sum_{i=0}^{v-1}\boldsymbol{\Theta}^{i}\boldsymbol{\Sigma}\boldsymbol{\Theta}^{i\prime}$$

where $\mathbf{Y}_{t+v|t}$ represents the optimal v-period ahead forecast conditional on information at time t. Now consider the analogous mse matrix evaluated at the OLS estimated parameter estimates;

$$(3.0.5)\quad \tilde{\boldsymbol{\Xi}}^{v} \equiv \mathrm{E}\left[\left(\mathbf{Y}_{t+v} - \tilde{\mathbf{Y}}_{t+v|t}\right)\left(\mathbf{Y}_{t+v} - \tilde{\mathbf{Y}}_{t+v|t}\right)'\right] \text{ with } \tilde{\mathbf{Y}}_{t+v|t} = \tilde{\boldsymbol{\Theta}}^{v}\mathbf{Y}_{t} \text{ so}$$

$$\begin{aligned}
\tilde{\boldsymbol{\Xi}}^{v} &= \mathrm{E}\left[\left(\mathbf{Y}_{t+v} + \boldsymbol{\Theta}^{v}\mathbf{Y}_{t} - \boldsymbol{\Theta}^{v}\mathbf{Y}_{t} - \tilde{\boldsymbol{\Theta}}^{v}\mathbf{Y}_{t}\right)\left(\mathbf{Y}_{t+v} + \boldsymbol{\Theta}^{v}\mathbf{Y}_{t} - \boldsymbol{\Theta}^{v}\mathbf{Y}_{t} - \tilde{\boldsymbol{\Theta}}^{v}\mathbf{Y}_{t}\right)'\right] \\
&= \sum_{i=0}^{v-1}\boldsymbol{\Theta}^{i}\boldsymbol{\Sigma}\boldsymbol{\Theta}^{i\prime} + \mathrm{E}\left[\left(\left(\boldsymbol{\Theta}^{v} - \tilde{\boldsymbol{\Theta}}^{v}\right)\mathbf{Y}_{t}\right)\left(\left(\boldsymbol{\Theta}^{v} - \tilde{\boldsymbol{\Theta}}^{v}\right)\mathbf{Y}_{t}\right)'\right] \\
&= \sum_{i=0}^{v-1}\boldsymbol{\Theta}^{i}\boldsymbol{\Sigma}\boldsymbol{\Theta}^{i\prime} + \mathrm{E}\left[\left(\left(\boldsymbol{\Theta}^{v} + \mathrm{E}\left(\tilde{\boldsymbol{\Theta}}^{v}\right) - \mathrm{E}\left(\tilde{\boldsymbol{\Theta}}^{v}\right) - \tilde{\boldsymbol{\Theta}}^{v}\right)\mathbf{Y}_{t}\right)\left(\left(\boldsymbol{\Theta}^{v} + \mathrm{E}\left(\tilde{\boldsymbol{\Theta}}^{v}\right) - \mathrm{E}\left(\tilde{\boldsymbol{\Theta}}^{v}\right) - \tilde{\boldsymbol{\Theta}}^{v}\right)\mathbf{Y}_{t}\right)'\right]
\end{aligned}$$

$$\begin{aligned}
(3.0.6)\quad &= \underbrace{\sum_{i=0}^{v-1}\boldsymbol{\Theta}^{i}\boldsymbol{\Sigma}\boldsymbol{\Theta}^{i\prime}}_{\text{term1}} \\
&+ \underbrace{\mathrm{E}\left[\left(\left(\mathrm{E}\left(\tilde{\boldsymbol{\Theta}}^{v}\right) - \tilde{\boldsymbol{\Theta}}^{v}\right)\mathbf{Y}_{t}\right)\left(\left(\mathrm{E}\left(\tilde{\boldsymbol{\Theta}}^{v}\right) - \tilde{\boldsymbol{\Theta}}^{v}\right)\mathbf{Y}_{t}\right)'\right]}_{\text{term2}} + \underbrace{\mathrm{E}\left[\left(\left(\boldsymbol{\Theta}^{v} - \mathrm{E}\left(\tilde{\boldsymbol{\Theta}}^{v}\right)\right)\mathbf{Y}_{t}\right)\left(\left(\boldsymbol{\Theta}^{v} - \mathrm{E}\left(\tilde{\boldsymbol{\Theta}}^{v}\right)\right)\mathbf{Y}_{t}\right)'\right]}_{\text{term3}}
\end{aligned}$$

Equation 3.06 differs from 3.0.4 by the second two terms of 3.0.6. The first of the latter two terms (term 2) captures the (square of the) difference between the OLS estimate of $\boldsymbol{\Theta}^{v}$, denoted $\tilde{\boldsymbol{\Theta}}^{v}$, and its expectation. Stated otherwise, it captures the variance of $\tilde{\boldsymbol{\Theta}}^{v}$ about its expectation or simply the variance of $\tilde{\boldsymbol{\Theta}}^{v}$. The third term captures the (square of the) difference between $\boldsymbol{\Theta}^{v}$ and the expectation of $\tilde{\boldsymbol{\Theta}}^{v}$, i.e. the bias component. We now turn to the issue of lag length selection and how it relates to these two terms.

3.1 Inefficiency; The consequence of setting k too large

With too many lags, there is a loss in efficiency as degrees of freedom (d.f.) are quickly eroded. As a result term 2 in equation 3.0.6 will be unnecessarily large. Recall that with p variables, the inclusion of one extra lag consumes kp d.f. The practical consequence of this will be large out-of-sample forecast errors and wide confidence bands around impulse responses despite an improvement in the in-sample fit. This begs an interesting question; why does the in-sample performance improve at the expense of the model's forecasting ability? The in-sample fit improves as we begin fitting parameters to the stochastic component of the dependent variable, i.e. the error term.[8] In fact if we estimate an equation that contains an arbitrary set of regressors equal to the number of observations we can achieve an R-square of one. This will be true despite the fact that the dependent variable contains a component that is random. In this instance the estimated relationship is unique to that particular sample, that is, to those particular realisations of the disturbance term. Consequently, it will be of little help when it comes to forecasting.

Since efficiency is a finite-sample property, the preceding argument is only relevant in finite and particularly small samples. As the sample size, T, approaches infinity $E\left(\tilde{\boldsymbol{\Theta}}^{v} - E\left(\tilde{\boldsymbol{\Theta}}^{v}\right)\right) \rightarrow \mathbf{0}$ so the first term in 3.0.6 disappears.

In order to illustrate how important the effects of over fitting a model can be consider the following experiment. We begin by estimating a 8 lag VAR using Canadian output, inflation, interest rates and the exchange rate as well as U.S. output and interest rates from 1972 to 1993 (about 85 observations). We then simulate the model dynamically starting in 1974 (1972q1 plus 8 lags) in order to assess what the conditional expectation for inflation would look like at this particular point in time. Recall that the information set is 2 years of data and every variable in the system is endogenous. Therefore, one should expect to see some near term variation in inflation owing to the lagged effects of, for instance, the output gap and exchange rate. As the effects of these shocks dissipate inflation should converge to its unconditional mean.

Error! Not a valid link.

[8] Here it is useful to think of the regressors as being non-stochastic and hence the only random component of the dependent variable is an additive disturbance term. This is not, however, a necessary requirement.

Turning to Figure 1.0 we see that the VAR is able to forecast the decline in inflation in 1974-75, the slight increase in 1976 and the continued decline into 1978 (the end of the first oil-price shock). Moreover, the model does an excellent job of forecasting the inflationary consequences of the second oil price shock that occurred some six years after the beginning of the simulation. This despite the fact that oil prices are not included in the model. Clearly the parameter values of this model are unique to this particular sample and would be of no practical use for forecasting.

3.2 Inconsistency: The consequence of setting k too small

With too few lags, the model is an inconsistent estimator of the true d.g.p.[9] A consistent estimator, $\tilde{\beta}$, of β has the property that $\underset{T\to\infty}{\text{plim}}\ \tilde{\beta}_T = \beta$. Consistency is an asymptotic property of an estimator. One can reasonably think of an inconsistent estimator as being biased even in large samples. If k is too small, term 3 in equation 3.0.6 will be positive thereby increasing the mse. This problem, unlike the loss in efficiency associated with too many lags, will not go away as more observations are added to the sample. At best we can hope to have a reasonable approximation of the actual process, however, we will never uncover the truth.

3.3 Bias; The consequence of using OLS with VARs

When the four Gauss-Markov (GM) conditions are satisfied, OLS is said to be BLUE or Best (i.e. minimum variance) Linear Unbiased Estimator. The (strong) GM condition of non-stochastic regressors ensures that OLS is unbiased. An estimator $\tilde{\beta}$ for β said to be unbiased if $E(\tilde{\beta}) = \beta$. Clearly when one or more of the regressors are lags of the dependent variable, this condition is not met.

Since it is possible to efficiently estimate a VAR by performing OLS on each equation individually, it is sufficient here to show that OLS is biased for a dynamic single-equation model. Consider the AR(k) specification;

(3.3.1) $y_t = \sum_{i=1}^{k} \beta_i y_{t-i} + u_t$ or $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$

in matrix notation. The OLS estimate of $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Consequently,

(3.3.2) $\mathrm{E}(\tilde{\boldsymbol{\beta}}) = \mathrm{E}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})\right) = \boldsymbol{\beta} + \mathrm{E}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\right) \neq \boldsymbol{\beta}$ since $\mathrm{E}(\mathbf{X}'\mathbf{u}) \neq 0$

While the regressors are assumed independent of the contemporaneous residual, the same cannot be said for lags of the residual. Consequently, OLS is consistent but biased. Unfortunately, the bias is a function of the unknown model parameters and therefore no simple correction exists. However, it can be shown that;

(3.3.3) $$\frac{\partial \mathbf{E}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} > \mathbf{0}$$

Coefficient bias stemming from OLS estimation of a VAR will also increase the mse matrix through term 3. Moreover, this bias is increasing in the persistence of the process.[10] Since macroeconomic time series are often characterised as having roots close to or equal to one, this bias can be substantial. Moreover, it exists even when k is chosen correctly.

3.4 'Optimal' versus 'True' models

Based on the preceding analysis, one is tempted to conclude that it is always preferable to select the same number of lags as in the true model. However, this need not be the case if the researcher's sole objective is to minimise forecast errors. For instance, if the loss in efficiency stemming from estimating an additional q lags reduces forecast accuracy by more than the bias introduced by incorrectly setting those parameters to zero, then clearly one is better off working with the smaller, albeit misspecified model. In other words, if the increase in term 2 exceeds the decline in term 3, then one is better working with the smaller model.

To help solidify the proposition that optimality (defined to mean mse minimising) need not imply truth, consider the following model selection problem; suppose the true model is given by

$$y_t = \beta x_t + u_t \quad u_t \sim (0, \sigma_u^2) \quad \beta \neq 0$$

---

[9] This will manifest itself in serially correlated errors.

[10] It is often noted in the time series literature that impulse response functions generated from estimated VAR models appear convergent even when the underlying series have unit roots. This stems from fact that OLS tends to impose stationarity on the series.

and r and T realisations of y and x, respectively, are observed by the researcher, T>r. Furthermore, suppose the researcher knows the functional form of the true model but does not know β. The problem is to forecast y_t from r+1 to T so as to minimise the mean squared forecast error.[11] The most obvious method is to condition one's choice of y on x since x is given. However, following this route requires an estimate of β. The second option is to impose β=0 and form the unconditional expectation of y, which eliminates the need to estimate β.[12] The question is, which approach will yield better forecasts?

It is reasonably straightforward to show that the condition which equilibrates the mse arising from the two approaches is $s.e.(\tilde{\beta}) = \beta$ [13], where $s.e.(\tilde{\beta})$ denotes the standard error of any unbiased estimator of β. If $s.e.(\tilde{\beta}) < \beta$ then option one is preferable whereas if $s.e.(\tilde{\beta}) > \beta$ the second choice is optimal. The intuition is as follows; choosing to estimate β introduces parameter uncertainty that raises the forecasting errors. Setting $\tilde{\beta} = 0$, on the other hand, throws out useful information regarding the relationship between y and x. That is, if $\tilde{\beta} = 0$ then $\tilde{\beta}$ is biased and the bias is equal to $\beta \ \forall \ t \in [r+1,T]$. Now, $s.e.(\tilde{\beta})$ may be thought of as representing the average deviation of $\tilde{\beta}$ from β. Now the previously stated equality condition should be clear. If the average deviation of $\tilde{\beta}$ from β is greater (less) than β, forming the unconditional (conditional) expectation is preferable.

The point of preceding exercise was to show analytically that the optimal model, from a forecaster's perspective, will not necessarily correspond with the true model in small samples. This comes as a direct consequence of the overparameterization problem highlighted in section 1.0. To convince ourselves that such a situation can indeed arise, we employ monte carlo techniques to generate artificial forecasts from models with varying lag lengths. Specifically, we

---

[11] The point of making T>r is to abstract from the uncertainty introduced by not knowing the future values of *x*. Here we just want to focus on parameter and innovation uncertainty so we pretend that the future path of *x* is known.

[12] This is equivalent to dropping a lag in an autoregression model.

[13] This is equivalent to a t-value of one.

begin by writing down the true model which in this case is a VAR(4) of dimension 5 (5 variables, 4 lags) and generate a multitude of sequences for each variable using normal innovations from an identity covariance matrix. We can then calculate analytically the mse arising from this model.[14]

Once the mse is calculated for the true model[15] we can explore the marginal effects of lag length and parameter uncertainty by varying both the sample size, T, and the number of lags in the VAR, k. Figure 2.0 graphs the ratio

$$(3.4.1) \quad \Pi(T,k) \equiv \frac{MSE(T,k)}{MSE(\infty,4)}$$

which is referred to as the inflation factor, $\Pi(T,k)$. $(1-\Pi(T,k))*100$ measures the percentage increase in the mse arising from parameter uncertainty when k=4 and parameter uncertainty/incorrect lag length when $k \neq 4$. This particular example illustrates a number of interesting points including;

(a) For fixed k, $\Pi(T,k)$ is strictly decreasing in T which reflects the fact that the degree of uncertainty surrounding parameter estimates is a function of T, the information set.

(b) When the sample size is very small, lower values of k tend to produce smaller forecast errors. Indeed, $\Pi(50,2) < \Pi(50,3) < \Pi(50,4) < \Pi(50,5)$ even though k=4 is the true model. So the bias introduced by incorrectly setting those parameters to zero must be small relative to the sampling variability associated with their least squares estimates. However, as T grows the ranking quickly changes. For T=100, the true model now dominates all but the VAR(3) (not shown) and for T>200, the correct model dominates.

---

[14] Why does the true model make forecast errors? Recall that these forecasts represent conditional expectations at some fixed point in time, which implies that future innovations are unknown and their rational expectation is zero. Consequently, forecast errors arise purely from what is referred to as innovation uncertainty.

[15] The mse for the model is the simple average of the mse for the individual series contained in the VAR. Naturally, the mse is calculated in the same manner for each of the estimated models.

(c) For k>4, plim $\Pi(T,k)=\Pi(4,T)=1$. But for k<4, this condition does not hold. This illustrates the consistency of any VAR($\{l\in \mathbf{R}^{+};l\geq k\}$) for VAR(k) and the inconsistency of any VAR($\{l\in \mathbf{R}^{+}; l<k\}$). Graphically, we note that $\Pi(5,T)$ gradually converges toward $\Pi(4,T)$ which itself is converging on one as T becomes large. It turns out that this result will hold regardless of how many lags we included in the model, provided it is greater than the true number. Also noteworthy is the fact that for the VAR(2), Π does not converge to one, regardless of how many observations one includes.

Error! Not a valid link.

4.0 How do lag length selection procedures differ?

In single equation models the most common way of testing single (multiple) variable exclusions is with a t (F) test, the outcome of which is determined by the change in the residual variance across the null and alternative hypotheses. However, in a multiple equation framework these tests are inadequate because one must consider the impact of a given variable or lag on the statistical behaviour of the whole system rather than any one particular equation. Thus we consider the multi-dimensional analogue of the residual variance, the residual covariance matrix and perform a likelihood ratio (LR) test. For instance, one could test the null that the data were generated by a VAR(k) against the alternative that they were generated by a VAR(k+n) for n>0. One strategy is to set k to an arbitrarily high number, set n=1 and then sequentially test down until the null is rejected. The likelihood ratio statistic has an asymptotic chi-square distribution under the null with degrees of freedom equal to the total number of restrictions. Since the maximum value of the likelihood function is always lower for the restricted model, this statistic is always positive. However, we are concerned with how likely it is that one should obtain a particular positive value for the statistic if the set restrictions are indeed true. If the probability is quite low, one is inclined to reject the restrictions. Unfortunately, the specified size of any single hypothesis test is conditional on the all of the previous nulls being true. Furthermore, these probabilities are typically only valid as T goes to infinity. In small samples, this statistic may provide little guidance in selecting the correct lag length even if all previous nulls are true (Lütkepohl (1985)).

Sims(1980) offers a correction that is designed to improve the finite sample properties of the test. This correction effectively reduces the incidence of type 1 error in small samples[16].

This correction has not, however, limited the proliferation of competing criteria into mainstream econometrics (for a review of these criteria see Lütkepohl (1985)). Here we consider two of the more popular approaches, the Akaike Information Criterion (AIC) (Akaike 1973, 1974) and the Schwarz Bayesian Criterion (SBC) (Schwarz (1978)). Essentially these tests incorporate an explicit penalty function for additional parameters which is designed to reduce the probability of selecting an over-fit model. For any reasonable sample size the penalty function is greater for SBC than AIC. Since these tests have been offered as alternatives to the LR statistic, it would useful to know by how much they actually differ in practice. To get a flavour of this consider the following example that compares LR with SBC. Suppose we have 50 observations for 5 variables and wish to test the hypothesis that a VAR(5) and a VAR(4) are not statistically different. Thus the unrestricted model (VAR(5)) has N=5×5×5=125 parameters whereas the restricted model (VAR(4)) has N_r = 5×5×4=100. Employing the LR test we would use the following formula;

(4.0.1) $$\lambda = (T - c)\left[\log\left|\Sigma_r\right| - \log\left|\Sigma\right|\right]$$

where c=25 is the number of parameters in each of the unrestricted equations and $\left|\Sigma\right|$, $\left|\Sigma_r\right|$ are respectively the determinants of the unrestricted and restricted covariance matricies of residuals. We will reject the restriction (accept the VAR(5)) if $\lambda > \lambda_{crit}$ at some significance level.

The SBC test statistic is given as;

(4.0.2) $$SBC = T \log\left|\Sigma\right| + N\log(T)$$

for the unrestricted model and;

(4.0.3) $$SBC_r = T\log\left|\Sigma_r\right| + N_r\log(T)$$

for the restricted model . In the case of the SBC criterion, we reject the restricted model (VAR(4)) if $SBC<SBC_r$.

---

[16] The effect of the correction quickly diminishes as T becomes large.

Equivalently, we reject the restriction if;

$$T \log |\Sigma_r| + N_r \log(T) - (T \log |\Sigma| + N \log(T)) > 0$$

$$(\log |\Sigma_r| - \log |\Sigma|) + T^{-1} \log(T)(N_r - N) > 0$$

In this example $N_r - N = -25$ and $T = 50$ so we obtain the result that;

$$(4.0.4) \quad \log |\Sigma_r| - \log |\Sigma| > \log(T) T^{-1} (N - N_r) = 1.96$$

If we assume that the asymptotic distribution of λ is a good approximation for T=50, then we can use it to calculate the implied size of the SBC test for this particular experiment. That is, if $\log |\Sigma_r| - \log |\Sigma| = 1.96$, then using the definition of λ we have;

$$(4.0.5) \quad \lambda(25) = (T - c)\left[\log |\Sigma_r| - \log |\Sigma|\right] = (50 - 25) * 1.96 = 49$$

using the cumulative χ^2 distribution we get $\Pr(\lambda(25) \geq 49) \sim 0.005$. Consequently, the implied size of this test is about 0.5% which is considerably smaller than the 5% or 1% levels typically used in applied work.[17] This means that the researcher is far less likely to select the larger model using SBC than using LR when the sample size is small.[18] Hence SBC will tend to select more parsimonious models when the data set is small.

The preceding example was intended to show differences in the behaviour of selection criteria when the sample is small. We can now generalise this intuition to make the following proposition. Given T, c, N, $|\Sigma_r|$ and $|\Sigma|$ one can calculate the area of the region (or cumulative probability) of rejecting (the null with size α) with the LR test and not rejecting using SBC or AIC for that matter. This probability is given as;

---

[17] This should not be taken to mean that if the chosen size is 5% the researcher is necessarily better off using the LR test. Recall that these sizes are based on the asymptotic distribution of λ. There may be large size distortions introduced by using the chi-square distribution in small samples. Furthermore, given that the LR test is employed in a sequential testing framework, the conditional and unconditional sizes will often be quite different (see, for instance, Lütkepohl (1991). So it may be case that the SBC comes closer to a 5% prob. of type 1 error than does the LR test.

[18] Provided a reasonable significance level is used.

(4.0.6) $\vartheta = \alpha - \int_{o}^{a} f(x)dx = \alpha - F(a)$ with $a = \frac{(T-c)\log T(N-N_R)}{T}$

where $F(a)$ is the cumulative χ^2 function evaluated at a. However, given the consistency of SBC and LR, these small sample differences will tend to zero as T becomes large;

(4.0.7) $\lim_{T\to\infty} [F(a)-\alpha] = 0$

It is worth noting that consistency does not hold for AIC as it asymptotically over estimates the true lag order with positive probability (see Quinn(1980) for a proof).

5.0 An alternative model selection procedure

All of the lag length procedures discussed in the previous section are based on the in-sample fit of the model relative to the number of freely estimated parameters. Minimising the log the determinant of the residual covariance matrix is equivalent to minimising $\log\left(\sum_{i=0}^{v-1}\Theta^i\tilde{\Sigma}\Theta^{i\prime}\right)$ for v=1 which is (asymptotically) equivalent to the first term in equation 3.0.6. These criteria explicitly consider the in-sample analogue of one of the three terms in this mse matrix expression. In an attempt to account for the second term, they apply a penalty function that is increasing in the number of estimated parameters. This correction attempts to reduce the likelihood of over-fitting the model. Unfortunately, they are only asymptotically valid[19] and therefore serve as approximations only in small samples. Furthermore, all 3 corrected statistics ignore the third term.

Rather than attaching an asymptotically-valid penalty function that ignores the problem of bias, why not attempt to generate a statistic that is valid in small samples. Specifically, why not divide the sample into two parts; an estimation and evaluation sub-sample? Then choose the lag structure that yields the lowest dynamic forecast errors over the evaluation sub-sample conditional on the OLS parameters estimated over the estimation sub-sample. Hereafter this estimator is referred to as the out-of-sample forecast or **OSF** estimator.

To make this proposition concrete consider the following example; suppose one has T observations on p variables and wishes to form a v period ahead conditional expectation for each variable. As illustrated in Figure 3.0, the OSF procedure suggests first dividing the T observations into sub-samples [1,s] and [s+1,T]. Second, estimate a VAR(1), VAR(2),…,VAR(j) over the sample [1,s] and generate the mse matrix for each model over the sample [s+1,T]. Finally, select the number of lags, k, that minimises (a transformation of) the mse matrix and then re-estimate this model over [1,T]. Formally, the OSF problem is given as;

$$(5.0.1) \qquad \min_{k} \left[\sum_{i=s}^{T-v} \left(\mathbf{Y}_{t+v} - \tilde{\mathbf{\Theta}}^{v}(k)\mathbf{Y}_{t} \right)' \mathbf{\Omega} \left(\mathbf{Y}_{t+v} - \tilde{\mathbf{\Theta}}^{v}(k)\mathbf{Y}_{t} \right) \right]$$

$\mathbf{\Omega}$ is a user specified diagonal weighting matrix that describes the relative importance of each variable's forecast errors at horizon v in the loss function, normalised such that $\mathrm{Tr}(\mathbf{\Omega}) = 1$. For a fixed point horizon v, $\mathbf{\Omega}$ will have non-zero entries along the first p diagonal entries only. That is, if $\mathbf{\Omega}(i)$ is the i^{th} diagonal entry of $\mathbf{\Omega}$ then;

$$\mathbf{\Omega}(i) > 0 \;\; i = 1,2,\ldots,p \;\text{ and }\; \mathbf{\Omega}(i) = 0 \;\; i = p+1, p+2, \ldots, p(k+1).$$

While a well-specified model contains many variables, the user may only care about accurately forecasting one or two of them. The fiscal agent, for example, may wish to place relatively more weight on interest rates and output growth.

The intuition behind the proposed model selection strategy is quite straightforward, we wish to select the lag structure that will yield the lowest out-of-sample population mse, given T observations. However, since we do not observe the population mse, we must be content with an estimator of this statistic. Obviously, the smaller is s the more precise our estimate of the population mse will be, ceteris paribus. However, we must remain mindful of the fact that as s becomes small parameter uncertainty increases. This underscores the key drawback of dividing the sample, i.e. there is loss in efficiency associated with excluding information from the estimation sample. However, the gains from evaluating the model out-of-sample may prove to exceed this cost. Based on the results of this paper **s = .7T** seems to work quite well.

---

[19] With the potential exception of AIC.

Since we are ultimately interested in this procedure's ability to select the model with lowest population mse, it is worth investigating the small sample properties of the suggested estimator. Here it will useful to revisit the inflation factor, Π(T,k), which is modified slightly to Π(T,k,p) where p is again the number of variables in the model.[20] Addressing the issue of the small sample properties of the proposed estimator is closely linked to the functional form of Π(T,k,p). This stems from the fact that while we want to estimate the mse for a particular lag structure based on T observations, our estimator uses only s observations. Since T is an argument in Π and

$$(5.0.2) \quad \mathrm{E}\left[\frac{\partial \Pi(\mathrm{T},\mathrm{k},\mathrm{p})}{\partial \mathrm{T}}\right] < 0$$

we get the result $\mathrm{E}[\Pi(\mathrm{T},\mathrm{k},\mathrm{p}) - \Pi(\mathrm{s},\mathrm{k},\mathrm{p})] < 0$ since T must be greater than s. Since the expected mse is strictly decreasing in the number of observations included in the estimation period, reducing this number will induce an upward bias in Π(T,k,p). However, if the bias is equal across all values of k, that is;

$$(5.0.3) \quad \mathrm{E}\left[\frac{\partial^2 \Pi(\mathrm{T},\mathrm{k},\mathrm{p})}{\partial \mathrm{T}\partial \mathrm{k}}\right] = 0$$

then the ranking of lag structures in terms of Π(T,k,p) will remain unaffected. Lütkepohl(1991) derives the approximate mse matrix for a forecast horizon of one period and shows that Π(T,k,p) may be written as

$$(5.0.4) \quad \Pi(\mathrm{T},\mathrm{k},\mathrm{p}) = \frac{\mathrm{MSE}(\mathrm{T},\mathrm{k},\mathrm{p})}{\mathrm{MSE}(\infty,\mathrm{k},\mathrm{p})} \approx 1 + \frac{\mathrm{kp}+1}{\mathrm{T}}$$

Consequently, the cross partial is given as

$$(5.0.5) \quad \mathrm{E}\left[\frac{\partial^2 \Pi(\mathrm{T},\mathrm{k},\mathrm{p})}{\partial \mathrm{T}\partial \mathrm{k}}\right] = -\frac{\mathrm{p}}{\mathrm{T}^2} < 0$$

---

[20] Previously the number of variables was taken as fixed so we were in fact evaluating $\Pi(T,k)|p$

This function indicates that the higher is k, the greater is the increase in bias as T decreases. The non-constancy of the bias may introduce a change in the rankings of the estimates of Π. Specifically, this problem will tend to introduce a downward bias in the optimal value of k. As a remedy to this problem, we propose the following bias corrected minimisation problem

$$(5.0.6)\quad \min_{k} \left[\left[\left(1+\frac{(pk+1)(T-s)}{Ts}\right)\varphi\right]^{-1}\sum_{i=s}^{T-v}\left(\mathbf{Y}_{t+v}-\tilde{\boldsymbol{\Theta}}^{v}(k)\mathbf{Y}_{t}\right)'\boldsymbol{\Omega}\left(\mathbf{Y}_{t+v}-\tilde{\boldsymbol{\Theta}}^{v}(k)\mathbf{Y}_{t}\right)\right]$$

with φ = T-s+v. Since this correction is based on asymptotic theory, however, its benefit in small samples unclear. In the next section we investigate the properties of both the corrected and uncorrected estimators.

At this point it is worth investigating what this methodology implies about users' preferences regarding forecast outcomes. For instance, minimising mse suggests the following;

(a) Losses arising from forecast errors are symmetric and quadratic

(b) The user cares only about the fixed point forecast horizon v

5.1 Symmetric and Quadratic Loss - While these two characteristics are common to most loss functions in (partially because the make possible closed-form solutions) one may wish to treat positive errors differently than negative ones or alternatively attach a linear (rather than quadratic) loss to errors. For example, the Fiscal authority may wish to attach a relatively high (low) loss to errors arising from over (under) predicting output growth. Each of these possibilities is feasible within the OSF framework.

5.2 Interval Forecast Horizons - Rather than specifying a fixed-point horizon, one may prefer instead to define an interval such as [v-q,v+q] for some value of q or [1,v]. For instance, the user may wish to place the greatest weight on the one-period ahead forecast error and then specify geometrically declining weights for horizons 2,3, …,v. In this instance, the weighting matrix $\boldsymbol{\Omega}$ is modified such that $\boldsymbol{\Omega}(i) > 0 \;\; i = 1,2,\ldots,p\times v$. Of course optimising at some horizon(s) other than one period is only useful if it actually results in better forecasts at those horizons. In other words, does optimising at horizon v≠1 lead to a different lag structure and hence better forecasts than

using v=1? It can be demonstrated that if the estimated model is a linear approximation to a non-linear system or if the true model is non-nested then there can be gains in forecast accuracy at a horizon v≠1 (and perhaps at all horizons) from optimising at that horizon.

6.0 Monte Carlo Design

We have argued that explicitly considering out-of-sample forecasting properties may facilitate the selection of more accurate forecasting models than those based on in-sample fit. In order to test this hypothesis we set up the following experiment;

Step 1: Write down the "true" model and corresponding covariance matrix. Here we assume the system is driven by a multivariate normal shock process.

Step 2: Using this covariance matrix, $\mathbf{\Sigma}$, generate n (at least 1000) stochastic time series of length T+q where q is a large number and then estimate the model using the first T observations. Calculate the mse over the remaining q observations. Calculate MSE(T,k,p) by averaging across the 1000 mse statistics. Since there are no structural breaks or non-linearities in the underlying d.g.p., we choose a forecast horizon of v=1.

Step 3: Repeat Step 2 by varying T=50,75,…,150 and k=1,2,…,7.[21] Form the 7×5 matrix $\mathbf{R}$ from these results. So for instance, $\mathbf{R}_{23}$ is the mse with 75 observations used for estimation and 3 lags used in the estimated VAR. Form the diagonal matrix ξ such that ξ_{ii} is the minimum value contained in the i^{th} row of $\mathbf{R}$. Hence, each element on the principal diagonal of this array gives the lowest possible mse given a particular number of observations.

Step 4: Again generate n time series, this time of length T=50 and using one of the criteria, select the best model. For instance, calculate the AIC statistic for k=1,2, …,7 and choose the model that makes the AIC statistic the smallest and keep track of the chosen lag length. After doing this n times repeat the procedure for T=75,100,…,150. Form the frequency matrix $\mathbf{V}_{\text{AIC}}$ such that the ij^{th} entry corresponds to the number of time i lags were chosen using 50+25j observations, so $\overline{\mathbf{V}}_{\text{AIC}} = \text{n}^{-1}\mathbf{V}_{\text{AIC}}$ will yield the corresponding probability matrix. Repeat this procedure for LR, SBC and the proposed test procedure, OSF.

Step 5: Using the AIC example, calculate

$$\hat{\mathbf{\Pi}}_{\text{AIC}} = \mathbf{R}_{\text{AIC}}\overline{\mathbf{V}}\xi^{-1}$$

[21] The choice of 7 (6 for the small VAR) as the upper limit was somewhat arbitrary. Basically, you want to choose a number high enough such that the probability of any criteria selecting a number higher than this is small. Of course, choosing too high a number will needlessly increase computation time.

The principal diagonal of this matrix contains the percent increase in mse relative to the optimal model for different values of T. For example, jj^{th} entry will yield the percent increase in mse using the AIC criteria with 50+25j observations. For instance, if this value is zero, the AIC criterion chooses the optimal forecasting model with probability one. On the other hand, a value of 10 indicates that using AIC will yield forecasts errors 10% higher, on average, than if the best forecasting model was used. Obviously, the lower the value the better the procedure.

7.0 Results

In this section we analyse the results using both a small and a large VAR model. The small VAR has 3 lags of 3 variables and is typical of these types of experiments (see, for instance, Lütkepohl (1985)). The large model is intended as a more realistic representation of the Canadian economy. In constructing this d.g.p. we had in mind a structural model containing some measure of inflation, monetary policy stance, the output gap, a long-term nominal interest rate and the exchange rate as endogenous variables. In addition, a well specified model would likely include the U.S. output gap (or output) and commodity prices as strongly exogenous variables. Consequently, the large model has 5 endogenous and 2 exogenous variables with 4 lags. Both models contain roots near, but inside, the unit circle. This feature is intended to capture the level of persistence usually observed in macro time series.

Table 1.0 gives the small VAR probability distributions, $\overline{\mathbf{V}}$**,** for each estimator and for each value of T, the sample size. For instance, with 50 observations, the probability that AIC will select 3 lags (the truth) is 0.26. In each column the **bold** entry represents the optimal (mse minimising) lag length. For this particular dgp, the optimal lag length with T=50 is 2. For all other sample sizes the optimal lag equals 3, the true lag.

Table 1.0 Small VAR Probability Densities

Pr(y=k | T)		T=50	T=75	T=100	T=125	T=150
AIC	*k=1*	0.09	0.01	0.00	0.00	0.00
	k=2	**0.44**	0.48	0.42	0.31	0.22
	k=3	0.26	**0.40**	**0.50**	**0.63**	**0.72**
	k=4	0.08	0.06	0.05	0.04	0.05
	k=5	0.05	0.03	0.02	0.01	0.01
	k=6	0.08	0.02	0.01	0.01	0.00
LR		0.00	0.00	0.00	0.00	0.00
		0.50	0.50	0.40	0.29	0.21
		0.29	**0.36**	**0.46**	**0.57**	**0.65**
		0.06	0.04	0.04	0.04	0.04
		0.07	0.05	0.06	0.06	0.05
		0.08	0.05	0.05	0.04	0.05
OSF		0.43	0.23	0.13	0.09	0.05
		0.36	0.42	0.41	0.39	0.34
		0.14	**0.23**	**0.31**	**0.38**	**0.42**
		0.04	0.06	0.10	0.09	0.09
		0.02	0.03	0.04	0.03	0.06
		0.00	0.03	0.02	0.03	0.03
SBC		0.73	0.49	0.27	0.12	0.04
		0.26	0.50	0.71	0.87	0.93
		0.01	**0.01**	**0.02**	**0.02**	**0.03**
		0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00	0.00

Turning to the results we observe that overall AIC tends to pick the optimal model with the highest probability and this probability is generally increasing in T. Conversely, owing to its high penalty function, SBC tends to select under parameterised models and this tendency dissipates only as T becomes very large. Overall the LR test using a 5% critical prob. value performs quite well, particularly when one considers that critical values are only valid asymptotically. However, LR and OSF both tend to choose too small a model on average. Although this merely reflects an asymmetry in mse arising from choosing $k<3$ versus $k>3$.

Table 2.0 lists the expected mse (see Step 5 of Section 6.0 for the various criteria divided by the optimal mse (the mse which corresponds to picking the optimal model with probability one). For instance, with 100 observations using the LR statistic will result in forecast errors that are about 1% higher than if the optimal lag length was a known quantity. From a criteria evaluation perspective, these are the most informative statistics. For T=50, OSF clearly dominates with an

inflation factor half as big as LR or AIC. However, for T>50 AIC dominates and its inflation factor is quite small reflecting the high probability with which it chooses the correct model.

Table 2.0 Small VAR Expected Loss

	T=50	T=75	T=100	T=125	T=150
AIC	3.4%	**0.7%**	**0.6%**	**0.4%**	**0.3%**
LR	3.3%	1.1%	1.0%	0.6%	0.5%
OSF	**1.5%**	1.9%	1.7%	1.2%	1.0%
SBC	1.9%	3.4%	2.4%	1.6%	1.1%

To summarise, if the model is small and the there exists a reasonable amount of data, one is likely best off using AIC as it does a very good job and is almost costless to use in terms of computer time. However, if the data set is quite small there appear to be significant benefits to using the more computer intensive OSF criterion.

We now investigate the relative performances of the selection procedures with the more realistic sized VAR. Table 3.0 provides the probability densities, $\overline{\mathbf{V}}$, for the large model. Here we have increased the minimum sample to 75 observations and increased the maximum lag to 7. Again, the **bold** entries represent the optimal lag length for a given T. The true lag length is 4. Hence for T=75 the optimal lag (2) is less than the true lag, but for T>75 the optimal and true lags are the same.

Table 3.0 Large VAR Probability Densities

Pr(y=k | T)		T=75	T=100	T=125	T=150
AIC	*k=1*	0.00	0.00	0.00	0.00
	k=2	**0.00**	0.00	0.00	0.00
	k=3	0.00	0.00	0.00	0.00
	k=4	0.01	**0.65**	**0.93**	**0.98**
	k=5	0.00	0.06	0.05	0.02
	k=6	0.01	0.04	0.01	0.00
	k=7	0.99	0.25	0.02	0.00
LR		0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00
		0.47	**0.70**	**0.77**	**0.82**
		0.10	0.09	0.07	0.06
		0.17	0.08	0.07	0.05
		0.26	0.13	0.09	0.07
OSF		0.33	0.12	0.03	0.01
		0.36	0.22	0.12	0.05
		0.18	0.16	0.14	0.10
		0.13	**0.45**	**0.62**	**0.68**
		0.00	0.04	0.07	0.13
		0.00	0.01	0.01	0.03
		0.00	0.00	0.00	0.01
SBC		0.98	0.92	0.73	0.41
		0.02	0.08	0.24	0.36
		0.00	0.00	0.00	0.02
		0.00	**0.00**	**0.03**	**0.21**
		0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00
		0.00	0.00	0.00	0.00

For T=75, OSF selects the optimal model with by far the highest probability (0.36) whereas the LR test is choosing the true model most consistently (0.47). With its high penalty function, SBC chooses the smallest model possible with probability close to one whereas AIC selects the opposite corner solution (k=7 with prob. 0.99). Clearly these information criteria exhibit strange properties when the model is large relative to the data set. For SBC, this tendency to select too small a model diminishes quite slowly as T increases. Conversely, AIC's bias toward selecting too many parameters diminishes rapidly as the sample size increases. For example, with T=125 it selects the optimal model with probability 0.93.

Table 4.0 provides the ratios of expected and optimal mean squared errors for each selection procedures. From a forecasters perspective it is this expected value that should be of most concern. Here we see that for T<125 OSF dominates by a wide margin. Indeed, with 100 observations (about the most any applied forecaster will have at a quarterly frequency) OSF's mse increase is 42% lower than the next best procedure (LR test). With 75 observations, OSF beats SBC by 26%. In fact this number would likely be substantially higher except for the fact that by coincidence the optimal lag (2) is close to the lag length of 1 that SBC always selects. As T grows, SBC's relative and absolute performance is extremely poor. In this sense, SBC cannot be taken that seriously as a contender. If we ignore SBC then OSF beats LR by a factor of 16!

As the sample sizes becomes large we once again observe AIC dominating. Indeed, owing to the fact that with T=150 it chooses the optimal model with probability 0.98, its expected loss is almost zero.

Table 4.0 Large VAR Expected Loss

	T=75	T=100	T=125	T=150
AIC	64.4%	7.4%	**0.5%**	**0.1%**
LR	27.0%	5.3%	2.1%	1.4%
OSF	**1.7%**	**3.1%**	2.3%	1.9%
SBC	2.3%	10.2%	12.1%	10.6%

With respect to the aforementioned bias correction (see Section 5.0), we have again employed monte carlo techniques to measure the exact finite sample bias and then compared it to the bias correction. It appears that the asymptotic correction represents a reasonable approximation in moderate sized samples. Consequently, the expected loss from using the bias corrected OSF estimator outperforms the uncorrected in most situations. However, when the model is small relative to the sample size it makes little difference.

8.0 Extensions

8.1 Tests of equal forecast accuracy

So far it has been argued that specifically considering out-of-sample information may lead to more accurate forecasting models. Specifically, we have suggested that one should select the lag length that minimises the mse in the evaluation sample. However, one must remain aware that these statistics are only estimates of the population mse and as such are subject to sampling variability, i.e. they are themselves random variables. Consequently, it would be useful to know

whether, for instance, the reduction in mse arising from estimating a VAR(k) as opposed to a VAR(k+1) is statistically significant at some level. The simplest way to test such a hypothesis is to assume that the forecast error process, ε_t, is independent normal with constant variance, that is $\varepsilon_t \sim IN(\mu,\sigma^2)$ so that $\varepsilon_t - \mu \sim IN(0,\sigma^2)$. Under these strict assumptions we get the usual chi-square distribution for the sum of squared forecast errors, i.e. $\sum_{t=1}^{T}(\varepsilon_t - \mu)^2 \sim \chi^2(T)$. Thus with two forecast error processes $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$ we can use a simple F test;

$$(8.1.1) \quad \frac{\sum_{t=1}^{T}(\varepsilon_{i,t} - \mu_i)^2}{\sum_{t=1}^{T}(\varepsilon_{j,t} - \mu_j)^2} \sim F(T,T)$$

Unfortunately, the assumption of independence is highly unrealistic in this context and consequently the F test is of little use in practice (see for instance Howrey, Klein and McCarthy (1974)). Many alternative (and more complicated) tests have been suggested (see Diebold and Mariano (1995) for a brief review). Unfortunately, many of these tests are invalidated, even asymptotically, by nested models and are therefore useless in this particular application. Recently, Clark and McCracken (1999) and McCracken (1999) proposed an F-style test of equal forecast accuracy that is well suited to nested hypotheses.

The addition of a reasonably efficient test of equal forecast accuracy will no doubt lead to further improvements to the already promising results presented in this paper. We hope to have one of these tests coded into the optimisation algorithm in the near future.

8.2 Encompassing and Granger causality

Recently it has been argued (see Diebold and Mariano(1995) and Ashley, Granger, and Schmalensee (1980)) that out-of-sample forecast comparisons represent a more stringent test of Granger causality than does the standard in-sample F-test. Recall that the simplest test of Granger causality is $H_0 : B_1 = B_2 = \ldots = B_n = 0$ in the regression

$$(8.2.1) \quad y_t = A(L)y_{t-1} + B(L)x_{t-1} + u_t$$

The out-of-sample alternative is to calculate the mean squared forecast errors under the null (restricted) and alternative (unrestricted) and then test the null that they are not statistically different. Rejection of the null would indicate a Granger relationship going from x to y.

The concept of forecast encompassing refers to the ability of a given model's forecasts help to predict the forecast errors of a competing model, but not visa versa (see Mizon and Richard (1986) and Marquez and Ericson (1990)). Given two models, A and B, the standard test[22] of encompassing is given as;

(8.2.2) $u_t^A = \alpha + \beta f^B_t + \varepsilon_t$ and $u_t^B = \varphi + \lambda f^A_t + \upsilon_t$

where f_t and u_t are respectively the forecasts and forecast errors from the two models. If $\beta \neq 0$ but $\lambda = 0$ (using OLS) then model B is said to encompass model A. With a nested model, this can be thought of as alternative to the out-of-sample Granger causality test described above.

9.0 Conclusion

This paper has two main goals. First, we have attempted to provide an intuitive explanation for why unrestricted econometric models often perform poorly as forecasting devices. This stems from the number of free parameters relative to the sample size embodied in such models. As a consequence of this the confidence bands around forecasts are implausibly large. This in turn makes definitive statements regarding ranges of outcomes very difficult.

Second, we demonstrate that it is possible to improve the forecasting ability of a model by explicitly considering its out-of-sample mse characteristics when choosing lags. This theme is formalised and an out-of-sample forecast (OSF) model selection strategy is developed. Using monte carlo methods, we show that OSF produces lower expected loss, compared to three popular criteria, when the model is of a reasonable dimension and the sample size is less than 125 quarterly observations.

[22] There are now a number of related tests. However, this is the most prevalent and easy to implement test

References

Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principal. 2nd International Symposium on Information Theory. Edited by B.N. Petrov and F. Csaki. Budapest. pp. 267-281

Akaike, H. (1974) A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control AC-19. Pp. 716-723

Ashley, R., C.W.J. Granger, and R. Schmalensee. Advertising and Aggregate Consumption: An analysis of Causality. Econometrica, v.48, no.5 (July 1980). pp. 1149-67

Clark, Todd E., and M.W. McCracken. Tests of Equal Forecast Accuarcy and Encompassing for Nested Models. Unpublished Manuscript (April 1999)

Diebold, F.X., and R.S. Mariano. Comparing Predictive Accuracy. Journal of Business & Economic Statistics, July 1995, Vol.13. No. 3

Howrey, E.P., Klein, L.R., and M.D. McCarthy. Notes on Testing the Predictive Performance of Econometric Models. International Economic Review, 15, 1974. pp. 366-383

Lütkepohl, H. (1991) Introduction to Multiple Time Series Analysis 2nd Ed. Springer-Verlag New York

Lütkepohl, H. (1985) Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process. Journal of Time Series Analysis, Volume 6, No. 1

Marquez, J., and N.R. Ericsson. Evaluating the Predictive Performance of Trade-account Models. Board of Governors of the Federal Reserve System, International Finance Discussion Paper, No. 377, 1990.

McCracken, M.W. Asymptotics for Out-of-Sample Tests of Causality. Manuscript, Louisiana State University, January 1999.

Mizon, G.E., and J.F. Richard. The Encompassing Principal and Its Application to Testing Non-nested Hypotheses. Econometrica, Vol. 54, 1986. Pp. 657-678

Quinn, B.G. Order Determination for a Multivariate Autoregression. Journal of the Royal Statistical Society, B42 ,1980. pp. 182-185

Schwarz, G. Estimating the Dimension of a Model. The Annals of Statistics, 6 , 1978. pp. 461-464

Sims, C. Macroeconomics and Reality. Econometrica, 48, 1980. pp. 1-48