JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/j.1467-6419.2011.00704.x

USING MODEL SELECTION ALGORITHMS TO OBTAIN RELIABLE COEFFICIENT ESTIMATES

Jennifer L. Castle
Oxford University

Xiaochuan Qin
University of Colorado

W. Robert Reed
University of Canterbury

Abstract. This review surveys a number of common model selection algorithms (MSAs), discusses how they relate to each other and identifies factors that explain their relative performances. At the heart of MSA performance is the trade-off between type I and type II errors. Some relevant variables will be mistakenly excluded, and some irrelevant variables will be retained by chance. A successful MSA will find the optimal trade-off between the two types of errors for a given data environment. Whether a given MSA will be successful in a given environment depends on the relative costs of these two types of errors. We use Monte Carlo experimentation to illustrate these issues. We confirm that no MSA does best in all circumstances. Even the worst MSA in terms of overall performance – the strategy of including all candidate variables – sometimes performs best (viz., when all candidate variables are relevant). We also show how (1) the ratio of relevant to total candidate variables and (2) data-generating process noise affect relative MSA performance. Finally, we discuss a number of issues complicating the task of MSAs in producing reliable coefficient estimates.

Keywords. AIC, AICc; *Autometrics*; Bayesian model averaging; General-to-specific modelling; Information criteria; Model selection algorithms; Monte Carlo analysis; Portfolio models; SIC; SICc

1. Introduction

When modelling economic phenomena there is a great deal of uncertainty regarding which variables to include, what functional form is appropriate, what lag length captures dynamic responses, whether there are non-stationarities such as unit roots or structural breaks, etc. Economic theory informs the model specification, but there are aspects that must be data based.

In practice, many empirical papers report results based on an *ad hoc* selection procedure, trying many specifications and selecting the 'best'. Without some objective model selection algorithm, non-systematic efforts may, at best, innocently miss superior specifications, or, at worst, strategically select results to support the researcher's preconceived biases. A substantial literature demonstrates that model selection matters. For example, many studies of economic growth find that results that are economically and statistically significant in one study are not robust to alternative specifications (cf. Levine and Renelt, 1992; Fernandez *et al.*, 2001; Hendry and Krolzig 2004; Hoover and Perez; 2004; Sala-i-Martin *et al.*, 2004). For these and related reasons, there is interest in automated model selection algorithms (MSAs) that can point researchers to the best model specification (Oxley, 1995; Phillips, 2005).

MSAs are designed with different goals in mind. These include selecting a model or models that (1) best represent the true data-generating process (DGP), (2) have desirable inference properties and (3) are best able to forecast out-of-sample observations. The main focus of this review is the estimation of model coefficients. We do not address estimation of coefficient standard errors. As is well known, procedures that produce accurate coefficient estimates do not necessarily produce accurate standard errors (Reed and Ye, 2011). Accordingly, our paper focuses on the performance of MSAs with respect to producing reliable coefficient estimates. We restrict our review to algorithms that can be easily automated to ensure transparency and replicability.

Although the list of MSAs available for use in applied work is large, there are few studies that compare MSA performance. This review provides a conceptual framework for comparing different types of MSAs. We then conduct an empirical review of these MSAs in a simple data environment to illustrate determinants of relative performance associated with coefficient estimation. Monte Carlo simulation is employed because MSAs are often complex and not amenable to theoretical analysis, especially with respect to their finite sample properties. In so doing, we address Owen's (2003, p. 622) call for evidence on the head-to-head performance of rival model selection methods. We then identify some challenging issues for MSAs that have been only partially addressed in the literature.

2. Competing Model Selection Algorithms

2.1 *The Data-Generating Process*

The framework that is commonly taken as a starting point for MSAs is one in which there is a well-defined DGP that can be consistently estimated.[1] The researcher has a set of J candidate variables ($\boldsymbol{x} = x_1, x_2, \ldots, x_J$) from which a model or models are selected. The DGP is given by:

$$y_t = \gamma + \sum_{i=0}^{P} \sum_{j=1}^{J} \beta_{ij} x_{j,t-i} + \varepsilon_t, \quad t = 1, 2, \ldots, T \tag{1}$$

where there are $L = PJ$ regressors (excluding the intercept) with the $x_{j,t-i}$ including lagged and nonlinear transformations of regressors such as interaction terms and polynomials.[2] A subset K of the regressors are 'relevant', defined by non-zero β's, with the remaining L-K coefficients equal to zero, $0 \leq \mathrm{K} \leq \mathrm{L}$. The ε_t are independently and identically distributed (i.i.d.), with $\varepsilon_t \sim \mathrm{IN}(0, \sigma^2)$.

This review excludes selection algorithms over models that are nonlinear in the parameters, or that have non-spherical errors, to focus on standard estimation procedures. Further, we assume that many of the key difficulties in modelling are known. Namely, we assume that the data are accurately measured over the sample period, the initial specification nests the DGP, the parameters are constant, there are no unmodelled structural breaks and conditioning on the set of regressors $\boldsymbol{x}$ is valid.

Given the setup described above, there are 2^L possible variable combinations, each constituting a separate model. The researcher's task is to choose the model, or models, that produce the most reliable coefficient estimates.[3] This leads to the large literature on automated MSAs.

2.2 *Consistency and Asymptotic Efficiency*

Goodness of fit and model selection are closely related. However, there are well-known pitfalls associated with choosing models based solely on goodness of fit (Dhrymes, 1970; Pesaran, 1974). For example, an MSA that chooses the model with the highest R^2 value will always select the specification containing all L variables.

Two widely employed properties for evaluating model selection are (1) consistency (Hannan, 1980) and (2) asymptotic efficiency (Shibata, 1980, 1981). Consider the case when the true model is one of the candidate models being evaluated by the MSA. An MSA is said to be consistent if it chooses the true model with probability one as the sample size increases to infinity. Alternatively, suppose the true model is not among the set of candidate models. An MSA is said to be asymptotically efficient if it selects the model having the smallest expected prediction error with probability approaching one as the sample size increases (Kuha, 2004). The two criteria correspond to different objectives.

Both criteria are asymptotic, and the finite sample behaviour of MSAs may differ significantly from their asymptotic optimality properties. The preferred asymptotic criterion will depend on the researcher's view of the DGP. If the DGP is thought to be infinitely complicated or comprises latent variables, then efficiency would be preferred (the concept was introduced in the context of infinite autoregressive processes, Shibata, 1980, later extended to linear regression models, Shibata, 1981). In contrast, if the DGP is thought to comprise observables that are nested within the model then consistency should be the chosen criterion (McQuarrie and Tsai, 1998).

2.3 *SIC, AIC and Related Information Criteria MSAs*

Two MSAs that receive considerable attention are based on information criteria (IC): the Schwarz information criterion (*SIC*, Schwarz, 1978) and the Akaike information criterion (*AIC*, Akaike 1973). Both the *AIC* and the *SIC* have the same general form: $ln(\hat{\sigma}^2) + Penalty$, where $\hat{\sigma}^2$ is the maximum-likelihood estimate of the variance of the error term for a given specification, and *Penalty* is a function that monotonically increases in the number of coefficients to be estimated.

If we assume that: (1) there are no data measurement errors; (2) the set of L regressors nests the DGP specification, including any nonlinear and interaction effects; (3) the parameters of the DGP are constant and there are no unmodelled structural breaks and (4) conditioning on the set of regressors $\boldsymbol{x}$ is valid, then the *SIC* and *AIC* are consistent and asymptotically efficient, respectively. Assumption (2) is fundamental here; if the DGP is infinite dimensional then *AIC* provides an asymptotically efficient selection of a finite dimensional approximating model.

It is well known that both the *SIC* and *AIC* tend to 'overfit' (i.e. include more variables than the DGP) in small samples. As a result, small-sample corrections for these have been developed by Hurvich and Tsai (1989) and McQuarrie (1999). These are denoted by *SICc* and *AICc*, respectively. These corrections adjust the penalty functions to include an additional argument for sample size, correcting the second-order bias. They are asymptotically equivalent to their uncorrected namesakes.

A number of other IC MSAs are related to either the *SIC* or the *AIC*. These also follow the same general form: $ln(\hat{\sigma}^2)+ Penalty$. Hannan and Quinn's (1979) *HQ* IC was developed as a consistent model selection criterion in response to *AIC*. *HQ* is asymptotically equivalent to *SIC*, though Monte Carlo experimentation by Hannan and Quinn (1979) suggests that *HQ* performs better than *SIC* in large samples when selecting the order for an autoregressive model. The key difference with *SIC* is that the penalty function decreases faster, resulting in the minimum rate at which additional parameters must be penalized in order to still ensure consistency.

Akaike's final prediction error criterion (Akaike, 1969), which preceded Akaike's *AIC*, computes the mean square prediction error when a model fitted to in-sample data is fitted to another independent observation. The model within the candidate set which has the smallest prediction error is chosen. If the objective of modelling is not prediction, then *AIC* (an in-sample criterion) is preferred. Similarly, Mallows' *Cp* (Mallows, 1973) uses a penalized mean square prediction error, and is asymptotically equivalent to the *AIC*. The *Cp* criterion is often used as a stopping rule for stepwise regression. The key difference between goodness-of-fit measures and ICs is that the latter measure the distance between the selected model and the true model using the Kullback–Leibler distance. As the adjusted *R*-squared criterion does not assume a 'true model' to compare to the selected model, it is neither consistent nor asymptotically efficient, and is therefore not asymptotically related to either the *SIC* or the *AIC*.

One noteworthy variant of IC MSAs is the informational complexity criterion of Bozdogan (2000). Like other IC MSAs, informational complexity includes a

goodness-of-fit element and a function that penalizes the inclusion of additional parameters. However, it also adds a third component that takes into account interdependencies of model parameter estimates and dependencies of model residuals. There are yet other variants of IC MSAs. For further discussions, see Gouriéroux and Monfort, (1995, section 2.3), Amemiya (1980), Chow (1981) and Phillips (1994, 1995).

Model selection using the *AIC*, *SIC* and related IC MSAs consists of estimating all possible models and then choosing the single best model (for example, the single model with the smallest IC value). The estimated coefficients from this 'single best model' then become the 'final' coefficient estimates for use in policy analysis. If a variable is not included in the single best model, then the associated coefficient is 'estimated' to be zero.

2.4 *Portfolio Selection*

All of the MSAs above involve selection of a single best model based on a sample IC value. However, these sample IC values are themselves random variables. Under certain conditions, the distribution of these measures can be calculated. This has led some researchers to advocate choosing a set of models, rather than a single best model. For example, Mallows (1973) advocates plotting the *Cp* measure for individual models against the number of explanatory variables in the model to choose a best subset of models.

Poskitt and Tremayne (1987) derive a measure based on the posterior odds ratio, $\Re_{\mathrm{m}} = exp[-\frac{1}{2}(IC_{min} - IC_m)]$, where IC_{min} is the minimum *IC* value among all 2^L models, and IC_m is the value of the respective *IC* in model m, $m = 1,2,\ldots,$ 2^L. They argue that a $\Re_{\mathrm{m}}$value greater than 100 is decisive evidence that the competing model should be discarded. If $\sqrt{10} < \Re_{\mathrm{m}} \leq 10$, there is 'no substantial evidence' in favour of the model minimizing the IC. And if $1 < \Re_{\mathrm{m}} < \sqrt{10}$, then the alternative model is said to be a 'close competitor' to the IC-minimizing model.[4] They suggest forming a portfolio of all models having $\Re_{\mathrm{m}} \leq \sqrt{10}$. Jeffreys (1961, p. 432) notes that $\Re_{\mathrm{m}}$ is used to grade the decisiveness of the evidence and has no physical meaning. Hence, the intervals are rules of thumb rather than based on any optimality properties.

Burnham and Anderson (2004) present a somewhat different set of recommendations. They categorize models as follows: (1) $\Re_{\mathrm{m}} < 2$ indicates that the competing model has 'substantial support'; (2) $4 < \Re_{\mathrm{m}} < 7$ indicates that the model has 'considerably less support' and (3) $\Re_{\mathrm{m}} > 10$ indicates that the model has 'no support'. These rough guidelines have similar counterparts in the Bayesian literature (e.g. Raftery, 1996). Less clear is how the respective models should be combined to obtain a single coefficient estimate.

2.5 *Path Reduction MSAs*

One problem with the previous MSAs is that they require all possible models to be estimated. When the number of candidate variables is large, this becomes

computationally unfeasible. This has led to MSAs that use various strategies to reduce the number of models to be compared. Four very common path reduction MSAs are backward selection, backward stepwise, forward selection and forward-stepwise (FW) model searches. Backward (forward) selection MSAs work by sequentially dropping (adding) variables one by one according to a specified significance criterion. Backward-stepwise (forward-stepwise) MSAs allow previously discarded variables to be added back into the model (previously included variables to be removed from the model). The cost of not estimating all paths is that superior specifications may be undetected.

An alternative type of path reduction strategy consists of dividing the set of all possible models into various subsets. By judiciously constructing the subsets, one can avoid estimating large swaths of the regression tree and still obtain the optimal IC model (such as *SIC* and *AIC*). These MSAs are commonly called 'branch and bound' MSAs (Hocking and Leslie, 1967; Gatu and Kontoghiorghes, 2006). Other algorithms proposed to undertake exhaustive searches include Schatzoff, Tsao and Fienberg (1968) and Furnival (1971).

The logic of these path reduction strategies can best be illustrated if we think in terms of R^2. Suppose there are 10 candidate variables and we wish to find the three-variable model with the highest R^2. An inefficient strategy is to estimate all 120 possible, three-variable models. However, if we compare the model $\{1,2,3\}$ and find that it has a higher R^2 than model $\{1,2,4,5,6,7,8,9,10\}$, then we know that model $\{1,2,3\}$ has a higher R^2 than the models $\{1,2,4\}$, $\{1,2,5\}, \ldots \{1,2,10\}$. Thus, judicious selection of models with more than three variables can reduce the number of three-variable models that need to be searched. Although this example is in terms of R^2, the logic applies directly to searching for models with minimum IC values. Unlike the backward and forward MSAs described above, branch and bound MSAs are able to achieve the best IC model without estimating all possible models.

Yet another variant of a path reduction strategy is general-to-specific model selection. This technique, which simplifies a general model that captures the salient characteristics of the data, has a long history and has been known as the LSE approach due to its proponents at the London School of Economics in the 1960s. Hendry (2003) discusses the origins of the LSE methodology and Mizon (1995) provides a history. See *inter alia*, Anderson (1962), Pagan (1987), Phillips (1988) and Campos *et al.* (2005) for reviews.

The latest generation of general-to-specific automatic model selection is embodied in *Autometrics* within the software package PcGive (Doornik, 2007, 2009a; Hendry and Doornik, 2009). *Autometrics* undertakes a multi-path tree search, commencing from the general model with all potential variables. It eliminates insignificant variables while ensuring a set of pre-specified diagnostic tests are satisfied in the reduction procedure by checking the subsequent reductions with encompassing tests.

As noted above, the desirable properties of IC such as the *AIC* and *SIC* require the validity of a number of assumptions. If these assumptions are not satisfied, these MSAs will select misspecified models. *Autometrics* refines the path reduction

algorithm by eliminating branches of the regression tree that violate underlying assumptions of the DGP (e.g. non-spherical errors). Multi-path reductions are undertaken to avoid path dependence and either a single best model is found or competing models are retained. The latter are then evaluated using encompassing tests (Doornik, 2008) to result in a final model. If a variable is not included in the single best model, then the associated coefficient is estimated to be zero.

2.6 *Bayesian Model Averaging*

Bayesian model averaging (BMA) MSAs employ a different conceptual framework than MSAs that select a single best model or portfolio of models, see, for example, Hoeting *et al.* (1999). Rather than assuming each model is 'true', and then comparing model diagnostics (such as IC) to select the best model or models, BMA estimates all models, attaching a posterior probability that any given model is the DGP. The final coefficient estimate for a given variable is calculated using a weighted average of individual coefficient estimates for that variable across all models, with individual coefficient estimates being weighted by their posterior model probabilities.

Strictly speaking, BMA is not a model selection tool; it is an estimation method. The appeal of BMA MSAs is that they are claimed to directly address model uncertainty by basing estimates on a weighted average over the model space, which accounts for uncertainty in both predictions and parameter estimates, see Hoeting (2002). Bayesian models require the specification of prior model probabilities, as well as prior distributions for the parameters.

A drawback of BMA models is that – like IC MSAs – they require estimation of all models. In practice, sophisticated sampling algorithms are employed to make BMA MSAs computationally feasible, (e.g. Raftery *et al.*, 1997; George and Foster, 2000, who explore the space of models stochastically via a Markov chain Monte Carlo). The end result is selection over a large subset – but not all – possible models. The individual models in this subset are given weights that sum to one over the subset. Coefficients for variables that do not appear in a given model are set equal to zero. Final coefficient estimates consist of a weighted average of zero and estimated coefficients.[5]

The extreme bounds literature of Leamer (1978, 1983, 1985) is a form of Bayesian analysis but requires a great deal of prior information to be assumed known. See McAleer *et al.* (1985) and Breusch (1990) for criticisms.

3. Performance in Finite Samples

Although properties such as consistency and asymptotic efficiency are conceptually useful, it is unclear how these properties map over to finite sample performance. There are many examples of estimators with desirable asymptotic properties being dominated by asymptotically inferior estimators in finite samples (e.g. the 'shrinkage principle', Diebold, 2007).

Interacting with sample size is the noisiness of the DGP via the variance of the error term. This introduces two kinds of bias. In all of the MSAs above, a better fit results in a higher probability of a model being selected, *ceteris paribus*. Spurious correlations will enhance a model's explanatory power, and thus the likelihood that it is selected. This results in coefficients being biased away from zero. On the other hand, setting estimated coefficient values to zero when a variable does not appear in a model biases coefficient estimates towards zero. It is not clear how these two biases balance out in finite samples.

There are numerous ways to measure the sample performance of MSAs, and the measure will necessarily depend on the modelling purpose. For example, a model may be assessed on its out-of-sample forecasting performance if it is intended to be used for forecasting, but this is a poor measure if the model is being used to test an economic theory (cf. Clements and Hendry, 2005). Castle *et al.* (2011) provide a range of possible performance measures. Some common measures of MSA performance for in-sample model selection include:

(1) Frequency of retaining the DGP,
(2) Retention rate of relevant variables, denoted Potency,
(3) Retention rate of irrelevant variables, denoted Gauge,
(4) Unconditional mean square error (UMSE) and
(5) Conditional mean square error (CMSE).

Let us suppose we are using Monte Carlo experiments to evaluate the performance of a given MSA, with $m = 1, \ldots, M$ replications. Further, suppose there are L ($=PJ$, see equation 1) total candidate variables: K are relevant (i.e. non-zero β's in the DGP); L-K are irrelevant; and let the variables be ordered so that the first K are relevant.[6] The first measure, frequency of retaining the DGP, counts the number of times the MSA chooses the DGP. A deficiency of this measure for our purposes is that it does not directly assess the accuracy of coefficient estimates. A further deficiency is that this can be an unreliable measure of MSA performance when the number of candidate models is large and there is a substantial degree of DGP noise (McQuarrie and Tsai, 1998).

'Potency' and 'Gauge' calculate average retention rates over relevant and irrelevant variables, respectively. Define the retention rate for a given variable i across all M replications as $\tilde{p}_i$: $\tilde{p}_i = \frac{1}{M}\sum_{m=1}^{M} 1(\tilde{\beta}_{i,m} \neq 0)$, $i = 1, \ldots, L$, where 1(.) denotes the indicator function. Then

$$Potency = \frac{1}{K}\sum_{i=1}^{K} \tilde{p}_i \text{and} \tag{2}$$

$$Gauge = \frac{1}{L-K}\sum_{i=K+1}^{L} \tilde{p}_i \tag{3}$$

Although potency and gauge are useful measures of the ability of MSAs to keep and omit the appropriate variables, they also are crude measures of coefficient accuracy. For example, an MSA may select relevant variables whose coefficients

are far from the true values, and may omit irrelevant variables whose estimated values are close to zero.

Denote $\tilde{\beta}_{i,m}$ as the ordinary least squares coefficient estimate associated with variable i in replication m as determined by a given MSA. For a given variable coefficient, UMSE and CMSE are calculated as:

$$UMSE_i = \frac{1}{M}\sum_{m=1}^{M}(\tilde{\beta}_{i,m} - \beta_i)^2 \tag{4}$$

$$CMSE_i = \frac{\sum_{m=1}^{M}(\tilde{\beta}_{i,m} - \beta_i)^2.1(\tilde{\beta}_{i,m} \neq 0)}{\sum_{m=1}^{M}1(\tilde{\beta}_{i,m} \neq 0)} \tag{5}$$

where $i = 1,2,\ldots,L$.[7] Note that both $UMSE_i$ and $CMSE_i$ set $\tilde{\beta}_{i,m} = 0$ when variable i is not included in the selected model.

There is some dispute whether CMSEs or UMSEs are preferable. Much of the literature focuses on UMSE, although consideration of the set of retained variables is closer to what is observed in empirical applications. MSEs are often used as a measure of MSA performance because they coincide with a key goal of estimation: that of producing accurate coefficient estimates. Other performance measures, such as predictive efficiency, may accept biased estimates of individual coefficients as long as accurate predictions are produced.[8] Another argument in favour of using UMSE is that it can be decomposed into (1) bias and (2) variance components, which are in turn related to type I and type II errors. As noted above, $UMSE_i$ and $CMSE_i$ are specific for a given MSA and variable i. In general, it is not meaningful to sum or average $UMSE_i$ and $CMSE_i$ across variables. This is a problem if our goal is to have a summary measure of MSA performance. We revisit this problem below

3.1 *Type I/Type II Errors*

At the heart of MSA performance is the trade-off between type I and type II errors. Some relevant variables will be mistakenly excluded, and some irrelevant variables will be mistakenly retained. Note that these outcomes map onto the measures of *Potency* and *Gauge* above. A successful MSA will find the optimal – as defined by the respective performance measure above – trade-off between the two types of errors for a given data environment.

IC model selection procedures explicitly define a penalty function that penalizes the inclusion of additional variables. In turn, the penalty function can be mapped into an implicit significance level, which measures the rejection frequency per candidate variable (Campos *et al.*, 2003). Thus, MSAs that allow the user to explicitly set the significance level, or IC MSAs in which the significance level can be inferred, can be advantageous when the modeller has a loss function that dictates their preferred type I/II trade-off, and hence their preferred penalty function.

It is straightforward to reduce type II error by using a sufficiently tight penalty function. If there are 100 irrelevant variables, a penalty function that maps to a significance level of 1% would result in only 1 irrelevant variable being retained, on average. Retaining relevant variables depends on the amount of signal relative to noise. If non-centralities are high, i.e. the variables are highly significant, then a tight significance level will not be too costly. We illustrate this below.

It follows that, in general, MSAs with tighter penalty functions/lower significance levels will perform well when there are many irrelevant variables and the relevant variables have high non-centralities. In contrast, MSAs with looser penalty functions/higher significance levels will perform better when there are few irrelevant variables and the non-centralities of the relevant variables are small. Of course, specific results will depend on the performance measure used.

4. Monte Carlo Comparison of Competing MSAs

4.1 *The Model Selection Algorithms*

Having discussed in general how the different MSAs relate to each other, we now engage in an examination of their relative performances in finite samples. We use UMSE as our measure of MSA performance because (1) it directly assesses the accuracy of coefficient estimates, (2) it allows interpretation in terms of bias and variance and (3) it acknowledges that accurate coefficient estimation for irrelevant variables may also be important to policy makers. As the previous discussion makes clear, there are virtually an infinite number of possible MSAs. We study 21 different MSAs, selecting representatives from each of the different categories defined above. These are listed in Table 1, along with a brief description.

The first four MSAs are *IC* based: *AIC*, *AICc*, *SIC* and *SICc*. The extra 'C' indicates that the respective MSA is the small-sample corrected version of its namesake. As is clear from Table 1, these all have the general form: $ln(\hat{\sigma}^2)+$ *Penalty*, and differ only in their penalty functions. The *SICc* imposes the harshest penalty for the inclusion of additional variables, followed by the *SIC*, *AICc* and *AIC*. Each MSA chooses the specification with the smallest IC sample value.

Our procedure for identifying the best model consists of calculating all 2^L possible models.[9] Coefficient estimates are taken from the model with the lowest *IC* value. If a variable does not appear in that model, then the associated estimate of that coefficient is set equal to zero.

The next eight MSAs are based on the idea of selecting – not a single best model – but a 'portfolio' of models that are all 'close' as measured by their *IC* values. Poskitt and Tremayne (1987) derive a measure based on the posterior odds ratio, $\Re_m = exp[-\frac{1}{2}(IC_{min} - IC_m)]$, where $IC_{\min}$ is the minimum *IC* value among all 2^L models, and IC_m is the value of the respective *IC* in model m, $m = 1,2,\ldots,2^L$. They suggest forming a portfolio of models all having $\Re_{\mathrm{m}} \leq \sqrt{10}$. Alternatively, Burnham and Anderson (2004) suggest a threshold $\Re_{\mathrm{m}}$ value of 2.

Our procedure estimates all 2^L possible models. The MSAs $AIC < 2$, $AICc < 2$, $SIC < 2$ and $SICc < 2$ each construct portfolios of models that have *AIC*, *AICc*, *SIC*

Table 1. Description of Model Selection Algorithms (MSAs).

Information Criterion (IC) Algorithms:			
(1)	*AIC*	$AIC = ln(\hat{\sigma}^2) + \frac{2(\tilde{K}+1)}{T}$	$\hat{\beta}_k$ is the estimate of β_k in the model with the minimum IC value. If X_k does not appear in that model, $\hat{\beta}_k = 0$. NOTE: $\hat{\sigma}^2$ is the maximum likelihood estimate of the variance of the error term; $\tilde{K}$ is the number of coefficients in the model excluding the intercept; and T is the number of observations.
(2)	*AICc*	$AICc = ln(\hat{\sigma}^2) + \frac{(T+\tilde{K}+1)}{(T-\tilde{K}-3)}$	
(3)	*SIC*	$SIC = ln(\hat{\sigma}^2) + \frac{(\tilde{K}+1)\cdot ln(T)}{T}$	
(4)	*SICc*	$SICc = ln(\hat{\sigma}^2) + \frac{(\tilde{K}+1)\cdot ln(T)}{(T-\tilde{K}-3)}$	
Portfolio algorithms:			
(5)	$AIC < 2$	$\hat{\beta}_k$ is the average value of β_k estimates from the portfolio of models that lie within a distance $\Re = 2$ of the respective minimum IC model, where $\Re_m = exp[-\frac{1}{2}(IC_{min} - IC_m)]$, IC_{min} is the minimum IC value among all 2^L models, and IC_m is the value of the respective IC in model m, $m = 1,2,\ldots,2^L$. If X_k does not appear in any of the portfolio models, $\hat{\beta}_k = 0$.	
(6)	$AICc < 2$		
(7)	$SIC < 2$		
(8)	$SICc < 2$		
(9)	$AIC < \sqrt{10}$	Same as above, except $\Re = \sqrt{10}$.	
(10)	$AICc < \sqrt{10}$		
(11)	$SIC < \sqrt{10}$		
(12)	$SICc < \sqrt{10}$		

Journal of Economic Surveys (2011) Vol. 00, No. 0, pp. 1–33

Table 1. *Continued.*

		Information Criterion (IC) Algorithms:
General-to-specific regression algorithms (*Autometrics*):		
(13)	*AUTO_1%*	$\hat{\beta}_k$ is the estimate of β_k in the best model as selected by the Autometrics program in PcGive, with the significance level, α, set equal to 1%, 5% and $\frac{1.6}{T^{0.9}} \cdot 100\%$, respectively. If $X_k$does not appear in that model, $\hat{\beta}_k = 0$. $\hat{\beta}_k$ is bias corrected using a two-step procedure.
(14)	*AUTO_5%*	
(15)	*AUTO_Variable*	
Forward-stepwise (FW) regression algorithms		
(16)	*FW_1%*	$\hat{\beta}_k$ is the estimate of β_k in the best model as selected by the FW program in PcGive, with the significance level, α, set equal to 1%, 5% and $\frac{1.6}{T^{0.9}} \cdot 100\%$, respectively. If $X_k$does not appear in that model, $\hat{\beta}_k = 0$.
(17)	*FW_5%*	
(18)	*FW_Variable*	
Bayesian model averaging algorithms:		
(19)	*LLWeighted_All*	$\hat{\beta}_k$ is the weighted average value of β_k estimates over all 2^L models, where model weights are determined according to $\omega_m = \frac{\ell_m}{\sum_{m=1}^{2^L} \ell_m}$, $m = 1,2,\ldots,2^L$, and ℓ is the maximized value of the log likelihood function for model m. For the 2^{L-1} models where $X_k$does not appear in any of the portfolio models, $\hat{\beta}_k = 0$.
(20)	*LLWeighted_Selected*	$\hat{\beta}_k$ is the weighted average value of β_k estimates over the 2^{L-1} models where X_k is included in the regression equation. Model weights are determined according to $\omega_m = \frac{\ell_m}{\sum m \in \{\text{models that contain the variable } X_k\} \ell_m}$.
All variables:		
(21)	*ALLVARS*	$\hat{\beta}_k$ is the estimate of β_k in the specification in which all variables are included.

and *SICc* values that lie within 2 of the minimum value model. The next four MSAs ($AIC < \sqrt{10}$, $AICc < \sqrt{10}$, $SIC < \sqrt{10}$ and $SICc < \sqrt{10}$) do the same for models lying within $\sqrt{10}$ of the respective minimum value model. Coefficient estimates are set equal to zero for variables that never appear in the portfolio. For variables that appear at least once in the portfolio of models, the respective coefficient estimates are calculated as the arithmetic average of all non-zero coefficient estimates.

The next three MSAs use an automated general-to-specific regression algorithm (*AUTO*). These are taken from the *Autometrics* program available in PcGive (Doornik, 2009a). *Autometrics* allows researchers to set their preferred significance level. We select 1% and 5% (*AUTO*_1% and *AUTO*_5%), as these are most common in the applied economics literature. We also allow a variable significance level that adjusts for the number of observations, with a lower significance level being used for larger *T*. We follow Hendry's suggestion (Hendry, 1995, p. 490) and set this variable significance level equal to $\frac{1.6}{T^{0.9}} \cdot 100\%$ (*AUTO_Variable*).[10] All three *Autometrics* MSAs apply bias-correction *ex post* (Hendry and Krolzig, 2005).[11]

Next are three FW algorithms. The particular versions that we employ also come from PcGive and use the same three significance levels as the preceding AUTO algorithms (*FW*_1%, *FW*_5% and *FW_Variable*). Variables are added to the model in order of significance, one at a time, until no further significant regressors are found. If included variables become insignificant as others are added, they are removed from the model. Both the *AUTO* and *FW* algorithms produce a single best model and assign a coefficient estimate of zero to those variables that are not retained in the final model.

The next two MSAs are examples, albeit highly simplified, of BMA (Hoeting *et al.*, 1999). Our procedure estimates all 2^L possible models. A composite model is constructed in which each of the variable coefficients equals a weighted average of individual estimated coefficients for that variable across models. For a given model, the weight is $\omega_m = \frac{\ell_m}{\sum_{m=1}^{2^L} \ell_m}$, $m = 1,2,\ldots,2^L$, where ℓ is the maximized value of the log likelihood function for the regression model from which the coefficient estimate is taken. For the 2^{L-1} models where the variable is excluded, the coefficient estimate is set equal to zero. We analyse two versions: (1) *LLWeighted_All*, which uses the full set of 2^L models to construct weighted average coefficient estimates and (2) *LLWeighted_Selected*, which restricts itself to the set of all 2^{L-1} models where the given variable appears. Note that in both cases the emphasis is on 'model averaging' rather than 'Bayesian', as we do not assign prior subjective values to the coefficients.

The final MSA (*ALLVARS*) selects the full set of potential variables for inclusion in the 'final model'. As should be apparent, the great disparity in approaches underlying these MSAs makes it difficult to analytically compare the performance of all 21 MSAs, and this is all the more true with respect to their performance in finite samples.

As a result, our analysis turns to Monte Carlo experimentation. Our experiments are conducted using a simple simulation design in which the DGP is nested. We assume a static linear model with weakly exogenous, orthogonal regressors,

Table 2. Retention Probabilities as a Function of ψ and α (for $T = 75$).

	$\mathbf{P(t_k \geq c_\alpha \mid E[t_k] = \psi)}$			
$\boldsymbol{\psi_k}$	$\alpha = 50\%$	$\alpha = 20\%$	$\alpha = 5\%$	$\alpha = 1\%$
1	62.6%	38.5%	16.1%	5.0%
2	90.7%	76.0%	50.3%	26.0%
3	99.0%	95.6%	84.3%	63.9%
4	100%	99.7%	97.8%	91.3%
5	100%	100%	99.9%	99.1%
6	100%	100%	100%	100%

constant parameters and spherical error terms. We recognize that this design is not representative of general economic data environments. Among other things, many would argue that it is unrealistic to assume that the DGP lies within the set of models being evaluated. However, it will serve our purpose of illustrating a number of key issues associated with the relative performances of MSAs.

4.2 *Description of Experiments*

The DGP is given by (1), where $P = 1$ (i.e. no lags), $\gamma = 5$, and $\beta_j = 1$, $\forall j = 1, ..., K$. $x_{j,t} \sim IN[0, 1]$ $\forall j$, and are fixed both within and across experiments. $\varepsilon_t \sim IN[0, \sigma^2]$.[12] σ^2 is fixed within an experiment, but variable across experiments. We vary σ^2 across experiments depending on the value we desire for the non-centrality parameter, $\psi \equiv \mathrm{E[t]}$, which is a measure of DGP noise. Specifically, σ^2 is adjusted to produce target values of ψ according to the relationship: [13]

$$\sigma^2 = \frac{T}{\psi^2} \tag{6}$$

Note that ψ is independent of K and L for a given sample size, and represents the expected value of the sample t-statistic for any of the relevant variables. Our experiments let ψ range from 1 to 6.

Table 2 identifies the relationship between ψ, our measure of DGP noise, and the probability of retaining a relevant variable using a single t-test, when the retention decision is determined by the significance level, α. A range of non-centralities and significance levels are reported. For example, a 5% significance level will result in a relevant variable with a non-centrality of 1 being retained 16% of the time. This increases to 50% for $\psi = 2$ and 100% for $\psi = 6$.[14] Although the values vary by T, they change only slightly even when the sample size becomes very large.

Our experiments are designed to allow four factors to vary across experiments: K, L, T and ψ.[15] MSAs that tend to underfit (overfit) will perform relatively well when there are few (many) relevant variables in the DGP. To illustrate this for given L, we run L consecutive experiments where K starts at 1 and progresses through

L. We set *L* equal to 5, 10 and 15. Although larger values would be desirable, we are limited by computational constraints because many of the MSAs require estimation of all possible 2^L models. As some MSAs have established asymptotic properties, we show the effect of increasing sample size by letting *T* take on the values 75, 150, 500 and 1500. Table A1 in the Appendix summarizes the 360 experiments.

As discussed above, we use UMSE of the coefficient estimates to compare MSA performance.[16] Each experiment consists of 1000 simulated data sets/replications *r*. For each replication of each experiment, and for each MSA, we produce a set of *L* coefficient estimates, $(\hat{\beta}_{1,r}^{MSA}, \hat{\beta}_{2,r}^{MSA}, ..., \hat{\beta}_{L,r}^{MSA})$.[17] We aggregate over replications to calculate a *UMSE* value for each coefficient and for each MSA in that experiment: $UMSE_i^{MSA} = \frac{\sum_{r=1}^{1000} (\hat{\beta}_{i,r}^{MSA} - \beta_i)^2}{1000}$, $i = 1,2,\ldots, L$.

It is easily seen that the $UMSE_i$ cannot generally be aggregated across coefficients within an experiment because they depend on the nominal sizes of the coefficients. And they cannot be aggregated across experiments because they depend on the variance of the error term. Accordingly, we assign a ranking from 1 to 21 for each $UMSE_i$, with the MSA producing the lowest *UMSE* for that coefficient receiving a rank of 1, the MSA with the next smallest *UMSE* receiving a rank of 2, and so on. These rankings are then averaged across all *L* coefficients to produce an overall MSA ranking for that experiment. For example, if $L = 5$ and a given MSA has individual coefficient rankings {10, 10, 12, 13, 10}, this MSA would receive an average rank of 11 for that experiment.[18] We then compare experiment-specific, average *UMSE* rankings of MSAs to illustrate how they vary across *K*,*L*,*T* and ψ.

5. Results

Table 3 summarizes the results over all experiments. The columns report mean, median, minimum and maximum rankings for all 360 experiments in ascending order, with the best MSA (as measured by mean rank) listed first.

In terms of overall performance, the top three MSAs, as measured by both mean and median rankings, are the three *Autometrics* MSAs. The best of the three, *AUTO_5%*, has an average ranking a full rank better than its next best, non-*Autometrics* competitor. Portfolio MSAs sometimes perform better than their non-portfolio analogues (cf. $AICc < \sqrt{10}$ and $AICc < 2$ versus *AICc*) and sometimes worse (cf. *SIC* versus $SIC < \sqrt{10}$ and $SIC < 2$). Model averaging over all possible models (*LLWeighted_All*) is generally superior in our experiments to model averaging over only those models in which the respective variable appears (*LLWeighted_Selected*). That being said, there are data environments where *LLWeighted_Selected* does better. The worst-performing MSA is *ALLVARS*. This highlights the fact it is generally not a good idea – as a general strategy – to include all potential variables in a regression specification.

The wide range of minimum and maximum values indicates that no single MSA always performs best, or worst. For example, while *ALLVARS* generally performs

Table 3. Comparison of MSA Performance: All Experiments (Sorted By Mean UMSE Rank in Ascending Order).

MSA	Mean	Median	Minimum	Maximum
AUTO_5%	9.4	9.4	3.7	17.6
AUTO_Variable	9.7	9.2	1.7	21.0
AUTO_1%	9.9	9.2	1.1	21.0
FW_1%	10.6	9.9	2.3	21.0
SIC	10.6	10.6	4.7	18.1
FW_Variable	10.8	9.8	3.4	21.0
SICc	10.9	10.3	4.0	19.2
$SIC < 2$	10.9	10.8	5.8	18.4
FW_5%	11.0	10.7	7.3	20.2
$SICc < 2$	11.1	11.0	5.4	18.8
$AICc < \sqrt{10}$	11.1	11.2	3.4	18.5
$AICc < 2$	11.1	11.5	3.3	15.6
$SIC < \sqrt{10}$	11.2	11.0	6.7	20.0
$AIC < 2$	11.2	11.8	2.6	16.9
LLWeighted_All	11.2	11.9	1.0	16.6
$SICc < \sqrt{10}$	11.3	11.1	5.6	20.1
AICc	11.3	11.2	3.7	19.2
$AIC < \sqrt{10}$	11.4	11.8	3.0	18.5
AIC	11.6	12.1	3.1	19.0
LLWeighted_Selected	11.8	12.5	1.9	19.7
ALLVARS	12.7	14.0	1.0	20.9

poorly, it does better than any other MSA when all the candidate variables are relevant ($K = L$) because the estimated model is the DGP for this specification.[19]

5.1 *Identifying the Determinants of Relative Performance of MSAs*

As noted above, measures of overall performance mask substantial differences between MSAs across different data environments. Table 4 illustrates the important role that the ratio (K/L) plays in determining MSA performance. It compares rankings for *SIC* and *AIC* as K changes, holding L, T and ψ constant (here set equal to $L = 10$, $T = 75$ and $\psi = 2$). Columns 1 and 4 report the average rank (over the 10 coefficients) for each of the respective experiments (where each experiment consists of 1000 replications). Columns 2/3 and 5/6 decompose these into average ranks over irrelevant and relevant variables.

When the number of relevant variables is relatively small, *SIC* outperforms *AIC*. As (K/L) increases, *SIC* monotonically loses ground to *AIC*. When $K = 5$, the relative rankings of the two MSAs switch positions, with *AIC* outperforming *SIC*. Note that average performance within the sets of irrelevant and relevant variables is little affected by increases in (K/L).

Table 4. Experimental Results for the Case: $L = 10$, $T = 75$, $\psi = 2$.

	Mean Ranking of *SIC* Algorithm Over . . .			Mean Ranking of *AIC* Algorithm Over . . .		
Number of Relevant Variables (K)	All Variables (1)	Irrelevant Variables (2)	Relevant Variables (3)	All Variables (4)	Irrelevant Variables (5)	Relevant Variables (6)
1	8.0	7.1	16.0	13.6	14.1	9.0
2	8.9	7.0	16.5	13.2	14.3	9.0
3	9.9	7.1	16.3	12.7	14.3	9.0
4	10.7	7.0	16.3	12.1	14.2	9.0
5	11.7	7.4	16.0	11.4	14.2	8.6
6	12.7	7.5	16.2	10.6	13.8	8.5
7	13.5	7.3	16.1	10.0	14.3	8.1
8	14.6	8.0	16.3	9.4	15.0	8.0
9	15.4	8.0	16.2	8.6	14.0	8.0
10	16.2	–	16.2	8.0	–	8.0

SIC outperforms *AIC* on irrelevant variables (cf. columns 2 and 5). *AIC* outperforms *SIC* on relevant variables (cf. columns 3 and 6). The switch in relative performance occurs because of changes in the weights of these two components. When there are many irrelevant variables and few relevant variables, *SIC*'s advantage on the former causes its overall performance to dominate *AIC*. As K increases, *AIC*'s advantage on relevant variables allows it to overtake *SIC*.

The explanation for *SIC*'s advantage (disadvantage) on irrelevant (relevant) variables is due to the penalty function, because this is the only characteristic that distinguishes the two MSAs. *SIC* has a larger penalty function than *AIC* and therefore selects, on average, fewer irrelevant variables. This will result in smaller bias for the *SIC* specification, because the estimated coefficients for selected, irrelevant variables from the *AIC* MSA will suffer from pre-testing bias. The *SIC* estimates will also be characterized by lower variance, because omitted variables are assigned coefficient values of 0. Of course, *SIC* also admits fewer relevant variables. This biases coefficient estimates of the relevant variables because their population values are non-zero. Therefore, *SIC*'s larger penalty function harms its performance with respect to relevant variables.

Figure 1 illustrates the principle. As noted above, the four IC MSAs can be strictly ordered in terms of the size of their penalty functions: *SICc* > *SIC* > *AICc* > *AIC*. Figure 1 reports the performance results for all 180 experiments where $T = 75$ (cf. Table A1). The vertical axes report MSA rankings (from 1 to 21). The horizontal axes are ordered by K (from 1 to L). There are three columns of figures, corresponding to $L = 5$, 10 and 15; and six rows for ψ from 1 to 6 (with DGP noise greatest for smallest ψ). The four boldfaced lines indicate the rankings for *SICc*/*SIC*/*AICc*/*AIC*, with the dotted lines becoming increasingly

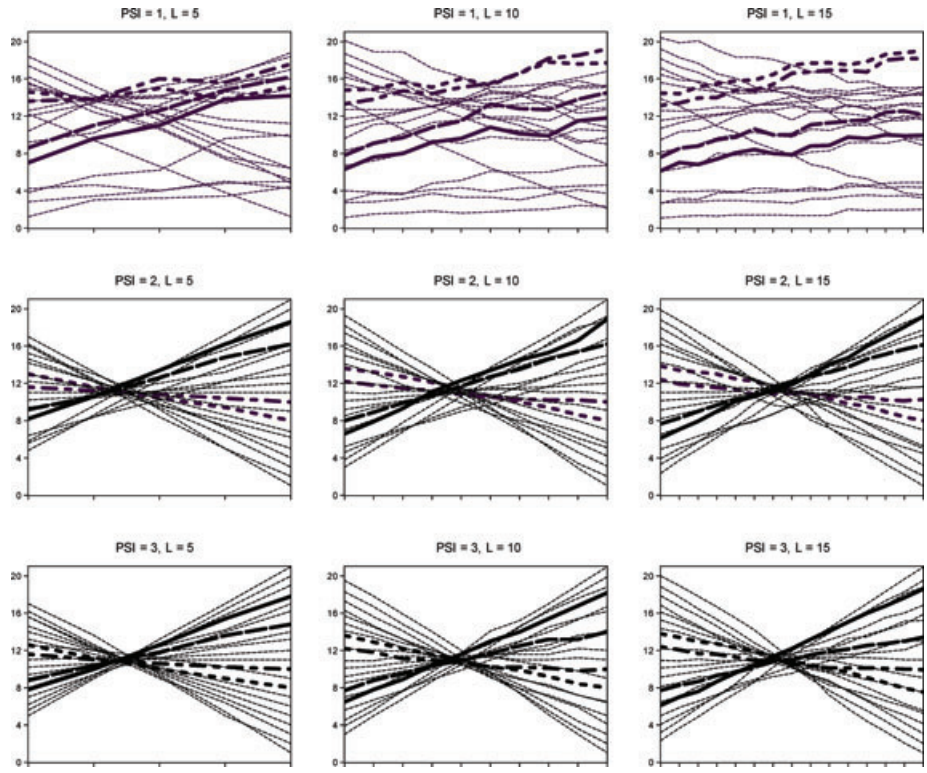

Figure 1. Rankings of MSAs as a Function of K, ψ and L ($T = 75$).

solid for IC with larger penalty functions. The performances of the other seventeen MSAs are indicated by dotted, non-boldfaced lines. Visual inspection indicates that, generally, the MSAs with larger penalty functions do relatively better (have lower rank) when (K/L) is small; and relatively worse when (K/L) is large, except when $\psi = 1$.

Figure 1 also highlights two other results. First, a similar relationship seems to be at work with respect to many of the other MSAs. Second, it is clear that other factors, such as DGP noise, as represented by ψ, also affect relative MSA performance.

We pursue these observations by regressing average experimental ranking as a function of the share of relevant variables (K/L), the degree of DGP noise (ψ) and the number of observations in the data set (T). We estimate separate regressions for each MSA, with 360 observations, one for each experiment.

$$\text{Average experimental ranking}_i^{MSA} = \beta_0 + \beta_1(K/L)_i + \beta_2\psi_i + \beta_3 T_i + \varepsilon_i \quad (7)$$

The results are reported in Table 5. Confirming our visual inspection of Figure 1, we see that the variable (K/L) is statistically significant in 20 of the 21 regressions,

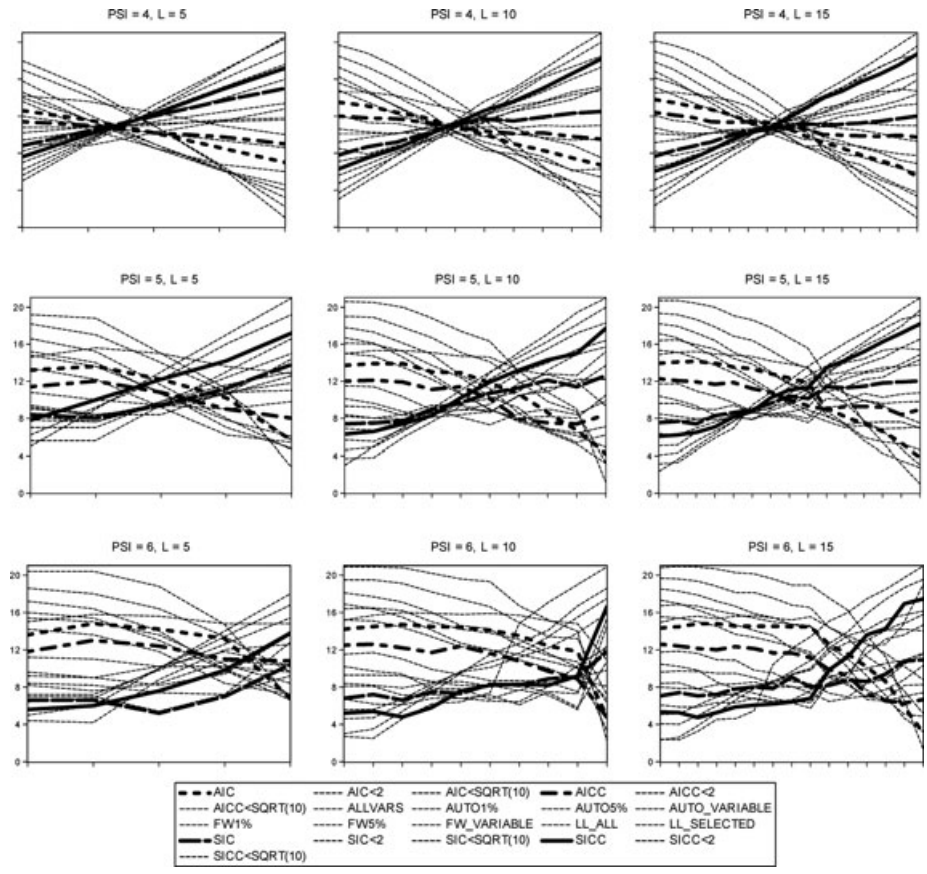

Figure 1. *Continued.*

indicating that the share of relevant variables is an important determinant of relative MSA performance, with the effect being evenly split (not surprisingly) as to whether (K/L) positively or negatively affects relative performance. DGP noise (ψ) is also an important determinant, being significant in 19 of the 21 regressions.

Number of observations in the data set (T) is significant in 10 of the 21 regressions, but the estimated effects are relatively small. The largest estimated effect in absolute value ($= -0.0013$ for the *ALLVARS* MSA), implies that increasing sample size by a 1000 observations improves its relative rank by a little over 1. In contrast, both (K/L) and (ψ) are estimated to have large impacts. Using the average of the absolute values of the coefficients in Table 5, we estimate that increasing the share of relevant variables by 50% causes a 4.6 change in relative rankings, on average. Increasing DGP noise by three causes a 1.8 change in relative rankings, on average.

Table 5. The Relationship Between MSA Ranking and (K/L), ψ and T.

MSA	(K/L)	ψ	T	R-squared
AIC	−5.849	−0.4606	−0.0006	0.425
	(−13.39)	(−5.27)	(−3.09)	
AICc	−3.074	−0.6635	−0.0004	0.378
	(−8.12)	(−8.86)	(−2.37)	
SIC	7.730	−0.6402	−0.0002	0.715
	(24.84)	(−11.05)	(−1.29)	
SICc	11.06	−0.4071	−0.0004	0.698
	(26.75)	(−4.72)	(−1.76)	
AIC < 2	−11.32	0.0065	−0.0003	0.747
	(−31.16)	(0.08)	(−1.99)	
AICc < 2	−8.747	−0.2002	−0.0003	0.716
	(−29.23)	(−3.16)	(−2.42)	
SIC < 2	3.056	−0.8734	0.0002	0.671
	(11.87)	(−18.74)	(1.63)	
SICc < 2	6.245	−0.7500	−0.0001	0.737
	(23.69)	(−14.37)	(−0.55)	
AIC < $\sqrt{10}$	−14.28	0.3083	−0.0001	0.867
	(−48.69)	(4.91)	(−0.85)	
AICc <$\sqrt{10}$	−12.35	0.2120	0.0002	0.871
	(−48.14)	(4.05)	(1.91)	
SIC < $\sqrt{10}$	−0.4820	−0.7342	0.0006	0.445
	(−1.49)	(−12.84)	(3.88)	
SICc < $\sqrt{10}$	2.697	−0.7382	0.0004	0.538
	(8.99)	(−13.56)	(3.59)	
AUTO_1%	15.04	1.133	0.0004	0.726
	(25.74)	(9.93)	(2.12)	
AUTO_5%	7.328	0.8267	0.0010	0.634
	(18.45)	(11.13)	(6.18)	
AUTO_Variable	13.66	1.036	0.0008	0.725
	(25.02)	(10.13)	(3.07)	
FW_1%	13.93	0.0596	0.0002	0.640
	(23.03)	(0.50)	(1.17)	
FW_5%	3.910	−0.7399	0.0006	0.492
	(10.31)	(−12.50)	(2.92)	
FW_Variable	11.98	−0.2686	−0.0001	0.671
	(24.63)	(−2.92)	(−0.37)	
LLWeighted_All	−7.444	1.363	−0.0001	0.833
	(−22.80)	(24.23)	(−0.34)	
LLWeighted_Selected	−15.42	0.8158	−0.0007	0.855
	(−40.97)	(11.37)	(−5.14)	
ALLVARS	−17.67	0.7145	−0.0013	0.7784
	(−32.55)	(6.69)	(−6.80)	

Note: The coefficients in the table are derived from ordinary least squares estimation of the regression equation, $\mathrm{Y_i} = \beta_0 + \beta_1(\mathrm{K/L})_\mathrm{i} + \beta_2\psi_\mathrm{i} + \beta_3\mathrm{T_i} + \varepsilon_\mathrm{i}$, $i = 1,2,\ldots,360$, where the dependent variable is the experiment-specific, rank value for the respective MSA. White-adjusted t-statistics are reported in parentheses below the respective coefficient estimates. We emphasize that each MSA equation was estimated separately, and that no tests for congruency were undertaken for the respective regression equations.

The simple specification of equation (8) will fail to capture complex relationships that may exist between these variables and relative performance. Even so, the three variables are able to explain an impressive amount of the variation in relative rankings. The average *R*-squared across the 21 regressions of Table 5 is 0.674, and the median value 0.715.

Although the results from Table 3 make it clear that no single MSA will dominate in all data environments, the results from Table 5 suggest that there may be certain data environments where one or more MSAs can consistently outperform the others. This raises the possibility that, for practical purposes – that is, for data environments where model selection is likely to be of greatest value to researchers – it may yet be possible to make MSA recommendations.

We can illustrate this through our experiments. For example, one might argue that the data environments where MSAs are most likely to be valuable are where:

(1) The researcher believes, on the basis of *a priori* judgment, that there are many more candidate than relevant variables, making it difficult to decide which ones to select.
(2) There is a substantial degree of DGP noise, so that many variables are on the edge of statistical significance.

In the context of our experiments, let us map these two conditions to (1) $\frac{K}{L} \leq 0.5$ and (2) $\psi \leq 2$. Table 6 analyses MSA performance for the 58 experiments where (1) half or less of the candidate variables are relevant and (2) the sample *t*-statistics for the relevant variables have an expected value of either 1 or 2. Panel A repeats the analysis of Table 3 for the restricted set of 58 experiments. As before, MSAs are ranked in ascending order, with the best performers listed first. The three *Autometrics* MSAs are (again) the top performers, but this time *AUTO*_1% and *AUTO_Variable* are virtually tied for best. Substantially further back (over two full ranks higher), are the two FW algorithms, *FW*_1% and *FW_Variable*. Still further back are the IC MSAs.

Another look at the superior performance of the *Autometrics* MSAs is provided by panel B of Table 6. These results report the frequency at which the respective *Autometrics* MSAs perform as well or better than all other MSAs – where 'as well or better' means that the respective MSA has a rank equal to or lower than all other, non-*Autometrics* MSAs. *AUTO*_1% did at least as well as all other non-*Autometrics* MSAs in 54 out of 58 experiments (93.1%). *AUTO_Variable* did at least as well in 53 of the 58 experiments (91.4%).

These results are suggestive that it may be possible to identify MSAs that dominate in particular data environments. Admittedly, our experimental results assume a rarefied data environment unlikely to be encountered in actual empirical work. Further research in more general data environments could prove useful. The last section of our review discusses some issues that complicate the task that MSAs face in choosing the best model/models.

Table 6. Comparison of MSA Performance: Experiments where $\frac{K}{L} \leq 0.5$ and $\psi \leq 2$ (Sorted in Ascending Order of Mean UMSE Rank).

A. Comparison of UMSE Ranks

MSA	Mean	Median	Minimum	Maximum
AUTO_1%	4.6	4.0	1.1	10.6
AUTO_Variable	4.8	3.8	1.7	10.6
AUTO_5%	6.4	6.3	3.7	9.3
FW_1%	7.0	7.0	2.7	12.4
FW_Variable	7.9	7.9	3.4	12.4
SICc	8.5	8.0	5.0	12.3
SIC	9.2	9.2	5.2	12.1
SICc < 2	10.6	10.4	8.2	14.0
FW_5%	10.9	10.7	8.1	16.3
SIC < 2	11.1	10.8	8.6	14.5
LLWeighted_All	11.6	11.8	7.8	14.2
SICc $< \sqrt{10}$	11.9	11.5	10.5	15.2
SIC $< \sqrt{10}$	12.3	12.1	10.9	15.3
AICc	12.7	12.8	10.8	15.4
AIC	13.5	13.6	10.9	16.5
AICc < 2	13.7	14.0	11.0	15.6
AIC < 2	14.0	14.5	11.0	16.5
AICc $< \sqrt{10}$	14.2	14.2	10.9	18.3
AIC $< \sqrt{10}$	14.8	14.8	10.9	18.1
LLWeighted_Selected	14.9	15.0	10.5	19.2
ALLVARS	16.3	16.8	10.3	20.4

B. Percentage of Experiments where *Autometrics* MSAs Perform as Well or Better Than All Other MSAs

MSA	Percentage
AUTO_1%	93.1
AUTO_Variable	91.4
AUTO_5%	46.6

Note: There are a total of 58 experiments where $\psi \leq 2$and $\frac{K}{L} \leq 0.5$.

6. Complications Facing MSAs

6.1 *Collinearity*

If the L variables were perfectly orthogonal in the sample, many MSAs would perform equally well. Eliminating or adding variables would have no impact on

the estimated coefficients of the other variables in the model. Castle *et al.* (2011) describe the optimal MSA algorithm in this context, where a one-cut decision rule is all that is needed.

Collinearity results in disrupted information attribution, which will increase null rejection frequencies and reduce non-null rejection frequencies. This will affect the variables chosen and the associated coefficient estimates. There is no simple solution to collinearity when variables are highly correlated, such as when variables are alternative measures of the same phenomena. However, there is reason to believe that some MSAs are more likely to perform poorly than others in the face of collinearity. MSAs that do not estimate all possible models are particularly vulnerable when data are collinear. For example, two regressors that are negatively correlated but must be included jointly to be significant would not be detected under a FW MSA.

Because collinearity is not invariant under linear transformations, linear models, which can be defined by various isomorphic representations, can deliver very different inter-correlations. As collinearity is a property of the parameterization of the model, and not the variables *per se*, re-parameterizing the model to a more orthogonal transformation can improve the performance of the MSA, for example, by taking differences.

Ridge regression is often seen as a solution to collinearity (Hoerl and Kennard, 1970a, 1970b). By allowing for some bias in the estimation, the variance of the estimated model coefficients is reduced. However, a ridge constant is needed to determine the bias/variance trade-off, and this requires *a priori* knowledge of the unknown coefficients.

High levels of correlation are often judged using the Variance Inflation Factor, which is the inverse of tolerance. Principal components have been suggested as a solution to high levels of collinearity, which also enables a dimension reduction. However, if the objective of model selection is to identify reliable coefficients to guide resource allocations by policy makers, principal components would not be a viable method.

6.2 *Non-Spherical Errors*

Non-spherical errors are generally thought of as indicating model mis-specification relative to the DGP, and can be interpreted as omitted variables, incorrect functional form, omitted dynamics, etc. Systematic components of the error term should properly enter the DGP as explanatory variables.[20] The theory of reduction (Hendry, 1995, chapter 9) describes the operations implicitly applied to the DGP to obtain the local DGP (LDGP, the generating process in the space of variables under analysis: see, for example, Hendry, 2009). The DGP is a highly complex joint distribution but can be simplified to the LDGP as long as there is no loss of information when applying the reductions, which is established by ensuring the LDGP satisfies properties such as constant parameters and innovation errors. Thus, a model that aims to approximate the LDGP should also have innovation errors. This is the concept of congruency.

The relative importance of congruency will depend on the objectives of the MSA. If the objective is to obtain the best forecasts, then congruency is not essential. If, on the other hand, the objective is to locate the best approximation to the DGP then not requiring congruency will imply that mis-specified models could be retained, which must be a poor approximation to the DGP as the unmodelled residuals contain aspects of the DGP.

The *Autometrics* MSAs are the only automatic algorithms that test for congruency during the selection procedure by undertaking a range of residual-based diagnostic tests, ensuring the overall test significance level is controlled. Other MSAs do not ensure congruency at any stage. This raises the likelihood that a mis-specified model could be selected, see Bontemps and Mizon (2003).

6.3 *Dimensionality Constraints*

Unfortunately, in economics, the number of candidate variables L is likely to be large due to uncertainty over relevant variables, lags and nonlinear transformations.[21] The 2^L models quickly become a binding constraint for MSAs that search over all models, due either to insufficient computing power or insufficient observations, or both. *Ad hoc* reductions in the number of models can be imposed to address these problems. This can be done, for example, by eliminating long lags, or variables with small t-statistics. However, this is unsatisfying because it removes large sets of models from consideration by the model selection procedure. This is an unavoidable cost of MSAs that compute all possible models.

In this sense, there is a computational advantage for MSAs that do not compute all possible models, such as stepwise MSAs, or *Autometrics*. *Autometrics* handles cases where there are more variables than observations by undertaking expanding and contracting searches so that the choice of L candidate variables need not be constrained by the number of observations T (Doornik, 2009b).

6.4 *Pre-Testing*

Applying model selection is known as pre-testing, and the process of model selection affects the validity of inference in finite samples. Pre-testing has been one of the main criticisms of model selection, see, for example, Judge and Bock (1978). Asymptotic distributions are unaltered by consistent model selection so asymptotic inferences are valid. However, in finite samples the distribution of estimators and test statistics can differ significantly from their limit distributions, see, for example, Leeb and Pötscher (2005).

Hendry and Krolzig (2003) distinguish between costs of inference and costs of search. Costs of inference are the costs associated with estimating the DGP. Even if the model is the DGP, estimation of a relevant variable may produce a low t-statistic, leading to the conclusion that the variable is 'insignificant'. This has nothing to do with model selection *per se*. No selection has taken place but the coefficient estimates of the model are interpreted as being insignificant according to a specified significance level.

The costs of search refer to the costs associated with searching for a specific model over and above the inevitable costs of inference. It is useful to separate the two costs because a measure of search costs will be contaminated if the DGP variables have low signal-to-noise ratios and would not be interpreted as significant even with no selection. Many evaluations of MSAs do not distinguish between these costs which result in misinterpreting the performance of MSAs. Specifically, an MSA may be concluded to perform poorly because it omits many relevant variables, even though these variables would be concluded to be insignificant if the DGP were the only model estimated.

Little can be done to correct for the omission of relevant variables due to costs of search and costs of inference. However, it may be possible to correct pre-test bias for selected variables depending on the specific search procedures of a given MSA. For example, *Autometrics* in Version 14 of PcGive automatically bias-corrects estimated coefficients after model selection.

6.5 *Endogeneity*

Most MSAs rely on weak exogeneity of regressors, unless the instrument set is known. Some MSAs can be applied to systems of equations, enabling tests of weak exogeneity (e.g. Krolzig, 2003). Key difficulties include the validity and significance of instrumental variables and identification of the simultaneous equations. In the absence of that knowledge, endogeneity can cause MSAs to produce biased and inconsistent coefficient estimates.

6.6 *Nonlinearity*

Economic relationships may be nonlinear, and a proliferation of nonlinear econometric models supports this view, ranging from nonlinear ARMA and bilinear models to random coefficient models, regime-switching models and artificial neural networks. MSAs that focus on variable selection often postulate a linear model. If this is a poor approximation, the selected model will not capture the key characteristics of the DGP. However, models that are nonlinear in the parameters can often be reparameterized to models that linear in parameters but nonlinear in variables. There are numerous nonlinear approximating classes including polynomials, orthogonal polynomials, Fourier series, asymptotic series and confluent hypergeometric functions. One problem with nonlinear functions is that they can generate substantial collinearity, an issue we identify above. Another problem is that generalizations can quickly produce a number of candidate variables that exceed the number of observations, another issue we discuss above.

7. Conclusion

This review endeavours to survey a number of common MSAs, discuss how they relate to each other, and identify factors that explain their relative performances. We categorize MSAs into four broad classes. The first class consists of *AIC*, *SIC* and related IC MSAs. These select a single best model based upon a search of all

possible models, with the best model being the one with the lowest IC value. The final model determines the value of the estimated coefficient. If the variable appears in the final model, the MSA assigns an estimated value equal to the estimated value of the corresponding coefficient in the final, best equation. If the variable does not appear in the final model, the MSA assigns an estimated coefficient value equal to zero.

The second class also uses IC criteria, but selects a portfolio of models, rather than a single best model. The range of the models included in the portfolio is derived from the sampling behaviour of the sample IC value. The literature is less clear on how the individual models in the portfolio should be combined. This class of models also searches over all possible models.

We denote the third class of models as 'path reduction' MSAs because they have the property that they do not search over all possible paths of the regression tree. Backwards- and forwards-stepwise MSAs fit into this class. 'Branch and bound' MSAs also do not search over all possible paths of the regression tree, but are still able to obtain the same outcome as MSAs that do. They do this by judiciously partitioning variables into various subsets, and then searching over these reduced subsets. Another type of path reduction MSAs is the general-to-specific modelling approach of the LSE school, of which *Autometrics* is the most modern version. This approach is distinguished by its emphasis on congruency and the use of multi-path searches to select variables based on significance rather than goodness of fit.

The final class of MSAs are BMA models. These MSAs, in principle, estimate all possible models and weight individual coefficients from a given model by the posterior probability that that model is 'correct'. These are then summed to produce estimated coefficients that are weighted averages of the estimated coefficients from individual models. If a variable does not appear in a given model, then its corresponding coefficient estimate is set equal to zero. In practice, sophisticated sampling algorithms are used to select a large subset from the full set of all possible models.

In order to illustrate the factors that affect relative MSA performance, we perform a large number of Monte Carlo experiments. The experiments vary over (1) the number of relevant variables (K); (2) the total number of candidate variables (L); (3) the degree of DGP noise, as measured by the non-centrality parameter (ψ) and (4) the number of observations in the data set, T. Twenty-one different MSAs are compared, representing a variety of approaches. The experiments illustrate the importance of (1) the ratio of relevant to total variables (K/L) and (2) DGP noise, as measured by the non-centrality parameter, as key determinants of relative MSA performance.

Our comparison of different MSAs highlights the fact that MSAs differ in the weights they place on type I and type II errors. MSAs with loose criteria place more weight on type II errors and are less concerned with type I errors, retaining irrelevant variables with a very high probability. MSAs with tight criteria place a lot of weight on type I errors, controlling the null-rejection frequency at a cost of failing to retain relevant variables when they have low non-centralities. It is this trade-off that is at the heart of MSA performance.

Our experiments confirm that no MSA does best in all circumstances. This follows directly from the fact that different MSAs place different weights on type I and type II errors. These weights will be advantageous or disadvantageous depending on the data environment. Even the worst MSA in terms of overall performance – the strategy of including all candidate variables – sometimes performs best (viz., when all candidate variables are relevant). Although no single MSA consistently dominates in all data environments, our experiments indicate that *Autometrics* does especially well when the ratio of relevant to irrelevant variables is less than 0.5, and the non-centrality parameter is equal to or less than 2. This case has particular interest because these conditions arguably define data environments where MSAs are likely to be most valuable to researchers. However, additional experimentation is required to determine whether these results are valid beyond the relatively simple testing environments we simulate here.

Finally, we discuss a number of further issues associated with the challenge of using MSAs to produce reliable coefficient estimates. These include: (1) collinearity, (2) non-spherical errors, (3) dimensionality constraints, (4) pre-testing, (5) endogeneity and (6) nonlinearity.

MSAs hold much promise to improve upon the method of *ad hoc* specification searches currently employed by most practitioners of empirical research. However, as this review makes clear, the choice of which MSA to use is not clear cut. Each has strengths and weaknesses that make it attractive in some, but not all, data environments. Additional research needs to delineate when, and which, MSAs may provide a useful alternative to current practice.

Acknowledgements

We acknowledge helpful comments from David F. Hendry, Jeffrey Wooldridge, participants at the 2008 Econometrics Society Australasian Meetings (Wellington, New Zealand) and the 2009 New Zealand Econometrics Study Group Meetings (Christchurch, New Zealand).

Notes

1. See Leeb and Pötscher (2003), for a case in which the DGP is not consistently estimated.
2. With an appropriate adjustment in notation, equation (1) could be modified to include lagged values of the dependent variable as explanatory variables, as well as allowing the explanatory variables to have different lag lengths, P_j.
3. Although our analysis assumes the researcher is interested in all the β's, it is straightforward to modify the analysis for when a given subset of the β's is of interest.
4. Although the authors do not list the fourth case, it should be noted that when $10 < \Re_m \leq 100$ the alternative model can again be discarded as non-competing.
5. There are other issues associated with BMA MSAs. One of these concerns the specification of the prior distribution. Magnus *et al.* (2010), propose an alternative to BMA that they call 'weighted average least squares', which utilizes the Laplace estimator. They claim two advantages for weighted average least squares over BMA:

(1) it adopts a more intuitive prior specification of parameter ignorance, and (2) it requires far less computational time, being linear, rather than exponential, in the number of regressors.

6. To ease notation, and without loss of generality, we treat lags of regressors as separate variables.
7. When a given variable i is not selected in any of the m replications ($\sum_{m=1}^{M} 1(\tilde{\beta}_{i,m} \neq 0) = 0$), it is conventional to set $CMSE = \beta_i^2$.
8. The difference between these two measures can be considerable when there is substantial multicollinearity. When this occurs, omitted variable bias may cause coefficients to differ substantially from their population values with little cost in predictive accuracy.
9. The intercept, γ, is fixed to enter all models.
10. $\frac{1.6}{T^{0.9}} \cdot 100\% \cong 5\%$ when $T = 47$, and $\frac{1.6}{T^{0.9}} \cdot 100\% \cong 1\%$ when $T = 281$.
11. Selection results in 'pre-test' biases (Judge and Bock, 1978). Hendry and Krolzig (2005) propose a bias correction procedure based on a truncated normal distribution for the post-selection coefficient estimates which can be easily implemented in a general-to-specific framework. Castle *et al.* (2011) motivate why bias correction is an integral aspect of the *Autometrics* algorithm, and the bias correction will be available in Version 14.
12. Even though the x's are orthogonal in the DGP, they will display non-zero correlations in the sample. Although this may affect relative MSA performance in any given experiment, it should not affect our cross-experiment results because the associated biases will differ as L and T are varied across experiments. A fuller examination of the role of collinearity on relative MSA performance is beyond the scope of this survey.
13. A t-test of H_0:$\beta_j = 0$ is given by $\mathrm{t_j} = \frac{\hat{\beta}_j}{\sigma_{\hat{\beta}_j}}$. If the x_j's are i.i.d., then $\sigma^2_{\hat{\beta}_j} = \frac{\sigma^2}{T\sigma^2_{x_j}}$. It follows that $\psi_j = \frac{\beta_j}{\sqrt{\sigma^2/T\sigma^2_{x_j}}}$. In our experimental design, $\beta_j = 1$ and $\sigma^2_{x_j} = 1$.
14. The power to reject the null hypothesis $H_0 : \beta_j = 0$ can be calculated as a function of ψ and α by $P(t \geq c_\alpha | E[t] = \psi) \approx P(t - \psi \geq c_\alpha - \psi | H_0)$, where c_α is the critical value for a given significance level, α. The associated retention rates are largely independent of T, except to the extent that T affects the critical value, c_α. Table A1 records powers for a single t-test for different values of ψ and α when $T = 75$.
15. See McQuarrie and Tsai (1998) for the importance of 'signal-to-noise' ratio as a determinant of MSA performance for IC algorithms.
16. Earlier analyses also compared MSA performance based on mean absolute deviations. We found little difference between these two performance measures and thus only report the UMSE results.
17. The intercept is omitted in the calculations as it is imposed in the selected model for all MSAs.
18. Ties were handled as follows. Let the MSAs be ranked in ascending order, MSA_1, $MSA_2, \ldots, MSA_j, MSA_{j+1}, \ldots, MSA_{j+m}, \ldots, MSA_{21}$; and suppose MSA_{j+1} to MSA_{j+m} are tied. Each of these receive rank $\sum_{\mathrm{i}=1}^{\mathrm{m}} (\mathrm{j} + \mathrm{i})/\mathrm{m}$.
19. The median ranking for *ALLVARS* over the 36 experiments where $K = L$ is 1.20. The next closest MSA has a median rank of 3.15.
20. Not everyone agrees that the DGP itself will necessarily have spherical errors. For example, Granger (2005, p. 5) argues that homoskedasticity need not be a necessary feature of the errors in a model for conditional expectations.

21. In our simulations the maximum number of regressors considered was $L = 15$, which equates to 32,768 possible models. Some of our individual experiments took more than a week to run on high-powered laptop computers.

References

Amemiya, T. (1980) Selection of regressors. *International Economic Review* 21: 331–354.

Akaike, H. (1969) Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21: 243–247.

Akaike, H. (1973) Information theory and an extension to the maximum likelihood principle. In B.N. Petrov and F. Csaki (eds), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akademia Kiado.

Anderson, T.W. (1962) The choice of the degree of a polynomial regression as a multiple-decision problem. *Annals of Mathematical Statistics* 33: 255–265.

Bontemps, C. and Mizon, G.E. (2003) Congruence and encompassing. In B.P. Stigum (ed.), *Econometrics and the Philosophy of Economics* (pp. 354–378). Princeton: Princeton University Press.

Bozdogan, H. (2000) Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology* 44: 62–91.

Breusch, T.S. (1990) Simplified extreme bounds. In C.W.J. Granger, (ed.), *Modelling Economic Series* (pp. 72–81). Oxford: Clarendon Press.

Burnham, K.P. and Anderson, D.R. (2004) Multimodel inference: understanding the AIC and BIC in model selection. *Sociological Methods & Research* 33(2): 261–304.

Campos, J., Hendry, D.F. and Krolzig, H. (2003) Consistent model selection by an automatic Gets approach. *Oxford Bulletin of Economics and Statistics* 65: 803–819.

Campos, J., Ericcson, N.R. and Hendry, D.F. (2005) General-to-specific modeling: an overview and selected bibliography. FRB International Finance Discussion Paper 838. Available at http://www.federalreserve.gov/pubs/ifdp/2005/838/default.htm (Last accessed 13 August 2011).

Castle, J.L., Doornik, J.A. and Hendry, D.F. (2011) Evaluating automatic model selection. *Journal of Time Series Econometrics* 3(1): Article 8.

Chow, G.C. (1981) Selection of econometric models by the information criteria. In E.G. Charatsis (ed.), *Proceedings of the Econometric Society European Meeting 1979*, Ch. 8. Amsterdam: North-Holland.

Clements, M.P. and Hendry, D.F. (2005) Evaluating a model by forecast performance. *Oxford Bulletin of Economics and Statistics* 67: 931–965.

Dhrymes, P.J. (1970) On the game of maximising R^2. *Australian Economic Papers* 9: 177–185.

Diebold, F.X. (2007) *Elements of Forecasting*, 4th edn. Cincinnati: South-Western.

Doornik, J.A. (2007) *An Object-Oriented Matrix Language Ox 5*. London: Timberlake Consultants Press.

Doornik, J.A. (2008) Encompassing and automatic model selection. *Oxford Bulletin of Economics and Statistics* 70: 915–925.

Doornik, J.A. (2009a) Autometrics. In J.L. Castle and N. Shephard (eds), *The Methodology and Practice of Econometrics* (pp. 88–121). Oxford: Oxford University Press.

Doornik, J.A. (2009b) Econometric model selection with more variables than observations. Nuffield College Working Paper. Available at http://www.ucl.be/cps/ucl/doc/core/documents/doornik.pdf (Last accessed 13 August 2011).

Fernández, C., Ley, E. and Steel, M.F.J. (2001) Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16: 563–576.

Furnival, G.M. (1971) All possible regressions with less computation. *Technometrics* 13: 403–408.
Gatu, C. and Kontoghiorghes, E.J. (2006) Branch-and-bound algorithms for computing the best subset regression models. *Journal of Computational and Graphical Statistics* 15: 139–156.
George, E.I. and Foster, D.P. (2000) Calibration and empirical Bayes variable selection. *Biometrika* 87(4): 731–747.
Gouriéroux, C. and Monfort, A. (1995) *Statistics and Econometric Models* (Vols. 1 and 2). Cambridge: Cambridge University Press.
Granger, C.W.J. (2005) Modeling, evaluation, and methodology in the new century. *Economic Inquiry* 43 (1): 1–12.
Hannan, E.J. (1980) The estimation of the order of an ARMA process. *Annals of Statistics* 8: 1071–1081.
Hannan, E.J. and Quinn, B.G. (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, B* 41: 190–195.
Hendry, D.F. (1995) *Dynamic Econometrics*. Oxford: Oxford University Press.
Hendry, D.F. (2003) Denis Sargan and the origins of the LSE econometric methodology. *Econometric Theory* 19: 456–480.
Hendry, D.F. (2009) The methodology of empirical econometric modeling: applied econometrics through the looking-glass. In T.C. Mills and K.D. Patterson (eds), *Palgrave Handbook of Econometrics* (pp. 3–67). Basingstoke: Palgrave MacMillan.
Hendry, D.F. and Doornik, J.A. (2009) *Empirical Econometric Modelling using PcGive* (Vol. I). London: Timberlake Consultants Press.
Hendry, D.F. and Krolzig, H. (2003) New developments in automatic general-to-specific modelling. In B.P. Stigum (ed.), *Econometrics and the Philosophy of Economics* (pp. 379–419). Princeton, New Jersey: Princeton University Press.
Hendry, D.F. and Krolzig, H. (2004) We ran one regression. *Oxford Bulletin of Economics and Statistics* 66(5): 799–810.
Hendry, D.F. and Krolzig, H. (2005) The properties of automatic GETS modelling. *The Economic Journal* 115: C32–C61.
Hocking, R.R. and Leslie, R.N. (1967) Selection of the best subset in regression analysis. *Technometrics* 9: 531–540.
Hoerl, A.E. and Kennard, R.W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1): 55–67.
Hoerl, A.E. and Kennard, R.W. (1970b) Ridge regression: applications to nonorthogonal problems. *Technometrics* 12(1): 69–82.
Hoeting, J.A. (2002) Methodology for Bayesian model averaging: an update. *Proceedings of the XXIst International Biometric Conference*: 231–240. Available at http://www.stat.colostate.edu/~jah/papers/ibcbma.pdf (Last accessed 19 August 2011).
Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) Bayesian model averaging: a tutorial. *Statistical Science* 14(4): 382–417.
Hoover, K.D. and Perez, S.J. (2004) Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics* 66(5): 765–798.
Hurvich, C.M. and Tsai, C.L. (1989) Regression and time series model selection in small samples. *Biometrika* 76: 297–307.
Jeffreys, H. (1961) *Theory of Probability*, 3rd edn. London: Oxford University Press.
Judge, G.G. and Bock, M.E. (1978) *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.
Krolzig, H. (2003) General-to-specific model selection procedures for structural vector autoregressions. *Oxford Bulletin of Economics and Statistics* 65: 769–801.
Kuha, J. (2004) AIC and BIC: comparisons of assumptions and performance. *Sociological Methods & Research* 33(2): 188–229.

Leamer, E.E. (1978) *Specification Searches. Ad Hoc Inference with Non-Experimental Data*. New York: John Wiley & Sons.
Leamer, E.E. (1983) Let's take the con out of econometrics. *American Economic Review* 73: 31–43.
Leamer, E.E. (1985) Sensitivity analyses would help. *American Economic Review* 75: 308–313.
Leeb, H. and Pötscher, B.M. (2003) The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. *Econometric Theory* 19: 100–142.
Leeb, H. and Pötscher, B.M. (2005) Model selection and inference: facts and fiction. *Econometric Theory* 21: 9–59.
Levine, R. and Renelt, D. (1992) A sensitivity analysis of cross-country growth regressions. *American Economic Review* 82(4): 942–963.
Magnus, J.R., Powell, O. and Prüfer, P. (2010) A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* 154: 139–153.
Mallows, C.L. (1973) Some comments on C_p. *Technometrics* 15: 661–675.
McAleer, M., Pagan, A.R. and Volker, P.A. (1985) What will take the con out of econometrics? *American Economic Review* 95: 293–301.
McQuarrie, A.D. (1999) A small-sample correction for the Schwarz SIC model selection criterion. *Statistics & Probability Letters* 44: 79–86.
McQuarrie, A.D. and Tsai, C. (1998) *Regression and Time Series Model Selection*. Singapore: World Scientific Publishing.
Mizon, G.E. (1995) Progressive modelling of macroeconomic time series: the LSE methodology. In K.D. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects* (pp. 107–170). Boston: Kluwer.
Owen, P.D. (2003) General-to-specific modelling using PcGets. *Journal of Economic Surveys* 17(4): 609–628.
Oxley, L.T. (1995) An expert systems approach to econometric modelling. *Mathematics and Computers in Simulation* 39: 379–383.
Pagan, A.R. (1987) Three econometric methodologies: a critical appraisal. *Journal of Economic Surveys* 1: 3–24.
Pesaran, M.H. (1974) On the general problem of model selection. *The Review of Economic Studies* 41(2): 153–171
Phillips, P.C.B. (1988) Reflections on econometric methodology. *Economic Record* 64: 344–359.
Phillips, P.C.B. (1994) Bayes models and forecasts of Australian macroeconomic time series. In C. Hargreaves (ed.), *Non-stationary Time Series Analyses and Cointegration* (pp. 53–86). Oxford: Oxford University Press.
Phillips, P.C.B. (1995) Automated forecasts of Asia-Pacific economic activity. *Asia-Pacific Economic Review* 1: 92–102.
Phillips, P.C.B. (2005) Automated discovery in econometrics. *Econometric Theory* 21(1): 3–20.
Poskitt, D.S. and A.R. Tremayne. (1987) Determining a portfolio of time series models. *Biometrika* 74(1): 125–137.
Raftery, A.E. (1996) Approximate Bayes factors and accounting for model uncertainty in generalized linear regression models. *Biometrika* 83: 251–266.
Raftery, A.E., Madigan, D. and Hoeting, J. (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92 : 179–191.
Reed, W.R. and Ye, H. (2011) Which panel data estimator should I use? *Applied Economics* 43(8): 985–1000.
Sala-i-Martin, X., Doppelhofer, G. and Miller, R.I. (2004) Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94(4): 813–835.

Schatzoff, M., Tsao, R. and Fienberg, S. (1968) Efficient calculation of all possible regressions. *Technometrics* 10: 769–779.
Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6(2): 461–464.
Shibata, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8: 147–164.
Shibata, R. (1981) An optimal selection of regression variables. *Biometrika* 68: 45–54. Correction Vol. 69 (1982): 492.

Table A1. Total Number of Experiments by ψ and T.

	$T = 75$	$T = 150$	$T = 500$	$T = 1500$	TOTAL
$\psi = 1$	$L = 5,10,15$ (30 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	60
$\psi = 2$	$L = 5,10,15$ (30 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	60
$\psi = 3$	$L = 5,10,15$ (30 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	60
$\psi = 4$	$L = 5,10,15$ (30 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	60
$\psi = 5$	$L = 5,10,15$ (30 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	60
$\psi = 6$	$L = 5,10,15$ (30 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	$L = 10$ (10 experiments)	60
Total	180	60	60	60	360

Journal of Economic Surveys (2011) Vol. 00, No. 0, pp. 1–33