Econometrics Journal (1999), volume **2**, pp. 192–201.

Discussion of 'Data mining reconsidered'

BRUCE E. HANSEN

University of Wisconsin, Department of Economics, Social Science Building, Madison, WI 53706, USA
E-mail: www.ssc.wisc.edu/~bhansen.

Received: December 1999

Summary The Hoover–Perez description of the LSE [London School of Economics] general-to-specific methodology of model selection is formalized and analysed using the theory of model selection. Numerical evidence is provided to justify the claim that simple and elegant information criteria (which are easy to implement in applications) work at least as well, if not better, than the complicated algorithm attributed to the LSE methodology.

Keywords: *BIC, Model selection, Forecast error.*

1. INTRODUCTION

Hoover and Perez (1999) have written a novel investigation into the LSE methodology of model selection. Perhaps the most interesting aspect of their paper is that they write down a formal algorithm which they claim mimics the behavior of LSE researchers. Once such an algorithm has been written down, it can be studied using standard statistical techniques, including Monte Carlo simulation.

I find that some formalization is often clarifying. In this discussion, I attempt to place the Hoover–Perez (HP) algorithm into a formal framework, in which we can compare their method with alternative model selection methods, and thereby deduce asymptotic properties of the procedure. I also provide my own simulation evidence, comparing their method with simple BIC-type (Bayesian information criteria) selection rules, and find that the simpler BIC-type rules can work much better than the complicated HP algorithm.

2. A FORMAL DESCRIPTION OF THE HP ALGORITHM

Hoover and Perez describe a search algorithm which they believe approximates the LSE general-to-specific modeling approach. This algorithm selects one model from a large space of possible regression models and is thus a particular solution to the general problem of model selection. What is important and distinguishing about the LSE modeling approach, and is properly modeled by Hoover and Perez, is that classic diagnostic tests are instrumentally used as a part of model selection. Formally, the HP search algorithm can be described by the choice of a *test*, a *measure of fit*, and a *search path*. Briefly, the HP search algorithm examines all models along the search path, and among those models, selects the best-fitting model among all those which are not rejected by the test.

It is helpful to introduce some formal notation. Let $\mathcal{M}$ be the space of possible models and let $m \in \mathcal{M}$ denote a generic element of this set. For example, in the context of the Monte Carlo study conducted by the authors, $\mathcal{M}$ is the set of 2^{40} linear regressions made up of the various subsets of the 40 regressors, and m is a generic regression model involving some subset of the regressors. The HP algorithm selects a single element m from $\mathcal{M}$, based on the observed dataset. We can denote this selection mechanism formally as $\hat{m} = s(\mathcal{M})$, where the function s implicitly depends on the data.

We said above that the algorithm can be described by a test, a measure of fit, and a search path. Formally, let $t(m)$ denote the test, where $t(m) = 1$ indicates 'rejection' and $t(m) = 0$ indicates otherwise, let $f(m)$ denote the measure of fit, and let $\mathcal{M}_P$ denote the set of models along the search path. The set of models which 'pass' the test t can be written as

$$\mathcal{M}_0 = \{m \in \mathcal{M}_P : t(m) = 0\}.$$

These are the models along the search path which are not rejected by the test.

The HP search algorithm sets the selected model $\hat{m}$ to be the element of $\mathcal{M}_0$ which has the smallest value of $f(m)$. Thus

$$\begin{aligned} \hat{m} &= \underset{m \in \mathcal{M}_0}{\operatorname{argmin}} \, f(m) \\ &= \underset{m \in \mathcal{M}_P : t(m) = 0}{\operatorname{argmin}} \, f(m). \end{aligned} \tag{1}$$

We now describe the HP test function in more detail. For some nominal test size α, the HP test function $t(m)$ can be written as

$$t(m) = 1\{\min(p_1, p_2, p_3, p_4, p_5, p_6, p_7) \leq \alpha\} \tag{2}$$

where the p_i, $i = 1, \ldots, 7$, are asymptotic p-values (based on chi-squared or F tables) for the following test statistics:

(1) Jarque and Bera (1980) normality of residuals, based on the initial 90% of observations.
(2) Breusch and Pagan (1980) residual autocorrelation, based on the initial 90% of observations.
(3) Engle (1982) autocorrelated conditional heteroskedasticity (ARCH), based on the initial 90% of observations.
(4) Chow (1960) sample-split parameter stability, based on the initial 90% of observations.
(5) Chow (1960), out-of-sample stability, based on the initial 90% of observations versus final 10%.
(6) General F test for exclusion restrictions implied by model m against the general model, based on the initial 90% of observations.
(7) F test for exclusion restrictions implied by model m_0 against model m, where m_0 contains all regressors in model m with 'significant' t-statistics, based on estimation using all observations.

Test (2) rejects if one of the seven test statistics is individually significant at the nominal level α. The authors vary the nominal level α among 0.01, 0.05 and 0.10, as they suggest that the LSE methodology does not dictate a particular choice. Furthermore, while this is not stated explicitly, it appears that all of these test statistics are calculated using the 'homoskedastic' formula for standard errors and covariance matrices, rather than a method which is valid in the presence of conditional heteroskedasticity.

For the measure of fit, Hoover and Perez use the standard error of regression, or

$$f(m) = \sqrt{\frac{1}{T - k(m)} \hat{e}_m' \hat{e}_m} \tag{3}$$

where $k(m)$ is the number of regressors in the model m and $\hat{e}_m$ is the OLS (ordinary least squares) residual vector from model m. They argue that in the class of linear regression models, the LSE methodology embraces this measure of fit.

Finally, the HP algorithm involves a particular search path $\mathcal{M}_P = (m_i : i = 1, \ldots, N)$ which is determined by a sequence of individual t-tests and application of the test $t(m)$, and is described in detail in their paper.

3. SEARCH PATH

Hoover and Perez devote considerable space to describing their search path. From their language, descriptions, and arguments, it appears that they consider this to be an essential part of their search algorithm. I would like to take a contrary view and suggest that the particular search path mechanism plays a relatively small role in determination of the search algorithm. Rather, what is important is that the search path includes a large number of distinct models m, and is disposed to examining models m which are most likely to succeed on the merits (pass the test $t(m)$ and have low measure of fit $f(m)$). That is, suppose that we define the model selection rule

$$\tilde{m} = \underset{m \in \mathcal{M}:\ t(m)=0}{\operatorname{argmin}} f(m), \tag{4}$$

which is only different from (1) in that (4) searches over all elements of $\mathcal{M}$, while (1) looks only at models on the search path. If $\mathcal{M}_P$ is sufficiently large (and is likely to contain the elements of $\mathcal{M}$ which pass the test and have good fit), then $\tilde{m}$ is likely to equal $\hat{m}$.

It seems reasonable to view $\tilde{m}$ in (4) as an idealized version of $\hat{m}$. Indeed, I think the view is that the ideal is to compute $\tilde{m}$, but this is infeasible, and $\hat{m}$ is a convenient computational shortcut to $\tilde{m}$. To be sure, the explicit computation suggested in (4) is prohibitively expensive, as it requires computing all $2^{40} = 1\,099\,511\,627\,776$ distinct models in $\mathcal{M}$. My point is that the HP proposed search path should be viewed by how closely it approximates the ideal (4) using minimal computational resources. On this measure their proposal appears quite reasonable.

4. TESTING

4.1. Non-normality and conditional heteroskedasticity

Hoover and Perez incorporate in their algorithm tests for normality of the residuals and for autocorrelated conditional hetereskedasticity (ARCH). They argue that this is consistent with the standard practice of researchers following the LSE methodology. They also implicitly indicate that LSE researchers routinely use test statistics which are not robust to heteroskedasticity (i.e. they use so-called 'conventional' standard errors). It is not clear to me if this is an accurate portrayal of the current econometric practice of LSE researchers, but it is clear to me that this is a portrayal of poor econometric practice.

First, why test residuals for normality? Normality is neither a good nor bad property of a regression error. The interpretation of the regression function as a conditional expectation has no relationship to an auxiliary assumption of normality. I have heard the argument that error normality is desirable because the distribution of test statistics (t and F) are based on the assumption of normality, but this argument is false in the case of time-series data. The conventional justification is asymptotic theory, where normality is rarely a necessary assumption.

Second, why test residuals for conditional heteroskedasticity? The true regression will have errors which may or may not be conditionally heteroskedastic. The finding of conditional homoskedasticity sheds no information concerning whether or not the conditional mean is correctly specified. Of course, conditional heteroskedasticity might be interesting on its own merits. For example, if the goal is to produce out-of-sample prediction intervals, then knowledge of the conditional mean is insufficient, knowledge of the conditional variance is necessary, and knowledge of the entire conditional distribution is best! However in this case the interest would not be to test for conditional heteroskedasticity, rather it would be to build a model of the conditional variance function, which is a completely different exercise. The observation remains that tests for conditional heteroskedasticity have no role in determination of the correct model of the conditional mean.

Third, once we acknowledge that the regression errors may be conditionally heteroskedastic, the use of conventional standard errors and covariance matrices appears unwise. Reported test statistics will not have the asymptotic distributions claimed, and the divergence can be quite large. Decisions will be based on inappropriate test statistics, and hence will be misinformed. Strangely enough, this should not be viewed as a problem or dilemma, since heteroskedasticity-robust methods to calculate standard errors and covariance matrices are well known, trivial to calculate, and widely used.

To summarize, my recommendation to applied researchers would be to omit the tests of normality and conditional heteroskedasticity, and replace all conventional standard errors and covariance matrices with heteroskedasticity-robust versions.

4.2. Parameter stability

Hoover and Perez incorporate tests for in-sample and out-of-sample stability, as they argue that these tests are routinely used by LSE practitioners. While testing for parameter instability has a long and solid tradition in econometric practice, I think it is fair to credit the LSE researchers for paying particular attention to this issue, and routinizing stability testing into all their econometric analysis. I applaud this effort, for the following reasons. First, parameter change is a fundamental violation of the goal of model construction. In the presence of unrecognized parameter change, least-squares estimates are not very informative and can be quite misleading concerning the objects of fundamental interest. Second, parameter instability testing may be viewed as a test of overidentifying restrictions, and this information is not taken into account by model selection criterion based on the conventional just-identifying restrictions. Another way of stating the latter is that the BIC criterion is an adjustment to the likelihood, but the latter is calculated for a model with constant parameters so may not properly convey the information that the model is misspecified. A proper way to solve this problem is to test for parameter instability and only adopt models which pass this test (as is done in the HP algorithm).

My trouble with the HP implementation, however, concerns the choice of test statistic. Their choice seems to suggest that there has been no progress in the testing issue since Chow (1960),

which is a poor characterization of the current literature. I note that Hoover and Perez incorporate two tests: an in-sample test, and an out-of-sample test. In-sample and out-of-sample stability may seem like useful complements, but they are really getting at the same issue: parameter stability. The best approach is to employ the most powerful test for structural instability, and ignore the less powerful tests. (If that is not obvious, reconsider the notion of a *best test*.) The literature suggests that the current best test is that of Andrews (1993), which is a full-sample (or in-sample) test for parameter change. This test is easy to implement and interpret in linear regressions, and should replace the older Chow-type tests in applied time series.

4.3. Significance levels

A test is described by a statistic and a critical value. HP suggest that the LSE methodology sets the critical value by reference to conventional asymptotic theory, but that the methodology does not provide a particular guideline concerning the significance level. They therefore experiment with three significance levels—10%, 5%, and 1%—but take 5% as the benchmark case corresponding to conventional applied practice.

The challenge for model selection based on hypothesis testing is that there is no good rule for the selection of the significance level. From a decision-theoretic framework, the 'optimal' significance level will depend on the power of the test in the specific context of interest, and the cost of incorrect decision. It follows that in some contexts, a 'liberal' significance level such as 50% may be appropriate, while in other contexts a 'conservative' critical value such as 1% or even 0.1% may be optimal. The difficulty is to know how to pick the level in a specific application.

The choice of significance level would not be a concern if it did not have a meaningful impact on the results, but as shown in the Monte Carlo experiments reported in the paper, it can have an enormous impact. (Compare Tables 4, 6, and 7 of Hoover and Perez.) This appears to be a fairly generic problem with model selection based on testing, and is a strong reason for avoiding such methodologies. This argument has been made forcefully in a recent paper by Granger *et al.* (1995), who argue that the indeterminacy of the nominal size renders the hypothesis testing approach ill-suited for model selection. These authors conclude 'that it is better to use model selection procedures rather than formal hypothesis testing when deciding on model specification.' We take up this recommendation in the next section.

5. MEASURING FIT

5.1. The HP measure

Some manipulation of the HP measure-of-fit formula (3) provides some insight into its interpretation. Let $\hat{\sigma}^2(m) = T^{-1}\hat{e}_m'\hat{e}_m$ be the quasi-likelihood estimator of the error variance. Then

$$2\log f(m) = \log\hat{\sigma}^2(m) - \log\left\{\frac{T-k(m)}{T}\right\} \simeq \log\hat{\sigma}^2(m) + \frac{k(m)}{T}.$$

Thus minimization of Hoover–Perez's $f(m)$ is equivalent to minimization of the criterion function

$$HP(m) = \log\hat{\sigma}^2(m) + \frac{k(m)}{T}. \tag{5}$$

The criterion $HP(m)$ has a similar form to the Akaike information criteria (AIC), which is

$$AIC(m) = \log \hat{\sigma}^2(m) + 2\frac{k(m)}{T} \tag{6}$$

but puts a smaller penalty on over-parameterization. It follows that model selection based on minimization of $HP(m)$ will produce larger parameterizations than model selection based on minimization of $AIC(m)$. Both place a smaller penalty[1] on large parameterizations than the Bayesian information criteria (BIC), also known as the Schwarz criteria, which is

$$BIC(m) = \log \hat{\sigma}^2(m) + \frac{\log(T)k(m)}{T}. \tag{7}$$

On the basis of this comparison alone, we have considerable reason to be skeptical of models selected by the HP criteria $HP(m)$.

5.2. Consistent model selection

The HP algorithm is a model selection device. One measure of 'success' is how frequently the true model is selected. A procedure is called *consistent* if the frequency of correct selection converges to one as the sample size increases. The consistency properties of various model selection methods have been studied in the theoretical literature. (See, for example, Nishi (1988), White (1990), Potscher (1991), Granger *et al.* (1995).)

One approach to model selection is hypothesis testing. One difficulty is that if the nominal size of the tests are held fixed, hypothesis testing yields inconsistent model selection. The problem is that an over-parameterized model is always 'accepted' with probability equaling the size of the test. Asymptotically, the selected model is either the true model or an over-parameterized model (but not an incorrect or under-parameterized model). Technically, a solution (as shown by White (1990)) is to let the nominal size shrink to zero as the sample size expands, in which case model selection can be consistent. In practice, however, this result does not give a rule to actually pick the nominal size in a particular application. Without a well-grounded (and presumably data-dependent) rule, this procedure is not operational. This has led most researchers (e.g. Granger *et al.* (1995)) to argue that hypothesis testing is ill-suited for model selection, and to alternatively advocate model selection procedures based on information criteria.

Two popular information criteria are the AIC and the BIC. As shown by Nishi (1988), model selection based on the AIC is inconsistent. The result is quite similar to that of model selection based on hypothesis testing: incorrect models are never selected asymptotically, but over-parameterized models are asymptotically selected with positive probability. This inconsistency is shared by any criterion of the form $C(m) = \log \hat{\sigma}^2(m) + c\frac{k(m)}{T}$ with $c > 0$. Both the AIC and HP criterion take this form, from which we conclude that model selection based on the HP criterion (5) is inconsistent.

Since the HP algorithm is a merger of model selection based on testing and on minimization of the HP criterion (5), and these two methods have similar asymptotic properties, there is no reason to suppose that the HP algorithm will be any different. That is, we conclude that their algorithm is inconsistent, and will over-select with positive probability, even in large samples.

[1] When $T \geq 8$.

Table 1. Consistent model selection. Percentage of searches for which the selected specification is the true model.

	Model								
Method	1	2	3	4	5	6	7	8	9
BIC	45	40	41	44	46	1	43	45	0
BIC*, $c = 1.4$	79	74	70	74	76	1	75	76	0
BIC*, $c = 1.7$	91	87	80	86	88	1	87	89	0
BIC*, $c = 2.0$	95	95	83	93	94	1	94	94	0

On the positive side, we can also conclude that their algorithm will not (asymptotically) select an incorrect (under-parameterized) model.

In contrast, model selection based on minimizing the BIC (7) is consistent. This observation suggests that LSE-type researchers might benefit from the BIC in model selection. The consistency property is shared by all criteria functions of the form

$$BIC^*(m) = \log \hat{\sigma}^2(m) + c\frac{\log(T)k(m)}{T} \tag{8}$$

with $c > 0$.

When the number of models is large (as in the Monte Carlo exercise), it may be impossible to actually implement the BIC-minimizing procedure, as this would require estimation of every possible model. So just as HP found it necessary to perform a computational shortcut by employing a linear search algorithm, it is also necessary when picking models via the BIC. There is no obvious unique shortcut, but the following is a fairly straightforward approximation. Start with the most general model, estimate this model, and calculate the t-ratios on the individual regressors. Then eliminate the regressor with the smallest t-ratio, and re-estimate the model with the remaining regressors. Eliminate the next regressor with the smallest t-ratio, and continue the process until just a manageable set of k regressors remains (perhaps set $k = 10$). At this point, sequentially estimate all 2^k models which can be formed from these k regressors. For each regression which has been run on this search, calculate and store the BIC. The model with the smallest BIC is the selected model. This method is likely to pick the true BIC-minimizing model if the manageable number k is sufficiently large. The total number of regressions estimated is on the order of 2^k. This approximation may appear to be a mixture of 'pre-selection' and BIC minimization, but it is more properly viewed as a numerical approximation to true BIC minimization.

We now numerically contrast this implementation of the BIC with the HP algorithm. We followed the HP simulation design, drawing 1000 samples from each of the nine models, and then found the model which minimized this implementation of the BIC criteria, setting $k = 10$.

The percentage of successes for the BIC selection methods are reported in Table 1. This should be contrasted with the first row of Tables 4, 6, and 7. For all models the BIC method is more successful than the benchmark HP method (with a 5% significance level). For some models, such as models 2 and 7, the difference is quite dramatic.

Hoover and Perez found that they could dramatically improve their success by decreasing the significance level of their tests to 1%. Indeed, using this level, their methodology does better than the BIC method. The reason for this improved performance is that the smaller significance level favors the null hypothesis, and hence more parsimonious specifications. We can do the

same with our information criteria method, by using (8) with $c > 1$. We tried three alternative values, $c = 1.4$, 1.7, and 2.0. The success of these criterion functions are reported in Table 1 as well. Similarly to the improvements in the HP algorithm, we find dramatic improvements in the performance of the BIC criterion by increasing the parameterization penalty c. Specifically, BIC* with $c = 1.4$ performs quite similarly to the HP algorithm with a level of 1% (their Table 7), with the notable exceptions of model 2 (the HP success rate is 1%, and the BIC* success rate is 74%) and model 7 (the HP success rate is 25%, and the BIC* success rate is 75%). By setting $c = 1.7$ or $c = 2.0$, BIC* does even better, with success rates above 94% for six of the nine models.

It is tempting to conclude from Table 1 that the BIC should be modified as in (8) to set $c > 1$. This would be an incorrect conclusion. The superiority of $c > 1$ to $c = 1$ is an artifact of the particular simulation study, and could easily be reversed using different models and parameterizations.

The point of this exercise rather is to show that simple and elegant information criteria, which are easy to implement in applications, perform at least as well, if not better, than complicated search algorithms.

5.3. Model evaluation

Hoover and Perez focus on the issue of correct model identification. However, that is not always the most appropriate measure of success for model selection. Rather, the measure of success depends on the purpose of the model selection exercise. In the standard language of statistical decision theory, it depends on the loss function.

A typical use of econometric models is forecasting. When that is the goal, a typical measure of success is out-of-sample forecast accuracy. While there are many measures of forecast accuracy, the most commonly used is expected root-mean-squared error.

In the Hoover and Perez Monte Carlo exercise it is quite straightforward to assess the out-of-sample forecasting performance of alternative model selection methods. For each simulated sample, there are 999 other simulated samples which can be spliced together to make a very large out-of-sample population. Using the fitted estimates from the selected model, for each sample 135 out-of-sample one-step-ahead prediction residuals can be calculated, and hence the mean-squared error. The square root of this quantity is the root-mse, and is associated with the particular simulated sample. As it is estimated on 1000×135 out-of-sample observations, this quantity is quite accurately estimated. Since there are 1000 simulated samples, there will be 1000 root-mse. (Each is the forecast root-mse for an individual fitted model.) The expected value is estimated by averaging over these 1000 observations.

To compare alternative forecast methods including the HP algorithm, I needed to be able to replicate the HP estimation algorithm itself. While they have provided the MATLAB code for their calculations, I did my calculations in GAUSS, and did not have the time to transcribe their entire algorithm. Rather, I wrote a poor man's approximation to their algorithm. My algorithm follows exactly all steps described in their paper, with the notable exception that I did not utilize the tests for normality, residual autocorrelation, ARCH, and the Chow stability tests (tests (1)–(5) listed in Section 2 above). This means that my poor-man's approximation relied exclusively on the sequential t-tests and joint F-tests for exclusion on the initial 90% of the data, and re-estimation and re-searching using the entire 100% of the data. This poor-man's approximation worked extremely well, and appears to mimic the HP algorithm quite closely, as measured by the accuracy information reported in their Tables 4, 6, and 7. (In fact, my poor-man's approximation

Table 2. Forecast accuracy. Percentage increase in average forecast root-mean-squared error (relative to forecasts using true specification).

	Model								
Method	1	2	3	4	5	6	7	8	9
HP*, 10% level	5.03	5.27	5.52	4.80	4.84	4.50	4.91	4.75	4.35
HP*, 5% level	2.89	3.17	3.48	2.78	2.76	2.46	2.92	2.84	2.55
HP*, 1% level	1.65	1.81	2.33	1.61	1.51	1.31	1.70	1.69	1.38
BIC	2.21	2.59	2.71	2.20	2.10	1.93	2.40	2.33	1.95
BIC*, $c = 1.4$	0.79	1.05	1.40	0.99	0.92	0.66	1.00	0.97	0.71
BIC*, $c = 1.7$	0.35	0.52	1.15	0.55	0.46	0.20	0.58	0.49	0.34
BIC*, $c = 2.0$	0.19	0.29	1.22	0.28	0.23	−0.04	0.32	0.27	0.09
AR(4)	1.40	1.06	0.72	39.89	213.98	39.26	61.83	413.88	60.64

actually selected the true model slightly more often than the full-blown HP algorithm.) I will denote this HP approximate method as HP*, and will use it for the subsequent analysis as if it were actually the HP method.

To introduce another forecast method for comparison, I included an AR(4). That is, for all models, the AR(4) was used for out-of-sample forecasting, without any data-dependent model selection. It is often argued in the forecasting literature that simple autoregressive models can out-perform, or come close to, more complicated models, partly due to their simplicity and parsimony.

For a baseline comparison model, the true model was also estimated and used for out-of-sample forecasts. As this is a baseline, the expected out-of-sample root-mse were all re-expressed in percentage deviations from this baseline.

The results are reported in Table 2. In all cases, BIC beats the benchmark HP* algorithm which uses a 5% significance level. However, in all cases, neither method performs much worse that using the 'true' specification, and the percentage loss (relative to using the true specification) does not vary meaningfully across models. Perhaps the comparison with the AR(4) is more interesting. For models 1 through 3, the AR(4) forecasts are much better than those from the BIC and HP* specifications, but this is not surprising, since these models are nested in the AR(4). For all other models, however, the AR(4) model performs quite badly in out-of-sample forecasting. In the worse case of model 8, the AR(4) model produces forecasts whose expected root-mse is over four times as large as that obtained from the other techniques.

As we found in the previous section, further improvements can be made by either decreasing the significance level used in the HP algorithm, or increasing the penalty c in the BIC* criterion. We find that by setting $c \geq 1.4$, the BIC* criterion performs much better in all models than the best HP algorithm, and the performance improves for all models by increasing c to 1.7 and 2.0. For model 6, BIC* with $c = 2.0$ actually produces better forecasts than using the true specification (this is possible as estimation of the true specification may be imprecise).

6. CONCLUSION

Hoover and Perez describe a model selection and testing methodology which they associate with the LSE style of econometrics. I am not sure if theirs is an accurate description of all LSE researchers—this is not my concern. Rather, my concern is that their implicit recommendations do not always seem prudent.

Constructively, if I were forced to give a stylized description of a 'modern' approach to model selection, these might be my rough guidelines:

(1) Pick the model by minimization of a reasonably-motivated information criteria such as the BIC.
(2) Subject your model to the Andrews (1993) test for structural change.
(3) Use test statistics and standard errors which are valid under heteroskedasticity.
(4) For inference, consider using bootstrap rather than asymptotic approximations.

ACKNOWLEDGEMENTS

This research was supported by a grant from the National Science Foundation. I thank Neil Shephard and two reviewers for helpful comments.

REFERENCES

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica 61*, 821–56.

Breusch, T. S. and A. R. Pagan (1980). The Lagrange multiplier test and its application to model specification in econometrics. *Review of Economic Studies 47*, 239–53.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica 28* 591–603.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica 50*, 987–1008.

Granger, C. W., M. K. King and H. White (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics 67*, 173–87.

Jarque, C. M. and A.K. Bera (1980). Efficient tests for normality, homoskedasticity and serial independence of regression residuals. *Economic Letters 6*, 255–9.

Nishi, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis 27*, 392–403.

Potscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory 7*, 163–85.

White, H. (1990). A consistent model selection procedure based on m-testing. In C. W. J. Granger (ed.), *Modeling Economic Series: Reading in Econometric Methodology*, pp. 369–83. Oxford: Oxford University Press.