ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

A Model Validation Procedure

Julia Polak, Maxwell L. King and Xibin Zhang

October 2014

Working Paper 21/14

A Model Validation Procedure

Julia Polak
Department of Econometrics and Business Statistics,
Monash University,
Clayton, VIC 3800, Australia.
Email: Julia.Polak@monash.edu

Maxwell L. King
Department of Econometrics and Business Statistics,
Monash University,
Clayton, VIC 3800, Australia.
Email: Max.King@monash.edu

Xibin Zhang
Department of Econometrics and Business Statistics,
Monash University,
Caulfield East, VIC 3145, Australia.
Email: Xibin.Zhang@monash.edu

6 October 2014

JEL classification: C12, C14, C52, C53

A Model Validation Procedure

Abstract

Statistical models can play a crucial role in decision making. Traditional model validation tests typically make restrictive parametric assumptions about the model under the null and the alternative hypotheses. The majority of these tests examine one type of change at a time. This paper presents a method for determining whether new data continues to support the chosen model. We suggest using simulation and the kernel density estimator instead of assuming a parametric distribution for the data under the hull hypothesis. This leads to a more versatile testing procedure, one that can be applied to test different types of models and look for a variety of different types of divergences from the null hypothesis. Such a flexible testing procedure, in some cases, can also replace a range of tests that each test against particular alternative hypotheses. The procedure's ability to recognize a change in the underlying model is demonstrated through AR(1) and linear models. We examine the power of our procedure to detect changes in the variance of the error term and the AR coefficient in the AR(1) model. In the linear model, we examine the performance of the procedure when there are changes in the error variance and error distribution, and when an economic cycle is introduced into the model. We find that the procedure has correct empirical size and high power to recognize the changes in the data generating process after 10 to 15 new observations, depending on the type and extent of the change.

Keywords: Chow test, model validation, *p*-value, multivariate kernel density estimation, structural break.

1 Introduction

Statistical models are widely used as an aid to decision making. Prior to using a model as a decision tool, it is carefully selected from a range of alternative models using various criteria, such as the ability to minimize the sum of squared errors or a cost function, having the lowest AIC or BIC value or having the best forecast performance over a testing sample. In addition, different parts of the model are examined, such as the residual behavior, the significance of coefficients, the logic of included and excluded explanatory variables and so on.

In this paper we look at the model, which already has been selected, and examine whether it is still supported by new data that has become available. We do not aim to select a model from several competing models, but to evaluate the selected model's likelihood to have generated the new data.

Once the best possible model is selected and 'brought down to production', there are few statistical tools to clarify its adequateness under the changing conditions of reality. Moreover, most of these tools are designed to detect specific types of change, require making restrictive assumptions on the data generating process (DGP) or a long waiting period before they can be applied.

One of the best known tools for model validation is the test proposed by Chow (1960), known as the Chow test. This test examines whether the m additional observations belong to the same data generating process as the previous n observations. The Chow test is widely used in testing for structural breaks. Although it provides a statistic with a known distribution, it has some limitations. The statistic follows a standard F distribution only for linear models where the errors are independently and normally distributed with a constant variance. For example, Ghilagaber (2004) provided a numerical examination of the performance of Chow test in the presence of heteroscedasticity. He found that the test shows very poor performance in this case unless the sample sizes before and after the break point are equal and the form of heteroscedasticity remains the same.

During the past half a century, many Chow-type tests have been proposed. Each test deals with some limitations of the Chow test, such as the poor performance in small samples, the effect of an increasing number of explanatory variables, the presence of autocorrelation in the residuals or heteroscedasticity, non-linear models and non-continuous data. See Ghilagaber (2004) for a comprehensive review. When the change time is unknown, but assumed to happen in a certain time interval, the tests proposed by Andrews (1993) and Andrews and Ploberger (1994) are very commonly used.

Another frequently used tool is to define confidence intervals. A plot of the confidence interval provides a good visualisation of the uncertainty level about the proposed statistic in the sense that wide intervals are associated with a high level of uncertainty. A special issue of the International Journal of Forecasting on time series monitoring published in 2009, summarizes the most influential developments in this area (see Gorr and Ord, 2009, for details). The majority of the existing methods for confidence interval calculation deals with one observation at a time.

Two popular approaches to construct the confidence interval non-parametrically are the bootstrap approach (see for example Hall, 1991; Diciccio and Romano, 1988) and the empirical likelihood (introduced by Owen, 1988, 1990, as the alternative to the bootstrap approach). Research by Chen (1994a,b) and Chen (1996) showed the advantages of the empirical likelihood confidence intervals over the bootstrap confidence intervals. Chen and Van Keilegom (2009) provides a review of empirical likelihood confidence interval methods and links the constructed confidence intervals to hypothesis testing.

This paper presents a very general model validation procedure (MVP). The aim of the MVP is to answer the question of whether the model under test is supported by new data. The most important property of this tool is its ability to provide an answer with only a few new observations using an approach developed by King, Zhang, and Akram (2011) for multiple hypothesis testing. A further advantage is that it does not require an alternative hypothesis to be specified.

The remainder of the paper is organized as follows. Section 2 formally describes the proposed procedure. Section 3 provides a simulation study to demonstrate the capabilities of the suggested procedure on several simple but very commonly used examples. The study includes an examination of the size and power of the proposed procedure. Finally, Section 4 concludes the paper.

2 The Model Validation Procedure (MVP)

Statistical models are typically fitted using the available data, the learning data set. Many different models can be fitted to the same data. During the fitting process we aim to approximate the DGP that generated the learning data set in the best possible way. Many statistical tools are available to examine how well the selected model fits the data. However, the further in time we get from the model fitting, the higher the probability of a change in reality[1]. As long as the DGP stays unchanged, the selected model keeps its desirable properties. Therefore, given a short sequence of new observations, we are in fact interested in being able to examine if the same DGP generated the data used for model development and the new sequence.

Supposed the chosen statistical model is of the form

$$y_t = m(\boldsymbol{x}_t, \boldsymbol{\gamma}, \varepsilon_t), \quad t = 1, \ldots, T, \tag{1}$$

where m is a known function, $\boldsymbol{x}_t$ is a vector of explanatory variable, $\boldsymbol{\gamma}$ is a vector of unknown parameters and ε_t is a parameter-free disturbance term[2]. Typically $\boldsymbol{\gamma}$ is unknown and is estimated by an appropriate estimation method based on observations $t = 1, \ldots, T$, so the working model becomes

$$y_t = m(\boldsymbol{x}_t, \hat{\boldsymbol{\gamma}}, \varepsilon_t). \tag{2}$$

We assume that N additional observations on y_t and $\boldsymbol{x}_t$ have become available and we wish to check whether our model is still supported by this new data. In what follows, let $\boldsymbol{y}_T^N$ denote the vector of new observations of $\boldsymbol{y}_t$, namely $\boldsymbol{y}_T^N = (y_{T+1}, \ldots, y_{T+N})'$ and let $\tilde{\boldsymbol{y}}_T^N$ denote the actual observed value of $\boldsymbol{y}_T^N$. If the new observations were generated by the working model, they would be drawings from the joint distribution of $\boldsymbol{y}_T^N$ implied by the working model whose density we will denote by $f(\boldsymbol{y}_T^N)$.

Our approach is to ask if the new observations could have come from this joint distribution. As noted by King, Zhang, and Akram (2011), the p-value is a useful device to use to answer this question. In our case, it is the probability under $f(\boldsymbol{y}_T^N)$ of finding a vector of observations of $\boldsymbol{y}_T^N$ as or more extreme than $\tilde{\boldsymbol{y}}_T^N$. In other words, it is the probability under $f(\boldsymbol{y}_T^N)$ that $f(\boldsymbol{y}_T^N) < f(\tilde{\boldsymbol{y}}_T^N)$ (see Hyndman, 1996).

Typically $f(\boldsymbol{y}_T^N)$ is unknown but we can readily simulate values of $\boldsymbol{y}_T^N$ by using an appropriate random number generator to generate ε_t, $t = T+1, \ldots, N$, values which are then substituted into equation (2) to obtain simulated values[3] of $\boldsymbol{y}_T^N$. Any number of $\boldsymbol{y}_T^N$ can be independently simulated which we will denote by $\boldsymbol{y}_{T,i}^N$, $i = 1, \ldots, M$. $f(\boldsymbol{y}_T^N)$ can then be estimated by any of a

[1] The original DGP that in fact generated the observed data.

[2] For example, if $u_t \sim N(0, \sigma^2)$ is an error component of the model then $\varepsilon_t = u_t/\sigma$ is the parameter-free disturbance term and u_t can be replaced by $\sigma\varepsilon_t$ in our characterisation of the model.

[3] We are effectively assuming that the distribution of ε_t is known. This assumption could be weakened to the distribution being unknown with this simulation step being replaced by an appropriate bootstrap sampling step.

number of density estimators. We favour the use of a multivariate kernel density estimator whose general form is

$$\hat{f}_{M,K}(\boldsymbol{y}_T^N) = \frac{1}{M}\sum_{i=1}^{M}|H|^{-1/2}K\left(|H|^{-1/2}(\boldsymbol{y}_T^N - \boldsymbol{y}_{T,i}^N)\right)$$

where $K(\cdot)$ is a kernel function and H is an $N \times N$ positive definite matrix of bandwidths known as the bandwidth matrix (see for example Wand and Jones, 1995).

Our proposed procedure involves calculating an estimated p-value, $\hat{p}$, by the relative frequency of

$$\hat{f}_{M,K}(\boldsymbol{y}_{T,i}^N) < \hat{f}_{M,K}(\boldsymbol{y}_T^N)$$

within the M independent drawings of $\boldsymbol{y}_T^N$ based on the working model. $\hat{p}$ can be interpreted as a p-value for a test (with no particular alternative) of the null hypothesis that $\tilde{\boldsymbol{y}}_T^N$ was generated by the working model (2). The smaller the value of $\hat{p}$, the less confidence we have in the proposition that $\tilde{\boldsymbol{y}}_T^N$ is compatible with working model (2). A reasonable interpretation would be that if $\hat{p} \geq 0.05$, then the new data has validated the working model. Because of how the test is constructed, it is optimal in the sense that it has the largest possible rejection region and therefor the smallest acceptance region with in y_t^N space for the desired level of significance.

Some users will wait for only a few new observations to become available and will choose a relatively high confidence level. Others will be more conservative, choosing to wait longer and using a lower confidence level before considering any changes in their models.

To develop some intuition about the procedure, we use a graphical illustration. Let's focus on the examination of the model's prediction capability after two new observations ($N = 2$). Figure 1 shows 100 potential realizations of a model (round and square dots) on a two dimensional plane, where values on the X axis denote one-step ahead potential realizations $\{m(\boldsymbol{x}_{T+1}, \hat{\gamma}, \varepsilon_{T+1})_n\}$ and values on the Y axis denote two-step ahead potential realizations $\{m(\boldsymbol{x}_{T+2}, \hat{\gamma}, \varepsilon_{T+2})_n\}$. The 3-dimensional mesh is the estimated joint density of the potential realizations. The asterisk denotes the actual new observation $(\tilde{y}_{T+1}, \tilde{y}_{T+2})$. The round dots represent potential realizations with higher estimated density than that of the actual observation. Similarly, square dots represent potential realizations with lower estimated density than that of the actual observation. In this example, the number of square dots is two which leads to a $\hat{p}$ of 0.02. Based on this $\hat{p}$ and choosing a significant level of 5%, we can conclude that it is likely the two new observations $(\tilde{y}_{T+1}, \tilde{y}_{T+2})$ do not belong to the same DGP as the estimated model. We have only two sequences

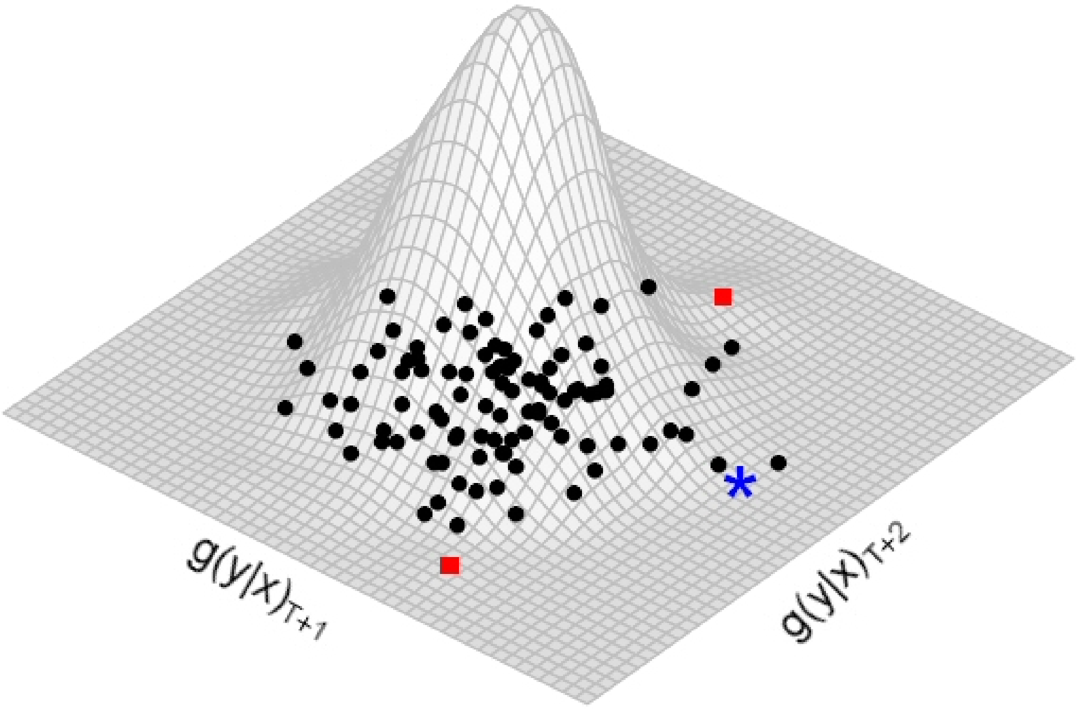

Figure 1: *Model validation procedure visualization - 2 steps ahead.*
The asterisk represents the real observation ($\tilde{y}_{T+1}, \tilde{y}_{T+2}$)*. The 3-dimensional mesh represents the estimated density of the potential realizations. The round dots represent the potential realizations with higher estimated density than the density of the actual new observation. The square dots represent the potential realizations with lower estimated density than the density of the real observation.*

of potential realizations with estimated densities lower (and therefore more extreme) than the density of ($\tilde{y}_{T+1}, \tilde{y}_{T+2}$). In other words, we can say that 98 out of 100 potential realizations being more likely than the actual two observation, is evidence against the model still being a good fit to the DGP.

3 Simulation Experiments

In order to evaluate the small sample properties of the MVP, we conducted a simulation study based on two models, the linear regression model and the autoregressive model of order one (AR(1)). The size and the power of the MVP was compared with that of Chow test (Chow, 1960) and the *AveLM* and *ExpLM* tests (Andrews and Ploberger, 1994). Because the basic assumptions of the standard Chow test do not hold for the AR(1) model, we followed Diebold and Chen (1996) and used a bootstrapped version of Chow test.

The *AveLM* and *ExpLM* tests are asymptotically optimal in the sense of maximizing weighted average power. They test for structural change at an unknown time point assumed to be within a certain interval. The p-values for each of those two tests was calculated by a two-stage approximation technique suggested by Hansen (1997). All four tests were applied by first calculating the (approximate) p-value and using it to determine acceptance or rejection of the null hypothesis at the desired significance level.

The linear model used takes the form

$$y_t = \alpha + \beta x_{1,t} + \delta x_{2,t} + \sigma \varepsilon_t, \quad t = 1, \dots, T, \tag{3}$$

where $\alpha = 3$, $\beta = 0.4$, $\delta = 0.6$, $\sigma = 0.25$, $x_{1,t} \sim t(2)$, $x_{2,t} \sim \chi^2(4)$ and $\varepsilon_t \sim N(0,1)$. In the context of model (1), m is the linear function, $\boldsymbol{x}_t = (1, x_{1,t}, x_{2,t})'$ and $\boldsymbol{\gamma} = (\alpha, \beta, \delta, \sigma)'$.

We examined three changes in this DGP starting at time $T+1$, namely

$$H_1^A: \quad y_t = \alpha + \beta x_{1,t} + \delta x_{2,t} + 0.6 \sin(0.3\pi t) + \sigma \varepsilon_t, \quad t = T+1, \dots, T+N,$$

which incorporates a new economic cycle explanatory variable into the model;

$$H_1^B: \quad y_t = 5 + \beta x_{1,t} + \delta x_{2,t} + \varepsilon_t, \quad t = T+1, \dots, T+N,$$

where $\varepsilon_t + 2 \sim \chi^2(2)$ thus incorporating a shift in the intercept plus a change in disturbance distribution; and

$$H_1^C: \quad y_t = \alpha + \beta x_{1,t} + \delta x_{2,t} + \sigma \varepsilon_t, \quad t = T+1, \dots, T+N,$$

where $\sigma = 0.5$ which means the standard deviation of the error term (σ) has been doubled. The linear model simulation experiment involved $N = 2, 5, 10$ and $T = 100, 200, 500, 1000$. The $[\pi_1, \pi_2]$ for Hansen's (1997) approximation technique were chosen so that represent the starting point for the structural break search window is $t = T+1$ and ending point of this window at $t = T+N-4$. This search window is comparable with the procedure that assumes that the break may happen only after $T+1$. Also the window leaves at least four data points to re-estimate the model after the last potential break points in the window. We ended up choosing this particular window after trialling a range of different windows. It means that the *AveLM* and *ExpLM* tests were not applied for $N = 2$ which does question the suitability of these procedures when there are only two new observations.

The AR(1) model used took the form

$$y_t = \rho y_{t-1} + \sigma \varepsilon_t,$$

where $\rho = 0.6$, $\sigma^2 = 0.25$ and $\varepsilon_t \sim N(0,1)$. In the notation of model (1), m is the linear function $\boldsymbol{x}_t = y_{t-1}$ and $\boldsymbol{\gamma} = (\rho, \sigma)'$. The changes in the DGP that the four tests were compared against involved two changes in the error variance, namely

$$H_1^I : \sigma^2 = 0.5$$

and

$$H_1^{II} : \sigma^2 = 1$$

and two in which the autoregressive coefficient changes, namely

$$H_1^{III} : \rho = 0.9$$

and

$$H_1^{IV} : \rho = -0.9.$$

For these simulations, we used $N = 2, 5, 10, 15, 20, 30$ and $T = 100, 200, 500, 1000$.

In all cases, the MVP was applied using the kernel density estimation method with diagonal bandwidth matrix (H) selected through the normal reference rule (NRR). See for example, Scott (1992, p.152)[4]. Because the generated disturbance was normally distributed, the normal distribution assumption required by the NRR is very reasonable. $M = 10,000$ simulations from the working model were used to compute the kernel distribution required to apply the MVP.

The four tests' sizes and powers were compared using 1000 Monte Carlo replications at the nominal 5% and 1% significance levels.

The size and the power results for the four tests procedures, namely the MVP, Chow test, ExpLM and AveLM tests in the context of the linear regression are presented in Tables 1 and 2.

[4] An alternative data-driven bandwidth selector is given by Zhang et al. (2006).

Table 1: *Size and power under H_1^A of the MVP, bootstrap Chow, exponential LM and average LM tests (linear model).*

Test	Size						Power under H_1^A					
	$N = 2$		$N = 5$		$N = 10$		$N = 2$		$N = 5$		$N = 10$	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
Sample size = 100												
MVP	0.060	0.017	0.073	0.019	0.073	0.022	0.768	0.564	0.895	0.738	0.969	0.912
Chow	0.051	0.014	0.047	0.007	0.043	0.009	0.741	0.533	0.579	0.409	0.457	0.277
ExpLM	-	-	0.052	0.007	0.044	0.006	-	-	0.672	0.414	0.587	0.308
AveLM	-	-	0.046	0.008	0.044	0.005	-	-	0.626	0.369	0.484	0.240
Sample size = 200												
MVP	0.057	0.017	0.063	0.014	0.066	0.015	0.782	0.559	0.877	0.713	0.957	0.881
Chow	0.051	0.018	0.058	0.014	0.054	0.008	0.766	0.555	0.565	0.401	0.554	0.333
ExpLM	-	-	0.051	0.010	0.047	0.010	-	-	0.721	0.468	0.700	0.438
AveLM	-	-	0.047	0.012	0.043	0.009	-	-	0.698	0.423	0.600	0.336
Sample size = 500												
MVP	0.055	0.015	0.054	0.012	0.061	0.015	0.781	0.555	0.902	0.752	0.976	0.904
Chow	0.056	0.019	0.055	0.011	0.043	0.007	0.791	0.563	0.646	0.462	0.560	0.380
ExpLM	-	-	0.050	0.008	0.040	0.008	-	-	0.729	0.495	0.734	0.491
AveLM	-	-	0.054	0.007	0.038	0.006	-	-	0.69 0	0.459	0.646	0.373
Sample size = 1000												
MVP	0.050	0.009	0.057	0.011	0.047	0.009	0.749	0.540	0.923	0.766	0.962	0.889
Chow	0.047	0.012	0.046	0.011	0.039	0.005	0.751	0.534	0.646	0.473	0.570	0.385
ExpLM	-	-	0.055	0.010	0.057	0.010	-	-	0.764	0.535	0.752	0.506
AveLM	-	-	0.052	0.011	0.055	0.009	-	-	0.729	0.497	0.655	0.397

H_0: $y_t = 3 + 0.4x_1 + 0.6x_2 + e_t$ with $x_1 \sim t_2$, $x_2 \sim \chi^2_{(4)}$ and $e_t \sim N(0, 0.25^2)$.
H_1^A: $y_t = 3 + 0.4x_1 + 0.6x_2 + 0.6\sin(0.3\pi t) + e_t$ with $x_1 \sim t_2$, $x_2 \sim \chi^2_{(4)}$ and $e_t \sim N(0, 0.25^2)$.

Table 2: *Powers under H_1^B and H_1^C of the MVP, bootstrap Chow, exponential LM and average LM tests (linear model).*

Test	Power under H_1^B						Power under H_1^C					
	$N = 2$		$N = 5$		$N = 10$		$N = 2$		$N = 5$		$N = 10$	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
Sample size = 100												
MVP	0.942	0.918	0.997	0.996	1.000	1.000	0.493	0.329	0.749	0.571	0.900	0.811
Chow	0.935	0.919	0.998	0.996	1.000	1.000	0.473	0.317	0.535	0.358	0.467	0.291
ExpLM	-	-	0.980	0.971	0.999	0.999	-	-	0.586	0.391	0.646	0.430
AveLM	-	-	0.980	0.970	0.998	0.994	-	-	0.571	0.357	0.567	0.339
Sample size = 200												
MVP	0.949	0.917	0.998	0.997	1.000	1.000	0.476	0.310	0.704	0.535	0.904	0.807
Chow	0.947	0.916	0.998	0.997	1.000	1.000	0.457	0.317	0.515	0.369	0.527	0.351
ExpLM	-	-	0.799	0.753	0.999	0.997	-	-	0.628	0.429	0.715	0.499
AveLM	-	-	0.980	0.970	0.998	0.997	-	-	0.605	0.403	0.642	0.407
Sample size = 500												
MVP	0.941	0.921	0.998	0.996	1.000	1.000	0.483	0.336	0.720	0.573	0.908	0.815
Chow	0.940	0.921	0.999	0.997	1.000	1.000	0.472	0.332	0.566	0.417	0.547	0.366
ExpLM	-	-	0.406	0.394	1	1	-	-	0.658	0.476	0.715	0.536
AveLM	-	-	0.995	0.993	1	1	-	-	0.630	0.442	0.645	0.459
Sample size = 1000												
MVP	0.951	0.919	0.999	0.999	1.000	1.000	0.449	0.303	0.725	0.544	0.902	0.792
Chow	0.949	0.919	1.000	0.999	1.000	1.000	0.456	0.308	0.589	0.425	0.587	0.404
ExpLM	-	-	0.329	0.318	0.999	0.999	-	-	0.617	0.468	0.752	0.558
AveLM	-	-	0.995	0.990	1	1	-	-	0.593	0.444	0.676	0.455

H_1^B: $y_t = 5 + 0.4x_1 + 0.6x_2 + e_t$ with $x_1 \sim t_2$, $x_2 \sim \chi^2_{(4)}$ and $e_t \sim \chi^2_{(2)} - 2$.
H_1^C: $y_t = 3 + 0.4x_1 + 0.6x_2 + e_t$ with $x_1 \sim t_2$, $x_2 \sim \chi^2_{(4)}$, $e_t \sim N(0, 0.5^2)$.

Looking initially at the size results, for sample sizes $T = 200$, 500 and 1000, all estimated sizes are within three standard deviations of the nominal size. In the case of the smallest sample size, namely $T = 100$, we see evidence of the MVP having slightly higher than nominal size for 5 and 10 new observations.

Turning to the power results for H_1^A, the case in which a business cycle variable has been introduced into the relationship at time $T + 1$, the MVP has the best overall power, particulary for $N = 5$ and 10 when it is often more than 50% above the corresponding power of the other tests for smaller sample sizes. As one would expect, the power of the MVP increases as N increases. The Chow test is relatively competitive with the MVP for $N = 2$ but shows an unfortunate tendency to lose power as the number of new observations increases. The ExpLM test is the next best performer for $N = 5, 10$ with the Chow test having the lowest power.

Both the MVP and Chow test have excellent power against H_1^B, particularly for the larger values of N. The main change for this alternative is the shift in the intecept which the Chow test is designed to pick up. The AveLM test has the next best power which rises to one (like the MVP and the Chow test) when N increases to 10. The power of the ExpLM test declines as T increases when $N = 5$ but it largely behaves like those of the other three tests when $N = 10$.

Against H_1^C in which the standard deviation of the regression errors jumps from 0.25 to 0.5 at time $T + 1$, the MPV clearly has the best power, particularly for $N = 5$ and 10. There are a small number of occurrences for $N = 2$ when the Chow test has best power. As might be expected, the power of the MPV test clearly improves as N increases. This is also true of the ExpLM and AveLM tests but not necessarily true of the Chow test. The ExpLM test is the second most powerful for the $N = 5$ and $N = 10$ with the Chow test being the least powerful.

In summary, the MVP has a slight question mark against its size for large N and small T but otherwise is the best test overall in terms of power, in many cases by a very big margin. The extra computational required to apply the MVP seems to be well rewarded.

The size results for the MVP, Chow, ExpLM and AveLM tests in context of the AR(1) model are presented in Table 3. For the two largest sample sizes $T = 500$ and 1000, all estimated sizes of the MVP and Chow tests are within three standard deviations of the nominal size as are all estimates sizes of the Chow test for $T = 100$ and 200. The MVP shows a tendency to have larger than nominal sizes for small T values and large N values, although there is a clear pattern of this tendency diminishing as T increases. On the other hand, the AveLM test and particularly

Table 3: *Sizes of MVP, bootstrap Chow, exponential LM and average LM tests (AR model)*

Test	$N = 2$		$N = 5$		$N = 10$		$N = 15$		$N = 20$		$N = 30$	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
Sample size = 100												
MVP	0.062	0.015	0.065	0.018	0.085	0.030	0.086	0.029	0.104	0.042	0.118	0.043
Chow	0.056	0.011	0.048	0.012	0.054	0.012	0.054	0.012	0.047	0.014	0.047	0.008
ExpLM	-	-	0	0	0.010	0.001	0.013	0.003	0.018	0	0.024	0.002
AveLM	-	-	0	0	0.015	0.001	0.022	0.003	0.028	0.003	0.041	0.003
Sample size = 200												
MVP	0.055	0.012	0.064	0.014	0.070	0.010	0.080	0.028	0.090	0.017	0.073	0.022
Chow	0.048	0.011	0.056	0.015	0.049	0.007	0.052	0.014	0.051	0.016	0.069	0.015
ExpLM	-	-	0	0	0.011	0	0.017	0.001	0.010	0	0.020	0.001
AveLM	-	-	0	0	0.016	0	0.026	0.002	0.023	0.004	0.030	0.003
Sample size = 500												
MVP	0.053	0.010	0.053	0.011	0.048	0.005	0.055	0.008	0.044	0.014	0.058	0.009
Chow	0.049	0.008	0.045	0.009	0.048	0.013	0.057	0.011	0.041	0.011	0.036	0.002
ExpLM	-	-	0	0	0.007	0	0.009	0.001	0.021	0	0.018	0.002
AveLM	-	-	0	0	0.011	0	0.023	0.001	0.029	0.001	0.029	0.003
Sample size = 1000												
MVP	0.059	0.010	0.057	0.011	0.054	0.011	0.055	0.015	0.044	0.013	0.046	0.012
Chow	0.061	0.010	0.049	0.015	0.057	0.013	0.047	0.015	0.056	0.014	0.050	0.009
ExpLM	-	-	0	0	0.012	0	0.009	0.001	0.023	0.001	0.022	0.001
AveLM	-	-	0	0	0.019	0	0.028	0.002	0.034	0.002	0.032	0.005

H_0: $y_t = 0.6y_{t-1} + e_t$ with $e_t \sim N(0, 0.25)$.

the ExpLM test tend to have smaller than nominal sizes, often being significantly smaller for smaller values of N.

Tables 4 and 5 present the powers of the four procedures for H_1^I and H_1^{II}, respectively. These two alternatives involve the variance increasing from 0.25 to 0.5 and to 1 and show similar patterns of power but with powers being higher under H_1^{II} as would be expected. For all values of N except $N = 2$, the MVP is the most powerful test and for medium and large N values, can be twice as, or even up to ten times more powerful than the next best test which is the Chow test. The MVP and the Chow test have rather similar powers when $N = 2$. The MVP increases in power as N increases while the power of Chow test typically declines as N increases. The ExpLM and AveLM tests are not designed to detect a change in the variance and not unexpectedly have power rather similar to their nominal significance levels.

Table 4: *Powers under H_1^I of the MVP, bootstrap Chow, exponential LM and average LM tests (AR model)*

Test	$N = 2$		$N = 5$		$N = 10$		$N = 15$		$N = 20$		$N = 30$	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
Sample size = 100												
MVP	0.218	0.103	0.359	0.194	0.505	0.304	0.607	0.419	0.675	0.443	0.770	0.577
Chow	0.208	0.106	0.222	0.098	0.194	0.084	0.185	0.080	0.158	0.067	0.123	0.035
ExpLM	-	-	0.002	0	0.008	0.001	0.021	0.003	0.017	0	0.029	0.003
AveLM	-	-	0.008	0	0.018	0.002	0.027	0.002	0.029	0.001	0.038	0.003
Sample size = 200												
MVP	0.212	0.090	0.364	0.185	0.505	0.302	0.579	0.342	0.675	0.446	0.749	0.539
Chow	0.207	0.092	0.214	0.090	0.228	0.093	0.201	0.088	0.175	0.073	0.148	0.060
ExpLM	-	-	0.001	0	0.016	0	0.029	0.003	0.021	0.004	0.037	0.003
AveLM	-	-	0.002	0	0.028	0	0.042	0.003	0.032	0.004	0.047	0.009
Sample size = 500												
MVP	0.242	0.104	0.367	0.190	0.485	0.271	0.614	0.378	0.663	0.439	0.769	0.554
Chow	0.244	0.105	0.250	0.130	0.229	0.098	0.231	0.095	0.195	0.075	0.212	0.101
ExpLM	-	-	0	0	0.005	0.001	0.012	0	0.024	0.002	0.025	0.003
AveLM	-	-	0	0	0.011	0.001	0.019	0	0.033	0.003	0.045	0.005
Sample size = 1000												
MVP	0.217	0.099	0.369	0.183	0.478	0.254	0.573	0.458	0.656	0.421	0.765	0.519
Chow	0.213	0.100	0.261	0.126	0.242	0.110	0.202	0.096	0.235	0.111	0.201	0.101
ExpLM	-	-	0	0	0.014	0	0.014	0	0.025	0.001	0.027	0.002
AveLM	-	-	0	0	0.022	0	0.024	0	0.039	0.004	0.042	0.005

H_1^I: $y_t = 0.6y_{t-1} + e_t$ with $e_t \sim N(0, 0.5)$.

The powers of the four procedures against H_1^{III} and H_1^{IV} are reported in Tables 6 and 7. This pair of alternatives involve ρ changing from 0.6 to 0.9 and -0.9. After acknowledging that powers against H_1^{IV} are typically higher than the corresponding power against H_1^{III}, the general pattern in both cases is as follows. Typically the Chow test is most powerful, particularly against H_1^{IV}. For small N (like $N = 2$), the MVP can sometime be the most powerful test. This is also the case against H_1^{III} for $T = 100$ and 200. The ExpLM and AveLM tests perform relatively poorly for small N but as N increases, their relative power (and particularly that of AveLM) improves with the AveLM test typically being the more powerful of the two. When the MVP is not most powerful, it is typically the second best most powerful against H_1^{III}. Against H_1^{IV}, the MVP is

Table 5: *Powers under H_1^{II} of the MVP, bootstrap Chow, exponential LM and average LM tests (AR model).*

Test	$N = 2$		$N = 5$		$N = 10$		$N = 15$		$N = 20$		$N = 30$	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
Sample size = 100												
MVP	0.478	0.300	0.740	0.591	0.911	0.789	0.948	0.888	0.979	0.928	0.998	0.988
Chow	0.460	0.288	0.467	0.288	0.377	0.222	0.321	0.183	0.289	0.138	0.198	0.090
ExpLM	-	-	0.013	0.001	0.019	0.002	0.027	0.003	0.024	0.001	0.045	0.005
AveLM	-	-	0.014	0.001	0.019	0.002	0.036	0.002	0.034	0.001	0.043	0.005
Sample size = 200												
MVP	0.476	0.312	0.733	0.563	0.894	0.784	0.965	0.892	0.983	0.956	0.993	0.975
Chow	0.478	0.313	0.477	0.311	0.435	0.283	0.401	0.248	0.339	0.205	0.296	0.145
ExpLM	-	-	0.006	0	0.024	0.002	0.031	0.003	0.027	0.005	0.044	0.010
AveLM	-	-	0.008	0	0.029	0.003	0.044	0.003	0.038	0.005	0.050	0.009
Sample size = 500												
MVP	0.482	0.345	0.735	0.576	0.915	0.795	0.951	0.894	0.984	0.951	0.997	0.989
Chow	0.476	0.346	0.498	0.353	0.480	0.340	0.435	0.294	0.412	0.269	0.420	0.251
ExpLM	-	-	0.003	0	0.009	0.002	0.015	0	0.027	0.002	0.028	0.004
AveLM	-	-	0	0	0.015	0.002	0.016	0	0.036	0.003	0.036	0.005
Sample size = 1000												
MVP	0.458	0.308	0.726	0.571	0.900	0.768	0.961	0.889	0.986	0.952	0.997	0.988
Chow	0.454	0.300	0.514	0.358	0.503	0.340	0.459	0.290	0.446	0.310	0.405	0.270
ExpLM	-	-	0	0	0.014	0.001	0.018	0	0.020	0.001	0.033	0.002
AveLM	-	-	0	0	0.022	0.002	0.026	0.001	0.038	0.001	0.046	0.004

H_1^{II}: $y_t = 0.6y_{t-1} + e_t$ with $e_t \sim N(0,1)$.

typically the second most powerful test for $T = 1000$, for $N = 2, 5, 10, 15$ as well as for $N = 20$ but only at the 1% significance level.

4 Conclusion

We have proposed a model validation procedure that is able to recognize changes in the underlying model without the need to be specific about the form of the change. The test is optimal in the sense that it has the smallest acceptance region within the sample space of the new observations. The simulation experiment reported in the previous section suggests the new procedure has comparative power against tests designed to look for certain structural changes when these

Table 6: *Powers under H_1^{III} of the MVP, bootstrap Chow, exponential LM and average LM tests (AR model).*

Test	$N = 2$		$N = 5$		$N = 10$		$N = 15$		$N = 20$		$N = 30$	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
Sample size = 100												
MVP	0.086	0.026	0.167	0.083	0.284	0.180	0.403	0.285	0.469	0.352	0.596	0.487
Chow	0.074	0.021	0.181	0.084	0.273	0.141	0.337	0.214	0.406	0.250	0.495	0.319
ExpLM	-	-	0.055	0.002	0.202	0.059	0.295	0.132	0.302	0.153	0.450	0.244
AveLM	-	-	0.064	0.002	0.228	0.065	0.316	0.129	0.296	0.142	0.434	0.210
Sample size = 200												
MVP	0.087	0.029	0.157	0.086	0.277	0.173	0.345	0.240	0.485	0.361	0.584	0.489
Chow	0.075	0.027	0.188	0.089	0.293	0.184	0.371	0.241	0.458	0.323	0.591	0.417
ExpLM	-	-	0.015	0	0.185	0.045	0.283	0.130	0.368	0.193	0.533	0.338
AveLM	-	-	0.021	0	0.213	0.049	0.303	0.130	0.385	0.185	0.538	0.325
Sample size = 500												
MVP	0.109	0.035	0.170	0.078	0.271	0.169	0.360	0.241	0.443	0.318	0.589	0.484
Chow	0.100	0.040	0.210	0.123	0.327	0.204	0.387	0.268	0.487	0.358	0.634	0.492
ExpLM	-	-	0.003	0	0.169	0.020	0.272	0.105	0.387	0.204	0.524	0.353
AveLM	-	-	0.004	0	0.210	0.029	0.289	0.119	0.404	0.223	0.547	0.344
Sample size = 1000												
MVP	0.090	0.029	0.154	0.103	0.240	0.159	0.344	0.245	0.442	0.330	0.565	0.439
Chow	0.081	0.028	0.193	0.100	0.308	0.201	0.403	0.270	0.475	0.340	0.619	0.484
ExpLM	-	-	0.001	0	0.177	0.018	0.290	0.096	0.356	0.151	0.531	0.347
AveLM	-	-	0	0	0.218	0.023	0.319	0.119	0.400	0.180	0.548	0.333

H_1^{III}: $y_t = 0.9y_{t-1} + e_t$ with $e_t \sim N(0, 0.25)$.

structural changes occur while also having much higher power than the established test against other structural changes. The only negative to emerge from the simulation experiment is that the MVP can have higher than nominal size for small T values and high N values. There could be two reasons for this. The first is that as N increases, we have greater difficulty accurately estimating $f(y_T^N)$, particulary the tails of this N-dimensional distribution. The second possible cause of the higher size is that the working model is using estimated parameters and if these are poor estimates, then the null may be rejected because the estimated parameters are significantly different from the true values. The fact that this problem goes away as the sample size used for estimation, T, increases does suggest this latter cause. It also, suggests that if the MPV rejects the null hypothesis, then one should check, as best one can, whether this rejection might have

Table 7: *Powers under H_1^{IV} of the MVP, bootstrap Chow, exponential LM and average LM tests (AR model).*

Test	$N = 2$		$N = 5$		$N = 10$		$N = 15$		$N = 20$		$N = 30$	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
Sample size = 100												
MVP	0.442	0.335	0.642	0.556	0.813	0.717	0.880	0.801	0.929	0.868	0.965	0.930
Chow	0.439	0.327	0.702	0.627	0.939	0.886	0.992	0.977	0.999	0.997	1.000	1.000
ExpLM	-	-	0.128	0.008	0.629	0.298	0.877	0.597	0.968	0.843	0.999	0.990
AveLM	-	-	0.155	0.007	0.679	0.302	0.898	0.598	0.969	0.849	0.999	0.979
Sample size = 200												
MVP	0.444	0.390	0.621	0.528	0.777	0.694	0.883	0.799	0.922	0.850	0.951	0.905
Chow	0.445	0.339	0.706	0.618	0.936	0.880	0.987	0.971	0.998	0.994	1.000	0.998
ExpLM	-	-	0.062	0.002	0.531	0.166	0.841	0.494	0.955	0.785	0.998	0.979
AveLM	-	-	0.087	0.002	0.604	0.189	0.878	0.547	0.974	0.810	0.999	0.976
Sample size = 500												
MVP	0.462	0.345	0.661	0.549	0.797	0.698	0.862	0.790	0.918	0.838	0.964	0.932
Chow	0.464	0.346	0.752	0.666	0.937	0.892	0.992	0.976	0.998	0.996	1.000	1.000
ExpLM	-	-	0.015	0	0.485	0.100	0.815	0.380	0.957	0.707	0.997	0.951
AveLM	-	-	0.026	0	0.590	0.135	0.866	0.417	0.983	0.754	0.997	0.964
Sample size = 1000												
MVP	0.437	0.388	0.668	0.557	0.794	0.704	0.863	0.778	0.913	0.856	0.946	0.904
Chow	0.444	0.339	0.728	0.682	0.946	0.909	0.990	0.975	0.999	0.998	1.000	1.000
ExpLM	-	-	0.004	0	0.470	0.039	0.791	0.335	0.935	0.643	0.998	0.935
AveLM	-	-	0	0	0.573	0.072	0.859	0.386	0.959	0.701	0.997	0.952

H_1^{IV}: $y_t = -0.9y_{t-1} + e_t$ with $e_t \sim N(0, 0.25)$.

been caused by poor parameter estimates. This could involve looking for outliers in the original sample that are highly influential, or seeing if adding the new observations to the sample used for estimation clearly improves the fit of the model under investigation. How best to modify the procedure to mitigate against this problem when T is small and N is large is a topic we are currently researching and leave for a future paper.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica 51*, 821–856.

Andrews, D. W. K. and W. Ploberger (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica 52*, 1383–1414.

Chen, S. X. (1994a). Comparing empirical likelihood and bootstrap hypothesis tests. *Journal of Multivariate Analysis 51*, 277–293.

Chen, S. X. (1994b). Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis 49*, 24–40.

Chen, S. X. (1996). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika 83*, 329–341.

Chen, S. X. and I. Van Keilegom (2009). A review on empirical likelihood methods for regression. *Test 18*, 415–447.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica 28*, 591–605.

Diciccio, T. J. and J. P. Romano (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 50*, 338–354.

Diebold, F. X. and C. Chen (1996). Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. *Journal of Econometrics 70*, 221–241.

Ghilagaber, G. (2004). Another look at Chow's test for the equality of two heteroscedastic regression models. *Quality & Quantity 38*, 81–93.

Gorr, W. L. and J. K. Ord (2009). Introduction to time series monitoring. *International Journal of Forecasting 25*, 463–466.

Hall, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics: A Journal of Theoretical and Applied Statistics 22*, 215–232.

Hansen, B. E. (1997). Approximate asymptotic p values for structural-change tests. *Journal of Business & Economic Statistics 15*, 60–67.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *American Statistician*, 120–126.

King, M. L., X. Zhang, and M. Akram (2011). A new procedure for multiple testing of econometric models. *Working paper No. 7-11, Monash University*.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*, 237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics 18*, 90–120.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. Wiley-Interscience, New York.

Wand, M. P. and M. C. Jones (1995). *Kernel smoothing*. Chapman & Hall, London.

Zhang, X., M. L. King, and R. J. Hyndman (2006). A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics & Data Analysis 50*, 3009–3031.