10ISSN 1183-1057

SIMON FRASER UNIVERSITY

Department of Economics

Working Papers

12-07

“Learning and Model Validation”

In-Koo Cho and Ken Kasa

February, 2011

SFU Economics

LEARNING AND MODEL VALIDATION

IN-KOO CHO AND KENNETH KASA

Abstract. This paper studies adaptive learning with multiple models. An agent operating in a self-referential environment is aware of potential model misspecification, and tries to detect it, in real-time, using an econometric specification test. If the current model passes the test, it is used to construct an optimal policy. If it fails the test, a new model is selected from a fixed set of models. As the rate of coefficient updating decreases, one model becomes dominant, and is used 'almost always'. Dominant models can be characterized using the tools of large deviations theory. The analysis is applied to Sargent's (1999) Phillips Curve model.

JEL Classification Numbers: C120, E590

If only good things survive the tests of time and practice, evolution produces intelligent design. – Sargent (2008, p.6)

1. Introduction

Macroeconomic policymaking confronts two challenges. First, policymakers do not know the model. That is, they do not know the economy's actual data-generating process (DGP). Second, their beliefs typically feedback to influence the model. That is, policy operates within a 'self-referential' system.

Considered separately, each of these difficulties can be handled using standard methods. For example, during the past two decades classical econometricians have developed methods that permit testing and comparison of misspecified models.[1] At the same time, there has been an explosion of recent work on Bayesian macroeconometrics.[2] However, neither of these approaches, at least as conventionally implemented, permits feedback from beliefs about models to the actual DGP.

Feedback is not new to economists either. It lies at the heart of the Lucas Critique, and macroeconometricians have developed methods for coping with it (Lucas and Sargent (1981)). However, these methods presume that agents within the model already know the model. It is only the outside econometrician who doesn't know the model. Although the macroeconomic learning literature is designed to relax this assumption, to date it

Date: February, 2011.

We thank Jim Bullard, Lars Hansen, Seppo Honkapohja, Albert Marcet, and Tom Sargent for helpful discussions. Financial support from the National Science Foundation (ECS-0523620, SES-0720592) is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[1]A highly selected sample includes: White (1982), Vuong (1989), and Sin and White (1996). White (1994) and Burnham and Anderson (2002) contain textbook treatments.

[2]An and Schorfheide (2007) contains a summary.

either presumes a priori knowledge of the correct model (up to a small handful of unknown parameters), or it assumes that any model misspecification goes unquestioned and undetected.[3]

Our paper attempts to confront both these challenges simultaneously. Its goal is to understand how feedback between model evaluation procedures and the DGP influences macroeconomic model selection. In his response to the Lucas Critique, Sims (1982) argued that policymakers do not need to know the correct model of the economy in order to formulate good policy. According to Sims, as long as policy is flexible and adaptive, policymakers will learn to do the right thing.[4] However, it is not at all obvious whether this is indeed the case when the data are endogenous. Sargent (1999) and Sargent and Williams (2005) study this question in a setting where the model is fixed, but parameters can vary. Their findings point to the importance of priors concerning parameter drift. Our analysis pursues the same question, but in a setting with multiple models. Here there is greater scope for getting stuck in suboptimal Self-Confirming Equilibria (SCE). As in Sargent (2008) and Fudenberg and Levine (2009), we study the role of experimentation in selecting favorable SCE.

Our particular strategy for doing this consists of the following four steps: (1) We assume an agent is endowed with a fixed *set* of models, each containing a collection of unknown parameters. The models may be misspecified and non-nested; (2) Each period the agent tests the specification of his current model; (3) If the current model survives the test, the model is updated and used to formulate a policy function, under the provisional assumption that the model will not change in the future, and (4) If the model is rejected, a new model is randomly selected. We refer to this combined process of estimation, testing, and selection as *model validation*. Our goal is to characterize the dynamics of this model validation process.

This is a challenging exercise. With endogenous data and multiple models, each with adaptively estimated coefficients, the underlying state of the economy is of very high dimension, and it evolves nonlinearly. The key to making this system tractable is to exploit the fact that under certain conditions subsets of the variables evolve on different time-scales. By appropriately averaging over each subset, we can simplify the analysis to one of studying the interactions between lower dimensional subsystems. This is a commonly employed strategy in science, going back to early 19th century celestial mechanics. It also received some attention from economists during the Cowles Commision era (e.g., Simon and Ando (1961)), but it wasn't until Marcet and Sargent (1989) that it began to be applied in the macroeconomic learning literature.

Our analyis extends the work of Marcet and Sargent (1989). We show that model validation dynamics feature a hierarchy of *three* time scales. This hierarchy of time-scales permits us to focus separately on the problems of control, model revision, and model selection. As in Marcet and Sargent (1989), economic variables evolve on a 'fast', calendar time-scale, whereas coefficient estimates evolve on a 'slow', model revision time-scale. The new element here is that under appropriate assumptions on specification testing, model selection occurs on a 'really slow', model switching time-scale. Model switches are rare

[3]See Evans and Honkapohja (2001) for a survey of the macro learning literature.

[4]See, in particular, pg. 117 of Sims (1982).

here, because they are triggered by departures from a model's self-confirming equilibrium, and are therefore 'large deviation' events. The fact that each model's coefficients can be adapted to fit the data it generates is crucial to this result, and it illustrates a key difference between specification testing with endogenous data and specification testing with exogenous data.

We show that model selection dynamics can be approximated by a low dimensional Markov chain, in which each model's coefficients are fixed at their self-confirming values, and the economic data are fixed at the mean of the invariant distribution associated with these values. In the limit, as the update gain parameter converges to zero, the invariant distribution of this Markov chain collapses onto a *single* model. We can identify this model from its large deviations rate function. In a sense, therefore, our analysis provides an equilibrium selection criterion for recursive learning models. It can also be interpreted as a refinement of the concept of self-confirming equilibria.

Large deviation methods provide an interesting and useful interpretation of this limiting model. Specifically, we show that it is the model possessing the largest so-called 'rate function'. A key result in the theory of large deviations (i..e, Sanov's theorem) links this rate function to relative entropy and the Kullback-Leibler Information Criterion (KLIC). This is interesing, since the KLIC is a pervasive concept in the econometrics literature on model testing and selection. The relative entropy that is being captured by each model's rate function is the KLIC distance between the probability distribution associated with its SCE and the distribution associated with the closest model that triggers a rejection or escape. This extends the results of White (1982) in a natural way to the case of endogenous data.

The remainder of the paper is organized as follows. Section 2 provides a brief overview of some new issues that arise when combining model uncertainty with adaptive learning. Section 3 maps our model validation approach into a standard Stochastic Recursive Algorithm. Section 4 uses results from the large deviations literature to characterize model validation dynamics. Section 5 provides an example. We revisit Sargent's (1999) well known *Conquest* model. We show when central banks are aware of model uncertainty, and respond to it by testing the specification of their models, inflation dynamics depend critically on the initial model class. Section 6 contains a few concluding remarks, and an appendix contains proofs of some technical results.

2. Overview

Incorporating model uncertainty into the learning literature raises a host of new questions and issues. This section briefly outlines how our approach to model validation addresses these issues.

2.1. Parameter Uncertainty vs. Model Uncertainty

The existing learning literature focuses on parameter uncertainty. For Bayesians, however, there is no essential difference between parameter uncertainty and model uncertainty. By formulating a single, all-encompassing, 'hypermodel', Bayesians convert model uncertainty into parameter uncertainty. We do not do this. Our agent uses his model to

construct a policy function. To make this tractable, models must be relatively simple and parsimonious.[5]

2.2. The Model Class

While our agent is not committed to a single model, he is committed to a single set of models, called the *model class*. This set is exogenously specified and fixed over time. Where does it come from? That's an important question we do not address. Although our agent can effectively dispose of (relatively) bad models, he cannot create new models in response to unanticipated events.[6]

2.3. Feedback

The fact that the data-generating process responds to the agent's own beliefs is of course a crucial issue even without model uncertainty. It means that all the classical econometric results on convergence and consistency of least-squares estimators go out the window. Developing methods that allow one to rigorously study the consequences of feedback has been a central accomplishment of the macroeconomic learning literature, at least from a technical standpoint. (Evans and Honkapohja (2001) summarize this literature.)

When one turns to *inference*, however, new issues arise. First, the presence of feedback means that we cannot directly apply recent econometric advances in testing and comparing misspecified models (White (1994)). Although we assume the agent is aware of these advances, and tries to implement them, we cannot appeal to known results to study their consequences. Second, traditionally it has been assumed that agents are unaware of feedback. Although beliefs are revised in an adaptive and purposeful manner, this adaptation is strictly passive.[7] This is a reasonable assumption in the context of learning the parameters of a single model, mainly because one is already confined to a local analysis. With multiple models, however, the distinction between local and global analysis becomes far more important, and therefore, assumptions about the agent's awareness of feedback become more important. We depart from tradition here by assuming that the agent is aware feedback. In particular, he realizes that with model uncertainty he confronts a difficult counterfactual - How would things have been different if instead a different model

[5]Of course, it is possible to use Bayesian methods to select among a set of simple models. Bayesian purists tend to frown upon this practice, however. Also, as discussed below, our model validation approach is based more on specification testing than on model comparison.

[6]Jovanovic (2009) discusses how one might expand a model class in response to unforseen events.

[7]There have been a few notable exceptions. The early work of Bray and Savin (1986) touched on this issue, asking whether agents could use standard diagnostics, like Chow tests and Durbin-Watson statistics, to detect the parameter variation that their own learning behavior generates. Bray and Savin (1986) found that when convergence is slow, agents are generally able to detect the misspecification of their models. Bullard (1992) and McGough (2003) studied convergence and stability when the agent's Perceived Law of Motion allows for time-varying parameters. McGough (2003) showed that convergence to Rational Expectations can still occur as long as this time-variation is expected to damp out at a sufficiently rapid rate. Finally, and perhaps most closely related to our own work, Sargent and Williams (2005) showed that priors about parameter drift have a strong influence on the large deviation properties of constant gain learning algorithms. However, all this prior work takes place within the confines of a single model.

had been used in the past? Fitting a model to data that was generated while a *different* model was in use could produce misleading inferences about the prospects of a given model. For the questions that we address, it is not important how exactly the agent responds to this counterfactual. What's important is that he is aware of its dangers, and takes steps to avoid becoming trapped in suboptimal self-confirming equilibria.

2.4. Model Comparison vs. Specification Testing

We assume the agent sticks with a model until sufficient evidence mounts against it. An alternative strategy would be to run a (recursive) horserace between models, by continuously comparing their relative performance. In this case, one might switch models even if the currently used model appears to be well specified. Our choice of specification testing reflects three main factors: (1) We think it is an accurate description of policymaking in most cases, (2) Specification testing can be easily embedded within a standard Stochastic Recursive Algorithm. In particular, the orthogonality condition that drives parameter updating can be interpreted as a score statistic, or equivalently, a localized likelihood ratio statistic, which can be used as the basis of a sequential Lagrange Multiplier test. (See, e.g., Chapter 5 of Benveniste, Metivier, and Priouret (1990)), and (3) The resulting analysis is easier. In Cho and Kasa (2009) we consider the case of recursive model comparison.

2.5. Model Switching vs. Parameter Revision

Adding model uncertainty does not eliminate parameter uncertainty. We continue to assume that each model's parameters are adaptively updated using a constant gain stochastic approximation algorithm. A constant gain recognizes the potential existence of slow parameter drift. What is new here is the agent's recognition that more drastic and sudden changes to the underlying environment may also occur. These are signalled by an excessively large score statistic. When the score statistic exceeds a given threshold, it indicates that required parameter changes are faster and larger than specified by the underlying null hypothesis of gradual parameter drift.[8]

2.6. Escape Dynamics, Type I Errors, and the Robustness of Self-Confirming Equilibria

We assume for simplicity that each model, when used, has a unique, stable, self-confirming equilibrium. This means that each model, if given the chance, is capable of passing the specification test. Of course, this does not imply that it is the 'true' data-generating process. In fact, the entire model class may be misspecified. However, with endogenous data, each model can adapt to fit the data that it itself generates. It is this possibility that wreaks havoc with the application of traditional statistical results.

Although all models are capable of passing the test, they are not all equally likely to do so on a repeated basis. Some models are more attached to their self-confirming equilibrium, while others are more apt to drift away. Model drift is driven by the fact

[8]Another possible response to an excessively large score statistic would be to allow the update gain to increase. See Kostyshyna (2010) for an analysis of this possibility.

that coefficient estimates drift in response to constant gain updating. We calibrate the testing threshold so that this kind of normal, gradual, parameter drift does not trigger model rejection. However, as first noted by Sargent (1999), constant gain algorithms also feature rare, but recurrent, 'large deviations' in their sample paths. These large deviations can be characterized analytically by the solution of a *deterministic* control problem. It is these rare 'escapes' from the self-confirming equilibrium that trigger model rejections. In a sense then, model rejections here are Type I errors.[9]

The value function of the large deviations control problem is called the 'rate function', and as you would expect, it depends sensitively on the tails of the score statistic. In Section 4 we show that as the update gain decreases the model with the largest rate function becomes dominant, in the sense that it is used 'almost always'. This bears some resemblence to results in the evolutionary game theory literature (Kandori, Mailath, and Rob (1993)). It also provides a selection criterion for models with multiple stable self-confirming equilbria.

2.7. Experimentation

When a model is rejected we assume the agent randomly selects a new model (which may turn out to be the existing model). This randomness is *deliberate*. It does not reflect capriciousness or computational errors, but instead reflects a strategic response to model uncertainty (Foster and Young (2003)). It can also be interpreted as a form of experimentation. Of course, macroeconomic policymakers rarely conduct explicit experiments, but they do occasionally try new things. Although the real-time dynamics of model selection naturally depend on the details of the experimentation process, our main conclusions about the stability and robustness of self-confirming equilibria do not.

3. A General Framework

Consider an agent who must solve a Linear Quadratic Regulator (LQR) without knowing the state transition equation. Traditional adaptive control methods are designed to handle coefficient uncertainty, but not model uncertainty. Recently developed robust control methods are designed to handle general forms of model uncertainty, *assuming* the agent starts with an exogenously specified reference model against which all uncertainty can be measured. (See Hansen and Sargent (2007b)). Bayesian methods could in principle be applied, but would require specification of high dimensional prior distributions. Moreover, Bayesian methods lead naturally to model averaging, and encounter well known difficulties when used to select models. None of these methods easily accommodate feedback. Our approach is to assume the agent starts with a *model class*, and then tries to sort out, in real-time, among this class using traditional specification testing methods.[10]

[9]Note, however, that with endogenous data the concept of Type I error becomes somewhat ambiguous, since the 'true model' depends on the agent's beliefs.

[10]Gilboa, Postlewaite, and Schmeidler (2008) argue that classical, frequentist-based econometric methods are actually more in keeping with recent developments in decision theory than are Bayesian methods.

3.1. **Objective Function.** The agent cares about an $s \times 1$ vector of state variables, x_t, and believes he can influence them by choosing a $c \times 1$ vector of control variables, u_t. Preferences over these variables are ordered by the following Linear-Quadratic loss function,

$$E_n \sum_{j=0}^{\infty} \delta^j \{x'_{n+j} Q x_{n+j} + u'_{n+j} R u_{n+j}\} \tag{3.1}$$

where Q and R are $s \times s$ and $c \times c$ positive definite matrices, and δ is a discount factor. In what follows, it is important to keep in mind that the expectations operator here pertains to the beliefs of the decision maker. The probability measure associated with these beliefs may differ from the probability measure generating the data, and it may evolve over time.

3.2. **Models.** The agent interprets the mapping between u_n and x_n using a set of candidate models, $\mathcal{M}$, containing m elements. Each model is linear, and is described by a finite collection of unknown parameters. The models may be nested or non-nested. Our maintained assumption is that the agent believes there to be a single 'best' model within this class, and the goal is to find it, while at the same time balancing the ongoing pressures of meeting the control objective.

We assume the agent entertains the possibility that the state variables of interest, x_n, are embedded within some larger system, which in addition to x_n, contains an $n \times 1$ vector of exogenous variables, z_n. It is possible these exogeneity restrictions are misspecified. In effect then, each model $i = 1, 2, \cdots, m$ is defined by a set of s variable selection matrices, e^i_j $j = 1, 2, \cdots, s$, specifying which of the elements of x_n and z_n it contains, and a set of s control selection matrices, $e^i_{u,j}$, specifying which of the control variables are relevant for each equation. If we then define the $(s+n) \times 1$ stacked vector $y_n = (x_n^{'}, z_n^{'})'$, we can express the elements of $\mathcal{M}$ as follows

$$x^i_{jn} = \alpha^i_j e^i_j y_{n-1} + \gamma^i_j e^i_{u,j} u_{n-1} + \varepsilon^i_{jn} \qquad i = 1, 2, \cdots, m \qquad j = 1, 2, \cdots, s \tag{3.2}$$

As usual, there is no loss in generality in assuming a single lag of y_t in each state transition equation, as higher order lags can be accommodated by increasing the dimensionality of the state. What is essential, and in fact what is the defining feature of each model, is the assumption that the error terms, ε^i_{jn}, are i.i.d. and uncorrelated with the regressors. The specification test will be designed to check the accuracy of this assumption.

3.3. **Data-Generating Process.** We assume there is an underlying time invariant data-generating process (DGP). The agent is trying to learn about this process. What makes this problem difficult is that the DGP *responds* to the agent's model selection and control decisions. Of course, if the agent knew the model, he could follow the advice of Lucas (1976) and account for this endogeneity in the formulation of his policies. Without knowing the model, however, it is not at all clear how he should proceed. The danger is that he gets trapped in a 'bad' self-confirming equilibium. One response would be to assume the agent is simply unaware of this danger. Our view is that policymakers are in fact often aware of feedbacks, but respond to them informally. Good policymakers know intuitively

that is important to occasionally try new things. We model this process very simply by assuming that model selection is subject to some deliberate randomness.[11]

Letting $\beta_n = (\alpha^i_{jn}, \gamma^i_{jn})$ be the full vector of model coefficient estimates at time-n, we can capture the feedbacks from beliefs to the DGP by writing it as follows,

$$\begin{aligned} x_n &= A_1(s_{n-1}, \beta_{n-1})x_{n-1} + B_1(s_{n-1}, \beta_{n-1})z_{n-1} + C_1(s_{n-1}, \beta_{n-1})u_n + D_1(s_{n-1}, \beta_{n-1})v_{1n} \\ z_n &= A_2(s_{n-1}, \beta_{n-1})z_{n-1} + B_2(s_{n-1}, \beta_{n-1})x_{n-1} + C_2(s_{n-1}, \beta_{n-1})u_n + D_2(s_{n-1}, \beta_{n-1})v_{2n} \end{aligned} \tag{3.3}$$

where s_n is an indictor for the model being used at time-n. The coefficient functions encoding the feedback from beliefs to the actual DGP are case specific, and can be quite complex and highly nonlinear. Fortunately, all we need is the following assumption,

Assumption 3.1. *For fixed $(\beta_n, s_n) = (\bar{\beta}, \bar{s})$, the joint distribution of x_n and z_n is stationary and ergodic.*

Our analytical methods rely heavily on the ability to 'average out' fluctuations in x_n and z_n for given values of the model coefficients and model indicator. They will not work without this assumption.

3.4. **Model Updating.** Models are updated in the standard way, using a (discounted) recursive least-squares algorithm. Unfortunately, writing this algorithm in the conventional way requires a little notation. Start by writing 3.2 as follows

$$\begin{aligned} x^i_{jn} &= (\alpha^i_j, \gamma^i_j) \begin{pmatrix} e^i_j & 0 \\ 0 & e^i_{uj} \end{pmatrix} \begin{pmatrix} y_{n-1} \\ u_{n-1} \end{pmatrix} + \varepsilon^i_{jn} \\ &= \beta^i_j E^i_j \phi_{n-1} + \varepsilon^i_{jn} \end{aligned}$$

where E^i_j collects all the regressors in the jth equation of model i, and β^i_j collects all their coefficients. Then, if we place all s equations into the $1 \times s$ vector x^i_n, we can write

$$x^i_n = \beta^{i\prime} \Phi^i_{n-1} + \varepsilon^i_n \qquad i = 1, 2, \cdots, m$$

where $\Phi^i_{n-1} = \text{diag}[E^i_1 \phi_{n-1}, E^i_2 \phi_{n-1}, \cdots, E^i_s \phi_{n-1}]$. Using this notation, we have the update equations

$$\hat{\beta}^i_n = \hat{\beta}^i_{n-1} + \epsilon (R^i_{n-1})^{-1} \Phi^i_{n-1} (x^{i\,\prime}_n - \Phi^{i'}_{n-1} \hat{\beta}^i_{n-1}) \tag{3.4}$$

$$R^i_n = R^i_{n-1} + \epsilon (\Phi^i_{n-1} \Phi^{i'}_{n-1} - R^i_{n-1}) \tag{3.5}$$

where ϵ is a constant gain parameter, assumed to be common across equations and models. To facilitate the analysis, we assume that the set of feasible coefficients for each model satisfies some regularity conditions.

Assumption 3.2. *Let $\mathcal{B}^i$ be the set of all feasible coefficients for model i. We assume that $\mathcal{B}^i$ is compact and convex.*

[11] Fudenberg and Levine (2009) discuss the importance of experimentation in macroeconomic policy. On the other hand, Cogley, Colacito, Hansen, and Sargent (2008) show that a concern for robustness in the face of model uncertainty may temper the motive to experiment.

The compactness assumption can be interpreted as a priori knowledge about the DGP. Although the agent does not know the precise value of the coefficients, he can rule out outrageously large coefficients. These bounds can be enforced algorithmically by a 'projection facility' (see, e.g., Kushner and Yin (1997)). Convexity is mainly a technical assumption, designed to address the learning dynamics along the boundary of the parameter space.

3.5. **Mean ODEs and Self-Confirming Equilibria.** Notice the x_n vector on the right-hand side of (3.4) corresponds to the actual law of motion given by (3.3), which depends on both the current model and the agent's control and the estimation efforts. This makes the agent's problem self-referential. It also makes the analysis of this problem difficult. To simplify, we exploit a two-tiered time-scale separation, one between the evolution of the data, Φ_n, and the evolution of each model's coefficient estimates, $\hat{\beta}^i_n$, and another between the evolution of the coefficient estimates and the rate of model switching, s_n. A key concept when doing this is the notion of a mean ODE, which is obtained by following four steps: (1) Substitute the actual law for x_n given by (3.3) into the parameter update equations in (3.4), (2) Freeze the coefficient estimates and model indicator at their current values, (3) Average over the stationary distribution of the 'fast' variables, Φ_n, which exists by Assumption 1, and (4) Form a continuous time interpolation of the resulting autonomous difference equation, and then obtain the mean ODE by taking limits as $\epsilon \to 0$.

Assumption 3.1 assures us that this averaging is well defined. We also need to make sure it is well behaved. To facilitate notation, write the update equations for model-i as follows,

$$\beta^i_n = \beta^i_{n-1} + \epsilon H_i(\beta^i_{n-1}, s_{n-1}, \Phi^i_{n-1}, v_n)$$

where $v_n = (v_{1n}, v_{2n})$ is the shock vector. Let $P^k_{\beta,s}(\Phi_{n+k} \in \cdot | \Phi_n)$ be the k-step transition probability of the data for fixed values $\beta_n = \beta$ and $s_n = s$, and let $\Gamma_{\beta,s}(d\xi) = \lim_{k\to\infty} P^k(\cdot)$ be its ergodic limit. Define the function

$$h_i(\beta^i, s) = \int H_i(\beta^i, s, \xi)\Gamma_{\beta,s}(d\xi) \tag{3.6}$$

We impose the following regularity condition on $h(\cdot)$,

Assumption 3.3. *For all i and s, $h_i(\beta^i, s)$ is a bounded, Lipschitz continuous function of β^i.*

Continuity is essential for our averaging strategy to work. Note that since the parameter space is bounded, Assumption 3.3 implies h is bounded.

A subtlety arises here from model switching. We assume that after a model is rejected it continues to be fit to data generated by other models. As a result, there is no guarantee that while a model is not being used its coefficient estimates remain in the neighborhood of its self-confirming equilibrium. Instead, its estimated coefficients tend to gravitate to some *other* self-confirming equilibrium, one that satisfies the model's orthogonality condition given that data are being generated by another model. However, as long as model rejections occur on a slower time-scale than coefficient updating, we can apply the same averaging principle as before, the only difference is that now for each model we obtain a *set* of m mean ODEs (the next section contains a formal proof),

$$\dot{\beta}^i_s = h_i(\beta^i_s, s) \qquad s = 1, 2, \cdots m \tag{3.7}$$

where β_s^i denotes model i's coefficient estimates given that model s is generating the data. Note that when model $s \neq i$ generates the data, its coefficent estimates influence model-i's mean ODE. This is because the rate of coefficient updating is assumed to be the same for all models. A self-confirming equilibrium for model-i, $\beta_{i,s}^*$, is defined to be a stationary point of the mean ODE in (3.7), i.e., $h_i(\beta_{i,s}^*, s) = 0$. Note that in general it depends on which model is generating the data. To simplify the analysis, we impose the following assumption

Assumption 3.4. *Each model $i = 1, 2, \cdots, m$ has a unique vector of globally asymptotically stable Self-Confirming Equilibria, $\beta_{i,s}^*$, $s = 1, 2, \cdots, m$*

This is actually stronger than required. All we really need is global asymptotic stability of $\beta_{i,i}^*$, $\forall i$. This guarantees that after a period of disuse, each model converges back to its own unique self-confirming equilibrium. If this *weren't* the case, then the problem would lose its Markov structure. However, given the boundedness we've already imposed on the parameter space, we don't need to worry too much about how a model's coefficient estimates behave while other models generate the data.[12]

3.6. **Model Validation.** There is no single best way to validate a model. The right approach depends on what the model is being used for, and the nature of the relevant alternatives. In this paper we apply a Lagrange Multiplier (LM) approach. LM tests can be interpreted as Likelihood Ratio tests against local alternatives, or as first-order approximations of the Kullback-Leibler Information Criterion (KLIC). Their defining feature is that they are based solely on estimation of the null model, and do not require specification of an explicit alternative. As a result, they are often referred to as *mis*specification tests. Benveniste, Metivier, and Priouret (1990) (BMP) outline a recursive validation procedure based on LM testing principles. Their method is based on the observation that the innovation in a typical stochastic approximation algorithm is proportional to the score vector. Essentially then, what is being tested is the significance of the algorithm's update term.

Our approach is similar to that of BMP, except our null and alternative hypotheses are slightly different. BMP fix a model's coefficients and adopt the null hypothesis that the score vector is zero when evaluated at these fixed values. A rejection indicates that the coefficients (or something else) must have changed. In our setting, with multiple models and endogenous data, it is not always reasonable to interpret nonzero score vectors as model rejections. It takes time for a new model to converge to its own self-confirming equilibrium. While this convergence is underway, a model's score vector will be nonzero, as it reflects the presence of nonzero mean dynamics. We want to allow for this drift in our null hypothesis. One possible way to do this would be to incorporate a 'burn in' period after model switching, during which no testing takes place. The idea would be to give new models a chance to adapt to their own data. Another possibility would be to only update models while they are in use. Neither of these approaches seem to be widely applied in practice. Instead, we incorporate drift into the null by using a decreasing test threshold. The initial value must be sufficiently tolerant that new models are not immediately rejected, despite having drifted away from their own self-confirming equilibrium values while other models are used. On the other hand, as a model converges

[12]Of course, this wouldn't be true if the agent recursively *compared* models (Cho and Kasa (2009)).

the test becomes more stringent and the threshold decreases, as confidence in the model grows. We assume the threshold's initial level and rate of decrease are calibrated so that model rejections are rare events.

To be more explicit, let $\Lambda_{i,n}$ denote the sequence of model-i's (scaled) scores, given by

$$\Lambda_{i,n} = (R^i_{n-1})^{-1}\Phi^i_{n-1}(x^{i\,\prime}_n - \Phi^{i\prime}_{n-1}\hat{\beta}^i_{n-1})$$

The null hypothesis is then, $H_0 : \Lambda_{i,n} \leq \bar{\theta}_n$, where the threshold, $\bar{\theta}_n$, decreases with n. Keep in mind this is a *sequential* test, much like the well know CUSUM test of Brown, Durbin, and Evans (1975), or the 'monitoring stuctural change' approach of Chu, Stinchcombe, and White (1996). Hence, another reason to have a tolerant threshold is to control for the obvious size distortions induced by repeated testing.[13]

3.7. **Model Selection.** When the LM test is rejected a new model is randomly selected. Our main conclusions are robust to the details of this selection process. The only essential feature is that the support of the distribution remain full, i.e., all models must remain on the table. This ensures a form of ergodicity that is crucial for our results. In what follows we simply assume that post-rejection model selection probabilities are denoted by π^i_n, where $\pi^i_n > 0\ \forall i, n$. Hence, the most recently rejected model may be reselected, although this probability could be made arbitrarily small.

4. Analysis

Let $\mathcal{X}_n = (x_n, z_n, \hat{\beta}^1, \hat{\beta}^2, \cdots \hat{\beta}^m, \theta_n, \bar{\theta}_n, s_n)$ denote the period-n state of the economy. It consists of: (i), the current values of all endogenous and exogenous variables, (ii), the current values of all coefficient estimates in all models, (iii), the test statistic for the currently used model, (iv), the current test threshold, and (v), the current model indicator. Clearly, in general this is a high dimensional vector. In this section, we show how the dimensionality of the state can be greatly reduced. In particular, we show how the evolution of the state can be described by the interaction of three smaller subsystems. None of these subsystems are Markov when studied in isolation on a calendar time scale. However, they can be shown to be *approximately* Markov when viewed on the appropriate time-scales. The analysis therefore consists of a sequence of convergence proofs.

We begin by casting model validation in the form of a Stochastic Recursive Algorithm (SRA). We then approximate the evolution of the coefficient estimates under the assumption that model switching takes place at a slower rate than coefficient updating. Next, conditions are provided under which this assumption is valid, in the sense that model rejections become large deviation events. Fourth, we prove that model switching dynamics can be approximated by a low-dimensional Markov chain. Finally, using this Markov chain approximation, we show that in general the limiting distribution across models collapses onto a single, dominant model, which we then characterize using the tools of large deviations theory.

[13]As stressed by both Brown, Durbin, and Evans (1975) and Chu, Stinchcombe, and White (1996), an *optimal* threshold would distribute Type I error probabilities evenly over time, and would result in an increasing threshold. In fact, with an infinite sample, the size is always one for any fixed threshold. The fact that our agent discounts old data effectively delivers a constant sample size, and diminishes the gains from an increasing threshold.

4.1. **Representation as a Stochastic Recursive Algorithm.** . We have purposely stayed as close as possible to the standard SRA framework.[14] These models feature an interplay between beliefs and outcomes. Our model validation framework features these same elements, but incorporates model testing and selection dynamics as well. It is useful to begin by collecting together the model's equations:

We first have a set of model update equations,

$$\hat{\beta}_n^i = \hat{\beta}_{n-1}^i + \epsilon \Lambda_n^i \tag{4.8}$$

$$\Lambda_n^i = (R_{n-1}^i)^{-1} \Phi_{n-1}^i (x_n^{i\,\prime} - \Phi_{n-1}^{i'} \hat{\beta}_{n-1}^i) \tag{4.9}$$

$$R_n^i = R_{n-1}^i + \epsilon(\Phi_{n-1}^i \Phi_{n-1}^{i'} - R_{n-1}^i) \tag{4.10}$$

Through feedback, these help to determine the actual DGP,

$$x_n = A_1(s_{n-1}, \beta_{n-1}) x_{n-1} + B_1(s_{n-1}, \beta_{n-1}) z_{n-1} + C_1(s_{n-1}, \beta_{n-1}) u_n + D_1(s_{n-1}, \beta_{n-1}) v_{1n} \tag{4.11}$$

$$z_n = A_2(s_{n-1}, \beta_{n-1}) z_{n-1} + B_2(s_{n-1}, \beta_{n-1}) x_{n-1} + C_2(s_{n-1}, \beta_{n-1}) u_n + D_2(s_{n-1}, \beta_{n-1}) v_{2n}$$

Models are tested by forming the recursive LM test statistics

$$\theta_n^i = \theta_{n-1}^i + \epsilon[\Lambda_n^{i\,\prime} \hat{\Omega}_{i,n}^{-1} \Lambda_n^i - \theta_{n-1}^i] \tag{4.12}$$

where $\hat{\Omega}_{i,n} = (R_{n-1}^i)^{-1} \Phi_{n-1}^i \hat{\Sigma}_n^i \Phi_{n-1}^{i'} (R_{n-1}^i)^{-1}$ is a recursive estimate of the variance of Λ_n^i, and where where $\hat{\Sigma}_n^i$ is an $s \times s$ diagonal matrix containing recursive estimates of the variances in model-i's equations, estimated by recursively summing each equation's squared residuals. Hence, θ_n^i is just a recursively estimated χ^2 statistic. If a model contains all variables, its degrees of freedom would be $s(s+n+c)$, but in practice it would be much smaller than this if models are constrained to be small.

Finally, the model indicator, s_n, evolves as an m-state Markov Chain. Its evolution depends on the evolution of the test statistic relative to the threshold, as well as the model selection distribution

$$s_{n+1} = \mathcal{I}_{(\theta_n \leq \bar{\theta}_n)} \cdot s_n + (1 - \mathcal{I}_{(\theta_n \leq \bar{\theta}_n)}) \cdot \Pi_{n+1} \tag{4.13}$$

where $\mathcal{I}$ is an indicator function, and Π_n is a model selection distribution with elements $\{\pi_n^i\}$, $i = 1, 2, \cdots m$. Let $p_n \in \Delta^{m-1}$ be the time-n probability distribution over models, and let $\mathcal{P}_n$ be an $m \times m$ state transistion matrix, where $\mathcal{P}_{ij,n}$ is the time-n probability of switching from model i to model j. Model selection dynamics can then be represented as follows

$$p_{n+1}^{'} = p_n^{'} \mathcal{P}_n \tag{4.14}$$

The diagonal elements of $\mathcal{P}_n$ are given by

$$\text{Prob}[\theta_n^i \leq \bar{\theta}_n)] + \text{Prob}[\theta_n^i > \bar{\theta}_n)] \cdot \pi_n^i \tag{4.15}$$

and the off-diagonals are given by

$$\text{Prob}[\theta_n^i > \bar{\theta}_n] \cdot \pi_n^j \tag{4.16}$$

where $\bar{\theta}_n$ is a decreasing test threshold.

[14] Benveniste, Metivier, and Priouret (1990) and Evans and Honkapohja (2001) contain good textbook treatments of SRA methods.

4.2. **Time-Scale Separation.** Equations (4.8) - (4.16) constitute a high-dimensional system of nonlinear stochastic difference equations. Shedding analytical light on this system would seem to be a hopeless task. The key to making the system tractable is the application of so-called 'singular perturbation' methods, which exploit the fact that subsets of the variables evolve on different time-scales. By appropriately averaging over subsets of the variables, we can simplify the analysis to one of studying the interactions between smaller subsystems, each of which can be studied in isolation.

We shall show that model validation dynamics feature a hierarchy of three time scales. The state and control variables evolve on a 'fast', calendar time-scale. The coefficients of each model evolve on a 'slow', model revision time-scale, where each unit of time corresponds to $1/\epsilon$ units of calendar time. Finally, model switching occurs on a 'really slow', large deviations time-scale, where each unit of model time corresponds to $\exp[S^*/\epsilon]$ units of coefficient time, where S^* is a model specific 'rate function', summarizing how difficult it is to escape from each model's self-confirming equilibrium. This hierarchy of time-scales greatly simplifies the analysis of model validation, as it permits us to focus separately on the problems of control, model revision, and model selection. The novel aspect of our analysis is the ultimate, large deviations time scale. It involves rare but recurrent Markov switching among the finite set of models, each with coefficients fixed at their self-confirming values, and with the underlying data fixed at the mean of a model specific invariant distribution. In other words, we are going to replace the above time-varying Markov transition matrix, $\mathcal{P}_n$, with a *constant*, state-independent, transition matrix, $\bar{\mathcal{P}}$, with elements determined by the large deviations properties of each of the models. A key feature of these large deviation properties is that model switches are approximately exponentially distributed, thus validating the homogeneous Markov chain structure of the approximation. In the spirit of Kandori, Mailath, and Rob (1993), it will turn out that as $\epsilon \to 0$ the stationary distribution across models will collapse onto a single model.

4.3. **Mean ODE Approximation of Model Revisions.** We need to characterize the dynamic interactions among three classes of variables: (1) The state and control variables that appear as regressors within each model, Φ_n, (2) The coefficient estimates, β_n, and (3) The model indicator, s_n. We start in the middle, with the coefficient estimates. Their dynamics can be approximated by averaging over the Φ_n variables for given values of the model coefficients and model indicator. This produces equation 3.6. We want to show that as $\epsilon \to 0$, the random path of each model's coefficient estimates can be approximated by the path of a deterministic ODE when viewed on a sufficiently long time-scale. To define this time-scale, let $\beta_{i,n}$ be the real-time sequence of coefficient estimates of model-i, and let $\beta^*_{i,n}$ be its SCE, given whatever model is generating the data during period-n. Let $\beta^\epsilon_i(t)$ be the piecewise-constant continuous-time interpolation of the model-i's coefficient estimates, $\beta^\epsilon_i(t) = \beta_{i,n} \ \ \forall t \in [\epsilon n, \epsilon(n+1))$, and let $\beta^{*\epsilon}_i(t)$ be the continuous-time interpolation of the sequence of SCEs. Finally, define $\tilde{\beta}_{i,n} = \beta_{i,n} - \beta^*_{i,n}$ and $\tilde{\beta}^\epsilon_i(t) = \beta^\epsilon_i(t) - \beta^{*\epsilon}_i(t)$ as the real- and continuous-time sequences of deviations between model coefficient estimates and their SCE values. The following result describes the sense in which the random paths of $\tilde{\beta}^\epsilon_i(t)$ can be approximated by a deterministically switching ODE,

Proposition 4.1. *Given Assumptions 3.3 and 4.4, as $\epsilon \to 0$, $\tilde{\beta}^\epsilon_i(t)$ converges weakly (in the space, $D([0,\infty))$ of right-continuous functions with left-hand limits endowed with the*

Skorohod topology)) to the deterministic process $\beta_i(t)$, where $\beta_i(t)$ solves the mean ODE,

$$\dot{\beta}_i = h_i(\beta_i(t)) \tag{4.17}$$

Proof. See Appendix A. □

Figure 1 illustrates the nature of this approximation for model-i, for the case of three models, where the set of its self-confirming equilibria are assumed to be: $\beta^*_{i,1} = 1$, $\beta^*_{i,2} = 2$, and $\beta^*_{i,3} = 3$. Over time, model-i's coefficients converge to the self-confirming equilibrium pertaining to whichever model is currently generating the data. The ODEs capture the mean path of the coefficient estimates in response to rare model switches. Later we shall see that on a logarithmic, large deviations time scale we can approximate these paths by their limit points, since the relative amount of time they remain in the neighborhood of each self-confirming equilibrium grows as $\epsilon \to 0$.

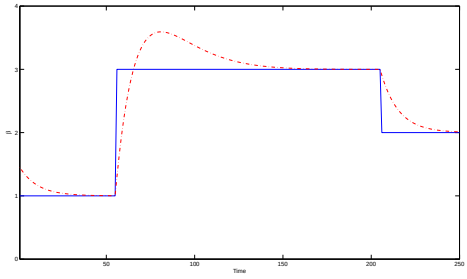

FIGURE 1. Mean ODE Approximation

Proposition 4.1 can be interpreted as a function space analog of the Law of Large Numbers. The scaled process $\beta^\epsilon(t)$ plays the role of "observations", and the mean ODE $\beta(t)$ plays the role of the "expected value" to which $\beta^\epsilon(t)$ converges as the number of observations $[t/\epsilon]$ increases. It implies that on any finite time interval the path of the interpolated process $\beta^\epsilon_i(t)$ closely shadows the solution of the mean ODE with arbitrarily high probability as $\epsilon \to 0$.

4.4. **Diffusion Approximations.** We can also obtain a function space analog of the Central Limit Theorem by studying the fluctuations of $\beta^\epsilon_i(t)$ around the mean ODE $\beta_i(t)$. To do this, define the scaled difference between the interpolated deviation process, $\tilde{\beta}^\epsilon_i(t)$, and the mean ODE

$$U^\epsilon_i(t) = \frac{\tilde{\beta}^\epsilon_i(t) - \beta_i(t)}{\sqrt{\epsilon}}$$

we can state the following result

Proposition 4.2. *Conditional on the event that model i continues to be used, as $\epsilon \to 0$ $U^\epsilon_i(t)$ converges weakly to the solution of the stochastic differential equation*

$$dU(t) = h^i_\beta(\beta^i(t))U(t)dt + \mathcal{R}^{1/2}(\beta^i(t))dW$$

where $h^i_\beta(\cdot)$ is the Jacobian of $h^i(\cdot)$ and $\mathcal{R}(\cdot)$ is the stationary long-run covariance matrix with elements

$$\mathcal{R}_{ij}(\beta) = \sum_{k=-\infty}^{\infty} cov[H_i(\beta, \Phi_k(\beta), x_k(\beta)), H_j(\beta, \Phi_0(\beta), x_0(\beta))]$$

Proof. See Appendix B. □

This result can be used to calibrate the test threshold, $\bar{\theta}_n$. The sequence of test statistics in 4.12 can also be approximated by a diffusion. Under the null, the mean dynamics are simple, $E(\Lambda'\Omega^{-1}\Lambda) = \kappa$, the number of degrees of freedom of the test (i.e., the number of model coefficients). Letting $\tilde{\theta}^\epsilon(t) = \theta^\epsilon(t) - k$, we get the following Ornstein-Uhlenbeck approximation for the path of $1/\sqrt{\epsilon}$ scaled test statistic,

$$d\tilde{\theta} = -\tilde{\theta}dt + \mathcal{R}dW$$

where $\mathcal{R}^2$ is the variance of the centered test statistic, which depends on the *fourth* moments of the data. This implies the test statistic exhibits typical fluctuations of order $\sqrt{\epsilon \cdot \mathcal{C}}$, where $\mathcal{C}$ is given by,

$$\mathcal{C} = \int_0^\infty e^{-s}\mathcal{R}^2 e^{-s} ds = \frac{1}{2}\mathcal{R}^2$$

If we want model rejections to be rare events, the limiting test threshold needs to be comfortably above this, so that isolated shock realizations do not trigger model rejections.

4.5. **Markov Chain Approximation of Model Switching.** Propositions 4.1 and 4.2 describe the *average* behavior of each model's coefficient estimates. Both are conditioned on a *fixed* time horizon. Eventually, however, for any $\epsilon > 0$, the coefficient estimates will wander a significant distance from the SCE (significant, that is, relative to the $\sqrt{\epsilon}$ Central Limit scaling). We have in mind a situation where this potentially triggers a model switch. These switches are *rare*, in the sense that they occur in response to tail events in the model revision process. We must now characterize these tail events. We do this using the tools of large deviations (LD) theory.

The analysis consists of four main steps. First, using results from Dupuis and Kushner (1989) and Cho, Williams, and Sargent (2002), we provide conditions under which each model's sequence of coefficient estimates satisfies a Large Deviations Principle. Second, we use the Contraction Principle to link the LD properties of the coefficient estimates to the LD properties of the LM test statistics. Third, we use the LD properties of the test statistics to construct a homogeneous Markov Chain approximation of the model selection process. Finally, using this approximation, we characterize the limiting model distribution, and identify a 'dominant' model in terms of its LD rate function.

We begin with a definition

Definition 4.3. *Let E be a separable Banach space. Suppose $\mathbf{S}_n, n > 0$ are E-valued random variables. It is said that $\{n^{-1}\mathbf{S}_n\}$ satisfies a Large Deviations Principle if there is a lower semicontinuous rate function $I : E \to [0, \infty]$, with compact level sets $I^{-1}([0, a])$ for all $a > 0$, such that*

$$\liminf_{n\to\infty} n^{-1} \log P(n^{-1}\mathbf{S}_n \in A) \geq -\inf_{x\in A} I(x)$$

for all open subsets $A \subset E$*, and*

$$\limsup_{n\to\infty} n^{-1} \log P(n^{-1}\mathbf{S}_n \in B) \leq -\inf_{x\in B} I(x)$$

for all closed subsets $B \subset E$

In our setting, $\mathbf{S}_n$ will either be defined as a sequence of coefficient estimates, or as a sequence of test statistics, with E then corresponding to the relevant path space. The crucial object here is the rate function, $I(x)$. Definition 4.3 shows precisely the sense in which large deviation events are rare, i.e., their probability declines *exponentially* with n, and the rate function plays the role of a scale factor in this decline. If one process has a uniformly larger rate function than another, the relative frequency of its escapes will vanish. This will be of some importance when characterizing the limiting properties of model validation dynamics.

Large deviations calculations have three components: (1) An H-functional, (2) the Legendre transformation of the H-functional, and (3) an action functional used to determine the large deviations rate function. The H-functional is the log moment generating function of the martingale difference component of the least-squares orthogonality conditions. Existence of the H-functional is the key existence condition of our model. Write the parameter update equations for each model as (since the same condition applies for each $i \in \mathcal{M}$, we omit superscripts for simplicity),

$$\begin{aligned}\beta_n &= \beta_{n-1} + \epsilon h(\beta, s) + \epsilon[H(\beta_{n-1}, s_{n-1}, \Phi_{n-1}, v_n) - h(\beta, s)] \\ &= \beta_{n-1} + \epsilon h(\beta, s) + \epsilon \tilde{H}(\beta_{n-1}, s_{n-1}, \Phi_{n-1}, v_n)\end{aligned}$$

so that $\tilde{H}(\cdot)$ represents the martingale difference component of the update algorithm. We assume $\tilde{H}(\cdot)$ satisfies the following assumption,

Assumption 4.4. *For all* $i \in \mathcal{M}$*, the following limit exists uniformly in* β *and* s *(with probability one),*

$$\lim_{k,n} \frac{1}{k} \log E_n \exp\left[a' \sum_{j=0}^{k-1} \tilde{H}_i(\beta^i, s, \Phi^i_{n+j}, v_{n+1+j})\right]$$

where $\lim_{k,n}$ *means the limit exists as* $k \to \infty$ *and* $n \to \infty$ *in any way at all.*

This limit defines the H-functional, and we denote it as $\mathcal{H} : \mathcal{B} \times \mathcal{M} \times \mathbb{R}^d_{++} \mapsto \mathbb{R}_+$, where d is the dimensionality of the parameter space. Existence of $\mathcal{H}(\beta, s, a)$ imposes restrictions on the tails of the data and the underlying shocks, and must be verified on a case-by-case basis.[15]

The Legendre transform of $\mathcal{H}(\beta, s, a)$ is defined as follows,

$$L(\beta, s, \lambda) = \sup_a [\lambda \cdot a - \mathcal{H}(\beta, s, a)] \tag{4.18}$$

In static, i.i.d., environments this is the end of the story. The probability of witnessing a large deviation of λ from the mean would be of order $\exp[-nL(\lambda)]$. However, in dynamic settings things are more complicated. The relevant sample space is now a function space,

[15]In Cho and Kasa (2011) we provide an example for the case of univariate linear regression models and Gaussian data and shocks.

and large deviations consist of sample *paths*. Calculating the probability of a large deviation involves solving a dynamic optimization problem. The Legendre transformation $L(\beta, s, \lambda)$ now plays the role of a flow cost function, summarizing the instantaneous probabilistic cost of any given path away from the self-confirming equilibrium. For a given boundary, the value function of this control problem captures the probability of escaping from the self-confirming to any given point on the boundary. If only the radius of the boundary is specified, as in our specification testing problem, then one must also minimize over the boundary. The control problem characterizing the large deviation properties of the estimated coefficients can now be written as the minimization of the following action functional:

$$S(\beta_0, s_0) = \inf_{T>0} \inf_{\beta} \int_0^T L(\beta, s, \dot{\beta}) dt \tag{4.19}$$

subject to the boundary conditions $\beta(0) = \beta_0$, $s(0) = s_0$, and $\beta(T) \in \partial G$, where ∂G denotes the escape boundary. Since the action functional is stationary and T is free, the solution is characterized by the following Hamilton-Jacobi-Bellman equation,

$$\inf_{\dot{\beta}} \{L(\beta, s, \dot{\beta}) + \nabla S \cdot \dot{\beta}\} = 0$$

where ∇S denotes the gradient of S with respect to β. This can equivalently be written,

$$\sup_{\dot{\beta}} \{-\nabla S \cdot \dot{\beta} - L(\beta, s, \dot{\beta})\} = 0 \tag{4.20}$$

We now make an important observation. The Legendre transform in (4.18) defines a convex duality relationship between $\mathcal{H}(\beta, s, a)$ and $L(\beta, s, \lambda)$. This means the HJB equation in (4.20) can be written compactly as,

$$\mathcal{H}(\beta, s, -\nabla S) = 0 \tag{4.21}$$

The solution of this problem depends on both the model being estimated and the model being used to generate the data. Denote its solution by S^*. The following proposition links S^* to the large deviation properties of each model's sequence of coefficient estimates

Proposition 4.5. *Fix $s = s_0$, and let $\beta_i^\epsilon(t)$ be the continuous-time interpolation of model-i's estimated coefficient vector. Let S_i^* denote the solution of the control problem in 4.19, and G be a set containing model-i's unique SCE (given $s = s_0$). Then, given Assumptions 3.1-3.4 and Assumption 4.4, model-i's large deviation properties are given by:*

(1) *If the exogenous shocks, $v = (v_1, v_2)$ are i.i.d. and unbounded, and there exist constants $\kappa > 1$ and $Q < \infty$ such that $\forall n$ and $s \geq 0$*

$$P(|H_i(\cdot)| \geq s|\mathcal{F}_n) < Qe^{-s^\kappa} (w.p.1)$$

then, for $\beta^\epsilon(0) \in G$

$$\limsup_{\epsilon \to 0} \epsilon \log P(\beta_i^\epsilon(t) \notin G \quad \text{for some} \quad 0 < t \leq T) \leq -S_i^*$$

(2) *If the exogenous shocks, v, are bounded, and S_i^* is continuous on ∂G, then*

$$\lim_{\epsilon \to 0} \epsilon \log P(\beta_i^\epsilon(t) \notin G \quad \text{for some} \quad 0 < t \leq T) = -S_i^*$$

(3) *Given the assumptions of part 2, and letting* $\tau_i^\epsilon = \inf_{t \leq T}(\beta_i^\epsilon(t) \notin G)$ *then*

$$\lim_{\epsilon \to 0} \epsilon \log E(\tau_i^\epsilon) = S_i^*$$

If the shocks are unbounded then $\lim_{\epsilon \to 0} \epsilon \log E(\tau_i^\epsilon) \geq S_i^*$

Proof. For part (1) see Dupuis and Kushner (1989). For parts (2) and (3) see Kushner and Yin (1997) and Dupuis and Kushner (1987) (Theorem 5). □

The are several noteworthy features of this result. First, note that the escape probabilities and mean escape times are independent of $\beta_i^\epsilon(0) \in G$. This reflects the fact that the mean dynamics are stabilizing for all $\beta^\epsilon(t) \in G$, so it is very likely that $\beta_i^\epsilon(t)$ converges to a small neighborhood of β_i^* before it succeeds in escaping.[16] Second, and closely related, the escape times are approximately exponentially distributed. This is important in delivering a homogeneous Markov Chain approximation to the model switching dynamics. Again, this reflects the fact that points within G are very likely to converge to the SCE before escaping. This makes each escape attempt independent from its predecessors, which eliminates 'duration dependence' and makes waiting times exponential. Third, note that we have said nothing about the evolution of the second moment matrix, R. Remember that it is being updated at the same time (and at the same rate) as $\beta(t)$. However, its evolution is deterministic, and does not introduce additional sources of noise that can drive escape. Consequently, the dynamics of R are tied to those of β (assuming of course the path of R is well behaved, which rules out linear dependencies among each model's regressors). Fourth, since S^* depends on G, the escape boundary, so do the escape probabilities and mean escape times. The 'bigger' G is, the less likely an escape.[17]

The remarkable thing about Propositions 4.1 and 4.5 is that together they characterize the sample paths of a nonlinear stochastic dynamic process in terms of the solutions of two *deterministic* differential equations; one characterizing the mean dynamics and the other characterizing the escape dynamics.

Solution of the large deviations control problem in 4.19 involves a minimization over points on the boundary, ∂G, of the parameter space. Since with overwhelming probability the escape path hits the boundary at a unique point, one could in principle calculate test statistics based directly on fluctuations in the coefficient estimates. However, a better approach is to base inferences on the sequence of estimated scores. Under the null, these behave as innovations, and therefore will more clearly reveal alternatives featuring breaks or other structural changes.[18] Hence, we must now translate the LD results for the coefficients into LD results for the LM test statistics in equation (4.12).

To do this, define the function, $\mathcal{Z}^i : \mathbb{R}^{d_i} \to \mathbb{R}_+$, where d_i is the number of variables in model i, as the score function, $\mathcal{Z}^i(\beta_n^i) = \Lambda_n^{i\prime} \hat{\Omega}_{i,n}^{-1} \Lambda_n^i$. Next, define the continuous-time

[16] A little more formally, given two initial conditions, $(\beta_1(0), \beta_2(0))$, within some ρ_1-neighborhood of β^*, then for any $\rho_2 < \rho_1$, the probability that one of them escapes to ∂G before both get within a ρ_2-neighborhood of β^* goes to zero as $\epsilon \to 0$.

[17] Technically, this is only true of uniform expansions of G, e.g., increasing the radius of a symmetric ball around β^*. Since escapes are very likely to occur in a single particular direction, expanding G in other directions will have no effect on escape probabilities.

[18] Benveniste, Metivier, and Priouret (1990) emphasize this point. See p. 182.

interpolation of the LM test statistic

$$\theta^{\epsilon}(t) = \theta^{\epsilon}(0) + \epsilon \sum_{i=0}^{[t/\epsilon]} [\mathcal{Z}(\beta^{\epsilon}(i)) - \theta^{\epsilon}(i)]$$

and then define $\theta(t) = \lim_{\epsilon \to 0} \theta_{\epsilon}(t)$ as its limit. Observe that any given continuous-time path, $\beta(t)$ for a model's coefficients induces a corresponding path for its test statistic. (Remember that the state variables are being averaged out here). Define this mapping, $F : C[0,T] \to C[0,T]$, as follows

$$\begin{aligned} \theta(t) &= \theta(0) + \int_0^t [\mathcal{Z}(\beta(s)) - \theta(s)] ds \\ &\equiv F(\beta) \end{aligned} \tag{4.22}$$

We now have the following result

Proposition 4.6. *Each model's LM test statistic process, $\theta_i^{\epsilon}(t)$, has a locally stable equilibrium at $\theta_i^* = \mathcal{Z}(\beta_i^*) = 0$, and it satisfies a large deviations principle with rate function given by*

$$V_i(\theta) = \inf_{T>0} \inf_{\{\beta : \theta = F(\beta)\}} \int_0^T L(\beta, s, \dot{\beta}) dt$$

subject to $\theta(t^e) \notin G^{\theta}$ for some $0 < t^e < T$, where ∂G^{θ} defines a rejection threshold.

Proof. The stability of θ^* is clear from inspection of (4.22). The proof of the LDP is based on the following useful result from the theory of large deviations (see Dembo and Zeitouni (1998), p. 126) □

Theorem 4.7. (Contraction Principle) *Let X and and Y be Hausdorff topological spaces and $f : X \to Y$ a continuous function. Consider a (good) rate function $S : X \to [0, \infty]$.*

(a): For each $y \in Y$ define

$$S'(y) = \inf\{S(x) : x \in X, \quad y = f(x)\}$$

Then S' is a (good) rate function on Y, where the infimum over the empty set is taken as ∞

(b): If S controls the LDP associated with a family of probability measures μ_{ε} on X, then S' controls the LDP associated with the family of probability measures $\mu_{\varepsilon} \circ f^{-1}$ on Y.

Loosely speaking, the contraction principle tells us that large deviations principles are preserved by continuous mappings. Of course, depending on the properties of f, the rate function S' might be quite different from the rate function S, so the large deviation *properties* of x and y themselves (e.g., escape times and escape routes) might be quite different. However, the contraction principle provides a means for translating between the two. To apply this theorem, we must establish that $F(\beta)$ is continuous.

Lemma 4.8. *If $\mathcal{Z}(\beta_n)$ is Lipschitz continuous, then $F(\beta(t))$ is continuous.*

Proof. See Appendix C □

Lipschitz continuity of $\mathcal{Z}(\beta_n)$ is a rather weak assumption. It will be satisfied as long as second moments remain bounded and there are no linear dependencies among the regressors.

The analysis so far has exploited a time-scale separation between the data and each model's coefficient estimates. We've studied the evolution of a model's coefficients by averaging out the dynamics of the state variables. Everything has been conditional on s_n, i.e., the current model. The next step in our analysis exploits a different kind of time-scale separation; namely, between the coefficient estimates and the frequency of model switching. After a new model is selected, its coefficients can be well away from their new SCE values. Applying the LM test with a fixed threshold would lead to instantaneous rejection. As noted earlier, we assume the agent in our model is relatively sophisticated, and is aware of feedback. Specifically, he knows that it takes time for a new model to converge to its own SCE. While this convergence is underway, a model's score vector will be nonzero, as it reflects the presence of nonzero mean dynamics. The agent wants to incorporate this drift into the null hypothesis. One way to do this would be to incorporate a 'burn in' period after model switching, during which no testing takes place. Another possibility would be to only estimate models while they are in use. Neither are widely applied in practice (although, of course, practitioners do occasionally add regime-specific dummies). Instead, we assume that drift is incorporated into the null via a declining test threshold, $\bar{\theta}_{n+1} < \bar{\theta}_n$. In other words, the test becomes more stringent the longer a model has been in use. To ensure that model rejections remain large deviation events, even in the presence of recurrent model switching, we make the following assumption,

Assumption 4.9. *There exists a deterministically declining test threshold, $\bar{\theta}_n$, such that Assumption 4.4 and Proposition 4.5 remain valid for all $i \in \mathcal{M}$ and all $s \in \mathcal{M}$.*

Given global asymptotic stability, existence is not really an issue, as long as each model's coefficients converge rapidly enough to their SCE values. (See Appendix D for a sufficient condition). However, actually computing $\bar{\theta}_n$ will be unavoidably model specific.[19]

Since model rejections are rare in the large deviations sense, we can now average out the dynamics in $\beta^i(t)$ and focus on switches *between* models. To do this we define a new logarithmic time-scale, $\tau = \epsilon \log(t)$, where τ can be interpreted as the time-scale over which model switching occurs. In other words, each unit of model switching time, τ, corresponds to $\exp[\epsilon^{-1}]$ units of model revision time. Large deviation events only become 'visible' on this scale. Over this length of time we can average over the excursions that $\beta(t)$ takes away from the SCE, and fix its value at β^* (its long-run average), just as we fixed the values of the state variables at their stationary equilibrium values when studying the dynamics of $\beta(t)$. In fact, to obtain the necessary averaging for *all* models, we must actually employ the time-scale $(\epsilon/\bar{V})\log(t)$, where $\bar{V}$ is the largest LD rate function among all the models.

As when studying the evolution of a model's coefficients, we start by defining a continuous-time interpolation of the discrete distribution over the models, p_n. Over short horizons, the transition probability matrix, $\mathcal{P}_n$, of this Markov Chain is quite complex (see eqs. (4.15)-(4.16)). Our goal is to simplify this matrix by applying singular perturbation methods.

[19] Again, see Cho and Kasa (2011) for an example.

Define the continuous-time interpolation of p_n as usual, i.e., $p^\epsilon(t) = p_n$ for $t \in [\epsilon n, \epsilon(n+1))$. Next, use the change of variables $\tau = (\epsilon/\bar{V})\log(t)$, and consider the rescaled process, $p^\epsilon(\tau)$. This process can be characterized as an m-state homogeneous Markov Chain.

Proposition 4.10. *Assume $\forall i \in \{1, 2, \cdots m\}$ that $\theta^i(t)$ is calibrated to reject during escapes of Model i. Assume $\pi^i(t) \in [\underline{a}, \bar{a}]$ $\forall i, t$, where $\underline{a} > 0$ and $\bar{a} < 1$. Then as $\epsilon \to 0$, $p^\epsilon(\tau)$ converges weakly to a homogenous m-state Markov Chain with generator Q,*

$$q_{ij} = \pi_j^* e^{(\bar{V}-V_i^*)/\epsilon} \qquad q_{ii} = -\left(\sum_{j\neq i}^{m} \pi_j^*\right) e^{(\bar{V}-V_i^*)/\epsilon}$$

which possesses a unique invariant distribution

$$\bar{p}_i^\epsilon = \frac{\pi_i^* e^{V_i^*/\epsilon}}{\sum_{j=1}^m \pi_j^* e^{V_i^*/\epsilon}} \tag{4.23}$$

where π_i^* is model i's selection probability defined at its SCE.

Proof. See Appendix D. □

4.6. **Dominant Models.** Proposition 4.10 asks what happens when $\tau \to \infty$ 'faster' than $\epsilon \to 0$. It's also useful to ask what happens when $\epsilon \to 0$ for a given (large) τ. It's clear from equation (4.23) that this limit is degenerate.

Proposition 4.11. *As $\epsilon \to 0$ the model distribution collapses onto the model with the largest LD rate function.*

This means that in the limit the agent uses one of the models 'almost always'. The dominant model will be the model with the largest LD rate function. This model survives specification testing longer than any other model. Interestingly, the dominant model may not be the best fitting model. Of course, all else equal, poorly fitting models will have smaller rate functions and will not endure specification testing for long. A large residual variance generates a lot of 'noise' around the SCE, and therefore, makes escape easier. However, the *precision* of a model's estimates also matters. Precise estimates are less liable to wander from their SCE values. Hence, overfitted models can escape just as quickly as underfitted models. In fact, recursive testing based on one-step ahead forecast errors embodies a model complexity cost that resolves the bias/variance trade-off that inevitably arises when attempting to discriminate among models (Hansen and Yu (2001)).

The reader may have noticed that we have not paid much attention to the details of randomization. Propositions 4.10 and 4.11 show why. It turns out that our LD approach is robust with respect to the details of experimentation. All that matters is that each model's chances of being selected remain strictly bounded between 0 and 1.

Corollary 4.12. *As long as the experimentation probabilities, π_t^i, remain strictly bounded between 0 and 1, the identity of the dominant SCE is independent of the details of randomization.*

Proof. Follows directly from equation (4.23). □

4.7. **An Information-Theoretic Interpretation.** We have defined a validated self-confirming equilibrium as an outcome generated by a model which survives specification testing longer than any other model in $\mathcal{M}$. We have identified this model as the model with the maximum large deviations rate function, defined at its own self-confirming equilibrium. To readers familiar with information theory and statistics, this may appear to be a puzzling result. From Sanov's Theorem we know rate functions are connected to relative entropies, and then, from either Stein's lemma (classical) or Chernoff bounds (Bayesian), we know that relative entropies are connected to detection error probabilities. In particular, larger relative entropies should make it easier to detect discrepancies between a model and the true DGP. That is, larger relative entropies reduce the probabilities of Type I and Type II errors. Why then are models with large rate functions *more* durable?

This apparent contradiction illustrates a key difference between model validation with exogenous data and model validation with endogenous data. With endogenous data, each model has the capacity to mimic the true DGP. In this case, rejecting a model constitutes a Type I error, and as usual, a larger rate function implies a smaller Type I error probability (or more precisely, it increases the rate at which it converges to zero).

4.8. **Example.** Suppose $\dim(\mathcal{M}) = 3$, i.e., there are 3 possible models. Let V_i^* be the large deviations rate function for model-i, evaluated at its unique stable SCE. The combination of constant gain learning, specification testing, and random model selection induces an approximating 3-state ergodic Markov chain across models. Model switches are triggered by escapes from each model's SCE. As $\epsilon \to 0$, these escape probabilities are of order $e^{-V_i^*/\epsilon}$. Model selection dynamics can therefore be approximated by the 3-state transition matrix, $\bar{\mathcal{P}} = I + Q^\epsilon$, where Q^ϵ is the generator

$$Q^\epsilon = \begin{pmatrix} -(\pi_2^* + \pi_3^*)e^{-V_1^*/\epsilon} & \pi_2^* e^{-V_1^*/\epsilon} & \pi_3^* e^{-V_1^*/\epsilon} \\ \pi_1^* e^{-V_2^*/\epsilon} & -(\pi_1^* + \pi_3^*)e^{-V_2^*/\epsilon} & \pi_3^* e^{-V_2^*/\epsilon} \\ \pi_1^* e^{-V_3^*/\epsilon} & \pi_2^* e^{-V_3^*/\epsilon} & -(\pi_1^* + \pi_2^*)e^{-V_3^*/\epsilon} \end{pmatrix} \tag{4.24}$$

and where $\pi_i^* \in (0,1)$ are parameters determining which model is more likely to be selected following a given model rejection.

The stationary distribution is as follows,

$$\begin{aligned} \bar{p_1} &= \Delta^{-1} a_1 e^{-(V_2^* + V_3^*)/\epsilon} \\ \bar{p_2} &= \Delta^{-1} a_2 e^{-(V_1^* + V_3^*)/\epsilon} \\ \bar{p_3} &= \Delta^{-1} a_3 e^{-(V_1^* + V_2^*)/\epsilon} \end{aligned}$$

where

$$\Delta = a_1 e^{-(V_2^* + V_3^*)/\epsilon} + a_2 e^{-(V_1^* + V_3^*)/\epsilon} + a_3 e^{-(V_1 + V_2^*)/\epsilon}$$

and where a_i are constants that are independent of ϵ. Therefore,

$$\begin{aligned} \frac{\bar{p_2}}{\bar{p_1}} &\propto e^{-(V_1^* - V_2^*)/\epsilon} \\ \frac{\bar{p_3}}{\bar{p_1}} &\propto e^{-(V_1^* - V_3^*)/\epsilon} \end{aligned}$$

Suppose Model 1 is dominant, so that $V_1^* > V_2^*$ and $V_1^* > V_3^*$. Then notice that as $\epsilon \to 0$, Model 1 is used almost always, and this conclusion does not depend on the experimentation probabilities. This independence derives from the fact that once a model starts to be used, its coefficent estimates converge to their self-confirming equilibrium values.

5. An Example

This section illustrates our results with an example based on Sargent's (1999) well known Phillips Curve model. It demonstrates the importance of incorporating specification testing and model validation procedures into the macroeconomic learning literature.[20]

5.1. **Sargent's Conquest Model.** Sargent studied the problem of a Central Bank that wants to minimize a quadratic loss function in inflation and unemployment, but is unsure about the true model. The Bank posits a reduced form regression model of the form,

$$u_n = \beta_0 + \beta_1 \pi_n + \epsilon_n$$

and then tries to learn about this relationship by adaptively updating its coefficients. To account for potential drift in the underlying relationship, the Central Bank discounts past data when updating its estimates. Simulations of this model produce the following time paths of inflation. (Here, and in what follows, we use the same parameter values as in Sargent). The striking feature here is the recurring cycle of gradually rising inflation, and

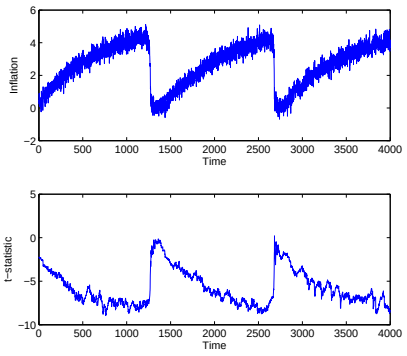

Figure 2. Simulated Inflation Path in Sargent's Conquest Model

then occasional sharp inflation stabilizations. As noted by Cho, Williams, and Sargent (2002) (CWS), this cycle represents the interplay between the model's *mean dynamics* and its so-called *escape dynamics*. The mean dynamics reflect the Central Bank's efforts to eliminate systematic forecast errors. These errors are eliminated once inflation reaches its Self-Confirming Equilibrium value of 5%. The escape dynamics are more exotic. In Sargent's model, the true data-generating process is a natural rate model, containing no exploitable Phillips Curve trade-off. At the SCE, the Central Bank's beliefs are free to wander in any direction, and when sequences of positively correlated inflation and Phillips Curve shocks occur, they cause the Central Bank to revise downward its Phillips Curve

[20]See Cho and Kasa (2011) for another example, based on the cobweb model.

slope estimate, and therefore, its inflation target. Since in truth there is no tradeoff, these inflation reductions produce further downward revisions, and the process feeds on itself until inflation reaches the Ramsey outcome of zero inflation. From this point, with no further changes in the inflation target, the Central Bank begins to rediscover the Phillips Curve, due to the presence of inflation shocks acting within the model's natural rate structure. This produces a gradual pull back up to the Nash inflation outcome.

5.1.1. *Sequential t-testing in Sargent's Model.* A natural question at this point is - To what extent is the Central Bank really learning anything here? True, it's revising estimates of a model in light of new data, but in practice policy makers spend most of their time looking for new and improved models, not refining estimates of a *given* model. In Sargent's analysis, the Central Bank never really evaluates the Phillips Curve as theoretical model of inflation and unemployment; it simply reconsiders the strength of an unquestioned trade-off. What if the Central Bank engages in a traditional process of hypothesis testing and model selection? In particular, suppose the Bank entertains the *possibility* that there is no trade-off, perhaps because someone in the research department read a recent presidential address by Milton Friedman. In response, the Bank decides to sequentially test the hypothesis that $\beta_1 = 0$, and if the hypothesis is not rejected, they switch to a vertical Phillips Curve model and set the inflation target to zero. Looking at the t-statistics reported in the bottom panel of Figure 2 suggests that this might produce a different result. Once inflation is stabilized, the Bank would clearly switch to a vertical Phillips Curve. In fact, the actual outcome is reported in Figure 3.

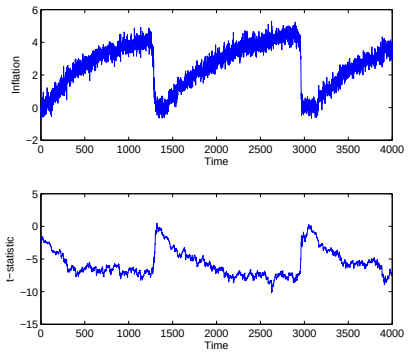

FIGURE 3. Sequential t-tests in Sargent's Conquest Model

Clearly, allowing the Bank to test hypotheses on the slope coefficient makes no difference, aside from a slight delay in the return to the SCE. The fact is, there *is* a correlation between inflation and unemployment, albeit not an exploitable one, and this correlation will lead the Bank to quickly reject the null hypothesis $\beta_1 = 0$, even after it has switched to the vertical Phillips Curve model. The problem, of course, is that the Bank's model is subject to a fundamental misspecification, based on a misinterpretation of the role of the private sector's expectations in the inflation process. To break out of its inflation cycle, the Central Bank must engage in a more sophisticated form of specification testing.

5.1.2. *Expected vs. Unexpected Inflation.* Now suppose a young, newly minted PhD arrives in the research department, and having read recent papers by Lucas, Sargent, and Barro, convinces the Bank to try to distinguish between expected and unexpected inflation when estimating its model. That is, it fits an 'expectations-augmented Phillips Curve'. It turns out, this is quite simple to do in Sargent's *Conquest* model, since the private sector's expectations are based directly on the Bank's own inflation target. Hence, rather than regress unemployment on realized inflation, the Bank just needs to run a multiple regression on its target and the ex post realized inflation shock. Assume the Bank has been conducting business-as-usual until mid-sample, observation 2500, when it suddenly decides to give the new researcher's ideas a try. The outcome is depicted in Figure 4

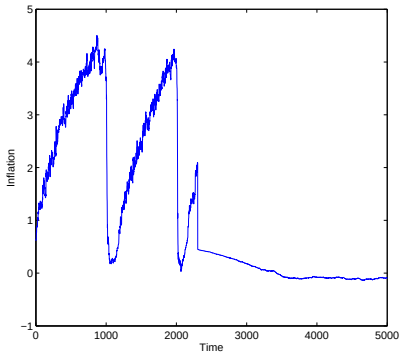

FIGURE 4. Expected vs. Unexpected Inflation in Sargent's Conquest Model

Evidently, the Bank quickly discovers the folly of its ways, and inflation is quickly stabilized, *forever*. In the words of Sargent (1999), this would truly constitute a 'triumph of the natural rate hypothesis'.

5.1.3. *Model Validation and the Phillips Curve.* We've seen that testing hypotheses within the context of a single model would not produce good outcomes. The Bank continues to rediscover a statistical Phillips Curve, and continues to let inflation slip out of control. We've also seen how good theory, in the form of a distinction between expected and unexpected inflation, could produce good outcomes. Suppose, however, the bright young theorist above never gets hired, but instead the Bank follows the kind of model validation procedure we've just outlined. As Sims (1982) argued, could the Bank learn to adopt the right policy even without a priori knowledge of the DGP?

Figure 5 shows what would happen if the Central Bank were to apply a recursive LM testing strategy in the context of Sargent's (1999) *Conquest* model.[21]

It seems clear that during escapes the Central Bank would have cause to doubt the specification of its model. Suppose the Bank begins with a model class comprised of *two* elements: (1) a statistical Phillips curve, as in Sargent (1999), and (2) a vertical Phillips

[21]Some details: (1) Let $x_n = (1, \pi_n)$ be the regression vector, R_n be its second moment matrix, ξ_n be the time-n model residual, and let $\hat{\sigma}_n^2 = \hat{\sigma}_{n-1}^2 + \eta(\xi_n^2 - \hat{\sigma}_{n-1}^2)$, (2) The bottom panel of Figure 5 then reports the recursively estimated statistic, $\theta_n = \theta_{n-1} + \epsilon[(x_n'\xi_n')R_n^{-1}(x_n\xi_n)/\hat{\sigma}_n^2 - \theta_{n-1}]$.

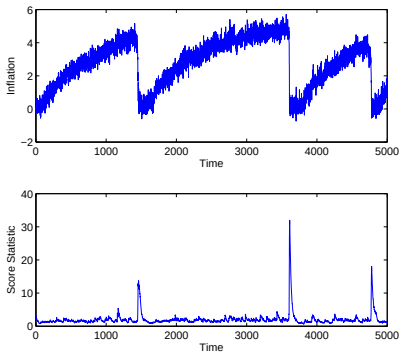

FIGURE 5. Sequential LM tests in Sargent's Conquest Model

curve. Our theory predicts that the dominant model will be the model with biggest large deviations rate function. In this particular case, this is an easy comparison to make. We know from Cho, Williams, and Sargent (2002) that the static Phillips Curve rate function is approximately $\bar{S}^* = .0005$. The mean escape time is approximately $\exp[\bar{S}^*/\epsilon]$ continuous time units, or $\epsilon^{-1}\exp[\bar{S}^*/\epsilon]$ discrete time units.[22] Hence, when $\epsilon = .01$ we should expect to observe escapes every 105 periods.[23] If the LM test is calibrated to reject only during escapes, then the static Phillips Curve would be expected to last about 100 periods. The vertical Phillips Curve case is especially simple. Since the sequence of coefficient estimates becomes Gaussian, the rate function is well known to be $\bar{S}^*(x) = .5(x-\bar{u})^2/(\sigma_1^2+\sigma_2^2)$, where σ_1^2 and σ_2^2 are the variances of the shocks to the Phillips Curve and the inflation target, and $\bar{u}$ is the natural rate of unemployment. Note that in this linear setting, the rate function is symmetric, and escapes are equally likely to occur in either direction. To maintain comparability with the static Phillips Curve we need to calibrate the boundary point, x, so that model rejections occur only during escapes, and with approximately equal statistical evidence. From Figure 5, rejections of the static Phillips Curve occur when the LM test reaches levels of approximately 16. Since in the case of a vertical Phillips Curve, the LM test essentially becomes a recursive F-test, or a squared t-statistic, this suggests a mean escape time of approximately $\epsilon^{-1}\exp[8]$ discrete time units; that is, about once every 300,000 periods! Clearly, the vertical Phillips Curve would dominate, and for all practical purposes the Bank would stick to a low inflation policy forever.

6. Concluding Remarks

This paper has attempted to model macroeconomic policymakers as econometricians. We've done this by combining recent work in both macroeconomics and econometrics. From macroeconomics, we've borrowed from the work of Sargent (1999) and Evans and Honkapohja (2001) on boundedly rational learning dynamics. From econometrics, we've

[22]Warning: For Gaussian shocks this formula provides only an upper bound. It becomes an equality only when shocks are bounded.

[23]Another warning: The distribution of escape times is not symmetric. It is exponential, with a long right tile. Hence, the median escape time will be less than this.

borrowed from recent work on the analysis of misspecified models (White (1994)). As it turns out, this produces a rather difficult marriage.

From a macroeconomic standpoint, it is difficult because we abandon the Rational Expectations Hypothesis, thereby putting ourselves into the 'wilderness of bounded rationality'. We do this not because we like to analyze difficult and ill-posed problems, but simply because of the casual observation that, as econometricians, macroeconomic policymakers do not spend their time refining estimates of a known model, but instead spend most of their time searching for new and better models. Of course, it is not *necessary* to abandon Rational Expectations and traditional Bayesian decision theory when confronting model uncertainty.[24] However, we think there are good reasons to explore alternative approaches.[25]

The marriage between macroeconomics and econometrics is difficult from an econometric standpoint because, presumably, policymakers have some influence over the data-generating processes they are attempting to learn about. The econometric analysis of misspecified models with endogenously generated data is truly uncharted territory.

We make progress on this problem by relating it to a problem that *is* relatively well understood, namely, the dynamics of constant gain recursive learning algorithms. We describe the sense in which the dynamics generated by a process of testing and model revision can be approximated by the dynamics generated by recursive learning models. This is a useful connection to make, because it enables us to apply the tools of large deviations theory to help us understand a potentially wide range of Markov-switching macroeconomic dynamics. Looking at it from the other side, a second payoff from making this connection is that it provides a more secure behavioral foundation for recursive learning models.

Although we feel this paper takes a significant step forward in understanding the interplay between macroeconomics and econometrics, there are certainly many loose ends and unexplored avenues remaining. One possibility is to consider alternative specification tests. Here we focused on LM tests. However, there are many possibilities, depending on what sort of potential misspecification is of most concern. Another possible extension would be to consider alternative model classes. The analysis here was confined to simple atheoretical linear regression models. However, it is becoming increasingly common to base policy discussions on more sophisticated structural models. The second step of our validation process, which involves comparing alternative models in the event the current reference model is rejected, might benefit from recent advances in Bayesian model comparison (Schorfheide (2000)). Finally, perhaps the most interesting and important extension would be to allow the agent to entertain doubts about the entire model class itself. The work of Hansen and Sargent (2007a) on robust filtering of discrete hidden states offers one route toward such an extension.

[24]See, e.g., Brock, Durlauf, and West (2007) for an application of Bayesian model averaging to macroeconomic policy.

[25]See Sargent (1993), Hansen and Sargent (2007b), Kreps (1998), Bray and Kreps (1987), and Gilboa, Postlewaite, and Schmeidler (2008).

Appendix A. Proof of Proposition 4.1

There are two key steps to any weak convergence argument: (1) Establish tightness, and (2) Identify the limit. Tightness delivers compactness (in the space, $D([0,\infty)$, of right-continuous functions with left-hand limits, endowed with the Skorohod topology) via Prohorov's Theorem, which then guarantees existence of a weakly convergent subsequence. Proving tightness can be challenging. However, given our level of generality, we simply assume it, since the details of any proof are unavoidably case specific. One can always guarantee it by resort to a projection facility.

To identify the limit, we employ the martingale method (Kushner and Yin (1997)). The martingale method is based on the following definition:

Definition A.1: *Let $\mathcal{S}$ be a metric space, and $\mathcal{A}$ be a linear operator on $B(\mathcal{S})$ (the set of Borel measurable functions on $\mathcal{S}$). Let $x(\cdot) = \{x(t) : t \geq 0\}$ be a right-continuous process with values in $\mathcal{S}$ such that for each $f(\cdot)$ in the domain of $\mathcal{A}$,*

$$f(x(t)) - \int_0^t \mathcal{A}f(x(s))ds$$

is a martingale with respect to the filtration $\mathcal{F}_t = \sigma\{x(s) : s \leq t\}$. Then $x(\cdot)$ is said to be a solution of the martingale problem with operator $\mathcal{A}$.

The definition of the operator $\mathcal{A}$ will depend on the nature of the underlying process, which in turn depends on the relevant time-scale. In this section, when considering weak convergence to an ODE, $\mathcal{A}$ will be the simple differential generator, $\mathcal{A}f(x) = \dot{x}'\nabla \cdot f(x)$. However, when considering the rate of convergence to this ODE, $\mathcal{A}$ will be the generator of a diffusion process. Later, when considering convergence to a Markov chain on a logarithmic time scale, $\mathcal{A}$ will be the infinitesimal generator of a jump process.

From the above definition, it is clear that application of the martingale method requires a way of verifying that a process is a martingale. The following is a key result in the theory of Markov processes,

Theorem A.2: (Ethier and Kurtz (1986)) *A right-continuous process $x(t)$ is a solution of the martingale problem for operator $\mathcal{A}$ if and only if*

$$E\left(\prod_{j=1}^{i} g_j(x(t_j))\left(f(x(t_{i+1})) - f(x(t_i)) - \int_{t_i}^{t_{i+1}} \mathcal{A}f(x(s))ds\right)\right) = 0 \tag{A.25}$$

for each $0 \leq t_1 < t_2 < \cdots < t_{i+1}$, $f(\cdot)$ in the domain of $\mathcal{A}$, and $g_1, \cdots g_i \in \mathcal{C}_b$, the space of continuous, bounded functions.

Hence, saying that a process solves a martingale problem is a statement about its finite-dimensional distributions. The logic of the martingale method can now be described as follows. We have a family of stochastic processes, $\{x^\epsilon(t)\}$, characterized by a parameter, ϵ. For us, ϵ is the update gain, and $\{x^\epsilon(t)\}$ are continuous-time interpolations of the paths of the coefficient estimates, test statistics, and model indicators. Given tightness, we know that as $\epsilon \to 0$ there is subsequence of $\{x^\epsilon(t)\}$ that converges weakly to a limit. Call it $\tilde{x}(t)$. Depending on the case in hand, our claim will be that $\tilde{x}(t)$ is given by a particular kind of stochastic process, e.g., an ODE (ie, a degenerate process), a diffusion, or a Markov chain. To establish this claim we show that $\tilde{x}(t)$ solves the martingale problem for the generator, $\mathcal{A}$, associated with this process. This involves substituting the ϵ-indexed process into (A.25) and verifying that as $\epsilon \to 0$ the expectation converges to zero. For this logic to work there must be a sense in which the solution of the martingale problem is unique. Fortunately, this is the case:

Theorem A.3: (Ethier and Kurtz (1986)) *Let $x(\cdot)$ and $y(\cdot)$ be two stochastic processes in $D([0,\infty))$, and let $\mathcal{A}$ be an infinitesimal generator. If for any f in the the domain of $\mathcal{A}$*

$$f(x(t)) - f(x(0)) - \int_0^t \mathcal{A}f(x(s))ds \qquad \text{and} \qquad f(y(t)) - f(y(0)) - \int_0^t \mathcal{A}f(y(s))ds$$

are $\mathcal{F}_t$-martingales, and $x(t)$ and $y(t)$ have the same distribution for each $t \geq 0$, then $x(\cdot)$ and $y(\cdot)$ have the same distribution on the path space $D([0,\infty))$.

We can now prove the proposition. Let $\beta^*_{i,n}$ be the SCE of model-i. From Assumption 3.4, specification testing and model switching cause $\beta^*_{i,n}$ to exhibit jumps, since the SCE depends on the model used to generate the data. Let $\beta_{i,n}$ be the real-time sequence of coefficient estimates, and $\beta^\epsilon_i(t)$ be its piecewise-constant continuous time interpolation. Similarly, let $\beta^{*\epsilon}_i(t)$ be the continuous time interpolation of $\beta^*_{i,n}$. If we then define $\tilde\beta_{i,n} = \beta_{i,n} - \beta^*_{i,n}$ and $\tilde\beta^\epsilon_i(t) = \beta^\epsilon_i(t) - \beta^{*\epsilon}_i(t)$ as the deviations of the estimates from the current SCE, we want to show

$$\lim_{\epsilon\to 0}\tilde\beta^\epsilon_i(t) \Rightarrow \beta_i(t)$$

where $\beta_i(t)$ solves to the mean ODE, $\dot\beta_i = h_i(\beta_i(t))$. Let $\beta^o_i(t)$ be the weak sense limit of $\beta^\epsilon_i(t)$. Based on the above results, we must therefore show that

$$f(\tilde\beta^o_i(t+s)) - f(\tilde\beta^o_i(t)) - \int_t^{t+s} h_i(\tilde\beta^o_i(u))\cdot\nabla f(\tilde\beta^o_i(u))du$$

is a martingale for $f \in \mathcal{C}^2_b$, the space of bounded, twice continuously differentiable functions. From Theorem A.2, this requires showing (omitting i-subscripts for simplicity)

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^i g_j(\tilde\beta^\epsilon(t_j))\left(f(\tilde\beta^\epsilon(t+s)) - f(\tilde\beta^\epsilon(t)) - \int_t^{t+s} h(\tilde\beta^\epsilon(u))\cdot\nabla f(\tilde\beta^\epsilon(u))du\right)\right) = 0 \tag{A.26}$$

where $f \in \mathcal{C}^2_b$ and $g_j \in \mathcal{C}_b$, and $0 < t_j \le t$. First, by virtue of the properties of g_j and f, the definition of weak convergence, and the Skorohod representation (which allows us to assume w.p.1 convergence on finite time intervals), we have

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^i g_j(\tilde\beta^\epsilon(t_j))\left(f(\tilde\beta^\epsilon(t+s)) - f(\tilde\beta^\epsilon(t))\right)\right) = E\left(\prod_{j=1}^i g_j(\tilde\beta^o(t_j))\left(f(\tilde\beta^o(t+s)) - f(\tilde\beta^o(t))\right)\right)$$

Choose a sequence n_ϵ such that $n_\epsilon \to \infty$ as $\epsilon \to 0$, and at the same time $\epsilon\cdot n_\epsilon \to 0$. We shall use n_ϵ to perform the requisite averaging. Next, divide the interval $[t, t+s]$ into subintervals of length $\delta_\epsilon \equiv \epsilon\cdot n_\epsilon$, and the discrete-time interval, $[t/\epsilon, (t+s)/\epsilon]$, into steps of size n_ϵ. By definition we have,

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^i g_j(\tilde\beta^\epsilon(t_j))\left(f(\tilde\beta^\epsilon(t+s)) - f(\tilde\beta^\epsilon(t)\right)\right) = \lim_{\epsilon\to 0} E\left(\prod_{j=1}^i g_j(\tilde\beta^\epsilon(t_j))\left(\sum_{k=t/\epsilon}^{(t+s)/\epsilon-1}[f(\tilde\beta_{k+n_\epsilon}) - f(\tilde\beta_k)]\right)\right)$$

Using the law of iterated expectations and the fact that $f \in \mathcal{C}^2_b$, a Taylor series approximation yields,

$$E\left(\prod_{j=1}^i g_j(\tilde\beta^\epsilon(t_j))\left(\sum_{k=t/\epsilon}^{(t+s)/\epsilon-1}[f(\tilde\beta_{k+n_\epsilon}) - f(\tilde\beta_k)]\right)\right) = E\left(\prod_{j=1}^i g_j(\tilde\beta^\epsilon(t_j))\left(\sum_{k=t/\epsilon}^{(t+s)/\epsilon-1}[\nabla f(\tilde\beta_k)\epsilon\sum_{r=k}^{k+n_\epsilon-1} E_k(H(\tilde\beta_r,\Phi_r,\bar s)]\right)\right) + O(\epsilon)$$

where we've used the large deviations result from section 4.5 that $\beta^*_{t+s} - \beta^*_t$ is $o(\epsilon)$ for $s \sim O(\epsilon^{-1})$. Now the key step is to average over the last term of the previous equation. Write this term as,

$$\epsilon\sum_{k=t/\epsilon}^{(t+s)/\epsilon-1}[\nabla f(\tilde\beta_k)\sum_{r=k}^{k+n_\epsilon-1} E_k(H(\tilde\beta_r,\Phi_r,s_k))] = \delta_\epsilon\sum_{k=t/\epsilon}^{(t+s)/\epsilon-1}\nabla f(\tilde\beta_k)\left[\frac{1}{n_\epsilon}\sum_{r=k}^{k+n_\epsilon-1} E_k(H(\tilde\beta_r,\Phi_r,\bar s))\right]$$

Using Assumption 3.3, the properties of δ_ϵ and n_ϵ, and the continuity of ∇f then implies

$$\delta_\epsilon\sum_{k=t/\epsilon}^{(t+s)/\epsilon-1}\nabla f(\tilde\beta_k)\left[\frac{1}{n_\epsilon}\sum_{r=k}^{k+n_\epsilon-1} E_k(H(\tilde\beta_r,\Phi_r,\bar s)\right] \to \int_t^{t+s}\nabla f(\tilde\beta^o(u))h(\tilde\beta^o(u))du \tag{A.27}$$

The final step is to again use the continuity and boundedness of h and ∇f, along with the definition of weak convergence, to show that

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^i g_j(\tilde\beta^\epsilon(t_j))\left(\int_t^{t+s} h(\tilde\beta^\epsilon(u))\nabla f(\tilde\beta^\epsilon(u))du\right)\right) = E\left(\prod_{j=1}^i g_j(\tilde\beta^o(t_j))\left(\int_t^{t+s} h(\tilde\beta^o(u))\nabla f(\tilde\beta^o(u))du\right)\right) \tag{A.28}$$

Combining (A.27) with (A.28) establishes the equality in (A.26) and the proposition is proved. □

Appendix B. Proof of Proposition 4.2

Again we can apply the martingale method. The logic and the steps are the same as in Appendix A. There are only two noteworthy differences. First, due to the $\sqrt{\epsilon}$ Central Limit scaling here, the operator $\mathcal{A}$ becomes the generator of a diffusion,

$$\mathcal{A}f(x) = \nabla f(x) \cdot \nabla h(x)x + \frac{1}{2}\epsilon \cdot \text{tr}[\nabla^2 f(x) \cdot \mathcal{R}(x)]$$

where derivatives are evaluated along the path of the mean ODE. Second, when performing the averaging as in (A.27) one must invoke the invariance principle

$$\lim_{\epsilon \to 0} \sqrt{\epsilon} \sum_{j=0}^{t/\epsilon - 1} \varepsilon_j \Rightarrow \int_0^t \mathcal{R}^{1/2} dW$$

where ε is the martingale difference component of the least squares orthogonality conditions, and $\mathcal{R}$ is its variance-covariance matrix. The details are left to the interested reader. (Proof of a similar result is contained in Yin and Krishnamurthy (2005). The only difference is that in their problem the slow Markov switching process is exogenous).

Appendix C. Proof of Lemma 4.8

We must show that $\forall \epsilon > 0$ there exists a $\delta > 0$ such that $||\beta_1 - \beta_2|| < \delta$ implies $||\theta_1 - \theta_2|| < \epsilon$, where $||\cdot||$ denotes the sup norm on $C[0,T]$. Suppose $|\mathcal{Z}(\beta_{1,n}) - \mathcal{Z}(\beta_{2,n})| < M|\beta_{1,n} - \beta_{2,n}|$. By definition,

$$\begin{aligned} |\theta_1(t) - \theta_2(t)| &= \left| \int_0^t [\mathcal{Z}(\beta_1(s)) - \theta_1(s)]ds - \int_0^t [\mathcal{Z}(\beta_2(s) - \theta_2(s)]ds \right| \\ &\leq \int_0^t |\mathcal{Z}(\beta_1(s)) - \mathcal{Z}(\beta_2(s))|ds + \int_0^t |\theta_2(s) - \theta_1(s)|ds \\ &\leq M\delta t + \int_0^t |\theta_2(s) - \theta_1(s)|ds \end{aligned}$$

Applying Gronwall's inequality to the last inequality gives

$$|\theta_1(t) - \theta_2(t)| \leq M\delta t + M\delta e^t \int_0^t s e^{-s} ds$$

and the result follows. □

Appendix D. Proof of Proposition 4.10

Since experimentation probabilities are bounded away from 0 and 1, the chain is recurrent and ergodic. Tightness is therefore not an issue. What needs to be done is to identify the limit. Again we can apply the martingale method. The steps are the same as in Appendix A, except now, on an exponentially long time-scale, the averaging and Taylor series approximations work somewhat differently. On this time-scale, the dynamics of $\beta^\epsilon(t)$ average out completely, and $\beta^\epsilon(t)$ becomes pinned to its (model-dependent) SCE value. In contrast, the Markov switching dynamics of s_n now become visible.

Since $\beta(t)$ and $\theta(t)$ live on the same time-scale, it is notationally convenient to define the vector, $\varphi^\epsilon(t) = (\beta^\epsilon(t), \theta^\epsilon(t))$. It is also convenient to introduce the change of variable, $\tau = \epsilon \log(t)$. Given tightness, let $\varphi^o(\tau)$ and $s^o(\tau)$ be the weak sense limits of $\varphi^\epsilon(\tau)$ and $s^\epsilon(\tau)$. The proposition asserts the following process is a martingale,

$$f(\varphi^o(\tau + s), s^o(\tau + s)) - f(\varphi^o(\tau), s^o(\tau)) - \int_\tau^{\tau+s} \mathcal{A}f(\varphi^o(u), s^o(u))du$$

where the operator, $\mathcal{A}$, is now given by

$$\mathcal{A}f(\varphi(u), s(u)) = \sum_{j=1}^m q_{s(u)j} f(\varphi^*_{s(u)}, j)$$

where

$$q_{s(u)j} = \pi_j \cdot e^{-V^*(s(u))/\epsilon} \qquad \text{and} \qquad q_{s(u)s(u)} = -\left(\sum_{j \neq s(u)} \pi_j\right) \cdot e^{-V^*(s(u))/\epsilon} \tag{D.29}$$

From Theorem A.2, we must show

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau+s)) - f(\varphi^\epsilon(\tau), s^\epsilon(\tau)) - \int_\tau^{\tau+s} \sum_{j=1}^{m} q_{s(u)j} f(\varphi_s^*, j) du\right)\right) = 0 \tag{D.30}$$

where $f \in \mathcal{C}_b^2$, $g_j \in \mathcal{C}_b$, and $0 < t_j \leq \tau$. Again by virture of the properties of g_j and f, the definition of weak convergence, and the Skorohod representation, we have

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau+s)) - f(\varphi^\epsilon(\tau), s^\epsilon(\tau))\right)\right) = E\left(\prod_{j=1}^{i} g_j(\varphi^o(t_j), s^o(t_j))\left(f(\varphi^o(\tau+s), s^o(\tau+s)) - f(\varphi^o(\tau), s^o(\tau))\right)\right)$$

Hence, we must show the left-hand side converges to the stated operator. To begin, decompose the left-hand side as

$$\begin{aligned} & \lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau+s)) - f(\varphi^\epsilon(\tau), s^\epsilon(\tau))\right)\right) \\ = & \lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left\{[f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau+s)) - f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau))] + [f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau)) - f(\varphi^\epsilon(\tau), s^\epsilon(\tau))]\right\}\right) \end{aligned}$$

Consider the second term,

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau)) - f(\varphi^\epsilon(\tau), s^\epsilon(\tau))\right)\right)$$

Our first goal is to show that this is zero. Using the law of iterated expectations and a second-order Taylor series approximation we can write this as (higher order terms are neglible),

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j)) E_\tau\left\{\nabla f(\varphi^\epsilon(\tau), s^\epsilon(\tau)) \cdot (\varphi^\epsilon(\tau+s) - \varphi^\epsilon(\tau)) + \frac{1}{2}\nabla^2 f(\varphi^\epsilon(\tau), s^\epsilon(\tau))(\varphi^\epsilon(\tau+s) - \varphi^\epsilon(\tau))^2\right\}\right)$$

On exponentially long time-scales, paths of $\varphi(\tau)$ can be decomposed into two 'regimes'. One regime consists of fluctuations around the neighborhood of a SCE while a given model is in use. The second regime consists of transits *between* SCE following model rejections. From Proposition 4.5, the expected duration of the first regime is of order $\exp(V^*/\epsilon)$ in units of t, or just V^* in units of τ. From Proposition 4.2, $E_\tau(\varphi^\epsilon(\tau+s) - \varphi^\epsilon(\tau)) = 0$ and $E_\tau(\varphi^\epsilon(\tau+s) - \varphi^\epsilon(\tau))^2 \sim O(\epsilon)$ during this regime. In general, it is difficult to say anything precise about mean transit times between SCE. Clearly, they will depend on the gradient of the mean dynamics. Fortunately, all we really need is the following assumption,

Assumption D.1. *Mean transit times between neighborhoods of SCE are bounded, and independent of ϵ (in units of t).*

Note that this implies mean transit times are $O(\epsilon^{-1})$ in calendar time, n. The following restriction on the mean dynamics provides a simple sufficient condition that validates assumption D.1,

Contractivity Condition: *If $\mathcal{D}$ is the domain of φ, then the mean dynamics, $h(\cdot)$, satisfy the contractivity condition if $\forall \varphi_1, \varphi_2 \in \mathcal{D}$*

$$\langle h(\varphi_1) - h(\varphi_2), \varphi_1 - \varphi_2 \rangle \leq -\alpha \cdot |\varphi_1 - \varphi_2|^2$$

To see how this delivers assumption D.1, let $|\mathcal{B}|$ be the size of $\mathcal{D}$ (in the Euclidean metric), and let ρ be the size of the neighborhood around each SCE. We then have

Lemma D.2. *Let $\bar{t}$ denote the mean transit time between all SCE, and let $\bar{t}_m$ denote its maximum. Then given assumption D.1 we have*

$$\bar{t} \leq \bar{t}_m \leq \frac{1}{\alpha} \ln\left(\frac{|\mathcal{B}|}{\rho}\right)$$

Proof. Let φ^* denote the new SCE, and $\varphi(0)$ denote the initial value of φ. By direct calculation,

$$\begin{aligned} \frac{d}{dt}\frac{1}{2}|\varphi-\varphi^*|^2 &= |\varphi-\varphi^*|\cdot\dot{\varphi} \\ &= \langle \varphi-\varphi^*, h(\varphi)-h(\varphi^*)\rangle \\ &\leq -\alpha|\varphi-\varphi^*|^2 \end{aligned}$$

where the second line uses the fact that, by definition, $h(\varphi^*)=0$. Then, by Gronwall's inequality, $|\varphi-\varphi^*| \leq e^{-\alpha t}|\varphi(0)-\varphi^*|$, and the result follows. □

Now let $\mathcal{I}_\rho$ be the indicator function for the event $|\varphi-\varphi^*| \leq \rho$. We can use $\mathcal{I}_\rho$ to decompose $E_\tau(\varphi^\epsilon(\tau+s)-\varphi^\epsilon(\tau))$ as follows,

$$\begin{aligned} E_\tau(\varphi^\epsilon(\tau+s)-\varphi^\epsilon(\tau)) &= E_\tau[(\varphi^\epsilon(\tau+s)-\varphi^\epsilon(\tau))\cdot\mathcal{I}_\rho] + E_\tau[(\varphi^\epsilon(\tau+s)-\varphi^\epsilon(\tau))\cdot(1-\mathcal{I}_\rho)] \\ &= \left(\frac{e^{V^*/\epsilon}}{\bar{t}+e^{V^*/\epsilon}}\right)\cdot 0 + \left(\frac{\bar{t}}{\bar{t}+e^{V^*/\epsilon}}\right)\cdot\mu_{\text{transit}} \end{aligned}$$

where μ_{transit} denotes the mean distance between SCE, and where the second line uses Proposition 4.5. The point to notice here is that the second term in the second line becomes neglible as $\epsilon \to 0$. It is clear that the same logic applies to the second-order term in the Taylor series. The only difference is that the dominating term is $O(\epsilon)$ rather than zero. Hence, using the fact that ∇f and $\nabla^2 f$ are bounded, we've established $\lim_{\epsilon\to 0} E_\tau(\varphi^\epsilon(\tau+s)-\varphi^\epsilon(\tau))=0$.

We must now go back to consider the first term in the decomposition,

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau+s)) - f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau))\right)\right) \tag{D.31}$$

Our goal is to show that this converges to

$$E\left(\prod_{j=1}^{i} g_j(\varphi^o(t_j), s^o(t_j))\left(\int_\tau^{\tau+s}\sum_{j=1}^{m} q_{s(u)j} f(\varphi^*_{s(u)}, j)du\right)\right)$$

with q given by D.29. As in Appendix A, choose a sequence n_ϵ such that $n_\epsilon \to \infty$ as $\epsilon \to 0$, and at the same time $\epsilon \cdot n_\epsilon \to 0$. We use n_ϵ to perform the requisite averaging. Next, divide the interval $[t, t+s]$ into subintervals of length $\delta_\epsilon \equiv \epsilon \cdot n_\epsilon$, and the discrete-time interval, $[t/\epsilon, (t+s)/\epsilon]$, into steps of size n_ϵ. To facilitate notation, define the function $t(\tau) = \exp(\tau/\epsilon)$. Then, by definition, we can then write D.31 as

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon}[f(\varphi^\epsilon_{k+n_\epsilon}, s^\epsilon_{k+n_\epsilon}) - f(\varphi^\epsilon_{k+n_\epsilon}, s^\epsilon_k)]\right)\right) \tag{D.32}$$

By Proposition 4.2, and the continuity and boundedness of $f(\cdot)$ and $\varphi(\cdot)$, we have

$$\lim_{\epsilon\to 0} E_k\{[f(\varphi^\epsilon(k+n_\epsilon), s^\epsilon(k+n_\epsilon)) - f(\varphi^\epsilon(k+n_\epsilon), s^\epsilon(k))]\} = E_k\{[f(\varphi^\epsilon(k), s^\epsilon(k+n_\epsilon)) - f(\varphi^\epsilon(k), s^\epsilon(k))]\}$$

Hence, by the law of iterated expectations, we can replace D.32 with

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon}[f(\varphi^\epsilon_k, s^\epsilon_{k+n_\epsilon}) - f(\varphi^\epsilon_k, s^\epsilon_k)]\right)\right) \tag{D.33}$$

Now, as before, we divide the above sum into segments of length n_ϵ over which we average, and again exploit the law of iterated expectations, to replace D.33 with

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon}\sum_{n=k}^{k+n_\epsilon-1} E_n[f(\varphi^\epsilon_k, s^\epsilon_{n+1}) - f(\varphi^\epsilon_k, s^\epsilon_n)]\right)\right) \tag{D.34}$$

Next, we can use the Markov transition probabilities to replace the inner expectation in D.34 to get

$$\lim_{\epsilon\to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon}\sum_{j_0=1}^{m}\sum_{j=1}^{m}\sum_{n=k}^{k+n_\epsilon-1}[f(\varphi_k^\epsilon, j)P_n(s_{n+1}^\epsilon = j|s_n^\epsilon = j_0) - f(\varphi_k^\epsilon, j_0)]\mathcal{I}_{\{s_n=j_0\}}\right)\right) \tag{D.35}$$

From Section 4, the transition probability matrix can be written,

$$P_n = I + Q_n^\epsilon$$

where Q_n^ϵ is the generator of a continuous-time Markov chain,

$$Q_n^\epsilon = \text{diag}[c_{1,n}, c_{2,n}, \cdots, c_{m,n}]\cdot\Pi \equiv C_n\cdot\Pi$$

where $c_{i,n} = \text{Prob}[\theta_{i,n}^\epsilon > \bar{\theta}_n^\epsilon]$, and where the $m\times m$ selection matrix, Π has the form

$$\pi_{ij} = \begin{cases} \pi_j & i\neq j \\ -\sum_{j\neq i}\pi_{ij} & i = j \end{cases}$$

.

From Section 4 we know $\lim_{\epsilon\to 0} c_{i,n} \sim \epsilon\cdot\exp[-V_i^*/\epsilon]$, where V_i^* is the rate function for model-i. To average over such rare events, we scale each $c_{i,n}$ by $\epsilon^{-1}\exp[\bar{V}^*/\epsilon]$, where $\bar{V}^*$ is the maximum rate function, and then write[26]

$$P_n = I + \epsilon e^{-\bar{V}^*/\epsilon}\cdot\tilde{Q}_n^\epsilon$$

where the rescaled generator, $\tilde{Q}_n^\epsilon$, is given by

$$\tilde{Q}_n^\epsilon \equiv \frac{1}{\epsilon}e^{\bar{V}^*/\epsilon}\cdot C_n\cdot\Pi$$

Now choose n_ϵ so that $n_\epsilon\to\infty$ as $\epsilon\to 0$, and at the same time $\delta_\epsilon \equiv \epsilon e^{-\bar{V}^*/\epsilon}\cdot n_\epsilon \to 0$. That is, n_ϵ cannot increase faster than the maximum (discrete) large deviations time-scale. This allows us to write D.35 as

$$\begin{aligned}
&\lim_{\epsilon\to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon}\sum_{j_0=1}^{m}\sum_{j=1}^{m}\sum_{n=k}^{k+n_\epsilon-1}[f(\varphi_k^\epsilon, j)P_n(s_{n+1}^\epsilon = j|s_n^\epsilon = j_0) - f(\varphi_k^\epsilon, j_0)]\mathcal{I}_{\{s_n=j_0\}}\right)\right)\\
=\ &\lim_{\epsilon\to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(\delta_\epsilon\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon}\sum_{j_0=1}^{m}\sum_{j=1}^{m}\frac{1}{n_\epsilon}\sum_{n=k}^{k+n_\epsilon-1}[f(\varphi_k^\epsilon, j)P_n(s_{n+1}^\epsilon = j|s_n^\epsilon = j_0) - f(\varphi_k^\epsilon, j_0)]\mathcal{I}_{\{s_n=j_0\}}\right)\right)\\
=\ &\lim_{\epsilon\to 0} E\left(\prod_{j=1}^{i} g_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(\delta_\epsilon\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon}\sum_{j_0=1}^{m}\sum_{j=1}^{m}\frac{1}{n_\epsilon}\sum_{n=k}^{k+n_\epsilon-1}[\tilde{q}_{n,j_0,j}^\epsilon f(\varphi_k^\epsilon, j)]\mathcal{I}_{\{s_n=j_0\}}\right)\right)\\
=\ &E\left(\prod_{j=1}^{i} g_j(\varphi^o(t_j), s^o(t_j))\left(\int_\tau^{\tau+s}\sum_{j=1}^{m} q_{s(u)j}f(\varphi_{s(u)}^*, j)du\right)\right)
\end{aligned}$$

with $q_{s(u)j}$ given by D.29, and where the bottom line follows from ergodicity and Proposition 4.5. The bottom line establishes the equality in D.30 and the proof is complete. □

[26]Note, this rescaling bears some resemblance to the strategy of 'Importance Sampling' in the simulation of rare event probabilities. See, e.g., Bucklew (2004).

References

An, S., and F. Schorfheide (2007): "Bayesian Analysis of DSGE Models," *Econometric Reviews*, 26, 113–72.

Benveniste, A., M. Metivier, and P. Priouret (1990): *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, Berlin.

Bray, M., and N. Savin (1986): "Rational Expectatons Equilibria, Learning, and Model Specification," *Econometrica*, 54, 1129–60.

Bray, M. M., and D. M. Kreps (1987): "Rational Learning and Rational Expectations," in *Arrow and the Ascent of Modern Economic Theory*, ed. by G. R. Feiwel, pp. 597–625. New York University Press.

Brock, W. A., S. N. Durlauf, and K. D. West (2007): "Model Uncertainty and Policy Evaluation: Some Theory and Empirics," *Journal of Econometrics*, 136, 629–64.

Brown, R., J. Durbin, and J. Evans (1975): "Techniques for Testing the Constancy of Regression Relationships over Time," *Journal of the Royal Statistical Society, Series B*, 37, 149–72.

Bucklew, J. A. (2004): *Introduction to Rare Event Simulation*. Springer.

Bullard, J. (1992): "Time-Varying Parameters and Nonconvergence to Rational Expectations under Least-Squares Learning," *Economics Letters*, 40, 159–66.

Burnham, K., and D. Anderson (2002): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2nd edn.

Cho, I.-K., and K. Kasa (2009): "Recursive Model Selection with Endogenous Data," University of Illinois.

——— (2011): "Learning and Model Validation: An Example," forthcoming.

Cho, I.-K., N. Williams, and T. J. Sargent (2002): "Escaping Nash Inflation," *Review of Economic Studies*, 69, 1–40.

Chu, J., M. Stinchcombe, and H. White (1996): "Monitoring Structural Change," *Econometrica*, 64, 1045–1065.

Cogley, T., R. Colacito, L. P. Hansen, and T. J. Sargent (2008): "Robustness and U.S. Monetary Policy Experimentation," *Journal of Money, Credit, and Banking*, 40, 1599–1623.

Dembo, A., and O. Zeitouni (1998): *Large Deviations Techniques and Applications*. Springer-Verlag, New York, 2nd edn.

Dupuis, P., and H. J. Kushner (1987): "Asymptotic Behavior of Constrained Stochastic Approximations vis the Theory of Large Deviations," *Probability Theory and Related Fields*, 75, 223–44.

——— (1989): "Stochastic Approximation and Large Deviations: Upper Bounds and w.p.1 Convergence," *SIAM Journal of Control and Optimization*, 27, 1108–1135.

Ethier, S., and T. Kurtz (1986): *Markov Processes: Characterization and Convergence*. Wiley-Interscience.

Evans, G. W., and S. Honkapohja (2001): *Learning and Expectations in Macroeconomics*. Princeton University Press.

Foster, D. P., and H. P. Young (2003): "Learning, Hypothesis Testing and Nash Equilibrium," *Games and Economic Behavior*, 45, 73–96.

Fudenberg, D., and D. K. Levine (2009): "Self-Confirming Equilibrium and the Lucas Critique," *Journal of Economic Theory*, 144, 2354–71.

Gilboa, I., A. W. Postlewaite, and D. Schmeidler (2008): "Probability and Uncertainty in Economic Modeling," *Journal of Economic Perspectives*, 22, 173–188.

Hansen, L. P., and T. J. Sargent (2007a): "Recursive Robust Estimation and Control Without Commitment," *Journal of Economic Theory*, 136, 1–27.

——— (2007b): *Robustness*. Princeton University Press.

Hansen, M. H., and B. Yu (2001): "Model Selection and the Principle of Minimum Description Length," *Journal of the American Statistical Association*, 96, 746–774.

Jovanovic, B. (2009): "Learning and Discovery," New York University.

Kandori, M., G. Mailath, and R. Rob (1993): "Learning, Mutation and Long Run Equilibria in Games," *Econometrica*, 61, 27–56.

KOSTYSHYNA, O. (2010): "Application of an Adaptive Step-Size Algorithm in Models of Hyperinflation," forthcoming in *Macroeconomic Dynamics*.

KREPS, D. M. (1998): "Anticipated Utility and Dynamic Choice," in *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures, 1983-1997*. Cambridge University Press.

KUSHNER, H. J., AND G. G. YIN (1997): *Stochastic Approximation Algorithms and Applications*. Springer-Verlag.

LUCAS, R. E., AND T. J. SARGENT (1981): *Rational Expectations and Econometric Practice*. University of Minnesota Press.

LUCAS, JR., R. E. (1976): "Econometric Policy Evaluation: A Critique," in *The Phillips Curve and Labor Markets*, ed. by K. Brunner, and A. Meltzer. Carnegie-Rochester Conf. Series on Public Policy.

MARCET, A., AND T. J. SARGENT (1989): "Convergence of Least Squares Learning Mechanisms in Self Referential Linear Stochastic Models," *Journal of Economic Theory*, 48, 337–368.

MCGOUGH, B. (2003): "Statistical Learning with Time-Varying Parameters," *Macroeconomic Dynamics*, 7, 119–39.

SARGENT, T. J. (1993): *Bounded Rationality in Macroeconomics*. Clarendon Press.

——— (1999): *The Conquest of American Inflation*. Princeton University Press.

——— (2008): "Evolution and Intelligent Design," *American Economic Review*, 98, 5–37.

SARGENT, T. J., AND N. WILLIAMS (2005): "Impacts of Priors on Convergence and Escapes from Nash Inflation," *Review of Economic Dynamics*, 8, 360–391.

SCHORFHEIDE, F. (2000): "A Loss-Function Based Evaluation of DSGE Models," *Journal of Applied Econometrics*, 15, 645–70.

SIMON, H. A., AND A. ANDO (1961): "Aggregation of Variables in Dynamic Systems," *Econometrica*, 29, 111–38.

SIMS, C. A. (1982): "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, 1:1982, 107–164.

SIN, C.-Y., AND H. WHITE (1996): "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, 71, 207–225.

VUONG, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometica*, 57(2), 307–333.

WHITE, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

——— (1994): *Estimation, Inference and Specification Analysis*. Cambridge University Press.

YIN, G. G., AND V. KRISHNAMURTHY (2005): "LMS Algorithms for Tracking Slow Markov Chains With Applications to Hidden Markov Estimation and Adaptive Multiuser Detection," *IEEE Transactions on Information Theory*, 51(7), 2475–91.