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. 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 used to address two questions posed by 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

This paper offers fresh insight into an age-old question - How should policymakers balance theory and empirical evidence? We study one particular approach to answering this question. It consists of the following four-step trial-and-error strategy: (1) An agent entertains a competing set of models, $\mathcal{M}$, called the 'model class', each containing a collection of unknown parameters. The agent suspects that all his models are misspecified; (2) As a result, 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, the agent experiments by selecting a new model from $\mathcal{M}$. 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.

Our paper builds on the previous work of Sargent (1999). Sargent compares two alternative histories of the rise and fall of postwar U.S. inflation. These histories differ in the roles played by theory and empirical evidence in macroeconomic policy. According to the "Triumph of the Natural Rate Hypothesis", inflation was conquered by a Central Bank that listened to theorists. Theorists convinced the Bank to incorporate the public's expectations into its model. According to the "Vindication of Econometric Policy Evaluation", inflation was instead conquered by a Central Bank that adapted a simple reduced form statistical model to evolving conditions. Our model validation approach blends elements of both the 'Triumph' story and the 'Vindication' story. According to model validation,

Date: June, 2014.

We thank the editor and three anonymous referees for helpful comments. We are also grateful to Jim Bullard, Steve Durlauf, 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.

the role of theorists is to convince policymakers to add (good) models to $\mathcal{M}$. The role of econometricians is to then evaluate these models empirically. We argue this blending of theory and evidence is reasonably descriptive of actual practice. Policymakers rarely trust models that have obvious data inconsistencies. However, good policymakers know that it is all too easy to make bad models fit the data, so it is important that models be based on sound economic theory, even if there is disagreement about the right theory. Despite its descriptive realism, the normative implications of model validation are not well understood, and so one of our goals is to shed light on the conditions under which it produces good outcomes, and just as important, when it does not. Given the assumed endogeneity of the data-generating process, this kind of question has been neglected by the traditional econometrics literature.

1.1. **A Motivating Example.** It is useful to begin by illustrating model validation in action. This will highlight both its strengths and its potential weaknesses. We do this by revisiting Sargent's (1999) *Conquest* model. Sargent studied the problem of a Central Bank that wants to minimize a quadratic loss function in unemployment and inflation, $E(u_n^2+\pi_n^2)$, but is unsure about the true model. The Bank posits a reduced form regression model of the form, $u_n = \gamma_0+\gamma_1\pi_n$, and then tries to learn about it by adaptively updating the parameter estimates using a (discounted) least-squares algorithm. The Bank's optimal inflation target, $x_n = -\hat{\gamma}_{0,n}\hat{\gamma}_{1,n}/(1+\hat{\gamma}_{1,n}^2)$, evolves along with its parameter estimates. Unbeknownst to the Bank, the true relationship between u_n and π_n is governed by a Natural Rate model, in which only unanticipated inflation matters, $u_n = u^* - \theta(\pi_n - x_n) + v_{1,n}$, where u^* is the natural rate of unemployment, and $\pi_n - x_n = v_{2,n}$ represents unexpected inflation. The inflation shock, $v_{2,n}$, is i.i.d. Notice that the Bank's model is misspecified, since it neglects the role of expectations in shifting the Phillips Curve. Evolving expectations manifest themselves as shifts in the estimated intercept of the reduced form Phillips Curve.

The top left panel of Figure 1 illustrates the resulting inflation dynamics, using the same parameter values as Sargent (1999).

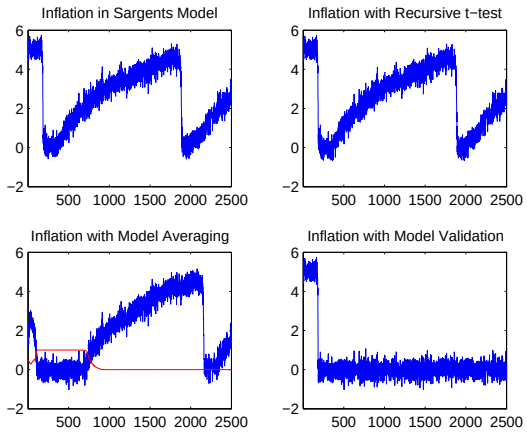

FIGURE 1. Model Averaging vs. Model Validation in Sargent's Conquest Model

The striking feature here is the recurring cycle of gradually rising inflation, and occasional sharp inflation stabilizations. As noted by Cho, Williams, and Sargent (2002), this cycle represents the interplay between the model's *mean dynamics* and its *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 (SCE) value of 5%. The escape dynamics are more exotic. At the SCE, the 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 Bank to revise downward its Phillips Curve slope estimate, and therefore, its inflation target. Since in truth there is no expoitable trade-off, these inflation reductions produce further downward revisions, and the process feeds on itself until inflation reaches the Ramsey outcome of zero inflation. From here, with no further changes in the inflation target, the 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 to the SCE.

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 policymakers spend most of their time looking for new and improved models, not refining estimates of a *given* model. In Sargent (1999), the Central Bank never really evaluates the Phillips Curve as a theoretical model of inflation and unemployment; it merely reconsiders the strength of an unquestioned trade-off. Evidently, this produces a bad outcome, as the Bank repeatedly succumbs to the temptation to try to exploit the Phillips Curve.

The remaining three panels of Figure 1 therefore explore the consequences of more sophisticated learning strategies, assuming the Bank confronts the same sequence of exogenous shocks. The top right panel assumes the Bank engages in a traditional process of hypothesis testing. In particular, suppose the Bank entertains the *possibility* that there is no trade-off. In response, the Bank decides to sequentially test the hypothesis that $\gamma_1 = 0$, and if the hypothesis is not rejected, it sets the inflation target to zero. Clearly, this makes virtually no difference, other than 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 causes the Bank to quickly reject the null hypothesis that $\gamma_1 = 0$. 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 public's expectations in the inflation process. To break out of its inflation cycle, the Bank must consider other models.

The bottom two panels of Figure 1 assume the Bank has *two* models: (1) the statistical Phillips Curve, as before, and (2) a vertical Phillips Curve, $u_n = \gamma_0 + \varepsilon_n$. The second model calls on the Bank to always set target inflation to zero. The problem is that it is not sure which model is correct. The lower left panel assumes the Bank responds to this uncertainty in traditional Bayesian fashion, by averaging across the two models. That is, it assigns a prior to the two models, updates its prior as new data come in, and each period sets its inflation target as the current probability weighted average of the target recommended by each of the two models, which is optimal given its quadratic loss function.[1] The red line plots the recursively estimated probability of the Natural Rate model. The initial prior is assumed to be (.5, .5), with parameter values initialized at their respective SCE values.

[1]Note, the Bank is not fully complying with an optimal Bayesian strategy, since we assume it does not actively experiment.

Although early on the Bank has confidence for awhile in the Natural Rate Hypothesis, it eventually comes to believe that the misspecified Phillips Curve is the correct model, and it never regains any confidence in the Natural Rate Hypothesis. How can this be? How could a Bayesian ever settle on the wrong model when the true model is in the support of his prior? The usual Bayesian consistency theorems do not apply here because the vertical Phillips Curve does in fact contain a subtle misspecification, since the data continue to be generated by Sargent's expectations-augmented Phillips Curve, in which one of the shocks is unexpected inflation. This introduces feedback from the Bank's policy to the actual data-generating process, which the vertical Phillips curve neglects. In contrast, the statistical Phillips curve exploits this feedback to improve its relative fit, and so it eventually drives out the vertical Phillips Curve.[2]

The lower right panel of Figure 1 illustrates what happens under model validation. The Bank has the same two models as before, but now selects just one model when formulating its inflation target. The selection is based on the outcome of a recursive Lagrange Multiplier test (discussed in more detail below). The current model continues to be used as long as it appears to be well specified. If the current model is rejected, a new model is randomly selected, with selection probabilities determined by historical relative forecast accuracy. (The figure uses a logit function with a 'choice intensity parameter' of 2). The simulation is initialized by assuming the Bank begins with Sargent's statistical Phillips Curve, with parameters set at their SCE values. As before, parameter estimates eventually escape from the neighborhood of the SCE, and move toward the Ramsey outcome. In Sargent's analysis, this large and rapid movement does not lead the Bank to reconsider the validity of the Phillips Curve. In contrast, under model validation, the escape triggers a rejection of the specification test, and the Bank switches to the vertical Phillips Curve. Once it does so, it 'never' goes back, and inflation remains at the Ramsey outcome. Note, the Bank does not rule out the possibility of an expoitable trade-off once it switches to the vertical Phillips Curve. The vertical Phillips Curve continues to be tested just as the exploitable Phillips Curve was tested. The difference is that the likelihood of escape and rejection is orders of magnitude smaller for the vertical Phillips Curve, and so for all practical purposes the Bank learns not to exploit the Phillips Curve. The analysis in the paper will explore in detail why some models are more resilient to repeated specification testing than others. We shall see that a key part of the story lies in the strength of their self-referential feedback.

1.2. **Lessons.** So what lessons have been learned here? First, the comparison between model validation and recursive t-testing highlights the importance of allowing agents to

[2]Evans, Honkapohja, Sargent, and Williams (2013) contains a similar result. They consider a standard cobweb model, in which agents average between a time-varying parameter specification, and a constant-parameter specification. They assume the true model has constant parameters, but find that Bayesian model averaging often converges to the time-varying parameter model, even when the initial prior puts significant weight on the true constant parameter model. This occurs when self-referential feedback from beliefs to outcomes is sufficiently strong. Cogley, Colacito, and Sargent (2007) is also similar. They consider a Central Bank that averages between a Natural Rate model and a statistical Phillips Curve. In their model, the Central Bank always learns the true model. Two factors explain the difference: (1) Their Natural Rate model is not misspecified, as it correctly conditions on expected inflation. This eliminates one source of feedback. (2) They assume the Bank knows each model's parameter values, so that policy only depends on model weights, not parameter estimates. This eliminates the other source of feedback.

entertain multiple models. A statistician might argue that the difference between the statistical Phillips Curve and the vertical Phillips Curve cannot possibly be relevant, since the vertical Phillips Curve is nested within the statistical Phillips Curve. Isn't there really just one model here? By starting with the more general specification, wouldn't a good econometrician eventually discover the right model? Although this is a valid argument when the data are exogenous and the general model encompasses the true model, it does *not* apply when the data are endogenous and all models are misspecified, e.g., when alternative models respond differently to underlying feedback in the data. Fully capturing the intricate feedbacks that exist between macroeconomic policy and time-series data is a challenging exercise, to say the least, and so it is important to devise learning strategies that are reliable even when all models potentially misspecify this feedback. Second, the comparison between model validation and model averaging simply reinforces this point. A Bayesian would never commit to a single model on the basis of a hypothesis test. Why not hedge your bets and average? Again, this makes sense when the prior encompasses the truth, but there is no guarantee it produces good outcomes when priors are misspecified. The above example illustrates the dangers of model averaging with endogenous data and misspecified models.

Although suggestive, the above simulations are just an example. How, if at all, do they generalize? As in Sargent (2008) and Fudenberg and Levine (2009), our goal in this paper is to understand how feedback and experimentation interact to influence macroeconomic model selection. Addressing this question poses serious technical challenges. With endogenous data and multiple models, each with adaptively estimated coefficients, the underlying state of the economy is of 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 19th century celestial mechanics. Marcet and Sargent (1989) were the first to apply it in the macroeconomic learning literature.

Our analysis 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 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. Our analysis therefore 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 interpretation of this limiting model. We show that it is the model possessing the largest 'rate function'. A key result in the theory of large deviations (Sanov's theorem) links this rate function to relative entropy and the Kullback-Leibler Information Criterion (KLIC). 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 an 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 derives explicit expressions for the case of linear Gaussian models. These expressions show that feedback is a key determinant of a model's durability. Section 6 returns to Sargent's (1999) *Conquest* model. We first use our large deviations analysis to explain the results in Figure 1. We then go on to consider a second example. In this example the Bank is unsure about identification; in particular, whether to impose a Classical or Keynesian identification restriction. Here model validation leads to the 'wrong' model. Section 7 briefly discusses some related literature, while Section 8 offers a few concluding remarks. An appendix contains proofs of some technical results.

2. Overview

Incorporating multiple models into the learning literature raises a host of new questions and issues. This section briefly outlines how model validation addresses these issues. Many of the ingredients are inspired by the discussion in Sims (2002), who visited the world's major Central Banks, and described their basic policymaking strategies. Interestingly, these strategies are quite similar. They share the following features: (1) They all use multiple models, (2) The models have evolved over time in response to both theory and data, (3) At a given point in time, there is a reliance on a 'primary model', and (4) The process itself is decentralized between a professional staff that develops and monitors the models, and a smaller group of appointed policymakers who make decisions by combining model projections with other (more subjective) data. Sims (2002) goes on to criticize many of these practices, and advocates a more Bayesian approach to policy. In contrast, we adopt a less normative and more descriptive viewpoint, and seek to characterize the outcomes produced by this process.

2.1. Why Not Bayesian?

Bayesian decision theory offers an elegant and theoretically coherent methodology for dealing with model uncertainty.[3] From a Bayesian perspective, there is no meaningful

[3]See, e.g., Brock, Durlauf, and West (2007) for an application of Bayesian decision theory to model uncertainty and macroeconomic policy.

distinction between model uncertainty and parameter uncertainty. A Bayesian would proceed as follows: (1) Formulate a single all-encompassing 'hypermodel', which nests all possible models. This converts model uncertainty into parameter uncertainty, (2) Update beliefs about parameters (including those that index alternative models) using Bayes rule, (3) Recognize that your own beliefs are part of the current state, and actively experiment in order to improve future decisions, (4) Do *not* select a single model when formulating policy. Instead, hedge your bets by averaging across them using estimates of their current probabilities, and finally (5) If for some reason you decide to select a single model, base your selection on expected losses, not statistical fit. Since model validation violates *all* of these precepts, before doing anything we should offer a few words of explanation.

We depart from Bayesian decision theory for three reasons. First, as noted above, our approach here is more positive than normative. Our goal is not to recommend an optimal strategy, but rather to study the properties of a strategy that is used in practice. It is important to know when such a strategy produces good outcomes and when it does not. Second, although a Bayesian approach is attractive in theory, there are serious practical difficulties associated with it. Of course, many of the *computational* challenges associated with Bayesian methods have been overcome, thanks to fast computers and clever monte carlo simulation algorithms.[4] However, those aren't the challenges that concern us. The challenges that concern us were illustrated in the above example. Bayesian model averaging converged to the 'wrong' model there because *all* models were misspecified. The fact is, most of the normative claims of Bayesian methods are lost when one entertains the possibility that priors are misspecified.[5] Finally, our third reason for not being Bayesian is more philosophical, and goes back to the roots of Bayesian decision theory. Savage (1972) himself, the leading light of Bayesian decision theory, was careful to caution against the misapplication of 'small worlds' Bayesian methods to 'large worlds' problems. On page 16 he states - "*It is even utterly beyond our power to plan a picnic or to play a game of chess in accordance with the principle, even when the world of states and the set of available acts to be envisaged are artificially reduced to the narrowest reasonable limits*". Since macroeconomic stabilization policy is every bit as difficult as planning a picnic, it is not unreasonable to consider nonBayesian approaches. Actually, this third reason is not so different from the second. 'Large world' problems are presumably those with infinite-dimensional parameter spaces, where prior misspecification occurs almost surely. Even without feedback, it has long been known that infinite dimensional parameter spaces create problems for Bayesian methods (Diaconis and Freedman (1986)). The problem is even worse with feedback (Nachbar (1997)).

2.2. Fit vs. Losses

Given that one is going to select, even if provisionally, a single model as the basis for policy, there remains the question of how to do this. According to Bayesian decision theory, the choice should be based on expected losses. The institutional structure of most policy

[4]See Schorfheide (2013) for a recent survey.

[5]Bayesian practitioners are well aware of this problem. For example, Schorfheide (2000) and Geweke (2010) propose strategies designed to make Bayesian methods less sensitive to prior misspecification. These strategies involve expanding the (theoretically motivated) model class to include (atheoretical) models that fit the historical data. Unfortunately, as Lucas (1976) and Sargent (2008) have observed, fitting the historical data does not immunize you against misspecification.

environments makes this difficult in practice. Policy environments in economics often mimic those in the natural sciences, which feature a division of labor between technicians, who build and validate models, and policymakers, who evaluate costs and benefits and make decisions based (partly) on model projections. The fact is, model builders often have imperfect knowledge of the objectives and contraints of policymakers, which makes loss-based model selection procedures problematic. Model validation focuses on the problem of the technicians. These technicians blend theory and evidence in an effort to give the best advice possible to policymakers. It is not too surprising that a separation between model validation and decision-making can produce bad outcomes. Section 6 provides an example.[6] Perhaps more surprising is the observation that it sometimes *does* produce good outcomes, as we saw in Figure 1.

2.3. Counterfactuals

The fact that the data-generating process responds to the agent's own beliefs is a crucial issue even without model uncertainty. It means 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. 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. 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 more important. We depart from tradition here by assuming the agent is aware of feedback, even though he responds to it in a less than optimal manner. In particular, he realizes that with multiple models he confronts a difficult counterfactual - How would things have been different if instead a different 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 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.

2.4. Specification Testing

We assume the agent sticks with a model until sufficient evidence mounts against it. This evidence takes the form of an econometric specification test. It turns out specification testing can be 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,

[6]Kocherlakota (2007) provides more examples of the dangers of basing model selection on model fit.

and Priouret (1990)). When the score statistic exceeds a time-varying theshold tuned to the environment's feedback, it indicates that required parameter changes are faster and larger than specified by the constant gain null hypothesis of gradual parameter drift.[7]

2.5. 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. This does not imply 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 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 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.

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.

2.6. 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

Traditional learning models focus on the dynamic interaction between beliefs and observed data. Beliefs are summarized by the current estimates of a model's parameters,

[7]Another possible response to an excessively large score statistic would be to allow the update gain to increase. See Kostyshyna (2012) for an analysis of this possibility.

$\beta_n \in R^{d_1}$, and are updated recursively as follows,

$$\beta_n = \beta_{n-1} + \epsilon_n G(\beta_{n-1}, X_n)$$

where $G : R^{d_1} \times R^{d_2} \to R^{d_1}$ captures some notion of the model's fit or performance. In many applications G is determined by least squares orthogonality conditions. The 'gain sequence', ϵ_n dictates how sensitive beliefs are to new information. In stationary environments $\epsilon_n \to 0$, as each new observation becomes less important relative to the existing stock of prior knowledge. However, in nonstationary environments it is optimal to keep ϵ bounded away from zero, and that is the case we consider. For simplicty, we assume it is a constant. The data, $X_n \in R^{d_2}$, evolve according to a stationary Markov process

$$X_n = F(\beta_{n-1}, X_{n-1}, W_n)$$

where $W_n \in R^{d_3}$ is a vector of exogenous shocks. The function $F : R^{d_1} \times R^{d_2} \times R^{d_3} \to R^{d_2}$ captures the interaction between the agent's model, his beliefs, and the underlying environment. The key feature here is the dependence of F on β. This makes the environment 'self-referential', in the sense that the underlying data-generating process depends on the agent's beliefs.

Most of the existing learning literature has been devoted to studying systems of this form. This is not an easy task, given that it represents a nonlinear, dynamic, stochastic system. Until recently attention primarily focused on long-run convergence and stability issues. A more recent literature has focused on sample path properties. As noted earlier, the key to making the analysis feasible is to exploit the fact that β_n and X_n operate on different time-scales, due to the presence of ϵ in the parameter update equations. As $\epsilon \to 0$, β_n evolves more and more slowly relative to X_n. This allows us to decouple the analysis of the interaction between beliefs and outcomes into two separate steps: (1) First examine the long-run *average* behavior of the data for *given* beliefs, $\bar{X}(\beta)$, and then substitute this averaged behavior into the belief update equations, which then produces an *autonomous* equation for β_n. In the economics literature, this 'mean ODE approach' was pioneered by Marcet and Sargent (1989). Evans and Honkapohja (2001) provide a detailed summary.

Model validation represents a natural extension of these ideas. When agents entertain multiple candidate models, one can simply index them by a discrete model indicator, $s_n \in \mathcal{M}$, and then think of s_n as just another parameter. However, in contrast to Bayesian decision theory, the agent is assumed to display a greater reluctace to revise beliefs about s_n than about other parameters.[8] Our maintained assumption is that the agent believes there to be a single best model within $\mathcal{M}$, and the goal is to find it, while at the same time balancing the ongoing pressures of meeting control objectives.

Technically, the key issue is to show that specification testing introduces a *third* time-scale, in the sense that the evolution of s_n can itself be decoupled from the evolution of each model's parameter estimates. So, to begin the analysis, we just need to augment the updating and data processes as follows:

$$\beta_n^i = \beta_{n-1}^i + \epsilon G_i(s_{n-1}, \beta_{n-1}, X_n) \qquad \forall i \in \mathcal{M} \tag{3.1}$$

$$X_n = F(s_{n-1}, \beta_{n-1}, X_{n-1}, W_n) \tag{3.2}$$

[8]We thank an anonymous referee for providing this interpretation. Note, this kind of distinction between models and parameters is also present in Hansen and Sargent's (2008, chpt. 18) work on robust learning.

The first equation implies the parameter estimates of each model will in general depend on which model is currently being used. This reflects our assumption that each model is continuously updated, even when it is not being used to make decisions. The second equation makes clear that now the economy is operating with a new layer of self-reference. Not only does policy potentially depend on beliefs about a particular model, but also on the model itself. The feedback from beliefs to the actual DGP is 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 state variables X_n possess a unique invariant distribution.*

Our analytical methods rely heavily on the ability to 'average out' fluctuations in X_n for given values of the model coefficients and model indicator. Assumption 3.1 guarantees these averages are well defined. Evans and Honkapohja (2001) provide sufficient conditions. To facilitate the analysis, we further 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.*

This compactness assumption can be interpreted as a priori knowledge about the DGP. Although the agent does not know the precise values 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.

As noted earlier, we assume the agent operates under a 'if it's not broke, don't fix it' principle, meaning that models are used until they appear to be misspecified, as indicated by a statistical specification test. The test is performed recursively, in real-time, and for the same reasons that past data are discounted when computing parameter estimates, we assume estimates of the test statistic, $\theta_n \in R^1_+$, are updated using a constant gain stochastic approximation algorithm,

$$\theta_n = \theta_{n-1} + \epsilon^\alpha [Q(s_{n-1}, \beta_{n-1}, X_n) - \theta_{n-1}] \tag{3.3}$$

where $Q(s_{n-1}, \beta_{n-1}, X_n)$ is the time-n value of the statistic. Averaging by letting $\alpha > 0$ is important, as it prevents single isolated shock realizations from triggering model rejections. At the same time, letting $\alpha < 1$ is convenient, since otherwise θ_n evolves more slowly than β_n, making the test statistic depend on the history of the coefficient estimates, rather than just their current magnitudes. This would complicate the derivation of the large deviation properties of the test statistic.

Finally, if $|\mathcal{M}| = m$, then the model indicator parameter, s_n, follows an m-state Markov chain that depends on the current value of the test statistic relative to an evolving test threshold, $\bar{\theta}_n$,

$$s_{n+1} = \mathcal{I}_{(\theta_n \leq \bar{\theta}_n)} \cdot s_n + (1 - \mathcal{I}_{(\theta_n \leq \bar{\theta}_n)}) \cdot \Pi_{n+1} \tag{3.4}$$

where $\mathcal{I}$ is an indicator function, and Π_n is a post-rejection model selection distribution with elements $\{\pi^i_n\}$, $i = 1, 2, \cdots m$. The only restriction that must be placed on Π_n is that it have full support $\forall n$. In practice, one would expect Π_n to evolve, and to reflect the historical relative performance of the various models, with better performing models

receiving higher weights. Our analysis permits this, as long as the elements of Π_n remain strictly positive. In our simulations we use the logit function,

$$\pi_n^i = \frac{e^{-\phi\omega_{i,n}}}{\sum_{j=1}^m e^{-\phi\omega_{j,n}}}$$

where $\omega_{i,n}$ is an estimate of model-i's one-step ahead forecast error variance, and ϕ is a 'choice intensity' parameter. As ϕ increases, the agent is less prone to experiment.

Equations (3.1)-(3.4) summarize our model validation framework. Our goal is to understand the asymptotic properties of this system (as $\epsilon \to 0$), with a special focus on the asymptotic properties of the model selection parameter, s_n.

3.1. **Mean ODEs and Self-Confirming Equilibria.** Notice the X_n vector on the right-hand side of (3.1) corresponds to the actual law of motion given by (3.2), which depends on both the current model and the agent's control and estimation efforts. This makes the agent's problem self-referential. It also makes analysis of this problem difficult. To simplify, we exploit a two-tiered time-scale separation, one between the evolution of the data, X_n, and the evolution of each model's coefficient estimates, $\hat{\beta}_n^i$, 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 (3.2) into the parameter update equations in (3.1), (2) Freeze the coefficient estimates and model indicator at their current values, (3) Average over the stationary distribution of the 'fast' variables, X_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. Let $P_{\beta,s}^k(X_{n+k} \in \cdot | X_n)$ 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

$$g_i(s, \beta^i) = \int G_i(s, \beta^i, \xi)\Gamma_{\beta,s}(d\xi) \tag{3.5}$$

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

Assumption 3.3. *For all i and s, $g_i(s, \beta^i)$ is a 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 g 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),

$$\dot{\beta}_s^i = g_i(s, \beta_s^i) \qquad s = 1, 2, \cdots m \tag{3.6}$$

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 (3.6), i.e., $g_i(s, \beta_{i,s}^*) = 0$. 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. Finally, note that if a model's own self-confirming equilibrium is *unstable*, one would expect its relative use to shrink rapidly to zero as $\epsilon \to 0$, since in this case model rejections are *not* rare events triggered by large deviations. Instead, they occur rapidly in response to a model's own mean dynamics.

3.2. **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 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 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, consider the case of a linear VAR model, $X_n = \beta X_{n-1} + \varepsilon_n$, where X_n is $s \times 1$, and where the true DGP is also linear, $X_n = T(\beta)X_{n-1} + C(\beta)v_n$. (We study this case in more detail in Section 5). Letting Λ_n be an $s \times s$ matrix with columns containing each equation's score vector, we have

$$\Lambda_n = R_{n-1}^{-1} X_{n-1}[X'_{n-1}(T(\beta_{n-1})' - \beta'_{n-1}) + v'_n C(\beta)']$$

where R_n is a recursive estimate of the second moment matrix of X_n. The null hypothesis is then, $H_0 : vec(\Lambda_n)'\Omega_n^{-1} vec(\Lambda_n) \leq \bar{\theta}_n$, where Ω_n is a recursive estimate of $\text{var}(vec(\Lambda_n))$

$$\Omega_n = \Omega_{n-1} + \epsilon^{\alpha}[vec(\Lambda_n)vec(\Lambda_n)' - \Omega_{n-1}]$$

and 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.[9]

3.3. **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. For example, it could be based on a deliberate search process which favors models that have performed relatively well historically. The only essential feature is that the support of the distribution remain full. 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_n^i, where $\pi_n^i > 0 \ \forall i, n$. Hence, the most recently rejected model may be re-selected, although this probability could be made arbitrarily small.

4. Analysis

Let $\mathcal{X}_n = (X_n, \hat{\beta}^1, \hat{\beta}^2, \cdots \hat{\beta}^m, \theta_n, \bar{\theta}_n, s_n)$ 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,

[9]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.

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.

4.1. **Representation as a Stochastic Recursive Algorithm.** We have purposely stayed as close as possible to the standard SRA framework.[10] 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 for the case of linear VAR models and a linear DGP:

We first have a set of model update equations,

$$\hat{\beta}_n^i = \hat{\beta}_{n-1}^i + \epsilon \Lambda_n^i \tag{4.7}$$

$$\Lambda_n^i = (R_{n-1}^i)^{-1} X_{n-1}^i [(X_n^i)' - (X_{n-1}^i)' \hat{\beta}_{n-1}^i] \tag{4.8}$$

$$R_n^i = R_{n-1}^i + \epsilon [X_{n-1}^i (X_{n-1}^i)' - R_{n-1}^i] \tag{4.9}$$

Through feedback, these determine the actual DGP,

$$X_n = A(s_{n-1}, \beta_{n-1}) X_{n-1} + C(s_{n-1}, \beta_{n-1}) v_n \tag{4.10}$$

Models are tested by forming the recursive LM test statistics

$$\theta_n^i = \theta_{n-1}^i + \epsilon^\alpha [vec(\Lambda_n^i)' \hat{\Omega}_{i,n}^{-1} vec(\Lambda_n^i) - \theta_{n-1}^i] \tag{4.11}$$

$$\Omega_{i,n} = \Omega_{i,n-1} + \epsilon^\alpha [vec(\Lambda_n^i) vec(\Lambda_n^i)' - \Omega_{i,n-1}] \tag{4.12}$$

Hence, θ_n^i is just a recursively estimated χ^2 statistic. If a model contains p parameters, its degrees of freedom would be p.

Finally, the model indicator, s_n, evolves as an m-state Markov Chain, where $m = |\mathcal{M}|$. 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

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

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

where $\bar{\theta}_n$ is a sequence of test thresholds (to be discussed below).

4.2. **Time-Scale Separation.** Equations (4.7) - (4.16) constitute a high-dimensional system of nonlinear stochastic difference equations. 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, X_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 X_n variables for given values of the model coefficients and model indicator. This produces equation (3.5). 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,

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

Proof. See Appendix A. □

Figure 2 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 $\epsilon \to 0$.

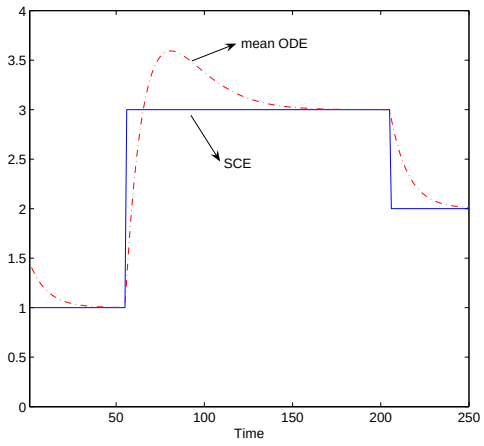

FIGURE 2. Mean ODE Approximation

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 $\epsilon \to 0$ $U_i^\epsilon(t)$ converges weakly to the solution of the stochastic differential equation*

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

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

$$\mathcal{R}_{ij}(\beta) = \sum_{k=-\infty}^{\infty} cov[G_i(\beta, X_k(\beta), x_k(\beta)), G_j(\beta, X_0(\beta), x_0(\beta))]$$

Proof. See Appendix B. □

This result can be used to calibrate the test threshold, $\bar{\theta}_n$. The sequence of test statistics in 4.11 can also be approximated by a diffusion on the time-scale $t_\theta = t \cdot \epsilon^{\alpha-1}$. Under the null, the mean dynamics are simple, $E(\Lambda'\Omega^{-1}\Lambda) = p$, the number of degrees of freedom of the test (i.e., the number of model coefficients). Letting $\tilde{\theta}^\epsilon(t_\theta) = \theta^\epsilon(t_\theta) - p$, we get the following Ornstein-Uhlenbeck approximation for the path of the test statistic,

$$d\tilde{\theta} = -\tilde{\theta}dt_\theta + \sqrt{\epsilon^\alpha}\mathcal{R}dW$$

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 $\sqrt{\epsilon^\alpha \cdot \mathcal{C}}$, 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 $\mathcal{E}$ be a separable Banach space. Suppose $\mathbf{S}_n, n > 0$ are $\mathcal{E}$-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 $I : \mathcal{E} \to [0, \infty]$, 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 $A \subset \mathcal{E}$, 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 $B \subset \mathcal{E}$

In our setting, $\mathbf{S}_n$ will either be a sequence of coefficient estimates, or a sequence of test statistics, with $\mathcal{E}$ 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.

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),

$$\begin{aligned} \beta_n &= \beta_{n-1} + \epsilon g(s, \beta) + \epsilon[G(s_{n-1}, \beta_{n-1}, X_{n-1}, W_n) - g(s, \beta)] \\ &= \beta_{n-1} + \epsilon g(s, \beta) + \epsilon \tilde{G}(s_{n-1}, \beta_{n-1}, X_{n-1}, W_n) \end{aligned}$$

so that $\tilde{G}(\cdot)$ represents the martingale difference component of the update algorithm. We assume $\tilde{G}(\cdot)$ 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),*

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

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 $\mathcal{H} : \mathcal{M} \times \mathcal{B} \times \mathbb{R}^d_{++} \mapsto \mathbb{R}_+$, where d is the dimensionality of the parameter space. Existence of $\mathcal{H}(s, \beta, a)$ imposes restrictions on the tails of the data and the shocks, and must be verified on a case-by-case basis.[11]

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

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

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,

[11] In Cho and Kasa (2013) 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 $L(s, \beta, \lambda)$ 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 equilibrium 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:

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

subject to the boundary conditions $\beta(0) = \beta_0$, $s(0) = s_0$, and $\beta(T) \in \partial B$, where ∂B 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,

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

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

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

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

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

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 B 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, W^i are i.i.d. and unbounded, and there exist constants $\kappa > 1$ and $Q < \infty$ such that $\forall n$ and $s \geq 0$*

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

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

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

(2) *If the exogenous shocks, W^i, are bounded, and S_i^* is continuous on ∂B, then*

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

(3) *Given the assumptions of part (2), and letting* $\tau_i^\epsilon = \inf_{t \leq T}(\beta_i^\epsilon(t) \notin B)$ *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 $\beta_i^\epsilon(0) \in B$. This reflects the fact that the mean dynamics are stabilizing for all $\beta^\epsilon(t) \in B$, so it is very likely that $\beta_i^\epsilon(t)$ converges to a small neighborhood of β_i^* before it succeeds in escaping.[12] 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 B 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. Fourth, since S^* depends on B, the escape boundary, so do the escape probabilities and mean escape times. The 'bigger' B is, the less likely an escape.[13]

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, ∂B, 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.[14] Hence, we must now translate the LD results for the coefficients into LD results for the LM test statistics in equation (4.11).

To do this, we make use of the following result (Dembo and Zeitouni (1998), p. 126):

Theorem 4.6. (Contraction Principle) *Let* X *and and* Y *be Hausdorff topological spaces and* $f : X \to Y$ *a continuous function. Consider a rate function* $S : X \to [0, \infty]$.

[12] 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 ∂B before both get within a ρ_2-neighborhood of β^* goes to zero as $\epsilon \to 0$.

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

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

(a): For each $y \in Y$ define $S'(y) = \inf\{S(x) : x \in X, \ \ y = f(x)\}$. Then S' is a rate function on Y.

(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.

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 use this result we must establish that θ_n is in some sense a continuous function of β_n. This requires two steps. First, define the function, $F^i : \mathbb{R}^{d_i} \to \mathbb{R}_+$, where d_i is the number of variables in model i, as the score function, $F^i(\beta^i_{n-1}) = vec(\Lambda^i_n)'\hat{\Omega}^{-1}_{i,n} vec(\Lambda^i_n)$, and then form the continuous-time interpolation of the recursive LM test statistic

$$\theta^\epsilon(t) = \theta^\epsilon(0) + \epsilon^\alpha \sum_{i=0}^{[t/\epsilon]} [F(\beta^\epsilon(i)) - \theta^\epsilon(i)] \tag{4.22}$$

As usual, average out the state dynamics by defining $\theta(t) = \lim_{\epsilon \to 0} \theta^\epsilon(t)$ as its limit. The second step is to note that for $0 < \alpha < 1$ the $\theta^\epsilon(t)$ process evolves faster than the $\beta^\epsilon(t)$ process. More precisely, $\theta^\epsilon(t)$ is 'exponentially equivalent' to $\theta^\epsilon(F(\beta^\epsilon(t))$.[15] This means the LD properties of θ are driven by the LD properties of $F(\beta)$ via the contraction principle. (See Theorem 4.2.13 in Dembo and Zeitouni (1998)). Hence, we have:

Proposition 4.7. *Each model's LM test statistic process, $\theta^\epsilon_i(t)$, has a locally stable equilibrium at $\theta^*_i = F(\beta^*_i) = 0$, and it satisfies a large deviations principle with rate function given by*

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

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

Proof. The stability of θ^* is clear from inspection of (4.22). The proof then just consists in verifying that $F(\beta)$ is continuous. □

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 quite 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,

[15] $\theta^\epsilon(t)$ and $\theta^\epsilon(F(\beta^\epsilon(t)))$ are exponentially equivalent if for each $\delta > 0$, $\limsup_{\epsilon \to 0} \epsilon \log P[d(\theta^\epsilon(t), \theta^\epsilon(F(\beta^\epsilon(t)))) > \delta] = -\infty$, where d is the sup norm on the space of continuous, bounded functions.

as it reflects the presence of nonzero mean dynamics. The agent wants to incorporate this drift into the null hypothesis. We assume that drift is incorporated into the null via a declining test threshold. In other words, the test becomes more stringent the longer a model has been in use.

To be more precise, let $\{n_k\}$ be a sequence of random model switching times, i.e., $n_{k+1} = \inf\{n > n_k : s_n \neq s_{n_k}\}$. Define an 'epoch' as an interval during which a single model is in use. To ensure new models are given a chance to fit their own data, we assume the test threshold begins each epoch at a sufficiently high value that model rejections continue to be rare events, in the above large deviations sense. Since the contribution of the mean dynamics to the test statistic is given by the difference in the gradients of the g^i functions, $\bar{\theta}_0$ should be of the same order as the maximum distance between the gradients. If convergence between SCE is monotone, then $\bar{\theta}_0 \approx \max_{i,j} ||g^i_\beta(i, \beta^{i,*}) - g^i_\beta(j, \beta^{i,*})||$. (If convergence is not monotone, then one would also need to maximize over the paths of β). In principle, one could allow model specific $\bar{\theta}_0$, but the extra conservativism associated with maximizing over i has no bearing on inferences about model dominance. Then, over time, as the currently used model converges to its SCE, the agent can afford to increase the test's power by letting the test threshold decline, i.e., $\bar{\theta}_{n+1} < \bar{\theta}_n \ \ \forall \ \ n \in \{n_k, \cdots n_{k+1} - 1\}$. Note that an *optimal* threshold sequence would require detailed knowledge of the dynamic properties of each model. Such knowledge is not likely to be possessed by actual policymakers. Fortunately, none of our conclusions depend on the assumption that the threshold sequence is optimal in any sense, or that the test's nominal size corresponds to its actual size. To summarize, we impose the following assumption

Assumption 4.8. *There exists a deterministic sequence of test thresholds, $\bar{\theta}_n$, 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 an issue. However, actually computing $\bar{\theta}_n$ will be unavoidably model specific. (See Cho and Kasa (2013) for an example).

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. 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 4.9. *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 for τ fixed, $p^\epsilon(\tau)$ converges weakly as $\epsilon \to 0$ 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 as $\tau \to \infty$,

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

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

Proof. See Appendix C. □

Note that for τ to remain constant, t must increase very rapidly as $\epsilon \to 0$. This reflects the rarity of the escapes.

4.6. **Dominant Models.** The invariant distribution in (4.23) shows what happens when $\tau \to \infty$ 'slower' (or after) $\epsilon \to 0$. It's also useful to ask what happens when $\tau \to \infty$ 'faster' (or before) $\epsilon \to 0$. It's clear from equation (4.23) that this limit is degenerate.

Proposition 4.10. *As $\epsilon \to 0$ the invariant model distribution collapses onto the model with the largest LD rate function.*

This means that in the limit, and over *very* long time horizons, 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 4.9 and 4.10 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 4.11. *As long as the experimentation probabilities, π_t^i, remain strictly bounded between 0 and 1 as $\epsilon \to 0$, 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 $|\mathcal{M}| = 3$. 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,

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

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.

4.9. **Real-Time Relevance of Model Switching.** Model switches are rare events here, which raises the question of whether they occur sufficiently often to be of empirical relevance. In the context of monetary policy, evidence in Romer and Romer (2002) suggests the Fed switched models every 10-20 years during the postwar era, or once every 150-250 periods if the discrete time interval is a month. Can model validation generate switches this often? On the one hand, it almost certainly can. Suppose T is an observed mean switching frequency. The above results imply $T \sim \epsilon \cdot e^{-V^*/\epsilon}$, which increases in ϵ. Hence, one can typically match an observed mean escape time by selecting a sufficiently high gain. On the other hand, remember these are *limit* results, as $\epsilon \to 0$. For a given $\epsilon > 0$, they are just approximations, and the quality of the approximation is case-specific. In some cases $\epsilon = .10$ might be sufficiently small to provide an adequate approximation. In others, $\epsilon = .001$ might be too large. In practice, one could first calibrate ϵ to observed model switches, and then conduct simulations to verify that the predicted escape times match the simulations.

One should keep in mind that the results here are also useful when observed model switching takes place too *infrequently* to be of empirical relevance. That is, one can adopt the perspective of a game-theorist, or of the early advocates of the Rational Expectations Hypothesis, and view our analysis as providing an equilibrium selection criterion. From this perspective, learning has already taken place at some prior stage, and the task is to explain why a given equilibrium/model is observed.

5. The Linear Gaussian Case

The previous section showed that the asymptotic properties of model validation hinge critically on a set of model specific large deviation rate functions. Unfortunately, numerical methods are typically required to compute these rate functions. For example, even if a closed-form expression for the H-functional can be obtained, one must still proceed to solve a nonlinear control problem to find the rate function. In this section, we consider a class of models where some analytical headway is possible. These models feature conditionally linear state dynamics and Gaussian disturbances. In this case, least squares orthogonality conditions are quadratic forms of Gaussian random variables, and we can use the results of Bryc and Dembo (1997) to simplify the resulting calculations.

Consider then a model class where each model is a vector autoregression,

$$X_n = \beta X_{n-1} + \varepsilon_n \tag{5.25}$$

and the actual law given these beliefs is

$$X_n = T_{11}(\beta)X_{n-1} + T_{12}(\beta)Z_{n-1} + v_{1,n} \tag{5.26}$$

where X_n is an $s \times 1$ vector, β and $T_{11}(\beta)$ are $s \times s$ matrices, $T_{12}(\beta)$ is an $s \times q$ matrix, and $v_{1,n} \sim i.i.d.N(0, \Sigma)$. The $T_{ij}(\beta)$ functions encode the feedback between beliefs and outcomes, and can be highly nonlinear. For Assumption 3.1 to be satisfied, they must be appropriately bounded (uniformly). The presence of Z_{n-1} allows for the possibility that the model in question is underparameterized. Alternative models can be represented by alternative specifications of the X and Z vectors. It is assumed that omitted variables also follow a Gaussian vector autogression, $Z_n = T_{21}(\beta)X_{n-1} + T_{22}(\beta)Z_{n-1} + v_{2,n}$. Notice that model specification determines the dynamic properties of a model's error term.

The linear Gaussian model in (5.25)-(5.26) has been a workhorse in applied macroeconomics, both under Rational Expectations and under adaptive learning. For example, it has been used in present value asset pricing models and New Keynesian Phillips curve models. (See Evans and Honkapohja (2001) for a full catalogue of examples). As a simple example, if the true model takes the form of a Lucas supply curve (or cobweb model),

$$X_n = \delta E_{n-1} X_n + \gamma X_{n-1} + v_n$$

and the Perceived Law of Motion (PLM) is correctly specified, then the Actual Law of Motion (ALM) is

$$X_n = (\delta\beta + \gamma) X_{n-1} + v_n = T(\beta) X_{n-1} + v_n$$

Hence, the results derived here are of wide applicability.

5.1. **Calculation of H-functional and Large Deviation Rate Function.** Define $(\Phi_n)' = ((X_n)', (Z_n)')$ as the combined vector of included and excluded variables, so that $\Phi_n = T(\beta)\Phi_{n-1} + v_n$, with $\text{var}(v_n) = \Sigma$. Let $F(\omega)$ be the spectral density matrix of Φ_n, and define the $s \times s$ matrix of co-states, α, where columns are the co-states pertaining to the coefficients in each equation. The rate function is then the solution of the following calculus of variations problem $S(\beta) = \inf_{\dot\beta} \int L(\beta, \dot\beta)$, where L is the Legendre transform of the H-functional. From Bryc and Dembo (1997), the solution of this problem can be stated as follows:

Proposition 5.1. *Assume the following Riccati equation has a unique positive definite solution* $\forall\ \beta \in B$

$$P = \Sigma + T(\beta)' P T(\beta) + T(\beta)' P [(2W(\alpha, \beta)^{-1} - P]^{-1} P T(\beta)$$

where the weighting matrix, $W(\alpha, \beta)$*, is given as follows:*

$$W(\alpha, \beta) = \begin{bmatrix} (T_{11}(\beta) - \beta)' \alpha' R^{-1} + \frac{1}{2} R^{-1} \alpha \Sigma \alpha' R^{-1} & \frac{1}{2} T_{12}(\beta)' \alpha' R^{-1} \\ \frac{1}{2} R^{-1} \alpha T_{12}(\beta) & 0 \end{bmatrix} \tag{5.27}$$

and is assumed to be uniformly positive semi-definite. Then for the case of linear Gaussian VAR models the LD rate function solves the following nonlinear PDE

$$det \mathcal{F}_0(-S_\beta, \beta) = 1 \tag{5.28}$$

where $\mathcal{F}_0 = \mathcal{F}(0)$ *is given by the following (canonical) spectral factorization*

$$I_{s+q} - 2W F(z) F(z^{-1})' = \mathcal{F}(z) \mathcal{F}_0 \mathcal{F}(z^{-1})'$$

Proof. See Appendix D. □

This nonlinear PDE must be solved subject to the boundary conditions, $S(\beta^*) = 0$, where β^* are the self-confirming values of the model's coefficients (presumed to be unique), and $\beta(T) = \partial B$, where ∂B defines the relevant escape boundary. The value of the rate function is then found by taking the minimum over the escape boundary. Clearly, this sort of problem cannot in general be solved with pencil and paper. In practice, since it's first-order, one would use the 'method of characteristics' to convert it to a system of ODEs, which can then be solved relatively easily with standard algorithms. However, rather than pursue a numerical approach, we consider a few special cases, which *can* be solved by hand. This will provide some intuition about what determines a dominant model.

5.2. **Case 1: Correctly Specified Univariate Model.** In this case, $s = 1$, $q = 0$, and only the upper left element of W in (5.27) is relevant. One can then readily verify that (5.28) is solved by

$$-(T_{11}(\beta) - \beta) = \frac{1}{2} S_\beta R^{-1} \Sigma$$

Using the fact that R evolves deterministically, we can first integrate this with respect to β and then evaluate R at β. This yields the following rate function

$$S(\beta) = 2\Sigma^{-1} R(\beta) \int_{\beta^*}^{\beta} [s - T_{11}(s)] ds \tag{5.29}$$

where β^* is the SCE value of β and $R(\beta) = \Sigma/(1 - T_{11}^2(\beta))$. Note that a model will be resilient (i.e., have a large rate function) when the mean dynamics are strong, as represented by $s - T_{11}(s)$. When this term is large, deviations from the SCE produce large discrepancies between the ALM and the PLM, which are easily identified and corrected. This makes escapes difficult. Escapes are also difficult when parameters are precisely estimated, which occurs when R (the second moment matrix of X) is large relative to Σ. Variation in the explanatory variables produces precise parameter estimates. Notice that Σ cancels out of (5.29). On the one hand, a higher Σ adds noise, which makes escape easier. On the other hand, a higher Σ increases variation of the explanatory variables. This makes parameter estimates more precise, and escapes more difficult. In the VAR case, the two effects exactly offset.

Although this is a special case, it does reveal a crucial point that applies more generally, namely, there is no guarantee that true models will be dominant. Misspecified models can dominate either because their feedback is stronger, or because their parameters are more precisely estimated. This highlights a key distinction between specification testing with endogenous data and specification testing with exogenous data.[16]

Finally, note that in the above cobweb example, $T(\beta) = \delta\beta + \gamma$ is *linear*, and the integral in (5.29) can actually be evaluated to give: $S(\beta) = \Sigma^{-1} R(\beta)(1 - \delta)(\beta - \beta^*)^2$, where $R(\beta) = \Sigma/[1 - (\gamma + \delta\beta)^2)]$, and $\beta^* = \gamma/(1 - \delta)$ is the SCE in this case. Note that the feedback parameter, δ, has an ambiguous effect on the rate function. On the one hand, a larger δ produces stronger feedback between the PLM and ALM. This makes specification errors harder to identify, and weakens the pull back to the SCE, which makes escapes easier. On the other hand, a larger δ increases the variance of the regressor, which increases the precision of the parameter estimate, which makes escape more difficult.

5.3. **Case 2: Correctly Specified Multivariate Model With Potential Function.** Again $q = 0$, and only the upper left part of W in (5.27) is relevant. Now, however, this is an $s \times s$ matrix, and the PDE in (5.28) becomes

$$-(T(\beta) - \beta)' S_\beta' R^{-1} = \frac{1}{2} R^{-1} S_\beta \Sigma S_\beta' R^{-1}$$

Assuming S_β is invertible, we can write this as

$$S_\beta = 2R(\beta - T(\beta)' \Sigma^{-1} \tag{5.30}$$

[16]Although more precise estimates yield a lower test threshold, making model rejection *easier*, remember we assume this effect is $O(\text{var}(\hat{\beta}))$, whereas the escape time to a given boundary point is $O(\exp(1/\text{var}(\hat{\beta})))$.

where $R(\beta)$ is the solution of the Lyapunov equation $R = T(\beta)RT'(\beta) + \Sigma$. Although in the scalar case we can always integrate this to derive the rate function, this integration strategy won't work in general in the vector case. However, it will work if appropriate symmetry conditions are satisfied, which guarantees the existence of a *potential function*. To derive these conditions, vectorize the HJB equation in (5.30) to get

$$vec(S_\beta) = 2(\Sigma^{-1} \otimes R(\beta))vec(\beta - T(\beta) \tag{5.31}$$

This gives us

Proposition 5.2. *If there exists a function $V(\beta)$ such that $\nabla V(\beta) = (\Sigma^{-1} \otimes R(\beta))vec(\beta - T(\beta)$ and $V(\beta^*) = 0$ then the rate function for a correctly specified multivariate model is $S(\beta) = 2V(\beta)$.*

When will this potential function exist? Besides the usual differentiability conditions, a necessary and sufficient condition is that $\partial\Psi_i(\beta)/\partial\beta_k = \partial\Psi_k/\partial\beta_i \ \ \forall i, k$ where Ψ_i is the ith component of the vector function on the right-hand side of (5.31). A simple example occurs when both T_{11} and Σ are diagonal, in which case all cross partials are zero. When this is the case, the rate function takes the form

$$S(\beta) = 2\sum_{i=1}^{s} \frac{\int_{\beta^*}^{\beta}[s - T_i(s)]ds}{1 - T_i^2(\beta)}$$

which is an obvious vector analog of the previous scalar example.

5.4. **Case 3: Misspecified Models.** So far we've only considered correctly specified models, which of course is rather limiting given our purposes. The easiest case of misspecification to consider is when relevant variables are omitted. Consider the case of a univariate model ($s = 1$), which can always be integrated, and assume there is a single omitted variable ($q = 1$), which is both exogenous and i.i.d ($T_{21} = T_{22} = 0$). In this case, misspecification just produces an error term with larger variance. However, as noted above, in VAR models this has offsetting effects on the rate function. Hence, if two VAR models differ only in which i.i.d. exogenous variables they exclude, their relative rate functions will depend only on the their own T-mappings.

5.5. **Large Deviations of the Test Statistic.** The previous section showed that the contraction principle can be used to deduce the rate function for θ from the rate function for β. Consider again the case of a correctly specified univariate VAR. Averaging over (x_n, v_n) while keeping β_n fixed, we have $E(\Lambda^2) = [T(\beta) - \beta]^2 + \sigma^2 R^{-1}$ and $\Omega = \sigma^2 R^{-1}$. Hence, $F(\beta) = 1 + \sigma^{-2}R[T(\beta) - \beta]^2$. Proposition 4.7 then yields the following expression for the rate function of θ,

$$V(\theta) = \inf_{\{\beta:\theta=F(\beta)\}} 2\left\{\sigma^{-2}R\int_{\beta^*}^{\beta}[s - T(s)]ds\right\}$$

Suppose $T(\beta)$ is linear, e.g., $T(\beta) = a + b\beta$, so that $\int_{\beta^*}^{\beta}[s - T(s)]ds = \frac{1}{2}[\beta - T(\beta)]^2/(1-b)$. Notice then that the rate function takes the extremely simple form, $V(\theta) = (\theta - 1)/(1 - b)$. Minimization over the escape boundary, ∂B, is also simple, since θ is univariate and obviously positive, so there is only one direction to escape. If $\bar{\theta}_\infty$ is the limiting value of the threshold, the rate function in the neighborhood of the SCE is just $V^* = (\bar{\theta}_\infty - 1)/(1 - b)$. As noted in section 4.4, $\bar{\theta}_\infty$ should be set so that routine fluctuations in θ_n do not trigger

model rejections. These fluctuations are of order $\epsilon^{(1-\alpha)/2}\mathcal{R}$, where $\mathcal{R}$ is the standard deviation in the diffusion approximation of θ (around its SCE). Thus, $\bar{\theta}_\infty > 1+\epsilon^{(1-\alpha)/2}\mathcal{R}$.

Unfortunately, matters aren't quite so simple when T is nonlinear, or when the model is multivariate. In these more general cases, the tight connection between the form of the test statistic and the form of the coefficient rate function will be severed, and V^* will depend on $\bar{\theta}_\infty$ nonlinearly. In practice, it is easier to leave the calibration of the test statistic implicit, and to assume that model rejections are triggered by escapes of the coefficients. Inferences about dominant models can then be based directly on the coefficient rate functions.

6. Applications

This section applies our results to two examples based on Sargent's (1999) *Conquest* model. The first example illustrates the importance of the model class. The second example shows that model validation can be used to learn about identification.

6.1. **Importance of the Model Class.**

Let's return to the example in the Introduction. That example showed, as Sims (1982) argued long ago, that adaptive policymakers could learn to do the right thing even without *a priori* knowledge of the true model. The key was to expand the model class, and consider multiple models. We saw that averaging across these models in a Bayesian fashion, or formulating a single large encompassing model, produced *worse* outcomes than selecting a single small model. We are now in a position to understand why.

Model validation predicts that a dominant model will be the model with maximum large deviations rate function. In this case, this is an easy comparison to make. We know from CWS (2002) that the static Phillips Curve rate function is $\bar{S}^* = .0005$, using the same parameter values as in Sargent (1999). The mean escape time is $\exp[\bar{S}^*/\epsilon]$ continuous time units, or $\epsilon^{-1}\exp[\bar{S}^*/\epsilon]$ discrete time units, where ϵ is the gain. Hence, when $\epsilon = .01$ we should observe escapes every 105 periods.[17] Figure 3 plots a representative inflation path from Sargent's model, along with its associated LM specification test statistic.[18]

With two degrees of freedom a conventional critical value would be in the neighborhood of 6.0. Clearly, during escapes the Bank would have reasons to doubt the specification of its model. Notice, however, that escapes occur less frequently than predicted. This discrepancy could arise for two reasons. First, as noted in Proposition 4.5, with Gaussian shocks the above formula merely provides a lower bound on mean escape times. Second, keep in mind these are all asymptotic results. It could well be that $\epsilon = .01$ is too large to provide an accurate approximation.[19]

The calculation of the rate function for the vertical Phillips Curve is especially simple. Since the sequence of coefficient estimates becomes Gaussian, the rate function is well known to be $\bar{S}^*(x) = .5(x-u^*)^2/(\sigma_1^2+\sigma_2^2)$. Note that in this linear setting, the rate function is symmetric, and escapes are equally likely to occur in either direction. To

[17]Warning: The distribution of escape times is not symmetric. It is exponential, with a long right tail. Hence, the median escape time is less than this.

[18]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 3 then reports the recursively estimated statistic, $\theta_n = \theta_{n-1} + \epsilon[(x_n'\xi_n')R_n^{-1}(x_n\xi_n)/\hat{\sigma}_n^2 - \theta_{n-1}]$.

[19]Kolyuzhnov, Bogomolova, and Slobodyan (2014) show that in dynamic Phillips Curve models the gain must be much smaller than $\epsilon = .01$ before large deviations approximations become accurate.

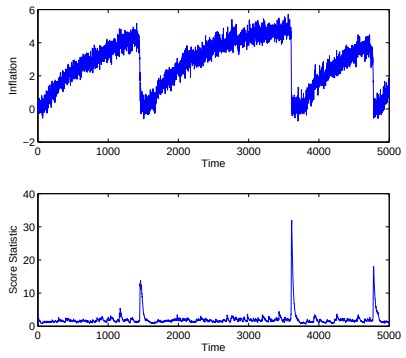

FIGURE 3. Sequential LM tests in Sargent's Model

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 3, 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.

The intuition for why the vertical Phillips Curve dominates follows from our previous results. Earlier we saw that models exhibiting strong self-referential feedback will be relatively fragile. Strong feedback makes it difficult to identify discrepancies between the Perceived and Actual laws of motion, and so coefficient estimates more easily escape from their SCE values. Feedback strength is determined by the slope (or Jacobian) of a model's T-map. When this is large, $\int[s - T(s)]ds$ is small, and so the rate function is also small. The static Phillips Curve is relatively fragile because its T-map is steeper (Evans and Honkapohja (2001, pgs. 325-28) provide a detailed derivation). In fact, the T-map of the vertical Phillips Curve is flat, since there is no feedback whatsoever from beliefs to the actual law of motion in this case.

6.2. **Classical vs. Keynesian Identification Restrictions.** The Phillips curve can be estimated two ways, depending on which variable is assigned as the dependent variable. This choice reflects an assumption about causation and identification. Sargent (1999) and CWS (2002) assume unemployment is the dependent variable,

$$u_n = \gamma_0 + \gamma_1 \pi_n \tag{6.32}$$

They call this the "Classical fit", since it regards inflation as an exogenous process that is under the control of the Central Bank. In practice, a more common specification assumes inflation is the dependent variable. This is called the "Keynesian fit", since it regards

inflation as being determined by aggregate demand.[20]

$$\pi_n = \beta_0 + \beta_1 u_n. \tag{6.33}$$

It is important to note that here both the Classical fit and the Keynesian fit are misspecified models of the true expectations augmented Phillips curve

$$u_n = u^* - \theta(\pi_n - x_n) + v_{1,n} \qquad \pi_n = x_n + v_{2,n} \tag{6.34}$$

since they both neglect the role of the public's expectations in the inflation process. However, there is a sense in which the Keynesian fit is 'more' misspecified, since its implicit identification restriction that unemployment is uncorrelated with the error term is invalid.

The key difference between the Classical and Keynesian models stems from alternative identification restrictions, which are based on alternative assumptions about the nature of the inflation process. Different assumptions about identification produce different beliefs about the 'sacrifice ratio', and hence, produce different self-confirming inflation rates. An important question is to what extent deliberate or 'natural' experiments can be used to learn about which identification restriction is valid. The usual claim that identification restrictions are untestable does *not* apply here, since the data are endogenous and the decisionmaker receives information about identification by witnessing the economy's response to policies that are based on alternative identification restrictions.

As before, suppose the Central Bank updates each model's coefficients using a recursive least square algorithm, and then selects the inflation target x_n by minimizing $Eu_n^2 + \pi_n^2$. The two models lead to different optimal policies, and more importantly, different self-confirming equilibrium inflation rates. Under the Classical fit, optimal policy is

$$x_{c,n} = -\frac{\gamma_{0,n}\gamma_{1,n}}{1+\gamma_{1,n}^2}. \tag{6.35}$$

and estimates converge to the self-confirming equilibrium

$$\bar{\gamma}_0 = u^*(1+\bar{\gamma}_1^2), \quad \bar{\gamma}_1 = -\theta \tag{6.36}$$

Optimal policy in the self-confirming equilibrium is $\bar{x}_c = \theta u^*$. Under the Keynesian fit in (6.33), optimal policy is

$$x_{k,n} = \frac{\beta_{0,n}}{1+\beta_{1,n}^2}. \tag{6.37}$$

and estimates converge to the self-confirming equilibrium

$$\bar{\beta}_0 = -\frac{u^*(1+\bar{\beta}_1^2)}{\bar{\beta}_1}, \quad \bar{\beta}_1 = -\frac{\theta\sigma_2^2}{(\sigma_1^2+\theta^2\sigma_2^2)} \tag{6.38}$$

Optimal policy in the self-confirming equilibrium is $\bar{x}_k = \theta u^* \left(1 + \frac{\sigma_1^2}{\theta^2\sigma_2^2}\right)$. Note that inflation is higher in the Keynesian self-confirming equilibrium. Comparing (6.36) and (6.38), it is clear the Keynesian fit produces a flatter Phillips curve, which then produces a larger estimate of the sacrifice ratio. Fears about the consequences of inflation stabilization cause the Keynesian Central Banker to tolerate a higher average inflation rate.

[20]See King and Watson (1994) for a detailed account of the debate between these two schools of thought, and the evidence each used to bolster its case. Romer and Romer's (2002) evidence points to a third possibility. They argue that by the early 1970s the Fed believed in the Natural Rate Hypothesis, but adopted a Keynesian identification scheme, which required an estimate of the natural rate. They argue the inflation of the 1970s was caused by an underestimate of the natural rate.

A natural question at this point is why a Central Bank would use the Keynesian model, since it's 'more' misspecified than the Classical model, and produces larger social costs. The validation process offers a clue.

6.2.1. *Calculation of H-functionals.* Both models here are linear and Gaussian, so in principle the H-functionals follow directly from the results in Section 5. However, in the neighborhood of their respective SCE, they are also *static*, since neither contains lags and both imply a constant inflation target at the SCE. The i.i.d./Gaussian nature of the disturbances then means we can compute the H-functionals with pencil and paper, via a standard complete-the-squares calculation. The following lemma summarizes the results.

Lemma 6.1. *Define the change of variables, $z' = \alpha' R^{-1}$, where α are the original coefficients. For the Classical model define $\delta_{c,0} = u^* + \theta x^c(\gamma) - \gamma_0$ and $\delta_{c,1} = (-\theta - \gamma_1)$, where $x_c(\gamma)$ is given by* (6.35)*. For the Keynesian model define $\delta_{k,0} = u^*\theta^{-1} + x_k(\beta) - \beta_0$ and $\delta_{k,1} = (-\theta^{-1} - \beta_1)$, where $x_k(\beta)$ is given by* (6.37)*. The Classical and Keynesian H-functionals are then given as follows:*

$$H_c(\gamma, z) = z_1\delta_{c,0} + \frac{1}{2}\sigma_1^2 z_{c,1}^2 + \frac{A_c x_c + \frac{1}{2}\sigma_2^2 A_c^2 + x_c^2 B_c}{1 - 2\sigma_2^2 B_c} - \frac{1}{2}\log(1 - 2\sigma_2^2 B_c) \tag{6.39}$$

$$H_k(\beta, z) = z_1\delta_{k,0} - \lambda z_1 u^* + \frac{1}{2}\sigma_1^2 z_1^2 + \frac{A_k u^* + \frac{1}{2}\sigma_u^2 A_k^2 + u^{*2} B_k}{1 - 2\sigma_u^2 B_k} - \frac{1}{2}\log(1 - 2\sigma_u^2 B_k)$$

where

$$A_c = (z_1\delta_{c,1} + z_2\delta_{c,0} + \sigma_1^2 z_1 z_2) \qquad B_c = (z_2\delta_{c,1} + \frac{1}{2}\sigma_1^2 z_2^2)$$

$$A_k = z_1\delta_{k,1} + z_2\delta_{k,0} + \lambda(z_1 - u^* z_2) + \sigma_\varepsilon^2 z_1 z_2 \qquad B_k = (z_2\delta_{k,1} + \frac{1}{2}\sigma_1^2 z_2^2)$$

and where $\sigma_\varepsilon^2 = \theta^{-2}\chi^2\sigma_1^2 + (1-\chi)^2\sigma_2^2$ and $\chi = \theta^2\sigma_2^2/(\theta^2\sigma_2^2 + \sigma_2^2)$.

Proof. See Appendix E □

6.2.2. *Calculation of Rate Functions.* The rate function solves a control problem with flow cost given by the Legendre transform of the H-functional. First consider the Classical model. Exploiting duality, we can write the HJB equation as, $H_c(\gamma, \alpha) = 0$. This is a first-order nonlinear PDE. An analogous PDE characterizes the Keynesian rate function. The solutions of these PDEs are described in the following proposition:

Proposition 6.2. *The Classical model's large deviation rate function is given by,*

$$S_c(\gamma_0, \gamma_1) = \frac{1}{\sigma_1^2}\left(u^* - \frac{\gamma_0}{1 + \gamma_1^2}\right)^2 + \frac{\sigma_2^2}{\sigma_1^2}(-\theta - \gamma_1)^2. \tag{6.40}$$

while the Keynesian model's large deviation rate function is given by,

$$S_k(\beta_0, \beta_1) = \frac{1}{\sigma_\varepsilon^2}\left(\frac{\beta_0\beta_1^2}{1 + \beta_1^2} + u^*\beta_1\right)^2 + \frac{\sigma_u^2}{\sigma_\varepsilon^2}\left(\lambda - \theta^{-1} - \beta_1\right)^2. \tag{6.41}$$

Proof. See Appendix F. □

Notice that both $S_c = 0$ and $S_k = 0$ at the SCE. However, due to correlation between the error term and the regressor in the Keynesian fit, the self-confirming value of β_1 is biased away from $-1/\theta$. Since $\lambda > 0$, the Keynesian Central Bank thinks $|\theta|$ is bigger than it really is, which leads it to set a higher inflation target.

6.2.3. *Interpretation.* Equations (6.40) and (6.41) are the bottom-line of the analysis. They offer the following clues about which model is likely to dominate:

(1) Assuming units are such that variances are small numbers, as in Sargent (1999), the escape dynamics in *both* models are driven by the first term in (6.40) and (6.41), since the second term is multiplied by a (small) variance. Hence, the magnitude of the rate function is dominated by small changes in the first term.
(2) Notice that in the Classical fit the first term remains nearly zero as the Ramsey outcome is approached (i.e., as $|\gamma_1|$ goes to zero and γ_0 goes to u^*). Although the decline in $|\gamma_1|$ causes the second term in S_c to increase, it is multiplied by σ_2^2, which tempers its influence.
(3) In the Keynesian model, notice that since $\beta_1 < 0$, the first term remains small when $|\beta_1|$ *increases*. It is much less sensitive to changes in β_0. Via the 'inverse Phelps problem', increases in $|\beta_1|$ cause the central bank to reduce inflation.
(4) For Sargent's (1999) calibration, the Keynesian rate function is larger than the Classical rate function. When $\theta = 1$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$ we have $\sigma_u^2 = 2\sigma^2$ and $\sigma_\varepsilon^2 = \sigma^2$. Hence, the denominator variances are the same in both, but the numerator variance multiplying the second term is twice as large in the Keynesian model. This implies escapes are much rarer in the Keynesian model. (This has been confirmed by our simulations, available upon request). Intuitively, the regressor in the Keynesian model has a higher variance, because the variance of u is double the variance of π. This causes the Keynesian slope coefficient to be relatively precisely estimated, which makes escape difficult. Interestingly, the relative stability of the Keynesian fit was emphasized by King and Watson (1994).

6.2.4. *Escape boundary.* Note that (6.40) induces an ellipse in the space of coefficients (γ_0, γ_1), with a center at the SCE (6.36). By changing coordinates we can convert the minimization problem over the boundary to an eigenvalue problem. Specifically, notice that if we set $y_1 = u^* + \theta x - \gamma_0$ and $y_1 = -\theta - \gamma_1$, then finding the escape point to a disk of radius ρ around the SCE becomes the eigenvalue problem: $\min y' R_c y$ subject to $y'y = \rho^2$. The escape direction is given by the eigenvector associated with the smallest eigenvalue. To determine which of the two endpoints is the minimum we can just substitute into S_c. Letting λ be the smallest eigenvalue of R_c, escape occurs along the nonlinear trajectory $(1-\lambda)(u^* - \gamma_0) - \lambda\theta x_c(\gamma) = x_c(\gamma)\gamma_1$. Escape is dictated by the *smallest* eigenvalue of R_c, since this is the *largest* eigenvalue of R_c^{-1}, and R_c^{-1} is proportional to the variance-covariance matrix of the regression coefficients. Hence, escapes occur in the direction along which coefficient estimates are most variable.

6.2.5. *Comparison.* In our first example, model validation led to the 'right' model, whereas in this second example it led to the 'wrong' model. Why the difference? With endogenous data, where models can adapt to their own data, a statistical specification test will *not* necessarily identify the true model. Nor will it necessarily identify the best model in terms of the policymaker's objectives. In the first example it did, but in the second it didn't.

Clearly, if the test were based on *outcomes*, the Classical identification scheme would prevail, despite its more frequent statistical rejection. Not only does it produce lower self-confirming inflation (with the same average unemployment rate), its more frequent escapes would actually enhance its relative fitness, since escapes to Ramsey lead to better outcomes. As discussed in Section 2, we base specification testing on statistical fit not because we think it is normatively appealing, but rather because we think it is often more descriptively accurate. Still, it would be interesting to adopt a more normative approach, and study the implications of a validation process that is based on outcomes rather than on fit. It could also be interesting to *combine* the previous two examples, by adding a vertical Phillips Curve to $\mathcal{M}$ in the second example. One might conjecture that it would again dominate, and model validation would again lead to the right model.

7. Related Literature

Our proposed model validation framework departs from the existing literature in two respects. First and foremost, it allows agents to consider more than one model. Second, the agents in our approach are somewhat more sophisticated than in conventional macroeconomic learning models, in the sense that they are assumed to be aware of their own influence over the DGP. Here we briefly review some prior work that has examined each of these issues separately.

From the beginning, researchers have worried about the passive nature of recursive least-squares learning. For example, 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. Perhaps more 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, this prior work all takes place within the confines of a single model.

More recently, a number of papers have begun to consider adaptive learning with multiple models. For example, in a repeated game context, Foster and Young (2003) allow players to construct, test, and revise simple models of their opponent's behavior. Hypothesis testing produces convergence to Nash equilibria in a relatively strong sense, although testing errors produce rare but recurrent experimentation phases. Our paper shares many of these same features, but focuses on equilibrium selection rather than convergence. Adam (2005) studies a Central Bank that selects between two inflation forecasting models. He shows that a misspecified model can be dominant. With feedback, use of a misspecified model can place correctly specified models at a forecasting disadvantage. In Adam (2007), he presents experimental evidence suggesting that subjects do indeed switch between forecasting models. We show that similar results can arise with real-time hypothesis testing and model validation. Our approach is also similar to Brock and Hommes (1997) and Branch and Evans (2007). However, their goal is quite different. They posit a large

collection of agents who randomly select between two models, with weights determined by recent forecasting performance. In contrast, we posit a single agent who continuously challenges the existing model, and where hypothesis testing generates model switches.

Finally, one criticism that could be made of our approach is that it lacks formal decision-theoretic foundations. Interestingly, there has been some recent work along these lines. Gilboa, Postlewaite, and Schmeidler (2008) argue that hypothesis testing and model selection are actually more consistent with recent developments in decision theory than are Bayesion methods. Ortoleva (2012) proposes a formal axiomatic justification of hypothesis testing based on a 'prior over priors'. Selecting a new prior in response to a low probability event is analogous to selecting a new model. However, his framework is essentially static, and therefore does not incorporate the feedback that is so central to our problem.

8. Concluding Remarks

Macroeconomic policymakers use multiple models. These models evolve over time, and there appear to be switches between them. The narrative evidence in Romer and Romer (2002) provides a fascinating description of this process. This paper has tried to model this process and evaluate its properties. 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 borrowed from 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. Although it is not *necessary* to abandon Rational Expectations and traditional Bayesian decision theory when confronting model uncertainty, we think there are good reasons to explore alternative approaches.[21]

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.

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. As noted earlier, it would be useful to study the implications of a validation process that is based on economic objectives, rather than on measures of statistical fit. 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 (2008) on robust filtering of discrete hidden states offers one route toward such an extension.

[21]See Sargent (1993), Hansen and Sargent (2008), 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 by imposing Assumption 3.2, 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, $\mathcal{A}f(x) = \dot{x}'\nabla f(x)$. 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*

$$E\left(\prod_{j=1}^{i} h_j(x(t_j))\left(f(x(t_{i+1})) - f(x(t_i)) - \int_{t_i}^{t_{i+1}} \mathcal{A}f(x(s))ds\right)\right) = 0 \tag{A.42}$$

for each $0 \leq t_1 < t_2 < \cdots < t_{i+1}$, $f(\cdot)$ in the domain of $\mathcal{A}$, and $h_1, \cdots h_i \in \mathcal{C}_b$, 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.42) 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 $\beta_i^\epsilon(t)$ be its piecewise-constant continuous time interpolation. Similarly, let $\beta_i^{*\epsilon}(t)$ be the continuous time interpolation of $\beta_{i,n}^*$. If we then define $\tilde{\beta}_{i,n} = \beta_{i,n} - \beta_{i,n}^*$ and $\tilde{\beta}_i^\epsilon(t) = \beta_i^\epsilon(t) - \beta_i^{*\epsilon}(t)$ as the deviations of the estimates from the current SCE, we want to show

$$\tilde{\beta}_i^\epsilon(t) \Rightarrow \beta_i(t) \qquad as \quad \epsilon \to 0$$

where $\beta_i(t)$ solves to the mean ODE, $\dot{\beta}_i = g_i(\beta_i(t))$. Let $\beta_i^o(t)$ be the weak sense limit of $\beta_i^\epsilon(t)$. Based on the above results, we must therefore show that

$$f(\tilde{\beta}_i^o(t+s)) - f(\tilde{\beta}_i^o(t)) - \int_t^{t+s} g_i(\tilde{\beta}_i^o(u)) \cdot \nabla f(\tilde{\beta}_i^o(u)) du$$

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

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} h_j(\tilde{\beta}^\epsilon(t_j)) \left(f(\tilde{\beta}^\epsilon(t+s)) - f(\tilde{\beta}^\epsilon(t)) - \int_t^{t+s} g(\tilde{\beta}^\epsilon(u)) \cdot \nabla f(\tilde{\beta}^\epsilon(u)) du \right) \right) = 0 \tag{A.43}$$

where $f \in \mathcal{C}_b^2$ and $h_j \in \mathcal{C}_b$, and $0 < t_j \leq t$. First, by virtue of the properties of h_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 (see Kushner and Yin (1997, chpt. 7), we have

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

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,

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

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

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

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,

$$\epsilon \sum_{k=t/\epsilon}^{(t+s)/\epsilon - 1} [\nabla f(\tilde{\beta}_k) \sum_{r=k}^{k+n_\epsilon - 1} E_k(G(\tilde{\beta}_r, X_r, s_k))] = \delta_\epsilon \sum_{k=t/\epsilon}^{(t+s)/\epsilon - 1} \nabla f(\tilde{\beta}_k) \left[\frac{1}{n_\epsilon} \sum_{r=k}^{k+n_\epsilon - 1} E_k(G(\tilde{\beta}_r, X_r, \bar{s})) \right]$$

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

$$\delta_\epsilon \sum_{k=t/\epsilon}^{(t+s)/\epsilon - 1} \nabla f(\tilde{\beta}_k) \left[\frac{1}{n_\epsilon} \sum_{r=k}^{k+n_\epsilon - 1} E_k(G(\tilde{\beta}_r, X_r, \bar{s}) \right] \to \int_t^{t+s} \nabla f(\tilde{\beta}^o(u)) g(\tilde{\beta}^o(u)) du \tag{A.44}$$

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

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} h_j(\tilde{\beta}^\epsilon(t_j)) \left(\int_t^{t+s} g(\tilde{\beta}^\epsilon(u)) \nabla f(\tilde{\beta}^\epsilon(u)) du \right) \right) = E\left(\prod_{j=1}^{i} h_j(\tilde{\beta}^o(t_j)) \left(\int_t^{t+s} g(\tilde{\beta}^o(u)) \nabla f(\tilde{\beta}^o(u)) du \right) \right) \tag{A.45}$$

Combining (A.44) with (A.45) establishes the equality in (A.43) 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,

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

where derivatives are evaluated along the path of the mean ODE. Second, when performing the averaging as in (A.44) one must invoke the invariance principle

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

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 Proposition 4.9

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

$$q_{s(u)j} = \pi_j \cdot e^{-V^*(s(u))/\epsilon} \qquad \text{and} \qquad q_{s(u)s(u)} = -\left(\sum_{j \neq s(u)} \pi_j\right) \cdot e^{-V^*(s(u))/\epsilon} \tag{C.46}$$

From Theorem A.2, we must show

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^i h_j(\varphi^\epsilon(t_j), s^\epsilon(t_j))\left(f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau+s)) - f(\varphi^\epsilon(\tau), s^\epsilon(\tau)) - \int_\tau^{\tau+s}\sum_{j=1}^m q_{s(u)j} f(\varphi^*_s, j)du\right)\right) = 0 \tag{C.47}$$

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

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

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

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

Consider the second term,

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

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),

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

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 C.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 C.1,

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

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

To see how this delivers assumption C.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 C.2. *Let $\bar{t}$ denote the mean transit time between all SCE, and let $\bar{t}_m$ denote its maximum. Then given assumption C.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, $g(\varphi^*) = 0$. 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,

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} h_j(\varphi^\epsilon(t_j), s^\epsilon(t_j)) \left(f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau+s)) - f(\varphi^\epsilon(\tau+s), s^\epsilon(\tau))\right)\right) \tag{C.48}$$

Our goal is to show that this converges to

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

with q given by C.46. 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 C.48 as

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} h_j(\varphi^\epsilon(t_j), s^\epsilon(t_j)) \left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon} [f(\varphi^\epsilon_{k+n_\epsilon}, s^\epsilon_{k+n_\epsilon}) - f(\varphi^\epsilon_{k+n_\epsilon}, s^\epsilon_k)]\right)\right) \tag{C.49}$$

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 C.49 with

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} h_j(\varphi^\epsilon(t_j), s^\epsilon(t_j)) \left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon} [f(\varphi^\epsilon_k, s^\epsilon_{k+n_\epsilon}) - f(\varphi^\epsilon_k, s^\epsilon_k)]\right)\right) \tag{C.50}$$

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 C.50 with

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} h_j(\varphi^\epsilon(t_j), s^\epsilon(t_j)) \left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon} \sum_{n=k}^{k+n_\epsilon-1} E_n[f(\varphi^\epsilon_k, s^\epsilon_{n+1}) - f(\varphi^\epsilon_k, s^\epsilon_n)]\right)\right) \tag{C.51}$$

Next, we can use the Markov transition probabilities to replace the inner expectation in C.51 to get

$$\lim_{\epsilon \to 0} E\left(\prod_{j=1}^{i} h_j(\varphi^\epsilon(t_j), s^\epsilon(t_j)) \left(\sum_{k=t(\tau)/\epsilon}^{t(\tau+s)/\epsilon} \sum_{j_0=1}^{m} \sum_{j=1}^{m} \sum_{n=k}^{k+n_\epsilon-1} [f(\varphi^\epsilon_k, j) P_n(s^\epsilon_{n+1} = j | s^\epsilon_n = j_0) - f(\varphi^\epsilon_k, j_0)] \mathcal{I}_{\{s_n = j_0\}}\right)\right) \tag{C.52}$$

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

$$P_n = I + Q^\epsilon_n$$

where Q^ϵ_n is the generator of a continuous-time Markov chain,

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

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

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

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[22]

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

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

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

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 C.52 as

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

with $q_{s(u)j}$ given by C.46, and where the bottom line follows from ergodicity and Proposition 4.5. The bottom line establishes the equality in C.47 and the proof is complete. □

Appendix D. Proof of Proposition 5.1

To start, write the update equations as follows:

$$\begin{aligned}
\beta_{n+1} &= \beta_n + \epsilon G(\beta_n, X_n) \\
&= \beta_n + \epsilon \bar{g}(\beta_n) + \epsilon [G(\beta_n, X_n) - \bar{g}(\beta_n)]
\end{aligned}$$

where $\bar{g}()$ picks up the mean dynamics by averaging over the ergodic distribution of X_n for fixed β. For this example it is just $\bar{g}(\beta) = (T_{11}(\beta) - \beta)' + \bar{R}^{-1} \Omega_{yz} T_{12}(\beta)'$, where $\bar{R}$ is the steady state second moment matrix of X_n, and Ω_{xz} is the steady state cross second moment matrix between X_n and Z_n. The martingale difference component, $G - \bar{g}$, can then be written,

$$G - \bar{g} = (R_n^{-1} X_n (X_n)' - I)(T_{11}(\beta) - \beta)' + (R_n^{-1} X_n (Z_n)' - \bar{R}^{-1} \Omega_{xz}) T_{12}(\beta)' + R_n^{-1} X_n v'_{n+1}$$

To derive the H-functional, define the the $s \times s$ matrix of co-states, α, where the columns are co-states pertaining to the coefficients in each equation. Using the fact that $vec(ABC) = (C' \otimes A) \cdot vec(B)$, we can then write the H-functional as

$$\begin{aligned}
H(\beta_n, \alpha) = \log E_n \{ \exp[& vec(\alpha)' [(T_{11}(\beta) - \beta) \otimes R^{-1} \sum_{j=0}^{\infty} vec(X_{n+j} X'_{n+j}) + T_{12}(\beta) \otimes R^{-1} \sum_{j=0}^{\infty} vec(X_{n+j} Z'_{n+j}) \\
& - vec(T_{11}(\beta) - \beta)' - vec(\bar{R}^{-1} \Omega_{yz} T_{12}(\beta)') + (I_s \otimes R^{-1}) \sum_{j=0}^{\infty} vec(X_{n+j} v'_{n+j})]] \}
\end{aligned}$$

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

Since X, Z, and v are jointly gaussian, we can use the results in Bryc and Dembo (1997) to evaluate this. To do so, we must write the summation as a quadratic form. The following result is useful when doing this (Magnus & Neudecker (2007, p. 35))

$$\text{tr}(ABCD) = (vecD')'(C' \otimes A)vec(B)$$

This implies the equality

$$vec(\alpha)'[(T(\beta) - \beta) \otimes R^{-1}]vec(XX') = \text{tr}(R^{-1}XX'(T(\beta) - \beta)'\alpha') = X'[(T(\beta) - \beta)'\alpha'R^{-1}]X$$

where the second equality uses the fact that $\text{tr}(ab') = b'a$ when a and b are $s \times 1$. Using this result and sequentially conditioning on X_{n+j} allows us to write the last term in the H-functional as the following quadratic form:

$$E_n\{\exp[vec(\alpha)'(I_s \otimes R^{-1})\sum_{j=0}^{\infty} vec(X_{n+j}v'_{n+j})]\} = \exp\left[\frac{1}{2}\sum_{j=0}^{\infty} X'_{n+j}R^{-1}\alpha\Sigma\alpha'R^{-1}X_{n+j}\right]$$

since $vec(\alpha)'[I_s \otimes R^{-1}]vec(Xv') = \text{tr}(R^{-1}Xv'I\alpha') = v'I\alpha'R^{-1}X = X'R^{-1}\alpha v$ and $Ee^{\mu'v} = e^{.5\mu'\Sigma\mu}$.

Finally, letting $(\Phi_n)' = ((Y_n)', (Z_n)')$, we can appeal to Theorem 2.2 in Bryc and Dembo (1997) to get

$$\begin{aligned}
& H(\beta, \alpha) \\
= \; & -vec(\alpha)' \cdot [vec(T(\beta) - \beta) + vec(\bar{R}^{-1}\Omega_{yz}T_{12}(\beta)')] + \log E_n \exp\left\{\sum_{j=0}^{\infty} \Phi'_{n+j}W(\alpha, \beta)\Phi_{n+j}\right\} \\
= \; & -vec(\alpha)' \cdot [vec(T(\beta) - \beta) + vec(\bar{R}^{-1}\Omega_{yz}T_{12}(\beta)')] - \frac{1}{4\pi}\int_0^{2\pi} \text{logdet}\left\{I_{s+q} - 2W(\alpha, \beta)F(\omega)\right\} d\omega \\
= \; & -vec(\alpha)'vec(\bar{g}) - \frac{1}{4\pi i}\oint \text{logdet}\left\{I_{s+q} - 2W(\alpha, \beta)F(z)F(z^{-1})'\right\}\frac{dz}{z}
\end{aligned}$$

where $F(\omega)$ is the spectral density matrix of the joint Φ_n process, and the weighting matrix, $W(\alpha, \beta)$, is given by eq. (5.27) in the text. The intuition behind the appearance of the spectral density here comes from the 'Toeplitz distribution theorem', which implies that as $n \to \infty$ the matrix of eigenvectors in the diagonalization of the variance-covariance matrix converges to the discrete Fourier transform matrix. Note, for the integral to be well defined (i.e., for the rate function to exist) we must in general impose parameter restrictions. These restrictions are easily derived. (See below).

To get the rate function, we must now compute the Legendre transform of $H(\beta, \alpha)$.

$$\begin{aligned}
L(\beta, \lambda) &= \sup_{\alpha}\{vec(\alpha)'[\lambda - vec(\bar{g}(\beta))] - H(\beta, \alpha)\} \\
&= \sup_{\alpha}\{vec(\alpha)'\lambda - \tilde{H}\}
\end{aligned}$$

where

$$\tilde{H} = vec(\alpha)'vec(\bar{g}) + H = -\frac{1}{4\pi i}\oint \text{logdet}\left\{I_{s+q} - 2W(\alpha, \beta)F(z)F(z^{-1})'\right\}\frac{dz}{z}$$

The rate function is then the solution of the following calculus of variations problem, $S(\beta) = \inf_{\dot{\beta}} \int L(\beta, \dot{\beta})$ The HJB equation for this problem can be written

$$\begin{aligned}
0 &= \inf_{\dot{\beta}}\{L + vec(S_\beta)' \cdot vec(\dot{\beta})\} \\
&= \tilde{H}(\beta, -S_\beta) \\
&= -\frac{1}{4\pi i}\oint \text{logdet}\left\{I_{s+q} - 2W(-S_\beta, \beta)F(z)F(z^{-1})'\right\}\frac{dz}{z}
\end{aligned}$$

where S_β is an $s \times s$ matrix whose columns are the gradients of each equation's coefficients. To evaluate this integral, we need to perform the following (canonical) spectral factorization

$$I_{s+q} - 2WF(z)F(z^{-1})' = \mathcal{F}(z)\mathcal{F}_0\mathcal{F}(z^{-1})' \tag{D.53}$$

where $\mathcal{F}_0 = \mathcal{F}(0)$, which is a function of both S_β and β. The evaluation of the above log integral follows immediately from the well known formula for the innovation variance of a stochastic process, and we obtain

the following compact expression for the rate function:

$$\text{logdet}\mathcal{F}_0(-S_\beta,\beta)=0 \qquad \Rightarrow \qquad \det\mathcal{F}_0(-S_\beta,\beta)=1$$

As before, define $(\Phi_n)'=((Y_n)',(Z_n)')$ as the combined vector of included and excluded variables, and assume $\Phi_n=T(\beta)\Phi_{n-1}+v_n$, with $\text{var}(v_n)=\Sigma$. In this case, the spectral density of Φ_n is $F(z)F(z^{-1})'=(I-Tz)^{-1}\Sigma(I-z^{-1}T')^{-1}$, and solving the spectral factorization problem in (D.53) can be converted into the problem of solving a Riccati equation. In particular, we can write[23]

$$2W[(2W)^{-1}-(I-Tz)^{-1}\Sigma(I-Tz^{-1})^{-1}]=2W[I-z(I-Az)^{-1}K][(2W)^{-1}-P][I-K'z^{-1}(I-z^{-1}A')^{-1}]$$

where $K=A'P[(2W)^{-1}-P]^{-1}$, and where P is the solution to the following Riccati equation,

$$P=\Sigma+T'PT+T'P[(2W)^{-1}-P]^{-1}PT \tag{D.54}$$

From this it is clear that $F(0)=2W[(2W)^{-1}-P]=I-2WP$, and so the PDE determining the rate function is $|I-2WP|=1$, where both W and P are functions of S_β. Existence conditions take the form of parameter restrictions that guarantee this Riccati equation has a unique positive definite solution.

Appendix E. Proof of Lemma 6.1

First consider the Classical fit. Its H-functional is,

$$H_c(\gamma,\alpha)=\log Ee^{\alpha' R^{-1}(1,\pi)'[u_n-\gamma_0-\gamma_1\pi]}=\log Ee^{z_1\delta_{c,0}+(z_1\delta_{c,1}+z_2\delta_{c,0})\pi+(z_1+\pi z_2)v_1+z_2\delta_{c,1}\pi^2}=H_c(\gamma,z)$$

where the second equality follows by substituting in the true law for u_n, and then simplifying using the changes of variables described in the statement of Lemma 6.1. The expectation here is with respect to the ergodic joint distribution of (π,v_1), *conditional* on the coefficient estimates. This makes the expectation especially easy to evaluate, since this distribution is i.i.d Gaussian, with zero cross correlation. We can evaluate the expectation by first conditioning on π. Using the formula for a log-normal, we get

$$\begin{aligned}H_c(\gamma,z) &= z_1\delta_{c,0}+\frac{1}{2}\sigma_1^2z_1^2+\log Ee^{(z_1\delta_{c,1}+z_2\delta_{c,0}+\sigma_1^2z_1z_2)\pi+(z_2\delta_{c,1}+\frac{1}{2}\sigma_1^2z_2^2)\pi^2}\\ &= z_1\delta_{c,0}+\frac{1}{2}\sigma_1^2z_1^2+\log Ee^{A_c\pi+B_c\pi^2}\end{aligned}$$

We must now take expectations with respect to the distribution of π. Note that (conditional on γ) π is i.i.d normal with mean $\bar{\pi}=x(\gamma)$ and variance $\sigma_\pi^2=\sigma_2^2$. Evaluating the expectation using the usual 'complete-the-square' trick gives equation (6.39).

For the Keynesian fit, we can follow exactly the same steps. There is one important difference, however. Now there is correlation between the error term and the regressor, so we must be a little careful when computing the expectation. Begin with the H-functional for the Keynesian fit,

$$H_k(\beta,z)=\log Ee^{z_1\delta_{k,0}+(z_1\delta_{k,1}+z_2\delta_{k,0})u+(z_1+uz_2)v_1/\theta+z_2\delta_{k,1}u^2}$$

What is new here is the third term in the exponent. In the Keynesian fit, u and v_1 are correlated, because $u=u^*-\theta v_2+v_1$. Thus, we cannot condition first on v_1, treating u as a constant. Instead we need to compute the mean and variance of v_1 *conditional* on u. Things still work out nicely since the conditional distribution of v_1 is normal. We just need to compute the following regression,

$$E(v_1|z_1+uz_2)=\hat{\phi}_0+\hat{\phi}_1(z_1+uz_2)=\hat{\phi}_1(u-u^*)z_2=\lambda(u-u^*)$$

where $\lambda=\theta^{-1}\sigma_1^2/(\theta^2\sigma_1^2+\sigma_1^2)$. The second equality follows from the fact that the regression intercept is $\hat{\phi}_0=-\hat{\phi}_1(z_1+u^*z_2)$, and the third equality just simplifies notation, using $\hat{\phi}_1=\theta^{-1}\sigma_1^2/[z_2(\theta^2\sigma_2^2+\sigma_1^2)]$. The conditional variance we need is just the variance of the regression residual. This residual is

$$\varepsilon=\theta^{-1}\chi v_1+(1-\chi)v_2$$

[23]The form of this Riccati equation is closely related to those arising in the robust control literature. See Hansen and Sargent (2007b, pgs. 182-195) for details on this derivation.

Its variance is $\sigma_\varepsilon^2 = \theta^{-2}\chi^2\sigma_1^2 + (1-\chi)^2\sigma_2^2$, where $\chi \equiv \theta^2\sigma_2^2/(\theta^2\sigma_2^2+\sigma_2^2)$. Now apply the same sequential conditioning strategy as before. First condition on u, and take expectations with respect to v_1. This gives us,

$$\begin{aligned} H_k(\beta, z) &= z_1\delta_{k,0} - \lambda z_1 u^* + \frac{1}{2}\sigma_\varepsilon^2 z_1^2 + \log E e^{(z_1\delta_{k,1}+z_2\delta_{k,0}+\lambda(z_1-u^*z_2)+\sigma_\varepsilon^2 z_1 z_2)u + (z_2\delta_1 + \frac{1}{2}\sigma_\varepsilon^2 z_2^2)u^2} \\ &\equiv z_1\delta_{k,0} - \lambda z_1 u^* + \frac{1}{2}\sigma_\varepsilon^2 z_1^2 + \log E e^{A_k u + B_k u^2}. \end{aligned}$$

Again using a 'complete-the-squares' trick, we can evaluate this to get the expression for H_k in Lemma 6.1. □

Appendix F. Proof of Proposition 6.2

We can first solve $H_c(\gamma, z) = 0$, and then find α by unwinding the transformation $z' = \alpha' R^{-1}$. Given the form of the H-functional, it is natural to seek a solution where $B_c = 0$. This implies

$$z_2 = -\frac{2}{\sigma_1^2}\delta_{c,1}$$

(Note, we must avoid the trivial solution $z_1 = z_2 = 0$). Substituting this back into H_c we find the following solution for z_1,

$$z_1 = -\frac{2}{\sigma_1^2}\delta_{c,0}$$

Reversing the transformation gives us $\alpha = R_c z$, where

$$R_c = \begin{bmatrix} 1 & x_c \\ x_c & x_c^2 + \sigma_2^2 \end{bmatrix}$$

Therefore, we have

$$\alpha_1 = -\frac{2}{\sigma_1^2}\left[\delta_{c,0} + x_c\delta_{c,1}\right] \qquad \alpha_2 = -\frac{2}{\sigma_1^2}\left[x\delta_{c,0} + (x_c^2+\sigma_2^2)\delta_{c,1}\right].$$

Remember, (α_1, α_2) are the derivatives of the rate function with respect to γ_0 and γ_1, so we must integrate these back to get the rate function. When integrating these we must substitute for $\delta_{c,0}$ and $\delta_{c,1}$ in terms of γ_0 and γ_1, but we do *not* want to substitute in for x_c in terms of the γ's, since this term arises from the second moment matrix, R_c, which is *reacting* to the escape, but not *driving* the escape. Substituting in for the δ's gives us

$$\alpha_1 = -\frac{2}{\sigma_1^2}\left[u^* - \gamma_0 - x\gamma_1\right] \qquad \alpha_2 = -\frac{2}{\sigma_1^2}\left[x(u^* - \gamma_0 - x\gamma_1) + \sigma_2^2(-\theta - \gamma_1)\right].$$

This system is easily integrated, and yields

$$S_c(\gamma_0, \gamma_1) = \frac{1}{\sigma_1^2}\left(u^* - \gamma_0 - x_c\gamma_1\right)^2 + \frac{\sigma_2^2}{\sigma_1^2}\left(-\theta - \gamma_1\right)^2.$$

Finally, substituting in $x_c = -\gamma_0\gamma_1/(1+\gamma_1^2)$ gives us the classical rate function in equation (6.40).

We follow exactly the same steps to calculate the Keynesian rate function. Solving $H_k(z, \beta) = 0$ yields

$$z_1 = -\frac{2}{\sigma_\varepsilon^2}\left[\delta_{k,0} - \lambda u^*\right] \qquad z_2 = -\frac{2}{\sigma_\varepsilon^2}\left[\delta_{k,1} + \lambda\right]$$

Next, we can use the fact that

$$R_k = \begin{bmatrix} 1 & u^* \\ u^* & u^{*2} + \sigma_u^2 \end{bmatrix}$$

to unwind the transformation, and solve for the α's. This gives us,

$$\alpha_1 = -\frac{2}{\sigma_\varepsilon^2}\left[\delta_{k,0} + u^*\delta_{k,1}\right] \qquad \alpha_2 = -\frac{2}{\sigma_\varepsilon^2}\left[u^*(\delta_{k,0} + u^*\delta_{k,1}) + \sigma_u^2(\delta_{k,1} + \lambda)\right].$$

Substituting in for the δ's and integrating back gives,

$$S_k(\beta_0, \beta_1) = \frac{1}{\sigma_\varepsilon^2}\left(x_k - \beta_0 - u^*\beta_1\right)^2 + \frac{\sigma_u^2}{\sigma_\varepsilon^2}\left(\lambda - \theta^{-1} - \beta_1\right)^2.$$

Finally, substituting in for $x_k = \beta_0/(1+\beta_1^2)$, we get the Keynesian rate function in equation (6.41).

Appendix G. Simulations

This appendix provides details of the calibrations used in the simulations contained in Figures 1 and 3. Matlab programs are available upon request. The calibrations for both Figures are the same, and for the most part, are identical to those in Sargent (1999). The parameter values are as follows:

Parameter Values Used in Figures 1 and 3

u^*	θ	σ_1	σ_2	ϵ	ϕ	$\bar{\tau}$
5.0	1.0	0.3	0.3	.015	2.0	10.0

The new elements here are in the final two columns. ϕ indexes the 'choice intensity' parameter used in the logit function for model selection (see page 12 in the main text). As ϕ increases, the agent is less prone to experiment. The results are reasonably robust for moderate values of ϕ. (As noted in the main text, none of the paper's asymptotic results depend on ϕ). $\bar{\tau}$ is the test threshold for the score statistic in the neighborhood of the self-confirming equilibrium. A value of 10 is motivated by the fact that the statistic has two degrees of freedom, and rejections are assumed to be triggered by large deviations. The simulations in both figures were initialized at the self-confirming equilibrium.

References

Adam, K. (2005): "Learning to Forecast and Cyclical Behavior of Output and Inflation," *Macroeconomic Dynamics*, 9, 1–27.

——— (2007): "Experimental Evidence on the Persistence of Output and Inflation," *Economic Journal*, 117, 603–36.

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

Branch, W. A., and G. W. Evans (2007): "Model Uncertainty and Endogenous Volatility," *Review of Economic Dynamics*, 10, 207–37.

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.

Brock, W. A., and C. Hommes (1997): "A Rational Route to Randomness," *Econometrica*, 65, 1059–1095.

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.

Bryc, W., and A. Dembo (1997): "Large Deviations for Quadratic Functionals of Gaussian Processes," *Journal of Theoretical Probability*, 10, 307–332.

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.

Cho, I.-K., and K. Kasa (2013): "Learning and Model Validation: An Example," in *Macroeconomics at the Service of Public Policy*, ed. by T. J. Sargent, and J. Vilmunen. Oxford University Press.

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, and T. J. Sargent (2007): "Benefits from U.S. Monetary Policy Experimentation in the Days of Samuelson and Solow and Lucas," *Journal of Money, Credit, and Banking*, 39, 67–99.

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

Diaconis, P., and D. Freedman (1986): "On the Consistency of Bayes Estimates," *Annals of Statistics*, 14, 1–26.

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.

Evans, G. W., S. Honkapohja, T. J. Sargent, and N. Williams (2013): "Bayesian Model Averaging, Learning, and Model Selection," in *Macroeconomics at the Service of Public Policy*, ed. by T. J. Sargent, and J. Vilmunen. Oxford 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.

Geweke, J. (2010): *Complete and Incomplete Econometric Models*. Princeton University Press.

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 (2008): *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.

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

King, R. G., and M. W. Watson (1994): "The Post-War U.S. Phillips Curve: A Revisionist Econometric History," *Carnegie-Rochester Conference Series on Public Policy*, 41, 157–219.

Kocherlakota, N. R. (2007): "Model Fit and Model Selection," *Federal Reserve Bank of St. Louis Review*, 89, 349–60.

Kolyuzhnov, D., A. Bogomolova, and S. Slobodyan (2014): "Escape Dynamics: A Continuous-Time Approximation," *Journal of Economic Dynamics and Control*, 38, 161–83.

Kostyshyna, O. (2012): "Application of an Adaptive Step-Size Algorithm in Models of Hyperinflation," *Macroeconomic Dynamics*, 16, 355–75.

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, 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.

Magnus, J. R., and H. Neudecker (2007): *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley and Sons, third edn.

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.

Nachbar, J. H. (1997): "Prediction, Optimization, and Learning in Repeated Games," *Econometrica*, 65, 275–309.

Ortoleva, P. (2012): "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News," *American Economic Review*, 102, 2410–36.

Romer, C. D., and D. H. Romer (2002): "The Evolution of Economic Understanding and Postwar Stabilization Policy," *Proceedings of the 2002 Jackson Hole Conference*, pp. 11–78.

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.

Savage, L. J. (1972): *The Foundations of Statistics*. Dover Publications, second revised edn.

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

Schorfheide, F. (2013): "Estimation and Evaluation of DSGE Models: Progress and Challenges," in *Advances in Economics and Econometrics, 10th World Congress, Volume III*, ed. by D. Acemoglu, M. Arrelano, and E. Deckel. Cambridge University Press.

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

——— (2002): "The Role of Models and Probabilities in the Monetary Policy Process," *Brookings Papers on Economic Activity*, 33(2), 1–62.

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.

In-Koo Cho
Department of Economics
University of Illinois
email: inkoocho@uiuc.edu

Kenneth Kasa
Department of Economics
Simon Fraser University
email: kkasa@sfu.ca