Learning and Model Validation

In-Koo Cho[1] Kenneth Kasa[2]

[1]Department of Economics
University of Illinois

[2]Department of Economics
Simon Fraser University

October 30, 2008

Evolution of the Macro Learning Literature

1. Convergence to Rational Expectations.
   - Singular Perturbation/Mean ODE Methods.

Evolution of the Macro Learning Literature

1. Convergence to Rational Expectations.
    - Singular Perturbation/Mean ODE Methods.
2. Stability. Equilibrium Selection.
    - E-stability.

Evolution of the Macro Learning Literature

1. Convergence to Rational Expectations.
    - Singular Perturbation/Mean ODE Methods.
2. Stability. Equilibrium Selection.
    - E-stability.
3. Constant Gain Learning.
    - Large Deviations.

Evolution of the Macro Learning Literature

1. Convergence to Rational Expectations.
    - Singular Perturbation/Mean ODE Methods.
2. Stability. Equilibrium Selection.
    - E-stability.
3. Constant Gain Learning.
    - Large Deviations.
4. Policy Design When Agents are Learning. Normative.
    - “Learnability"’.

Parameter Uncertainty vs. Model Uncertainty

This literature shares one crucial assumption.

Parameter Uncertainty vs. Model Uncertainty

This literature shares one crucial assumption.

Agents know the model.

Agents update beliefs about the parameters of a *given* model.

Parameter Uncertainty vs. Model Uncertainty

This literature shares one crucial assumption.

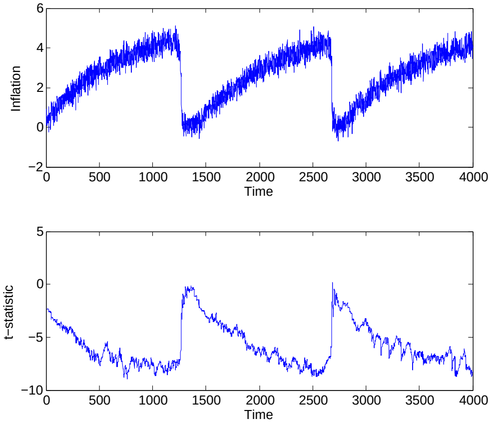
Agents know the model.

Agents update beliefs about the parameters of a *given* model.

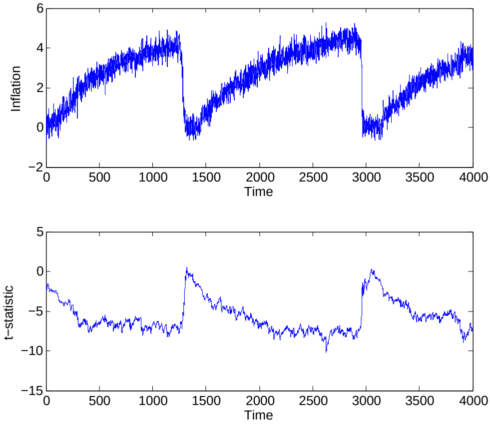
Our goal is to take the next natural step in the learning literature, by incorporating *model uncertainty*.

We do this by continuing the tradition of modeling agents as econometricians.

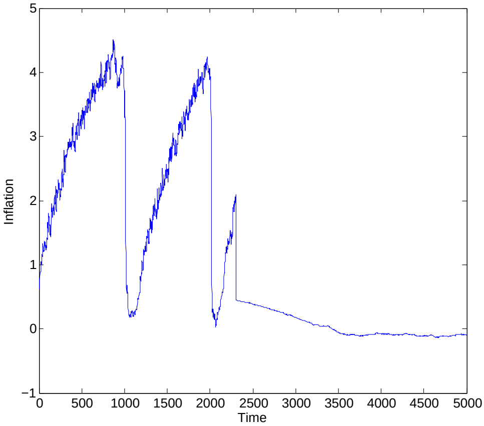
EXAMPLE 1: SARGENT'S *Conquest* MODEL



Sequential t-tests in Sargent's Model

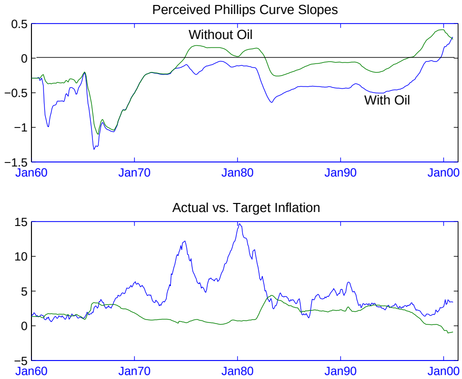


Expected vs. Unexpected Inflation



Example 2: Phillips Curve With Supply Shocks

Model Switch Dates: (1) Late 1966, (2) End of 1969, (3) Early 1975, (4) Mid 2000



A Puzzle

Why did the Fed wait so long to bring down inflation?

Potential Explanations

1. Bayesian model uncertainty (Cogley & Sargent (RED, 2005)). High inflation Keynesian policy risk-dominated the better-fitting low inflation natural rate policy. According to this account, the Fed stopped believing in an exploitable Phillips Curve in the mid 1970s.

A PUZZLE

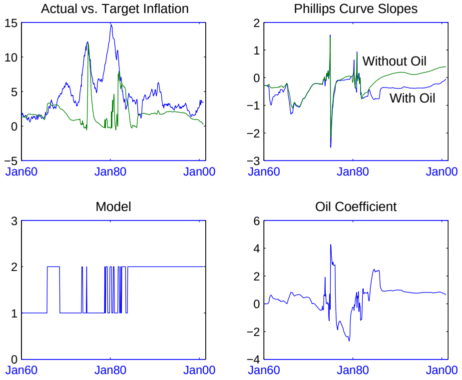
Why did the Fed wait so long to bring down inflation?

Potential Explanations

1. Bayesian model uncertainty (Cogley & Sargent (RED, 2005)). High inflation Keynesian policy risk-dominated the better-fitting low inflation natural rate policy. According to this account, the Fed stopped believing in an exploitable Phillips Curve in the mid 1970s.

2. Model Validation. Fed looked for (and found) additional variables that explained the apparent breakdown of the Phillips Curve. This allowed the Fed to sustain its belief in an exploitable Phillips Curve.

Supply Shocks + Variable Gain



Main Ingredients

1. Model Uncertainty: Agent must solve LQR, but doesn't know the state transition equation.

Main Ingredients

1. Model Uncertainty: Agent must solve LQR, but doesn't know the state transition equation.

2. Sequential, real-time decision-making: Usual validation procedures (e.g., hold-out samples) impractical and inefficient.

Main Ingredients

1. Model Uncertainty: Agent must solve LQR, but doesn't know the state transition equation.
2. Sequential, real-time decision-making: Usual validation procedures (e.g., hold-out samples) impractical and inefficient.
3. Model Class: Agent commits to a finite collection of linear, Gaussian models.

Main Ingredients

1. Model Uncertainty: Agent must solve LQR, but doesn't know the state transition equation.
2. Sequential, real-time decision-making: Usual validation procedures (e.g., hold-out samples) impractical and inefficient.
3. Model Class: Agent commits to a finite collection of linear, Gaussian models.
4. Model Selection: Agent commits to a single model each period.

Model Selection vs. Model Averaging

How many models does a Bayesian consider?

Model Selection vs. Model Averaging

How many models does a Bayesian consider?

One!

A Bayesian must construct a “model of the models”’.

Model Selection vs. Model Averaging

How many models does a Bayesian consider?

One!

A Bayesian must construct a “model of the models”'.

For a Bayesian, there is no difference between model uncertainty and parameter uncertainty.

Bayesians do *not* select models!

Problems

1. Lindley's (1957) Paradox/Prior Sensitivity
2. Infinite Dimensional Parameter Spaces (Diaconis & Freedman (1986))
3. Ambiguity/Breakdown of Savage Axioms (Hansen & Sargent (2007))
4. Multiagent Learning/Absolute Continuity (Nachbar (1997), Foster & Young (2001))

Main Ingredients (cont.)

5. Feedback: Actual DGP depends on agent's beliefs and actions.

Main Ingredients (cont.)

5. Feedback: Actual DGP depends on agent's beliefs and actions.
6. Learning: Agent updates models using discounted recursive least squares.

Counterfactual Dilemma

Feedback confronts the agent with a difficult counterfactual:

How would things have been different if instead a different model had been used in the past?

Counterfactual Dilemma

Feedback confronts the agent with 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 <u>different</u> model was in use could produce misleading inferences.

Assumption: Models are only updated while in use.

Main Ingredients (cont.)

5. Feedback: Actual DGP depends on agent's beliefs and actions.

6. Learning: Agent updates models using discounted recursive least squares.

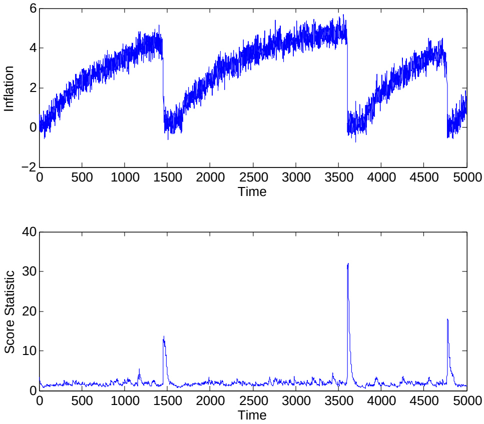
Main Ingredients (cont.)

5. Feedback: Actual DGP depends on agent's beliefs and actions.
6. Learning: Agent updates models using discounted recursive least squares.
7. Self-Confirming Equilibria: Each model has a unique, stable SCE.

MAIN INGREDIENTS (CONT.)

5. Feedback: Actual DGP depends on agent's beliefs and actions.
6. Learning: Agent updates models using discounted recursive least squares.
7. Self-Confirming Equilibria: Each model has a unique, stable SCE.
8. Model Validation: While in use, models are subject to sequential LM test. Agent sticks with a model until it is rejected.

Sequential LM tests in Sargent's Model



Specification Testing vs. Model Comparison

- **Specification Testing**. Stick with a model until it is statistically rejected by a general specification test (no explicit alternative). Cho & Kasa (2008), "*Learning and Model Validation*".
- **Model Comparision**. Continuously (or periodically) compare a model against alternative models. Cho & Kasa (2008), "*Calibrated Learning and Robust Model Validation*".

Choice of specification testing reflects implicit model switching costs.

Main Ingredients (cont.)

9. Randomization: If a model is rejected, another model is *randomly* selected.

Randomization vs. Model Averaging

- Although our agent commits to a model when formulating policy, he randomizes when selecting a model.
- This randomization does not reflect capriciousness. It is based on strategic considerations.
- With model uncertainty, random model selection is *robust*.

  It prevents 'nature' from exploiting your ignorance (e.g., it prevents you from getting stuck in a bad, yet self-confirming, equilibrium.
- Technically, randomization is important for delivering ergodicity. All models remain on the table.

Main Ingredients (cont.)

9. Randomization: If a model is rejected, another model is *randomly* selected.

Main Ingredients (cont.)

9. Randomization: If a model is rejected, another model is *randomly* selected.

10. Bounded Rationality
    - Agent neglects future model revisions when formulating policy.
    - Agent interprets escapes as model rejections.

SRA Representation

$$
\begin{aligned}
\hat{\beta}^i_n &= \hat{\beta}^i_{n-1} + \eta s^i_{n-1} (R^i_{n-1})^{-1} \Lambda^i_n \\
\Lambda^i_n &= \Phi^i_{n-1}(x_n - \Phi^{i\prime}_{n-1}\hat{\beta}^i_{n-1}) \\
R^i_n &= R^i_{n-1} + \eta s^i_{n-1}(\Phi^i_{n-1}\Phi^{i\prime}_{n-1} - R^i_{n-1})
\end{aligned}
$$

$$
y_n = A_1(s_{n-1}, \beta_{n-1}) y_{n-1} + B_1(s_{n-1}, \beta_{n-1}) \epsilon_{1n}
$$

$$
\begin{aligned}
\theta^i_n &= \theta^i_{n-1} + \eta s^i_{n-1}[\Lambda^{i\prime}_n \Sigma^{-1/2}_{in} (R^i_n)^{-1} \Sigma^{-1/2}_{in} \Lambda^i_n - \theta^i_{n-1}] \\
\omega^i_n &= \omega^i_{n-1} + \eta s^i_{n-1} \left(\nu^i_n(\beta^i_{n-1})' \nu^i_n(\beta^i_{n-1}) - \omega^i_{n-1}\right) \\
\pi^i_n &= \frac{e^{-\rho \omega^i_n}}{\sum_{j=1}^m e^{-\rho \omega^j_n}}
\end{aligned}
$$

The model indicator, s_{it}, follows an m-state Markov Chain

$$p'_{n+1} = p'_n \mathcal{P}_n$$

Diagonal Elements

$$\text{Prob}[\theta^i_n \leq \bar{\theta}(n))] + \text{Prob}[\theta^i_n > \bar{\theta}(n))] \cdot \pi^i_n$$

Off-Diagonals

$$\text{Prob}[\theta^i_n > \bar{\theta}(n)] \cdot \pi^j_n$$

TIME-SCALE SEPARATION

Traditional analysis of recursive learning exploits a time-scale separation between the evolution of the data (a 'fast' time-scale), and the evolution of a model's coefficient estimates (a 'slow' time-scale).

Specification testing adds a *third* time-scale, pertaining to the frequency of model switching.

Model rejections occur when parameters 'escape' their SCE values. Hence, model switching is *rare*.

Model Revision Time $\sim 1/\eta$
Model Switching Time $\sim \exp[S^*/\eta]$

Large Deviations I

For small η, dynamics of $\beta^i(t)$ is described by *two* ode's.

1. The mean dynamics, $\dot{\beta}^i = h^i(\beta^i)$. This describes agents' efforts to eliminate systematic forecast errors.

2. The escape dynamics. Rare but recurrent escapes away from the Self-Confirming Equilibrium. The escape dynamics solve the control problem

$$S^i(\beta_0, G) = \inf_{T>0} \inf_{\beta \in A} \int_0^T L^i(\dot{\beta}, \beta, t)dt$$

where

$$A = \{\beta \in C[0, T] : \beta(0) = \beta_0, \beta(T) \in \partial G\}$$

Large Deviations II

- The frequency of escapes is governed by the value function, $S^i(\beta_0, G)$. It is called the *rate function*.

$$\lim_{\eta \to 0} \eta \log E(\tau^\eta) = S^{*i}$$

- The function $L^i(\dot{\beta}, \beta, t)$ is called the *action functional*. It is the Legendre transform of the log moment generating function of the martingale difference component of the least squares orthogonality conditions for model-i.

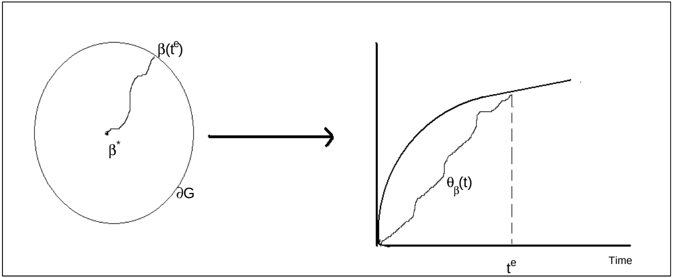
- For LQG problems L is quadratic. Williams (2001).

$$L^i(\dot{\beta}, \beta) = v' Q(\beta, R)^\dagger v \qquad v = \dot{\beta} - R^{-1}\bar{g}(\beta)$$

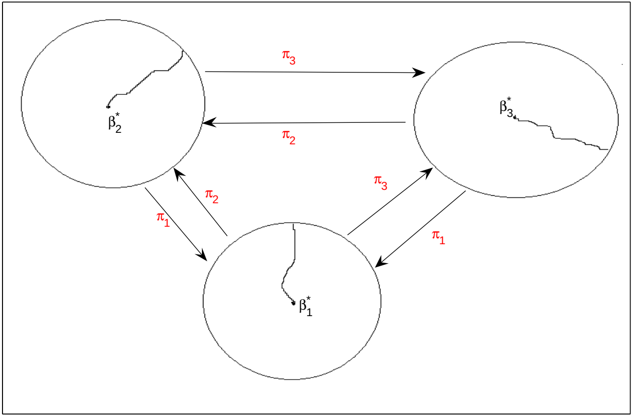
- The instantaneous escape direction is the eigenvector associated with the smallest eigenvalue of $Q^\dagger$.

THE CONTRACTION PRINCIPLE

We assume the specification test is calibrated to reject during escapes. This requires relating the LD properties of $\beta(t)$ to the LD properties of $\theta(t)$.



Model Switching Dynamics



Markov Chain Approximation

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

$$q_{ij} = \pi_j^* e^{(\bar{S}_{\max} - \bar{S}^i)/\eta} \qquad q_{ii} = -\left(\sum_{j \neq i}^{m} \pi_j^*\right) e^{(\bar{S}_{\max} - \bar{S}^i)/\eta}$$

which possesses a unique invariant distribution

$$\bar{p}_i^\eta = \frac{\pi_i^* e^{\bar{S}^i/\eta}}{\sum_{j=1}^m \pi_j^* e^{\bar{S}^i/\eta}}$$

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

Dominant Models

Proposition 4.12: *As $\eta \to 0$ the model distribution collapses onto the model with the largest LD rate function*.

The dominant model survives specification testing longer than any other model.

The dominant model may not be the best-fitting model.

Corollary 4.13: *As long as the experimentation probabilities, π_t^i, remain strictly bounded between 0 and 1, the identity of the dominant SCE is independent of the details of randomization*.

INFORMATION-THEORETIC INTERPRETATION

The dominant model is characterized by the *largest* LD rate function (defined at each model's SCE). Why?

Sanov's Theorem

$$\text{LD Rate Function} \Longleftrightarrow \text{Relative Entropy}$$

Stein's Lemma/Chernoff Bounds

$$\text{Relative Entropy} \uparrow \Longleftrightarrow \text{Detection Error Prob.} \downarrow$$

With endogenous data, models have the capacity to mimic the true DGP. Model rejection $=$ Type I error.

$$\begin{aligned} \text{LD rate function} \uparrow \quad &\Rightarrow \quad \text{Type I error rate} \downarrow \\ &\Rightarrow \quad \text{More durable model} \end{aligned}$$

Back to the Phillips Curve

Suppose the agent entertains *two* models:

- A static Phillips Curve, $u_t = \gamma_0 + \gamma_1 \pi_t + \varepsilon_t$
- A vertical Phillips Curve, $u_t = \gamma_0 + \varepsilon_t$.

LD Rate Function for static Phillips Curve $\approx .0005$
$\Rightarrow$ Mean escape time $\approx$ 100 discrete time periods.

LD Rate Function for vertical Phillips Curve $\approx \frac{.5(x-\bar{u})^2}{\sigma_1^2+\sigma_2^2}$
$\Rightarrow$ Mean escape time $\approx$ 300,000 discrete time periods!

Extensions

1. Alternative Specification Tests
2. Alternative Model Classes
3. Model Comparison vs. Specification Testing
4. Distrust of the Model Class. Robustness.