

Model selection and forecast comparison
in unstable environments

Raffaella Giacomini, UCLA

Barbara Rossi, Duke

Duke Forecasting Conference, March 9-10, 2007

Motivation

- Question: how to compare the performance of competing models in the presence of *misspecification* and *structural instability*?
- Main idea: structural instability \implies the relative performance of the models can change over time (supported by empirical evidence in Stock and Watson, 2003)
- Goal: propose formal techniques to test whether the relative performance of misspecified models is stable over time

Motivation

- Existing econometric tools inadequate:
 - Previous model selection and forecast comparison techniques allow for misspecification but not for instability
 - ⇒ they compare average performance ⇒ loss of information if relative performance varies over time
 - Previous analysis of structural instability focused on the parameters of one model, assuming correct specification
 - ⇒ parameters may vary but the models' relative performance be constant
 - ⇒ parameters may be constant, but the models' relative performance be time-varying

Contributions

- We propose two tests:
 - **Fluctuation test** to analyze the evolution of the models' relative performance over historical samples. Two measures of performance:
 - * In-sample: Kullback-Leibler Information Criterion (KLIC) \implies choose model that is closer to the true unknown data-generating process \iff model with largest expected log-likelihood
 - * Out-of-sample: choose model with lowest expected forecast loss (general loss)
 - **Sequential test** to monitor the models' relative performance in real time, as new data becomes available

Related literature

- In-sample fluctuation test:
 - Vuong (1989), Rivers and Vuong (2002) \implies test for equal full-sample average KLIC of misspecified models
 - Rossi (2005) \implies test for nested model selection under instability, but assumes correct specification
- Out-of-sample fluctuation test:
 - Diebold and Mariano (1995); West (1996); McCracken (2000) etc. \implies test for equal out-of-sample average forecast loss
 - Giacomini and White (2006) \implies test whether relative performance is different in different states of the economy (i.e. related to economic variables)

Related literature

- Sequential test:
 - Chu, Stinchcombe and White (1996) \implies real-time parameter instability in a correctly specified model
 - Inoue and Rossi (2005) \implies real-time nested model selection under instability but correct specification

Outline of the talk

- Motivating example - In-sample fluctuation test
- Theory
- Monte Carlo evidence
- Empirical application to DSGE vs. VAR
- Conclusion

Example - DGP and models

- True conditional density for y_t :

$$h_t : N(\theta_t x_t + \gamma_t z_t, \mathbf{1})$$

- $x_t \sim N(0, \text{var}(x_t))$, $z_t \sim N(0, \text{var}(z_t))$ independent

- Two competing misspecified models:

$$f_t : N(\theta_t x_t, \mathbf{1})$$

$$g_t : N(\gamma_t z_t, \mathbf{1})$$

Example - In-sample fluctuation test

- Goal: analyze relative in-sample performance over the historical sample
- Measure of relative performance = relative distance (measured by the KLIC) of f and g from h :

$$\begin{aligned}\Delta KLIC &= E [\log h_t/g_t] - E [\log h_t/f_t] \\ &= E [\log f_t - \log g_t]\end{aligned}$$

$$\text{for } t = 1, \dots, T$$

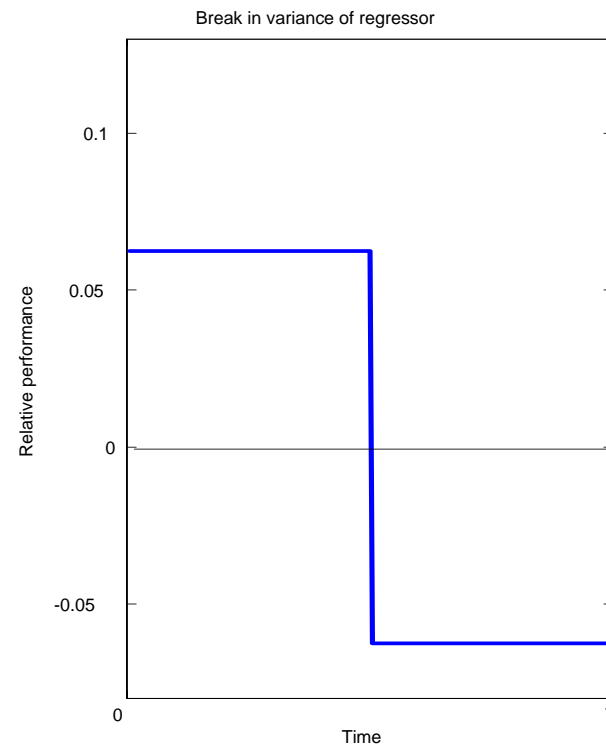
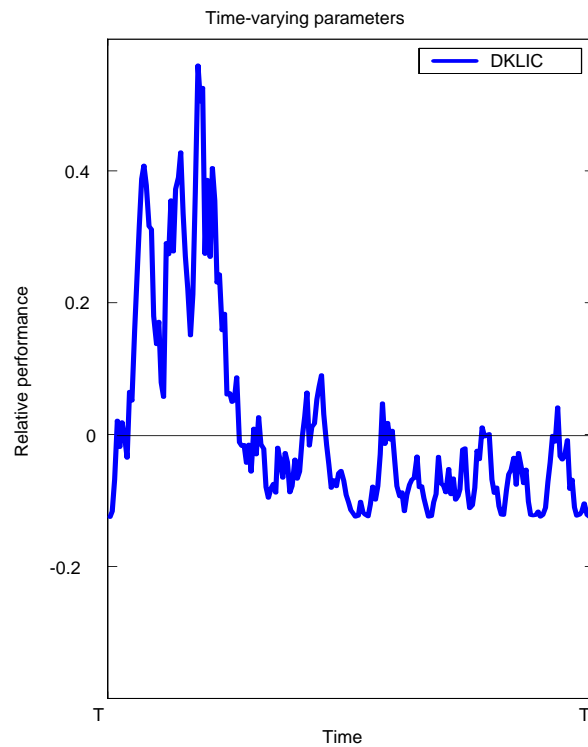
- If $\Delta KLIC > 0$, f performs better than g

Example - In-sample fluctuation test

- In the example, $\Delta KLIC = \frac{1}{2}(\theta_t^2 \text{var}(x_t) - \gamma_t^2 \text{var}(z_t))$
- Intuition: $\theta_t x_t =$ part of the error of g due to misspecification $\implies f$ better if the contribution of its misspecification term to the variance of error is smaller than the same for g
- Time variation in $\Delta KLIC \iff$ time variation in relative misspecification
 - $\Delta KLIC$ changes if θ_t, γ_t change in different ways
 - $\Delta KLIC$ changes if θ, γ constant but $\text{var}(x_t), \text{var}(z_t)$ change in different ways
 - $\Delta KLIC$ constant if $\theta_t^2 \text{var}(x_t)$ and $\gamma_t^2 \text{var}(z_t)$ change in the same way

Example - Two scenarios with time-varying $\Delta KLIC$

- θ_t varies as a random walk; γ , $var(x_t)$, $var(z_t)$ constant
- θ , γ , $var(z_t)$ constant; $var(x_t)$ changes at $T/2$



Example - The "smoothed" $\Delta KLIC$

- We would like to estimate $\Delta KLIC$ but it depends on unknown $\theta_t, \gamma_t \implies$ estimate a "smoothed" version of $\Delta KLIC$ computed over moving windows of size m

$$\text{Smoothed } \Delta KLIC : E \left[\frac{1}{m} \sum_{j=t-m/2+1}^{t+m/2} \left(\log f_j(\theta_{t,m}^*) - \log g_j(\gamma_{t,m}^*) \right) \right]$$

$$\text{for } t = \frac{m}{2} + 1, \dots, T - \frac{m}{2}$$

$\theta_{t,m}^*$ and $\gamma_{t,m}^*$ are pseudo-true parameters, e.g.,

$$\theta_{t,m}^* = \max_{\theta} E \left[m^{-1} \sum_j \log f_j(\theta) \right]$$

Example - The "smoothed" $\Delta KLIC$

- In the example $\theta_{t,m}^* = \sum_j \theta_j \text{var}(x_j) / \sum_j \text{var}(x_j)$
- Smoothed $\Delta KLIC$ is

Smoothed $\Delta KLIC$

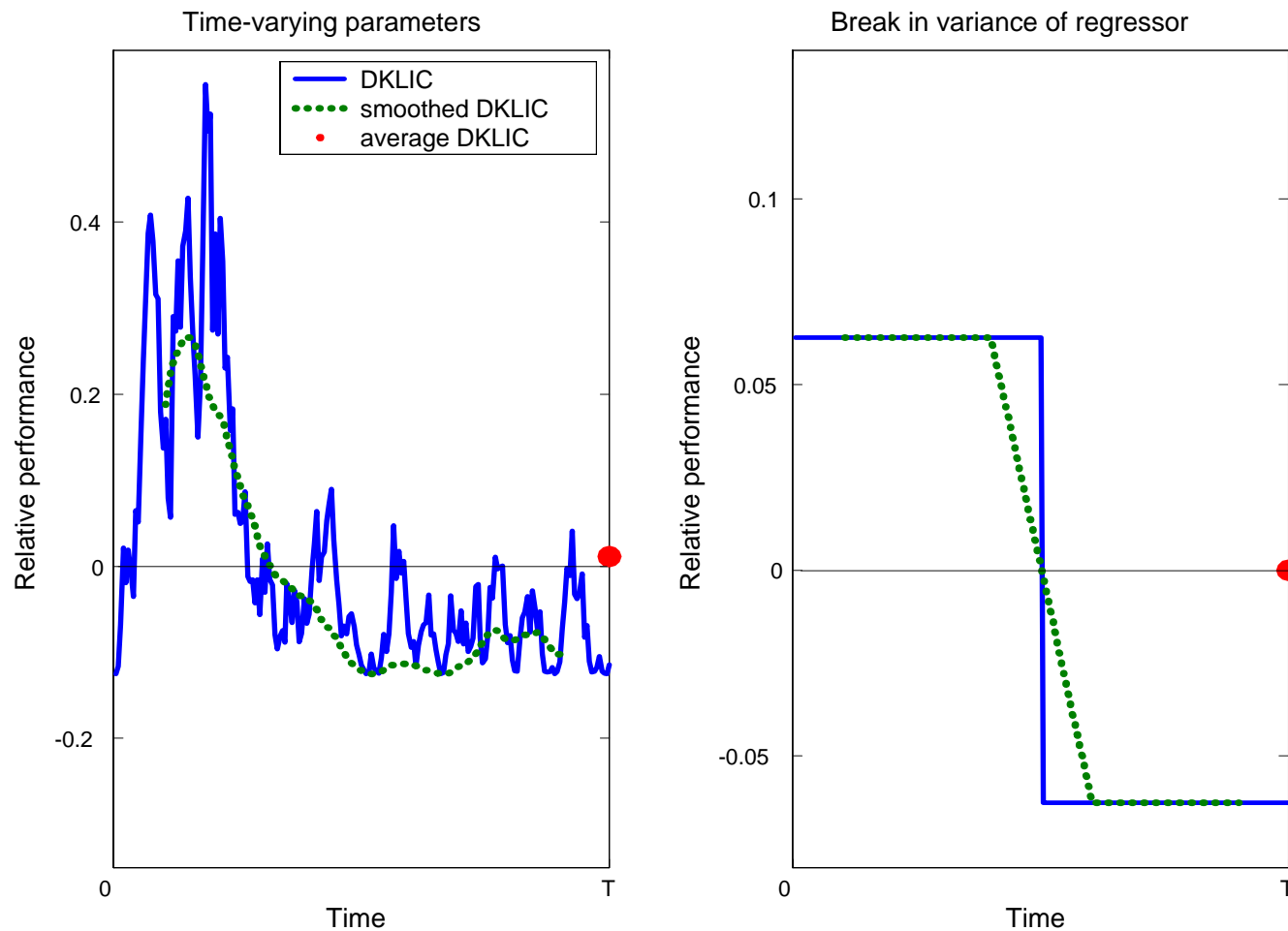
$$= \frac{1}{2} \left[\theta_{t,m}^{*2} \frac{1}{m} \sum_j \text{var}(x_j) - \gamma_{t,m}^{*2} \frac{1}{m} \sum_j \text{var}(z_j) \right]$$

$$\text{for } t = \frac{m}{2} + 1, \dots, T - \frac{m}{2}$$

- If variation in parameters and variance of regressors is small within the moving window, smoothed $\Delta KLIC \approx \Delta KLIC$

Example - Comparison with previous approaches

- In the example, for a moving window of size $m = T/5$



Example - Implementation of the fluctuation test

- Compute the sample analog of the smoothed $\Delta KLIC$
- Normalize it to obtain a sequence of fluctuation statistics

$$F_t^{IS} = \hat{\sigma}^{-1} m^{-1/2} \sum_j \left(\log f_j(\hat{\theta}_{t,m}) - \log g_j(\hat{\gamma}_{t,m}) \right)$$

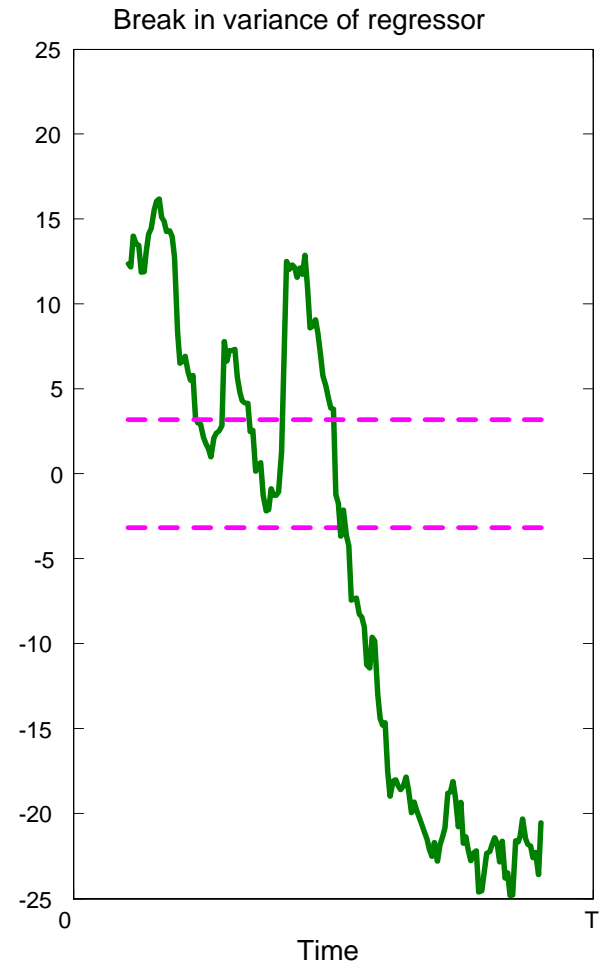
$$\text{for } t = \frac{m}{2} + 1, \dots, T - \frac{m}{2},$$

where $\hat{\sigma}^2$ = estimate of the asymptotic variance and $\hat{\theta}_{t,m}$ and $\hat{\gamma}_{t,m}$ = ML estimates computed over each moving window

Example - Implementation of the fluctuation test

- Derive the asymptotic distribution of F_t^{IS} under the null hypothesis the models perform equally well at each point in time
- We provide boundary lines - depending on m/T - that are crossed by the sample path of the limiting process with small probability under the null
- Reject the null if the sample path of the fluctuation statistics crosses boundaries

Example - The fluctuation test in practice



Example - Open issues

- General issues with fluctuation test:
 1. Tradeoffs in choice of moving window size m .
 - E.g., for larger m the smoothed $\Delta KLIC$ better approximated by its sample analog but smoothed $\Delta KLIC$ may appear constant whereas true $\Delta KLIC$ time-varying
 2. Fluctuation test does not specify an alternative hypothesis \implies flexible but may have low power \implies think of optimal tests against specific alternatives

Outline of the talk

- Motivating example - In-sample fluctuation test
- Theory
- Monte Carlo evidence
- Empirical application to DSGE vs. VAR
- Conclusion

Assumptions

- Two competing (misspecified) models depending on parameters θ and γ
 - Possibly non-linear, dynamic. For in-sample fluctuation test, sequential test \implies could be multivariate
 - In-sample fluctuation test, sequential test \implies non-nested models only
 - Estimation methods allowed:
 - * In-sample fluctuation test, sequential test \implies maximum likelihood
 - * Out-of-sample fluctuation test \implies general estimation procedure

Assumptions

- Primitive assumption: Functional Central Limit Theorem for $T^{-1/2} \sum_{j=1}^t \Delta L_t$. ($L_t = \log$ -likelihood for in-sample; forecast loss for out-of-sample)
- "Global covariance stationarity" under $H_0 = \text{variance of } \Delta L_t$ may be unstable in finite samples but instability vanishes asymptotically (could be relaxed, but complicates statement of FCLT. See Wooldridge and White, 1988) \implies satisfied in two scenarios above
- $m/T \rightarrow \mu$ finite and positive
- Out-of-sample fluctuation test: in-sample size R fixed (same as Giacomini and White, 2006)

In-sample fluctuation test - Null hypothesis

- Smoothed $\Delta KLIC$ is zero at each point in time

$$H_0 : E \left[\frac{1}{m} \sum_{j=t-m/2+1}^{t+m/2} \Delta L_j(\theta_{t,m}^*, \gamma_{t,m}^*) \right] = 0$$

$$\text{for all } t = \frac{m}{2} + 1, \dots, T - \frac{m}{2}$$

- $\theta_{t,m}^*, \gamma_{t,m}^*$ pseudo-true parameters for each moving window of size m
- Joint hypothesis: equal performance + performance stable over time

In-sample fluctuation test - Implementation

- Compute sequence of statistics

$$F_{t,m}^{IS} = \hat{\sigma}^{-1} m^{-1/2} \sum_j \Delta L_j (\hat{\theta}_{t,m}, \hat{\gamma}_{t,m})$$

$$\text{for } t = \frac{m}{2} + 1, \dots, T - \frac{m}{2}$$

- $\hat{\theta}_{t,m}, \hat{\gamma}_{t,m}$ ML estimators over moving window of size m
- $\hat{\sigma}^2$ is a HAC estimator of the global asymptotic variance

$$\sigma^2 = \lim_{m \rightarrow \infty} \text{var} \left(m^{-1/2} \sum_j \Delta L_j (\theta_{t,m}^*, \gamma_{t,m}^*) \right)$$

Out-of-sample fluctuation test

- Same as in-sample, except first divide sample into in-sample portion (data $1, \dots, R$) and out-of-sample ($R + 1, \dots, T$) and compute forecast losses for out-of-sample data using fixed, rolling or recursive scheme

- Test statistic

$$F_{t,m}^{OOS} = \hat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m+1}^t \Delta L_j \left(\hat{\theta}_{j,R}, \hat{\gamma}_{j,R} \right), \quad t = R + m + 1, \dots, T.$$

$\hat{\theta}_{j,R}, \hat{\gamma}_{j,R}$ in-sample parameter estimates for the j -th out-of-sample forecast

Fluctuation test - Implementation

- For both in-sample and out-of-sample tests, under H_0 :

$$F_{t,m} \implies [\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)] / \sqrt{\mu},$$

where $t = [\tau T]$, $m = [\mu T]$, and $\mathcal{B}(\cdot)$ is a Brownian motion. The boundary lines for a significance level α are $\pm k_\alpha$, where k_α solves

$$P \left\{ \sup_{\tau} |[\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)] / \sqrt{\mu}| > k_\alpha \right\} = \alpha.$$

- We give a table with k_α for several values of a and m/T (obtained by simulation)
- H_0 is rejected when $\max_{m/2+1 \leq t \leq T-m/2} |F_{t,m}| > k_\alpha$

Sequential test

- Monitor the model-selection decision in the post-historical sample period
- Suppose model f was best over the historical sample up to time $T \implies E \left[T^{-1} \sum_{j=1}^T \Delta L_j(\theta_T^*, \gamma_T^*) \right] > 0$.
- Null hypothesis: model f is the best performing model for all post-historical sample points:

$$H_0 : E \left[t^{-1} \sum_{j=1}^t \Delta L_j(\theta_{t,m}^*, \gamma_{t,m}^*) \right] \geq 0 \text{ for } t = T + 1, T + 2, \dots,$$

- One-sided alternative $H_1 : E \left[t^{-1} \sum_{j=1}^t \Delta L_j(\theta_{t,m}^*, \gamma_{t,m}^*) \right] < 0$ at some $t \geq T$.

Sequential test

- Doing a sequence of Vuong's (1989) tests for each t rejects too often \implies we give critical values that control the overall size of the procedure
- Construct sequence of test statistics

$$J_t = \hat{\sigma}_t^{-1} t^{-1/2} \sum_{j=1}^t \Delta L_j(\hat{\theta}_{t,m}, \hat{\gamma}_{t,m}), t = T + 1, T + 2, \dots,$$

- The critical value at time t for a level α test is $c_\alpha = -\sqrt{r_\alpha^2 + \ln(t/T)}$, with, e.g., $r_\alpha = 2.7955$ for $\alpha = .05$

Outline of the talk

- Motivating example - In-sample fluctuation test
- Theory
- Monte Carlo evidence
- Empirical application to DSGE vs. VAR
- Conclusion

Monte Carlo evidence

- Compare in-sample fluctuation test and sequential test to Vuong's (1989) test

- DGP with parameter variation:

$$y_t = \theta_t x_t + \gamma_t z_t + \varepsilon_t, \quad t = 1, \dots, 400$$

$$\theta_t = 1 + \theta \cdot \mathbf{1}(200 < t \leq 250) + (1 - \theta) \cdot \mathbf{1}(t > 250)$$

$$\gamma_t = 1 + \gamma \cdot \mathbf{1}(200 < t \leq 250) + (1 - \gamma) \cdot \mathbf{1}(t > 250).$$

- Model 1: $y_t = \beta_1 x_t + u_{1t}$. Model 2: $y_t = \beta_2 z_t + u_{2t}$.
- Size: $\theta = \gamma = 0.5 \implies$ models are equally good.
- Power: $\theta = 0.95, \gamma = 0.4 \implies$ time variation in relative performance

Monte Carlo evidence

Rejection frequencies of nominal 5% tests.

		$F_{t,m}^{IS}$	Vuong
(a) Historical sample	Size	0.051	0.047
	Power	0.449	0.047
(b) Post-historical sample	t/T	J_t	Vuong
	1.5	0.010	0.121
	1.75	0.020	0.152
	2	0.032	0.179

Outline of the talk

- Motivating example - In-sample fluctuation test
- Theory
- Monte Carlo evidence
- Empirical application to DSGE vs. VAR
- Conclusion

Application: DSGE vs. VAR

- Smets and Wouters (2003) (SW): “An estimated DSGE model of the Euro Area”: estimation of a 7-equation linearized DSGE model with sticky prices and wages, habit formation, capital adjustment costs and variable capacity utilization.
- They find that the DSGE model has comparable fit to that of atheoretical VARs

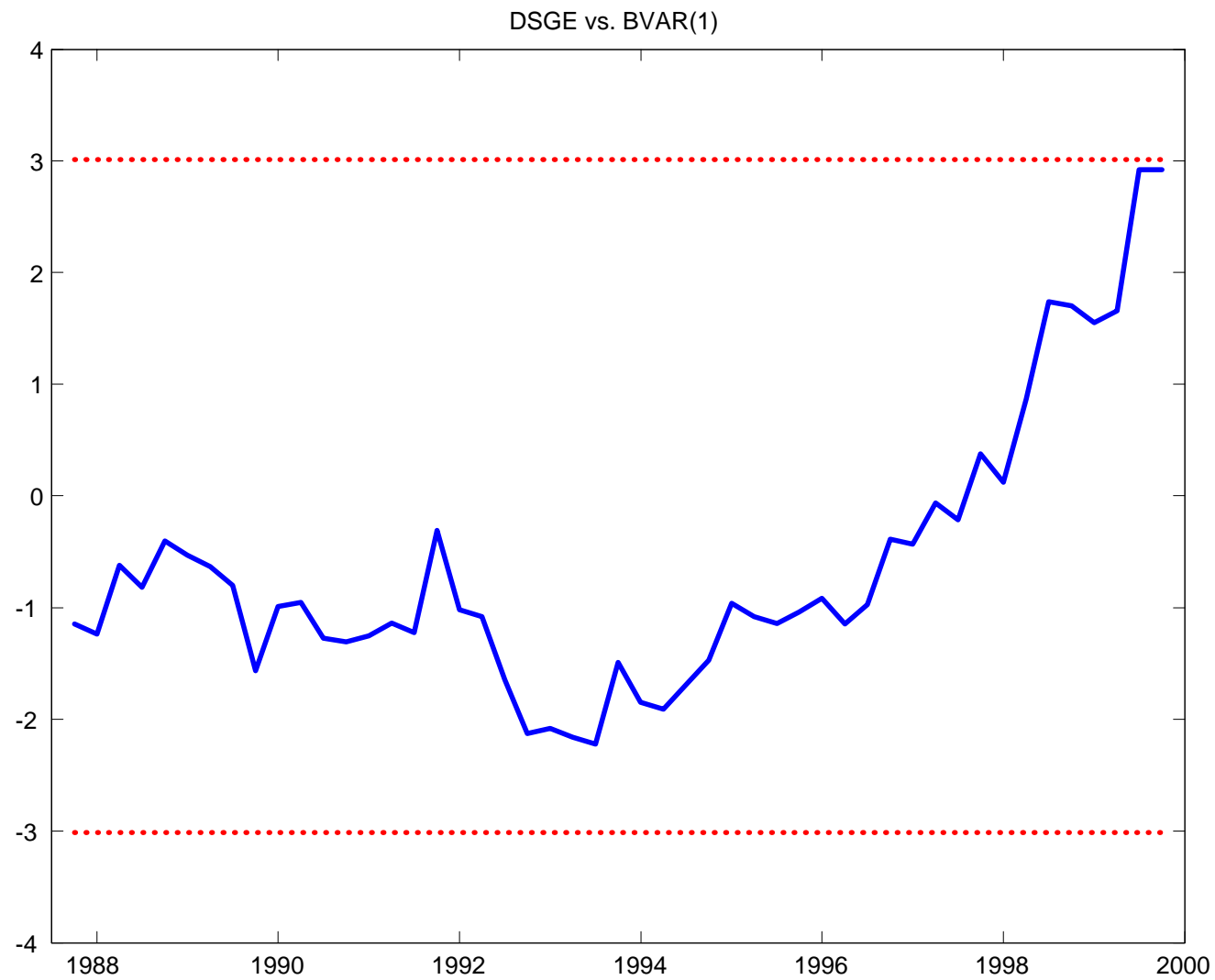
Application: DSGE vs. VAR

- Open questions:
 - Have the parameters been stable? Perhaps not. Possible structural changes in the economy (European union introduction, productivity changes, etc.)
 - If the parameters have changed \implies the performance of the DSGE model may have changed too... so SW's result only holds on average
 - Can we say that the performance of the DSGE and the VAR was equal at each point in time?

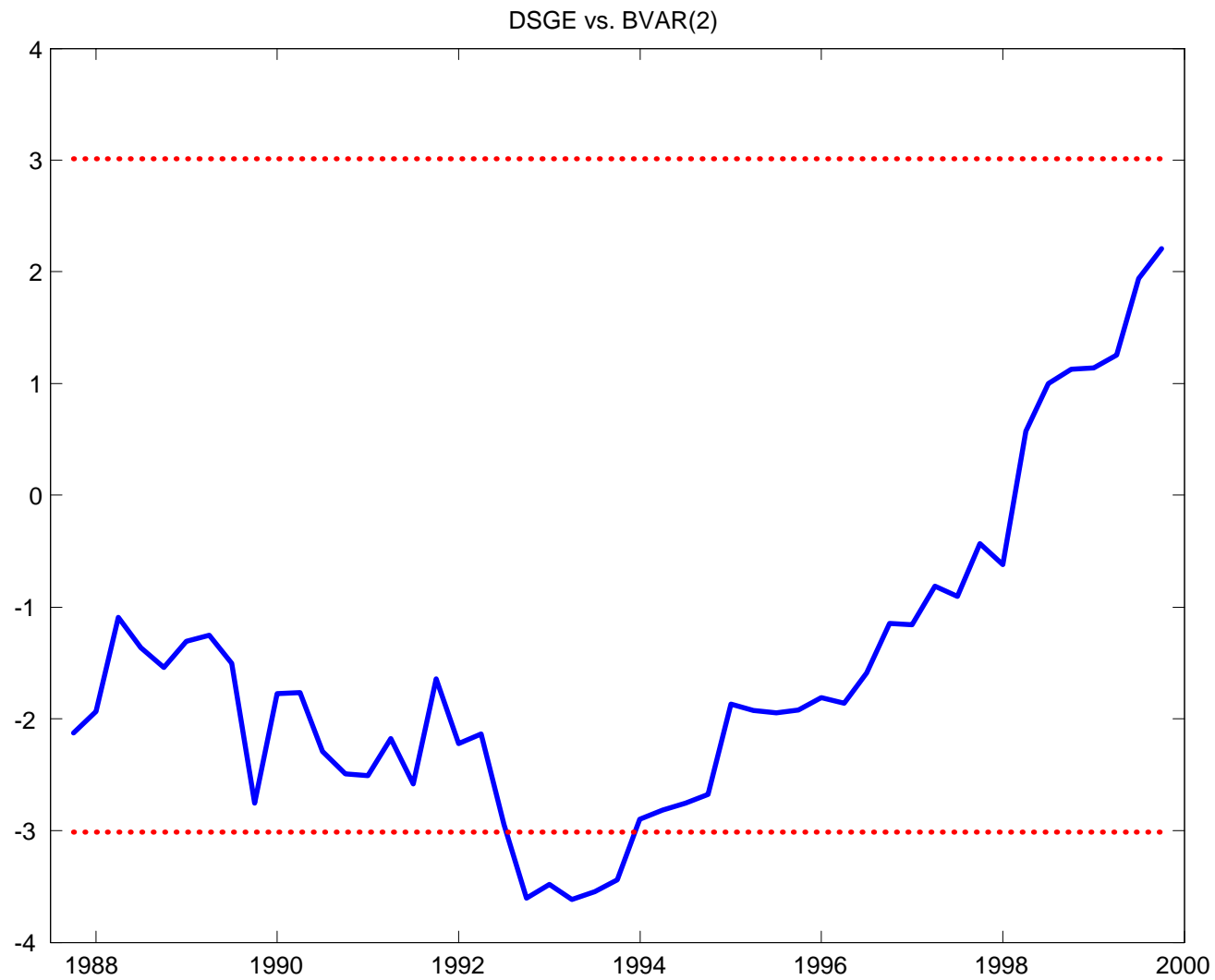
Application: DSGE vs. VAR

- SW sample: quarterly data 1970:2-1999:4 ($T = 118$) on DGP, consumption, investment, prices, real wages, employment, real interest rate
- Estimate DSGE model recursively by Bayesian methods (following SW) using moving windows of size $m = 70$
- Compare DSGE to BVAR(1), BVAR(2) with Minnesota priors
- In the fluctuation test statistic $F_{t,m}^{IS}$, $\hat{\theta}_{t,m}$, $\hat{\gamma}_{t,m}$ are the posterior modes (consistent estimators of pseudo-true parameters)

Fluctuation test - DSGE vs. BVAR(1)



Fluctuation test - DSGE vs. BVAR(2)



Conclusion and extensions

- Proposed a formal method for evaluating time-variation in relative performance of misspecified models
 - Two tests: Fluctuation (focus on historical samples) and Sequential (for real-time applications)
 - Two measures of performance: in-sample fit and out-of-sample forecast performance
- Empirical application confirmed SW's result that a DSGE has comparable performance to a BVAR in recent years
- Extension: optimal test against specific forms of time variation in relative performance