AEA Continuing Education Program

Time Series Econometrics

James H. Stock, Harvard

January 5-7, 2015

AEA Continuing Education Course

Time Series Econometrics

Lecture 4

Heteroskedasticity- and Autocorrelation-Robust Inference

or

Three Decades of HAC and HAR: What Have We Learned?

James H. Stock

Harvard University

January 6, 2015

Outline

HAC = Heteroskedasticity- and Autocorrelation-Consistent
HAR = Heteroskedasticity- and Autocorrelation-Robust

1) HAC/HAR Inference: Overview
2) Notational Preliminaries: Three Representations, Three Estimators
3) The PSD Problem and Equivalence of Sum-of-Covariance and Spectral Density Estimators
4) Three Approaches to the Bandwidth Problem
5) Application to Flat Kernel in the Frequency Domain
6) Monte Carlo Comparisons
7) Panel Data and Clustered Standard Errors
8) Summary

1) HAC/HAR Inference: Overview

The task: valid inference on β when X_t and u_t are possibly serially correlated:

$$Y_t = X_t'\beta + u_t,\ E(u_t|X_t) = 0,\ t = 1,\dots, T$$

Asymptotic distribution of OLS estimator:

$$\sqrt{T}(\hat{\beta}-\beta) = \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t u_t\right)$$

Assume throughout that WLLN and CLT hold:

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} \Sigma_{XX} \text{ and } \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t u_t \xrightarrow{d} \mathrm{N}(0, \Omega),$$

so $$\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} N\left(0, \Sigma_{XX}^{-1}\Omega\Sigma_{XX}^{-1}\right).$$

Σ_{XX} is easy to estimate, but what is Ω and how should it be estimated?

Ω: The Long-Run Variance of $X_t u_t$

Let $Z_t = X_t u_t$. Note that $EZ_t = 0$ (because $E(u_t|X_t) = 0$). Suppose Z_t is second order stationary. Then

$$\Omega_T = \mathrm{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\right) = E\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\right)^2$$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} E\left(Z_t Z_s'\right)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \Gamma_{t-s} \text{ (Z_t is second order stationary)}$$

$$= \frac{1}{T}\sum_{j=-(T-1)}^{T-1} \left(T - |j|\right)\Gamma_{t-s} \text{ (adding along the diagonals)}$$

$$= \sum_{j=-(T-1)}^{T-1} \left(1 - \left|\frac{j}{T}\right|\right)\Gamma_j \to \sum_{j=-\infty}^{\infty} \Gamma_j$$

so

$$\Omega = \sum_{j=-\infty}^{\infty} \Gamma_j = 2\pi S_Z(0) \quad \text{(recall that } S_Z(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \Gamma_j e^{-i\omega j}\text{)}$$

Standard approach: Newey-West Standard Errors

- HAC/HAR SEs are generically needed in time series regression. The most common method (by far) for computing HAC/HAR SEs is to use the Newey-West (1987) estimator.
- Newey-West estimator: declining average of sample autocovariances

$$\hat{\Omega}^{NW} = \sum_{j=-m}^{m} \left(1 - \left|\frac{j}{m}\right|\right) \hat{\Gamma}_j$$

where $\hat{\Gamma}_j = \frac{1}{T}\sum_{t=1}^{T} \hat{Z}_t \hat{Z}_{t-j}'$, where $\hat{Z}_t = X_t \hat{u}_t$.

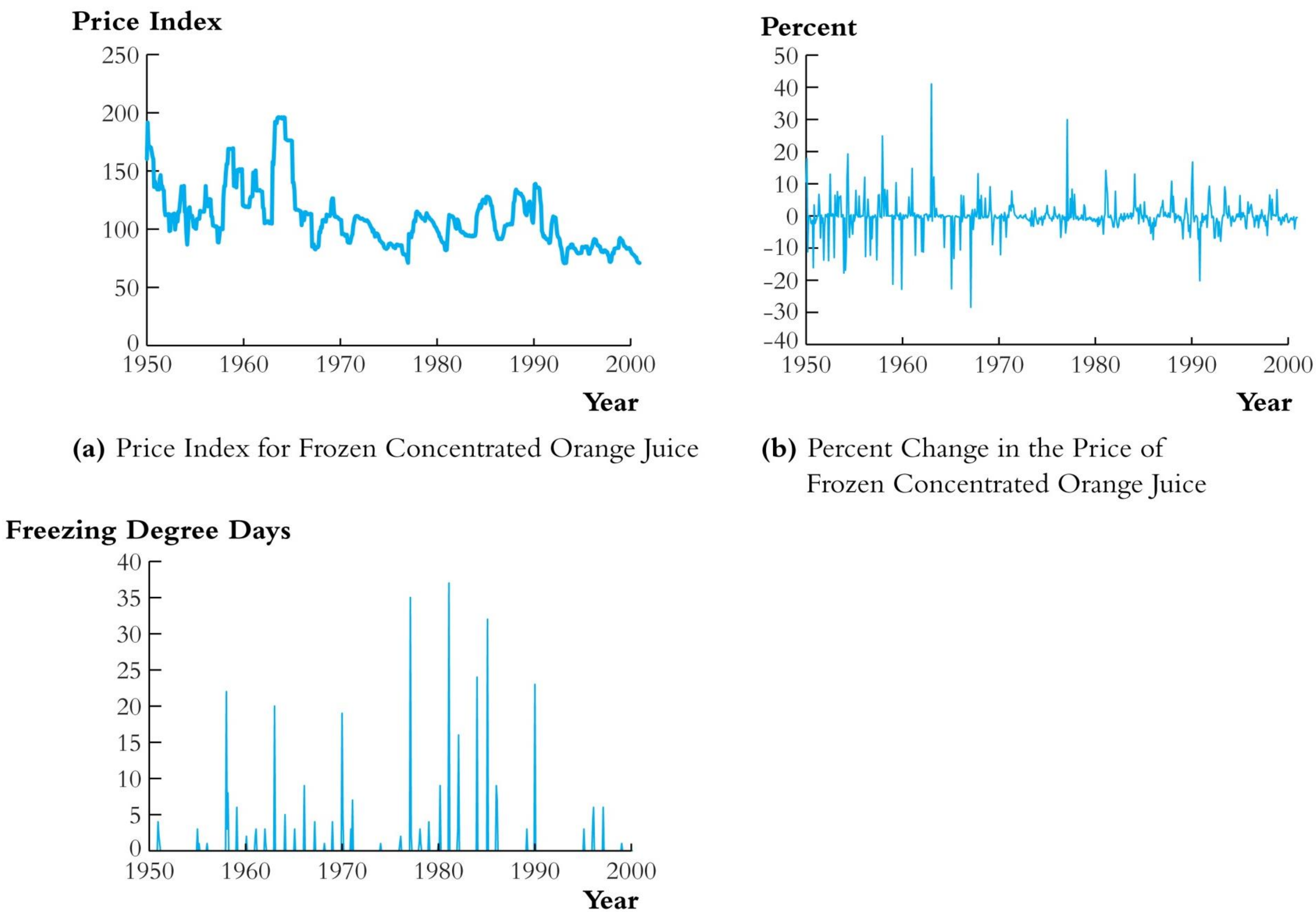
- Rule-of-thumb for m: $m = m_T = .75T^{1/3}$ (e.g. Stock and Watson, *Introduction to Econometrics*, 3rd edition, equation (15.17).
  - This rule-of-thumb dates to the 1990s. More recent research suggests it needs updating – and that, perhaps, the NW weights need to be replaced.

Four examples…


Florida's Orange Groves

Source: “USDA Assesses Freeze Damage of Florida Oranges,” Feb. 1, 2011 at
http://blogs.usda.gov/2011/02/01/usda-assesses-freeze-damage-of-florida-oranges/

FIGURE 15.1 Orange Juice Prices and Florida Weather, 1950–2000



(a) Price Index for Frozen Concentrated Orange Juice

(b) Percent Change in the Price of Frozen Concentrated Orange Juice

(c) Monthly Freezing Degree Days in Orlando, Florida

Example 1: OJ prices and Freezing degree-days:

$$\Delta\ln P_t = \alpha + \beta(\mathrm{L})FDD_t + u_t$$

Example 2: GDP growth and monetary policy shock:

$$\Delta\ln GDP_t = \alpha + \beta(\mathrm{L})\varepsilon_t^m + u_t$$

Example 3: Multiperiod asset returns:

$$\Delta\ln(P_{t+k}/P_t) = \alpha + \beta X_t + u_t^{t+l}, \text{ e.g. } X_t = \text{dividend yield}_t$$

Example 4: (GMM) Hybrid New Keynesian Phillips Curve:

$$\pi_t = \lambda x_t + \gamma_f E_t\pi_{t+1} + \gamma_b\pi_{t-1} + \eta_t$$

where x_t = marginal cost/output gap/unemployment gap and π_t = inflation. Suppose $\gamma_b + \gamma_f = 1$ (empirically supported); then

$$\Delta\pi_t = \lambda x_t + \gamma_f(E_t\pi_{t+1} - \pi_{t-1}) + \eta_t$$

Instruments: $\{\pi_{t-1}, x_{t-1}, \pi_{t-2}, x_{t-2},\ldots\}$

- η_t could be serially correlated by omission of supply shocks

Digression: Why not just use GLS?

The path to GLS: suppose u_t follows an AR(1)

$$Y_t = X_t'\beta + u_t,$$

$$u_t = \rho u_{t-1} + \varepsilon_t,\ \varepsilon_t \text{ serially uncorrelated}$$

This suggests Cochrane-Orcutt quasi-differencing:

$$(1\text{-}\rho L)Y_t = ((1\text{-}\rho L)X_t)' + \varepsilon_t \text{ or } \tilde{y}_t = \tilde{x}_t'\beta + \varepsilon_t$$

(Feasible GLS uses an estimate of ρ – not the issue here)

Validity of the quasi-differencing regression requires $E(\varepsilon_t|\tilde{x}_t) = 0$:

$$E(\varepsilon_t|\tilde{x}_t) = E(u_t - \rho u_{t-1}|x_t - \rho x_{t-1}) = 0$$

For general ρ, this requires all the cross-terms to be zero:

(i) $E(u_t|x_t) = E(u_{t-1}|x_{t-1}) = 0$

(ii) $E(u_t|x_{t-1}) = 0$

(iii) $E(u_{t-1}|x_t) = 0$ – this condition fails in examples 1-4

2) Notational Preliminaries: Three Representations, Three Estimators

The challenge: estimate $\Omega = \sum_{j=-\infty}^{\infty} \Gamma_j$

- This is hard: the sum has ∞'s!

- Draw on the literature on estimation of the spectral density to estimate Ω

- Three estimators of the spectral density:

(1) Sum-of-covariances: $$\hat{\Omega}^{sc} = \sum_{j=-(T-1)}^{T-1} k_T(j)\hat{\Gamma}_j$$

(2) Weighted periodogram: $$\hat{\Omega}^{wp} = 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) I_{\hat{Z}\hat{Z}}(2\pi l / T)$$

(3) VARHAC: $$\hat{\Omega}^{VARHAC} = \hat{A}(1)^{-1}\hat{\Sigma}_{\hat{u}\hat{u}}\hat{A}(1)^{-1}$$

We follow the literature and focus on (1) and (2)

(1) Sum-of-covariances estimator of Ω

$$\Omega = \sum_{j=-\infty}^{\infty} \Gamma_j$$

Because Z_t is stationary and Ω exists, Γ_j dies off. This suggests and estimator of Ω based a weighted average of the first few sample estimators of Γ:

$$\hat{\Omega}^{sc} = \sum_{j=-(T-1)}^{T-1} k_T(j)\hat{\Gamma}_j$$

where $\hat{\Gamma}_j = \frac{1}{T}\sum_{t=1}^{T} Z_t Z_{t-j}{}'$ (throughout, use the convention $Z_t = 0$, $t<1$ or $t > T$)

$k_T(.)$ is the weighting function or "kernel":

- Example: $k_T(j) = 1 - |j/m_T|$ = "triangular weight function" = "Bartlett kernel" = "Newey-West weights" with truncation parameter m_T
- We return to kernel and truncation parameter choice problem below

(2) Smoothed periodogram estimator of Ω

The periodogram as an inconsistent estimator of the spectral density:

- Fourier transform of Z_t at frequency ω: $d_Z(\omega) = \dfrac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T} Z_t e^{-i\omega t}$
- The periodogram is $I_{ZZ}(\omega) = d_Z(\omega)\overline{d_Z(\omega)}'$

Asymptotically, $I_{ZZ}(\omega)$ is distributed as $S_Z(0)\times(\chi_2^2/2)$ (scalar case)

- Mean:

$$E\, I_{ZZ}(\omega) = E(d_Z(\omega)\overline{d_Z(\omega)}')$$

$$= \frac{1}{2\pi} E\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t e^{i\omega t}\right|^2$$

$$= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \Gamma_j e^{-i\omega j} = S_Z(\omega)$$

- Distribution (Brillinger (1981), Priestley (1981), Brockwell and Davis (1991)):

$$d_Z(\omega) = \frac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T} Z_t e^{i\omega t}$$

$$= \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \cos\omega t + i\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \sin\omega t\right)$$

$= z_1 + iz_2$, say, where z_1 and z_2 are i.i.d. mean zero normal

So

$$I_{ZZ}(\omega) = d_Z(\omega)\overline{d_Z(\omega)}' = z_1^2 + z_2^2 \xrightarrow{d} S_Z(\omega)\times(\chi_2^2/2)$$

- For ω evaluated at $\omega_j = 2\pi j/T$, $j = 0, 1,\ldots, T$, $d_Z(\omega_j)$ and $d_Z(\omega_k)$ are asymptotically independent (orthogonality of sins and cosines).
- The weighted periodogram estimator averages the periodogram near zero:

$$\hat{\Omega}^{wp} = 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) I_{ZZ}(2\pi l / T)$$

(3) VAR-HAC estimator of Ω

Approximate the dynamics of Z_t by a vector autoregression: A(L)Z_t = u_t

so Z_t has the vector MA representation, Z_t = A(L)$^{-1}u_t$

Thus

$$S_Z(\omega) = \frac{1}{2\pi} A\left(e^{i\omega}\right)^{-1} \Sigma_{uu} \overline{A\left(e^{i\omega}\right)}^{-1\prime}$$

so

$$S_Z(0) = \frac{1}{2\pi} A\left(1\right)^{-1} \Sigma_{uu} A\left(1\right)^{-1\prime}$$

This suggests the VAR-HAC estimator (Priestley (1981), Berk (1974); den Haan and Levin (1997),

$$\hat{\Omega}^{VARHAC} = \hat{A}(1)^{-1} \hat{\Sigma}_{\hat{u}\hat{u}} \hat{A}(1)^{-1}$$

where $\hat{A}(1)$ and $\hat{\Sigma}_{\hat{u}\hat{u}}$ are obtained from a VAR estimated using $\hat{Z}_t$.

3) The PSD Problem and Equivalence of Sum-of-Covariance and Spectral Density Estimators

Not all estimators of Ω are positive semi-definite – including some natural ones. Consider the m-period return problem – so under the null $\beta = 0$, u_t is a MA(m-1). This suggests using a specific sum of covariances estimator:

$$\tilde{\Omega} = \sum_{j=-(m-1)}^{m-1} \hat{\Gamma}_j .$$

But $\tilde{\Omega}$ isn't psd with probability one! Consider $m = 2$ and the scalar case:

$$\tilde{\Omega} = \sum_{j=-1}^{1} \hat{\gamma}_j = \hat{\gamma}_0 \left(1 + 2\frac{\hat{\gamma}_1}{\hat{\gamma}_0}\right) < 0 \text{ if } \frac{\hat{\gamma}_1}{\hat{\gamma}_0} = \text{first sample autocorrelation} < \text{-}0.5$$

Solutions to the PSD problem

- Restrict kernel/weight function so that estimator is PSD with probability one (standard method)
- Hybrid, e.g. use $\tilde{\Omega}$ but switch to PSD method if $\tilde{\Omega}$ isn't psd – won't pursue (not used in empirical work)

Choice of kernel so that $\hat{\Omega}^{sc}$ is psd w.p.1

Step 1:

Note that $\hat{\Omega}^{wp}$ is psd w.p.1 if the frequency-domain weight function is non-negative. Recall that $\hat{\Omega}^{wp}$ is psd if $\lambda'\hat{\Omega}^{wp}\lambda \geq 0$ for all λ. Now

$$\begin{aligned}
\lambda'\hat{\Omega}^{wp}\lambda &= 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l)\left(\lambda' I_{ZZ}(2\pi l/T)\lambda\right) \\
&= 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l)\left(\lambda' d_Z(\omega_l)\overline{d_Z(\omega_l)}'\lambda\right) \\
&= 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l)\left|\lambda' d_Z(\omega_l)\right|^2 \geq 0
\end{aligned}$$

with probability 1 if $K_T(l) \geq 0$ for all l.

- $K_T(l) \geq 0$, all l, is necessary and sufficient for $\hat{\Omega}^{wp}$ to be psd

Step 2: $\hat{\Omega}^{wp}$ and $\hat{\Omega}^{sc}$ are equivalent!

$$\hat{\Omega}^{wp} = 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) I_{ZZ}(2\pi l / T)$$

$$= 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) \left(\frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} Z_t e^{i2\pi lt/T} \right) \left(\frac{1}{\sqrt{2\pi T}} \sum_{s=1}^{T} Z_s e^{-i2\pi ls/T} \right)$$

$$= \sum_{l=-(T-1)}^{T-1} K_T(l) \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} Z_t Z_s{}' e^{-i2\pi l(s-t)/T}$$

$$= \sum_{l=-(T-1)}^{T-1} K_T(l) \sum_{j=-(T-1)}^{T-1} \frac{1}{T} \sum_{t=1}^{T} Z_t Z_{t-j}{}' e^{-i2\pi lj/T}$$

$$= \sum_{j=-(T-1)}^{T-1} \frac{1}{T} \sum_{t=1}^{T} Z_t Z_{t-j}{}' \sum_{l=-(T-1)}^{T-1} K_T(l) e^{-i(2\pi j/T)l}$$

$$= \sum_{j=-(T-1)}^{T-1} \hat{\Gamma}_j k_T(j) = \hat{\Omega}^{sc}, \text{ where } k_T(j) = \sum_{l=-(T-1)}^{T-1} K_T(l) e^{-i(2\pi j/T)l}$$

Result: $\hat{\Omega}^{sc}$ is psd w.p.1 if and only if k_T is the (inverse) Fourier transform of a nonnegative frequency domain weight function K_T. Also, k_T is real if K_T is symmetric (then $k_T(j) = K_T(0) + 2\sum_{l=1}^{T-1} K_T(l) \cos\left[(2\pi j / T)l\right]$).

Kernel and bandwidth choice

The class of estimators here is very large. What is a recommendation for empirical work?

Two distinct questions:

(i) What kernel to use?

(ii) Given the kernel, what bandwidth to use?

It turns out that problem (ii) is more important in practice than problem (i).

Some final preliminaries

- Closer look at four kernels:
  - Newey-West (triangular in time domain)
  - Flat in time domain
  - Flat in frequency domain
  - Epinechnikov (Quadratic Spectral) – certain optimality properties
- Link between time domain and frequency domain kernels

Flat kernel in frequency domain

In general:

$$\hat{\Omega}^{wp} = 2\pi \sum_{l=-(T-1)}^{T-1} K_T(l) I_{ZZ}(2\pi l / T)$$

Flat kernel:

$$K_T(l) = \begin{cases} \dfrac{1}{2B_T + 1} & \text{if } |l| \le B_T \\ 0 & \text{if } |l| > B_T \end{cases}$$

Then $\hat{\Omega}^{wp}$ becomes

$$\hat{\hat{\Omega}} = \frac{2\pi}{2B_T + 1} \sum_{l=-B_T}^{B_T} I_{ZZ}\left(\frac{2\pi l}{T}\right)$$

The time-domain kernel corresponding to the flat frequency-domain kernel is

$$k_T(j) = \sum_{l=-(T-1)}^{T-1} K_T(l)e^{-i(2\pi j/T)l}$$

$$= \frac{1}{2B_T+1}\sum_{l=-B_T}^{B_T} e^{-i(2\pi j/T)l}$$

$$= \ldots \to_{T\to\infty} \frac{\sin(2\pi j/m_T)}{2\pi j/m_T}$$, where $m_T = T/B_T$

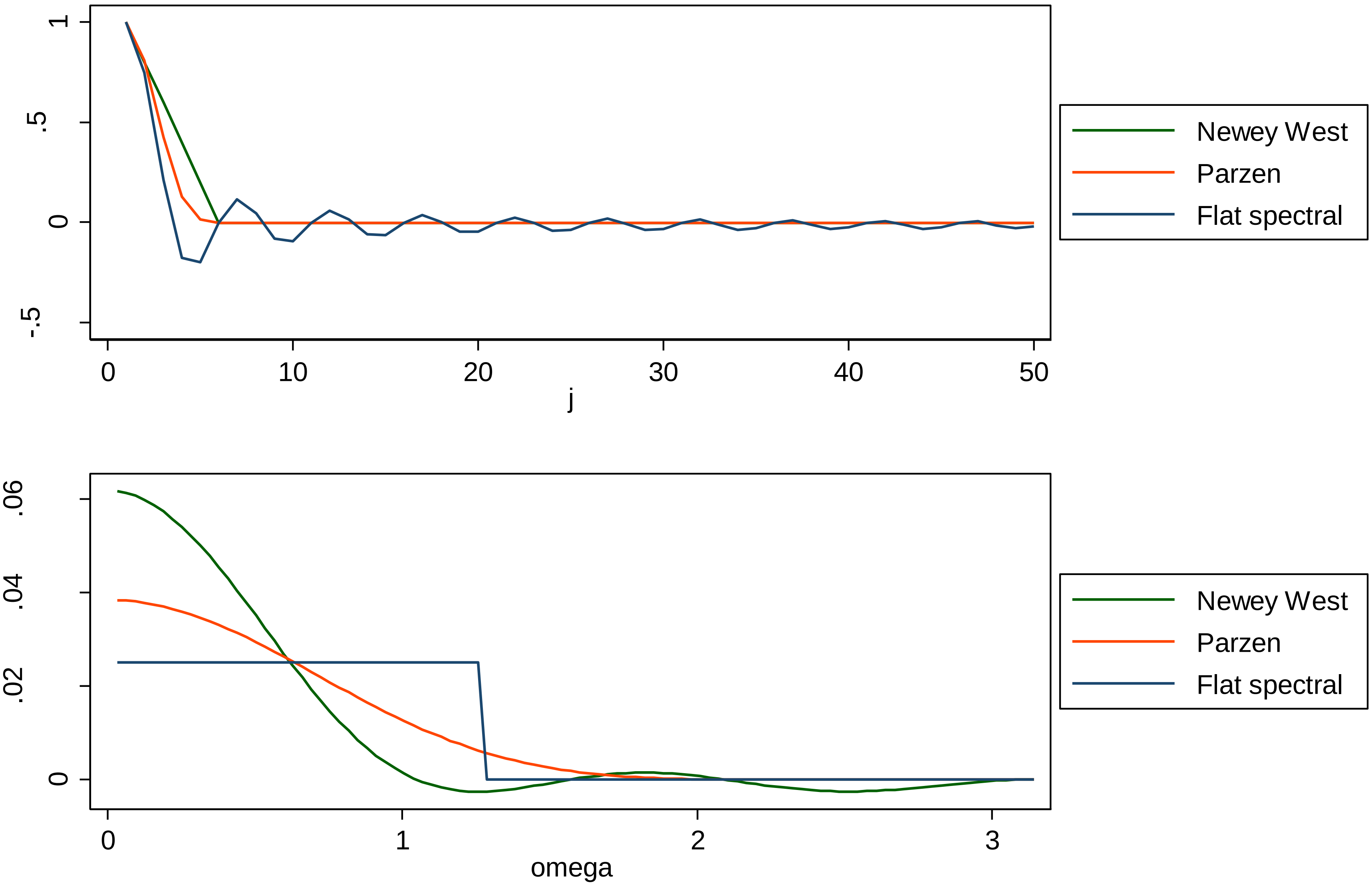
Important points:

- $m_T B_T = T$: using few periodogram ordinates corresponds to using many covariances
- Flat in frequency domain (which is psd) produces some negative weights in the sum-of-covariance kernel

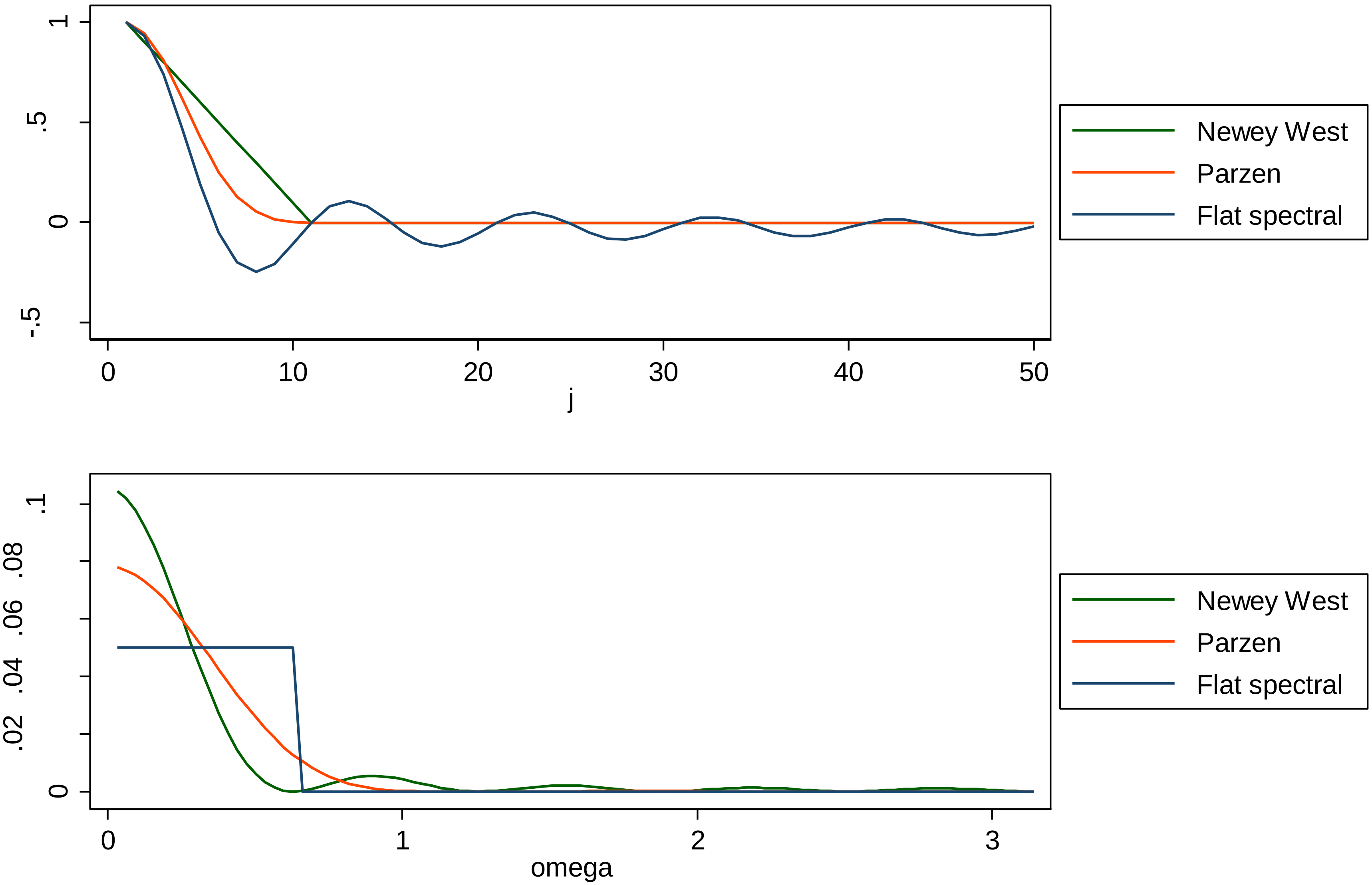
Three PSD kernels in pictures

Kernel	$k(x)$, $x = \|j\|/m$	$K(u)$, $u = \|l\|/B$
Newey-West	$1-\|x\|$ if $\|x\| \le 1$	
Parzen	$1 - 6x^2 + 6\|x\|^3$ if $\|x\|<.5$ $2(1-\|x\|)^3$ if $.5\le \|x\| \le 1$	
Flat spectral		1 if $\|u\| \le 1$

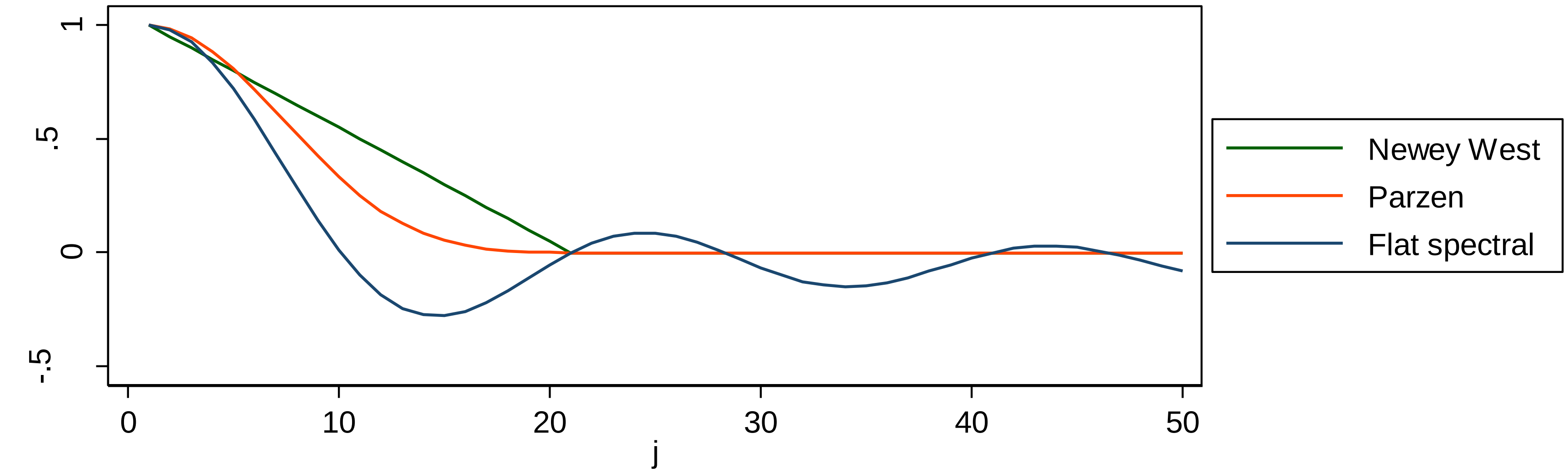
Three PSD Kernels: m = 5, B = 40, T = 200

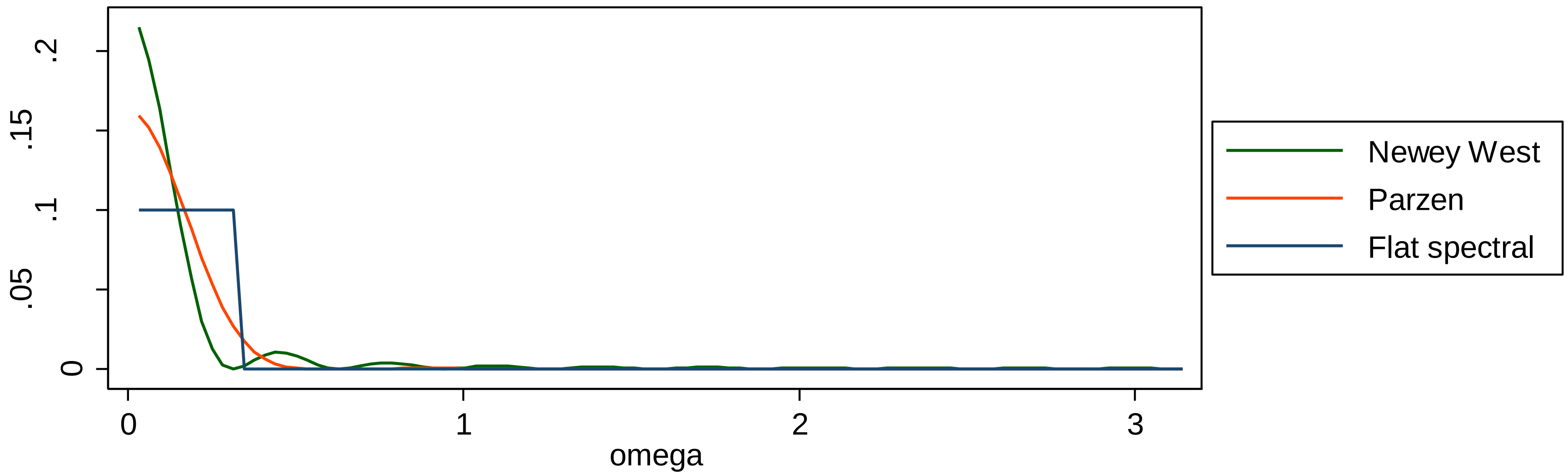


Three PSD Kernels: m = 10, B = 20, T = 200

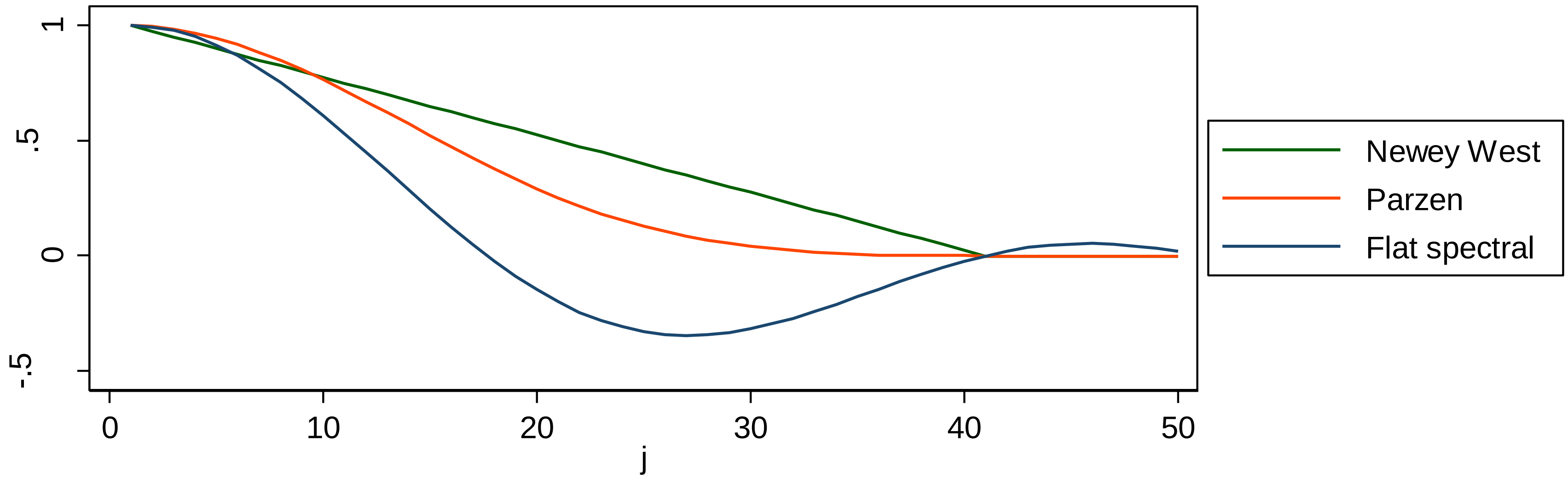


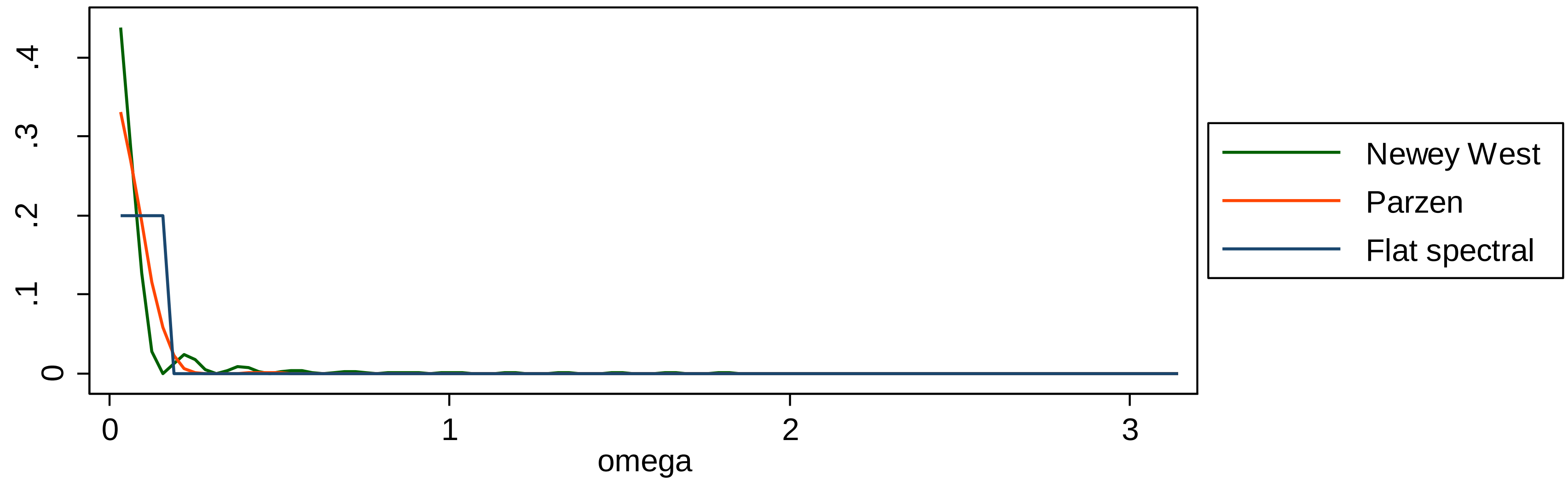
Three PSD Kernels: m = 20, B = 10, T = 200





Three PSD Kernels: m = 40, B = 5, T = 200





4) Three Approaches to the Bandwidth Problem

As in all nonparametric problems, there is a fundamental tradeoff between bias and variance when choosing smoothing parameters.

- In frequency domain:

$$\hat{\Omega}^{wp} = 2\pi \sum_{l=-B}^{B} K_T(l) I_{ZZ}(2\pi l / T)$$

Larger B decreases variance, but increases bias

- In time domain:

$$\hat{\Omega}^{sc} = \sum_{j=-m}^{m} k_T(j) \hat{\Gamma}_j$$

Larger m increases variance, but decreases bias

- Recall $m_T B_T = T$

How should this bias-variance tradeoff be resolved?

First generation answer:

Obtain as good an estimate of Ω as possible (Andrews [1991])

- "Good" means:
  - psd with probability 1
  - consistent (HAC)
  - minimize mean squared error:

    $$\text{MSE}(\hat{\Omega}) = E(\hat{\Omega} - \Omega)^2 = \text{bias}(\hat{\Omega})^2 + \text{var}(\hat{\Omega})$$
  - This yields a bandwidth m_T that increases with, but more slowly than, T
- Practical issue:
  - if true spectral density is flat in neighborhood of zero, you should include many periodogram ordinates (large B); equivalently, if true Γ_j's are small for $j\neq0$ then you should include few $\hat{\Gamma}_j$'s
  - But, you don't know the true spectral density!!
  - So, in practice you can estimate and plug in, or use a rule-of-thumb.
  - The $m = .75T^{1/3}$ rule of thumb assumes X_t and u_t are AR(1) with coefficient 0.5
- Then use asymptotic chi-squared critical values to evaluate test statistics.

Big problem with the first generation answer

- The resulting estimators do a very bad job of controlling size when the errors are in fact serially correlated, even with a modest amount of serial correlation
    - den Haan and Levin (1997) provided early complete Monte Carlo assessment
    - We will look at MC results later

- Why? The key insight is that the min MSE problem isn't actually what we are interested in – we are actually interested in size control or equivalently coverage rates of confidence intervals.
    - For coverage rates of confidence intervals, what matters is not bias^2, but bias (Velasco & Robinson [2001]; Kiefer & Vogelsang [2002]; Sun, Phillips, and Jin (2008))

- Practical implication: use fewer periodogram ordinates (smaller *B*) i.e. more autocovariances (larger *m*).

Approach #2: Retain consistency, but minimize size distortion

Sketch of asymptotic expansion of size distortion

for details see Velasco and Robinson (2001), Sun, Phillips, and Jin (2008)

Consider the case of a single X and the null hypothesis $\beta = \beta_0$. Then $u_t = Y_t - X_t\beta_0$, and $Z_t = X_t u_t$, so the Wald test statistic is,

$$W_T = \frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\hat{\Omega}}$$

The probability of rejection under the null thus is,

$$\Pr[W_T < c] = \Pr\left[\frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\hat{\Omega}} < c\right]$$

where c is the asymptotic critical value (3.84 for a 5% test). The size distortion is obtained by expanding this probability…

First, note that $T^{-1/2}\sum_1^T Z_t$ and $\hat{\Omega}$ are asymptotically independent. Now

$$\Pr[W_T < c] = \Pr\left[\frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\hat{\Omega}} < c\right] = \Pr\left[\frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\Omega} < c\frac{\hat{\Omega}}{\Omega}\right]$$

$$= E\left\{\Pr\left[\frac{\left(T^{-1/2}\sum_1^T Z_t\right)^2}{\Omega} < c\frac{\hat{\Omega}}{\Omega}\right] \middle| \hat{\Omega}\right\}$$

$$\approx E\left[F\left(c\frac{\hat{\Omega}}{\Omega}\right)\right]$$, where F = chi-squared c.d.f

$$= E\left[F(c) + cF'(c)\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right) + \frac{1}{2}cF''(c)\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2 + \ldots\right]$$

so the size distortion approximation is,

$$\Pr[W_T < c] - F(c) \approx cF'(c)\frac{bias(\hat{\Omega})}{\Omega} + \frac{1}{2}cF''(c)\frac{MSE(\hat{\Omega})}{\Omega^2}$$

or

$$\Pr[W_T < c] - F(c) \approx cF'(c)\frac{bias(\hat{\Omega})}{\Omega} + \frac{1}{2}cF''(c)\frac{\mathrm{var}(\hat{\Omega})}{\Omega^2} + \text{smaller terms}$$

Thus minimizing the size distortion entails minimizing a linear combination of bias and variance – *not* bias^2 and variance

Approach #3: "Fixed b" asymptotics

- Drop consistency – but use correct critical values that account for additional variance (HAR)
  - This decision has a cost – consistency provides first-order asymptotic efficiency of tests – but this isn't worth much if you don't have size control

- Fixed *b* corresponds in our notation to fixed *B* (or, equivalently, to $m \propto T$)
  - The fixed-*b* calculations typically use a FCLT approach, see Kiefer-Vogelsang (2002), Müller (2007), Sun (2013).
  - We will sidestep the FCLT results by using classical results from the spectral density estimation literature for the flat kernel in the frequency domain.

5) Application to Flat Kernel in the Frequency Domain

Consider scalar X_t and flat-kernel in frequency domain:

$$\hat{\hat{\Omega}} = \frac{2\pi}{2B_T}\sum_{l=-B}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right) = \frac{2\pi}{B_T}\sum_{l=1}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right)$$

- This adjusts the kernel to drop $\omega = 0$ since $I_{\hat{Z}\hat{Z}}(0) = 0$ (OLS residuals are orthogonal to X)
- The second equality holds because
  (i) in scalar case, $I_{ZZ}(\omega) = I_{ZZ}(-\omega)$, and
  (ii) $I_{\hat{Z}\hat{Z}}(0) = 0$ because $d_{\hat{Z}}(0) = 0$ ($\hat{u}_t$ are OLS residuals)
- This kernel plays a special historical role in frequency domain estimation.

We now provide explicit results for the three approaches:

i. Fixed B (this kernel delivers asymptotic t_{2B} inference!)
ii. Min MSE
iii. Min size distortion

i. Fixed *b*

- For this kernel, you don't need to use FCLT approach – the result for its fixed-*B* distribution is very old and is a cornerstone of classical theory of frequency domain estimation (e.g. Brillinger (1981)). For X_t, u_t stationary, with suitable moment conditions,

  (a) $\hat{\hat{\Omega}} \xrightarrow{d} \Omega \times (\chi^2_{2B} / 2B)$, that is,

  $$\hat{\hat{\Omega}} \sim \Omega \times (\chi^2_{2B} / 2B)$$

  (b) Moreover $\hat{\hat{\Omega}}$ is asymptotically independent of $T^{-1/2}\sum_1^T Z_t \sim \mathrm{N}(0,\Omega)$

- It follows that, for *B* fixed, the *t* statistic has an asymptotic t_{2B} distribution:

  $$t = \frac{T^{-1/2}\sum_1^T Z_t}{\hat{\Omega}^{1/2}} \xrightarrow{d} t_{2B}$$

- This result makes the size/power tradeoff clear – using t_{2B} distribution has power loss relative to asymptotically efficient normal inference – but the power loss is slight for $B \geq 10$ (say).

Sketch of (a) and (b):

Consider scalar case, and recall that $I_{\hat{Z}\hat{Z}}(0) = 0$ (OLS residuals), so

(a) Distribution of $\hat{\hat{\Omega}}$ with B fixed:

$$\hat{\hat{\Omega}} = \frac{2\pi}{B}\sum_{l=1}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right)$$

$$\sim \frac{2\pi}{B}\sum_{l=1}^{B} S_{ZZ}\left(\frac{2\pi l}{T}\right)\xi_l \text{, where } \xi_l \sim \chi_2^2 / 2$$

$$= \frac{2\pi}{B}\sum_{l=1}^{B}\left[S_{ZZ}(0) + \frac{1}{2}\left(\frac{2\pi l}{T}\right)^2 S_{ZZ}''(0) + \ldots\right]\xi_l$$

$$\approx \frac{2\pi}{B}\sum_{l=1}^{B} S_{ZZ}(0)\xi_l$$

$$= 2\pi S_{zz}(0) \times (\chi_{2B}^2 / 2B)$$

$$= \Omega \times (\chi_{2B}^2 / 2B)$$

(b) $\hat{\hat{\Omega}}$ is independent of $T^{-1/2}\sum_1^T Z_t$. This follows from the result above that $d_Z(\omega_l)$ and $d_Z(\omega_k)$ are asymptotically independent, applied here to $d_Z(0)$ (the numerator) and d_Z at other ω_l's (the denominator)

ii. and iii. – Preliminaries for the asymptotic expansions

Bias

$$E\left(\hat{\hat{\Omega}}-\Omega\right) = E\left[\frac{2\pi}{B}\sum_{l=1}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right) - S_{ZZ}(0)\right]$$

$$\approx \frac{2\pi}{B}\sum_{l=1}^{B}\left[S_{ZZ}\left(\frac{2\pi l}{T}\right) - S_{ZZ}(0)\right]$$

$$= \frac{2\pi}{B}\sum_{l=1}^{B}\left\{\left[S_{ZZ}(0) + \frac{2\pi l}{T}S_{ZZ}{}'(0) + \frac{1}{2}\left(\frac{2\pi l}{T}\right)^2 S_{ZZ}{}''(0) + \ldots\right] - S_{ZZ}(0)\right\}$$

$$= \frac{2\pi}{B}\sum_{l=1}^{B}\left\{\left[S_{ZZ}(0) + \frac{2\pi l}{T}S_{ZZ}{}'(0) + \frac{1}{2}\left(\frac{2\pi l}{T}\right)^2 S_{ZZ}{}''(0) + \ldots\right] - S_{ZZ}(0)\right\}$$

Because $S_{ZZ}(\omega) = S_{ZZ}(-\omega)$, $S_{ZZ}{}'(0) = 0$, and after dividing by Ω,

$$E\left(\hat{\hat{\Omega}}-\Omega\right)\Big/\Omega = \left[\frac{2\pi}{B}\sum_{l=1}^{B}\frac{1}{2}\left(\frac{2\pi l}{T}\right)^2\right]S_{ZZ}{}''(0)\Big/2\pi S_{ZZ}(0) = \frac{1}{2d}\left(\frac{B}{T}\right)^2$$

where $d = \dfrac{3S_{ZZ}(0)}{4\pi^2 S_{ZZ}{}''(0)}$.

Variance

$$\frac{\mathrm{var}(\hat{\hat{\Omega}})}{\Omega^2} = \mathrm{var}\left[\frac{2\pi}{B}\sum_{l=1}^{B} I_{\hat{Z}\hat{Z}}\left(\frac{2\pi l}{T}\right)\right]\Big/\Omega^2$$

$$\approx \frac{4\pi^2}{B^2}\sum_{l=1}^{B}\mathrm{var}\left[I_{ZZ}\left(\frac{2\pi l}{T}\right)\right]\Big/\left(2\pi S_{ZZ}(0)\right)^2$$

$$= \frac{4\pi^2}{B^2}\sum_{l=1}^{B} S_{ZZ}\left(\frac{2\pi l}{T}\right)^2 \Big/ 4\pi^2 S_{ZZ}(0)^2 = \ldots = \frac{1}{B}$$

(keeping only the leading term in the Taylor series expansion).

Summary: relative bias and relative variance:

$$\frac{\mathrm{var}(\hat{\hat{\Omega}})}{\Omega^2} = \frac{1}{B} \quad \text{and} \quad \frac{E(\hat{\hat{\Omega}} - \Omega)}{\Omega} = \frac{1}{2d}\left(\frac{B}{T}\right)^2, \text{ where } d = \frac{3S_{ZZ}(0)}{4\pi^2 S_{ZZ}''(0)}$$

Special case: Z_t is AR(1) with autoregressive parameter $\alpha \neq 0$:

$$d = -\frac{3}{8\pi^2}\frac{(1-\alpha)^2}{\alpha}$$

ii. Min MSE

$$\text{Min}_B \text{ MSE}(\hat{\hat{\Omega}}) = \text{Min}_B \text{ bias}^2(\hat{\hat{\Omega}}) + \text{var}(\hat{\hat{\Omega}})$$

$$= \text{Min}_B \left[\frac{1}{2d}\left(\frac{B}{T}\right)^2 \Omega \right]^2 + \frac{\Omega^2}{B}$$

Solution:

$$B_T^{MinMSE}(\hat{\alpha}) = \left[d\right]^{2/5} T^{4/5}, \text{ where } d = \frac{3S_{ZZ}(0)}{4\pi^2 S_{ZZ}''(0)} = -\frac{3}{8\pi^2}\frac{(1-\alpha)^2}{\alpha}$$

iii. Min Size Distortion

$$\text{Min}_B \Pr[W_T < c] - F(c) \approx \text{Min}_B \, cF'(c)\frac{bias(\hat{\Omega})}{\Omega} + \frac{1}{2}cF''(c)\frac{\text{var}(\hat{\Omega})}{\Omega^2}$$

Solution (for $\alpha > 0$):

$$B_T^{1stOrderSize}(\hat{\alpha}) = \left[\frac{cF''(c)}{2F'(c)}d\right]^{1/3} T^{2/3}$$

where c = 3.84 for 5% tests and F is χ_1^2 cdf.

Optimal HAC Bandwidths for flat spectral kernel: Z_t AR(1) with parameter α

	T = 100				T = 800			
Minimize:	*MSE*		*Size distortion*		*MSE*		*Size distortion*	
α	B	m	B	m	B	m	B	m
.1	43	5	25	8	131	6	62	13
.2	30	7	18	11	90	9	45	18
.3	23	9	14	14	69	12	36	22
.4	18	11	12	17	54	15	30	27
.5	14	14	10	21	43	19	25	33
.6	11	18	8	25	33	24	20	40
.7	8	24	6	32	25	32	16	51
.8	6	35	5	44	17	47	11	70
.9	3	65	3	73	9	85	7	116

Notes: b = bandwidth in frequency domain, m = lag truncation parameter in time domain.

- The rule-of-thumb $m = .75T^{1/3}$ corresponds to m = 4 for T = 100 and m = 7 for T = 800 (however not directly comparable since the rule-of-thumb is for the Newey-West kernel).

6) Monte Carlo Comparisons

Illustrative results:

- Design: $X_t = 1$, u_t AR(1)
- Flat spectral kernel (so that t_{2B} inference is asymptotically valid under fixed-b asymptotics)
- Two bandwidth choices: min MSE and minimize size distortion
- Bandwidths chosen using plug-in formula based on estimated α (formula given above, with $\hat{\alpha}$ replacing α)
- Additional MC results: den Haan and Levin (1997), Kiefer and Vogelsang (2002), Kiefer, Vogelsang and Bunzel (2000), Sun (2013).

NULL REJECTION RATE

		χ^2 c.v.		t c.v.	
φ	T	B_T^{MinMSE}	$B_T^{\text{1stOrderSize}}$	B_T^{MinMSE}	$B_T^{\text{1stOrderSize}}$
0.00	100	0.055	0.055	0.050	0.049
	400	0.052	0.052	0.051	0.050
0.50	100	0.094	0.088	0.075	0.066
	400	0.068	0.064	0.061	0.055
0.90	100	0.216	0.212	0.141	0.132
	400	0.111	0.107	0.083	0.073
0.95	100	0.310	0.309	0.195	0.190
	400	0.149	0.144	0.102	0.092

Table 1: Null rejection rates for tests based on χ^2 and t critical values, and on two different bandwidth formulas. 50,000 Monte Carlo repetitions.

7) Panel Data and Clustered Standard Errors

Clustered standard errors are an elegant solution to the HAC/HAR problem in panel data.

- Although the original proofs of clustered SEs used large N and small T (Arellano [2003]) in fact they are valid for small N if T is large (Hansen [2007], Stock and Watson [2008]), but using t or F (not normal or chi-squared) inference.
- The standard fixed effects panel data regression model

$$Y_{it} = \alpha_i + \beta' X_{it} + u_{it},\ i = 1,\ldots,N,\ t = 1,\ldots,\ T,$$

  where $E(u_{it}|X_{i1},\ldots,\ X_{iT},\ \alpha_i) = 0$ and u_{it} is uncorrelated across i but possibly serially correlated, with variance that can depend on t; assume i.i.d. over i
- The discussion here considers the special case $X_t = 1$– the ideas generalize

Clustered SEs with $X_t = 1$

$$Y_{it} = \alpha_i + \beta + u_{it},\ i = 1,\ldots,N,\ t = 1,\ldots,T,$$

The fixed effects (FE) estimator is

$$\hat{\beta}^{FE} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}$$

Thus

$$\sqrt{NT}(\hat{\beta}^{FE} - \beta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_{it}\right)$$

$$= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} v_i,\ v_i = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_{it}$$

For fixed N and large T, $v_i \xrightarrow{d} N(0,\Omega)$, $i = 1,\ldots, N$ (i.i.d.). Thus the problem is asymptotically equivalent to having N observations on v_i, which is i.i.d. $N(0,\Omega)$.

$X_t = 1$ case, continued:

Clustered variance formula: $\hat{\Omega}^{cluster} = \frac{1}{N}\sum_{i=1}^{N}(\hat{v}_i - \bar{\hat{v}})^2$, $\hat{v}_i = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{u}_{it}$

By standard normal/t arguments: $\hat{\Omega}^{cluster} \xrightarrow{d} \frac{\Omega\chi^2_{N-1}}{N} = \frac{\Omega\chi^2_{N-1}}{N-1} \times \frac{N-1}{N}$

and $t = \frac{\hat{\beta}^{FE} - \beta_0}{\sqrt{\hat{\Omega}^{cluster}}} \xrightarrow{d} \sqrt{\frac{N}{N-1}}t_{N-1}$

- Note the complication of the degrees of freedom correction – this is because the standard definition of $\hat{\Omega}^{cluster}$ has N, not N-1, in the denominator.
- Extension to multiple X: The F-statistic testing p linear restrictions on β, computed using $\hat{\Omega}^{cluster}$, is distributed $\frac{N}{N-p}F_{p,N-p}$
- For N very small, the power loss from t_{N-1} inference can be large – so for very small N it might be better to use HAC/HAR methods, not clustered SEs (not much work has been done on this tradeoff, however).

8) Summary

- Applications of HAC/HAR methods are generic in time series. GLS is typically not justified because it requires strict exogeneity (no feedback from u to X)
- Choice of the bandwidth is critical and reflects a tradeoff between bias and variance.
- The rule-of-thumb $m = .75T^{1/3}$ uses too few autocovariances (m is too small) – overweights variance at the expense of bias
- However, inference becomes complicated when large m (small B) is used, because this increases the variance of $\hat{\Omega}$.
- In general (including for N-W weights), fixed-b inference is complicated and requires specialized tables (e.g. Kiefer-Vogelsang inference).
- However, in the special case of the flat spectral kernel, asymptotically valid fixed-B inference is based on t_{2B}. Initial results for size control (and power) using this approach are promising.

AEA Continuing Education Course
Time Series Econometrics

Lectures 5 and 6

Weak Identification & Many Instruments in IV Regression and GMM

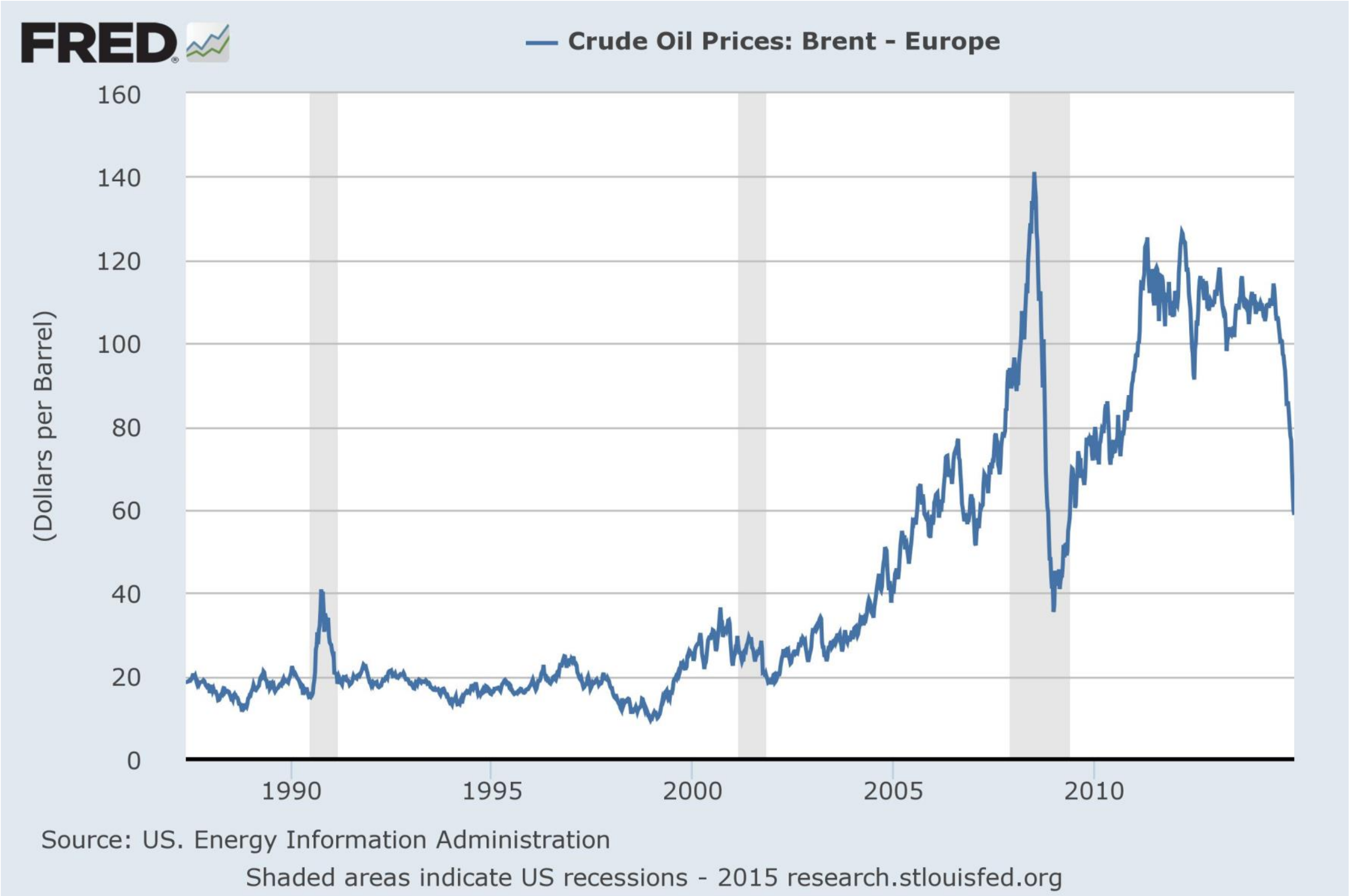
James H. Stock
Harvard University

January 6 & 7, 2015

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: Hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

Introductory Application

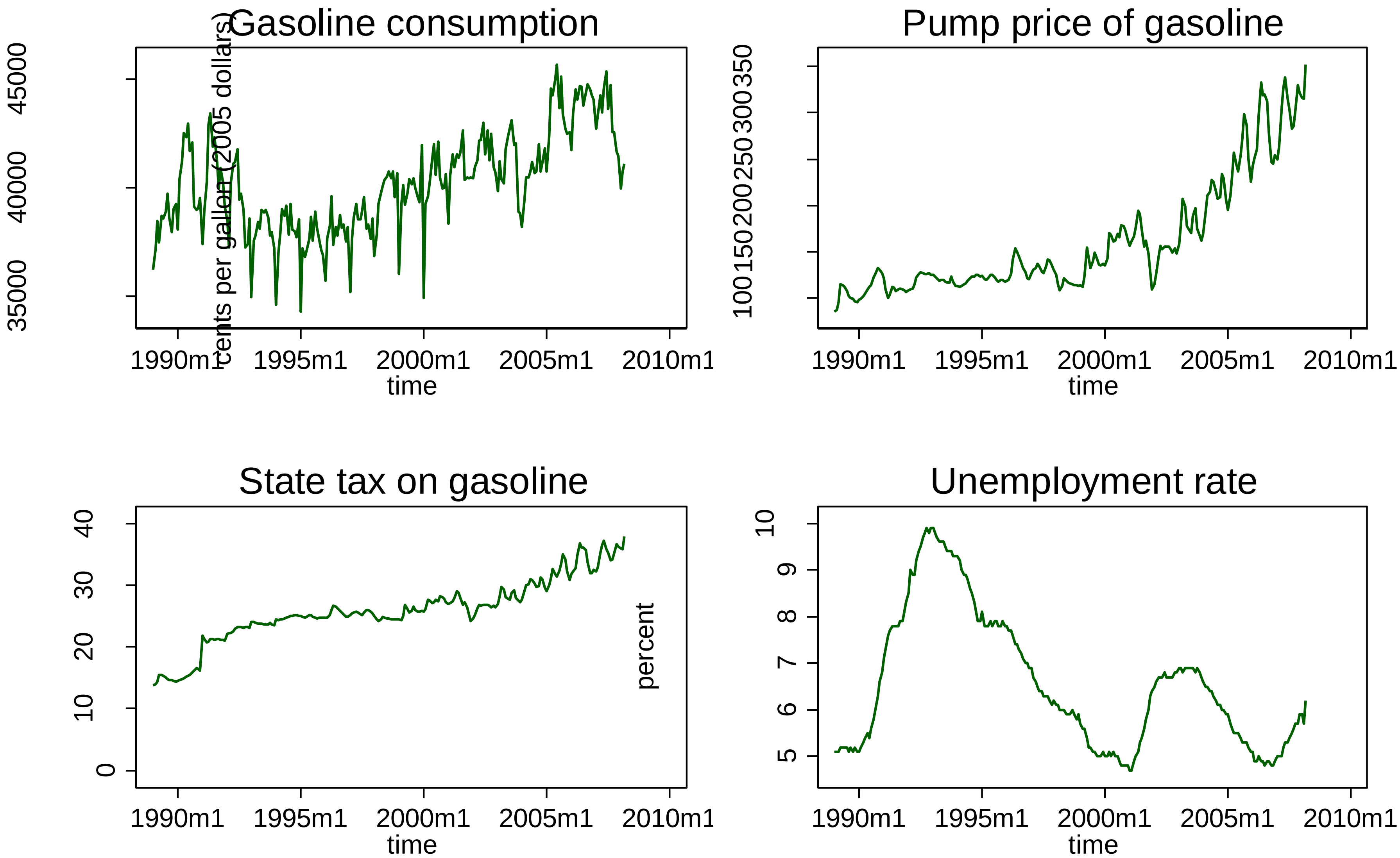


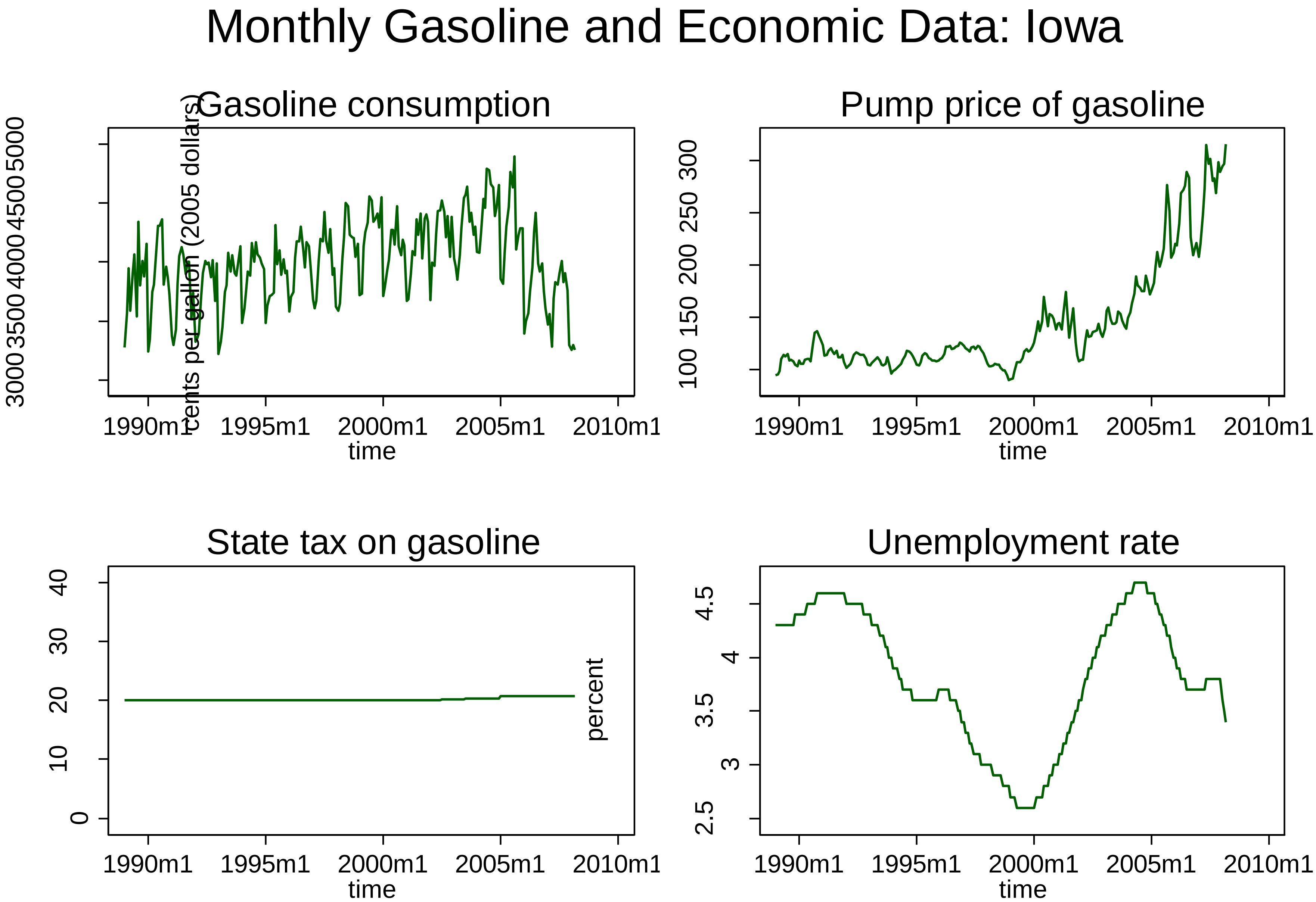
What is the price elasticity of demand for gasoline?

Data:

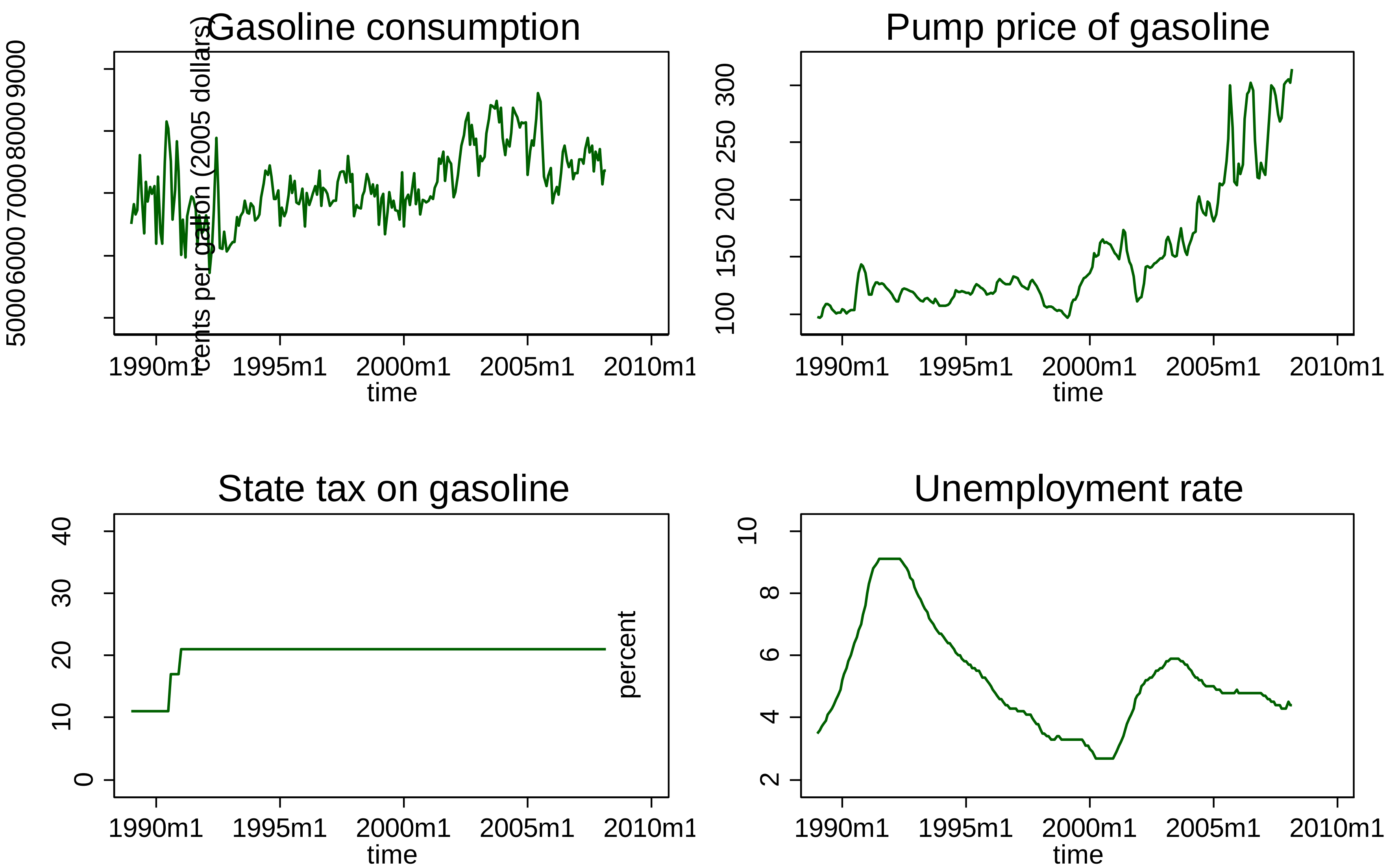
- 48 continental U.S. states, January 1989-March 2008, monthly
- volume, pump prices (nominal and real), state taxes, unemployment rates
- Source: Davis and Kilian, *J. Appl. Econometrics* (2011), augmented with unemployment rates (nicely documented replication files at http://qed.econ.queensu.ca/jae/2011-v26.7/davis-kilian/)

Monthly Gasoline and Economic Data: California

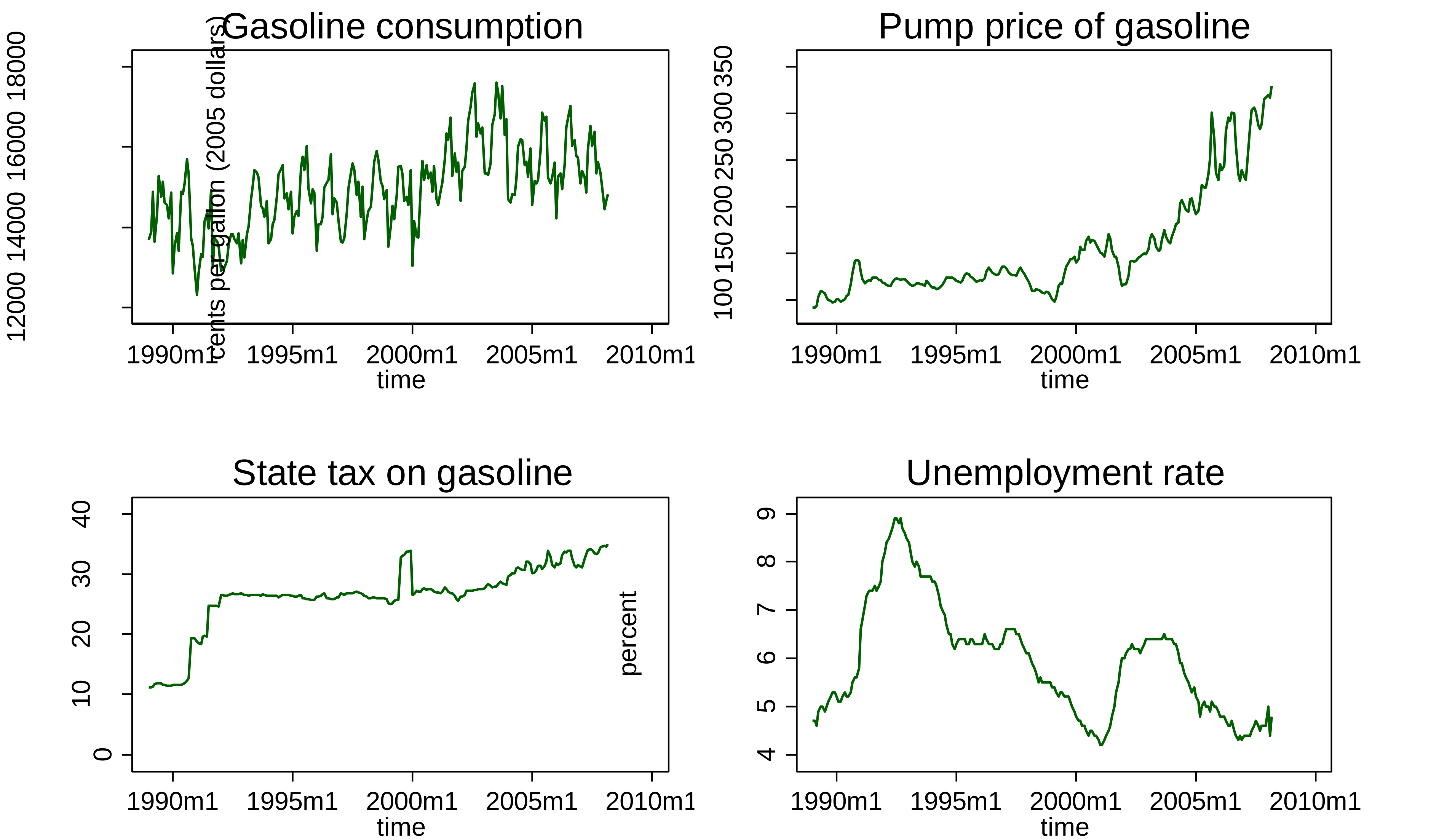



Monthly Gasoline and Economic Data: Iowa
Gasoline consumption
cents per gallon (2005 dollars)
3000 3500 4000 4500 5000
1990m1 1995m1 2000m1 2005m1 2010m1
time
Pump price of gasoline
100 150 200 250 300
1990m1 1995m1 2000m1 2005m1 2010m1
time
State tax on gasoline
0 10 20 30 40
percent
1990m1 1995m1 2000m1 2005m1 2010m1
time
Unemployment rate
2.5 3 3.5 4 4.5
1990m1 1995m1 2000m1 2005m1 2010m1
time

Monthly Gasoline and Economic Data: Massachusetts



Monthly Gasoline and Economic Data: New_York



All regressions in first differences with fixed effects (why)?

```
* (1) OLS, growth rates, HR SEs;
reg dlvolume dlrpumpprice unemployment i.statefip i.time, r;
*;
* (2) OLS, growth rates, cluster SEs;
reg dlvolume dlrpumpprice unemployment i.statefip i.time, cluster(statefip);
*;
* (3) 2SLS, contemporaneous pump price only;
ivregress 2sls dlvolume unemployment (dlrpumpprice = drstatetax_tot)
    i.statefip i.time, cluster(statefip);
*;
* (4) 2SLS, one lead and 0-2 lags of pump prices;
ivregress 2sls dlvolume unemployment (F.dlrpumpprice L(0/2).dlrpumpprice
  = F.drstatetax_tot L(0/2).drstatetax_tot) i.statefip i.time,
  cluster(statefip);
lincom F.dlrpumpprice + dlrpumpprice + L1.dlrpumpprice + L2.dlrpumpprice ;
*;
* (5) 2SLS, one lead and 0-3 lags of pump prices;
ivregress 2sls dlvolume unemployment (F.dlrpumpprice L(0/3).dlrpumpprice
  = F.drstatetax_tot L(0/3).drstatetax_tot) i.statefip i.time,
  cluster(statefip);
lincom F.dlrpumpprice + dlrpumpprice + L1.dlrpumpprice + L2.dlrpumpprice
     + L3.dlrpumpprice;
```

```
. reg dlvolume dlrpumpprice unemployment i.statefip i.time, r;

Linear regression                                      Number of obs =   11040
                                                       F(278, 10761) =   37.02
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4917
                                                       Root MSE      =  .04481

------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice | -.1960045    .019535    -10.03   0.000    -.2342967   -.1577123
unemployment | -.0009202   .0006881     -1.34   0.181    -.0022689    .0004286
             |
 ------------------------------------------------------------------------------
```

```
.  reg dlvolume dlrpumpprice unemployment i.statefip i.time, cluster(statefip);

Linear regression                                      Number of obs =   11040
                                                       F( 46,    47) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.4917
                                                       Root MSE      =  .04481

                              (Std. Err. adjusted for 48 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |  -.1960045   .0399006    -4.91   0.000    -.2762742   -.1157348
unemployment |  -.0009202   .0002402    -3.83   0.000    -.0014033    -.000437
------------------------------------------------------------------------------
```

```
. ivregress 2sls dlvolume unemployment (dlrpumpprice = drstatetax_tot)

Instrumental variables (2SLS) regression               Number of obs =   11040
                                                       Wald chi2(278)=22597.94
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.4593
                                                       Root MSE      =  .04562

                                (Std. Err. adjusted for 48 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |  -.7157622   .2239263    -3.20   0.001     -1.15465   -.2768747
unemployment |  -.0008435   .0002272    -3.71   0.000    -.0012888   -.0003983
             |
------------------------------------------------------------------------------
```

```
> ivregress 2sls dlvolume unemployment (F.dlrpumpprice L(0/2).dlrpumpprice =
F.drstatetax_tot L(0/2).drstatetax_tot)
>   i.statefip i.time, cluster(statefip);

Instrumental variables (2SLS) regression               Number of obs =   10896
                                                       Wald chi2(278)=12805.00
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.4565
                                                       Root MSE      =  .04562
                              (Std. Err. adjusted for 48 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |
         F1. |   .3718785   .1418534     2.62   0.009     .0938509    .6499061
         --. |  -.7353892    .233089    -3.15   0.002    -1.192235   -.2785432
         L1. |   .1886337   .1439397     1.31   0.190     -.093483    .4707504
         L2. |  -.1230229   .1116925    -1.10   0.271    -.3419363    .0958905
unemployment |  -.0009755   .0002183    -4.47   0.000    -.0014034   -.0005476
------------------------------------------------------------------------------
```

```
.    lincom F.dlrpumpprice + dlrpumpprice + L1.dlrpumpprice + L2.dlrpumpprice ;

 ( 1)  F.dlrpumpprice + dlrpumpprice + L.dlrpumpprice + L2.dlrpumpprice = 0

------------------------------------------------------------------------------
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.2978998   .1886253    -1.58   0.114    -.6675985    .0717989
------------------------------------------------------------------------------
```

```
. * 2SLS, one lead and 0-3 lags of pump prices;
. ivregress 2sls dlvolume unemployment (F.dlrpumpprice L(0/3).dlrpumpprice
dlrpumpprice = F.drstatetax_tot L(0/3).drstatetax_tot)
>   i.statefip i.time, cluster(statefip);

Instrumental variables (2SLS) regression              Number of obs =   10848
                                                       Wald chi2(278)=11495.52
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.4576
                                                       Root MSE      =  .04557
                              (Std. Err. adjusted for 48 clusters in statefip)
------------------------------------------------------------------------------
             |               Robust
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlrpumpprice |
         F1. |   .3724716   .1421469     2.62   0.009     .0938689    .6510744
         --. |  -.7289675   .2341491    -3.11   0.002    -1.187891   -.2700438
         L1. |    .186246   .1435427     1.30   0.194    -.0950925    .4675846
         L2. |  -.1219444   .1117365    -1.09   0.275     -.340944    .0970552
         L3. |  -.0012995   .1009509    -0.01   0.990    -.1991596    .1965605
unemployment |  -.0008956   .0002608    -3.43   0.001    -.0014068   -.0003844
------------------------------------------------------------------------------
```

```
.    lincom F.dlrpumpprice + dlrpumpprice + L1.dlrpumpprice + L2.dlrpumpprice +
L3.dlrpumpprice;

 ( 1)  F.dlrpumpprice + dlrpumpprice + L.dlrpumpprice + L2.dlrpumpprice +
L3.dlrpumpprice = 0

------------------------------------------------------------------------------
    dlvolume |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.2934937   .1789459    -1.64   0.101    -.6442212    .0572338
------------------------------------------------------------------------------
```

$-0.293 \times -0.30 = 2.8\% \times 1200$ mmt = + 105 mmt/year

Brief Review of IV Regression and Sources of Exogeneity

IV regression with one included endogenous variable *Y*, no included exogenous regressors:

$$y_t = \beta_0 + \beta_1 Y_t + u_t$$

- The problem: corr(Y,u) $\neq$ 0, possibly because of simultaneous causation, omitted variable bias, or errors in variables.
  - If corr(Y,u) $\neq$ 0 then OLS is biased and inconsistent

- Terminology: endogeneity and exogeneity
  - An ***endogenous*** variable is one that is correlated with u
  - An ***exogenous*** variable is one that is uncorrelated with u

The IV Estimator, one Y and one Z

$$y_t = \beta_0 + \beta_1 Y_t + u_t$$

Two conditions for a valid instrument

1. ***Instrument relevance***: $\text{corr}(Z,Y) \neq 0$

2. ***Instrument exogeneity***: $\text{corr}(Z,u) = 0$

By instrument exogeneity,

$$\text{cov}(u,Z) = \text{cov}(y - \beta_0 - \beta_1 Y, Z) = 0$$

so

$$\text{cov}(y,Z_i) = \beta_1 \text{cov}(Y,Z)$$

By instrument relevance,

$$\beta_1 = \frac{\text{cov}(y,Z)}{\text{cov}(Y,Z)}$$

The IV (2SLS) estimator:

$$\hat{\beta}_1^{IV} = \frac{s_{yZ}}{s_{YZ}}$$

Multiple instruments: $\boldsymbol{Z}_i$ is $k \times 1$

For all vectors $\boldsymbol{a}$, by *instrument exogeneity*,

$$\text{cov}(u,\boldsymbol{a}'\boldsymbol{Z}) = \text{cov}(y - \beta_0 - \beta_1 Y,\boldsymbol{a}'\boldsymbol{Z}) = 0$$

or

$$\text{cov}(y,\boldsymbol{a}'\boldsymbol{Z}) = \text{cov}(\beta_1 Y,\boldsymbol{a}'\boldsymbol{Z}) = \beta_1\text{cov}(Y,\boldsymbol{a}'\boldsymbol{Z})$$

By *instrument relevance*, $\beta_1 = \dfrac{\text{cov}(y,\boldsymbol{a}'\boldsymbol{Z})}{\text{cov}(Y,\boldsymbol{a}'\boldsymbol{Z})}$

Which choice of $\boldsymbol{a}$ is the best?

- when $k > 1$, different IV estimators are available
- What is the value of $\boldsymbol{a}$ that results in the most efficient (lowest variance) estimator asymptotically?
- Result is TSLS (or others! LIML, k-class,…)

Two Stage Least Squares (TSLS)

Suppose you have k valid instruments, $\boldsymbol{Z}$.

Stage 1:Regress Y on $\boldsymbol{Z}$, obtain the predicted values $\hat{Y}$

Stage 2:Regress y on $\hat{Y}$; the coefficient on $\hat{Y}$ is

the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

- Intuitively, the first stage isolates part of the variation in Y that is uncorrelated with u
- In terms of the previous slide, $\boldsymbol{a}'\boldsymbol{Z}$ is constructed to be the linear combination of instruments that is the predicted value of Y
- This is the linear combination that maximizes the sample correlation between Y and $\boldsymbol{a}'\boldsymbol{Z}$.

The General IV Regression Model

Extension to:

- multiple endogenous regressors ($Y_1,\ldots,Y_m$)
- multiple instrumental variables ($Z_1,\ldots,Z_k$)
- multiple included exogenous variables ($W_1,\ldots,W_r$)

Why use multiple instruments?

- More relevant instruments means more variation in $\hat{Y}$ which means smaller variance

Why include the W's?

- For instrument exogeneity, you need corr(u,Z) = 0. The definition of u depends on what variables are included – u might only be uncorrelated with Z, conditional on the W's (you still need control variables!)

Terminology: identification & overidentification

- In general, a parameter is ***identified*** if different values of the parameter produce different distributions of the data.
- In IV regression, the coefficients $\beta_1,\ldots, \beta_m$ are:
  - ***exactly identified*** if #IVs = $k = m$.
  - ***overidentified*** if $k > m$

    Then there are more than enough instruments – you can test the validity of redundant instruments (more on this shortly)
  - ***underidentified*** if $k < m$

    Then there are too few instruments – you need more!

More terminology: strong and weak instruments

- Strong instruments: partial correlation corr(Z,Y|W) is "large"
- Weak instruments: partial correlation corr(Z,Y|W) is "small"

The IV regression model in matrix form

$$\boldsymbol{y} = \boldsymbol{Y\beta} + \boldsymbol{W\gamma} + \boldsymbol{U}$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{Y}$ is $n \times m$, and $\boldsymbol{W}$ is $n \times r$ and the $n \times k$ matrix of k instruments is $\boldsymbol{Z}$

TSLS in general IV regression

Stage 1: Regress $\boldsymbol{Y}$ on $\boldsymbol{Z}$ and $\boldsymbol{W}$ to obtain the predicted values $\hat{\boldsymbol{Y}}$

Stage 2: Regress $\boldsymbol{y}$ on $\hat{\boldsymbol{Y}}$ and $\boldsymbol{W}$; the coefficient vector on $\hat{\boldsymbol{Y}}$ is the TSLS estimator, $\hat{\boldsymbol{\beta}}^{TSLS}$

Conventional asymptotic results for the TSLS estimator:

- If the instruments are strong and exogenous, plus some moments exist, then TSLS is consistent ($\hat{\beta}_1^{TSLS} \xrightarrow{p} \beta_1$)
- If the data are i.i.d. (e.g. cross-sectional) ***and homoskedastic****, then TSLS estimator is asymptotically normal:

$$\sqrt{n}\,(\hat{\beta}_1^{TSLS} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma}^{TSLS})$$

where

$$\boldsymbol{\Sigma}^{TSLS} = \left(\boldsymbol{Q}_{\boldsymbol{YZ}}\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{Q}_{\boldsymbol{ZY}}\right)^{-1}\sigma_u^2$$

where $\boldsymbol{Q}_{\boldsymbol{YZ}} = E(Y_t Z_t')$, etc.

***Homoskedasticity**: $E(u_t^2|Z_t) = \sigma_u^2$ = constant

$$\sqrt{n}\,(\hat{\beta}_1^{TSLS} - \boldsymbol{\beta}) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma}^{TSLS})$$

$$\boldsymbol{\Sigma}^{TSLS} = \left(\boldsymbol{Q}_{\boldsymbol{YZ}}\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{Q}_{\boldsymbol{ZY}}\right)^{-1}\sigma_u^2$$

- Note that $\boldsymbol{Q}_{\boldsymbol{YZ}}\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{Q}_{\boldsymbol{ZY}}$ is the (population) variance of the predicted value of Y from the first stage regression – so the higher the first-stage R^2, the smaller the TSLS variance
- Because of the asymptotic normal distribution, inference is conventional – confidence intervals are ± 1.96 standard errors, F-tests are justified, etc.
- The linear combination of $\boldsymbol{Z}$ ($\boldsymbol{a}'\boldsymbol{Z}$ in previous slide) estimated in the first stage is the "right" one –TSLS is asymptotically efficient (under strong instruments)
- Heteroskedasticity:
  - To guard against heteroskedasticity in TSLS, use "heteroskedasticity-robust" (HR) standard errors
  - Under heteroskedasticity, IV is no longer efficient – the efficient estimator is the efficient GMM estimator (more on this shortly)

Checking Overidentifying Restrictions: the *J*-test

Consider the simplest case:

$$y_t = \beta_0 + \beta_1 Y_t + u_t,$$

- Suppose there are two valid instruments: Z_{1t}, Z_{2t}
- Then you could compute two separate TSLS estimates.
- Intuitively, if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
- The *J*-test of overidentifying restrictions makes this comparison in a statistically precise way.
- This can only be done if #Z's > #Y's (overidentified).

Sources of Exogeneity (where do instruments come from?)

General comments

The hard part of IV analysis is finding valid instruments

- Traditional (simultaneous equation) method: "variables that are excluded from the equation of interest and enter another equation in the system"
    - e.g. supply shifters that do not affect demand
- More general (contemporary) view: look for exogenous variation (Z) that is "as if" randomly assigned (does not directly affect y) but affects Y.
- Formally these are the same but they suggest different empirical strategies.

- Stinebrinckner and Stinebrinckner (2008) is a great example for teaching…
  - Individual student data, 210 (first semester freshman wave of a multiyear panel data set), Berea College (Kentucky), 2001
  - Y = first-semester GPA
  - X = average study hours per day (time use survey)
  - Z = 1 if roommate brought video game, = 0 otherwise

Table 2
First Stage Regressions
The effect of instruments (and other variables) on study hours

Independent Variable	**estimate (std error) n=210**	**estimate (std error) n=176**
INSTRUMENTS		
video game TREATMENT	-.668 (.252)**	-.658 (.268)**
RSTUDYHS		.028 (.013)**
REXSTUDY		.049 (.074)
OTHER VARIABLES		
MALE	-.155 (.244)	-.204 (.263)
BLACK	.417 (.341)	.549 (.350)
ACT	-.019 (.036)	-.016 (.038)
$MAJOR_1$	1.423 (.828)*	1.230 (.816)
$MAJOR_2$	1.421 (.783)*	1.015 (.772)
$MAJOR_3$	1.120 (.811)	.891 (.789)
$MAJOR_4$	1.637 (.784)**	1.410 (.782)*
$MAJOR_5$	1.575 (.776)**	1.375 (.762)*
$MAJOR_6$	1.777 (.806)**	1.604 (.797)**
$MAJOR_7$	2.128 (.836)**	2.006 (.827)**
HEALTH_BAD	.209 (.463)	.221 (.478)
HEALTH_EXC	.095 (.241)	.010 (.258)
	R^2=.092	R^2=.179

Note: The first column uses the entire sample of individuals with randomly assigned roommates. The second column which takes advantage of roommates' reports of how many hours they studied per week in high school (RSTUDYHS) and how many hours they expect to study per day in college (REXSTUDY) uses the subset of these students whose roommates are also members of the sample and are not missing values of RSTUDYHS and REXSTUDY.
*significant at .10
**significant at .05

Table 5
Causal impact in reduced form: The direct effect of treatment on first semester grades

Independent Variable	Dependent Variable GPA first semester grades estimate (std error)
CONSTANT	.793 (.398)**
TREATMENT	-.241 (.089)**
MALE	-.079 (.086)
BLACK	-.209 (.120)*
ACT	.062 (.012)**
$MAJOR_1$	.906 (.293)**
$MAJOR_2$	.868 (.277)**
$MAJOR_3$	.739 (.287)**
$MAJOR_4$	.889 (.277)**
$MAJOR_5$	.741 (.274)**
$MAJOR_6$	.731 (.285)**
$MAJOR_7$	1.002 (.295)**
HEALTH_BAD	.045 (.164)
HEALTH_EXC	.149 (.085)*
	R^2=.289

*significant at .10
**significant at .05

Table 4
Estimates of the effect of studying on grade performance:
Ordinary Least Squares, Instrumental Variables, Fixed Effects

Independent Variable	OLS n=210 estimate (std. error)	IV instrument: video game TREATMENT n=210 estimate (std. error)	IV instruments: video game TREATMENT, RSTUDYHS, REXSTUDY n=176 estimate (std. error)	Fixed Effects n=210 estimate (std. error)
CONSTANT	.719 (.408)*	-.073 (.709)	-.062 (.638)	-.050 (.047)
STUDY	.038 (.025)	.360 (.183)**	.291 (.121)**	-.043 (.027)*
SEX	-.132 (.084)	-.023 (.129)	-.010 (.126)	
BLACK	-.220 (.122)*	-.356 (.183)*	-.334 (.176)*	
ACT	.062 (.013)**	.069 (.018)**	.072 (.018)**	
$MAJOR_1$	.834 (.298)**	.393 (.474)	.576 (.410)	
$MAJOR_2$	.793 (.282)**	.356 (.454)	.475 (.380)	
$MAJOR_3$	.725 (.292)**	.335 (.452)	.467 (.389)	
$MAJOR_4$	.796 (.283)**	.298 (.474)	.411 (.403)	
$MAJOR_5$	.643(.280)**	.174 (.462)	.366 (.389)	
$MAJOR_6$	.664(.292)**	.091 (.510)	.143 (.427)	
$MAJOR_7$	.901 (.304)**	.235 (.555)	.243 (.468)	
HEALTH_BAD	.019(.166)	-.029 (.226)	-.020 (.219)	
HEALTH_EXC	.127 (.086)	.115 (.117)	.158 (.118)	
	R^2=.273			

1) What is weak identification, and why do we care?

1a) Four examples

Example #1: Philip G. Wright and the supply and demand for flaxseed

$$\ln(Q_i^{flaxseed}) = \beta_0 + \beta_1 \ln(P_i^{flaxseed}) + u_i$$

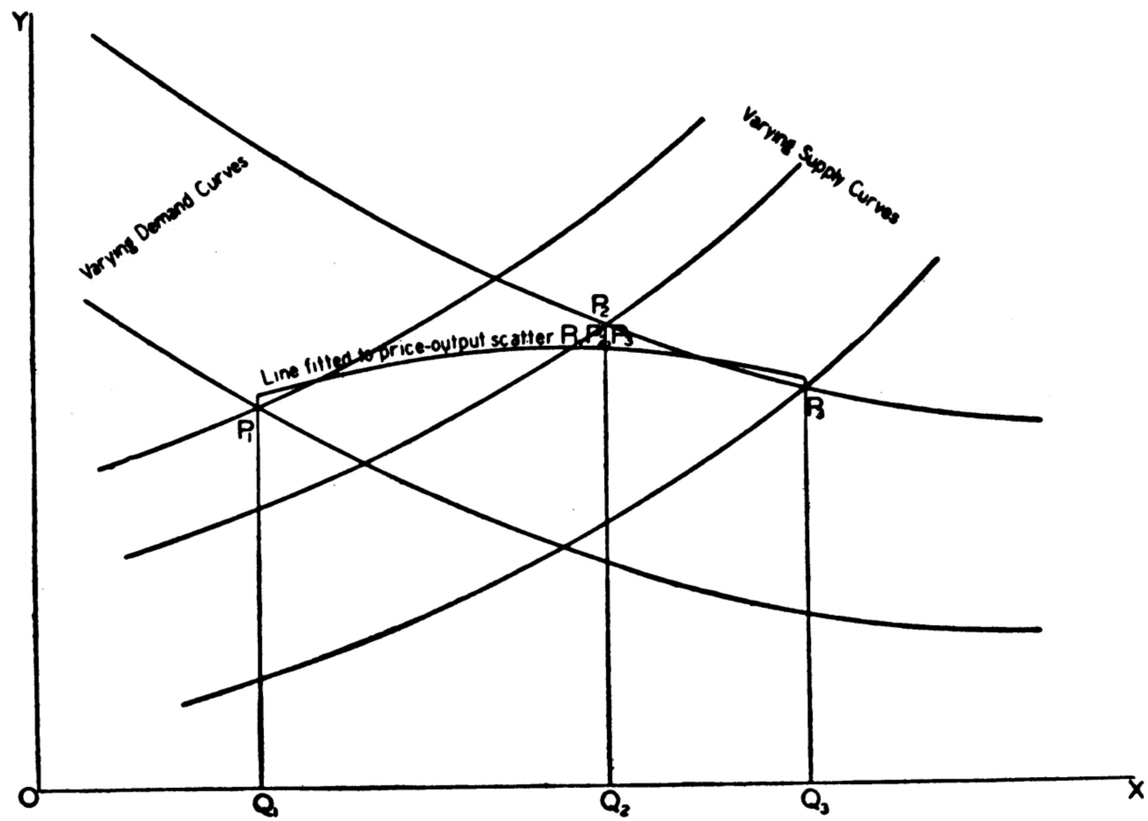
The first application of IV regression was to estimate the supply elasticity of flaxseed.

Flaxseed was used around the turn of the century for production of linseed oil – used (pre-petroleum derivatives) as a paint binder or wood finish.

Philip G. Wright (1928), "The Tariff on Animal and Vegetable Oils," App. B.

Figure 4, p. 296, from P.G. Wright, Appendix B (1928):

FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



Philip Wright (1861-1934)
Economist, teacher, poet
MA Harvard, Econ, 1887
Lecturer, Harvard, 1913-1917

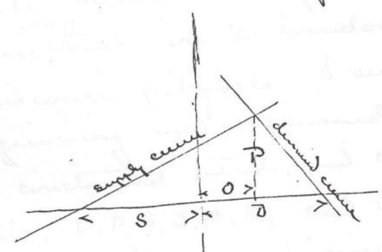
Sewall Wright (1889-1988)
genetic statistician
ScD Harvard, Biology, 1915
Prof., U. Chicago, 1930-1954

The Wrights' letters, December 1925 - March 1926

March 4, 1926.

Dear Bewall:

It may interest you to see a very simple geometric demonstration which I have worked out for your method of estimating supply and demand curves without reference to the theory of path coefficients.



P is price, O is output, S is supply under mean price, and D is demand under mean price, all expressed as percentage deviations from mean.

Then $e = \frac{O-S}{P}$

A is factor uncorrelated with S

$$eP_1 = O_1 - S_1$$
$$eA_1P_1 = A_1O_1 - A_1S_1$$
$$eA_2P_2 = A_2O_2 - A_2S_2$$
$$eA_3P_3 = A_3O_3 - A_3S_3$$
$$\text{-- -- --}$$
$$e\Sigma AP = \Sigma AO - \Sigma AS$$
$$= \Sigma AO \text{ [since } A \text{ is uncorrelated with } S\text{]}$$

$$\therefore e = \frac{\Sigma AO}{\Sigma AP}$$

$\eta = \frac{O-D}{P}$ [since η is negative]

B is factor uncorrelated with D

$$\eta P = O - D$$
$$\eta B_1P_1 = B_1O_1 - B_1D_1$$
$$\eta B_2P_2 = B_2O_2 - B_2D_2$$
$$\eta B_3P_3 = B_3O_3 - B_3D_3$$
$$\text{-- -- --}$$
$$\eta\Sigma BP = \Sigma BO - \Sigma BD$$
$$= \Sigma BO \text{ [and } B \text{ is uncorrelated with } D\text{]}$$

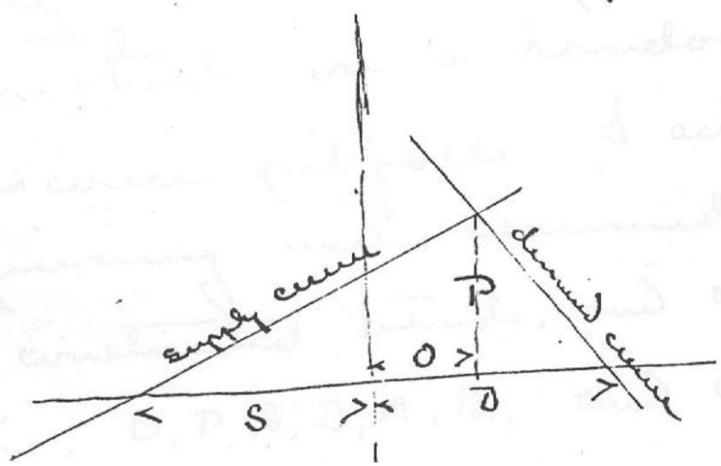
$$\therefore \eta = \frac{\Sigma BO}{\Sigma BP}$$

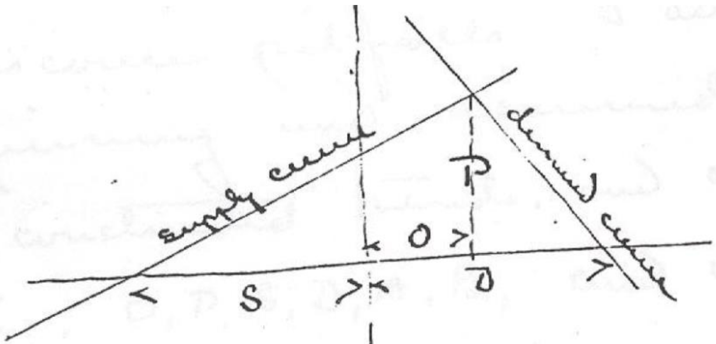
MARCH 4 1926 - 1

March 4, 1926.

Dear Sewall:

It may interest you to see a very simple geometric demonstration which I have worked out for your method of estimating supply and demand curves without reference to the theory of path coefficients.





P is price, O is output, S is supply under mean price, and D is demand under mean price, all expressed as percentage deviations from mean.

Then $e = \dfrac{O - S}{P}$

$\eta = \dfrac{O - D}{P}$ [since η is negative]

A is factor uncorrelated with S

B is factor uncorrelated with D

$$\begin{aligned}
eP_1 &= O_1 - S_1 \\
eA_1P_1 &= A_1O_1 - A_1S_1 \\
eA_2P_2 &= A_2O_2 - A_2S_2 \\
eA_3P_3 &= A_3O_3 - A_3S_3 \\
&\;\;-\;-\;- \\
e\Sigma AP &= \Sigma AO - \Sigma AS \\
&= \Sigma AO \text{ [since A is uncorrelated with S]}
\end{aligned}$$

$$\begin{aligned}
\eta P &= O - D \\
\eta B_1P_1 &= B_1O_1 - B_1D_1 \\
\eta B_2P_2 &= B_2O_2 - B_2D_2 \\
\eta B_3P_3 &= B_3O_3 - B_3D_3 \\
&\;\;-\;-\;- \\
\eta \Sigma BP &= \Sigma BO - \Sigma BD \\
&= \Sigma BO \text{ [and B is uncorrelated with D]}
\end{aligned}$$

Then $e = \dfrac{O - S}{P}$

A is factor uncorrelated with S

$$\underbrace{eP_1 = O_1 - S_1}$$

$$eA_1P_1 = A_1O_1 - A_1S_1$$

$$eA_2P_2 = A_2O_2 - A_2S_2$$

$$eA_3P_3 = A_3O_3 - A_3S_3$$

$$\text{— . . . —} \quad \text{— — —}$$

$$e\Sigma AP = \Sigma AO - \Sigma AS$$

$$= \Sigma AO \quad [\text{since A is uncorrelated with S}]$$

$$\therefore e = \frac{\Sigma AO}{\Sigma AP}$$

March 15, 1926

Dear Sewall:

I have just received your letters of the 11th and 13th.
and have read them carefully. I think now that the
theory of the method is pretty clear in my mind
though I shall need to give a little more study to
the problem of long time elasticity of output. Just now,
however, a difficulty more fundamental than any that
has as yet arisen has occurred to me. I am somewha
haunted by the suspicion that we may be arguin
in a circle. Suppose we take any price-output scatter whate
the respective

of the economic "universe".

The problem, therefore, boils down to this: In the case of any specific commodity is it possible to find factors which have such distinct causal relations with output or demand conditions that the values of <u>e</u> and η computed from them can be accepted with any confidence as having any relation with actuality. Such factors, I fear, especially in the case of demand conditions, it is not easy to find. I have been experimenting with flaxseed and so far have arrived at no results in which I can place much confidence.

The most likely data which I have been able to secure

MARCH 15 1926

Notes: *e* = supply elasticity, η = demand elasticity; by "output" in this paragraph PGW means supply.

Supply)	Real prices 1/ [Money price ÷ index no.]	Output 2/ (million bu.)	B Acreage 2/ (millions)	Yield	B Rainfall 3/ (inches)	B Ratio value Flaxseed to Spring wheat 4/	A Building permits (thousands)
1903	126	27.3	3.23	8.4	3.40	93	128
4	153	23.4	2.26	10.3	2.19	75	140
5	123	28.5	2.53	11.2	4.27	95	186
6	126	25.6	2.51	10.2	3.30	93	181
7	133	25.9	2.86	9.0	2.66	119	187
8	157	25.8	2.68	9.6	3.38	76	175
9	204	19.7	2.08	9.5	3.10	95	213
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

…

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
4	171	31.7	3.47	[illegible]	3.03	188	
5							

1/ Average for crop year beginning Sep. 1. The Minneapolis price was divided by wholesale price-index all commodities to get "real price".

2/ Figures are for calendar year.

3/ Figures are a simple average for rainfall (May, June, and July) for Duluth, Minn., Bismarck, N.D., Pierre, S.D.

4/ The ratios of the values of flaxseed per acre to spring wheat per acre lagged 1 year. i.e. the ratios for the year shown in the table are really the ratios for the preceding year

Acreage, rainfall, and ratios of value I assumed might be used as factor B. Building permits as factor A. I have not yet worked with the B factor, but A gave a very unsatisfactory result. I fitted price, output, and permits scatters to straight line trends and computed percentage deviations. The form $\frac{\Sigma AO}{\Sigma AP}$ gave $-.88$ as the value of e – a result obviously absurd.

I have not, as I said, tried the B factor. I think it not unlikely that they might give values of η ~~that~~ looked reasonable and possibly values ~~that~~ would approximate one another. I am, however, chiefly interested in finding the value of e. The only factor which I have thought of which would on a priori grounds affect demand conditions and not ~~out~~output conditions (the same year) was building permits as affecting primarily the demand for linseed oil and hence flaxseed. "Consumption of Linseed Oil" would be more direct but data are available only from 1912 to 1924 inclusive with 1913 and 1915 missing. There is no substitute for linseed oil of sufficient importance to make its ... business condition substitute for [illegible] 10

index of general business conditions are defensible - Possibly some might give results, but such a factor seems rather remote and I don't know what "general business factor" would be most ap- propriate. Again, as you notice I used a straight line trend The fluctuations are so violent that a curve fitted by eye might be preferable. But could more confidence be placed in the results obtained from such a curve than from a general estimate of elasticity based on a price-output scatter?

We heard from Quincy that your Louise had been having chicken pox and you feared the children might also have it. We were very sorry to hear this and extend our sympathies. We wish you were not so far off as to prevent us from extending anything more tangible than sympathies. That was one of the advantages of being in Washington. We could help each other out in time of stress. What with all the fevers and contagious diseases Chicago is cer

MARCH 15 1936

- Flaxseed was grown mainly in the upper Midwest (can plant in April and harvest in August)
- PGW data:
  - Prices are Minneapolis fall prices
  - Rainfall is average in Bismark ND, Duluth MN, Minneapolis MN
  - Data are annual, 1904-1923
  - PGW deviated all data from a linear trend
  - $Y = Q$ (% deviation from trend)
  - $X = P$ (% deviation from trend)
  - Z = building permits (deviation from trend)
    - Exogeneity: corr(u_i, $Building\ Permits_i$) = 0?
    - Relevance: corr(P_i, $Building\ Permits_i$) $\neq$0?

Checking for Instrument Relevance: Wright's Flaxseed Data

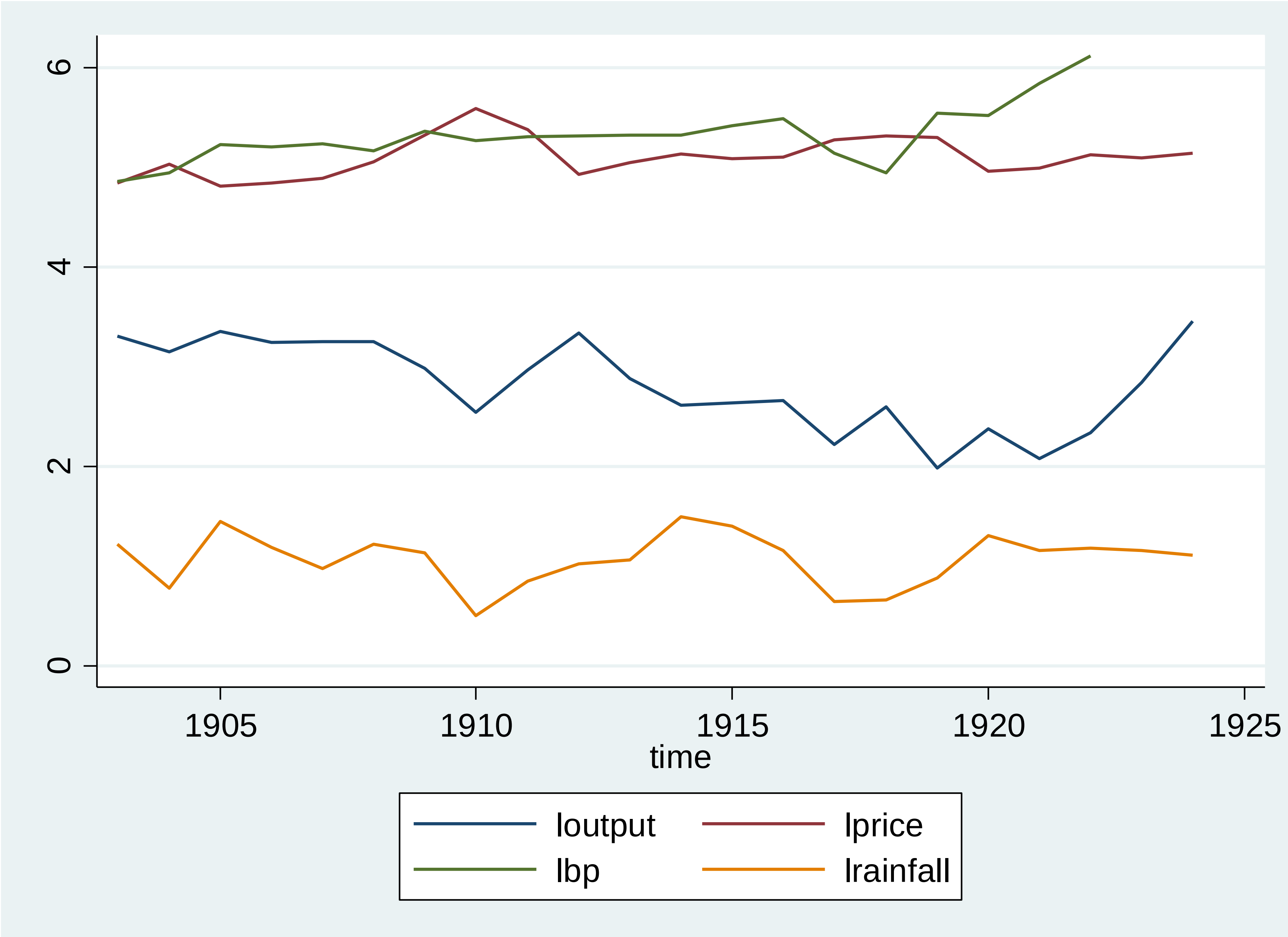
What went wrong with PGW's supply elasticity regression?

Z = deviation of building permits from trend = *bp_dev*

```
. ivregress 2sls output_dev (price_dev = bp_dev), first;

First-stage regressions
-----------------------
                                                  Number of obs   =         20
                                                  F(   1,     18) =       1.25
                                                  Prob > F        =     0.2783
                                                  R-squared       =     0.0649
                                                  Adj R-squared   =     0.0130
                                                  Root MSE        =     0.2168

------------------------------------------------------------------------------
   price_dev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      bp_dev |  -.2732793   .2444394    -1.12   0.278    -.7868275    .2402689
       _cons |   .0077936   .0484871     0.16   0.874     -.094074    .1096612
------------------------------------------------------------------------------
```

```
Instrumental variables (2SLS) regression               Number of obs =       20
                                                       Wald chi2(1)  =     0.72
                                                       Prob > chi2   =   0.3974
                                                       R-squared     =   0.1641
                                                       Root MSE      =   .21633

------------------------------------------------------------------------------
  output_dev |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   price_dev |  -.7553123   .8925526    -0.85   0.397    -2.504683    .9940587
       _cons |  -.0906035   .0487388    -1.86   0.063    -.1861299    .0049228
------------------------------------------------------------------------------
Instrumented:  price_dev
Instruments:   bp_dev
```


Price and building permits, deviated from trend
.6
.4
.2
0
-.2
-.5
0
.5
bp_dev



0
2
4
6
1905
1910
1915
1920
1925
time
loutput
lprice
lbp
lrainfall

Example #2 (cross-section IV): Angrist-Kreuger (1991)

What are the returns to education?

y = log(earnings)

Y = years of education

Z = quarter of birth; k = #IVs = 3 binary variables or up to 178

(interacted with year-of-birth, state-of-birth)

n = 329,509

A-K results: $\hat{\beta}^{TSLS}$ = .081 (SE = .011)

Then came Bound, Jaeger, and Baker (1995)…

$\Rightarrow$ The problem is that $\boldsymbol{Z}$ (once you include all the interactions) **is weakly correlated with $\boldsymbol{Y}$**

Example #3 (linear GMM): New Keynesian Phillips Curve

e.g. Gali and Gertler (1999), where x_t = labor share; see survey by Mavroeidis, Plagborg-Møller, and Stock (*JEL*, 2014). Hybrid NKPC with shock η_t:

$$\pi_t = \lambda x_t + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} + \eta_t$$

Rational expectations: $E_{t-1}(\pi_t - \lambda x_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1}) = 0$

GMM moment condition: $E[(\pi_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1} - \lambda x_t)Z_t] = 0$

Instruments: $Z_t = \{\pi_{t-1}, x_{t-1}, \pi_{t-2}, x_{t-2},\ldots\}$ (GG: 23 total)

Issues:

- Z_t needs to predict π_{t+1} – beyond π_{t-1} (included regressor)
- But predicting inflation is really hard! Atkeson-Ohanian (2001), Stock and Watson (2007), recent literature on backwards-looking Phillips curve

Example #4 (nonlinear GMM): Estimating the elasticity of intertemporal substitution, nonlinear Euler equation

With CRRA preferences, in standard GMM notation,

$$h(Y_t,\theta) = \delta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} \overset{G\times 1}{R_{t+1}} - \iota_G$$

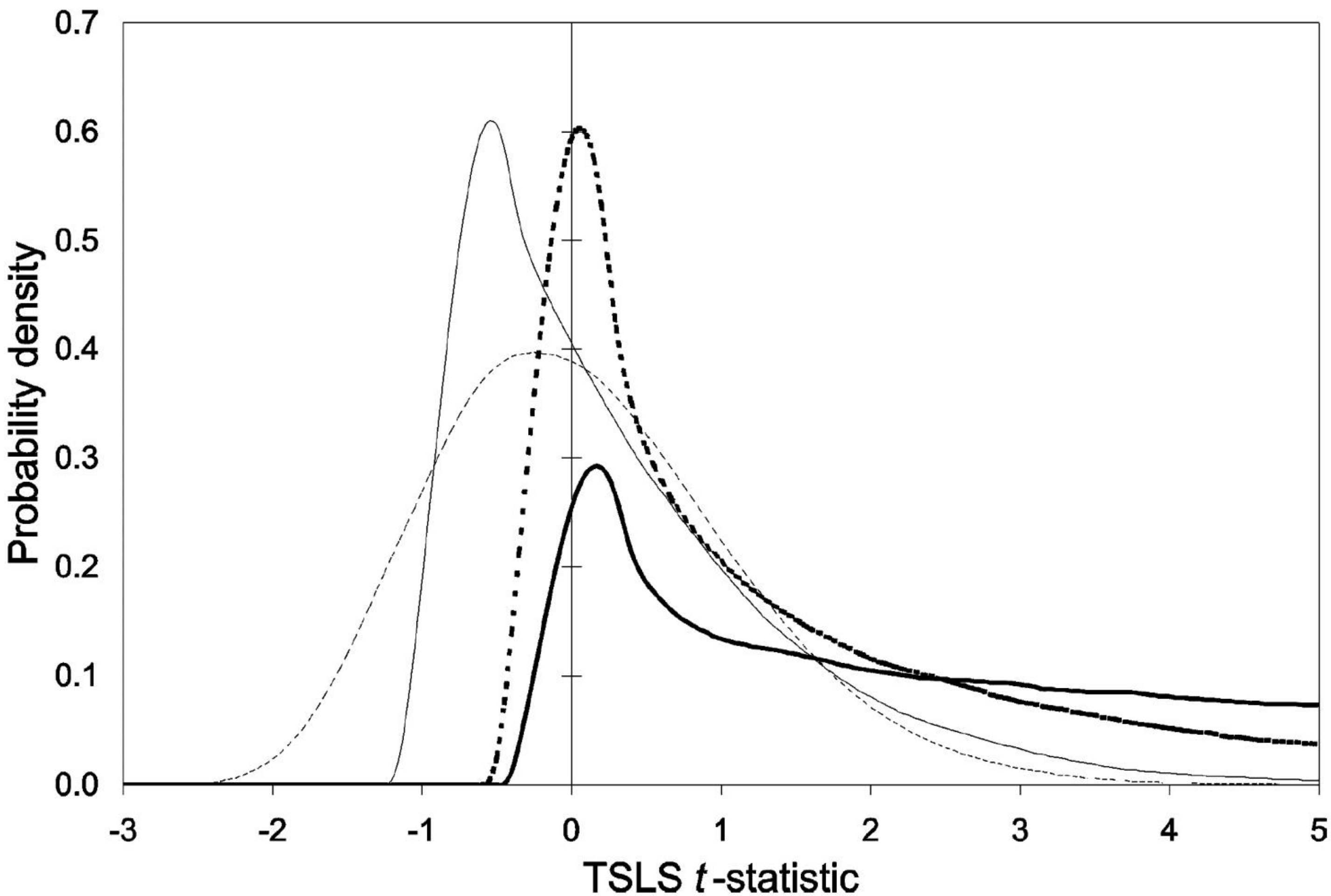
where R_{t+1} is a $G\times 1$ vector of asset returns and ι_G is the G-vector of 1's.
GMM moment conditions (Hansen-Singleton (1982)):

$$\mathrm{E}[h(Y_t,\theta) \otimes Z_t] = 0 \text{ where } Z_t = \Delta c_t,\ R_t, \text{ etc.}$$

$\Rightarrow$ Z_t must predict consumption growth (and stock returns) using past data

How important are these deviations from normality quantitatively?

Nelson-Startz (1990a,b) plots of the distribution of the TSLS *t*-statistic:



Dark line = irrelevant instruments; dashed light line = strong instruments; intermediate cases: weak instruments

Working definition of weak identification

We will say that θ is ***weakly identified*** if the distributions of GMM or IV estimators and test statistics are not well approximated by their standard asymptotic normal or chi-squared limits because of limited information in the data.

- Departures from standard asymptotics are what matters in practice
- The source of the failures is limited information, not (for example) heavy tailed distributions, near-unit roots, unmodeled breaks, etc.
- We will focus on large samples - the source of the failure is not small-sample problems in a conventional sense. In fact most available tools for weak instruments have large-sample justifications. This is not a theory of finite sample inference (although it is closely related, at least in the linear model.)
- Throughout, we assume instrument exogeneity – weak identification is about instrument relevance, not instrument exogeneity

Some special cases:

- Special cases we will come back to
  - θ is unidentified
  - Some elements of θ are strongly identified, some are weakly identified
- A special cases we won't come back to
  - θ is ***partially identified***, i.e. some elements of θ are identified and the rest are not identified
- Not a special case
  - θ is ***set identified***, i.e. the true value of θ is identified only up to a set within Θ. Weak identification and set identification could be married in theory, but they haven't been.
  - Inference when there is set identification is a hot topic in econometric theory. Set identification will come up in SVARs.

Additional preparatory comments

- The literature has differing degrees of maturity and completion:
  - Testing and confidence intervals in classical (cross-sectional) IV regression model with a single included endogenous regressor: a mature area in which the first order problems are solved
  - Estimation in general nonlinear GMM – little is known
- These lectures focus on:
  - explaining how weak identification arises at a general level;
  - providing practical tools and advice ("state of the art")
  - providing references to the most recent literature (untested methods)
- Literature reviews:
  - Mikusheva (2013) – focuses on linear IV, comprehensive
  - Andrews and Stock (2007) (comprehensive but technical)

Outline

1) What is weak identification, and why do we care?
2) **Classical IV regression I: Setup and asymptotics**
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

2) Classical IV regression I: Setup and asymptotics

Classical IV regression model & notation

Equation of interest: $y_t = Y_t\beta + u_t$, $m = \dim(Y_t)$

k exogenous instruments Z_t: $E(u_t Z_t) = 0$, $k = \dim(Z_t)$

Auxiliary equations: $Y_t = \Pi' Z_t + v_t$, $\text{corr}(u_t, v_t) = \rho$ (vector)

Sampling assumption (y_t, Y_t, Z_t) are i.i.d.

Equations in matrix form:

$$\mathbf{y} = \mathbf{Y}\beta + \mathbf{u}$$

$$\mathbf{Y} = \mathbf{Z}\Pi + \mathbf{v}$$

Comments:

- We assume throughout the instrument is exogenous ($E(u_t Z_t) = 0$)
- Included exogenous regressors have been omitted without loss of generality
- Auxiliary equation is just the projection of Y on Z

IV regression with one Y and a single irrelevant instrument

$$\hat{\beta}^{TSLS} = \frac{\mathbf{Z}'\mathbf{y}}{\mathbf{Z}'\mathbf{Y}} = \frac{\mathbf{Z}'(\mathbf{Y}\beta + \mathbf{u})}{\mathbf{Z}'\mathbf{Y}} = \beta + \frac{\mathbf{Z}'\mathbf{u}}{\mathbf{Z}'\mathbf{Y}}$$

If Z is irrelevant (as in Bound et. al. (1995)), then $\mathbf{Y} = \mathbf{Z}\Pi + \mathbf{v} = \mathbf{v}$, so

$$\hat{\beta}^{TSLS} - \beta = \frac{\mathbf{Z}'\mathbf{u}}{\mathbf{Z}'\mathbf{v}} = \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t u_t}{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t v_t} \xrightarrow{d} \frac{z_u}{z_v}, \text{ where } \begin{pmatrix} z_u \\ z_v \end{pmatrix} \sim N\left(0, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}\right)$$

Comments:

- $\hat{\beta}^{TSLS}$ isn't consistent (this should make sense)
- Distribution of $\hat{\beta}^{TSLS}$ is Cauchy-like (ratio of correlated normals)

- The distribution of $\hat{\beta}^{TSLS}$ is a ***mixture of normals with nonzero mean***: write $z_u = \delta z_v + \eta$, $\eta \perp z_v$ where $\delta = \sigma_{uv}/\sigma_v^2$. Then

$$\frac{z_u}{z_v} = \frac{\delta z_v + \eta}{z_v} = \delta + \frac{\eta}{z_v}, \text{ and } \frac{\eta}{z_v}|z_v \sim N(0, \frac{\sigma_\eta^2}{z_v^2})$$

so **the asymptotic distribution** of $\hat{\beta}^{TSLS} - \beta_0$ is the mixture of normals,

$$\hat{\beta}^{TSLS} - (\beta_0 + \delta) \xrightarrow{d} \int N(0, \frac{\sigma_\eta^2}{z_v^2}) f_{z_v}(z_v) dz_v \text{ (1 irrelevant instrument)}$$

- heavy tails (mixture is based on inverse chi-squared)
- center of distribution of $\hat{\beta}^{TSLS}$ is $\beta_0 + \delta$. But

$$\hat{\beta}^{OLS} - \beta_0 = \frac{\mathbf{Y}'\mathbf{u}/n}{\mathbf{Y}'\mathbf{Y}/n} = \frac{\mathbf{v}'\mathbf{u}/n}{\mathbf{v}'\mathbf{v}/n} \xrightarrow{p} \frac{\sigma_{uv}}{\sigma_v^2} = \delta, \text{ so } plim(\hat{\beta}^{OLS}) = \beta_0 + \delta$$

Thus $\hat{\beta}^{TSLS}$ is centered around $plim(\hat{\beta}^{OLS})$

This is one end of the spectrum; the usual normal approximation is the other. If instruments are weak the distribution is somewhere in between…

TSLS with possibly weak instruments, 1 included endogenous regressor

Suppose that **Z** is fixed and **u**, **v** are normally distributed. Then the sample size enters the distribution of $\hat{\beta}^{TSLS}$ only through the ***concentration parameter*** μ^2, where

$$\mu^2 = \Pi'\mathbf{Z}'\mathbf{Z}\Pi/\sigma_v^2 \text{ (concentration parameter).}$$

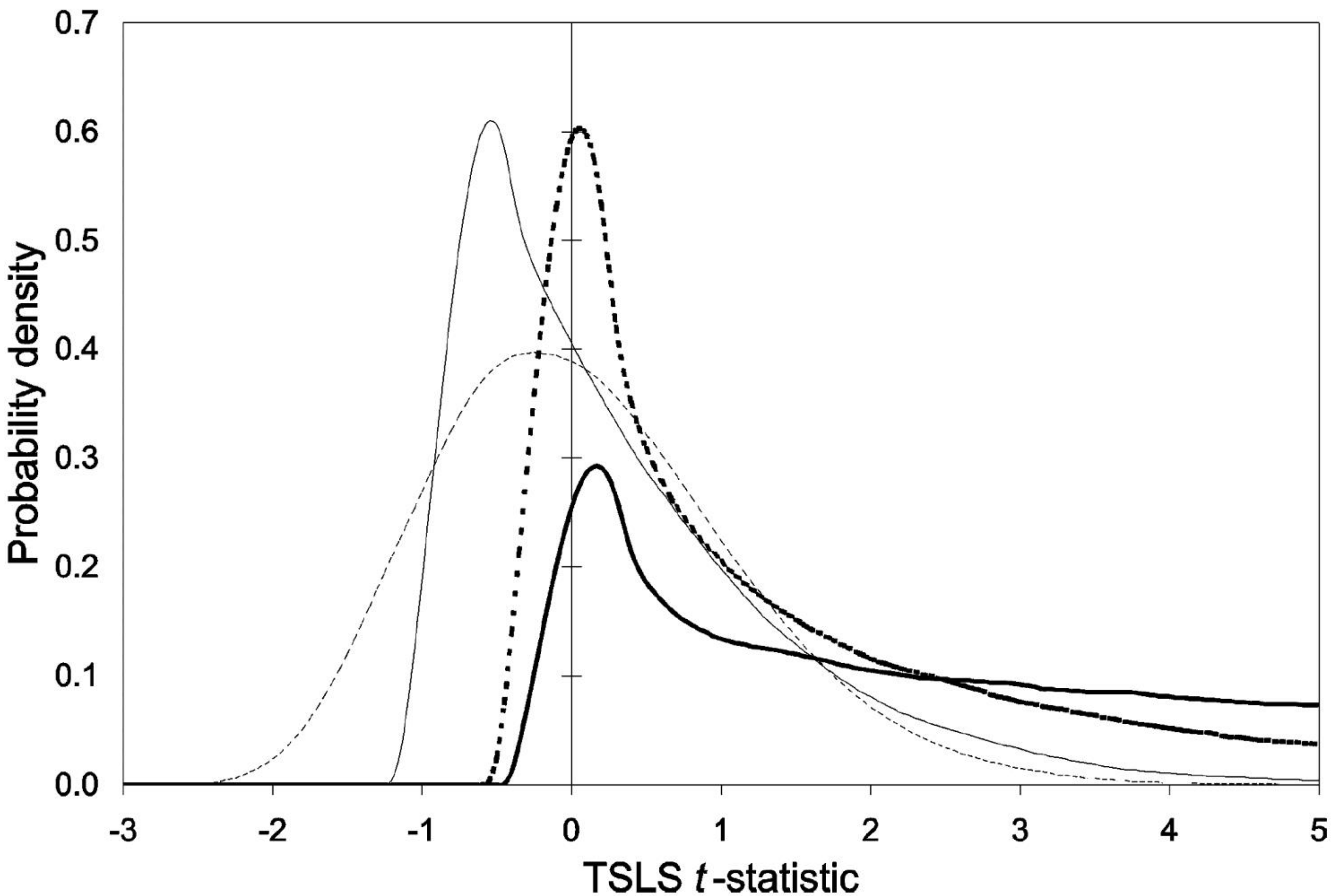
- μ^2 plays the role usually played by n
- As $\mu^2 \to \infty$, the usual asymptotic approximation obtains:

$$\text{as } \mu^2 \to \infty,\ \mu(\hat{\beta}^{TSLS} - \beta) \xrightarrow{d} N(0,\sigma_u^2/\sigma_v^2)$$

  (the σ_v^2 terms in μ and limiting variance cancel)
- for small values of μ^2, the distribution is nonstandard
- *Digression*: for a possibly helpful expansion of TSLS estimator in terms of μ^2 in the classical case, see Rothenberg (1984)

How important are these deviations from normality quantitatively?

Nelson-Startz (1990a,b) plots of the distribution of the TSLS *t*-statistic:



Dark line = irrelevant instruments; dashed light line = strong instruments; intermediate cases: weak instruments

Four approaches to computing distributions of IV statistics with weak IVs

The goal: a distribution theory that is tractable; provides good approximations uniformly in μ^2; and can be used to compare procedures

1. Finite sample theory?
   - large literature in 70s and 80s under the strong assumptions that $\boldsymbol{Z}$ is fixed (strictly exogenous) and (u_t, v_t) are i.i.d. normal
   - literature died – distributions aren't tractable, results aren't useful

2. Edgeworth expansions?
   - expand dist^n in orders of $T^{-1/2}$ – requires consistent estimability
   - work poorly when instruments are very weak (Rothenberg (1984))

3. Bootstrap and subsampling?
   - Neither work uniformly (irrelevant to weak to strong) in general
   - We return to these later (recent interesting literature)

4. Weak instrument asymptotics

Adopt nesting that makes the concentration parameter tend to a constant as the sample size increases; that is, model F as not increasing with the sample size.

This is accomplished by setting $\Pi = C/\sqrt{T}$

- This is the Pitman drift for obtaining the local power function of the first-stage F.
- This nesting holds $E\mu^2$ constant as $T \to \infty$.
- Under this nesting, $F \xrightarrow{d}$ noncentral χ_k^2/k with noncentrality parameter $E\mu^2/k$ (so $F = O_p(1)$)
- Letting the parameter depend on the sample size is a common ways to obtain good approximations – e.g. local to unit roots (Bobkoski 1983, Cavanagh 1985, Chan and Wei 1987, and Phillips 1987)

Weak IV asymptotics for TSLS estimator, 1 included endogenous vble:

$$\hat{\beta}^{TSLS} - \beta_0 = (\mathbf{Y}'\mathrm{P}_{\mathbf{Z}}\mathbf{u})/(\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y})$$

Now

$$\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y} = \left(\frac{(\mathbf{Z}\Pi + \mathbf{v})'\mathbf{Z}}{\sqrt{T}}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1}\left(\frac{\mathbf{Z}'(\mathbf{Z}\Pi + \mathbf{v})}{\sqrt{T}}\right)$$

$$= \left(\frac{\Pi\mathbf{Z}'\mathbf{Z}}{\sqrt{T}} + \frac{\mathbf{v}'\mathbf{Z}}{\sqrt{T}}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1/2\prime}\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1/2}\left(\frac{\mathbf{Z}'\mathbf{Z}\Pi}{\sqrt{T}} + \frac{\mathbf{Z}'\mathbf{v}}{\sqrt{T}}\right)$$

$$= \left[C'\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{1/2} + \frac{\mathbf{v}'\mathbf{Z}}{\sqrt{T}}\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1/2\prime}\right]\left[\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{1/2\prime}C + \left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1/2}\frac{\mathbf{Z}'\mathbf{v}}{\sqrt{T}}\right]$$

$$\xrightarrow{d} (\lambda + z_v)'\,(\lambda + z_v),$$

where

$$\lambda = C'Q_{ZZ}^{1/2},\ Q_{ZZ} = EZ_tZ_t',\ \text{and}\ \begin{pmatrix} z_u \\ z_v \end{pmatrix} \sim N\left(0, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}\right)$$

Similarly,

$$\mathbf{Y}'P_{\mathbf{Z}}\mathbf{u} = \left(\frac{(\mathbf{Z}\Pi + \mathbf{v})'\mathbf{Z}}{\sqrt{T}}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{u})}{\sqrt{T}}\right)$$

$$= \left(C'\frac{\mathbf{Z}'\mathbf{Z}}{T} + \frac{\mathbf{v}'\mathbf{Z}}{\sqrt{T}}\right)\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{u}}{\sqrt{T}}\right)$$

$$\xrightarrow{d} (\lambda + z_v)'z_u,$$

so

$$\hat{\beta}^{TSLS} - \beta_0 \xrightarrow{d} \frac{(\lambda + z_v)'z_u}{(\lambda + z_v)'(\lambda + z_v)}$$

- Under weak instrument asymptotics, $\mu^2 \xrightarrow{p} C'Q_{ZZ}C/\sigma_v^2 = \lambda'\lambda/\sigma_v^2$
- Unidentified special case: $\hat{\beta}^{TSLS} - \beta_0 \xrightarrow{d} \dfrac{z_v{}'z_u}{z_v{}'z_v}$ (obtained earlier)
- Strong identification: $\sqrt{\lambda'\lambda}\,(\hat{\beta}^{TSLS} - \beta_0) \xrightarrow{d} N(0,\sigma_u^2)$ (standard limit)

Summary of weak IV asymptotic results:

- Resulting asymptotic distributions are the same as in the exact normal classical model with fixed Z – but with ***known*** covariance matrices.
- IV estimators are not consistent (and are biased) under this nesting

Digression: Identification and consistency

- Identification means (loosely) that if you change a parameter, the distribution of the data changes. Because you can estimate the distribution of the data, this means you can work backwards to the parameter.
- Identification does not imply consistency. Consider the regression model, with $T \to \infty$:

$$Y_t = \beta_0 D_t + \beta_1(1 - D_t) + u_t, \text{ where } D_t = \begin{cases} 1, & t = 1,\ldots,10 \\ 0, & t = 11,\ldots,T \end{cases}$$

Both β_0 and β_1 are identified, but only β_1 is consistently estimable.

Summary of weak IV asymptotic results, ctd:

- IV estimators are nonnormal ($\hat{\beta}^{TSLS}$ has mixture of normals with nonzero mean, where mean $\propto k/\mu^2$)
- Test statistics (including the J-test of overidentifying restrictions) do not have normal or chi-squared distributions
- Conventional confidence intervals do not have correct coverage (coverage can be driven to zero)
- Provide good approximations to sampling distributions uniformly in μ^2 for T moderate or greater (say, 100+ observations).
- Remember, μ^2 is unknown – so these distributions can't be used directly in practice to obtain a "corrected" distribution….

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) **Classical IV regression II: Detection of weak instruments**
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

3) Classical IV regression II: Detection of weak instruments

Bound et. al. revisited

- n = 329,509 (it is μ^2, or μ^2/k, not sample size that matters!)
- for K = 3 (quarter of birth only), F = 30.53,
  - Recall that $E(F) = 1 + \mu^2/k$
  - Estimate of μ^2/k is 29.53
  - Estimate μ^2 as $k(F-1) = 3\times(30.53-1) = 88.6$
- for K = 178 (all interactions), F = 1.869
  - Estimate of $\mu^2 = 178\times(1.869-1) = 154.7$
  - Estimate of μ^2/k is 0.869
- We will see that numerical work suggests that
  - $\mu^2/k = 29.53$: strong instruments
  - $\mu^2/k = 0.869$: very weak instruments

How weak is weak? Need a cutoff value for μ^2

The basic idea is to compare F to some cutoff. But how should that cutoff be chosen? In general, this depends on the statistic you are using (different statistics have different sensitivities to μ^2). TSLS is among the worst (most sensitive) – and is also most frequently used. So, it is reasonable to develop a cutoff for F assuming use of TSLS.

Various procedures:

- First stage $F > 10$ rule of thumb
- Stock-Yogo (2005a) bias method
- Stock-Yogo (2005a) size method

TSLS bias cutoff method (Stock-Yogo (2005a))

Let $\mu^2_{10\%bias}$ be the value of μ^2 such that, if $\mu^2 \geq \mu^2_{10\%bias}$, the maximum bias of TSLS will be no more than 10% of the bias (inconsistency) of OLS.

Stock-Yogo (2005a): decision rule of the form:

$$\text{if} \quad F \begin{pmatrix} \leq \\ > \end{pmatrix} \kappa_{.10}(k), \text{ conclude that instruments are } \begin{pmatrix} \text{weak} \\ \text{strong} \end{pmatrix}$$

where F is the first stage F-statistic* and $\kappa_{.10}(k)$ is chosen so that P($F > \kappa_{.10}(k)$; $\mu^2 = \mu^2_{10\%bias}$) = .05 (so that the rule acts like a 5% significance test at the boundary value $\mu^2 = \mu^2_{10\%bias}$).

***F** = F-statistic testing the hypothesis that the coefficients on $Z_t = 0$ in the regression of Y_t on Z_t, W_t, and a constant, where W_t = the exogenous regressors included in the equation of interest.

TSLS bias cutoff method (Stock-Yogo (2005a)), ctd

Some background:

The relative squared normalized bias of TSLS to OLS is,

$$B_n^2 = \frac{(E\hat{\beta}^{\mathrm{IV}} - \beta)'\Sigma_{YY}(E\hat{\beta}^{\mathrm{IV}} - \beta)}{(E\hat{\beta}^{\mathrm{OLS}} - \beta)'\Sigma_{YY}(E\hat{\beta}^{\mathrm{OLS}} - \beta)}$$

The square root of the maximal relative squared asymptotic bias is:

$$B^{max} = \max_{\boldsymbol{\rho}:\, 0 < \boldsymbol{\rho}'\boldsymbol{\rho} \le 1} \lim_{n\to\infty}|B_n|, \text{ where } \boldsymbol{\rho} = \mathrm{corr}(u_t, v_t)$$

This maximization problem is a ratio of quadratic forms so it turns into a (generalized) eigenvalue problem; algebra reveals that the solution to this eigenvalues problem depends only on μ^2/k and k; this yields the cutoff μ_{bias}^2.

Critical values

One included endogenous regressor

The 5% critical value of the test is the 95% percentile value of the noncentral χ_k^2/k distribution, with noncentrality parameter μ_{bias}^2/k

Multiple included endogenous regressors

The Cragg-Donald (1993) statistic is:

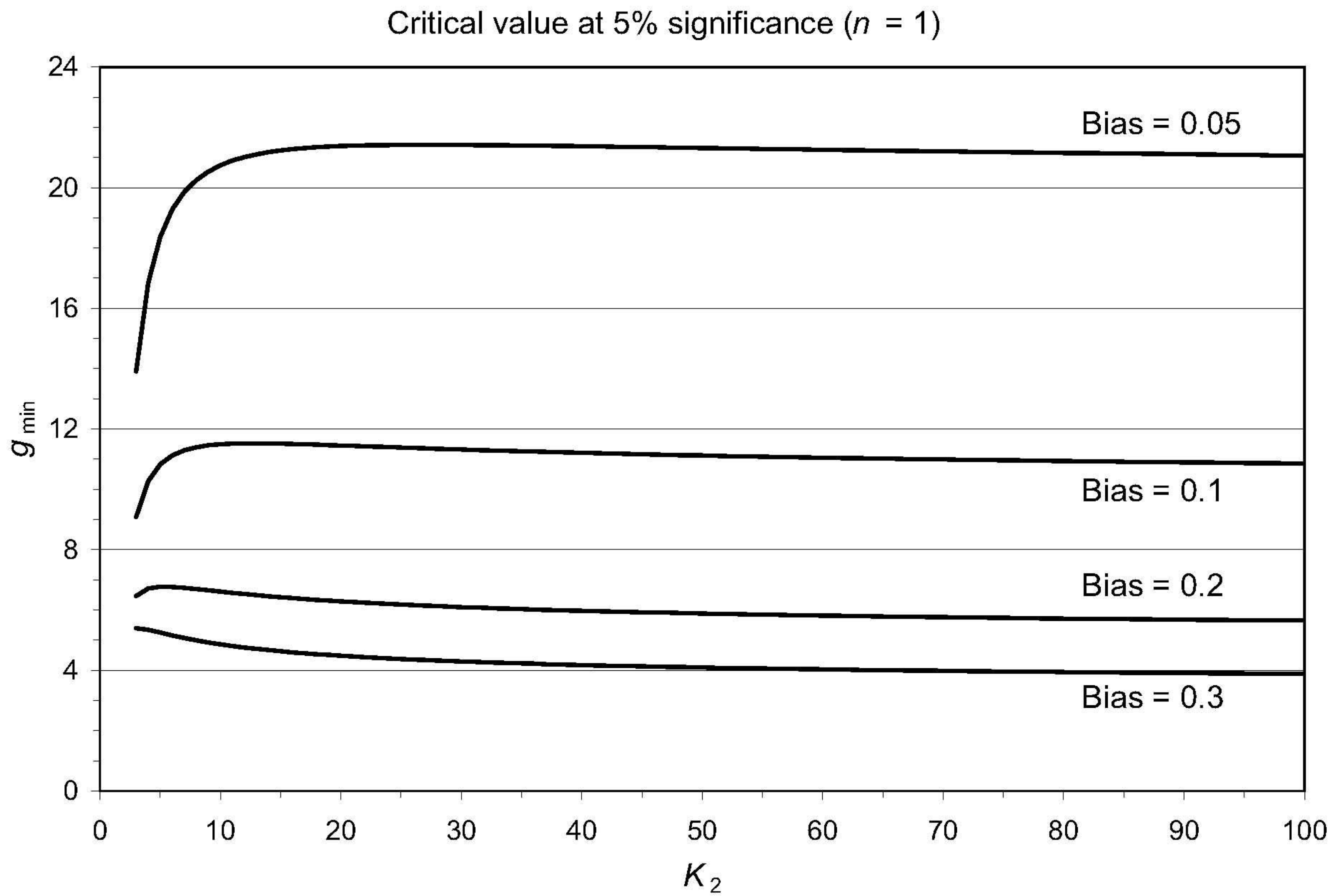
$$g_{min} = \text{mineval}(G_T), \text{ where } G_T = \hat{\Sigma}_{\mathbf{VV}}^{-1/2}{}'\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y}\hat{\Sigma}_{\mathbf{VV}}^{-1/2}/k,$$

- G_T is essentially a matrix first stage F statistic
- Critical values are given in Stock-Yogo (2005a)

Software

STATA (ivreg2),…

5% critical value of F to ensure indicated maximal bias
(Stock-Yogo, 2005a)



To ensure 10% maximal bias, need $F \geq 11.52$; $F \geq 10$ is a rule of thumb

5% critical values for Weak IV test statistic g_{min} ,
for 10% maximal TSLS Bias (Stock-Yogo (2005), Table 1) m = dim(Y_t)

k	m = 1	m = 2	m = 3
3	9.08	–	–
4	10.27	7.56	–
5	10.83	8.78	6.61
6	11.12	9.48	7.77
7	11.29	9.92	8.50
8	11.39	10.22	9.01
9	11.46	10.43	9.37
10	11.49	10.58	9.64
15	11.51	10.93	10.33
20	11.45	11.03	10.60
25	11.38	11.06	10.71
30	11.32	11.05	10.77

Other methods for detecting weak instruments

Stock-Yogo (2005a) size method

- Instead of controlling bias, control the size of a Wald test of $\beta = \beta_0$
- Less frequently used
- Not really relevant (any more) since fully robust methods for testing exist

Recent work has focused on extention to heteroskedasticity and serial correlation

- The problem: With heteroskedasticity, except in special cases the concentration parameter for 2SLS and the noncentrality parameter of the first-stage F (either hetero-robust or nonrobust) don't coincide
- The solution: ongoing research. See Olea Montiel and Pflueger (2013) , I. Andrews (2014)

Other methods for detecting weak instruments

Examination of R^2, partial R^2, or adjusted R^2

- None of these are a good idea, more precisely, what needs to be large is the concentration parameter, not the R^2. An $R^2 = .10$ is small if $T = 50$ but is large if $T = 5000$.
- The first-stage R^2 is especially uninformative if the first stage regression has included exogenous regressors (W's) because it is the marginal explanatory content of the Z's, given the W's, that matters.

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) **Classical IV regression III: hypothesis tests and confidence intervals**
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

4) Classical IV regression III:
Hypothesis tests and confidence intervals

There are two approaches to improving inference (providing tools):

Fully robust methods:

- Inference that is valid for any value of the concentration parameter, including zero, at least if the sample size is large, under weak instrument asymptotics
  - For tests: asymptotically correct size (and good power!)
  - For confidence intervals: asymptotically correct coverage rates
  - For estimators: asymptotically unbiased (or median-unbiased)

Partially robust methdos:

- Methods are less sensitive to weak instruments than TSLS – e.g. bias is "small" for a "large" range of μ^2

Fully Robust Testing

- The TSLS *t*-statistic has a distribution that depends on μ^2, which is unknown
- Approach #1: use a statistic whose distribution depends on μ^2, but use a "worst case" conservative critical value
    - This is unattractive – substantial power loss
- Approach #2: use a statistic whose distribution does not depend on μ^2 (two such statistics are known)
- Approach #3: use statistics whose distribution depends on μ^2, but compute the critical values as a function of another statistic that is sufficient for μ^2 under the null hypothesis.
    - Both approaches 2 and 3 have advantages and disadvantages – we discuss both

Approach #2: Tests that are valid unconditionally

(that is, the distribution of the test statistic does not depend on μ^2)

The Anderson-Rubin (1949) test

Consider H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in $\mathbf{y} = \mathbf{Y}\boldsymbol{\beta} + \mathbf{u}$,

$$\mathbf{Y} = \mathbf{Z}\mathbf{\Pi} + \mathbf{v}$$

The Anderson-Rubin (1949) statistic is the F-statistic in the regression of $\mathbf{y} - \mathbf{Y}\beta_0$ on $\mathbf{Z}$.

$$\text{AR}(\beta_0) = \frac{(\mathbf{y} - \mathbf{Y}\beta_0)' P_{\mathbf{Z}}(\mathbf{y} - \mathbf{Y}\beta_0)/k}{(\mathbf{y} - \mathbf{Y}\beta_0)' M_{\mathbf{Z}}(\mathbf{y} - \mathbf{Y}\beta_0)/(T-k)}$$

$$\text{AR}(\beta_0) = \frac{(\mathbf{y}-\mathbf{Y}\beta_0)' P_{\mathbf{Z}} (\mathbf{y}-\mathbf{Y}\beta_0)/k}{(\mathbf{y}-\mathbf{Y}\beta_0)' M_{\mathbf{Z}} (\mathbf{y}-\mathbf{Y}\beta_0)/(T-k)}$$

<u>Comments</u>

- AR($\hat{\beta}^{TSLS}$) = the *J*-statistic
- Null distribution doesn't depend on μ^2:

  Under the null, $\mathbf{y} - \mathbf{Y}\beta_0 = \mathbf{u}$, so

  $$\text{AR} = \frac{\mathbf{u}' P_{\mathbf{Z}} \mathbf{u}/k}{\mathbf{u}' M_{\mathbf{Z}} \mathbf{u}/(T-k)} \sim F_{k,n-k} \qquad \text{if } u_t \text{ is normal}$$

  $$\text{AR} \xrightarrow{d} \chi^2_k/k \qquad \text{if } u_t \text{ is i.i.d. and } Z_t u_t \text{ has 2 moments (CLT)}$$

- The distribution of AR under the alternative depends on μ^2 – more information, more power (of course)

The AR statistic if there are included endogenous regressors

Let $\mathbf{W}$ denote the matrix of observations on included exogenous regressors, so the structural equation and first stage regression are,

$$\mathbf{y} = \mathbf{Y}\beta + \mathbf{W}\gamma + \mathbf{u}$$
$$\mathbf{Y} = \mathbf{Z}\Pi + \mathbf{W}\Pi_W + \mathbf{v}$$

The AR statistic is the *F*-statistic testing the hypothesis that the coefficients on $\mathbf{Z}$ are zero in the regression of $\mathbf{y} - \mathbf{Y}\beta_0$ on $\mathbf{Z}$ and $\mathbf{W}$.

Advantages and disadvantages of AR

Advantages

- Easy to use – entirely regression based
- Uses standard F critical values
- Works for $m > 1$ (general dimension of Z) (see Kleibergen and Mavroeidis (2009) for subset inference when $m > 1$)

Disadvantages

- Difficult to interpret: rejection arises for two reasons: β_0 is false *or* Z is endogenous
- Power loss relative to other tests (we shall see)
- Is not efficient if instruments are strong – under strong instruments, not as powerful as TSLS Wald test (power loss because AR(β_0) has k degrees of freedom)

Kleibergen's (2002) LM test

Kleibergen developed an LM test that has a null distribution that is χ^2_1 - doesn't depend on μ^2.

Advantages

- Fairly easy to implement
- Is efficient if instruments are strong

Disadvantages

- Has very strange power properties – power function isn't monotonic
- Its power is dominated by the conditional likelihood ratio test

Approach #3: Conditional tests

Conditional tests have rejection rate 5% for all points under the null (β_0, μ^2) ("similar tests")

Recall your first semester probability and statistics course…

- Let S be a statistic with a distribution that depends on θ
- Let T be a sufficient statistic for θ
- Then the distribution of $S|T$ does not depend on θ

Here (Moreira (2003)):

- LR will be a statistic testing $\beta = \beta_0$ (LR is "S" in notation above)
- Q_T will be sufficient for μ^2 under the null (Q_T is "T")
- Thus the distribution of $LR|\ Q_T$ does not depend on μ^2 under the null
- Thus valid inference can be conducted using the quantiles of $LR|\ Q_T$ – that is, critical values that are a function of Q_T

Moreira's (2003) conditional likelihood ratio (CLR) test

$$LR = \max_\beta \text{log-likelihood}(\beta) - \text{log-likelihood}(\beta_0)$$

After lots of algebra, this becomes:

$$LR = \tfrac{1}{2}\{\hat{Q}_S - \hat{Q}_T + [(\hat{Q}_S - \hat{Q}_T)^2 + 4\hat{Q}_{ST}^2]^{1/2}\}$$

where

$$\hat{Q} = \begin{bmatrix} \hat{Q}_S & \hat{Q}_{ST} \\ \hat{Q}_{ST} & \hat{Q}_T \end{bmatrix} = \hat{J}_0{}'\hat{\Omega}^{-1/2}\mathbf{Y}^{+\prime}P_{\mathbf{Z}}\mathbf{Y}^{+}\hat{\Omega}^{-1/2\prime}\hat{J}_0$$

$$\hat{\Omega} = \mathbf{Y}^{+\prime}M_{\mathbf{Z}}\mathbf{Y}^{+}/(T-k), \quad \mathbf{Y}^{+} = (\mathbf{y} \;\; \mathbf{Y})$$

$$\hat{J}_0 = \left[\frac{\hat{\Omega}^{1/2\prime}b_0}{\sqrt{b_0{}'\hat{\Omega}b_0}} \quad \frac{\hat{\Omega}^{-1/2}a_0}{\sqrt{a_0{}'\hat{\Omega}^{-1}a_0}} \right], \; b_0 = \begin{pmatrix} 1 \\ -\beta_0 \end{pmatrix} \; a_0 = \begin{pmatrix} \beta_0 \\ 1 \end{pmatrix}.$$

CLR test, ctd.

Implementation:

- Q_T is sufficient for μ^2 (under weak instrument asymptotics)
- The distribution of $LR|Q_T$ does not depend on μ^2
- *LR* proc exists in STATA (condivreg), GAUSS
- STATA (condivreg), Gauss code for computing LR and conditional *p*-values exists

Advantages and disadvantages of the CLR test

Advantages

- More powerful than AR or LM
- In fact, effectively uniformly most powerful among valid tests that are invariant to rotations of the instruments (Andrews, Moreira, Stock (2006) – among similar tests; Andrews, Moreira, Stock (2008) – among nonsimilar tests)
- Implemented in software (STATA,…)

Disadvantages

- More complicated to explain and write down
- Only developed (so far) for a single included endogenous regressor
- As written, the software requires homoskedastic errors; extensions to heteroskedasticity and serial correlation have been developed but are not in common statistical software

Confidence Intervals

(a) A 95% confidence set is a function of the data contains the true value in 95% of all samples

(b) A 95% confidence set is constructed as the set of values that cannot be rejected as true by a test with 5% significance level

Usually (b) leads to constructing confidence sets as the set of β_0 for which $-1.96 < \dfrac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} < 1.96$. Inverting this t-statistic yields $\hat{\beta} \pm 1.96SE(\hat{\beta})$

- This won't work for TSLS – t^{TSLS} isn't normal (the critical values of t^{TSLS} depend on μ^2)
- Dufour (1997) impossibility result for weak instruments: unbounded intervals must occur with positive probability.
- However, you can compute a valid, fully robust confidence interval by inverting a fully robust test!

(1) Inversion of AR test: AR Confidence Intervals

$$95\%\ \mathrm{CI} = \{\beta_0 \text{: } \mathrm{AR}(\beta_0) < F_{k,T-k;.05}\}$$

Computational issues:

- For $m = 1$, this entails solving a quadratic equation:

$$\mathrm{AR}(\beta_0) = \frac{(\mathbf{y}-\mathbf{Y}\beta_0)' P_{\mathbf{Z}} (\mathbf{y}-\mathbf{Y}\beta_0)/k}{(\mathbf{y}-\mathbf{Y}\beta_0)' M_{\mathbf{Z}} (\mathbf{y}-\mathbf{Y}\beta_0)/(T-k)} < F_{k,T-k;.05}$$

- For $m > 1$, solution can be done by grid search or using methods in Dufour and Taamouti (2005)

- Sets for a single coefficient can be computed by projecting the larger set onto the space of the single coefficient (see Dufour and Taamouti (2005)), also see Kleibergen and Mavroeidis (2009)

AR confidence intervals, ctd.

$$95\%\text{ CI} = \{\beta_0: \text{AR}(\beta_0) < F_{k,T-k;.05}\}$$

Four possibilities:

- a single bounded confidence interval
- a single unbounded confidence interval
- a disjoint pair of confidence intervals
- an empty interval

Note:

- Difficult to interpret
- Intervals aren't efficient (AR test isn't efficient) under strong instruments

(2) Inversion of CLR test: CLR Confidence Intervals

$$95\%\ \text{CI} = \{\beta_0 : \text{LR}(\beta_0) < \text{cv}_{.05}(Q_T)\}$$

where $\text{cv}_{.05}(Q_T)$ = 5% conditional critical value

Comments:

- Efficient GAUSS and STATA (condivreg) software
- Will contain the LIML estimator (Mikusheva (2005))
- Has certain optimality properties: nearly uniformly most accurate invariant; also minimum expected length in polar coordinates (Mikusheva (2005))
- Only available for m = 1

Extensions to >1 included endogenous regressor

- Usually the extension to higher dimensions is easy – standard normal *t*-ratios, chi-squared *F*-tests, etc. But once normality of estimators and chi-squared distribution of tests are gone, the extensions are not easy.
- CLR exists in theory, but unsolved computational issues because the conditioning statistic has dimension $m(m+1)/2$ (Kleibergen (2007))
- Can test joint hypothesis H_0: $\beta = \beta_0$ using the AR statistic:

$$\text{AR}(\beta_0) = \frac{(\mathbf{y}-\mathbf{Y}\beta_0)'P_{\mathbf{Z}}(\mathbf{y}-\mathbf{Y}\beta_0)/k}{(\mathbf{y}-\mathbf{Y}\beta_0)'M_{\mathbf{Z}}(\mathbf{y}-\mathbf{Y}\beta_0)/(T-k)}$$

under H_0, AR $\xrightarrow{d}$ χ_k^2/k

Recent references on testing in linear IV case, including robustifying (heteroskedasticity, autocorrelation):

I. Andrews (2013)

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) **Classical IV regression IV: Estimation**
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

5) Classical IV regression IV: Estimation

Estimation is much harder than testing or confidence intervals

- Uniformly unbiased estimation is impossible (among estimators with support on the real line), uniformly in μ^2
- Estimation must be divorced from confidence intervals

<u>Partially robust estimators (with smaller bias/better MSE than TSLS):</u>

Remember *k*-class estimators?

$$\hat{\beta}(\underline{k}) = [\mathbf{Y}'(I - \underline{k}M_{\mathbf{Z}})\mathbf{Y}]^{-1}[\mathbf{Y}'(I - \underline{k}M_Z)\mathbf{y}]$$

TSLS: $\underline{k} = 1$,

LIML: $\underline{k} = \hat{\underline{k}}_{LIML}$ = smallest root of $\det(Y^{\perp\prime}Y^{\perp} - \underline{k}Y^{\perp\prime}M_Z Y^{\perp}) = 0$

Fuller: $\underline{k} = \hat{\underline{k}}_{LIML} - c/(T\text{–}k\text{–}\#\textit{included exog.}),\ c > 0$

Comparisons of *k*-class estimators

Anderson, Kunitomo, and Morimune (1986) – using second order theory
Hahn, Hausman, and Kuersteiner (2004) – using MC simulations

LIML

- median unbiased to second order
- HHK simulations – LIML exhibits very low median bias
- no moments exist! There can be extreme outliers
- LIML also can be shown to minimize the AR statistic:

$$\hat{\beta}^{LIML}\text{: } \min_{\beta} \text{AR}(\beta) = \frac{(\mathbf{y}-\mathbf{Y}\beta_0)'P_{\mathbf{Z}}(\mathbf{y}-\mathbf{Y}\beta_0)/k}{(\mathbf{y}-\mathbf{Y}\beta_0)'M_{\mathbf{Z}}(\mathbf{y}-\mathbf{Y}\beta_0)/(T-k)}$$

so LIML necessarily falls in the AR confidence set if it is nonempty

Comparisons of *k*-class estimators, ctd.

Fuller

- With c = 1, lowest RMSE to second order among a certain class (Rothenberg (1984))
- In simulation studies (m=1), Fuller performs very well with c = 1

Others

- (Jacknife TSLS; bias-adjusted TSLS) are dominated by Fuller, LIML

LIML (and other) estimators with heterogeneous treatment effects.

Kolesár (2013) shows that a class of minimum distance estimators, which includes LIML and the Hausman et. al. (2012) many instrument estimator, can have an estimand that is outside the convex hull of the individual treatment effects – that is, it estimates an object which is not a treatment effect for anyone, or a (convex) average of anyone's. A big problem for LIML and related estimators – making them much less attractive as a solution to the weak (or many) instrument problem.

Summary and recommendations

- Under strong instruments, LIML, TSLS, *k*-class will all be close to each other.
- under weak instruments, TSLS has greatest bias and large MSE
- LIML has the advantage of minimizing AR – and thus always falling in the AR (and CLR) confidence set. LIML is a reasonable (good) choice as an alternative to TSLS.
- But LIML is not well-suited to situations in which there are heterogeneous treatment effects, such as individual-level program evaluation studies.

What about the bootstrap or subsampling?

The bootstrap is often used to improve performance of estimators and tests through bias adjustment and approximating the sampling distribution.

A straightforward bootstrap algorithm for TSLS:

$$y_t = \beta' Y_t + u_t$$

$$Y_t = \Pi' Z_t + v_t$$

i) Estimate β, Π by $\hat{\beta}^{TSLS}$, $\hat{\Pi}$

ii) Compute the residuals $\hat{u}_t$, $\hat{v}_t$

iii) Draw T "errors" and exogenous variables from $\{\hat{u}_t, \hat{v}_t, Z_t\}$, and construct bootstrap data $\tilde{y}_t$, $\tilde{Y}_t$ using $\hat{\beta}^{TSLS}$, $\hat{\Pi}$

iv) Compute TSLS estimator (and t-statistic, etc.) using bootstrap data

v) Repeat, and compute bias-adjustments and quantiles from the boostrap distribution, e.g. bias = bootstrap mean of $\hat{\beta}^{TSLS} - \hat{\beta}^{TSLS}$ using actual data

Bootstrap, ctd.

- Under strong instruments, this algorithm works (provides second-order improvements).
- Under weak instruments, this algorithm (or variants) does not even provide first-order valid inference

  The reason the bootstrap fails here is that $\hat{\Pi}$ is used to compute the bootstrap distribution. The true pdf depends on μ^2, say $f_{TSLS}(\hat{\beta}^{TSLS};\mu^2)$ (e.g. Rothenberg (1984 exposition above, or weak instrument asymptotics). By using $\hat{\Pi}$, μ^2 is estimated, say by $\hat{\mu}^2$. The bootstrap correctly estimates $f_{TSLS}(\hat{\beta}^{TSLS};\hat{\mu}^2)$, but $f_{TSLS}(\hat{\beta}^{TSLS};\hat{\mu}^2) \neq f_{TSLS}(\hat{\beta}^{TSLS};\mu^2)$ because $\hat{\mu}^2$ is not consistent for μ^2. ‘

Bootstrap, ctd.

- This is simply another aspect of the nuisance parameter problem in weak instruments. If we could estimate μ^2 consistently, the bootstrap would work – but we if so wouldn't need it anyway (at least to first order) since we would have operational first order approximating distributions!
- This story might sound familiar – it is the same reason the bootstrap fails in the unit root model, and in the local-to-unity model, which led to Hansen's (1999) grid bootstrap, which has been shown to produce valid confidence intervals for the AR(1) coefficient by Mikusheva (2007).
- Failure of bootstrap in weak instruments is related to failure of Edgeworth expansion (uniformly in the strength of the instrument), see Hall (1992) in general, Moreira, Porter, and Suarez (2005a,b) in particular.
- One way to avoid this problem is to bootstrap test statistics with null distributions that do not depend on μ^2. Bootstrapping AR and LM *does* result in second order improvements, see Moreira, Porter, and Suarez (2005a,b).

What about subsampling?

Politis and Romano (1994), Politis, Romano and Wolf (1999)

Subsampling uses smaller samples of size *m* to estimate the parameters directly. If the CLT holds, the distribution of the subsample estimators, scaled by $\sqrt{m/T}$, approximates the distribution of the full-sample estimator.

A subsampling algorithm for TSLS:

(i) Choose subsample of size *m* and compute TSLS estimator

(ii) Repeat for all subsamples of size *m* (in cross-section, there are $\binom{T}{m}$ such subsamples; in time series, there are $T–m$)

(iii) Compute bias adjustments, quantiles, etc. from the rescaled empirical distribution of the subsample estimators.

Subsampling, ctd.

- Subsampling works in some cases in which bootstrap doesn't (Politis, Romano, and Wolf (1999))
- However, it doesn't work (doesn't provide first-order valid approximations to sampling distributions) with weak instruments (Andrews and Guggenberger (2007a,b)).
- The subsampling distribution estimates $f_{TSLS}(\hat{\beta}^{TSLS};\mu_m^2)$, where μ_m^2 is the concentration parameter for *m* observations. But this is less (on average, by the factor *m*/*T*) than the concentration parameter for *T* observations, so the scaled subsample distribution does not estimate $f_{TSLS}(\hat{\beta}^{TSLS};\mu_T^2)$.
- Subsampling can be size-corrected (in this case) but there is power loss relative to CLR; see Andrews and Guggenberger (2007b)

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) **GMM I: Setup and asymptotics**
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

6) GMM I: Setup and asymptotics

GMM notation and estimator

GMM "error" term (G equations): $h(Y_t;\theta)$; θ_0 = true value

Errors times k instruments: $\phi_t(\theta) = \overset{G\times 1}{h(Y_t,\theta_0)} \otimes \overset{k\times 1}{Z_t}$

Moment conditions - k instruments: $E\phi_t(\theta) = E[\overset{G\times 1}{h(Y_t,\theta_0)} \otimes \overset{k\times 1}{Z_t}] = 0$

GMM objective function: $S_T(\theta) = \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta)\right]' W_T \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta)\right]$

GMM estimator: $\hat{\theta}$ minimizes $S_T(\theta)$

Linear GMM: $h(Y_t;\theta) = y_t - \theta' Y_t$

(linear GMM is the IV regression model, allowing for possible heteroskedasticity and/or serial correlation in the errors h)

Efficient GMM

Centered sample moments: $\Psi_T(\theta) = T^{-1/2}\sum_{t=1}^{T}\left(\phi_t(\theta) - E\phi_t(\theta)\right)$

Efficient (infeasible) GMM: $W_T = \Omega^{-1}$, $\Omega = E[\Psi_T(\theta)\Psi_T(\theta)'] = 2\pi S_{\phi_t(\theta)}(0)$

Feasible GMM

Estimator of Ω: $\hat{\Omega}(\theta)$ = HAC estimator of $\Omega = \sum_{j=-S}^{S}\kappa_j\hat{\Gamma}_j(\theta)$,

where $\hat{\Gamma}_j(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(\phi_t(\theta) - \overline{\phi_t(\theta)}\right)\left(\phi_{t-j}(\theta) - \overline{\phi_{t-j}(\theta)}\right)'$

$\{\kappa_j\}$ are kernel weights (e.g. Newey-West)

Feasible GMM variants

One-step: W_T = fixed matrix (e.g. $W_T = I$)

Two-step efficient: $W_T^{(1)} = I$, $W_T^{(2)} = \hat{\Omega}(\hat{\theta}^{(1)})^{-1}$

Iterated: continue iterating, with $W_T^{(i+1)} = \hat{\Omega}(\hat{\theta}^{(i)})^{-1}$

CUE (Hansen, Heaton, Yaron 1996): $W_T = \hat{\Omega}(\theta)^{-1}$ (evaluate $\hat{\Omega}$ at every θ!)

Standard GMM asymptotics

1) Establish consistency by showing the minimum of S_T will occur local to the true value θ_0: Pr[$S_T(\theta) < S_T(\theta_o)$] $\to 0$ for $|\theta - \theta_0| > \varepsilon$

so by smoothness of the objective function, Pr[$|\hat{\theta} - \theta_0| > \varepsilon$] $\to 0$

2) Establish normality by making quadratic approximation to S_T, based on consistency (which justifies dropping the higher order terms in the Taylor expansion):

$$S_T(\hat{\theta}) \approx S_T(\theta_0) + \sqrt{T}\,(\hat{\theta} - \theta_0)' \frac{1}{\sqrt{T}} \left.\frac{\partial S_T(\theta)}{\partial \theta}\right|_{\theta_0}$$

$$+ \tfrac{1}{2}\sqrt{T}\,(\hat{\theta} - \theta_0)' \left[\frac{1}{T} \left.\frac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right|_{\theta_0} \right] \sqrt{T}\,(\hat{\theta} - \theta_0)$$

so $$\sqrt{T}\,(\hat{\theta} - \theta_0) \approx \left[\frac{1}{T} \left.\frac{\partial^2 S_T(\theta)}{\partial \theta \partial \theta'}\right|_{\theta_0} \right]^{-1} \frac{1}{\sqrt{T}} \left.\frac{\partial S_T(\theta)}{\partial \theta}\right|_{\theta_0}$$

If $W_T \xrightarrow{p} W$ (say), then

$$\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta_0} \xrightarrow{p} DWD', \text{ where } D = E\frac{\partial\phi_t(\theta)}{\partial\theta}\bigg|_{\theta_0}$$

$$\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial\theta}\bigg|_{\theta_0} \xrightarrow{d} N(0,DW\Omega W'D')$$

so

$$\sqrt{T}(\hat{\theta} - \theta_0) \approx \left[\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\right]^{-1}\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial\theta}\bigg|_{\theta_0}$$

$$\xrightarrow{d} N(0,[DWD']^{-1}DW\Omega W'D'[DWD']^{-1})$$

<u>Feasible efficient GMM</u>

For two-step, iterated, and CUE, $W_T \xrightarrow{p} \Omega^{-1}$, so $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma)$

where $\Sigma = (\mathrm{D}\Omega^{-1}\mathrm{D}')^{-1}$

Estimator of variance matrix: $\hat{\Sigma} = [\hat{D}(\hat{\theta})\hat{\Omega}(\hat{\theta})\hat{D}(\hat{\theta})']^{-1}$

Weak identification in GMM – what goes wrong in the usual proof?

Digression:

- We will use the term "weak identification" because "weak instruments" is not precise in the nonlinear setting
- In the linear case, the strength of the instruments doesn't depend on θ
- In nonlinear GMM, the strength of the instruments can depend on θ: they can be weak for some departures $h(Y_t,\theta) - h(Y_t,\theta_0)$, but strong for others

When identification is weak, there are 2 problems with the usual proof:

(a) The curvature, which reflects the amount of information, is small, so the maximizer of S_T might not be close to θ_0.

(b) The curvature matrix is not well-approximated as nonrandom (I. Andrews and Mikusheva (2014a, b))

(c) The linear term, $\left.\frac{\partial S_T(\theta)}{\partial \theta}\right|_{\theta_0}$, is not approximately normal with mean 0

Illustration: linear IV in the GMM framework

The TSLS objective function (two-step GMM) is exactly quadratic:

$$S(\theta) = (\mathbf{y} - \mathbf{Y}\theta)'P_{\mathbf{Z}}(\mathbf{y} - \mathbf{Y}\theta)$$

$$= [\mathbf{u} - \mathbf{Y}(\theta - \theta_0)]'P_{\mathbf{Z}}[\mathbf{u} - \mathbf{Y}(\theta - \theta_0)]$$

$$= \mathbf{u}'P_{\mathbf{Z}}\mathbf{u} + (2\mathbf{u}'P_{\mathbf{Z}}\mathbf{Y})(\theta - \theta_0) - \tfrac{1}{2}(\theta - \theta_0)'(2\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y})(\theta - \theta_0)$$

or

$$S_T(\hat{\theta}) = S_T(\theta_0) + \sqrt{T}(\hat{\theta} - \theta_0)'\frac{1}{\sqrt{T}}\left.\frac{\partial S_T(\theta)}{\partial \theta}\right|_{\theta_0}$$

$$+ \tfrac{1}{2}\sqrt{T}(\hat{\theta} - \theta_0)'\left[\frac{1}{T}\left.\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]\sqrt{T}(\hat{\theta} - \theta_0)$$

where $$S_T(\theta_0) = \mathbf{u}'P_{\mathbf{Z}}\mathbf{u}$$

$$\frac{1}{\sqrt{T}}\left.\frac{\partial S_T(\theta)}{\partial \theta}\right|_{\theta_0} = 2\mathbf{u}'P_{\mathbf{Z}}\mathbf{Y}/\sqrt{T}$$

$$\frac{1}{T}\left.\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\right|_{\theta_0} = 2\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y}/T$$

Illustration: linear IV in the GMM framework, ctd.

(a) The curvature is small (so estimator need not be local)

$$\frac{1}{T}\frac{\partial^2 S_T(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta_0} = 2\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y}$$

$$= 2\frac{\mathbf{Y}'P_{\mathbf{Z}}\mathbf{Y}/k}{\mathbf{Y}'M_{\mathbf{Z}}\mathbf{Y}/(T-k)}\mathbf{Y}'M_{\mathbf{Z}}\mathbf{Y}/(T-k)$$

$$= 2kFs_v^2,$$

where F is the first-stage F and s_v^2 is the estimator of σ_v^2.

(b) The curvature is random – not well approximated by a constant

$F/\mu^2 \to 1$ as $\mu^2 \to \infty$, but for small μ^2, $F = \mu^2 + o_p(1)$

(c) Under weak instrument asymptotics, the linear term is non-normal:

$$\frac{1}{\sqrt{T}}\frac{\partial S_T(\theta)}{\partial\theta}\bigg|_{\theta_0} = 2\mathbf{u}'P_{\mathbf{Z}}\mathbf{Y}/\sqrt{T} \xrightarrow{d} 2(\lambda + z_v)'z_u,$$

which has a mixture-of-normals distribution with a nonzero mean (recall the distribution of TSLS under weak instrument asymptotics)

Alternative asymptotics for weak identification

As in the linear case, we need asymptotics for GMM that are tractable; that provide a good approximations uniformly in strength of identification; and that can be used to compare procedures.

Alternative approaches:

1. Finite sample – good luck!
2. Edgeworth and related expansions – useful for developing partially robust procedures but won't cover complete range through unidentified case
3. Bootstrap & resampling – doesn't work in linear IV special case
4. Weak identification asymptotics – provide nesting (parameter sequence) that provides an approximation uniformly in strength of identification

Weak ID asymptotics in GMM

(Stock and Wright (2000); Cheng and Andrews (2012))

Use local sequence (sequence of mean functions) to provide non-quadratic global approximation to $S_T(\theta)$:

$$S_T(\theta) = \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) \right]' W_T \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) \right]$$

Write

$$\begin{aligned} T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) &= T^{-1/2} \sum_{t=1}^{T} \left[\phi_t(\theta) - E\phi_t(\theta) \right] + T^{-1/2} \sum_{t=1}^{T} E\phi_t(\theta) \\ &= \Psi_T(\theta) + \sqrt{T}\, E\phi_t(\theta) \\ &= \Psi_T(\theta) + m_T(\theta) \end{aligned}$$

Weak ID asymptotics in GMM, ctd.

Applied to the linear IV regression model, this reorganization yields,

$$T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta) = T^{-1/2}\sum_{t=1}^{T}(y_t-\theta'Y_t)Z_t$$

$$= T^{-1/2}\sum_{t=1}^{T}\left(u_t-(\theta-\theta_0)'Y_t\right)Z_t$$

$$= T^{-1/2}\sum_{t=1}^{T}\zeta_t \; - E\left(T^{-1/2}\sum_{t=1}^{T}(\theta-\theta_0)'Y_tZ_t\right)$$

$$= \Psi_T(\theta) + m_T(\theta)$$

where $\zeta_t = u_tZ_t - \left[(\theta-\theta_0)'Y_tZ_t - E(\theta-\theta_0)'Y_tZ_t\right]$. Now:

- $\Psi_T(\theta) = T^{-1/2}\sum_{t=1}^{T}\zeta_t \xrightarrow{d}$ N(0, Ω) (because ζ_t is mean zero and i.i.d. – instrument strength doesn't enter this limit (subtracted out))
- The mean function $m_T(\theta)$ is a finite nonrandom (linear) function under the local nesting $\Pi = T^{-1/2}C$

Weak ID asymptotics in GMM, ctd.

$$T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta) = T^{-1/2}\sum_{t=1}^{T}\left[\phi_t(\theta)-E\phi_t(\theta)\right]+T^{-1/2}\sum_{t=1}^{T}E\phi_t(\theta) = \Psi_T(\theta) + m_T(\theta)$$

Suppose:

1. $m_T \xrightarrow{p} m$ uniformly in θ, where $m(\theta)$ is a limiting (finite continuous differentiable) function.

   This is the extension to a function of assuming $\Pi = T^{-1/2}C$

2. $\Psi_T(\bullet) \Rightarrow \Psi(\bullet)$, where $\Psi(\theta)$ is a Gaussian stochastic process on Θ with mean zero and covariance function $\Omega(\theta_1,\theta_2) = E\,\Psi(\theta_1)\,\Psi(\theta_2)'$

Weak ID asymptotics in GMM, ctd.

2. $\Psi_T \Rightarrow \Psi$, where $\Psi(\theta)$ is a Gaussian stochastic process on Θ with mean zero and covariance function $\Omega(\theta_1,\theta_2) = E\Psi(\theta_1)\Psi(\theta_2)'$

Digression on $\Psi_T \Rightarrow \Psi$:

Item #2 is an extension of the FCLT. Generally, the FCLT talks about convergence in distribution of a sequence of random functions, to a limiting function, which has a (limiting) distribution. In the more familiar time series FCLT, the function is indexed by s = τ/T $\in$ [0,1], and the limiting process has the covariance matrix of Brownian motion (it is Brownian motion). Here, the function is indexed by θ, and the limiting process has the covariance matrix $\Omega(\theta_1,\theta_2)$. The proof of the FCLT entails proving:

Weak ID asymptotics in GMM, ctd.

(a) *Convergence of finite dimensional distributions*. Here, this corresponds the joint distributions of $\Psi_T(\theta_1)$, $\Psi_T(\theta_2)$,…, $\Psi_T(\theta_r)$. But $\Psi_T(\theta) = T^{-1/2}\sum_{t=1}^{T}\left[\phi_t(\theta) - E\phi_t(\theta)\right]$, so it is a weak (standard) assumption that $\Psi_T(\theta_1)$, $\Psi_T(\theta_2)$,…, $\Psi_T(\theta_r)$ will converge jointly to a normal; the covariance matrix is filled out using $\Omega(\theta_1,\theta_2)$ (applied to all the points).

(b) *Tightness (or stochastic equicontinuity)*. That is, for θ_1 and θ_2 close, that $\Psi_T(\theta_1)$ and $\Psi_T(\theta_2)$ must be close (with high probability). This allows going from the function evaluated at finitely many points, to the function itself. Proving this is application specific (depends on $h(Y_t,\theta)$). Proof in the linear GMM case is in Stock and Wright (2000).

Weak ID asymptotics in GMM, ctd.

Back to main argument…

Under 1 and 2, $$T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta) \Rightarrow \Psi(\theta) + m(\theta)$$

3. $W_T(\theta) \xrightarrow{p} W(\theta)$ uniformly in θ, where $W(\theta)$ is psd, continuous in θ

Under 1, 2, and 3, $$S_T(\theta) = \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta)\right]' W_T(\theta)\left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta)\right]$$

$$\Rightarrow S(\theta) = [\psi(\theta) + m(\theta)]'W[\psi(\theta) + m(\theta)]$$

and

$$\hat{\theta} \Rightarrow \theta^*, \text{ where } \theta^* = \operatorname{argmin} S(\theta)$$

Weak ID asymptotics in GMM, ctd.

$$\hat{\theta} \Rightarrow \theta^* = \text{argmin } \{S(\theta) = [\Psi(\theta) + m(\theta)]'W[\Psi(\theta) + m(\theta)]\}$$

Comments

- With $\phi_t(\theta) = (y_t - \theta Y_t)Z_t$ and $W_T = (\mathbf{Z'Z}/T)^{-1}$, this yields the weak IV asymptotic distribution of TSLS obtained earlier.
- $S_T(\theta)$ is not well approximated by a quadratic (is not quadratic in the limit) with a nonrandom curvature matrix that gets large – instead, $S_T(\theta)$ is $O_p(1)$
- $\hat{\theta}$ is not consistent in this setup
- $\hat{\theta}$ has a nonstandard limiting distribution
- Standard errors of $\hat{\theta}$ aren't meaningful ($\pm 1.96SE$ isn't valid conf. int.)
- J-statistic doesn't have chi-squared distribution
- Well-identified elements of $\hat{\theta}$ have the usual limiting normal distributions, under the true values of the weakly identified elements
- Extensions and proofs are in Stock and Wright (2000)
- What about intermediate "semi-strong" cases? Chen and Andrews (2012)

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) **GMM II: Detection of weak identification**
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) Many instruments

7) GMM II: Detection of weak identification

This is an open area of research with no best solution. Some thoughts:

1. In linear GMM, the noncentrality parameter of the ***first-stage F*** and the concentration parameter are no longer the same thing if there is heteroskedasticity and/or serial correlation in $h(Y_t,\theta)$. With heteroskedasticity, the first-stage F still provides a reasonable guide (MC findings) but with serial correlation the first stage F isn't very reliable.

2. ***Wright (2003)*** provides a test for weak instruments, based on the extension of the Cragg-Donald (1993) using the estimated curvature of the objective function. The test is a test of non-identification (contrast with Stock-Yogo, testing whether μ^2 exceeds a critical cutoff; in

Wright (2003), the cutoff is taken to be $\mu^2 = 0$ in linear IV case). The test is conservative, which gives it low power against weak identification – a benefit in this instance. Important drawback is that it is only local (multiple peak problem).

3. ***Some symptoms of weak identification*:**

- CUE, two-step, and iterated GMM converge to quite different values (see Hansen, Heaton, Yaron (1996) MC results)
- for two-step and iterated, the normalization matters
- multiple valleys in the CUE objective function
- Significant discrepancies between GMM-AR confidence sets (discussed below) and conventional Wald confidence sets

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) **GMM III: Hypothesis tests and confidence intervals**
9) GMM IV: Estimation
10) Many instruments

8) GMM III: Hypothesis tests and confidence intervals

Extensions of methods in linear IV:

(1) The GMM-Anderson Rubin statistic

(Kocherlakota (1990); Burnside (1994), Stock and Wright (2000)) The extension of the AR statistic to GMM is the CUE objective function evaluated at θ_0:

$$S_T^{CUE}(\theta_0) = \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta_0)\right]'\hat{\Omega}(\theta_0)^{-1}\left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta_0)\right]$$

$$\xrightarrow{d} \psi(\theta_0)'\Omega(\theta_0)^{-1}\Psi(\theta_0) \sim \chi_k^2$$

- Thus a valid test of H_0: $\theta = \theta_0$ can be undertaken by rejecting if $S_T(\theta_0) > 5\%$ critical value of χ_k^2.

The GMM-Anderson Rubin statistic, ctd

- The statistic above tests all elements of θ. If some elements are strongly identified, they can be concentrated out (estimated under the null) for valid subset inference. Specifically, let $\theta = (\alpha, \beta)$, and let α be weakly identified and β be strongly identified. Fix α at the hypothesized value α_0 and let $\hat{\beta}^{GMM}$ be an efficient GMM estimator of β, at the given value of α_0. Then construct the CUE objective function, using the hypothesized value of α and the estimated value of β:

$$S_T^{CUE}(\alpha_0, \hat{\beta}^{GMM}) = \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\alpha_0, \hat{\beta}^{GMM}) \right]' \hat{\Omega}(\alpha_0, \hat{\beta}^{GMM})^{-1} \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\alpha_0, \hat{\beta}^{GMM}) \right]$$

The statistic $S_T^{CUE}(\alpha_0, \hat{\beta}^{GMM})$ has a $\chi^2_{k-\dim(\beta)}$ distribution under H_0: $\alpha = \alpha_0$, and is a weak-identification robust test statistic for H_0: $\alpha = \alpha_0$.

GMM-Anderson-Rubin, ctd.

In the homoskedastic linear IV model, the GMM-AR statistic simplifies to the AR statistic (up to a degrees of freedom correction):

$$S_T^{CUE}(\theta_0) = \left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta_0)\right]'\hat{\Omega}(\theta_0)^{-1}\left[T^{-1/2}\sum_{t=1}^{T}\phi_t(\theta_0)\right]$$

$$= \left[T^{-1/2}\sum_{t=1}^{T}(y_t-\theta_0'Y_t)Z_t\right]'\left(\frac{\mathbf{Z}'\mathbf{Z}}{T}s_v^2\right)^{-1}\left[T^{-1/2}\sum_{t=1}^{T}(y_t-\theta_0'Y_t)Z_t\right]$$

$$= \frac{(\mathbf{y}-\mathbf{Y}\theta_0)'P_{\mathbf{Z}}(\mathbf{y}-\mathbf{Y}\theta_0)}{(\mathbf{y}-\mathbf{Y}\theta_0)'M_{\mathbf{Z}}(\mathbf{y}-\mathbf{Y}\theta_0)/(T-k)} = k\times \mathrm{AR}(\theta_0)$$

Comments:

- The statistic, $S_T^{CUE}(\theta_0)$, is called various things in the literature, including the *S*-statistic, the CUE objective function statistic, the nonlinear AR statistic, and the GMM-AR statistic. I think GMM-AR is the most descriptive and we will use that term here.

GMM-Anderson-Rubin, ctd.

- The GMM-AR statistic has the same issues of interpretation issues as the AR, specifically, the GMM-AR rejects because of endogenous instruments and/or incorrect θ
- With little information, the GMM-AR can fail to reject any values of θ (remember the Dufour (1997) critique of Wald tests)

(2) GMM-LM

Kleibergen (2005) – develops score statistic (based on CUE objective function – details of construction matter) that provides weak-identification valid hypothesis testing for sets of variables

(3) GMM-CLR

Andrews, Moreira, Stock (2006) – extension of CLR to linear GMM with a single included endogenous regressor, also see Kleibergen (2007). Very limited evidence on performance exists; also problem of dimension of conditioning vector

(4) Other methods

Guggenberger-Smith (2005) objective-function based tests based on Generalized Empirical Likelihood (GEL) objective function (Newey and Smith (2004)); Guggenberger-Smith (2008) generalize these to time series data. Performance is similar to CUE (asymptotically equivalent under weak instruments)

Confidence sets

- Fully-robust 95% confidence sets are obtained by inverting (are the acceptance region of) fully-robust 5% hypothesis tests
- Computation is by grid search in general: collect all the points θ which, when treated as the null, are not rejected by the GMM-AR statistic.
- Subsets by projection (see Kleibergen and Mavroeidis (2009) for an application of GMM-AR confidence sets and subsets)
- Valid tests must be unbounded (contain Θ) with finite probability with weak instruments

Bottom line recommendation

Work is under way in this area, but the best thing for now is to use the GMM-AR statistic to test $\theta = \theta_0$, and to invert the GMM-AR statistic to construct the GMM version of the AR confidence set. The GMM-AR statistic must in general be inverted by grid search. The GMM-AR confidence set, if nonempty, will contain the CUE estimator.

Example (linear GMM): New Keynesian Phillips Curve

See the survey by Mavroeidis, Plagborg-Møller, and Stock (2014)

Hybrid NKPC: $\pi_t = \lambda x_t + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} + \eta_t$

Rational expectations: $E_t(\pi_t - \lambda x_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1}) = 0$

GMM moment condition: $E[(\pi_t - \gamma_f \pi_{t+1} - \gamma_b \pi_{t-1} - \lambda x_t) Z_t] = 0$

Instruments: $Z_t = \{\pi_{t-1}, x_{t-1}, \pi_{t-2}, x_{t-2}, \ldots\}$

$m = 2$, so AR sets are needed. Confidence intervals can be computed by projecting the sets to the axes.

minev(μ^2) = 1.8 minev(μ^2) = 108

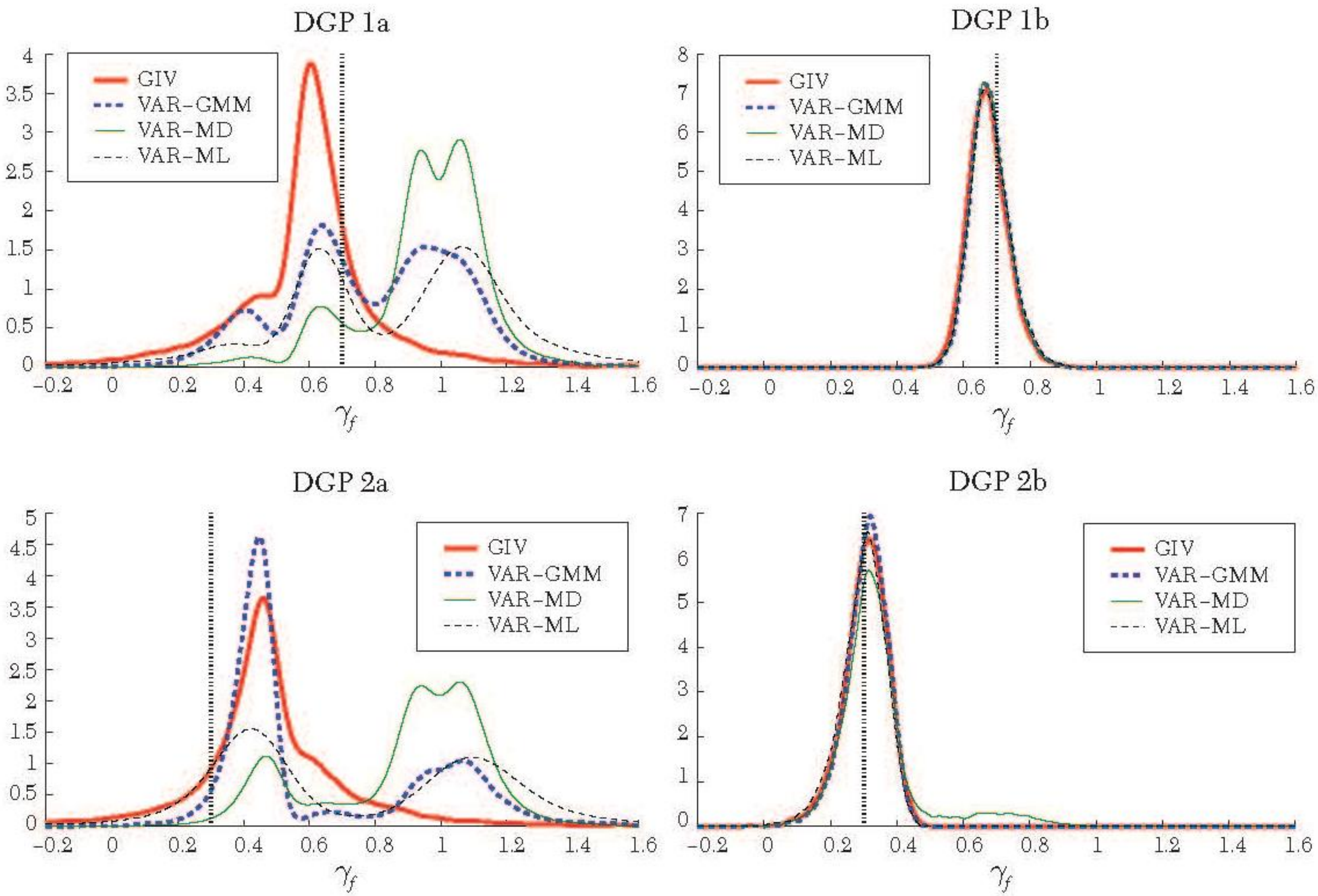


Figure 2. Sampling Distribution of γ_f Estimators

Notes: Kernel-smoothed density estimates of the sampling distribution of γ_f estimators in the hybrid NKPC model (21) for the DGPs listed in table 1. The dotted vertical line marks the true parameter value.

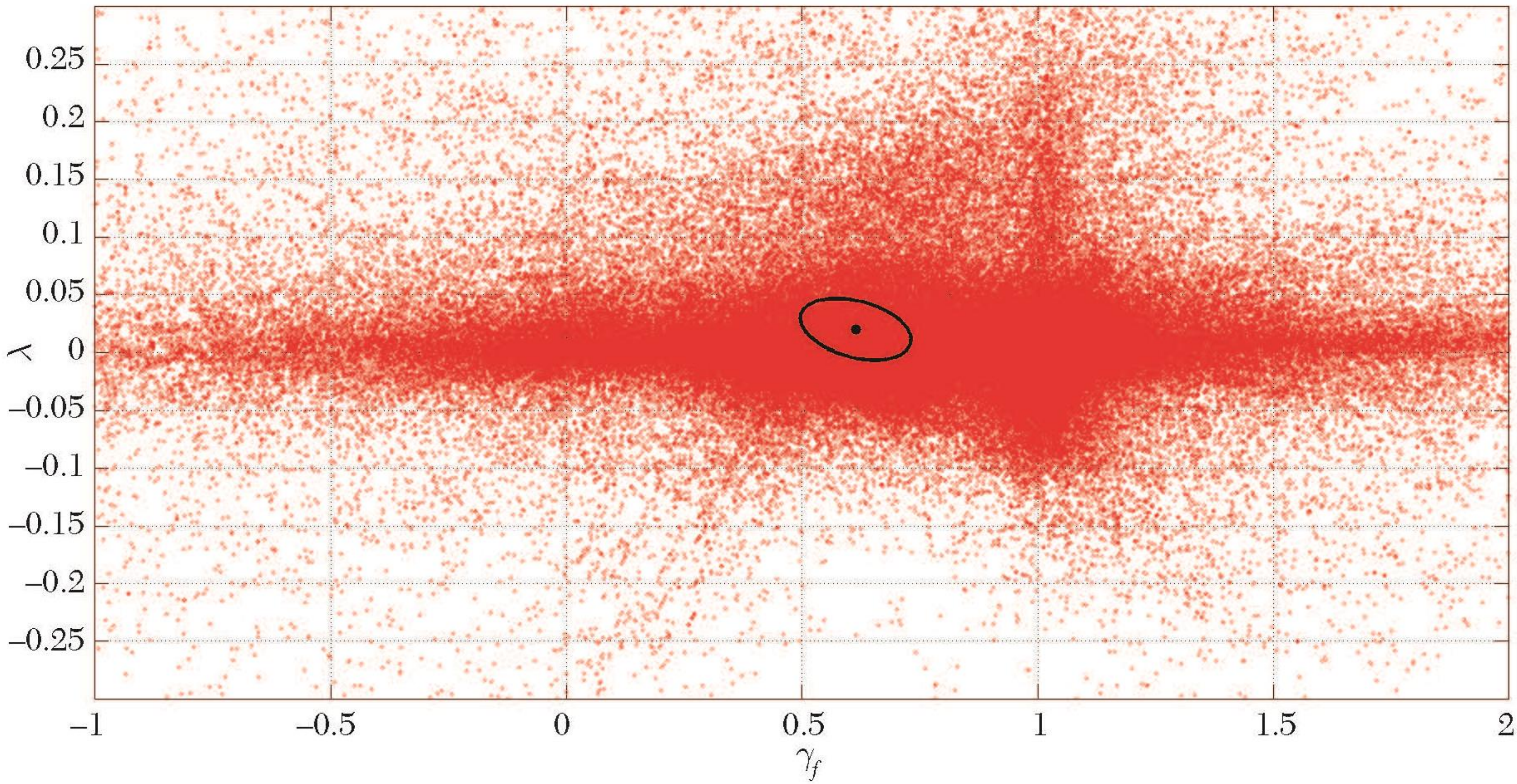
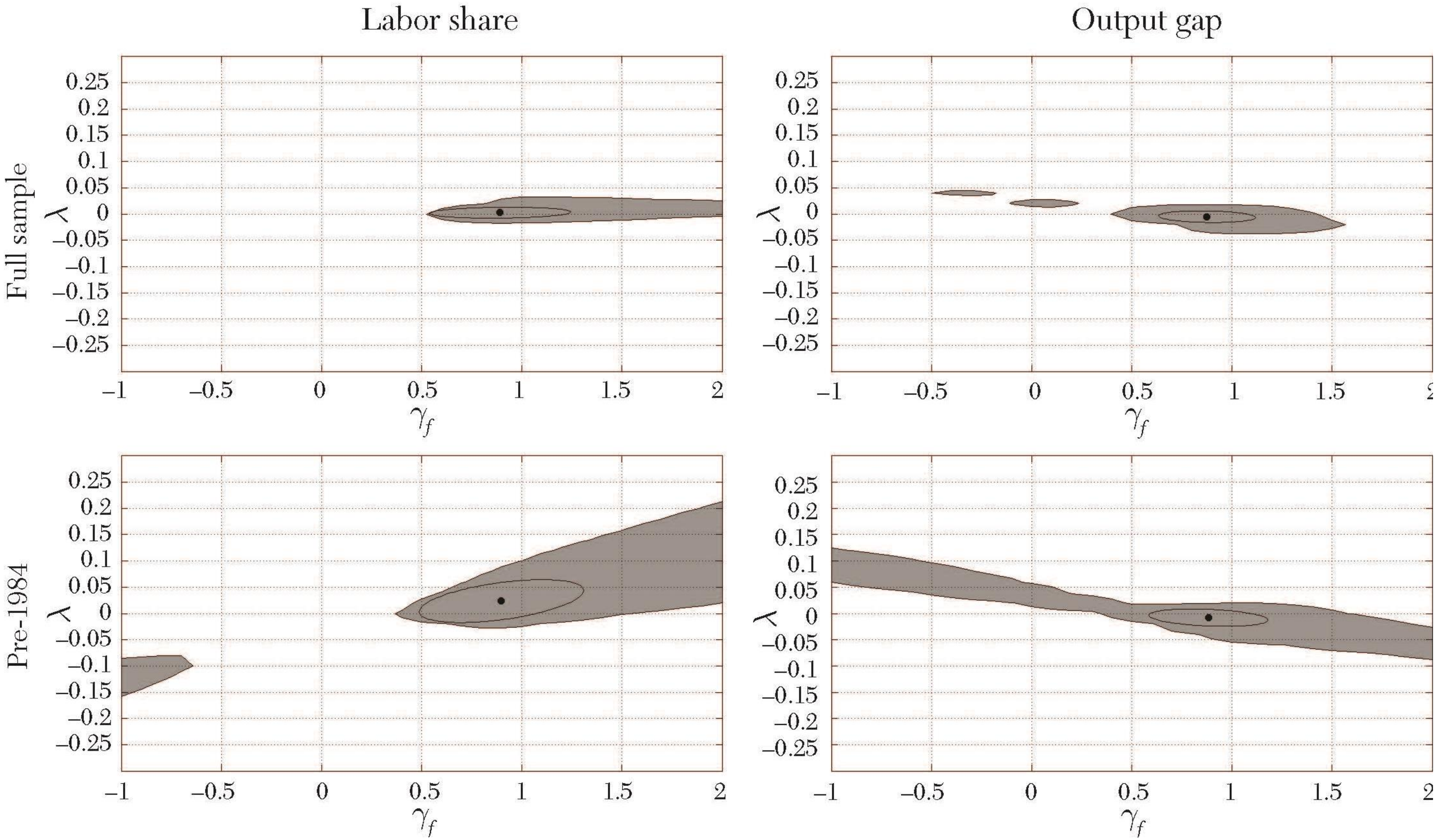


Figure 4. Point Estimates: Labor Share Specifications

Notes: Point estimates of λ, γ_f from the various specifications listed in table 4 that use the labor share as forcing variable, excluding real-time and survey instrument sets. The black dot and ellipse represent the point estimate and 90 percent joint Wald condence set from the 1998 vintage results in table 3.


Labor share
Output gap
Full sample
Pre-1984
λ
γ_f
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
−0.25
−1
−0.5
0
0.5
1
1.5
2

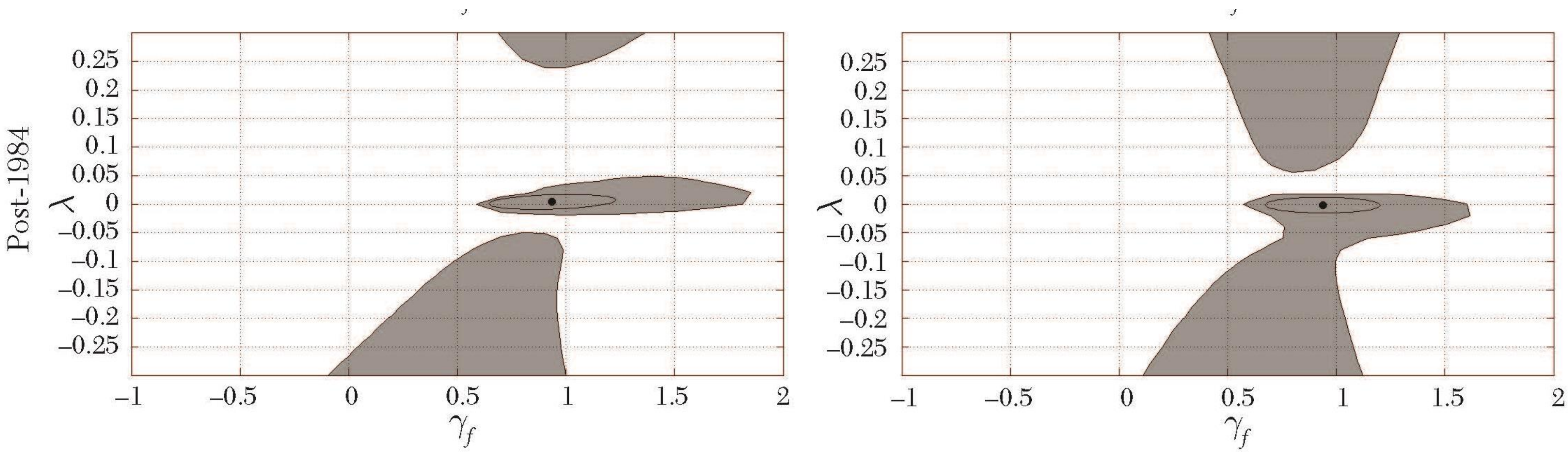


Figure 11. Robust Confidence Regions: RE Specifications

Notes: 90 percent S set (gray), 90 percent Wald ellipse, and CUE GMM point estimate (bullet) of the coefficients of the labor share and future inflation in the hybrid NKPC specification with one lag of inflation, where inflation coefficients sum to one. Inflation: GDP deflator. Forcing variable: NFB labor share (left panels), CBO output gap (right panels). Instruments: three lags of $\Delta\pi_t$ and the forcing variable. Sample: starts 1948q2 (labor share), 1949q4 (output gap), ends 2011q3; full sample (top row), pre-1983q4 (middle row), post-1984q1 (bottom row). Weight matrix: Newey–West with automatic lag truncation.

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) **GMM IV: Estimation**
10) Many instruments

9) GMM IV: Estimation

- Impossibility of a (data-based) fully robust estimators are available – just as in linear case
- The challenge is to find partially robust estimators – estimators that improve upon 2-step and iterated GMM (which perform terribly – just like TSLS)

<u>(a) The continuous updating estimator (CUE)</u>

Hansen, Heaton, Yaron (1996). The CUE minimizes,

$$S_T^{CUE}(\theta) = \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) \right]' \hat{\Omega}(\theta)^{-1} \left[T^{-1/2} \sum_{t=1}^{T} \phi_t(\theta) \right]$$

Basic idea: "same θ in the numerator and the denominator".

CUE, ctd

Comments

- The CUE might seem arbitrary but actually it isn't. In fact, it was shown above that in the linear model with spherical errors, the CUE objective function *is* the AR statistic, $S_T^{CUE}(\theta)$ = AR(θ). It was stated above (without proof) that LIML minimizes the AR statistic. So in the special case of linear GMM when there is no heteroskedasticity or serial correlation, the CUE estimator is LIML (asymptotically under weak instrument asymptotics if Ω is estimated).
- CUE will always be contained in the GMM-AR set
- The CUE seems to inherit median unbiasedness of LIML (MC result; for some theory see Hausman, Menzel, Lewis, and Newey (2007))
- CUE (like LIML) exhibits wide dispersion in MC studies (Guggenberger 2005)

(b) Other estimators

- Generalized empirical likelihood (GEL) family. Interestingly, GEL estimators are asymptotically equivalent to CUE under weak instrument asymptotics (Guggenberger and Smith (2005))
- Fuller-*k* type modifications explored in Hausman, Menzel, Lewis, and Newey (2007), with some simulation evidence.
- These alternative estimators are promising but preliminary and their properties, including the extent to which they are robust to weak instruments in practice, are not yet fully understood.

Outline

1) What is weak identification, and why do we care?
2) Classical IV regression I: Setup and asymptotics
3) Classical IV regression II: Detection of weak instruments
4) Classical IV regression III: hypothesis tests and confidence intervals
5) Classical IV regression IV: Estimation
6) GMM I: Setup and asymptotics
7) GMM II: Detection of weak identification
8) GMM III: Hypothesis tests and confidence intervals
9) GMM IV: Estimation
10) **Many instruments**

10) Many Instruments

<u>The appeal of using many instruments</u>

- Under standard IV asymptotics, more instruments means greater efficiency.
- This story is not very credible because

  (a) the instruments you are adding might well be weak (you already have used the first two lags, say) and

  (b) even if they are strong, this requires consistent estimation of increasingly many parameter to obtain the efficient projection – hence slow rates of growth of the number of instruments in efficient GMM literature.

Example of problems with many weak instruments – TSLS

Recall the TSLS weak instrument asymptotic limit:

$$\hat{\beta}^{TSLS} - \beta_0 \xrightarrow{d} \frac{(\lambda + z_v)' z_u}{(\lambda + z_v)'(\lambda + z_v)}$$

with the decomposition, $z_u = \delta z_v + \eta$. Suppose that k is large, and that $\lambda'\lambda/k \to \Lambda_\infty$ (one way to implement "many weak instrument asymptotics"). Then as $k \to \infty$,

$\lambda' z_v/k \xrightarrow{p} 0$ and $\lambda' z_u/k \xrightarrow{p} 0$

$z_v' z_v/k \xrightarrow{p} 1$ and $z_v'\eta/k \xrightarrow{p} 0$ (z_v and η are independent by construction)

Putting these limits together, we have, as $k \to \infty$,

$$\frac{(\lambda + z_v)' z_u}{(\lambda + z_v)'(\lambda + z_v)} \xrightarrow{p} \frac{\delta}{1 + \Lambda_\infty}$$

In the limit that $\Lambda_\infty = 0$, as $k \to \infty$ TSLS is <u>*consistent*</u> for the *plim* of OLS!

Comments

- This calculation cuts a corner – it uses sequential asymptotics ($T \to \infty$, then $k \to \infty$). However the sequential asymptotics is justified under certain (restrictive) conditions on K/T (specifically, $k^4/T \to 0$)
- Typical conditions on k are $k^3/T \to 0$ (e.g. Newey and Windmeijer (2004))
- Many instruments can be turned into a blessing (if they are not too weak! They can't push the scaled concentration parameter to zero) by exploiting the additional convergence across instruments. This can lead to bias corrections and corrected standard errors. There is no single best method at this point but there is promising research, e.g. Newey and Windmeijer (2004), Chao and Swanson (2005), and Hansen, Hausman, and Newey (2006))

Comments, ctd.

- For testing, the AR, LM, and CLR are all valid under many instruments (again, slow rate: $k \to \infty$ but $k^3/T \to 0$) in the classical IV regression model; the CLR continues to be essentially most powerful (the power of the AR deteriorates substantially because of the large number of restrictions being tested)
- An important caveat in all of this is that the rates suggest that the number of instruments must be quite small compared to the number of observations. (The specific rate at which you can add instruments depends on their strength – the stronger the instruments, the more you can add; see the discussion in Hansen, Hausman, and Newey (2006) for example.) Consider the $k^3/T \to 0$ rate:

  with T = 200 and k = 6, k^3/T = 1.08.

  with T = 329,509 and k = 178, k^3/T = 17 (!)

Instrument selection

- Donald and Newey (2001) provide an information criterion instrument selection method in the classical linear IV model that applies when some instruments are strong (θ strongly identified) and others possibly weak. Problem with is that you need to know which are strong.
- Unaware of instrument selection methods that are appropriate when all instruments are possibly weak.

Final comments on many instruments

- Strong instruments: more instruments, more efficiency
- Weak instruments: more weak instruments, less reliable inference – more bias, size distortions (using standard estimators – two-step and iterated GMM)
- Don't be fooled by standard errors that get smaller as you add instruments.

  Remember the result that $\hat{\beta}^{TSLS} - \hat{\beta}^{OLS} \xrightarrow{p} 0$ as $k \to \infty$ (and $k^3/T \to 0$) when all but a few instruments are irrelevant.
- Some gains seem to be possible in theory (papers cited above) by exploiting the idea of many instruments but the theory is delicate: bias adjustments and size corrections that hold for rates such as $k \to \infty$ but $k^3/T \to 0$, but break down for k too large. Work needs to be done before these are ready for implementation
- For now, the best advice is to restrict attention to relatively few instruments, to use judgment selecting the strongest (recent lags, not distant ones), and to use relatively well understood.

Bottom line recommendations

- Weak instruments/weak identification comes up in a lot of applications
- In the linear case, it is helpful to check the first-stage F to see if weak instruments are plausibly a problem.
- TSLS and 2-step efficient GMM can give highly misleading estimates if instruments are weak.
- TSLS and 2-step GMM confidence intervals, constructed in the usual way ($\pm$ 1.96 standard errors) are highly unreliable (can have very low true coverage rates) if instruments are weak.
- If you have weak instruments, the best thing to do is to get stronger instruments, but barring that you should use econometric procedures that are robust to weak instruments. Robust procedures give valid inference even if the instruments are weak.

Bottom line recommendations, ctd.

- In the linear case with m=1 and no serial correlation, the CLR and CLR confidence intervals are recommended. Estimation by LIML is preferred to TSLS, but LIML can deliver very large outliers. Fuller is also a plausible option (see above).
- In the general nonlinear GMM case, GMM-AR confidence sets are recommended, but care must be taken in interpreting these (see discussion above). If you must compute an estimator, CUE seems to be the best choice given the current state of knowledge.

AEA Continuing Education Course

Time Series Econometrics

Lecture 7

Structural Vector Autoregressions: Recent Developments

James H. Stock

Harvard University

January 6 & 7, 2015

Outline

1) VARs, SVARs, and the Identification Problem

2) Classical approaches to identification

  2a) Identification by Short Run Restrictions

  2b) [Identification by Long Run Restrictions]

3) New approaches to identification (post-2000)

  3a) Identification from Heteroskedasticity

  3b) Direct Estimation of Shocks from High Frequency Data

  3c) External instruments

  3d) Identification by Sign Restrictions

1) VARs, SVARs, and the Identification Problem

A classic question in empirical macroeconomics: what is the effect of a policy intervention (interest rate increase, fiscal stimulus) on macroeconomic aggregates of interest – output, inflation, etc?

Let Y_t be a vector of macro time series, and let ε_t^r denote an unanticipated monetary policy intervention. We want to know the ***dynamic causal effect*** of ε_t^r on Y_t:

$$\frac{\partial Y_{t+h}}{\partial \varepsilon_t^r},\ h = 1, 2, 3,\ldots.$$

where the partial derivative holds all other interventions constant. In macro, this dynamic causal effect is called the ***impulse response function*** (***IRF***) of Y_t to the "shock" (unexpected intervention) ε_t^r.

The challenge is to estimate $\left\{\frac{\partial Y_{t+h}}{\partial \varepsilon_t^r}\right\}$ from observational macro data.

Two conceptual approaches to estimating dynamic causal effects (*IRF*)

1) Structural model (Cowles Commission): DSGE or SVAR
2) Quasi-Experiments

The identification problem. *Consider the Reduced form VAR(p)*:

$$Y_t = A_1 Y_{t-1} + \dots + A_p Y_{t-p} + u_t$$

or $\quad A(L)Y_t = u_t$, where $A(L) = I - A_1 L - A_2 L^2 - \dots - A_p L^p$

where A_i are the coefficients from the (population) regression of Y_t on $Y_{t-1},\dots,Y_{t-p}$.

- $u_t = Y_t - \text{Proj}(Y_t|Y_{t-1},\dots, Y_{t-p})$ are the innovations, and are identified.
- If u_t were the shocks, then we could compute the structural IRF using the MA representation of the VAR, $Y_t = A(L)^{-1}u_t$.
- But in general u_t is affected by multiple shocks: in any given quarter, GDP changes unexpectedly for a variety of reasons.
- For example, if $n = 2$,

$$u_{1t} = \mathbf{R_{12}}\boldsymbol{u_{2t}} + \varepsilon_{1t}$$

$$u_{2t} = \mathbf{R_{21}}\boldsymbol{u_{1t}} + \varepsilon_{2t}$$

  - To identify R we need an instrument Z_t or a restriction on the parameters.
  - For example, $R_{12} = 0$ identifies R (Cholesky decomposition)

Reduced form to structure:

Suppose: (i) A(L) is finite order p (known or knowable)

(ii) u_t spans the space of structural shocks ε_t, that is, $\varepsilon_t = \mathrm{R}u_t$, where R is square (equivalently, Y_t is linear in the structural shocks & the model is invertible)

(iii) A(L), Σ_u , and R are time-invariant, e.g. A(L) is invariant to policy changes over the relevant period

Because $\varepsilon_t = \mathrm{R}u_t$,

$$\mathrm{RA(L)}Y_t = \mathrm{R}u_t = \varepsilon_t.$$

Letting RA(L) = B(L), this delivers the structural VAR,

$$\mathrm{B(L)}Y_t = \varepsilon_t,$$

The MA representation of the SVAR delivers the structural IRFs:

$$Y_t = D(\mathrm{L})\varepsilon_t,\ D(\mathrm{L}) = \mathrm{B(L)}^{-1} = \mathrm{A(L)}^{-1}\mathrm{R}^{-1}$$

Impulse response: $$\frac{\partial Y_{t+h}}{\partial \varepsilon_t} = D_h$$

Summary of VAR and SVAR notation

Reduced form VAR	Structural VAR
$A(L)Y_t = u_t$ $Y_t = A(L)^{-1}u_t = C(L)u_t$ $A(L) = I - A_1L - A_2L^2 - \ldots - A_pL^p$ $Eu_tu_t' = \Sigma_u$ (unrestricted)	$B(L)Y_t = \varepsilon_t$ $Y_t = B(L)^{-1}\varepsilon_t = D(L)\varepsilon_t$ $B(L) = B_0 - B_1L - B_2L^2 - \ldots - B_pL^p$ $E\varepsilon_t\varepsilon_t' = \Sigma_\varepsilon = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_k^2 \end{pmatrix}$
$Ru_t = \varepsilon_t$ $B(L) = RA(L) \quad (B_0 = R)$ $D(L) = C(L)R^{-1}$	

- Note the assumption that the structural shocks are uncorrelated
- D(L) is the structural IRF of Y_t w.r.t. ε_t.
- structural forecast error variance decompositions are computed from D(L) and Σ_ε

Identification of R and identification of shocks: Two equivalent views

1. ***Identification of R.*** In population, we can know A(L). If we can identify R, we can obtain the SVAR coefficients, B(L) = RA(L).

2. ***Identification of shocks.*** If you knew (or could estimate) one of the shocks, you could estimate the structural IRF of Y w.r.t. that shock. Partition Y_t into a policy variable r_t and all other variables:

$$Y_t = \begin{pmatrix} \overset{(k-1\times 1)}{X_t} \\ \overset{(1\times 1)}{r_t} \end{pmatrix},\ u_t = \begin{pmatrix} u_t^X \\ u_t^r \end{pmatrix},\ \varepsilon_t = \begin{pmatrix} \varepsilon_t^X \\ \varepsilon_t^r \end{pmatrix},$$

The IRF/MA form is $Y_t = \mathrm{D(L)}\varepsilon_t$, or

$$Y_t = \begin{pmatrix} D_{YX}(L) & D_{Yr}(L) \end{pmatrix} \begin{pmatrix} \varepsilon_t^X \\ \varepsilon_t^r \end{pmatrix} = D_{Yr}(\mathrm{L})\varepsilon_t^r + v_t,$$

where $v_t = D_{\mathrm{Y}X}(\mathrm{L})\varepsilon_t^X$. Because $E\varepsilon_t^r v_t = 0$, the IRF of Y_t w.r.t. ε_t^r, $D_{Yr}(\mathrm{L})$ is identified by the population OLS regression of Y_t onto ε_t^r.

A word on "invertibility":

Recall the SVAR assumption:

(ii) u_t spans the space of structural shocks ε_t, that is, $\varepsilon_t = Ru_t$, where R is square

- This is often called the assumption of invertibility: the VAR can be inverted to span the space of structural shocks. If there are more structural shocks than u_t's, then condition (ii) will not hold.
- One response is to add more variables so that u_t spans ε_t. This response is an important motivation of the FAVAR approach (references below)
- If agents see future shocks, invertibility fails. Or, does the definition of shock just become more subtle (an expectations shock)?
- See Lippi and Reichlin (1993, 1994), Sims and Zha (2006b), Fernandez-Villaverde, Rubio-Ramirez, Sargent, and Watson (2007), Hansen and Sargent (2007), E. Sims (2012), Blanchard, L'Huillier, and Lorenzoni (2012), Forni, Gambetti, and Sala (2012), and Gourieroux and Monfort (2014)

This talk

- Early promise of SVARs

  Surveys of classical methods: Christiano, Eichenbaum, and Evans (1999), Lütkepohl (2005), Stock and Watson (2001), Watson (1994)

  Survey of new ideas about how to tackle the identification problem

- Critiques of the 1990s

- This talk focuses on the interesting new work on identification – much of it quite recent – in response to those critiques

Outline

1) VARs, SVARs, and the Identification Problem
2) Classical approaches to identification
    2a) Identification by Short Run Restrictions
    2b) [Identification by Long Run Restrictions]
3) New approaches to identification (post-2000)
    3a) Identification from Heteroskedasticity
    3b) Direct Estimation of Shocks from High Frequency Data
    3c) External instruments
    3d) Identification by Sign Restrictions

2a) Identification by Short Run Restrictions

<u>Overview: the traditional SVAR identification approach</u>

Bernanke (1986), Blanchard and Watson (1986), Sims (1986)

***(a) 2-variable example*.**

$$u_{1t} = \mathbf{R_{12}}\boldsymbol{u_{2t}} + \varepsilon_{1t}$$

$$u_{2t} = \mathbf{R_{21}}\boldsymbol{u_{1t}} + \varepsilon_{2t}$$

- Suppose $R_{12} = 0$. E.g. Blanchard and Galí (2007) for oil price shocks.
- Then $\varepsilon_{1t} = u_{1t}$ so R_{21} can be estimated by OLS (u_{1t} is uncorrelated with ε_{2t}).
- How credible is the Blanchard-Galí assumption?

(b) System identification. In general, the SVAR is fully identified if

$$R\Sigma_u R' = \Sigma_\varepsilon$$

can be solved for the unknown elements of R and Σ_ε.. Recall that Σ_u is identified.

- There are $k(k+1)/2$ distinct equations in the matrix equation above, so the order condition says that you can estimate (at most) $k(k+1)/2$ parameters.
- If we set $\Sigma_\varepsilon = I$ (just a normalization), there are k^2 parameters
- So we need $k^2 - k(k+1)/2 = k(k-1)/2$ restrictions on R.
- If $k = 2$, then $k(k-1)/2 = 1$, which is delivered by imposing a single restriction (commonly, that R is lower or upper triangular).
- This ignores rank conditions, which can matter.
- This description of identification is via method of moments, however identification can equally be described via IV, e.g. see Blanchard and Watson (1986).

(c) Identification of only one shock or IRF. Many applications now take a limited information approach, in which only a row of R is identified. Partition $\varepsilon_t = Ru_t$, and partition Y_t so that:

$$\begin{pmatrix} \varepsilon_t^X \\ \varepsilon_t^r \end{pmatrix} = \begin{pmatrix} R_{XX} & R_{Xr} \\ R_{rX} & R_{rr} \end{pmatrix} \begin{pmatrix} u_t^X \\ u_t^r \end{pmatrix}$$

If R_{rX} and R_{rr} are identified, then (in population) ε_t^r can be computed using just the final row and $D_{Yr}(L)$ can be computed by the regression of Y_t on $\varepsilon_t^r, \varepsilon_{t-1}^r, \ldots$.

(d) The "fast-r-slow" scheme. Almost all short-run restriction applications can be written as "fast-r-slow." Following CEE (1999), the benchmark timing identification assumption is

$$\begin{pmatrix} \varepsilon_t^S \\ \varepsilon_t^r \\ \varepsilon_t^f \end{pmatrix} = \begin{pmatrix} R_{SS} & 0 & 0 \\ R_{rS} & R_{rr} & 0 \\ R_{fS} & R_{fr} & R_{ff} \end{pmatrix} \begin{pmatrix} u_t^S \\ u_t^r \\ u_t^f \end{pmatrix} \text{ where } Y_t \text{ is partitioned } \begin{pmatrix} X_{St} \\ r_t \\ X_{ft} \end{pmatrix}$$

which identifies ε_t^r as the residual from regressing u_t^r on u_t^S.

<u>Selected criticisms of timing restrictions (Rudebusch (1998), others)</u>

- The implicit policy reaction function doesn't accord with theory or practical experience (does Fed ignore the stock market?)
- Implementations often ignore changes in policy reaction functions
- questionable credibility of lack of in-period response of X_{st} to r_t
- VAR information is typically far less than standard information sets
- Estimated monetary policy shocks don't match futures market data

Outline

1) VARs, SVARs, and the Identification Problem
2) Classical approaches to identification
    2a) Identification by Short Run Restrictions
    2b) [Identification by Long Run Restrictions]
3) New approaches to identification (post-2000)
    3a) Identification from Heteroskedasticity
    3b) Direct Estimation of Shocks from High Frequency Data
    3c) External instruments
    3d) Identification by Sign Restrictions

2b) [Identification by Long Run Restrictions]

This approach identifies R by imposing restrictions on the long run effect of one or more ε's on one or more Y's.

Reduced form VAR: $\quad A(L)Y_t = u_t$

Structural VAR: $\quad B(L)Y_t = \varepsilon_t, \quad Ru_t = \varepsilon_t, \quad B(L) = RA(L)$

Long run variance matrix from VAR: $\quad \Omega = A(1)^{-1}\Sigma_u A(1)^{-1\prime}$

Long run variance matrix from SVAR: $\quad \Omega = B(1)^{-1}\Sigma_\varepsilon B(1)^{-1\prime}$

Digression: $B(1)^{-1} = D(1)$ is the long-run effect on Y_t of ε_t; this can be seen using the Beveridge-Nelson decomposition,

$$\sum_{s=1}^{t} Y_s = D(1)\sum_{s=1}^{t}\varepsilon_s + D^*(L)\varepsilon_t, \text{ where } D_i^* = -\sum_{j=i+1}^{\infty} D_j$$

Notation: think of Y_t as being growth rates, e.g. if Y_t is employment growth, $\Delta \ln N_t$, then $\sum_{s=1}^{t} Y_s$ is log employment, $\ln N_t$

Long run restrictions, ctd.

From VAR: $\Omega = A(1)^{-1}\Sigma_u A(1)^{-1\prime}$

From SVAR: $\Omega = B(1)^{-1}\Sigma_\varepsilon B(1)^{-1\prime} = RA(1)^{-1}\Sigma_\varepsilon A(1)^{-1\prime}R'$

System identification by long run restrictions. The SVAR is identified if

$$RA(1)^{-1}\Sigma_\varepsilon A(1)^{-1\prime}R' = \Omega \qquad (*)$$

can be solved for the unknown elements of R and Σ_ε..

- There are $k(k+1)/2$ distinct equations in (*), so the order condition says that you can estimate (at most) $k(k+1)/2$ parameters. If we set $\Sigma_\varepsilon = I$ (just a normalization), it is clear that we need $k^2 - k(k+1)/2 = k(k-1)/2$ restrictions on R.
- If $k = 2$, then $k(k-1)/2 = 1$, which is delivered by imposing a single exclusion restriction (that is, R is lower or upper triangular).
- This ignores rank conditions, which matter
- This is a moment matching approach; an IV interpretation comes later

Long run restrictions, ctd.

The long run neutrality restriction. The main way long restrictions are implemented in practice is by setting $\Sigma_\varepsilon = I$ and imposing zero restrictions on D(1). Imposing $D_{ij}(1) = 0$ says that the effect the long-run effect on the i^{th} element of Y_t, of the j^{th} element of ε_t is zero

If $\Sigma_\varepsilon = I$, the moment equation above can be rewritten,

$$\Omega = \mathrm{D}(1)\mathrm{D}(1)'$$

where $\mathrm{D}(1) = \mathrm{B}(1)^{-1}$. Because $R\mathrm{A}(1) = \mathrm{B}(1)$, R is obtained from D(1) as $R = \mathrm{A}(1)^{-1}\mathrm{B}(1)$, and B(L) = RA(L) as above.

Comments:

- If the zero restrictions on D(1) make D(1) lower triangular, then D(1) is the Cholesky factorization of Ω.
- Blanchard-Quah (1989) had 2 variables (unemployment and output), with the restriction that the demand shock has no long-run effect on the unemployment rate. This imposed a single zero restriction, which is all that is needed for system identification when $k = 2$.
- King, Plosser, Stock, and Watson (1991) work through system and partial identification (identifying the effect of only some shocks), things are analogous to the partial identification using short-run timing.
- This approach was at the center of a debate about whether technology shocks lead to a short-run decline in hours, based on long-run restrictions (Galí (1999), Christiano, Eichenbaum, and Vigfusson (2004, 2006), Erceg, Guerrieri, and Gust (2005), Chari, Kehoe, and McGrattan (2007), Francis and Ramey (2005), Kehoe (2006), and Fernald (2007))
- More generally, the theoretical grounding of long-run restrictions is often questionable; for a case in favor of this approach, see Giannone, Lenza, and Primiceri (2014)

Long run restrictions, ctd.

In this literature, Ω is estimated using the VAR-HAC estimator,

VAR-HAC estimator of Ω: $\hat{\Omega} = \hat{A}(1)^{-1}\hat{\Sigma}_u\hat{A}(1)^{-1\prime}$

D(1) and R are estimated as: $\hat{D}(1) = \text{Chol}(\hat{\Omega}), \hat{R} = \left[\hat{D}(1)\hat{A}(1)\right]^{-1}$

Comments:

- A recurring theme is the sensitivity of the results to apparently minor specification changes, in Chari, Kehoe, and McGrattan's (2007) example results are sensitive to the lag length. It is unlikely that $\hat{\Sigma}_u$ is sensitive to specification changes, but $\hat{A}(1)$ is much more difficult to estimate.
- These observations are closely linked to the critiques by Faust and Leeper (1997), Pagan and Robertson (1998), Sarte (1997), Cooley and Dwyer (1998), Watson (2006), and Gospodinov (2008), which are essentially weak instrument concerns.
- One alternative is to use medium-run restrictions, see Uhlig (2004)

Outline

1) VARs, SVARs, and the Identification Problem

2) Classical approaches to identification

    2a) Identification by Short Run Restrictions

    2b) [Identification by Long Run Restrictions]

3) New approaches to identification (post-2000)

    3a) Identification from Heteroskedasticity

    3b) Direct Estimation of Shocks from High Frequency Data

    3c) External instruments

    3d) Identification by Sign Restrictions

3a) Identification from Heteroskedasticity

Suppose:

(a) The structural shock variance breaks at date s: $\Sigma_{\varepsilon,1}$ before, $\Sigma_{\varepsilon,2}$ after.

(b) R doesn't change between variance regimes.

(c) normalize R to have 1's on the diagonal, but no other restrictions; thus the unknowns are: R ($k^2–k$); $\Sigma_{\varepsilon,1}$ (k), and $\Sigma_{\varepsilon,2}$(k).

First period:	$R\Sigma_{u,1}R' = \Sigma_{\varepsilon,1}$	$k(k+1)/2$ equations, k^2 unknowns
Second period:	$R\Sigma_{u,2}R' = \Sigma_{\varepsilon,2}$	$k(k+1)/2$ equations, k more unknowns

Number of equations = $k(k+1)/2 + k(k+1)/2 = k(k+1)$

Number of unknowns = $k^2 – k + k + k = k(k+1)$

Rigobon (2003), Rigobon and Sack (2003, 2004)

ARCH version by Sentana and Fiorentini (2001)

Identification from Heteroskedasticity,ctd.

Comments:

1. There is a rank condition here too – for example, identification will not be achieved if $\Sigma_{\varepsilon,1}$ and $\Sigma_{\varepsilon,2}$ are proportional.
2. The break date need not be known as long as it can be estimated consistently
3. Different intuition: suppose only one structural shock is homoskedastic. Then find the linear combination without any heteroskedasticity!
4. This idea also can be implemented exploiting conditional heteroskedasticity (Sentana and Fiorentini (2001))
5. But, some cautionary notes:
    a. R must remain constant despite change in Σ_ε (think about it…)
    b. Strong identification will come from large differences in variances

Example: Wright (2012), Monetary Policy at ZLB

Outline

1) VARs, SVARs, and the Identification Problem
2) Classical approaches to identification
    2a) Identification by Short Run Restrictions
    2b) [Identification by Long Run Restrictions]
3) New approaches to identification (post-2000)
    3a) Identification from Heteroskedasticity
    3b) Direct Estimation of Shocks from High Frequency Data
    3c) External instruments
    3d) Identification by Sign Restrictions

3b) Direct Estimation of Shocks from High Frequency Data

Monetary shock application: Estimate ε_t^r directly from daily data on monetary announcements or policy-induced FF rate changes:

Recall,

$$Y_t = \begin{pmatrix} D_{YX}(L) & D_{Yr}(L) \end{pmatrix} \begin{pmatrix} \varepsilon_t^X \\ \varepsilon_t^r \end{pmatrix} = D_{Yr}(\mathrm{L})\varepsilon_t^r + v_t,$$

where $v_t = D_{YX}(\mathrm{L})\varepsilon_t^X$, so if you observed ε_t^r you could estimate $D_{Yr}(\mathrm{L})$.

- Cochrane and Piazessi (2002)

  aggregates daily ε_t^r (Eurodollar rate changes after FOMC announcements) to a monthly ε_t^r series

- Faust, Swanson, and Wright (2003, 2004)

  estimates IRF of r_t wrt ε_t^r from futures market, then matches this to a monthly VAR IRF (results in set identification – discuss later)

- Bernanke and Kuttner (2005)

Outline

1) VARs, SVARs, and the Identification Problem
2) Classical approaches to identification
    2a) Identification by Short Run Restrictions
    2b) [Identification by Long Run Restrictions]
3) New approaches to identification (post-2000)
    3a) Identification from Heteroskedasticity
    3b) Direct Estimation of Shocks from High Frequency Data
    3c) External Instruments
    3d) Identification by Sign Restrictions

3c) External Instruments

The external instrument approach entails finding some external information (outside the model) that is relevant (correlated with the shock of interest) and exogenous (uncorrelated with the other shocks).

***Example* 1**: The Cochrane- Piazessi (2002) shock (Z^{CP}) measures the part of the monetary policy shock revealed around a FOMC announcement – but not the shock revealed at other times. If CP's identification is sound, $Z^{CP} \neq \varepsilon_t^r$ but

(i) corr(ε_t^r,Z^{CP}) ≠ 0 (relevance)

(ii) corr(other shocks, Z^{CP}) = 0 (exogeneity)

***Example* 2**: Romer and Romer (1989, 2004, 2008); Ramey and Shapiro (1998); Ramey (2009) use the narrative approach to identify moments at which fiscal/monetary shocks occur. If identification is sound, $Z^{RR} \neq \varepsilon_t^r$ but

(i) corr(ε_t^r,Z^{RR}) ≠ 0 (relevance)

(ii) corr(other shocks, Z^{RR}) = 0 (exogeneity)

Selected empirical papers that can be reinterpreted as external instruments

- **Monetary shock**: Cochrane and Piazzesi (2002), Faust, Swanson, and Wright (2003. 2004), Romer and Romer (2004), Bernanke and Kuttner (2005), Gürkaynak, Sack, and Swanson (2005)

- **Fiscal shock**: Romer and Romer (2010), Fisher and Peters (2010), Ramey (2011)

- **Uncertainty shock**: Bloom (2009), Baker, Bloom, and Davis (2011), Bekaert, Hoerova, and Lo Duca (2010), Bachman, Elstner, and Sims (2010)

- **Liquidity shocks**: Gilchrist and Zakrajšek’s (2011), Bassett, Chosak, Driscoll, and Zakrajšek’s (2011)

- **Oil shock**: Hamilton (1996, 2003), Kilian (2008a), Ramey and Vine (2010)

The method of External Instruments

Stock (2007), Stock and Watson (2012); Mertens and Ravn (2013);Gertler and P. Karadi (2014); for IV in VAR (not full method) see Hamilton (2003), Kilian (2009).

Additional notation: *focus on shock* 1

Reduced form VAR: $A(\mathrm{L})Y_t = u_t$

Structural errors ε_t: $\mathrm{R}u_t = \varepsilon_t$ or $u_t = \mathrm{R}^{-1}\varepsilon_t$, or $u_t = \boldsymbol{H}\varepsilon_t$

Structural MAR: $Y_t = \mathrm{A(L)}^{-1}u_t = \mathrm{C(L)}u_t = \mathrm{C(L)}H\varepsilon_t$

Partitioning notation: $u_t = H\varepsilon_t = \begin{bmatrix} H_1 & \cdots & H_r \end{bmatrix}\begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{rt} \end{pmatrix} = \begin{bmatrix} H_1 & H_\bullet \end{bmatrix}\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{\bullet t} \end{pmatrix}$

Structural MAR: $Y_t = C(\mathrm{L})H\varepsilon_t = C(\mathrm{L})H_1\varepsilon_{1t} + C(\mathrm{L})H_\bullet\varepsilon_{\bullet t}$

Structural MAR for j^{th} variable: $Y_{jt} = \sum_{k=0}^{\infty} C_{k,j}H_1\varepsilon_{1t-k} + \sum_{k=0}^{\infty} \overset{1\times r}{C}_{k,j}\, H_\bullet\varepsilon_{\bullet t-k}$

Identification of H_1

$$A(L)Y_t = u_t, \quad u_t = H\varepsilon_t = \begin{bmatrix} H_1 & \cdots & H_r \end{bmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{rt} \end{pmatrix}$$

Suppose you have k instrumental variables Z_t (not in Y_t) such that

(i) $E\left(\varepsilon_{1t}Z_t'\right) = \alpha' \neq 0$ (relevance)

(ii) $E\left(\varepsilon_{jt}Z_t'\right) = 0$, $j = 2,\ldots, r$ (exogeneity)

(iii) $E\left(\varepsilon_t\varepsilon_t'\right) = \Sigma_{\varepsilon\varepsilon} = D = diag(\sigma_{\varepsilon_1}^2,...,\sigma_{\varepsilon_r}^2)$

Under (i) and (ii), you can identify H_1 up to sign & scale

$$E(u_tZ_t') = E(H\varepsilon_tZ_t') = \begin{bmatrix} H_1 & \cdots & H_r \end{bmatrix} \begin{pmatrix} E(\varepsilon_{1t}Z_t') \\ \vdots \\ E(\varepsilon_{rt}Z_t') \end{pmatrix} = \begin{bmatrix} H_1 & \cdots & H_r \end{bmatrix} \begin{pmatrix} \alpha' \\ 0 \\ 0 \end{pmatrix} = H_1\alpha'$$

Identification of H_1, ctd.

$$E(u_t Z_t') = E(H\varepsilon_t Z_t') = \begin{bmatrix} H_1 & H_\bullet \end{bmatrix} \begin{pmatrix} E(\varepsilon_{1t} Z_t') \\ E(\varepsilon_{\bullet t} Z_t') \end{pmatrix} = H_1 \alpha'$$

<u>Normalization</u>

- The scale of H_1 and $\sigma_{\varepsilon_1}^2$ is set by a normalization subject to

$$\Sigma_{uu} = HDH' \quad \text{where } D = diag(\sigma_{\varepsilon_1}^2, ..., \sigma_{\varepsilon_r}^2)$$

- Normalization used here: a unit positive value of shock 1 is defined to have a unit positive effect on the innovation to variable 1, which is u_{1t}. This corresponds to:

(iv) $H_{11} = 1$ (unit shock normalization)

where H_{11} is the first element of H_1

Identification of H_1, ctd.

Impose normalization (iv):

$$E(u_t Z_t') = \begin{pmatrix} Eu_{1t}Z_t' \\ Eu_{\bullet t}Z_t' \end{pmatrix} = H_1\alpha' = \begin{pmatrix} H_{11} \\ H_{1\bullet} \end{pmatrix}\alpha' = \begin{pmatrix} 1 \\ H_{1\bullet} \end{pmatrix}\alpha'$$

So

$$\begin{pmatrix} H_{1\bullet}Eu_{1t}Z_t' \\ Eu_{\bullet t}Z_t' \end{pmatrix} = \begin{pmatrix} H_{1\bullet}\alpha' \\ H_{1\bullet}\alpha' \end{pmatrix}$$

or

$$H_{1\bullet}Eu_{1t}Z_t' = Eu_{\bullet t}Z_t'$$

If Z_t is a scalar (k = 1): $$H_{1\bullet} = \frac{Eu_{\bullet t}Z_t}{Eu_{1t}Z_t}$$

Identification of ε_{1t}

$$\varepsilon_t = H^{-1}u_t = \begin{bmatrix} H^{1\prime} \\ \vdots \\ H^{r\prime} \end{bmatrix} u_t$$

- Identification of first column of H and $\Sigma_{\varepsilon\varepsilon} = D$ identifies first row of H^{-1} up to scale (can show via partitioned matrix inverse formula).

- Alternatively, let Φ be the coefficient matrix of the population regression of Z_t onto u_t:

$$\Phi = E(Z_t u_t{}')\Sigma_u^{-1} = \alpha H_1{}'(HDH')^{-1} = \alpha H_1{}' H'^{-1} D^{-1} H^{-1} = (\alpha/\sigma_{\varepsilon_1}^2)H^{1\prime}$$

because $H^{-1}H_1 = (1\ 0\ \ldots\ 0)'$. Thus ε_{1t} is identified up to scale by

$$\Phi u_t = \frac{\alpha}{\sigma_{\varepsilon_1}^2} H^{1\prime} u_t = \frac{\alpha}{\sigma_{\varepsilon_1}^2}\varepsilon_{1t}$$

Identification of ε_{1t}, ctd

Φu_t is the predicted value from the population projection of Z_t on η_t:

$$\tilde{\varepsilon}_{1t} = \Phi u_t = E(Z_t u_t{}')\Sigma_u^{-1} u_t = \frac{\alpha}{\sigma_{\varepsilon_1}^2}\varepsilon_{1t}$$

- Φ has rank 1 (in population), so this is a (population) reduced rank regression
- 2 instruments identify 2 shocks. Suppose they are shocks 1 and 2, identified by Z_{1t} and Z_{2t}. Then

$$E(\tilde{\varepsilon}_{1t}\tilde{\varepsilon}_{2t}) = E(Z_{1t}u_t{}')\Sigma_u^{-1}E(u_t Z_{2t})$$

which = 0 if both instruments satisfy (i) – (iii)

Estimation

Recall notation: $H_1 = \begin{bmatrix} H_{11} \\ H_{1\bullet} \end{bmatrix}$, $u_t = \begin{bmatrix} u_{1t} \\ u_{\bullet t} \end{bmatrix}$

Impose the normalization condition (iv) $H_{11} = 1$, so

$$E(u_t Z_t') = H_1 \alpha' = \begin{pmatrix} 1 \\ H_{1\bullet} \end{pmatrix} \alpha \text{ or } E(u_t \otimes Z_t) = \begin{pmatrix} 1 \\ H_{1\bullet} \end{pmatrix} \otimes \alpha$$

High level assumption (assume throughout)

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left([u_t \otimes Z_t] - [H_1 \otimes \alpha] \right) \xrightarrow{d} \mathrm{N}(0,\Omega)$$

Estimation of H_1

Efficient GMM objective function:

$S(H_{1\bullet},\alpha;\hat{\Omega})$

$$= \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left((\hat{u}_t \otimes Z_t) - (\begin{bmatrix}1\\ H_{1\bullet}\end{bmatrix} \otimes \alpha)\right)\right)' \hat{\Omega}^{-1} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left((\hat{u}_t \otimes Z_t) - (\begin{bmatrix}1\\ H_{1\bullet}\end{bmatrix} \otimes \alpha)\right)$$

$k = 1$ (exact identification): $E(u_t Z_t') = H_1\alpha' = \begin{pmatrix}\alpha \\ \alpha H_{1\bullet}\end{pmatrix}$

so GMM estimator solves, $T^{-1}\sum_{t=1}^{T}\hat{u}_t Z_t = \begin{pmatrix}\hat{\alpha} \\ \hat{\alpha}\hat{H}_{1\bullet}\end{pmatrix}$

GMM estimator:
$$\hat{H}_{1\bullet} = \frac{T^{-1}\sum_{t=1}^{T}\hat{u}_{\bullet t}Z_t}{T^{-1}\sum_{t=1}^{T}\hat{u}_{1t}Z_t}$$

IV interpretation:
$$\hat{u}_{jt} = H_{1j}\hat{u}_{1t} + u_{jt},$$
$$\hat{u}_{1t} = \Pi_j' Z_t + v_{jt}$$

GMM estimation of $H^{1}{}'$ and ε_{1t}

Recall $$\tilde{\varepsilon}_{1t} = E(Z_t u_t{}')\Sigma_u^{-1} u_t = \Phi u_t$$

<u>Estimator:</u>

- $k = 1$:

  $\hat{\varepsilon}_{1t}$ is the predicted value (up to scale) in the regression of Z_t on $\hat{u}_t$

- $k > 1$(no-HAC):

  Absent serial correlation/no heteroskedasticity, the GMM estimator simplifies to reduced rank regression:

$$Z_t = \Phi \hat{u}_t + \nu_t \qquad \text{(RRR)}$$

- If Z_t is available only for a subset of time periods, estimate (RRR) using available data, compute predicted value over full period

Strong instrument asymptotics

- $k = 1$ case:

$$\sqrt{T}\left(\hat{H}_{1\bullet} - H_{1\bullet}\right) \xrightarrow{d} \mathrm{N}(0, \Gamma'\Omega\Gamma), \text{ where } \Gamma = \begin{bmatrix} -H_{1\bullet}{}' \\ I_{r-1} \end{bmatrix}$$

- Overidentified case ($k > 1$):
  - usual GMM formula
  - J-statistics, etc. are standard textbook GMM

Weak instrument asymptotics: k = 1

(Stock and Watson (2012b)) Weak IV asymptotic setup – local drift (limit of experiments, etc.):

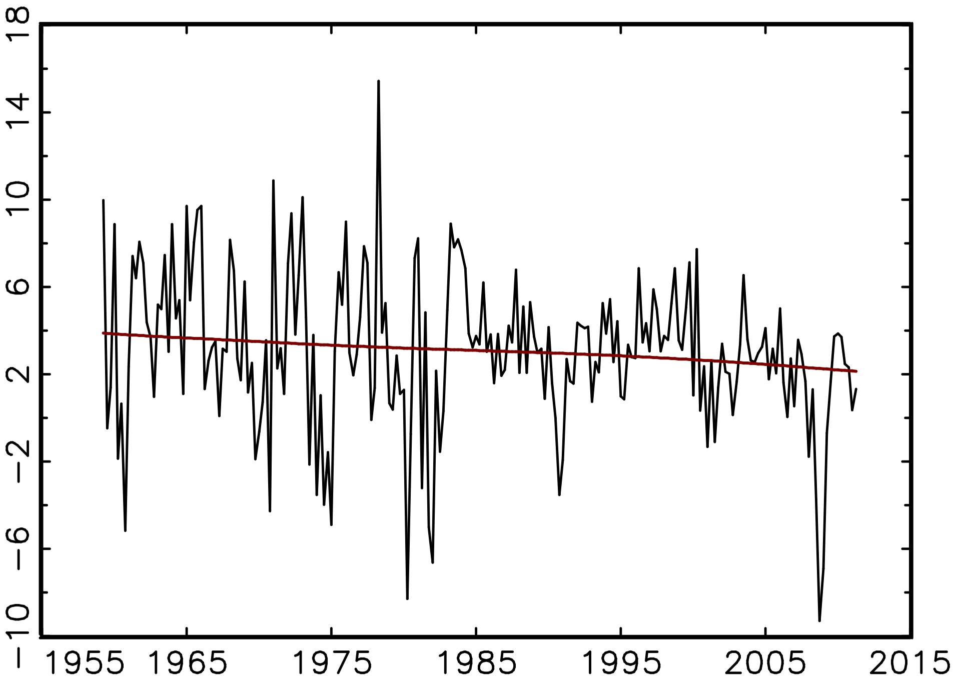
$$\alpha = \alpha_T = a/\sqrt{T}$$

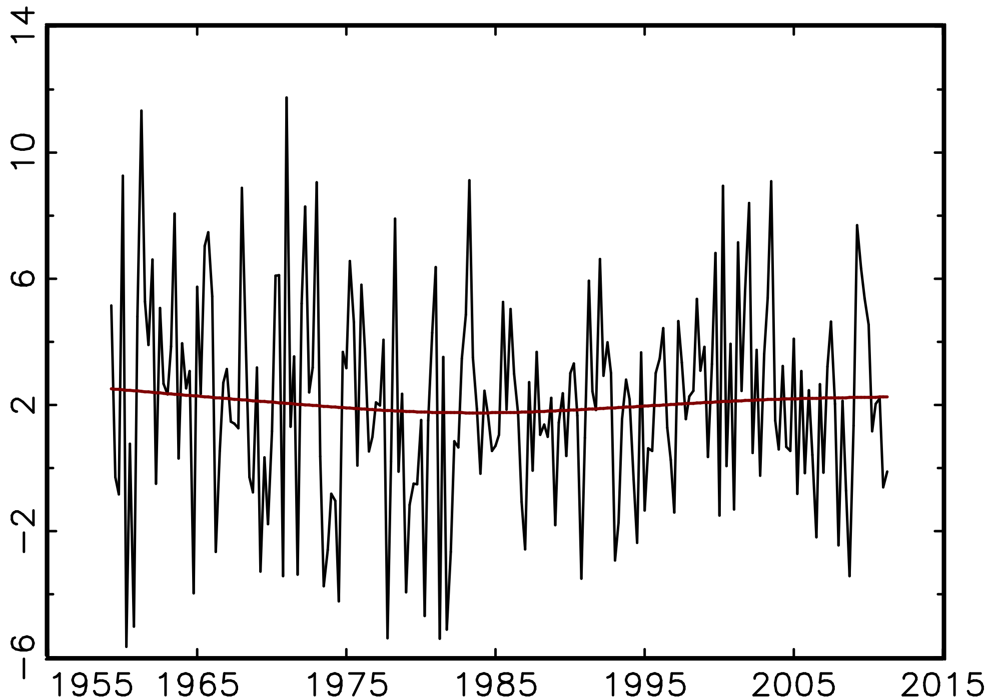
Obtain weak instrument distribution

Empirical Application: Stock-Watson (BPEA, 2012)

Dynamic factor model identified by external instruments:

- U.S., quarterly, 1959-2011Q2, 200 time series
- Almost all series analyzed in changes or growth rates
- All series detrended by local demeaning – approximately 15 year centered moving average:





Quarterly GDP growth (a.r.) — Quarterly productivity growth

Trend: 3.7% → 2.5% — 2.3% → 1.8% → 2.2%

Instruments

1. Oil Shocks
   - a. Hamilton (2003) net oil price increases
   - b. Killian (2008) OPEC supply shortfalls
   - c. Ramey-Vine (2010) innovations in adjusted gasoline prices

2. Monetary Policy
   - a. Romer and Romer (2004) policy
   - b. Smets-Wouters (2007) monetary policy shock
   - c. Sims-Zha (2007) MS-VAR-based shock
   - d. Gürkaynak, Sack, and Swanson (2005), FF futures market

3. Productivity
   - a. Fernald (2009) adjusted productivity
   - b. Gali (200x) long-run shock to labor productivity
   - c. Smets-Wouters (2007) productivity shock

Instruments, ctd.

4. Uncertainty
    a. VIX/Bloom (2009)
    b. Baker, Bloom, and Davis (2009) Policy Uncertainty

5. Liquidity/risk
    a. Spread: Gilchrist-Zakrajšek (2011) excess bond premium
    b. Bank loan supply: Bassett, Chosak, Driscoll, Zakrajšek (2011)
    c. TED Spread

6. Fiscal Policy
    a. Ramey (2011) spending news
    b. Fisher-Peters (2010) excess returns gov. defense contractors
    c. Romer and Romer (2010) “all exogenous” tax changes.

"First stage": F_1: regression of Z_t on u_t, F_2: regression of u_{1t} on Z_t

Structural Shock	F_1	F_2
1. Oil		
Hamilton	2.9	**15.7**
Killian	1.1	1.6
Ramey-Vine	1.8	0.6
2. Monetary policy		
Romer and Romer	4.5	**21.4**
Smets-Wouters	9.0	5.3
Sims-Zha	6.5	**32.5**
GSS	0.6	0.1
3. Productivity		
Fernald TFP	**14.5**	**59.6**
Smets-Wouters	7.0	**32.3**
Structural Shock	F_1	F_2
4. Uncertainty		
Fin Unc (VIX)	**43.2**	**239.6**
Pol Unc (BBD)	**12.5**	73.1

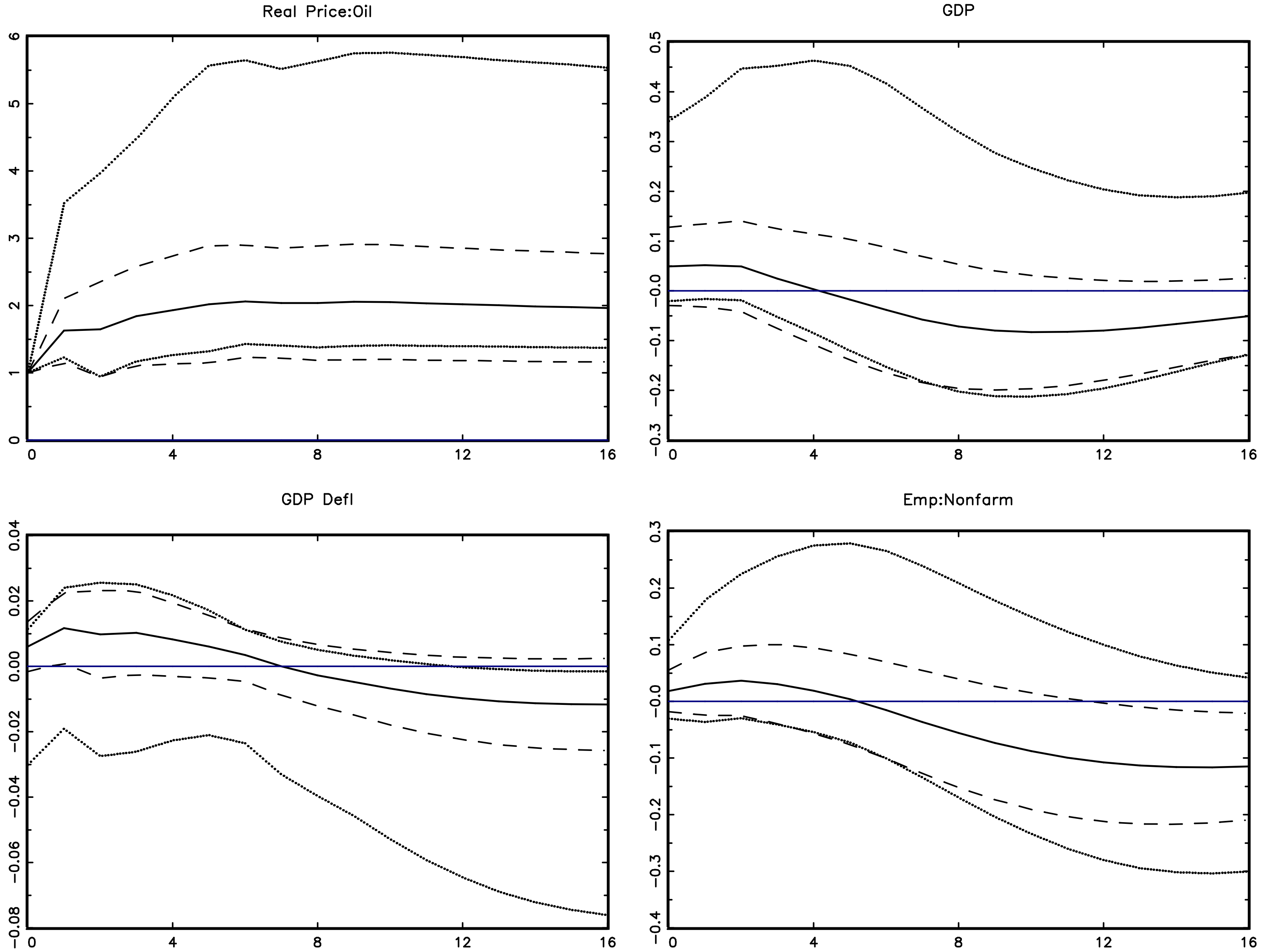
5. Liquidity/risk	F_1	F_2
GZ EBP Spread	4.5	**23.8**
TED Spread	**12.3**	**61.1**
BCDZ Bank Loan	4.4	4.2
6. Fiscal policy		
Ramey Spending	0.5	1.0
Fisher-Peters Spending	1.3	0.1
Romer-Romer Taxes	0.5	2.1

Correlations among selected structural shocks

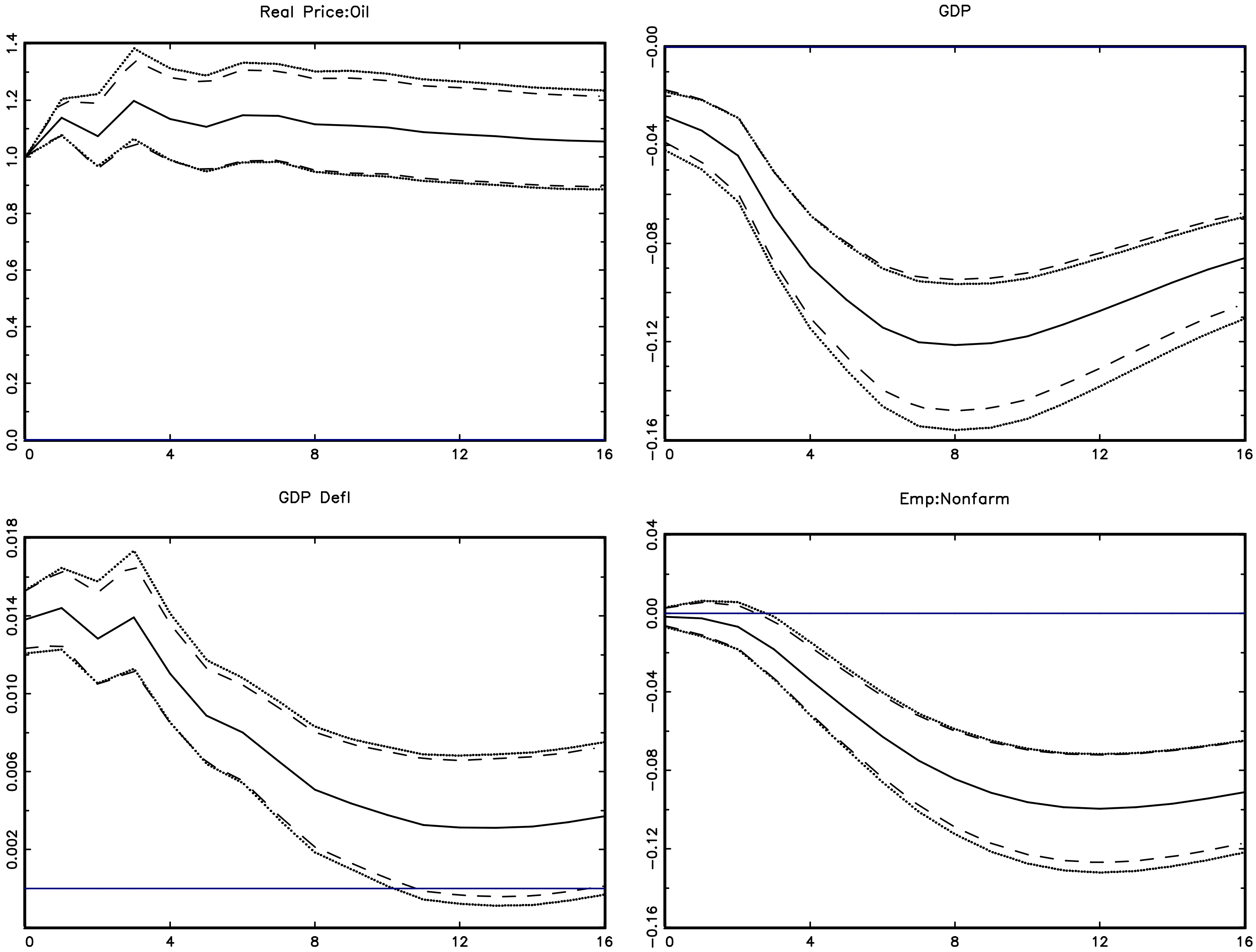
	O_K	M_{RR}	M_{SZ}	P_F	U_B	U_{BBD}	S_{GZ}	B_{BCDZ}	F_R	F_{RR}
O_K	1.00									
M_{RR}	**0.65**	1.00								
M_{SZ}	0.35	0.93	1.00							
P_F	0.30	0.20	0.06	1.00						
U_B	-0.37	-0.39	-0.29	0.19	1.00					
U_{BBD}	0.11	-0.17	-0.22	-0.06	0.78	1.00				
L_{GZ}	-0.42	-0.41	-0.24	0.07	**0.92**	**0.66**	1.00			
L_{BCDZ}	0.22	0.56	0.55	-0.09	**-0.69**	-0.54	-0.73	1.00		
F_R	**-0.64**	**-0.84**	**-0.72**	-0.17	0.26	-0.08	0.40	-0.13	1.00	
F_{RR}	0.15	**0.77**	**0.88**	0.18	0.01	-0.10	0.02	0.19	-0.45	1.00

Oil_{Kilian} oil – Kilian (2009)
M_{RR} monetary policy – Romer and Romer (2004)
M_{SZ} monetary policy – Sims-Zha (2006)
P_F productivity – Fernald (2009)
U_B Uncertainty – VIX/Bloom (2009)
U_{BBD} uncertainty (policy) – Baker, Bloom, and Davis (2012)
L_{GZ} liquidity/risk – Gilchrist-Zakrajšek (2011) excess bond premium
L_{BCDZ} liquidity/risk – BCDZ (2011) SLOOS shock
F_R fiscal policy – Ramey (2011) federal spending
F_{RR} fiscal policy – Romer-Romer (2010) federal tax

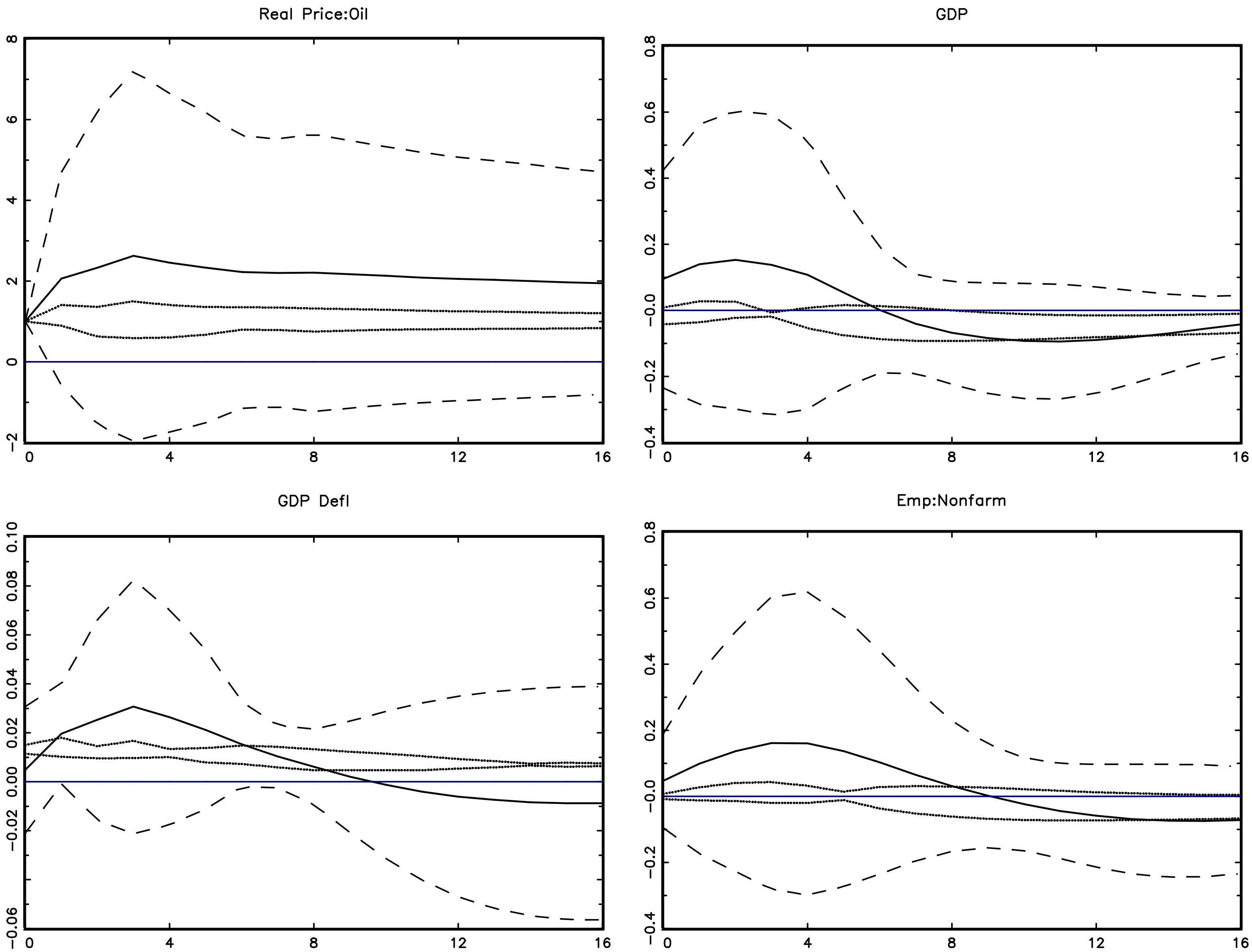
IRFs: strong-IV (dashed) and weak-IV robust (solid) pointwise bands



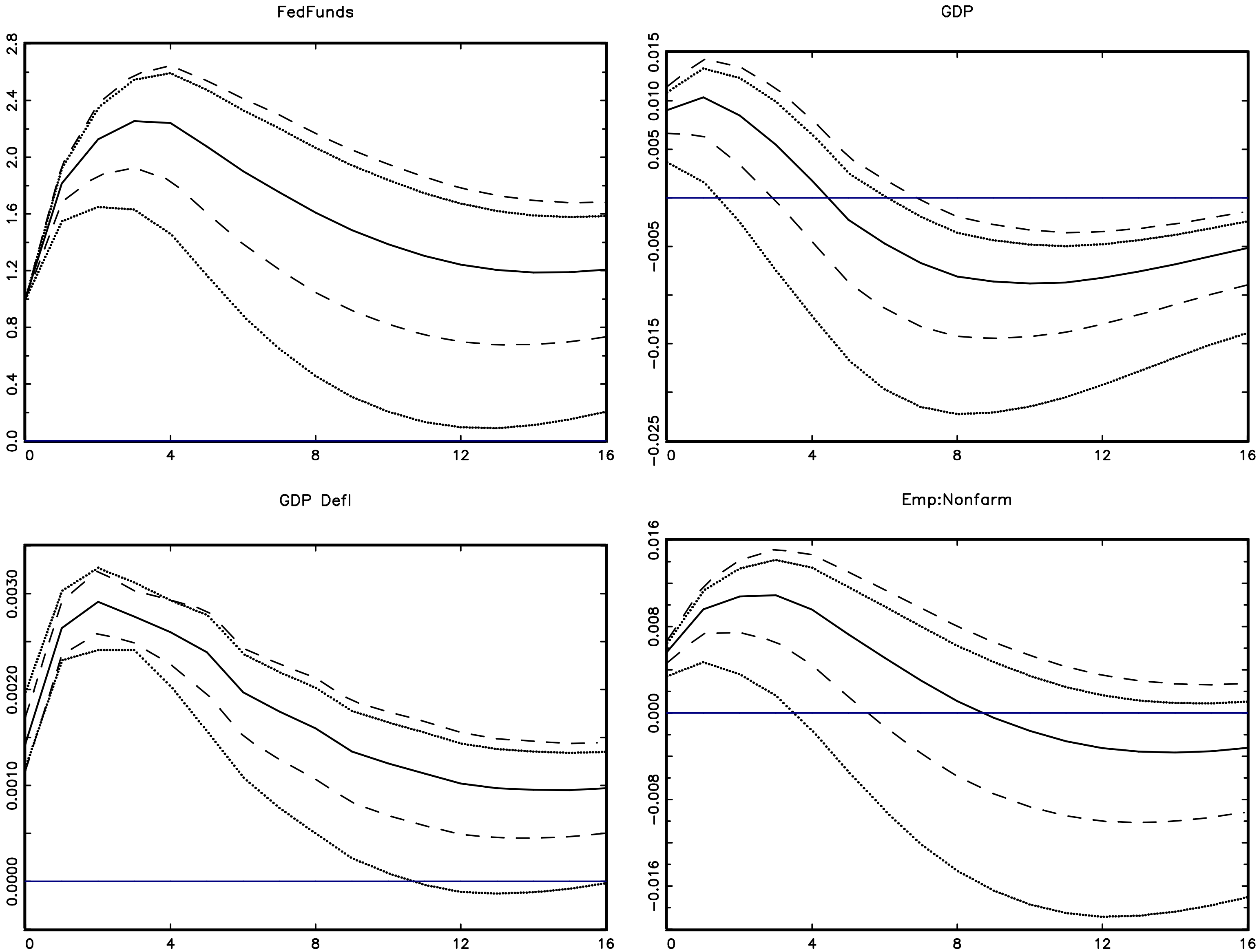
Kilian (2008) oil shock ($F_2 = 1.6$)



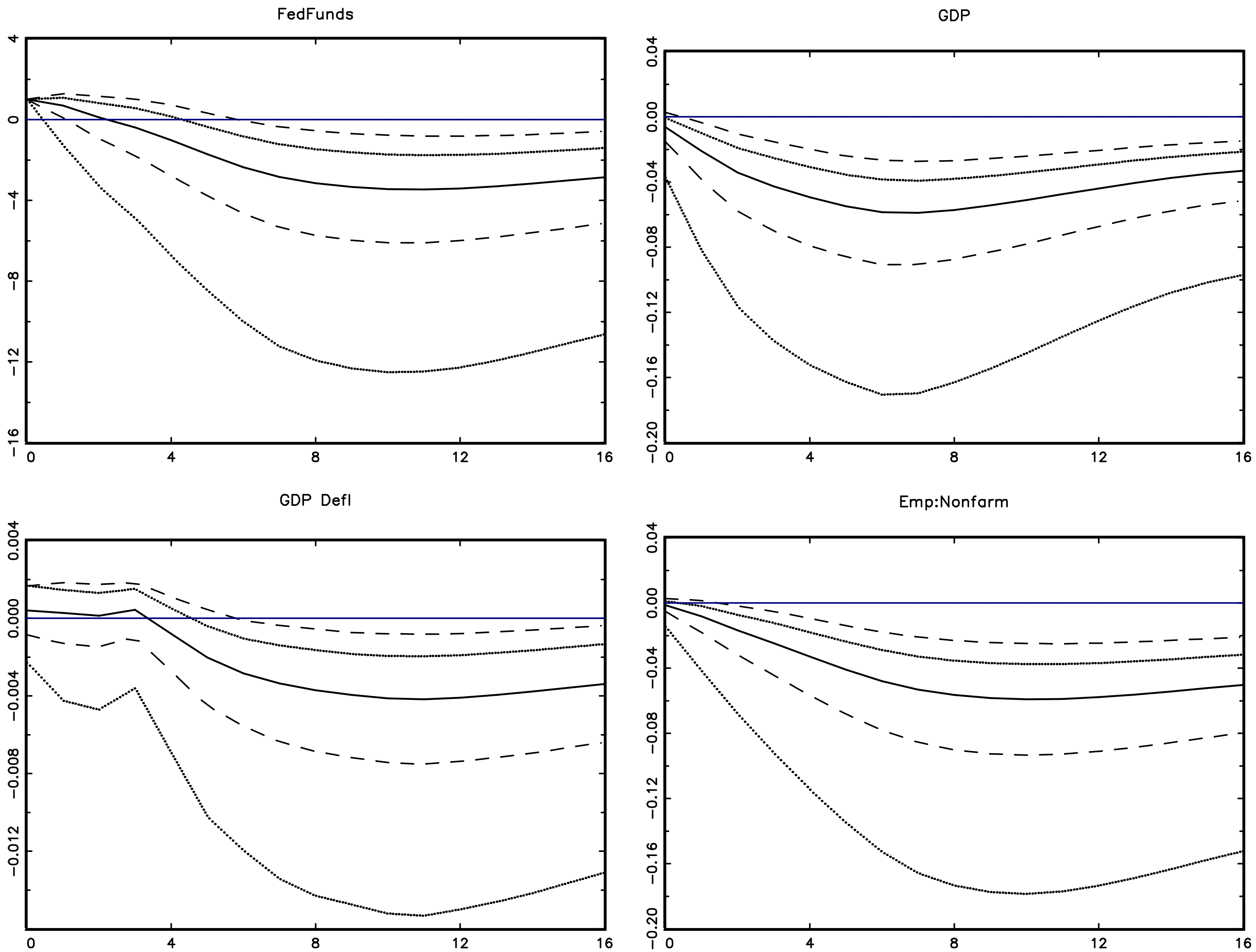
Hamilton (1996, 2003) oil shock (F_2 = 15.7)



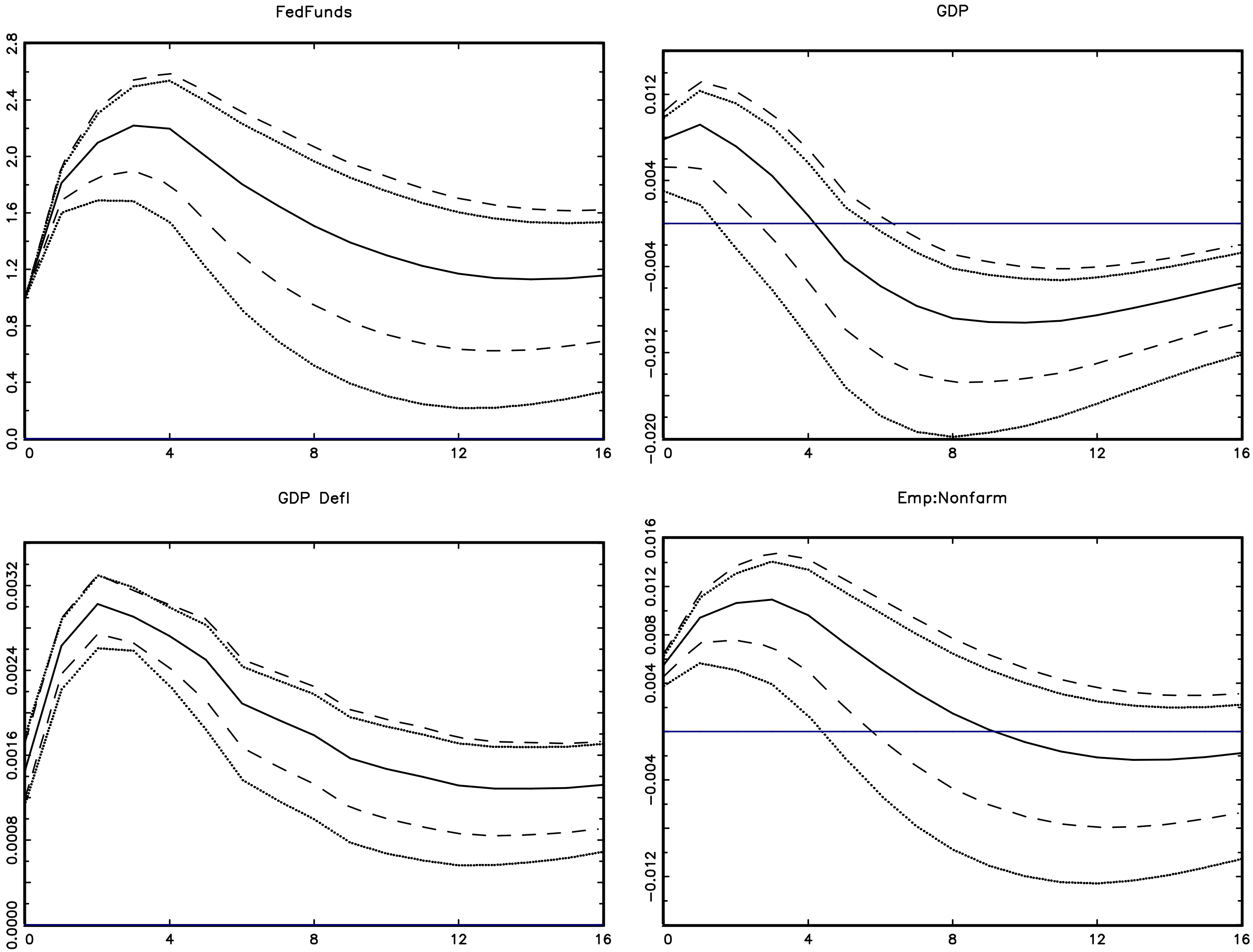
Ramey-Vine (2010) oil shock ($F_2 = 0.6$)



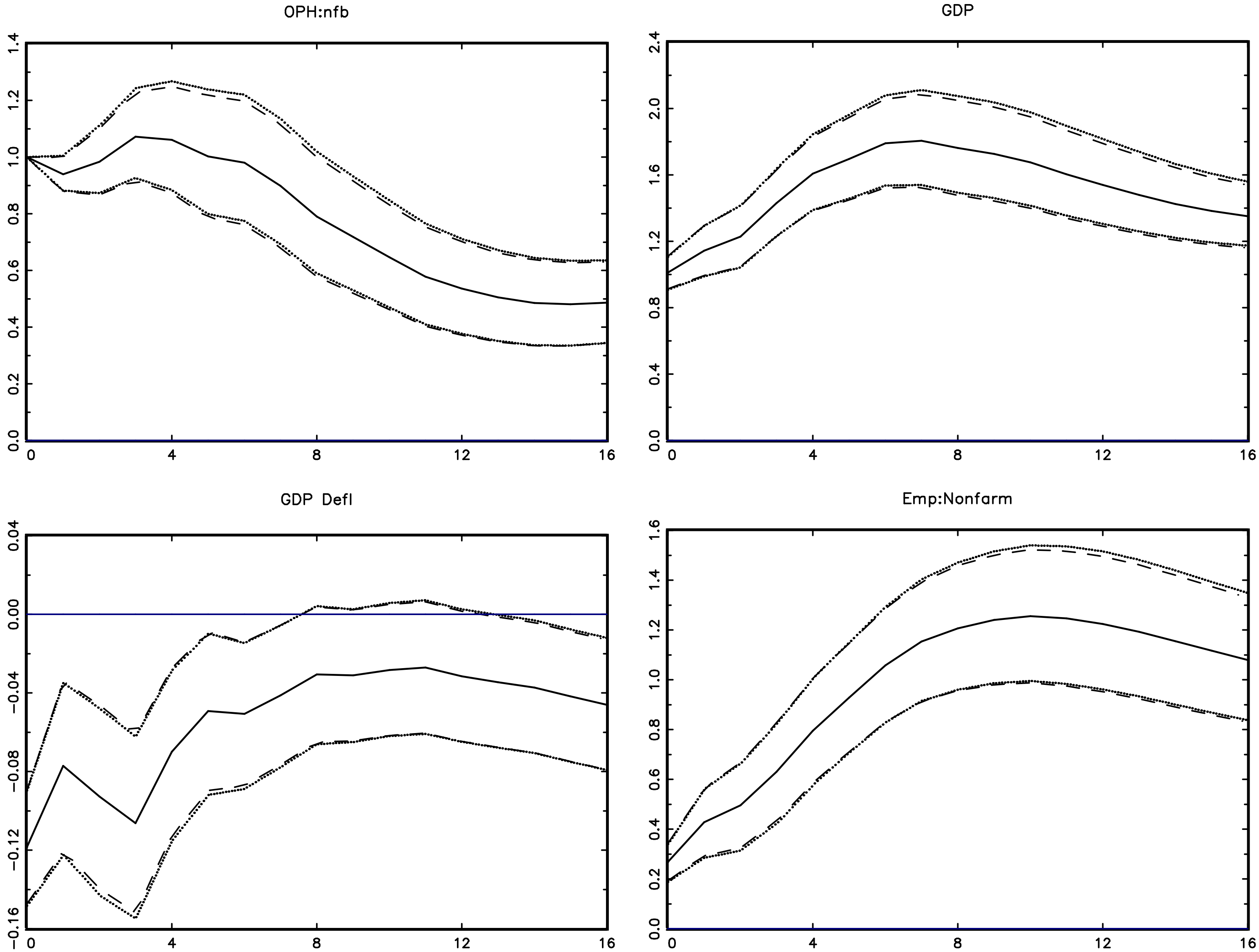
Romer and Romer (2004) monetary policy shock (F_2 = 21.4)



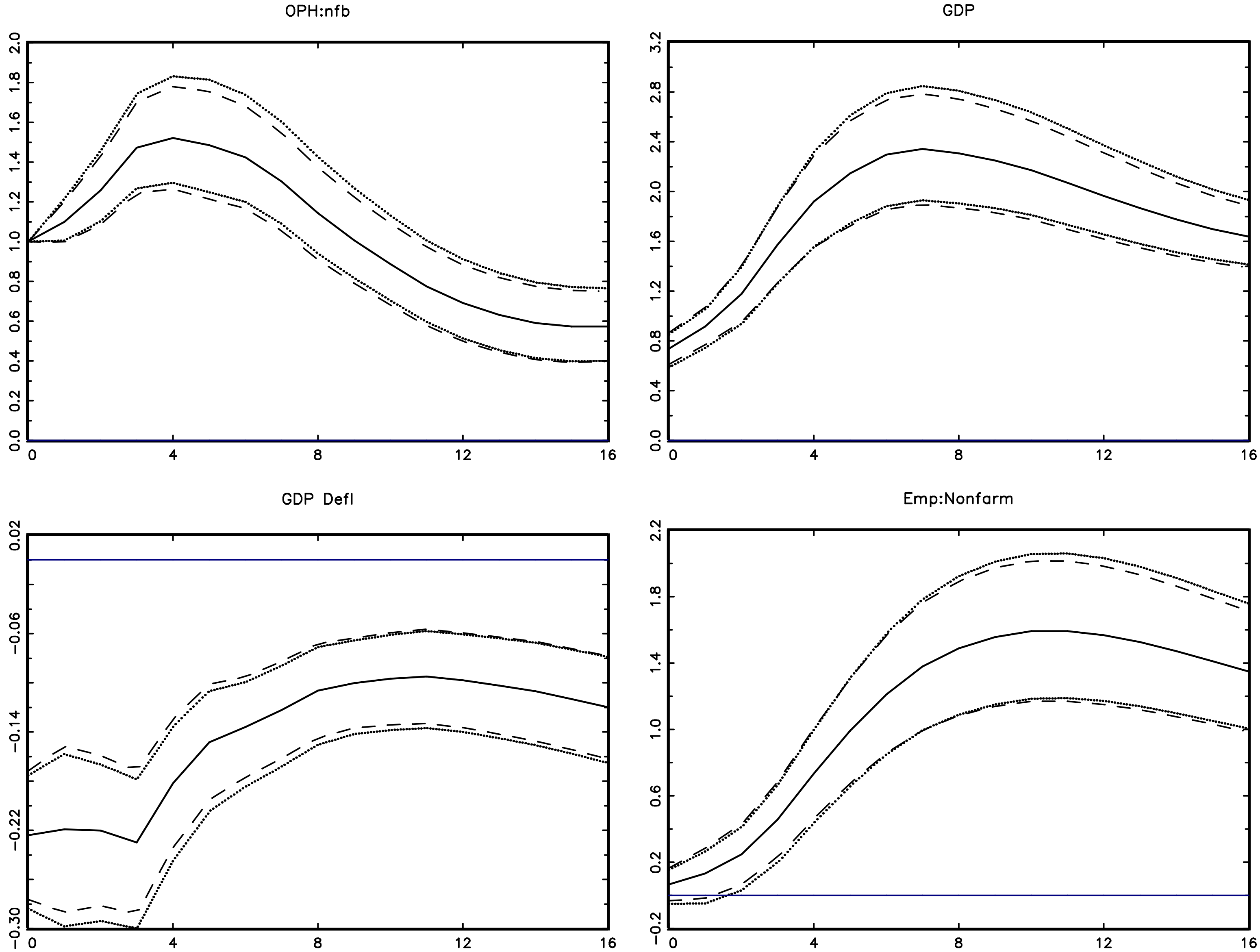
Smets-Wouters (2007) monetary policy shock ($F_2 = 5.3$)



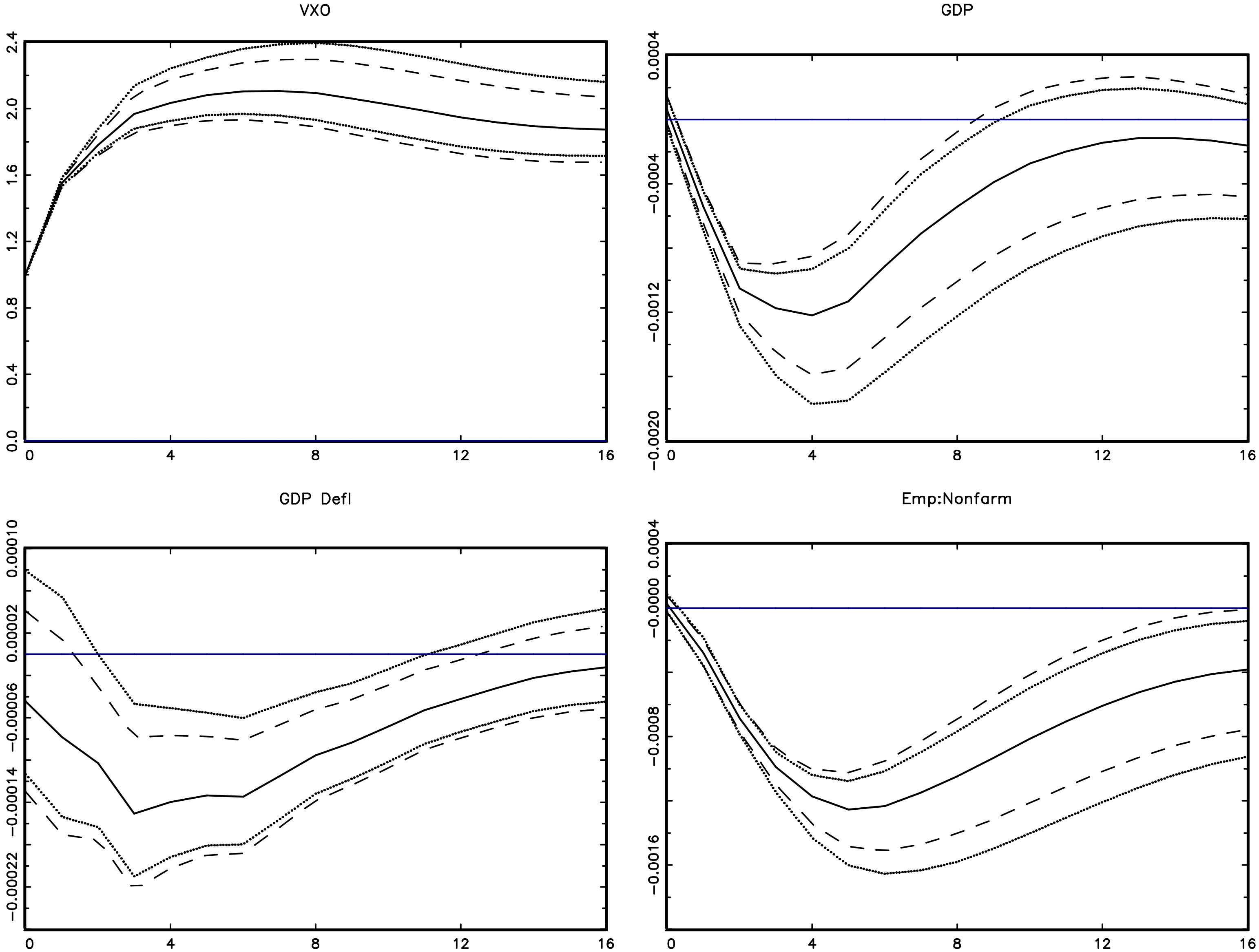
Sims-Zha (2006) monetary policy shock (F_2 = 32.5)



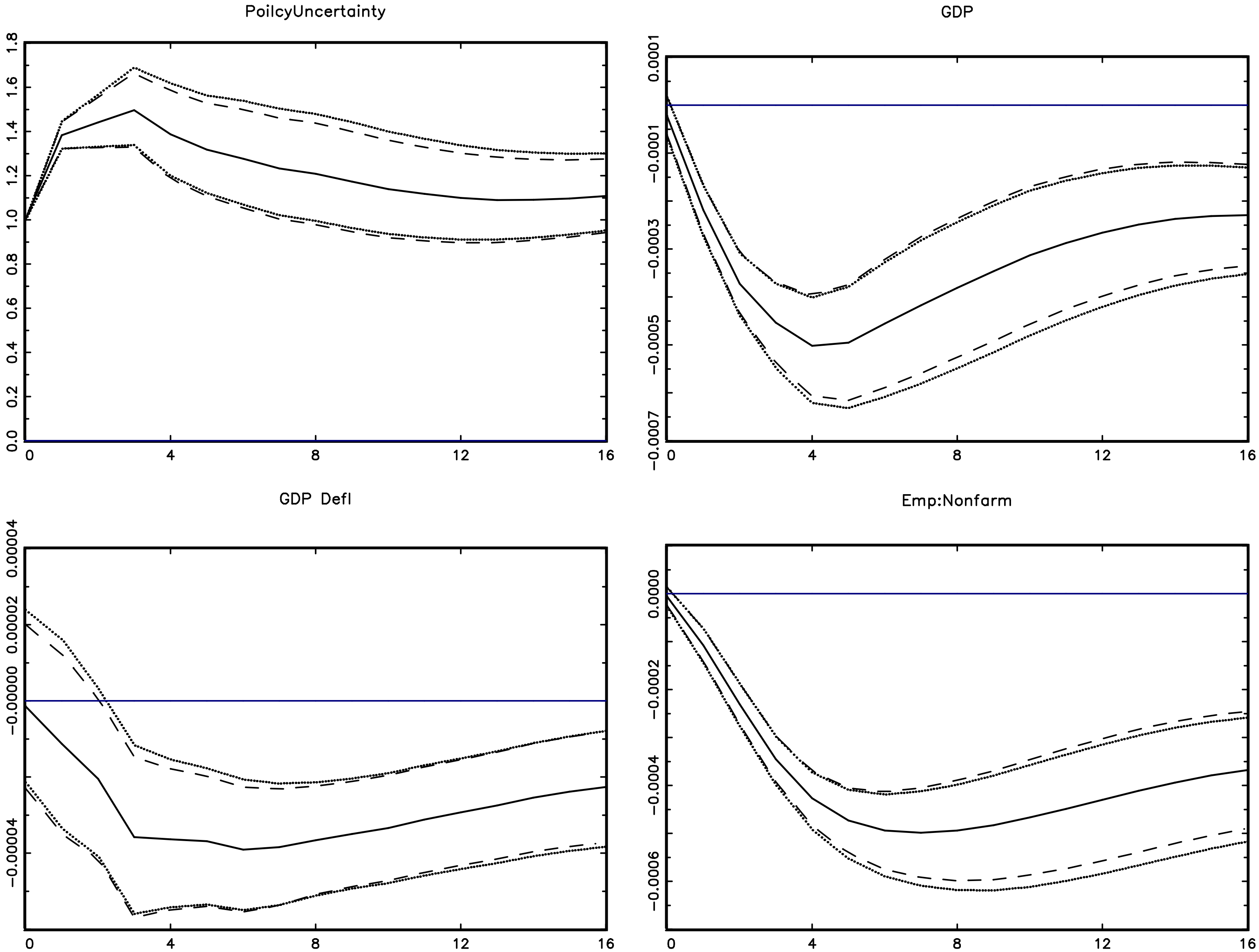
Fernald (2009) productivity shock ($F_2 = 59.6$)



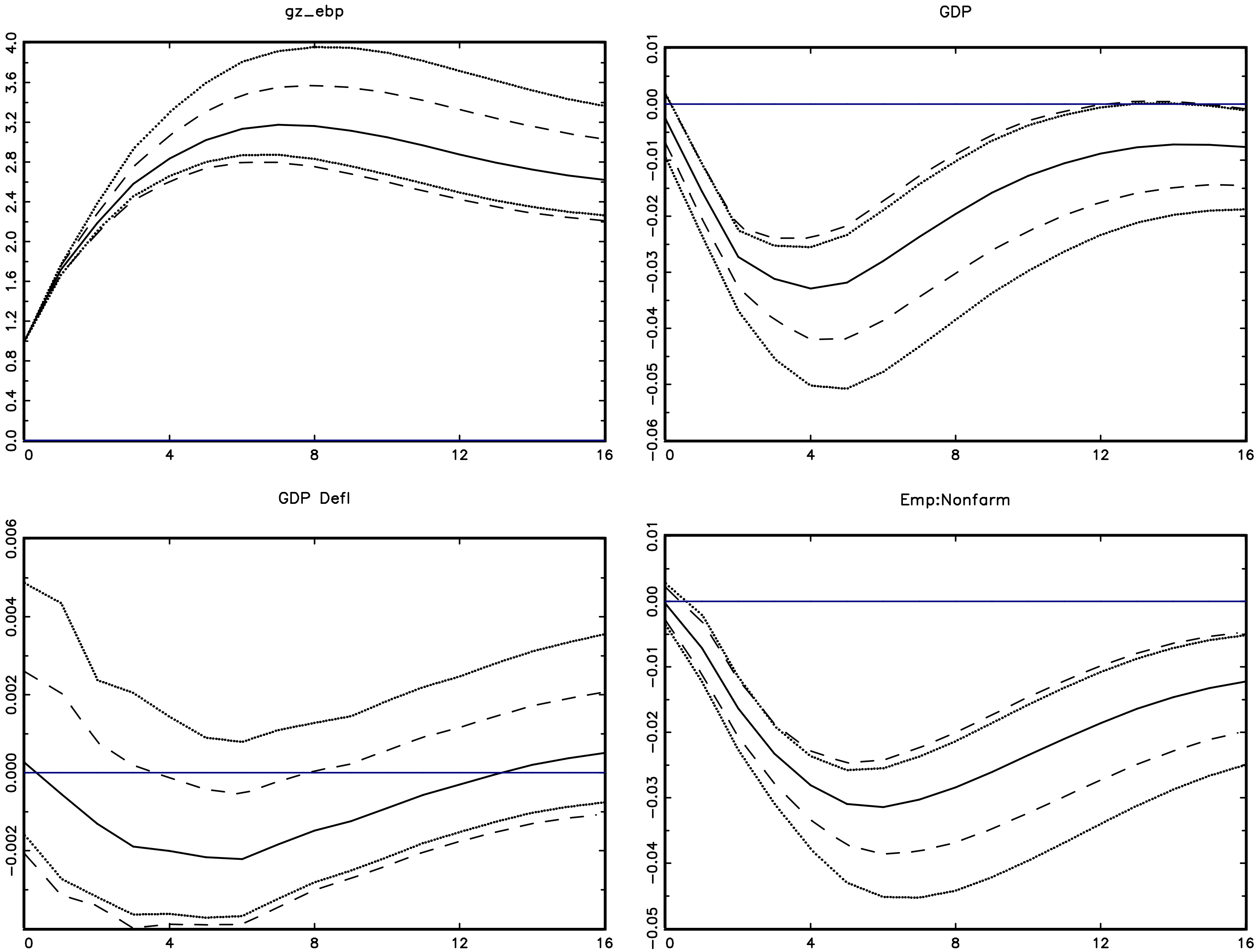
Smets-Wouters (2007) productivity shock ($F_2 = 32.3$)



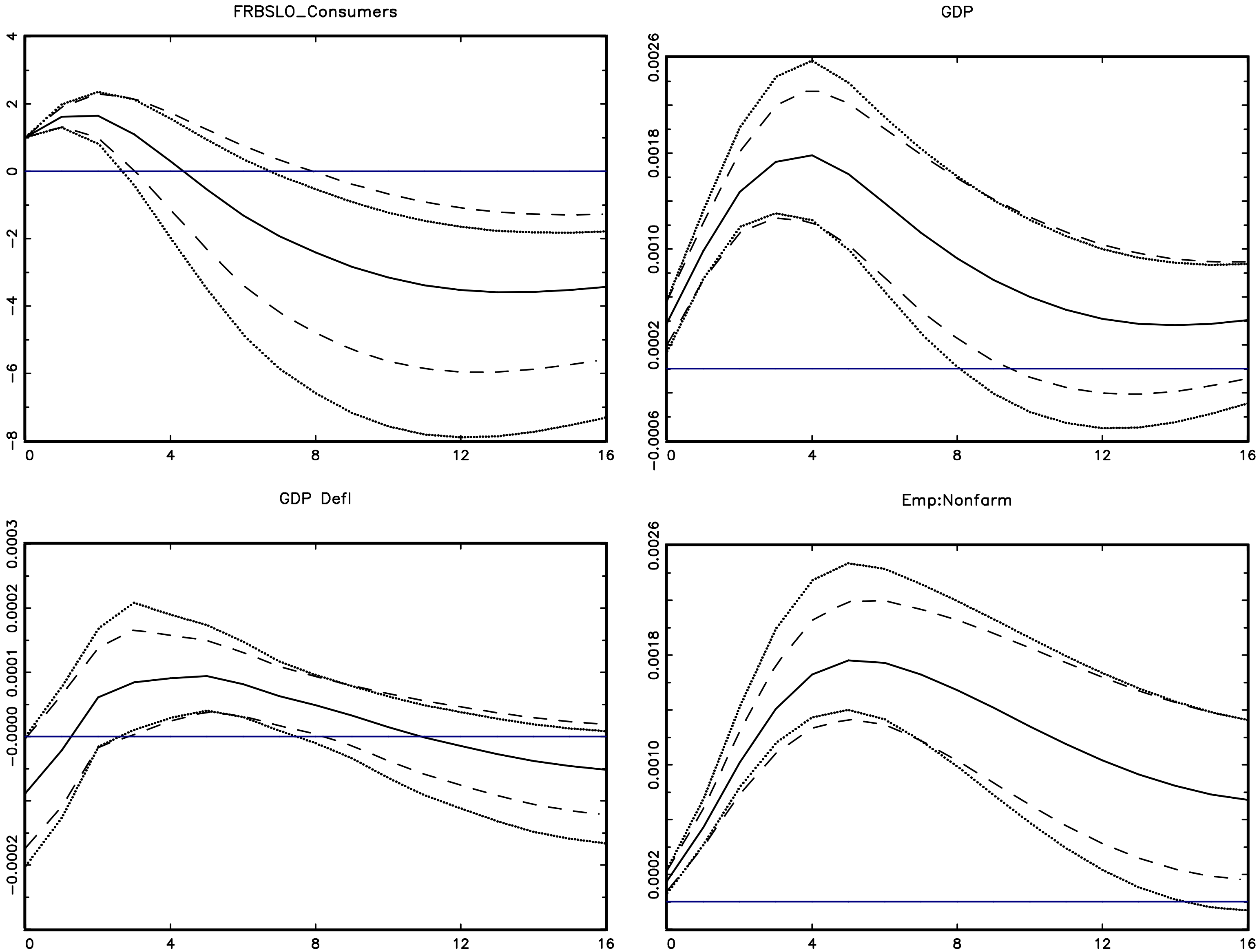
Bloom (2009) (VIX) uncertainty shock (F_2 = 239.6)



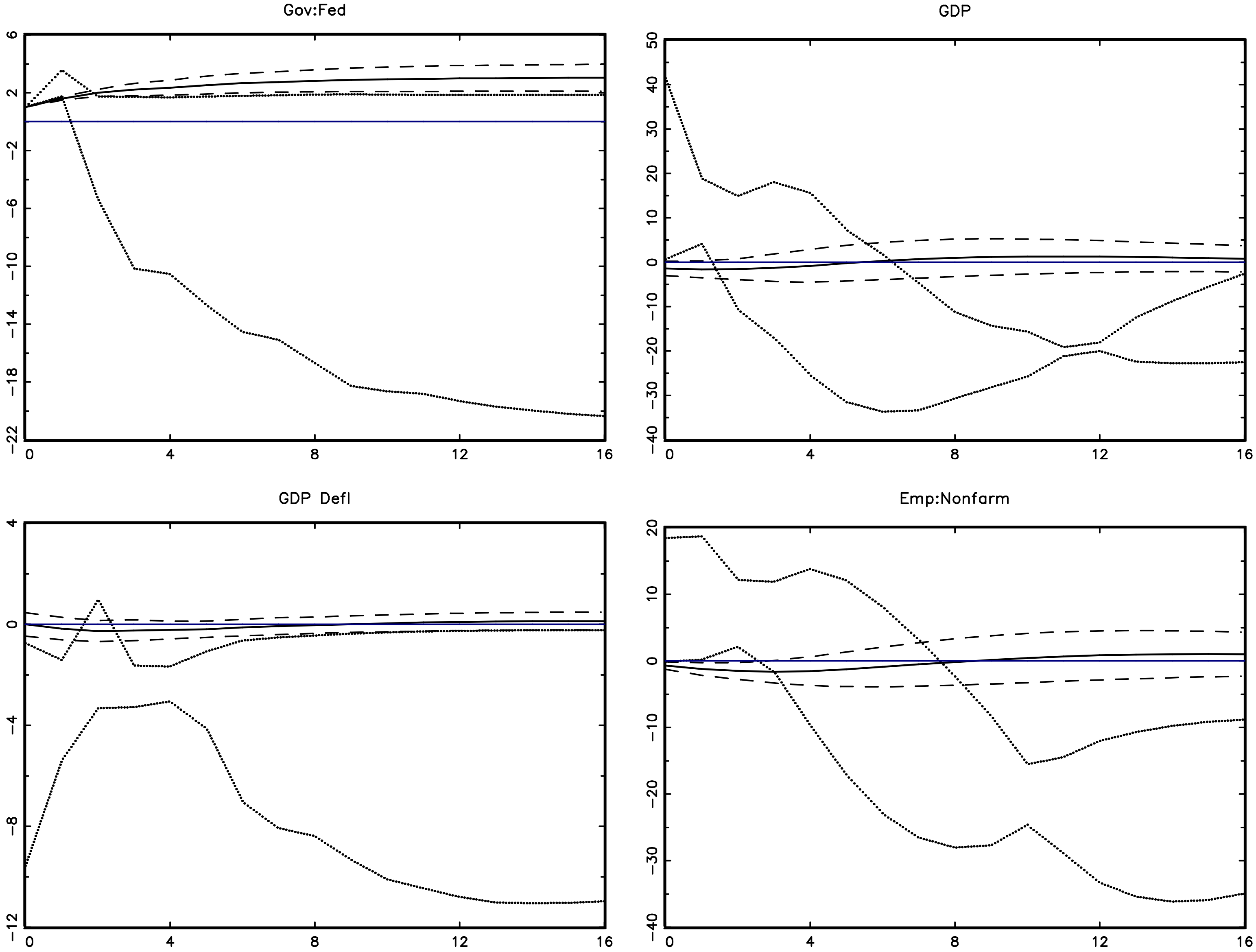
Baker, Bloom, Davis (2012) policy uncertainty shock (F_2 = 73.1)



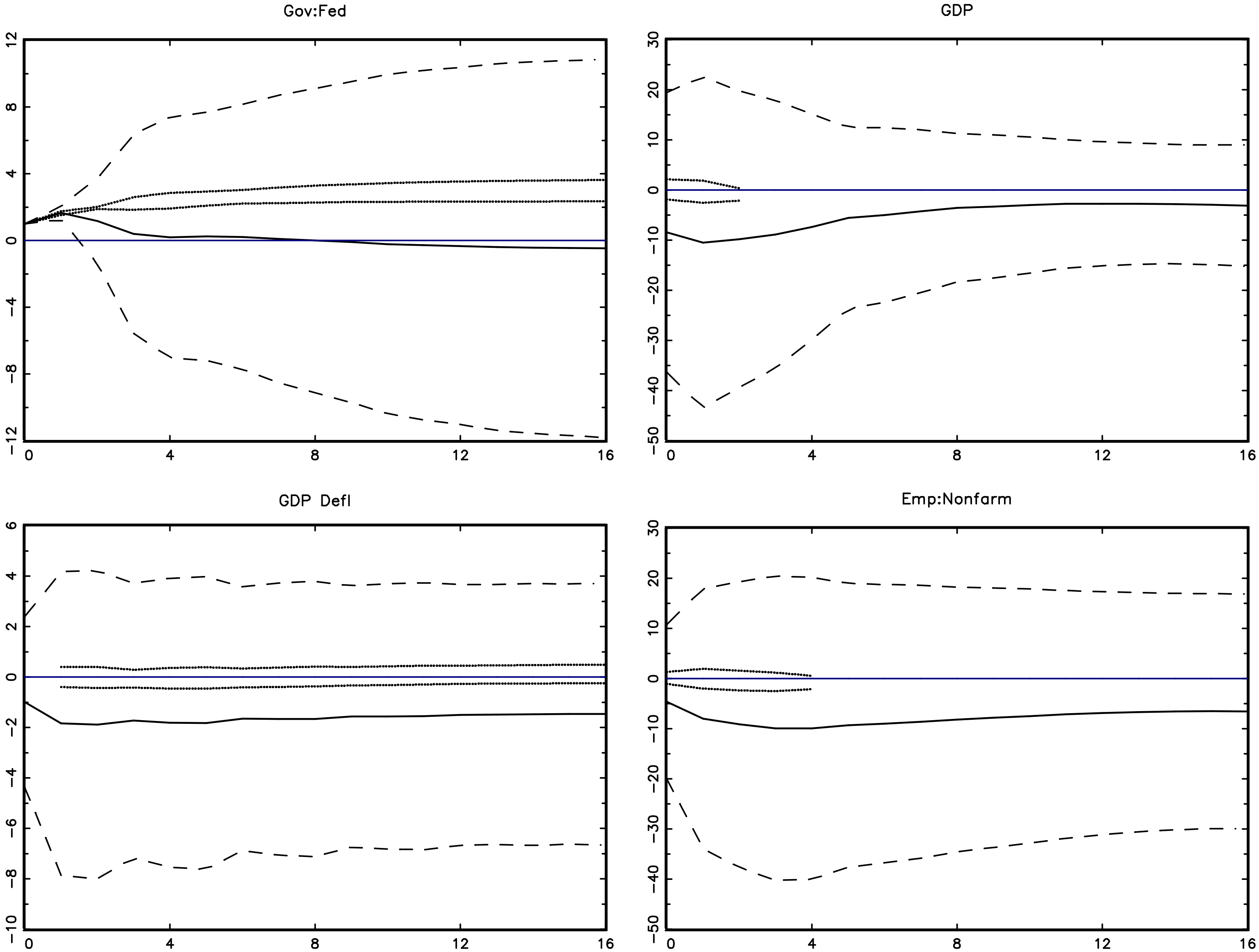
Gilchrist and Zakrajšek (2011) excess bond premium liquidity/risk shock (F_2 = 23.8)



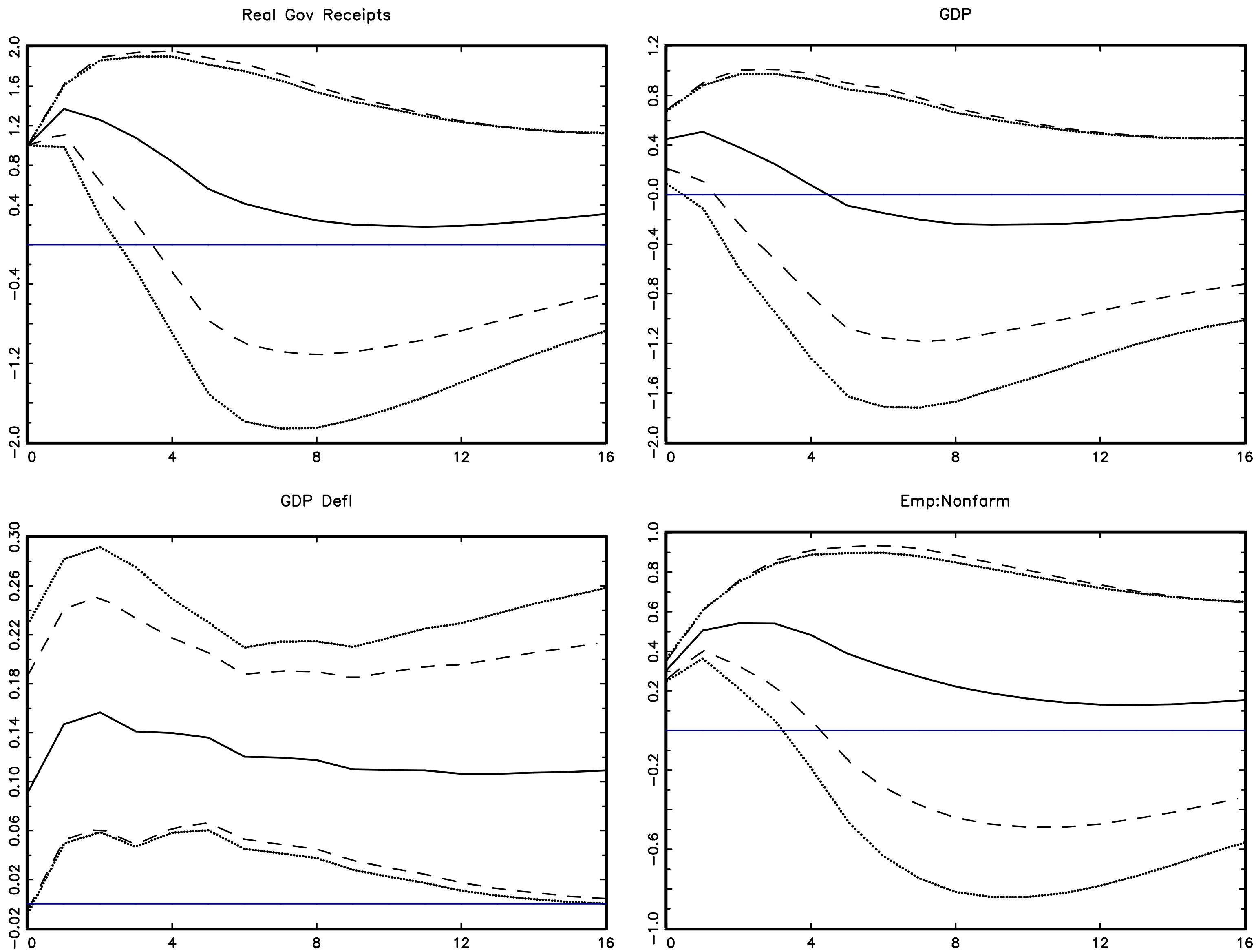
Bassett, Chosak, Driscoll, and Zakrajšek (2011) bank loan supply liquidity/risk shock (F_2 = 4.2)



Ramey (2011) fiscal (spending) shock ($F_2 = 1.0$)



Fisher and Peters (2010) fiscal (spending) shock ($F_2 = 0.1$)



Romer and Romer (2010) fiscal (tax) schock (F_2 = 2.1)

Outline

1) VARs, SVARs, and the Identification Problem
2) Classical approaches to identification
    2a) Identification by Short Run Restrictions
    2b) [Identification by Long Run Restrictions]
3) New approaches to identification (post-2000)
    3a) Identification from Heteroskedasticity
    3b) Direct Estimation of Shocks from High Frequency Data
    3c) External Instruments
    3d) Identification by Sign Restrictions
4) Inference: Challenges and Recently Developed Tools

3d) Identification by Sign Restrictions

Consider restrictions of the form: a monetary policy shock…

- does not decrease the FF rate for months 1,…,6
- does not increase inflation for months 6,..,12

These are restrictions on the sign of elements of D(L).

Sign restrictions can be used to set-identify D(L). Let D denote the set of D(L)'s that satisfy the restriction. There are currently three ways to handle sign restrictions:

1. Faust's (1998) quadratic programming method
2. Uhlig's (2005) Bayesian method
3. Uhlig's (2005) penalty function method

I will describe #2, which is the most popular method (the first steps are the same as #3; #1 has only been used a few times)

Sign restrictions, ctd.

It is useful to rewrite the identification problem after normalizing by a Cholesky factorization (and setting $\Sigma_\varepsilon = I$):

SVAR identification: $R\Sigma_u R' = \Sigma_\varepsilon$

Normalize $\Sigma_\varepsilon = I$; then $\Sigma_u = R^{-1}R^{-1\prime} = R_c^{-1}QQ'R_c^{-1\prime}$

Where $R_c^{-1} = Chol(\Sigma_u)$ and Q is a $n\times n$ orthonormal matrix so $QQ' = I$. Then

Structural errors: $u_t = R_c^{-1}Q\varepsilon_t$

Structural IRF: $\text{D(L)} = \text{C(L)}R_c^{-1}Q$

Let **D** denote the set of acceptable IRFs (IRFs that satisfy the sign restrictions)

Sign restrictions, ctd.

Structural IRF: $D(L) = C(L)R_c^{-1}Q$

Uhlig's algorithm (slightly modified):

(i) Draw $\tilde{Q}$ randomly from the space of orthonormal matrices

(ii) Compute the IRF $\tilde{D}(L)$ = D(L) = $C(L)R_c^{-1}\tilde{Q}$

(iii) If $\tilde{D}(L) \notin \mathbf{D}$, discard this trial $\tilde{Q}$ and go to (i). Otherwise, if $\tilde{D}(L) \in$ D, retain $\tilde{Q}$ then go to (i)

(iv) Compute the posterior (using a prior on A(L) and Σ_u, plus the retained $\tilde{Q}$'s) and conduct Bayesian inference, e.g. compute posterior mean (integrate over A(L), Σ_u, and the retained $\tilde{Q}$'s), compute credible sets (Bayesian confidence sets), etc.

This algorithm implements Bayes inference using a prior proportional to

$$\pi(A(L), \Sigma_u)\times\mathbf{1}(\tilde{D}(L) \in \mathbf{D})\mu(Q)$$

where $\mu(Q)$ is the distribution from which Q is drawn.

n = 2 example

Consider a n = 2 VAR: A(L)Y_t = u_t and structural IRF

$$\mathrm{D(L)} = \begin{pmatrix} D_{11}(L) & D_{12}(L) \\ D_{21}(L) & D_{22}(L) \end{pmatrix} = \mathrm{A(L)}^{-1} R_c^{-1} Q.$$

The sign restriction is $D_{21,I} \geq 0$, I = 1,…, 4 (shock 1 has a positive effect on variable 2 for the first 4 quarters).

Suppose the population reduced form VAR is A(L)Y_t = u_t where

$$\mathrm{A(L)} = \begin{pmatrix} (1-\alpha_1 L)^{-1} & 0 \\ 0 & (1-\alpha_2 L)^{-1} \end{pmatrix} \text{ and } \Sigma_u = I \text{ so } R_c^{-1} = I.$$

What does set-identified Bayesian inference look like for this problem, in a large sample?

- With point-identified inference and nondogmatic priors, it looks like frequentist inference (Bernstein-von Mises theorem)

n = 2 example, ctd.

Step 1: use n =2 to characterize Q

In the n = 2 case, the restriction QQ' = I implies that there is only one free parameter in Q, so that all orthonormal Q can be written,

$$Q = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \text{ [check: } \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = I\text{]}$$

- The standard method, used here, is to draw Q by drawing $\theta \sim$ U[0,2π]
- The main point of this example is that the uniform prior on θ ends up being informative for what matters, D(L), so much so that the prior induced a Bayesian posterior coverage region strictly inside the identified set.

Step 2: Condition for checking whether Q is retained:

$$\hat{D}_{21}(L) = \left[\hat{A}(L)^{-1}\hat{R}_c^{-1}Q\right]_{21} \geq 0 \text{ for first 4 lags}$$

Step 3: In a very large sample, A(L) and Σ_u will be essentially known (WLLN), so that

$$\hat{A}(L)^{-1}\hat{R}_c^{-1}Q \approx \begin{pmatrix} (1-\alpha_1 L)^{-1} & 0 \\ 0 & (1-\alpha_2 L)^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

$$= \begin{pmatrix} (1-\alpha_1 L)^{-1}\cos\theta & -(1-\alpha_1 L)^{-1}\sin\theta \\ (1-\alpha_2 L)^{-1}\sin\theta & (1-\alpha_2 L)^{-1}\cos\theta \end{pmatrix}$$

so $\hat{D}_{21}(L) = \left[\hat{A}(L)^{-1}\hat{R}_c^{-1}Q\right]_{21} \approx (1-\alpha_2\mathrm{L})^{-1}\sin\theta$

Thus the step, keep Q if $\hat{D}_{21,i} \geq 0$, $i = 1,\ldots,4$ reduces to keep Q if $\sin\theta \geq 0$, which is equivalent to $0 \leq \theta \leq \pi$.

Thus, in large samples the posterior of $\hat{D}_{21}(L)$ is $\approx (1-\alpha_2\mathrm{L})^{-1}\sin\theta$, for $\theta \sim U[0,\pi]$.

Characterization of posterior

A draw from the posterior (for a retained θ is): $D_{21}(\mathrm{L}) = (1-\alpha_2\mathrm{L})^{-1}\sin\theta$

Posterior mean for $D_{21,i}$:

$$
\begin{aligned}
E[D_{21,i}] &= E\left(\alpha_2^i \sin\theta\right) \\
&= \alpha_2^i E\left(\sin\theta\right) \\
&= \alpha_2^i \int_0^\pi \frac{1}{\pi}\sin\theta d\theta \\
&= \frac{\alpha_2^i}{\pi}(-\cos\theta\big|_0^\pi) = \frac{2}{\pi}\alpha_2^i \approx .637\alpha_2^i
\end{aligned}
$$

Posterior distribution: drop scaling by α_2^i and focus on $\sin\theta$ part

$$\Pr[\sin\theta \le x] = \Pr[\theta \le \mathrm{Sin}^{-1}(x)] \text{ for } \theta \sim U[0,\pi/2]$$

$$= 2\mathrm{Sin}^{-1}(x)/\pi$$

So the pdf of x is: $f_X(x) = \dfrac{d}{dx}\dfrac{2}{\pi}\mathrm{Sin}^{-1}(x) = \dfrac{2}{\pi\sqrt{1-x^2}}$

So the posterior of $\hat{D}_{21,i}$ is: $p(\hat{D}_{21,i}|Y) \propto \dfrac{2}{\pi\sqrt{1-x^2}}\alpha_2^i$

67% posterior probability interval with equal mass in each tail:

Lower cutoff:

$\Pr[\sin\theta \le x] = 1/6 \rightarrow x_{lower} = \sin(\pi/12) = .259$

$\Pr[\sin\theta \le x] = 5/6 \rightarrow x_{upper} = \sin(5\pi/12) = .966$

so 67% posterior coverage interval is $[.259\alpha_2^i, .966\alpha_2^i]$, with mean $.637\alpha_2^i$

What's wrong with this picture?

- Posterior coverage interval: $[.259\alpha_2^i, .966\alpha_2^i]$, with mean $.637\alpha_2^i$
- Identified set is $[0, \alpha_2^i]$
- What is the frequentist confidence interval here?
- Why don't Bayesian and frequentist coincide?

Recent references on sign-restriction VARs:

Baumeister and Hamilton (WP, 2014)

Fry and Pagan (2011)

Kilian and Murphy (*JEEA*, 2012)

Moon and Schorfheide (*ECMA*, 2012)

Moon, Schorfheide, and Granziera (WP, 2013)

ASSA 2015 Continuing Education Course: Time Series
References (Stock Lectures on HAC & HAR, Weak ID, and SVARs)

I. Heteroskedasticity- and Autocorrelation-Robust Standard Errors

Andrews, D.W.K. (1991), “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation,” *Econometrica* 59, 817-858.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford, U.K.: Oxford University Press.

Brillinger, D.R. (1981), *Time Series Data Analysis and Theory, second edition*. New York: Holt, Rinehart and Winston.

Brockwell, P.J. and R.A. Davis (1991). *Time Series: Theory and Methods*, 2nd Edition. New York: Springer-Verlag.

den Haan, W.J. and A.T. Levin (1997), “A Practioner’s Guide to Robust Covariance Matrix Estimation,” in Maddala, G.S. and C.R. Rao (eds), *Handbook of Statistics*, Vol. 15, Elsevier, Amsterdam, 309-327.

Hansen, C. (2007). “Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T Is Large,” *Journal of Econometrics*, 141, 597–620.

Ibragimov, R. and Müller, U.K. (2010), “*t*-statistic based correlation and heterogeneity robust inference,” *Journal of Business and Economic Statistics* 28, 453-468.

Kiefer, N. and T.J. Vogelsang (2002), “Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel Without Truncation,” *Econometrica*, 70, 2093-2095, 2002

Kiefer, N. and T.J. Vogelsang (2005), “A New Asymptotic Theory for Heteroskesdacity-Autocorrelation Robust Tests,” *Econometric Theory*, 121, 110-1164.

Kiefer, N., T.J. Vogelsang, and H. Bunzel (2000), “Simple Robust Testing of Regression Hypotheses,” *Econometrica*, 69, 695-714.

Müller, Ulrich (2007), “A Theory of Robust Long-Run Variance Estimation,” *Journal of Econometrics*, 141, 1331-1352.

Müller, U. K. (2014). “HAC Corrections for Strongly Autocorrelated Time Series”, *J. Bus. Econ. Stat.*, vol. 32, no. 3, pp. 311–322.

Newey, W.K. and K.D. West (1987), “A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,” *Econometrica* 55, 703-708.

Politis, D.N. (2011). “Higher-Order Accurate, Positive Semidefinite Estimation of Large-Sample Covariance and Spectral Density Matrices,” *Econometric Theory* 27, 703-744.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Sun, Y. (2013). “A Heteroskedasticity and Autocorrelation Robust *F* test Using an Orthonormal Series Variance Estimator,” *The Econometrics Journal* 16, 1-26.

Sun, Y. (2014). “Let's Fix It: Fixed-b Asymptotics versus Small-b Asymptotics in Heteroscedasticity and Autocorrelation Robust Inference,” *Journal of Econometrics*, Vol. 178(3), pp. 659-677

Sun, Y., P.C.B. Phillips, and S. Jin (2008), “Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing,” *Econometrica*, 76(1): 175-194.

Sun, Y., Phillips, P.C.B. and Jin, S. (2011), “Power Maximization and Size Control in Heteroscedasticity and Autocorrelation Robust Tests with Exponentiated Kernels,” *Econometric Theory*, Vol. 27(6), pp. 1320-1368.

Velasco, C. and P.M. Robinson (2001). “Edgeworth Expansions f or Spectral Density Estimates and Studentized Sample Mean,” *Econometric Theory*, 17, 497-539.

II. Weak Identification

Anderson, T.W., and H. Rubin (1949). “Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations,” *Annals of Mathematical Statistics*, 21, 570-582.

Andrews, D.W.K. and Z. Cheng (2012). "Estimation and Inference with Weak, Semi-Strong, and Strong Identification," *Econometrica* 80(5): 2153-2211.

Andrews, D.W.K., M. Moreira, and J.H. Stock (2006). “Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression”, *Econometrica* 74, 715-752.

Andrews, D.W.K., M. Moreira, and J.H. Stock (2008). “Efficient Two-Sided Nonsimilar Invariant Tests in IV Regression with Weak Instruments,” *Journal of Econometrics* 146: 241-254.

Andrews, D.W.K., and J.H. Stock (2006), “Testing with Many Weak Instruments,” *Journal of Econometrics*, 138, 24-46.

Andrews, D.W.K. and J.H. Stock (2007). “Inference with Weak Instruments,” in *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol. III*, ed. by R. Blundell, W. K. Newey, and T. Persson. Cambridge, UK: Cambridge University Press.

Andrews, I. (2014) “Conditional Linear Combination Tests for Weakly Identified Models”, manuscript.

Andrews, I. (2014). “Robust Two-Step Confidence Sets, and the Trouble with the First Stage F-Statistic”, manuscript.

Andrews, I. and A. Mikusheva (2013), “A Geometric Approach to Weakly Identified Econometric Models”, manuscript.

Andrews, I. and A. Mikusheva (2014a), “Weak Identification in Maximum Likelihood: A Question of Information”, *American Economic Review* 104: 195-199.

Andrews, I. and A. Mikusheva (2014b), “Maximum Likelihood Inference in Weakly Identified DSGE models”, *Quantitative Economics* (forthcoming).

Angrist, J. D., and A.B. Krueger (1991), “Does Compulsory School Attendance Affect Schooling and Earnings?” *Quarterly Journal of Economics*, 106, 979-1014.

Atkeson, A., and L.E. Ohanian (2001), “Are Phillips Curves Useful for Forecasting Inflation?,” *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1): 2-11.

Bekker, P. (1994), “Alternative Approximations to the Distributions of Instrumental Variables Estimators,” *Econometrica*, 62, 657-681.

Bound, J., D. A. Jaeger, and R. M. Baker (1995), “Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak,” *Journal of the American Statistical Association*, 90, 443-450.

Chao, J.C., and N.R. Swanson (2005), “Consistent Estimation with a Large Number of Weak Instruments,” *Econometrica*, 73, 1673-1692.

Dufour, J.-M. (1997), “Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models,” *Econometrica*, 65, 1365-1387.

Dufour, J.-M. (2003), “Identification, Weak Instruments, and Statistical Inference in Econometrics,” *Canadian Journal of Economics*, 36, 767-808.

Dufour, J.M., and M. Taamouti (2005), “Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments,” *Econometrica*, 73, 1351-1365.

Dufour, J.M., L. Khalaf, and M. Kichian (2006), “Inflation Dynamics and the New Keynesian Phillips Curve: An Identification Robust Econometric Analysis,” *Journal of Economic Dynamics and Control,* 30, 1707-1727.Guggenberger, P. and R.J. Smith (2005), “Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification,” *Econometric Theory* 21, 667-709.

Kolesár, M. (2013). “Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity,” manuscript, Princeton University

Gali, J., and M. Gertler (1999), “Inflation Dynamics: A Structural Econometric Analysis,” *Journal of Monetary Economics*, 44, 195-222.

Guggenberger, P. (2005), “Monte Carlo Evidence Suggesting a No Moment Problem of the Continuous Updating Estimator,” *Economics Bulletin*, 3, 1-6.

Guggenberger, P. (2007), “Generalized Empirical Likelihood Tests in Time Series Models With Potential Identification Failure,” *Journal of Econometrics*, 142, 134-161.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012). “On the Asymptoic Sizes of Subset Anderson-Rubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression.” *Econometrica* 80: 2649-2666.

Guggenberger, P., and R.J. Smith (2005), “Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification,” *Econometric Theory*, 21, 667-709.

Hansen, L.P. and Singleton, K. (1982), “Generalized Instrumental Variable Estimation of Nonlinear Rational Expectation Models,” *Econometrica* 1269-1286 and errata, January 1984, 267-68.

Hansen, L.P., J. Heaton, and A. Yaron (1996), “Finite Sample Properties of Some Alternative GMM Estimators,” *Journal of Business and Economic Statistics* 14, 262-280.

Hausman, J., W.K. Newey, T. Woutersen, J.C. Chao, and N.R. Swanson (2012), “IV Estimation with Heteroskedasticity and Many Instruments,” *Quantitative Economics* 3: 211-255.

Hausman, J., K. Menzel, R. Lewis, and W. Newey (2007), “ A Reduced Bias GMM-like Estimator with Reduced Estimator Dispersion,” CEMMAP working paper CWP24/07.

Kleibergen, F.R. (2002), “Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression,” *Econometrica*, 70, 1781-1803.

Kleibergen, F.R. (2007), “Generalizing Weak Instrument Robust IV Statistics Towards Multiple Parameters, Unrestricted Covariance Matrices and Identification,” *Journal of Econometrics*, 139, 181-216.

Kleibergen, F.R., and S. Mavroeidis (2009), “Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve,” *Journal of Business and Economic Statistics* 27” 293-339.

Kleibergen, F.R., and S. Mavroeidis (2009), “Inference on Subsets of Parameters in GMM without Assuming Identification,” manuscript, Brown University.

Kocherlakota, N. (1990), “On Tests of Representative Consumer Asset Pricing Models,” *Journal of Monetary Economics*, 26, 285-304.

Mavroeidis, S. (2005), “Identification Issues in Forward-Looking Models Estimated by GMM, with an Application to the Phillips Curve,” *Journal of Money, Credit and Banking*, 37, 421-449.

Mavroeidis, S., M. Plagborg-Møller, and J.H. Stock (2014). "Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve," *Journal of Economic Literature* 52: 124-188.

Mikusheva, A (2013). "Survey on statistical inferences in weakly-identified instrumental variable models", *Applied Econometrics*, vol. 29, no. 1, 117–131.

Montiel Olea, J.L. and C. Pflueger (2013). "A Robust Test for Weak Instruments," *Journal of Business and Economic Statistics* 31: 358-369.

Moreira, M.J. (2003), "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027-1048.

Nason, J.M., and G.W. Smith (2007), "Identifying the New Keynesian Phillips Curve," manuscript, Federal Reserve Bank of Atlanta.

Nelson, C.R., and Startz, R., (1990a), "The Distribution of the Instrumental Variable Estimator and Its t Ratio When the Instrument Is a Poor One," *Journal of Business*, 63, S125-S140.

Nelson, C.R., and Startz, R., (1990b), "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator," *Econometrica*, 58, 967-976.

Newey, W.K., and R.J. Smith, (2004), Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica*, 72, 219.255.

Newey, W.K., and F. Windmeijer (2004), "GMM with Many Weak Moment Conditions," *Econometrica* 77: 687-719.

Politis, D. N., and J. P. Romano (1994), "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions," *Annals of Statistics* 22, 2031-2050.

Politis, D. N., J. P. Romano, and M. Wolf (1999), *Subsampling*. New York: Springer.

Rothenberg, T.J. (1973), *Efficient Estimation with A Priori Information*. New Haven: Yale University Press.

Rothenberg, T.J. (1984), "Approximating the Distributions of Econometric Estimators and Test Statistics," ch. 15 in *Handbook of Econometrics*, *Vol. II*, ed. by Z. Griliches and M.D. Intriligator. Amsterdam: North Holland, 881-935.

Staiger, D. and J.H. Stock (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica* 65, no. 3, 557-586

Stock, J.H. and J. Wright (2000), "GMM With Weak Identification," *Econometrica* 68, 1055 – 1096.

Stock, J.H., J. Wright, and M. Yogo (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20, 518 – 529.

Stock, J.H., and M. Yogo (2005a), "Testing for Weak Instruments in Linear IV Regression," ch. 5 in *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, ed. by J.H. Stock and D.W.K. Andrews, Cambridge, UK: Cambridge University Press.

Stock, J.H., and M. Yogo (2005b), "Asymptotic Distributions of Instrumental Variables Statistics with Many Weak Instruments," ch. 6 in *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, ed. by J.H. Stock and D.W.K. Andrews, Cambridge, UK: Cambridge University Press.

Yogo, M. (2004), "Estimating the Elasticity of Intertemporal Substitution when the Instruments are Weak," *Review of Economics and Statistics* 86, 797-810.

III. Structural Vector Autoregressions

Auerbach, A. and Y. Gorodnichenko (2012). "Measuring the Output Responses to Fiscal Policy," *American Economic Journal – Economic Policy* 4: 1–27.

Baumeister, C. and J. Hamilton, "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information", WP, 2014.

Bernanke, Ben S. (1986), "Alternative Explanations of the Money-Income Correlation," *Carnegie-Rochester Conference Series on Public Policy*, Autumn, 25, 49-99.

Bernanke, B.S., and M. Gertler (1995), "Inside the Black Box: The Credit Channel of Monetary Policy Transmission," *Journal of Economic Perspectives* 9, 27-48.

Bernanke, B., M. Gertler, and M. Watson (1997), "Systematic Monetary Policy and the Effects of Oil Price Shocks," *Brookings Papers on Economic Activity* 1997:1, 91-158 (with discussion)

Bernanke, B.S., and K.N. Kuttner (2005), "What Explains the Stock Market's Reaction to Federal Reserve Policy?," *Journal of Finance* 40, 1221-1257.

Blanchard, O. and J. Galí (2007). "The Macroeconomic Effects of Oil Price Shocks: Why are the 2000s so Different from the 1970s?" in J. Galí and M.J. Gertler (eds), *International Dimensions of Monetary Policy*, University of Chicago Press for the NBER.

Blanchard, O., J. L'Huillier, and G. Lorenzoni (2012), "News, Noise, and Fluctuations: An Empirical Exploration," WP.

Blanchard, O.J., and D. Quah (1989), "Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79, 655-673.

Blanchard, O.J., and M.W. Watson (1986), "Are Business Cycles All Alike?" in R.J. Gordon (ed.), *The American Business Cycle*, University of Chicago Press: Chicago.

Canova, F., and M. Ciccarelli (2008), "Estimating Multi-Country VAR models," manuscript.

Chari, V.V., P.J. Kehoe, and E. McGrattan (2007), "A Critique of Structural VARs Using Real Business Cycle Theory" (aka "Are Structural VARs with Long-Run Restrictions Useful in Developing Business Cycle Theory?"), Federal Reserve Bank of Minneapolis Working Paper Series 634.

Christiano, L.J., M.S. Eichenbaum, and C.L. Evans (1999), "Monetary Policy Shocks: What Have We Learned and to What End?" in *Handbook of Macroeconomics*, ed. by J.B. Taylor and M. Woodford, Amsterdam: Elsevier Science, North-Holland.

Christiano, L., M. Eichenbaum, and C. Evans (2005), "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113(1): 1-45.

Christiano, L., M. Eichenbaum, and R. Vigfusson (2004), "The Response of Hours to a Technology Shock: Evidence Based on Direct Measures of Technology," *Journal of the European Economic Association*, 2, 381-395.

Christiano, L., M. Eichenbaum, and R. Vigfusson (2006) "Assessing Structural VARs," *NBER Macroeconomics Annual* 2006, 1-72.

Cochrane, J.H., and M. Piazzesi (2002), "The Fed and Interest Rates: A High-Frequency Identification," *American Economic Review* 92, 90-95

Cooley, T., and M. Dwyer (1998), "Business Cycle Analysis without Much Theory: A Look at Structural VARs," *Journal of Econometrics*, 83, 57-88.

Del Negro, M., and F. Schorfheide, (2004), "Priors From General Equilibrium Models for VARs," *International Economic Review*, 45, 643-673.

Erceg, C.J., L. Guerrieri, and C. Gust (2005), "Can Long-Run Restrictions Identify Technology Shocks?," *Journal of the European Economic Association*, 3, 1237-1278.

Faust, J., (1998), "The Robustness of Identified VAR Conclusions About Money," *Carnegie-Rochester Series on Public Policy* 49, 207–244.

Faust. J., and E. Leeper (1997), "When Do Long-Run Identifying Restrictions Give Reliable Results?," *Journal of Business and Economic Statistics* 15, 345-353.

Faust, J., and J.H. Rogers (2003), "Monetary Policy's Role in Exchange Rate Behavior," *Journal of Monetary Economics* 50, 1403-1424.

Faust, J., J.H. Rogers, E. Swanson, and J.H. Wright (2003), "Identifying the Effects of Monetary Policy Shocks on Exchange Rates Using High Frequency Data," *Journal of the European Economic Association*, 1(5): 1031-1057.

Faust, J., E. Swanson, and J. Wright (2004), "Identifying VARs Based on High-Frequency Futures Data," *Journal of Monetary Economics* 51(6): 1107-1131.

Fernandez-Villaverde, J., J.F. Rubio-Ramirez, T.J. Sargent, and M.W. Watson (2007), "ABCs (and Ds) for Understanding VARS," *American Economic Review* 97, 1021-1026.

Forni, M., L. Gambetti, and L. Sala, "No News in Business Cycles", WP, 2012

Francis, Neville, and Valerie A. Ramey (2005), "Is the Technology-Driven Real Business Cycle Hypothesis Dead? Shocks and Aggregate Fluctuations Revisited," *Journal of Monetary Economics* 52, 1379-99.

Fry, R. and A. Pagan (2011). "Sign Restrictions in Structural Vector Autoregressions: A Critical Review", *Journal of Economic Literature*, vol. 49, no. 4: 938–960.

Gali, J. (1999), "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?," *American Economic Review* 89, 249-271.

Gertler M. and P. Karadi (2014), "Monetary Policy Surprises, Credit Costs and Economic Activity", WP, 2014.

Giannone, D., M. Lenza, and G. Primiceri (2012). "Prior Selection for Vector Autoregressions", WP.

Giannone, D., M. Lenza, and G.E. Primiceri (2014). "Prices for the Long Run," manuscript.

Gospodinov, N. (2010), "Inference in Nearly Nonstationary SVAR Models with Long-Run Identifying Restrictions," *Journal of Business and Economic Statistics*, 28(1): 1-12.

Ingram, B., and C. Whiteman (1994), "Supplanting the Minnesota Prior: Forecasting Macroeconomic Time Series Using Real Business Cycle Model Priors," *Journal of Monetary Economics*, 34, 497–510.

Kehoe, P.J. (2006), "Comment on Christiano, Eichenbaum, and Vigfusson's 'Assessing Structural VARs'," *NBER Macroeconomics Annual* 2006, 97-102.

Kilian, L. (1998a), "Small-sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics* 80, 218-230.

Kilian, L. (1998b), "Confidence Intervals for Impulse Responses Under Departures from Normality," *Econometric Reviews*, 17, 1-29.

Kilian, L. (1999), "Finite-Sample Properties of Percentile and Percentile-t Bootstrap Confidence Intervals for Impulse Responses," *Review of Economics and Statistics* 81, 652-660.

Kilian, L. (2001), "Impulse Response Analysis in Vector Autoregressions with Unknown Lag Order," *Journal of Forecasting* 20, 161-179.

Kilian, L., and P.-L. Chang (2000), "How Accurate Are Confidence Intervals for Impulse Responses in Large VAR Models?," *Economics Letters*, 69, 299-307.

Kilian, L. and D.P. Murphy (2012). "Why Agnostic Sign Restrictions are not enough: Understanding the Dynamics of Oil Market VAR Models," *Journal of the European Economics Association* 10: 1166-1188.

King, Robert G., C.I. Plosser, J.H. Stock, and M.W. Watson (1991), "Stochastic Trends and Economic Fluctuations," *American Economic Review* 81(4): 819-840.

Lippi, M., and L. Reichlin (1993), "The dynamic effects of aggregate demand and supply disturbances: comment," *American Economic Review* 83, 644-652.

Lippi, M., and L. Reichlin (1994), "VAR Analysis, Nonfundamental Representations, Blaschke Matrices," *Journal of Econometrics*, 63, 307–25.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, New York: Springer Verlag.

Mertens, K. and M.O. Ravn (2013). "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States," *American Economic Review* 103: 1212-1247.

Moon, H.R., F. Schorfheide, E. Granziera, and M. Lee (2013). "Inference for VARs Identified with Sign Restrictions," manuscript, University of Pennsylvania.

Pagan, A.R., and J.C. Robertson (1998), "Structural Models of the Liquidity Effect," *Review of Economics and Statistics*, 80, 202-217.

Pesavento, E., and B. Rossi (2005), "Do Technology Shocks Drive Hours Up or Down?," *Macroeconomic Dynamics*, 9, 478-488.

Pesavento, E., and B. Rossi (2006), "Small-Sample Confidence Intervals for Multivariate Impulse Response Functions at Long Horizons," *Journal of Applied Econometrics* 21, 1135-1155.

Phillips, P.C.B (1998), "Impulse Response and Forecast Error Variance Asymptotics in Nonstationary VARs," *Journal of Econometrics*, 83, 21-56.

Pope, A.L. (1990), "Biases of Estimators in Multivariate non-Gaussian Autoregressions," *Journal of Time Series Analysis*, 11, 249-258.

Ramey, V.A., 2009. "Identifying Government Spending Shocks: It's All in the Timing," NBER Working Paper 15464.

Ramey, V.A., and M. Shapiro (1998), "Costly Capital Reallocation and the Effects of Government Spending" (with discussion), *Carnegie Rochester Conference on Public Policy* 48, 145-209.

Rigobon, R. (2003), "Identification through Heteroskedasticity," *Review of Economics and Statistics* 85, 777-792.

Rigobon, R., and B. Sack (2003), "Measuring the Reaction of Monetary Policy to the Stock Market," *Quarterly Journal of Economics* 118, 639-669

Rigobon, R., and B. Sack (2004), "The Impact of Monetary Policy on Asset Prices," *Journal of Monetary Economics*, 51, 1553-1575.

Romer, C.D., and D.H. Romer (1989), "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz," *NBER Macroeconomics Annual* 4, 121-170.

Romer, C.D., and D.H. Romer (2004), "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review* 94, 1055-1084.

Romer, C.D., and D.H. Romer (2008), "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks," manuscript, University of California – Berkeley.

Rudebusch, G.D. (1998), "Do Measures of Monetary Policy in a VAR Make Sense?," *International Economic Review*, 39, 907-931.

Runkle, D.E. (1987), "Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models," *Review of Economics and Statistics* 72, 116-125.

Sarte, P.-D. G. (1997), "On the Identification of Structural Vector Autoregressions," *Richmond Fed Economic Quarterly* 83(3): 45-67.

Sentana, E., and G. Fiorentini (2001), "Identification, Estimation, and Testing of Conditionally Heteroskedastic Factor Models," *Journal of Econometrics*, 102, 143-164.

Shapiro, M.D., and M.W. Watson (1988), "Sources of Business Cycle Fluctuations," *NBER Macroeconomics Annual*, 3, 111-156.

Sims, C.A. (1980), "Macroeconomics and Reality," *Econometrica*, 48, 1-48.

Sims, Christopher A. (1986), "Are Forecasting Models Usable for Policy Analysis?," *Federal Reserve Bank of Minneapolis Quarterly Review*, Winter, 10(1): 2-16.

Sims, C.A. and T. Zha (1999), "Error Bands for Impulse Responses," *Econometrica*, 1113-1155.

Sims, C.A., and T. Zha (2006), "Does Monetary Policy Generate Recessions?" *Macroeconomic Dynamics* 10, 231–272.

Sims, E.R. (2012). "News, Non-Invertibility, and Structural VARs," in N. Balke, F. Canova, F. Milani, and M.A. Wynne (eds), *DSGE Models in Macroeconomics: Estimation, Evaluation, and New Developments (Advances in Econometrics, Volume 28)* Emerald Group Publishing Limited, pp.81 – 135.

Stock, J.H. (1996), "VAR, Error Correction and Pretest Forecasts at Long Horizons," *Oxford Bulletin of Economics and Statistics* 58, 685-701

Stock, J.H. (1997), "Cointegration, Long-Run Comovements, and Long-Horizon Forecasting," in *Advances in Econometrics: Proceedings of the Seventh World Congress of the Econometric Society*, *vol. III*. ed. by D. Kreps and K.F. Wallis, Cambridge, UK: Cambridge University Press, 34-60.

Stock, J.H., and M.W. Watson (2001), "Vector Autoregressions," *Journal of Economic Perspectives* 15, 101-116.

Stock, J.H., and M.W. Watson (2005), "Understanding Changes in International Business Cycle Dynamics," *Journal of the European Economic Association*, 5, 968-1006.

Stock, J.H., and M.W. Watson (2005), "Disentangling the Channels of the 2007-09 Recession," *Brookings Papers on Economic Activity*: Spring 2012, 871-156 (including discussion).

Uhlig, H. (2005), "What Are the Effects of Monetary Policy on Output? Results From an Agnostic Identification Procedure," *Journal of Monetary Economics* 52, 381-419.

Uhlig, H. (2004), "Do Technology Shocks Lead to a Fall in Total Hours Worked?," *Journal of the European Economic Association* 2, 361-371.

Watson, M.W. (1993), "Measures of Fit for Calibrated Models, *Journal of Political Economy* 101, 1011-1041

Watson, M.W. (1994), "Vector Autoregressions and Cointegration," *Handbook of Econometrics*, *v. IV,* 2844-2915 (section 3).

Watson, M.W. (2006), “Comment on Christiano, Eichenbaum, and Vigfusson’s ‘Assessing Structural VARs’,” *NBER Macroeconomics Annual* 2006, 97-102.

Wright, J.H. (2000), “Confidence Intervals for Univariate Impulse Responses with a Near Unit Root,” *Journal of Business and Economic Statistics* 18, 368-373.

Wright, J. (2012), “What Does Monetary Policy to do Long-Term Interest Rates at the Zero Lower Bound?” *Economic Journal* 122, F447-F466.

AEA Continuing Education Program

Time Series Econometrics

Mark W. Watson, Princeton

January 5-7, 2015

AEA Continuing Education Course

Time Series Econometrics

Lecture 1: Time series refresher and inference tools

Mark W. Watson

January 5, 2015

4:00PM – 6:00PM

Course Topics

1. Time series refresher and inference tools (MW)
2. The Kalman filter, nonlinear filtering, and Markov chain monte carlo (MW)
3. Prediction with large datasets (MW)
4. Heteroskedasticity and autocorrelation consistent/robust (HAC, HAR) standard errors (JS)
5. Many instruments/weak identification in IV and GMM (JS)
6. Structural VARs: Recent Developments (JS)

Lecture Outline

1. Time Series Basics
2. Spectral representation of stationary process
3. Approximation tools (CLT, FCLT, etc.).

Time Series Basics (and notation)

(References: Hayashi (2000), Hamilton (1994), Brockwell and Davis (1991)… , lots of other books)

1. $\{Y_t\}$: a sequence of random variables

2. Stochastic Process: The probability law governing $\{Y_t\}$

3. Realization: One draw from the process, $\{y_t\}$

4. Strict Stationarity: The process is strictly stationary if the probability distribution of $(Y_t, Y_{t+1}, ..., Y_{t+k})$ is identical to the probability distribution of $(Y_\tau, Y_{\tau+1}, ..., Y_{\tau+k})$ for all t, τ, and k. (Thus, all joint distributions are time invariant.)

5. Autocovariances: $\gamma_{t,k} = cov(Y_t, Y_{t+k})$

6. Autocorrelations: $\rho_{t,k} = cor(Y_t, Y_{t+k})$

7. Covariance Stationarity: The process is covariance stationary if $\mu_t = E(Y_t) = \mu$ and $\gamma_{t,k} = \gamma_k$ for all t and k.

8. White noise: A process is called white noise if it is covariance stationary and $\mu = 0$ and $\gamma_k = 0$ for $k \neq 0$.

9. Martingale: Y_t follows a martingale process if $E(Y_{t+1} \mid \boldsymbol{F}_t) = Y_t$, where $\boldsymbol{F}_t \subseteq \boldsymbol{F}_{t+1}$ is the time t information set.

10. Martingale Difference Process: Y_t follows a martingale difference process if $E(Y_{t+1} \mid \boldsymbol{F}_t) = 0$. $\{Y_t\}$ is called a martingale difference sequence or "mds."

11. The Lag Operator: "L" lags the elements of a sequence by one period. $\mathrm{L}y_t = y_{t-1}$, $\mathrm{L}^2 y_t = y_{t-2}$,. If b denotes a constant, then $b\mathrm{L}Y_t = \mathrm{L}(bY_t) = bY_{t-1}$.

12. Linear filter (moving averages): Let $\{c_j\}$ denote a sequence of constants and

$$c(\mathrm{L}) = c_{-r}\mathrm{L}^{-r} + c_{-r+1}\mathrm{L}^{-r+1} + \ldots + c_0 + c_1\mathrm{L} + \ldots + c_s\mathrm{L}^s$$

denote a polynomial in L. Note that $X_t = c(\mathrm{L})Y_t = \sum_{j=-r}^{s} c_j Y_{t-j}$ is a moving average of Y_t. $c(\mathrm{L})$ is sometimes called a linear filter (for reasons discussed below) and X is called a filtered version of Y.

13. AR(p) process: $\phi(\mathrm{L})Y_t = \varepsilon_t$ where $\phi(\mathrm{L}) = (1 - \phi_1\mathrm{L} - \ldots - \phi_p\mathrm{L}^p)$ and ε_t is white noise.

14. MA(q) process: $Y_t = \theta(\mathrm{L})\varepsilon_t$ where $\theta(\mathrm{L}) = (1 - \theta_1\mathrm{L} - \ldots - \theta_q\mathrm{L}^q)$ and ε_t is white noise.

15. ARMA(p,q): $\phi(\mathrm{L})Y_t = \theta(\mathrm{L})\varepsilon_t$.

16. Wold decomposition theorem (e.g., Brockwell and Davis (1991))
Suppose Y_t is generated by a "linearly indeterministic" covariance stationary process. Then Y_t can be represented as

$$Y_t = \varepsilon_t + c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2} + \ldots ,$$

where ε_t is white noise with variance σ_ε^2, $\sum_{i=1}^{\infty} c_i^2 < \infty$, and

$\varepsilon_t = Y_t - Proj(Y_t \mid \text{lags of } Y_t)$ (so that ε_t is "fundamental").

17. The autocovariance generating function for a covariance stationary process is given by $\gamma(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j$, so the autocovariances are given by the coefficients on the argument z^j.

(a) With x represented as $x_t = c(L)\varepsilon_t$, the ACGF is

$\gamma(z) = \sigma_\varepsilon^2 c(z)c(z^{-1})$.

Example: For the MA(1) model $x_t = (1 - c_1 L)\varepsilon_t$

$\gamma_0 = \sigma_\varepsilon^2(1 + c_1^2)$, $\gamma_{-1} = \gamma_1 = -\sigma_\varepsilon^2 c_1$, and $\gamma_k = 0$ for $|k| > 1$. Thus

$$
\begin{aligned}
\gamma(z) &= \sum_{j=-\infty}^{\infty} \gamma_j z^j \\
&= \gamma_{-1} z^{-1} + \gamma_0 z^0 + \gamma_1 z^1 \\
&= \sigma_\varepsilon^2 \left(-c_1 z^{-1} + (1 + c_1^2) - c_1 z \right) \\
&= \sigma_\varepsilon^2 (1 - c_1 z)(1 - c_1 z^{-1})
\end{aligned}
$$

18. Spectral Representation Theorem(e.g, Brockwell and Davis (1991)): Suppose Y_t is a discrete time covariance stationary zero mean process, then there exists an orthogonal-increment process $Z(\omega)$ such that

(i) $\mathrm{Var}(Z(\omega)) = F(\omega)$

and

(ii) $Y_t = \int_{-\pi}^{\pi} e^{it\omega}\, dZ(\omega)$

where F is the spectral distribution function of the process. (The spectral density, $S(\omega)$, is the density associated with F.)

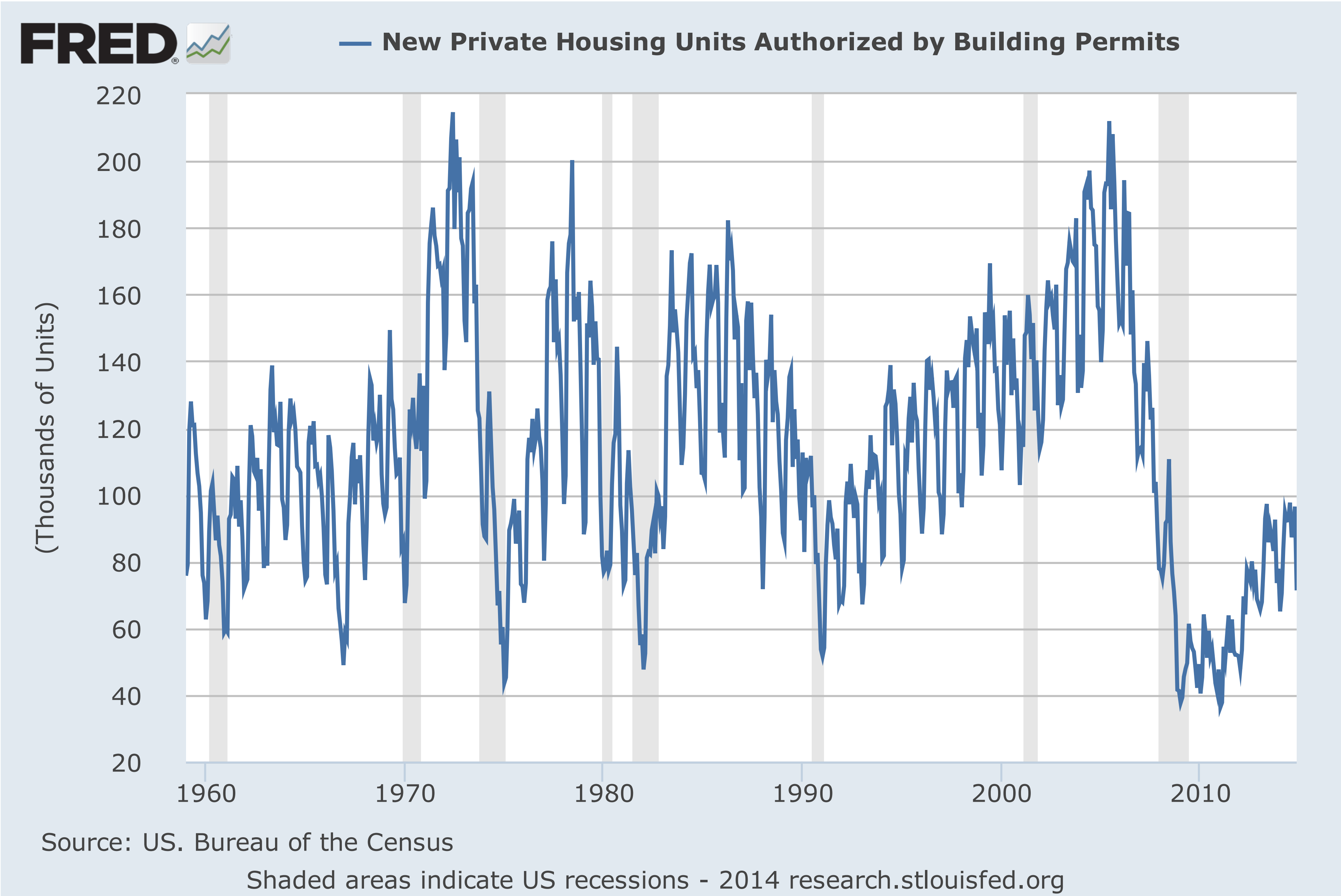
This is a useful and important decomposition, and we'll spend some time discussing it.

Lecture Outline

~~1. Time Series Basics~~

2. Spectral representation of stationary process

3. Approximation tools (CLT, FCLT, etc.)



FRED
New Private Housing Units Authorized by Building Permits
220
200
180
160
140
120
100
80
60
40
20
(Thousands of Units)
1960
1970
1980
1990
2000
2010
Source: US. Bureau of the Census
Shaded areas indicate US recessions - 2014 research.stlouisfed.org

Some questions

1. How important are the "seasonal" or "business cycle" components in Y_t?

2. Can we measure the variability at a particular frequency? Frequency 0 (long-run) will be particularly important as that is what HAC/HAR Covariance matrices are all about.

3. Can we isolate/eliminate the "seasonal" ("business-cycle") component? (Ex-Post vs. Real Time).

2.1 Spectral representation of a covariance stationary stochastic process

Deterministic processes:

(a) $Y_t = \cos(\omega t)$, strictly periodic with $period = \dfrac{2\pi}{\omega}$,

$Y_0 = 1$

amplitude = 1.

(b) $Y_t = a\times\cos(\omega t) + b\times\sin(\omega t)$, strictly period with $period = \dfrac{2\pi}{\omega}$,

$Y_0 = a$

amplitude $= \sqrt{a^2 + b^2}$

Stochastic process:

$Y_t = a \times \cos(\omega t) + b \times \sin(\omega t)$, a and b are random variables, 0-mean, mutually uncorrelated, with common variance σ^2.

2^{nd} - moments :

$\mathrm{E}(Y_t) = 0$

$\mathrm{Var}(Y_t) = \sigma^2 \times \{\cos^2(\omega t) + \sin^2(\omega t)\} = \sigma^2$

$$
\begin{aligned}
Cov(Y_t, Y_{t-k}) &= \sigma^2 \{\cos(\omega t)\cos(\omega(t-k)) + \sin(\omega t)\sin(\omega(t-k))\} \\
&= \sigma^2 \cos(\omega k)
\end{aligned}
$$

Stochastic process with more components:

$$Y_t = \sum_{j=1}^{n} \{a_j \cos(\omega_j t) + b_j \sin(\omega_j t)\},$$ $\{a_j, b_j\}$ are uncorrelated 0-mean random variables, with $\mathrm{Var}(a_j) = \mathrm{Var}(b_j) = \sigma_j^2$

2^{nd} - moments :

$\mathrm{E}(Y_t) = 0$

$$\mathrm{Var}(Y_t) = \sum_{j=1}^{n} \sigma_j^2 \qquad \text{(Decomposition of variance)}$$

$$\mathrm{Cov}(Y_t Y_{t-k}) = \sum_{j=1}^{n} \sigma_j^2 \cos(\omega_j k) \qquad \text{(Decomposition of auto-covariances)}$$

Stochastic Process with even more components:

$$Y_t = \int_0^{\pi} \cos(\omega t)\, da(\omega) + \int_0^{\pi} \sin(\omega t)\, db(\omega)$$

$da(\omega)$ and $db(\omega)$: random variables, 0-mean, mutually uncorrelated, uncorrelated across frequency, with common variance that depends on frequency. This variance function, say $S(\omega)$, is called the spectrum.

.. **Digression:** A convenient change of notation:

$$Y_t = a\times\cos(\omega t) + b\times\sin(\omega t)$$

$$= \frac{1}{2}e^{i\omega}(a - ib) + \frac{1}{2}e^{-i\omega}(a + ib)$$

$$= e^{i\omega}Z + e^{-i\omega}\bar{Z}$$

where $i = \sqrt{-1}$ and $e^{i\omega} = cos(\omega) + i\times sin(\omega)$, $Z = \frac{1}{2}(a - ib)$ and $\bar{Z}$ is the complex conjugate of Z.

Similarly

$$Y_t = \int_0^{\pi} \cos(\omega t)\, da(\omega) + \int_0^{\pi} \sin(\omega t)\, db(\omega)$$

$$= \frac{1}{2}\int_0^{\pi} e^{i\omega t}(da(\omega) - i\, db(\omega)) + \frac{1}{2}\int_0^{\pi} e^{-i\omega t}(da(\omega) + i\, db(\omega))$$

$$= \int_{-\pi}^{\pi} e^{i\omega t}\, dZ(\omega)$$

where $dZ(\omega) = \frac{1}{2}(da(\omega) - idb(\omega))$ for $\omega \geq 0$ and $dZ(-\omega) = \overline{dZ(\omega)}$ for $\omega >$ 0.

Because *da* and *db* have mean zero, so does *dZ*. Denote the variance of $dZ(\omega)$ as Var($dZ(\omega)$) = E($dZ(\omega)\overline{dZ(\omega)}$)=$S(\omega)d\omega$, and using the assumption that *da* and *db* are uncorrelated across frequency E($dZ(\omega)\overline{dZ(\omega)'}$)=0 for $\omega \neq \omega'$.

Second moments of Y:

$$E(Y_t) = E\left\{\int_{-\pi}^{\pi} e^{i\omega t}\, dZ(\omega)\right\} = \int_{-\pi}^{\pi} e^{i\omega t} E(dZ(\omega)) = 0$$

$$\gamma_k = E(Y_t Y_{t-k}) = E(Y_t \bar{Y}_{t-k}) = E\left\{\int_{-\pi}^{\pi} e^{i\omega t}\, dZ(\omega) \int_{-\pi}^{\pi} e^{-i\omega(t-k)} \overline{dZ(\omega)}\right\}$$

$$= \int_{-\pi}^{\pi} e^{i\omega t} e^{-i\omega(t-k)} E(dZ(\omega)\overline{dZ(\omega)})$$

$$= \int_{-\pi}^{\pi} e^{i\omega k} S(\omega)\, d\omega = 2\int_{0}^{\pi} \cos(\omega k) S(\omega)\, d\omega$$

where the last equality follows from $S(\omega) = S(-\omega)$.

Setting $k = 0$, $\gamma_0 = \mathrm{Var}(Y_t) = \int_{-\pi}^{\pi} S(\omega)\, d\omega$

… End of Digression

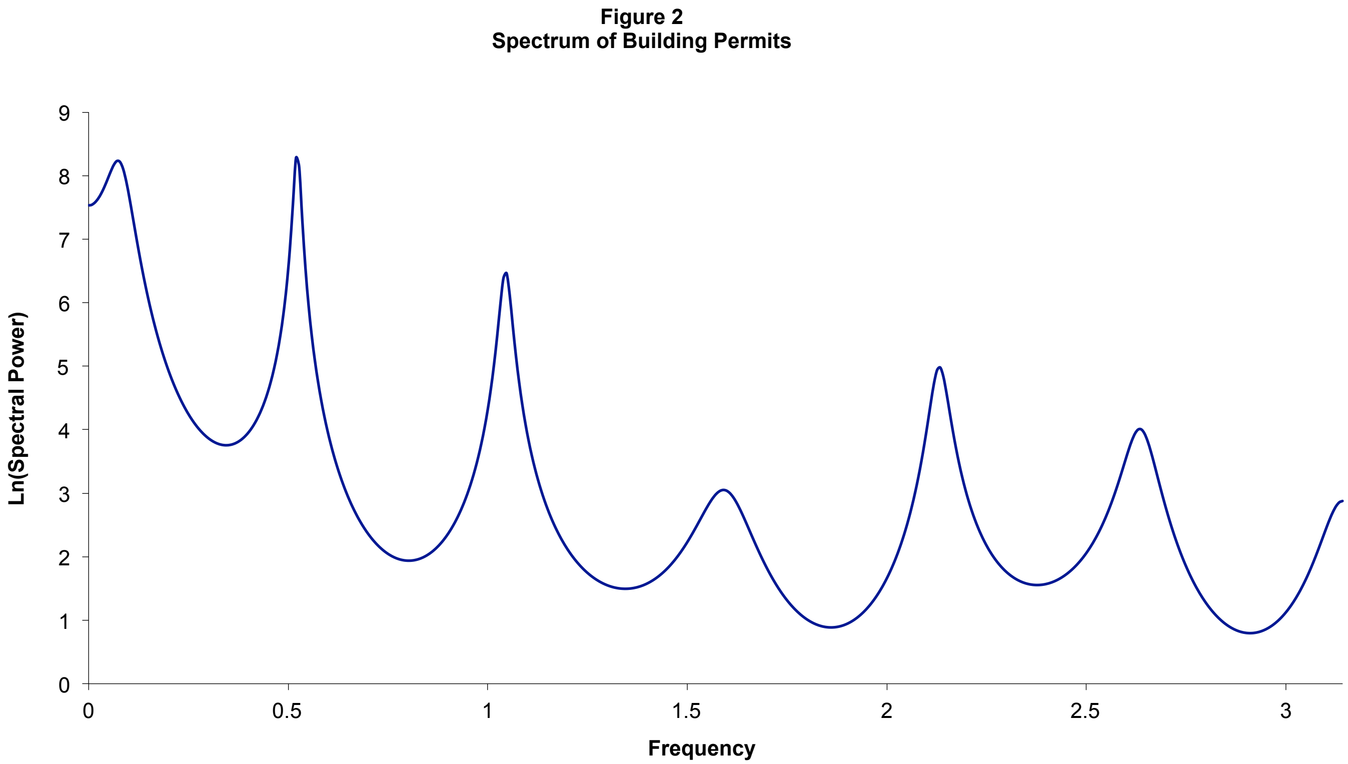
Summarizing

1. $S(\omega)d\omega$ can be interpreted as the variance of the cyclical component of Y corresponding to the frequency ω. The period of this component is $period = 2\pi/\omega$.

2. $S(\omega) \geq 0$ (it is a variance)

3. $S(\omega) = S(-\omega)$. Because of this symmetry, plots of the spectrum are presented for frequencies $0 \leq \omega \leq \pi$.

Example: The Spectrum of Building Permits



Most of the mass in the spectrum is concentrated around the seven peaks evident in the plot. (These peaks are sufficiently large that spectrum is plotted on a log scale.) The first peak occurs at frequency $\omega = 0.07$ corresponding to a period of 90 months. The other peaks occur at frequencies $2\pi/12$, $4\pi/12$, $6\pi/12$, $8\pi/12$, $10\pi/12$, and π. These are peaks for the seasonal frequencies: the first corresponds to a period of 12 months, and the others are the seasonal "harmonics" 6, 4, 3, 2.4, 2 months. (These harmonics are necessary to reproduce an arbitrary − not necessary sinusoidal − seasonal pattern.)

4. $\gamma_k = \int_{-\pi}^{\pi} e^{i\omega k} S(\omega)\, d\omega = 2\int_0^{\pi} \cos(\omega k) S(\omega)\, d\omega$ can be inverted to yield

$$S(\omega) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} e^{-i\omega k}\gamma_k = \frac{1}{2\pi}\left\{\gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \cos(\omega k)\right\}$$

“Long-Run Variance”

The long-run variance is $S(0)$, the variance of the 0-frequency (or ∞-period component).

Since $S(\omega) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} e^{-i\omega k}\gamma_k$, then $S(0) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma_k e^{-ik0} = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma_k$.

As we will see, this plays an important role in statistical inference because (except for the factor 2π) it is the large-sample variance of the sample mean.

5. Recall that the ACGF is $\gamma(z) = \sum \gamma_j z^j$.

Thus, $S(\omega) = (2\pi)^{-1}\gamma(z)$, with $z = e^{i\omega}$.

Application: If x_t follows an ARMA process, then it can be represented as

$\phi(\mathrm{L})x_t = \theta(\mathrm{L})\varepsilon_t$, or $x_t = c(\mathrm{L})\varepsilon_t$ with $c(\mathrm{L}) = \theta(\mathrm{L})/\phi(\mathrm{L})$.

The ACGF is therefore $\gamma(z) = \sigma_\varepsilon^2 c(z)c(z^{-1}) = \sigma_\varepsilon^2 \dfrac{\theta(z)}{\phi(z)}\dfrac{\theta(z^{-1})}{\phi(z^{-1})}$, or

$$\gamma(z) = \sigma_\varepsilon^2 \frac{(1-\theta_1 z - ... - \theta_q z^q)(1-\theta_1 z^{-1} - ... - \theta_q z^{-q})}{(1-\phi_1 z - ... - \phi_p z^p)(1-\phi_1 z^{-1} - ... - \phi_p z^{-p})}$$

and the spectrum is

$$S_y(\omega) = (2\pi)^{-1}\sigma_\varepsilon^2 \frac{(1-\theta_1 e^{i\omega} - ... - \theta_q e^{iq\omega})(1-\theta_1 e^{-i\omega} - ... - \theta_q e^{-iq\omega})}{(1-\phi_1 e^{i\omega} - ... - \phi_p e^{ip\omega})(1-\phi_1 e^{-i\omega} - ... - \phi_p e^{-ip\omega})}$$

This suggests a simple (parametric) method for estimating the spectrum of a series:

(1) Estimate an appropriate ARMA model, say
$\hat{\phi}(L)y_t = \hat{\theta}(L)\varepsilon_t$

(2) Plug in estimated ARMA parameter values to form

$$\hat{S}_y(\omega) = (2\pi)^{-1}\hat{\sigma}_\varepsilon^2 \frac{(1-\hat{\theta}_1 e^{i\omega} - \ldots - \hat{\theta}_q e^{iq\omega})(1-\hat{\theta}_1 e^{-i\omega} - \ldots - \hat{\theta}_q e^{-iq\omega})}{(1-\hat{\phi}_1 e^{i\omega} - \ldots - \hat{\phi}_p e^{ip\omega})(1-\hat{\phi}_1 e^{-i\omega} - \ldots - \hat{\phi}_p e^{-ip\omega})}$$

Non-parametric estimators based on the "Periodogram" will be discussed in the lecture on HAC/HAR standard errors.

Lecture Outline

~~1. Time Series Basics~~

~~2. Spectral representation of stationary process~~

3. Approximation tools (CLT, FCLT, etc.)

3 familiar notions

1. Convergence in distribution or "weak convergence": ξ_T, $T = 1, 2, \ldots$ is a sequence of random variables.

$\xi_T \Rightarrow \xi$ (or $\xi_T \xrightarrow{d} \xi$) means that the probability distribution function (PDF) of ξ_T converges to the PDF of ξ. (Equivalently, $\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X))$ for any continuous bounded function g.)

As a practical matter this means that we can approximate the PDF of ξ_T using the PDF of ξ when T is large.

2. Central Limit Theorem: Let ε_t be a mds$(0,\sigma_\varepsilon^2)$ with $2+\delta$ moments and $\xi_T = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t$. Then $\xi_T \Rightarrow \xi \sim$ N$(0, \sigma_\varepsilon^2)$.

(Digression − Additional persistence …

Suppose $a_t = \varepsilon_t - \theta\varepsilon_{t-1} = (1-\theta\text{L})\varepsilon_t = \theta(\text{L})\varepsilon_t$. Then

$$T^{-1/2}\sum_{t=1}^{T} a_t = T^{-1/2}\sum_{t=1}^{T}(\varepsilon_t - \theta\varepsilon_{t-1}) = T^{-1/2}\sum_{t=1}^{T}\varepsilon_t - \theta\sum_{t=0}^{T-1}\varepsilon_t = (1-\theta)T^{-1/2}\sum_{t=1}^{T}\varepsilon_t + \theta T^{-1/2}(\varepsilon_T - \varepsilon_0)$$

But $\theta T^{-1/2}(\varepsilon_T - \varepsilon_0)$ is negligible, so that

$$T^{-1/2}\sum_{t=1}^{T} a_t = (1-\theta)T^{-1/2}\sum_{t=1}^{T}\varepsilon_t + o_p(1) \Rightarrow (1-\theta)\xi \sim N\left(0, \sigma_\varepsilon^2(1-\theta)^2\right)$$

Note: $\sigma_\varepsilon^2(1-\theta)^2 = \sigma_\varepsilon^2\,\theta(1)^2 = \sigma_\varepsilon^2\,\theta(e^{i\omega})\theta(e^{-i\omega})$ with $\omega = 0$

and is the “long-run” variance of a.

This generalizes: suppose $a_t = \theta(\mathrm{L})\varepsilon_t$ and $\sum_{i=0}^{\infty} i\,|\theta_i| < \infty$ (so that the MA coefficients are "one-summable"), then

$$T^{-1/2}\sum_{t=1}^{T} a_t = \theta(1)T^{-1/2}\sum_{t=1}^{T}\varepsilon_t + o_p(1) \Rightarrow \theta(1)\xi \sim N\left(0,\sigma_\varepsilon^2\theta(1)^2\right)$$

and $\sigma_\varepsilon^2\theta(1)^2$ is the long-run variance of a.

… End of Digression)

3. Continuous mapping theorem. Let g be a continuous function and $\xi_T \Rightarrow \xi$, then $g(\xi_T) \Rightarrow g(\xi)$.

Example ξ_T is the usual t-statistic, and $\xi_T \Rightarrow \xi \sim N(0, 1)$, then $\xi_T^2 \Rightarrow \xi^2 \sim \chi_1^2$.

These ideas can be extended to random functions:

A particular random function: The Wiener Process, a continuous-time stochastic process sometimes called Standard Brownian Motion that will play the role of a "standard normal" in the relevant function space.

Denote the process by $W(s)$ defined on $s \in [0,1]$ with the following properties

1. $W(0) = 0$

2. For any dates $0 \leq t_1 < t_2 < \ldots < t_k \leq 1$, $W(t_2)-W(t_1)$, $W(t_3)-W(t_4)$, … , $W(t_k)-W(t_{k-1})$ are independent normally distributed random variables with $W(t_i)-W(t_{i-1}) \sim N(0, t_i-t_{i-1})$.

3. Realizations of W(s) are continuous w.p. 1.

From (1) and (2), note that $W(1) \sim N(0,1)$.

Another Random Function: Suppose $\varepsilon_t \sim$ iidN(0,1), $t = 1, \ldots, T$, and let $\xi_T(s)$ denote the function that linearly interpolates between the points $\xi_T(t/T) = \frac{1}{\sqrt{T}}\sum_{i=1}^{t}\varepsilon_i$.

Can we use W to approximate the probability law of $\xi_T(s)$ if T is large?

More generally, we want to know whether the probability distibution of a random function can be well approximated by the PDF of another (perhaps simpler, maybe Gaussian) function when T is large. Formally, we want to study weak convergence on function spaces.

Useful References: Hall and Heyde (1980), Davidson (1994), Andrews (1994)

Suppose we limit our attention to continuous functions on $s \in [0,1]$ (the space of such functions is denoted C[0,1]), and we define the distance between two functions, say x and y as $d(x,y) = sup_{0 \leq s \leq 1} |x(s) - y(s)|$.

Three important theorems (Hall and Heyde (1980) and Davidson (1994, part VI):

Important Theorm 1: (Hall and Heyde Theorem A.2) Weak Convergence of random functions on C[0,1]

Weak convergence follows from (i) and (ii), where

(i) Let $0 \le s_1 < s_2 \ldots < s_k \le 1$, a set of k points. Suppose that $(\xi_T(s_1), \xi_T(s_2), \ldots, \xi_T(s_k)) \Rightarrow (\xi(s_1), \xi(s_2), \ldots, \xi(s_k))$ for any set of k points, $\{s_i\}$.

(ii) The function $\xi_T(s)$ is "tight" (or more generally satisfies "stochastic equicontinuity" as discussed in Andrews (1994)), meaning

(a) For each $\varepsilon > 0$, $\text{Prob}[\sup_{|s-t|<\delta}|\xi_T(s) - \xi_T(t)| > \varepsilon\,] \to 0$ as $\delta \to 0$ uniformly in T. (This says that the function ξ_T does not get too "wild" as T grows.)

(b) $\text{Prob}[|\xi_T(0)| > \delta] \to 0$ as $\delta \to \infty$ uniformly in T. (This says the function ξ_T can't get too crazy at the origin ast T grows.)

Important Theorem 2: (Hall on Heyde Theorem A.3) Continuous Mapping Theorem

Let g: C[0,1] $\rightarrow \mathbb{R}$ be a continuous function and suppose $\xi_T(.) \Rightarrow \xi(.)$.

Then $g(\xi_T) \Rightarrow g(\xi)$.

Important Theorem 3: (Hall and Heyde) Functional Central Limit Theorem:

Suppose $\varepsilon_t \sim$ mds with variance σ_ε^2 and bounded 2+δ moments for some $\delta > 0$.

(a) Let $\xi_T(s)$ denote the function that linearly interpolates between the points $\xi(t/T) = \frac{1}{\sqrt{T}}\sum_{i=1}^{t}\varepsilon_i$. Then $\xi_T \Rightarrow \sigma_\varepsilon W$, where W is a Wiener process (standard Brownian motion).

(b) The results can be extended to $\xi_T(s) = \frac{1}{\sqrt{T}}\sum_{i=1}^{[sT]}\varepsilon_i$, the step-function interpolation, where [.] is the "less than or equal to integer function" (so that [3.1] = 3, [3.0] = 3, [3.9999] = 3, and so forth).

See Davidson Ch. 29 for extensions.

An Example:

(1): Let $x_t = \sum_{i=1}^{t} \varepsilon_i$, where ε_i is $mds(0,\sigma_\varepsilon^2)$, and let $\xi_T(s) = \frac{1}{\sqrt{T}} \sum_{i=1}^{[sT]} \varepsilon_i = \frac{1}{\sqrt{T}} x_{[sT]}$

be a step function approximation of $W(s)$.

Then

$$v_T = \frac{1}{T^{3/2}} \sum_{t=1}^{T} x_t = \frac{1}{T} \sum_{t=1}^{T} \left[\frac{1}{T^{1/2}} \sum_{i=1}^{t} \varepsilon_i \right] = \sigma_\varepsilon \int_0^1 \xi_T(s) ds \Rightarrow \sigma_\varepsilon \int_0^1 W(s) ds = v$$

What does this all mean?

Suppose I want to approximate the 95$^{\text{th}}$ quantile of the distribution of, say,

$v_T = \frac{1}{T^{3/2}} \sum_{t=1}^{T} x_t$. Because $v_T \Rightarrow v = \sigma_\varepsilon \int_0^1 W(s)ds$, I can use the 95$^{\text{th}}$ quantile of v are the approximator.

How do I find (or approximate) the 95$^{\text{th}}$ quantile of v?

Use Monte Carlo draws of $\sigma_\varepsilon N^{-3/2} \sum_{t=1}^{N} \sum_{i=1}^{t} z_i$ where $z_i \sim$ iidN(0,1) and N is very large.

This approximation works well when T is reasonably large, and does not require knowledge of the distribution of x.

(Digression − Additional persistence …

Suppose $a_t = \theta(\mathrm{L})\varepsilon_t$, where the θ-coefficients are 1-summable.

And, suppose $x_t = x_{t-1} + a_t$.

Then $T^{-1/2}x_{[sT]} \Rightarrow \theta(1)\sigma_\varepsilon W(s)$.

Note: $\theta(1)\sigma_\varepsilon$ is the "long-run" standard deviation of a.

… End of Digression)

Application: Testing for a "Break"

Model: $y_t = \beta_t + \varepsilon_t$, where $\varepsilon_t \sim$ iid $(0, \sigma_\varepsilon^2)$

$$\beta_t = \begin{cases} \beta \text{ for } t \leq \tau \\ \beta + \delta \text{ for } t > \tau \end{cases}$$

Null and alternative: H_o: $\delta = 0$ vs. H_o: $\delta \neq 0$

Tests for H_o vs. H_a depends on whether τ is known or unknown.

Chow Tests (known break date)

Least squares estimator of δ: $\hat{\delta} = \bar{Y}_2 - \bar{Y}_1$

where $\bar{Y}_1 = \frac{1}{\tau}\sum_{t=1}^{\tau} y_t$ and $\bar{Y}_2 = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} y_t$

Wald statistic: $\xi_W = \frac{1}{\hat{\sigma}_\varepsilon^2}\frac{\hat{\delta}^2}{(\frac{1}{\tau}+\frac{1}{T-\tau})} \Rightarrow \xi \sim \chi_1^2$

Follows from $\bar{Y}_1 \overset{a}{\sim} N(\beta, \frac{\sigma_e^2}{\tau})$ and $\bar{Y}_2 \overset{a}{\sim} N(\beta+\delta, \frac{\sigma_\varepsilon^2}{T-\tau})$ and they are

independent so that $\hat{\delta} \overset{a}{\sim} N\left(\delta, \sigma_\varepsilon^2\left(\frac{1}{\tau}+\frac{1}{T-\tau}\right)\right)$

Under H_o ξ_W is distributed as a χ_1^2 random variable in large (τ and $T-\tau$) samples. Thus, critical values for the test can be determined from the χ^2 distribution.

Quandt Tests (Sup Wald or QLR) (unknown break date)

Quandt (1960) suggested computing the Chow statistic for a large number of possible values of τ and using the largest of these as the test statistics.

$$\text{QLR statistic: } \xi_Q = \max_{\tau_1 \le \tau \le \tau_2} \xi_W(\tau)$$

where the Chow statistic ξ_W is now indexed by the break date.

The problem is then to find the distribution of ξ_Q under the null (it will not be χ^2), so that the critical value for the test can be determined.

Let $s = \tau/T$. Under the null $\delta = 0$, and (now using s as the index), we can then write ξ_W as

$$\xi_{W,T}(s) = \frac{1}{\hat{\sigma}_e^2}\frac{\left[\frac{1}{[sT]}\sum_{t=1}^{[sT]} y_t - \frac{1}{[(1-s)T]}\sum_{t=[sT]+1}^{T} y_t\right]^2}{\frac{1}{[sT]}+\frac{1}{[(1-s)T]}}$$

$$\overset{H_o}{=} \frac{1}{\hat{\sigma}_e^2}\frac{[\frac{1}{[sT]}\sum_{t=1}^{[sT]} \varepsilon_t - \frac{1}{[(1-s)T]}\sum_{t=[sT]+1}^{T} \varepsilon_t]^2}{\frac{1}{[sT]}+\frac{1}{[(1-s)T]}}$$

$$= \frac{1}{\hat{\sigma}_e^2}\frac{[\frac{1}{s}\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]} \varepsilon_t - \frac{1}{(1-s)}\frac{1}{\sqrt{T}}\sum_{t=[sT]+1}^{T} \varepsilon_t]^2}{\frac{1}{s}+\frac{1}{(1-s)}}$$

$$= \frac{[\frac{1}{s}W_T^a(s) - \frac{1}{(1-s)}(W_T^a(1) - W_T^a(s))]^2}{\frac{1}{s}+\frac{1}{(1-s)}} = \frac{[W_T^a(s) - sW_T^a(1)]^2}{s(1-s)}$$

where $W_T^a(s) = \dfrac{1}{\hat{\sigma}_\varepsilon}\dfrac{1}{\sqrt{T}}\sum_{t=1}^{[sT]}\varepsilon_t$, and the last equality follows from multiplying the numerator and denominator by $s^2(1-s)^2$ and simplifying.

Thus, using FCLT, $\xi_{W,T} \Rightarrow \xi$, where $\xi(s) = \dfrac{[W(s) - sW(1)]^2}{s(1-s)}$.

Suppose that τ_1 is chosen as $[\lambda T]$ and τ_2 is chosen as $[(1-\lambda)T]$, where $0 < \lambda < 0.5$. Then

$$\xi_Q = \sup_{\lambda \le s \le (1-\lambda)} \xi_{W,T}(s), \text{ and } \xi_Q \Rightarrow \sup_{\lambda \le s \le (1-\lambda)} \xi(s)$$

It has become standard practice to use a value of $\lambda = 0.15$.

The results have been derived here for the case of a single constant regressor. Exensions to the case of multiple (non-constant) regressors can be found in Andrews (1993) (Critical values for the test statistic are also given in Andrews (1993) with corrections in Andrews (2003), reprinted in Stock and Watson (2014).)

TABLE 14.6 Critical Values of the QLR Statistic with 15% Trimming

Number of Restrictions (q)	10%	5%	1%
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23
11	2.40	2.62	3.09
12	2.33	2.54	2.97
13	2.27	2.46	2.87
14	2.21	2.40	2.78
15	2.16	2.34	2.71
16	2.12	2.29	2.64
17	2.08	2.25	2.58
18	2.05	2.20	2.53
19	2.01	2.17	2.48
20	1.99	2.13	2.43

These critical values apply when $\tau_0 = 0.15T$ and $\tau_1 = 0.85T$ (rounded to the nearest integer), so that the F-statistic is computed for all potential break dates in the central 70% of the sample. The number of restrictions q is the number of restrictions tested by each individual F-statistic. Critical values for other trimming percentages are given in Andrews (2003).

Lecture Outline

~~1. Time Series Basics~~

~~2. Spectral representation of stationary process~~

~~3. Approximation tools (CLT, FCLT, etc.).~~

Course Topics

~~1. Time series refresher and inference tools (MW)~~

2. The Kalman filter, nonlinear filtering, and Markov chain monte carlo (MW)
3. Prediction with large datasets (MW)
4. Heteroskedasticity and autocorrelation consistent (HAC) standard errors (JS)
5. Many instruments/weak identification in IV and GMM (JS)
6. Structural VAR modeling (JS)

References for Lecture 1

Andrews, D.W.K. (1993), "Tests for Parameter Instability and Structural Change with Unknown Change Point" *Econometrica*, 61(4): 821-856.

Andrews, D.W.K. (1994), "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Vol. 4., Robert F. Engle and Daniel L. McFadden (eds), Amstedam: Elsevier.

Andrews, D.W.K. (2003), "Tests for Parameter Instability and Structural Change with Unknown Change Point: A Corrigendum," *Econometrica*, 71(1): 395-397.

Brockwell, P.J., and R.A. Davis (1991), *Time Series*: *Theory and Methods*, 2nd Edition, New York: Springer Verlag.

Chow, Gregory (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28(3): July, 591-605.

Davidson, J. (1994), *Stochastic Limit Theory*, Oxford: Oxford University Press.

Hall, P., and C. Heyde (1980), *Martingale Limit Theory and Its Application (Probability and Mathematical Statistics)*, Academic Press.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton: Princeton University Press

Hayashi, F. (2000), *Econometrics*. Princeton: Princeton University Press.

Priestly, M.B. (1981), *Spectral Analysis and Time Series*, London: Academic Press.

Quandt, Richard (1960), "Tests of the Hypothesis That a Linear Regression System Obeys Two Separate Regimes," *Journal of the American Statistical Association*, 55(290): 324-330.

Stock, James H., and Mark W. Watson (2014), *Introduction to Econometrics*, 3rd Updated Edition, Prentice Hall.

AEA Continuing Education Course

Time Series Econometrics

Lecture 2: The Kalman filter, nonlinear filtering, and Markov chain monte carlo

Mark W. Watson

January 6, 2015

8:15AM-10:15AM

Outline

1. A motivating example
2. Models, objects of interest, and general formulae
3. Special Cases
4. MCMC (Gibbs)
5. Likelihood Evaluation

1. A motivating example: Cogley and Sargent (2014)

How "uncertain" and "instable" have prices been in the U.S. from 1850-2012, and how did uncertainty/instability change over this historical period?

Price level and inflation: $p_t = \ln(P_t)$ and $\pi_t = p_t - p_{t-1}$

Changes: $p_{t+h} - p_t = \pi_{t+1} + \pi_{t+2} + \ldots + \pi_{t+h}$

Uncertainty: $\mathrm{Var}(p_{t+h} - p_t \,|\boldsymbol{Y_t})$

Instability: $\mathrm{E}(p_{t+h} - p_t \,|\boldsymbol{Y_t})^2 + \mathrm{Var}(p_{t+h} - p_t \,|\boldsymbol{Y_t})$

A Model:

UC/Local-Level/IMA(1,1) model

(Nelson-Schwert (1977), Harvey (1989), others)

$$\pi_t \quad = \quad \tau_t \quad + \quad \varepsilon_t$$

Persistent (martingale)   Not-Persistent (white noise)

($\Delta\tau$, ε): heteroskedastic ("UCSV")

(Stock-Watson (2007), Shephard (2013), Cogley-Sargent (2014), others).

Note: CS also incorporate a "measurement error" component.

$$\pi_t \quad = \quad \tau_t \quad + \quad \varepsilon_t$$

Challenges:

(1) Estimation of τ_t and ε_t ?

(2) Estimation of $\sigma_{\Delta\tau}$ and σ_ε ?

(3) Estimation of $\sigma_{\Delta\tau}(t)$ and $\sigma_\varepsilon(t)$?

(4) $\mathrm{E}(p_{t+h} - p_t \mid \boldsymbol{Y_t})$ and $\mathrm{Var}(p_{t+h} - p_t \mid \boldsymbol{Y_t})$

~~1. A motivating example~~

2. Models, objects of interest, and general formulae

3. Special Cases

4. MCMC (Gibbs)

5. Likelihood Evaluation

2. General Model (Nonlinear, non-Gaussian state-space model)

$$y_t = H(s_t, \varepsilon_t)$$

$$s_t = F(s_{t-1}, \eta_t)$$

$$\varepsilon \text{ and } \eta \sim \text{iid}$$

Example 1: Linear Gaussian Model

$$y_t = Hs_t + \varepsilon_t$$

$$s_t = Fs_{t-1} + \eta_t$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim iidN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix} \right)$$

Applications:

- Unobserved component models (s is serially correlated part of y)
- Factor Models (many y's, few s's)
- TVP Regression models ($H = H_t = x_t$, $s_t = \beta_t$)

Example 2: Hamilton Regime-Switching Model

$$y_t = \mu(s_t) + \sigma(s_t)\varepsilon_t$$

$s_t = 0$ or 1 with $P(s_t = i \mid s_{t-1} = j) = p_{ij}$

(using $s_t = F(s_{t-1},\eta_t)$ notation:

$s_t = \mathbf{1}(\eta_t \le p_{10} + (p_{11} - p_{10})s_{t-1})$, where $\eta \sim \mathrm{U}[0,1]$)

Example 3: Stochastic volatility model

$$y_t = e^{s_t}\varepsilon_t$$

$$s_t = \mu + \phi(s_{t-1} - \mu) + \eta_t$$

with, say, $\varepsilon_t \sim \text{iid}(0,1)$ and $e^{s_t} = \sigma_t$, the model for y is

$$y_t \mid s_t \sim \text{N}(0, \sigma_t^2)$$

Some things you might want to calculate

Notation: $y_{1:t} = (y_1, y_2, \ldots, y_t)$, $s_{1:t} = (s_1, s_2, \ldots, s_t)$,

$f(. \mid .)$ a generic density function.

A. Prediction and Likelihood

(i) $f(s_t \mid y_{1:t-1})$

(ii) $f(y_t \mid y_{1:t-1})$ … Note $f(y_{1:\mathrm{T}}) = \prod_{t=1}^{T} f(y_t \mid y_{1:t-1})$ is the likelihood

B. Filtering: $f(s_t \mid y_{1:t})$

C. Smoothing: $f(s_t \mid y_{1:T})$.

General Recursive Formulae (Kitagawa (1987)):

Model: $y_t = H(s_t, \varepsilon_t)$, $s_t = F(s_{t-1}, \eta_t)$, ε and η ~ iid

A. Prediction of s_t and y_t given Y_{t-1}.

(i)

$$\begin{aligned} f(s_t \mid y_{1:t-1}) &= \int f(s_t, s_{t-1} \mid y_{1:t-1})\, ds_{t-1} \\ &= \int f(s_t \mid s_{t-1}, y_{1:t-1}) f(s_{t-1} \mid y_{1:t-1})\, ds_{t-1} \\ &= \int f(s_t \mid s_{t-1}) f(s_{t-1} \mid y_{1:t-1})\, ds_{t-1} \end{aligned}$$

(ii) $f(y_t \mid y_{1:t-1}) = \int f(y_t \mid s_t) f(s_t \mid y_{1:t-1})\, ds_t$ ("t" component of likelihood)

Model: $y_t = H(s_t, \varepsilon_t)$, $s_t = F(s_{t-1}, \eta_t)$, ε and η ~ iid

B. Filtering

$$f(s_t \mid y_{1:t}) = f(s_t \mid y_t, y_{1:t-1}) = \frac{f(y_t \mid s_t, y_{1:t-1}) f(s_t \mid y_{1:t-1})}{f(y_t \mid y_{1:t-1})} = \frac{f(y_t \mid s_t) f(s_t \mid y_{1:t-1})}{f(y_t \mid y_{1:t-1})}$$

C. Smoothing

$$f(s_t \mid y_{1:T}) = \int f(s_t, s_{t+1} \mid y_{1:T})\, ds_{t+1} = \int f(s_t \mid s_{t+1}, y_{1:T}) f(s_{t+1} \mid y_{1:T})\, ds_{t+1}$$

$$= \int f(s_t \mid s_{t+1}, y_{1:t}) f(s_{t+1} \mid y_{1:T})\, ds_{t+1} = \int \left[\frac{f(s_{t+1} \mid s_t) f(s_t \mid y_{1:t})}{f(s_{t+1} \mid y_{1:t})} \right] f(s_{t+1} \mid y_{1:T})\, ds_{t+1}$$

$$= f(s_t \mid y_{1:t}) \int f(s_{t+1} \mid s_t) \frac{f(s_{t+1} \mid y_{1:T})}{f(s_{t+1} \mid y_{1:t})}\, ds_{t+1}$$

Outline

1. ~~A motivating example~~
2. ~~Models, objects of interest, and general formulae~~
3. **Special Cases**
4. MCMC (Gibbs)
5. Likelihood Evaluation

3. Special Cases

Model: $y_t = H(s_t, \varepsilon_t)$, $s_t = F(s_{t-1}, \eta_t)$, ε and η ~ iid

General Formulae depend on H, F, and densities of ε and η.

Well-known special case: Linear Gaussian Model

$$y_t = Hs_t + \varepsilon_t$$

$$s_t = Fs_{t-1} + \eta_t$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim iidN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix}\right)$$

In this case, all joint, conditional distributions and so forth are Gaussian, so that they depend only on mean and variance, and these are readily computed.

Digression: Recall that if

$$\begin{pmatrix} a \\ b \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right),$$

then $(a|b) \sim N(\mu_{a|b}, \Sigma_{a|b})$

where $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(b - \mu_b)$ and $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$.

Interpreting a and b appropriately yields the Kalman Filter and Kalman Smoother.

Model: $y_t = Hs_t + \varepsilon_t,\ \ s_t = Fs_{t-1} + \eta_t, \begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim iidN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix} \right)$

Let $s_{t/k} = E(s_t \mid y_{1:k}),\ P_{t/k} = \text{Var}(s_t \mid y_{1:k}),$

$\mu_{t/t-1} = E(y_t \mid y_{1:t-1}),\ \Sigma_{t/t\text{-}1} = \text{Var}(y_t \mid y_{1:t-1}).$

Deriving Kalman Filter:

Starting point: $s_{t-1} \mid y_{1:t-1} \sim N(s_{t-1/t-1}, P_{t-1/t-1})$. Then

$$\begin{pmatrix} s_t \\ y_t \end{pmatrix} \mid y_{1:t-1} \sim N\left(\begin{pmatrix} s_{t/t-1} \\ y_{t/t-1} \end{pmatrix}, \begin{pmatrix} P_{t/t-1} & P_{t/t-1}H' \\ HP_{t/t-1} & HP_{t/t-1}H' + \Sigma_\varepsilon \end{pmatrix} \right)$$

interpreting s_t as "a" and y_t as "b" yields the Kalman Filter.

Model: $y_t = Hs_t + \varepsilon_t,\ \ s_t = Fs_{t-1} + \eta_t, \begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim iidN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix}\right)$

Details of KF:

(i) $s_{t/t-1} = Fs_{t-1/t-1}$

(ii) $P_{t/t-1} = FP_{t-1/t-1}F' + \Sigma_\eta,$

(iii) $\mu_{t/t-1} = Hs_{t/t-1},$

(iv) $\Sigma_{t/t-1} = HP_{t/t-1}H' + \Sigma_\varepsilon$

(v) $K_t = P_{t/t-1}H'\Sigma_{t/t-1}^{-1}$

(vi) $s_{t/t} = s_{t/t-1} + K_t(y_t - \mu_{t/t-1})$

(vii) $P_{t/t} = (I - K_t)P_{t/t-1}.$

The log-likelihood is

$$L(Y_{1:T}) = \text{constant } -0.5\sum_{t=1}^{T}\left\{\ln|\Sigma_{t|t-1}| + (y_t - \mu_{t|t-1})'\Sigma_{t|t-1}^{-1}(y_t - \mu_{t/t-1})\right\}$$

The Kalman Smoother (for $s_{t|T}$ and $P_{t|T}$) is derived in analogous fashion (see Anderson and Moore (2005), or Hamilton (1990).)

5. A Stochastic Volatility Model (Linear, but non-Gaussian Model)

(With a slight change of notation)

$$x_t = \sigma_t e_t$$

$$\ln(\sigma_t) = \ln(\sigma_{t-1}) + \eta_t$$

or, letting $y_t = \ln(x_t^2)$, $s_t = \ln(\sigma_t)$ and $\varepsilon_t = \ln(e_t^2)$

$$y_t = 2\, s_t + \varepsilon_t$$

$$s_t = s_{t-1} + \eta_t$$

Complication: $\varepsilon_t \sim \ln(\chi_1^2)$

3 ways to handle the complication

(1) Ignore it (KF is Best Linear Filter. Gaussian MLE is QMLE)
Reference: Harvey, Ruiz, Shephard (1994)

(2) Work out analytic expressions for all the filters, etc. (Uhlig (1997) does this in a VAR model with time varying coefficients and stochastic volatility. He chooses densities and priors so that the recursive formulae yield densities and posteriors in the same family.)

(3) Numerical approximations to (2).

Numerical Approximations: A trick and a simulation method.

Trick: Shephard (1994), Approximate the distribution of $\varepsilon_t \sim \ln(\chi_1^2)$ by a mixture of normals, $\varepsilon_t = \sum_{i=1}^{n} q_{it} v_{it}$, where $v_{it} \sim$ iid$N(\mu_i, \sigma_i^2)$, and $P(q_{it}=1)=p_i$.

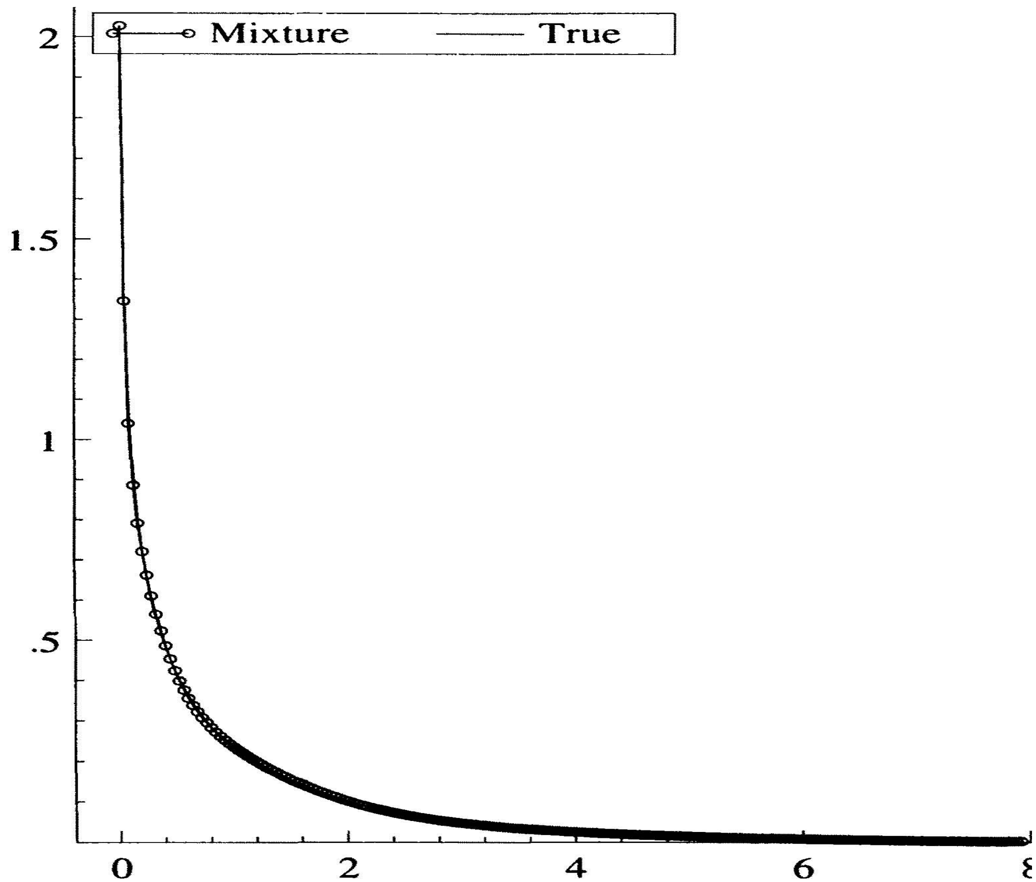
Table 1
Selection of $(p_j, m_j, v_j^2, a_j, b_j)$

j	KSC			$K = 10$		
	p_j	m_j	v_j^2	p_j	m_j	v_j^2
1	0.04395	1.50746	0.16735	0.00609	1.92677	0.11265
2	0.24566	0.52478	0.34023	0.04775	1.34744	0.17788
3	0.34001	−0.65098	0.64009	0.13057	0.73504	0.26768
4	0.25750	−2.35859	1.26261	0.20674	0.02266	0.40611
5	0.10556	−5.24321	2.61369	0.22715	−0.85173	0.62699
6	0.00002	−9.83726	5.17950	0.18842	−1.97278	0.98583
7	0.00730	−11.40039	5.79596	0.12047	−3.46788	1.57469
8				0.05591	−5.55246	2.54498
9				0.01575	−8.68384	4.16591
10				0.00115	−14.65000	7.33342

(numbers taken from Omori, Chib, Shephard, and Nakajima (2007)

χ_1^2 density and n= 7 mixture approximation

(picture taken from Kim, Shephard and Chib (1998))



Simulation method: MCMC methods (here Gibbs Sampling)

Some References: Casella and George (1992), Chib (2001), Fernandez-Villaverde (2014), Geweke (2005), Koop (2003).

Outline

1. ~~A motivating example~~
2. ~~Models, objects of interest, and general formulae~~
3. ~~Special Cases~~
4. **MCMC (Gibbs)**
5. Likelihood Evaluation

Markov Chain Monte Carlo (MCMC) methods

<u>Monte Carlo method:</u> Let a denote a random variable with density $f(a)$, and suppose you want to compute E$g(a)$ for some function g. (Mean, standard deviation, quantile, etc.)

Suppose you can simulate from $f(a)$. Then $\widehat{Eg(a)} = \frac{1}{N}\sum_{i=1}^{N} g(a_i)$, where a_i are draws from $f(a)$. If the Monte Carlo stochastic process is sufficiently well behaved, then $\widehat{Eg(a)} \xrightarrow[N]{p} = Eg(a)$ by the LLN.

<u>Markov Chains:</u> Methods for obtaining draws from $f(a)$. Suppose that it is difficult to draw from $f(a)$ directly. Choose draws a_1, a_2, a_3, … using a Markov chain.

Draw a_{i+1} from a conditional distribution, say $h(a_{i+1}|a_i)$, where h has the following properties:

(1) $f(a)$ is the invariant distribution associated with the Markov chain. (That is, if a_i is draw from f, then $a_{i+1}|a_i$ is a draw from f.)

(2) Draws can't be too dependent (or else $\widehat{Eg(a)} = \frac{1}{N}\sum_{i=1}^{N} g(a_i)$ will not be a good estimator of $Eg(a)$.)

Markov chain theory (see refs above) gives sufficient conditions on h that imply consistency and asymptotic normality of $\widehat{Eg(a)}$. In practice, diagnostics are used on the MC draws to see if there are problems.

How can $h(a_{i+1}|a_i)$ be constructed so that f is invariant distribution. Gibbs sampling is one way. (Others …)

Gibbs idea: partition a as $a = (a^1, a^2)$.

Then $f(a^1, a^2) = f(a^2|a^1)f(a^1) = f(a^1|a^2)f(a^2)$

This suggests the following: given the *i'th* draw of a, say $a_i = (a_i^1, a_i^2)$, generate a_{i+1} in two steps:

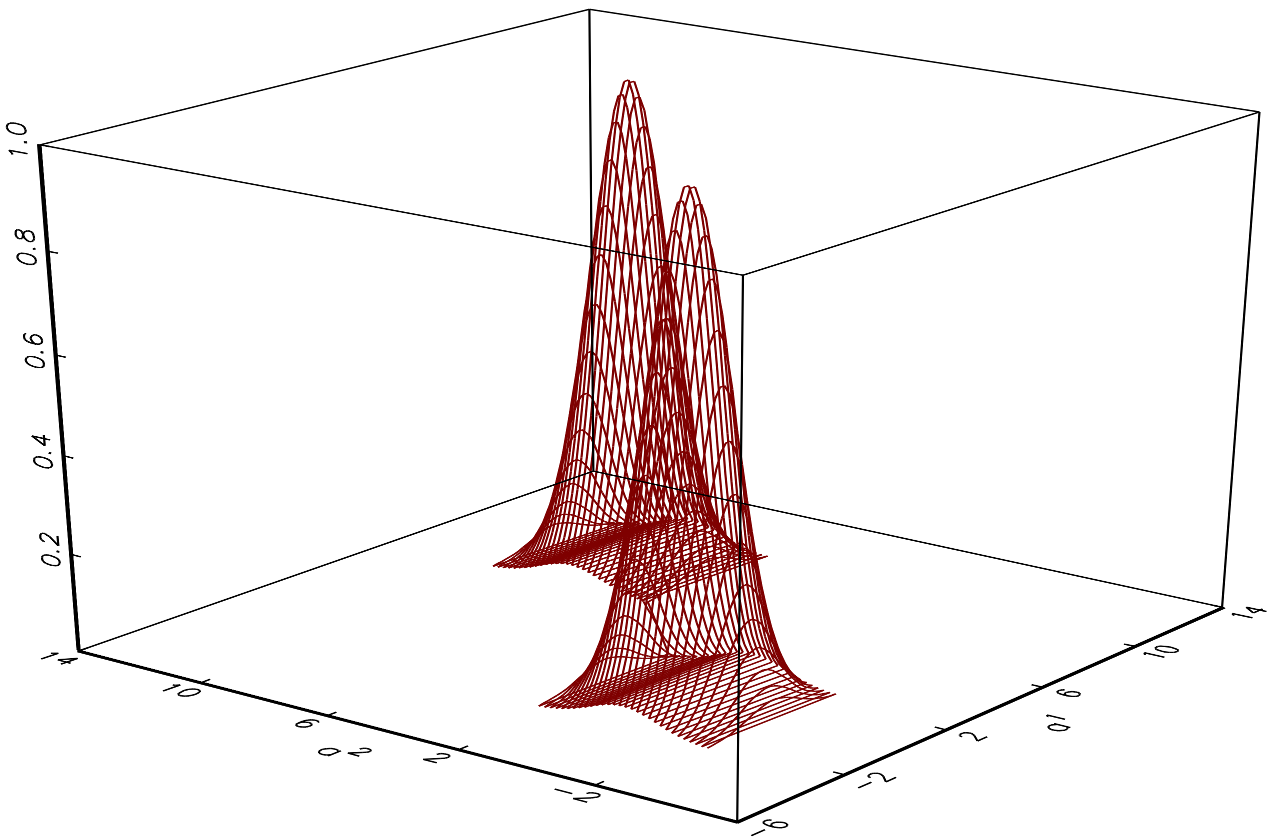
(i) draw a_{i+1}^1 from $f(a^1|a_i^2)$

(ii) draw a_{i+1}^2 from $f(a^2|a_{i+1}^1)$

Gibbs sampling is convenient when draws from $f(a^1|a_i^2)$ and $f(a^2|a_{i+1}^1)$ are easy.

Issues: When will this work (or when will it fail) … draws are too correlated (requiring too many Gibbs draws for accurate Monte Carlo sample averages).

Example: Bimodality



Checking quality of approximation: $\widehat{Eg(a)} = \frac{1}{N}\sum_{i=1}^{N} g(a_i)$

$$\sqrt{N}(\widehat{Eg(a)} - Eg(a)) \xrightarrow{d} N(0,V)$$

(1) 95% CI for $Eg(a)$ = $\widehat{Eg(a)} \pm 1.96\sqrt{\hat{V} / N}$

(2) Multiple runs from different starting values (should not differ significantly from one another)

(3) Compare $\widehat{Eg(a)}$ based on N_{first} draws and last N_{last} draws (say first 1/3 and last 1/3 … middle 1/3 left out). The estimates should not differ significantly from one another.

Returning to the Stochastic Volatility Model

$$x_t = \sigma_t e_t, \;\; \ln(\sigma_t) = \ln(\sigma_{t-1}) + \eta_t$$

or

$$y_t = 2\, s_t + \varepsilon_t, \quad s_t = s_{t-1} + \eta_t$$

$y_t = \ln(x_t^2)$, $\varepsilon_t = ln(\chi_1^2) \approx \sum_{i=1}^{n} q_{it} v_{it}$, where $v_{it} \sim \text{iidN}(\mu_i, \sigma_i^2)$, and $P(q_{it}=1)=p_i$.

Smoothing Problem: $E(\sigma_t \mid y_{1:T}) = E(g(s_t) \mid y_{1:T})$ with $g(s) = e^s$:

Let $a = \left(\{s_t\}_{t=1}^{T}, \{q_{it}\}_{i=1,t=1}^{10,T} \right) = (a_1, a_2)$

Jargon: "Data Augmentation" … add a_2 to problem even though it is not of direct interest.)

Model: $y_t = 2\, s_t + \sum_{i=1}^{n} q_{it} v_{it}$, $s_t = s_{t-1} + \eta_t$, $v_{it} \sim \text{iidN}(\mu_i, \sigma_i^2)$, and $P(q_{it}=1)=p_i$.

Gibbs Draws (throughout condition on $y_{1:T}$)

(i) $(a_1 \mid a_2)$: $\{s_t\}_{t=1}^{T} \mid \{q_{it}\}_{i=1,t=1}^{10,T}$

With $\{q_{it}\}_{i=1,t=1}^{10,T}$ known, this is a linear Gaussian model (with known time varying "system" matrices).

$\{s_t\}_{t=1}^{T} \mid (\{q_{it}\}_{i=1,t=1}^{10,T}, y_{1:T})$ is normal with mean and variance easily determined by formulae analogous to Kalman-filter (see Carter, C.K. and R. Kohn (1994)).

(ii) $(a_2 \mid a_1)$: $\left\{q_{it}\right\}_{i=1,t=1}^{10,T} \mid \left\{s_t\right\}_{t=1}^{T}$

With s_t known, $\varepsilon_t = y_t - 2s_t$ can be calculated. So

$$\text{Prob}(q_{it} = 1 \mid \left\{s_t\right\}_{t=1}^{T}, Y_T) = \frac{f_i(\varepsilon_t)p_i}{\sum_{j=1}^{10} f_j(\varepsilon_t)p_j}$$

where f_i is the $N(\mu_i, \sigma_i^2)$ density.

More Complicated Examples:

TVP-VAR-SV Model: $y_t = \sum_{i=1}^{p} \Phi_t y_{t-i} + e_t$ ($e_t \sim$ SV)

(VAR) Cogley and Sargent (2005), Uhlig (1997), (SVAR) Primiceri (2005), Del Negro and Primiceri (2014), (Markov Switching VAR) Sims and Zha (2006) ... many others

UC-SV: $Y_t = \tau_t + \varepsilon_t$, $\tau_t = \tau_{t-1} + \eta_t$ (ε_t and $\eta_t \sim$ SV)

Cogley and Sargent (2104), Garnier, Mertens, and Nelson (2013), Shephard (2013), Stock and Watson (2007) … others

$$Y_t = \tau_t + \varepsilon_t, \qquad \tau_t = \tau_{t-1} + \eta_t$$

$$\ln(\varepsilon_t^2) = 2\ln(\sigma_{\varepsilon,t}) + \sum_{i=1}^{10} q_{\varepsilon,i,t} v_{\varepsilon,i,t}, \quad \ln(\eta_t^2) = 2\ln(\sigma_{\eta,t}) + \sum_{i=1}^{10} q_{\eta,i,t} v_{\eta,i,t}$$

$$\ln(\sigma_{\varepsilon,t}) = \ln(\sigma_{\varepsilon,t-1}) + \upsilon_{\varepsilon,t}, \qquad \ln(\sigma_{\eta,t}) = \ln(\sigma_{\eta,t-1}) + \upsilon_{\eta,t},$$

$$a = \left(\left\{\tau_t\right\}, \left\{\sigma_{\varepsilon,t}, \sigma_{\eta,t}\right\}, \left\{q_{\varepsilon,i,t}, q_{\eta,i,t}\right\}\right) = (a_1, a_2, a_3)$$

Gibbs Draws:

(1) $\{\tau_t\}\{q_{\varepsilon,i,t}, q_{\eta,i,t}\} \mid \{\sigma_{\varepsilon,t}, \sigma_{\eta,t}\}, y_{1:T}$

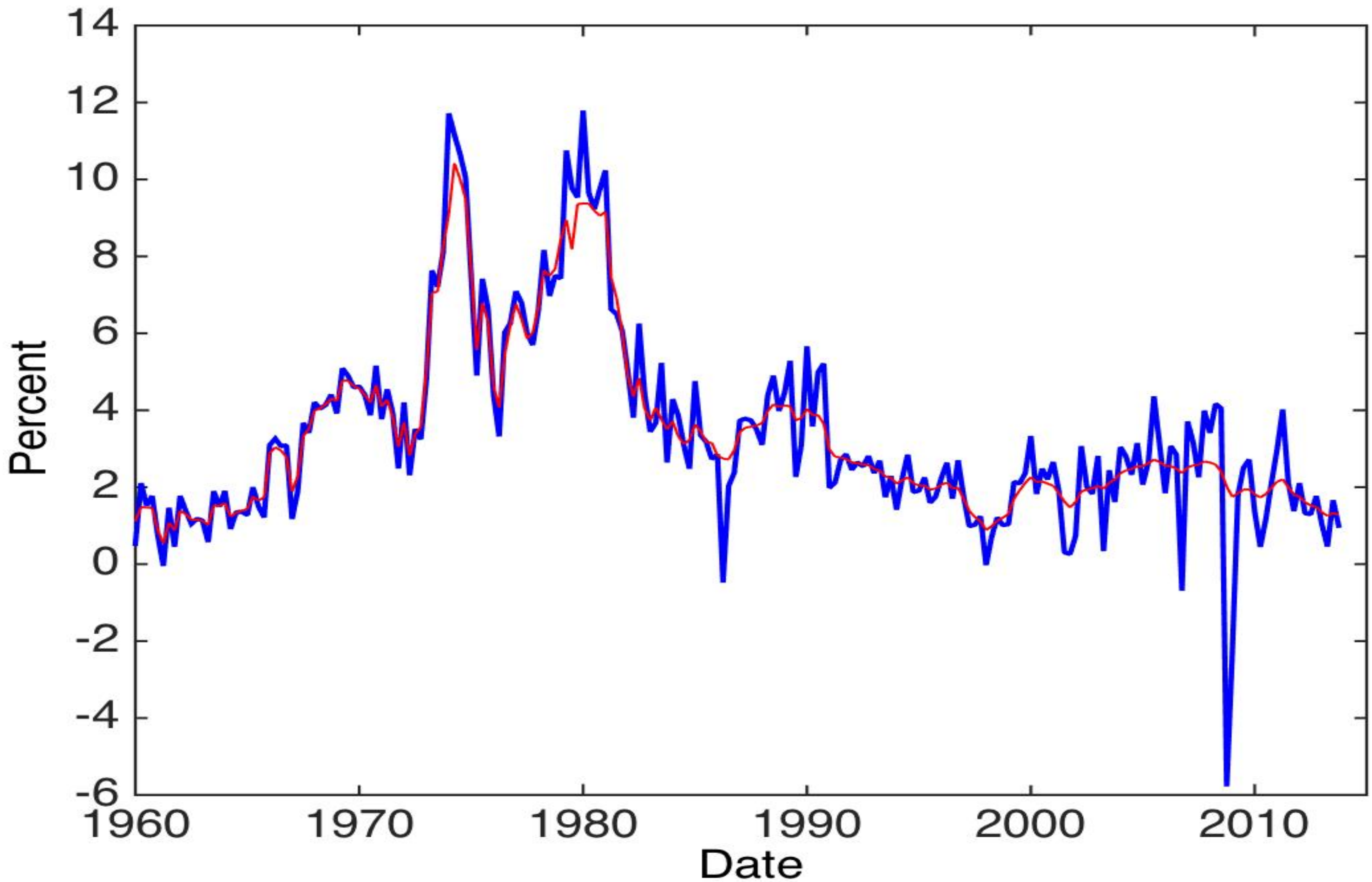
(a) $\{\tau_t\} \mid \{\sigma_{\varepsilon,t}, \sigma_{\eta,t}\}, y_{1:T}$: Kalman Filter (UC Model)

(b) $\{q_{\varepsilon,i,t}, q_{\eta,i,t}\} \mid \{\tau_t\}\{\sigma_{\varepsilon,t}, \sigma_{\eta,t}\}, y_{1:T}$: Multinomial Mixture

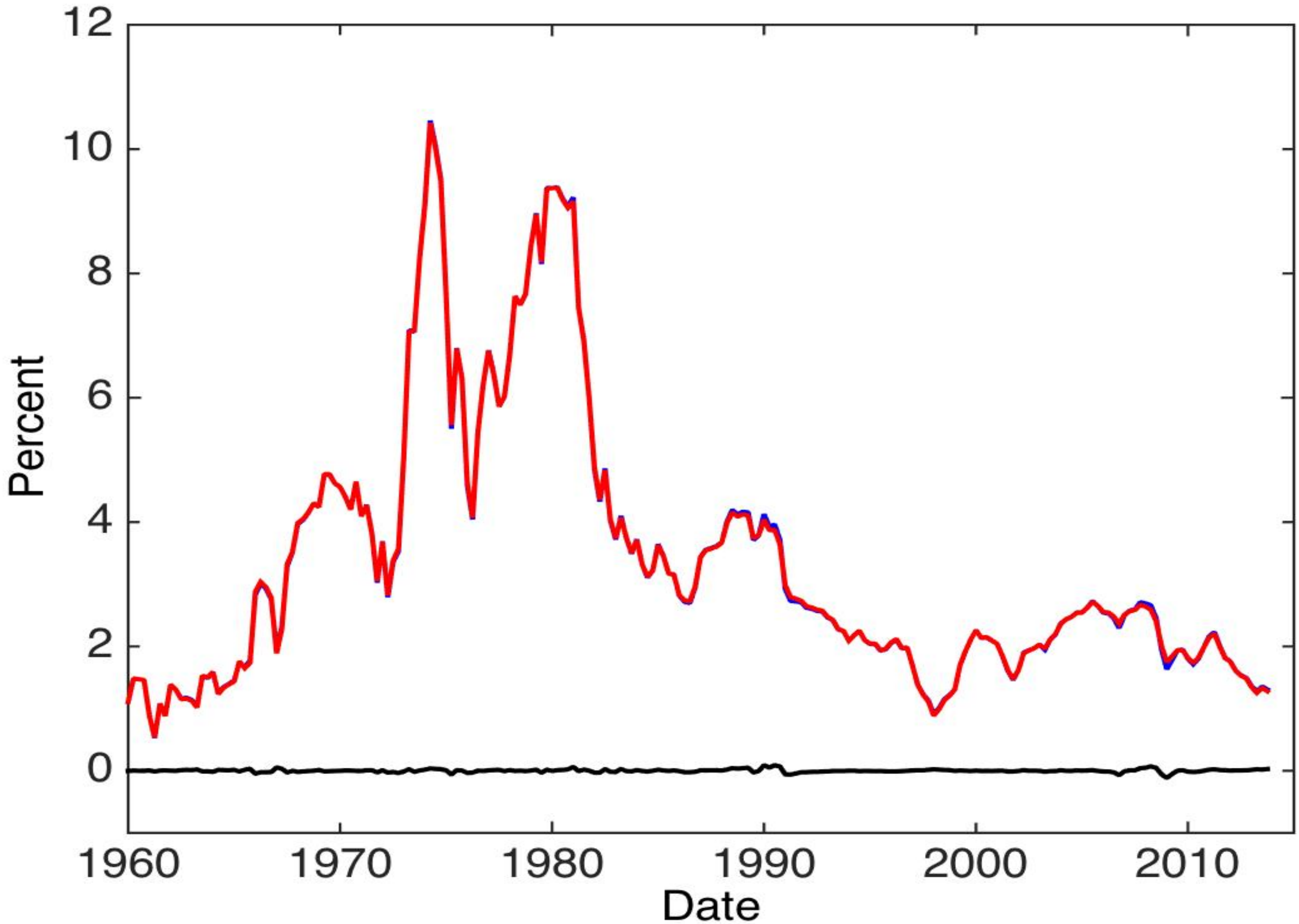
(2) $\{\sigma_{\varepsilon,t}, \sigma_{\eta,t}\} \mid \{\tau_t\}, \{q_{\varepsilon,i,t}, q_{\eta,i,t}\}, y_{1:T}$: "Kalman filter" – SV (as above)

(Placement of q-draws is important – Del Negro and Primiceri (2014))

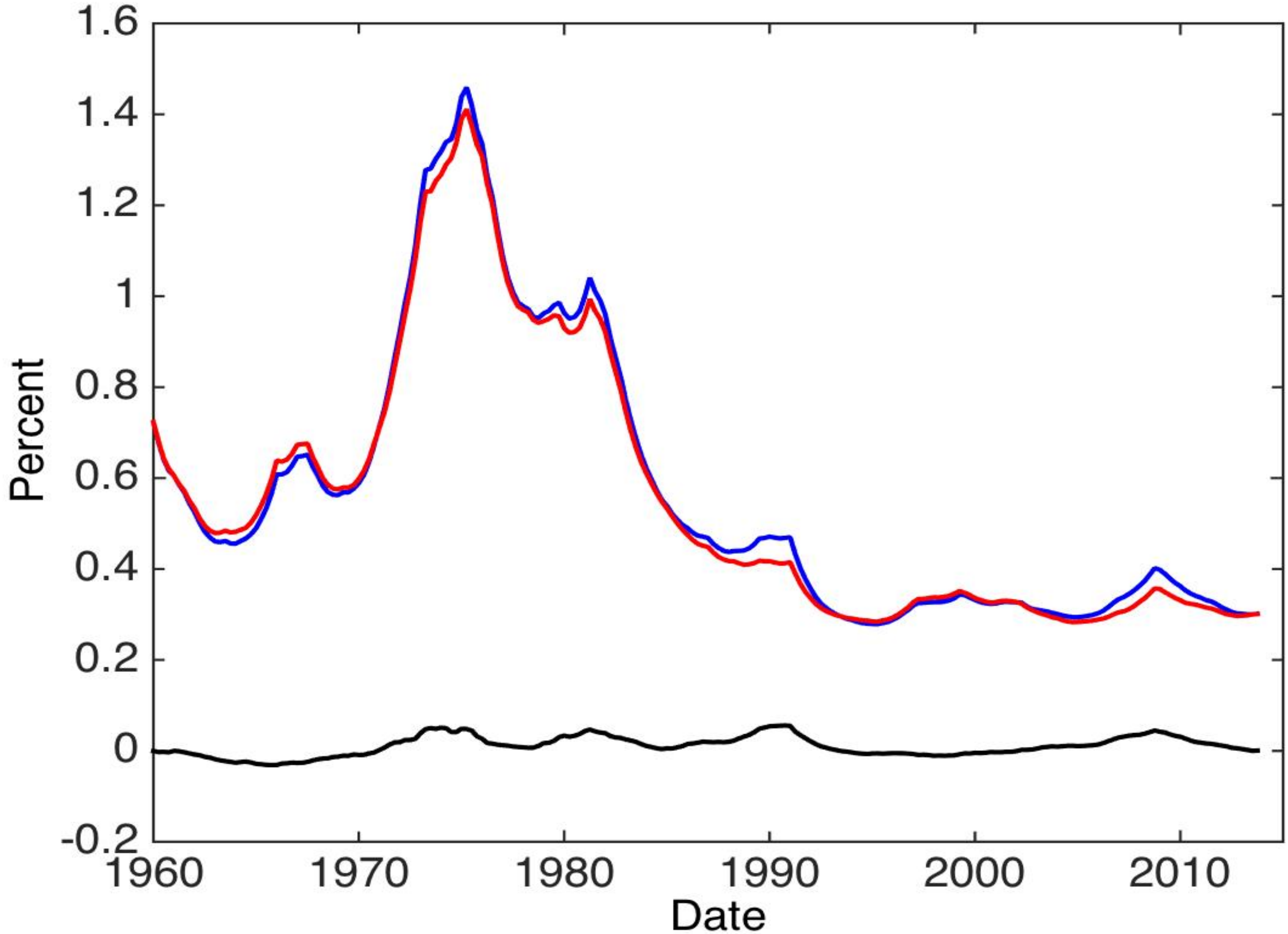
Inflation (PCE Deflator) and smoothed estimate of τ
(N =10,000, burnin = 2,000)



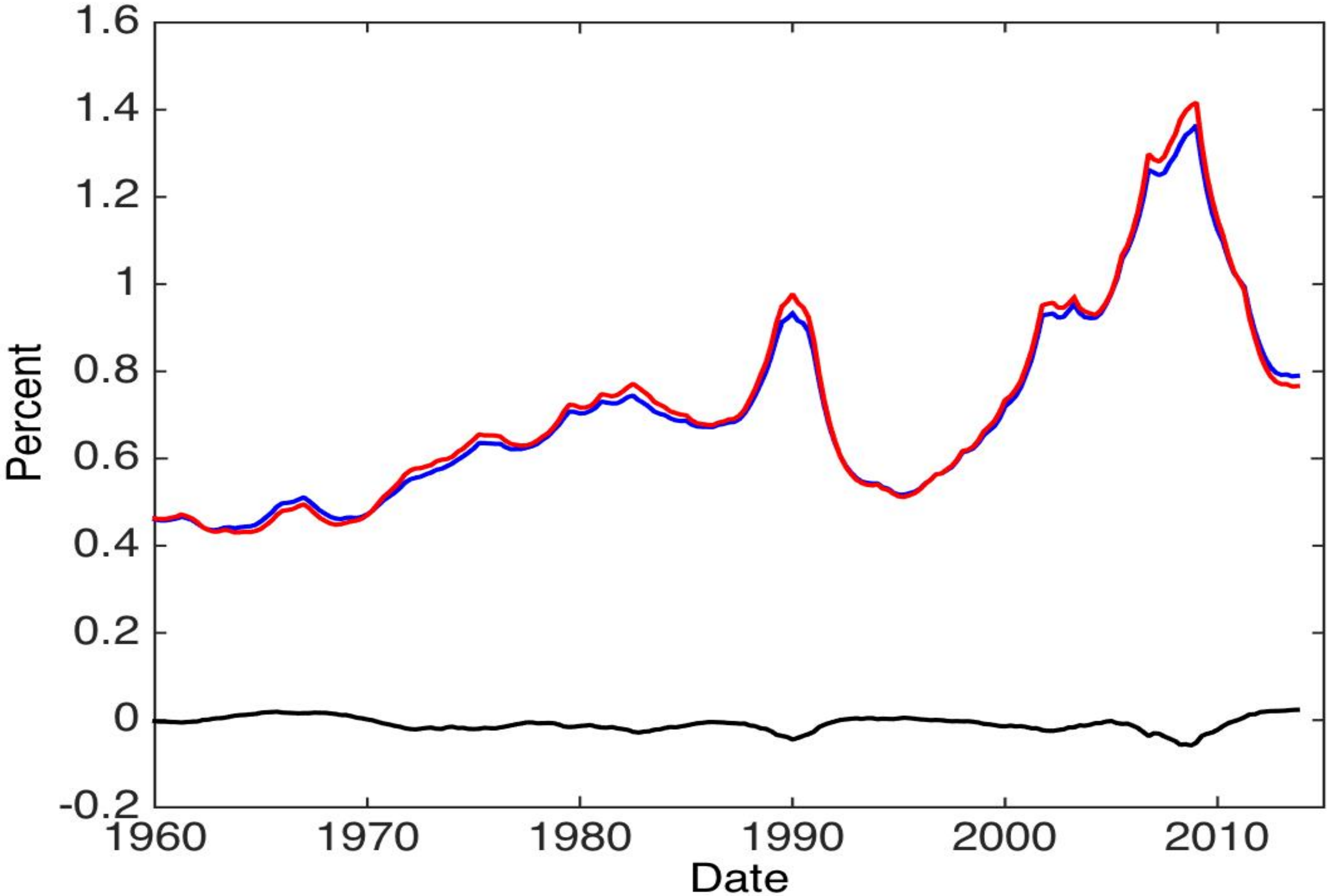
Estimates of τ from two independent sets of draws



Estimates of $\sigma_{\Delta\tau}$ from two independent sets of draws



Estimates of σ_ε from two independent sets of draws



$$\widehat{Eg(a)} = \frac{1}{N}\sum_{i=1}^{N} g(a_i); \quad \sqrt{N}(\widehat{Eg(a)} - Eg(a)) \xrightarrow{d} N(0,V)$$

Average values over all dates

	Serial Correlation in $g(a_i)$	$\sqrt{V/N}$	$\frac{\sqrt{V/n}}{\widehat{Eg(a)}}$
τ	0.15	0.0066	0.3%
$\sigma_{\Delta\tau}$	0.80	0.0073	1.6%
σ_{ε}	0.72	0.0058	0.9%

Outline

~~1. A motivating example~~

~~2. Models, objects of interest, and general formulae~~

~~3. Special Cases~~

~~4. MCMC (Gibbs)~~

5. Likelihood Evaluation

Computing the likelihood: Particle filtering

Model: $y_t = H(s_t, \varepsilon_t)$, $s_t = F(s_{t-1}, \eta_t)$, ε and η ~ iid

The "t'th component" of likelihood: $f(y_t \mid y_{1:t-1}) = \int f(y_t \mid s_t) f(s_t \mid y_{1:t-1})\, ds_t$

Often $f(y_t|s_t)$ is known, and the challenge is $f(s_t \mid y_{1:t-1})$. Particle filters use simulation methods to draw samples from $f(s_t \mid y_{1:t-1})s_t \mid Y_{t-1})$, say ($s_{1t}$, s_{2t}, … s_{nt}), where s_{it} is a called a "particle." The t'th component of the likelihood can then be approximated as $\widehat{f(y_t \mid y_{1:t-1})} = \frac{1}{n}\sum_{i=1}^{n} f(y_t \mid s_{it})$.

Methods for computing draws utilize the structure of the particular problem under study. Useful references include Kim, Shephard and Chib (1998), Chib, Nardari and Shephard (2002), Pitt and Shephard (1999), and Fernandez-Villaverde and Rubio-Ramirez (2007), Fernandez-Villaverde (2014).

Outline

1. A motivating example

~~2. Models, objects of interest, and general formulae~~

~~3. Special Cases~~

~~4. MCMC (Gibbs)~~

~~5. Likelihood Evaluation~~

Returning to the Cogley-Sargent motivating example:

Uncertainty: $\mathrm{Var}(p_{t+h} - p_t \,|\, \boldsymbol{Y_t})$

Instability: $\mathrm{E}(p_{t+h} - p_t \,|\, \boldsymbol{Y_t})^2 + \mathrm{Var}(p_{t+h} - p_t \,|\, \boldsymbol{Y_t})$

Uncertainty: $\sqrt{Var(p_{t+h} - p_t \mid p_{1:t})}$

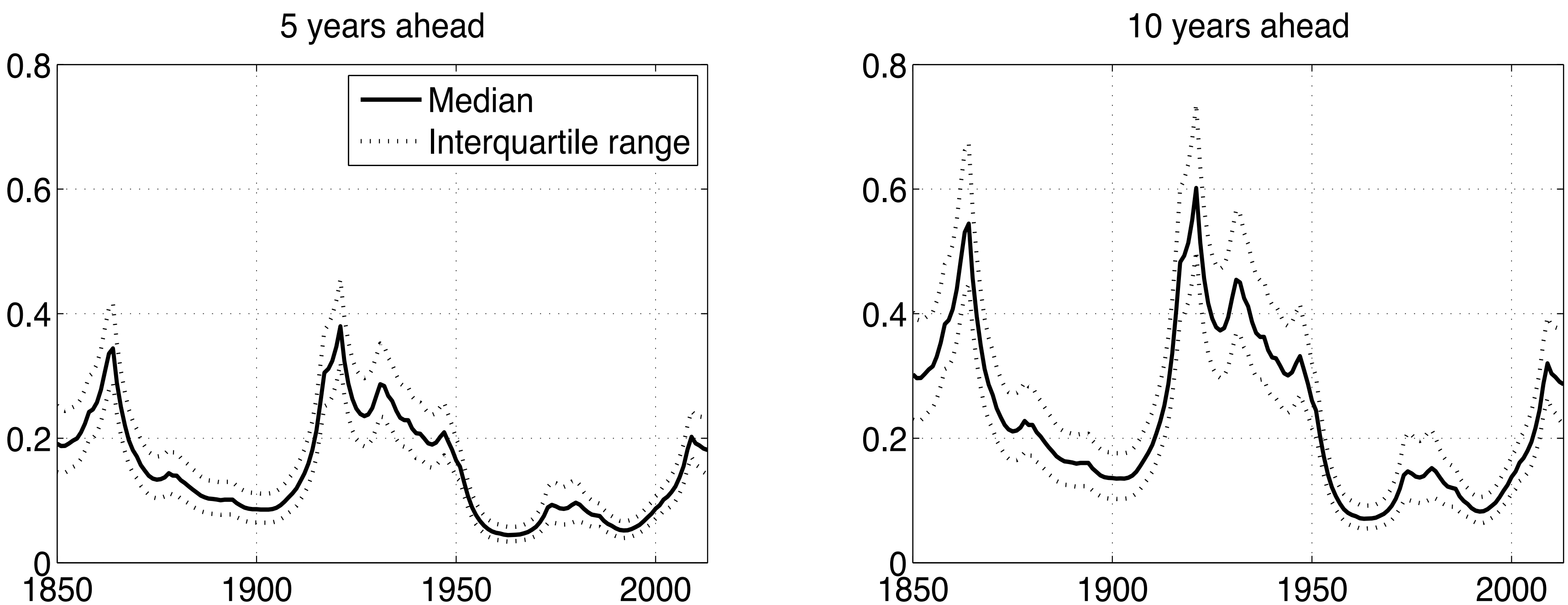


Figure 8: Posteriors for smoothed conditional volatilities 5 and 10 years ahead

Instability: $\sqrt{E(p_{t+h} - p_t \mid p_{1:t})^2 + Var(p_{t+h} - p_t \mid p_{1:t})}$

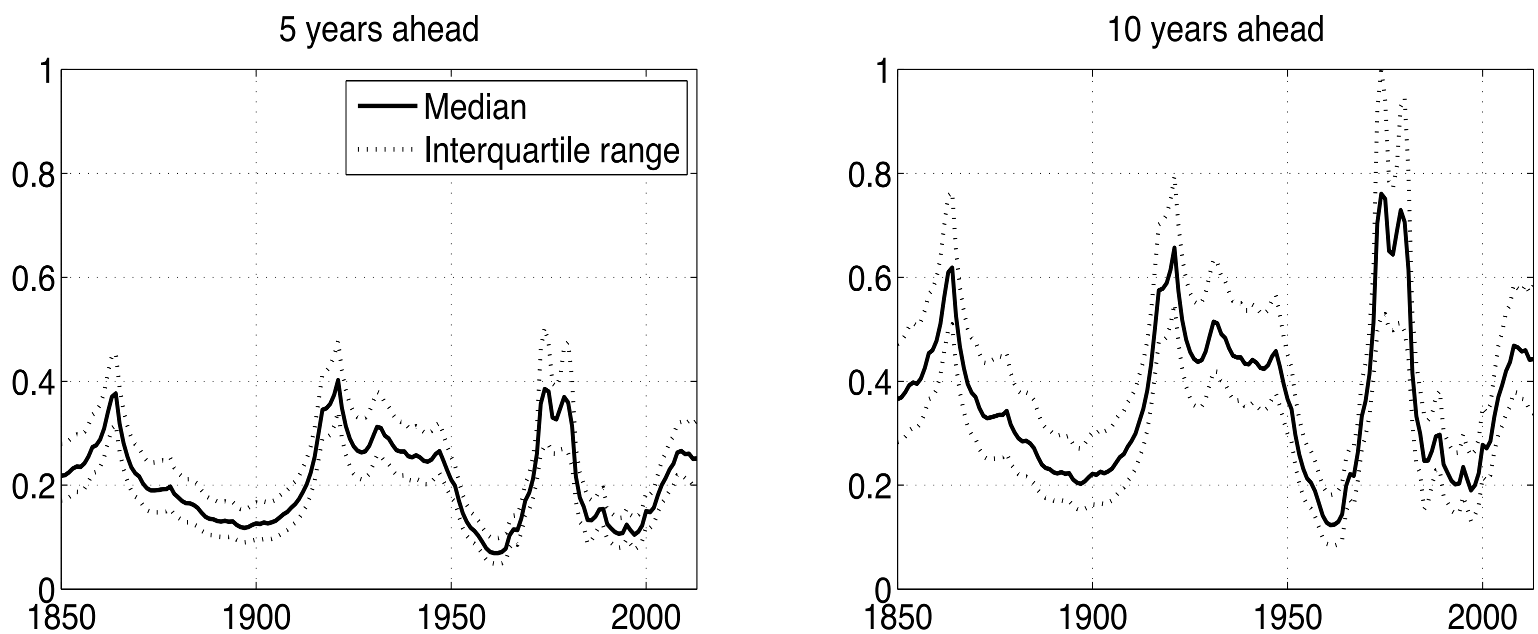


Figure 9: Posteriors for smoothed conditional root mean square statistics 5 and 10 years ahead

Outline

~~1. A motivating example~~

~~2. Models, objects of interest, and general formulae~~

~~3. Special Cases~~

~~4. MCMC (Gibbs)~~

~~5. Likelihood Evaluation~~

Course Topics

1. ~~Time series refresher and inference tools (MW)~~
2. ~~The Kalman filter, nonlinear filtering, and Markov chain monte carlo (MW)~~
3. Prediction with large datasets (MW)
4. Heteroskedasticity and autocorrelation consistent (HAC) standard errors (JS)
5. Many instruments/weak identification in IV and GMM (JS)
6. Structural VAR modeling (JS)

References for Lecture 2

Anderson, B.D.O., and J.B. Moore (2005), *Optimal Filtering*, Dover Publishing.

Carter, C.K., and R. Kohn (1994), "On Gibbs Sampling for State Space Models," *Biometrika*, 81, 541-553.

Cassella, G., and E.I. George (1992), "Explaining the Gibbs Sampler," *American Statistician*, 26, 167-174.

Chib, S. (2001), "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Economics Vol. 5*, ed. by J.J. Heckman and E. Leamer, Amsterdam: Elsevier.

Chib, C., N. Nardari, and N. Shephard (2002), "Markov Chain Monte Carlo Methods for Stochastic Volatility Models," *Journal of Econometrics*, 108, 281-316.

Cogley, T., and T.J. Sargent (2005), "Drifts and Volatilities: Monetary Policy and Outcomes in the Post WWII U.S.," *Review of Economic Dynamics*, 8, 262-302.

Cogley, T., and T.J. Sargent (2014), "Measuring Price-Level Uncertainty and Instability in the U.S., 1850-2012," *Review of Economics and Statistics*, forthcoming.

Del Negro, M. and G.E. Primiceri (2014), "Time Varying Structural Vector Autoregressions and Monetary Policy: A Corrigendum," manuscript Northwestern University.

Fernandez-Villaverde, J., (2014), *Lecture notes: Methods in Macroeconomic Dynamics*, available at http://economics.sas.upenn.edu/~jesusfv/teaching.html

Fernandez-Villaverde, J., and J.F. Rubio-Ramirez (2007), "Estimating Macroeconomic Models: A Likelihood Approach," *Review of Economic Studies*, 74, 1059-1087.

Garnier, C., E. Mertens, and E. Nelson (2013), "Trend Inflation in Advanced Economies," FEDS working paper 2013-74.

Geweke, John (2005), *Contemporary Bayesian Econometrics and Statistics*, New York: Wiley

Hamilton, J.D. (1989), "A New Approach to the Economic Analysis of Changes in Regime," *Econometrica*, March.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton: Princeton University Press

Harvey, A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge, UK: Cambridge University Press.

Harvey, A.C., E. Ruiz, and N. Shephard (1994), “Multivariate Stochastic Variance Models,” *Review of Economic Studies*, 61, 247-264.

Kim, S., N. Shephard, and S. Chib (1998), “Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models,” *Review of Economic Studies*, 65, 361-393.

Kitagawa, G. (1987), “Non-Gaussian State-Space Modeling of Nonstationary Time Series,” *Journal of the American Statistical Association*, 82(4): 1032 -1041.

Koop, Gary (2003), *Bayesian Econometrics*, New York: Wiley

Nelson, C.R and G.W. Schwert (1977), “Short-Term Interest Rates as Predictors of Inflation: On Testing the Hypothesis that the Real Rate of Interest in Constant,” *American Economic Review*, 67, 478-86,

Omori, G., S. Chib, N. Shephard, and J. Nakajima (2007), “Stochastic Volatility with Leverage: Fast and Efficient Likelihood Inference,” *Journal of Econometrics*, 140, 425-449.

Pitt, M.K., and N. Shephard (1999), “Filtering via Simulation: Auxiliary Particle Filters,” *Journal of the American Statistical Association*, 94(446), 590-599.

Primiceri, G. (2005), “Time Varying Structural VAR Autoregressions and Monetary Policy,” *Review of Economic Studies*, 72, 821-852.

Shephard, N. (1994), "Partial non-Gaussian state space", *Biometrika*, 81, 115-131.

Shephard, N. (2013), “Martingale Unobserved Component Models,” manuscript, Harvard University.

Sims, C.A., and T. Zha (2006), “Were There Regime Switches in US Monetary Policy?" *American Economic Review*, 96(1): 54-81.

Stock, James H., and Mark W. Watson (2007), “Has Inflation Become Harder to Forecast?” *Journal of Money, Credit and Banking,* 2007, 39(1): 3-34.

Uhlig, H. (1997), “Bayesian Vector Autoregressions with Stochastic Volatility,” *Econometrica*, 65(1): 59-73.

AEA Continuing Education Course

Time Series Econometrics

Lecture 3: Prediction with large datasets

Mark W. Watson
January 6, 2015
10:30AM – 12:30PM

Outline

1. Motivation and Setup
2. Dynamic Factor Models
3. Shrinkage
4. Sparse Models

1. Motivation and setup

"Linear" prediction problem: $y_{t+1} = x_t{'}\beta + \varepsilon_{t+1} = \sum_{i=1}^{n} x_{it}\beta_i + \varepsilon_{t+1}$

Sample size is T.

Forecast: $\hat{y}_{T+1} = \sum_{i=1}^{n} x_{iT}\hat{\beta}_i$

Forecast error: $y_{T+1} - \hat{y}_{T+1} = \sum_{i=1}^{n} x_{iT}(\beta_i - \hat{\beta}_i) + \varepsilon_{T+1}$

MSFE: $E\left(\sum_{i=1}^{n} x_{iT}(\beta_i - \hat{\beta}_i)\right)^2 + \sigma^2$

Suppose: (1) T is large and n is small
(2) T is large and n is large

"Linear" prediction problem: $y_{t+1} = x_t'\beta + \varepsilon_{t+1} = \sum_{i=1}^{n} x_{it}\beta_i + \varepsilon_{t+1}$

Suppose: (2) T is large and n is large

Approaches:

(1) Use "small-n" estimators (e.g. OLS)

(2) Impose some structure

(a) Common "Factors" (Dynamic Factor Model)

(b) β_i's are "small" (Shrinkage)

(c) There are only a few non-zero β_i's (Sparsity)

Outline

~~1. Motivation and Setup~~

2. Dynamic Factor Models

3. Shrinkage

4. Sparse Models

Dynamic Factor Models (DFMs)

Forecasting setup: $y_{t+1} = \alpha(\mathrm{L})f_t + \varepsilon_{t+1}$

$$x_{it} = \lambda_i(\mathrm{L})f_t + e_{it}$$

$$\Psi(\mathrm{L})f_t = \eta_t$$

"f_t" are latent factors.

x is useful for forecasting for y because x provides information about f:

$$E(y_t \mid x_t) = E\left(\alpha(\mathrm{L})f_t \mid x_t\right)$$

DFM: Use x to estimate f. Use this to forecast y. Estimated factors are also useful for other purposes.

DFMs: A brief survey

$$x_{it} = \lambda_i(\mathrm{L})f_t + e_{it}$$
$$\Psi(\mathrm{L})f_t = \eta_t$$

(1) Large T, small n DFMs: (Geweke (1977), Sargent and Sims (1977), Engle and Watson (1981), Stock and Watson (1989)). Parametric model:

$$\rho_i(\mathrm{L})e_{it} = a_{it}$$

$$\begin{bmatrix} a_t \\ \eta_t \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D_a & 0 \\ 0 & \Sigma_{\eta\eta} \end{bmatrix} \right), \quad D_a \text{ diagonal.}$$

Estimation via ML (Kalman filtering, etc.).

Conceptually and computationally difficult with large n. (Quah and Sargent (1989).

(2) Large-n "Approximate" factor models: Chamberlain-Rothschild (1983), Connor and Korajczyk (1986), Forni, Hallin, Lippi, Reichlin (2000, 2004), Stock and Watson (2002), Bai-Ng (2002, 2006), many others …

An example following Forni and Reichlin (1998): Suppose f_t is scalar and $\lambda_i(L) = \lambda_i$ ("no lags in the factor loadings"):

$$X_{it} = \lambda_i f_t + e_{it}$$

Then

$$\frac{1}{n}\sum_{i=1}^{n} X_{it} = \frac{1}{n}\sum_{i=1}^{n}(\lambda_i f_t + e_{it}) = \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right) f_t + \frac{1}{n}\sum_{i=1}^{n} e_{it}$$

If the errors e_{it} have *limited dependence* across series, then as n gets large,

$$\frac{1}{n}\sum_{i=1}^{n} X_{it} \xrightarrow{p} \bar{\lambda} f_t$$

In this special case, a very easy nonparametric estimator (the cross-sectional average) is able to recover f_t – as long as n is large

A convenient representation for the DFM: $X_t = \lambda(\mathrm{L})f_t + e_t$

$$\Psi(\mathrm{L})f_t = \eta_t,$$

Suppose that $\lambda(\mathrm{L})$ has at most p_f lags. Then the DFM can be written,

$$\begin{pmatrix} X_{1t} \\ \vdots \\ X_{nt} \end{pmatrix} = \begin{pmatrix} \lambda_{10} & \cdots & \lambda_{1p_f} \\ \vdots & \ddots & \vdots \\ \lambda_{n0} & \cdots & \lambda_{np_f} \end{pmatrix} \begin{pmatrix} f_t \\ \vdots \\ f_{t-p_f} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ \vdots \\ e_{nt} \end{pmatrix}$$

or $$\underset{n\times 1}{X_t} = \underset{n\times r}{\Lambda} \underset{r\times 1}{F_t} + \underset{n\times 1}{e_t}$$

F_t are sometimes called "*static factors*". But, they aren't static: the VAR for f_t implies that there is a VAR for F_t

$$\Phi(\mathrm{L})F_t = G\eta_t$$

where G is a matrix of 1's and zeros and Φ consists of 1's, 0's, and Ψ's.

Principal Components (estimating the factors by least squares)

$$X_t = \Lambda F_t + e_t$$

$$\Phi(\mathrm{L})F_t = G\eta_t$$

Consider estimating Λ and $\{F_t\}$ by least squares:

$$\min_{F_1,\ldots,F_T,\Lambda} T^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t) \qquad (1)$$

subject to $\Lambda'\Lambda = I_r$ (identification). Given Λ, the (infeasible) OLS estimator of F_t is:

$$\hat{F}_t(\Lambda) = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_t$$

Now substitute $\hat{F}_t(\Lambda)$ into (1) to concentrate out $\{F_t\}$:

$$\min_\Lambda T^{-1}\sum_{t=1}^{T} X_t'[I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda]X_t$$

$$\min_\Lambda T^{-1}\sum_{t=1}^{T} X_t'[I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda]X_t$$

$$\rightarrow \max_\Lambda T^{-1}\sum_{t=1}^{T} X_t'\Lambda(\Lambda'\Lambda)^{-1}\Lambda X_t$$

$$\rightarrow \max_\Lambda \mathrm{tr}\{(\Lambda'\Lambda)^{-1/2\prime}\Lambda'\left(T^{-1}\sum_{t=1}^{T} X_t X_t'\right)\Lambda(\Lambda'\Lambda)^{-1/2}\}$$

$$\rightarrow \max_\Lambda \mathrm{tr}\{\Lambda'\hat{\Sigma}_{XX}\Lambda\} \text{ s.t. } \Lambda'\Lambda = I_r, \text{ where } \hat{\Sigma}_{XX} = T^{-1}\sum_{t=1}^{T} X_t X_t'$$

$\rightarrow$ $\hat{\Lambda}$ = first r eigenvectors of $\hat{\Sigma}_{XX}$ (corresponding to largest eigenvalues)

Remember $\hat{F}_t(\Lambda) = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_t$, so

$$\hat{F}_t(\hat{\Lambda}) = \left(\hat{\Lambda}'\hat{\Lambda}\right)^{-1}\hat{\Lambda}' X_t = \hat{\Lambda}' X_t \quad (\text{because } \hat{\Lambda}'\hat{\Lambda} = I_r)$$

= first r principal components of X_t.

Distribution theory for PC as factor estimator

Selected results for the approximate DFM: $X_t = \Lambda F_t + e_t$

Typical conditions (Stock-Watson (2002), Bai-Ng (2002, 2006),…):

(a) $\frac{1}{T}\sum_{i=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F$ (stationary factors)

(b) $\Lambda'\Lambda/n \to$ (or $\xrightarrow{p}$) Σ_Λ Full rank factor loadings

(c) e_{it} are weakly dependent over time and across series (approximate DFM)

(d) F, e are uncorrelated at all leads and lags

(e) n, $T \to \infty$ plus Bai-Ng (2006) rate condition: $n^2/T \to \infty$

Selected results for the approximate DFM, ctd.

Stock and Watson (2002), Bai and Ng (2006):

- consistency of $\hat{F}_t$ for F_t (up to a $r\times r$ rotation)
- $\hat{F}_t$ converges at a sufficiently fast rate that $\hat{F}_t$ can be used as a regressor (e.g. in forecasting equations) without adjusting standard errors – you can treat $\hat{F}_t$ as if it actually *is* F_t (up to a $r\times r$ rotation)
- The PCA estimator of the common component is asymptotically normal at rate $\min(n^{1/2}, T^{1/2})$
- Bai-Ng (2006) give a method for constructing confidence bands for predicted values (these are for predicted value [for example estimates of common components] – *not* forecast confidence bands)

Estimating the number of factors in F

Most widely used method: Bai-Ng (2002) propose an estimator of r based on an information criterion; their main result is $\hat{r} \xrightarrow{p} r_0$ for the approximate DFM

Digression on information criteria (IC) for lag length selection in an AR

Consider the AR(p): $y_t = a_1 y_{t-1} + \ldots + a_p y_{t-p} + \varepsilon_t$

- Why not just maximize the R^2?
- IC trades off estimator bias (too few lags) vs. estimator variance (too many lags), from the perspective of fit of the regression:

Bayes Information Criterion: $$\mathrm{BIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{\ln T}{T}$$

Akaike Information Criterion: $$\mathrm{AIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{2}{T}$$

The Bai-Ng (2002) information criteria have the same form:

$$\mathrm{IC}(r) = \ln\left(\frac{SSR(r)}{T}\right) + \text{penalty}(N, T, r)$$

Bai-Ng (2002) propose several IC's with different penalty factors that all produce consistent estimators of r. Here is the one that seems to work best in MCs (and is the most widely used in empirical work):

$$IC_{p2}(r) = \ln(\mathrm{V}(r,\hat{F}^r)) + r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right]$$

where
$$V(r,\hat{F}^r) = \min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \lambda_i^{r\prime}\hat{F}_t^r\right)^2$$

$$= \min_{F_1,\ldots,F_T,\Lambda}(NT)^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t)$$

$\hat{F}_t^r$ are the PC estimates of r

Bai-Ng (2002) IC_{p2}: $$IC_{p2}(r) = \ln(V(r,\hat{F}^r)) + r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right]$$

where $$V(r,\hat{F}^r) = \min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \lambda_i^{r\prime}\hat{F}_t^r\right)^2$$

Comments:

- $\ln(V(r,\hat{F}^r))$ is a measure of (trace) fit – generalizes $\ln(SSR/T)$ in AIC/BIC
- If $N = T$, then $r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right] = r\left(\frac{2T}{T^2}\right)\ln T = 2r\frac{\ln T}{T}$

  which is 2 × the usual BIC penalty factor
- Both N and T are in the penalty factor: you need N, $T \to \infty$.
- Bai-Ng's (2002) main result: $\hat{r} \xrightarrow{p} r_0$

Comments on Bai-Ng factor selection:

- Monte Carlo studies show B-N works well when n, T are large, and DFM model is correct.
- But in practice:
  - Different IC can yield substantially different answers
  - Adding series often increases the number of estimated factors (adding sectors should increase number of factors; adding series within sectors should not)
- Judgment is required
- There are several estimators that have been proposed and this is an on-going area of research.

Empirical Applications using DFMs – many, here are a few:

(1) Forecasting ... more on this below

(2) SVARs: Bernanke, Boivin, and Eliasz's (2005) is most famous example.

(3) Factors as instruments: Bai and Ng (2011)

(4) DSGE Modeling: Sargent (1989), Boivin-Giannoni (2006b).

(5) Real-time tracking: Stock and Watson (1989), Giannone, Reichling and Small (2008), Council of Economic Advisors (2012)

(6) Data Description: example follows …

Stock and Watson (2012) "Disentangling the Channels of the 2007-09 Recession"

$$X_t = \Lambda F_t + e_t$$

$$\Phi(\mathrm{L})F_t = G\eta_t$$

Were there new factors in the 2007-09 recession?

Were there instabilities in Λ?

Were there instabilities in $\Phi(\mathrm{L})$?

Were there unusually extreme values of η and/or e?

1. Structural breaks post 2007Q4

Empirical analysis

1. Estimate DFM parameters using data through 2007Q3
   a. Compute factors using "old" factor loadings:
   b. $\hat{F}_t = (\hat{\Lambda}'\hat{\Lambda})^{-1}\hat{\Lambda}'X_t$, where $\hat{\Lambda}$ are pre-07Q3 factor loadings
   c. How well do pre-07Q3 factors & factor loadings do in explaining post-07Q4 macro variables?

2. Formal stability tests:
   a. Stability of Λ
   b. Test for new factor (excess covariance among idiosyncratic disturbances)

1.1. Fit of pre-07Q3 parameters and factors, post-07Q4

Figures:

Plot of 4-Q growth $(100\ln(X_t/X_{t-4}))$ *or 4-Q change*:

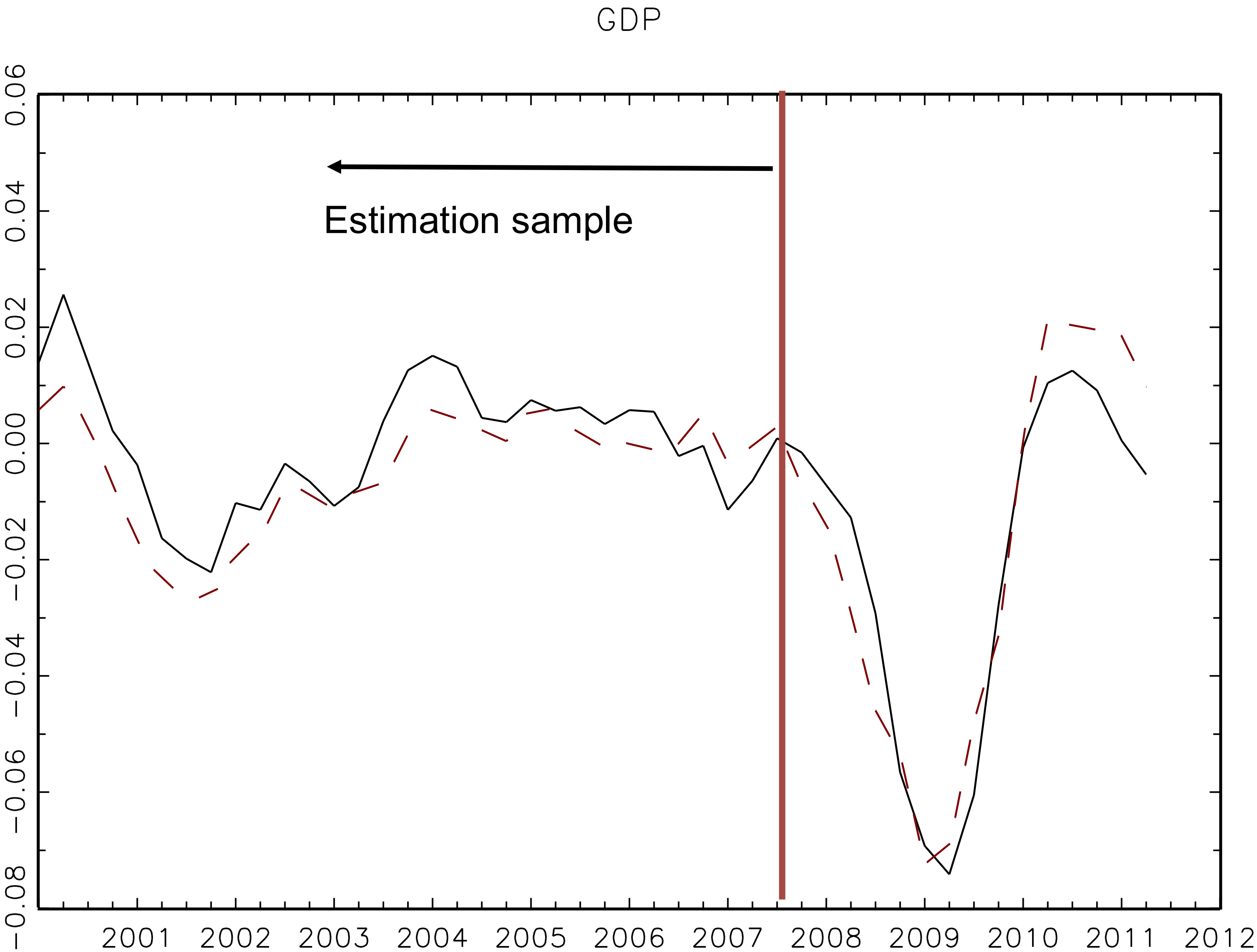
solid = actual

dashed = common component (pre-07Q3 model)

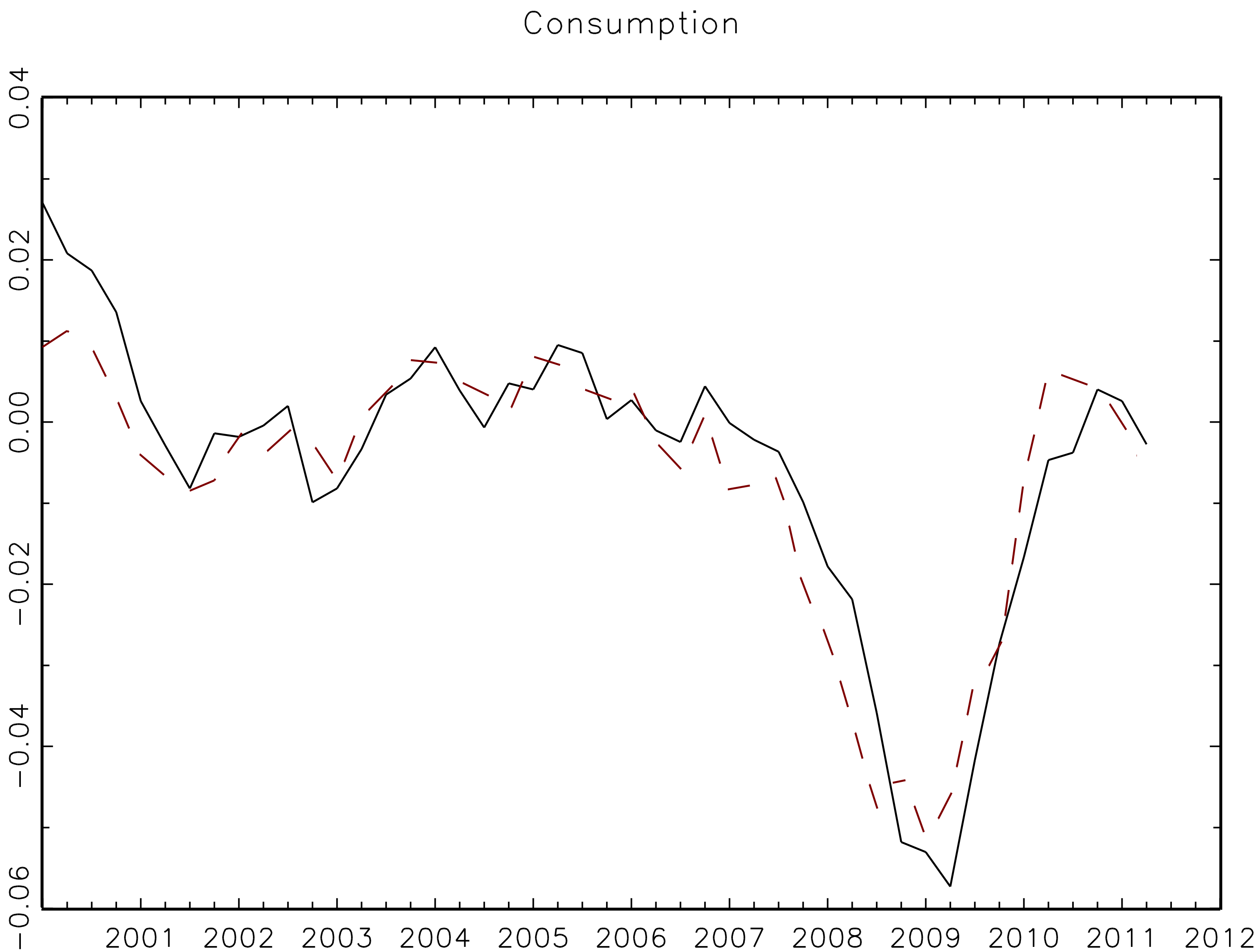
Average R^2 2007Q4 R^2

Average R^2 = 1-quarter R^2 of "ΛF_t", NBER peak to peak + 14 quarters, averaged over previous 7 recessions, 1960Q1,…, 2001Q1

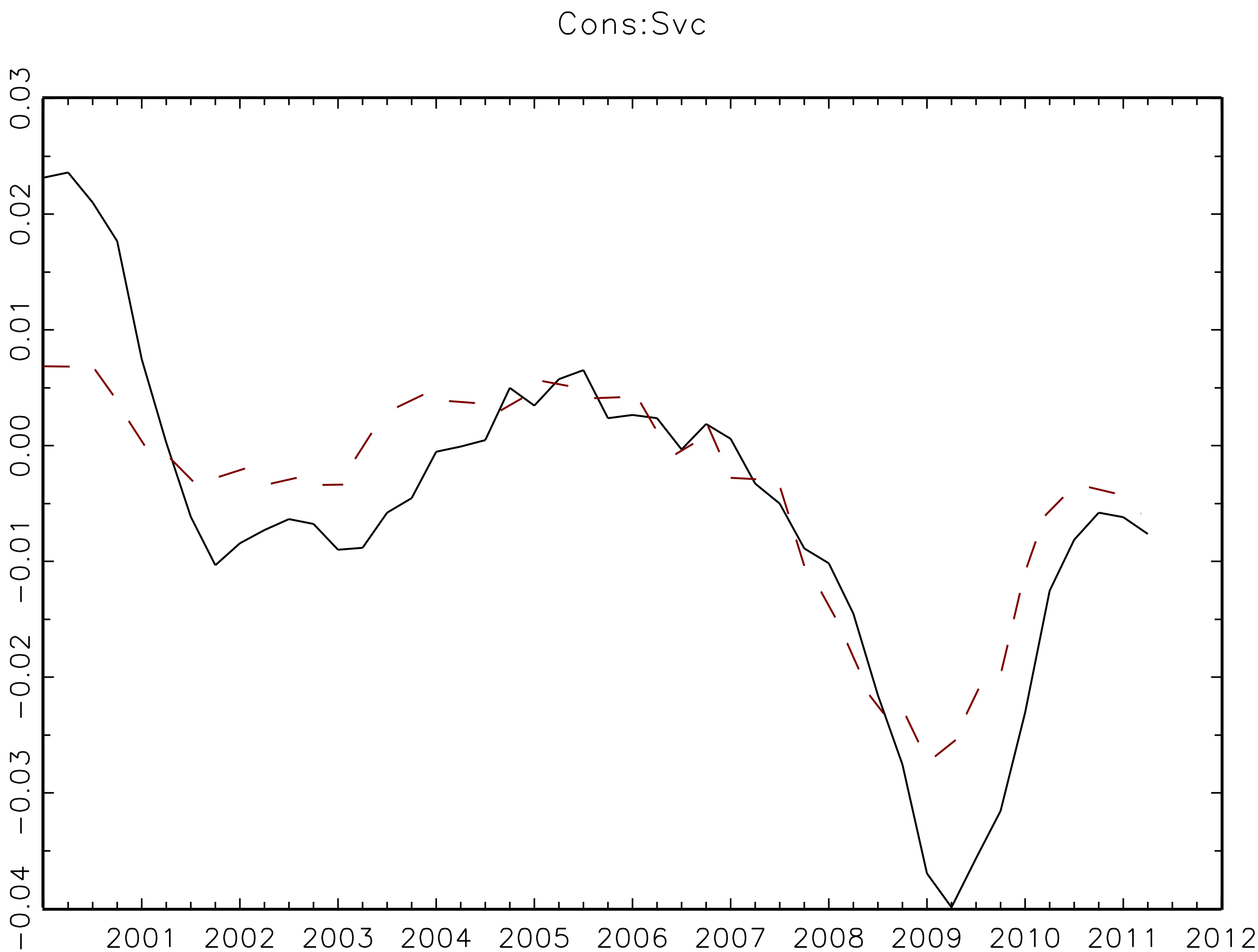
2007Q4 R^2 = value for 2007Q4 – 2011Q2.



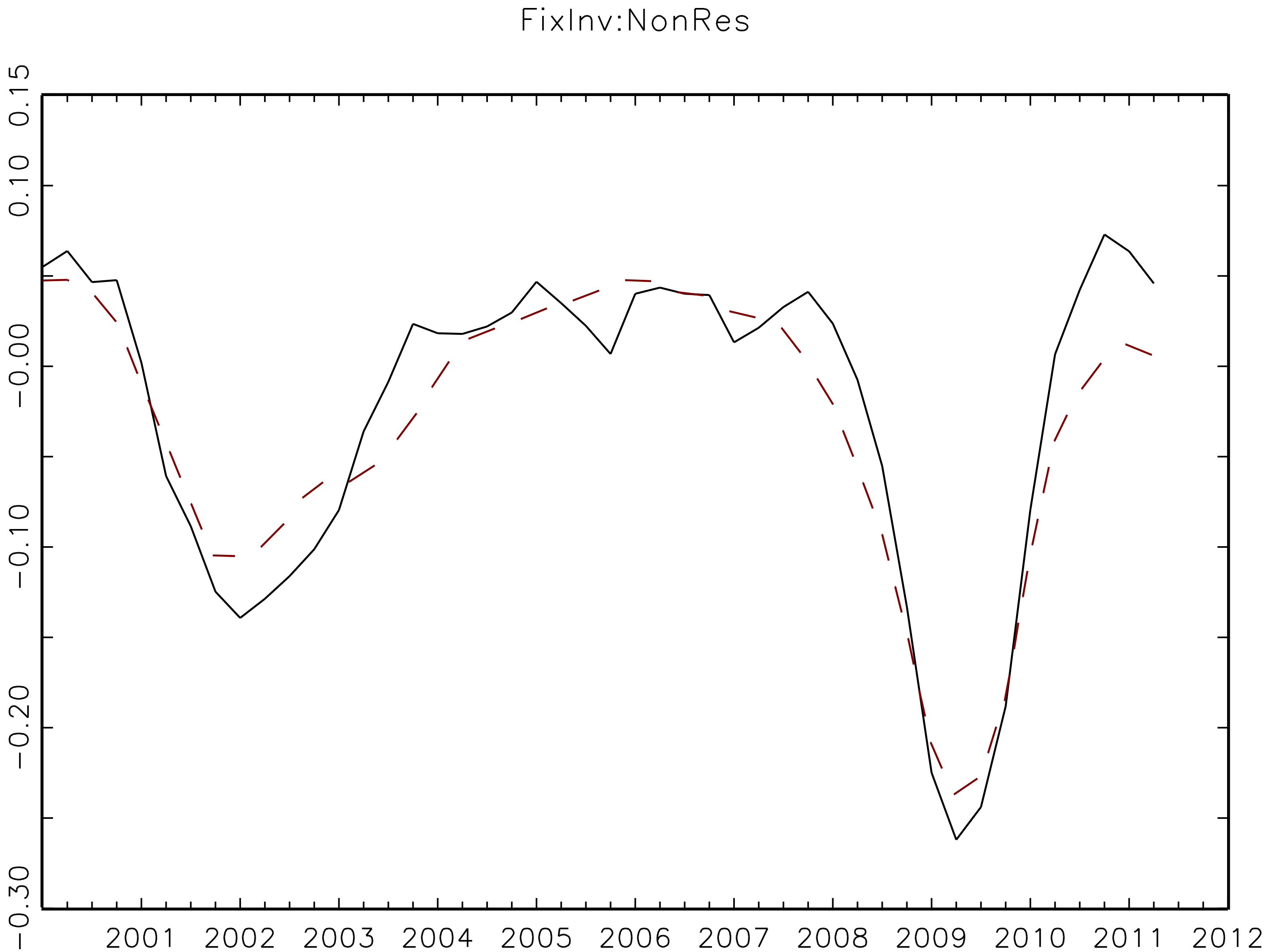
average R^2: 0.78 2007Q4 R^2: 0.64



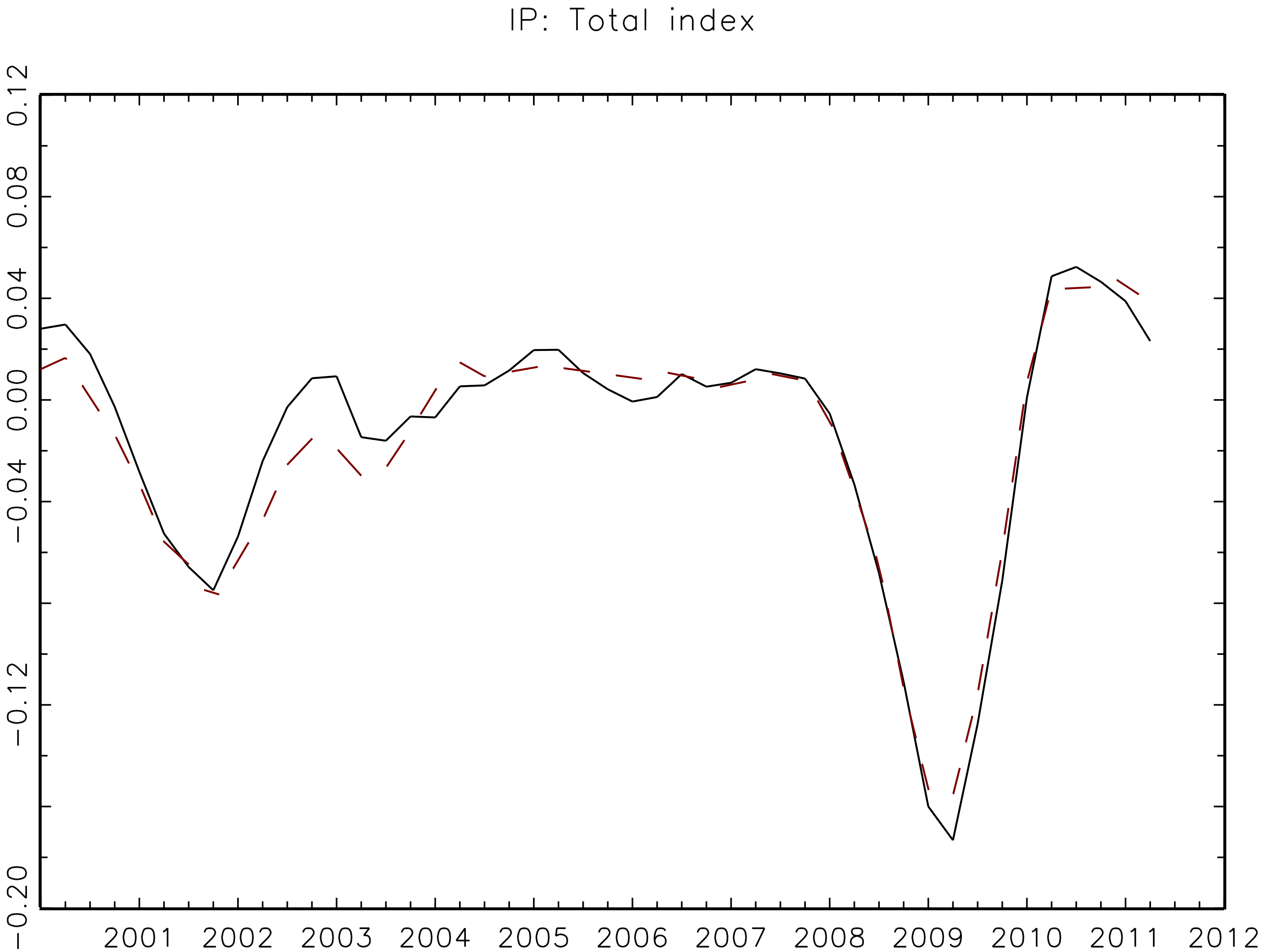
average R^2: 0.62 2007Q4 R^2: 0.56



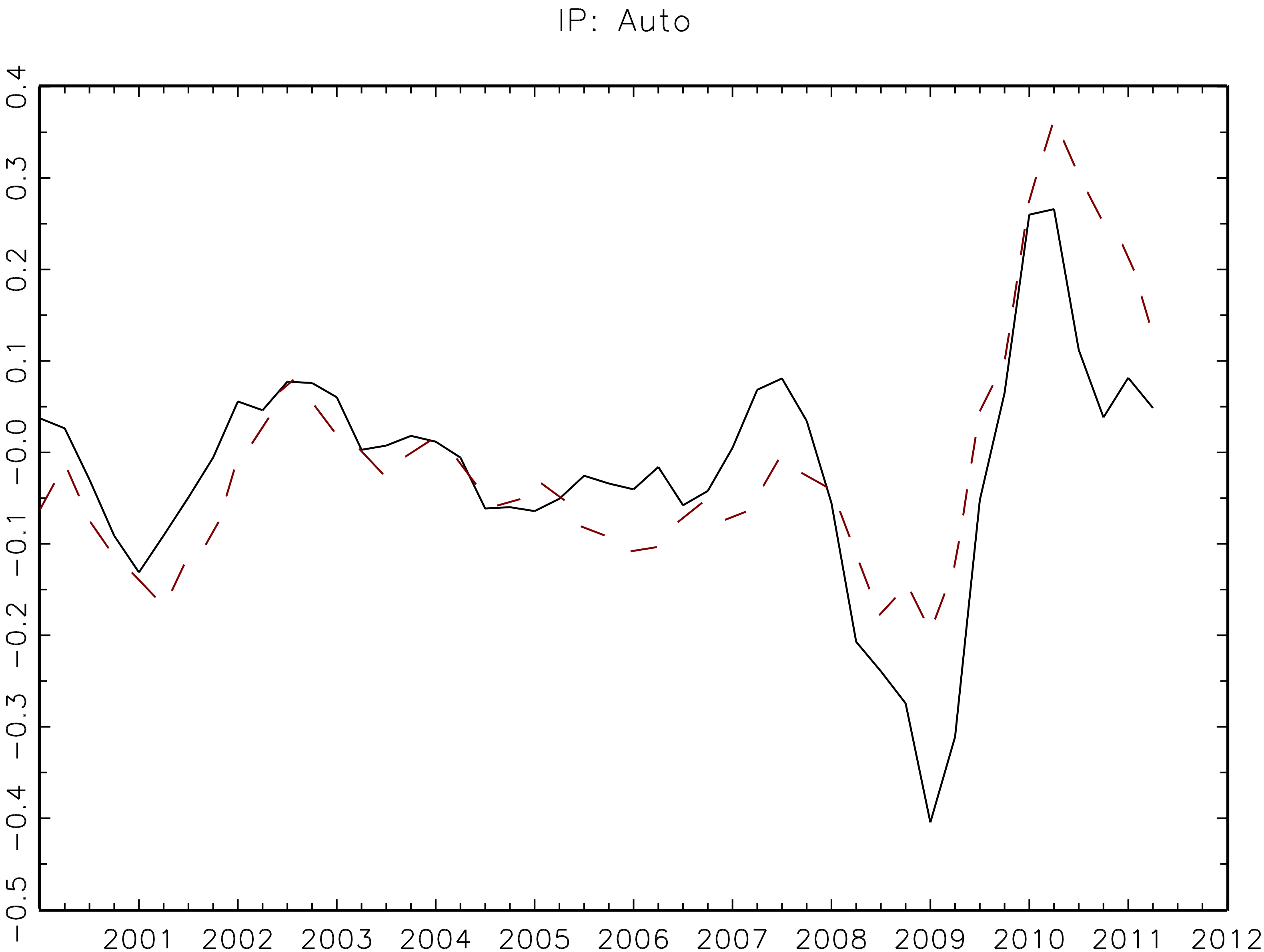
average R^2: 0.29 2007Q4 R^2: 0.83



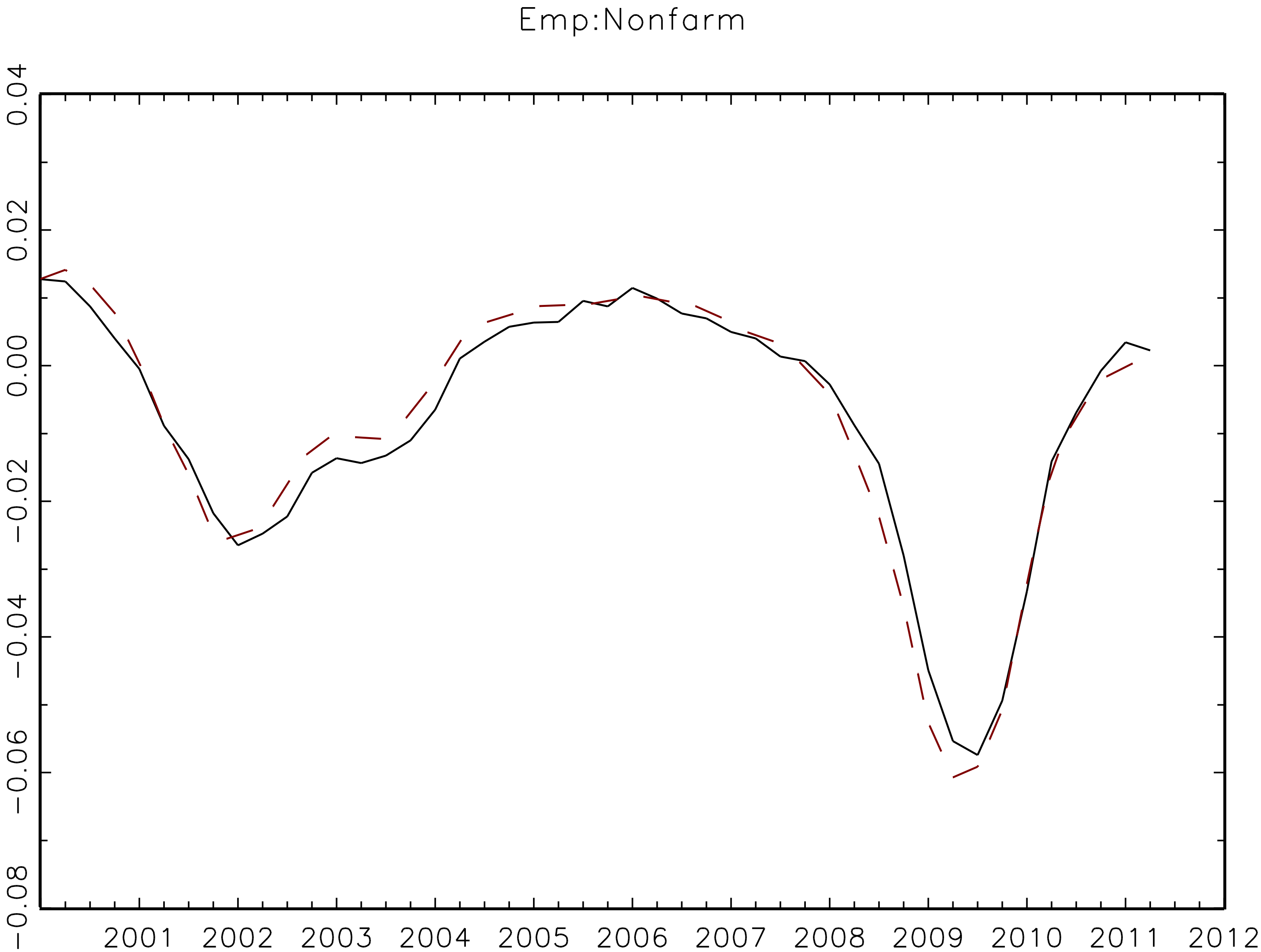
average R^2: 0.66 2007Q4 R^2: 0.86



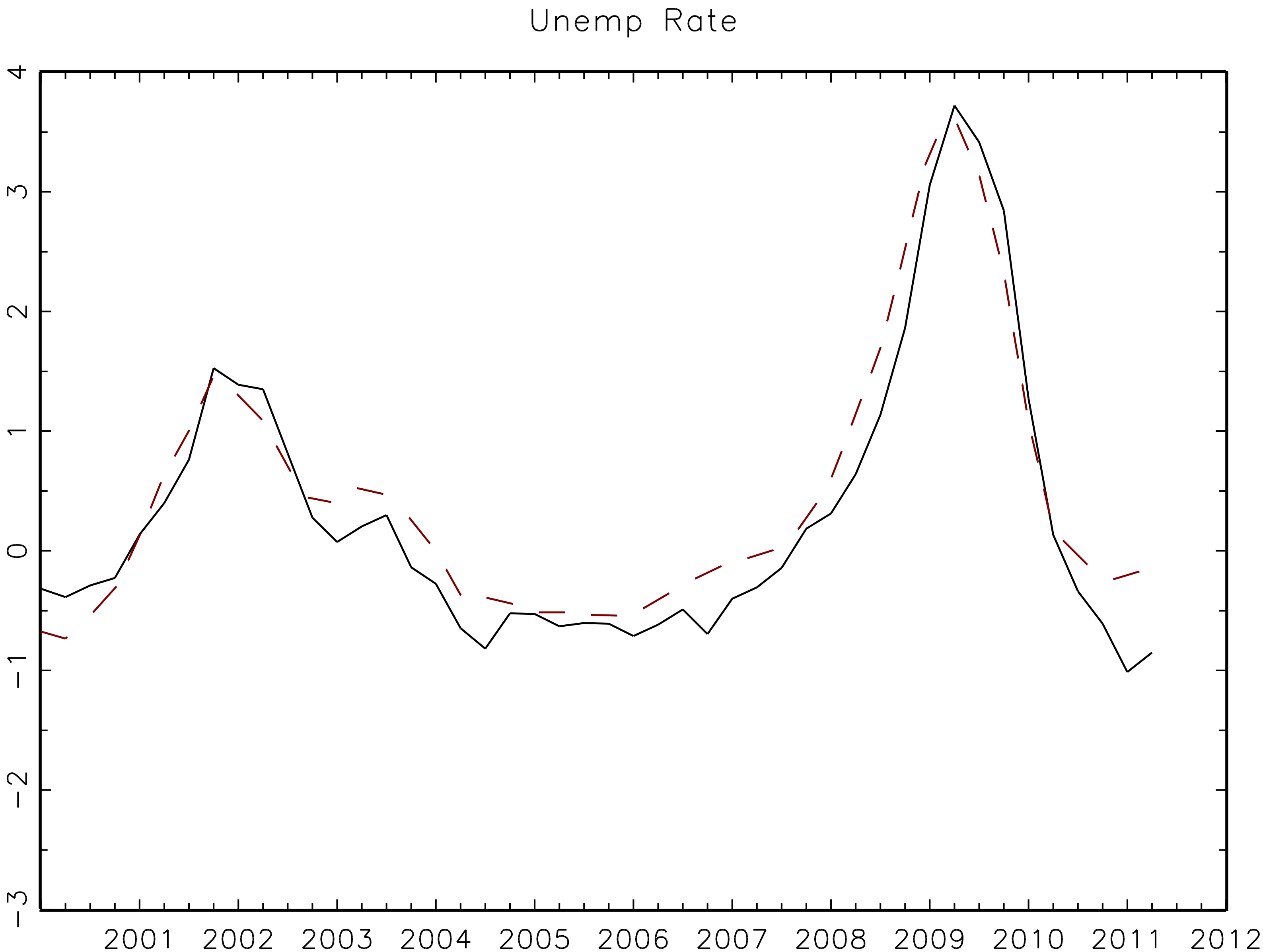
average R^2: 0.89 2007Q4 R^2: 0.95



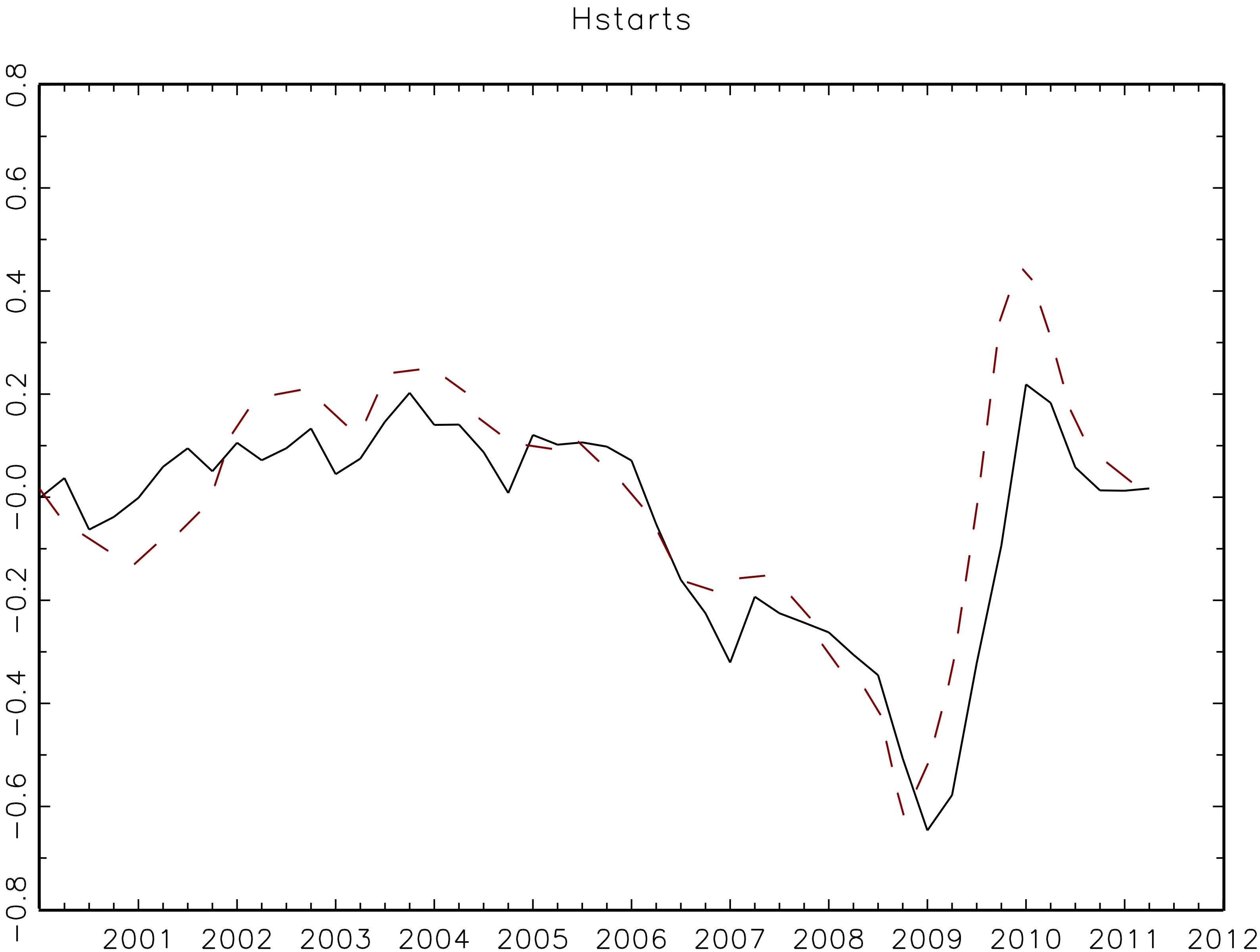
average R^2: 0.54 2007Q4 R^2: 0.62



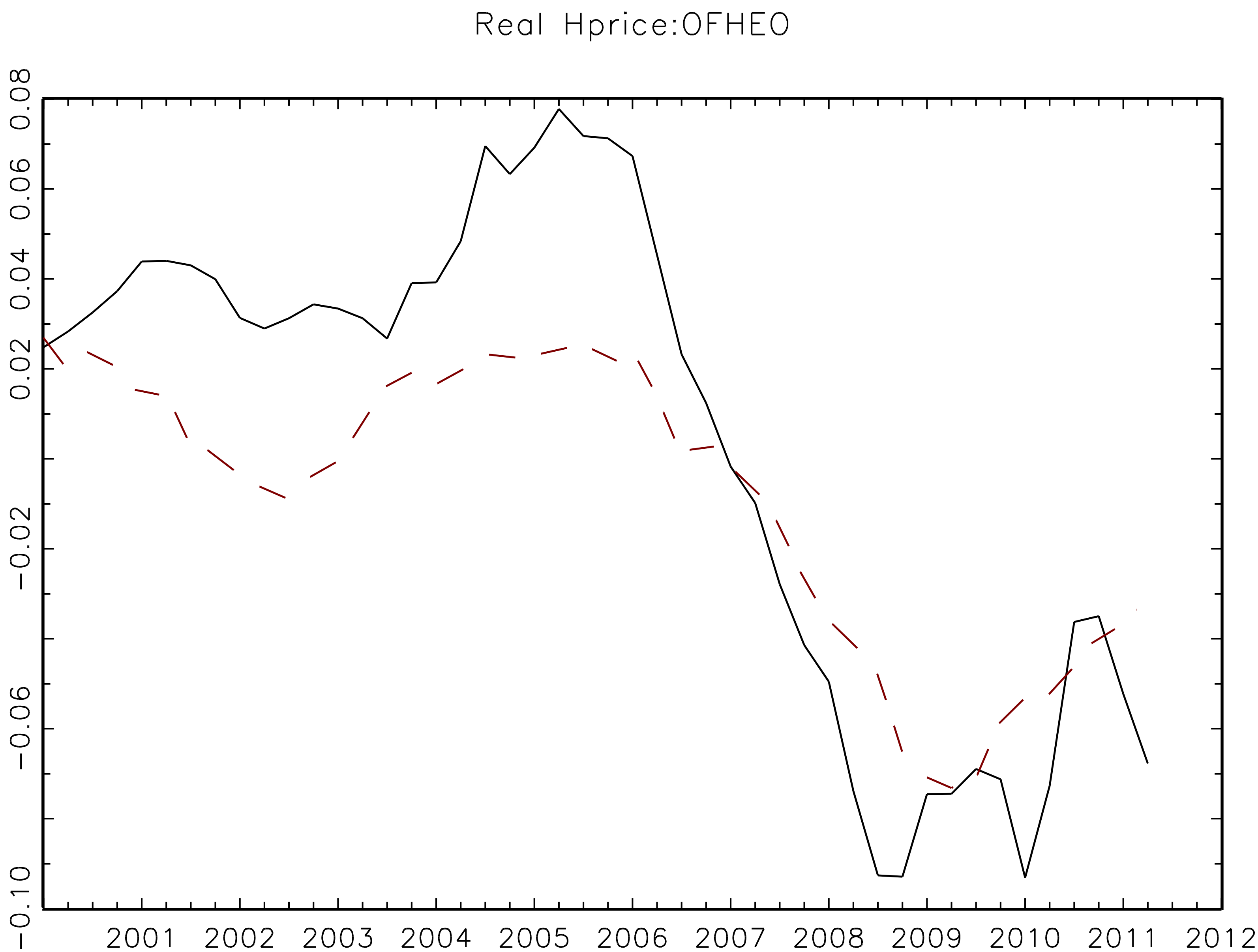
average R^2: 0.94 2007Q4 R^2: 0.96



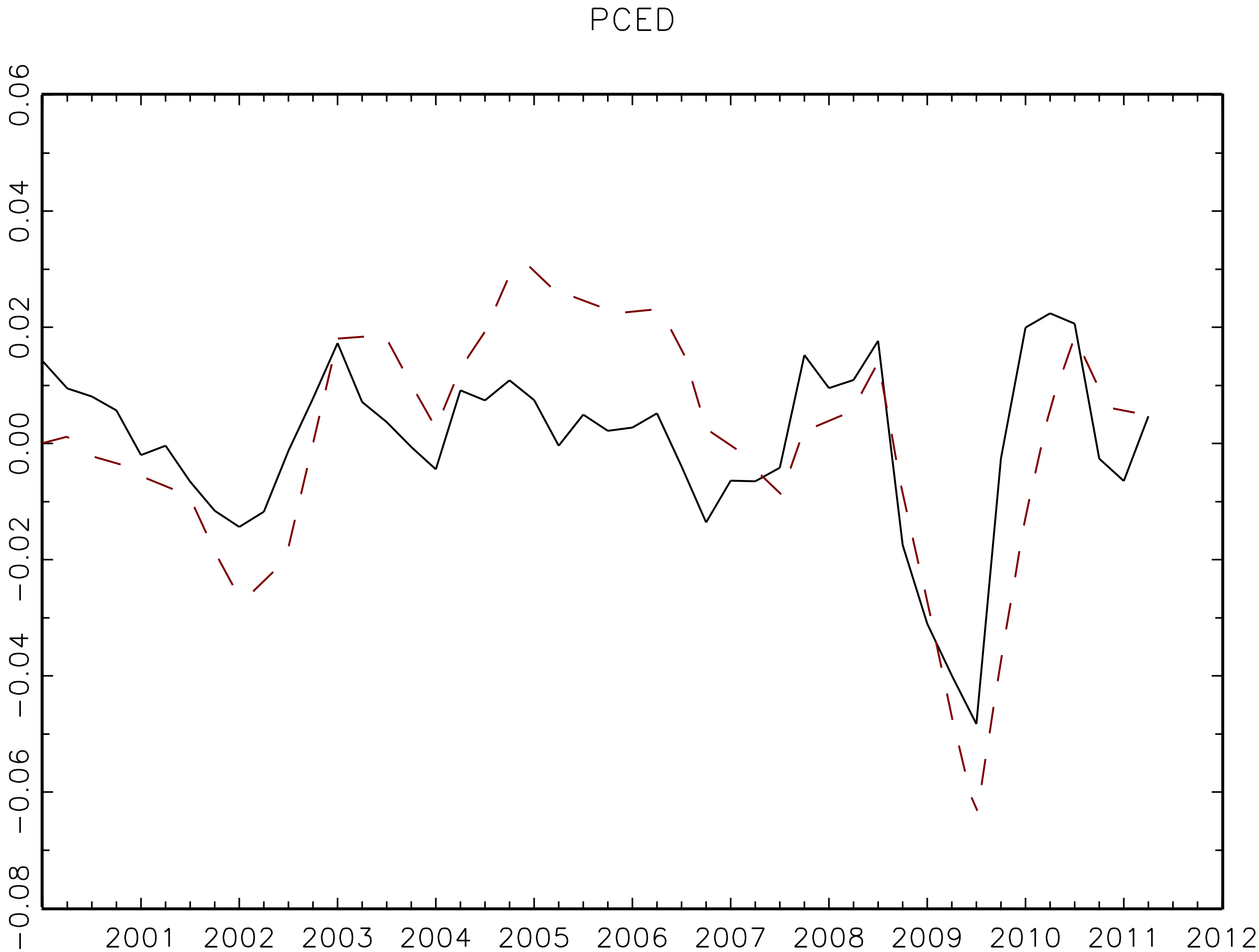
average R^2: 0.85 2007Q4 R^2: 0.89



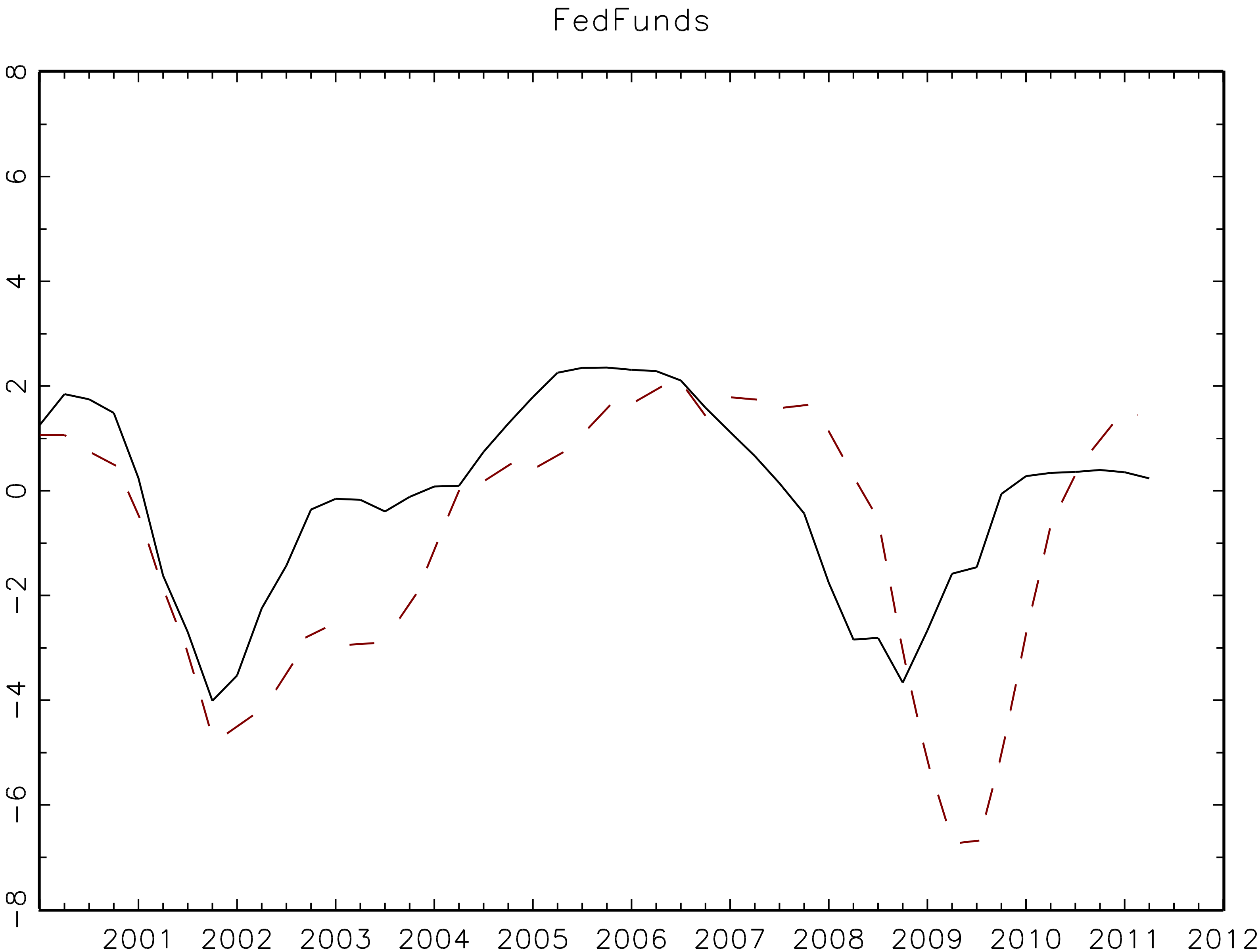
average R^2: 0.59 2007Q4 R^2: 0.53



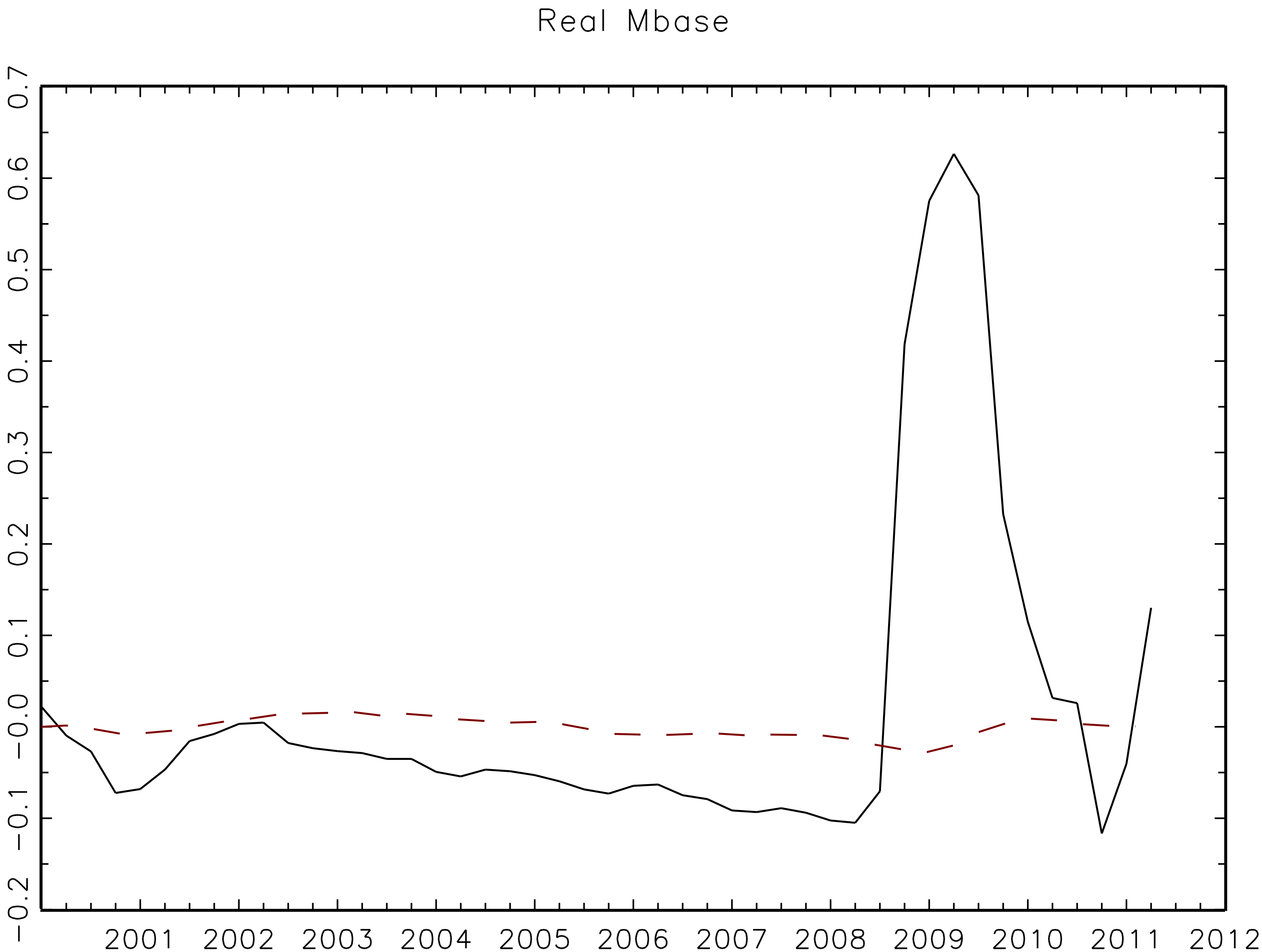
average R^2: 0.38 2007Q4 R^2: 0.55



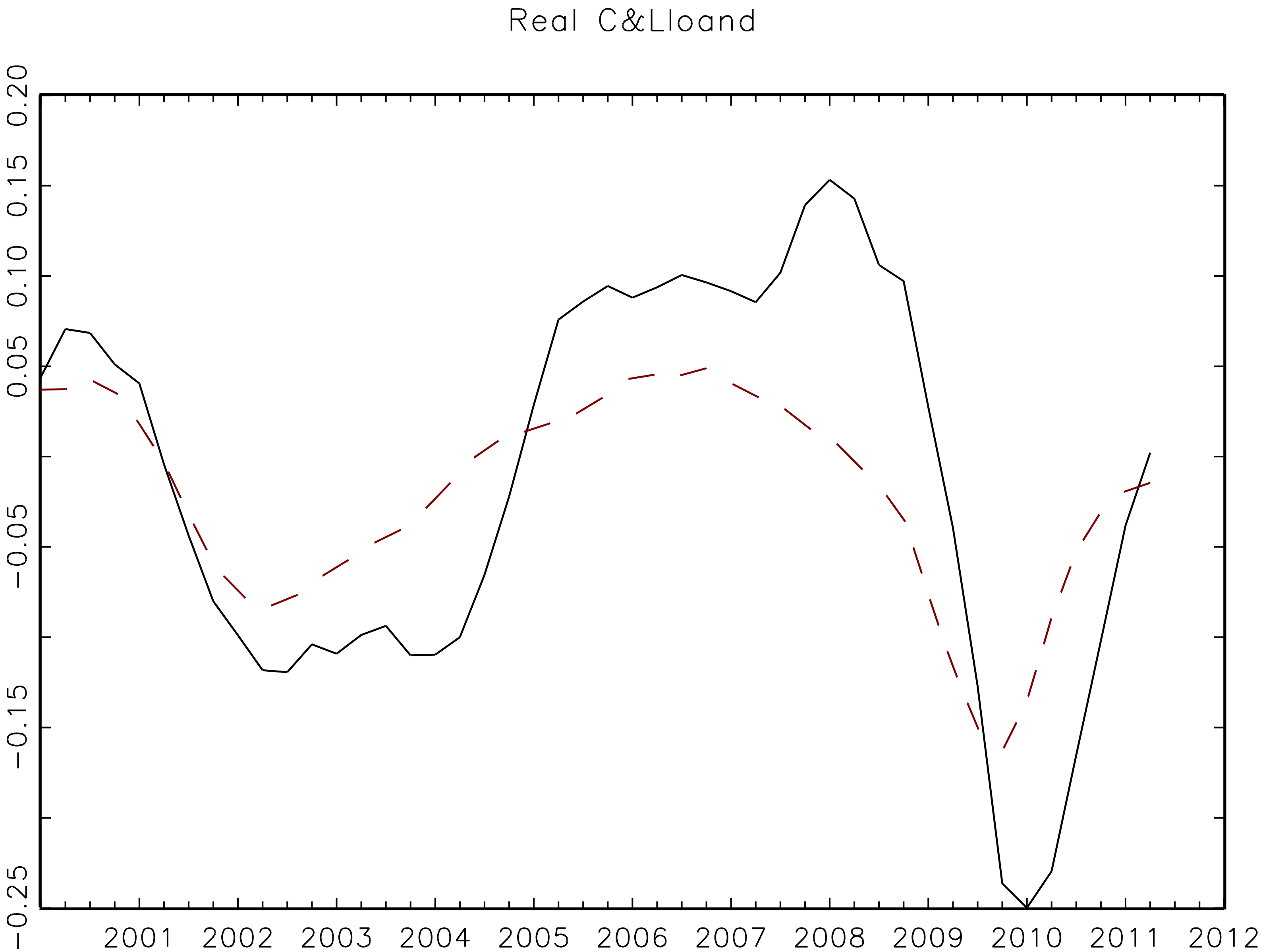
average R^2: 0.51 2007Q4 R^2: 0.82



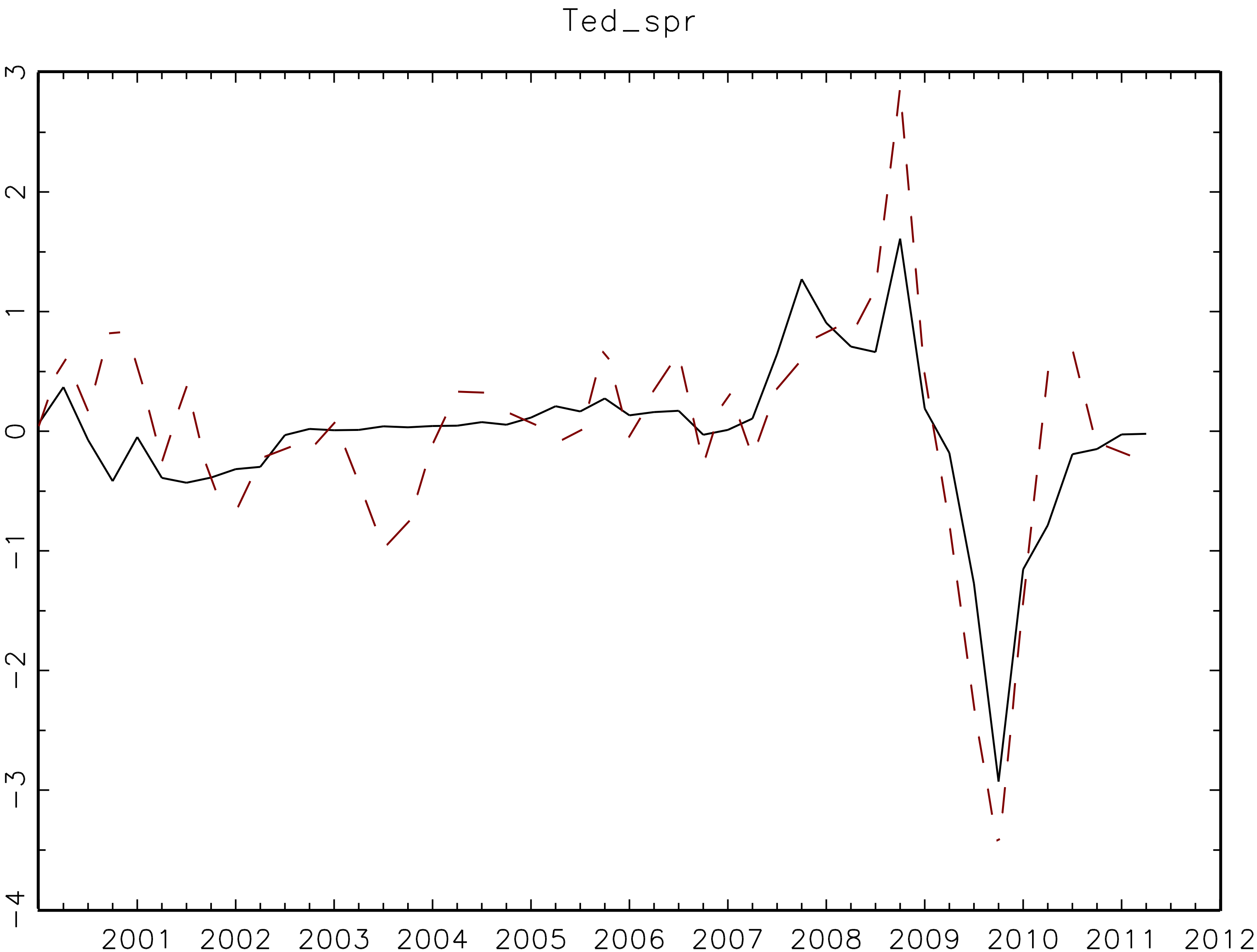
average R^2: 0.39 2007Q4 R^2: -1.54



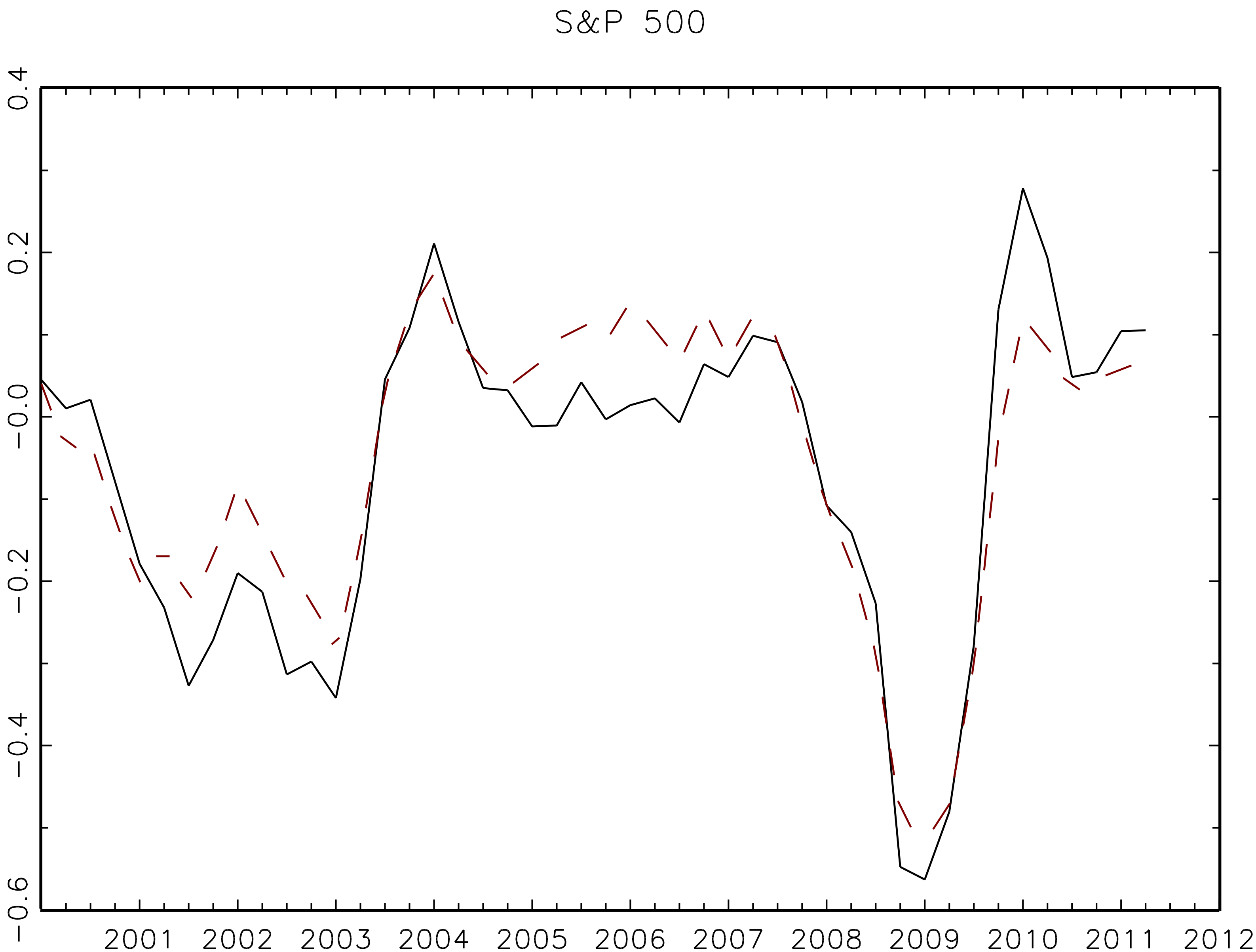
average R^2: 0.22 2007Q4 R^2: -0.03



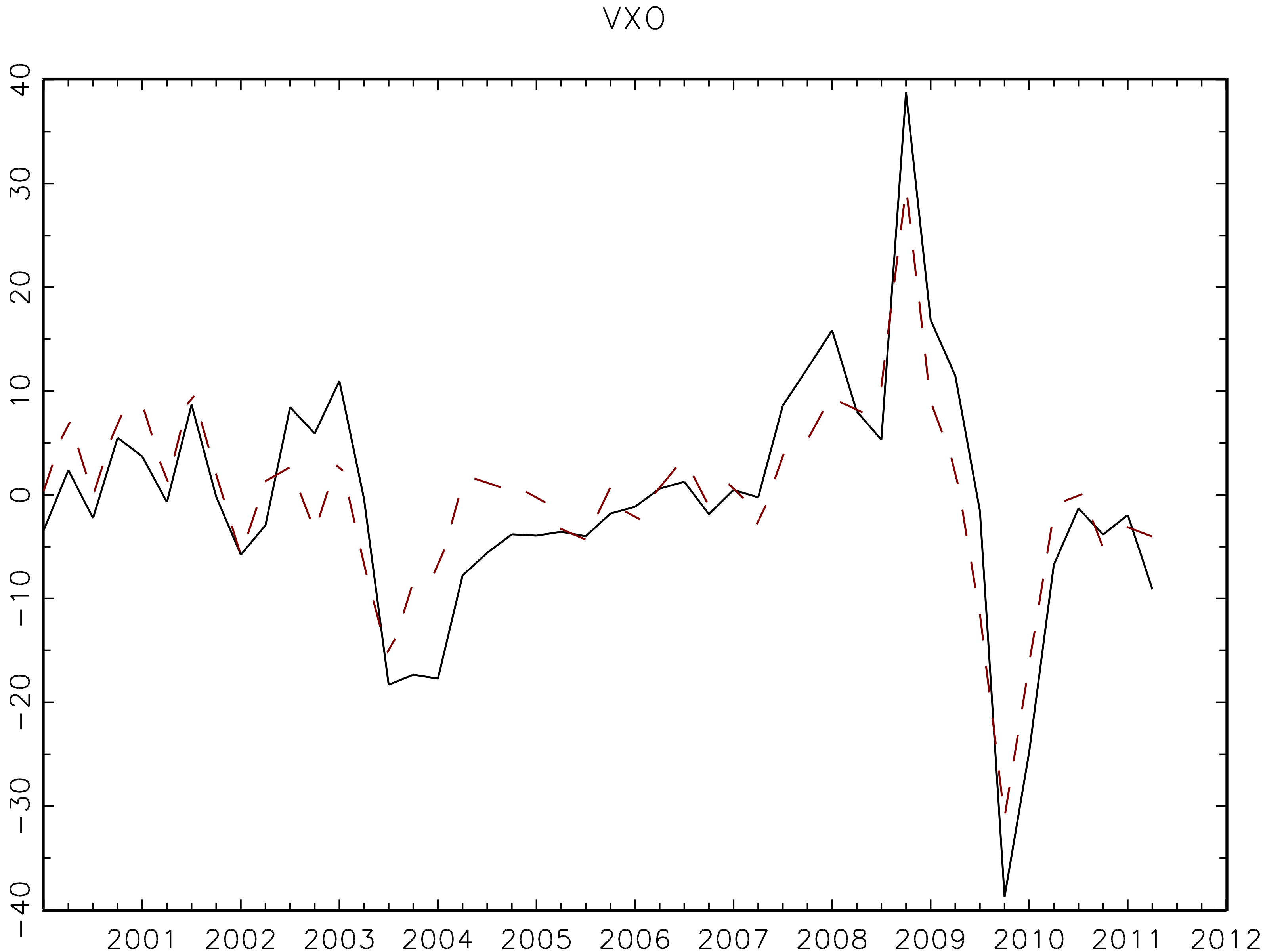
average R^2: 0.43 2007Q4 R^2: 0.78



average R^2: 0.43 2007Q4 R^2: 0.78



average R^2: 0.67 2007Q4 R^2: 0.87



average R^2: 0.12 2007Q4 R^2: 0.89

Stock and Watson (2012) "Disentangling the Channels of the 2007-09 Recession"

$$X_t = \Lambda F_t + e_t$$

$$\Phi(\mathrm{L})F_t = G\eta_t$$

Were there new factors in the 2007-09 recession? **No**

Were there instabilities in Λ? **Not much**

Were there instabilities in $\Phi(\mathrm{L})$? **Not much**

Were there unusually extreme values of η and/or e? **YES**

Returning to the Prediction Problem

Forecasting setup: $y_{t+1} = F_t'\alpha + \varepsilon_{t+1}$

$$X_t = \Lambda F_t + e_t$$

$$\Phi(\mathrm{L})F_t = G\eta_t$$

Use X to estimate F using $\hat{F}^{PC}$.

Use $\hat{F}^{PC}$ as if they were true values of F.

Result (Stock-Watson (2002)): $\hat{y}_{T+1}\left(\hat{F}^{PC}\right) - \hat{y}_{T+1}\left(F\right) \xrightarrow{ms} 0$

Outline

~~1. Motivation and Setup~~

~~2. Dynamic Factor Models~~

3. Shrinkage

4. Sparse Models

Linear prediction problem: $y_{t+1} = x_t{'}\beta + \varepsilon_{t+1}$

Simpler problem: Orthonormal regressors.

Transform regressors as $p_t = Hx_t$ where H is chosen so that

$$T^{-1}\sum_{t=1}^{T} p_t p_t{'} = T^{-1}P'P = I_n.$$ (Note: This requires $n \le T$)

Regression equation: $y_{t+1} = p_t{'}\alpha + \varepsilon_{t+1}$

OLS Estimator: $\hat{\alpha} = (P'P)^{-1}P'Y = T^{-1}P'Y$

so that $\hat{\alpha}_i = T^{-1}\sum_{t=1}^{T} p_{it} y_{t+1}$

Note: Suppose p_t are strictly exogenous and $\varepsilon_t \sim iidN(0,\sigma^2)$. (This will motivate the estimators .. more discussion below).

In this simple setting:

(1) $\hat{\alpha}$ are sufficient for α.

(2) $\left(\hat{\alpha}-\alpha\right) \sim N\left(0, T^{-1}\sigma^2 I_n\right)$

(3) MSFE: $E\left(\sum_{i=1}^{n} p_{iT}(\alpha_i - \tilde{\alpha}_i)\right)^2 + \sigma^2 \approx \frac{n}{T} MSE(\tilde{\alpha}) + \sigma^2$

So we can think about analyzing n-independent normal random variables, $\hat{\alpha}_i$, to construct estimators $\tilde{\alpha}(\hat{\alpha}_i)$ that have small MSE – shrinkage can help achieve this.

Shrinkage: Basic idea

Consider two estimators: (1) $\hat{\alpha}_i \sim \mathrm{N}(\alpha_i, T^{-1}\sigma^2)$

(2) $\tilde{\alpha}_i = ½\, \hat{\alpha}_i$

$\mathrm{MSE}(\hat{\alpha}_i) = T^{-1}\sigma^2$

$\mathrm{MSE}(\hat{\alpha}_i) = 0.25 \times (T^{-1}\sigma^2 + \alpha_i^2)$

$$\mathrm{MSFE}(\hat{\alpha}) = \frac{n}{T}\sigma^2 + \sigma^2$$

$$\mathrm{MSFE}(\tilde{\alpha}) = 0.25\times \left[\frac{n}{T}\sigma^2 + \sum_{i=1}^{n}\alpha_i^2\right] + \sigma^2$$

How big is $\sum_{i=1}^{n}\alpha_i^2$?

What is optimal amount (and form) of shrinkage?

It depends on distribution of $\{\alpha_i\}$

- Bayesian methods use priors for the distribution

- Empirical Bayes methods estimate the distribution

Examples 1: L_2 – Shrinkage

Bayes: Suppose $\alpha_i \sim \text{iidN}(0, T^{-1}\omega^2)$

Then, with $\hat{\alpha}_i|\alpha_i \sim \text{N}(\alpha_i, T^{-1}\sigma^2)$,

$$\begin{bmatrix} \alpha_i \\ \hat{\alpha}_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, T^{-1} \begin{bmatrix} \omega^2 & \omega^2 \\ \omega^2 & \sigma^2 + \omega^2 \end{bmatrix} \right)$$

so that $\alpha_i|\hat{\alpha}_i \sim N\left(\frac{\omega^2}{\sigma^2 + \omega^2}\hat{\alpha}_i, T^{-1}\frac{\omega^2\sigma^2}{\sigma^2 + \omega^2} \right)$

MSE minimizing estimator conditional mean: $\tilde{\alpha}_i = \frac{\omega^2}{\omega^2 + \sigma^2}\hat{\alpha}_i$

Empirical Bayes: Requires estimates of σ^2 and ω^2

If $T-n$ is large, then σ^2 can be accurately estimated.

If n is large, then ω^2 can be accurately estimated:

$$E(\hat{\alpha}_i^2) = T^{-1}(\sigma^2 + \omega^2), \text{ so } \hat{\omega}^2 = \frac{T}{n}\sum_{i=1}^{n}\hat{\alpha}_i^2 - \hat{\sigma}^2$$

(Extensions to more general distributions, etc. in this prediction framework – see Zhang (2005), and Knox, Stock and Watson (2004) and references therein.)

Alternative Formulation:

Write Joint density of data and α as

$$\text{constant} \times \exp\left\{-0.5\left[\frac{1}{\sigma^2}\sum_{t=1}^{T}(y_{t+1}-p_t'\alpha)^2+\frac{1}{\omega^2}\sum_{i=1}^{n}\alpha_i^2\right]\right\}$$

Which is proportional to posterior for α. Because posterior is normal, mean = mode, so $\tilde{\alpha}$ can be found by maximizing posterior. Equivalently by solving:

$$\min_{\tilde{\alpha}}\sum_{t=1}^{T}(y_{t+1}-p_t'\tilde{\alpha})^2+\lambda\sum_{i=1}^{n}\tilde{\alpha}_i^2 \quad \text{with } \lambda = \sigma^2/\omega^2$$

This is called “Ridge Regression”

In the original X – regressor model, the ridge estimator of

$$\tilde{\beta}^{Ridge} = \left(X'X + \lambda I_n \right)^{-1} (X'Y)$$

and λ can be determined by prior-knowledge, or estimated (empirical Bayes, cross-validation, etc.)

(Note this estimator allows $n > T$.)

Other shrinkage methods (There are many, of course, that depend on the assumed distribution of the regressions coefficients).

One of particular interest is ***Bayesian model averaging (BMA)***.

- ***References***
  - Leamer (1978); Min and Zellner (1990); Fernandez, Ley, and Steele (2001), Koop and Potter (2004)
  - Surveys: Hoeting, Madigan, Raftery, and Volinsky (1999), Geweke and Whiteman (2004)
- ***Basic idea***: there are many possible models (submodels); assign them prior probability and compute posterior means.
- ***The BMA setup*** (notation: using X_t, not P_t – this doesn't need orthogonalized regressors in theory).

$Y_{t+1} \mid X_t$ is given by one of K models, denoted by $M_1, \ldots, M_K$.

Models are linear, so M_k lists variables in model k

$\pi(M_k)$ = prior probability of model k

D_t denotes the data set through date t

The ***predictive density*** is the density of Y_{T+1} given the past data – the priors and the model are integrated out:

$$f(Y_{T+1}|D_T) = \sum_{k=1}^{K} f_k(Y_{T+1} \mid D_T)\Pr(M_k \mid D_T),$$

where $f_k(Y_{T+1}|D_T) = k^{\text{th}}$ predictive density

The ***posterior probability*** of model k is:

$$\Pr(M_k|D_T) = \frac{\Pr(D_T \mid M_k)\pi(M_k)}{\sum_{i=1}^{K}\Pr(D_T \mid M_i)\pi(M_i)},$$

where

$$\Pr(D_T|M_k) = \int \Pr(D_T \mid \theta_k, M_k)\pi(\theta_k \mid M_k)d\theta_k$$

θ_k = parameters in model k

$\pi(\theta_k|M_k)$ = prior for θ_k in model k

Under quadratic loss, optimal forecast is the mean of the predictive density, which is the weighted average of the forecasts you would make under each model, weighted by the posterior probability of that model:

$$\tilde{Y}_{T+1|T} = \sum_{k=1}^{K} \Pr(M_k \mid D_T)\tilde{Y}_{M_k,T+1|T},$$

where $\tilde{Y}_{M_k,T+1|T}$ = posterior mean of Y_{T+1} for model M_k.

Comments

- Akin to forecast combining – where there are K forecasts
- How many models are there? How many distinct subsets of 135 variables can you make?
- fun for computational Bayesians (MCMC, etc)
- This simplifies with orthogonal regressors however…
- Contrast with "Prediction Pools": Hall and Mitchel (2007), Geweke and Amisano (2011).

BMA with orthogonal regressors

Clyde, Desimone, and Parmigiani (1996), Clyde (1999):

- Variable j is in the model with probability π (coin flip)
- Given the model, the coefficients are distributed with a conjugate "*g*-prior" – and you get a closed form expression for posteriors (see Stock and Watson (2012))

More Comments:

1. Link to forecast combination – Bates and Granger (1969) … for an ambitious on-going application see Norges Bank (2014)

2. If the parameters of the prior (the "hyperparameters") are estimated, then this is parametric empirical Bayes.

3. All the theory and setup of BMA is for the cross-sectional case – the theoretical Bayes justification doesn't go through with predetermined regressors, nor for multistep forecasts. So its motivation is by analogy to to the i.i.d./exogenous regressor case.

Outline

~~1. Motivation and Setup~~

~~2. Dynamic Factor Models~~

~~3. Shrinkage~~

4. Sparse Models

Sparse models: Many/most values of β_i or α_i are zero.

Can be interpreted as shrinkage with lots of point mass at zero:

Approaches:

- BMA … (but can be computationally challenging … 2^n models): Hoeting, Madiga, Raftery, and Volinsky (1999))

- Hard thresholds (AIC/BIC) or smoothed out using "Bagging": (Breiman (1996), Bühlmann and Yu (2002); Inoue and Kilian (2008))

- L_1 penalization: Lasso ("Least Absolute Shrinkage and Selection Operator"): Tibshirani (1996)

Lasso: (With orthonormal regressors)

Ridge: $$\min_{\tilde{\alpha}} \sum_{t=1}^{T} (y_{t+1} - p_t '\tilde{\alpha})^2 + \lambda \sum_{i=1}^{n} \tilde{\alpha}_i^2$$

Lasso: $$\min_{\tilde{\alpha}} \sum_{t=1}^{T} (y_{t+1} - p_t '\tilde{\alpha})^2 + \lambda \sum_{i=1}^{n} \left|\tilde{\alpha}_i\right|$$

Equivalently: $$\min_{\tilde{\alpha}} \sum_{i=1}^{n} (\hat{\alpha}_i - \tilde{\alpha}_i)^2 + \lambda \sum_{i=1}^{n} \left|\tilde{\alpha}_i\right|$$

$$\min_{\tilde{\alpha}} \sum_{i=1}^{n} (\hat{\alpha}_i - \tilde{\alpha}_i)^2 + \lambda \sum_{i=1}^{n} |\tilde{\alpha}_i|$$

Notes:

- The solution yields $\text{sign}(\tilde{\alpha}_i) = \text{sign}(\hat{\alpha}_i)$

- Suppose $\hat{\alpha}_i > 0$. FOC … $2(\hat{\alpha}_i - \tilde{\alpha}_i) + \lambda = 0$
so solution is

$$\tilde{\alpha}_i = \begin{cases} \hat{\alpha}_i - \lambda / 2 \text{ if } (\hat{\alpha}_i - \lambda / 2) > 0 \\ 0 \text{ otherwise} \end{cases}$$

- Similarly for $\hat{\alpha}_i < 0$.

Comments:

(1) No closed form expression for estimator with non-orthogonal X, but efficient computational procedures using LARS (Efron, Johnstone, Hastie, and Tibshirani (2002), Hastie, Tibshirani, Friedman (2009)).

(2) "Oracle" Results: Fan and Li (2001), Zhao and Yu (2006), Zou (2006), Leeb and Pötscher (2008), Bickel, Ritov, and Tsybakov (2009).

(3) Nice overview for economists and economic research: Belloni, Chernozhukov, and Hansen (2014); application to choosing "controls" Belloni, Chernozhukov, and Hansen (2014b), and instruments Belloni, Chen, Chernozhukov, and Hansen (2012).

(4) Bayes Interpretation: Park and Casella (2008)

Suppose $\alpha_i \sim$ iid with $f(\alpha_i)$ = constant × $\exp\left(-\gamma|\alpha_i|\right)$

Then posterior is

$$\text{constant} \times \exp\left\{-0.5\left[\frac{1}{\sigma^2}\sum_{t=1}^{T}(y_{t+1}-p_t{}'\alpha)^2+2\gamma\sum_{i=1}^{n}|\alpha_i|\right]\right\}$$

The lasso estimator (with $\lambda = 2\gamma\sigma^2$) yields the posterior mode.

But note mode $\neq$ mean for this distribution.

Outline

~~1. Motivation and Setup~~

~~2. Dynamic Factor Models~~

~~3. Shrinkage~~

~~4. Sparse Models~~

Course Topics

~~1. Time series refresher and inference tools (MW)~~

~~2. The Kalman filter, nonlinear filtering, and Markov chain monte carlo (MW)~~

~~3. Prediction with large datasets (MW)~~

4. Heteroskedasticity and autocorrelation consistent (HAC) standard errors (JS)

5. Many instruments/weak identification in IV and GMM (JS)

6. Structural VAR modeling (JS)

References for Lecture 3: Prediction with large datasets

Bai, J., and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.

Bai, J., and S. Ng (2006), "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74,1133-1150.

Bai, J., and S. Ng (2010), "Instrumental Variable Estimation in a Data Rich Environment" *Econometric Theory*, 26:6, 1577-1606.

Bates, J.M., and Clive W.J. Granger (1969), "The Combination of Forecasts," *Operations Research Quarterly* 20, 451-468.

Belloni, A. V. D. Chen, Chernozhukov, and C. Hansen (2012) "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80:6, 657-681.

Belloni, A. V. Chernozhukov, and C. Hansen (2014a) "High-Dimensional Metho,ds and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives,* 28:2, 29-50.

Belloni, A. V. Chernozhukov, and C. Hansen (2014b) "Inference on Treatment Effects after Selection Amongst High-Dimensional Controls," *Review of Economic Studies,* 81, 608-650.

Bernanke, B.S., J. Boivin, and P. Eliasz (2005), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics*, 120, 387-422.

Bickel, P.J., Y. Rotov, and A.B. Tsybakov (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics,* 37:4, 1705-1732.

Boivin, J., and M.P. Giannoni (2006), "DSGE Models in a Data-Rich Environment," NBER WP12772.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 36, 105-139.

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *Annals of Statistics*, 30, 927-961.

Chamberlain, G., and M. Rothschild (1983), "Arbitrage Factor Structure, and Mean-Variance Analysis of Large Asset Markets," *Econometrica*, 51,1281-1304.

Clyde, M. (1999a), "Bayesian Model Averaging and Model Search Strategies" (with discussion) in *Bayesian Statistics 6*, ed. by J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith, Oxford: Oxford University Press.

Clyde, M., Desimone, H., and Parmigiani, G. (1996), "Prediction via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197-1208.

Connor, G., and R.A. Korajczyk (1986), "Performance Measurement with the Arbitrage Pricing Theory," *Journal of Financial Economics*, 15, 373-394.

Council of Economic Advisors, *Economic Activity During the Government Shutdown and Debt Limit Brinksmanship*, October 2013.

Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004). "*Least angle regression,*" 32:2, 407-499.

Engle, R.F., and M.W. Watson (1981), "A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates," *Journal of the American Statistical Association*, 76, 774-781.

Fan. J. and R. Li (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348-1360.

Fernandez, C., E. Ley, and M.F.J. Steele (2001), "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381-427.

Forni, M., and L. Reichlin (1998), "Let's Get Real: A Dynamic Factor Analytical Approach to Disaggregated Business Cycle," *Review of Economic Studies*, 65, 453-474.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), "The Generalized Factor Model: Identification And Estimation," *Review of Economics and Statistics*, 82, 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004), "The Generalized Factor Model: Consistency and Rates," *Journal of Econometrics*, 119, 231-255.

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio-Economic Models*, ed. by D.J. Aigner and A.S. Goldberger, Amsterdam: North-Holland.

Geweke, J. and G. Amisano (2011): "Optimal Prediction Pools," *Journal of Econometrics*, 164, 130–141.

Geweke, J., and C. Whiteman (2006), "Bayesian Forecasting," in *The Handbook of Economic* Sargent, T.J., and C.A. Sims (1977), "Business Cycle Modeling Without Pretending to Have Too Much A-Priori Economic Theory," in *New Methods in Business Cycle Research*, ed. by C. Sims et al., Minneapolis: Federal Reserve Bank of Minneapolis.

Giannone, D., L. Reichlin, and D. Small (2008), "Nowcasting: The Real-Time Informational Content of Macroeconomic Data," *Journal of Monetary Economics*, 55, 665-676.

Hall, S. G. and J. Mitchell (2007): "Combining Density Forecasts," *International Journal of Forecasting*, 23, 1–13.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning*, 2nd Edition, New York: Springer.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14(38): 382-401.

Inoue, Atsushi, and Lutz Kilian (2008), "How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation," *Journal of the American Statistical Association*, 103, 511 - 522

Knox, T., J.H. Stock, and M.W. Watson (2001), "Empirical Bayes Forecasts of One Time Series Using Many Regressors," Technical Working Paper No. 269 (NBER).

Koop, G., and S. Potter (2004), "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *Econometrics Journal*, 7, 550-565.

Leamer, E.E. (1978), *Specification Searches*, New York: Wiley.

Min, C., and A. Zellner (1993), "Bayesian and Non-Bayesian Methods for Combining Models And Forecasts with Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56, 89-118.

Norges Bank (2014), *Models for Short-term Forecasting*, available at http://www.norges-bank.no/en/Monetary-policy/Models-for-monetary-policy-analysis-and-forecasting/SAM/ .

Park, T., and Casella, G. (2008) The Bayesian Lasso, *Journal of the American Statistical Association*, 103:681-686.

Quah, D., and T.J. Sargent (1993), "A Dynamic Index Model for Large Cross Sections" (with discussion), in *Business Cycles, Indicators, and Forecasting*, ed. by J.H. Stock and M.W. Watson, Chicago: University of Chicago Press for the NBER, 285-310.

Sargent, T.J. (1989), "Two Models of Measurements and the Investment Accelerator," *Journal of Political Economy*, 97, 251-287.

Stock, J.H., and M.W. Watson (1989), "New Indexes of Coincident and Leading Economic Indicators," *NBER Macroeconomics Annual 1989*, 351-393.

Stock, J.H., and M.W. Watson (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167-1179.

Stock, J.H., and M.W. Watson (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors,"*Journal of Business and Economic Statistics*.

Stock, J.H and M.W. Watson (2012), "Disentangling the Channels of the 2007-2009 Recession", *Brookings Papers on Economic Activity,* Spring 2012.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B*., Vol. 58, No. 1, pages 267-288).

Zhang, C.-H. (2005), "General Empirical Bayes Wavelet Methods and Exactly Adaptive Minimax Estimation," *Annals of Statistics*, 33, 54-100.

Zhao, P. and B. Yu (2006), ′On Model Selection Consistency of Lasso," *Journal of Machine Learning*, 7, 2541-2563.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association,* 101, 1418-1429.

AEA Continuing Education Course

Time Series Econometrics

Lecture 3: Prediction with large datasets

Mark W. Watson
January 6, 2015
10:30AM – 12:30PM

Outline

1. Motivation and Setup
2. Dynamic Factor Models
3. Shrinkage
4. Sparse Models

1. Motivation and setup

"Linear" prediction problem: $y_{t+1} = x_t{'}\beta + \varepsilon_{t+1} = \sum_{i=1}^{n} x_{it}\beta_i + \varepsilon_{t+1}$

Sample size is T.

Forecast: $\hat{y}_{T+1} = \sum_{i=1}^{n} x_{iT}\hat{\beta}_i$

Forecast error: $y_{T+1} - \hat{y}_{T+1} = \sum_{i=1}^{n} x_{iT}(\beta_i - \hat{\beta}_i) + \varepsilon_{T+1}$

MSFE: $E\left(\sum_{i=1}^{n} x_{iT}(\beta_i - \hat{\beta}_i)\right)^2 + \sigma^2$

Suppose: (1) T is large and n is small
(2) T is large and n is large

"Linear" prediction problem: $y_{t+1} = x_t'\beta + \varepsilon_{t+1} = \sum_{i=1}^{n} x_{it}\beta_i + \varepsilon_{t+1}$

Suppose: (2) T is large and n is large

Approaches:

(1) Use "small-n" estimators (e.g. OLS)

(2) Impose some structure

(a) Common "Factors" (Dynamic Factor Model)

(b) β_i's are "small" (Shrinkage)

(c) There are only a few non-zero β_i's (Sparsity)

Outline

~~1. Motivation and Setup~~

2. Dynamic Factor Models

3. Shrinkage

4. Sparse Models

Dynamic Factor Models (DFMs)

Forecasting setup: $y_{t+1} = \alpha(\mathrm{L})f_t + \varepsilon_{t+1}$

$$x_{it} = \lambda_i(\mathrm{L})f_t + e_{it}$$

$$\Psi(\mathrm{L})f_t = \eta_t$$

"f_t" are latent factors.

x is useful for forecasting for y because x provides information about f:
$E(y_t \mid x_t) = E\left(\alpha(\mathrm{L})f_t \mid x_t\right)$

DFM: Use x to estimate f. Use this to forecast y. Estimated factors are also useful for other purposes.

DFMs: A brief survey

$$x_{it} = \lambda_i(L)f_t + e_{it}$$
$$\Psi(L)f_t = \eta_t$$

(1) Large T, small n DFMs: (Geweke (1977), Sargent and Sims (1977), Engle and Watson (1981), Stock and Watson (1989)). Parametric model:

$$\rho_i(L)e_{it} = a_{it}$$

$$\begin{bmatrix} a_t \\ \eta_t \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D_a & 0 \\ 0 & \Sigma_{\eta\eta} \end{bmatrix} \right), \quad D_a \text{ diagonal.}$$

Estimation via ML (Kalman filtering, etc.).

Conceptually and computationally difficult with large n. (Quah and Sargent (1989).

(2) Large-*n* "Approximate" factor models: Chamberlain-Rothschild (1983), Connor and Korajczyk (1986), Forni, Hallin, Lippi, Reichlin (2000, 2004), Stock and Watson (2002), Bai-Ng (2002, 2006), many others …

An example following Forni and Reichlin (1998): Suppose f_t is scalar and $\lambda_i(L) = \lambda_i$ ("no lags in the factor loadings"):

$$X_{it} = \lambda_i f_t + e_{it}$$

Then

$$\frac{1}{n}\sum_{i=1}^{n} X_{it} = \frac{1}{n}\sum_{i=1}^{n}(\lambda_i f_t + e_{it}) = \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right) f_t + \frac{1}{n}\sum_{i=1}^{n} e_{it}$$

If the errors e_{it} have *limited dependence* across series, then as *n* gets large,

$$\frac{1}{n}\sum_{i=1}^{n} X_{it} \xrightarrow{p} \bar{\lambda} f_t$$

In this special case, a very easy nonparametric estimator (the cross-sectional average) is able to recover f_t – as long as *n* is large

A convenient representation for the DFM: $X_t = \lambda(\mathrm{L})f_t + e_t$

$$\Psi(\mathrm{L})f_t = \eta_t,$$

Suppose that $\lambda(\mathrm{L})$ has at most p_f lags. Then the DFM can be written,

$$\begin{pmatrix} X_{1t} \\ \vdots \\ X_{nt} \end{pmatrix} = \begin{pmatrix} \lambda_{10} & \cdots & \lambda_{1p_f} \\ \vdots & \ddots & \vdots \\ \lambda_{n0} & \cdots & \lambda_{np_f} \end{pmatrix} \begin{pmatrix} f_t \\ \vdots \\ f_{t-p_f} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ \vdots \\ e_{nt} \end{pmatrix}$$

or $$\underset{t}{\overset{n\times 1}{X}} = \overset{n\times r}{\Lambda} \underset{t}{\overset{r\times 1}{F}} + \underset{t}{\overset{n\times 1}{e}}$$

F_t are sometimes called "*static factors*". But, they aren't static: the VAR for f_t implies that there is a VAR for F_t

$$\Phi(\mathrm{L})F_t = G\eta_t$$

where G is a matrix of 1's and zeros and Φ consists of 1's, 0's, and Ψ's.

Principal Components (estimating the factors by least squares)

$$X_t = \Lambda F_t + e_t$$

$$\Phi(\mathrm{L})F_t = G\eta_t$$

Consider estimating Λ and $\{F_t\}$ by least squares:

$$\min_{F_1,\ldots,F_T,\Lambda} T^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t) \qquad (1)$$

subject to $\Lambda'\Lambda = I_r$ (identification). Given Λ, the (infeasible) OLS estimator of F_t is:

$$\hat{F}_t(\Lambda) = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_t$$

Now substitute $\hat{F}_t(\Lambda)$ into (1) to concentrate out $\{F_t\}$:

$$\min_{\Lambda} T^{-1}\sum_{t=1}^{T} X_t'[I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda]X_t$$

$$\min_\Lambda T^{-1}\sum_{t=1}^{T} X_t'[I-\Lambda(\Lambda'\Lambda)^{-1}\Lambda]X_t$$

$$\to \max_\Lambda T^{-1}\sum_{t=1}^{T} X_t'\Lambda(\Lambda'\Lambda)^{-1}\Lambda X_t$$

$$\to \max_\Lambda \mathrm{tr}\{(\Lambda'\Lambda)^{-1/2\prime}\Lambda'\left(T^{-1}\sum_{t=1}^{T} X_t X_t'\right)\Lambda(\Lambda'\Lambda)^{-1/2}\}$$

$$\to \max_\Lambda \mathrm{tr}\{\Lambda'\hat{\Sigma}_{XX}\Lambda\} \text{ s.t. } \Lambda'\Lambda = I_r, \text{ where } \hat{\Sigma}_{XX} = T^{-1}\sum_{t=1}^{T} X_t X_t'$$

$\to$ $\hat{\Lambda}$ = first r eigenvectors of $\hat{\Sigma}_{XX}$ (corresponding to largest eigenvalues)

Remember $\hat{F}_t(\Lambda) = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_t$, so

$$\hat{F}_t(\hat{\Lambda}) = \left(\hat{\Lambda}'\hat{\Lambda}\right)^{-1}\hat{\Lambda}'X_t = \hat{\Lambda}'X_t \quad (\text{because } \hat{\Lambda}'\hat{\Lambda} = I_r)$$

= first r principal components of X_t.

Distribution theory for PC as factor estimator

Selected results for the approximate DFM: $X_t = \Lambda F_t + e_t$

Typical conditions (Stock-Watson (2002), Bai-Ng (2002, 2006),…):

(a) $\frac{1}{T}\sum_{i=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F$ (stationary factors)

(b) $\Lambda'\Lambda/n \rightarrow$ (or $\xrightarrow{p}$) Σ_Λ Full rank factor loadings

(c) e_{it} are weakly dependent over time and across series (approximate DFM)

(d) F, e are uncorrelated at all leads and lags

(e) n, $T \rightarrow \infty$ plus Bai-Ng (2006) rate condition: $n^2/T \rightarrow \infty$

Selected results for the approximate DFM, ctd.

Stock and Watson (2002), Bai and Ng (2006):

- consistency of $\hat{F}_t$ for F_t (up to a $r\times r$ rotation)
- $\hat{F}_t$ converges at a sufficiently fast rate that $\hat{F}_t$ can be used as a regressor (e.g. in forecasting equations) without adjusting standard errors – you can treat $\hat{F}_t$ as if it actually *is* F_t (up to a $r\times r$ rotation)
- The PCA estimator of the common component is asymptotically normal at rate $\min(n^{1/2}, T^{1/2})$
- Bai-Ng (2006) give a method for constructing confidence bands for predicted values (these are for predicted value [for example estimates of common components] – *not* forecast confidence bands)

Estimating the number of factors in *F*

Most widely used method: Bai-Ng (2002) propose an estimator of r based on an information criterion; their main result is $\hat{r} \xrightarrow{p} r_0$ for the approximate DFM

Digression on information criteria (IC) for lag length selection in an AR

Consider the AR(p): $y_t = a_1 y_{t-1} + \ldots + a_p y_{t-p} + \varepsilon_t$

- Why not just maximize the R^2?
- IC trades off estimator bias (too few lags) vs. estimator variance (too many lags), from the perspective of fit of the regression:

Bayes Information Criterion: $$\text{BIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{\ln T}{T}$$

Akaike Information Criterion: $$\text{AIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{2}{T}$$

The Bai-Ng (2002) information criteria have the same form:

$$\text{IC}(r) = \ln\left(\frac{SSR(r)}{T}\right) + \text{penalty}(N, T, r)$$

Bai-Ng (2002) propose several IC's with different penalty factors that all produce consistent estimators of r. Here is the one that seems to work best in MCs (and is the most widely used in empirical work):

$$IC_{p2}(r) = \ln(\mathrm{V}(r,\hat{F}^r)) + r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right]$$

where $$V(r,\hat{F}^r) = \min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \lambda_i^{r\prime}\hat{F}_t^r\right)^2$$

$$= \min_{F_1,\ldots,F_T,\Lambda}(NT)^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t)$$

$\hat{F}_t^r$ are the PC estimates of r

Bai-Ng (2002) IC_{p2}: $$IC_{p2}(r) = \ln(V(r,\hat{F}^r)) + r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right]$$

where $$V(r,\hat{F}^r) = \min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \lambda_i^{r\prime}\hat{F}_t^r\right)^2$$

Comments:

- $\ln(V(r,\hat{F}^r))$ is a measure of (trace) fit – generalizes $\ln(SSR/T)$ in AIC/BIC
- If $N = T$, then $r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right] = r\left(\frac{2T}{T^2}\right)\ln T = 2r\frac{\ln T}{T}$

  which is $2 \times$ the usual BIC penalty factor
- Both N and T are in the penalty factor: you need N, $T \to \infty$.
- Bai-Ng's (2002) main result: $\hat{r} \xrightarrow{p} r_0$

Comments on Bai-Ng factor selection:

- Monte Carlo studies show B-N works well when n, T are large, and DFM model is correct.
- But in practice:
    - Different IC can yield substantially different answers
    - Adding series often increases the number of estimated factors (adding sectors should increase number of factors; adding series within sectors should not)
- Judgment is required
- There are several estimators that have been proposed and this is an on-going area of research.

Empirical Applications using DFMs – many, here are a few:

(1) Forecasting ... more on this below

(2) SVARs: Bernanke, Boivin, and Eliasz's (2005) is most famous example.

(3) Factors as instruments: Bai and Ng (2011)

(4) DSGE Modeling: Sargent (1989), Boivin-Giannoni (2006b).

(5) Real-time tracking: Stock and Watson (1989), Giannone, Reichling and Small (2008), Council of Economic Advisors (2012)

(6) Data Description: example follows …

Stock and Watson (2012) "Disentangling the Channels of the 2007-09 Recession"

$$X_t = \Lambda F_t + e_t$$

$$\Phi(\mathrm{L})F_t = G\eta_t$$

Were there new factors in the 2007-09 recession?

Were there instabilities in Λ?

Were there instabilities in $\Phi(\mathrm{L})$?

Were there unusually extreme values of η and/or e?

1. Structural breaks post 2007Q4

Empirical analysis

1. Estimate DFM parameters using data through 2007Q3
   a. Compute factors using "old" factor loadings:
   b. $\hat{F}_t = (\hat{\Lambda}'\hat{\Lambda})^{-1}\hat{\Lambda}'X_t$, where $\hat{\Lambda}$ are pre-07Q3 factor loadings
   c. How well do pre-07Q3 factors & factor loadings do in explaining post-07Q4 macro variables?

2. Formal stability tests:
   a. Stability of Λ
   b. Test for new factor (excess covariance among idiosyncratic disturbances)

1.1. Fit of pre-07Q3 parameters and factors, post-07Q4

Figures:

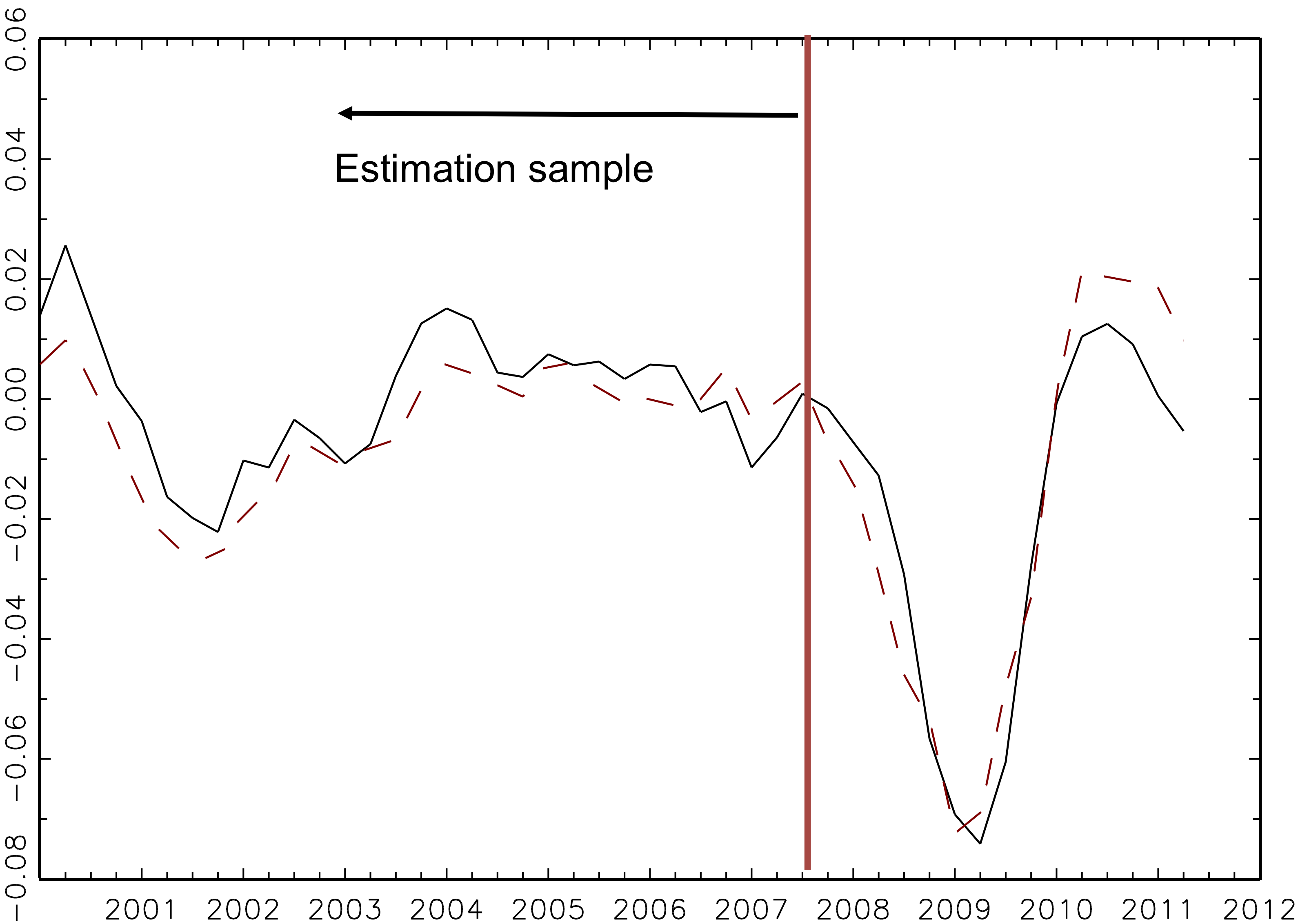
Plot of 4-Q growth ($100ln(X_t/X_{t-4})$) or 4-Q change:
solid = actual
dashed = common component (pre-07Q3 model)

Average R^2 2007Q4 R^2

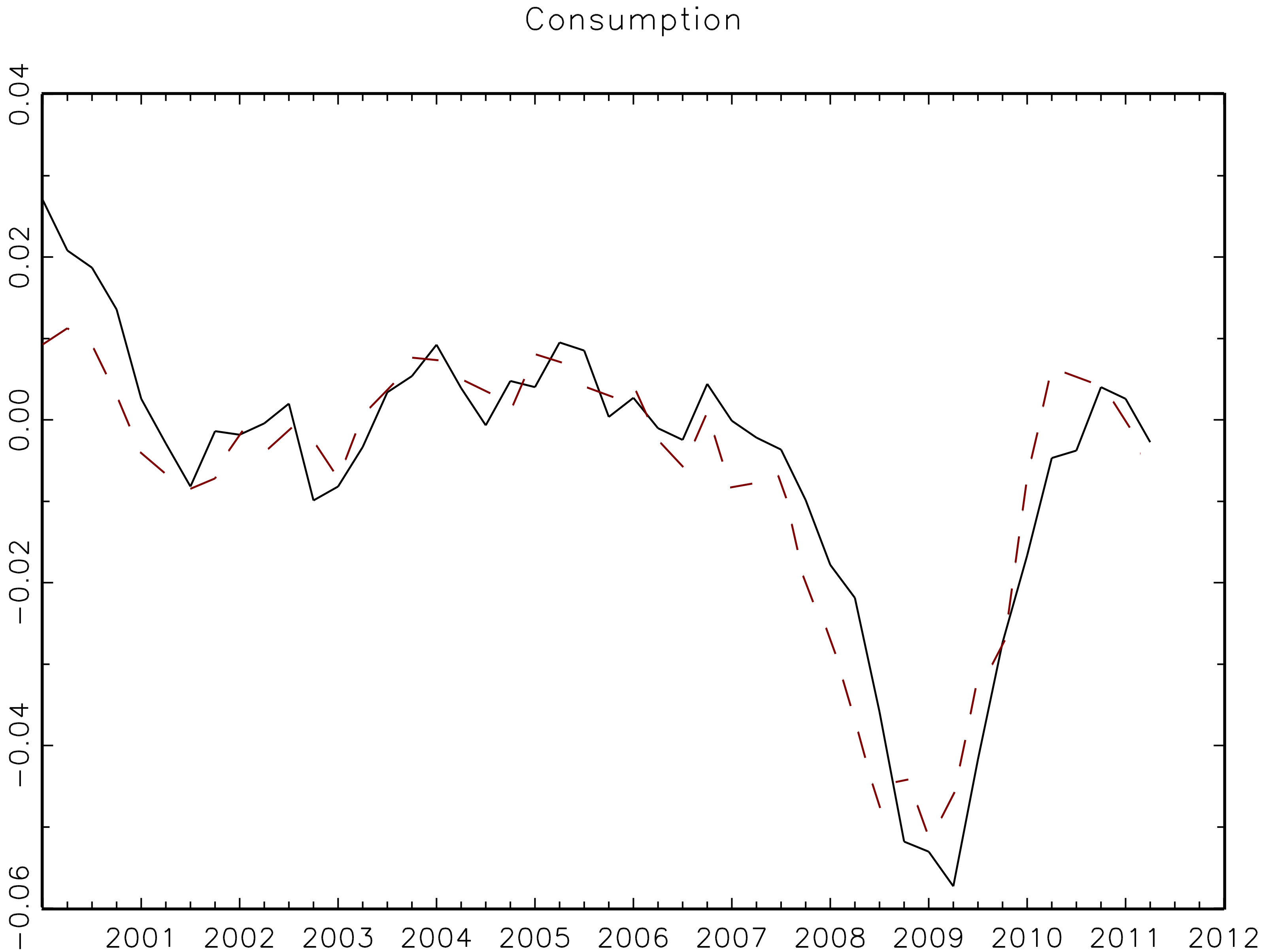
Average R^2 = 1-quarter R^2 of "ΛF_t", NBER peak to peak + 14 quarters, averaged over previous 7 recessions, 1960Q1,…, 2001Q1

2007Q4 R^2 = value for 2007Q4 – 2011Q2.

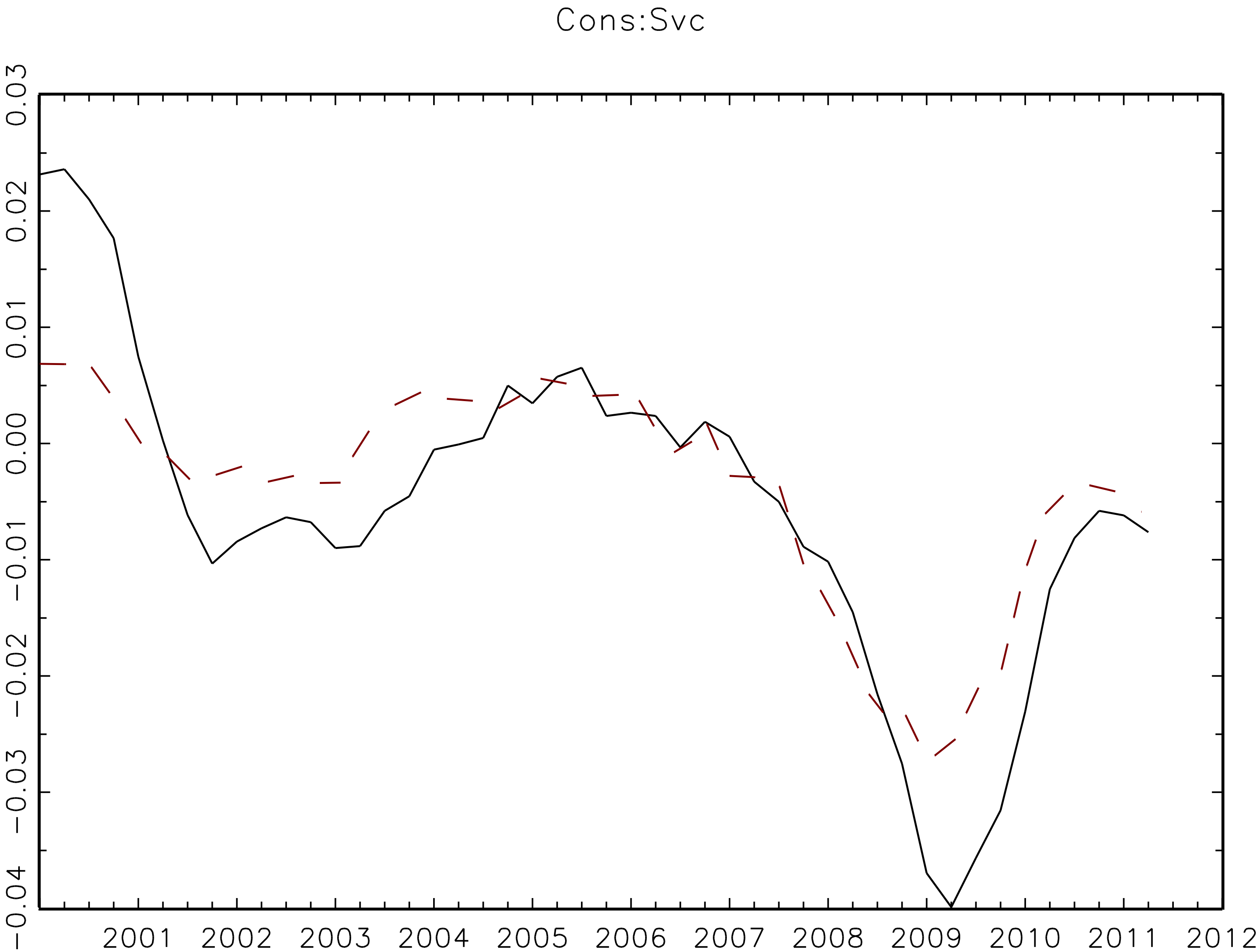
GDP



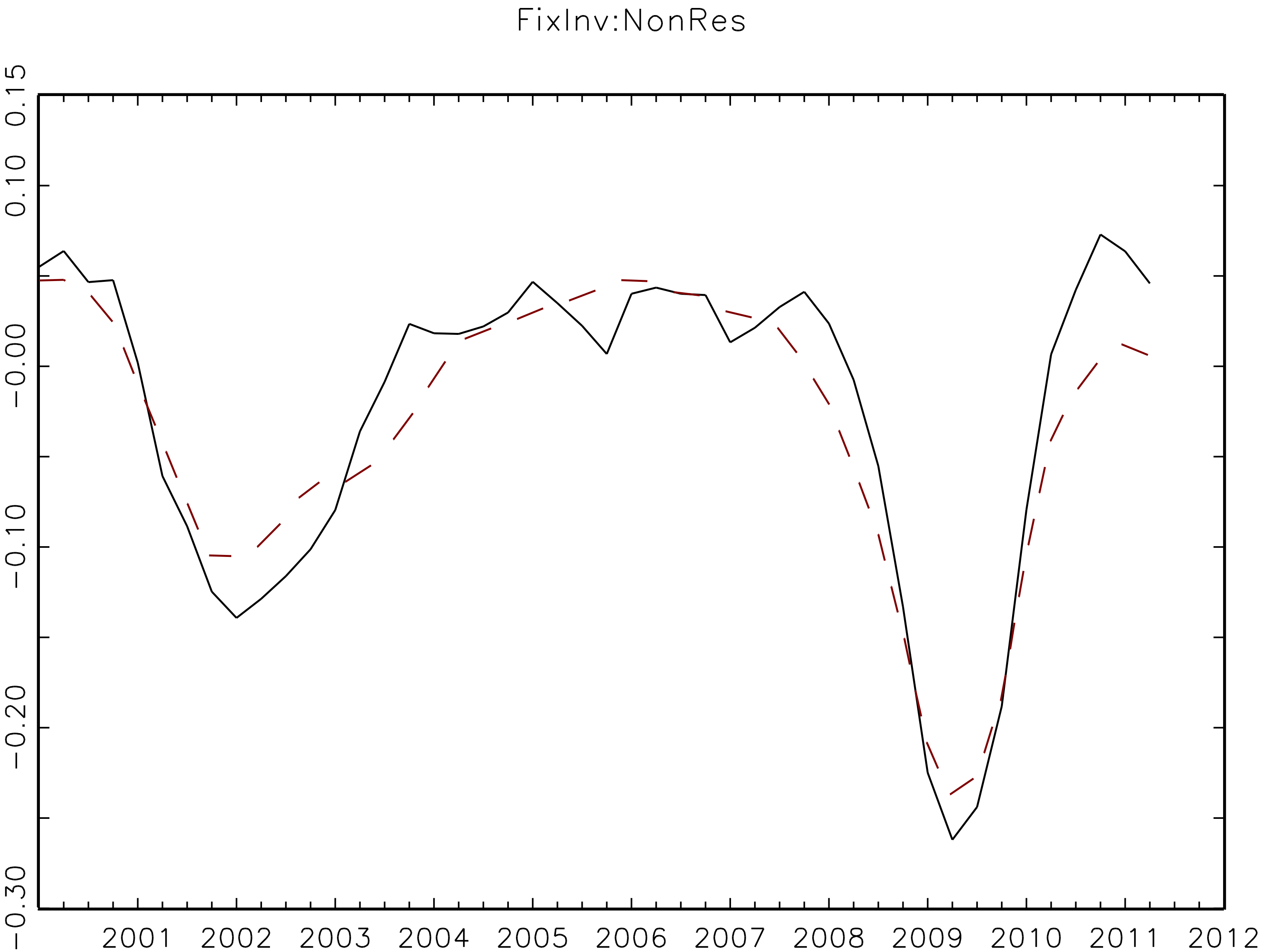
average R^2: 0.78 2007Q4 R^2: 0.64



average R^2: 0.62 2007Q4 R^2: 0.56

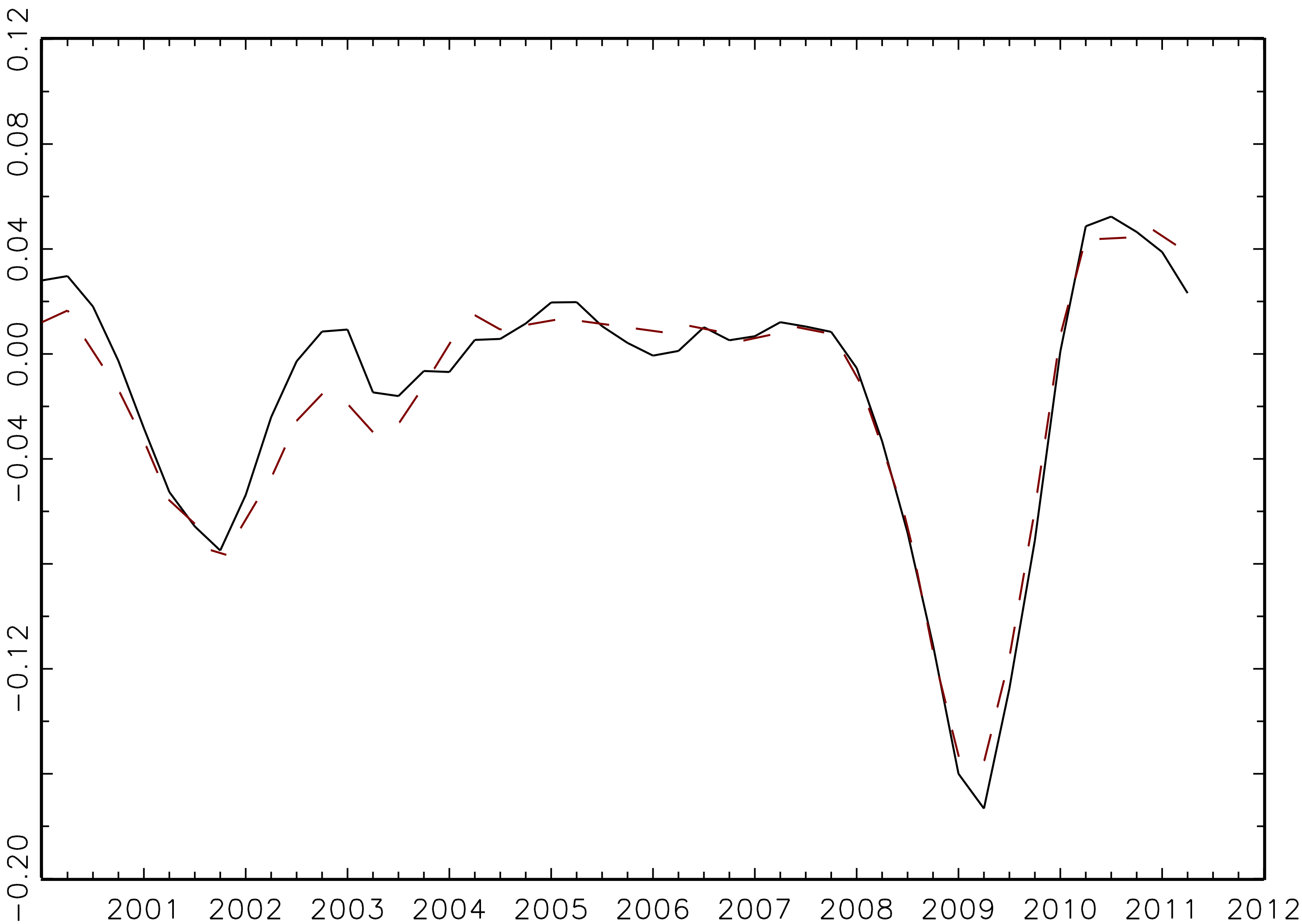


average R^2: 0.29 2007Q4 R^2: 0.83

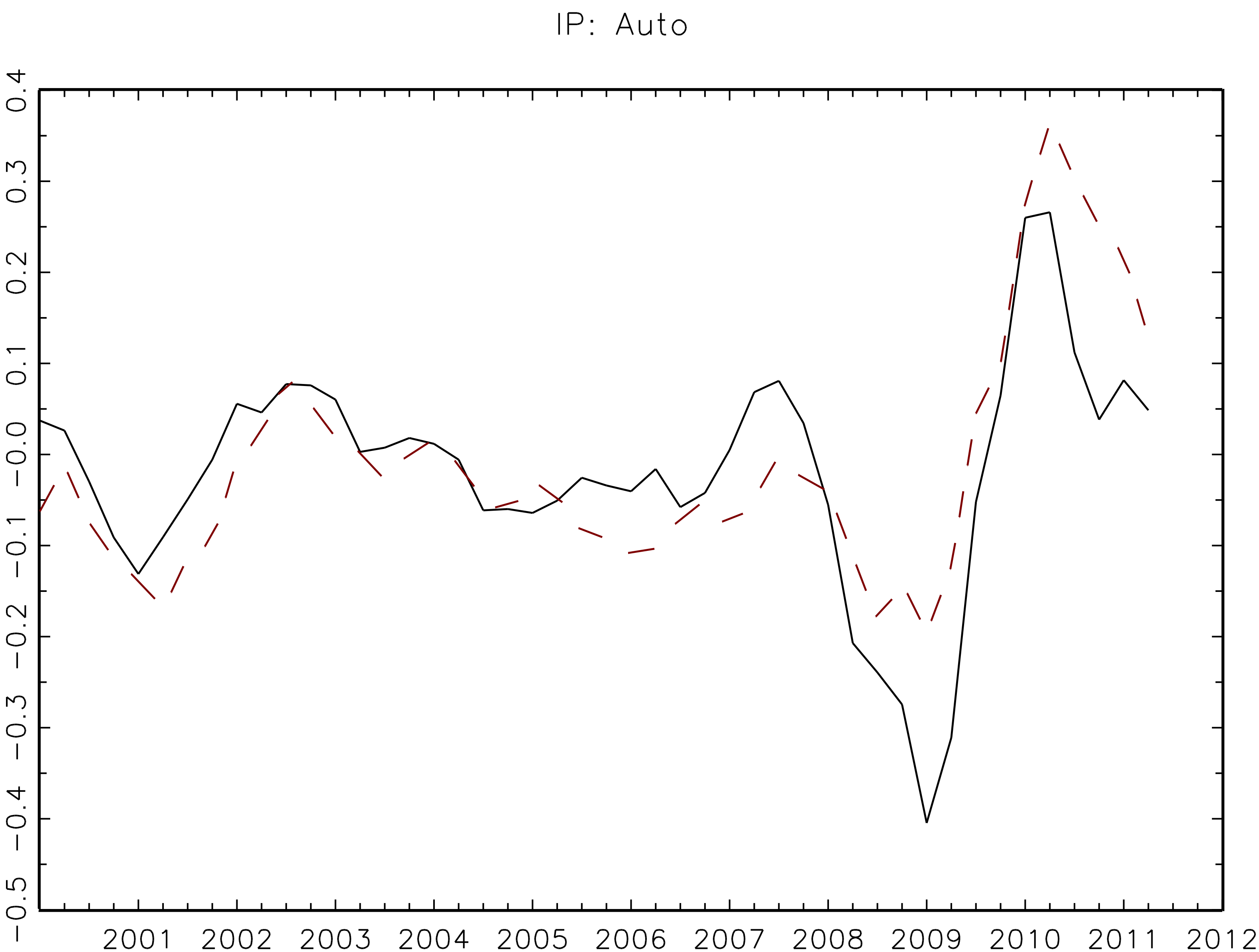


average R^2: 0.66 2007Q4 R^2: 0.86

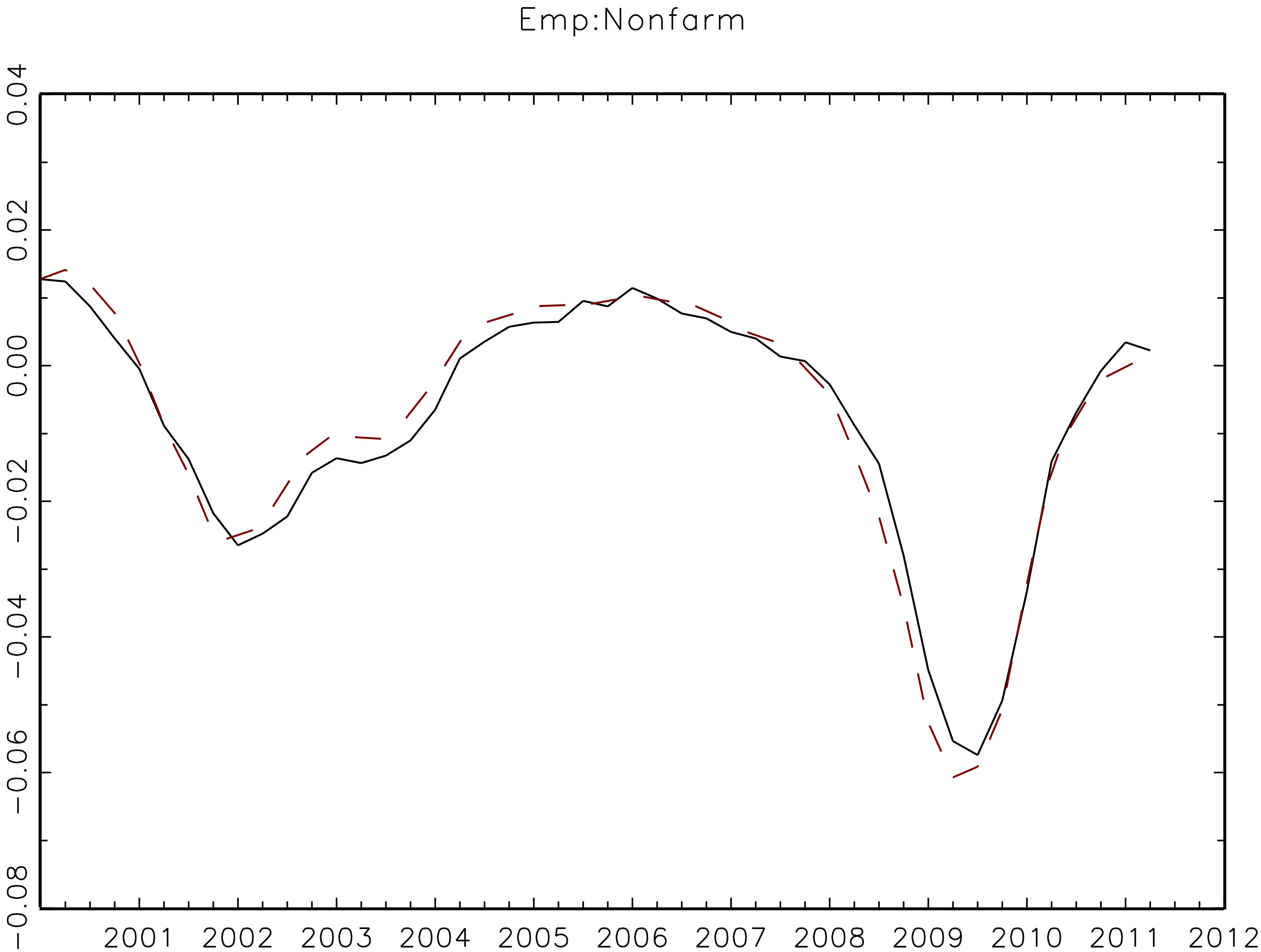
IP: Total index



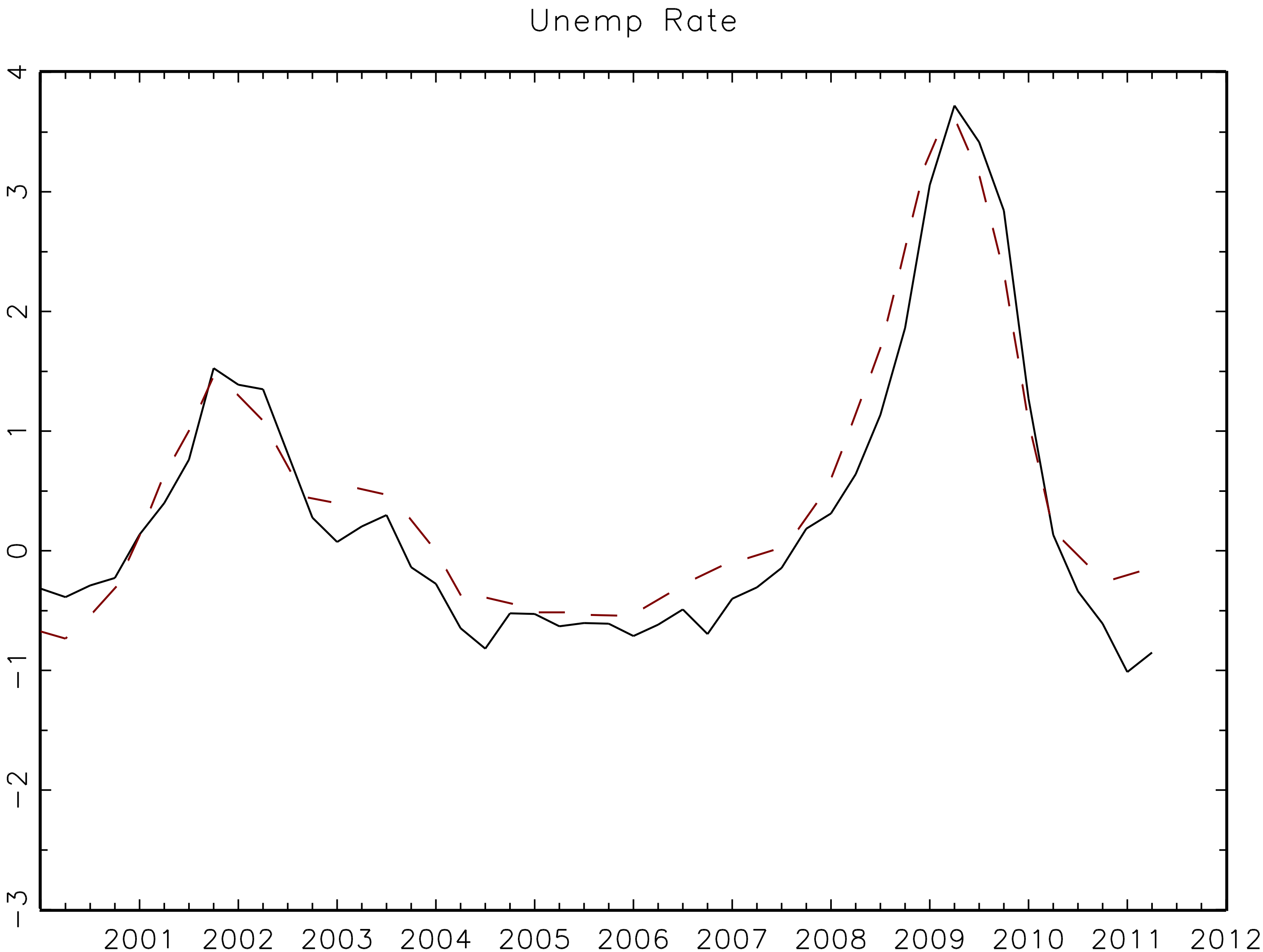
average R^2: 0.89 2007Q4 R^2: 0.95



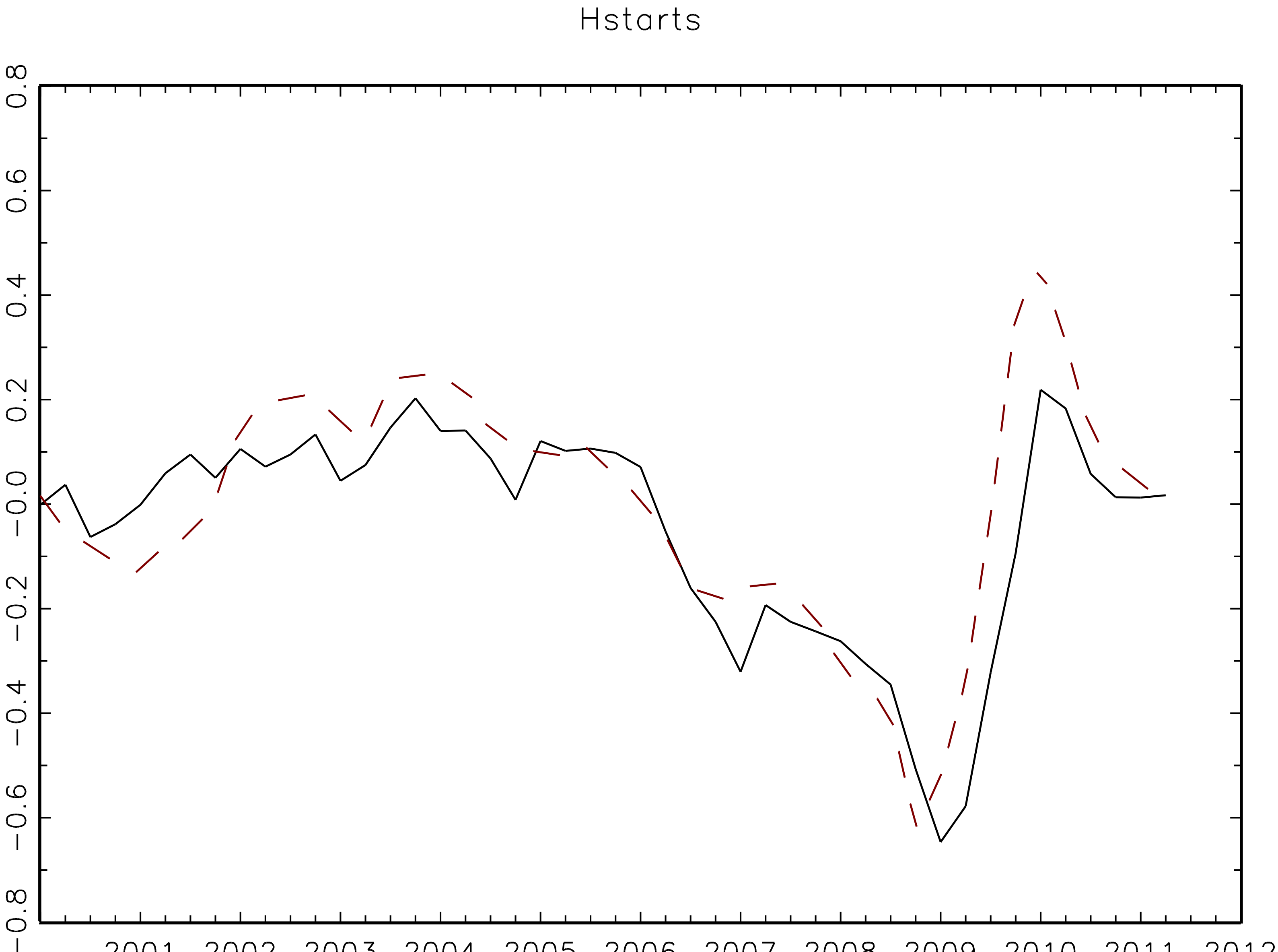
average R^2: 0.54 2007Q4 R^2: 0.62



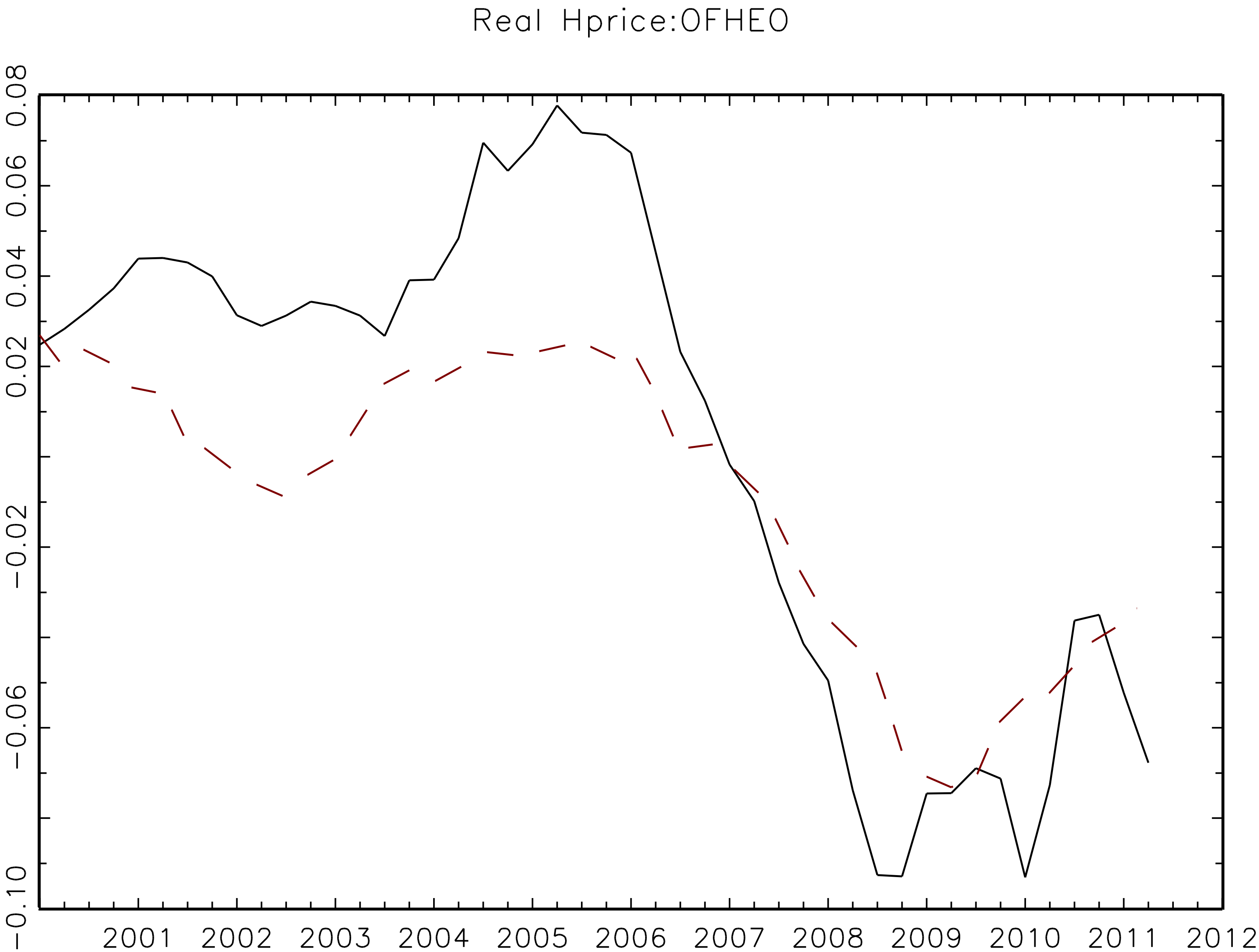
average R^2: 0.94 2007Q4 R^2: 0.96



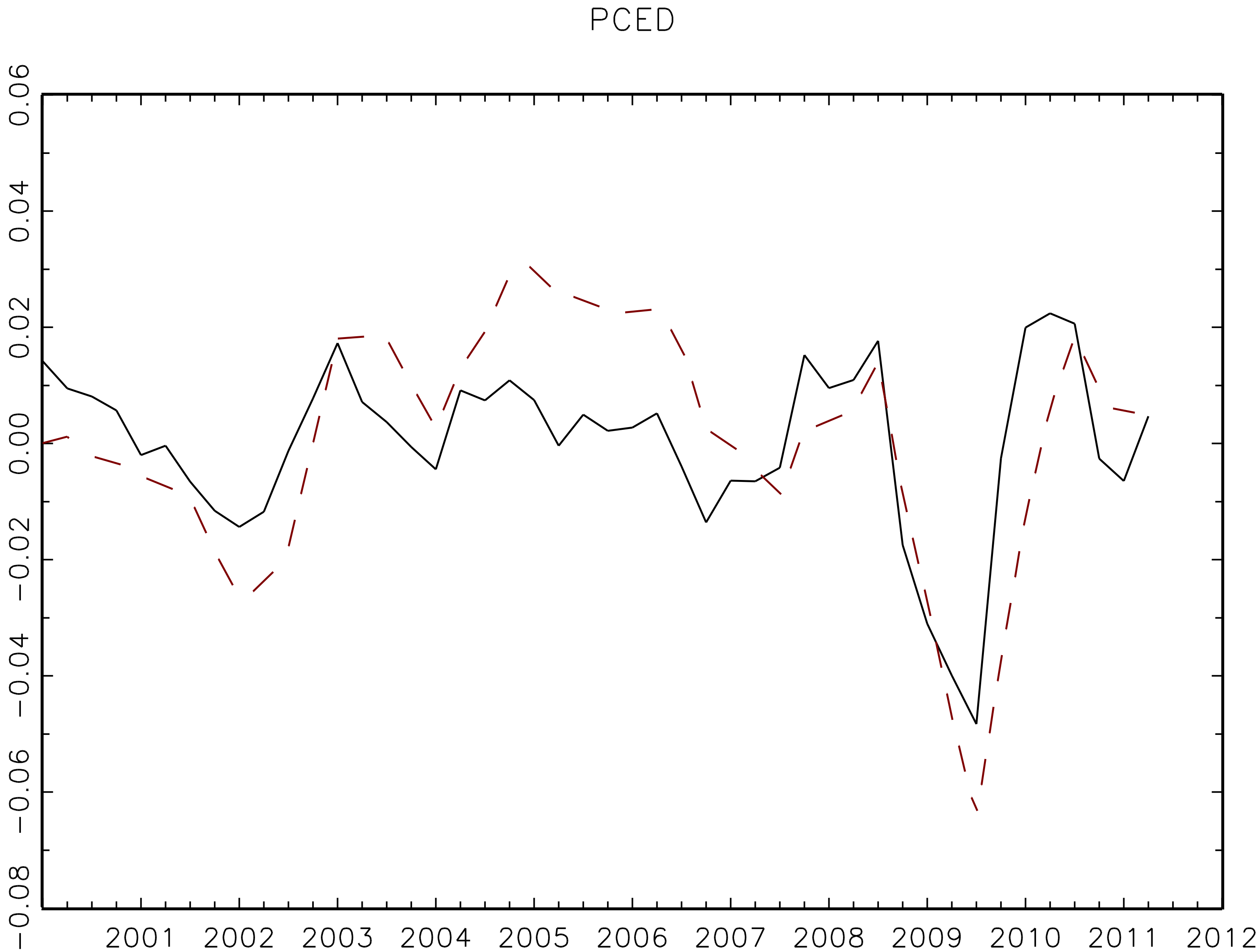
average R^2: 0.85 2007Q4 R^2: 0.89



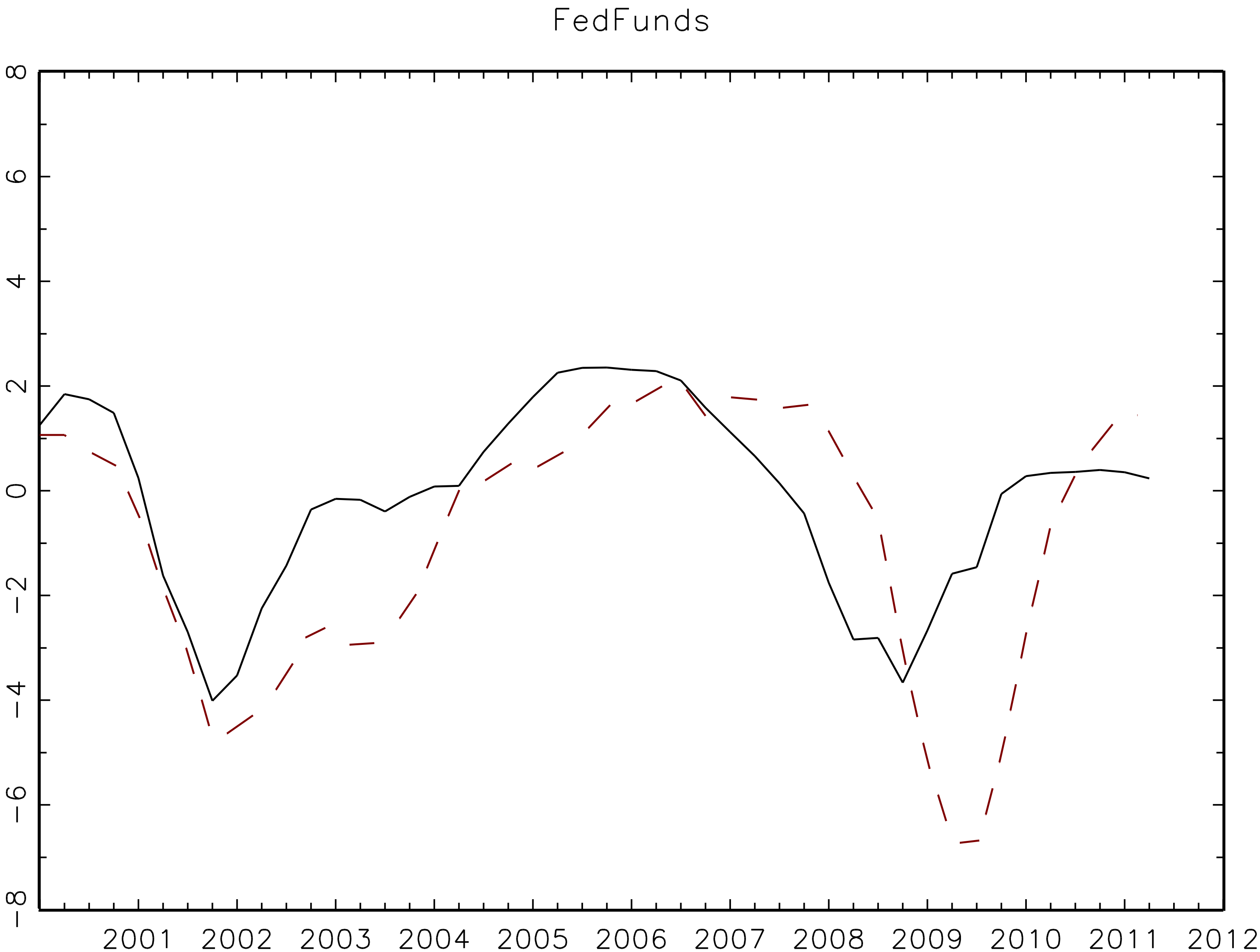
average R^2: 0.59 2007Q4 R^2: 0.53



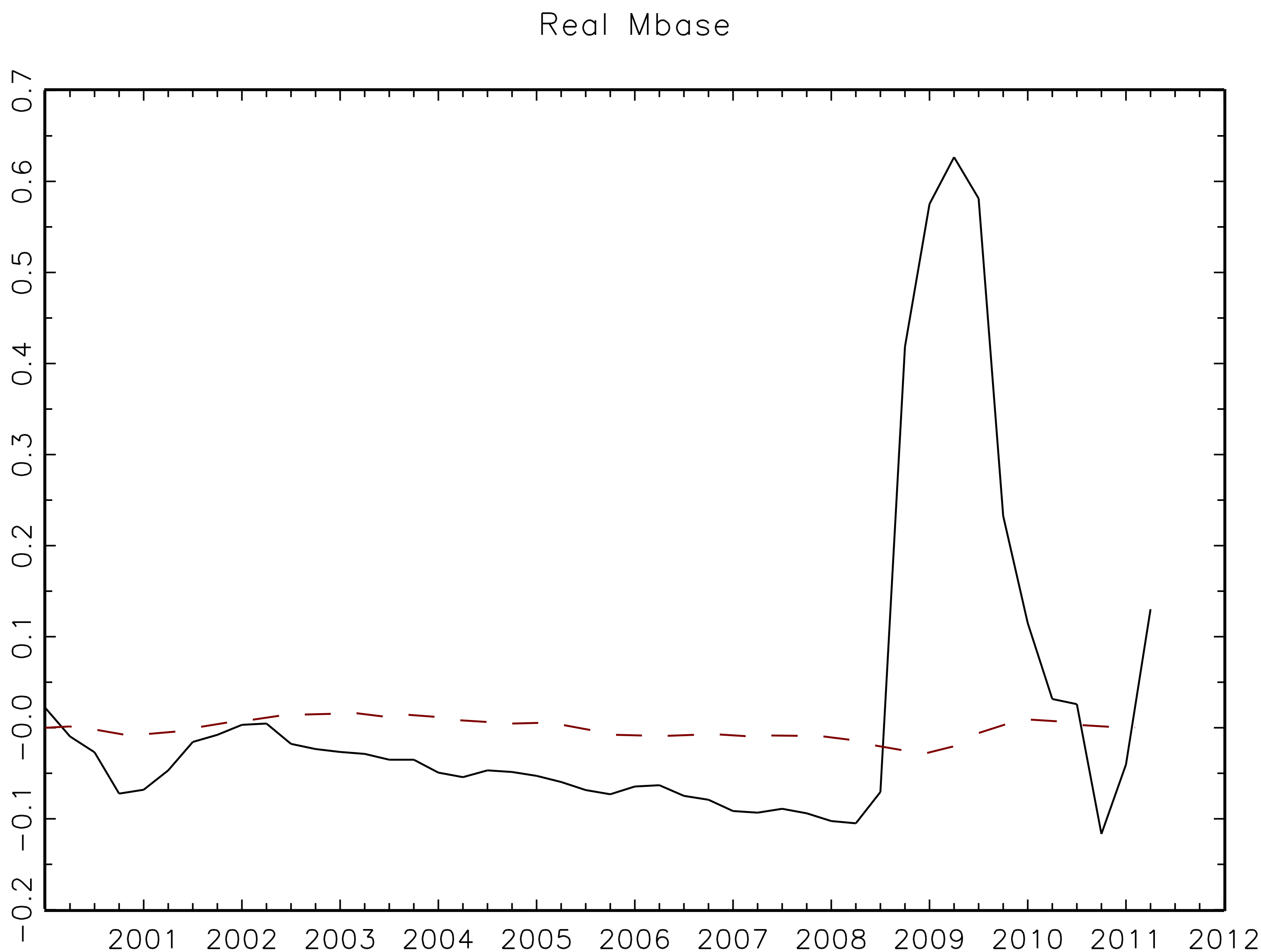
average R^2: 0.38 2007Q4 R^2: 0.55



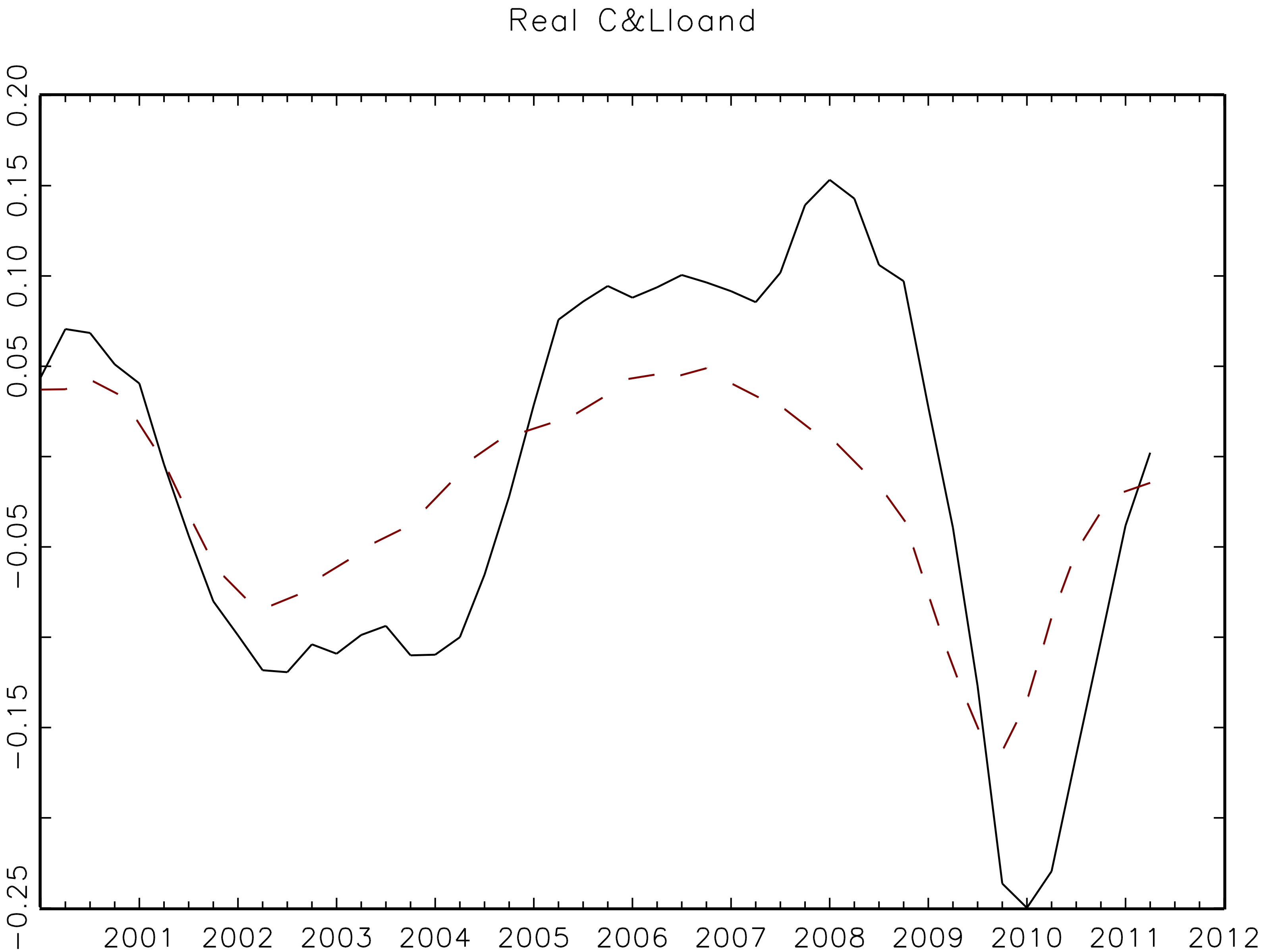
average R^2: 0.51 2007Q4 R^2: 0.82



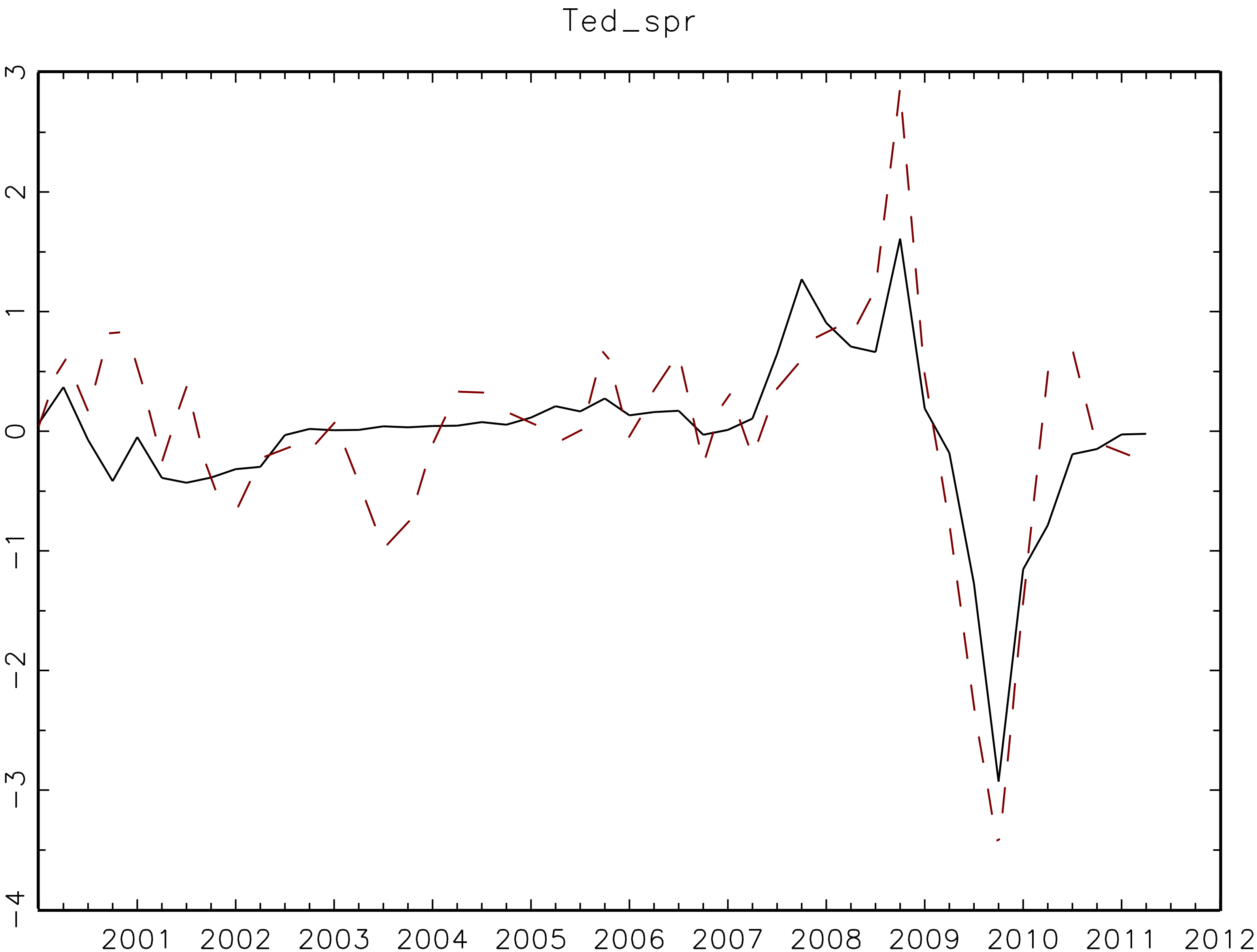
average R^2: 0.39 2007Q4 R^2: -1.54



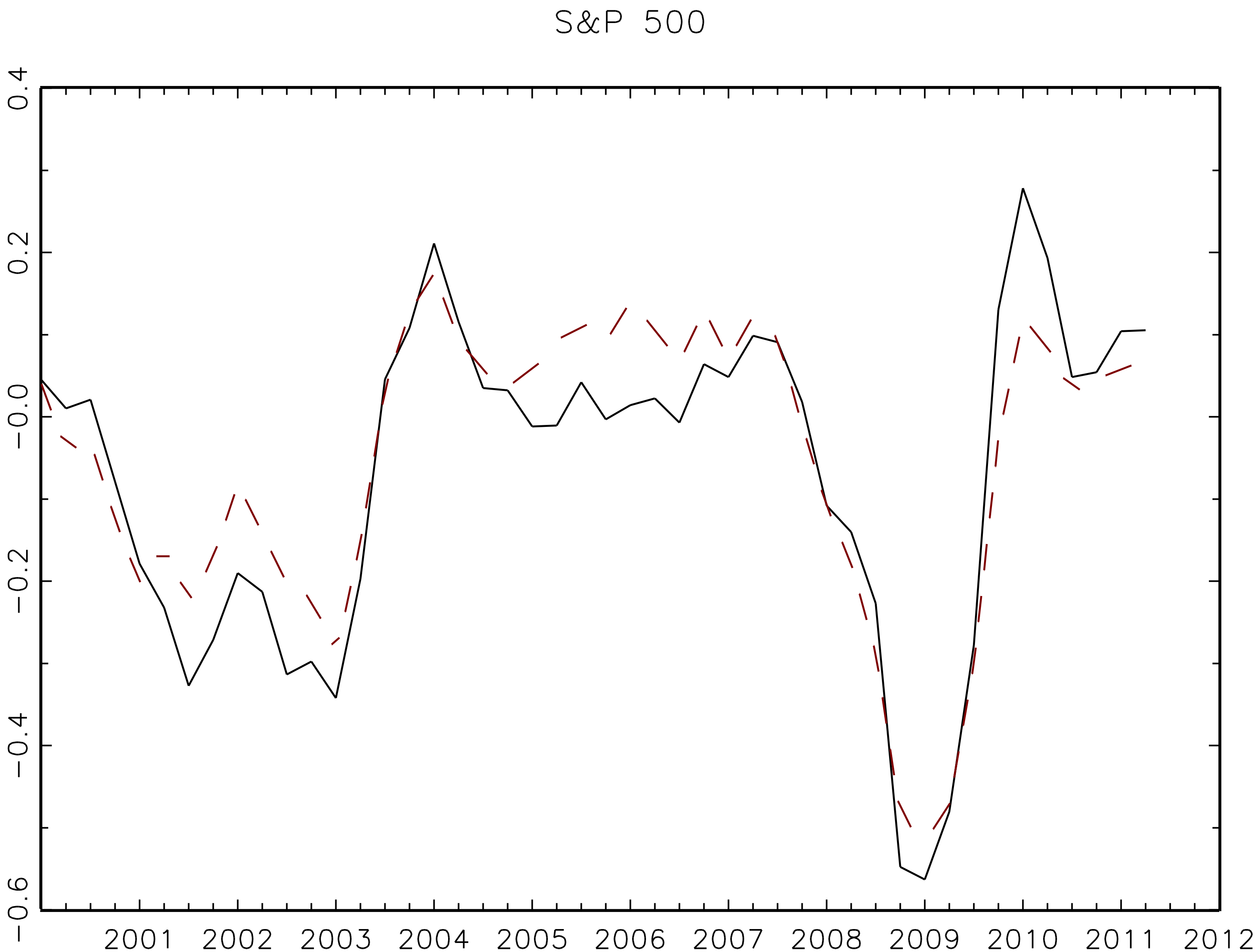
average R^2: 0.22 2007Q4 R^2: -0.03



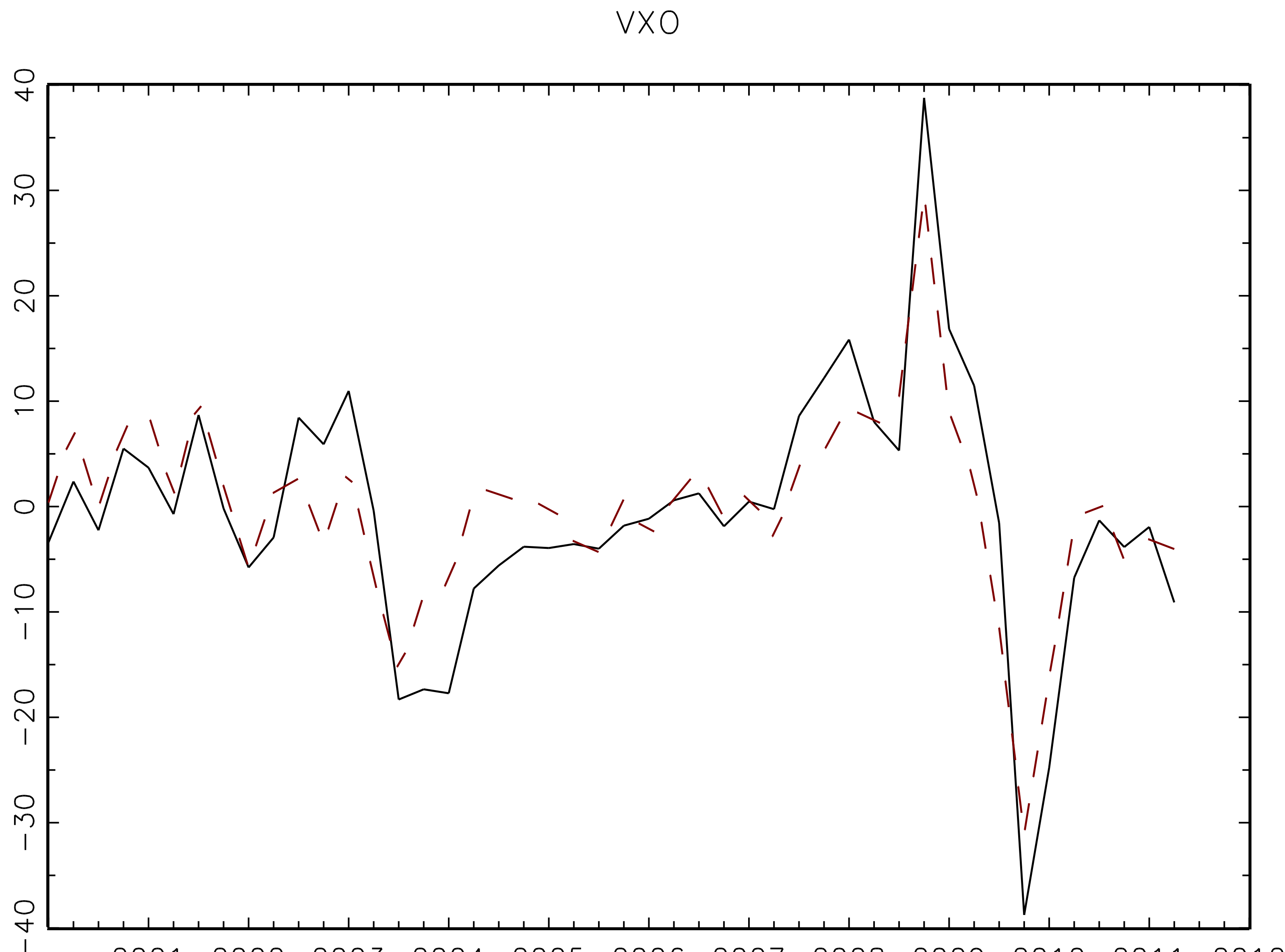
average R^2: 0.43 2007Q4 R^2: 0.78



average R^2: 0.43 2007Q4 R^2: 0.78



average R^2: 0.67 2007Q4 R^2: 0.87



average R^2: 0.12 2007Q4 R^2: 0.89

Stock and Watson (2012) “Disentangling the Channels of the 2007-09 Recession”

$$X_t = \Lambda F_t + e_t$$

$$\Phi(\mathrm{L})F_t = G\eta_t$$

Were there new factors in the 2007-09 recession? **No**

Were there instabilities in Λ? **Not much**

Were there instabilities in Φ(L)? **Not much**

Were there unusually extreme values of η and/or e? **YES**

Returning to the Prediction Problem

Forecasting setup: $y_{t+1} = F_t{}'\alpha + \varepsilon_{t+1}$

$$X_t = \Lambda F_t + e_t$$

$$\Phi(\mathrm{L})F_t = G\eta_t$$

Use X to estimate F using $\hat{F}^{PC}$.

Use $\hat{F}^{PC}$ as if they were true values of F.

Result (Stock-Watson (2002)): $\hat{y}_{T+1}\left(\hat{F}^{PC}\right) - \hat{y}_{T+1}\left(F\right) \xrightarrow{ms} 0$

Outline

~~1. Motivation and Setup~~

~~2. Dynamic Factor Models~~

3. Shrinkage

4. Sparse Models

Linear prediction problem: $y_{t+1} = x_t'\beta + \varepsilon_{t+1}$

Simpler problem: Orthonormal regressors.

Transform regressors as $p_t = Hx_t$ where H is chosen so that

$$T^{-1}\sum_{t=1}^{T} p_t p_t' = T^{-1}P'P = I_n.$$ (Note: This requires $n \le T$)

Regression equation: $y_{t+1} = p_t'\alpha + \varepsilon_{t+1}$

OLS Estimator: $\hat{\alpha} = (P'P)^{-1}P'Y = T^{-1}P'Y$

so that $\hat{\alpha}_i = T^{-1}\sum_{t=1}^{T} p_{it} y_{t+1}$

Note: Suppose p_t are strictly exogenous and $\varepsilon_t \sim iidN(0,\sigma^2)$. (This will motivate the estimators .. more discussion below).

In this simple setting:

(1) $\hat{\alpha}$ are sufficient for α.

(2) $\left(\hat{\alpha}-\alpha\right) \sim N\left(0, T^{-1}\sigma^2 I_n\right)$

(3) MSFE: $E\left(\sum_{i=1}^{n} p_{iT}(\alpha_i - \tilde{\alpha}_i)\right)^2 + \sigma^2 \approx \frac{n}{T} MSE(\tilde{\alpha}) + \sigma^2$

So we can think about analyzing n-independent normal random variables, $\hat{\alpha}_i$, to construct estimators $\tilde{\alpha}(\hat{\alpha}_i)$ that have small MSE – shrinkage can help achieve this.

Shrinkage: Basic idea

Consider two estimators: (1) $\hat{\alpha}_i \sim \mathrm{N}(\alpha_i, T^{-1}\sigma^2)$

(2) $\tilde{\alpha}_i = ½\, \hat{\alpha}_i$

$$\mathrm{MSE}(\hat{\alpha}_i) = T^{-1}\sigma^2$$

$$\mathrm{MSE}(\hat{\alpha}_i) = 0.25 \times (T^{-1}\sigma^2 + \alpha_i^2)$$

$$\mathrm{MSFE}(\hat{\alpha}) = \frac{n}{T}\sigma^2 + \sigma^2$$

$$\mathrm{MSFE}(\tilde{\alpha}) = 0.25 \times \left[\frac{n}{T}\sigma^2 + \sum_{i=1}^{n}\alpha_i^2\right] + \sigma^2$$

How big is $\sum_{i=1}^{n}\alpha_i^2$?

What is optimal amount (and form) of shrinkage?

It depends on distribution of $\{\alpha_i\}$

- Bayesian methods use priors for the distribution

- Empirical Bayes methods estimate the distribution

Examples 1: L_2 – Shrinkage

Bayes: Suppose $\alpha_i \sim \text{iidN}(0, T^{-1}\omega^2)$

Then, with $\hat{\alpha}_i | \alpha_i \sim \text{N}(\alpha_i, T^{-1}\sigma^2)$,

$$\begin{bmatrix} \alpha_i \\ \hat{\alpha}_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, T^{-1} \begin{bmatrix} \omega^2 & \omega^2 \\ \omega^2 & \sigma^2 + \omega^2 \end{bmatrix} \right)$$

so that $\alpha_i | \hat{\alpha}_i \sim N\left(\frac{\omega^2}{\sigma^2 + \omega^2} \hat{\alpha}_i, T^{-1} \frac{\omega^2 \sigma^2}{\sigma^2 + \omega^2} \right)$

MSE minimizing estimator conditional mean: $\tilde{\alpha}_i = \frac{\omega^2}{\omega^2 + \sigma^2} \hat{\alpha}_i$

Empirical Bayes: Requires estimates of σ^2 and ω^2

If $T-n$ is large, then σ^2 can be accurately estimated.

If n is large, then ω^2 can be accurately estimated:

$$\mathrm{E}(\hat{\alpha}_i^2) = T^{-1}(\sigma^2 + \omega^2), \text{ so } \hat{\omega}^2 = \frac{T}{n}\sum_{i=1}^{n}\hat{\alpha}_i^2 - \hat{\sigma}^2$$

(Extensions to more general distributions, etc. in this prediction framework – see Zhang (2005), and Knox, Stock and Watson (2004) and references therein.)

Alternative Formulation:

Write Joint density of data and α as

$$\text{constant} \times \exp\left\{-0.5\left[\frac{1}{\sigma^2}\sum_{t=1}^{T}(y_{t+1} - p_t'\alpha)^2 + \frac{1}{\omega^2}\sum_{i=1}^{n}\alpha_i^2\right]\right\}$$

Which is proportional to posterior for α. Because posterior is normal, mean = mode, so $\tilde{\alpha}$ can be found by maximizing posterior. Equivalently by solving:

$$\min_{\tilde{\alpha}} \sum_{t=1}^{T}(y_{t+1} - p_t'\tilde{\alpha})^2 + \lambda\sum_{i=1}^{n}\tilde{\alpha}_i^2 \quad \text{with } \lambda = \sigma^2/\omega^2$$

This is called “Ridge Regression”

In the original X – regressor model, the ridge estimator of

$$\tilde{\beta}^{Ridge} = \left(X'X + \lambda I_n \right)^{-1} (X'Y)$$

and λ can be determined by prior-knowledge, or estimated (empirical Bayes, cross-validation, etc.)

(Note this estimator allows $n > T$.)

Other shrinkage methods (There are many, of course, that depend on the assumed distribution of the regressions coefficients).

One of particular interest is ***Bayesian model averaging (BMA)***.

- ***References***
  - Leamer (1978); Min and Zellner (1990); Fernandez, Ley, and Steele (2001), Koop and Potter (2004)
  - Surveys: Hoeting, Madigan, Raftery, and Volinsky (1999), Geweke and Whiteman (2004)
- ***Basic idea***: there are many possible models (submodels); assign them prior probability and compute posterior means.
- ***The BMA setup*** (notation: using X_t, not P_t – this doesn't need orthogonalized regressors in theory).

$Y_{t+1} \mid X_t$ is given by one of K models, denoted by $M_1,\ldots, M_K$.

Models are linear, so M_k lists variables in model k

$\pi(M_k)$ = prior probability of model k

D_t denotes the data set through date t

The ***predictive density*** is the density of Y_{T+1} given the past data – the priors and the model are integrated out:

$$f(Y_{T+1}|D_T) = \sum_{k=1}^{K} f_k(Y_{T+1} \mid D_T)\Pr(M_k \mid D_T),$$

where $f_k(Y_{T+1}|D_T) = k^{\text{th}}$ predictive density

The ***posterior probability*** of model k is:

$$\Pr(M_k|D_T) = \frac{\Pr(D_T \mid M_k)\pi(M_k)}{\sum_{i=1}^{K}\Pr(D_T \mid M_i)\pi(M_i)},$$

where

$$\Pr(D_T|M_k) = \int \Pr(D_T \mid \theta_k, M_k)\pi(\theta_k \mid M_k)d\theta_k$$

θ_k = parameters in model k

$\pi(\theta_k|M_k)$ = prior for θ_k in model k

Under quadratic loss, optimal forecast is the mean of the predictive density, which is the weighted average of the forecasts you would make under each model, weighted by the posterior probability of that model:

$$\tilde{Y}_{T+1|T} = \sum_{k=1}^{K} \Pr(M_k \mid D_T)\tilde{Y}_{M_k,T+1|T},$$

where $\tilde{Y}_{M_k,T+1|T}$ = posterior mean of Y_{T+1} for model M_k.

Comments

- Akin to forecast combining – where there are K forecasts
- How many models are there? How many distinct subsets of 135 variables can you make?
- fun for computational Bayesians (MCMC, etc)
- This simplifies with orthogonal regressors however…
- Contrast with "Prediction Pools": Hall and Mitchel (2007), Geweke and Amisano (2011).

BMA with orthogonal regressors

Clyde, Desimone, and Parmigiani (1996), Clyde (1999):

- Variable j is in the model with probability π (coin flip)
- Given the model, the coefficients are distributed with a conjugate "*g*-prior" – and you get a closed form expression for posteriors (see Stock and Watson (2012))

More Comments:

1. Link to forecast combination – Bates and Granger (1969) … for an ambitious on-going application see Norges Bank (2014)

2. If the parameters of the prior (the "hyperparameters") are estimated, then this is parametric empirical Bayes.

3. All the theory and setup of BMA is for the cross-sectional case – the theoretical Bayes justification doesn't go through with predetermined regressors, nor for multistep forecasts. So its motivation is by analogy to to the i.i.d./exogenous regressor case.

Outline

~~1. Motivation and Setup~~

~~2. Dynamic Factor Models~~

~~3. Shrinkage~~

4. Sparse Models

Sparse models: Many/most values of β_i or α_i are zero.

Can be interpreted as shrinkage with lots of point mass at zero:

Approaches:

- BMA … (but can be computationally challenging … 2^n models): Hoeting, Madiga, Raftery, and Volinsky (1999))

- Hard thresholds (AIC/BIC) or smoothed out using "Bagging": (Breiman (1996), Bühlmann and Yu (2002); Inoue and Kilian (2008))

- L_1 penalization: Lasso ("Least Absolute Shrinkage and Selection Operator"): Tibshirani (1996)

Lasso: (With orthonormal regressors)

Ridge: $\min_{\tilde{\alpha}} \sum_{t=1}^{T} (y_{t+1} - p_t{}'\tilde{\alpha})^2 + \lambda \sum_{i=1}^{n} \tilde{\alpha}_i^2$

Lasso: $\min_{\tilde{\alpha}} \sum_{t=1}^{T} (y_{t+1} - p_t{}'\tilde{\alpha})^2 + \lambda \sum_{i=1}^{n} |\tilde{\alpha}_i|$

Equivalently: $\min_{\tilde{\alpha}} \sum_{i=1}^{n} (\hat{\alpha}_i - \tilde{\alpha}_i)^2 + \lambda \sum_{i=1}^{n} |\tilde{\alpha}_i|$

$$\min_{\tilde{\alpha}} \sum_{i=1}^{n} (\hat{\alpha}_i - \tilde{\alpha}_i)^2 + \lambda \sum_{i=1}^{n} |\tilde{\alpha}_i|$$

Notes:

- The solution yields $\text{sign}(\tilde{\alpha}_i) = \text{sign}(\hat{\alpha}_i)$

- Suppose $\hat{\alpha}_i > 0$. FOC … $2(\hat{\alpha}_i - \tilde{\alpha}_i) + \lambda = 0$
  so solution is

$$\tilde{\alpha}_i = \begin{cases} \hat{\alpha}_i - \lambda/2 \text{ if } (\hat{\alpha}_i - \lambda/2) > 0 \\ 0 \text{ otherwise} \end{cases}$$

- Similarly for $\hat{\alpha}_i < 0$.

Comments:

(1) No closed form expression for estimator with non-orthogonal X, but efficient computational procedures using LARS (Efron, Johnstone, Hastie, and Tibshirani (2002), Hastie, Tibshirani, Friedman (2009)).

(2) “Oracle” Results: Fan and Li (2001), Zhao and Yu (2006), Zou (2006), Leeb and Pötscher (2008), Bickel, Ritov, and Tsybakov (2009).

(3) Nice overview for economists and economic research: Belloni, Chernozhukov, and Hansen (2014); application to choosing “controls” Belloni, Chernozhukov, and Hansen (2014b), and instruments Belloni, Chen, Chernozhukov, and Hansen (2012).

(4) Bayes Interpretation: Park and Casella (2008)

Suppose $\alpha_i \sim$ iid with $f(\alpha_i)$ = constant × $\exp\left(-\gamma|\alpha_i|\right)$

Then posterior is

$$\text{constant} \times \exp\left\{-0.5\left[\frac{1}{\sigma^2}\sum_{t=1}^{T}(y_{t+1} - p_t{}'\alpha)^2 + 2\gamma\sum_{i=1}^{n}|\alpha_i|\right]\right\}$$

The lasso estimator (with $\lambda = 2\gamma\sigma^2$) yields the posterior mode.

But note mode $\neq$ mean for this distribution.

Outline

~~1. Motivation and Setup~~

~~2. Dynamic Factor Models~~

~~3. Shrinkage~~

~~4. Sparse Models~~

Course Topics

~~1. Time series refresher and inference tools (MW)~~

~~2. The Kalman filter, nonlinear filtering, and Markov chain monte carlo (MW)~~

~~3. Prediction with large datasets (MW)~~

4. Heteroskedasticity and autocorrelation consistent (HAC) standard errors (JS)

5. Many instruments/weak identification in IV and GMM (JS)

6. Structural VAR modeling (JS)

References for Lecture 3: Prediction with large datasets

Bai, J., and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.

Bai, J., and S. Ng (2006), "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74,1133-1150.

Bai, J., and S. Ng (2010), "Instrumental Variable Estimation in a Data Rich Environment" *Econometric Theory*, 26:6, 1577-1606.

Bates, J.M., and Clive W.J. Granger (1969), "The Combination of Forecasts," *Operations Research Quarterly* 20, 451-468.

Belloni, A. V. D. Chen, Chernozhukov, and C. Hansen (2012) "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80:6, 657-681.

Belloni, A. V. Chernozhukov, and C. Hansen (2014a) "High-Dimensional Metho,ds and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives,* 28:2, 29-50.

Belloni, A. V. Chernozhukov, and C. Hansen (2014b) "Inference on Treatment Effects after Selection Amongst High-Dimensional Controls," *Review of Economic Studies,* 81, 608-650.

Bernanke, B.S., J. Boivin, and P. Eliasz (2005), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics*, 120, 387-422.

Bickel, P.J., Y. Rotov, and A.B. Tsybakov (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics,* 37:4, 1705-1732.

Boivin, J., and M.P. Giannoni (2006), "DSGE Models in a Data-Rich Environment," NBER WP12772.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 36, 105-139.

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *Annals of Statistics*, 30, 927-961.

Chamberlain, G., and M. Rothschild (1983), "Arbitrage Factor Structure, and Mean-Variance Analysis of Large Asset Markets," *Econometrica*, 51,1281-1304.

Clyde, M. (1999a), "Bayesian Model Averaging and Model Search Strategies" (with discussion) in *Bayesian Statistics 6*, ed. by J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith, Oxford: Oxford University Press.

Clyde, M., Desimone, H., and Parmigiani, G. (1996), "Prediction via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197-1208.

Connor, G., and R.A. Korajczyk (1986), "Performance Measurement with the Arbitrage Pricing Theory," *Journal of Financial Economics*, 15, 373-394.

Council of Economic Advisors, *Economic Activity During the Government Shutdown and Debt Limit Brinksmanship*, October 2013.

Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004). "*Least angle regression,*" 32:2, 407-499.

Engle, R.F., and M.W. Watson (1981), "A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates," *Journal of the American Statistical Association*, 76, 774-781.

Fan. J. and R. Li (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348-1360.

Fernandez, C., E. Ley, and M.F.J. Steele (2001), "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381-427.

Forni, M., and L. Reichlin (1998), "Let's Get Real: A Dynamic Factor Analytical Approach to Disaggregated Business Cycle," *Review of Economic Studies*, 65, 453-474.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), "The Generalized Factor Model: Identification And Estimation," *Review of Economics and Statistics*, 82, 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004), "The Generalized Factor Model: Consistency and Rates," *Journal of Econometrics*, 119, 231-255.

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio-Economic Models*, ed. by D.J. Aigner and A.S. Goldberger, Amsterdam: North-Holland.

Geweke, J. and G. Amisano (2011): "Optimal Prediction Pools," *Journal of Econometrics*, 164, 130–141.

Geweke, J., and C. Whiteman (2006), "Bayesian Forecasting," in *The Handbook of Economic* Sargent, T.J., and C.A. Sims (1977), "Business Cycle Modeling Without Pretending to Have Too Much A-Priori Economic Theory," in *New Methods in Business Cycle Research*, ed. by C. Sims et al., Minneapolis: Federal Reserve Bank of Minneapolis.

Giannone, D., L. Reichlin, and D. Small (2008), "Nowcasting: The Real-Time Informational Content of Macroeconomic Data," *Journal of Monetary Economics*, 55, 665-676.

Hall, S. G. and J. Mitchell (2007): "Combining Density Forecasts," *International Journal of Forecasting*, 23, 1–13.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning*, 2nd Edition, New York: Springer.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14(38): 382-401.

Inoue, Atsushi, and Lutz Kilian (2008), "How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation," *Journal of the American Statistical Association*, 103, 511 - 522

Knox, T., J.H. Stock, and M.W. Watson (2001), "Empirical Bayes Forecasts of One Time Series Using Many Regressors," Technical Working Paper No. 269 (NBER).

Koop, G., and S. Potter (2004), "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *Econometrics Journal*, 7, 550-565.

Leamer, E.E. (1978), *Specification Searches*, New York: Wiley.

Min, C., and A. Zellner (1993), "Bayesian and Non-Bayesian Methods for Combining Models And Forecasts with Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56, 89-118.

Norges Bank (2014), *Models for Short-term Forecasting*, available at http://www.norges-bank.no/en/Monetary-policy/Models-for-monetary-policy-analysis-and-forecasting/SAM/ .

Park, T., and Casella, G. (2008) The Bayesian Lasso, *Journal of the American Statistical Association*, 103:681-686.

Quah, D., and T.J. Sargent (1993), "A Dynamic Index Model for Large Cross Sections" (with discussion), in *Business Cycles, Indicators, and Forecasting*, ed. by J.H. Stock and M.W. Watson, Chicago: University of Chicago Press for the NBER, 285-310.

Sargent, T.J. (1989), "Two Models of Measurements and the Investment Accelerator," *Journal of Political Economy*, 97, 251-287.

Stock, J.H., and M.W. Watson (1989), "New Indexes of Coincident and Leading Economic Indicators," *NBER Macroeconomics Annual 1989*, 351-393.

Stock, J.H., and M.W. Watson (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167-1179.

Stock, J.H., and M.W. Watson (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors,"*Journal of Business and Economic Statistics*.

Stock, J.H and M.W. Watson (2012), “Disentangling the Channels of the 2007-2009 Recession”, *Brookings Papers on Economic Activity,* Spring 2012.

Tibshirani, R. (1996). “Regression shrinkage and selection via the lasso,” *J. Royal. Statist. Soc B*., Vol. 58, No. 1, pages 267-288).

Zhang, C.-H. (2005), “General Empirical Bayes Wavelet Methods and Exactly Adaptive Minimax Estimation,” *Annals of Statistics*, 33, 54-100.

Zhao, P. and B. Yu (2006), ′On Model Selection Consistency of Lasso,” *Journal of Machine Learning*, 7, 2541-2563.

Zou, H. (2006), “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association,* 101, 1418-1429.