Topics in Bayesian Econometrics
November 2012
Central Bank of Chile
Fabio Canova

Outline

The course presents a self-contained exposition of numerical Bayesian methods applied to reduced form models, to structural VARs, to a class of state space models (including TVC models, factor models, stochastic volatility models, Markov switching models) and to DSGE models.

It is assumed that participants are familiar with the following topics: (a) Basic VAR techniques: in particular, the identification of shocks and calculation of standard errors of impulse responses; (b) Kalman Filtering techniques; (c) Current models used in dynamic macroeconomics. In addition, a working knowledge of Matlab (and Dynare) programming language is required.

The lectures are based on chapters 9 to 11 of my book: Methods for Applied Macroeconomic Research, Princeton University Press, 2007, and on additional new material.

Program

Day 1 Bayesian estimation and inference. Posterior simulators. Robustness.
Day 2: Bayesian methods for VARs models
Day 3: Bayesian methods for dynamic panel VARs, for state space and factor models.
Day 4: Bayesian methods for DSGE Models. Evaluation techniques for DSGE models

Good textbooks

- Berger, J. and Wolpert, R. (1998), *The Likelihood Principle*, Institute of Mathematical Statistics, Hayward, Ca., 2nd edition
- Bauwens, L., M. Lubrano and J.F. Richard (1999) *Bayesian Inference in Dynamics Econometric Models*, Oxford University Press.
- Robert, C. and Casella, G. (2003) Monte Carlo Statistical Methods, Springer Verlag.
- Gelman, A., J. B. Carlin, H.S. Stern and D.B. Rubin (1995), *Bayesian Data Analysis*, Chapman and Hall, London.
- Poirier, D. (1995) Intermediate Statistics and Econometrics, MIT Press.
- Kim, C. and Nelson, C. (1999), *State Space Models with Regime Switching*, MIT Press, London, UK.

- Koop, G. (2004) Bayesian Econometrics, Wiley and Sons
- Zellner, A. (1971) Introduction to Bayesian Inference in Econometrics, Wiley and Sons

1) Introduction and posterior simulators

• Preliminaries : Bayes Theorem, Prior Selection, Nuisance Parameters.

• Inference, Uncertainty, Credible Intervals, (Asymptotic) Normal Approximations, Multiple models, Testing models, Forecasting.

• Hierarchical and Empirical Bayes Models, Meta-analysis.

• Normal Approximations

• Acceptance and Importance Sampling

• MCMC methods (Gibbs sampler and Metropolis-Hastings)

• Prior Robustness

References

- Carlin B.P. and Gelfand, A.E, Smith, A.F.M (1992) Hierarchical Bayesian Analysis of change point problem, *Journal of the Royal Statistical Society, C*, 389-405.
- Canova, F. and Pappa, E. (2007) Price Differential in Monetary Union: the role of fiscal shocks, *Economic Journal*, 117, 713-737.
- Canova, F (2005) The transmission of US shocks to Latin America, *Journal of Applied Econometrics*, 20, 229-251.
- Kass, R. and Raftery, A (1995), Empirical Bayes Factors, *Journal of the American Statistical Association*, 90, 773-795.
- Sims, C. (1988) " Bayesian Skepticism on unit root econometrics", *Journal of Economic Dynamics and Control*, 12, 463-474.
- Casella, G. and George, E. (1992) Explaining the Gibbs Sampler *American Statistician*, 46, 167-174.

- Chib, S. and Greenberg, E. (1995) Understanding the Hasting-Metropolis Algorithm, *The American Statistician*, 49, 327-335.
- Chib, S. and Greenberg, E. (1996) Markov chain Monte Carlo Simulation methods in Econometrics, *Econometric Theory*, 12, 409-431.
- Geweke, J. (1995) Monte Carlo Simulation and Numerical Integration in Amman, H., Kendrick, D. and Rust, J. (eds.) *Handbook of Computational Economics* Amsterdaam, North Holland, 731-800.
- Smith, A.F.M. and Roberts, G.O, (1993), "Bayesian Computation via the Gibbs sampler and related Markov Chain Monte Carlo methods" *Journal of the Royal Statistical Society, B,* 55, 3-24.
- Tierney, L (1994) Markov Chains for Exploring Posterior Distributions (with discussion), *Annals of Statistics*, 22, 1701-1762.

2) VAR and dynamic regression models

- Likelihood function for an M variable VAR(q)
- Priors for VARs (Minnesota (Litterman), General, DSGE)
- Structural BVARs
- Bayesian dynamic panels
- Bayesian clustering

References

- Lindlay, D. V. and Smith, A.F.M. (1972) "Bayes Estimates of the Linear Model", *Journal of the Royal Statistical Association*, Ser B, 34, 1-18.
- Zellner, A., Hong, (1989) Forecasting International Growth rates using Bayesian Shrinkage and other procedures, *Journal of Econometrics*, 40, 183-202.
- Ballabriga, C. (1997) "Bayesian Vector Autoregressions", manuscript.
- Canova, F. (1992) " An Alternative Approach to Modelling and Forecasting Seasonal Time Series " *Journal of Business and Economic Statistics*, 10, 97-108.

- Canova, F. (1993a) " Forecasting time series with common seasonal patterns", *Journal of Econometrics*, 55, 173-200.

- Del Negro, M. and F. Schorfheide (2004), " Priors from General Equilibrium Models for VARs", *International economic Review*, 45, 643-673.

- Ingram, B. and Whitemann, C. (1994), "Supplanting the Minnesota prior. Forecasting macroeconomic time series using real business cycle priors, *Journal of Monetary Economics*, 34, 497-510.

- Kadiyala, R. and Karlsson, S. (1997) Numerical methods for estimation and Inference in Bayesian VAR models, *Journal of Applied Econometrics*, 12, 99-132.

- Koop, G.(1996) "Bayesian Impulse responses", *Journal of Econometrics*, 74, 119-147.

- Sims, C. and Zha T. (1998) "Bayesian Methods for Dynamic Multivariate Models", *International Economic Review*, 39, 949-968.

- Waggoner and T. Zha (2003) A Gibbs Simulator for Restricted VAR models, *Journal of Economic Dynamics and Control*, 26, 349-366.

- Zha, T. (1999) "Block Recursion and Structural Vector Autoregressions", *Journal of Econometrics*, 90, 291-316.

- Marcet, A. and M Jarocinski (2010) Autoregressions in small samples, prior about observables and initial conditions. UAB manuscript.

- Canova, F. (2004) Testing for Convergence Club: A Predictive Density Approach, *International Economic Review, 45,49-77.*

3) Bayesian Time series models

- State Space Models and Kalman filter. Classical Inference in state space models

- Gibbs sampler for state space models

- Applications: TVC- VARs, Factor models, Stochastic volatility, Markov switching models

References

- Albert, J. and Chib, S. (1993) Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts, *Journal of Business and Economic Statistics*, 11, 1-16.

- Chib, S. (1996) Calculating Posterior Distributions and Model Estimates in Markov Mixture Models, *Journal of Econometrics*, 75, 79-98.

- Fruhwirth-Schnatter, S (2001) MCMC estimation of classical and Dynamic switching and Mixture Models *Journal of the American Statistical Association*, 96, 194-209.

- Geweke, J. and Zhou, G. (1996) Measuring the Pricing Error of the Arbitrage Pricing Theory, *Review of Financial Studies*, 9, 557-587.

- Otrok, C. and Whitemann, C. (1998), "Bayesian Leading Indicators: measuring and Predicting Economic Conditions in Iowa", *International Economic Review*, 39, 997-1114.

- Jacquier, E., Polson N. and Rossi, P. (1994), " Bayesian Analysis of Stochastic Volatility Models", *Journal of Business and Economic Statistics*, 12, 371-417.

- McCulloch, R. and R. Tsay (1994) Statistical Analysis of Economic Time Series via Markov Switching Models", *Journal of Time Series Analysis*, 15, 521-539.

- Sims, C. and Zha, T. (2006) Were there regime switches in US monetary policy, *American Economic Review*, 96(1), 54-81.

- Sims, C. D. Waggoner, T. Zha (2008), Methods for Inference in Large Multiple-Equation Markov-Switching Models, *Journal of Econometrics*, 146(2) 255-274.

- Cogley, T. and Sargent, T. (2005) Drifts and Breaks in US Inflation, *Review of Economic Dynamics*, 8, 262-302.

- Cogley, T., Morozov, and Sargent, T. (2005) Bayesian fan charts for UK inflation: Forecasting and sources of uncertainty in evolving monetary systems, *Journal of Economic Dynamics and Control*, 29, 1893-1925.

- Canova, F. and Gambetti, L. (2009) Structural Changes in the US economy: is there a role for monetary policy? *Journal of Economic Dynamics and Control*, 33, 477-490.

- Canova, F. and Ciccarelli, M., (2004), "Forecasting and Turning Point Prediction in a Bayesian Panel VAR model", *Journal of Econometrics*, 120, 327-359.

- Canova, F. and Ciccarelli, M., (2009), "Estimating multicountry VAR models", *International Economic Review*, 50, 929-961.

- Benati, L. (2008) The great moderation in the UK: Good luck or good policy?, Journal of Money Credit and Banking, 40, 121-147.

- Carlin, B., Polson, N. and Stoffer, D. (1992)"A Monte Carlo Approach to nonnormal and nonlinear state-space modelling", *Journal of the American Statistical Association*, 87, 493-500

- Mumtaz, H. and Surico, P. (2009) Evolving International Inflation Dynamics: Evidence from a time varying Dynamic Factor Model, forthcoming, Journal of the European Economic Association

- Gambetti, L., Pappa, E. and Canova, F. (2008) The structural dynamics of Output and Inflation: what explains the changes?, *Journal of Money, Credit and Banking*, 40, 369-388.

- Canova, F., Ciccarelli, M. and Ortega, E. (2007), "Similarities and Convergence in G-7 Cycles", *Journal of Monetary Economics*, 54, 850-878.

- Canova, F. Ciccarelli, M. and Ortega, E. (2012), Do Institutional Changes affect Business Cycles? Evidence from Europe, Journal of Economic Dynamics and Control, 36, 1520-1533.

- Del Negro, M. and Schorfeide, F., (2010) Bayesian Macroeconometrics, in J. Geweke, G. Koop, and H. van Dijk (eds.) Handbook of Bayesian Econometrics.

4) Bayesian DSGE models

- Algorithms and examples.

- Prior elicitation and data-rich DSGE.

- Misspecified DSGE.

- Identification problems in DSGE.

- Evaluating DSGE models.

References

- An, S and Schorfheide, F. ,2007, Bayesian analysis of DSGE models, *Econometric Reviews*, 26, 113-172 (with discussion).

- Schorfheide, F, 2011 Estimation and Evaluation of DSGE models: Progress and challenges, NBER working paper 16781.

- Driffill, J, Pesaran, H. Smith, R. G. Ascari, M. Miller, R. Werner (2011) The future of macroeconomics, Manchester Journal, supplement, 1-38. (4 articles and an introduction).

- Fernandez Villaverde, J., 2009, The econometrics of DSGE models, NBER working paper 14677.
- Primiceri, G. and Justianiano, A., 2008, The time varying volatility of Macroeconomic Fluctuations, *American Economic Review*, 98, 604-641.
- Smets, F. and R. Wouters, 2003, An Estimated Stochastic DSGE model of the Euro Area, *Journal of the European Economic Association*, 5, 1123-1175.
- Smets, F. and R. Wouters, 2007, Shocks and Frictions in US Business cycles, *American Economic Review*, 97, 586-606.
- Schorfheide, F., 2000 Loss function based evaluation of DSGE models, *Journal of Applied Econometrics*, 15, 645-670.
- Adolfson, M, Laseen, S., Linde, J. and Villani, M., 2008, Evaluating an Estimated new Keynesian small open economy model, *Journal of Economic Dynamics and Control*, 32, 2690-2721.
- Canova, F. and Sala, L., 2009, Back to square one: Identification issues in DSGE models, *Journal of Monetary Economics*, 56(4), 431-449.
- Del Negro, M, Schorfheide, F., Smets, F. and Wouters, R., 2006, On the fit of New-keynesian models, *Journal of Business and Economic Statistics*, 25, 143-162.
- Iskrev, N., 2010, Local identification in DSGE models, *Journal of Monetary Economics*, 57, 189-202.
- Canova, F. and Paustian, M., 2011, Business cycle measurement with some theory, *Journal of Monetary Economics*, 48, 365-381.
- Chari, V., Kehoe, P. and McGrattan, E., 2009, "New Keynesian models: not yet useful for policy analysis, *American Economic Journal: Macroeconomics*, 1, 242-266.
- Canova, F., 2010, "Bridging DSGE models and the data", available at http://www.crei.cat/people/canova.
- Canova, F., and Ferroni, F., 2011, "Multiple filtering devices for the estimation of DSGE models", *Quantitative Economics*, 2, 73-98.
- Canova, F. , Ferroni, F. and Matthes, C. (2012) Choosing the variables to estimate singular DSGE models, EUI manuscript.

Identification issues in DSGE models

Fabio Canova
EUI and CEPR

October 2012

References

Canova, F. and Paustian, M (2011), Business cycle measurement with some theory, *Journal of Monetary Economics*, 48, 345-361.

Canova, F. (2009), How much structure in empirical models, T. Mills and K. Patterson (eds.), *Palgrave Handbook of Econometrics, volume 2*, 30-65.

Canova, F. (2009), Comment to Weak Instruments Robust tests in GMM and the New Keynesian Phillips curve, by Frank Kleibergen and Sophocles Mavroeidis, *Journal of Business and Economic Statistics*,27, 311-315.

Canova, F. and Sala, L. (2009) Back to square one: identification issues in DSGE models, *Journal of Monetary Economics*, 56(4), 431-449.

Chari, V, Kehoe, P. and McGrattan, E. (2007) Business cycle accounting, *Econometrica*, 75, 781-836.

Chari, V., Kehoe, P. and McGratten, E. (2009) "New Keynesian models: not yet useful for policy analysis, *American Economic Journal: Macroeconomics*, 1, 242-266.

Komunjer, I. and Ng, S. (2011) Dynamic Identification of DSGE models, *Econometrica*, 79, 1995-2032.

Koop, G. , H. Pesaran, R. Smith (2011) On the Bayesian identification of DSGE Models, Cambridge University, manuscript.

Iskrev,N. (2010) Local Identification in DSGE Models, *Journal of Monetary Economics*, 57, 189-202.

Mueller, U. (2010), Measuing Prior sensitivity and prior informativeness in large Bayesian models, Princeton University, manuscript.

Adolfson, M. and Linde', J. (2007), Parameter identification in an estimated New Keynesian Open Economy model, Riksbank, manuscript.

Qu, Z and Tkachenko, D. (2012) Identification and frequency domain maximum likelihood estimation of linearized dynamics stochastic general equilibrium models, Quantitative Economics, 3, 95-112.

Rubio, J., Waggonner, D, and Zha, T (2010) Structural vector autoregressions: theory of identification and algorithm for inference, Review of Economic Studies, 77, 665-696.

DSGE models:

$$E_t[A(\theta)x_{t+1} + B(\theta)x_t + C(\theta)x_{t-1} + D(\theta)z_{t+1} + F(\theta)z_t] = 0 \quad (1)$$
$$z_{t+1} - G(\theta)z_t - e_t = 0 \quad (2)$$

Stationary (log-linearized) RE solution:

$$x_t = J(\theta)x_{t-1} + K(\theta)e_t \quad (3)$$

• Solution is a restricted, singular VAR(1) or state space model (if some x_t are non-observable).

How are DSGE estimated/evaluated?

1. Limited information methods

i. GMM/ Indirect Inference (matching impulse responses) (CEE (2005).

ii. SVAR (magnitude and sign restrictions (Canova and Paustian (2011)).

2. Full Information methods:

i. Maximum Likelihood

ii. Bayesian methods

3. Business cycle accounting/calibration Chari et. al. (2007), (2009).

Matching impulse responses (conditional on some shock j):

Model responses: $X_t^M(\theta) = C(\theta)(\ell)e_t^j$
Data responses: $X_t = \hat{W}(\ell)e_t^j$ (after shock identification).

$\hat{\theta} = \underset{\theta}{argmin}\ g(\theta) = ||X_t - X_t^M(\theta)||_{W(T)}$, $W(T)$ weighting matrix.

ML: $\hat{\theta} = \underset{\theta}{argmax} \ln L(X, \theta)$ where $L(X, \theta)$ is computed using (3) and the normality of e_t.

Bayesian: $\hat{\theta} = \int \theta g(\theta|X)d\theta$ or
$\tilde{\theta} = \underset{\theta}{argmax}\, L(X, \theta)g(\theta)$ (constrained maximum likelihood) where $g(\theta)$ is the prior and $g(\theta|y)$ is the posterior.

Prior to estimation: can we recover structural parameters from the data?

- **Identifiability:** Mapping from objective function to the parameters needs to be well behaved.

To do this need:

- Objective function to have a unique minimum at $\theta = \theta_0$
- Hessian is positive definite and has full rank
- Curvature of the objective function is "sufficient"

Difficult to verify if these conditions hold in practice because:

- Mapping from structural parameters to solution coefficients is unknown (numerical solution).

- Objective function is typically nonlinear function of solution parameters.

- Different objective functions may have different "identification power" .

Standard rank and order conditions for linear models can't be used here.

Definitions

• i) Solution identification: can we recover structural θ from the aggregate decision rule matrices $J(\theta), K(\theta), G(\theta)$?

• ii) Objective function identification: can we recover aggregate decision rule matrices $J(\theta), K(\theta), G(\theta)$ from the objective function?

• iii) Population identification (convoluting i) and ii)): can we recover the structural parameters from the objective function in population?

• iv) Sample identification: can we recover structural parameters from the objective function, given a sample of data?

Note:

- i) and ii) can occur separately or in conjunction.

- i) is due to the model specification, ii) may result from an improper choice of objective functions.

- iv) may occur even if i) and ii) are fine.

- iv) object of much econometric literature. Here focus on i) and ii). **Problems with DSGE models are in i)-ii).**

What kind of population problems may DSGE models encounter?

• Observational equivalence. Two models may have the same (minimized) value of the objective function at two different vector of parameters (e.g. a sticky price and a sticky wage model).

• Observational equivalence within a model. Two vectors of parameters may give the same (minimized) value of the objective function, given a model (e.g. given a sticky price model, get the same likelihood if Calvo parameter is 0.25 or 0.75).

• Partial/under identification within a model. A subset of the structural parameter enters in a particular functional form in the solution/ may disappear from the solution.

• Weak/asymmetric identification within a model. The population mapping is very flat or asymmetric in some dimension.

• Limited Information identification. A subset of the parameters of the model can't be identified because the objective function uses only a portion of the restrictions of the solution.

- Problems may be local or global.

- First four issues refer to solution identification problems. The last to objective function identification.

1: Observational equivalence

1.1) Linear (log-linearized) RE forward looking models:

$$B(\theta)x_t = A(\theta)E_t x_{t+1} + e_t \tag{4}$$

where $e_t \sim (0, \Omega)$. Assume that B is non-singular.

• Solution is $x_t = \sum_{j=0}^{\infty} Q^j B^{-1} E_t e_{t+j}$ where $Q = B^{-1}A$. Since $E_t e_{t+j} = 0$, $E_t x_{t+1} = 0$, the unique RE equilibrium is $x_t = B^{-1}e_t$.

i) Model is observationally equivalent to a model with no dynamics, i.e. to a model of the type $y_t = Me_t$, where $M = B^{-1}$.

ii) Model is observationally equivalent to a model where the structural shocks are linear combination of the original structural shocks, i.e. $y_t = u_t$ where $u_t = B^{-1}e_t$.

iii) Model is observationally equivalent to a model with higher order degree of forward lookingness, i.e. $B(\theta)y_t = A(\theta)E_t y_{t+n} + e_t, n > 1$ or to a model with more complicated forward looking dynamics, e.g. $B(\theta)y_t = \sum_{n=1}^{p} A_n(\theta)E_t y_{t+n} + e_t$.

1.2) Linear RE forward and backward looking models

$$B(\theta)x_t = A(\theta)E_t x_{t+1} + Cx_{t-1} + e_t \tag{5}$$

where $e_t \sim (0, \Omega)$. Still maintain that B is non-singular.

• The solution is $x_t = Dx_{t-1} + B^{-1}e_t$ where D solves $AD^2 - BD + C = 0$.

• The solution is unique and stationary if all the eigenvalues of D and of $(B - AD)^{-1}A$ are all less than one in absolute value.

iv) The solution of the model is observationally equivalent to the one of the model with just backward looking dynamics.

Note: the parameter space may not be variational free, e.g. there may be restrictions on the parameter space $(A(\theta) + C(\theta) = 1)$ and restrictions due to eigenvalues constraints.

Example 1 *Consider the three processes ($\lambda_2 \geq 1 \geq \lambda_1 \geq 0$):*

1) $x_t = \frac{1}{\lambda_2+\lambda_1} E_t x_{t+1} + \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} x_{t-1} + v_t.$

2) $y_t = \lambda_1 y_{t-1} + w_t$

3) $y_t = \frac{1}{\lambda_1} E_t y_{t+1}$ *where* $y_{t+1} = E_t y_{t+1} + w_t$ *and* $w_t \; iid \; (0, \sigma_w^2)$.

Stable RE solution of 1) $x_t = \lambda_1 x_{t-1} + \frac{\lambda_2+\lambda_1}{\lambda_2} v_t.$

Stable RE solution of 3) is $y_t = \lambda_1 y_{t-1} + w_t.$

If $\sigma_w = \frac{\lambda_2+\lambda_1}{\lambda_2}\sigma_v$, *three processes have same impulse responses.*

- Bayer and Farmer (2004): $Ax_t + DE_t x_{t+1} = B_1 x_{t-1} + B_2 E_{t-1} x_t + Cv_t.$

- Kim (2001, JEDC); Ma (2002, EL); Lubik and Schoefheide (2004,AER) An and Schorfheide (2007,ER).

2: Underidentification

Example 2

$$
\begin{aligned}
R_t &= \psi\pi_t + e_{1t} && (6)\\
y_t &= E_t y_{t+1} - \sigma(R_t - E_t\pi_{t+1}) + e_{2t} && (7)\\
\pi_t &= \beta E_t\pi_{t+1} + \gamma y_t + e_{3t} && (8)
\end{aligned}
$$

Here $B = \begin{bmatrix} 1 & 0 & -\psi \\ \sigma & 1 & 0 \\ 0 & -\gamma & 1 \end{bmatrix}, A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & \sigma \\ 0 & 0 & \beta \end{bmatrix}$ *and* $Q = \frac{1}{\gamma\sigma\psi}\begin{pmatrix} 0 & \gamma\psi & \psi(\beta+\gamma\sigma) \\ 0 & 1 & \sigma(1-\beta\psi) \\ 0 & \gamma & \gamma\sigma \end{pmatrix}$.

The two nonzero eigenvalues of Q *are* $\kappa_i = \frac{(1+\beta+\gamma\sigma)\pm\Phi}{2(\gamma\sigma\psi+1)}, i = 1, 2$, *where* $\Phi = (\beta^2 - 2\beta + \gamma^2\sigma^2 + 2\gamma\sigma + 2\gamma\sigma\beta - 4\gamma\sigma\beta\psi + 1)^{0.5}$. *If* $\kappa_i < 1, \forall i$, *the solution is*

$$
\begin{aligned}
R_t &= \psi\pi_t + e_{1t} && (9)\\
y_t &= -\sigma R_t + e_{2t} && (10)\\
\pi_t &= \gamma y_t + e_{3t} && (11)
\end{aligned}
$$

• *β is not identifiable (it only appears in A_1, and this does not enter the likelihood function).*

• *Since the solution is valid for $\kappa_i < 1$, the formula for eigenvalues implies restrictions on all four parameters of the model. Thus, there are implicit restrictions in the parameter space: to keep $\kappa_i < 1$ as γ, ψ, σ vary, β needs to be correspondingly adjusted.*

- Even if β is calibrated, not all parameters are separately identifiable.

- Because of the stability restrictions, the posterior for β may be updated even if the likelihood is independent of β (see later).

Example 3 *Consider a version of the previous model*

$$R_t = \psi E_t \pi_{t+1} + e_{1t} \quad (12)$$
$$y_t = \delta E_t y_{t+1} - \sigma(R_t - E_t \pi_{t+1}) + e_{2t} \quad (13)$$
$$\pi_t = \beta E_t \pi_{t+1} + \gamma y_t + e_{3t} \quad (14)$$

The solution can we written (in MA format) as $x_t = \begin{bmatrix} 1 & 0 & 0 \\ \sigma & 1 & 0 \\ \sigma\gamma & \sigma & 1 \end{bmatrix} e_t$.

• δ, ψ, β *disapper from the solution; they are underidentified (we need a model with backward and forawrd looking dynamics to identify them).*

• *Different impulse responses have different "identification" information. Limited and full information objective functions have different information content. How do we maximize the identification information?*

• *Identification may be "local", i.e. it depends on the values* γ *and* σ.

Example 4 *Consider the partial equilibrium NK Phillips relations*

$$\pi_t = \omega\delta^{-1}\pi_{t-1} + \beta\alpha\delta^{-1}E_{t-1}\pi_{t+1} + \gamma x_t + e_{1t} \quad (15)$$

$$x_t = \rho x_{t-1} + e_{2t} \quad (16)$$

where $\delta = \alpha + \omega(1-\alpha(1-\beta))$, $1-\omega$ *fraction of agents chosing prices optimally among those allowed to change prices,* ω *fractions of firms using rules of thumb to set prices,* $1-\alpha$ *fraction of firms resetting prices in every period,* β *the discount factor,* $\gamma = (1-\omega)(1-\alpha)(1-\lambda\alpha)\delta^{-1} > 0$. *(WHAT is* λ *????).*

- Theory restricts the sum of the coefficient (backward plus forward part) to be equal to 1. The solution is

$$\pi_t = \kappa_b\pi_{t-1} + \frac{\gamma}{1-\kappa_b\omega\delta^{-1}}\sum_j(\kappa_f)^{-1}E_{t-1}x_{t+j} + \gamma(x_t - E_{t-1}x_t) + e_{1t} \quad (17)$$

where κ_b, κ_f *solve* $\beta\alpha\delta^{-1}\kappa^2 - \kappa + \omega\delta^{-1} = 0$. *The solution for the whole system is a VAR(1) of the form*

$$\pi_t = \zeta_1(\theta)\pi_{t-1} + \zeta_2(\theta)x_{t-1} + u_t \quad (18)$$

$$x_t = \rho x_{t-1} + e_{2t} \quad (19)$$

Here $\theta = (\beta, \alpha, \omega, \rho)$, $u_t = e_{1t} + \gamma e_{2t}$, $\zeta_1 = \frac{1-(1-4b_f b_b)^{0.5}}{2b_f}$, $\zeta_2 = \frac{\gamma\rho}{1-\beta_f(\kappa+\rho)}$, *where* $b_f = \beta\alpha\delta^{-1}$, $b_b = \omega\delta^{-1}$.

1) Can not separately identify β, α, ω. *At best we can identify* b_b, b_f.

2) Three reduced form parameters (ζ_1, ζ_2, ρ) *and four structural parameters* $\beta, \alpha, \omega, \rho$: *can not seperately identify the structrual parameters*

3) We can identify the four parameters independently if the process for x_t *is at least an AR(2) (see Mavroedis (2005)).However, if* ρ_2 *is small, identification can be weak (see later).*

4) Theory imposes restrictions on $b_b + b_f$*. Thus even if they can not separately identifed, their posterior distribution can be updated relative to the prior, even if they priors are independent.*

3: Weak and partial identification

$$\max \beta^t \sum_t \frac{c_t^{1-\phi}}{1-\phi}$$

$$c_t + k_{t+1} = k_t^{\eta} z_t + (1-\delta)k_t$$

R.E. solution for $w_{t+1} = [k_{t+1}, c_t, y_t, z_t] = Aw_t + Be_t$.

Select $\beta = 0.985, \phi = 2.0, \rho = 0.95, \eta = 0.36, \delta = 0.025, z^{ss} = 1$.

Strategy: simulate data, compute population objective function. Study its shape and features.

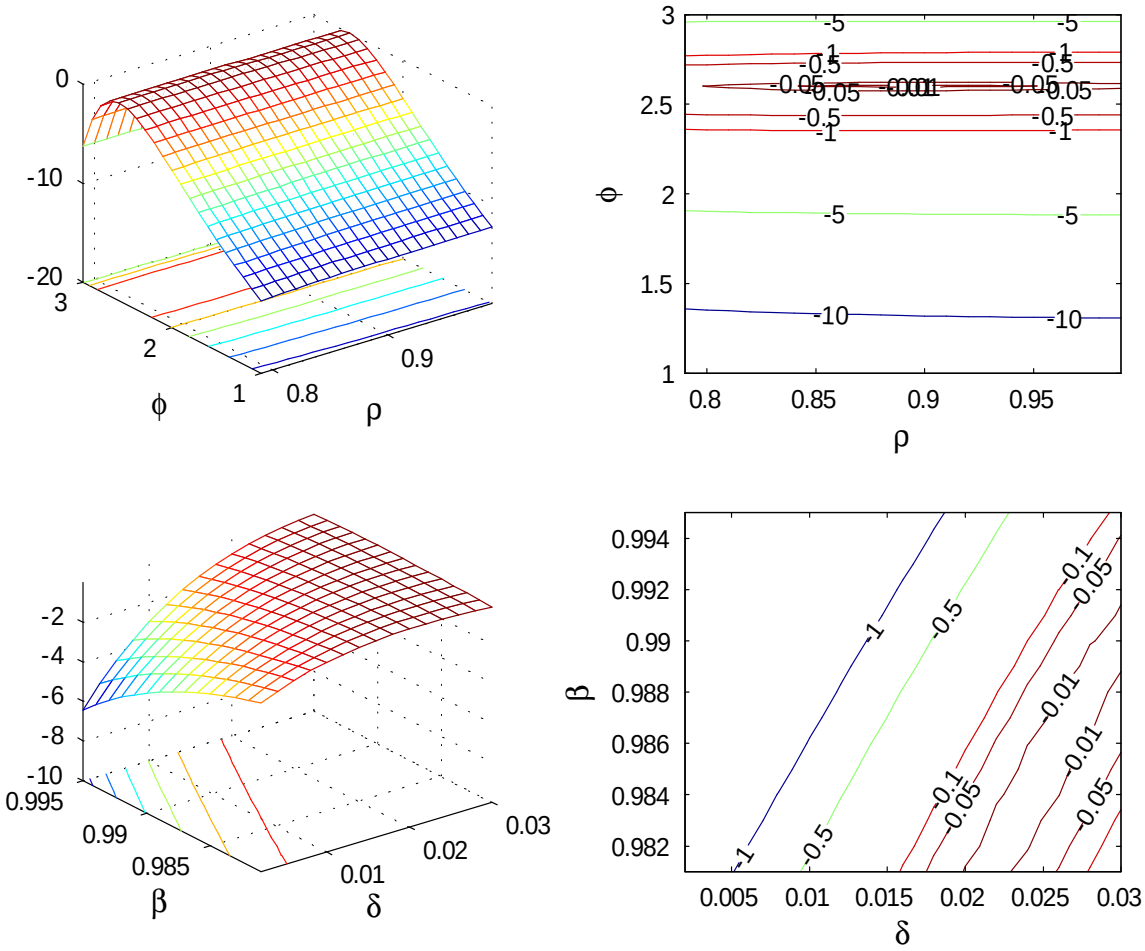


Figure 1: Distance surface for selected parameters

What causes the problems? Law of motion of the capital stock in almost invariant to :

(a) variations of η and ρ (weak identification).

(b) variations of β and δ are additive (partial under-identification).

Can we reduce problems by:

(i) Changing $W(T)$? (long horizon may have little information).

(ii) Matching VAR coefficients?

(iii) Altering the objective function?

In this specific case: NO.

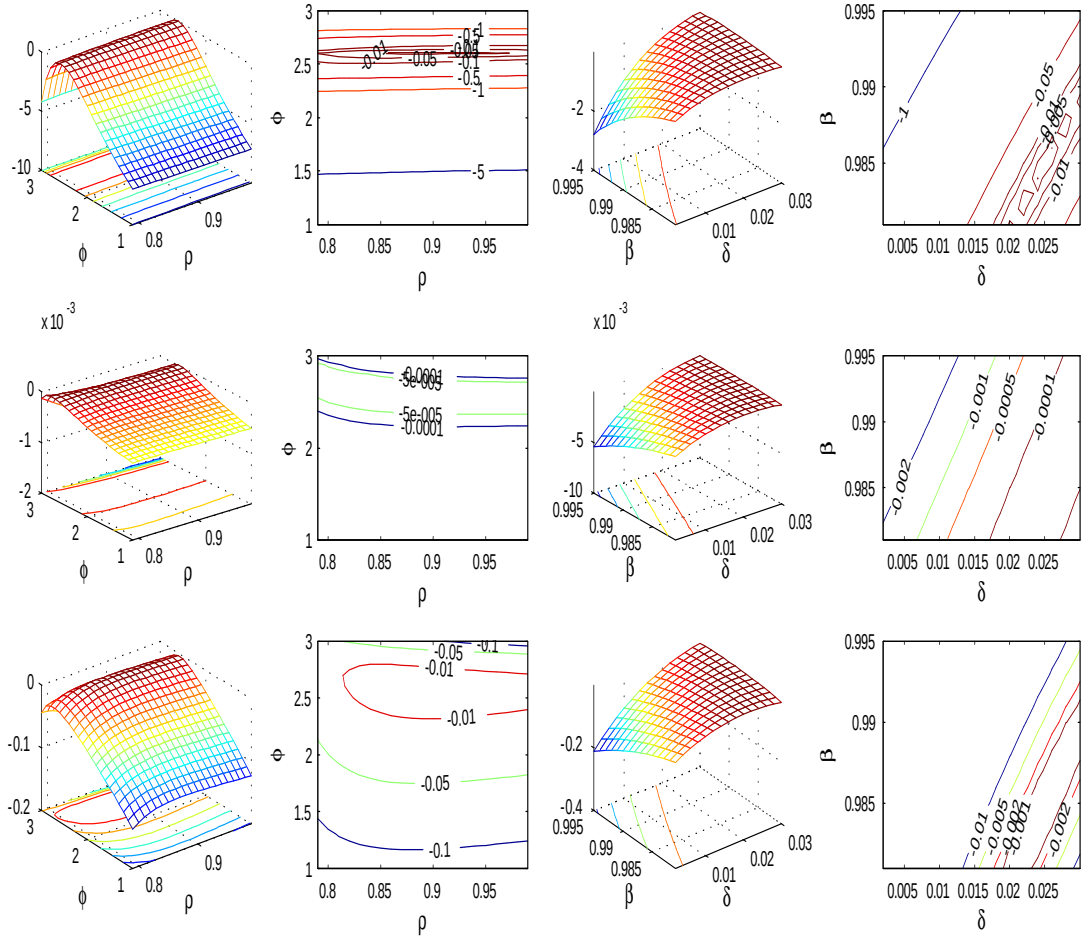


Figure 2: Distance surface for selected parameters

Can we eliminate weak identification problems?

- Change options in your optimization routine. Set tolerance level to 10^{-15} instead that standard 10^{-8}.

- Start optimization routine from many initial values.

Can we eliminate partial identification problems?

Standard solution: calibrate one of the two parameters.

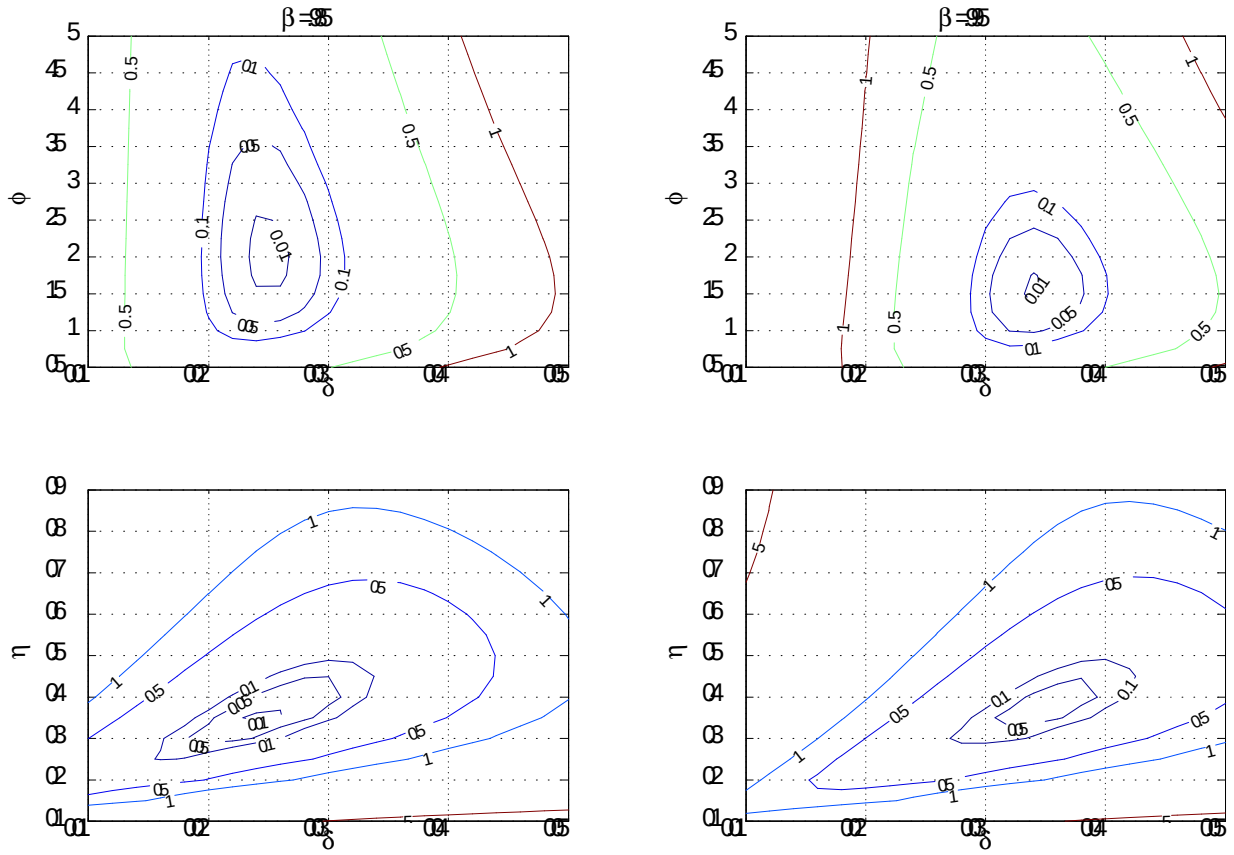


Figure 3: Fixing beta

Summing up

- Identification problems intrinsic to the models and their parameterization.

- Detecting them is complicated because structural parameters enter non-linearly and solution not analytically available.

- These are **population** problems. In small samples additional problems can emerge.

Identification and objective function

What objective function should one use? Likelihood!!

- It has all the information of the model.

- Using a distance function throws away potentially useful identification information. If you use a subset of impulse responses, problems could be compounded.

- Better to add steady states back to the solution. Many parameters may enter only the steady states.

- What does a prior do? Can help if small sample identification problems but not if they are there in population!!

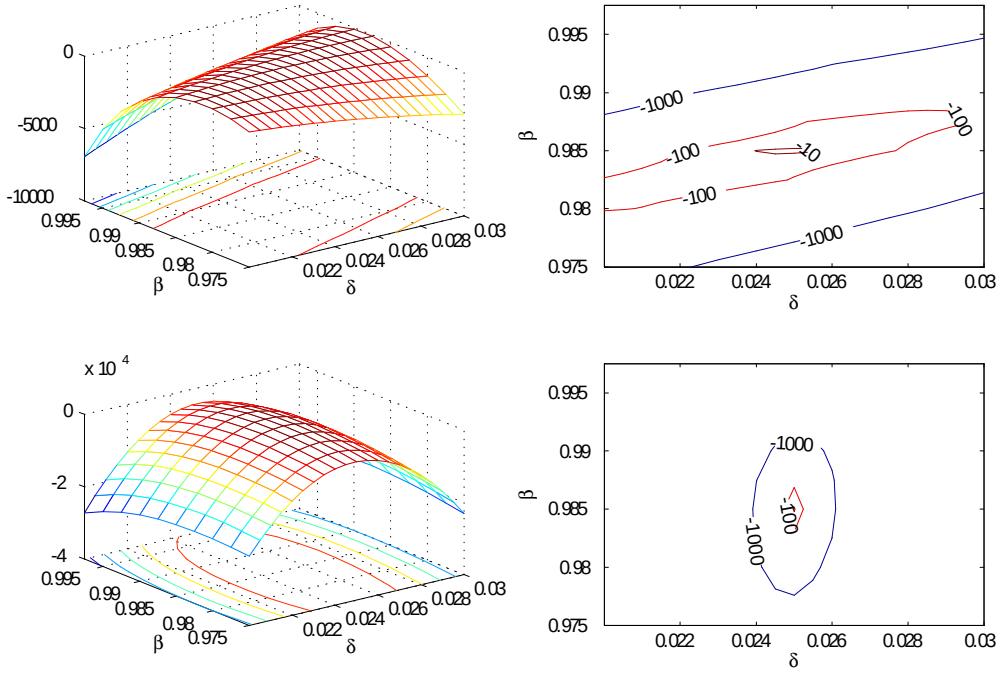


Figure 4: Likelihood and Posterior

Posterior not usually updated if likelihood has no information.

With stability constraints, updating is possible.

Identification and solution methods

• An-Schorfheide (2007) Likelihood function better behaved (in terms of identifying the parameters) if second order approximation is used. How about distance function?

$$\max E_0 \sum_t \beta^t [\log(c_t - b\bar{c}_{t-1}) - a_t N_t]$$

$$c_t = y_t = z_t N_t$$

$\bar{c}_t$ external habit; a_t stationary labor supply shock; $\ln(\frac{z_t}{z_{t-1}}) \equiv u_t^z$ technology shock.

Linear solution (only labor supply shocks):

$$\hat{N}_t = (b+\rho)\hat{N}_{t-1} - b\rho\hat{N}_{t-2} - (1-b)\hat{u}_t^a \tag{20}$$

- Sargent (1978), Kennan (1988): b and ρ are not separately identified.

Second order solution (only labor supply shocks):

$$\hat{N}_t = b\hat{N}_{t-1} + \frac{b(b-1)}{2}\hat{N}_{t-1}^2 - (1-b)\hat{a}_t - \frac{1}{2}(-(1-b)^2 + 1 - b)\hat{a}_t^2$$
$$\hat{a}_t = \rho\hat{a}_{t-1} + u_t^a$$

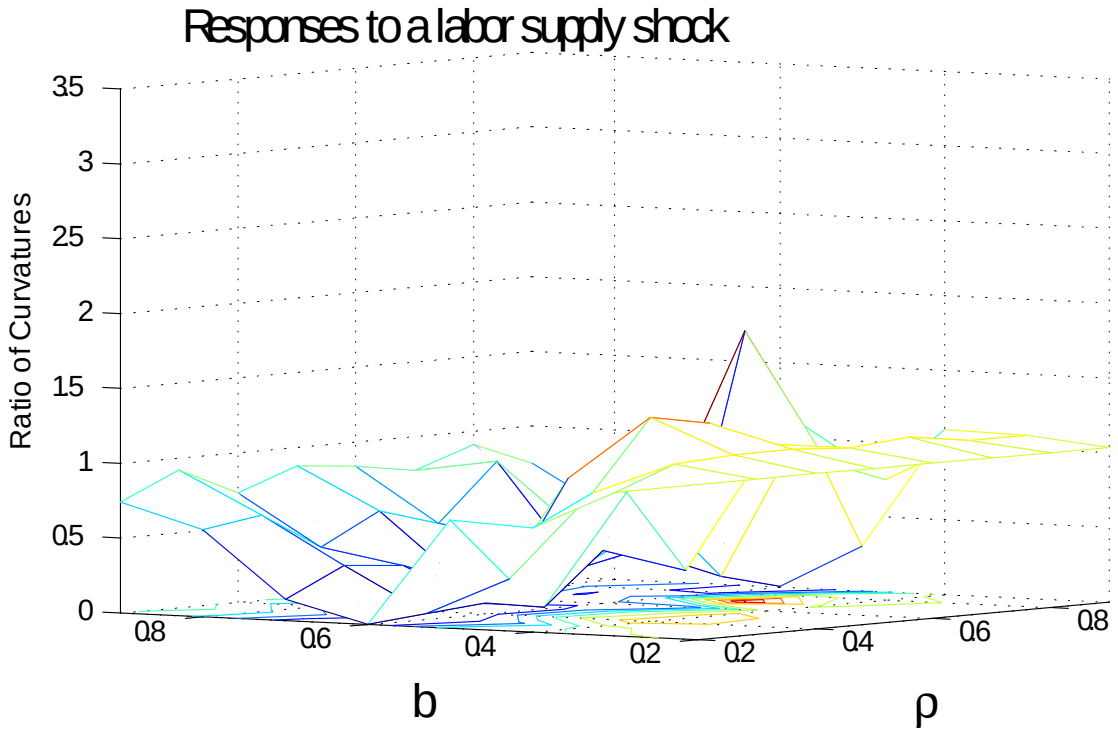


Figure 5: Distance function: linear vs. quadratic

Identification and estimation

What happens if we disregard identification issues and estimate models with a finite sample?

$$
\begin{aligned}
y_t &= \frac{h}{1+h} y_{t-1} + \frac{1}{1+h} E_t y_{t+1} + \frac{1}{\phi}(i_t - E_t \pi_{t+1}) + v_{1t} \\
\pi_t &= \frac{\omega}{1+\omega\beta} \pi_{t-1} + \frac{\beta}{1+\omega\beta} \pi_{t+1} + \frac{(\phi + 1.0)(1-\zeta\beta)(1-\zeta)}{(1+\omega\beta)\zeta} y_t + v_{2t} \\
i_t &= \lambda_r i_{t-1} + (1-\lambda_r)(\lambda_\pi \pi_{t-1} + \lambda_y y_{t-1}) + v_{3t}
\end{aligned}
$$

h: degree of habit persistence (.85); ϕ: relative risk aversion (2)

β: discount factor (.985); ω: degree of price indexation (.25)

ζ: degree of price stickiness (.68)

$\lambda_r, \lambda_\pi, \lambda_y$: policy parameters (.2, 1.55, 1.1)

v_{1t}: AR(ρ_1) (.65); v_{2t}: AR(ρ_2) (.65); v_{3t}: i.i.d.

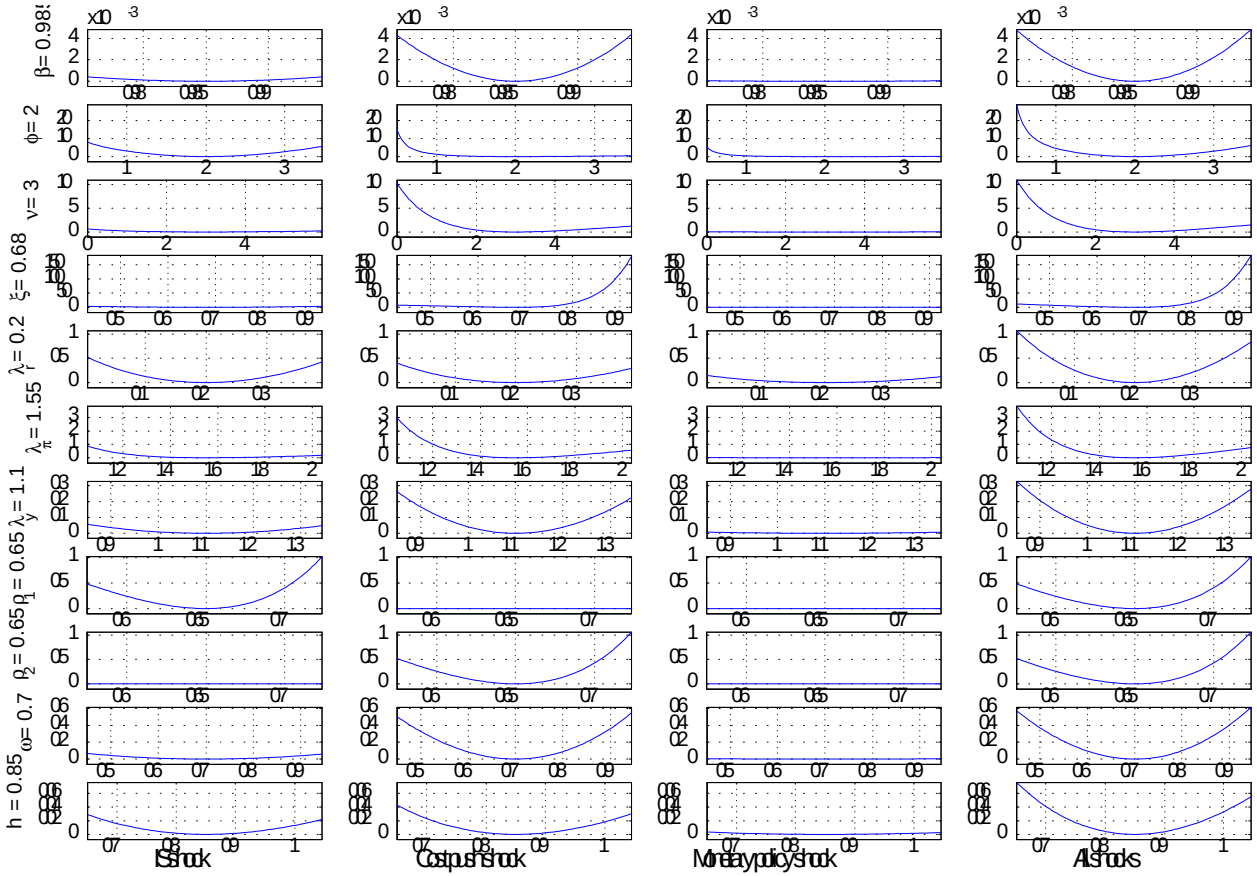


Figure 6: Distance function

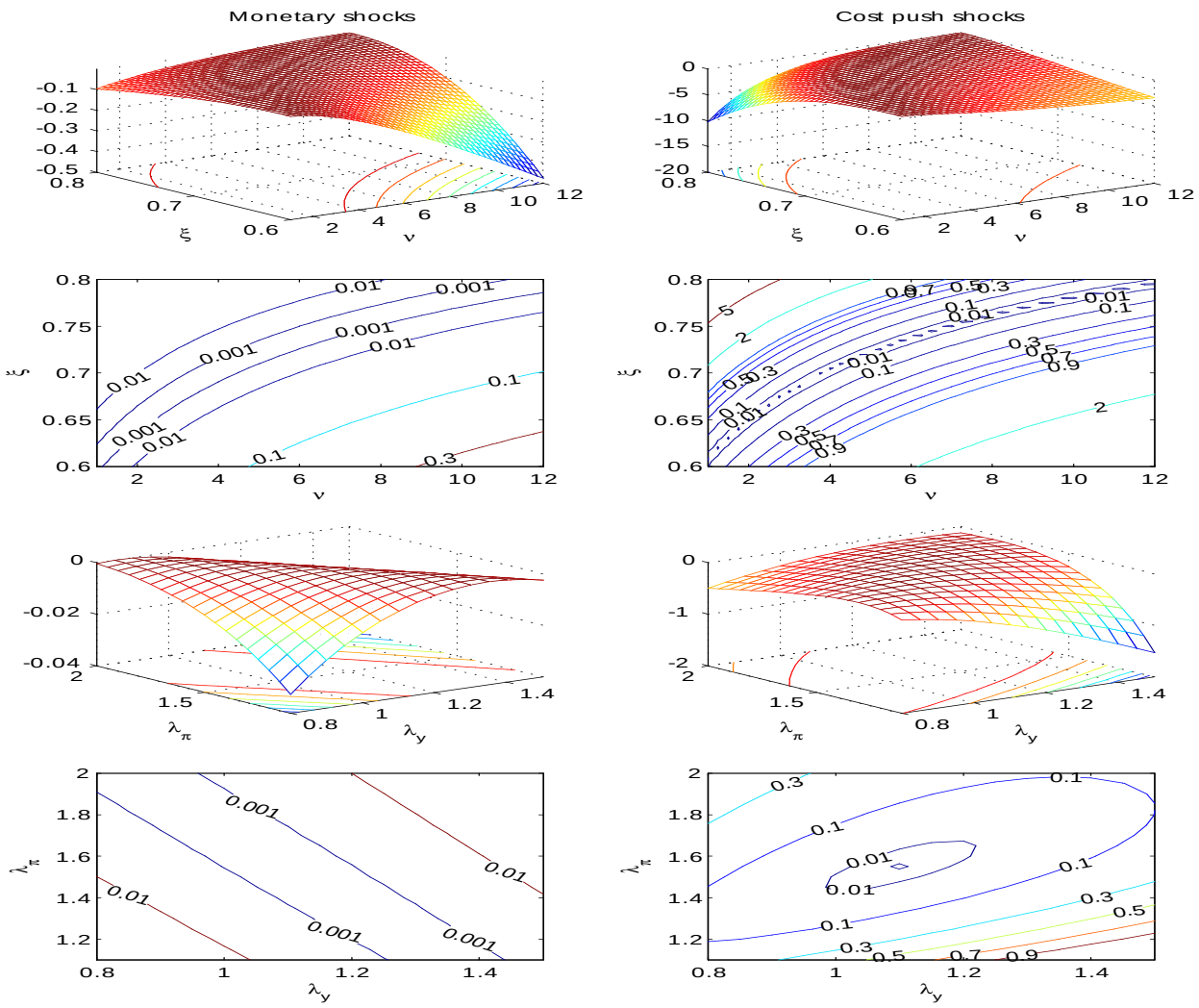


Figure 7: Distance function and contours plots

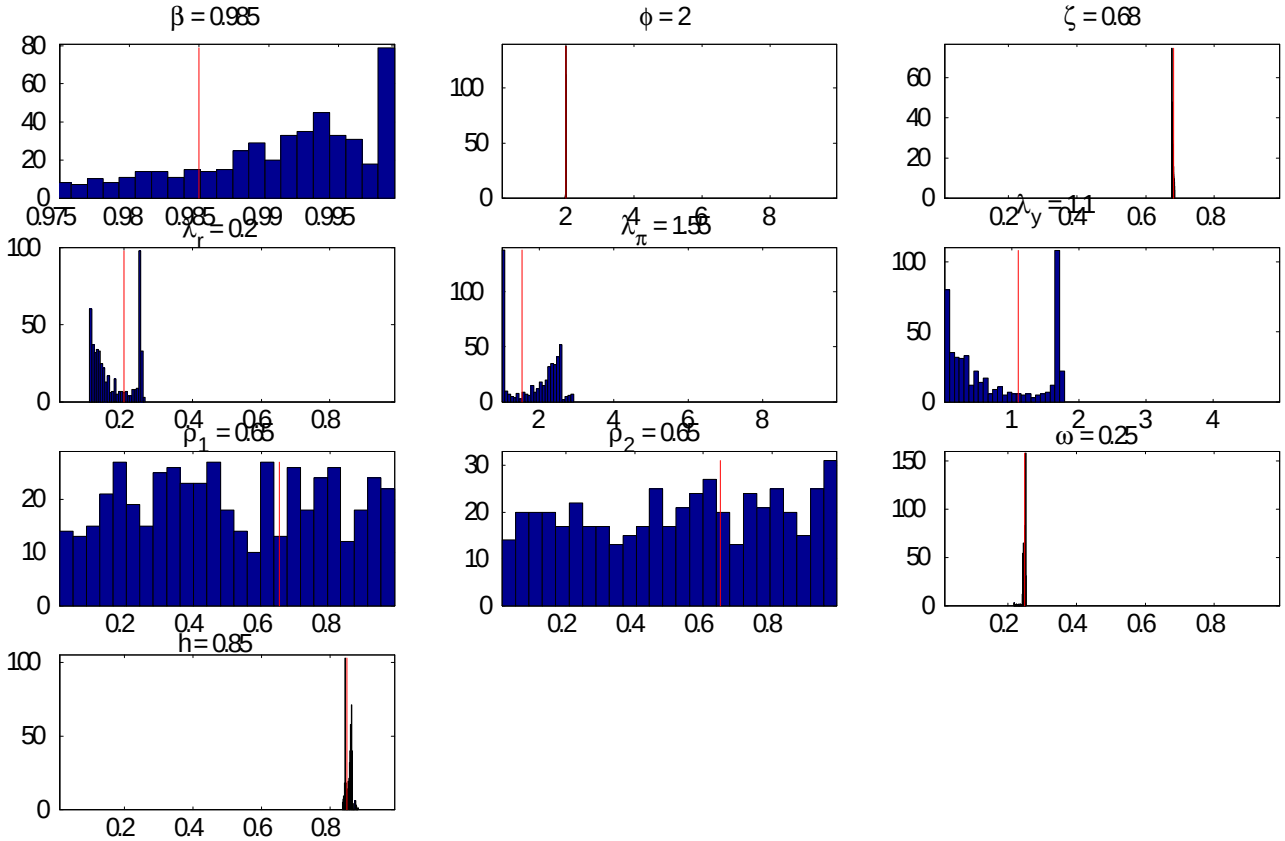


Figure 8: Density Estimates, Monetary Shocks

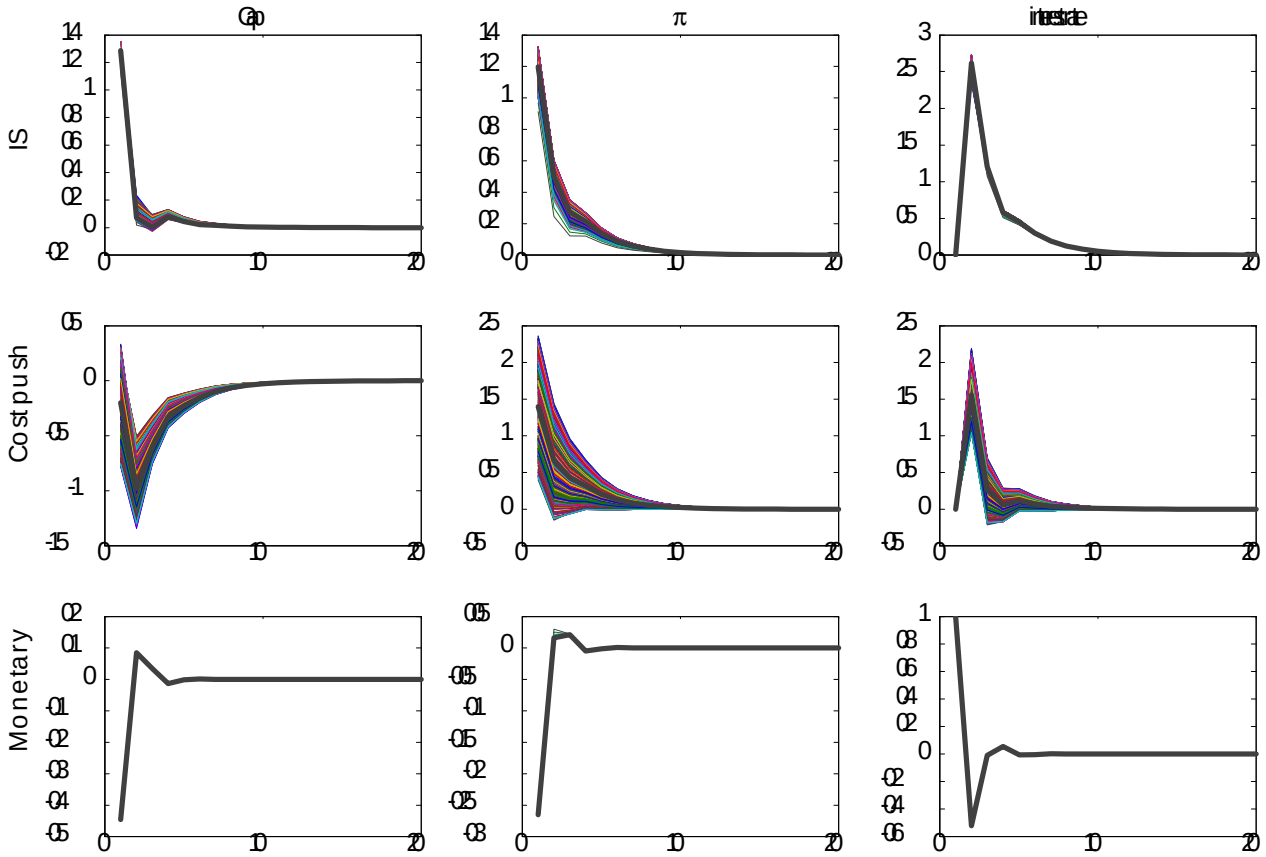


Figure 9: Impulse responses, Monetary Shocks

Table 1: NK model. Matching monetary policy shocks, bias

	True	Population	T = 120	T = 200	T=1000	T=1000 wrong
β	0.985	0.2	0.6	0.7	0.7	0.6
ϕ	2.00	0.7	95.2	70.6	48.6	400
ζ	0.68	0.1	19.3	17.5	23.5	23.7
λ_r	0.2	2.9	172.0	152.6	132.7	90.5
λ_π	1.55	32.5	98.7	78.4	74.5	217.5
λ_y	1.1	34.9	201.6	176.5	126.5	78.3
ρ_1	0.65	13.1	30.4	34.3	31.0	31.3
ρ_2	0.65	12.8	32.9	34.8	34.7	34.7
ω	0.25	0.01	238.9	232.3	198.1	284.0
h	0.85	0.04	30.9	32.4	21.3	100

Conclusions

- Population biases are present.

- Distribution of estimates far from normal.

- Impulse responses "close" to the true one not a criteria to judge how good is a model.

- Surface plots/ numerical analysis can help to detect potential problems.

Wrong inference

$$
\begin{aligned}
0 &= -k_{t+1} + (1-\delta)k_t + \delta x_t \\
0 &= -u_t + \psi r_t \\
0 &= \frac{\eta\delta}{\bar{r}}x_t + (1 - \frac{\eta\delta}{\bar{r}})c_t - \eta k_t - (1-\eta)N_t - \eta u_t - ez_t \\
0 &= -R_t + \phi_r R_{t-1} + (1-\phi_r)(\phi_\pi \pi_t + \phi_y y_t) + er_t \\
0 &= -y_t + \eta k_t + (1-\eta)N_t + \eta u_t + ez_t \\
0 &= -N_t + k_t - w_t + (1+\psi)r_t \\
0 &= E_t[\frac{h}{1+h}c_{t+1} - c_t + \frac{h}{1+h}c_{t-1} - \frac{1-h}{(1+h)\varphi}(R_t - \pi_{t+1})] \\
0 &= E_t[\frac{\beta}{1+\beta}x_{t+1} - x_t + \frac{1}{1+\beta}x_{t-1} + \frac{\chi^{-1}}{1+\beta}q_t + \frac{\beta}{1+\beta}ex_{t+1} - \frac{1}{1+\beta}ex_t] \\
0 &= E_t[\pi_{t+1} - R_t - q_t + \beta(1-\delta)q_{t+1} + \beta\bar{r}r_{t+1}] \\
0 &= E_t[\frac{\beta}{1+\beta\gamma_p}\pi_{t+1} - \pi_t + \frac{\gamma_p}{1+\beta\gamma_p}\pi_{t-1} + T_p(\eta r_t + (1-\eta)w_t - ez_t + ep_t)] \\
0 &= E_t[\frac{\beta}{1+\beta\gamma_p}w_{t+1} - w_t + \frac{1}{1+\beta}w_{t-1} + \frac{\beta}{1+\beta}\pi_{t+1} - \\
&\quad \frac{1+\beta\gamma_w}{1+\beta}\pi_t + \frac{\gamma_w}{1+\beta\gamma_w}{}_{t-1}(w_t - \sigma N_t - \frac{\varphi}{1-h}(c_t - hc_{t-1}) - ew_t)]
\end{aligned}
$$

δ	depreciation rate (.0182)	λ_w	wage markup (1.2)
ψ	parameter (.564)	$\bar{\pi}$	steady state π (1.016)
η	share of capital (.209)	h	habit persistence (.448)
φ	risk aversion (3.014)	σ_l	inverse elasticity of labor supply (2.145)
β	discount factor (.991)	χ^{-1}	investment's elasticity to Tobin's q (.15)
ζ_p	price stickiness (.887)	ζ_w	wage stickiness (.62)
γ_p	price indexation (.862)	γ_w	wage indexation (.221)
ϕ_y	response to y (.234)	ϕ_π	response to π (1.454)
ϕ_r	int. rate smoothing (.779)		
$T_p \equiv$	$\frac{(1-\beta\zeta_p)(1-\zeta_p)}{(1+\beta\gamma_p)\zeta_p}$		
$T_w \equiv$	$\frac{(1-\beta\zeta_w)(1-\zeta_w)}{(1+\beta)(1+(1+\lambda_w)\sigma_l\lambda_w^{-1})\zeta_w}$		

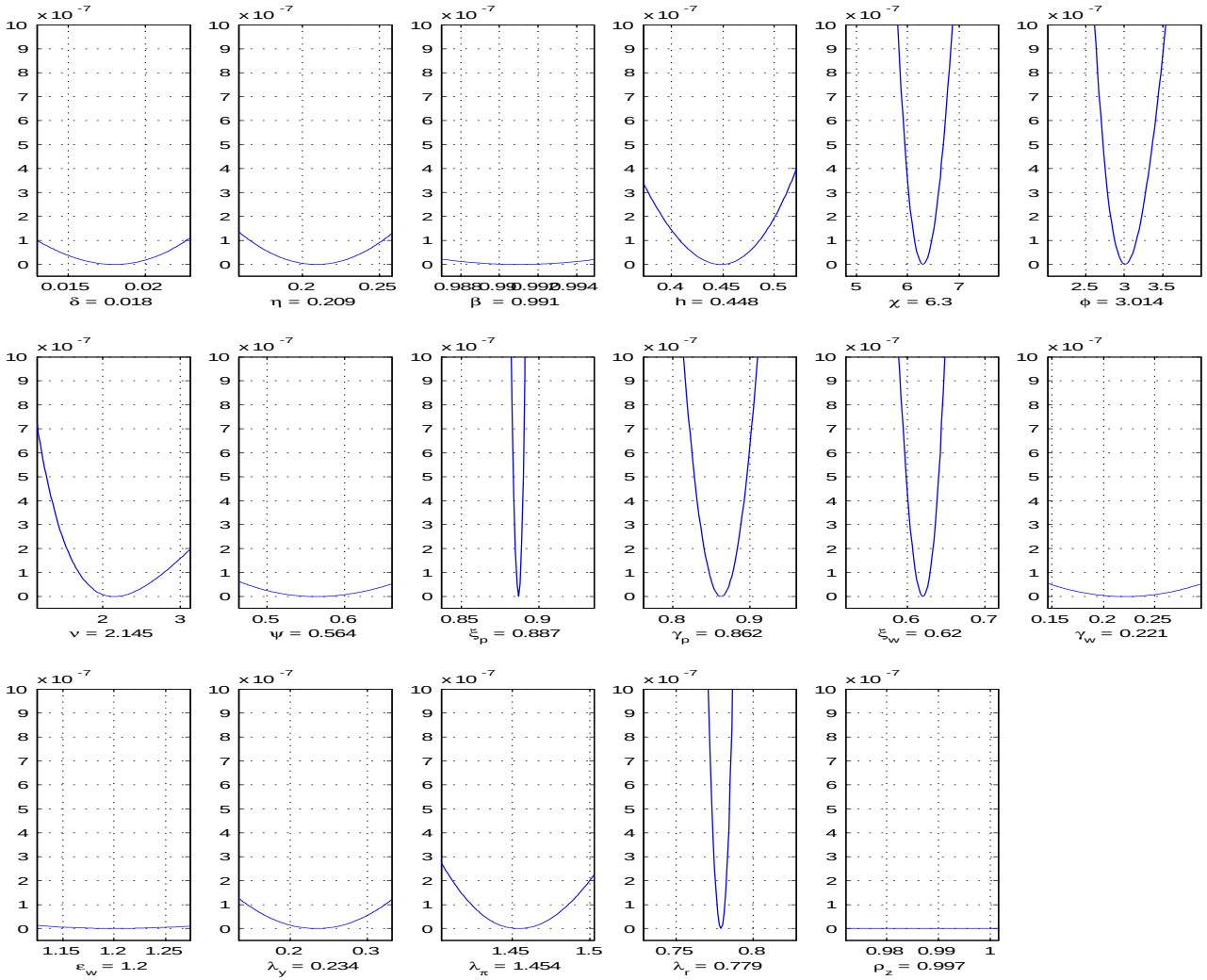


Figure 10: Objective function: monetary and technology shocks

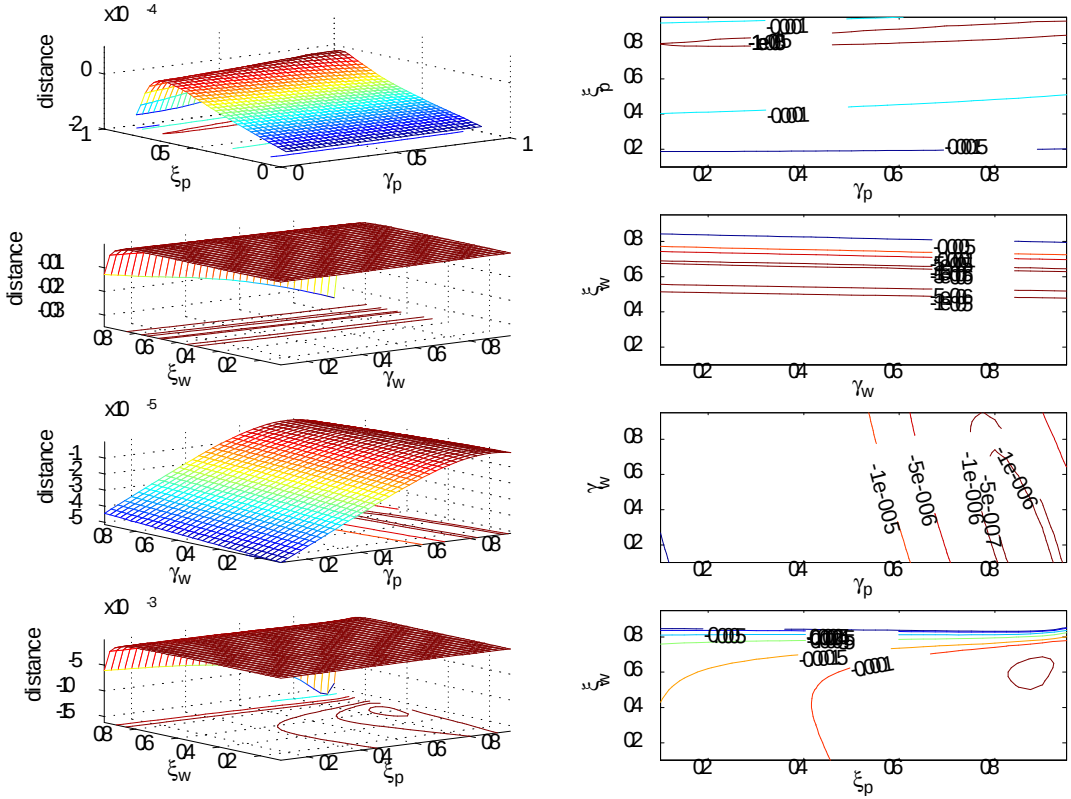


Figure 11: Distance surface and Contours Plots

Experiment:

- use population responses from a model with some features (e.g. with price stickiness and no price indexation).

- ask: is it possible for a model with has different features (e.g. no price stickiness and price indexation) to have impulse responses which are very close to the benchmark one?

- Do they imply different welfare properties?

	ζ_p	γ_p	ζ_w	γ_w	Obj.Fun.
Baseline	**0.887**	**0.862**	**0.62**	**0.221**	
x0 = lb + 1std	0.8944	0.8251	0.615	0	1.8235E-07
x0 = lb + 2std	0.8924	0.7768	0.6095	0.1005	3.75E-07
x0 = ub - 1std	0.882	0.7957	0.6062	0.1316	2.43E-07
x0 = ub - 2std	0.9044	0.7701	0.6301	0	8.72E-07
Case 1	**0**	**0.862**	**0.62**	**0.221**	
x0 = lb + 1std	0.1304	0.0038	0.6401	0.245	2.7278E-08
x0 = lb + 2std	0.1015	0.0853	0.6065	0.1791	4.84E-08
x0 = ub - 1std	0.0701	0.1304	0.6128	0.1979	4.72E-08
x0 = ub - 2std	0.0922	0.0749	0.618	0.215	3.05E-08
Case 2	**0**	**0.862**	**0.62**	**0**	
x0 = lb + 1std	0.0248	0	0.6273	0.029	7.437E-09
x0 = lb + 2std	0.4649	0	0.7443	0.4668	2.10E-06
x0 = ub - 1std	0.0652	0.0004	0.6147	0.0447	7.13E-08
x0 = ub - 2std	0.6463	0.2673	0.8222	0.3811	5.56E-06

	ζ_p	γ_p	ζ_w	γ_w	Obj.Fun.
Case 3	**0.887**	**0**	**0.62**	**0.8**	
x0 = lb + 1std	0.9264	0.3701	0.637	0.4919	3.5156E-07
x0 = lb + 2std	0.9076	0.2268	0.6415	0.154	3.51E-07
x0 = ub - 1std	0.9014	0.3945	0.6477	0	6.12E-07
x0 = ub - 2std	0.9263	0.3133	0.6294	0.4252	4.13E-07
Case 4	**0.887**	**0**	**0**	**0.221**	
x0 = lb + 1std	0.9186	0.3536	0.0023	0	4.7877E-07
x0 = lb + 2std	0.8994	0.234	0	0	3.06E-07
x0 = ub - 1std	0.905	0.3494	0.0021	0	4.14E-07
x0 = ub - 2std	0.9343	0.5409	0.0042	0	9.64E-07
Case 5	**0.887**	**0**	**0**	**0.221**	
x0 = lb + 1std	0.877	0.0123	0.0229	0	2.4547E-06
x0 = lb + 2std	0.8919	0.0411	0.0003	0	4.26E-07
x0 = ub - 1std	0.907	0.2056	0.001	0.0001	6.58E-07
x0 = ub - 2std	0.8839	0.0499	0.0189	0	2.46E-06

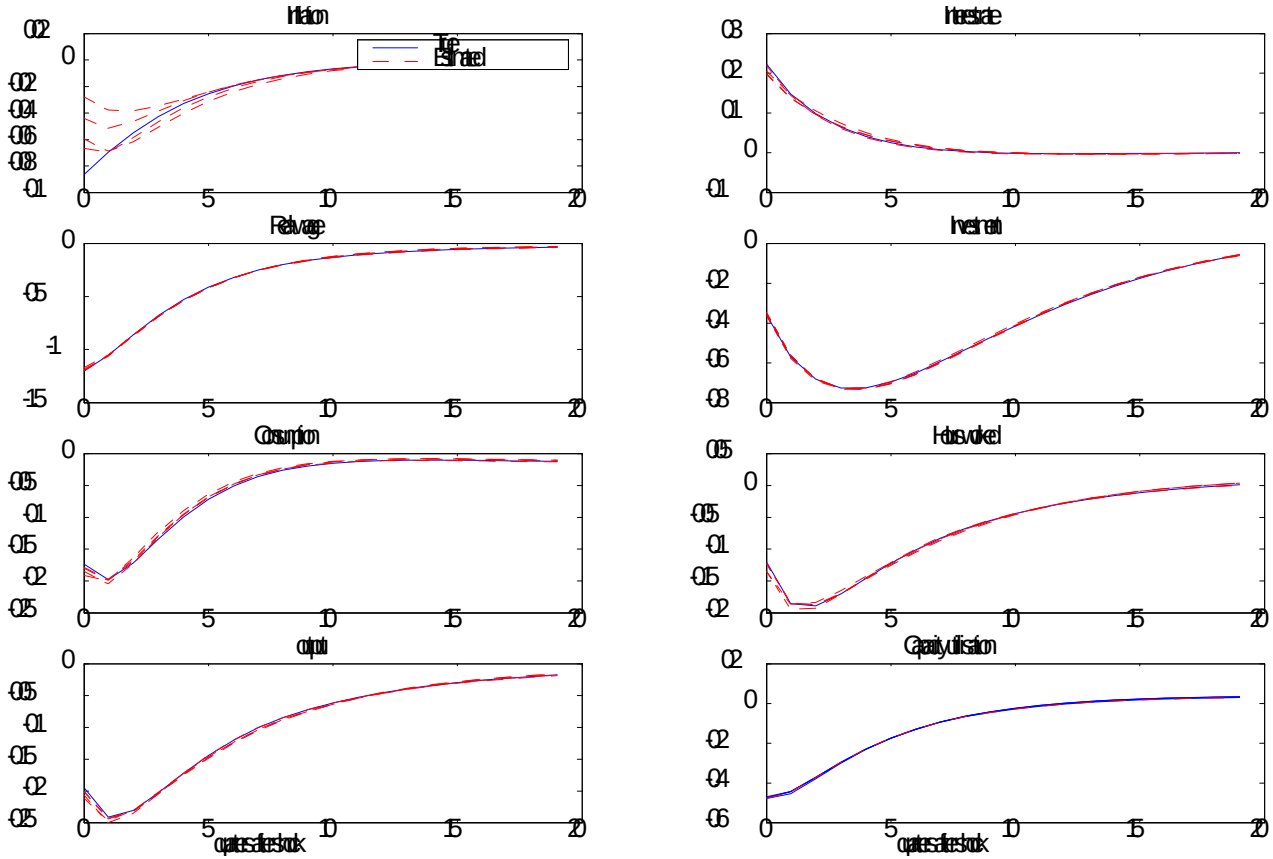


Figure 12: Impulse responses, Case 4.

Welfare costs different!

$L(\pi^2, y^2) = -0.0005$ with true parameters.

$L(\pi^2, y^2) = -0.0022$ with estimated parameters.

Detecting identification problems:

Ex-ante diagnostics:

- Graphical analysis. (Canova- Sala (2009))

- Numerical derivatives/elasticities of the solution/objective function at likely parameter values.

- Simulation analysis: check distribution of population estimates.

Ex-post diagnostics:

- Erratic parameter estimates as T increases.

- Large or non-computable standard errors.

- Crazy t-test (Choi and Phillips (1992), Stock and Wright (2003)).

5) Canova- Sala (2009) (graphical diagnostic)

- Perform prior predictive analysis (can do this prior to the estimation).

- Simulate the objective function (likelihood, posterior, distance function) drawing parameters from some (prior) distribution. Plot objective function against relevant parameters. Check if it is flat, if it displays, ridges, or other peculiarities.

- If the objective function does not change much when we vary a parameter, that parameters can not be identified.

- If the objective function does not change much when we vary a subset of the parameters, the parameters can not be separately identified.

2) Iskrev (2010): testing the rank of a matrix.

- Likelihood function of normal stationary data depends only on its auto-covariance function.

- The Jacobian of the transformation from the structural parameters of a model to the ACF of the data must be full rank at θ_0 for the model to be locally identifiable.

- Randomly draw θ_0 from the prior of the parameters.
Calculate the (analytical) Jacobian at θ_0. If less than full rank $\longrightarrow$ identification deficiencies.

Solution

$$y_{2t} = A_{22}(\theta)y_{2t-1} + A_{23}(\theta)y_{3t} \quad (21)$$
$$y_{1t} = A_{12}(\theta)y_{2t-1} + A_{13}(\theta)y_{3t} \quad (22)$$

where y_{2t} are the (endogenous and exogenous) states, y_{3t} are the shocks, y_{1t} are the controls and θ the $k \times 1$ vector of structural parameters.

- Let $x_t = H[y_{1t}, y_{2t}]'$ where H is a selection matrix.

- Let $m_x(\theta)$ be a vector of theoretical moments of x (in the case of Iskrev, $m_x = vec(E(x), ACF_x(j))$ where $j = 0, \ldots, J$). Let $\hat{m}_x$ be the vector of estimated moment in the actual data. We want $\hat{m}_x = m_x(\theta)$.

- Let $M_x = \frac{\partial m_x(\theta)}{\partial \theta}$.

- All the parameters are locally identifiable at θ_0 if rank$(M_x) \geq k$.

- How do you check the rank of the matrix? Can compute condition number of the eigenvalues and see if there is at least one less than a critical value. Or use Cragg-Donald approach (see below).

Note $M_x = M_a * M_\theta$ where M_θ is the matrix of derivatives of reduced form coefficients (decision rules) with respect to structural parameter, M_a derivative of the moments with respect to reduced form coefficients. Usually the problem is in M_θ.

- Problem 1: identification in DSGE models is not a either/or proposition e.g., the rank of M_x may be k and still some of its eigenvalues may be very small.

- Problem 2: A lots of parameters may not enter the ACF. Can't just use the ACF.

- Problem 3: If there are parameters entering only in A_{13} or A_{23} may not separately identifiable from the variance of the shocks (see Komunijer and Ng (2011).

3) Komunijer and Ng (2011): testing the rank of a matrix.

- Start from (21)-(22), where y_{3t} may also contain measurement errors and let $x_t = y_{1t}$ be the vector of observables.

- The MA representation for x_t is $x_t = H(L, \theta)y_{3t}$. The matrix $H(z, \theta) = \sum_{j=0}^{\infty} h_{y3}(\theta, j)z^{-j}$ is obtained as

$$H(z, \theta) = D(\theta) + C(\theta)[z * I_{N_{y_2}} - A(\theta)]^{-1}B(\theta) \qquad (23)$$

where N_{y_2} is the size of y_{2t}, $z \in C$.

- Define the spectral density of x_t by $s_x(\omega, \theta) = H(z, \theta)\Sigma_{y_3}(\theta)H(z, \theta)'$.

- Properties of the spectral density (or ACF) of x_t are determined by the the properties of H. H, in turn, is linked to the (Rosenbrook) system matrix $P(z,\theta) = \begin{bmatrix} z * I_{N_{y_2}} - A(\theta) & B(\theta) \\ -C(\theta) & D(\theta) \end{bmatrix}$. In particular, rank $(P(z,\theta)) = N_{y_2}+$ rank $(H(z,\theta))$

- For identification want $\frac{\partial S_x(\omega,\theta)}{\partial \theta}$ to have full column rank. To make sure that this is the case, Kommunjer and Ng derive conditions on the inputs of the matrix $P(z,\theta)$.

- Nice since P contains the mapping from DSGE parameters to the decision rules.

- Result 1 (case of $N_{y_3} < N_{y_1}$): Two vectors θ_1 and θ_0 are observationally equivalent if there exists T, U matrices of dimension $N_{y_2} \times N_{y_2}$ and $N_{y_3} \times$

N_{y_3} respectively, and the following hold

$$\begin{aligned} A(\theta_1) &= TA(\theta_0)T' && (24)\\ B(\theta_1) &= TB(\theta_0)U && (25)\\ C(\theta_1) &= C(\theta_0)T^{-1} && (26)\\ D(\theta_1) &= D(\theta_0)U && (27)\\ \Sigma_{y_3}(\theta_1) &= U^{-1}\Sigma_{y_3}(\theta_0)U^{-1} && (28) \end{aligned}$$

- Result 2: Let $\delta(T, U, \theta) = \begin{bmatrix} vec(TA(\theta_0)T') \\ vec(TB(\theta_0)U) \\ vec(C(\theta_0)T^{-1}) \\ vec(D(\theta_0)U) \\ vech(U^{-1}\Sigma_{y_3}(\theta_0)U^{-1}) \end{bmatrix}$.

The parameters of the model are locally identifiable at θ_0 if $\delta(T, U, \theta_1) = \delta(I, I, \theta_0)$has a unique solution at $(T, U, \theta_1) = (I, I, \theta_0)$.

- Practical implication: Compute $\frac{\partial \delta(T,U,\theta)}{\partial \theta}$. Check if it has full column rank at (I, I, θ_0).

i) Need to pick a θ_0

ii) Need to compute numerical derivatives (see matlab program in on-line appendices in Econometrica).

- For other cases ($N_{y_3} \geq N_{y_1}$) see paper for the changes needed.

How do you test the rank of matrix?

- Cragg and Donald (1997): Testing rank of Hessian. Under regularity conditions: $(vec(\hat{H}) - vec(H))'\Omega(vec(\hat{H}) - vec(H)) \sim \chi^2((N - L_0)(N - L_0))$ $N = dim(H), L_0 =$rank of H.

- Anderson (1984): Size of characteristic roots of Hessian. Under regularity conditions: $\frac{\sum_{i=1}^{N-m} \hat{\lambda}_i}{\sum_{i=1}^{N} \hat{\lambda}_i} \xrightarrow{D}$ Normal distribution.

Concentration Statistics: $\mathcal{C}_{\theta_0}(i) = \int_{j \neq i} \frac{g(\theta) - g(\theta_0) d\theta}{\int (\theta - \theta_0) d\theta}, i = 1, 2 \ldots$ (Stock, Wright and Yogo (2002)) = measures the global curvature of the objective function around θ_0.

Applied to SW model:

- rank of $H = 6$;

- sum of 12-13 characteristics roots is smaller than 0.01 of the average root; i.e. 12-13 dimensions with weak or partial identification problems.

Which are the parameters is causing problems? $\beta, h, \sigma_l, \delta, \eta, \psi, \gamma_p, \gamma_w, \lambda_w, \phi_\pi, \phi_y, \rho_z$. (consistent with graphical analysis).

Why? Variations of these parameters hardly affect law of motion of states!

Almost a rule: **for identification need states to react changes in structural parameters**.

4) Rubio, Waggoner, Zha (2010) testing the rank of matrix (global identification). DSGE model solution must have a restricted VAR representation.

5) Mueller (2010)

- How much do the results depend on the prior?

- How much prior information is there in the posterior?

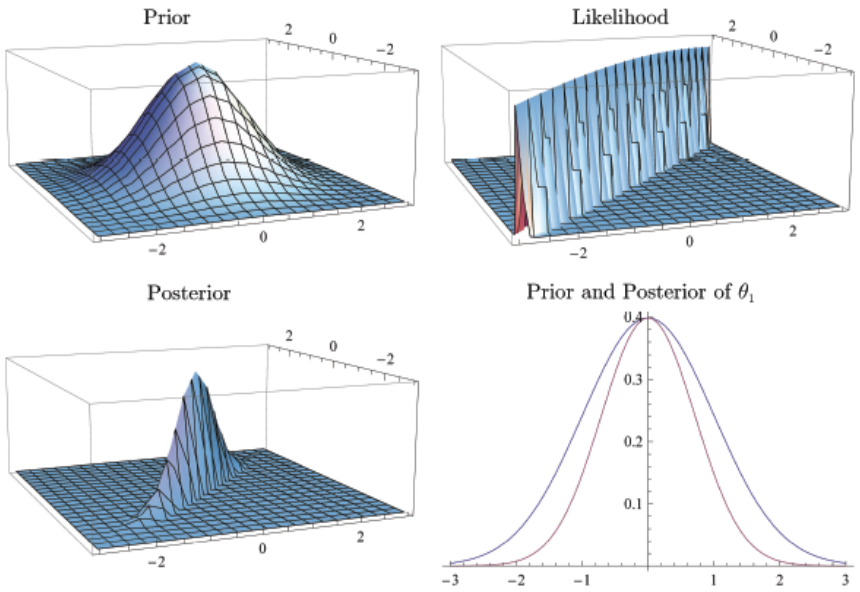
- Is the posterior reflecting mostly the likelihood or the prior?

Current answers based on plotting marginal/prior marginal posterior insufficient:

- Univariate representation.

- Prior of one parameter affect posterior of other parameters as well.

- Stability conditions imply that (marginal) priors and posteriors differ even though the likelihood has not information.


Prior
Likelihood
Posterior
Prior and Posterior of θ_1
2
0
−2
0.4
0.3
0.2
0.1
−3
−1
1
3

- Traditional approach: Compute the rank of the information matrix $I(\theta) = -E[\frac{\partial f(y,\theta)}{\partial \theta^2}]$.

- Problem: measure local; does not satisfy the likelihood principle (inference ins based on averages of hypothetical histories that never materialized).

Iskev (2010), Komunjer and Ng (2011): compute the rank of a matrix of derivatives.

- Not based on the (classical) likelihood. Still problems: local measure, classical inference.

General idea: Assume that θ is a scalar (for simplicity)

1) Start from $g(\theta)$ with mean μ and variance $\sigma^2_{g(\theta)}$.

2) Embed this prior in a family $g_\alpha(\theta)$ with mean $\mu + \alpha$ and scores $s_\alpha(\theta) = \frac{\partial \ln g_\alpha(\theta)}{\partial \alpha}$

3) The posterior for this class has mean $\mu_{g(\alpha|y)}(\theta) = \frac{\int \theta L(\theta|y) g_\alpha(\theta) d\theta}{\int L(\theta|y) g_\alpha(\theta) \theta d\theta}$ and

$$\frac{\partial \mu_{g(\alpha|y)}}{\partial \alpha}|_{(\alpha=0)} = E_{g(\alpha|y)}[(\theta - \mu_{g(\alpha|y)}) s_{(\alpha=0)}(\theta)] \tag{29}$$

4) If $g_\alpha(\theta) = g_{\alpha=0}(\theta)\exp[\alpha\frac{\theta-\mu}{\sigma^2} - C(\alpha)]$ where $C(\alpha)$ is independent of θ then

$$\frac{\partial \mu_{g(\alpha|y)}}{\partial \alpha}|_{\alpha=0} = J = \sigma^2_{g(\alpha)}(\theta)^{-1}\sigma^2_{g(\alpha|y)}(\theta) \tag{30}$$

- Prior sensitivity: $PS = J * \sigma_g = \frac{\sigma^2_{g(\alpha|y)}(\theta)}{\sigma_{g(\alpha)}(\theta)}$.

- Prior informativeness: $PI = min(1, J)$.

Interpretation:

1) PS measures the linear approximation to the change of the posterior mean that can be induced by increasing the prior mean by one prior standard deviation

2) If the likelihood pins down exactly θ than changing the prior mean leaves the posterior mean unchanged and $J = 0$. At the opposite extreme if the likelihood is flat $J = 1$. Thus, values of PI between zero and one may be thought of as a numerical measure for the relative importance of prior information for the posterior results.

Advantages:

1) Global rather than a local measure:compare $\sigma^2_{g(\alpha|y}(\theta)$ with $-E[\frac{\partial f(y,\theta)}{\partial \theta^2}|_{\hat{\theta}}$.

2) Joint as opposed to marginal: compare multivariate J (see Mueller) with $[\sigma^2_{g(\alpha)}(\theta)(j,j)]^{-1}[\sigma^2_{g(\alpha|y)}(\theta)(j,j)]$

3) Consistent with the likelihood principle, dependent on the prior

4) Deliver a measure rather than a yes/no answer (as it happens with testing)

5) Very easy to compute: just take a the posterior covariance matrix and compare it with prior covariance matrix.

Example: Smets and Wouters (2007)

Parameter	Prior Mean	Prior Std	Post Mean	Post Std	Post/Prior Std Ratio	PS	$J_{(i,i)}$
α	0.330	0.020	0.348	0.003	0.150	0.001	0.022
ζ_p	0.748	0.099	0.839	0.026	0.260	0.027	0.068
S''	4.007	1.502	4.098	0.933	0.621	0.799	0.386
h	0.700	0.050	0.659	0.049	0.978	0.053	0.955
a''	0.200	0.100	0.282	0.107	1.071	0.117	1.147
ν_l	2.000	0.750	1.241	0.490	0.654	0.645	0.431
ζ_w	0.749	0.100	0.781	0.137	1.370	0.310	1.866
r^*	1.507	1.003	0.359	0.207	0.206	0.084	0.042
ψ_1	2.001	0.250	2.218	0.181	0.723	0.142	0.523
ψ_2	0.201	0.101	0.246	0.080	0.794	0.068	0.630
ρ_r	0.502	0.200	0.792	0.025	0.123	0.012	0.016
π^*	2.001	0.250	2.397	0.183	0.733	0.146	0.537
γ	2.751	0.499	1.805	0.264	0.529	0.170	0.280

J(i,i) close to one for many parameters.

5.1) Koop, Pesaran and Smith (2011) (simulation approach)

- In large samples, the variance of the likelihood distribution must converge to zero at the rate T if a parameter is identified. Since, in large samples, the importance of the prior disappears, the variance of the posterior must have the same properties.

- In large samples, the variance of the posterior distribution of parameters with identification problems must converge to zero at the rate slower than T or may not converge at all.

- Simulate data from the model with different length. Check how the variance of the posterior of the parameters change.

What to do when identification problems exist?

1) Which type of problems?

- If population problems need respecify/reparameterize the model. For example, estimation of the following NK system has less identification problems.

$$
\begin{aligned}
y_t &= \frac{h}{1+h} y_{t-1} + \frac{1}{1+h} E_t y_{t+1} + \frac{1}{\phi}(i_t - E_t \pi_{t+1}) + v_{1t} \\
\pi_t &= a\pi_{t-1} + b\pi_{t+1} + \kappa y_t + v_{2t} \\
i_t &= \lambda_r i_{t-1} + (1-\lambda_r)(\lambda_\pi \pi_{t-1} + \lambda_y y_{t-1}) + v_{3t}
\end{aligned}
$$

- If are due to a particular objective function or to the use of limited information: use likelihood.

- If are due to small sample, add information (prior or other data).

- Don't proceed as if they do not exist. Estimates make no sense!!

- Careful with mixed calibration-estimation. It is preferable to use full calibration or Bayesian calibration (Canova and Paustian (2011)).

- Do you really need to estimate the model or can you do with reasonable calibration?

Bayesian Methods for DSGE models

Fabio Canova
EUI and CEPR
November 2012

Outline

- Bayes Theorem.
- Prior Selection.
- Posterior Simulators.
- Robustness.
- Estimation of DSGE.
- Topics: Prior elicitation, data selection, DSGE-VAR, data rich DSGE, dealing with trends, non-linear DSGE.

References

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer and Verlag.

Bauwens, L., M. Lubrano and J.F. Richard (1999) *Bayesian Inference in Dynamics Econometric Models*, Oxford University Press.

Faust, J. and Gupta, A. (2012) Posterior Predictive Analysis for Evaluating DSGE Models, NBER working paper 17906.

Gelman, A., J. B. Carlin, H.S. Stern and D.B. Rubin (1995), *Bayesian Data Analysis*, Chapman and Hall, London.

Poirier, D. (1995) Intermediate Statistics and Econometrics, MIT Press.

Casella, G. and George, E. (1992) Explaining the Gibbs Sampler *American Statistician*, 46, 167-174.

Chib, S. and Greenberg, E. (1995) Understanding the Hasting-Metropolis Algorithm, *The American Statistician*, 49, 327-335.

Chib, S. and Greenberg, E. (1996) Markov chain Monte Carlo Simulation methods in Econometrics, *Econometric Theory*, 12, 409-431.

Geweke, J. (1995) Monte Carlo Simulation and Numerical Integration in Amman, H., Kendrick, D. and Rust, J. (eds.) *Handbook of Computational Economics* Amsterdaam, North Holland, 731-800.

Kass, R. and Raftery, A (1995), Empirical Bayes Factors, *Journal of the American Statistical Association*, 90, 773-795.

Sims, C. (1988) " Bayesian Skepticism on unit root econometrics", *Journal of Economic Dynamics and Control*, 12, 463-474.

Tierney, L (1994) Markov Chains for Exploring Posterior Distributions (with discussion), *Annals of Statistics*, 22, 1701-1762.

Del Negro, M. and F. Schorfheide (2003), " Priors from General Equilibrium Models for VARs", *International Economic Review*, 45, 643-673.

Del Negro M. and Schorfheide, F. (2008) Forming priors for DSGE models (and how it affects the assessment of nominal rigidities), *Journal of Monetary Economics*, 55, 1191-1208.

Kadane, J., Dickey, J., Winkler, R. , Smith, W. and Peters, S., (1980), Interactive elicitation of opinion for a normal linear model, *Journal of the American Statistical Association*, 75, 845-854.

Beaudry, P. and Portier, F (2006) Stock Prices, News and Economic Fluctuations, *American Economic Review*, 96, 1293-1307.

Boivin, J. and Giannoni, M (2006) DSGE estimation in data rich environments, University of Montreal working paper

Canova, F., (1998), "Detrending and Business Cycle Facts", *Journal of Monetary Economics*, 41, 475-540.

Canova, F., (2010), "Bridging DSGE models and the data", manuscript.

Canova, F., and Ferroni, F. (2011), "Multiple filtering device for the estimation of DSGE models", *Quantitative Economics*, 2, 73-98.

Chari, V., Kehoe, P. and McGratten, E. (2009) "New Keynesian models: not yet useful for policy analysis, *American Economic Journal: Macroeconomics*, 1, 242-266.

Guerron Quintana, P. (2010), "What you match does matter: the effects of data on DSGE estimation", *Journal of Applied Econometrics*, 25, 774-804.

Ireland, P. (2004) A method for taking Models to the data, *Journal of Economic Dynamics and Control*, 28, 1205-1226.

Stock, J. and Watson, M. (2002) Macroeconomic Forecasting using Diffusion Indices, Journal of Business and Economic Statistics, 20, 147-162.

Smets, F. and Wouters, R (2003), An estimated dynamic stochastic general equilibrium model of the euro area, *Journal of European Economic Association*, 1, 1123–1175.

Watson, M. (1993) "Measures of Fit for Calibrated Models", *Journal of Political Economy*, 101, 1011-1041.

1 Preliminaries

Classical and Bayesian analysis differ on a number of issues

Classical analysis:

• Probabilities = limit of the relative frequency of the event.

• Parameters are fixed, unknown quantities.

• Unbiased estimators useful because average value of sample estimator converge to true value via some LLN. Efficient estimators preferable because they yield values closer to true parameter.

• Estimators and tests are evaluated in repeated samples (to give correct result with high probability).

Bayesian analysis:

• Probabilities = degree of (typically subjective) beliefs of a researcher in an event.

• Parameters are random with a probability distributions.

• Properties of estimators and tests in repeated samples uninteresting: beliefs not necessarily related to relative frequency of an event in large number of hypothetical experiments.

• Estimators are chosen to minimize expected loss functions (expectations taken with respect to the posterior distribution), conditional on the data. Use of probability to quantify uncertainty.

In large samples (under appropriate regularity conditions):

- Posterior mode $\alpha^* \xrightarrow{P} \alpha_0$ (Consistency)

- Posterior distribution converges to a normal with mean α_0 and variance $(T \times I(\alpha_0))^{-1}$, where $I(\alpha)$ is Fisher's information matrix (Asymptotic normality).

Classical and Bayesian analyses differ in small samples and for dealing with unit root processes.

Bayesian analysis requires:

- Initial information $\longrightarrow$ Prior distribution.
- Data $\longrightarrow$ Likelihood.
- Prior and Likelihood $\longrightarrow$ Bayes theorem $\longrightarrow$ Posterior distribution.
- Can proceed recursively (mimic economic learning).

2 Bayes Theorem

Parameters of interest $\alpha \in A$, A compact. Prior information $g(\alpha)$. Sample information $f(y|\alpha) \equiv \mathcal{L}(\alpha|y)$.

- Bayes Theorem.

$$g(\alpha|y) = \frac{f(y|\alpha)g(\alpha)}{f(y)} \propto f(y|\alpha)g(\alpha) = \mathcal{L}(\alpha|y)g(\alpha) \equiv \grave{g}(\alpha|y)$$

$f(y) = \int f(y|\alpha)g(\alpha)d\alpha$ is the unconditional sample density (Marginal likelihood), and it is constant from the point of view of $g(\alpha|y)$; $g(\alpha|y)$ is the posterior density, $\grave{g}(\alpha|y)$ is the posterior kernel, $g(\alpha|y) = \frac{\grave{g}(\alpha|y)}{\int \grave{g}(\alpha|y)d\alpha}$.

• $f(y)$ it is a measure of fit. It tells us how good the model is in reproducing the data, not at a single point, but on average over the parameter space.

• α are regression coefficients, structural parameters, etc.; $g(\alpha|y)$ is the conditional probability of α, given what we observe, y.

• Theorem uses rule: $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$. It says that if we start from some beliefs on α, we may modify them if we observe y. It does not says what the initial beliefs are, but how they should change is data is observed.

To use Bayes theorem we need:

a) Formulate prior beliefs, i.e. choose $g(\alpha)$.

b) Formulate a model for the data (the conditional probability of $f(y|\alpha)$).

After observing the data, we treat the model as the likelihood of α conditional on y, and update beliefs about α.

• Bayes theorem with nuisance parameters (e.g. α_1 long run coefficients, α_2 short run coefficients; α_1 regression coefficient; α_2 serial correlation coefficient in the errors).

Let $\alpha = [\alpha_1, \alpha_2]$ and suppose interest is in α_1. Then $g(\alpha_1, \alpha_2|y) \propto f(y|\alpha_1, \alpha_2)g(\alpha_1, \alpha_2)$

$$\begin{aligned} g(\alpha_1|y) &= \int g(\alpha_1, \alpha_2|y)d\alpha_2 \\ &= \int g(\alpha_1|\alpha_2, y)g(\alpha_2|y)d\alpha_2 \end{aligned} \tag{1}$$

Posterior of α_1 averages the conditional of α_1 with weights given by the posterior of α_2.

• Bayes Theorem with two (N) samples.

Suppose $y_t = [y_{1t}, y_{2t}]$ and that y_{1t} is independent of y_{2t}. Then

$$\breve{g} \equiv f(y_1, y_2|\alpha)g(\alpha) = f_2(y_2|\alpha)f_1(y_1|\alpha)g(\alpha) \propto f_2(y_2|\alpha)g(\alpha|y_1) \quad (2)$$

Posterior for α is obtained finding first the posterior of using y_{1t} and then, treating it as a prior, finding the posterior using y_{2t}.

- Sequential learning.

- Can use data from different regimes.

- Can use data from different countries.

2.1 Likelihood Selection

• It should reflect an economic model.

• It must represent well the data. Misspecification problematic since it spills across equations and makes estimates uninterpretable.

• For our purposes the likelihood is simply the theoretical (DSGE) model you write down.

2.2 Prior Selection

• Three methods to choose priors in theory. Two not useful for DSGE models since are designed for models which are linear in the parameters.

1) Non-Informative subjective. Choose **reference priors** because they are invariant to the parametrization.

- Location invariant prior: $g(\alpha)$ =constant (=1 for convenience). Scale invariant prior $g(\sigma) = \sigma^{-1}$.

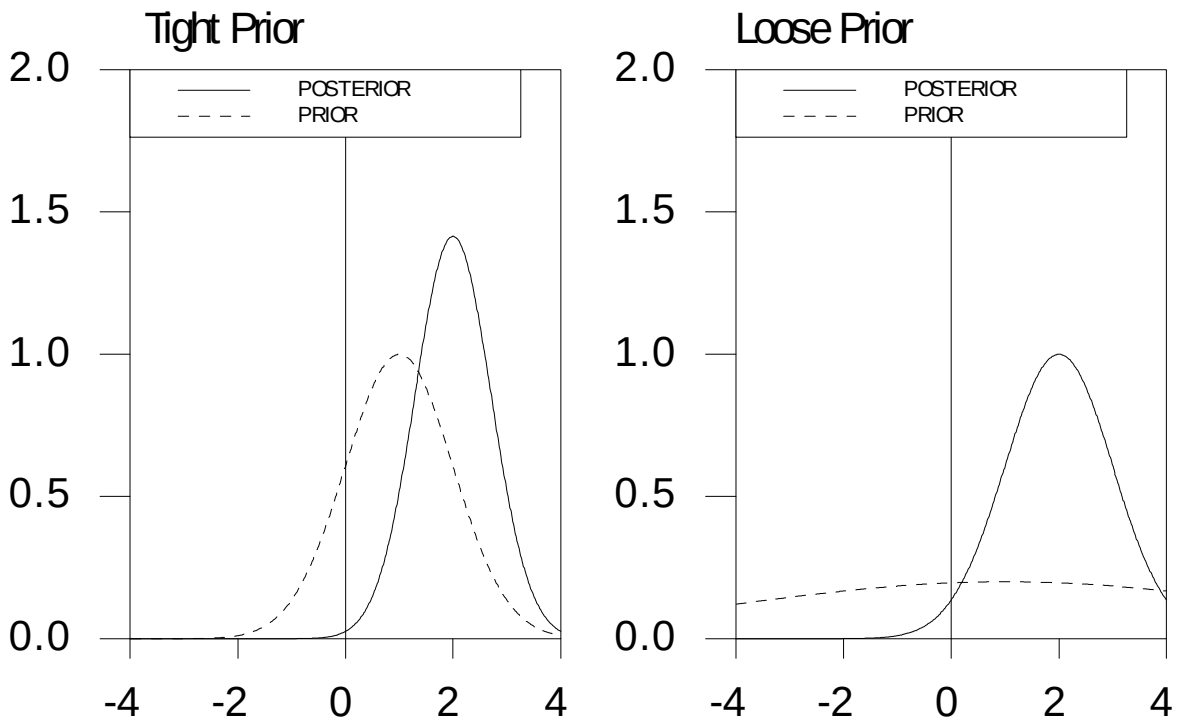
- Location-scale invariant prior : $g(\alpha, \sigma) = \sigma^{-1}$.

• Non-informative priors useful because many classical estimators (OLS, ML) are Bayesian estimators with non-informative priors

2) Conjugate Priors

A prior is conjugate if the posterior has the same form as the prior. Hence, the form posterior will be analytically available, only need to figure out its posterior moments.

• Important result in linear models with conjugate priors: Posterior moments = weighted average of sample and prior information. Weights = relative precision of sample and prior informations.


Tight Prior
Loose Prior
POSTERIOR
PRIOR
2.0
1.5
1.0
0.5
0.0
-4
-2
0
2
4

3) Objective priors and ML-II approach. Based on:

$$f(y) = \int \mathcal{L}(\alpha|y)g(\alpha)d\alpha \equiv \mathcal{L}(y|g) \tag{3}$$

Since $\mathcal{L}(\alpha|y)$ is fixed, $\mathcal{L}(y|g)$ reflects the plausibility of g in the data.
If g_1 and g_2 are two priors and $\mathcal{L}(y|g_1) > \mathcal{L}(y|g_2)$, there is better support for g_1. Hence, can estimate the "best" g using $\mathcal{L}(y|g)$.

In practice, set $g(\alpha) = g(\alpha|\theta)$, where θ= hyperparameters (e.g. the mean and the variance of the prior). Then $\mathcal{L}(y|g) \equiv \mathcal{L}(y|\theta)$.

The θ that maximizes $\mathcal{L}(y|\theta)$ is called ML-II estimator and $g(\alpha|\theta_{ML})$ is ML-II based prior.

Important:

- $y_1, \ldots y_T$ **should not** be the same sample used for inference.

- $y_1, \ldots y_T$ could represent past time series information, cross sectional/ cross country information.

- Typically $y_1, \ldots y_T$ is called "Training sample".

4) Priors for DSGE - similar to MLII priors.

- Assume that $g(\alpha) = g_1(\alpha_1)g_2(\alpha_2)....g_q(\alpha_q)$.

- Use a conventional format for the distributions: a Normal, Beta and Gamma for individual parameters. Choose moments in a data based fashion: mean = calibrated parameters, variance: subjective.

Problems:

• Independent priors typically inconsistent with any subjective prior beliefs over joint outcomes. In particular, multivariate priors are often too tight!!

• Calibrated value may be different for different purposes. For example, risk aversion mean is 6-10 to fit the equity premium; close to 1-2 if we

want to fit the reaction of consumption to changes in monetary policy; negative values to fit aggregate lottery revenues. Which one do we use? Same for habit parameters (see Faust and Gupta, 2012)

• Circularity: priors based on the same data used to estimate!! Use calibrated values in a " training sample".

See later del Negro and Schorfheide (2008) for formally choosing data based priors in training samples which are not independent.

Summary

Inputs of the analysis: $g(\alpha),\ f(y|\alpha)$.

Outputs of the analysis:

$g(\alpha|y) \propto f(y|\alpha)g(\alpha)$ (posterior),

$f(y) = \int f(y|\alpha)g(\alpha)$ (marginal likelihood), and

$f(y^{T+\tau}|y^T)$ (predictive density of future observations).

Likelihood should reflect data/ economic theory.

Prior could be non-informative, conjugate, data based (objective).

- In simple examples, $f(y)$ and $g(\alpha|y)$ can be computed analytically.

- In general, can only be computed numerically by Monte Carlo methods.

- If the likelihood is a (log-linearized) DSGE model: always need numerical computations.

3 Posterior simulators

Objects of interest for Bayesian analysis: $E(h(\alpha)) = \int h(\alpha)g(\alpha|y)d\alpha$. Occasionally, can evaluate the integral analytically. In general, it is impossible.

If $g(\alpha|y)$ were available: we could compute $E(h(\alpha))$ with MC methods:

- Draw α^l from $g(\alpha|y)$. Compute $h(\alpha^l)$

- Repeat draw L times. Average $h(\alpha^l)$ over draws.

Example 3.1 *Suppose we are interested in computing* $Pr(\alpha > 0)$. *Draw* α^l *from* $g(\alpha|y)$. *If* $\alpha^l > 0$, *set* $h(\alpha^l) = 1$, *else set* $h(\alpha^l) = 0$. *Draw L times and average* $h(\alpha^l)$ *over draws. The result is an estimate of* $Pr(\alpha > 0)$.

• Approach works because with iid draws the law of large numbers (LLN) insures that sample averages converge to population averages (ergodicity).

• By a central limit theorem (CLT) the difference between sample and population averages has a normal distribution with zero mean and some variance as L grows (numerical standard errors can be used a a measure of accuracy).

- Since $g(\alpha|y)$ is not analytically available, need to use a $g^{AP}(\alpha|y)$, which is similar to ($g(\alpha|y)$, and easy to draw from.

• Normal Approximation

• Basic Posterior simulators (Acceptance and Importance sampling).

• Markov Chain Monte Carlo (MCMC) methods

3.1 Normal posterior analysis

If T is large $g(\alpha|y) \approx f(\alpha|y)$. If $f(\alpha|y)$ is unimodal, roughly symmetric, and α^* (the mode) is in the interior of A:

$$\log g(\alpha|y) \approx \log g(\alpha^*|y) + 0.5(\alpha - \alpha^*)'[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha\partial\alpha'}|_{\alpha=\alpha^*}](\alpha - \alpha^*) \quad (4)$$

Since $g(\alpha^*|y)$ is constant, letting $\Sigma_{\alpha^*} = -[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha\partial\alpha'}{}^{-1}|_{\alpha=\alpha^*}]$

$$g(\alpha|y) \approx N(\alpha^*, \Sigma_{\alpha^*}) \quad (5)$$

- An approximate 100(1-ρ)% highest credible set is $\alpha^* \pm \Phi(\rho/2)I(\alpha^*)^{-0.5}$ where $\Phi(.)$ the CDF of a standard normal.

• Approximation is valid under regularity conditions when $T \to \infty$ or when the posterior kernel is roughly normal. It is highly inappropriate when:

- Likelihood function flat in some dimension ($I(\alpha^*)$ badly estimated).

- Likelihood function is unbounded (no posterior mode exists).

- Likelihood function has multiple peaks.

- α^* is on the boundary of A (quadratic approximation wrong).

- $g(\alpha) = 0$ in a neighborhood of α^* (quadratic approximation wrong).

How do we construct a normal approximation?

A) Find the mode of the posterior.

$$\max \log g(\alpha|y) = \max(\log L(\alpha|y) + \log g(\alpha))$$

- Problem is identical to the one of finding the maximum of a likelihood. The objective function differs.

Two mode finding algorithms:

i) Newton algorithm

- Let $L = \log g(\alpha|y)$ (or $L = \log \breve{g}(\alpha|y)$). Choose α_0.

- Calculate $L' = \frac{\partial L}{\partial \alpha}(\alpha_0)$ $L'' = \frac{\partial^2 L}{\partial \alpha \partial \alpha'}(\alpha_0)$. Approximate L quadratically.

- Set $\alpha^l = \alpha^{l-1} - \gamma(L''(\alpha^{l-1}|y))^{-1}(L'(\alpha^{l-1}|y))$ $\gamma \in (0, 1)$.

- Iterate until convergence i.e. until $||\alpha^l - \alpha^{l-1}|| < \iota$, ι small.

Fast and good if α_0 is good and L close to quadratic. Bad if L'' not positive definite.

ii) Conditional maximization algorithm.

Let $\alpha = (\alpha_1, \alpha_2)$. Start from some $(\alpha_{10}, \alpha_{20})$. Then

- Maximize $L(\alpha_1, \alpha_2)$ with respect to α_1 keeping α_{20} fixed. Let α_1^* the maximizer.

- Maximize $L(\alpha_1, \alpha_2)$ with respect to α_2 keeping $\alpha_1 = \alpha_1^*$ fixed. Let α_2^* the maximizer.

- Iterate on two previous steps until convergence.

- Start from different $(\alpha_{10}, \alpha_{20})$, check if maximum is global.

B) Compute the variance covariance matrix at the mode

- Use the Hessian $\Sigma_{\alpha^*} = -[\frac{\partial^2 \log g(\alpha|y)}{\partial\alpha\partial\alpha'}{}^{-1}|_{\alpha=\alpha^*}]$

C) Approximate the posterior density around the mode: $g^{AP}(\alpha|y) = \mathbb{N}(\alpha^*, \Sigma_{\alpha^*})$.

- If multiple modes are present, find an approximation to each mode, and set $g^{AP}(\alpha|y) = \sum_i \varrho_i \mathbb{N}(\alpha_i^*, \Sigma_{\alpha_i^*})$ where $0 \leq \varrho_i \leq 1$. If modes are clearly separated select $\varrho_i = g(\alpha_i^*|y)|\Sigma_{\alpha_i^*}|^{-0.5}$.

- If the sample is small, use a t-approximation i.e. $g^{AP}(\alpha|y) = \sum_i \varrho_i g(\tilde{\alpha}|y)[\nu + (\alpha - \alpha_i^*)'\Sigma_{\alpha_i}(\alpha - \alpha_i^*)]^{-0.5(k+v)}$ with small ν.

(If $\nu = 1$ t-distribution=Cauchy distribution, large overdispersion. Typically $\nu = 4, 5$ appropriate).

D) To conduct inference, draw α^l from $g^{AP}(\alpha|y)$.

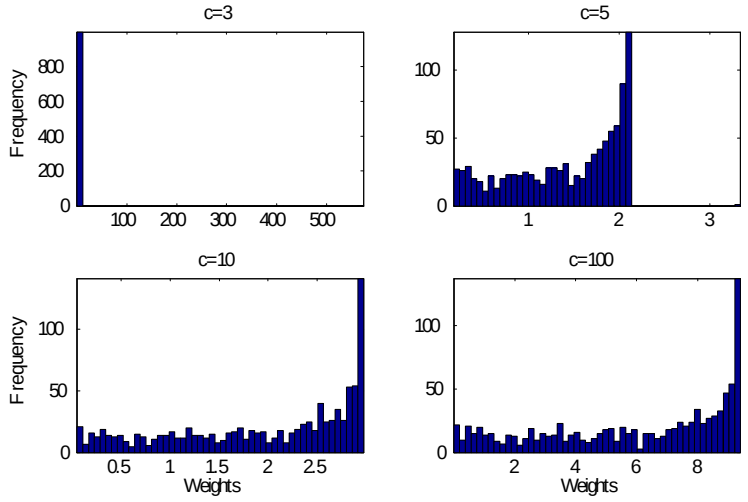
If draws are iid, $E(h(\alpha)) = \frac{1}{L}\sum_l h(\alpha^l)$. Use LLN to approximate any posterior probability contours of $h(\alpha)$, e.g. a 16-84 range is $[h(\alpha^{16}), h(\alpha^{84})]$.

E) Check accuracy of approximation.

Compute *Importance Ratio* $IR^l = \frac{\breve{g}(\alpha^l|y)}{g^{AP}(\alpha^l|y)}$. Accuracy is good if IR^l is constant across l. If not, need to use other techniques.

Note: Importance ratios are not automatically computed in Dynare. Need to do it yourself.

Example 3.2 *True:* $g(\alpha|y)$ *is t(0,1,2). Approximation: N(0,c), where* $c = 3, 5, 10, 100$.



Horizontal axis=importance ratio weights, vertical axis= frequency of the weights.

- Posterior has fat tails relative to a normal (poor approximation).

3.2 Basic Posterior Simulators

- Draw from a general $g^{AP}(\alpha|y)$ (not necessarily normal).
- Non-iterative methods - $g^{AP}(\alpha|y)$ is fixed across draws.
- Work well when IR^l is roughly constant across draws.

A) Acceptance sampling

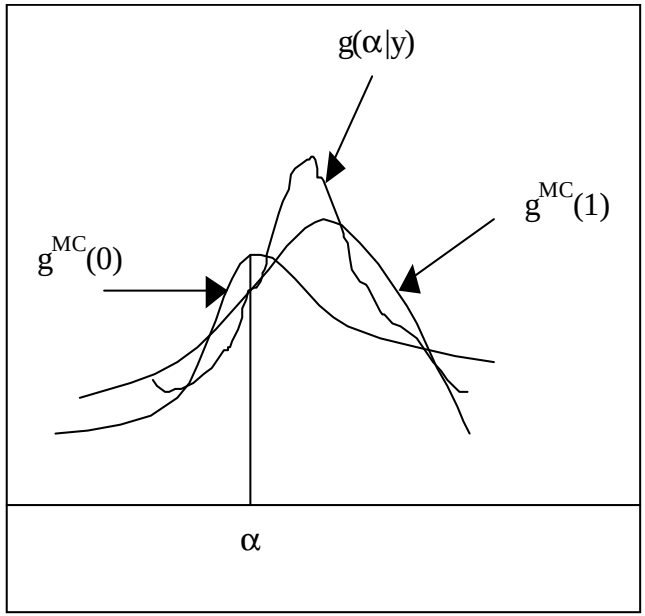
B) Importance sampling

3.3 Markov Chain Monte Carlo Methods

• Problem with basic simulators: approximating density is selected once and for all. If mistakes are made, they stay. With MCMC location of approximating density changes as iterations progress.

• Idea: Suppose n states $(x_1, \ldots x_n)$. Let $P(i,j) = Pr(x_{t+1} = x_j | x_t = x_i)$ and let $\mu(t) = (\mu_{1t}, \ldots \mu_{nt})$ be the unconditional probability at t of each state n. Then $\mu(t+1) = P\mu(t) = P^t\mu(0)$ and μ is an equilibrium (ergodic, steady state, invariant) distribution if $\mu = \mu P$.

Set $\mu = g(\alpha|y)$, choose some initial density $\mu(0)$ and some transition P across states. If conditions are right, iterate from $\mu(0)$ and limiting distribution is $g(\alpha|y)$, the unknown posterior.



• Under general conditions, the ergodicity of P insures consistency and asymptotic normality of estimates of any $h(\alpha)$.

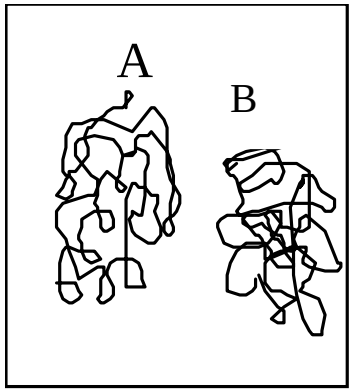
Need a transition $P(\alpha, A)$, where A is some set, such that $||P(\alpha, A) - \mu(\alpha)|| \rightarrow 0$ in the limit. For this need that the chain associated with P :

- is irreducible, i.e. it has no absorbing state.

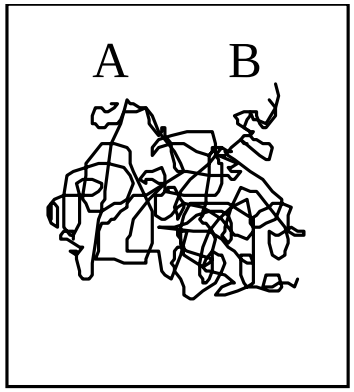
- is aperiodic, i.e. it does not cycle across a finite number of states.

- it is Harris recurrent, i.e. each cell is visited an infinite number of times with probability one.

Bad draws



Good draws



Result 1: A reversible Markov chain, has an ergodic distribution (existence). (if $\mu_i P_{i,j} = \mu_j P_{j,i}$ then $(\mu P)_j = \sum_j \mu_i P_{i,j} = \sum_i \mu_j P_{j,i} = \mu_j \sum_i P_{j,i} = \mu_j$.)

Result 2: (Tierney (1994)) (uniqueness) If a Markov chain is Harris recurrent and has a proper invariant distribution. $\mu(\alpha)$, $\mu(\alpha)$ is unique.

Result 3: (Tierney(1994)) (convergence) If a Markov chain with invariant $\mu(\alpha)$ is Harris recurrent and aperiodic, for all $\alpha_0 \in A$ and all A, as $L \to \infty$.

- $||P^L(\alpha_0, A) - \mu(\alpha)|| \to 0$, $||.||$ is the total variation distance.

- For all $h(\alpha)$ absolutely integrable with respect to $\mu(\alpha)$.

- $\lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} h(\alpha^l) \stackrel{a.s.}{\to} \int h(\alpha)\mu(\alpha)d\alpha$.

If chain has a finite number of states, it is sufficient for the chain to be irreducible, Harris recurrent and aperiodic that $P(\alpha^l \in A_1 | \alpha^{l-1} = \alpha_0, y) > 0$, all $\alpha_0, A_1 \in A$.

- Can dispense with the finite number of state assumption.

- Can dispense with the first order Markov assumption.

General simulation strategy:

- Choose starting values α_0, choose a P with the right properties.
- Run MCMC simulations.
- Check convergence.
- Summarize results i.e compute $h(\alpha)$.

1) MCMC methods generate draws which are *correlated* (with normal/basic simulators, posterior draws are iid).

2) MCMC methods generate draws from posterior only after a burn-in period (with normal/basic simulators, first draw is from the posterior).

3) MCMC can be used to explore intractable likelihoods using "data augmentation" technique (non-bayesian method).

4) MCMC methods only need the kernel $\breve{g}(\alpha|y)$ (no knowledge of the normalizing constants is needed).

3.3.1 Metropolis-Hastings algorithm

MH is a general purpose MCMC algorithm that can be used when faster methods (such as the Gibbs sampler) are either not usable or difficult to implement.

Starts from an arbitrary transition function $q(\alpha^{\dagger}, \alpha^{l-1})$, where $\alpha^{l-1},\ \alpha^{\dagger} \in A$ and an arbitrary $\alpha^{0} \in A$. For each $l = 1, 2, \ldots L$.

- Draw $\alpha^{\dagger}$ from $q(\alpha^{\dagger}, \alpha^{l-1})$ and draw $\varpi \sim U(0, 1)$.

- If $\varpi < \mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) = [\frac{\breve{g}(\alpha^{\dagger}|Y)q(\alpha^{\dagger}, \alpha^{l-1})}{\breve{g}(\alpha^{l-1}|Y)q(\alpha^{l-1}, \alpha^{\dagger})}]$, set $\alpha^{\ell} = \alpha^{\dagger}$.

- Else set $\alpha^{\ell} = \alpha^{\ell-1}$.

These iterations define a mixture of continuous and discrete transitions:

$$\begin{aligned} P(\alpha^{l-1},\ \alpha^{l}) &= q(\alpha^{l-1},\alpha^{l})\mathfrak{E}(\alpha^{l-1},\alpha^{l}) \qquad \text{if } \alpha^{l} \neq \alpha^{l-1} \\ &= 1 - \int_{A} q(\alpha^{l-1},\alpha)\mathfrak{E}(\alpha^{l-1},\alpha)d\alpha \quad \text{if } \alpha^{l} = \alpha^{l-1} \end{aligned} \tag{6}$$

$P(\alpha^{l-1},\ \alpha^{l})$ satisfies the conditions needed for existence, uniqueness and convergence.

• Idea: Want to sample from highest probability region but want to visit as much as possible the parameter space. How to do it? Choose an initial vector and a candidate, compute kernel of posterior at the two vectors. If you go uphill, keep the draw, otherwise keep the draw with some probability.

If $q(\alpha^{l-1}, \alpha^{\dagger}) = q(\alpha^{\dagger}, \alpha^{l-1})$, (Metropolis version of the algorithm) $\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) = \frac{\breve{g}(\alpha^{l-1}|Y)}{\breve{g}(\alpha^{\dagger}|Y)}$. If $\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) > 1$, the chain moves to $\alpha^{\dagger}$. Hence, keep the draw if you move uphill. If the draw moves you downhill stay at α^{l-1} with probability $1 - \mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger})$, and explore new areas with probability equal to $\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger})$.

Important: $q(\alpha^{l-1}, \alpha^{\dagger})$ is not necessarily equal (proportional) to posterior - histograms of draws not equal to the posterior. This is why we use a scheme which accepts more in the regions of high probability.

How do you choose $q(\alpha^{l-1}, \alpha^{\dagger})$ (the transition probability)?

- Typical choice: random walk chain. $q(\alpha^{\dagger}, \alpha^{l-1}) = q(\alpha^{\dagger} - \alpha^{l-1})$, and $\alpha^{\dagger} = \alpha^{l-1} + v$ where $v \sim \mathbb{N}(0, \sigma_v^2)$. To get "reasonable" acceptance rates adjust σ_v^2. Often $\sigma_v^2 = c * \Omega_\alpha, \Omega_\alpha = [-g''(\alpha^*|y)]^{-1}$. Choose c.

Alternatives:

- Reflecting random walk: $\alpha^{\dagger} = \mu + (\alpha^{l-1} - \mu) + v$

- Independent chain $q(\alpha^{\dagger}, \alpha^{l-1}) = \bar{q}(\alpha^{\dagger}), \mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) = min[\frac{w(\alpha^{\dagger})}{w(\alpha^{l-1})}, 1]$, where $w(\alpha) = \frac{g(\alpha|Y)}{\bar{q}(\alpha)}$. Monitor both the location and the shape of $\bar{q}$ to insure reasonable acceptance rates. Standard choices for $\bar{q}$ are normal and t.

• General rule for selecting q. A good q must:

a) be easy to sample from

b) be such that it is easy to compute $\mathfrak{E}$.

c) each move goes a reasonable distance in parameter space but does not reject too frequently (ideal rejection rate 30-50%).

Implementation issues

A) How to draw samples?

- Produce one sample (of dimension $n * L + \bar{L}$). Throw away initial $\bar{L}$ observations. Keep only elements $(L, 2L, \ldots, n * L)$ (to eliminate the serial correlation of the draws).

- Produces n samples of $\bar{L} + L$ elements. Use last L observations in each sample for inference.

- Dynare setup to produce n samples, keep the last 25 percent of the draws. **Careful: Need to make sure that with 75 percent of the draws the chain has converged**.

B) How long should be $\bar{L}$? How do you check convergence?

- Start from different α^0. Check if sample you keep, for a given $\bar{L}$, has same properties (Dynare approach).

- Choose two points, $\bar{L}_1 < \bar{L}_2$; compute distributions/moments of α after these points. If visually similar, algorithm has converged at $\bar{L}_1$. Could this recursively $\rightarrow$ CUMSUM statistic for mean, variance, etc.(checks if it settles down, no testing required).

For simple problems $\bar{L} \approx 50$ and $L \approx 200$.

For DSGEs $\bar{L} \approx 100,000 - 200,000$ and $L \approx 500,000$. If Multiple modes are present L could be even larger.

C) Inference : easy.

- Weak Law of Large Numbers $E(h(\alpha)) \approx \frac{1}{j}\sum_{j=1}^{n} h(\alpha^{jL})$, where α^{jL} is the $j * L$-th observation drawn after $\bar{L}$ iterations are performed.

- $E(h(\alpha)h(\alpha)') = \sum_{-J(L)}^{J(L)} w(\tau)ACF_h(\tau)$; $ACF_h(\tau) =$ autocovariance of $h(\alpha)$ for draws separated by τ periods; $J(L)$ function of L, $w(\tau)$ a set of weights.

- Marginal density $(\alpha_k^1, \ldots \alpha_k^L)$: $g(\alpha_k|y) = \frac{1}{L}\sum_{j=1}^{L} g(\alpha_k|y, \alpha_{k'}^j, k' \neq k)$.

- Predictive inference $f(y_{t+\tau}|y_t) = \int f(y_{t+\tau}|y_t, \alpha)g(\alpha|y_t)d\alpha$.

- Model comparisons: compute marginal likelihood numerically.

4 Robustness

• Typically prior chosen to make calculation convenient. How sensitive are results to prior choice?

• Typical (brute force) approach: repeat estimation for different priors (inefficient).

• Alternative.

i) Select an alternative prior $g_1(\alpha)$ with support included in $g(\alpha)$.

ii) Let $w(\alpha) = \frac{g(\alpha)}{g_1(\alpha)}$. Then any $h_1(\alpha) = \int (h(\alpha)w(\alpha)dg_1(\alpha)$ can be approximated using $h_1(\alpha) \approx \frac{\frac{1}{L}\sum_l w(\alpha^l)h(\alpha^l)}{\sum_l w(\alpha^l)}$.

•Just need the original output obtained and a set of weigths!

Example 4.1 $y_t = x_t\alpha + u_t$ $u_t \sim (0, \sigma^2)$. *Suppose* $g(\alpha)$ *is* $\mathbb{N}(0, 10)$. *Then* $g(\alpha|Y)$ *is normal with mean* $\tilde{\alpha} = \tilde{\Sigma}^{-1}(0.1 + \sigma^{-2}x'x\alpha_{ols})$ *and variance* $\tilde{\Sigma} = 0.1 + \sigma^{-2}x'x$, . *If one wishes to examine how forecasts of the model change when the prior variances changes (for example to 5) two alternatives are possible:*

(a) draw from normal $g(\alpha|Y)$ *which has mean* $\tilde{\alpha}_1 = \tilde{\Sigma}_1^{-1}(0.2+\sigma^{-2}x'x\alpha_{ols})$ *and variance* $\tilde{\Sigma} = 0.2 + \sigma^{-2}x'x$, *and compare forecasts.*

(b) Weight draws from the initial posterior distribution with $\frac{g(\alpha)}{g_1(\alpha)}$ *where* $g_1(\alpha)$ *is* $N(0, 5)$.

5 Bayesian estimation of DSGE models

Why using Bayesian methods to estimate DSGE models?

1) Hard to include non-sample information in classical ML (a part from range of possible values).

2) Classical ML is justified only if the model is the GDP of the actual data. Can use Bayesian methods for misspecified models (economic inference may be problematic, no problem for statistical inference).

3) Can incorporate prior uncertainty about parameters and models.

General Principles:

• Use the fact that (log-)linearized DSGE models are state space models whose reduced form parameters α are nonlinear functions of structural θ. Compute the likelihood via the Kalman filter.

• Posterior of θ can be obtained using MH algorithm.

• Use posterior output to compute the marginal likelihood, Bayes factors and any posterior function of the parameters (impulse responses, ACF, turning point predictions, forecasts, etc.).

• Check robustness to the choice of prior.

General algorithm: Given θ_0

[1.] Construct a log-linear solution of the DSGE economy.

[2.] Specify prior distributions $g(\theta)$.

[3.] Transform the data to make sure that is conformable with the model.

[4.] Compute likelihood via Kalman filter.

[5.] Draw sequences for θ using MH algorithm. Check convergence.

[6.] Compute marginal likelihood and compare it to the one of alternative models. Compute Bayes factors.

[7.] Construct statistics of interest. Use loss-based evaluation of discrepancy model/data.

[8.] Perform robustness exercises.

Step 1.: can have nonlinear state space models (see later and e.g. Amisano and Tristani (2006), Rubio and Villaverde (2009)) or value function problems (see Bi and Traum (2012)) but computations much more complex.

System are typically singular! Need to:

i) add measurement errors if want to use all observables (where to put measurement error? In all variables or just enough to complete the probability space?)

ii) find a way to reduce the dimensionality of the system (substituting equations before the solution is computed).

iii) choose the observables optimally (see Canova et al. (2012)).

iv) invent new structural shocks.

In Step 3. transformations are needed because the model is typically solved in deviation from the steady states. Need to eliminate from the data any long run component. How do you do it? Many ways of doing this (see Canova, 2010) all unsatisfactory.

Step 4 is typically the most computationally intensive step. Considerable gains if this is efficiently done.

In step 5. Given θ^l

i) Draw a $\theta^\dagger$ from the $\mathfrak{P}(\theta^\dagger \theta^l)$. Solve the model.

ii) Use the KF to compute the likelihood.

iii) Evaluate the posterior kernel at the draw $\breve{g}(\theta^{\dagger}|y) = f(y|\theta^{\dagger})g(\theta^{\dagger})$.

iv) Evaluate the posterior kernel at θ^{l} i.e $\breve{g}(\theta^{0}|y) = f(y|\theta^{l})g(\theta^{l})$.

v) Compute $IR = \frac{\breve{g}(\theta^{\dagger})}{\breve{g}(\theta^{l})}\frac{\mathfrak{P}(\theta^{l},\theta^{\dagger})}{\mathfrak{P}(\theta^{\dagger},\theta^{l})}$. If $IR > 1$ set $\theta^{l+1} = \theta^{\dagger}$.

vi) Else draw $\varpi \sim U(0,1)$. If $\varpi < IR$ set $\theta^{l+1} = \theta^{\dagger}$ otherwise set $\theta^{l+1} = \theta^{l}$.

vii) Repeat i)-vi) $\bar{L} + nL$ times. Throw away $\bar{L}$ draws. Keep one every n for inference.

In Step 6. use a modified harmonic mean estimator i.e. approximate $\mathcal{L}(y_t|\mathcal{M}_i)$ using $[\frac{1}{L}\sum_l \frac{f(\alpha_l^i)}{\mathcal{L}(y_t|\alpha_l^i,\mathcal{M}_i)g(\alpha_l^i|\mathcal{M}_i)}]^{-1}$ where α_i^l is the draw l of the

parameters α of model i and f is a density with tails thicker than a normal. If $f(\alpha_l^i) = 1$ we have a simple harmonic mean estimator.

Competitors could be a more densely parametrized structural model (nesting the interested one) or more densely parametrized reduced form model (e.g. VAR or a BVAR).

Bayes factors can be computed numerically or via Laplace approximations (to decrease computational burden in large scale systems).

In step 7 Estimate marginal/ joint posteriors using kernel methods. Compute point estimate and credible sets. Compute continuous functions $h(\theta)$ of interest. Set up a loss function. Compare models using the risk function.

In step 8. Reweight the draws appropriately.

Example 5.1 *(One sector growth model)*

- *Analytic solution if* $U(c,l) = \ln c$ *and* $\delta = 1$. *Equations are:*

$$\begin{aligned} K_{t+1} &= (1-\eta)\beta A K_t^{1-\eta}\zeta_t + u_{1t} && (7)\\ GDP_t &= A K_t^{1-\eta}\zeta_t + u_{2t} && (8)\\ c_t &= \eta\beta GDP_t + u_{3t} && (9)\\ r_t &= (1-\eta)\frac{GDP_t}{K_t} + u_{4t} && (10) \end{aligned}$$

- ζ_t *technology shock,* u_{jt} *measurement errors added to avoid singularity.*

Parameters: β*: is the discount factor,* $1-\eta$*: the share of capital in production,* σ^2*: variance of technology shock, A: constant in the production function.*

Simulate 1000 points from using $k_0 = 100.0$ *using* $A = 2.86; 1 - \eta = 0.36; \beta = 0.99, \sigma^2 = (0.07)^2$.

Assume $u_{1t} \sim \mathbb{N}(0, 0.1^2)$; $u^m_{2t} \sim \mathbb{N}(0, 0.06^2)$; $u^m_{3t} \sim \mathbb{N}(0, 0.02^2)$; $u^m_{4t} \sim \mathbb{N}(0, 0.08^2)$; *(Note: lots of measurement error!)*

- Keep last 160 as data (to mimic about 40 years of quarterly data).

Interested in $(1 - \eta), \beta$ *i.e (treat* σ^2, A *as fixed).*

Use (9)-(10) to identify the parameters from the data.

Priors: $(1-\eta) \sim$ *Beta(3,5);* $\beta \sim$ *Beta(98,2) (NOTATION DIFFERENT FROM DYNARE)*

Mean of a Beta(a,b) is $(a/a+b)$ *and the variance of a Beta(a,b)is* $ab/[(a+b)^2 * (a+b+1)]$. *Thus prior mean of* $1-\eta = 0.37$, *prior variance* 0.025; *prior mean of* $\beta = 0.98$, *prior variance* 0.0001.

Let $\theta = (1-\eta, \beta)$ *Use random walk to draw* $\theta^{\dagger}$, *i.e.* $\theta^{\dagger} = +\theta^{l-1} + e^{\dagger}$, μ *is the mean and* e_i *is* $U(-0.08, 0.08)$ *for* β *and* $U(-0.06, 0.06)$ *for* η *(roughly about 28% acceptance rate).*

Draw 10000 replications from the posterior kernel. Convergence is fast.

Keep last 5000; use one every 5 for inference.

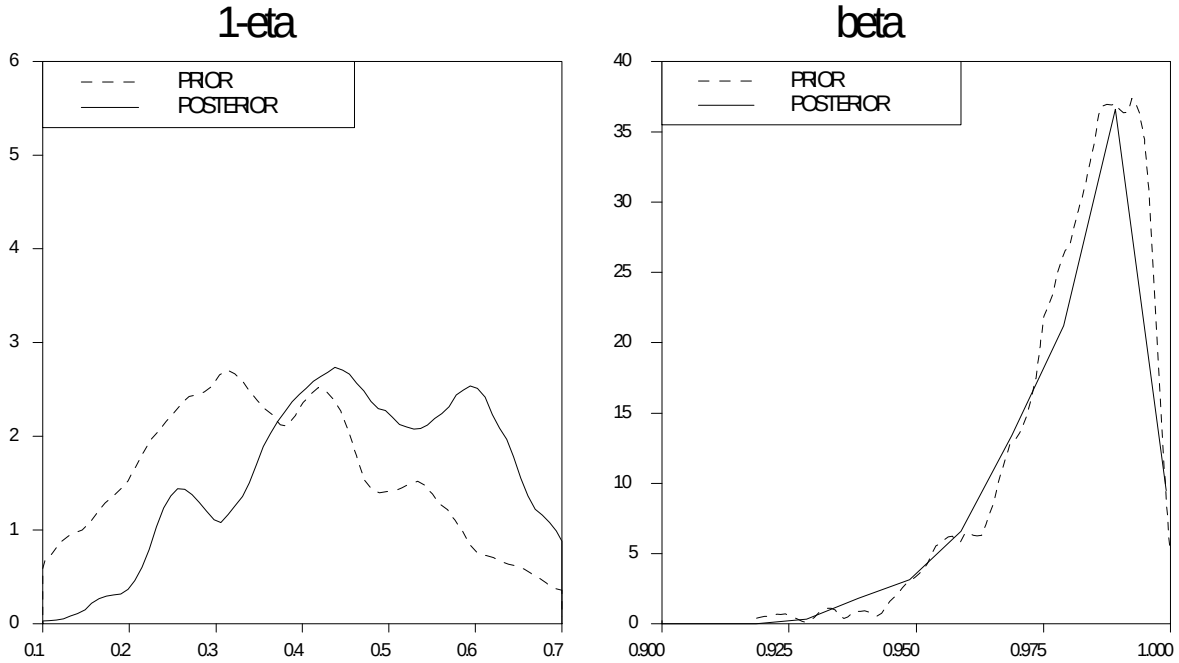


Figure 4: Priors and Posteriors, RBC model

- Prior for β sufficiently loose, posterior similar, data is not very formative.

-Posteriors centered around the true parameters,large dispersion.

Variances/covariances

	true	*posterior 68% range*
var(c)	*0.24*	*[0.11, 0.27]*
var(y)	*0.05*	*[0.03, 0.11]*
cov(c,y)	*0.0002*	*[0.0003, 0.0006]*

Wrong model

- Simulate data from model with habit $\gamma = 0.8$

- Estimate model conditioning on $\gamma = 0$.

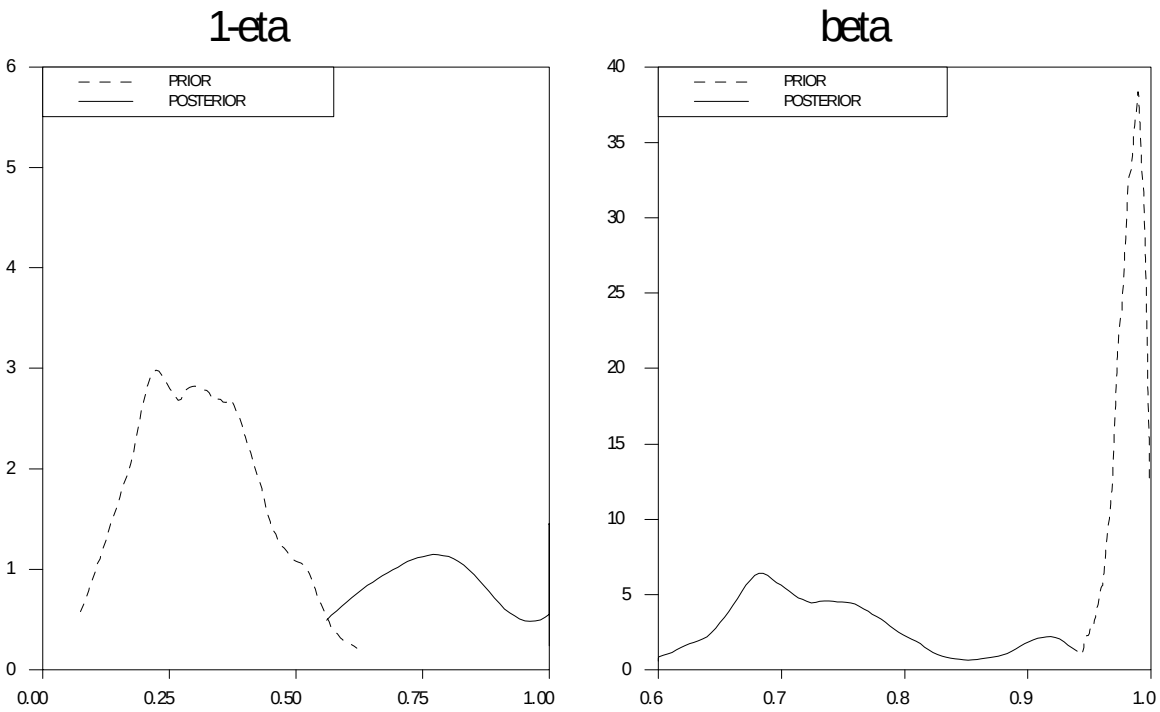


Figure 5: Priors and Posteriors, wrong model

Example 5.2 *(New Keynesian model)*

$$gap_t = E_t gap_{t+1} - \frac{1}{\varphi}(r_t - E_t\pi_{t+1}) + g_t \quad (11)$$

$$\pi_t = \beta E_t \pi_{t+1} + \kappa gap_t + v_t \quad (12)$$

$$r_t = \phi_r r_{t-1} + (1-\phi_r)(\phi_\pi \pi_{t-1} + \phi_{gap} gap_{t-1}) + e_t \quad (13)$$

$\kappa = \frac{(1-\zeta_p)(1-\beta\zeta_p)(\varphi+\vartheta_N)}{\zeta_p}$; ζ_p = *degree of (Calvo) stickiness,* β = *discount factor,* φ = *risk aversion,* ϑ_N = *elasticity of labor supply.* g_t *and* v_t *are AR(1) with persistence* ρ_g, ρ_v *and variances* σ_g^2, σ_v^2; $e_t \sim iid(0, \sigma_r^2)$.

$\theta = (\beta, \varphi, \vartheta_l, \zeta_p, \phi_\pi, \phi_{gap}, \phi_r, \rho_g, \rho_v, \sigma_v^2, \sigma_g^2, \sigma_r^2)$.

Assume $g(\theta) = \prod g(\theta_i)$

Assume $\beta \sim Beta(98, 3), \varphi \sim \mathbb{N}(1, 0.375^2), \vartheta_N \sim \mathbb{N}(2, 0.75^2), \quad \zeta_p \sim Beta(9, 3), \phi_r \sim Beta(6, 2), \phi_\pi \sim Normal(1.5, 0.1^2),$
$\phi_{gap} \sim \mathbb{N}(0.5, 0.05^2), \ \rho_g \sim Beta(17, 3), \rho_v \sim Beta(17, 3) \ \sigma_i^2 \sim IG(2, 0.01), i = g, v, r.$

Use US linearly detrended data from 1948:1 to 2002:1 to estimate the model.

Use random walk MH algorithm to draw candidates.

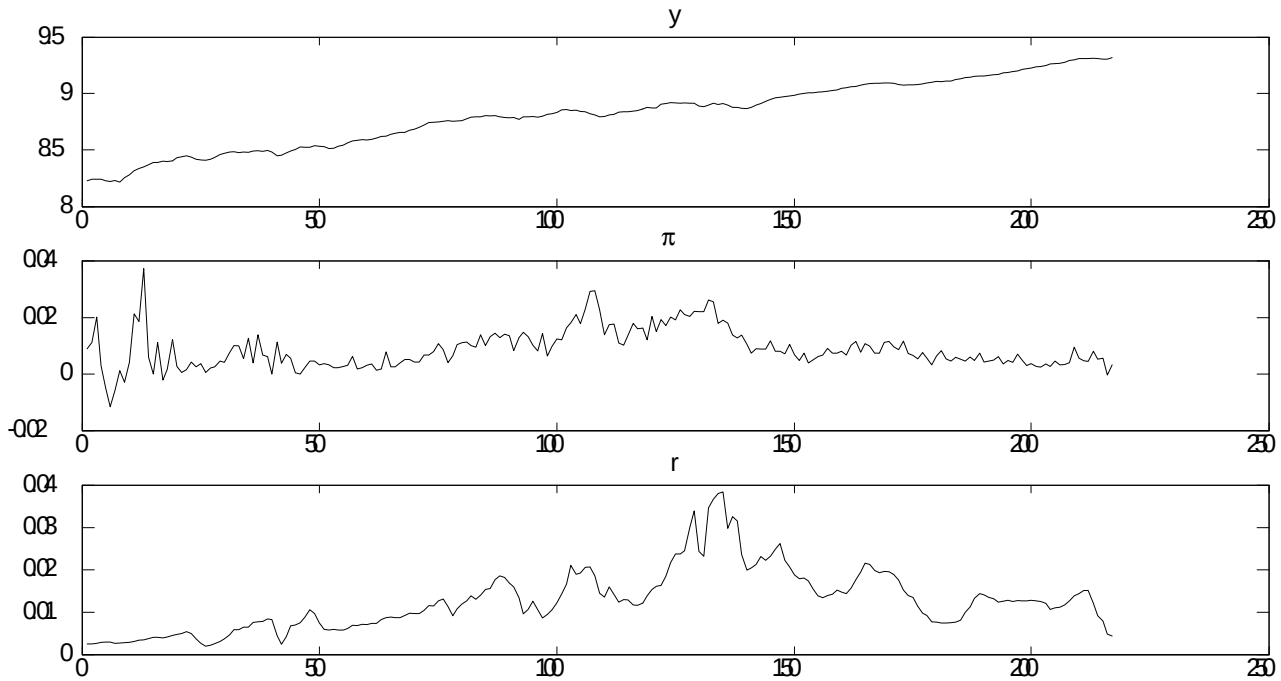


Figure 6: Raw Time series

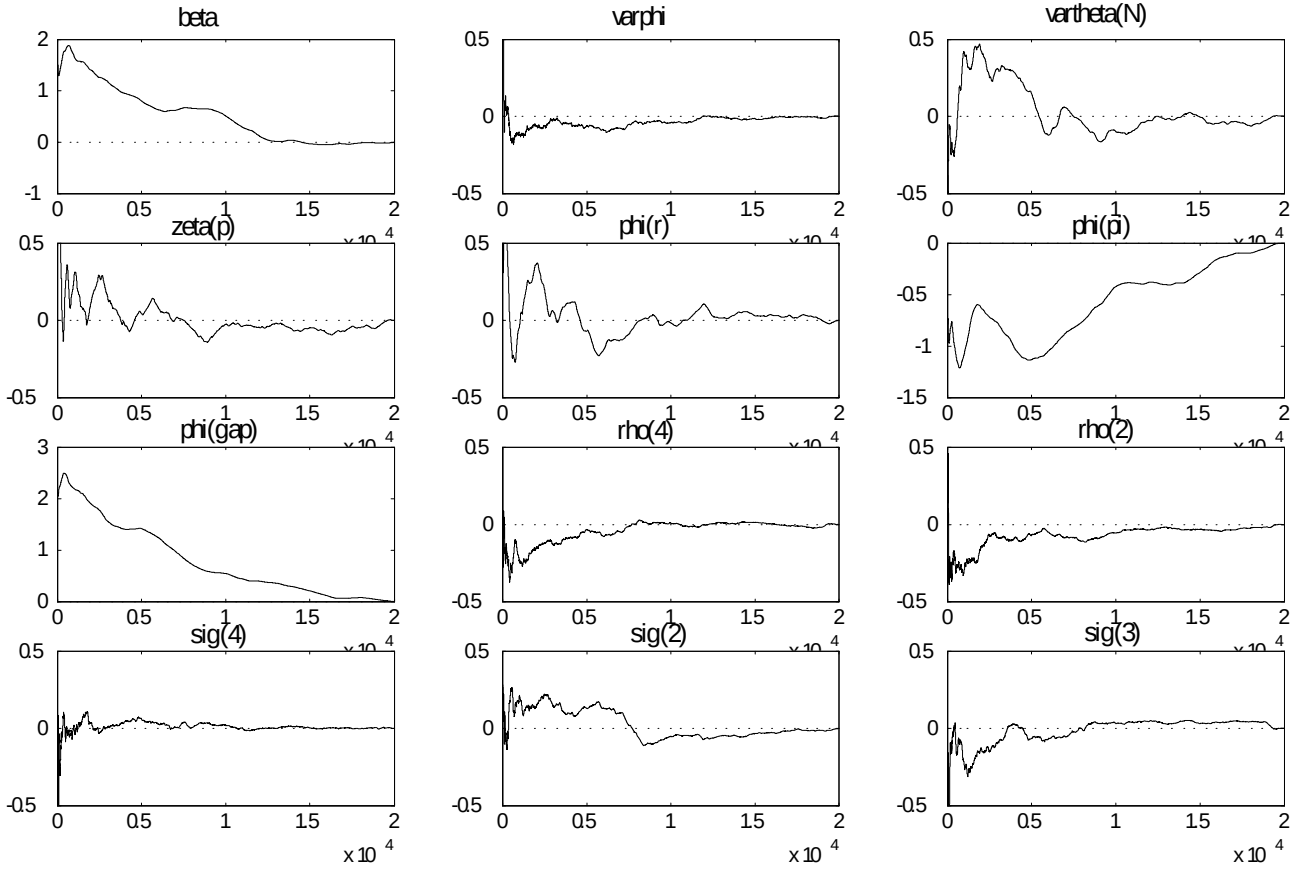


Figure 7: CUMSUM statistics

Prior and Posterior statistics

	Prior		Posterior				
	mean	*std*	*median*	*mean*	*std*	*max*	*min*
β	*0.98*	*0.01*	*0.992*	*0.991*	*0.003*	*0.999*	*0.998*
φ	*1.00*	*0.37*	*0.826*	*0.843*	*0.123*	*1.262*	*0.425*
ϑ_N	*2.00*	*0.75*	*1.825*	*1.884*	*0.768*	*3.992*	*0.145*
ζ_p	*0.75*	*0.12*	*0.743*	*0.696*	*0.195*	*0.997*	*0.141*
ϕ_r	*0.75*	*0.14*	*0.596*	*0.587*	*0.154*	*0.959*	*0.102*
ϕ_π	*1.50*	*0.10*	*1.367*	*1.511*	*0.323*	*2.33*	*1.042*
ϕ_{gap}	*0.5*	*0.05*	*0.514*	*0.505*	*0.032*	*0.588*	*0.411*
ρ_g	*0.85*	*0.07*	*0.856*	*0.854*	*0.036*	*0.946*	*0.748*
ρ_u	*0.85*	*0.07*	*0.851*	*0.851*	*0.038*	*0.943*	*0.754*
σ_g	*0.025*	*0.07*	*0.025*	*0.025*	*0.001*	*0.028*	*0.021*
σ_v	*0.025*	*0.07*	*0.07*	*0.07*	*0.006*	*0.083*	*0.051*
σ_r	*0.025*	*0.07*	*0.021*	*0.021*	*0.005*	*0.035*	*0.025*

- Little information in the data for some parameters (prior and posterior overlap).

- For parameters of the policy rule: posteriors move and not more concentrated.

-Posterior distributions roughly symmetric except for ϕ_π and ζ_p (mean and median coincide).

-Posterior distribution of economic parameters reasonable (except φ).

- Posterior for the AR parameters has a high mean, but no pile up at one.

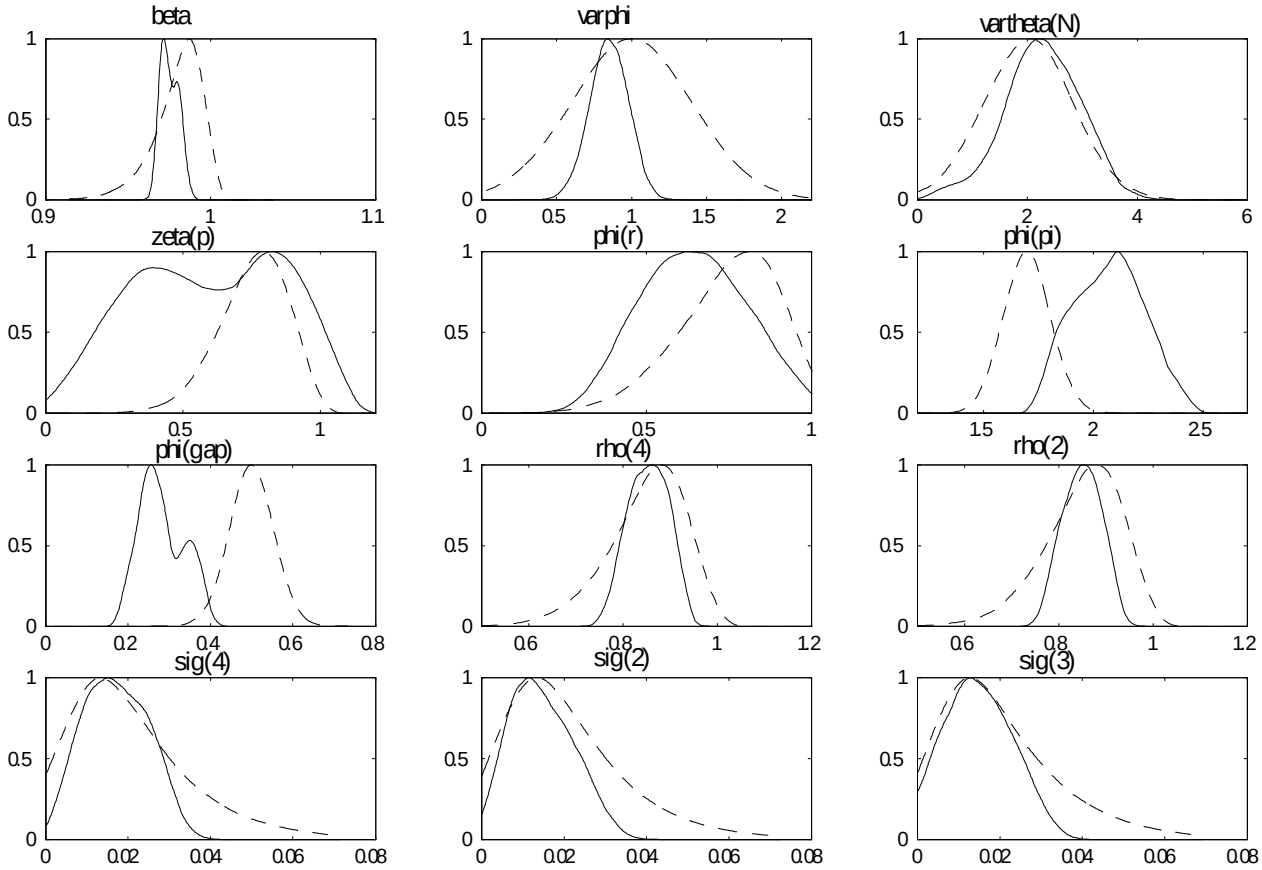


Figure 8: Priors and Posteriors, NK model

Model comparisons

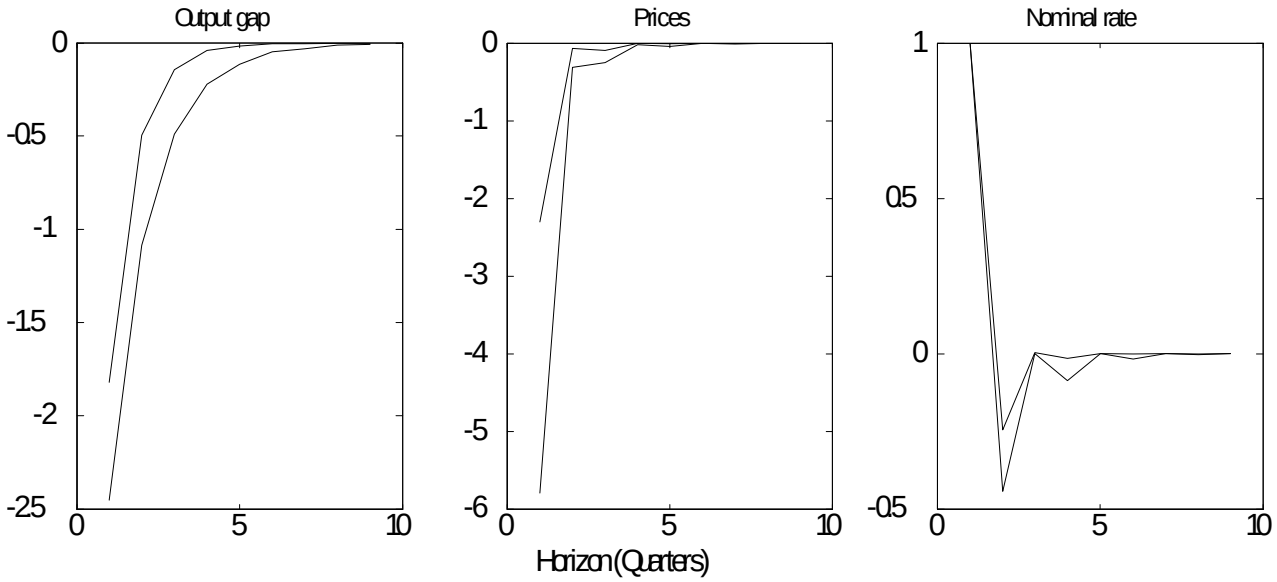
Compare ML against flat prior VAR(3) or a BVAR(3) with Minnesota prior and standard parameters (tightness=0.1, linear lag decay and weight on other variables equal 0.5), both with a constant.

Bayes factor are very small $\approx$ 0.02 in both cases.

- *The restrictions the model imposes are false. Need to add features to the model that make dynamics of the model more similar to those of a VAR(3).*

Posterior analysis

How do responses to monetary shocks look like? No persistence!



How much of the output gap and inflation variance explained by monetary shocks? Almost all!!

5.1 Interpreting results

- Most of the shocks of DSGE models are non-structural (alike to measurement errors). Careful with interpretation and policy analyses with these models (see Chari et al. (2009)).

- A model where "measurement errors" explain a large portion of main macro variables is very suspicious (e.g. in Smets and Wouters (2003) markup shocks dominate).

- If the standard error of one the shocks is large relative to the others: evidence of misspecification.

- Compare estimates with standard calibrated values. Are they sensible? Often yes, but because of tight priors are centered at calibrated values.

5.2 Bayesian methods and identification

Likelihood of a DSGE typically flat. Could be due to marginalization (use only a subset of economic relationships), or to lack of information. Difficult to say a-priori which parameters is underidentified and which is not (since we do not have an analytic solution).

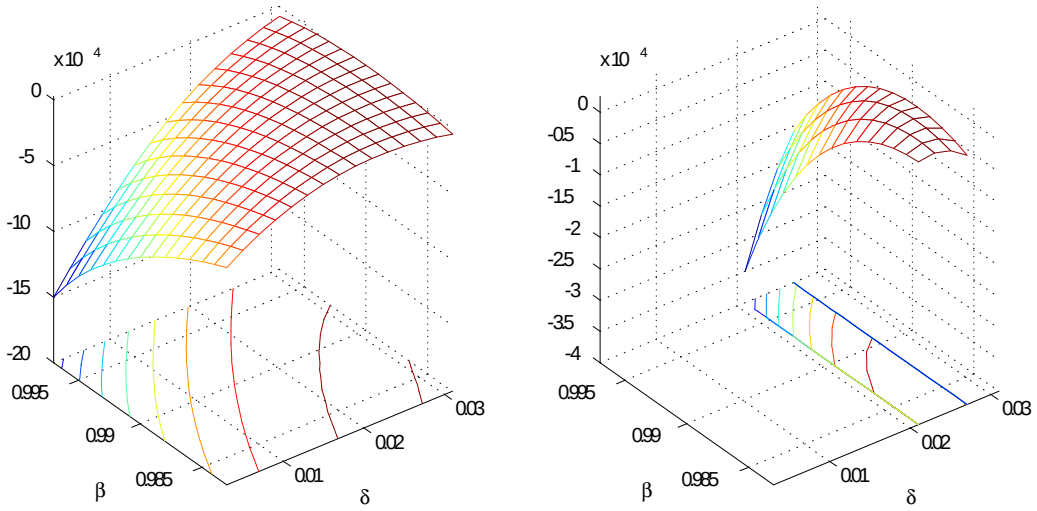
Could go a long way by numerically constructing the likelihood as a function of the parameters (see Canova and Sala (2009)).

Standard remedy when some parameters are hard to identify: calibrate. Problem if parameter not fixed at a consistent estimator $\longrightarrow$ biases could be extensive! (see Canova and Sala (2009)).

Alternative: add a prior. This increases the curvature of the likelihood $\rightarrow$ underidentification may be hidden!. Posterior look nice because the prior does the job!!.

In general if $\mathcal{L}(\theta_1, \theta_2|Y^T) = \bar{\mathcal{L}}(\theta_1|Y^T)$ then $g(\theta_1, \theta_2|Y^T) = g_1(\theta_1|Y^T)$ $g(\theta_2|\theta_1)$, i.e no updating of conditional prior of θ_2.

However, updating possible even if no sample information is present if θ_1, θ_2 are linked by economic or stability conditions!!



Likelihood and Posterior, δ and β in a RBC model

If prior $\approx$ posterior: weak identification or too much data based prior?

6 Topics

6.1 Eliciting Priors from existing information

- Prior distributions for DSGE parameters often arbitrary.

- Prior distribution for individual parameters assumed to be independent: the joint distribution may assign non-zero probability to " unreasonable" regions of the parameter space.

- Prior sometimes set having some statistics in mind (the prior mean is similar to the one obtained in calibration exercises).

- Same prior is used for the parameters of different models. Problem: same prior may generate very different dynamics in different models. Hard to compare the outputs.

Example 6.1 *Let* $y_t = \theta_1 y_{t-1} + \theta_2 + u_t,\ u_t \sim N(0,1)$.. *Suppose* θ_1 *and* θ_2 *are independent and* $p(\theta_1) \sim U(0, 1-\epsilon), \epsilon > 0$; $p(\theta_2|\theta_1) \sim N(\bar{\mu}, \lambda)$.

Since the mean of y_t *is* $\mu = \frac{\theta_2}{1-\theta_1}$, *the prior for* θ_1 *and* θ_2 *imply that* $\mu|\theta_1 \sim N(\bar{\mu}, \frac{\lambda}{(1-\theta_1)^2})$. *Hence, the prior mean of* y_t *has a variance which is increasing in the persistence parameter* θ_1*! Why? Reasonable ?*

Alternative: state a prior for μ, *derive the prior for* θ_1 *and* θ_2 *(change of variables). For example, if* $\mu \sim N(\bar{\mu}, \lambda^2)$ *then* $p(\theta_1) = U(0, 1-\epsilon),\ p(\theta_2|\theta_1) = N(\bar{\mu}(1-\theta_1), \lambda^2(1-\theta_1)^2)$. *Note here that the priors for* θ_1 *and* θ_2 *are correlated.*

Suppose you want to compare the model with $y_t = \theta + u_t,\ u_t \sim N(0,1)$. *If* $p(\theta) = N(\bar{\mu}, \lambda^2)$ *the two models are immediately comparable. If, instead, we had assumed independent priors for* $p(\theta_1)$ *and* $p(\theta_2)$, *the two models would not be comparable (standard prior has weird predictions for the prior of the mean of* y_t*).*

- Del Negro and Schorfheide (2008): elicit priors consistent with some distribution of statistics of actual data (see also Kadane et al. (1980)). Basic idea:

i) Let θ be a set of DSGE parameters. Let S_T be a set of statistics obtained in the data with T observations and σ_S be the standard deviation of these statistics (which can be computed using asymptotic distributions or small sample devices, such as bootstrap or MC methods).

ii) Let $S_N(\theta)$ be the same set of statistics which are measurable from the model once θ is selected using N observations. Then

$$S_T = S_N(\theta) + \eta \quad \eta \sim (0, \Sigma_{TN}) \tag{14}$$

where η is a set of measurement errors.

Note

i) in calibration exercises $\Sigma_{TN} = 0$ and S_T are averages of the data.

ii) in SMM: $\Sigma_{TN} = 0$ and S_T are generic moments of the data.

Then $L(S_N(\theta)|S_T) = p(S_T|S_N(\theta))$, where the latter is the conditional density in (14).

Given any other prior information $\pi(\theta)$ (which is not based on S_T) the prior for θ is

$$p(\theta|S_T) \propto L(S_N(\theta)|S_T)\pi(\theta) \tag{15}$$

- $dim(S_T) \geq dim(\theta)$: overidentification is possible.

- Even if Σ_{TN} is diagonal, $S_N(\theta)$ will induce correlation across θ_i.

-Information used to construct S_T should be **different** than information used to estimate the model. Could be data in a training sample or could be data from a different country or a different regime (see e.g. Canova and Pappa (2007)).

- Assume that η are normal why? Make life easy, Could also use other distributions, e.g. uniform, t.

- What are the S_T? Could be steady states, autocorrelation functions, etc. What S_T is depends on where the parameters enters.

Example 6.2

$$\max_{(c_t,K_{t+1},N_t)} \quad E_0 \sum_t \beta^t \frac{(c_t^{\vartheta}(1-N_t)^{1-\vartheta})^{1-\varphi}}{1-\varphi} \tag{16}$$

$$G_t + c_t + K_{t+1} = GDP_t + (1-\delta)K_t \tag{17}$$

$$\ln \zeta_t = \bar{\zeta} + \rho_z \ln \zeta_{t-1} + \epsilon_{1t} \quad \epsilon_{1t} \sim (0, \sigma_z^2) \tag{18}$$

$$\ln G_t = \bar{G} + \rho_g \ln G_{t-1} + \epsilon_{4t} \quad \epsilon_{4t} \sim (0, \sigma_g^2) \tag{19}$$

$$GDP_t = \zeta_t K_t^{1-\eta} N_t^{\eta} \tag{20}$$

K_0 are given, c_t is consumption, N_t is hours, K_t is the capital stock. Let G_t be financed with lump sum taxes and λ_t the Lagrangian on (17).

The FOC are ((24) and (25) equate factor prices and marginal products)

$$\lambda_t = \vartheta c_t^{\vartheta(1-\varphi)-1}(1-N_t)^{(1-\vartheta)(1-\varphi)} \quad (21)$$

$$\lambda_t \eta \zeta_t k_t^{1-\eta} N_t^{\eta-1} = -(1-\vartheta) c_t^{\vartheta(1-\varphi)}(1-N_t)^{(1-\vartheta)(1-\varphi)-1} \quad (22)$$

$$\lambda_t = E_t \beta \lambda_{t+1}[(1-\eta)\zeta_{t+1} K_{t+1}^{-\eta} N_{t+1}^{\eta} + (1-\delta)] \quad (23)$$

$$w_t = \eta \frac{GDP_t}{N_t} \quad (24)$$

$$r_t = (1-\eta)\frac{GDP_t}{K_t} \quad (25)$$

Using (21)-(22) we have:

$$-\frac{1-\vartheta}{\vartheta}\frac{c_t}{1-N_t} = \eta \frac{GDP_t}{N_t} \quad (26)$$

Log linearizing the equilibrium conditions

$$\hat{\lambda}_t - (\vartheta(1-\varphi)-1)\hat{c}_t + (1-\vartheta)(1-\varphi)\frac{N^{ss}}{1-N^{ss}}\hat{N}_t = 0 \quad (27)$$

$$\hat{\lambda}_{t+1} + \frac{(1-\eta)(GDP/K)^{ss}}{(1-\eta)(GDP/K)^{ss}+(1-\delta))}(\widehat{GDP}_{t+1} - \hat{K}_{t+1}) = \hat{\lambda}_t \quad (28)$$

$$\frac{1}{1-N^{ss}}\hat{N}_t + \hat{c}_t - \widehat{gdp}_t = 0 \quad (29)$$

$$\hat{w}_t - \widehat{GDP}_t + \hat{n}_t = 0 \quad (30)$$

$$\hat{r}_t - \widehat{GDP}_t + \hat{k}_t = 0 \quad (31)$$

$$\widehat{GDP}_t - \hat{\zeta}_t - (1-\eta)\hat{K}_t - \eta\hat{N}_t = 0 \quad (32)$$

$$(\frac{g}{GDP})^{ss}\hat{g}_t + (\frac{c}{GDP})^{ss}\hat{c}_t + (\frac{K}{GDP})^{ss}(\hat{K}_{t+1} - (1-\delta)\hat{K}_t) - \widehat{GDP}_t = 0 \quad (33)$$

(32) and (33) are the production function and resource constraint.

Four types of parameters appear in the log-linearized conditions:

i.) Technological parameters (η, δ).

ii) Preference parameters $(\beta, \varphi, \vartheta)$.

iii) Steady state parameters $(N^{ss}, (\frac{c}{GDP})^{ss}, (\frac{K}{GDP})^{ss}, (\frac{g}{GDP})^{ss})$.

iv) Parameters of the driving process $(\rho_g, \rho_z, \sigma_z^2, \sigma_g^2)$.

Question: How do we set a prior for these 13 parameters?

The steady state of the model (using (23)-(26)-(17)) is:

$$\frac{1-\vartheta}{\vartheta}\left(\frac{c}{GDP}\right)^{ss} = \eta\frac{1-N^{ss}}{N^{ss}} \quad (34)$$

$$\beta[(1-\eta)\left(\frac{GDP}{K}\right)^{ss}+(1-\delta)] = 1 \quad (35)$$

$$\left(\frac{g}{GDP}\right)^{ss}+\left(\frac{c}{GDP}\right)^{ss}+\delta\left(\frac{K}{GDP}\right)^{ss} = 1 \quad (36)$$

$$\frac{GDP}{wc} = \eta \quad (37)$$

$$\frac{K}{i} = \delta \quad (38)$$

Five equations in 8 parameters!! Need to choose.

For example: (34)-(38) determine $(N^{ss}, \left(\frac{c}{GDP}\right)^{ss}, \left(\frac{K}{GDP}\right)^{ss}, \eta, \delta)$ *given* $(\left(\frac{g}{GDP}\right)^{ss}, \beta, \vartheta)$.

Set $\theta_2 = [(\frac{g}{GDP})^{ss}, \beta, \vartheta]$ *and* $\theta_1 = [N^{ss}, (\frac{c}{GDP})^{ss}, (\frac{K}{GDP})^{ss}, \eta, \delta]$

Then if S_{1T} *are steady state relationships, we an use (34)-(38) to construct a prior distribution for* $\theta_1|\theta_2$.

How do we measure uncertainty in S_{1T}*?*

- *Take a rolling window to estimate* S_{1T} *and use uncertainty of the estimate to calibrate var*(η).

- *Bootstrap* S_{1T}*, etc.*

How do we set a prior for θ_2? Use additional information (statistics)!

- $(\frac{g}{GDP})^{ss}$ could be centered at the average G/Y in the data with standard error covering the existing range of variations

- $\beta = (1+r)^{-1}$ and typically $r^{ss} = [0.0075, 0.0150]$ per quarter. Choose a prior centered at around those values and e.g. uniformly distributed.

- ϑ is related to Frish elasticity of labor supply: use estimates of labor supply elasticity to obtain histograms and to select a prior shape.

Note: uncertainty in this case could be data based or across studies (meta uncertainty).

Parameters of the driving process $(\rho_g, \rho_z, \sigma_z^2, \sigma_g^2)$ *do not enter the steady state. Call them* θ_3*. How do we choose a prior for them?*

- ρ_z, σ_z^2 *can be backed out from moments of Solow residual i.e. estimate the variance and the AR(1) of* $\hat{z} = \ln GDP_t - (1-\eta)K_t - \eta N_t$*, once* η *is chosen. Prior for* η *induce a distribution for* $\hat{z}$

- ρ_g, σ_g^2 *backed out from moments government expenditure data.*

Prior standard errors should reflect variations in the data of these parameters.

- For φ (coefficient of relative risk aversion (RRA) is $1-\vartheta(1-\varphi)$) one has two options:

(a) appeal to existing estimates of RRA. Construct a prior which is consistent with the cross section of estimates (e.g. a $\chi^2(2)$ would be ok).

(b) select an interesting moment, say $var(c_t)$ and use

$$var(c_t) = var(c_t(\varphi)|\theta_1, \theta_2, \theta_3) + \eta \tag{39}$$

to back out a prior for φ.

For some parameters (call them θ_5) we have no moments to match but some micro evidence. Then $p(\theta_5) = \pi(\theta_5)$ could be estimated from the histogram of the estimates which are available.

In sum, the prior for the parameters is

$$\begin{aligned} p(\theta) &= p(\theta_1|S_{1T})p(\theta_2|S_{2T})p(\theta_3|S_{3T})p(\theta_4|S_{4T}) \\ & \quad \pi(\theta_1)\pi(\theta_2)\pi(\theta_3)\pi(\theta_4)\Pi(\theta_5) \end{aligned} \tag{40}$$

- If we had used a different utility function, the prior e.g. for θ_1, θ_4 would be different. **Prior for different models/parameterizations should be different**.

- To use these priors, need a normalizing constant ((15 is not necessarily a density). Need a RW metropolis to draw from the priors we have produced.

- Careful about multidimensional ridges: e.g. steady states are 5 equations, and there are 8 parameters - solution not unique, impossible to invert the relationship.

- Careful about choosing θ_3 and θ_4 when there are weak and partial identification problems.

6.2 Choice of data and estimation

- Does it matter which variables are used to estimate the parameters? Yes.

i) Omitting relevant variables may lead to distortions.

ii) Adding variables may improve the fit, but also increase standard errors if added variables are irrelevant.

iii) Different variables may identify different parameters (e.g. with aggregate consumption data and no data on who own financial assets may be very difficult to get estimate the share of rule-of-thumb consumers).

Example 6.3

$$
\begin{aligned}
y_t &= a_1 E_t y_{t+1} + a_2(i_t - E_t\pi_{t+1}) + v_{1t} && (41)\\
\pi_t &= a_3 E_t \pi_{t+1} + a_4 y_t + v_{2t} && (42)\\
i_t &= a_5 E_t \pi_{t+1} + v_{3t} && (43)
\end{aligned}
$$

Solution:

$$
\begin{bmatrix} y_t \\ \pi_t \\ i_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & a_2 \\ a_4 & 1 & a_2 a_4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_{1t} \\ v_{2t} \\ v_{3t} \end{bmatrix}
$$

- a_1, a_3, a_5 *disappear from the solution.*

- *Different variables identify different parameters (i_t identify nothing!!)*

iv) Likelihood function (Posterior) may change shape depending on the variables use. Bimodality or multimodality may be present if important variables are omitted (e.g. if y_t is excluded in above example).

- Using the same model and the same econometric approach Levin et al (2005, NBER macro annual) find habit in consumption is 0.30; Fernandez and Rubio (2008, NBER macro annual) find habit in consumption is 0.88. Why? They use different data sets to estimate the same model!

Can we say something systematic about the choice of variables?

Guerron-Quintana (2010); use Smets and Wouters model and different combinations of observable variables. Finds:

- Internal persistence of the model change if nominal rate, inflation and real wage are absent.

- Duration of price spells affected by the omission of consumption and real wage data.

- Responses of inflation, investment, hours and real wage sensitive to the choice of variables.

- " Best combination" of variables (use in-sample prediction and out-of-sample MSE): use $Y_t, C_t, I_t, R_t, H_t, \pi_t, W$.

Parameter	Wage stickiness	Price Stickiness	Slope Phillips
Data	Median (s.d.)	Median (s.d.)	Median (s.d.)
Basic	0.62 (0.54,0.69)	0.82 (0.80, 0.85)	0.94 (0.64,1.44)
Without C	0.80 (0.73,0.85)	0.97 (0.96, 0.98)	2.70 (1.93,3.78)
Without Y	0.34 (0.28,0.53)	0.85 (0.84, 0.87)	6.22 (5.05,7.44)
Without C,W	0.57 (0.46,0.68)	0.71 (0.63, 0.78)	2.91 (1.73,4.49)
Without R	0.73 (0.67,0.78)	0.81 (0.77, 0.84)	0.74 (0.53,1.03)

(in parenthesis 90% probability intervals)

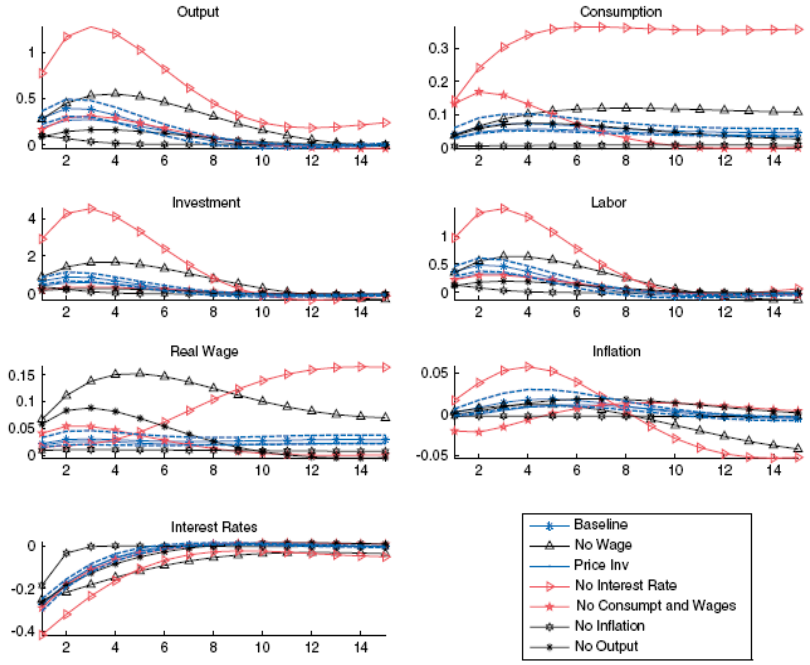


Figure 1. Responses to an expansionary monetary shock. This figure is available in color online at www.interscience.wiley.com/journal/jae

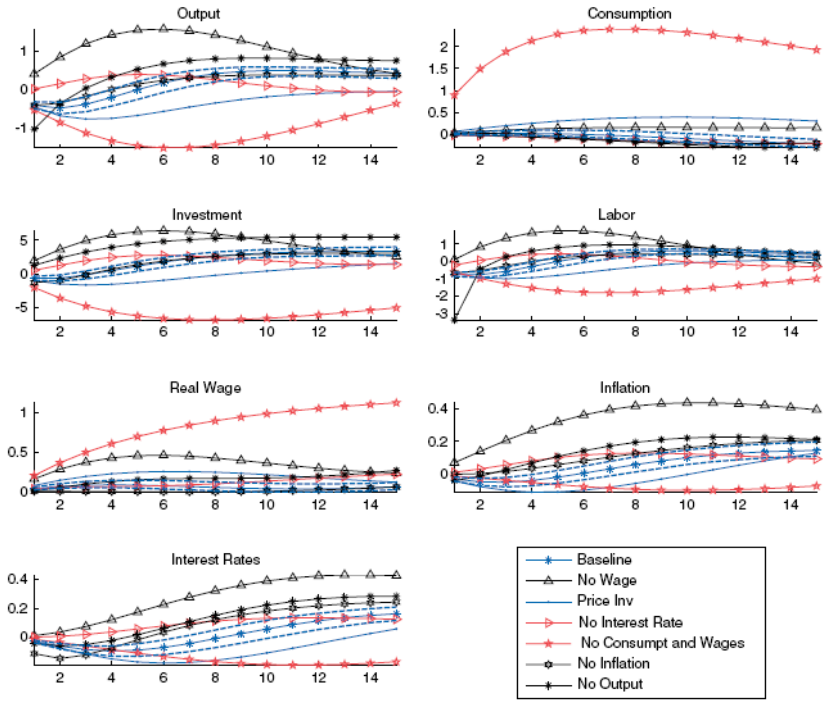


Figure 2. Responses to a positive investment-specific technology shock. This figure is available in color online at www.interscience.wiley.com/journal/jae

Output recession after an investments specific shock and no C and W.

Canova, Ferroni and Matthes (2012)

- Use statistical criteria to choose the variables in estimation

1) Choose the variables that maximize the identificability of relevant parameters.

Compute the rank of the derivative of the spectral density of the solution of the model with respect to the parameters

Komunjer and Ng (2011): have necessary and sufficient conditions for full identification of the parameters

Choose the combination of observables which gives you a rank as close as possible to the ideal.

2) Compare the curvature of the convoluted likelihood in the singular and the non-singular systems in the dimensions of interest.

3) Choose the variables that minimize the information loss going from the larger scale to the smaller scale system.

Loss of information is measured by

$$p_t^j(\theta, e^{t-1}, u_t) = \frac{\mathcal{L}(W_{jt}|\theta, e^{t-1}, u_t)}{\mathcal{L}(Z_t|\theta, e^{t-1}, u_t)} \tag{44}$$

where $\mathcal{L}(.|\theta, y_{1t})$ is the likelihood of Z_t, W_{jt}

$$Z_t = y_t + u_t \tag{45}$$

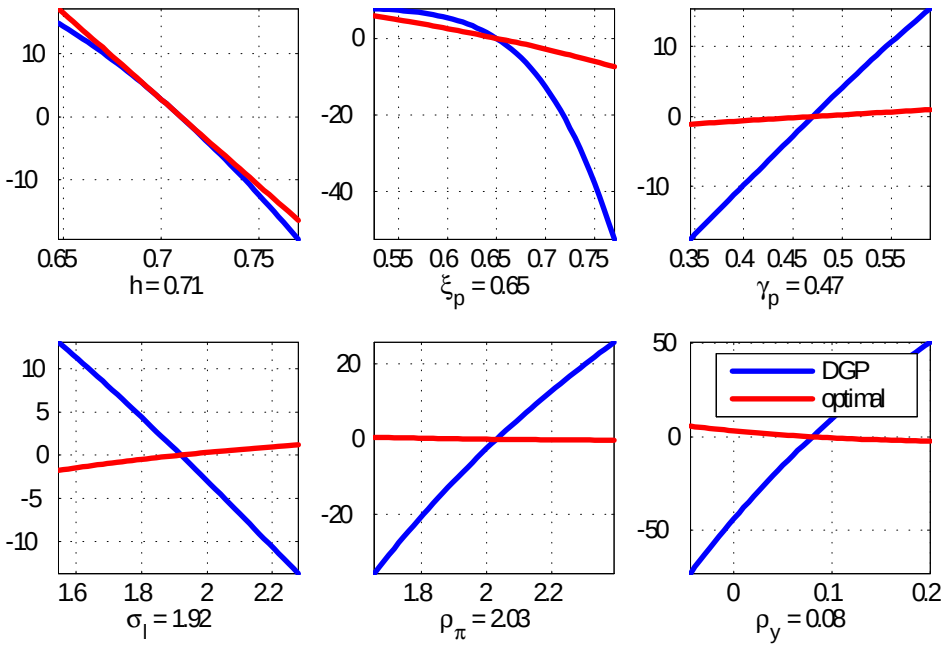
$$W_{jt} = Sy_{jt} + u_t \tag{46}$$

u_t is the convolution error, y_t the original set of variables and y_{jt} the j-th subset of that variables which produce a non-singular system.

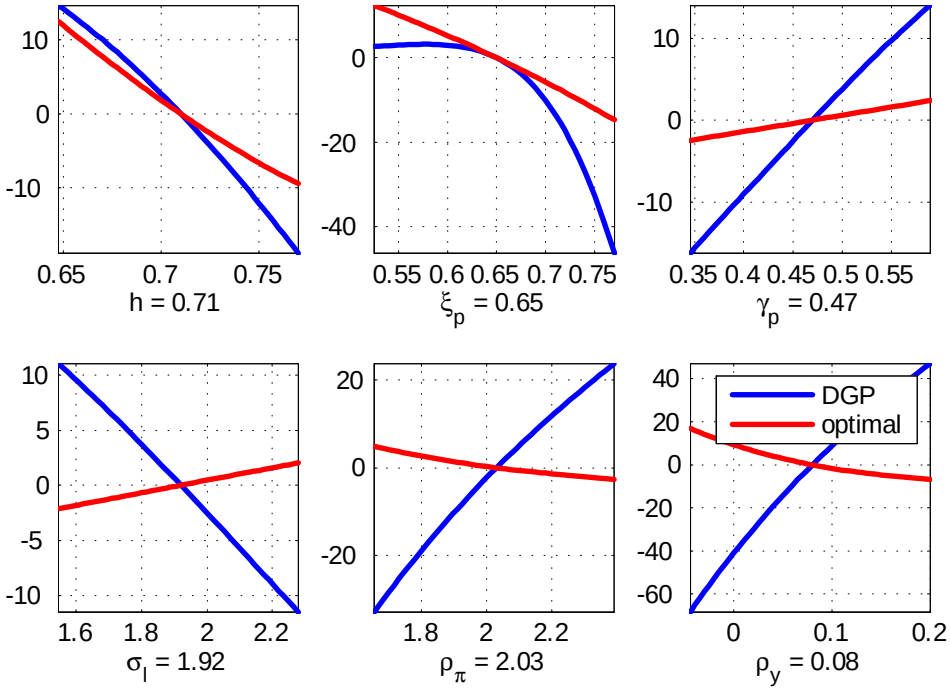
• Apply the procedures to choose the best combination of variables in a SW model driven by only 4 shocks and 7 potential observables.

Vector	Unrest Rank(Δ)	SW Restr Rank(Δ)	SW Restr and Sixth Restr
y, c, i, w	186	188	ψ
y, c, i, π	185	188	ψ
y, c, r, h	185	188	ψ
y, i, w, r	185	188	ψ
c, i, w, h	185	188	ψ, σ_c, ρ_i
c, i, π, h	185	188	ψ
c, i, r, h	185	188	$\zeta_\omega, \zeta_p, i_\omega$
y, c, i, r	185	187	
...			
c, w, π, r	183	187	
c, w, π, h	183	187	
i, w, π, r	183	187	
w, π, r, h	183	187	
c, i, π, r	183	186	
Required	189	189	

Rank conditions for all combinations of variables in the unrestricted SW model (columns 2) and in the restricted SW model (column 3), where five parameters are fixed $\delta = 0.025$, $\varepsilon_p = \varepsilon_w = 10$, $\lambda_w = 1.5$ and $c/g = 0.18$. The fourth columns reports the extra parameter restriction needed to achieve identification; a blank space means that there are no parameters able to guarantee identification.



Likelihood curvature: output and hours



Likelihood curvature: output and labor productivity

Order	Basic		T=1500		$\Sigma_u = 0.01 * I$	
	Vector	Relative Info	Vector	Relative info	Vector	Relative Info
1	(y, c, i, h)	1	(y, c, i, h)	1	(y, c, i, h)	1
2	(y, c, i, w)	0.89	(y, c, i, w)	0.87	(y, c, i, w)	0.86
3	(y, c, i, r)	0.52	(y, c, i, r)	0.51	(y, c, i, r)	0.51
4	(y, c, i, π)	0.5	(y, c, i, π)	0.5	(y, c, i, π)	0.5

Ranking based on the $p(\theta)$ statistic. The first two column present the results for the basic setup, the next six columns the results obtained altering some nuisance parameters. Relative information is the ratio of the $p(\theta)$ statistic relative to the best combination.

How different are good and bad combinations?

- Simulate 200 data points from the model with $[a_t, i_t, g_t, \epsilon_t^m]$ and estimate structural parameters using

(1) Model A: 4 shocks and (y, c, i, w) as observables (best rank analysis).

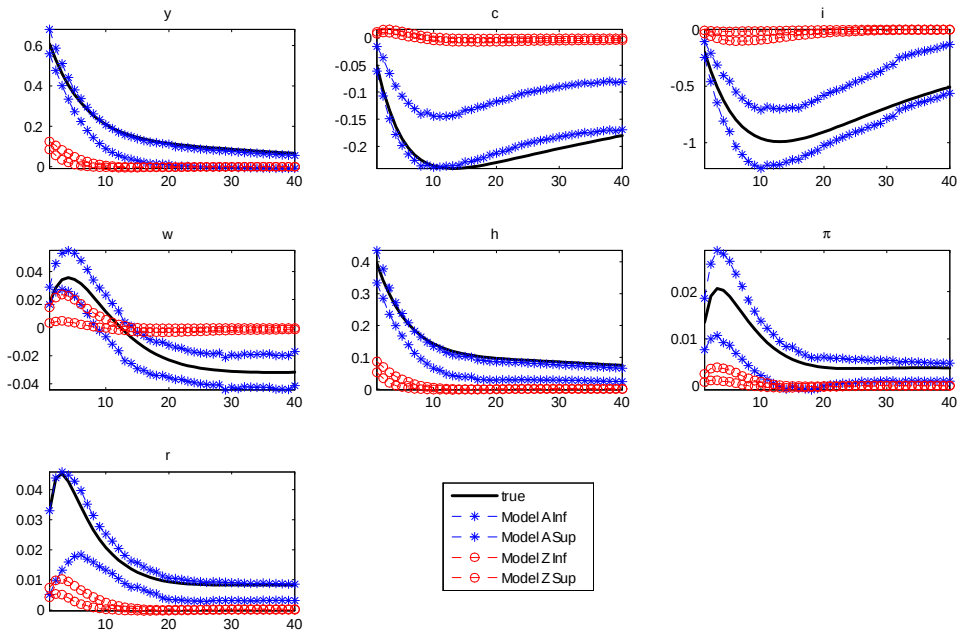
(2) Model B: 4 shocks and (y, c, i, w) as observables (best information analysis).

(3) Model Z: 4 shocks and (c, i, π, r) as observables(worst rank analysis).

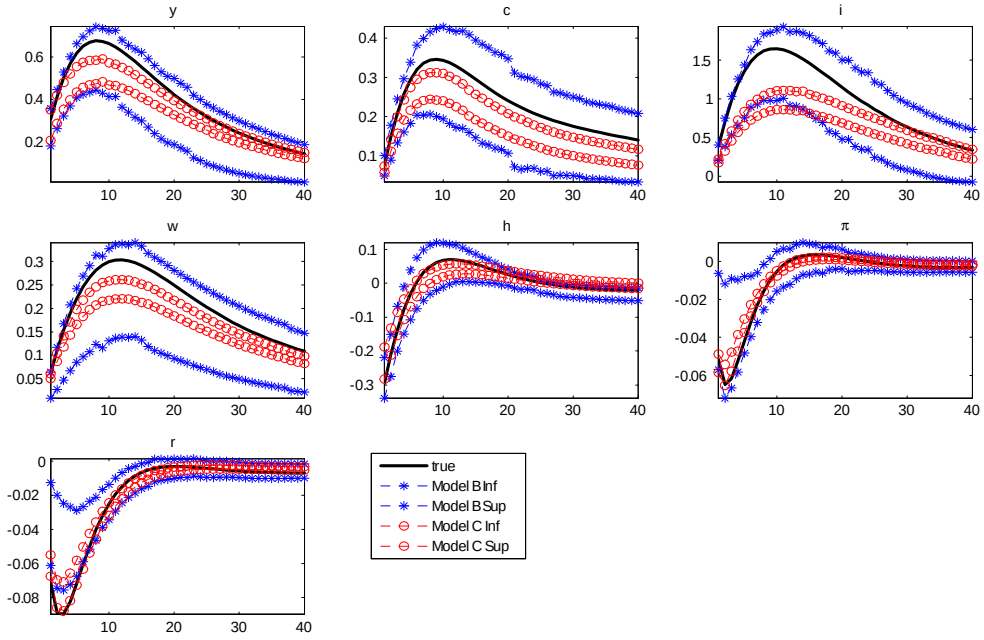
(4) Model C: 4 structural shocks, three measurement errors and $(y_t, c_t, i_t, w_t, \pi, r_t, h_t)$ as observables.

(5) Model D: 7 structural shocks and $(y_t, c_t, i_t, w_t, \pi, r_t, h_t)$ as observables.

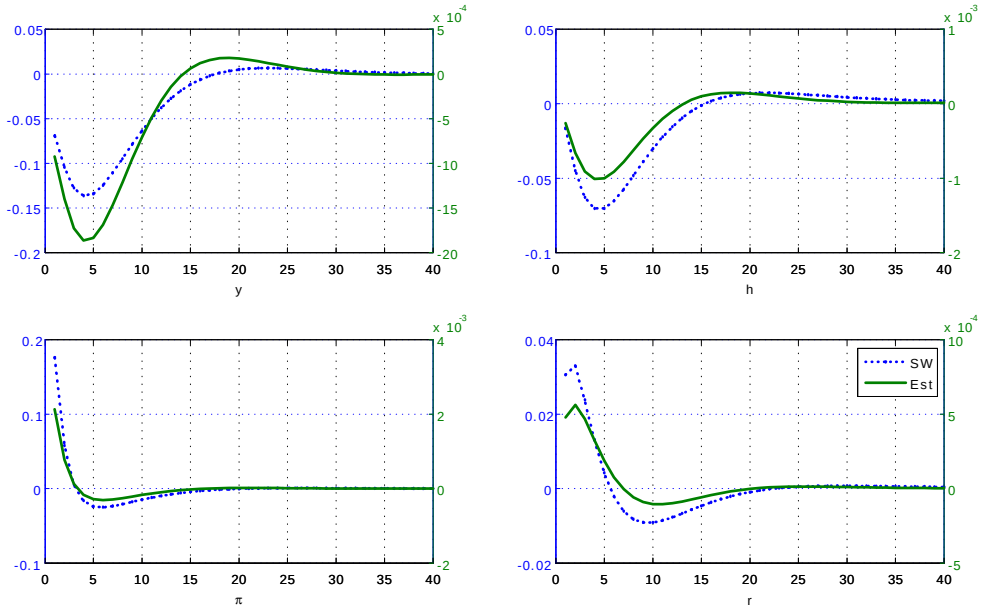
	True	Model A	Model B	Model Z	Model C	Model D
ρ_a	0.95	(0.920 , 0.975)	(0.905 , 0.966)	(0.946 , 0.958)	(0.951 , 0.952)	(0.939 , 0.943)
ρ_g	0.97	(0.930 , 0.969)	(0.930 , 0.972)	(0.601 , 0.856)	(0.970 , 0.971)	(0.970 , 0.972)
ρ_i	0.71	(0.621 , 0.743)	(0.616 , 0.788)	(0.733 , 0.844)	(0.681 , 0.684)	(0.655 , 0.669)
ρ_{ga}	0.51	(0.303 , 0.668)	(0.323 , 0.684)	(0.010 ,0.237)	(0.453 , 0.780)	(0.114 , 0.885)
σ_n	1.92	(1.750 , 2.209)	(1.040 , 2.738)	(0.942 , 2.133)	(1.913 , 1.934)	(1.793 , 1.864)
σ_c	1.39	(1.152 , 1.546)	(1.071 , 1.581)	(1.367 , 1.563)	(1.468 , 1.496)	(1.417 , 1.444)
h	0.71	(0.593 , 0.720)	(0.591 , 0.780)	(0.716 , 0.743)	(0.699 , 0.701)	(0.732 , 0.746)
ζ_ω	0.73	(0.402 , 0.756)	(0.242, 0.721)	(0.211 ,0.656)		(0.806 , 0.839)
ζ_p	0.65	(0.313 , 0.617)	(0.251 , 0.713)	(0.512 , 0.616)	(0.317 , 0.322)	(0.509 , 0.514)
i_ω	0.59	(0.694 , 0.745)	(0.663 , 0.892)	(0.532 ,0.732)	(0.728 , 0.729)	(0.683 , 0.690)
i_p	0.47	(0.571 , 0.680)	(0.564 , 0.847)	(0.613 , 0.768)	(0.625 , 0.628)	(0.606 , 0.611)
ϕ_p	1.61	(1.523 , 1.810)	(1.495 , 1.850)	(1.371 , 1.894)	(1.624 , 1.631)	(1.654 , 1.661)
φ	0.26	(0.145 , 0.301)	(0.153 , 0.343)	(0.255 , 0.373)	(0.279 , 0.295)	(0.281 , 0.306)
ψ	5.48	(3.289 , 7.955)	(3.253 , 7.623)	(2.932 , 7.530)	(11.376 , 13.897)	(4.332 , 5.371)
α	0.2	(0.189 , 0.331)	(0.167 , 0.314)	(0.136 , 0.266)	(0.177 , 0.198)	(0.174 , 0.199)
ρ_π	2.03	(1.309 , 2.547)	(1.277 , 2.642)	(1.718 , 2.573)	(1.868 , 1.980)	(2.119 , 2.188)
ρ_y	0.08	(0.001 , 0.143)	(0.001 , 0.169)	(0.012 , 0.173)	(0.124 , 0.162)	
ρ_R	0.87	(0.776 , 0.928)	(0.813 , 0.963)	(0.868 , 0.916)	(0.881 , 0.886)	
$\rho_{\Delta y}$	0.22	(0.001 , 0.167)	(0.010, 0.192)	(0.130 ,0.215)	(0.235 , 0.244)	
σ_a	0.46	(0.261 , 0.575)	(0.382 , 0.460)	(0.420 ,0.677)	(0.357 , 0.422)	(0.386 , 0.455)
σ_g	0.61	(0.551 , 0.655)	(0.551 , 0.657)	(0.071 ,0.113)	(0.536 , 0.629)	(0.585 , 0.688)
σ_i	0.6	(0.569 , 0.771)	(0.532 , 0.756)	(0.503 ,0.663)	(0.561 , 0.660)	(0.693 , 0.819)
σ_r	0.25	(0.100 , 0.259)	(0.078 , 0.286)	(0.225 ,0.267)	(0.226 , 0.265)	(0.222 , 0.261)



Impulse responses to a goverment spending shock



Impulse responses to a technology shock



Impulse responses to a price markup shock

6.3 Combining DSGE and VARs

Del Negro and Schorfheide(2004):

- $f(y|\alpha, \Sigma_u) =$ likelihood of the data conditional on the VAR parameters.

- $g(\alpha, \Sigma_u|\theta)$ prior for the VAR parameters, conditional on the DSGE model parameters (the hyperparameters)

- $g(\theta)$ the prior distribution for DSGE parameters $\rightarrow g(\alpha, \Sigma_u|\theta)$ is the prior on the reduced form parameters induced by the prior on the structural parameters and the structure of the DGSE model.

Joint posterior of VAR and structural parameters is

$g(\alpha, \Sigma_u, \theta|y) = g(\alpha, \Sigma_u, |\theta, y)g(\theta|y)$

$g(\alpha, \Sigma_u, |\theta, y)$ is of normal-inverted Wishart form: easy to compute.

Posterior kernel $\breve{g}(\theta|y) = f(y|\theta)g(\theta)$ where $f(y|\theta)$ is given by

$$\begin{aligned} f(y|\theta) &= \int f(y|\alpha, \Sigma_u)g(\alpha, \Sigma_u, \theta)d\alpha d\theta \\ &= \frac{f(y|\alpha, \Sigma_u)g(\alpha, \Sigma_u|\theta)}{g(\alpha, \Sigma_u|y)} \end{aligned}$$

Given that $g(\alpha, \Sigma_u, |\theta, y) = g(\alpha, \Sigma_u, |y)$. Then

$$
\begin{aligned}
f(y|\theta) &= \frac{|T_1 x^{s'}(\theta)x^s(\theta) + X'X|^{-0.5M}|(T_1+T)\tilde{\Sigma}_u(\theta)|^{-0.5(T_1+T-k)}}{|\tau x^{s'}(\theta)x^s(\theta)|^{-0.5M}|T_1\tilde{\Sigma}_u^s(\theta)|^{-0.5(T_1-k)}} \\
&\times \frac{(2\pi)^{-0.5MT}2^{-0.5M(T_1+T-k)}\prod_{i=1}^{M}\Gamma(0.5*(T_1+T-k+1-i))}{2^{-0.5M(T_1-k)}\prod_{i=1}^{M}\Gamma(0.5*(T_1-k+1-i))} \qquad (47)
\end{aligned}
$$

$T_1 =$ number of simulated observations, Γ is the Gamma function, X includes all lags of y and the superscript s indicates simulated data.

- Draw θ using an MH algorithm.

- Conditional on θ construct posterior of α (draw α from a Normal-Wishart).

Estimation algorithm:

1) Draw a candidate θ. Use MCMC to decide if accept or reject.

2) With the draw compute the model induced prior for the VAR parameters.

3) Compute the posterior for the VAR parameters (analytically if you have a conjugate structure or via the Gibbs sampler if you do not have one. Draw from this posterior

4) Repeat steps 1)-3) $NL + \bar{L}$ times. Check convergence

5) Repeat 1)-4) for different T_1. Choose the T_1 that maximizes the marginal likelihood.

6.4 Practical issues

Log-linear DSGE solution:

$$y_{1t} = \mathcal{A}_{11}(\theta) y_{1t-1} + \mathcal{A}_{13}(\theta) y_{3t} \quad (48)$$

$$y_{2t} = \mathcal{A}_{12}(\theta) y_{1t-1} + \mathcal{A}_{23}(\theta) y_{3t} \quad (49)$$

where y_{2t} are the control, y_{1t} the states (predetermined and exogenous), y_{3t} the shocks, θ are the structural parameters and $\mathcal{A}_{ij}$ the coefficients of the decision rules.

How to you put DSGE models on the data when:

a) the model implies that the covariance of $y_t = [y_{1t}, y_{2t}]$ is singular.

b) the variables are mismeasured relative to the model quantities.

c) have additional information one would like to use.

For a):

• Choose a selection matrix F_1 such that dim(x_{1t}) = dim $(F_1\ y_t)$ = dim (y_{3t}), i.e. throw away model information. Good strategy to follow if some component of y_t are non-observable.

• Explicitly solve out fraction of variables from the model. Format of the solution is no longer a restricted VAR(1).

• Adds measurement errors to the y_{2t} so that dim(x_{2t}) = dim $(F_2\ y_t)$ = dim (y_{3t})+dim (e_t), where e_t are measurement errors.

- If the model has two shocks and implications for four variables, we could add at least two and up to four measurement errors to the model.

Here (1)-(2) are the state equations and the measurement equation is

$$x_{2t} = F_2 y_t + e_t \tag{50}$$

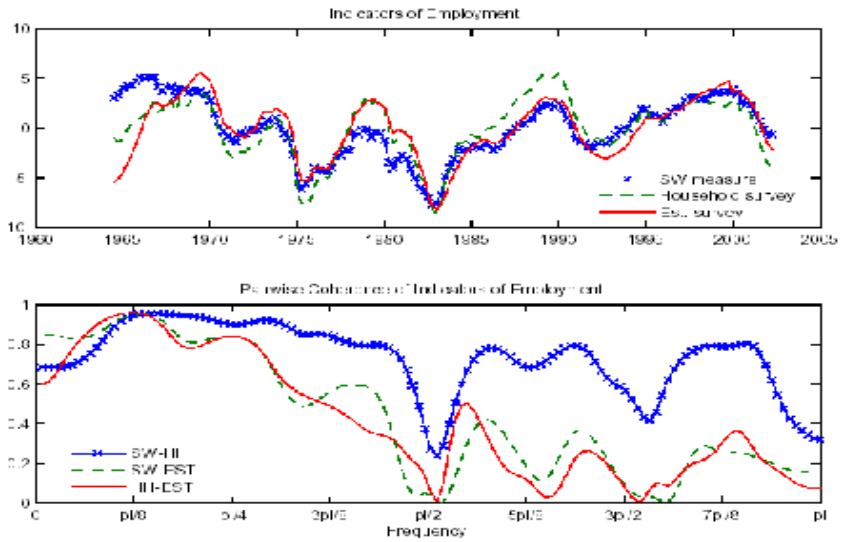
- Need to restrict time series properties of e_t. Otherwise difficult to distinguish dynamics induced by structural shocks and the measurement errors.

i) the measurement error is iid (since θ is identified from the dynamics induced by the reduced form shocks, if measurement error is iid, θ identified by the dynamics due to structural shocks).

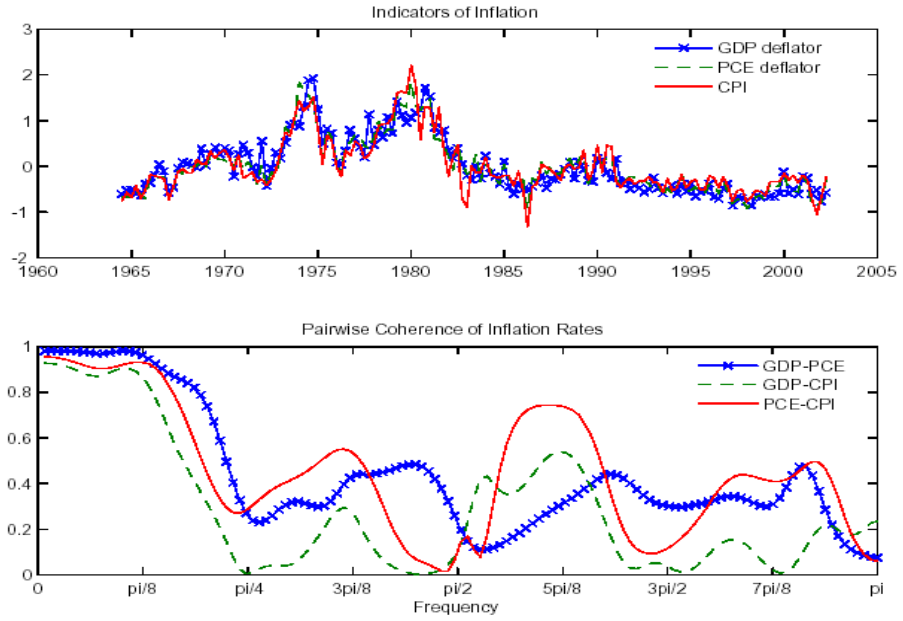
Ireland (2004): VAR(1) process for the measurement error; identification problems! Can be used to verify the quality of the model's approximation to the data - measurement error captures what is missing from the model to fit the data (see also Watson (1993)). Useful device when θ is calibrated. Less useful when θ is estimated.

For b): Recognize that existing measures of theoretical concepts are contaminated.

- How do you measure hours? Use establishment survey series? Household survey series? Employment?



- Do we use CPI inflation, GDP deflator or PCE inflation?



- Different measures contain (noisy) information about the true series. Not perfectly correlated among each other.

- Use ideas underlying factor models. Let x_{3t} be a $k \times 1$ vector of observable variables and x_{1t} be of dimension $N \times 1$ where dim (N) ¡ dim(k). Then:

$$x_{3t} = \Lambda_3 x_{1t} + u_t \tag{51}$$

where the first row of Λ_3 is normalized to 1. Thus:

$$\begin{aligned} x_{3t} &= \Lambda_3[F_1 y_{1t}, F_1 \mathcal{A}_{12}(\theta) y_{1t-1} + F_1 \mathcal{A}_{13}(\theta) y_{3t}]' + u_{3t} && (52) \\ &= \Lambda_3[F_1 y_{1t}, F_1 \mathcal{B}(\theta) y_{1t}]' + u_{3t} && (53) \end{aligned}$$

so that x_{3t} can be used to recover the vector of states y_{1t} and to estimate θ

- What is the advantage of this procedure? If only one component of x_{3t} is used to measure y_{1t}, estimate of θ will probably be noisy.

- By using a vector of information and assuming that the elements of u_t are idiosyncratic:

i) reduce the noise in the estimate of y_{1t} (the estimated variance of y_{1t} will be asymptotically of the order $1/k$ time the variance obtained when only one indicator is used (see Stock and Watson (2002)).

ii) estimates of θ more precise.

- How different is from factor models?. The DSGE model structure is imposed in the specification of the law of motion of the states (states have economic content). In factor models the states are assumed to follow is an assumed unrestricted time series specification, say an AR(1) or a random walk and are uninterpretable.

- How do we separately identify the dynamics induced by the structural shocks and the measurement errors? Since the measurement error is identified from the cross sectional properties of the variables in x_{3t}, possible to have structural disturbances and measurement errors to both be serially correlated of an unknown form.

For c): Sometimes we may have proxy measures for the unobservable states. (commodity prices are often used as proxies for future inflation shocks, stock market shocks are used as proxies for future technology shocks (Beaudry and Portier (2006)).

- Can use these measures to get information about the states. Let x_{4t} a $q \times 1$ vector of variables. Assume

$$x_{4t} = \Lambda_4 y_{1t} + u_{4t} \tag{54}$$

where Λ_4 is unrestricted. Combining all sources of information we have

$$X_t = \Lambda y_{1t} + u_t \tag{55}$$

where $X_t = [x_{3t}, x_{4t}]'$, $u_t = [u_{3t}, u_{4t}]$ and $\Lambda = [\Lambda_3 F, \Lambda_3 F \mathcal{B}(\theta), \Lambda_4]'$.

- The fact that we are using the DSGE structure ($\mathcal{B}$ depends on θ) imposes restrictions on the way the data behaves. (interpret data information through the lenses of the DSGE model).

- Can still jointly estimate the structural parameters and the unobservable states of the economy.

6.5 An example

Use a simple three equation New-keynesian model:

$$x_t = E_t(x_{t+1}) - \frac{1}{\phi}(i_t - E_t\pi_{t+1}) + e_{1t} \quad (56)$$

$$\pi_t = \beta E_t\pi_{t+1} + \kappa x_t + e_{2t} \quad (57)$$

$$i_t = \psi_r i_{t-1} + (1-\psi_r)(\psi_\pi \pi_t + \psi_x x_t) + e_{3t} \quad (58)$$

where β is the discount factor, ϕ the relative risk aversion coefficient, κ the slope of Phillips curve, $(\psi_r, \psi_\pi, \psi_x)$ policy parameters. Here x_t is the output gap, π_t the inflation rate and i_t the nominal interest rate. Assume

$$e_{1t} = \rho_1 e_{1t-1} + v_{1t} \quad (59)$$

$$e_{2t} = \rho_2 e_{2t-1} + v_{2t} \quad (60)$$

$$e_{3t} = v_{3t} \quad (61)$$

where $\rho_1, \rho_2 < 1$, $v_{jt} \sim (0, \sigma_j^2), j = 1, 2, 3$.

6.5.1 Contaminated data

- Ambiguities in linking the output gap, the inflation rate and the nominal interest rate to empirical counterparts. e.g. for the nominal interest rate: should we use a short term measure or a long term one? for the output gap, should we use a statistical based measure or a theory based measure? In the last case, what is the flexible price equilibrium?

The solution of the model can be written as

$$w_t = RR(\theta)w_{t-1} + SS(\theta)v_t \tag{62}$$

where w_t is a 8×1 vector including x_t, π_t, i_t, the three shocks and the expectations of x_t and π_t and $\theta = (\phi, \kappa, \psi_r, \psi_y, \psi_\pi, \rho_1, \rho_2, \sigma_1, \sigma_2, \sigma_3)$.

Let $x_t^j, j = 1, \ldots N_x$ be observable indicators for x_t, let $\pi_t^j, j = 1, \ldots N_\pi$ observable indicators for π_t, and $i_t^j, j = 1, \ldots N_i$ observable indicators for i_t. Let $W_t = [x_t^1, \ldots, x_t^{N_x}, \pi_t^1, \ldots, \pi_t^{N_\pi}, i_t^1, \ldots i_t^{N_i}]'$ be a $N_x + N_\pi + N_i \times 1$ vector.

Assume that (62) is the state equation of the system and that the measurement equation is

$$W_t = \Lambda w_t + e_t \tag{63}$$

where λ is $N_x + N_\pi + N_i \times 3$ matrix with at most one element different from zero in each row.

- Once we normalize the nonzero element of the first row of Λ to be one, we can estimate (62)-(63) with standard methods. The routines give us estimates of λ, RR, SS and of w_t which are consistent with the data.

6.5.2 Conjunctoral information

- Suppose we have available measures of future inflation (from surveys, from forecasting models) or data which may have some information about future inflation, for example, oil prices, housing prices, etc.

- Want to predict inflation h periods ahead, $h = 1, 2, \ldots$.

Let $\pi_t^j, j = 1, \ldots N_\pi$ be the observable indicators for π_t and let $W_t = [x_t, i_t, \pi_t^1, \ldots, \pi_t^{N_\pi}]'$ be a $2 + N_\pi \times 1$ vector.

The measurement equation is:

$$W_t = \Lambda w_t + e_t \tag{64}$$

where Λ is $2 + N_\pi \times 3$ matrix, $\Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_1 \\ \dots & \dots & \dots \\ 0 & 0 & \lambda_{N_\pi} \end{bmatrix}$.

- Estimates of the unobservable w_t can be obtained with the Kalman filter. Using estimates of $RR(\theta)$ and $SS(\theta)$ from the state equation we can unconditionally predict w_t h-steps ahead or predict its path conditional on a path for $v_{l,t+h}$.

- Forecast will incorporate information from the model, information from conjunctural data and from standard data and information about the path of the shocks. This information will be optimally mixed depending on their relative precision.

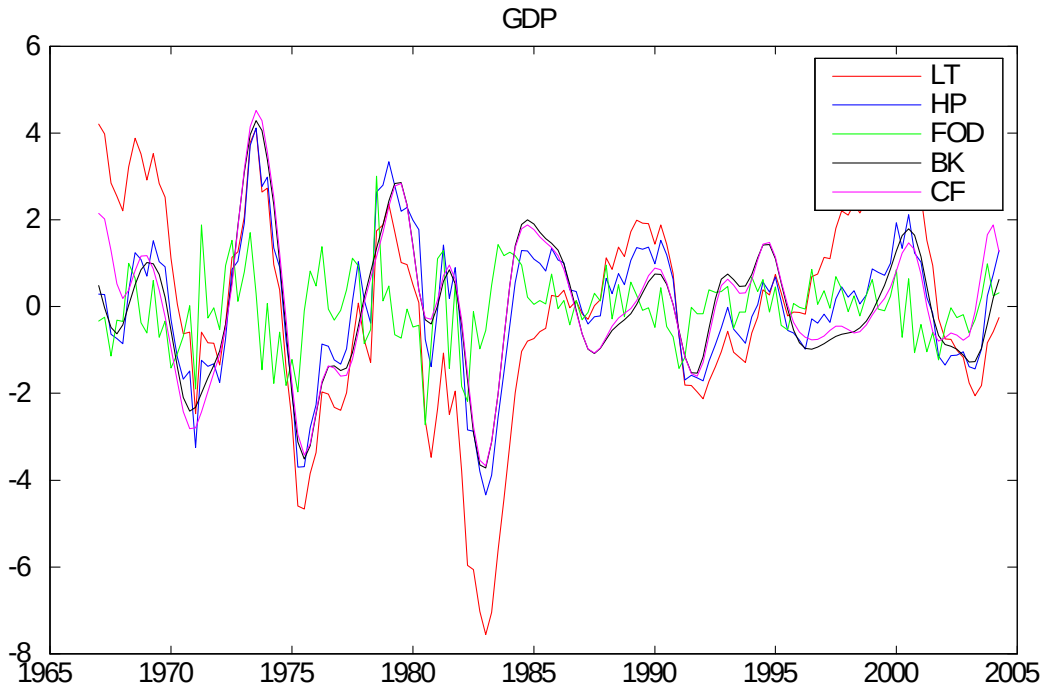
6.6 Dealing with trends

- Most of models available for policy are stationary and cyclical.

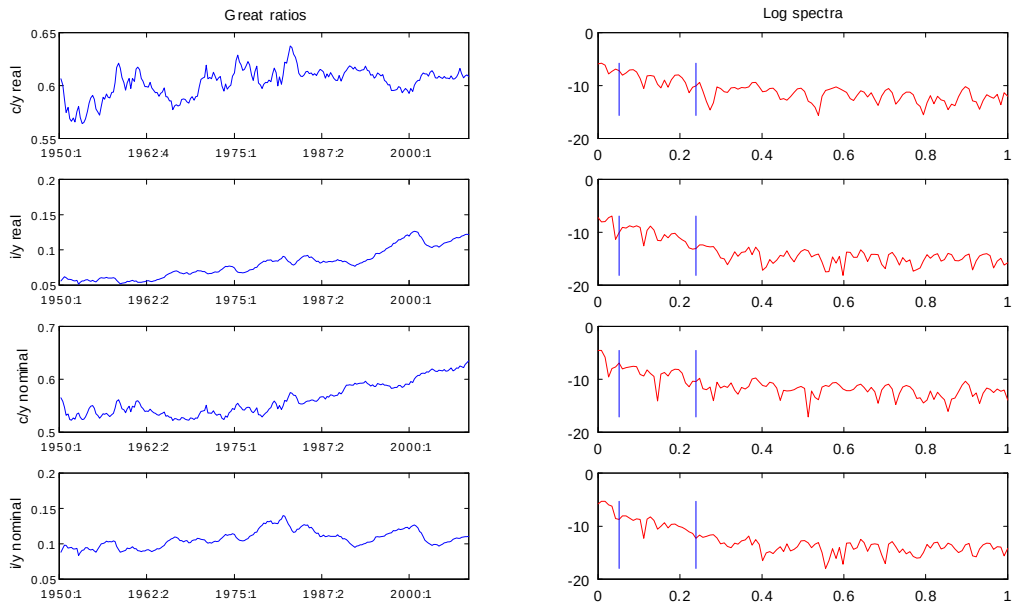
- Data is close to non-stationary, has trends and displays breaks.

- How to we match models to the data?

a) Detrend actual data. Model is a representation for detrended data standard approach. Problem: which detrended data is the model representing?


GDP
LT
HP
FOD
BK
CF
6
4
2
0
-2
-4
-6
-8
1965
1970
1975
1980
1985
1990
1995
2000
2005

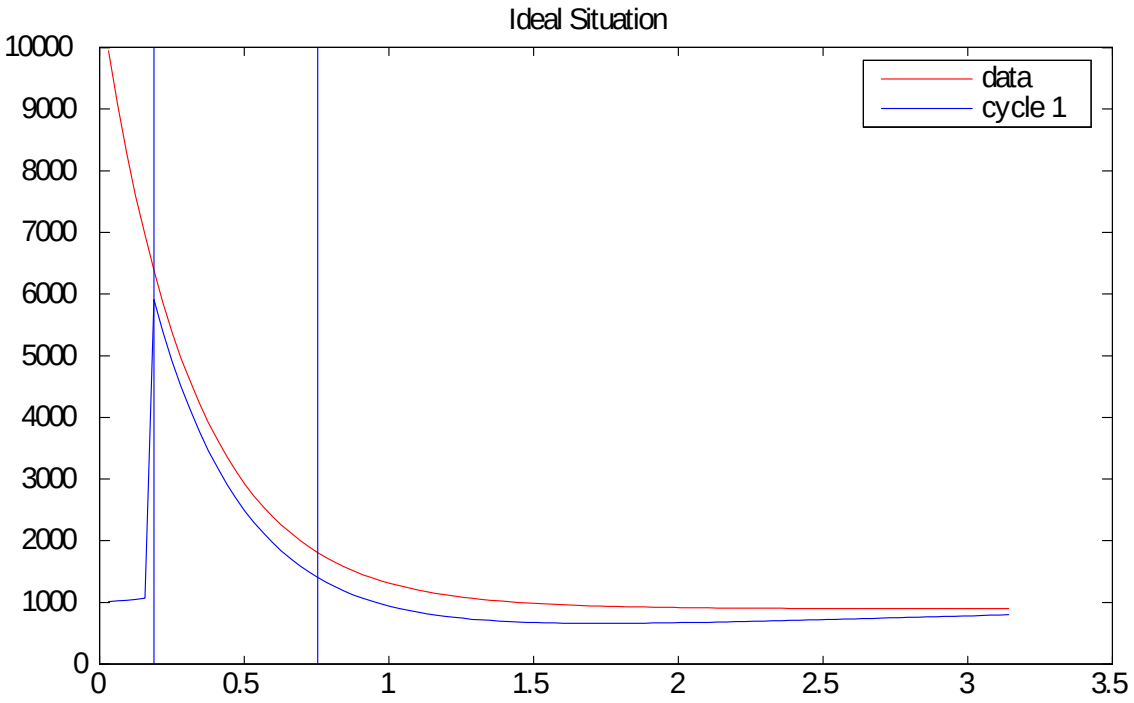
b) Build-in a trend into the model. Detrend the data with model-based-trend. Problem: data does not seem to satisfy balanced growth.



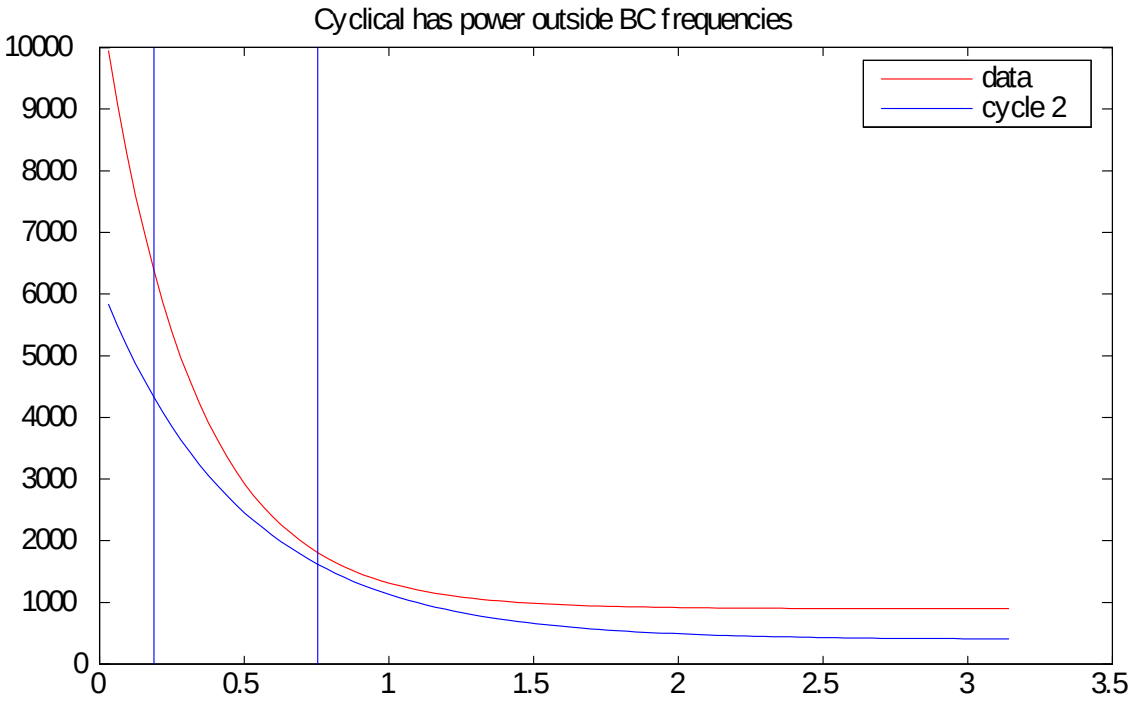
Real and nominal Great ratios in US, 1950-2008.

c) Use transformation of the data which allow you to estimate jointly cycle and the parameters trend (see e.g. growth rates in Smets and Wouter 2007). Problem: hard to fit models to quarterly growth rates

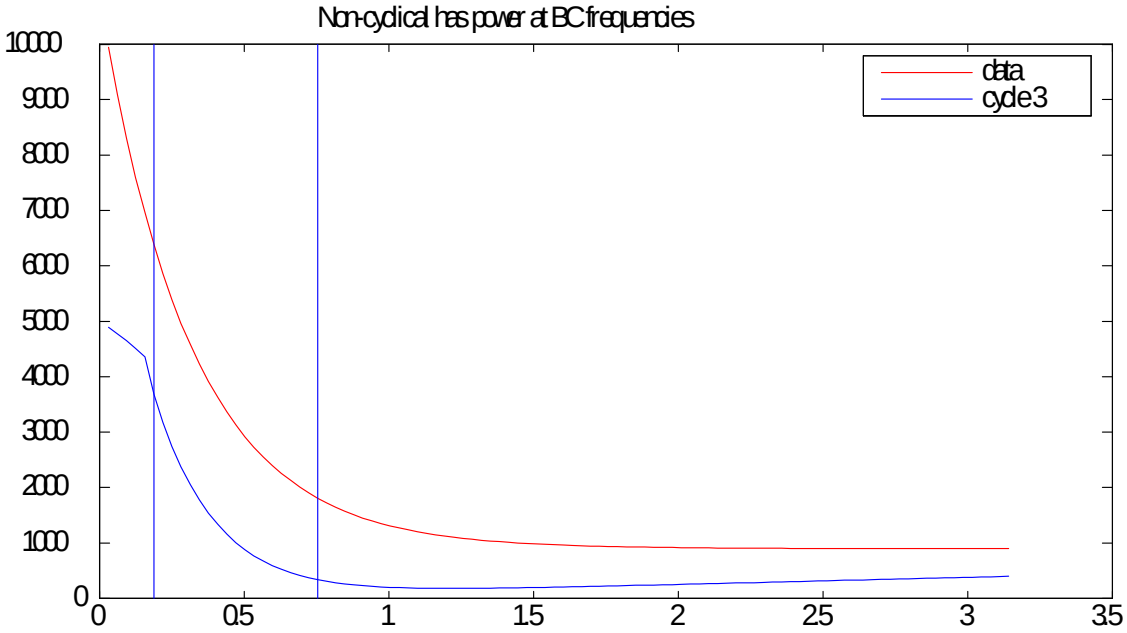
- General problem: statistical definition of cycles different than economic definition. All statistical approaches are biased even in large samples.


Ideal Situation
data
cycle 1
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
0.5
1
1.5
2
2.5
3
3.5

Ideal case



Realistic case



General case

- In developing countries most of cyclical fluctuations driven by trends (Aguiar and Gopinath (2007)).

Two potential approaches:

1) Data-rich environment (Canova and Ferroni (2011)). Let y_t^i be the actual data filtered with method $i = 1, 2, ..., I$ and $y_t^d = [y_t^1, y_t^2, \ldots]$. Assume:

$$y_t^d = \lambda_0 + \lambda_1 y_t(\theta) + u_t \tag{65}$$

where $\lambda_j, j = 0, 1$ are matrices of parameters, measuring bias and correlation between data and model based quantities, u_t are measurement errors and θ the structural parameters.

- Factor model setup a-la Boivin and Giannoni (2005).

- Can jointly estimate θ and λ's. Can obtain a more precise estimate of the unobserved $y_t(\theta)$ if measurement error is uncorrelated across methods.

- Same interpretation as GMM with many instruments.

2) Bridge cyclical model and the data with a flexible specification for the trend (Canova, 2010)).

$$y_t^d = c + y_t^T + y_t^m(\theta) + u_t \tag{66}$$

where $y_t^d \equiv \tilde{y}_t^d - E(\tilde{y}_t^d)$ the log demeaned vector of observables, $c = \bar{y} - E(\tilde{y}_t^d)$, y_t^T is the non-cyclical component, $y_t^m(\theta) \equiv S[y_t, x_t]'$, S is a selection matrix, is the model based- cyclical component, u_t is a iid $(0, \Sigma_u)$ (measurement) noise, $y_t^T, y_t^m(\theta)$ and u_t are mutually orthogonal.

- Model (linearized) solution: cyclical component

$$\begin{aligned} y_t &= RR(\theta)x_{t-1} + SS(\theta)z_t & (67)\\ x_t &= PP(\theta)x_{t-1} + QQ(\theta)z_t & (68)\\ z_{t+1} &= NN(\theta)z_t + \epsilon_{t+1} & (69) \end{aligned}$$

$PP(\theta), QQ(\theta), RR(\theta), SS(\theta)$ functions of the structural parameters $\theta = (\theta_1, \ldots, \theta_k)$, $x_t = \tilde{x}_t - \bar{x}$; $y_t = \tilde{y}_t - \bar{y}$; and z_t are the disturbances, $\bar{y}, \bar{x}$ are the steady states of $\tilde{y}_t$ and $\tilde{x}_t$.

- Non cyclical component

$$y_t^T = y_{t-1}^T + \bar{y}_{t-1} + e_t \qquad e_t \sim iid\,(0, \Sigma_e^2) \tag{70}$$

$$\bar{y}_t = \bar{y}_{t-1} + v_t \qquad v_t \sim iid\,(0, \Sigma_v^2) \tag{71}$$

$\Sigma_v^2 > 0$ and $\Sigma_e^2 = 0$, y_t^T is a vector of I(2) processes.

$\Sigma_v^2 = 0$, and $\Sigma_e^2 > 0$, y_t^T is a vector of I(1) processes.

$\Sigma_v^2 = \Sigma_e^2 = 0$, y_t^T is deterministic.

$\Sigma_v^2 > 0$ and $\Sigma_e^2 > 0$ and $\sigma_v^2 \sigma_e^2$ is large, y_t^t is "smooth" and nonlinear (as in HP).

- Jointly estimate structural θ and non-structural parameters.

Example 6.4 *The log linearized equilibrium conditions of basic NK model are:*

$$\begin{aligned}
\lambda_t &= \chi_t - \frac{\sigma_c}{1-h}(y_t - h y_{t-1}) && (72)\\
y_t &= z_t + (1-\alpha) n_t && (73)\\
w_t &= -\lambda_t + \sigma_n n_t && (74)\\
r_t &= \rho_r r_{t-1} + (1-\rho_r)(\rho_\pi \pi_t + \rho_y y_t) + v_t && (75)\\
\lambda_t &= E_t(\lambda_{t+1} + r_t - \pi_{t+1}) && (76)\\
\pi_t &= k_p(w_t + n_t - y_t + \mu_t) + \beta E_t \pi_{t+1} && (77)\\
z_t &= \rho_z z_{t-1} + \iota_t^z && (78)
\end{aligned}$$

where $k_p = \frac{(1-\beta\zeta_p)(1-\zeta_p)}{\zeta_p}\frac{1-\alpha}{1-\alpha+\varepsilon\alpha}$, λ *is the Lagrangian on the consumer budget constraint,* z_t *is a technology shock,* χ_t *a preference shock,* v_t *is an iid monetary policy shock and* ϵ_t *an iid markup shock.*

Filter		LT	HP	FOD	BP	Flexible
Parameter	True	Median (s.d.)	Median (s.d.)	Median (s.d.)	Median(s.d.)	Median(s.d.)
σ_c	3.00	2.08 (0.11)	2.08 (0.14)	1.89 (0.14)	2.13 (0.12)	3.68(0.40)
σ_n	0.70	1.72 (0.09)	1.36 (0.07)	1.24 (0.06)	1.58 (0.08)	0.54(0.14)
h	0.70	0.67 (0.02)	0.58 (0.03)	0.36 (0.03)	0.66 (0.02)	0.55(0.04)
α	0.60	0.28 (0.03)	0.15 (0.02)	0.14 (0.02)	0.17 (0.02)	0.19(0.03)
ϵ	7.00	3.19 (0.11)	5.13 (0.19)	3.76 (0.18)	3.80 (0.13)	6.19(0.07)
ρ_r	0.20	0.54 (0.03)	0.77 (0.03)	0.72 (0.04)	0.53 (0.03)	0.16(0.04)
ρ_π	1.20	1.69 (0.08)	1.65 (0.06)	1.65 (0.07)	1.63 (0.10)	0.30(0.04)
ρ_y	0.05	-0.14 (0.04)	0.45 (0.04)	0.63 (0.06)	0.40 (0.04)	0.07(0.03)
ζ_p	0.80	0.85 (0.03)	0.91 (0.03)	0.93 (0.03)	0.90 (0.03)	0.78(0.04)
ρ_χ	0.50	1.00 (0.03)	0.96 (0.03)	0.96 (0.03)	0.95 (0.03)	0.53(0.02)
ρ_z	0.80	0.84 (0.03)	0.96 (0.03)	0.97 (0.03)	0.96 (0.03)	0.71(0.03)
σ_χ	1.12	0.11 (0.02)	0.17 (0.02)	0.21 (0.03)	0.14 (0.02)	1.29(0.01)
σ_z	0.51	0.07 (0.01)	0.09 (0.01)	0.09 (0.01)	0.07 (0.01)	0.72(0.02)
σ_{mp}	0.10	0.05 (0.01)	0.05 (0.01)	0.05 (0.01)	0.05 (0.01)	0.22(0.004)
σ_μ	20.60	6.30 (0.50)	16.75 (0.62)	22.75 (0.83)	14.40 (0.58)	15.88(0.06)
σ_χ^{nc}	3.21					

σ_χ^{nc} is the standard deviation of the non-cyclical component. Parameters Estimates using different filters, small variance of non-cyclical shock

Filter		LT	HP	FOD	BP	Flexible
Parameter	True	Median (s.d.)	Median (s.d.)	Median (s.d.)	Median(s.d.)	Median(s.d.)
σ_c	3.00	1.89 (0.07)	1.89 (0.07)	1.87 (0.07)	2.03 (0.09)	3.26 (0.29)
σ_n	0.70	2.13 (0.08)	2.11 (0.08)	2.15 (0.08)	1.90 (0.08)	0.80 (0.13)
h	0.70	0.58 (0.02)	0.60 (0.02)	0.56 (0.02)	0.69 (0.02)	0.77 (0.04)
α	0.60	0.47 (0.02)	0.46 (0.02)	0.49 (0.02)	0.24 (0.03)	0.41 (0.04)
ϵ	7.00	3.85 (0.13)	3.92 (0.13)	3.46 (0.11)	4.16 (0.13)	6.95 (0.09)
ρ_r	0.20	0.68 (0.03)	0.59 (0.03)	0.43 (0.04)	0.50 (0.03)	0.31 (0.04)
ρ_π	1.20	1.14 (0.04)	1.25 (0.04)	1.25 (0.04)	1.23 (0.04)	1.25 (0.03)
ρ_y	0.05	-0.07 (0.00)	-0.01 (0.01)	-0.05 (0.02)	0.23 (0.01)	0.08 (0.10)
ζ_p	0.80	0.81 (0.03)	0.78 (0.03)	0.76 (0.03)	0.89 (0.03)	0.72 (0.02)
ρ_χ	0.50	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	0.97 (0.03)	0.69 (0.05)
ρ_z	0.80	0.90 (0.03)	0.92 (0.03)	0.91 (0.03)	0.98 (0.03)	0.90 (0.03)
σ_χ	1.12	0.09 (0.01)	0.31 (0.05)	0.61 (0.15)	1.87 (0.14)	1.28 (0.03)
σ_z	0.51	0.61 (0.07)	0.30 (0.04)	0.40 (0.05)	0.10 (0.01)	0.69 (0.01)
σ_{mp}	0.10	0.06 (0.01)	0.06 (0.01)	0.06 (0.01)	0.06 (0.01)	0.24 (0.004)
σ_μ	20.60	18.00 (0.74)	18.04 (0.61)	15.89 (0.83)	17.55 (0.57)	12.73 (0.04)
σ_χ^{nc}	23.21					

Parameters Estimates using different filters; σ_χ^{nc} is the standard deviation of the non-cyclical component.

Why are estimates distorted?

- Posterior proportional to likelihood times prior.

- Log-likelihood of the parameters (see Hansen and Sargent (1993))

$$L(\theta|y_t) = A_1(\theta) + A_2(\theta) + A_3(\theta)$$

$$A_1(\theta) = \frac{1}{\pi}\sum_{\omega_j} \log\det G_\theta(\omega_j)$$

$$A_2(\theta) = \frac{1}{\pi}\sum_{\omega_j} \text{ trace } [G_\theta(\omega_j)]^{-1} F(\omega_j)$$

$$A_3(\theta) = (E(y) - \mu(\theta)) G_\theta(\omega_0)^{-1} (E(y) - \mu(\theta))$$

where $\omega_j = \frac{\pi j}{T}, j = 0, 1, \ldots, T-1$,, $G_\theta(\omega_j)$ is the model based spectral density matrix of y_t, $\mu(\theta)$ the model based mean of y_t, $F(\omega_j)$ is the data based spectral density of y_t and $E(y)$ the unconditional mean of the data.

- first term: sum of the one-step ahead forecast error matrix across frequencies;

- the second a penalty function, emphasizing deviations of the model-based from the data-based spectral density at various frequencies.

- the third another penalty function, weighting deviations of model-based from data-based means, with the spectral density matrix of the model at frequency zero.

- Suppose that the actual data is filtered so that frequency zero is eliminated and low frequencies deemphasized. Then

$$L(\theta|y_t) = A_1(\theta) + A_2(\theta)^*$$

$$A_2(\theta)^* = \frac{1}{\pi} \sum_{\omega_j} \text{ trace } [G_\theta(\omega_j)]^{-1} F(\omega_j)^*$$

where $F(\omega_j)^* = F(\omega_j) I_\omega$ and I_ω is an indicator function.

Suppose that $I_\omega = I_{[\omega_1, \omega_2]}$, an indicator function for the business cycle frequencies, as in an ideal BP filter.

The penalty $A_2(\theta)^*$ matters only at these frequencies.

Since $A_2(\theta)^*$ and $A_1(\theta)$ enter additively in the log-likelihood function, there are two types of biases in $\hat{\theta}$.

- estimates $F_\theta(\omega_j)^*$ only approximately capture the features of $F(\omega_j)^*$ at the required frequencies - the sample version of $A_2(\theta)^*$ has a smaller values at business cycle frequencies and a nonzero value at non-business cycle ones.

- To reduce the contribution of the penalty function to the log-likelihood, parameters are adjusted to make $[G_\theta(\omega_j)]$ close to $F(\omega_j)^*$ at those frequencies where $F(\omega_j)^*$ is not zero. This is done by allowing fitting errors in $A_1(\theta)$ large at frequencies $F(\omega_j)^*$ is zero - in particular the low frequencies.

Conclusions:

1) The volatility of the structural shocks will be overestimated - this makes $[G_\theta(\omega_j)]$ close to $F(\omega_j)^*$ at the relevant frequencies.

2) Their persistence underestimated - this makes $G_\theta(\omega_j)$ small and the fitting error large at low frequencies.

Estimated economy very different from the true one: agents' decision rules are altered.

- Higher perceived volatility implies distortions in the aversion to risk and a reduction in the internal amplification features of the model.

- Lower persistence implies that perceived substitution and income effects are distorted with the latter typically underestimated relative to the former.

- Distortions disappear if:

i) the non-cyclical component has low power at the business cycle frequencies. Need for this that the volatility of the non-cyclical component is considerably smaller than the volatility of the cyclical one.

ii) The prior eliminates the distortions induced by the penalty functions.

Question: What if we fit the filtered version of the model to the filtered data? (CKM (2008))

- Log-likelihood=$A_1(\theta)^* = \frac{1}{\pi}\sum_{\omega_j} \log\det G_\theta(\omega_j) I_\omega + A_2(\theta)$. Suppose that $I_\omega = I_{[\omega_1,\omega_2]}$.

- $A_1(\theta)^*$ matters only at business cycle frequencies while the penalty function is present at all frequencies.

- If the penalty is more important in the low frequencies (typical case) parameters adjusted to make $[G_\theta(\omega_j)]$ close to $F(\omega_j)$ at these frequencies.

-Procedure implies that the model is fitted to the low frequencies components of the data!!!

i) Volatility of the shocks will be generally underestimated.

ii) Persistence overestimated.

iii) Since less noise is perceived, decision rules will imply a higher degree of predictability of simulated time series.

iv) Perceived substitution and income effects are distorted with the latter overestimated.

How can we avoid distortions?

- Build models with non-cyclical components (difficult).

- Use filters which flexibly adapt, see Gorodnichenko and Ng (2007) and Eklund, et al. (2008).

- ?

Advantages of suggested approach:

• No need to take a stand on the properties of the non-cyclical component and on the choice of filter to tone down its importance - specification errors and biases limited.

• Estimated cyclical component not localized at particular frequencies of the spectrum.

- Cyclical, non-cyclical and measurement error fluctuations driven by different and orthogonal shocks. But model is observationally equivalent to one where cyclical and non-cyclical are correlated.

An experiment, again

- Simulate data from the model, assuming that the preference shock has two components: a nonstationary one and a stationary one (the properties of the other three shocks are unchanged).

- Variance of the non-cyclical shock is large or small relative to the variance of the other shocks.

- Use same Bayesian approach, same prior for structural parameters and gamma priors with large variance for non-structural ones.

- Compute the model-based cyclical component; calculate the autocorrelation function and the log spectrum of output after passing it through LT, HP, FOD, BP.

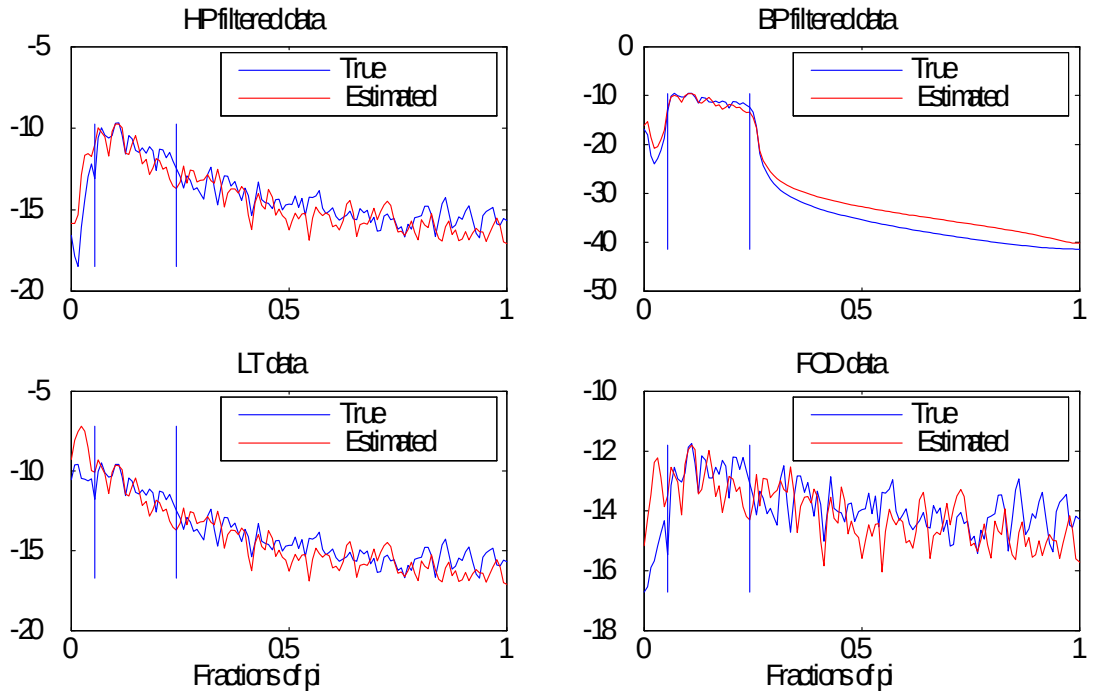
		Small variance			Large variance	
	True	Median	(s.e)	True	Median	(s.e)
σ_c	3.00	3.68	(0.40)	3.00	3.26	(0.29)
σ_n	0.70	0.54	(0.14)	0.70	0.80	(0.13)
h	0.70	0.55	(0.04)	0.70	0.77	(0.04)
α	0.60	0.19	(0.03)	0.60	0.41	(0.04)
ϵ	7.00	6.19	(0.07)	7.00	6.95	(0.09)
ρ_r	0.20	0.16	(0.04)	0.24	0.31	(0.04)
ρ_π	1.30	1.30	(0.04)	1.30	1.25	(0.03)
ρ_y	0.05	0.07	(0.03)	0.05	0.08	(0.10)
ζ_p	0.80	0.78	(0.04)	0.80	0.72	(0.02)
ρ_χ	0.50	0.53	(0.04)	0.50	0.69	(0.05)
ρ_z	0.80	0.71	(0.03)	0.80	0.90	(0.03)
σ_χ	0.011	0.012	(0.0003)	0.011	0.012	(0.0003)
σ_z	0.005	0.006	(0.0001)	0.005	0.007	(0.0001)
σ_{mp}	0.001	0.002	(0.0004)	0.001	0.002	(0.0004)
σ_μ	0.206	0.158	(0.0006)	0.206	0.1273	(0.0004)
σ_χ^{nc}	0.02			0.23		

Parameters estimates using flexible specification. σ_χ^{nc} is the standard error of the shock to the non-cyclical component.

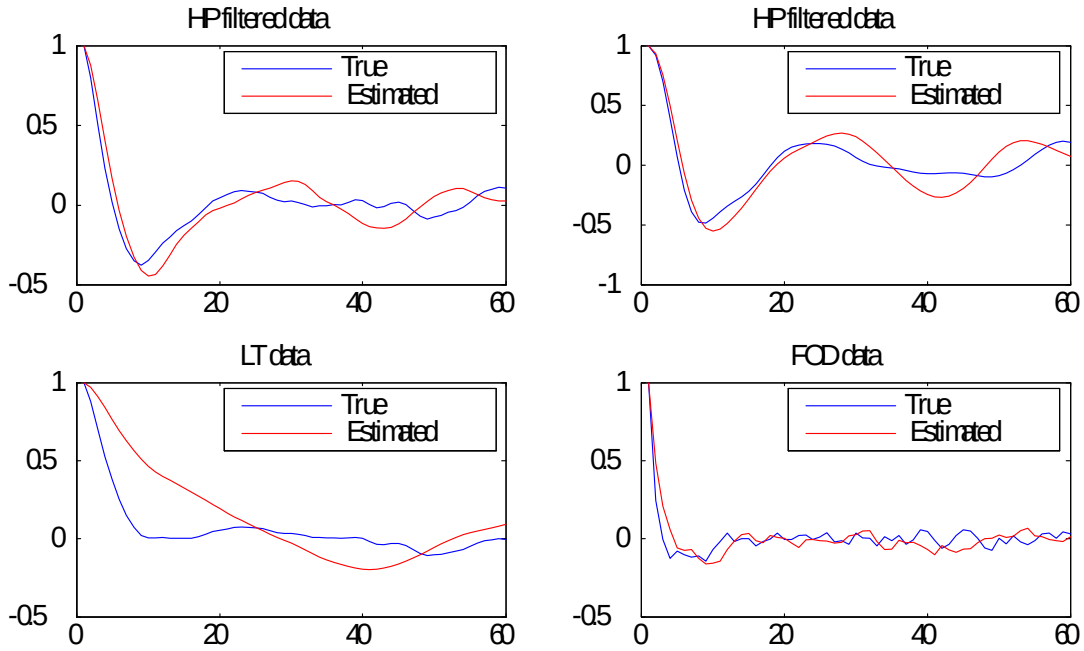
- Estimates of the structural parameters are roughly unchanged in two specifications.

- Estimates are precise but the median is not the true value (problem bigger for α or σ_n which are only weakly identified).

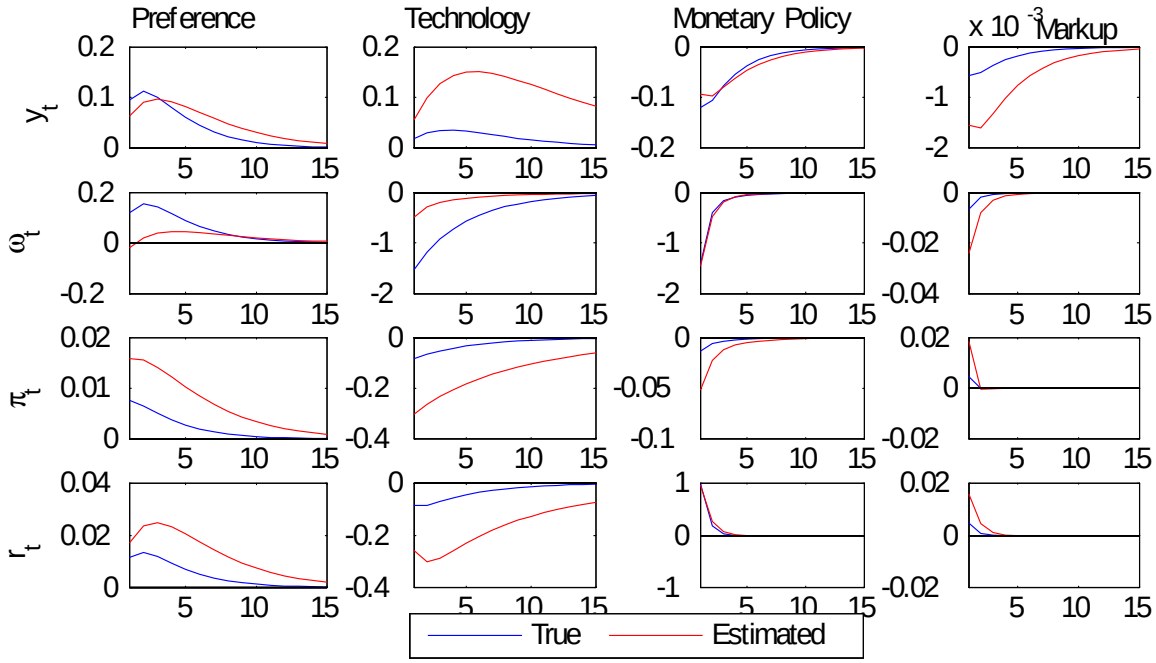
- The relative magnitude of the various shocks and their persistence is well estimated. Hence, true and estimated decision rules are similar.



Model based cyclical output spectra, true and estimated, different filtering. Vertical bars indicate the frequencies where cycles with 8-32 quarters periodicities are located



Autocorrelation function of filtered cyclical component, true and estimated



Model based IRF, true and estimated.

- The true and estimated log spectrum and the autocorrelation function of the model-based cyclical component close, regardless of the filter.

- Both true and estimate cyclical components have power at all frequencies of the spectrum.

Actual data: do we get a different story?

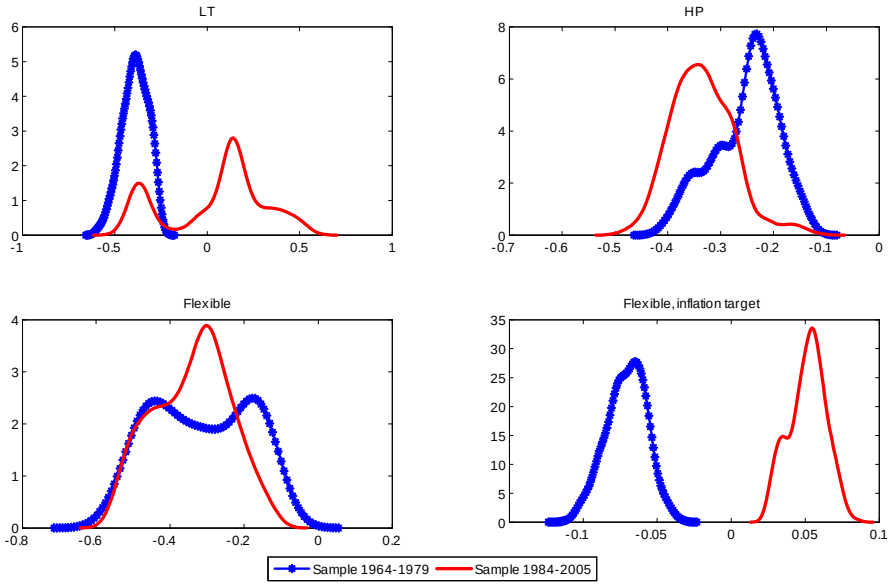


Figure 5: Posterior distributions of the policy activism parameter, samples 1964:1-1979:4 and 1984:1-2007:4. LT refers to linearly detrended data, HP to Hodrick and Prescott filtered data and Flexible to the approach the paper suggests

	LT		FOD		Flexible	
	Output	Inflation	Output	Inflation	Output	Inflation
TFP shocks	0.01	0.04	0.00	0.01	0.01	0.19
Gov. expenditure shocks	0.00	0.00	0.00	0.00	0.00	0.02
Investment shocks	0.08	0.00	0.00	0.00	0.00	0.05
Monetary policy shocks	0.01	0.00	0.00	0.00	0.00	0.01
Price markup shocks	0.75(*)	0.88(*)	0.91(*)	0.90(*)	0.00	0.21
Wage markup shocks	0.00	0.01	0.08	0.08	0.03	0.49(*)
Preference shocks	0.11	0.04	0.00	0.00	0.94(*)	0.00

Variance decomposition at the 5 years horizon. Estimates are obtained using the median of the posterior of the parameters. A (*) indicates that the 68 percent highest credible set is entirely above 0.10. The model and the data set are the same as in Smets Wouters (2007). LT refers to linearly detrended data, FOD to growth rates and Flexible to the approach this paper suggests.

Non linear DSGE models

$$y_{2t+1} = h_1(y_{2t}, \epsilon_{1t}, \theta) \quad (79)$$
$$y_{1t} = h_2(y_{2t}, \epsilon_{2t}, \theta) \quad (80)$$

ϵ_{2t} = measurement errors, ϵ_{1t} = structural shocks, θ = vector of structural parameters, y_{2t} = vector of states, y_{1t} = vector of controls. Let $y_t = (y_{1t}, y_{2t})$, $\epsilon_t = (\epsilon_{1t}, \epsilon_{2t})$, $y^{t-1} = (y_0, \ldots, y_{t-1})$ and $\epsilon^t = (\epsilon_1, \ldots, \epsilon_t)$.

• Likelihood is $\mathcal{L}(y^T, \theta y_{20}) = \prod_{t=1}^{T} f(y_t|y^{t-1}, \theta) f(y_{20}, \theta)$. Integrating the initial conditions y_{20} and the shocks out, we have:

$$\mathcal{L}(y^T, \theta) = \int [\prod_{t=1}^{T} \int f(y_t|\epsilon^t, y^{t-1}, y_{20}, \theta) f(\epsilon^t|y^{t-1}, y_{20}, \theta) d\epsilon^t] f(y_{20}, \theta) dy_{20} \quad (81)$$

(81) is intractable.

• If we have L draws for y_{20} from $f(y_{20}, \theta)$ and L draws for $\epsilon^{t|t-1,l}$, $l = 1, \ldots, L, \quad t = 1, \ldots, T$, from $f(\epsilon^t | y^{t-1}, y_{20}, \theta)$ approximate (81) with

$$\mathcal{L}(y^T, \theta) = \frac{1}{L}[\prod_{t=1}^{T} \frac{1}{L} \sum_{l} f(y_t | \epsilon^{t|t-1,l}, y^{t-1}, y_{20}^{l}, \theta)] \tag{82}$$

Drawing from $f(y_{20}, \theta)$ is simple; drawing from $f(\epsilon^t | y^{t-1}, y_{20}, \theta)$ complicated. Fernandez-Villaverde and Rubio-Ramirez (2004):
use $f(\epsilon^{t-1} | y^{t-1}, y_{20}, \theta)$ as importance sampling for $f(\epsilon^t | y^{t-1}, y_{20}, \theta)$:

- Draw y_{20}^l from $f(y_{20}, \theta)$. Draw $\epsilon^{t|t-1,l}$ L times from $f(\epsilon^t|y^{t-1}, y_{20}^l, \theta) = f(\epsilon^{t-1}|y^{t-1}, y_{20}^l, \theta) f(\epsilon_t|\theta)$.

- Construct $IR_t^l = \frac{f(y_t|\epsilon^{t|t-1,l}, y^{t-1}, y_{20}^l, \theta)}{\sum_{l=1}^L f(y_t|\epsilon^{t|t-1,l}, y^{t-1}, y_{20}^l, \theta)}$ and assign it to each draw $\epsilon^{t|t-1,l}$.

- Resample from $\{\epsilon^{t|t-1,l}\}_{l=1}^L$ with probabilities equal to IR_t^l.

- Repeat above steps for every $t = 1, 2, \ldots, T$.

Step 3) is crucial, if omitted, only one particle will asymptotically remain and the integral in (81) diverges as $T \to \infty$.

• Algorithm is computationally demanding. You need a MC within a MC. Fernandez-Villaverde and Rubio-Ramirez (2004): some improvements over linear specifications.

Bayesian methods for state space models

Fabio Canova
EUI and CEPR
November 2012

Outline

- State Space Models and Kalman filter
- Classical Inference in state space models
- Gibbs sampler for state space models
- Application 1: TVC-VARs
- Application 2: Factor models
- Application 3: Stochastic volatility
- Application 4: Markov switching models

References

Albert, J. and Chib, S. (1993) Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts, *Journal of Business and Economic Statistics*, 11, 1-16.

Chib, S. (1996) Calculating Posterior Distributions and Model Estimates in Markov Mixture Models, *Journal of Econometrics*, 75, 79-98.

Fruhwirth-Schnatter, S (2001) MCMC estimation of classical and Dynamic switching and Mixture Models *Journal of the American Statistical Association*, 96, 194-209.

Geweke, J. (1994), Comment to Jacquier, Polson and Rossi , *Journal of Business and Economic Statistics*, 12, 397-398.

Geweke, J. and Zhou, G. (1996) Measuring the Pricing Error of the Arbitrage Pricing Theory, *Review of Financial Studies*, 9, 557-587.

Otrok, C. and Whitemann, C. (1998), "Bayesian Leading Indicators: measuring and Predicting Economic Conditions in Iowa", *International Economic Review*, 39, 997-1114.

Jacquier, E., Polson N. and Rossi, P. (1994), " Bayesian Analysis of Stochastic Volatility Models", *Journal of Business and Economic Statistics*, 12, 371-417.

Kim, C. and Nelson, C. (1999), *State Space Models with Regime Switching*, MIT Press, London, UK.

McCulloch, R. and R. Tsay (1994) Statistical Analysis of Economic Time Series via Markov Switching Models", *Journal of Time Series Analysis*, 15, 521-539.

Roll, R. and Ross, S. (1980), " An empirical Investigation of the Arbitrage Pricing Theory, *Journal of Finance*, 35, 1073-1103.

Ross, S. (1976) "The Arbitrage Theory of the Capital Asset Pricing", *Journal of Economic Theory*, 13, 341-360.

Sims, C., Waggoner, D., and Zha, T. (2008) Methods for Inference in LargeMultiple-Equation Markov-Switching Models, Journal of Econometrics, 146, 255-274.

Cogley, T., Morozov, and Sargent, T. (2005). Bayesian Prediction Intervals in Evolving Monetary Systems, Journal of Economic Dynamics and Control.

1 State Space Models

$$y_t = x_t'\alpha_t + v_{1t} \quad (1)$$
$$\alpha_t = \mathbb{D}_1\alpha_{t-1} + \mathbb{D}_2 v_{2t} \quad (2)$$

x_t' is $m \times k$, $v_{1t} \sim \ iid\, \mathbb{N}(0, \Sigma_1)$; $\mathbb{D}_1, \mathbb{D}_2$ are $k \times k$ and $v_{2t} \sim \ iid\, \mathbb{N}(0, \Sigma_2)$. $E(v_{1t}v_{2\tau}') = 0$ and $E(v_{1t}\alpha_0') = 0\ \forall t, \tau$.

- The (1) is called measurement (observation) equation, (2) transition (state) equation.

- This class of models is general and flexible: many specifications fit (1)-(2).

Example 1 *An ARMA(2,1):* $y_t = A_1 y_{t-1} + A_2 y_{t-2} + e_t + D_1 e_{t-1}$ *can be written as:*

$$y_t = [1\ 0] \begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix}$$

$$\begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix} = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ A_2 y_{t-2} + D_1 e_{t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ D_1 \end{bmatrix} e_t$$

which fits (1)-(2) for $\alpha_t = \begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix}$, $\mathbb{D}_1 = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix}$, $\mathbb{D}_2 = \begin{bmatrix} 1 \\ D_1 \end{bmatrix}$, $x_t' = [1,\ 0]$, $\Sigma_1 = 0$, $\Sigma_2 = \sigma_e^2$.

In general, any ARMA model $\phi(\ell) y_t = \theta(\ell) \epsilon_t$ can be transformed into a state space model by setting $\phi(\ell) x_t = \epsilon_t$ (transition equation) and $y_t = \theta(\ell) x_t$ (measurement equation).

Example 2 *A VAR(q):* $y_t = A(L)y_{t-1} + e_t$ *is also a state space model. Use companion form representation* $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + \mathbb{E}_t$ *where* $\mathbb{E}_t = [e_t, 0, \ldots 0]'$ *and*

$$\mathbb{A} = \begin{bmatrix} A_1 & A_2 & \ldots & \ldots & A_q \\ I & 0 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & I & 0 \end{bmatrix}$$

Trivially fits (1)-(2) for $x_t' = [I, 0, \ldots 0], \alpha_t = [y_t', y_{t-1}', \ldots, y_{t-q}'],\ \mathbb{D}_1 = \mathbb{A},\ \mathbb{D}_2 = I,\ \Sigma_1 = 0,\ v_{2t} = \mathbb{E}_t,$

- A time varying coefficient model is a state space model:

$$y_t = x_t'\alpha_t + v_{1t} \quad (3)$$
$$\alpha_t = \mathbb{D}\alpha_{t-1} + v_{2t} \quad (4)$$

- A dynamic factor model is a state space model:

$$y_t = \alpha_0 + \alpha_1 x_t + v_{1t} \quad (5)$$
$$x_t = \rho x_{t-1} + v_{2t} \quad (6)$$

where y_t is a $M \times 1$ vector and x_t is a scalar.

- A stochastic volatility model is also a state space model

$$y_t = \sigma_t v_{1t} \quad (7)$$
$$\log \sigma_t = \log \sigma_{t-1} + v_{2t} \quad (8)$$

Very much used in finance, similar to GARCH but more flexible. (Careful: both σ_t and v_{1t} are random here so the model is non-linear).

• A Markov switching model has also the form of (5) and(6) but v_{1t} and v_{2t} are not normal.

Example 3 *(Ex-ante vs. ex-post real rate of interest) Here* $\alpha_t \equiv i_t - \pi_t^e = \phi\alpha_{t-1} + v_{2t}$ *(transition equation) The observed real rate* $y_t \equiv i_t - \pi_t = \alpha_t + v_{1t}$*, where* v_{1t} *is a measurement error (observable equation).*

Example 4 *(Common trends)* $\alpha_t = \alpha_{t-1} + v_{2t}$ *is a one dimensional process. Then* $x_t' = x'$ *are the loadings on the trend.*

• The log-linearized solution of a DSGE model is a state space model

$$y_{2t} = \mathcal{A}_{22}(\theta) y_{2t-1} + \mathcal{A}_{21}(\theta) y_{3t} \quad (9)$$

$$y_{1t} = \mathcal{A}_1(\theta) y_{2t} = \mathcal{A}_{11}(\theta) y_{2t-1} + \mathcal{A}_{12}(\theta) y_{3t} \quad (10)$$

y_{2t} = endogenous and exogenous states, y_{1t} = endogenous controls, y_{3t} the innovations in the driving forces. $\mathcal{A}_{ij}(\theta), i, j = 1, 2$ are time invariant (reduced form) matrices which depend on θ, the structural parameters of preferences, technologies, policies, etc.

• Since here we are typically interested in θ (and not in $\mathcal{A}$'s), the DSGE problem is slightly different from those typically considered here since the mapping from θ to $\mathcal{A}$ is non-linear.

2 Kalman filter

Let $y^{t-1} = (y_1, \ldots, y_{t-1})$. The Kalman filter (KF) computes optimal forecasts of y_t and recursive estimates of the mean and variance of α_t, given y^{t-1}, α_0 , for models like (1)-(2).

- The Kalman smoother (KS) computes estimates of the mean and variance of α_t for models like (1)-(2). given $y^T = (y_1, \ldots, y_{t-1}, y_t, y_{t+1} \ldots, y_T, \alpha_0)$.

Let $\alpha_{t|t}$ be the optimal (MSE) estimator of α_t using information up to t; and let $\Omega_{t|t}$ the MSE of α_t. Let $\mathbb{D}_1$ and $\mathbb{D}_2$ be known; assume $y_t, x_t, t = 1, \ldots T$ is available.

- Since the errors in (1) and (2) are normal $(y_t, \alpha_t | y^{t-1}, \alpha_0)$ are jointly normal and $(\alpha_t | y^t, \alpha_0)$ is also normal. This means that we need to keep track only of the mean and the variance to fully account for the distribution.

The KF requires the following steps:

• Initial Conditions: If all eigenvalues of $\mathbb{D}_1$ are less then one in absolute value then $\alpha_{1|0} = E(\alpha_1)$ and $\Omega_{1|0} = \mathbb{D}_1\Omega_{1|0}\mathbb{D}_1' + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2'$ or $vec(\Omega_{1|0}) = (I - (\mathbb{D}_1 \otimes \mathbb{D}_1')^{-1})vec(\mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2')$, i.e. choose as initial conditions the unconditional mean and variance of the process.

If initial conditions can not be drawn from the unconditional distribution: guess e.g. $\alpha_{1|0} = 0$, $\Omega_{1|0} = \kappa * I$, κ large.

- Forecast of y_t and prediction of the state α_t, given y^{t-1}.

$$
\begin{aligned}
E(y_{t|t-1}) &= x'\alpha_{t|t-1} && (11)\\
E(\alpha_{t|t-1}) &= \mathbb{D}_1\alpha_{t-1|t-1} && (12)\\
E(y_t - y_{t|t-1})(y_t - y_{t|t-1})' &= E(x'(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'x) + \Sigma_1 \\
&= x'\Omega_{t|t-1}x + \Sigma_1 \equiv \Sigma_{t|t-1} && (13)\\
E(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})') &\equiv \Omega_{t|t-1} = \mathbb{D}_1\Omega_{t-1|t-1}\mathbb{D}_1' + \mathbb{D}_2\Sigma_2\mathbb{D}_2' && (14)\\
E(y_t - y_{t|t-1})(\alpha_t - y_{t|t-1})' &= \Omega_{t|t-1}x' && (15)
\end{aligned}
$$

- Update estimates after observing y_t:

$$
\begin{aligned}
\alpha_{t|t} &= \alpha_{t|t-1} + \Omega_{t|t-1}x'\Sigma_{t|t-1}^{-1}(y_t - x\alpha_{t|t-1}) && (16)\\
\Omega_{t|t} &= \Omega_{t|t-1} - \Omega_{t|t-1}x\Sigma_{t|t-1}^{-1}x'\Omega_{t|t-1} && (17)
\end{aligned}
$$

where $\epsilon_t = y_t - x'\alpha_{t|t-1}$ is the one-step ahead forecast error, and $\mathfrak{K}_t = \Omega_{t|t-1}x\Sigma_{t|t-1}^{-1}$ is the Kalman gain.

- Forecast the state next period:

$$\alpha_{t+1|t} = \mathbb{D}_1 \alpha_{t|t} = \mathbb{D}_1 \alpha_{t|t-1} + \mathfrak{K}_t \epsilon_t \tag{18}$$

$$\Omega_{t+1|t} = \mathbb{D}_1 \Omega_{t|t} \mathbb{D}_1' + \mathbb{D}_2 \Sigma_{v_2} \mathbb{D}_2' \tag{19}$$

- Repeat previous steps until $t = T$.

Note: since $\Omega_{t|t-1} x' = E(\alpha_t - \alpha_{t|t-1})(y_t - y_{t|t-1})$, α_t is updated using linear OLS projection of $\alpha_t - \alpha_{t|t-1}$ on $y_t - y_{t|t-1}$ multiplied by the prediction error. Similarly, since $\Omega_{t|t} = E(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'$, it is updated using covariance between forecast errors in the two equations and the MSE error of the forecasts of y_t.

• The Kalman filter can be used to evaluate the likelihood function of a state space model since $L(y^T|\alpha) = L(y_0|\alpha)\prod_{t=1}^{T} L(y_t|y_{t-1},\alpha))$ and each $L(y_t|y_{t-1},\alpha)$ is Normal with mean $y_{t|t-1}$ and variance $\Sigma_{t|t-1}$, both of which are produced recursively by the Kalman filter (see (11)-(13)).

• The initialization of the KF is typically difficult. Better to use a large covariance matrix if you expect the model to be nearly non-stationary.

The assumption that errors are normal may not be great. If they are, the Kalman filter is Best predictor of α_t; otherwise it is only BLUP!

2.1 Kalman smoother

- Computes mean and variance of $(\alpha_t|y^T)$ with the output of the Kalman filter.

• Starting from y_T, and setting $t = T-1, \ldots, 1$, we have

$$\alpha_{t|T} = \alpha_{t|t} + (\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})(\alpha_{t+1|T} - \alpha_{t+1|t}) \quad (20)$$

$$\Omega_{t|T} = \Omega_{t|t} - (\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})(\Omega_{t+1|T} - \Omega_{t+1|t})(\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})' \quad (21)$$

where $\alpha_{t|t}, (\Omega_{t|t})$ are produced by the KF. Equations (20)-(21) define the Kalman smoother. They can be used for signal extraction problems, e.g. to find the state at t using the information available up to T.

Example 5 $y_t = A_1 y_{t-1} + A_2 y_{t-2} + e_t$. *Then* $\alpha = [y_t, y_{t-1}]'$, $v_{2t} = [e_t, 0]$, $\mathbb{D}_1 = \begin{bmatrix} A_1 & A_2 \\ 1 & 0 \end{bmatrix}$, $\Sigma_{v_2} = \begin{bmatrix} \sigma_e^2 & 0 \\ 0 & 0 \end{bmatrix}$, $\mathbb{D}_2 = I, v_{1t} = 0, x_1' = [1, 0]$.

Forecast of y_t $E_{t-1} y_t = A_1 y_{t-1} + A_2 y_{t-2}$; $E(y_t - E_{t-1} y_t)^2 = \sigma_e^2$. *Then* $\alpha_{t|t} = \alpha_{t|t-1} + \Omega_{t|t-1} \sigma_e^{-2} v_{2t}$.

Example 6 *(Trend):* $\alpha_t = \alpha_{t-1}$; *GDP is* $y_t = \alpha_t + v_{1t}$, v_{1t} iid $\mathbb{N}(0, \sigma_{v_1}^2)$. *Then* $\Omega_{t|t} = \Omega_{t|t-1} - \Omega_{t|t-1}(\Omega_{t|t-1} + \sigma_{v_1}^2)^{-1} \Omega_{t|t-1} = \dfrac{\Omega_{t|t-1}}{1 + \frac{\Omega_{t|t-1}}{\sigma_{v_1}^2}} = \dfrac{\Omega_{t-1|t-1}}{1 + \frac{\Omega_{t-1|t-1}}{\sigma_{v_1}^2}}$

and $\alpha_{t+1|t+1} = \alpha_{t|t} + \dfrac{\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}{1 + t\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}} (y_t - \alpha_{t|t})$. *As* $t \to \infty$, $\alpha_{t+1|t+1} = \alpha_{t|t} = \bar{\alpha}$.

3 Classical ML estimation of state space models

- EM algorithm:

1) Choose initial $\phi_0 = (\mathbb{D}_1, \mathbb{D}_2, \Sigma_1, \Sigma_2)$ and some $(\alpha_{1|0}, \Omega_{1|0})$.

2) Run the KF and, for each t, save $\epsilon_t = y_t - y_{t|t-1}$ and $\Sigma_{t|t-1}$. Construct the conditional likelihood $L(y_t|y_{t-1}, \phi)$ and $\prod_i L(y_i|y_{i-1}) * L(y_0)$.

3) Update initial estimates of ϕ using any methods (gradient, etc.).

4) Repeat steps 2) through 3) until $|\phi^l - \phi^{l-1}| \leq \iota$; $|\mathcal{L}(\phi^l) - \mathcal{L}(\phi^{l-1})| < \iota$; or $(\frac{\partial \mathcal{L}(\phi)}{\partial \phi})|_{\phi=\phi^l} < \iota$, or all of them, ι small.

• Once you have converged, compute standard errors for the estimates using square root of diagonal of Hessian $H(\phi_{ML}) = \frac{\partial^2 L}{\partial \phi \partial \phi'}|_{\phi_{ML}}$.

- For some state space models, direct maximization of the likelihood is too complicated. Use MCMC methods.

- For Bayesian analysis: $g(\phi|y^T) \propto L(\phi|y^T)g(\phi)$ so also in this case we even need to evaluate the likelihood function with Kalman filter to compute the kernel of the posterior.

- Bayesian estimation of certain state space models is relatively easy. In others estimation is more complicated.

4 Gibbs sampler for (linear) state space models

Consider the model:

$$y_t = x_t'\alpha_t + v_{1t} \quad (22)$$

$$\alpha_t = \mathbb{D}\alpha_{t-1} + v_{2t} \quad (23)$$

$v_1 \sim N(0, \Sigma_1),\ v_2 \sim N(0, \Sigma_2)$. How do you apply the Gibbs sampler?

- There are four groups of parameters $\mathbb{D}, \Sigma_1, \Sigma_2, \alpha^t = (\alpha_1, \ldots, \alpha_t)$. Need to find conditional posterior of each group.

- Easy for the first three. If $g(vec(\mathbb{D}))$ is normal and $g(vec(\Sigma_1))$ and $g(vec\Sigma_2))$ are inverted Wishart, $g(vec(\mathbb{D})|y^T, \Sigma_1, \Sigma_2, \alpha^t)$is normal, $g(\Sigma_1|vec(\mathbb{D}), y^T, \Sigma_2, \alpha^t)$and $g(\Sigma_2|vec(\mathbb{D}), y^T, \Sigma_1, \alpha^t)$are inverted Wishart.

- Slightly more complicated to find $g(\alpha^t|\Sigma_2, vec(\mathbb{D}), y^T, \Sigma_1)$. Note that

$$\begin{aligned} g(\alpha^t|\Sigma_2, vec(\mathbb{D}), y^T, \Sigma_1) &= \\ g(\alpha_t|\Sigma_2, vec(\mathbb{D}), y^T, \Sigma_1) \prod_{i=1}^{t-1} g(\alpha_i|\alpha_{i+1}, \Sigma_2, vec(\mathbb{D}), y^T, \Sigma_1) &= \\ g(\alpha_t|\Sigma_2, vec(\mathbb{D}), y^t, \Sigma_1) \prod_{i=1}^{t-1} g(\alpha_i|\alpha_{i+1}, \Sigma_2, vec(\mathbb{D}), y^i, \Sigma_1) & \qquad (24) \end{aligned}$$

since the model has a Markovian structure. Hence, to draw from $g(\alpha^t|\Sigma_2, vec(\mathbb{D}), y^T, \Sigma_1)$ we can draw recursively from the marginal and conditionals in (24).

- It is easy to draw from $g(\alpha_t|\Sigma_2, vec(\mathbb{D}), y^t, \Sigma_1)$ since this Normal with mean $\alpha_{t|t}$ and variance $\Omega_{t|t}$ and both these quantities are produced by the Kalman filter.

- How to draw from these conditionals? These are also normal since for any i, α_i, α_{i+1} are jointly normal (so conditionals are also normal). To characterize normal distributions we need just their first and second moments. What are they?

- They are $\alpha_{i|i+1}$ and $\Omega_{i|i+1}$ and are obtained from the Kalman smoother.

$$\alpha_{i|i+1} = \alpha_{i|i} + (\Omega_{i|i}\mathbb{D}'\Omega^{-1}_{i+1|i})(\alpha_{i+1} - \alpha_{i+1|i}) \quad (25)$$

$$\Omega_{i|i+1} = \Omega_{i|i} - (\Omega_{i|i}\mathbb{D}'\Omega^{-1}_{i+1|i})\mathbb{D}'\Omega^{-1}_{i|i})' \quad (26)$$

Conclusions:

• The Gibbs sampler can cycle using $g(vec(\mathbb{D})|y^T, \Sigma_1, \Sigma_2, \alpha^t)$ (a normal), $g(\Sigma_1|vec(\mathbb{D})|y^T, \Sigma_2, \alpha^t)$ and $g(\Sigma_2|vec(\mathbb{D})|y^T, \Sigma_1, \alpha^t)$ (two inverted Wishart) and $g(\alpha^t|\Sigma_2, vec(\mathbb{D}), y^T, \Sigma_1)$ (again a bunch of normals).

• To draw from the latter run the Kalman filter and Kalman smoother through the sample and save $\alpha_{t|t}, \Omega_{t|t}$, and $\alpha_{i|i+1}, \Omega_{i|i+1}$, $i = t-1, t-2, \ldots, 1$. Then a draw for α^t can be made by drawing separately each element from normals with the above means and variances.

Step 2 has do be done within each draw of the Gibbs sampler and this is time consuming. For complex problems may want to reduce costs by using steady state Kalman gain i.e. $\mathfrak{K}_t = \mathfrak{K}$; this cuts computation time dramatically.

• We estimate the vector of time varying parameters (α^t) and time-invariant ones ($\mathbb{D}, \Sigma_1, \Sigma_2$) jointly.

4.1 A TVC-AR model

$$y_t = y'_{t-1}\alpha_t + v_{1t} \quad (27)$$
$$\alpha_t = \alpha_{t-1} + v_{2t} \quad (28)$$

$v_1 \sim N(0, \Sigma_1),\ v_2 \sim N(0, \Sigma_2)$.

- This model is a special case of the previous general state space model. The object of interest $(\Sigma_e, \Sigma_a, \alpha^t)$ and to draw a sample say for α^t we simply need to go through the same steps as before.

- Note that each α_t could be a scalar or a vector. If (27) is a VAR, then α_t in (28) could be of large dimension.

• Needs to be careful with TVC models because of identification and pile up problems.

Example 7 *Suppose the true model is* $\pi_t = \pi_{t-1} + e_t$ *and suppose one estimates* $\pi_t = \alpha_{0t} + \alpha_{1t} + e_t$ *where* $\alpha_t = \alpha_{t-1} + v_t$ *and* $\alpha_t = (\alpha_{0t}, \alpha_{1t})$.

Identification problems!! Model with $\alpha_{1t} = 0, \ \forall t$ *is observationally equivalent to a model with* $\alpha_{0t} = 0, \ \forall t$ *and to the true model. Problems if one is interested in measuring e.g. time varying inflation persistence. The prior will determine which model will be chosen.*

Example 8 *Suppose* $y_t = \alpha_t + e_t$, $\alpha_t = \alpha_{t-1} + \gamma v_t$, *where* γ *is unknown, and the variance of* e_t *and* v_t *is unity. This model has an ARIMA representation* $\Delta y_t = \Delta e_t + \gamma v_t$ *or* $\Delta y_t = \epsilon_t - \theta \epsilon_{t-1}$. *where* $\sigma_\epsilon^2(1 + \theta^2) = 2 + \gamma^2$ *and* $-\theta\sigma_\epsilon^2 = -1$ *so that* $\frac{1+\theta^2}{\theta} = 2 + \gamma^2$. *Thus, if* γ^2 *tends to zero,* θ *tends to 1, so it can not be identified (it cancels out).*

This is called pile up problem (γ *has a point mass at zero). To avoid it, typical to choose priors that keep* γ *away from zero.*

• What kind of priors do you want to use for TVC-VAR?

Need good guesses if α is of large dimensions. Also, the choice of parameters regulating the prior for Σ_2 crucial since

i) too little time variation, easy to get stuck at no time variation.

ii) too much time variation, the unobserved states wander around too much to fit the data as best as possible.

Using a training sample to tune up the priors helps a lot here!

- Possible to have a non-linear TVC model of the form

$$y_t = h_{1t}(\alpha_t) + e_t \quad e_t \sim (0, \Sigma_e)$$
$$\alpha_t = h_{2t}(\alpha_{t-1}) + v_t \quad v_t \sim (0, \Sigma_a) \tag{29}$$

where h_{1t} and h_{2t} are given but perhaps depend on unknown parameters. The same logic applies.

- Possible to consider a TVC model with non-normal disturbances. Assume $(\alpha_t|\alpha_{t-1}, \varpi_{1t}, \Sigma_a) \sim N(\alpha_{t-1}, \varpi_{1t}\Sigma_a)$ where e.g. $\varpi_{1t} \sim exp(2)$. Since $g(\alpha_t|\alpha_{t-1}, \varpi_{1t}, \Sigma_a) \sim N(\alpha_{t-1}, \varpi_{1t}\Sigma_a)$; $g(\varpi_{1t}|y^t, \alpha^t, V) \propto (\frac{1}{\varpi_{1t}})^{0.5}$ exp $\{-0.5\varpi_{1t} + (\alpha_t - \alpha_{t-1})'\varpi_{1t}^{-1}V^{-1}(\alpha_t - \alpha_{t-1})\}$. This is reciprocal of the inverse Gaussian distribution.

If ϖ_{1t} is $\chi^2(\bar{\nu})$, then $g(\alpha|y)$ is a t-distribution, with $T + \bar{\nu}$ degrees of freedom

Example 9 *(Canova et al. (2007)) Use a TVC-VAR model with* (Y, π, M, R) *for the US, Euro area and the UK. Draw 20000 vectors for* α^t *and keep one out of 5 of the last 10000. Plot the dynamics of output and inflation volatility and persistence and the time varying contribution of each shock.*

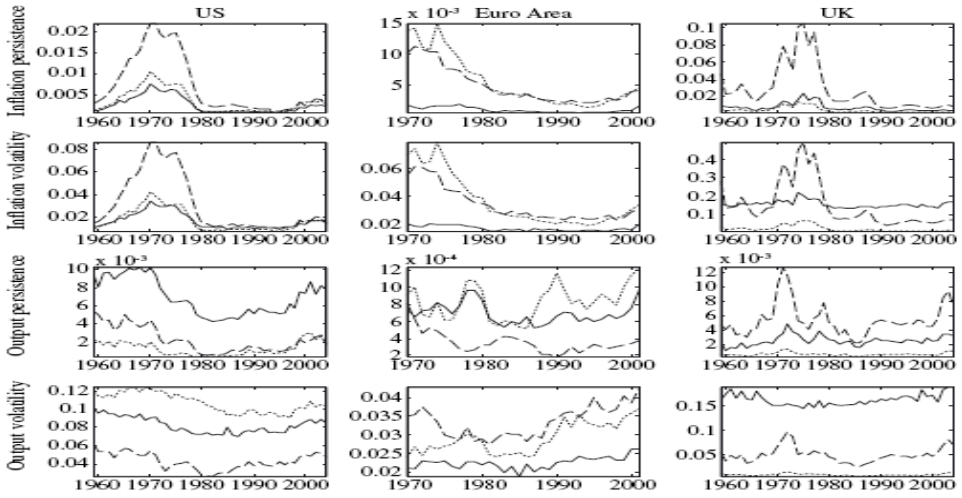


Fig. 4. *Sources of Output Growth and Inflation Dynamics*
Solid – supply shock. Dashed – demand shock. Dotted – monetary policy shock

4.2 Multi-country heterogeneous (panel) VARs.

Panel VARs (featuring dynamic interdependencies and heterogeneous dynamics) with some flexible restrictions on the coefficients generate observable factor models (Canova and Ciccarelli (2009)) or can be thought as hierarchical (state space) models. One such model is:

$$y_{it} = D_{it}(L)Y_{t-1} + F_{it}(L)W_{t-1} + e_{it}$$

$i = 1, ..., N$ countries, y_{it} is $G \times 1$, W_t are the exogenous variables, $Y_t = \left(y'_{1t}, \dots y'_{Nt}\right)'$.

- Parameter specific to each variable-country.

- Parameters time-varying.

- Allow for lagged and contemporaneous interdependencies $E(e_{it}, e_{jt}) \neq 0$.

- Impossible to estimate this model with classical unrestricted methods: each equation has $k = NGp + Mq$ time varying coefficients, and $r = NG$ equations. Even with fixed coefficients, T smaller than $k \times r$.

Short cuts:

- α_t does not depend on i (apart from fixed effect).

- no dynamic interdependencies (see Holtz Eakin et al. (1988) or Binder et al (2001)).

- use an indicator for interdependencies (see Pesaran et al. (2004))

-Parsimonious representation:

$$Y_t = X_t\alpha_t + E_t \qquad E_t \sim N(0, \Omega) \tag{30}$$

$$\alpha_t = \Xi\theta_t + u_t \qquad u_t \sim N(0, \Omega \otimes V) \tag{31}$$

where e.g. $\theta_t = [\lambda_t', \delta_t', \rho_t', \psi_t']'$.

- Factorize δ_t into components: θ_t is $s \times 1$ vector, $s << k * r$, Ξ_j are matrices with elements equal to zero or one.

Typical structure:

- λ_t captures movements in the coefficients vector δ_t common to all countries and variables (a 1×1 vector).

- δ_t is the country specific component (a $N \times 1$ vector).

- ρ_t is the variable specific component ($G \times 1$ vector).

- ψ_t is the exogenous variable component (1×1 vector).

- u_t captures unmodelled features of the coefficients vector.

- All factors in (31) time varying (for time invariant structures see below).

- Factorization is exact. Add error if omit some factors.

- With (31) over-parametrization dramatically reduced.

- Can treat (31) as part of the prior or part of the model. If the latter, we can test for the number of factors to be included.

Observable Index model

Using (31) into (30) we have

$$Y_t = \mathcal{Z}_{1t}\lambda_t + \mathcal{Z}_{2t}\delta_t + \mathcal{Z}_{3t}\rho_t + \mathcal{Z}_{4t}\psi_t + \upsilon_t = \mathcal{Z}_t\theta_t + \upsilon_t$$

$$\mathcal{Z}_{1t} = X_t\Xi_1, \mathcal{Z}_{2t} = X_t\Xi_2, \mathcal{Z}_{3t} = X_t\Xi_3, \mathcal{Z}_{4t} = X_t\Xi_4, \upsilon_t = E_t + X_t u_t$$

- Regressors of the model are averages of lags of the VAR variables. Dynamically span lagged interdependencies between variables and countries.

- $\lambda_t, \delta_t, \rho_t, \psi_t$ are the factor loadings. Time varying.

- Business cycle (common, country, variable) indicators easy to construct (observable and correlated).

- e.g.$\mathcal{Z}_{1t|t}\lambda_{t|t}$ is a coincident indicator of the common cycle in Y_t.

- Can make them leading, e.g. $\mathcal{Z}_{1t|t-1}\lambda_{t|t-1}$.

- Indicators emphasize low frequency movements, since they are average of lags of VAR variables. Good for medium term forecasting.

- Analysis feasible with small T and small N and when degrees of freedom in Panel VAR small. Estimate loadings θ_t not VAR coefficients α_t.

Example 10 $G = 2$ *variables,* $N = 2$ *countries,* 1 *lag, no exogenous:* α_t *is a vector* 16×1

$$\alpha_t = \Xi_1 \lambda_t + \Xi_2 \alpha_t + \Xi_3 \rho_t + u_t$$

λ_t *is scalar,* α_t *is* 2×1*,* ρ_t *is* 2×1*, and the VAR can be rewritten as*

$$\begin{bmatrix} y_t^1 \\ x_t^1 \\ y_t^2 \\ x_t^2 \end{bmatrix} = \begin{bmatrix} \mathcal{Z}_{1t} \\ \mathcal{Z}_{1t} \\ \mathcal{Z}_{1t} \\ \mathcal{Z}_{1t} \end{bmatrix} \lambda_t + \begin{bmatrix} \mathcal{Z}_{2,1,t} & 0 \\ \mathcal{Z}_{2,1,t} & 0 \\ 0 & \mathcal{Z}_{2,2,t} \\ 0 & \mathcal{Z}_{2,2,t} \end{bmatrix} \delta_t + \begin{bmatrix} \mathcal{Z}_{3,1,t} & 0 \\ 0 & \mathcal{Z}_{3,2,t} \\ \mathcal{Z}_{3,1,t} & 0 \\ 0 & \mathcal{Z}_{3,2,t} \end{bmatrix} \rho_t + v_t$$

e.g. $\mathcal{Z}_{1t} = y_{t-1}^1 + x_{t-1}^1 + y_{t-1}^2 + x_{t-1}^2$ *is the common information,* $\mathcal{Z}_{2,1,t} = y_{t-1}^1 + x_{t-1}^1$ *is country 1 information (across variables),* $\mathcal{Z}_{3,1,t} = y_{t-1}^1 + y_{t-1}^2$ *is variable* y *(across countries).*

- if λ_t *large relative to* δ_t*,* y_t^1 *and* x_t^1 *comove with* y_t^2 *and* x_t^2*.*

- *if* $\lambda_t = 0$*:* y_t^1 *and* x_t^1 *may drift apart from* y_t^2 *and* x_t^2*.*

- *A leading indicator for* Y_t *based on the common information is* $CLI_t = \mathcal{Z}_{1t}\lambda_t$*; a leading indicators based on common and unit specific information is* $CULI_t = \mathcal{Z}_{1t}\lambda_t + \mathcal{Z}_{2t}\delta_t$*, etc.*

• Del Negro and Schorfheide (2012): Second stage reduction is a cross sectional shrinkage prior.

Estimation (Hierarchical model)

$$
\begin{aligned}
Y_t &= X_t\alpha_t + E_t \qquad E_t \sim N(0, \Sigma) && (32)\\
\alpha_t &= \Xi\theta_t + u_t \qquad u_t \sim N(0, \Sigma \otimes W) && (33)\\
\theta_t &= \theta_{t-1} + \eta_t \qquad \eta_t \sim N(0, B_t) && (34)
\end{aligned}
$$

- E_t, u_t, η_t uncorrelated, $W = \sigma^2 I_k$, B_t could be time-varying, e.g. $B_t = \gamma_1 B_{t-1} + \gamma_2 B_0$, with $B_0 = diag(B_{01}, B_{02}, B_{03}, B_{04})$.

- Use Bayesian methods: get posterior distribution of $(\Omega, \{\theta_t\}_{t=1}^T, \sigma^2)$ and of transformations of interest: λ_t, α_t^j, $Y_{t+\tau|t}$, coincident and leading indicators, etc.

- Need prior densities for $(\Sigma, B_0, \theta_0, \sigma^2)$ (choose them proper but loose).

Estimation (state space model)

$$Y_t = \mathcal{Z}_t \theta_t + v_t \qquad v_t \sim (0, V_t \equiv \sigma_t \Sigma) \tag{35}$$

$$\theta_t = \theta_{t-1} + \eta_t \qquad \eta_t \sim N(0, B_t) \tag{36}$$

- Get posterior distribution of $(V_t, \{\theta_t\}_{t=1}^T)$, of the loading λ_t, α_t^j, and of $Y_{t+\tau|t}$, coincident and leading indicators, etc.

- Never estimate α directly, smaller computational burden.

- If $B_t = B$, this is a more complex state space model. Having time varying B_t rather than time varying Σ_t reduces computational burden.

- However, $v_t = E_t + X_t u_t$ so unless factorization is exact difficult to do structural analyses.

Priors for state space model: Still assume $B_t = \gamma_1 B_{t-1} + \gamma_2 B_0$, with $B_0 = diag\,(B_{01}, B_{02}, B_{03}, B_{04})$. Also make $B_{0i} = \phi_i * I,\ i = i, \ldots, f$.

$g(V_t^{-1}, \theta_0, \phi_i) = g(\sigma_t) g(\Sigma^{-1}) g(\theta_0) \prod_i g(\phi_i)$ and

- $g(\Sigma^{-1}) = W(\bar{\nu}_1, \bar{\Sigma}_1)$;

- $g(\phi_i) \propto (\phi_i)^{-1}$;

- $g(\theta_0) \propto 1$;

-$g(\sigma_t^{-1}) = G\left(\frac{\bar{\nu}_2}{2}, \frac{\bar{\nu}_2 \bar{s}_t}{2}\right)$; $\bar{s}_t^{-1} = E(\sigma_t^{-1})$.

- Treat $\bar{\nu}_1, \bar{\Sigma}_1, \bar{s}_t, \bar{\nu}_2$ as known (or estimable form a training sample).

- Note If $(v_t|\sigma_t) \sim N\,(0, \sigma_t \Sigma_E)$ given that $\sigma_t \sim$ Inv-$\chi^2\,(\bar{\nu}_2, \bar{s}_t)$, unconditionally v_t is t-distributed. As $\bar{\nu}_2 \to 0$, this prior becomes non-informative.

Conditional on θ, B_t, the likelihood is:

$$\propto \left(\prod_{t=1}^{T} \sigma_i\right)^{-NG/2} |\Sigma_E|^{-T/2} \exp\left[-\frac{1}{2}\sum_t \left(Y_t - X_t \Xi \theta_t\right)' \left(\sigma_t \Sigma_E\right)^{-1} \left(Y_t - X_t \Xi \theta_t\right)\right]$$

Given Y^T, the conditional posterior are:

$$\theta_t \mid Y^T, \psi_{-\theta_t}, \mathsf{I}_{t-1} \sim N\left(\bar{\theta}_{t|t}, \bar{\mathsf{R}}_{t|t}\right) \quad t \leq T,$$

$$\Sigma^{-1} \mid Y^T, \psi_{-\Sigma} \sim W\left(\nu_1 + T, \left[, \frac{\sum_t \left(Y_t - X_t \Xi \theta_t\right)\left(Y_t - X_t \Xi \theta_t\right)'}{\sigma_t} + \bar{\Sigma}_1^{-1}\right]^{-1}\right),$$

$$\phi_i \mid Y^T, \psi_{-b_i} \sim IG\left(T, \frac{\sum_t \left(\theta_t^i - \theta_{t-1}^i\right)' \left(\theta_t^i - \theta_{t-1}^i\right)}{2\xi_t}\right),$$

$$\sigma_t^{-1} \mid Y^T, \psi_{-\sigma_t} \sim G\left(\frac{\zeta + NG}{2}, \frac{\zeta s_t + \left(Y_t - X_t \Xi \theta_t\right)' \Omega^{-1} \left(Y_t - X_t \Xi \theta_t\right)}{2}\right),$$

where $\bar{\theta}_{t|t}$ and $\bar{\mathsf{R}}_{t|t}$ are obtained with the Kalman smoother, $\psi^*_{-i} = \psi^*$ minus parameter i. Here ξ_t is a function of γ_1 and γ_2.

• How do you decide the dimension of θ? Use Bayes factors.

- Model with i indices preferred to a model with $i+1$ indices, $i = 1, 2, \ldots, f_1 - 1$ if $\frac{f(Y^t|\mathcal{M}_i)}{f(Y^t|\mathcal{M}_{i+1})} > 1$; where $f(Y^t|M)$ is the marginal likelihood of model M.

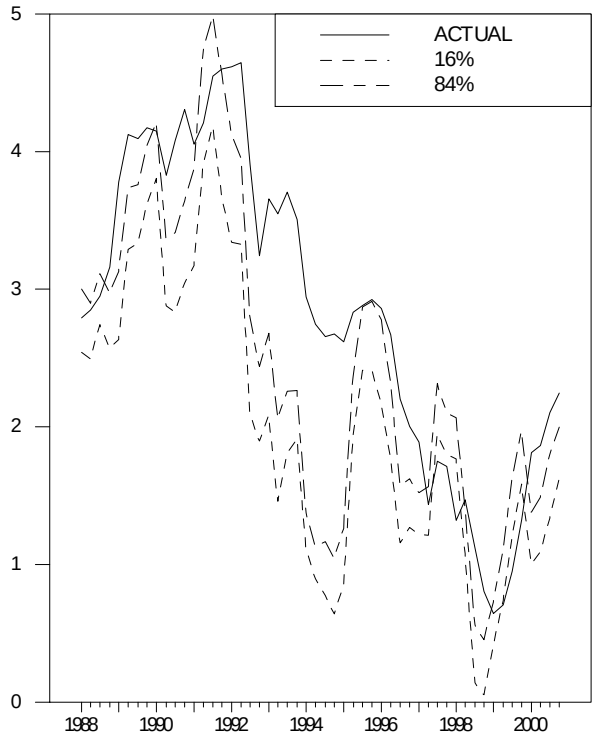
- Possibility of computing out-of-sample Bayes factors, i.e. instead of $f(Y^t|M)$ use $f(Y^{t+\tau}|\mathcal{M}_i) = \int f(Y^{t+\tau}|\theta_{t,i}, \mathcal{M}_i) g(\theta_{t,i}|\mathcal{M}_i) d\theta_{t,i}$ which is the predictive density of i indices for $Y^{t+\tau} = [y_{t+1}, \ldots, y_{t+\tau}]$, and $g(\theta_{t,i}|\mathcal{M}_i)$ the posterior density for θ_i.

Note that $f(Y^{t+\tau}|\mathcal{M}_i)$ can be computed using Gibbs sampler output: i.e. draw θ_t^l from the $g(\theta|y)$, construct $Y_{t+\tau}^l$ and prediction errors for each τ, use $\hat{f}(Y_{t+\tau}|\mathcal{M}_i) = [\frac{1}{J}\sum_j f(Y_{t+\tau}|\theta^{ji})^{-1}]^{-1}$; where θ^{ji} is j-th draws from model i (harmonic mean estimator).

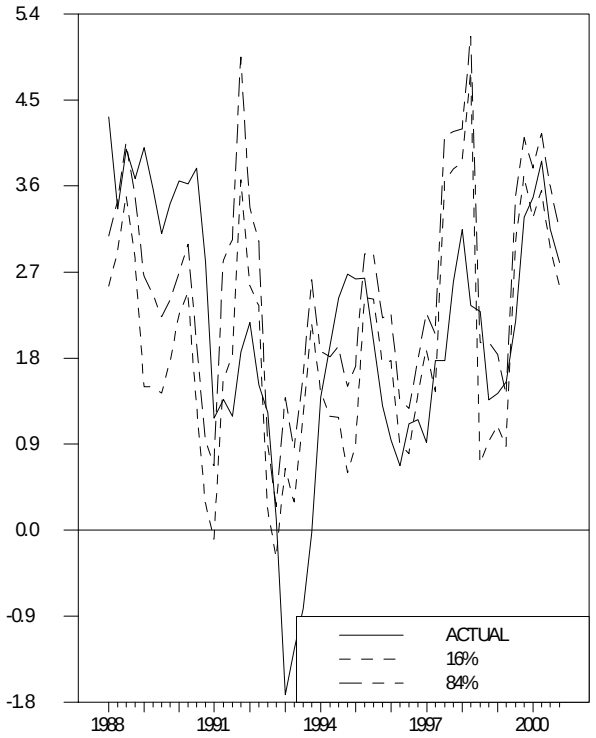
Example 11 *Use a heterogenous panel VAR model with dynamic interdependencies for G-7 countries with GDP growth, inflation, employment growth and the real exchange rate for each country. Specify: a* 2×1 *vector of common factors - (one EU and one non-EU), a* 7×1 *vector of country specific factors and a* 4×1 *vector of variables specific factors.*

Assume time variations in all factors, no exchangeable prior and non-informative priors on the hyperparameters. Calculate posterior distributions one year in advance constructed recursively at each t*. Figure plots leading indicator 68% bands for EU GDP growth and inflation (with actual values).*

Leading indicator = sum of the three estimated components. Model predicts the ups and downs of both series well using one year ahead info. Theil-U for 1996:1-2000:4 and 1991:1-1995:4 are 0.87 and 0.66, much lower than single country BVAR (0.96, 0.94) or univariate AR(0.98,0.96).



Inflation



GDP growth

4.3 Factor Models

$$y_{it} = \alpha_{0i} + \alpha_{1i} y_{0t} + u_{it} \quad (37)$$

$$\phi_i(L) u_{it} = v_{it} \quad (38)$$

$$\phi_0(L) y_{0t} = v_{0t} \quad (39)$$

Assume $E(v_{it}, v_{jt-s}) = 0,\ \forall i \neq j,\ i = 1, \ldots, M,\ E(v_{it}, v_{it-s}) = \sigma^2$ if $s = 0$ and zero otherwise, $E(v_{0t}, v_{0t-s}) = \sigma_0^2$ if $s = 0$ and zero otherwise.

- y_{0t} ($M_1 << M \times 1$) is unobservable and can have arbitrary serial correlation.

- u_{it} could be serially correlated; it could also be a VAR, i.e. $\phi(L) u_t = v_t$.

A factor model can be transformed into a state space model, where y_{0t} is the unobservable state using (38) into(37) (this becomes the measurement equation) and (39) is the state equation

$$\begin{aligned} \phi_i(L)y_{it} &= \phi_i(L)(\alpha_{0i} + \alpha_{1i}y_{0t}) + v_{it} && (40) \\ \phi_0(L)y_{0t} &= v_{0t} && (41) \end{aligned}$$

Example 12 *a)* y_{0t} *coincident business cycle indicator,* u_{it} *idiosyncratic fluctuations.*

b) y_{0t} *common stochastic trend,* u_{it} *is stationary part of* y_{it} *(Watson (1987)).*

c) y_{1t} *is a vector of stock returns,* y_{0t} *is a unobservable market portfolio (CAPM model).*

Example 13 *(a very special case)*

$$y_{it} = \alpha_{0i} + \alpha_{1i} y_{0t} + u_{it} \tag{42}$$

u_{it} *and* y_{0t} *are normal iid random variables, with variances* σ_i^2, σ_0^2 *and* $i = 1, \ldots, M$*. To apply the Gibbs sampler we need the conditional posteriors of (*y_{0t}*,* $\alpha_i = (\alpha_{0i}, \alpha_{1i})$ *and* σ_i^2*).*

The posteriors for $\alpha_i = (\alpha_{0i}, \alpha_{1i})$ *and* σ_i^2*, conditional on* y_{it}, y_{0t} *are standard Normal-inverted Gamma for any reasonable specification of the prior. The likelihood of* y_{0t} *for each* t *is Normal with mean zero and variance* σ_0^2 *(treat this as a prior). Hence, the conditional posterior for* $y_0 = (y_{01}, \ldots, y_{0t})$ *is Normal with mean* $\bar{y}_0 = \sigma_{y_0}^{-2} (\sum_{i=1}^{M} \frac{\alpha_{i1}(y_i - \alpha_{0i})}{\sigma_i^2})$ *and variance* $\sigma_{y_0}^2 = (\sum_{i=0}^{M} \frac{\alpha_{1i}^2}{\sigma_i^2})^{-1}$*.*

We complicate this (unrealistic) setup by allowing:

i) serial correlation in the factor.

ii) serial correlation in the error term (we know how to deal with this, see (41).

iii) do not condition on initial observations (i.e. we use full likelihood rather than conditional likelihood.

iv) impose stationarity: roots of $\phi_j(L), j = 0, i$ all less than one in absolute value.

Identifying Restrictions: $\alpha_{11} > 0$, σ_0^2 a fixed constant.

Let $\psi = (\alpha_{0i}, \alpha_{1i}, \sigma_i^2, \phi_{ij}),\ i = 1, \ldots, M; j = 1, \ldots p_i; h = 1, \ldots q;$

Let $y_i = (y_{i1}, \ldots, y_{it}),\ y = (y_1', \ldots, y_M')$.

Note: $g(\psi|y, y_0) \propto f(y|\psi, y_0)g(\psi)$ (posterior of the parameters)

$g(y_0|\psi, y) \propto f(y|\psi, y_0)f(y_0|\psi)$ (posterior of the latent factor).

So, given some $g(\psi)$, we need $f(y|\psi, y_0)$ and $f(y_0|\psi) = \int f(y, y_0|\psi)dy$ to apply the Gibbs sampler.

Road map:

1) Derivation of the distribution of the first p observations.

2) Derivation of the likelihood for p+T observations, transforming the model to take into account serial correlation of the shocks.

3) Set up priors for the parameters and find conditional posteriors.

4) Find the conditional posterior of y_0.

5) Run the Gibbs sampler.

1) Computation of $g(\psi|y, y_0)$.

Initial observations: $y_i^1 = (y_{i,1}, \ldots y_{i,p_i})'$; $y_0^1 = (y_{0,1}, \ldots y_{0,q})'$, y_0^1 given.

Let $\alpha_i = (\alpha_{0i}, \alpha_{1i})$, $\phi_i = (\phi_{i,1}, \ldots, \phi_{i,p_i})$, $\phi_i(L) = (\phi_{i,1}L, \ldots \phi_{i,p_i}L^{p_i})$, $x_i^1 = [1, y_0^1]$; $1 = [1, 1, \ldots, 1]'$ (a $p_i \times 1$ vector); $\Phi_i = \begin{bmatrix} \phi_1 & \cdots & \phi_{p_i} \\ I_{(p_i-1)} & & 0 \end{bmatrix}$, (a $(p_i \times p_i)$ matrix) .

If errors are normal, $(y_i^1|\alpha_i, \sigma_i^2, \Phi_i, y_0^1) \sim N(\alpha_{0i} + \alpha_{1i}y_0^1, \sigma_i^2\Sigma_i)$, where $\Sigma_i = (I \otimes \Phi_i'\Phi_i)(1, 0, \ldots 0)'(1, 0, \ldots 0)$.

2) Specification of the full model.

Let $\Sigma_i = P_i'P_i$ and set $y_{i1}^* = P_i^{-1}y_i^1$; $x_{i1}^* = P_i^{-1}x_i^1$. Let $u_i = [u_{i,p_i+1}, \ldots u_{i,T}]$ (this is $(T - p_i) \times 1$ vector); $U = [u_1, \ldots, u_p]$, $u_{it} = y_{it} - \alpha_{0i} - \alpha_{1i}y_{0t}$.

Let y_i^{2*} be a $T - p_i \times 1$ vector with the t-row equal to $\phi_i(L)y_{it}$ and let x_i^{2*} be a $(T - p_i \times 2)$ matrix with t-row equal to $(\phi_i(1), \phi_i(L)y_{0t})$. Let $x_i^* = [x_i^{1*}, x_i^{2*}]',\ y_i^* = [y_i^{1*}, y_i^{2*}]$.

The likelihood of $(y_i^*|x_i^*, \psi)$ normal.

3) Priors and conditional posteriors

Priors $g(\psi) = \prod_j g(\psi_j)$; $\alpha_i \sim N(\bar{\alpha}_i, \bar{\Sigma}_{\alpha_i})$; $\sigma_i^{-2} \sim G(\bar{a}_i, \bar{b}_i), i = 1, \ldots$ and $\phi_i \sim N(\bar{\phi}_i, \bar{\Sigma}_{\phi_i})\mathcal{I}_\phi, i = 0, 1, \ldots,$, where $\mathcal{I}_\phi$ is an indicator stationarity.

Conditional on y_{0t}, the conditional posteriors for the elements of ψ are the same as in the linear regression model with AR errors:

$$
\begin{aligned}
(\alpha_i|y_i,\ \psi_{-\alpha}) &\sim N(\tilde{\bar{\Sigma}}_{\alpha_i}^{-1}(\bar{\Sigma}_{\alpha_i}^{-1}\bar{\alpha}_i + \sigma_i^{-2}x_i^{*}y_i^{*}), \tilde{\bar{\Sigma}}_{\alpha_i}) \\
(\phi_i|y_i,\ y_0,\ \psi_{-\phi_i}) &\sim N(\tilde{\phi}_i, \tilde{\bar{\Sigma}}_{\phi_i})\mathcal{I}_\phi * \Upsilon(\phi_i) \\
(\sigma_i^{-2}|y_i,\ y_0,\ \psi_{-\sigma_i}) &\sim G(\bar{a}_i + T, \bar{b}_i + (y_i^{*} - x_i^{*}\alpha_i)^2) \qquad (43)
\end{aligned}
$$

where $\tilde{\bar{\Sigma}}_{\alpha_i} = (\bar{\Sigma}_{\alpha_i}^{-1} + \sigma_i^{-2}x_i^{*\prime}x_i^{*})^{-1}$; $\tilde{\phi}_i = \tilde{\bar{\Sigma}}_{\phi_i}^{-1}(\bar{\Sigma}_{\alpha_i}^{-1}\bar{\phi}_i + \sigma^{-2}U_i^{\prime}u_i)$, $\Upsilon(\phi_i) = |\Sigma_i(\phi_i)|^{-0.5}\exp\{-\frac{1}{2\sigma^2}(y_{i1}-x_{i1}\alpha_i)^{\prime}\Sigma_i(\phi_i)^{-1}(y_{i1}-x_{i1}\alpha_i)\}$, $\tilde{\bar{\Sigma}}_{\phi_i} = (\bar{\Sigma}_i^{-1} + \sigma^{-2}U_i^{\prime}U_i)^{-1}$.

Sampling α_i, σ_i^2 from (43) straightforward (discard the draws $\alpha_{11} \leq 0$).

Sampling ϕ_i is more complicated: indicator for stationarity plus initial p_i observations (otherwise straightforward). Use a MH step within Gibbs:

Algorithm 4.1 (1.) *Draw* ϕ_i^+ *from* $N(\tilde{\phi}_i, \tilde{\boldsymbol{\Sigma}}_{\phi_i})$. *If* $\sum \phi_i \geq 1$ *discard it*

[2.] If draw is not discarded, set draw $\zeta \sim U(0,1)$.

[3.] If $\zeta < \Upsilon(\phi_i^l)/\Upsilon(\phi_i^{l-1})$, *set* $\phi_i^l = \phi_i^+$. *Else set* $\phi_i^l = \phi_i^{l-1}$.

4) Computation of $g(y_0|\psi, y)$. Let $\mathcal{G}_i^{-1} = \begin{bmatrix} \hat{P}_i^{-1} & 0 \\ \hat{R}_i & \end{bmatrix}$) be a $T \times T$ matrix;

and $\hat{R}_i = \begin{bmatrix} -\phi_{i,p_i} & \dots & -\phi_{i,1} & 1 & 0 & \dots & 0 \\ 0 & -\phi_{i,p_i} & \dots & -\phi_{i,1} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\phi_{i,p_i} & \dots & \dots & 1 \end{bmatrix}$), where $\hat{P}_i$ is a $p_i \times p_i$ matrix.

Transform the model by letting $x_i^\dagger = \mathcal{G}_i^{-1} x_i$; $y_i^\dagger = \mathcal{G}_i^{-1} y_i$; $y_i^+ = y_i^\dagger - \mathcal{G}_i^{-1} \mathbf{1a_i}$; $\mathbf{1}=[1, \dots, 1]'$.

The likelihood function is $\prod_{i=1}^{M} f(y_i^+|\alpha_i, \sigma_i^2, \phi_i, y_0)$ where $f(y_i^+|\alpha_i, \sigma_i^2, \phi_i, y_0) = (2\pi\sigma_i^2)^{-0.5T} \exp\{-(2\sigma_i^2)^{-1}(y_i^+ - \alpha_i \mathcal{G}_i^{-1} y_0)'(y_i^+ - \alpha_i \mathcal{G}_i^{-1} y_0)\}$.

The likelihood of y_0 is $f(y_0|\phi_0) = (2\pi\sigma_0^2)^{-0.5T} \exp\{-(2\sigma_0^2)^{-1}(\mathcal{G}_0^{-1} y_0)'(\mathcal{G}_0^{-1} y_0)\}$.

The joint likelihood is $f(y^+, y_0|\psi) = \prod_{i=1}^{M} f(y_i^+|\alpha_i, \sigma_i^2, \phi_i, y_0) f(y_0|\phi_0)$.

Hence, completing the squares, the conditional posterior is:

$$g(y_0|y^+, \psi) \sim N(\tilde{y}_0, \tilde{\Sigma}_{y_0}) \tag{44}$$

$\tilde{y}_0 = \tilde{\Sigma}_{y_0}^{-1}[\sum_{i=1}^{M} \frac{\alpha_{1i}}{\sigma_i^2} \mathcal{G}_i^{-1'} \mathcal{G}_i^{-1}(y_i - \alpha_{0i}\mathbf{1})]$; $\tilde{\Sigma}_{y_0} = (\sum_{i=0}^{M} \frac{\alpha_{1i}^2}{\sigma_i^2} (\mathcal{G}_i^{-1})' \mathcal{G}_i^{-1})^{-1}$; $\alpha_{10} = 1$.

5) Run the Gibbs sampler

- Use $g(\psi|y, y_0),\ g(y_0|\psi, y)$ into a Gibbs sampler after setting σ_0^2 and the parameters of the prior distributions.

- To compute the predictive density of y_{0t}: draw ϕ_0 from posterior. Construct $y_{0t+\tau}$ taking y_{0t} as given and drawing u_{0t} from $N(0, \sigma_0^2)$.

4.3.1 A special factor model: a APT Model

$$r_t = \alpha_0 + \alpha_1 f_t + u_t \tag{45}$$

r_t is $N \times 1$ vector; f_t a $K \times 1$ vector of "pervasive" factors, $E(f) = 0, E(ff') = I$, $E(u|f) = 0, E(uu'|f) = \Sigma$; α_0 = conditional mean, α_1 vector of loadings; α_1 and f_t unknown.

- How do you estimate such a model? Traditionally two step estimation; first step: get the factor(s) with a cross sectional regression. Second, taking the estimated factor(s) as if they were true do a regression in time series to estimate (see e.g. Roll and Ross (1980)). Problem: error-in-variables unless cross section is very large.

- Ross (1976): as $N \to \infty$, absence of arbitrage opportunities implies $\alpha_{0i} \approx \lambda_0 + \sum_k \alpha_{1ik}\lambda_k$, ($\lambda_0$ is the intercept of the pricing relationship (the

so-called zero-beta rate) and λ_k is the risk premium on factor k, both unknown). In two step approach restrictions become linear. Easy to test (see Campbell, Lo and McKinley (1997)).

- Alternative test for the absence of arbitrage opportunities: check the size of the pricing errors relative to the average returns (large relative errors= inappropriate specifications) i.e use $\mathcal{Q}^2 = \frac{1}{N}\alpha_0'[I - \alpha_1^*(\alpha_1^{*\prime}\alpha_1^*)^{-1}\alpha_1^{*\prime}]\alpha_0$ where $\alpha_1^* = (\mathbf{1}, \alpha_1)$, $\mathbf{1} = (1, 1, \ldots 1)$. For fixed N, $\mathcal{Q}^2 \neq 0$; as $N \to \infty$, $Q^2 \to 0$. The sampling distribution of Q^2 is hard to compute.

- We can get the exact small sample posterior for this statistic.

- For identification we need: $K < \frac{N}{2}$ and α_1^k is lower triangular with $\alpha_{1ii} > 0$. $\alpha_1^k =$ matrix of α_1 containing the first k independent rows.

- We need $g(\psi|f_t, r_t)$; $g(f_t|\psi, r_t)$, where $\psi = (\alpha_0, \alpha_1, \Sigma = \text{diag}\{\sigma_i^2\})$.

Assume independence across i and

i) $\alpha_{1i} \sim N(\bar{\alpha}_{1i}, \zeta_1 I),\ \alpha_{1ii} > 0, i = 1, \ldots, K$;

ii) $\alpha_{1i} \sim N(\bar{\alpha}_{1i}, \zeta_2 I), i = K+1, \ldots, N$;

iii) $\bar{\nu}_i \frac{\bar{R}_i^2}{\sigma_i^2} \sim \chi^2(\bar{\nu}_i)$;

iv) $\alpha_{0i} \sim N(\bar{\alpha}_{0i}, \bar{\sigma}_{0i}^2)$ where $\bar{\alpha}_{0i} = \lambda_0 + \sum_k \lambda_k \alpha_{1ik}$, λ_i are constant and $\bar{\alpha}_{1i}, \zeta_1, \zeta_2, \bar{\nu}_i, \bar{s}_i^2, \bar{\sigma}_{0i}^2$ are given.

The conditional posteriors are:

•$g(\alpha_{0i}|r_t, f_t, \alpha_{1i}, \sigma_i) \sim N(\tilde{\alpha}_{0i}, \tilde{\sigma}^2_{0i})$; $\tilde{\alpha}_{0i} = \frac{(\bar{\sigma}^2_{0i}\hat{\alpha}_{0i}+\bar{\alpha}_{0i}(\sigma^2_i/T))}{(\sigma^2_i/T)+\bar{\sigma}^2_{0i}}$; $\tilde{\sigma}^2_{0i} = \frac{(\sigma^2_i\bar{\sigma}^2_{0i})/T}{\sigma^2_i/T+\bar{\sigma}^2_{0i}}$, $\hat{\alpha}_{0i} = \sum_t(r_{it} - \alpha_i f_t)$.

•$g(\alpha_{1i}|r_t, f_t, \alpha_{01}, \sigma_i) \sim N(\tilde{\alpha}_{1i}, \tilde{\tilde{\Sigma}}_{\alpha_{1i}})$, $\tilde{\tilde{\Sigma}}_{\alpha_{1i}} = (\zeta_1^{-1}I+\sigma_i^{-2}f_i^{*\prime}f_i^*)^{-1}$; $\tilde{\alpha}_{1i} = \Sigma^{-1}_{\alpha_{1i}}(\bar{\alpha}_{1i}\zeta_1^{-1}+f_i^{*\prime}f_i^*\hat{\alpha}^*_{1i}\sigma_i^{-2})$ for $i = 1, \ldots, k$ and $\tilde{\tilde{\Sigma}}_{\alpha_{1i}} = (\zeta_2^{-1}I+\sigma_i^{-2}f_i^{*\prime}f_i^*)^{-1}$; $\tilde{\alpha}_{1i}$ $\Sigma^{-1}_{\alpha_{1i}}(\bar{\alpha}_{1i}\zeta_2^{-1} + f_i^{*\prime}f_i^*\hat{\alpha}^*_{1i}\sigma_i^{-2})$ for $i = k+1, \ldots, N$ where $f_i^* = f$ matrix minus first column, $\hat{\alpha}^*_{1i}$ is OLS estimator of a regression of $r_i - \alpha_{0i}$ on the factors.

•$g(\tilde{\nu}\frac{\tilde{s}^2}{\sigma^2_i}|r_t, f_t, \alpha_i) \sim \chi^2(\tilde{\nu})$ where $\tilde{\nu} = \bar{\nu} + T$; $\tilde{s}^2_i = \frac{\bar{\nu}\bar{R}^2_i+TR^2_i}{\tilde{\nu}}$ and $R^2_i = \frac{1}{T}\sum_t(r_{it} - \alpha_{0i} - f_t\alpha_{1i})'(r_{it} - \alpha_{0i} - f_t\alpha_{1i})$.

- Joint distribution of the data and the factor is

$$\begin{bmatrix} f_t \\ r_t \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ \alpha_0 \end{pmatrix}, \begin{pmatrix} I & \alpha_1' \\ \alpha_1 & \alpha_1\alpha_1' + \Sigma \end{pmatrix}\right]$$

- Using the properties of conditional normals:

$$g(f_t|r_t, \psi) \sim N(\alpha_1'(\alpha_1'\alpha_1 + \Sigma)^{-1}(r_t - \alpha_0); I - \alpha_1'(\alpha_1'\alpha_1 + \Sigma)^{-1}\alpha_1).$$

• Put these four conditionals into the Gibbs sampler. Compute $\mathcal{Q}^2$ at every draw. Average over draws.

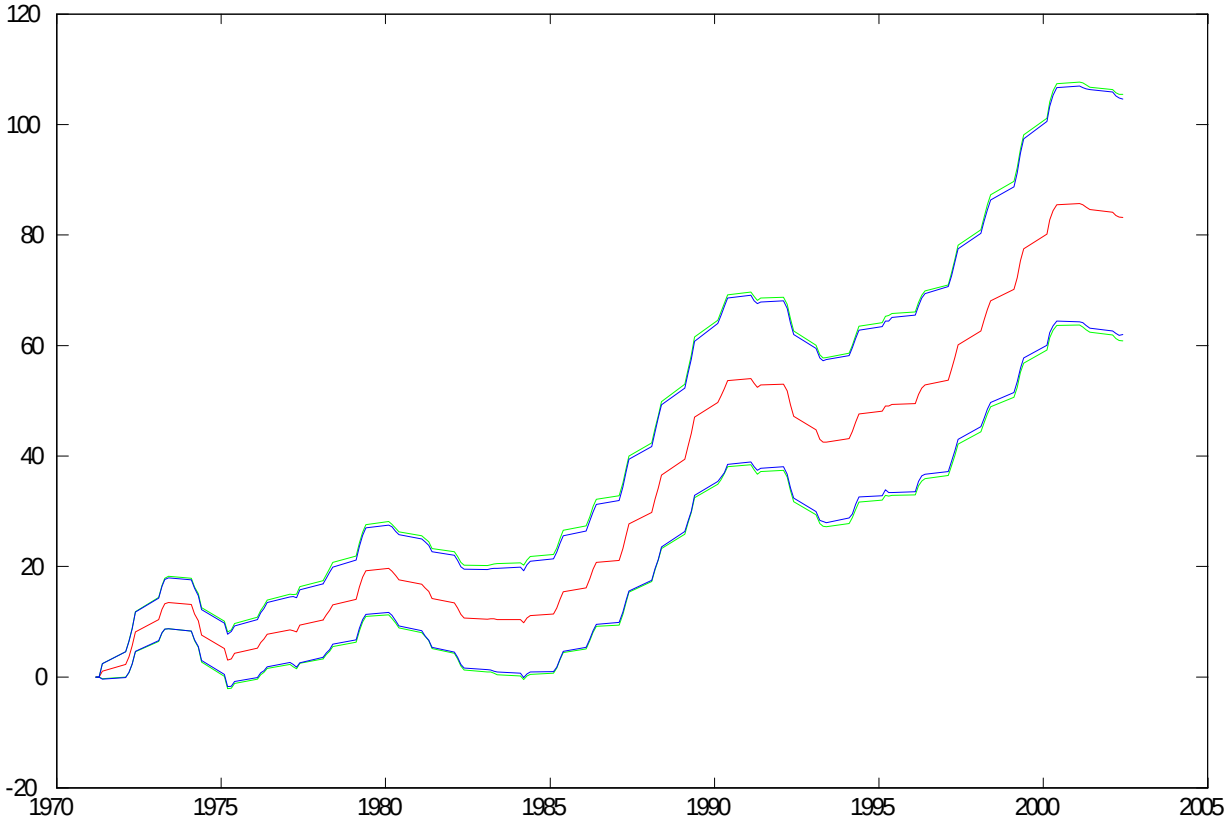
Example 14 *We construct a coincident indicator for the EU business cycle using quarterly data on real goverment consumption, real private investment, real employment and real GDP from 1970:1 to 2002:4.*

We allow an AR(2) structure on the unknown indicator y_{0t} *and and AR(1) structure on the errors of the model.*

Posterior estimates obtained with 10000 draws from the conditionals: 5000 are discarded, one every 5 is used to construct the indicator.

The posterior mean of the two AR oefficients of the factors are 0.711 and 0.025 respectively, the posterior standard errors are 0.177 and 0.134.

The coicident indicator seem to be doing its job: over the sample it shows (classical) recessions at roughly speaking the same dates the CEPR selected. Furthermore, it indicates that after 2001 a recession took place.



EU Coincident indicator

4.4 Stochastic Volatility Models

- Stochastic volatility models are alternatives to GARCH models to capture time variations in second moments.

- Produce time varying volatility and leptokurthosis in the endogenous y_t.

- They imply that observables driven by two independent shocks.

- Produce a non-linear state space model.

Simplest SV model:

$$y_t = h_t^{0.5} u_t \qquad u_t \sim N(0,1) \tag{46}$$

$$\log h_t = \rho_0 + \rho_1 \log h_{t-1} + \sigma v_t \qquad v_t \sim iidN(0,1) \tag{47}$$

where v_t and u_t independent.

Variance of y_t is changing over time. No level dynamics in y_t.

Let $h^t = (h_1, \ldots h_t)$; $\alpha = (\rho_0, \rho_1)$

The data density is $f(y^t|\alpha, \sigma) = \int f(y^t|h^t) f(h^t|\alpha, \sigma) dh^t$.

As in TVC models we treat h^t as a parameter whose conditional posterior needs to found. Once it is found, we use it into the Gibbs sampler, together with the conditional of α.

How do you compute the conditional posterior of h^t?

i) The joint of h^t is the product of conditionals of the form $g(h_t|h_{t-1}, h_{t+1}, \alpha, \sigma, y_t)$ and of a marginal.

ii) The conditional densities in i) are mixture of a normal and a log-normal.

$$\begin{aligned} g(h_t|h_{t-1}, h_{t+1}, \alpha, \sigma, y_t) &\propto f(y_t|h_t) f(h_t|h_{t-1}, \alpha, \sigma) f(h_{t+1}|h_t, \alpha, \sigma) \\ &\propto \frac{1}{h_t^{0.5}} \exp\{-\frac{y_t^2}{2h_t}\} \frac{1}{h_t} \times \exp\{-\frac{(\ln h_t - \mu_t)^2}{2\sigma_h^2}\} \end{aligned} \tag{48}$$

where $\mu_t = \frac{(\rho_0(1-\rho_1)+\rho_1(\ln h_{t+1}+\ln h_{t-1})}{1+\rho_1^2}$, $\sigma_h^2 = \frac{\sigma^2}{1+\rho^2}$.

How to draw form such a conditional?

• Approximate the log normal portion with inverted gamma Since the first portion can also be approximated with an inverted gamma, approximate the whole kernel with one $IG(a, b)$; where $a = \frac{(1-2\exp(\sigma^2))}{1-\exp(\sigma^2)} + 0.5$ and $b = [(a-1)(\exp(\mu + 0.5\sigma^2)) + 0.5y_t^2]^{-1}$.

Example 15 *We set* $\rho_0 = 0.0, \rho_1 = 0.8$ *and* $\sigma = 1.0$.

Table 1: Percentiles of the approximating distributions

	5th	*25th*	*median*	*75th*	*95*
Gamma	*0.11*	*0.70*	*1.55*	*3.27*	*5.05*
Normal	*0.12*	*0.73*	*1.60*	*3.33*	*5.13*

Alternatives:

• (47) is a particular nonlinear Gaussian model. To make life easier (i.e. to use the KF to evaluate the likelihood), we can transform it into a linear non-Gaussian state space model.

Set $z_t = \log y_t^2 + 1.27,\ x_t = \log h_t;\ e_t = \log u_t^2 + 1.27$, then (47) is

$$\begin{aligned} z_t &= x_t + e_t \\ x_{t+1} &= \rho_0 + \rho_1 x_t + \sigma v_t \end{aligned} \tag{49}$$

where e_t is non-normal.

- To approximate a non-normal distribution for e_t we can use a mixture of normals, i.e. $f(e_t) = \sum_j \pi_j f(e_t|j)$ where each $f(e_t|j) \sim N(\mu_j, \sigma_j^2),\ j = 1, \ldots J$ (see e.g. Chib (1996)). Conditional on $\log h_t$ the model is now linear and Gaussian.

• The kernel of $\log h_t$ is $-0.5y_t^2(\exp\{-\log h_t\}) - \frac{(\ln h_t - \mu_t)^2}{2\sigma_h^2}$. Then draw h_t from $N(\mu_t^*, \sigma^2)$ where $\mu_t^* = \mu_t - 0.5\sigma$ and accept the draw with probability $\exp\{-0.5\frac{y_t^2}{h_t}\}$ (Metropolis step) (here no approximation is taken).

- Can add regression terms to (46). Nothing changes. We can derive the posterior on these new parameters conditional on (h^t, α) and put all of them in the Gibbs sampler.

- We will show how to do this in the context of a TVC-BVAR.

4.4.1 TVC VAR with stochastic volatility

$$\begin{aligned} y_t &= (I \otimes X_t)\alpha_t + u_t \qquad u_t \sim N(0, \Sigma_t) \\ \Sigma_t &= B^{-1} H_t B^{-1'} \\ \alpha_t &= \mathcal{G}\alpha_{t-1} + v_t \qquad v_t \sim N(0, \Omega) \end{aligned} \tag{50}$$

where B is a lower triangular matrix with one on the main diagonal, $H_t = \text{diag}\{h_{it}\}$ and $\ln h_{it} = \ln h_{it-1} + \sigma_i \epsilon_{it}$; $\mathcal{G}$ such that α_t is a stationary.

New parameters relative to a standard TVC-VAR: B and h_t.

How do you conduct Bayesian estimation of this model? Construct the conditional posteriors for $(\Omega, \sigma_i^2, \alpha^t, B, h^t)$ jointly.

Priors:

-$g(\Omega) \sim iW(\bar{\Omega}^{-1}, \bar{\nu})$; where $\bar{\Omega} = \gamma * \bar{\Sigma}_a$; , $\bar{\nu} = dim(\alpha_0) + 1$.

-$g(\sigma_i^2) \sim IG(\bar{a}, \bar{b})$.

-$\alpha_0 \sim N(\bar{\alpha}, \bar{\Sigma}_a)$.

-$\beta \sim N(\bar{\beta}, \bar{\Sigma}_\beta)$, where β are the non zero elements of B.

-$g(\ln h_{i0}) \sim N(\ln \bar{h}_i, \bar{\Sigma}_h)$.

Conditional posteriors:

- Ω is $IW(\bar{\Omega}^{-1} + (\sum_t v_t v_t')^{-1}, \bar{\nu} + T)$.

- σ_i^2 is $IG(\bar{a} + T, \bar{b} + \sum_t (\ln h_{it} - \ln h_{it-1})^2)$.

- α_t is normal, each t. Use the Kalman Filter as in the standard TVC model to compute the moments.

- For the conditional posterior of β notice that $u_t = Be_t$. If $e_t \sim (0, H_t)$ and if u_t is known, B can be estimated from equations like

$$h_{mt}^{-0.5} u_{mt} = \beta_{m1}(-h_{mt}^{-0.5} u_{1t}) + \ldots + \beta_{m,m-1}(-h_{mt}^{-0.5} u_{m-1t}) + (h_{mt}^{-0.5} e_{mt}) \quad (51)$$

Let $Z_{mt} = (-h_{mt}^{-0.5} u_{1t}, \ldots, -h_{mt}^{-0.5} u_{mt})$ and $z_{mt} = -h_{mt}^{-0.5} e_{mt}$. Then:

$\beta_i \sim N(\tilde{\beta}_i, \tilde{B}_i)$ with $\tilde{\beta} = \tilde{B}_i^{-1}(Z_i' z_i + \bar{B}^{-1}\bar{\beta}), \tilde{B}_i = (\bar{B}_i + (Z_i Z_i')^{-1})$.

- Let h_{-it} be the sequence of h^t except its i-th element; and let $u^t = (u_1, \ldots u_t)$.

$g(h_{it}|h_{(-i)t}, \sigma_i, u^t) = g(h_{it}|h_{it-1}, h_{it+1}, u^t, \sigma_i)$ and

$g(h_{it}|h_{it-1}, h_{it+1}, u^t, \sigma_i) \propto h_{it}^{-1.5} \exp\{-\frac{u_{it}^2}{2h_{it}} \frac{(-\ln h_{it} - \mu_{it})^2}{2\sigma_c}\}$; where $\sigma_c = 0.5\sigma_i$, $\mu_{it} = 0.5(\ln h_{it+1} + \ln h_{t-1})$.

To draw from this conditional posterior we need to use MH step within the Gibbs sampler . That is, choose as candidate density $h_{it}^{-1} \exp\{\frac{(-\ln h_{it} - \mu_{it})^2}{2\sigma_c}\}$ and accept the draw with probability $\frac{(h_{it}^{+})^{-0.5} \exp\{-\frac{u_{it}^2}{2h_{it}^{+}}\}}{(h_{it}^{\ell-1})^{-0.5} \exp\{-\frac{u_{it}^2}{2h_{it}^{\ell-1}}\}}$.

- The predictive distribution of future y_t' can be computed using:

$$\begin{aligned} g(y^{t+\tau}|\alpha^t, H^t, \Omega, \beta, \sigma, y^t) &= g(\alpha^{t+\tau}|\alpha^t, H^t, \Omega, \beta, \sigma, y^t) \\ &\times \; g(H^{t+\tau}|\alpha^{t+\tau}, H^t, \Omega, \beta, \sigma, y^t) \\ &\times \; f(y^{t+\tau}|\alpha^{t+\tau}, H^{t+\tau}, \Omega, \beta, \sigma, y^t) \end{aligned}$$

- A point estimate of the conditional volatility (needed for example in option pricing formulas) can be obtained using the smoothed density $g(h_t|y^t)$. This is obtained using joint draws from $g(h_t, \alpha_t|y^t)$ [since $g(h_t|y^t) = \int g(h_t, |\alpha_t, y^t) g(\alpha_t|y^t) d\alpha_t$] and a non-parametric kernel.

Example 16 *(Canova and Gambetti (2009)). Use a TVC-VAR with stochastic volatility to compute the dynamics of the volatility of the monetary policy shock over the sample 1965-2005. The volatility of this shocks has been far from constant and was very high at the beginning of the 1980s.*

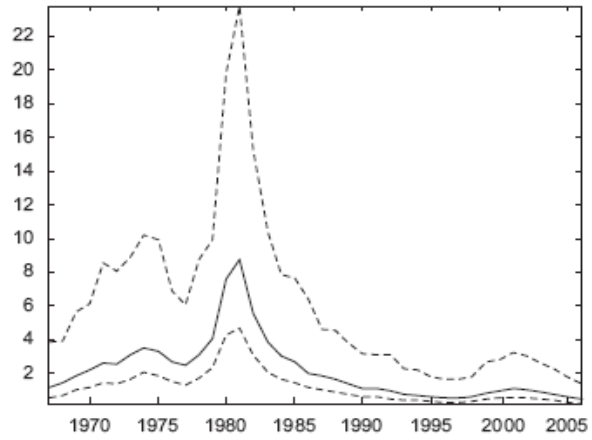


Fig. 3. Posterior standard deviation of the monetary policy shock; solid line median, dotted lines 68 percent interval.

Example 17 *(Bayesian-EGARCH model)*

$$y_t = x_t'\alpha + h_t^{0.5}u_t, \qquad u_t \sim N(0,1) \tag{52}$$

$$h_t = \exp\{\delta_0 + \delta_2 u_{t-1}^2 + \delta_1 h_{t-1}\}. \tag{53}$$

The likelihood function of this model is: $\prod_t h_t^{-0.5} \exp\{-0.5(y_t - x_t'\alpha)^2/h_t\}$.

Priors: $\alpha \sim N(\bar{\alpha}, \bar{\sigma}_\alpha^2)$; $\delta_0 \sim N(\bar{\delta_0}, \bar{\sigma}_{\delta_0}^2)$. $g(\delta_1,\ \delta_2)$ *is diffuse (uniform) over [0,1] and restricted so that* $\delta_1 + \delta_2 \leq 1$.

Let $I(.)$ *be the kernel of a t-distribution with location* $\tilde{\psi}$*, (the mode of the posterior), scale* $\tilde{H}_\psi^{-1}$ *(the hessian at the mode) and degrees of freedom* $\bar{\nu}$.

Use an independent Metropolis algorithm to draw from the posteriors: i.e. draw ψ^+ *from* $I(.)$ *and accept the draw with probability equal to* $\min[\frac{\tilde{g}(\psi^+|y_t)/I(\psi^+)}{\tilde{g}(\psi^{l-1}|y_t)/I(\psi^{l-1})}, 1]$.

5 Markov switching models

• Markov switching models have an unobservable state. This can be treated as "missing" data and "generated" with the Gibbs sampler.

$$y_t = x_{1t}\alpha_1 + x_{2t}\alpha_2(\mathcal{S}_t - 1) + u_t \quad u_t \sim N(0, \sigma^2) \tag{54}$$

$\mathcal{S}_t$ is a two states Markov switching indicator ($\mathcal{S}_t = 0$ or $\mathcal{S}_t = 1$). $\mathcal{S}_t = 1$ is normal state (so that for $\mathcal{S}_t = 0$ $y_t = x_{1t}\alpha_1 - x_{2t}\alpha_2 + u_t$).

We want the conditional posteriors of $(\alpha_1,\ \alpha_2,\ \sigma^2,\ \mathcal{S}_t)$.

Let $\eta_{11} = p(S_t = 1|S_{t-1} = 1)$; $\eta_{00} = p(S_t = 0|S_{t-1} = 0), \eta_{10} = 1 - \eta_{00}, \eta_{01} = 1 - \eta_{11}$ where η_{ij} is unknown.

Let $\theta = (\alpha_1, \alpha_2, \sigma^2)$; $y^{t-1} = (y_1, \ldots, y_{t-1}, x_{i1}, \ldots x_{it-1},\ i = 1, 2)$, $S^t = (S_1, \ldots, S_t)$ and $\psi = (\theta, S^t, \eta_{ij})$.

Priors:

- $g(\psi) = g(\theta)g(S^t|\eta_{ij})g(\eta_{ij})$ where

- $g(\theta) \propto N(\bar{\alpha}_1, \bar{\Sigma}_1)\ N(\bar{\alpha}_2, \bar{\Sigma}_2)\ IG(\bar{a}, \bar{b})$.

- $g(S^t|\eta_{ij}) = \eta_{00}^{d_{00}}\eta_{01}^{d_{01}}\eta_{10}^{d_{10}}\eta_{11}^{d_{11}}$, where d_{ij} is the a-priori proportions of i, j elements.

- $g(\eta_{i1}, \eta_{i0}) \propto (\eta_{i1}^{f_{i1}})(\eta_{i0}^{f_{i0}})$, i.e $\eta_{i.} \sim Beta(f_{i1}, f_{i0}),\ f_{ij} \geq 1$.

Conditional posteriors:

$\tilde{g}(\psi|y) = \sum_{t=1}^{T} f(y_t|\psi, y^{t-1})g(\psi)$ where $f(y_t|\psi, y^{t-1}) \sim N(\alpha x_t, \sigma^2)$.

Thus, given ψ_0 and S_0, we can sample the parameters of interest using

Algorithm 5.1 *[1.] Sample* α_i *from a normal with variance* $\tilde{\bar{\Sigma}}_a = \sum_t \frac{x_t' x_t}{\sigma^2} + \bar{\Sigma}^{-1})^{-1}$, $x_t = (x_{1t}, x_{2t}), \bar{\Sigma} = \text{diag}(\bar{\Sigma}_1, \bar{\Sigma}_2)$ *and mean* $\tilde{\alpha} = \tilde{\bar{\Sigma}}_a^{-1}(\sum_t \frac{x_t y_t}{\sigma^2} + \bar{\Sigma}^{-1}\bar{\alpha})$; $\bar{\alpha} = (\bar{\alpha}_1, \bar{\alpha}_2)$).

[2.] Sample σ^2 *from an IG with parameters* $a_T = \bar{a} + (T-1)/2$ *and* $b_T = \bar{b} + 0.5 \sum_t (y_t - \alpha_1 2x_{1t} + \alpha_2 x_{2t}(\mathcal{S}_t - 1))^2$.

[3.] Sample η_{ij} *from* $\tilde{g}(\eta_{ij}|S^T, y, \theta) \propto (\eta_{i1}^{\tilde{f}_{i1}} \eta_{i0}^{\tilde{f}_{i0}})$; $\tilde{f}_{ij} = f_{ij} + d_{ij}, i, j = 1, 2$.

[4.] To sample $\mathcal{S}^t$ from $\tilde{g}(\mathcal{S}^t|y,\theta,\eta_{ij})$ we need the following filter algorithm:

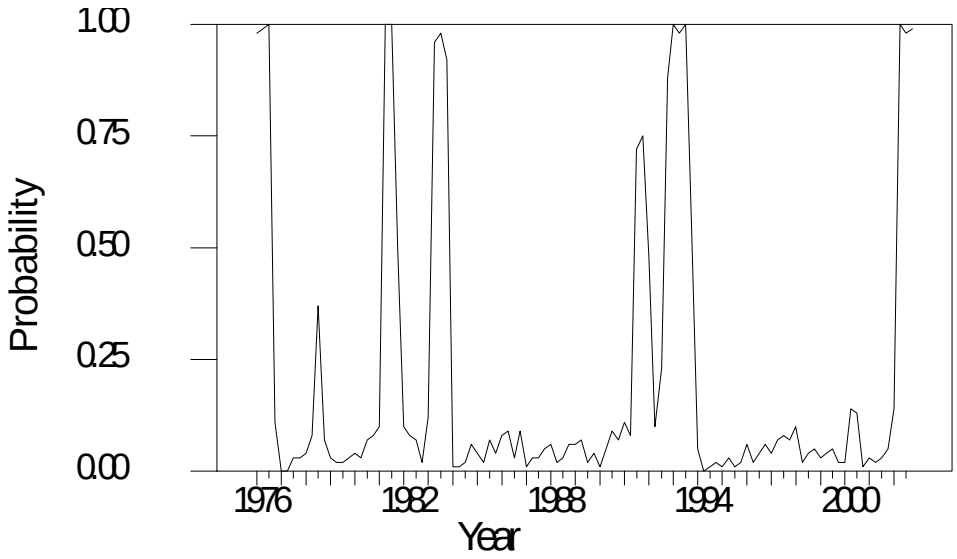
- given $g(\mathcal{S}_0)$ run forward using $g(\mathcal{S}_t|\theta,y^t,\eta_{ij}) \propto f(y_t|y^{t-1},\theta,\mathcal{S}_t)g(\mathcal{S}_t|\theta,y^{t-1}\eta_{ij})$, $f(y_t|y^{t-1},\theta,\mathcal{S}_t) \sim N(\alpha x_t,\sigma^2)$ where

$g(\mathcal{S}_t|\theta,y^{t-1},\eta_{ij}) = \sum_{\mathcal{S}_{t-1}=0}^{1} g(\mathcal{S}_{t-1}|\theta,y^{t-1}\eta_{ij})\eta_{(t_i,(t-1)_j)}; \eta_{(t_i,(t-1)_j)} = P(\mathcal{S}_t = i|\mathcal{S}_{t-1} = j)$.

- given $g(\mathcal{S}_t|y_t)$ run backward i.e. compute $g(\mathcal{S}_\tau|\mathcal{S}_{\tau+1},y^\tau,\theta,\eta_{ij}) \propto g(\mathcal{S}_\tau|\theta,y^\tau\eta_{ij})\eta_{(\tau_i,(\tau+1)_j)}$.

The approach is similar to the one used in state space model. In fact a two state Markov chain model is equivalent to an AR(1) model with AR coefficient $\eta_{00}+\eta_{11}-1$. Careful: the innovations of this model are binary and not normal.

Example 18 *Use (54) to the study IP fluctuations in EU (Germany, France and Italy aggregated with GDP weights).* $y_t =$ *yearly change in IP, data 1974.1-2001:4. The conditional mean estimates are* $\alpha_1 = 0.46$, $\alpha_2 = 0.96$, *the standard deviations 0.09 and 0.09 respectively. Thus, the growth rate in expansions two percent higher. Recession probabilities are below.*



Recession probabilities

5.1 A More complicated structure

$$A(\ell)(y_t - \mu(\mathcal{S}_t, x_t)) = \sigma(\mathcal{S}_t)^{0.5} u_t$$

all the roots of the $A(\ell)$ polynomial are outside the unit circle, and $\mathcal{S}_t$ is a two-state Markov chain with transition matrix η_{ij}. Here, $\sigma(\mathcal{S}_t) = \sigma^2 - \zeta^2 \mathcal{S}_t$; $\mu(\mathcal{S}_t, x_t) = x_t\beta_0 - \beta_1 \mathcal{S}_t$ with $\zeta^2 > 0,\ \beta_1 > 0$.

Let $y^t = (y_1, \ldots y_t), \mathcal{S}^t = (\mathcal{S}_1, \ldots, \mathcal{S}_t),\ \alpha = (A_1, \ldots A_p)$; set $\omega = \frac{\zeta^2}{\sigma^2}$ and let $\psi = (\beta_0, \beta_1, \alpha, \sigma^2, \omega, \eta_{ij})$.

The likelihood function is $f(y^t|S^t, \psi) = f(Y^p|S^p, \psi) \prod_{\tau=p+1}^{t} f(y_\tau|y^{\tau-1}, S^{t-1}, \psi)$.

The density of the first p observations is $N(x^p\beta_0 + S^p\beta_1, \sigma^2\Omega_P)$ where $\Omega_p = W_p \Sigma_p W_p$, $\Sigma_p = A^\dagger \Sigma_p A^{\dagger\prime} + e_t e_t'$, $W_p = \text{diag}\{(1 + \omega S_j)^{0,5},\ j = 1, \ldots, p\}$ $A^\dagger$ is the companion matrix of $A(\ell)$, $e_t = (1, 0, \ldots, 0)'$, a $p \times 1$ vector.

Using the prediction error decomposition of the likelihood we can write: $f(y_\tau|y^{\tau-1}, \mathcal{S}^{\tau-1}, \psi) \propto \exp\{-0.5\sigma(\mathcal{S}_\tau)^{-1}(y_\tau - y_{\tau|\tau-1})^2\}$; where $y_{\tau|\tau-1} = (1-A(\ell))y_t + A(\ell)(x_\tau\beta_0 + \beta_1\mathcal{S}_\tau)$. This implies that $y_t \sim N(y_{t|t-1}, \sigma(S_t))$.

The joint density is $f(y^t, S^t|\psi) = f(y^t|S^t, \psi)\prod_{\tau=2}^{t} f(S_\tau|S_{\tau-1})f(S_1)$

The likelihood of the data is $\int f(y^t S^t|\psi)dS^t$.

Hamilton (1994): get $\hat{\psi}_{ML}$, compute S^t conditional on $\hat{\psi}_{ML}$ (i.e. no uncertainty in $\hat{\psi}$ considered). That is, given $P(S_{\tau-1}, \ldots, S_{\tau-r}|y^\tau, \psi)$, compute

$$\begin{aligned} P(S_\tau, \ldots, S_{\tau-r+1}|y^\tau, \psi) &= \sum_{S_{\tau-r}=0}^{1} P(S_\tau, \ldots, S_{\tau-r}|y^{\tau-1}, \psi) \\ &\propto P(S_\tau|S_{\tau-1})P(S_{\tau-1}, \ldots S_{\tau-r}|y^{\tau-1}, \psi)f(y_\tau|y^{\tau-1}, S^\tau, \psi) \end{aligned} \tag{55}$$

where the proportionality factor is $f(y_\tau|y^{\tau-1},\psi) = \sum_{S_\tau},\ldots\sum_{S_{\tau-r}} f(y_\tau, S_\tau,\ldots S_{\tau-r}|y^{\tau-1},\psi)$. Since log $f(y_{p+1},\ldots y_t|y^p,\psi) = \sum_\tau$ log $f(y_\tau|y^{\tau-1},\theta)$, we can find the transition probabilities using $P(S_t|Y^t,\hat{\psi}_{ML})$.

How do you take into consideration uncertainty in ψ? Priors:

- $g(\beta,\sigma^2) \propto N(\bar{\beta}_0,\bar{\Sigma}_{\beta_0}^{-1})\ N(\bar{\beta}_1,\bar{\Sigma}_{\beta_1}^{-1})\mathcal{I}_{(\beta_1>0)}\ IG(\bar{a}_\sigma,\bar{b}_\sigma)$ where $\mathcal{I}_{(\beta_1>0)}$ is an indicator function . We assume $(\bar{\beta}_i,\bar{\Sigma}_{\beta_i},\bar{a}_\sigma,\bar{b}_\sigma)$ to be known.

- $g(1+\omega) \sim IG(\bar{a}_\omega, 0.5\bar{b}_\omega)\mathcal{I}_{(\omega>-1)}$.

- $g(\alpha) \sim N(\bar{\alpha},\bar{\Sigma}_\alpha)\mathcal{I}_{(stat)}$; where $\mathcal{I}_{(stat)}$ is an indicator for stationarity.

- $g(\eta_{ii})$ is Beta with parameters f_{ij}.

$(\bar{a}_\sigma,\bar{a}_\omega,\bar{b}_\sigma,\bar{b}_\omega,\bar{\phi},\bar{\Sigma}_\beta,f_{ij})$ are assumed to be known.

Let ψ_{-x} the vector of parameters ψ except for x and let $\beta = (\beta_0, \beta_1)$. Then conditional posterior are

$$\begin{aligned}
g(\beta|y^t, \mathcal{S}^t, \psi_{-\alpha}) &\sim N(\tilde{\beta}, \tilde{\bar{\Sigma}}_\beta)\mathcal{I}_{\beta_1>0} \\
g(\sigma^2|y^t, \mathcal{S}^t, \psi_{-\sigma^2}) &\sim IG((\bar{a}_\sigma + T), (\bar{b}_\sigma + (y^* - x^*\beta_0 + \mathcal{S}^*\beta_1)^2) \\
g(1+\omega|y^t, \mathcal{S}^t, \psi_{-\omega}) &\sim IG((\bar{a}_\omega + T_1), (\bar{b}_\omega + \sum_{t=1}^{T_1}(\frac{(y_t^{**} - x_t^{**}\beta_0 - \mathcal{S}_t^{**}\beta_1}{\sigma})^2)\mathcal{I}_{(\omega>-1)} \\
g(\alpha|y^t, \mathcal{S}^t, \psi_{-\alpha}) &\sim \psi(\alpha)N(\tilde{\alpha}, \tilde{\bar{\Sigma}}_\alpha)\mathcal{I}_{(stat)} \\
g(\eta_{ii}|y^t, \mathcal{S}^t, \psi_{-\eta}) &\sim \text{Beta}(f_{ii} + d_{ii}, f_{ij} + d_{ij}) \quad j, i = 1, 2 \\
g(\mathcal{S}_t|y^t, \mathcal{S}_{-t}) &\propto P(\mathcal{S}_t|\mathcal{S}_{t-1})P(\mathcal{S}_{t+1}|\mathcal{S}_t)\prod_k f(y_k|y^{k-1}, \mathcal{S}^k) \qquad (56)
\end{aligned}$$

where a $*$-variable is computed using $\Sigma_p = QQ'$; e.g. $y* = Q^{-1}y$; a $**$-variable is computed premultiplying the original variables by $(1 + \omega S_t)^{0.5}$, $T_1 =$ is the number of elements in T for which $S_t = 1$, $\psi(\alpha) = |\Omega_p|^{-0.5} \exp\{-0.5\sigma^{-2}(y^p - X^p\beta)'\Omega_p^{-1}(y^p - X^p\beta)\}$, and d_{ij} is the number of transitions from state i to state j.

5.2 Markov switching VARs

The simplest specification one can consider is:

$$A_1(\ell)r_t = c(\mathcal{S}_t) + b(\mathcal{S}_t)A_2(\ell)\pi_t + \sigma(\mathcal{S}_t)u_t \qquad u_t \sim N(0,1) \qquad (57)$$

where r_t is the nominal rate, π_t is price inflation; S_t has three states with transition $\eta_{ij} = \begin{bmatrix} \eta_1 & 1-\eta_1 & 0 \\ 0.5*(1-\eta_2) & \eta_2 & 0.5*(1-\eta_2) \\ 0 & 1-\eta_3 & \eta_3 \end{bmatrix}$.

We have restrictions in this model: i) r_t dynamics do not depend on the state; ii) lag distribution on π_t is the same across states except for a scale factor; iii) no possibility to jump from state 1 to state 3 (and viceversa) without passing through state 2; iv) η_{ij} depends only on three parameters.

Let $\psi = vec(A_1(\ell)), vec(A_2(\ell)), c(\mathcal{S}_t), b(\mathcal{S}_t), \sigma(\mathcal{S}_t), \eta_i, i = 1, 2, 3)$. Let I_t the information set available at time t.

To estimates of the unobserved state use the following algorithm:

Algorithm 5.2 *1.* $f(r_t, \mathcal{S}_t|I_{t-1}) = f(r_t|\mathcal{S}_t, I_{t-1})f(\mathcal{S}_t|I_{t-1})$ *where* $f(\mathcal{S}_t|I_{t-1}) = \eta_{ij}$ *and* $f(r_t|\mathcal{S}_t, I_{t-1}) \sim N(A_1(\ell)^{-1}(c(\mathcal{S}_t) + b(\mathcal{S}_t)A_2(\ell)\pi_t), \sigma(\mathcal{S}_t)^2)$.

2. $f(r_t|I_{t-1}) = \sum_{i=1}^{3} f(r_t, \mathcal{S}_t|I_{t-1})$.

3. $f(\mathcal{S}_t|I_t) = \frac{f(r_t, \mathcal{S}_t|I_{t-1})}{f(r_t|I_{t-1})}$.

4. $f(\mathcal{S}_{t+1}|I_t) = \begin{bmatrix} f(\mathcal{S}_t = 1|I_t) \\ f(\mathcal{S}_t = 2|I_t) \\ f(\mathcal{S}_t = 3|I_t) \end{bmatrix}' * \eta_{ij}$

5. *Repeat 1.-4. until t+1=T.*

Given a flat prior on ψ, the posterior is proportional to $f(\psi|r_t, \pi_t)$. Then, the posterior of the parameters and of the states immediate to compute.

If you have a VAR:

$y_t \mathcal{A}_0(\mathcal{S}_t) = x_t' \mathcal{A}_+(\mathcal{S}_t) + u_t'$; $x_t =$ lags of y_t; $u_t \sim N(0, I)$.

Assume $\mathcal{A}_+(\mathcal{S}_t) = D_+(\mathcal{S}_t) + J\mathcal{A}_0(\mathcal{S}_t)$ where $J = [I, 0]'$.

We need restrictions to estimate this model:

(i)$\mathcal{A}_0(\mathcal{S}_t) = \bar{A}_0 \Lambda(\mathcal{S}_t)$ and $D_+(\mathcal{S}_t) = D_t \Lambda(\mathcal{S}_t)$.

(ii) $\mathcal{A}_0(\mathcal{S}_t)$ free and $D_+(\mathcal{S}_t) = \bar{D}_+$
(proportional changes in the contemporaneous and lagged coefficients or state affects only the contemporaneous but not the lagged relationship).
(see Sims and Zha (2006) for details).

5.3 A general Markov Switching specification

General two-state Markov switching model

$$\begin{aligned} y_t &= x_t\alpha_{01} + Y_t\alpha_{02} + u_{0t} \qquad \text{if } S_t = 0 \\ &= x_t\alpha_{11} + Y_t\alpha_{12} + u_{1t} \qquad \text{if } S_t = 1 \end{aligned} \tag{58}$$

x_t : $q \times 1$ vector of exogenous, $Y_t = (y_{t-1}, \ldots y_{t-p})'$; $u_{jt},\ j = 0, 1$ iid $\sim (N(0, \sigma_j^2)$. Let the transition probability for S_t have elements η_{ij}.

This model has no restriction on the dynamics. For identification set $\alpha_{02} < \alpha_{12}$ and let

- $\theta_c =$ parameters common across states.

- $\theta_i =$ parameters unique to the state.

- $\theta_{ig} =$ parameters restricted to achieve identification.

Then the model can be rewritten as:

$$\begin{aligned} y_t &= w_{ct}\theta_c + w_{0t}\theta_0 + w_{gt}\theta_{0g} + u_{0t} \qquad \text{if } S_t = 0 \\ &= w_{ct}\theta_c + w_{1t}\theta_1 + w_{gt}\theta_{1g} + u_{1t} \qquad \text{if } S_t = 1 \end{aligned} \tag{59}$$

$(w_{ct}, w_{it}, w_{ig}) = (x_t, y_t)$; $(\theta_c, \theta_i, \theta_{ig}) = (\alpha_0, \alpha_1)$. Priors:

- $\theta_c \sim N(\bar{\theta}_c, \bar{\Sigma}_c)$, $\theta_i \sim N(\bar{\theta}_i, \bar{\Sigma}_i)$, $\theta_{gi} \sim N(\bar{\theta}_g, \bar{\Sigma}_g)\mathcal{I}_{(rest)}$

- $\sigma_i^2 \sim \frac{v_i \varphi_i}{\chi^2(v_i)}$

- $\eta_{ii} \sim \text{Beta}(f_{1i}, f_{2i})$ $i = 1, 2$,

We assume that $\bar{\theta}_c, \bar{\Sigma}_c, \bar{\theta}_i, \bar{\Sigma}_i, \bar{\theta}_g, \bar{\Sigma}_g, v_i, \varphi_i, f_{ji}$ are known and $\mathcal{I}_{(rest)}$ indicates identification restrictions.

Take the first $max[p, q]$ observations as given. The conditional posteriors are:

- $\theta_c \sim N(\tilde{\theta}_c, \tilde{\Sigma}_c)$ $\tilde{\theta}_c = \sum_{t=min[p,q]}^{T} \tilde{\Sigma}_c^{-1}((\sum_{t=min[p,q]}^{T} \frac{w_{ct} y'_{c,t}}{\sigma_t^2} + \bar{\Sigma}_c \bar{\theta}_c$; $y_{c,t} = y_t - w_{it}\theta_i - w_{gt}\theta_{ig}$; $\tilde{\Sigma}_c = (\sum_{t=min[p,q]}^{T} (\frac{w_{ct} w'_{ct}}{\sigma_t^2} + \bar{\Sigma}_c))^{-1}$.

- $\theta_i \sim N(\tilde{\theta}_i, \tilde{\Sigma}_i)$; $\tilde{\theta}_i = (\sum_{t=min[p,q]}^{T} \tilde{\Sigma}_i^{-1}(\sum_{t=1}^{T_i} \frac{w_{it} y'_{i,t}}{\sigma_t^2} + \bar{\Sigma}_i \bar{\theta}_i)$; $y_{i,t} = y_t - w_{ct}\theta_c - w_{gt}\theta_{ig}$ and $\tilde{\Sigma}_i = (\sum_{t=1}^{T_i} \frac{w_{it} w'_{it}}{\sigma_t^2} + \bar{\Sigma}_i)^{-1}$; $T_i =$ number of observations in state i.

- $\theta_g \sim N(\tilde{\theta}_g, \tilde{\Sigma}_g)$.

- $\sigma^2 \frac{(v_i \varphi_i + R_i^2)}{\sigma_i^2} \sim \chi^2(v_i + T_i - \max[p, q])$.

- $\eta_{ii} \sim$ Beta$(f_{1i} + d_i, f_{2i} + T_i - d_i)$ where $d_1(d_2)$ is the number of transitions from state 1 (2) to state 2 (1).

- Consider $S_{t,\tau} = (S_t, \ldots S_{t+\tau-1})$ and let $S_{t,(-\tau)}$ be S_t with $S_{t,\tau}$ subsequence removed. Then $g(S_{t,\tau}|y^t, S_{t,(-\tau)}) \propto f(y^t|S_t, \theta, \sigma^2)g(S_{t,\tau}|$ $S_{t,(-\tau)}, \eta_{ij})$, (a discrete distribution with 2^k outcomes).

- $g(S_{t,\tau}|S_{t,(-\tau)}, \eta_{ij}) = g(S_{t,\tau}|S_{t-1}, S_{t+\tau}, \eta_{ij})$ and $f(y^t|S_t, \theta) \propto \prod_{j=t}^{t+\tau-1}$ $\frac{1}{\sigma_j} \exp\{-0.5\frac{u_j^2}{\sigma_j^2}\}$.

How do we choose initial conditions?

- Assign all the observations to one state. Arbitrarily set the parameters of the other state equal to the estimates plus (or minus) 0.1.

- Split the points arbitrarily but equally across the two states.

Bayesian methods for VAR Models

Fabio Canova
EUI and CEPR
October 2012

Outline

- Introduction.
- Likelihood function for an M variable VAR(q).
- Priors for VARs (Minnesota (Litterman), General, Hierarchical, DSGE based).
- Forecasting with BVARs.
- Structural (overidentified) BVAR.
- Univariate dynamic panels; endogenous grouping, partial pooling of VARs.

References

Lutkepohl, H., (1991), *Introduction to Multiple Time Series Analysis*, Springer and Verlag.

Ballabriga, F. (1997) "Bayesian Vector Autoregressions", ESADE, manuscript.

Canova, F. (1992) " An Alternative Approach to Modelling and Forecasting Seasonal Time Series " *Journal of Business and Economic Statistics*, 10, 97-108.

Canova, F. (1993a) " Forecasting time series with common seasonal patterns", *Journal of Econometrics*, 55, 173-200.

Canova, F. (2004), "Testing for Convergence Clubs in Income per Capita: A Predictive Density Approach", *International Economic Review*, 45(1), 2004, 49-77.

Del Negro, M. and F. Schorfheide (2004), " Priors from General Equilibrium Models for VARs", *International Economic Review*, 45, 643-673.

Del Negro, M. and Schorfheide, F. (2012)," Bayesian macroeconometrics" in J. Geweke, G.Koop, H. Van Dijk (eds.) The Oxford Handbook of Bayesian econometrics, Oxford University Press.

Favero, C. (2001) Econometrics, Oxford University Press.

Giannone, D., Primiceri, G. and Lenza, M. (2012) Prior selection for vector autoregression, Northwestern University, manuscript.

Girlchrist, S. and Gertler, M., (1994), Monetary Policy, Business Cycles and the Behavior of Small Manufactoring Firms, *Quarterly Journal of Economics*, CIX, 309-340.

Kadiyala, R. and Karlsson, S. (1997) Numerical methods for estimation and Inference in Bayesian VAR models, *Journal of Applied Econometrics*, 12, 99-132.

Killian, L. (2011) Structural vector autoregressions, University of Michigan, manuscript.

Ingram, B. and Whitemann, C. (1994), "Supplanting the Minnesota prior. Forecasting macroeconomic time series using real business cycle priors, *Journal of Monetary Economics*, 34, 497-510.

Lindlay, D. V. and Smith, A.F.M. (1972) "Bayes Estimates of the Linear Model", *Journal of the Royal Statistical Association*, Ser B, 34, 1-18.

Robertson, J. and Tallman, E. (1999), 'Vector Autoregressions: Forecasting and Reality", *Federal Reserve Bank of Atlanta, Economic Review*, First quarter, 4-18.

Sims, C. and Zha T. (1998) "Bayesian Methods for Dynamic Multivariate Models", *International Economic Review*, 39, 949-968.

Waggoner and T. Zha (2003) A Gibbs Simulator for Restricted VAR models, *Journal of Economic Dynamics and Control*, 26, 349-366.

Zellner, A., Hong, (1989) Forecasting International Growth rates using Bayesian Shrinkage and other procedures, *Journal of Econometrics*, 40, 183-202.

Zha, T. (1999) "Block Recursion and Structural Vector Autoregressions", *Journal of Econometrics*, 90, 291-316.

1 Why BVAR?

- VARs have lots of parameters to be estimated. If they are used for forecasting, their performance is poor.

- Even if they are used for structural analysis, parameter uncertainty is a concern.

- Impossible to incorporate prior views of the client into classical analysis.

- BVAR are a flexible way to incorporate extraneous (client) information. They can also help to reduce the dimensionality of the parameter space.

2 Likelihood function of an M variable VAR(q)

Consider an $M \times 1$ VAR model with q lags each (k=Mq coefficients each equation, Mk total coefficients in total), no constant.

$$y_t = B(L)y_{t-1} + e_t \quad e_t \sim N(0, \Sigma_e)$$

Letting $B = [B_1, \ldots B_q]$; $X_t = [y_{t-1}, \ldots y_{t-p}]$, $\beta = vec(B)$, the VAR is:

$$y = (I_M \otimes X)\beta + e \quad e \sim (0, \Sigma_e \otimes I_T) \tag{1}$$

where y, e are $MT \times 1$ vectors, I_M is the identify matrix, and β is a $Mk \times 1$ vector. Conditioning on initial observations $y_p = [y_{-1}, \ldots, y_{-q}]$:

$$\begin{aligned} L(\beta, \Sigma_e | y, y_p) &= \frac{1}{(2\pi)^{0.5MT}} |\Sigma_e \otimes I_T|^{-0.5} \\ &\times \exp\{-0.5(y - (I_M \otimes X)\beta)'(\Sigma_e^{-1} \otimes I_T)(y - (I_M \otimes X)\beta)\} \end{aligned}$$

Some manipulations of the likelihood function:

$$\begin{aligned}(y-(I_M\otimes X)\beta)'(\Sigma_e^{-1}\otimes I_T)(y-(I_M\otimes X)\beta) &= \\ (\Sigma_e^{-0.5}\otimes I_T)(y-(I_M\otimes X)\beta)'(\Sigma_e^{-0.5}\otimes I_T)(y-(I_M\otimes X)\beta) &= \end{aligned}$$

$$[(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes X)\beta)]'[(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes X)\beta)]$$

Also $(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes X)\beta = (\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes X)\hat{\beta}+$ $(\Sigma_e^{-0.5}\otimes X)(\hat{\beta}-\beta)$ where $\hat{\beta}=(\Sigma_e^{-1}\otimes X'X)^{-1}(\Sigma_e^{-1}\otimes X)y$. Therefore:

$$\begin{aligned}(y-(I_M\otimes X)\beta)'(\Sigma_e^{-1}\otimes I_T)(y-(I_M\otimes X)\beta) &= \\ ((\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes X)\hat{\beta})'((\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes X)\hat{\beta}) &+ \\ (\hat{\beta}-\beta)'(\Sigma_e^{-1}\otimes X'X)(\hat{\beta}-\beta)&\end{aligned}$$

Putting the pieces together:

$$
\begin{aligned}
L(\beta, \Sigma_e) &\propto |\Sigma_e \otimes I_T|^{-0.5} exp\{-0.5((\beta - \hat{\beta})'(\Sigma_e^{-1} \otimes X'X)(\beta - \hat{\beta}) \\
&- 0.5[(\Sigma_e^{-0.5} \otimes I_T)y - (\Sigma_e^{-0.5} \otimes X)\hat{\beta})' \\
&\quad [(\Sigma_e^{-0.5} \otimes I_T)y - (\Sigma_e^{-0.5} \otimes X)\hat{\beta})]\} \\
&= |\Sigma_e|^{-0.5k} exp\{-0.5(\beta - \hat{\beta})'(\Sigma_e^{-1} \otimes X'X)(\beta - \hat{\beta})\} \\
&\times |\Sigma_e|^{-0.5(T-k)} exp\{-0.5tr[(\Sigma_e^{-0.5} \otimes I_T)y \\
&- (\Sigma_e^{-0.5} \otimes X)\hat{\beta})'(\Sigma_e^{-0.5} \otimes I_T)y - (\Sigma_e^{-0.5} \otimes X)\hat{\beta})]\} \\
&\propto N(\beta|\hat{\beta}, \Sigma_e, X, y, y_p) \times iW(\Sigma_e|\hat{\beta}, X, y, y_p, T - \nu) \qquad (2)
\end{aligned}
$$

where tr = trace of the matrix.

• The conditional likelihood of a VAR(q) is the product of Normal density for β conditional on $\hat{\beta}$ and Σ_e, and an inverted Wishart distribution for Σ_e, conditional on $\hat{\beta}$, with scale $(y - (x \otimes \Sigma_e)\hat{\beta})'(y - (x \otimes \Sigma_e)\hat{\beta})$, where $(T - \nu)$ degrees of freedom; $\nu = k + M + 1$.

• In classical inference $L(y|\beta, \Sigma_e, X, y_p)$ is normal (in large samples). Here $\hat{\beta} = \beta_{ols}$ if errors are independent and regressors are the same in each equation (so that ML=SUR=OLS). Typically, $\Sigma_e = (\Sigma_e)_{ols} = \hat{\Sigma}$, and thus $L(y|\beta, \Sigma, X, y_p) \approx (y|\beta, \hat{\Sigma}_e, X, y_p)$.

• Bayesian inference: combine likelihood with a prior.

i)If the prior is conjugate and the hyperparameters are known (or estimated): closed form solution for the conditional and marginal of β and marginal of Σ_e.

ii) If hyperparameters are random, need numerical MC methods to get conditional and marginal distributions, even if prior is conjugate in general.

What priors conjugate with (2)?

3 Conjugate priors for VARs

1. Normal prior for β with Σ_e fixed.

2. Diffuse prior for both β and Σ_e.

3. Normal prior for β, diffuse prior for Σ_e (semi-conjugate)

4. Normal for $\beta|\Sigma_e$, inverted Wishart for Σ_e (conjugate).

Case 1: $\beta = \bar{\beta} + v,\ v \sim N(0, \Sigma_b)$, where $\bar{\beta}, \Sigma_b$ known.

Then the prior is:

$$\begin{aligned} g(\beta) &\propto |\Sigma_b|^{-0.5} exp[-0.5(\beta - \bar{\beta})' \Sigma_b^{-1} (\beta - \bar{\beta})] \\ &= |\Sigma_b|^{-0.5} exp[-0.5(\Sigma_b^{-0.5}(\beta - \bar{\beta}))' \Sigma_b^{-0.5} (\beta - \bar{\beta})] \end{aligned} \tag{3}$$

Posterior:

$$
\begin{aligned}
g(\beta|y) &\propto g(\beta)L(\beta|y) \\
&= |\Sigma_b|^{-0.5}\,\mathsf{exp}\{-0.5(\Sigma_b^{-0.5}(\beta-\bar{\beta}))'\Sigma_b^{-0.5}(\beta-\bar{\beta})\} \times |\Sigma_e \otimes I_T|^{-0.5} \\
&\times \mathsf{exp}\,\{(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes X)\beta)'(\Sigma_e^{-0.5}\otimes I_T)y-(\Sigma_e^{-0.5}\otimes X)\beta)\} \\
&= \mathsf{exp}\,\{-0.5(z-Z\beta)'(z-Z\beta)\} \\
&= \mathsf{exp}\,\{-0.5(\beta-\tilde{\beta})'Z'Z(\beta-\tilde{\beta})+(z-Z\tilde{\beta})'(z-Z\tilde{\beta})\} \qquad (4)
\end{aligned}
$$

where $z=[\Sigma_b^{-0.5}\bar{\beta},(\Sigma_e^{-0.5}\otimes I_T)y]'$; $Z=[\Sigma_b^{-0.5},(\Sigma_e^{-0.5}\otimes X)]'$ and

$$\tilde{\beta}=(Z'Z)^{-1}(Z'z)=[\Sigma_b^{-1}+(\Sigma_e^{-1}\otimes X'X)]^{-1}[\Sigma_b^{-1}\bar{\beta}+(\Sigma_e^{-1}\otimes X)'y] \quad (5)$$

Since Σ_e and Σ_b are fixed, the second term in (4) is a constant and

$$g(\beta|y) \propto exp[-0.5(\beta-\tilde{\beta})'Z'Z(\beta-\tilde{\beta})] \quad (6)$$

$$\propto exp[-0.5(\beta-\tilde{\beta})'\tilde{\Sigma}_b^{-1}(\beta-\tilde{\beta})] \quad (7)$$

Conclusion: $g(\beta|y)$ is $N(\tilde{\beta},\tilde{\Sigma}_b)$ where $\tilde{\Sigma}_b = [\Sigma_b^{-1} + (\Sigma_e^{-1} \otimes X'X)]^{-1}$.

- If Σ_e is unknown, use $\hat{\Sigma}_e = \frac{1}{T-1}\hat{e}'\hat{e}$ in formulas, where $\hat{e}_t = y_t - (I \otimes X)\hat{\beta}$ and $\hat{\beta} = \beta_{ols}$.

- The $\tilde{\beta}$ obtained with this prior is related to the classical least square estimator under uncertain linear restrictions.

Model

$$\begin{aligned} y_t &= x_t B + e_t \qquad e_t \sim (0, \sigma^2) \\ \bar{B} &= B - \epsilon \quad \epsilon \sim (0, \Sigma_b) \end{aligned} \tag{8}$$

where $B = [B_1, \ldots B_q]', x_t = [y_{t-1}, \ldots y_{t-q}]$. Set $z_t = [y_t, \bar{B}]'$, $Z_t = [x_t, I]'$, $E_t = [e_t, \epsilon]'$. Then $z_t = Z_t B + E_t$ where $E_t \sim (0, \Sigma_E)$,Σ_E is known, $t = 1, \ldots, T$. Thus:

$B_{GLS} = (Z'\Sigma_E^{-1} Z)^{-1}(Z'\Sigma_E^{-1} z) = \tilde{B}$ (Theil' s mixed estimator).

• Prior on VAR coefficients can be treated as a dummy observation added to the system of VAR equations.

• Prior can treat it as initial condition. If we write the initial observation as $y_0 = x_0 B + e_0$, then $y_0 = \sigma^2 W^{-1} \bar{B}, x_0 = \sigma^2 W^{-1}$, $e_0 = \sigma^2 W^{-1} \epsilon$, $WW' = \Sigma_b$.

Special case 1: Ridge Estimator

Consider a univariate model. If $\bar{B} = 0$; $\Sigma_e = I * \sigma_e^2$, $\Sigma_b = I * \sigma_v^2$,

$$\tilde{B} = (I_q + \kappa(X'X)^{-1})^{-1}\hat{B} \tag{9}$$

where $\kappa = \frac{\sigma_e^2}{\sigma_v^2}$ and $\hat{B} = (X'X)^{-1}(X'Y)$.

- Prior reflects the belief that all the coefficients of an AR(q) are small.

- Posterior estimator increases the smallest eigenvalues of the data matrix by a factor κ (useful when q is large: $(X'X)$ matrix ill-conditioned)

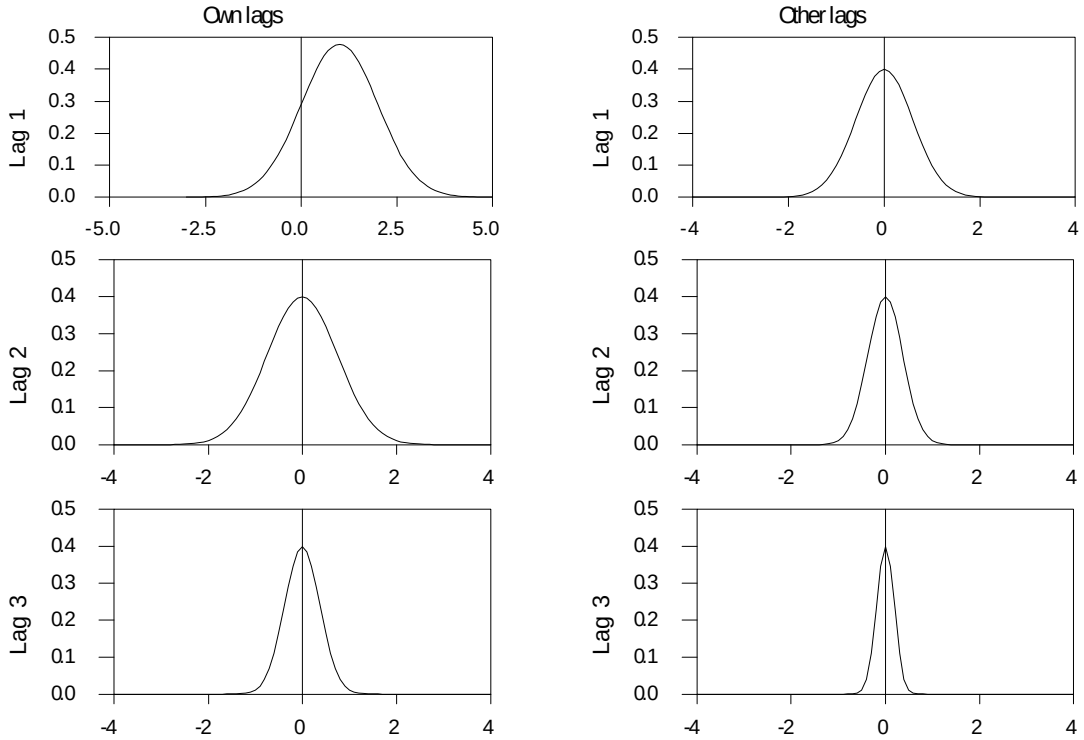
Special case 2: Litterman (Minnesota) setup

Multivariate setup. Now $\bar{\beta}, \Sigma_\beta$ have special structure: $\bar{\beta} = 0$ except $\bar{\beta}_{i1} = 1$. $\Sigma_b = \Sigma_b(\phi)$ where:

$$\begin{aligned} \sigma_{ij,\ell} &= \frac{\phi_0}{h(\ell)} \text{ if } i = j \\ &= \phi_0 \frac{\phi_1}{h(\ell)} * (\frac{\sigma_i}{\sigma_j})^2 \text{ otherwise} \quad (10) \\ &= \phi_0 * \phi_2 \text{ for exogenous variables} \quad (11) \end{aligned}$$

$\phi_0 =$ tightness on the variance of the first lag; $\phi_1 =$ relative tightness on other variables; $h(l) =$ tightness of the variance of lags other than the first one; (decay parameter); $(\frac{\sigma_i}{\sigma_j})^2$scaling factor.

Typically, $h(\ell)$ regulated by one (decay) parameter. Useful structures: harmonic decay $h(\ell) = l^{\phi_3}$; geometric decay $h(\ell) = \phi_3^{-\ell+1}$; linear decay $h(\ell) = l$.


Own lags
Other lags
Lag 1
Lag 2
Lag 3
0.0
0.1
0.2
0.3
0.4
0.5
-5.0
-2.5
0.0
2.5
5.0
-4
-2
0
2
4

Logic for this (shrinkage) prior:

- Mean chosen so that the VAR is M a-priori random walks (good for forecasting).

- Σ_b very big. Decrease dimensionality by setting $\Sigma_b = \Sigma_b(\phi)$.

- Σ_b is a-priori diagonal (no expected relationship among equations and coefficients); ϕ_0 is the relative importance of prior to the data.

- The variance of lags of LHS variables shrinks to zero as lags increase. Variance of lags of other RHS variables shrinks to zero at a different rate (governed by ϕ_1). $\phi_1 \leq 1$ relative importance of other variables.

- Variance of the exogenous variables is regulated by ϕ_2. If ϕ_2 is large, prior information on the exogenous variables diffuse.

Example 1 *Bivariate VAR(2) with* $h(\ell) = \ell$.

$$\bar{\beta} = [1, 0, 0, 0, 0, 1, 0, 0]$$

$$\Sigma_b = \begin{bmatrix} \phi_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \phi_0\phi_1(\frac{\sigma_1}{\sigma_2})^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\phi_0}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\phi_0}{2}\phi_1(\frac{\sigma_1}{\sigma_2})^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi_0\phi_1(\frac{\sigma_1}{\sigma_2})^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \phi_0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\phi_0}{2}\phi_1(\frac{\sigma_1}{\sigma_2})^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\phi_0}{2} \end{bmatrix}$$

- If Σ_b is diagonal, $\phi_1 = 1$ and the same variables belong to all equations, then $\tilde{\beta}$=vec($\tilde{\beta}_i$), where $\tilde{\beta}_i$ computed equation by equation. In other setups, Σ_b is not diagonal and this result does not hold.

- Let $\alpha = (\beta, vech(\Sigma_b))$. Minnesota prior makes $\alpha = \alpha(\phi)$, ϕ small dimension. Better estimates of ϕ than for α from the data. Better forecasts than univariate ARIMA models or traditional multivariate SES (see e.g. Robertson and Tallman (1999)).

- Standard approaches: "unimportant" lags purged using t-test. (see e.g. Favero (2001)). Strong a-priori restrictions on what variables and which lags enter in the VAR. Unpalatable.

- Minnesota prior imposes probability distributions on VAR coefficients (uncertain linear restrictions). It gives a reasonable account of the uncertainty faced by an investigator.

- How do we choose $\phi = (\phi_0,\ \phi_1,\ \phi_2, \ldots)$ and $(\frac{\sigma_i}{\sigma_j})^2$?

1) Use rules of thumb. Typical default values: $\phi_0 = 0.2,\ \phi_1 = 0.5, \phi_2 = 10^5$, an harmonic specification for $h(\ell)$ with $\phi_3 = 1$ or 2, implying loose prior on lagged coefficients and uninformative prior for the exogenous variables.

2) Estimate them using ML-II approach. That is, maximize $\mathcal{L}(\phi|y) = \int f(\beta|y,\phi)g(\beta|\phi)d\beta$ on training sample.

3) Set up prior $g(\phi)$, produce hierarchical posterior estimates. For this we need MCMC methods, see later.

Example 2 *Consider* $y_t = Bx_t + u_t$, B *scalar,* $u_t \sim N(0, \sigma_u^2)$, σ_u^2 *known and let* $B = \bar{B} + \upsilon$ *where* $\upsilon \sim N(0, \sigma_\upsilon^2)$, $\bar{B}$ *fixed and* $\sigma_\upsilon^2 = q(\phi)^2$, *where* ϕ *is a set of hyperparameters.*

Then $y_t = \bar{B}x_t + \epsilon_t$ *where* $\epsilon_t = e_t + \upsilon x_t$ *and posterior kernel is:*

$$\grave{g}(\beta, \phi|y) = \frac{1}{(2\pi)^{0.5}\sigma_u\sigma_\upsilon} \exp\{-0.5\frac{(y - Bx)^2}{\sigma_u^2} - 0.5\frac{(B - \bar{B})^2}{\sigma_\upsilon^2}\} \quad (12)$$

where $y = [y_1, \ldots y_t]'$, $x = [x_1, \ldots x_t]'$. *Integrating* B *out of (12):*

$$\tilde{g}(\phi|y) = \frac{1}{(2\pi q(\phi)^2 tr|X'X| + \sigma_u^2)^{0.5}} \exp\{-0.5\frac{(y - \bar{B}x)^2}{\sigma_u^2 + q(\phi)^2 tr|X'X|}\} \quad (13)$$

Maximize (13) using gradient or grid methods. Alternative: compute prediction error decomposition of $\grave{g}(\phi|y)$ *with the Kalman filter; find modal estimates of* ϕ.

- Recent applications of this method.

i) Giannone, Primiceri, Lenza (2012): employ marginal likelihood to choose the informativeness of prior restrictions.

Idea: $\beta \sim N(\bar{\beta}, \Sigma \otimes \Omega\zeta)$ where ζ is a scalar Σ the covariance matrix of VAR shocks and Ω a known scale matrix. problem choose ζ in an optimal way.

ii) Belmonte, Koop, Korobilis (2012): employ marginal likelihood to choose the informativeness of prior distribution for time variations in coefficients and in the variance.

iii) Carriero, Kapetanios, Marcellino (2011): employ marginal likelihood to select the variance of the prior from a grid.

Fully Hierachical VARs

Model is

$$y_t = (I \otimes X)\beta + e \quad e \sim N(0, \Sigma) \tag{14}$$
$$\beta = \bar{\beta} + v \quad v \sim N(0, \Sigma \otimes \Omega * \zeta) \tag{15}$$
$$\zeta = \bar{\zeta} + \epsilon \quad \epsilon \sim N(0, \eta) \tag{16}$$

$\bar{\beta},\ \Omega, \eta$ known (or estimable).

- Need to compute the joint posterior of β, ζ.

- Then $g(\zeta|\beta, y, X, y_p) = \int g(\zeta, \beta|y, X, y_p)d\beta$.

- Typically impossible to compute $g(\zeta|\beta, y, X, y_p)$ analytically. One example when this is possible is in Canova (2007, chapter 9). Otherwise use MCMC methods to get draws from this distribution.

Results for other prior structures (Kadiyala and Karlsson (1997)):

Case 2) $g(\beta, \Sigma_e)$ is diffuse, i.e. $g(\beta, \Sigma_e) \propto |\Sigma_e|^{-0.5(M+1)}$. Then

$$g(\beta|\Sigma_e, y) \sim N(\hat{\beta}, \Sigma_e \otimes (X'X)^{-1}) \quad (17)$$

$$g(\Sigma_e|y) \sim iW((y - X\hat{B})'(y - X\hat{B}), T - k) \quad (18)$$

Note: the marginal $g(\beta|y)$ is a t-distribution with parameters $((X'X), (y - X\hat{B})'(y - X\hat{B}), \hat{B}, T - k)$, where a $\hat{B} = (X'X)^{-1}(X'Y)$, $\beta = vec(B)$.

Case 3): $g(\beta, \Sigma_e)$ is Normal-diffuse, i.e. $g(\beta) \sim N(\bar{\beta}, \bar{\Sigma}_b)$; $\bar{\beta}$ and Σ_b known, and $g(\Sigma_e) \propto |\Sigma_e|^{-0.5(M+1)}$. This prior is semi-conjugate. This means that the conditional posteriors are of the same form as case 2) (moments of the normal are different) but the marginal posterior $g(\beta|y) \propto \exp\{0.5(\beta-\tilde{\beta})'\bar{\Sigma}_b^{-1}(\beta-\tilde{\beta})\} \times |(y-X\hat{B})'(y-X\hat{B})+(B-\hat{B})'(X'X)(B-\hat{B})|^{-0.5T}$ has an unknown format.

Case 4): $g(\beta|\Sigma_e) \sim N(\bar{\beta}, \Sigma_e \otimes \bar{\Omega})$ and $g(\Sigma_e) \sim iW(\bar{\Sigma}, \bar{\nu})$. Then $g(\beta|\Sigma_e, y) \sim N(\tilde{\beta}, \Sigma_e \otimes \tilde{\Omega})$, $g(\Sigma_e|y) \sim iW(\tilde{\Sigma}, T+\bar{\nu})$ where $\tilde{\Omega} = (\bar{\Omega}^{-1} + X'X)^{-1}$; $\tilde{\Sigma} = \hat{B}'X'X\hat{B} + \bar{B}'\bar{\Omega}^{-1}\bar{B} + \bar{\Sigma} + (y-X\hat{B})'(y-X\hat{B}) - \tilde{B}(\bar{\Omega}^{-1}+X'X)\tilde{B}$; $\tilde{\beta} = \tilde{\Omega}(\bar{\Omega}^{-1}\bar{\beta} + X'X\hat{\beta})$.

Marginal of β is t with parameters $(\tilde{\Omega}^{-1}, \tilde{\Sigma}_e, \tilde{B}, T+\bar{\nu})$.

- In cases 2)-4) there is posterior dependence among the equations (even with prior independence and $\phi_1 = 1$).

• Any additional uncertain restrictions on the coefficients can be tagged on to the system in exactly the same way as in case 1).

i) Quasi-deterministic seasonality

Example 3 *In quarterly data, a prior for a bivariate VAR(2) with 4 seasonal dummies has mean* $\bar{\beta} = [1, 0, 0, 0, 0, 0, 0, 0|0, 0, 1, 0, 0, 0, 0, 0]$ *and the block of* Σ_a *corresponding to the seasonal dummies has diagonal elements* $\sigma_{dd} = \theta_0\theta_s$ *where* θ_s *is the tightness of seasonal information (large* θ_s *means little prior information).*

ii) Stochastic seasonality: there is peak in spectrum at $\omega_q = \frac{\pi}{2}$ or π or both (quarterly data).

Let $y_t = D(\ell)e_t$. If there is a peak at ω_q : $|D(\omega_q)|^2$ is large or $|B(\omega_q)|^2$ small, where $B(\ell) = D(\ell)^{-1}$.

A small $|B(\omega_q)|^2$ implies $\sum_{j=1}^{\infty} B_j cos(j\omega_q) \approx -1$. This is a "sum-of-coefficients" restrictions.

In a VAR model: $1 + \sum_{j=1}^{\infty} B_j cos(j\omega_q) \approx 0$, B_j where AR coefficients in equation j. (see Canova, 1992).

Set $\mathcal{R}\beta = r + v$, $r = [-1, \ldots, -1]'$ and $\mathcal{R}$ is a $2 \times Mk$. For quarterly data, if the first variable of the VAR displays seasonality at $\frac{\pi}{2}$ and π:

$$\mathcal{R} = \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & -1 & \dots & 0 \\ -1 & 1 & -1 & 1 & -1 & 1 & \dots & 0 \end{bmatrix}$$

Add these restrictions to original prior. Use Theil's Mixed estimator.

iii) Trend restrictions on variable i: $\sum_{j=1}^{\infty} B_{ji} \approx -1$;

iv) Cyclical peak restriction $\sum_{j=1}^{\infty} B_{ji} cos(j\omega) \approx -1$ for all $\omega \in (\frac{2\pi}{d} \pm \delta)$, some d, δ small, $i = 1, 2, \dots$.

v) High coherence at frequency $\frac{\pi}{2}$ in series i and i' of a VAR implies that $\sum_{j=1}^{\infty}(-1)^j B_{i'i'(2j)} + \sum_{j=1}^{\infty}(-1)^j B_{ii(2j)} \approx -2$.

Some tips

- If hyperparameters are treated as fixed, we need some sensitivity analysis. Rule-of-thumb parameters work well for forecasting. Do they work well in structural estimation?

- You can set prior mean or prior variance as you wish (after all this is a prior!!). In all cases we consider, the covariance matrix has a Kroneker product form (easy to compute).

- What are the gains from using fully hierarchical methods (relative to empirical based or rules of thumb)? Not much is known (see Giannone et al. (2012), Carriero et al. (2012)).

Example 4 *(Forecasting inflation rates in Italy)*

- Large changes in the persistence of inflation: AR(1) coefficient is 0.85 in 1980s and 0.48 in 1990s.

- Which model to use? Univariate ARIMA; VAR(4) with annualized three month inflation, rent inflation and the unemployment rate; two trivariate BVAR(4) (one with arbitrary hyperparameters 0.2, 1, 0.5; one with optimal ones =(0.15, 2.0, 1.0)). Report one year ahead Theil-U Statistics.

Sample	*ARIMA*	*VAR*	*BVAR1*	*BVAR2*
1996:1-2000:4	*1.04*	*1.47*	*1.09 (0.03)*	*0.97 (0.02)*
1990:1-1995:4	*0.99*	*1.24*	*1.04 (0.04)*	*0.94 (0.03)*

- Difficult to forecast; VAR poor, BVAR better.

- Results robust to changes of the forecasting sample.

4 Forecasting with BVARs: Fan Charts

Let the VAR be written in a companion form:

$$\mathbb{Y}_t = \mathbb{B}\mathbb{Y}_{t-1} + \mathbb{E}_t \tag{19}$$

where $\mathbb{Y}_t$ and $\mathbb{E}_t$ are $Mq \times 1$ vectors, $\mathbb{B}$ is a $Mq \times Mq$ matrix.

Repeatedly substituting: $\mathbb{Y}_t = \mathbb{B}^\tau \mathbb{Y}_{t-\tau} + \sum_{j=0}^{\tau-1} \mathbb{B}^j \mathbb{E}_{t-j}$ or

$$y_t = \mathcal{J}\mathbb{B}^\tau \mathbb{Y}_{t-\tau} + \sum_{j=0}^{\tau-1} \mathbb{B}^j e_{t-j} \tag{20}$$

where $\mathcal{J}$ is such that $\mathcal{J}\mathbb{Y}_t = y_t$, $\mathcal{J}\mathbb{E}_t = e_t$ and $\mathcal{J}'\mathcal{J}E_t = E_t$.

- Unconditional point forecast for $y_{t+\tau}$

$$y_t(\tau) = J\mathbb{B}^{\tau}\mathbb{Y}_t \tag{21}$$

Use the posterior mean or median or mode,$\tilde{\mathbb{B}}$ depending on the loss function. Recall that if $\mathbb{E}_t$ is normal, mean, mode and median coincide.

Forecast error is $y_{t+\tau} - \hat{y}_t(\tau) = \sum_{j=0}^{\tau-1} \tilde{\mathbb{B}}^j e_{t+\tau-j} + [y_t(\tau) - \hat{y}_t(\tau)]$.

• Unconditional probability distributions for forecasts (fan charts).

Algorithm 4.1 *Assume* $\beta \sim N(\tilde{\beta}, \tilde{\Sigma}_b)$. *Set* $\tilde{\mathcal{P}}\tilde{\mathcal{P}}' = \tilde{\Sigma}_b$.

- *Draw a normal (0,1) random vector* v_t *and set* $\beta^\ell = \tilde{\beta} + \tilde{\mathcal{P}} v_t$

- *Construct point forecasts* $y_t(\tau),\ \tau = 1, 2, , \ldots$ *using* β^ℓ

- *Repeat previous steps L times.*

- *Construct distributions at each* τ *using kernel methods and extract percentiles (fan charts).*

Can also be used for recursive forecasts charts, only difference would be that $\tilde{\beta}$ and $\tilde{\Sigma}_b$ depend on t (they are recursively estimated).

• "Average" τ-step ahead forecasts

Construct $f\left(y_{t+\tau} \mid y_t\right) = \int f\left(y_{t+\tau} \mid y_t, \beta\right) g\left(\beta \mid y_t\right) d\beta$ where $f\left(y_{t+\tau} \mid y_t, \beta\right)$ is the conditional density of $y_{t+\tau}$ and $g\left(\beta \mid y_t\right)$ the posterior of β.

- Can calculate this numerically. Draw β^l from $g\left(\beta \mid y_t\right)$. Compute $f\left(y_{t+\tau} \mid y_t, \beta^l\right)$. Average over $y_{t+\tau}$ paths.

- Can use the above algorithm to calculate turning point probabilities

i) A upturn turn τ in $y_t(\tau)$ (typically, GDP) if $y_t(\tau-2) < y_t(\tau-1) < y_t(\tau) > y_t(\tau+1) > y_t(\tau+2)$.

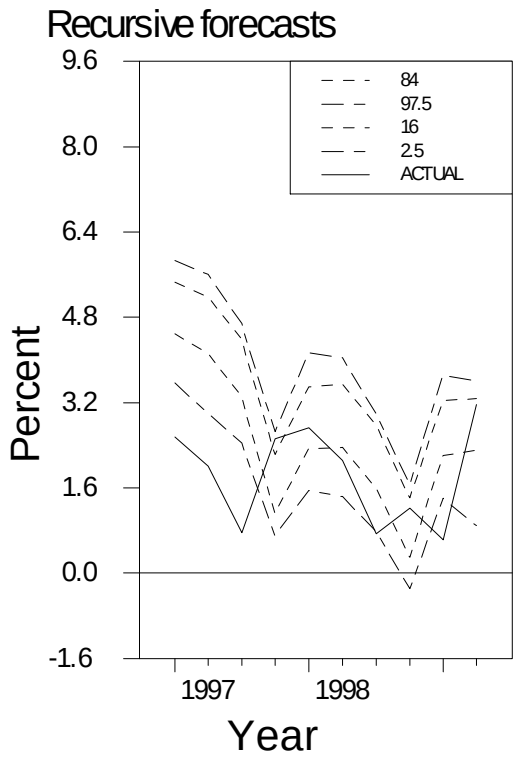
ii) A downturn at τ in $y_t^i(\tau)$ if $y_t(\tau-2) > y_t(\tau-1) > y_t(\tau) < y_t(\tau+1) < y_t(\tau+2)$.

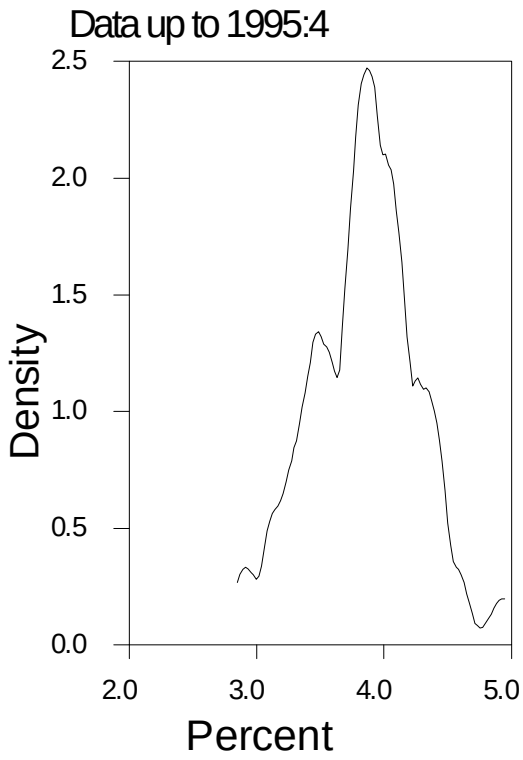
Implementation: draw β^ℓ, construct $(y_t(\tau))^\ell,\ \ell = 1, \ldots L$; apply above rule for each τ. The fraction of times for which the condition is satisfied at each t is an estimate of the probability of an upturn (downturn).

Example 5 *Use a BVAR to construct one year ahead bands for inflation, recursively updating posterior estimates over 1995:4-1998:2.*

- The bands are relatively tight: errors at the beginning. Distribution of one year ahead forecasts (based on 1995:4) also tight.

- Sample 1996:1 2002:4: 4 downturns. Median forecasted downturns 3; $Pr(n \leq 3) = 0.9$, $Pr(n > 4) = 0.0$.


Recursive forecasts
9.6
8.0
6.4
4.8
3.2
1.6
0.0
-1.6
Percent
84
97.5
16
2.5
ACTUAL
1997
1998
Year


Data up to 1995:4
2.5
2.0
1.5
1.0
0.5
0.0
Density
2.0
3.0
4.0
5.0
Percent

5 DSGE priors for VARs

Log linearized solution of a DSGE model:

$$y_{1t+1} = \Gamma(\theta) y_{1t} + e_{t+1} \quad (22)$$
$$y_{2t} = \Pi(\theta) y_{1t} \quad (23)$$

y_{1t} are exogenous and endogenous $(d_1 \times 1)$ states; y_{2t} are endogenous $(d_2 \times 1)$ controls; e_{t+1} are the innovations in the shocks; $\Gamma(\theta), \Pi(\theta)$ function of θ, structural parameters. Letting $y_t = [y_{2t}, y_{1t}]'$ the system is:

$$\begin{bmatrix} 0 & 0 \\ 0 & I_{d_1} \end{bmatrix} y_{t+1} = \begin{bmatrix} -I_{d_2} & \Pi(\theta) \\ 0 & \Gamma(\theta) \end{bmatrix} y_t + \begin{bmatrix} 0 \\ e_{t+1} \end{bmatrix} \quad (24)$$

or

$$\mathcal{B}_0 y_{t+1} = \mathcal{B}_1(\theta) y_t + u_{t+1}$$

- (Log-)linear DSGE solution is a restricted VAR!

Given $g(\theta)$, the model implies priors $g(\Pi(\theta))$ and $g(\Gamma(\theta))$ for the decision rule coefficients and thus a prior $\beta = [\mathcal{B}(\ell)]$.

• A DSGE model implies restrictions on the VAR coefficients. It can be used to link (in a hierarchical fashion) the VAR coefficients β and DSGE parameters θ).

Note: if $\theta \sim N(\bar{\theta}, \Sigma_\theta)$, $vec(\Pi(\theta)) \sim N(vec(\Pi(\bar{\theta})), \frac{\partial vec(\Pi(\theta))}{\partial \theta} \Sigma_\theta \frac{\partial vec(\Pi(\theta))}{\partial \theta}')$; $vec(\Gamma(\theta)) \sim N(vec(\Gamma(\bar{\theta})), \frac{\partial vec(\Gamma(\theta))}{\partial \theta} \Sigma_\theta \frac{\partial vec(\Gamma(\theta))}{\partial \theta}')$.

Example 6 *Consider a VAR(q)* $y_{t+1} = B(\ell)y_t + u_t$. *From (24)* $g(B_1)$ *is normal with mean* $\mathcal{B}_0^G \mathcal{B}_1(\bar{\theta})$, $\mathcal{B}_0^G$ *is the generalized inverse of* $\mathcal{B}_0$ *and variance* $\Sigma_b = \beta_0^G \Sigma_{b_1} \beta_0^{G\prime}$; $\beta_0^G = vec(\mathcal{B}_0^G)$; Σ_{b_1} *is the variance of* $vec(\mathcal{B}_1(\theta))$. *A DSGE prior on* $B_\ell, \ell \geq 2$ *has a dogmatic form: mean zero and zero variance.*

If there are unobservables, want a prior for VAR with observables only.

Example 7 *(RBC prior: Ingram and Whiteman, 1994). A RBC model with utility function* $U(c, n) = \log(c_t) + \log(1 - n_t)$ *implies*

$$\begin{bmatrix} K_{t+1} \\ \ln A_{t+1} \end{bmatrix} = \begin{bmatrix} \gamma_{kk} & \gamma_{ka} \\ 0 & \rho \end{bmatrix} \begin{bmatrix} K_t \\ \ln A_t \end{bmatrix} + \begin{bmatrix} 0 \\ e_{t+1} \end{bmatrix} \equiv \Gamma \begin{bmatrix} K_t \\ \ln A_t \end{bmatrix} + u_{t+1} \tag{25}$$

$$\begin{bmatrix} c_t & n_t & y_t & i_t \end{bmatrix}' = \Pi \begin{bmatrix} K_t \\ \ln A_t \end{bmatrix} \tag{26}$$

K_t *is the capital stock,* A_t *a technological disturbance;* c_t *consumption,* n_t *hours,* y_t *output and* i_t *investments.*

Here Π *and* Γ *are function of* α*, the share of labor in production;* β *the discount factor,* δ *the depreciation rate,* ρ *the AR parameter of the technology shock. Let* $y_{1t} = [c_t, n_t, y_t, i_t]'$ *and* $y_{2t} = [k_t, \ln A_t]'$*,* $\theta = (\alpha, \beta, \delta, \rho)$*.*

A VAR for y_{1t} *only is* $y_{1t} = H(\theta)y_{1t-1} + \epsilon_{1t}$ *where* $H(\theta) = \Pi(\theta)\Gamma(\theta)(\Pi(\theta)'\Pi(\theta))^{-1}\Pi(\theta)$; $\epsilon_{1t} = \Pi(\theta)u_t$ *and* $(\Pi(\theta)'\Pi(\theta)^{-1})\Pi(\theta)$ *is the generalized inverse of* $\Pi(\theta)$.

If $\theta \sim N(\begin{bmatrix} 0.58 \\ 0.988 \\ 0.025 \\ 0.95 \end{bmatrix}, \begin{bmatrix} 0.0006 & & & \\ & 0.0005 & & \\ & & 0.0006 & \\ & & & 0.00015 \end{bmatrix})$, *the model implies that the prior mean for* $H(\theta)$ *is* $\bar{H(\theta)} = \begin{bmatrix} 0.19 & 0.33 & 0.13 & -0.02 \\ 0.45 & 0.67 & 0.29 & -0.10 \\ 0.49 & 1.32 & 0.40 & 0.17 \\ 1.35 & 4.00 & 1.18 & 0.64 \end{bmatrix}$;

(Note substantial feedback from C, Y, N to I in the last row).

The prior variance for $H(\bar{\theta})$ *is* $\Sigma_H = \frac{\partial H}{\partial \theta} \Sigma_\theta \frac{\partial H}{\partial \theta}'$.

- *A Minnesota-style prior for* y_{1t} *consistent with the RBC is*

- *Coefficient on* $y_{1t-1} \sim N(H\bar{(\theta)}, \phi_0 * \Sigma_H)$.

- *Coefficients on* $y_{1t-j} \sim N(0, \frac{\phi_0}{h(\ell)} * \Sigma_H), j > 1$ *where* ϕ_0 *is a tightness parameter and* $h(l)$ *a decay function. Note that here* $\phi_1 = 1$.

- **Move from statistical to economic priors.**

Del Negro and Schorfheide (2004):

- DSGE model provides more than a "form" of the prior restrictions (zero mean on lags greater than one, etc.). It gives quantitative info.

- Exploit the idea that prior is an additional set of equations that can be appended to a model.

- Can make the DSGE prior more or less informative for the VAR depending on how much DSGE data is appended to the actual data.

- Setup a hierarchical model that allows us to compute the posterior of DGSE and VAR parameters jointly.

Idea of the approach:

- Given θ, simulate data from model. Append simulated data to actual data and estimate a VAR on extended data set.

- Estimates of the VAR coefficients and of the covariance matrix will reflect sample and model information. The weight will be given by the precision of the two types of information.

- Precision of data information depends on T (which is fixed). Precision of simulated information depends on T_1, which can be chosen by the investigator. By varying $\kappa = \frac{T_1}{T}$, one can make the prior more or less informative and thus assess of important the model is for the data.

- The model has restrictions. If κ large is optimal it means that the restrictions imposed by the model are not violated. If κ is small, restrictions are violated (test of the model).

• Let $g(\theta) = \prod_{i=1}^{k} g(\theta_k)$ be the prior on DGSE parameters.

• The DSGE model implies a prior $g(\beta|\theta) \sim N(\bar{\beta}(\theta), \bar{\Sigma}_b(\theta))$; $\Sigma_e \sim IW(T_1\bar{\Sigma}(\theta), T_1 - k)$ on the VAR parameters of the decision rule where

$$\begin{aligned} \bar{\beta}(\theta) &= (X^{s'}X^s)^{-1}(X^{s'}y^s) \\ \bar{\Sigma}_b(\theta) &= \Sigma_e(\theta) \otimes (T_1 X^{s'}X^s)^{-1} \\ \bar{\Sigma}(\theta) &= (y^{s'}y^s - (y^{s'}X^s)\bar{\beta}(\theta)) \end{aligned} \tag{27}$$

y^s simulated data, X^s lags in the VAR of simulated data, T_1 =length of simulated data.

Let $\kappa = \frac{T_1}{T}$ control the relative importance of two types of information. $\kappa \to 0$ ($\kappa \to \infty$) actual (simulated) data dominates.

• The VAR implies a density $f(\beta, \Sigma_u|y)$.

The model has a hierarchical structure: $f(\beta, \Sigma_e|y)g(\beta|\theta)g(\Sigma_e|\theta)g(\theta)$. Since likelihood and the prior are conjugate (see the Normal-IW assumption above); the conditional posteriors for VAR parameters are available in analytical format.

• $g(\beta|\theta, y, \Sigma_e) \sim N(\tilde{\beta}(\theta), \tilde{\Sigma}_b(\theta))$; $g(\Sigma_e|\theta, y) \sim iW((\kappa+T)\tilde{\Sigma}(\theta), T+\kappa-k)$ where

$$\begin{aligned} \tilde{\beta}(\theta) &= (T_1 X^{s'} X^s + X'X)^{-1}(T_1 X^{s'} y^s + X'y) \\ \tilde{\Sigma}_b(\theta) &= \Sigma_e(\theta) \otimes (T_1 X^{s'} X^s + X'X)^{-1} \\ \tilde{\Sigma}(\theta) &= \frac{1}{(1+\kappa)T}[(T_1 y^{s'} y^s + y'y) - (T_1 y^{s'} X^s + y'X)\tilde{\alpha}(\theta)] \end{aligned} \quad (28)$$

• If we pick a θ we can immediately construct these posteriors.

• $g(\theta|y) \propto g(\theta) \times |\Sigma_e|^{-0.5(T-M-1)} \exp\{-0.5 tr[\Sigma_e^{-1}(Y - X\beta)'(Y - X\beta)\} \times$

$|\Sigma_e(\theta)|^{-0.5(T_1-M-1)} \exp\{-0.5 tr[\Sigma_e(\theta)^{-1}(Y^s - X^s\beta(\theta))'(Y^s - X^s\beta(\theta))\}.$
This conditional posterior is non-standard: need Metropolis-Hasting step to calculate it.

- Use $g(\theta|y), g(\beta|\theta, y, \Sigma_e), g(\Sigma_e|\theta, y)$ in the Gibbs sampler to obtain a marginal for β.

- All posterior moments in (28) are conditional on κ. How do we select it? i) Use Rules of thumbs (e.g. $\kappa = 1$, T observation added). ii) Maximize the marginal likelihood.

Example 8 *(sticky price model) In a basic NK sticky price-sticky wage economy, set* $\eta = 0.66, \pi^{ss} = 1.005, N^{ss} = 0.33, \frac{c}{gdp} = 0.8, \beta = 0.99, \zeta_p = \zeta_w = 0.75, a_0 = 0, a_1 = 0.5, a_2 = -1.0, a_3 = 0.1$*. Run a VAR with output, interest rates, money and inflation using actual quarterly data from 1973:1 to 1993:4 and data simulated from the model conditional on these parameters. Overall, only a modest amount of simulated data (roughly, 20 data) should be used to set up a DSGE prior.*

ML: Sticky price sticky wage model.

$\kappa = 0$	$\kappa = 0.1$	$\kappa = 0.25$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
-1228.08	*-828.51*	*-693.49*	*-709.13*	*-913.51*	*-1424.61*

6 Structural BVARs

So far we set priors for reduced form VAR parameters. Can we set directly priors for structural VARs?

$$\mathcal{B}_0 y_t - \mathcal{B}(\ell) y_{t-1} = e_t \quad e_t \sim (0, I) \tag{29}$$

$$y_t - B(\ell) y_{t-1} = u_t \quad u_t \sim (0, \Sigma) \tag{30}$$

$\mathcal{B}(\ell) = \mathcal{B}_1 L + \ldots \mathcal{B}_q L^q$; $\mathcal{B}_0$ non singular; $B(\ell) = \mathcal{B}_0^{-1}\mathcal{B}(\ell)$; $\Sigma = \mathcal{B}_0^{-1}\mathcal{B}_0^{-1'}$.

- (29) is a structural system, while (30) is the corresponding VAR.

- Why do we want prior for (29)? We may have a-priori restrictions on the structural dynamics (output responses to a monetary shock have a hump).

- We may have a-priori restrictions on the structural impacts effects of shocks (output responses to a monetary shocks take time to materialize).

What priors do you use for $\mathcal{B}_0$ and $\mathcal{B}(\ell)$? How do you draw from their posterior?

• Standard approach (Canova (1991), Gordon and Leeper (1994)): Use Normal- inverted Wishart prior for reduced form coefficients $(B(\ell), \Sigma)$. This implies a Normal- inverted Wishart posterior. Draw $B(\ell)^l, \Sigma^l)$ and use identification restrictions to draw structural parameters i.e. $\Sigma^l = (\mathcal{B}_0^{-1})^l(\mathcal{B}_0^{-1'})^l$; $\mathcal{B}_j^l = \mathcal{B}_0^l B_j^l$.

- Procedure is OK for just-identified systems. For overidentified systems it does not take into account the extra restrictions.

• Sims and Zha (1998) work directly with the structural model (valid for both just-identified and over-identified systems). Staking the observations in (29):

$$Y\mathcal{B}_0 - X\mathcal{B}_+ = E \tag{31}$$

where Y is a $T \times M$, X is a $T \times k$ matrix of lagged variables; E is a $T \times M$ matrix. Setting $Z = [Y, -X]$; $\mathcal{B} = [\mathcal{B}_0, \mathcal{B}_+]'$, the likelihood is:

$$L(\mathcal{B}|y) \propto |\mathcal{B}_0|^T \exp\{-0.5tr(Z\mathcal{B})'(Z\mathcal{B})\} \propto |\mathcal{B}_0|^T \exp\{-0.5b'(I \otimes Z'Z)b\} \tag{32}$$

where $b = vec(\mathcal{B})$ is a $M(k+M) \times 1$ vector; $b_0 = vec(\mathcal{B}_0)$ is a $M^2 \times 1$ vector; $b_+ = vec(\mathcal{B}_+)$ is a $Mk \times 1$ vector, I a $(Mk \times Mk)$ identity matrix.

Priors:
- $g(b) = g(b_0)g(b_+|b_0)$, where $g(b_0)$ may have singularities (due to zero identification restrictions).
- $g(b_+|b_0) \sim N(h(b_0), \Sigma(b_0))$.

• Make prior on dynamics conditional on prior for contemporaneous effects.

Posterior kernel:

$$\begin{aligned} g(b|y) \quad \propto \quad & g(b_0)|\mathcal{A}_0|^T|\Sigma(b_0)|^{-0.5}\,\text{exp}\{-0.5[b'(I\otimes Z'Z)b\} \\ & exp\{(b_+ - h(b_0))'\Sigma(b_0)^{-1}(b_+ - h(b_0))\} \end{aligned} \tag{33}$$

- Since $b'(I\otimes Z'Z)b = b_0'(I\otimes Y'Y)b_0 + b_+'(I\otimes X'X)b_+ - 2b_+'(I\otimes X'Y)b_0$, conditional on b_0, the quantity in the exponent is quadratic in b_+, thus

- $g(b_+|b_0, y) \sim N(\tilde{b}_0, \tilde{\Sigma}(b_0)^{-1})$ where $\tilde{b}_0 = ((I\otimes X'X) + \Sigma(b_0)^{-1})^{-1}((I\otimes X'Y)b_0 + \Sigma(b_0)^{-1}h(b_0))$; $\tilde{\Sigma}(b_0) = ((I\otimes X'X) + \Sigma(b_0)^{-1})$.

- $g(b_0|y) \propto g(b_0)|B_0|^T|(I\otimes X'X)\Sigma(b_0) + I|^{-0.5}$
$exp\{-0.5[b_0'(I\otimes Y'Y)b_0 + h(b_0)'\Sigma(b_0)^{-1}\mu(b_0) - \tilde{b}_0\tilde{\Sigma}(b_0)\tilde{b}_0\}$

- $g(b_0|y)$ has unknown format!! In addition, $dim(b_+) = M(Mq+1)$ so the calculation of $g(b_+|b_0, y)$ is complicated.

To simplify the computations: Choose $\Sigma(b_0) = \Sigma_1 \otimes \Sigma_2$ and restrict $\Sigma_1 = \varphi * I$.

Then even if $\Sigma_{2i} \neq \Sigma_{2j}$, independence across equations is guaranteed since $(I \otimes X'X) + \Sigma(b_0)^{-1}) \propto (I \otimes X'X) + \text{diag}\{\Sigma_{21}, \ldots \Sigma_{2m}\} = \text{diag}\{\Sigma_{21} + X'X, \ldots \Sigma_{2m} + X'X\}$. This means that we can proceed to estimate the equations one by one, without worrying about simultaneity.

In general, if had we started from VAR then $\tilde{\Sigma}(b_0) = (\Sigma_e \otimes X'X) + \Sigma(b_0)^{-1}$ (correlation across equations).

- Structural Minnesota priors.

Given $\mathcal{B}_0$, let $y_t = B(\ell)y_{t-1} + C + e_t$. Let $\beta = vec[B_1, \ldots B_q, C]$. Since $\beta = [\mathcal{B}_+\mathcal{B}_0^{-1}]$; $E(\beta) = [I_m, 0, \ldots 0]$ and $var(\beta) = \Sigma_b$ imply

$$\begin{aligned} E(\mathcal{B}_+|\mathcal{B}_0) &= [\mathcal{B}_0, 0, \ldots, 0] \\ var(\mathcal{B}_+|\mathcal{B}_0) = \text{diag}(\Sigma_{+(ijl)}) &= \frac{\phi_0\phi_1}{h(\ell)\sigma_j} \quad i, j = 1, \ldots m, \ \ell = 1, \ldots, p \\ &= \phi_0\phi_2 \qquad \text{otherwise} \end{aligned} \tag{34}$$

where i stands for equation, j for variable, ℓ for lag.

(i) No distinction own vs. other coefficients (in SES no normalization with respect to one RHS variable).

(ii) Scale factor differ from reduced form Minnesota prior since $var(v_t) = I$.

(iii) Prior for constant independently parametrized.

(iv) Because $\beta = vec[\mathcal{B}_+\mathcal{B}_0^{-1}]$ there is a-priori correlation in the coefficients across equations (since they depend on the beliefs about $\mathcal{B}_0$). For example, if $\Sigma_{2i} = \Sigma_2 \ \forall i$, $g(\beta|\mathcal{B}_0)$ is normal with covariance matrix $\Sigma_e \otimes \Sigma_2$ (see Kadiyala and Karlsson (1997)).

• Additional restrictions for a structural system:

- Average value of lagged y_i's (say $\bar{y}_{io}$) is a good predictor of y_{it} for each equation. Then $Y_d\mathcal{B}_0 - X_d\mathcal{B}_+ = V$ where $Y_d = \{y_{ij}\} = \phi_3\bar{y}_{0i}$ if $i = j$ and zero otherwise, $i, j = 1, \ldots M$; $X_d = \{x_{is}\} = \phi_3\bar{y}_{0i}$ if $i = j, s < k$ and zero otherwise $i = 1, \ldots M,\ s = 1, \ldots k$.

Note that as $\phi_3 \to \infty$, this restriction implies model in first difference.

- Initial dummy restriction: suppose $Y_{dc}\mathcal{B}_0 - X_{dc}\mathcal{B}_+ = E$ where $Y_{dc} = \{y_j\} = \phi_4\bar{y}_{0j}$ if $j = 1, \ldots M$ $X_{dc} = \{x_s\} = \phi_4\bar{y}_{0j}$ if $s < k - 1$ and $= \phi_4$ if $s = k$.

If $\phi_4 \to \infty$, the dummy observation becomes $[I - A(1)]\bar{y}_0 + \mathcal{A}_0^{-1}C = 0$. If $C \neq 0$, this implies cointegration.

• How do we choose $g(b_0)$.

b_0 contains contemporaneous structural parameters. We need to make a distinction between soft vs. hard restrictions.

- Hard restrictions give you identification (possibly of blocks of equations).

- Select the prior for non-zero coefficients as non-informative i.e. if b_0^n are the non-zero elements of b_0, $g(b_0^n) \propto 1$ or normal.

Example 9 *Suppose* $M(M-1)/2$ *restrictions, e.g.* $\mathcal{B}_0$ *upper triangular. One prior could be to set* $g(\bar{b}_0)$ *to be independent normal with zero mean so* $E(\bar{b}_0(ij)\bar{b}_0(kh)) = 0$ *- no relationship across equations. The variance* $\sigma^2(\bar{b}_0(ij)) = (\frac{\phi_5}{\sigma_i})^2$ *i.e. all the elements of equation* i *have the same variance.*

An alternative would be to use a Wishart prior for Σ_e^{-1}*, i.e.* $g(\Sigma_e^{-1}) \sim IW(\bar{\nu}, \bar{\Sigma})$ *where* $\bar{\nu}$ *are dof and* $\bar{\Sigma}$ *the scale. If* $\bar{\nu} = M+1$, $\bar{\Sigma} = \text{diag}\ (\frac{\phi_5}{\sigma_i})^2$, *then a prior for* $\bar{b}_0$ *is the same as before except for the Jacobian* $|\frac{\partial \Sigma_e^{-1}}{\partial \mathcal{B}_0}| = 2^m \prod_{j=1}^m a_{jj}^j$. *Since likelihood contains a term* $|\mathcal{B}_0|^T = \prod_{j=1}^m b_{jj}^T$, *ignoring the Jacobian makes no difference if* $T >> m$.

How do we draw samples from $g(b_0|y)$? Need MC techniques:

Algorithm 6.1 *[1.] Calculated mode of* $g(b_0|y)$ *and the Hessian at the mode.*

[2.] Draw b_0 *from a normal centered at mode with covariance equal to the Hessian at the mode or a t-distribution with the same mean and covariance and* $\bar{\nu} = M + 1$ *degrees of freedom.*

[3.] Use importance sampling to weight the draw (use ratio $IR_l = \frac{g^{AP}(b_0^l)}{\xi(b_0^l)}$*), and check the magnitude of* IR *over* $l = 1, \ldots, L$.

Alternative: MH algorithm with a normal or a t-distribution as the candidates or a restricted Gibbs (Waggoner-Zha (2003)).

What if the identification restrictions are of non-contemporaneous form? Same idea. Long run restrictions imply special form of $\mathcal{B}_0$. Sign restrictions $\Sigma_e = PDP' = \tilde{P}\tilde{P}'$ and $\mathcal{B}_0 = \tilde{P}^{-1}$.

Extensions:

- VAR with exogenous variables: e.g. Oil prices in a VAR for domestic variables.

- partial VARs (different lag lengths) (special case of 1).

- VAR with block exogenous variables and overidentifying restrictions in some block, e.g. a two country VAR model where one is block exogenous.

General structure for last case

$$\mathcal{B}_i(\ell)y_t = v_{it} \qquad i = 1, \ldots, N \tag{35}$$

i is the number of blocks and $M = \sum_i M_i$ and M_i is the number of equations in each block. e_{it} is a $M_i \times 1$ vector for each i and $\mathcal{B}_i(\ell) = (\mathcal{B}_{i1}(\ell), \ldots, \mathcal{B}_{in}(\ell))$. Let $\mathcal{B}_0 = \text{diag}\{\mathcal{B}_{ii}(0)\}$ and rewrite (35) as

$$\mathcal{B}_0^{-1}\mathcal{B}_i(\ell)y_t = \mathcal{B}_0^{-1}v_{it}$$

or

$$y_{it} = B_i(\ell)y_{it-1} + e_{it} \tag{36}$$

where $B_i(\ell) = (0_-, I_i, 0_+) - \mathcal{B}_{ii}^{-1}(0)\mathcal{B}_i(\ell)$ 0_- is a $M_i \times M_{i_-}$ matrix of zeros of dimension , 0_+ is a $M_i \times M_{i_+}$ matrix of zeros, where $M_{i_-} = 0$ for $i = 1$ and $M_{i_-} = \sum_j^{i-1} M_j$ for $i = 2, \ldots, n$ $M_{i_+} = 0$ for $i = n$ and $M_{i_+} = \sum_{j=i+1}^{n} M_j$ for $i = 1, \ldots, n-1$ and where $E(e_t e_t') = \text{diag}\{\Sigma_{ii}\} = \text{diag}\{\mathcal{B}_{ii}(0)^{-1}\mathcal{B}_{ii}(0)^{-1'}\}$.

Stack all the observations to have

$$y_i = z_i b_i + E_i \tag{37}$$

where y_i and v_i are $T \times m_i$ matrices, x_i is a $T \times k$ matrix and k is the number of coefficients in each block, $b_i = vec(B_i)$, $z_i = (I \otimes X_i)$ The likelihood is

$$\begin{aligned} L(b_i|y_{-q} \ldots, y_0, y_1, \ldots, y_T) &\propto \prod_{i=1}^{n} |\mathcal{B}_{ii}(0)|^T \exp\{-0.5tr[(S_i(b_i)\mathcal{B}_{ii}(0)'\mathcal{B}_{ii}(0)]\} \\ &\propto \prod_{i=1}^{n} |\mathcal{B}_{ii}(0)|^T \exp\{-0.5tr[(S_i(\hat{b}_i)\mathcal{B}_{ii}(0)'\mathcal{B}_{ii}(0) \\ &+ (b_i - \hat{b}_i)'x_i'x_i(b_i - \hat{b}_i)\mathcal{B}_{ii}(0)'\mathcal{B}_{ii}(0)]\} \end{aligned} \quad (38)$$

where $\hat{b}_i = (z_i'z_i)^{-1}(z_i'y_i)$ and $S_i(b_i) = (y_i - z_ib_i)'(y_i - z_ib_i)$.

Suppose $g(\mathcal{B}_{ii}(0), b_i) \propto |\mathcal{B}_{ii}(0)|^k$. The posterior are

$$g(\mathcal{B}_{ii}(0)|y) \propto |\mathcal{B}_{ii}(0)|^T \exp\{-0.5tr[(S_i(\hat{b}_i)\mathcal{B}_{ii}(0)'\mathcal{B}_{ii}(0)]\} \quad (39)$$

$$g(b|\mathcal{B}_{ii}(0), y) \sim N(\hat{b}, (\mathcal{B}_{ii}(0)'\mathcal{B}_{ii}(0))^{-1} \otimes (X_i'X_i)^{-1}) \quad (40)$$

where $\hat{b}_i = vec(\hat{B}_i)$. To draw posterior sequences for $b, \mathcal{B}_{ii}(0)$ need to distinguish whether the system is just or over-identified.

i) If just-identified (e.g. $\mathcal{B}_{ii}(0)$ is the Choleski factor of Σ_{ii}), then there is one-to-one mapping between the prior and the posterior of $\mathcal{B}_{ii}(0)$ and for Σ_{ii}. So we can draw from (39) or from the posterior of Σ_{ii} and use identification restrictions to get a draw for $\mathcal{B}_{ii}(0)$.

ii) If $\mathcal{B}_{ii}(0)$ is overidentified, then we need to draw $\mathcal{B}_{ii}(0)^l,\ l = 1, \ldots J$ from the marginal posterior directly and use, e.g.

Algorithm 6.2 *[1.] Draw* $\mathcal{B}_{ii}(0)^l,\ l = 1, \ldots J$ *from* $N(B *_{ii} (0), H_{B^*_{ii}})$ *where* $B^*_{ii}(0)$ *is the mode of the posterior and* $H_{(B^*_{ii})}$ *Hessian at the mode.*

[2.] Draw b^l *from* $g(b_i|\mathcal{B}_{ii}(0)^l, y)$ *and calculate* $B_i(\ell)^l = \mathcal{B}_{ii}(0)^l(0_-, I_i, 0_+) - \mathcal{B}_i(\ell)^l$.

[3.] Calculate any function $h(B_i(\ell)^l)$, *weighting the* $B_i(\ell)^l$ *with the importance ratio* $IR^l = \frac{g(B_i(\ell)^l)}{N(B_i(\ell)^l}$).

An importance sampling algorithm can be inefficient when the degree of simultaneity is high and the number of degrees of freedom is small since the posterior of the parameters can be far from a normal (or even a t-distribution).

- Alternative 1: Use the Gibbs sampler of Waggoner and Zha (2003).

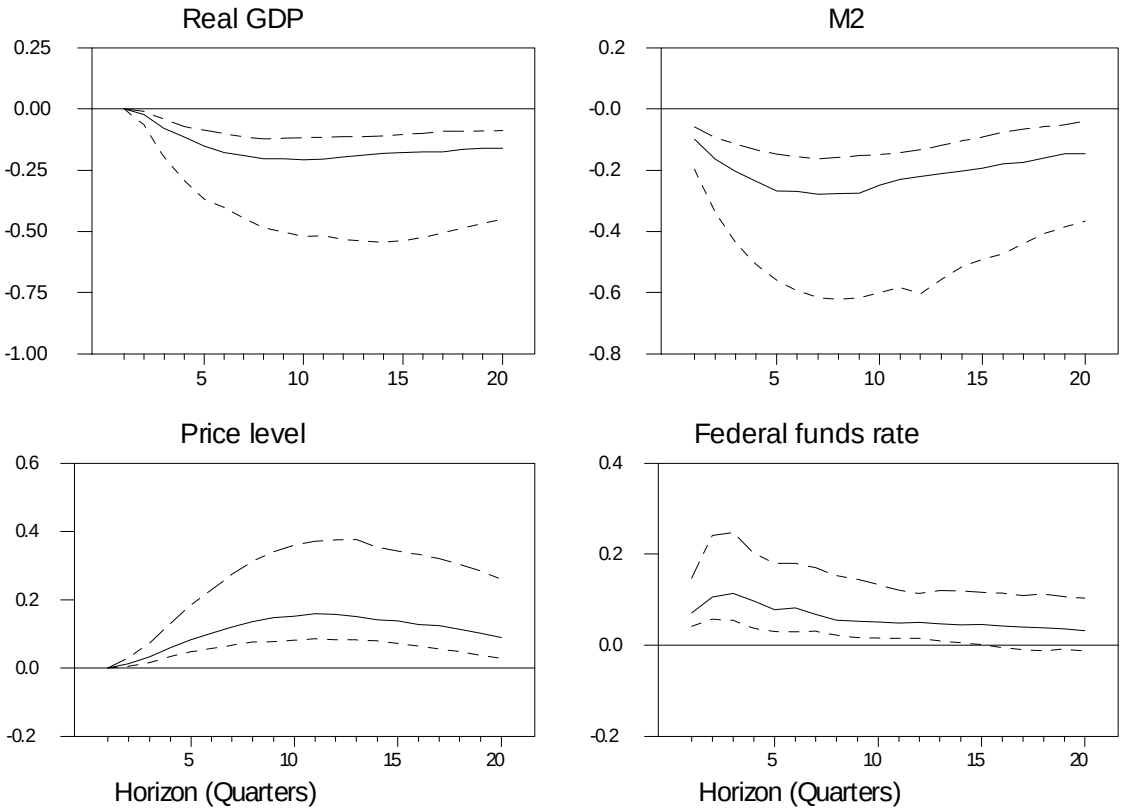
- Alternative 2: Use a Metropolis algorithm to draw $\mathcal{B}_{ii}(0)^l,\ l = 1, \ldots J$. That is, rather than step [1.] of the previous algorithm use the following:

[1'.] Draw $\mathcal{B}_{ii}(0)^l,\ l = 1, \ldots J$ from $\mathcal{B}_{ii}(0)^\dagger = \mathcal{B}_{ii}(0)^{l-1} + u_{ii}$ where var$(u_{ii}) \propto H_{(B^*_{ii})}$ and accept it if $\frac{g(\mathcal{B}_{ii}(0)^\dagger|y)}{g(\mathcal{B}_{ii}(0)^{l-1}|y)} > U(0,1)$ random variable. Otherwise set $\mathcal{B}_{ii}(0)^l = \mathcal{B}_{ii}(0)^{l-1}$.

Example 10 *(Transmission of monetary shocks) Use US data, 1960:1 to 2003:1 for the log IP, log of CPI, Fed Funds rate and the log of M2.*

Overidentify the system: the central bank only looks at money when manipulating the nominal rate, i.e. contemporaneous impact matrix is Choleski form except (3,1), (3,2) elements which are zero.

$g(\bar{b}_i) \sim N(0,1)$. *Use an importance sampling to draw from a normal centered at the mode and with dispersion equal to Hessian at the mode. Importance ratio: in 17 out of 1000 draws weight is large.*


Real GDP
0.25
0.00
-0.25
-0.50
-0.75
-1.00
5
10
15
20
M2
0.2
-0.0
-0.2
-0.4
-0.6
-0.8
5
10
15
20
Price level
0.6
0.4
0.2
0.0
-0.2
5
10
15
20
Horizon (Quarters)
Federal funds rate
0.4
0.2
0.0
-0.2
5
10
15
20
Horizon (Quarters)

• Both output and money persistently decline in response to an increase in interest rates. The response of price initially close to zero but turns positive and significant after about 5 months (price puzzle?).

• Monetary shocks explain 4-18% of var(Y) at the 48 month horizon and 0-7% of var(P).

7 Bayesian panel data analysis

7.1 Univariate dynamic panels

$$y_{it} = \varrho_i + B_{1i}(\ell)y_{it-1} + B_{2i}(\ell)x_t + e_{it} \quad e_{it} \sim (0, \sigma_i^2) \tag{41}$$

$B_{ji}(\ell) = B_{ji1}\ell + B_{ji2}\ell^2 + \ldots B_{jiq_j}\ell^{q_j}$, B_{jil} is a scalar, ϱ_i is the unit specific fixed effect, x_t are exogenous variables, common to all units. Assume $E(e_{it}e_{j\tau}) = 0 \; \forall i \neq j, \; \forall t, \tau$.

- Interesting quantities that can be computed: $b_{i1}(1) = (1 - B_{1i}(1))^{-1}$, $b_{i2}(1) = (1 - B_{1i}(1))^{-1}B_{2i}(1)$ (long run effects), $b_{i1}(\ell)$, $b_{i2}(\ell)$ (impulse responses).

- Stack the T observations to create $y_i, x, e_i, \mathbf{1}$. Let $X_i = (y_i, x, \mathbf{1})$, $X = diag\{X_i\}$, $\beta = [B_1, \ldots B_N]'$; $B_i = (\varrho_i, B_{1i1}, \ldots, B_{1iq_1}, B_{2i1}, \ldots, B_{2iq_2})$, $\Sigma_i = \sigma_i^2 * I_T, \Sigma = diag\{\Sigma_i\}$ then:

$$y = X\beta + e \quad e \sim (0, \Sigma) \tag{42}$$

$y = (y_1', \ldots y_N')', e = (e_1', \ldots e_N')'$.

- Comparing (42) with (1) one can see dynamic panel has same structure as a VAR but X_i are unit specific and the covariance matrix has a (block) heteroschedastic structure.

- Likelihood is of (42) is still the product of a normal for β, conditional on Σ, and N inverted gammas for σ_i^2. Note that since $var(e)$ is diagonal, ML=OLS equation by equation.

What kind of priors could be used?

- Semi-conjugate prior: $g(\beta) \sim N(\bar{\beta}, \bar{\Sigma}_b)$ and $g(\sigma_i^2) \sim IG(0.5a_1, 0.5a_2)$.

- Exchangeable prior: $g(\beta) = \prod_i g_i(\beta)$; $\beta_i \sim N(\bar{\beta}, \sigma_b)$, where σ_b measures a-priori heterogeneity. With exchangeability $\tilde{\beta}$ can be computed equation by equation.

- Exchangeable prior on the difference (Canova and Marcet (1998)): $\beta_i - \beta_j \sim N(0, \Sigma_b)$. Σ_b has a special structure.

- Depending on the choice of prior, the posterior will reflect prior and sample or prior and pooled info (see Zellner and Hong, 1989).

Example 11 *(Growth and convergence)*

$$Y_{it} = \varrho_i + B_i Y_{it-1} + e_{it} \qquad e_{it} \sim N(0, \sigma_i^2) \tag{43}$$

where $Y_{it} = \log(\frac{y_{it}}{y_t})$, y_t *is the average EU GDP.*

Let $\beta_i = (\varrho_i, B_i) = \bar{\beta} + v_i$, *where* $v_i \sim N(0, \sigma_b^2)$. *Assume* σ_i^2 *given,* $\bar{\beta}$ *known (if not get it from pooled regression on* $(-\tau, 0)$*) and treat* σ_b^2 *fixed.*

Let $\kappa_{i,j} = \frac{\sigma_i^2}{\sigma_{b_{jj}}^2}$, $j = 1, 2$ *be the relative importance of prior and sample information. Choose loose prior (*$\kappa_{i,j} = 0.5$*).*

Use income per-capita for 144 EU regions from 1980 to 1996 to construct $SS_i = \tilde{\varrho}_i \frac{1-\tilde{B}_i^T}{1-\tilde{B}_i} + \tilde{B}_i^{T+1} z_{i0}$ *where* $\tilde{\varrho}_i, \tilde{B}_i$ *are posterior mean, and* $CV_i = 1 - \tilde{B}_i$ *(the convergence rate).*

-Mode of CV_i distribution is 0.09: fast catch up. The highest 95% credible set is large (from 0.03 to 0.45).

- Distribution of SS has many modes (at least 2).

What can we say about the posterior of the cross sectional mean SS? Suppose $g(SS_i) \sim N(\mu, \zeta^2)$. Assume $g(\mu) \propto 1$ and $\zeta = 0.4$.

- $g(\mu|y)$ combines the prior and the data and the posterior of $g(SS_i|y)$ combines unit specific and pooled information.

- $\tilde{\mu} = -0.14$ (highly left skewed distribution); variance is 0.083; 95 percent credible interval is (-0.30, 0.02).

7.2 Endogenous grouping

• Are there groups in the cross section? Convergence clubs; credit constrained vs non-credit constrained consumers, large vs. small firms, etc. Classifications typically exogenous (see e.g., Gertler and Gilchrist (1991)).

- Want an approach that simultaneously allows for endogenous cross sectional grouping and Bayesian estimation of the parameters.

- Idea: if units i and j belong to a group, coefficients α_i and α_j have same distribution. If not, they have different distributions.

- Basic problem: what ordering of the cross section gives grouping? There are $\wp = 1, 2, \ldots N!$ orderings. How do you find groups?

• Suppose $\varsigma = 1, 2, \ldots, \bar{\varsigma}$ breaks, $\bar{\varsigma}$ given. For each $\varsigma + 1$ groups let the model be:

$$y_{it} = \varrho_i + B_{1i}(\ell) y_{it-1} + A_{2i}(\ell) x_{t-1} + e_{it} \quad (44)$$

$$\beta_i^j = \bar{\beta}^j + v^j \quad (45)$$

where $i = 1, \ldots, n^j(\wp)$; $n^j(\wp)$ is the number of units in group j, given the $\wp$-th ordering, $\sum_j n^j(\wp) = N$, each $\wp$ and $e_{it} \sim (0, \sigma^2_{e_i})$, $v^j \sim (0, \bar{\Sigma}_j)$ $\beta_i = [\varrho_i, B_{1i1}, \ldots, B_{1iq_1}, B_{2i1}, \ldots, , B_{2iq_2}]$. Let $h^j(\wp)$ be the location of the break for group $j = 1, \ldots, \varsigma + 1$.

Alternative to (45): $\bar{\varsigma} = 0$ and exchangeable structure $\forall i$, i.e

$$\beta_i = \bar{\beta} + v_i \qquad i = 1, \ldots, N \quad v_i \sim N(0, \bar{\Sigma}_i) \quad (46)$$

Want to evaluate (44)-(45) against (44)-(46) and estimate (β, σ_{e_i}) jointly with optimal $(\wp, \varsigma, h^j(\wp))$ (ordering, number of breaks, location of break).

- Given an ordering $\wp$, the number of breaks ς, and the location of the break point $h^j(\wp)$, rewrite (44) $-$ (45) as:

$$Y = X\beta + E \qquad E \sim (0, \Sigma_e) \tag{47}$$

$$\beta = \Xi\beta_0 + V \qquad V \sim (0, \Sigma_V) \tag{48}$$

where Σ_E is $(NTM) \times (NTM)$ and $\Sigma_V = diag\{\Sigma_i\}$ is $(Nk) \times (Nk)$.

- Specify priors for $(\beta_0, \Sigma_e, \Sigma_V)$. Construct posterior estimates for (β, Σ_E), (β_0, Σ_V) jointly with posterior estimates of $(\wp,\ \varsigma\ ,\ h^j(\wp))$. Problem complicated!

- Split the problem in three steps. Use Empirical Bayes techniques to construct posteriors estimates of β, conditional on optimal $(\wp,\ \varsigma\ ,\ h^j(\wp))$ and estimates of $(\beta_0, \Sigma_V, \Sigma_E)$.

• Step 1: How do you compute $\wp, \varsigma, h^j(\wp)$ optimally?

a) Given $(\beta_0,\ \mathbf{\Sigma}_V,\ \mathbf{\Sigma}_e)$, and a $\wp$, examine how many groups are present (select ς).

b) Given $\wp$ and $\hat{\varsigma}$, check for the location of the break points (select $h^j(\wp)$).

c) Iterate on the first two steps, altering $\wp$.

Conclusion: selected submodel maximizes the predictive density over orderings $\wp$, groups $\varsigma + 1$ and break points $h^j(\wp)$.

Let: $f(Y|H_0)$ be the predictive density under cross sectional homogeneity.

Let $f(Y|H_\varsigma; h^j(\wp), \wp) = \prod_{j=1}^{\varsigma+1} f(Y^j|H_\varsigma, h^j(\wp), \wp)$ the predictive density for group j, with ς break points at location $h^j(\wp)$, using ordering $\wp$.

Define: - I^ς: set of possible break points when there are ς groups

- J: set of possible orderings of the cross section.

- $\pi_h^j(\wp)$: (diffuse) prior of a break at location h for group j of ordering $\wp$.

- $f^-(Y|H_\varsigma, \wp) \equiv \sup_{h^j(\wp)\in I^\varsigma} f(Y|H_\varsigma, h^j(\wp), \wp)$ (max w.r. to break)

- $f^\dagger(Y|H_\varsigma) \equiv \sup_{\wp\in J} f^-(Y|H_\varsigma, \wp)$ (max w.r. to break and ordering)

- $f^0(Y|H_\varsigma, \wp) \equiv \sum_{h^j(\wp)\in I^\varsigma} \pi_h^j(\wp) f(Y|H_\varsigma, h^j(\wp), \wp)$ (average).

To test for breaks (set $\bar{\varsigma} << (N/2)^{0.5}$).

1) Given $\wp, H(0)$ no breaks, $H(1)$ ς breaks:

$$PO(\wp) = \frac{\pi_0 f^0(Y|H_0)}{\sum_\varsigma \pi_\varsigma f^0(Y|H_\varsigma, \wp)} \tag{49}$$

π_0 (π_ς) the prior probability that there are 0 (ς) breaks.

2) Given $\wp, H_0 : \varsigma - 1$ breaks $H(1) : \varsigma$ breaks.

$$PO(\wp, \varsigma - 1) = \frac{\pi_{\varsigma-1} f^{0(\varsigma-1)}(Y|H_{\varsigma-1}, \wp)}{\pi_\varsigma f^{0(\varsigma)}(Y|H_\varsigma, \wp)} \tag{50}$$

Given ς, assign units to j i.e find $f^-(Y|H_{\bar{\varsigma}}, \wp)$. Alter $\wp$ to get $f^\dagger(Y|H_{\bar{\varsigma}})$.

Questions:

i) Can we proceed sequentially to test for (cross sectional) breaks? Bai (1997) OK consistent. But estimated break point is consistent for *any* of the existing break points, location depends on the "strength" of the break.

ii) How to maximize predictive density over $\wp$ when N is large? Do we need N permutations? No, much less. Plus use economic theory to give you interesting ordering.

• Step 2: Given $(\wp,\ \varsigma,\ h^j(\wp))$ estimate $[\beta_0', vech(\Sigma_V)', vech(\Sigma_e)']'$ using $f^\dagger$ on a training sample.

If the e's are normally distributed, then

$$\begin{aligned}
\hat{\beta}_0^j &= \frac{1}{n^j(\wp)} \sum_{i=1}^{n^j(\wp)} \beta_{ols}^i \\
\hat{\Sigma}_j &= \frac{1}{n^j(\wp)-1} \sum_{i=1}^{n^j(\wp)} (\beta_{ols}^i - \hat{\beta}^i)(\beta_{ols}^i - \hat{\beta}^i)' - \frac{1}{n^j(\wp)} \sum_{i=1}^{n^j(m)} (X_i X_i')^{-1} \hat{\sigma}_i^2 \\
\hat{\sigma}_i^2 &= \frac{1}{T-k}(Y_i' Y_i - Y_i' X_i \beta_{ols}^i) \qquad (51)
\end{aligned}$$

$j = 1, \ldots, \varsigma + 1$; x_i regressors and y_i dependent variables for unit i of group j , and $\beta_{ols}^j = (x^{j'} x^j)^{-1}(x^{j'} y^j)$ is the OLS estimator for unit i (in group j).

• Step 3: Construct posterior estimates of β conditional on all other parameters.

- EB posterior point estimate $\hat{\beta} = (X'\hat{\mathbf{\Sigma}}_E^{-1}X + \hat{\mathbf{\Sigma}}_V^{-1})^{-1}(X'\hat{\mathbf{\Sigma}}_E^{-1}Y + \hat{\mathbf{\Sigma}}_V^{-1}A\hat{\beta}_0)$.

- Alternatively, joint estimation prior and posterior if e's and the v's are normal and the prior on hyperparameters diffuse (see Smith 1973).

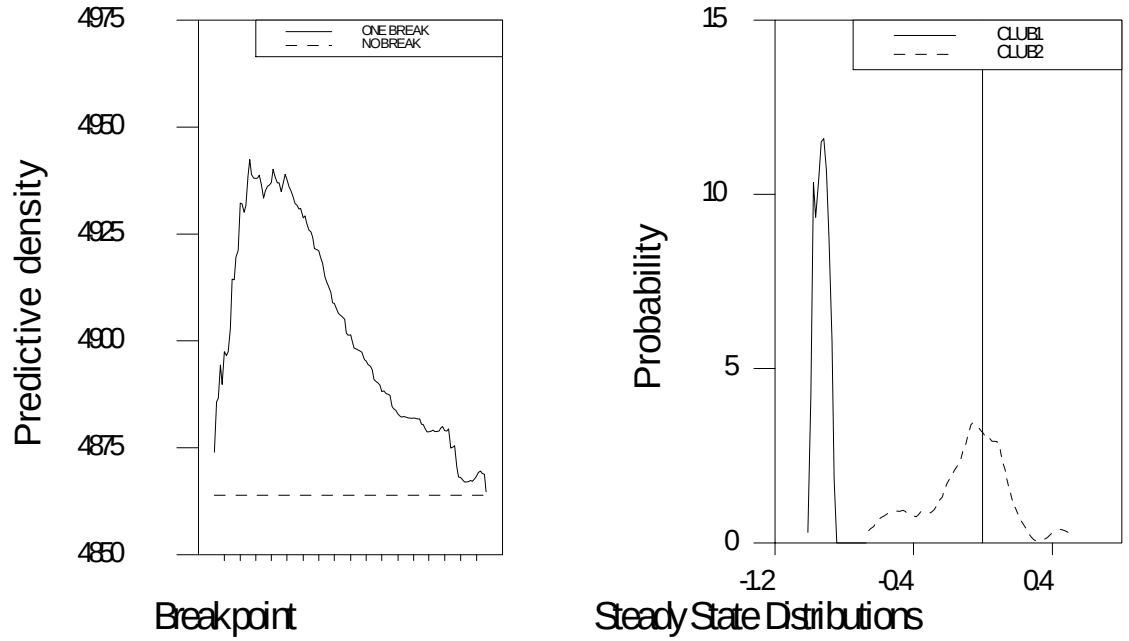
Example 12 *(Convergence clubs). The posterior in example 11 is multimodal. Are there at least two convergence clubs? Where is the break point? How different are convergence rates across groups?*

- Examine several ordering. More or less they give the same result. Best use initial conditions of relative income per-capita.

- Set $\hat{\varsigma} = 4$ *and sequentially examine* ς *against* $\varsigma + 1$ *breaks starting from* $\varsigma = 0$*. Three breaks, PO ratios of 0.06. 0.52, 0.66 respectively. Evidence in favour of two groups.*

- Figure reports the predictive density as a function of the break point (for $(\varsigma = 1)$ *and* $(\varsigma = 0)$*. Units up to 23 (poor, Mediterranean and peripheral regions in the EU) belong to the first group and from 24 to 144 to the second.*

The average CV of two groups are 0.78 and 0.20: faster convergence to below average steady state in the first group. Posterior distributions of the steady states for the two groups distinct.



8 Bayesian pooling

- Often in cross country studies we have only a few data points for a moderate number of countries.

- If dynamic heterogeneities are suspected, exact pooling of cross sectional information will lead to biases and inconsistencies.

- Any way to do some partial cross sectional pooling to improve over single unit estimators?

- How do you compute "average" effects in dynamic models which are heterogeneous in the cross section?

• Simple univariate model to set up ideas:

$$y_{it} = X_{it}\beta_i + e_{it} \qquad e_{it} \sim iid\ (0, \sigma^2 I) \tag{52}$$

where $X_{it} = [1, y_{it-1}, \ldots, y_{it-p}], \beta_i = [a_{i0}, A_{i1}, A_{i2}, \ldots, A_{ip}]$. Assume that T is short. Suppose

$$\beta_i = \bar{\beta} + v_i$$

where $v_i \sim (0, \Sigma_v)$.

• Coefficient of the dynamic model are drawn from the same distribution (they are different realizations of the same process).

• Σ_v controls degree of dispersion. $\Sigma_v = 0$ coefficients equal; $\Sigma_v \to \infty$ no relationship between the coefficients.

- Two interpretations of (**??**): i) uncertain linear restriction (classical approach); ii) prior which shrink coefficients of unit i and j toward a common mean.

• Bayesian Random coefficient estimator.

If e_i and v_i are normal, $\bar{\beta}$ and Σ_v known, $g(\beta_i|y)$ is normal with mean

$$(\frac{1}{\sigma_i^2}x_i'x_i + \Sigma_v^{-1})^{-1}(\frac{1}{\sigma_i^2}x_i'x_i\beta_{i,ols} + \Sigma_v^{-1}\bar{\beta})$$

where $\beta_{i,ols}$ is the OLS estimator of β_i and variance

$$(\frac{1}{\sigma_i^2}x_i'x_i + \Sigma_v^{-1})^{-1}$$

• Weighted mean of prior and sample information with weights given by the relative precision of the two informations!!

- Use $\sigma^2_{i,ols}$ in the formulas.

- If Σ_v is large, $\tilde{\beta}_i \rightarrow \beta_{i,ols}$.

• $\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\beta}_i = \beta_{GLS}$ applied to the uncertain linear model using Theil mixed estimator.

- If $\bar{\beta}$, σ_i^2, Σ_v are unknown, need a prior for these parameters. No analytical solution for the posterior mean of β_i exists.

- Approximate posterior modal estimates (see Smith (1973))

$$\bar{\beta}^* = \frac{1}{n}\sum_{i=1}^{n}\beta_i^* \tag{53}$$

$$(\sigma_i^*)^2 = \frac{1}{T+2}[(y_i - x_i\alpha_i^*)'(y_i - x_i\alpha_i^*)] \tag{54}$$

$$\Sigma_v^* = \frac{1}{n - \dim(\alpha) - 1}[\sum_i(\beta_i^* - \bar{\beta}^*)(\beta_i^* - \bar{\beta}^*) + \kappa \tag{55}$$

where "*" are modal estimates from a training sample, $\kappa = \text{diag}[0.001]$.

- Plug in these estimates in the posterior mean/variance formulas. Underestimate uncertainty (parameters treated as fixed when they are random).

- Two step estimator.

- Alternative estimator (see Rao (1975)) using a training sample:

$$\bar{\beta}_{EB} = \frac{1}{n}\sum_{i=1}^{n}\beta_{i,ols} \tag{56}$$

$$\sigma^2_{i,EB} = \frac{1}{T - dim(\alpha)}(y_i'y_i - y_i'x_i\beta_{i,ols}) \tag{57}$$

$$\hat{\Sigma}_{v,EB} = \frac{1}{n-1}\sum_{i=1}^{n}(\beta_{i,ols} - \bar{\beta}_{EB})(\beta_i - \bar{\beta}_{EB})' - \frac{1}{n}\sum_{i=1}^{n}(x_i'x_i)^{-1}\sigma^2_{i,ols} \tag{58}$$

• The two estimators of $\bar{\beta}$ are similar, but the first averages posterior modes, the second averages OLS estimates.

Can use the procedure to partially pool subsets of the cross sectional units. Assume (**??**) within each subset but not across subsets.

8.1 Bayesian pooling for VARs

- Can maintain same setup and same ideas. Approach is the same.

$$y_{it} = (I \otimes X_t)\beta_i + e_{it} \qquad e_{it} \sim iid\ (0, \Sigma_e) \tag{59}$$

where $X_t = [1, y_{t-1}, \ldots, y_{it-p}], \beta_i = [a_{i0}, A_{i1}, A_{i2}, \ldots, A_{ip}]$. Suppose

$$\beta_i \;=\; \bar{\beta} + v_i \quad v_i \sim (0, \Sigma_v) \tag{60}$$

Case 1: $\bar{\beta}, \Sigma_v$ known. Posterior for α_i is normal with mean and variance given by

$$\tilde{\beta}_i \;=\; (\frac{1}{\sigma_i^2} x_i' x_i + \Sigma_v^{-1})^{-1} (\frac{1}{\sigma_i^2} x_i' x_i \beta_{i,ols} + \Sigma_v^{-1} \bar{\beta}) \tag{61}$$

$$\tilde{\Sigma}_\alpha \;=\; (\frac{1}{\sigma_i^2} x_i' x_i + \Sigma_v^{-1})^{-1} \tag{62}$$

Case 2:$\bar{\beta}, \Sigma_v$ unknown fixed quantities estimable on a training sample.

- (61)-(62) still applicable with estimates of $\bar{\beta}, \Sigma_v$ in place of true ones.

Case 3: $\bar{\beta}, \Sigma_v$ unknown random quantities with prior distribution.

- Use MCMC to derive posterior marginals of the parameters.

- Cross sectional prior can be used in addition or in alternative to time series prior. Both have shrinkage features.

- Same logic can be applied if one expects impulse responses (rather than VAR coefficients) to be similar. Model in this case is

$$y_{it} = \sum_j \gamma_{ij} e_{it-j} \quad e_{it} \sim iid\,(0, \Sigma_e) \tag{63}$$

$$\gamma_i = \bar{\gamma} + v_i \tag{64}$$

where $v_i \sim (0, \Sigma_v)$ and $\gamma_i = [\gamma_{i1}, \gamma_{i2}, \ldots]$.

- Posterior distribution of impulse responses will reflect unit specific (sample) information and prior information. Weights will depend on the relative precision of the two information.

- Note that we treat Σ_e as fixed (known or estimable quantity). If it is a random variable we need to use some conjugate format to derive analytically the posterior; otherwise we need to use MCMC methods.

DSGE Models: Evaluation and forecasting

Fabio Canova
EUI and CEPR
October 2012

Outline

- Principles of policy evaluation
- In-sample evaluation
- Out-of-sample evaluation
- Evaluation via other features (internal propagation, co-cycles, cointegration).
- Evaluation via VARs.
- Evaluation via loss functions.

References

Adolfson, M., Laseen, S., Linde, J. and Villani, M., (2008), Evaluating an estimated new Keynesian small open economy model, Journal of Economic Dynamics and Control, 32, 2690-2721.

Box, G. (1980), Sampling and Bayes' inference in scientific modelling and robustness, Journal of the Royal Statistical Society, Seria A, 143, 383-430.

Brock, W. Durlauf, S. and West, K (2003), Policy analyses in an uncertain economic environment, *Brookings Papers on economic activity*, 1, 235-322.

Canova, F. and M. Paustian (2011), Business Cycle measurement with some theory, *Journal of Monetary Economics*, 48, 345-361.

Canova, F. and Ortega, E. (2000), "Testing Calibrated General Equilibrium Models", in Mariano, R., T. Shuermann and M. Weeks (eds.) *Inference using Simulation Techniques*, Cambridge University Press.

Canova, F., Finn, M. and Pagan, A. (1994), "Evaluating a Real Business Cycle Model", in C. Hargreaves (ed.), *Nonstationary Time Series Analyses and Cointegration*, Oxford, UK: Oxford University Press.

Del Negro, M. and Schorfheide, F. (2005), "Policy predictions if the model does not fit", *Journal of the European Economic Association*.

Del Negro, M., Schorfheide, F., Smets, F. and Wouters, R. (2006), On the fit of New-keynesian models, *Journal of Business and Economic Statistics*, 25, 143-162.

Ferroni, F. (2011), Trend agnostic, one step estimation of DSGE models, *The BE Journal of Macroeconomics*, volume 1, issue 1 (advances), article 25.

Faust, J. and Gupta, A. (2012) Posterior predictive analyses for evaluating DSGE models, NBER working paper 17906.

Favero, C. (2006), Model evaluation in Macroeconomics from Cowles foundation to DSGE models, IGIER manuscript.

Kapetanios, G., Pagan, A. and Scott, A. (2007), "Making a match: Combining theory and evidence in a policy-oriented macroeconomic modeling", *Journal of Econometrics*, 126, 565-594.

Kydland, F. and Prescott, E. (1996), "The Computational Experiment: An Econometric Tool", *Journal of Economic Perspective*, 10, 69-85

Pagan, A. (2003), Report on Modelling and forecasting at the Bank of England,*Bank of England Quarterly Bulletin*, Spring, 1-29.

Pappa, P. (2009), The effects of government shocks on employment and real wages, *International Economic Review*, 50, 217-244.

Schorfheide, F. (2000), Loss function based evaluation of DSGE models, *Journal of Applied Econometrics*. 15, 645-670.

Sims, C. (1996), "Macroeconomics and Methodology", *Journal of Economic Perspectives*, 10, 105-120.

Sims, C. (2002), "The Role of Models and Probabilities in the Monetary Policy Process, *Brookings Papers on Economic Activity*, 2, 1-40.

Zellner, A. (2007), Philosophy and objectives of econometrics, *Journal of Econometrics*, 136, 331-339.

1 Principles of policy evaluation: Theory

$$x = m(p, \beta_m, \eta) \tag{1}$$

$p =$ policy variable, $m =$ model, $\beta_m =$ parameters of model m, η random errors.

• Case 1: Model, parameters, errors are known.

- Set up a loss function $L(x)$.

- Find the p which minimizes the loss function.

Unrealistic setup!!

• Case 2: η unknown, but its pdf μ_η is available.

- $L(x)$ is a random variable and, for each p, the loss function has a distribution.

- Evaluate $\mu(L(x)|p, m, \beta_m)$ (the pdf of the loss function) or $L(\mu((x)|p, m, \beta_m))$ (the loss of the pdf of x), e.g. find the p which gives the $\mu(L(x))$ with the lower variability.

- Still unrealistic setup.

• Case 3: η, β_m unknown.

a) Evaluate policies using $\mu(L(x)|p, m, \hat{\beta}_m)$, where the data d is used estimate β_m.

b) Parameter averaging: use $\mu(L(x)|p, m) = \int \mu(L(x)|p, m, \beta_m) d\mu(\beta_m|d)$ where $d\mu(\beta_m|d)$ is the posterior distribution of β_m, conditional on the data d.

- Parameter uncertainty is small if T is sufficiently large.

• Case 4: η, β_m and m all unknown. Why is m unknown?

i) Unclear which economic theory one should use.

ii) Different functional forms can represent the same theory.

What can you do in this situation?

a) Use model selection criteria (AIC, BIC, etc.), i.e. use $\mu(L(x)|p, \hat{m}, \hat{\beta}_m)$ where $\hat{\beta}_m$ is chosen after $\hat{m}$ is selected.

Problems: (i) data mining; (ii) pre-testing matters (artificially small s.e.); (iii) a model could be good according to the chosen selection criteria but may have low or zero posterior probability.

b) Do model averaging ($m = 1, \ldots, M$), i.e. use $\mu(L(x)|p) = \sum_m \int \mu(L(x)|p, m, \beta_m) d\mu(\beta_m|m)\mu(m|d)$, where $\mu(m|d)$ is the posterior of model m, given the data.

In standard exercises one computes;
$E(L(x)|p, m, \hat{\beta}_m) = \int L(x)\mu(x|p, m, \hat{\beta}_m)dx$.

With model averaging (but not parameter averaging) one computes:
$E(L(x)|p, \hat{\beta}_m) = \int L(x)\mu(x|p, \hat{\beta}_m)dx$.

Example 1 *Suppose we care about the long run effect of a policy choice on the variability of* x*. One can compute:*

$$\begin{aligned} var(x_\infty|p, m, \beta_m) \\ var(x_\infty|p, m) \\ var(x_\infty|p) \end{aligned} \tag{2}$$

The latter is the effect of policy on long run variability of x *without assuming that the model selection exercise has identified the correct one.*

- Averaging is theoretically OK, but policymakers mainly interested in knowing whether and how alternative assumptions about policy affect loss function. (Calculating expected loss may not be that interesting for them). Alternatives statistics:

1) *Outcome Dispersion*

$$L(x|m_1, p) - L(x|m_2, p) \tag{3}$$

2) *Action Dispersion*

$$L(x|m_1, p(m_1)) - L(x|m_2, p(m_2)) \tag{4}$$

In 2) can ask: does conditioning the policy on a particular model changes the outcome of the experiment? By how much?

In practice it is common to proceed as in case 3. Policymakers informally do model averaging.

How do we use these principles for estimation purposes?

For estimation purposes, p is given (e.g. Taylor rule). Interested in measuring model fit. Can use the same ideas:

i) Set up a loss function.

ii) Condition on a model-parameter estimate.

iii) Measure discrepancy.

or

ii') Have an array of models and/or an array of potential estimates.

iii') Average over models-estimates and measure discrepancy.

2 DSGE Model evaluation

• Statistical vs. economic evaluation?

- Cowles: Evaluation = test of overidentifying (statistical) restrictions.

- Calibration: Evaluation = informal distance of moments with economic interpretation.

- DSGE-Bayesian: match conditional dynamics, measure credibility of models restrictions. Both statistical and economic evaluation are possible.

- Central Bank models: what criteria do they use?

2.1 In-sample evaluation

2.2 Graphical evaluation

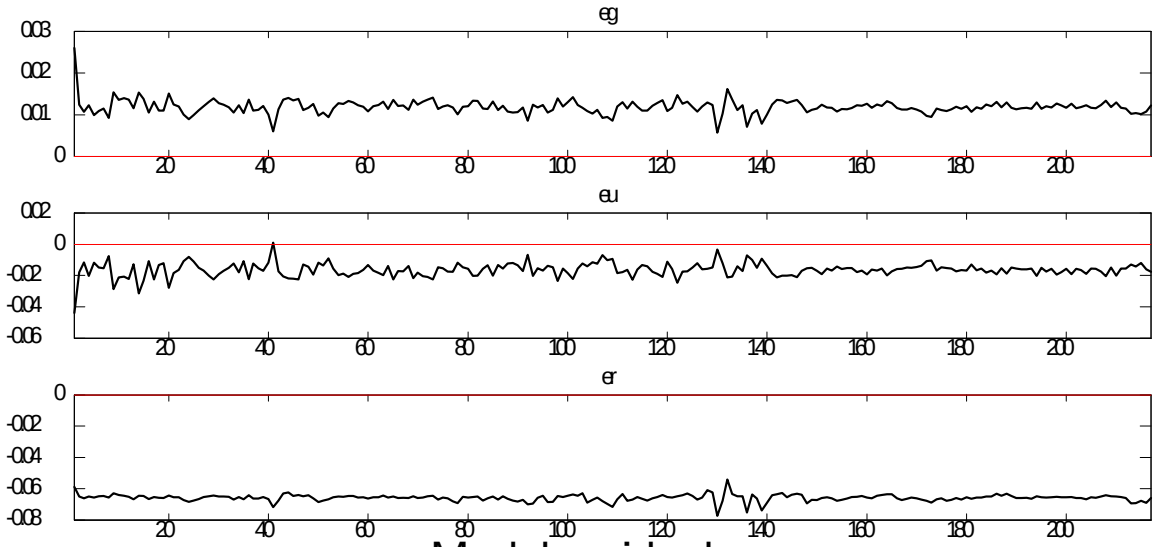
Let $y_t - \hat{y}_t$ the prediction error of the model. Prediction error should be:

- mean zero, iid (no trend or serial correlation should be detected).

- be homoskedastic (no clear break in the variance should be spotted).

- no shock should have " unreasonable" variance.

Residuals of a three equation sticky price model. Ok?



Model residuals

- Alternative plot actual and predicted values. Any interesting discrepancies? When?

2.2.1 Statistical tests

- Assume you have available a reference model ("the traditional one") and an alternative one.

• Test whether the Mean square Error (MSE) or Mean Absolute Error (MAE) of two (or more) models is the same.

Let y_t^1 and y_t^2 be the predicted value of y_t from models m_1 and m_2. Estimate jointly

$$y_t - y_t^1 = \mu + \epsilon_t^1 \quad (5)$$
$$y_t - y_t^2 = \mu + \epsilon_t^2 \quad (6)$$

where $\epsilon_t^1, \epsilon_t^2$ have the same variance, σ^2. Estimate the mean and the variance of each equation separately. Use a $\chi^2(2)$ test to verify if the restrictions hold (if they do than $MSE = \mu^2 + \sigma^2$ is the same for the two models, if they don't, the MSE is different).

- Compute unbiasedness regressions

$$y_t = a + by_t^* + u_t \tag{7}$$

y_t^* is the predicted value. Ideally $a = 0, b = 1$ for a " good" model.

- Compute predictive regressions

$$y_t = ay_t^b + by_t^* + e_t \tag{8}$$

y_t^* the predicted value for the (structural) model, y_t^b is the predicted value from a baseline (time series) model.

Check whether $b \neq 0$, i.e. does the new model adds information to the previous model?

• Estimate an unobservable factor model

$$y_t^* = \delta + \Lambda f_t^* + \epsilon_{1,t} \quad (9)$$

$$y_t = a + b f_t^* + \epsilon_{2,t} \quad (10)$$

f_t^* is the (unknown) predictable part of y_t. Here y_t^* is not an unbiased predictor for y_t but only a noisy measure of its predictable component. Does the first equation adds information to the estimates in the second?

• Add predicted values in a VAR and test significance (similar in spirit to predictive regressions - here looks at lagged info).

$$y_t = A(\ell) y_{t-1} + B(\ell) y_{t-1}^* + u_t \quad (11)$$

Test $B(\ell) = 0$, jointly or separately for each equation.

• Case studies: how does the model performs in particular episodes, i.e. a recession or an expansion; a period of high or low inflation, etc.

• If models have a Bayesian setup (with proper priors) can use the marginal likelihood. Priors can be non-informative but need to be proper.

Marginal likelihood of model j is $ML(j) = \int f(y, \theta) g(\theta | M(j)) d\theta$.

To compare alternatives: Posterior odds ratio/Bayes factor

$$PO = \frac{g(\mathcal{M}_j|y)}{g(\mathcal{M}_k|y)} = \frac{g(\mathcal{M}_j)}{g(\mathcal{M}_k)} \frac{ML(y|\mathcal{M}_j)}{ML(y|\mathcal{M}_k)} \tag{12}$$

The first term is the prior odds, the second the Bayes factor (BF).

Example 2 *Want to evaluate the stability of a fixed exchange rate agreement. Under H_0 (normal conditions) there is a 50-50 chance that the regime will be maintained (i.e. $f(y = 1|H_0) = f(y = 0|H_0) = 0.5$). Under H_1 (say, increasing oil prices) the probability that the fixed exchange*

rate regime will be maintained is 0.25 $f(y = 1|H_0) = 0.75$, $f(y = 0|H_0) = 0.25$. *Suppose* $g(H_0) = g(H_1) = 0.5$ *(equal prior probability),* $T = 100$, *and that the fixed exchange rate was maintained in 90 periods. Then:*

$$PO_{01} = \frac{(0.5)^{0.1}(0.5)^{0.9}}{(0.75)^{0.1}(0.25)^{0.9}} = \frac{0.5}{0.2790} = 1.79 \tag{13}$$

Hence, odds in favor of H_0 *increased from 1 to 1.79.*

• Bayes factors $=$ ratio of marginal likelihood of the two models. *It is different than LR statistic!*. What matters is the agreement of prior and likelihood and the least square fit of different models. LR does not integrate over α_j.

• BF implicitly discounts the fit of large scale models!

• Can also perform posterior predictive analysis (Box, 1980; Faust and Gupta, 2012).

Idea: provide a formal measure of how far a certain feature of the model is at odds with the data

- Can be applied to moments, impulse responses, autocovariances, spectral densities, etc. Only need the features to be a well defined continuous function of the data.

• Canova (1994),(1995) use prior predictive analysis to discard models which are going to be clearly at odds with the data.

Prior predictive analysis: For each θ simulate a sample $y(\theta)$ from the model, and compute statistics of interest. Can construct distribution of outcomes implied by the model and the prior. Check if the actual value of the statistic is within the range of values produced by the model for that statistic.

- If prior is sufficiently loose, values in the tails indicate that the model should not be used to study that particular phenomena.

Posterior predictive analysis: draw θ from the posterior and do the exercise as above.

Faust and Gupta (2012): alternative algorithm

- Draw θ from the posterior, compute $h(Y(\theta))$

- Simulate $Y^d(\theta)$ from every value of θ you have draws. Compute statistics $h(Y^d(\theta)$

- Plot joint distributions of $h(Y(\theta))$ and $h(Y^d(\theta))$. If they lie around the 45 degree line, data and model agree: otherwise data is unlikely form the point of view of the model.

- The share of points on the 45 degree line is a p-value for the hypothesis that model and data are from the same DGP.

- Apply the technique to SW model.

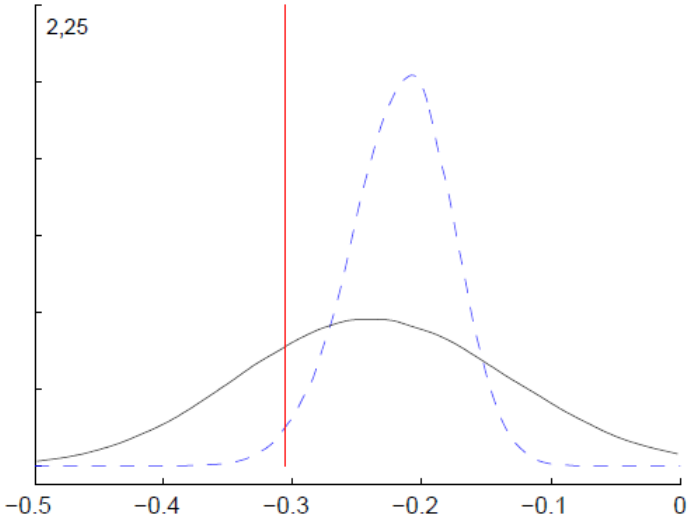


Figure 3: Posterior density for population, unconditional, correlation of output and inflation (dashed) and posterior predictive density for the sample correlation (solid). Vertical line is the realized value of sample correlation. The numbers in the upper left give the proportion of mass under the posterior and posterior predictive density, respectively, that is to the left of the realized value.

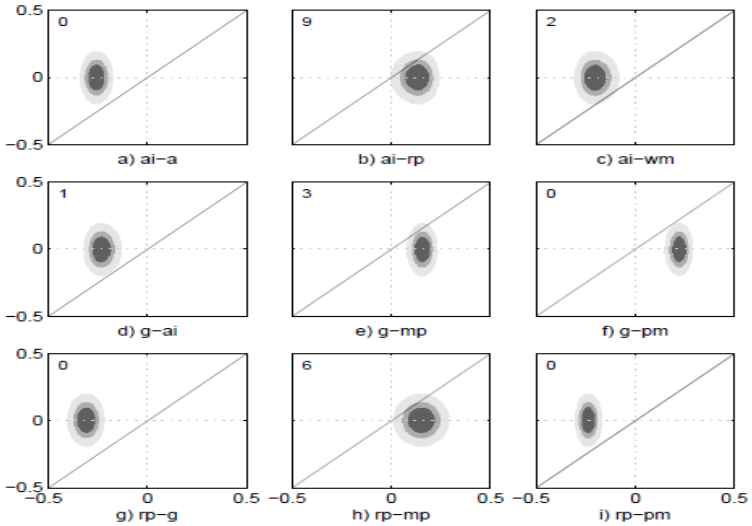


Figure 8: Posterior predictive contour plots for the correlations of various structural shocks. Each panel portrays the joint distribution of $h(Y(\theta), \theta)$ (vertical axis) and $h(Y^r, \theta)$ (horizontal) where θ is distributed according to the posterior, Y comprises two structural shocks, and h is the sample correlation between the two shocks. The shock labels are: a, productivity, ai: investment productivity, rp, risk premium; pm: price markup; wm: wage markup; g: government spending; mp: monetary policy. The number in the upper left is smaller of the share of points on either side of the 45 degree line.

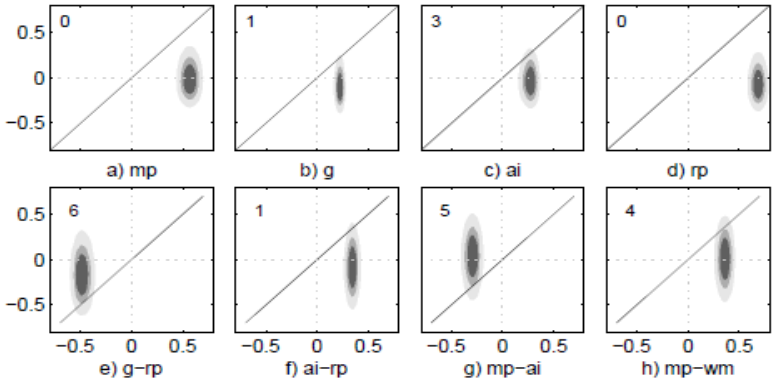


Figure 9: Posterior predictive contour plots for differences in properties of structural shock in recessions and expansions. Each panel portrays the joint distribution of $h(Y(\theta), \theta)$ (vertical axis) and $h(Y^r, \theta)$ (horizontal) where θ is distributed according to the posterior, Y is two smoothed structural shocks, and h is either the the proportional difference (recession minus expansion over expansion) in shock standard deviation in the top row, and the simple difference in correlation (recession minus expansion) in the second row. The shock labels are: a, productivity, ai: investment productivity, rp, risk premium; pm: price markup; wm: wage markup; g: government spending; mp: monetary policy. The number in the upper left is smaller of the share of points on either side of the 45 degree line.

Example 3 *Candidate models ARMA(1,1), BVAR-TVC with output, inflation and interest rate and a New-Keynesian model. Evaluation based on the fit of the inflation equation. Sample US 1955:1-2002:4.*

1) ARMA : $\pi_t = \rho_1 \pi_{t-1} + e_t + \rho_2 e_{t-1}$

2) TVC-BVAR ($\theta = vec(a_t, b_t(\ell))$*:*

$$y_t = a_t + b_t(\ell) y_{t-1} + e_t \tag{14}$$

$$\theta_t = \rho\theta_{t-1} + (1-\rho)\theta_0 + u_t, \quad \theta_t \sim N(0, \Omega) \tag{15}$$

3) NK model:

IS: $x_t = E_t x_{t+1} - \frac{1}{\phi}(r_t - E_t \pi_{t+1}) + g_t$

PC: $\pi_t = \beta E_t \pi_{t+1} + \frac{\phi(1-\zeta)(1-\beta\zeta)}{\zeta} x_t + u_t$

Taylor-Rule: $r_t = \psi_r r_t - 1 + (1 - \psi_r)(\phi_x x_{t-1} + \phi_p \pi_{t-1}) + e_t$

$v_t = (g_t, u_t) = \rho v_{t-1} + \eta_t;\ \eta_t\ iid\ N(0, \sigma^2).$

Pick estimates from Canova (2008) $\beta = 0.983(0.0008),\ \phi = 3.04(0.27), \zeta = 0.7709(0.185).$

In-sample RMSE, percentage points

Model	ARIMA	BVAR-TVC	NK
	1.88	1.04	1.33

In-sample, correlations: actual and predicted

Model	-1	0	1
ARIMA	0.67	0.88	0.76
BVAR-TVC	0.77	0.89	0.72
NK	0.56	0.68	0.51

Why is the MSE different from in-sample correlation analysis?

- MSE sum of square bias and variance. MSE could be small is variance small and bias not too large - a "straight line prediction " (a random walk) is typically good.

- Good MSE des not mean that actual and predicted go up and down together.

Unbiaseness regressions

Model	a	b	p-value $a = 0, b = 1$
ARIMA	0.159 (2.01)	0.79 (1.88)	0.03
BVAR-TVC	0.109 (1.56)	0.67 (2.06)	0.02
NK	0.035 (0.99)	0.56 (1.71)	0.01

Predictive regressions

Model	a	b
ARIMA-NK	0.82 (2.17)	0.23 (1.65)
BVAR-NK	0.73 (1.96)	0.35 (2.00)

Output growth VAR regressions

Model	p-value lags of predicted inflation
Just lagged output	0.00
Adding Inflation	0.15
Adding Nominal Rate	0.42

- Bayes factor: ML(NK)/ML(BVAR) $= 0.02$.

2.2.2 Economic tests

- Compute moments of actual data and predicted ones.
- Compute dynamic responses to shocks in the model and the data
- Compute turning points statistics, overall or at some dates.
- Compare favorite stylized facts with model implications.

Moments

Model	Mean	Variance	Corr(π^*, y)	Corr(π^*, R)
ARIMA	2.56	3.99	-0.23	0.65
BVAR	2.19	2.78	-0.39	0.67
NK	2.24	2.22	-0.37	0.88
Actual	2.08	2.57	-0.51	0.70

Peak inflation, late 1970s

Model	date	68% range
ARIMA	1979:2	[1978:4, 1980:2]
BVAR	1979:4	[1979:1, 1980:4]
NK	1981:2	[1979:4, 1982:2]
Actual	1980:1	

2.3 Forecasting with DSGE models

Recall: Log linearized aggregate decision rule of a DSGE model is:

$$y_{2t} = A_{22}(\theta)y_{2t-1} + A_{21}(\theta)y_{3t} \quad (16)$$

$$y_{1t} = A_1(\theta)y_{2t} = A_{11}(\theta)y_{2t-1} + A_{12}(\theta)y_{3t} \quad (17)$$

$y_{2t} =$ states and the driving forces, $y_{1t} =$ controls, y_{3t} shocks. $\mathcal{A}_{ij}(\theta), i, j = 1, 2$ are time invariant functions of θ, the structural parameters.

- There are cross equation restrictions since $\theta_i, i = 1, \ldots, n$ appears in more than one entry of these matrices.

- (17) is a state space or a restricted VAR(1) model

• Unconditional forecast: $y_{3t+\tau} = 0, \forall \tau > 0$, let the system run. With a VAR(1) representation: let $y_t = (y_{1t}, y_{2t})$. Then $y_{t+\tau} = \hat{A}^\tau y_t$ and $y_{2t+\tau} = S\hat{A}^\tau$, where $\hat{A}$ is an estimate of A and S is a selection matrix, picking up the second set of elements from A.

To calculate uncertainty around point forecasts.
If a distribution for $\hat{A}$ is available (asymptotic or posterior) then:

1. Draw A^l from this distribution, compute $y^l_{t+\tau}, \quad l = 1, 2, \ldots, L$, each horizon τ.

2. Order $y^l_{t+\tau}$ over l, each τ and extract 16-84 or 2.5-97.5 percentiles.

• Conditional forecast 1: Manipulating shocks.

This is the same as computing impulse responses, i.e. need to orthogonalize the disturbances if they are not orthogonal. Only difference is that here impulse may last more than one period. Choose $y_{3t+\tau} = \bar{y}_{3t+\tau}$, $\tau = 0, 1, 2, \ldots, \bar{\tau}$. Given $\hat{A}$ find $y_{2t+\tau} = \hat{A}_{22}(\theta) y_{2t+\tau-1} + \hat{A}_{21}(\theta) y_{3t+\tau}$ and $y_{1t+\tau} = \hat{A}_1(\theta) y_{2t+\tau}$.

To calculate uncertainty around the forecasted path, use same algorithm employed for unconditional forecasts (i.e. draw A's from their distributions).

• Conditional Forecast 2: Manipulating endogenous states
This requires backing out shocks needed to produce the path $\bar{y}_{2t+\tau},\ \tau = 0, 1, 2, \ldots$. Simply use the first equation of (17) to do this. Then $y_{1t+\tau} = A_1(\theta)\bar{y}_{2t+\tau},\ \tau = 1, 2, \ldots$. Same as above to compute uncertainty around the forecasted path.

Identification problem: there may be different elements of y_{3t} which may induce the require path for $y_{2t+\tau}$.

Example 4 *What is the range of paths for consumption from next quarter up to 10 years if the capital stock is higher by ten percent in all these periods? Question: how do we increase the capital stock? Via technology shocks? Via labor supply shocks?*

• Conditional Forecast 3: Manipulating endogenous controls. Separate $y_{1t} = [y_{1t}^A, y_{1t}^B]$ and $y_{1t+\tau}^A = \bar{y}_{1t+\tau}^A,\ \tau = 0, 1, 2, ...$ Back out the path of $y_{1t+\tau}^A$ needed to produce $\bar{y}_{1t+\tau}^A$. With this path compute $y_{1t+\tau}^B$. Same identification problems as above; less problematic in some cases.

Example 5 *Suppose that interest rates are (discretionarily) kept 50 basis point higher than the endogenous Taylor rule would imply. What is the effect on inflation?*

2.3.1 Out-of-sample evaluation

Use the same statistics employed for in-sample analysis. Now can evaluate forecasts at different horizons.

Out-of-sample RMSE, percentage points, unconditional forecasts

Model	1 quarter	4 quartesr	8 quarters
ARIMA	1.43	2.16	2.92
BVAR-TVC	1.21	1.72	1.89
NK	1.33	1.58	1.87

Unconditional Out-of-sample Predictive regressions, estimates of b

Model	1 quarter	4 quarters	8 quarters
ARIMA-NK	0.35 (1.71)	0.42 (1.97)	0.34 (2.00)
BVAR-NK	0.17 (1.66)	0.35 (1.89)	0.44 (2.06)

3 Exploiting other features for evaluation

• Distinguish between internal vs. external dynamics. Models driven almost entirely by external dynamics not very useful.

- Plot together shocks (not the residuals) and the data. What is the contribution of the model?

• Co-cycles analysis.

Log-linearized model (no distinction now states/controls):

$$y_t = B(\theta)E_t y_{t+1} + A(\theta)y_{t-1} + F(\theta)u_t \tag{18}$$

Solution:

$$y_t = Py_{t-1} + D\sum_j S^j E_t u_{t+j} \tag{19}$$

where $P - BP^2 - A = 0$, $D = (I - BP)^{-1}F$, $S = (I - BP)^{-1}B$. If $u_t = \Phi u_{t-1} + \eta_t$ and we let $G = D\sum_j S^j \Phi^j$

$$y_t = Py_{t-1} + Gu_t = Py_{t-1} + G\Phi u_{t-1} + G\eta_t \tag{20}$$

Interested in features of P and G.

If $rank(G) \leq dim(u)$, let $G^{+} = (G'G)^{-1}G'$ and $u_t = G^{+}(y_t - Py_{t-1})$ so that

$$\begin{aligned} y_t &= Py_{t-1} + G\Phi(G^{-1}(y_{t-1} - Py_{t-2}) + G\eta_t \quad (21) \\ &= (P + G\Phi G^{-1})y_{t-1} + G\Phi G^{-1}Py_{t-2} + G\eta_t \quad (22) \end{aligned}$$

Since rank(P)=min(rank($I - BP$),rank(A)), if A is of reduced rank (i.e. more states than controls), P will be of reduced rank, i.e. comovements in y_t driven by a reduced number of shocks. (Note this is different from the fact that dim(u) <dim(y)).

Is this true in the data? Can use factor models to verify this.

• When models feature both long run and short run dynamics Possibility of evaluation looking at long run features.

- Check if permanent component of the model has the same properties as permanent component in the data.

- Can use both cointegration or BQ decompositions. If cointegration: $\Delta y_t = C(\ell)\eta_t$. Interested in $C(1)$. Choose $\beta' C(1) = 0$. Partition $y_t = [y_{1t}, y_{2t}]$ where y_{2t} are I(0). Then $\phi_t = \beta' y_{1t}$ are cointegrating vectors and the VECM is

$$\begin{bmatrix} \Delta y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \alpha' \\ \gamma' \end{bmatrix} \phi_{t-1} + M y_{2t-1} + v_t$$

Use this to extract the permanent component of y_{1t} in model and data and to compare them.

Blanchard-Quah decomposition is:

$$\begin{pmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ 0 \end{pmatrix} + \begin{pmatrix} C_1(1) \\ 0 \end{pmatrix} e_t + \begin{pmatrix} (1-\ell)C_1^{\dagger}(\ell) \\ (1-\ell)C_2^{\dagger}(\ell) \end{pmatrix} e_t \qquad (23)$$

where $C_1^{\dagger}(\ell) = \frac{C_1(\ell)-C_1(1)}{1-\ell}$ $C_2^{\dagger}(\ell) = \frac{C_2(\ell)}{1-\ell}$, $0 < \text{rank}[C_1(1)] \leq m_1$ and $\Delta y_t^x = [\bar{y}_1 + C_1(1)e_t, 0]'$ is the permanent component of y_t.

• May want to choose (estimate) parameters so that the model tracks long run properties of the data. Then check if dynamics (univariate) properties fit.

A few issues to keep in mind

• Estimated DSGE models typically have driving forces that are correlated (theory assumes that they are not). Misspecification!!!

• If the number of variables is different from the number of shocks, solution is not a VAR but a VARMA: invertibility problems!

• Common to add measurement error but be careful:

$$y_t = y_t^* + e_t \tag{24}$$
$$y_t^* = P y_{t-1}^* + G u_t \tag{25}$$

a) if e_t is iid $\longrightarrow$ signal extraction problem, use Kalman Filter to get y_t^*.

b) If e_t is serially correlated ($e_t = \rho e_{t-1} + v_t$) then:

$$\Delta y_t = \Delta y_t^* + (1 - \rho)(y_{t-1}^* - y_{t-1}) + v_t \tag{26}$$

This is VECM linking observables y_t and unobservables y_t^*.

- On average $y_t = y_t^*$. In short run deviations are possible.

- Can't use the KF to construct y_t^* in this case.

4 Evaluation via VARs

- Long history in the literature

Canova, Finn, Pagan (1993): Evaluate quantitative properties of RBC models through VARs.

Ingram and Whiteman (1994): Use model to setup a prior for the VAR. Is it better than standard statistical priors?

Canova and Paustian (2007): Evaluate model using qualitative model-based sign restrictions to identify shocks in a VAR.

Procedure

- Start with a broad class of structural models. The class should nest submodels through parameter restrictions (price and wage stickiness, indexation, habit,...).

- Find implications that are robust to parameter variations.

(a) Some implications are robust across submodels.

(b) Some implications are robust within a particular submodel.

- Use a subset of implications robust across models to identify shocks in a VAR.

- Use implications that are robust within a submodel and different across submodels for evaluation.

- Do this qualitatively and quantitatively using probabilistic criteria.

Details

- What are robust restrictions? Magnitude restrictions not robust. Zeros not typically a feature of theory. Use sign of the impact response.

- Robust testing: sign and shape of dynamics of unrestricted variables to shocks.

- Produce a partially identified model: standard statistical criteria problematic (Moon and Schorfheide (2008)).

- Can be used **without** estimation of the parameters - good if there are big identification problems.

- VAR mispecification (relative to a DSGE) ok.

Why VAR misspecification not a problem?

- Use robust sign restriction.

- Shock identification robust to time series representation of decision rules.

$$\begin{aligned} x_{1t} &= A(\theta)x_{1t-1} + B(\theta)e_t \\ x_{2t} &= C(\theta)x_{1t-1} + D(\theta)e_t \end{aligned} \tag{27}$$

$$\begin{bmatrix} I - F_{11}\ell & F_{12}\ell \\ F_{21}\ell & I - F_{22}\ell \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} e_t$$

Representation for y_{2t} (integrating out y_{1t}):

$$(I - F_{22}\ell - F_{21}F_{12}(1 - F_{11}\ell)^{-1}\ell^2)y_{2t} = [G_2 - (F_{21}(1 - F_{11}\ell)^{-1}G_1\ell]e_t \tag{28}$$

ARMA(∞, ∞) but impact effects of e_t has correct sign and magnitude.

Example 6 *(use model based restrictions to robustify inference). Use Christiano, et. al. (2005) and Smets and Wouters (2003) class of models.*

$$
\begin{aligned}
y_t &= c_y c_t + i_y i_t + g_y e_t^g && (29)\\
c_t &= \frac{h}{1+h} c_{t-1} + \frac{1}{1+h} E_t c_{t+1} - \frac{1-h}{(1+h)\sigma_c}(R_t - E_t \pi_{t+1}) + \frac{1-h}{(1+h)\sigma_c}(e_t^b - E_t e_{t+1}^b) && (30)\\
i_t &= \frac{1}{1+\beta} i_{t-1} + \frac{\beta}{1+\beta} E_t i_{t+1} + \frac{\phi}{1+\beta} q_t - \frac{\beta E_t e_{t+1}^I - e_t^I}{1+\beta} && (31)\\
q_t &= \beta(1-\delta) E_t q_{t+1} - (R_t - \pi_{t+1}) + \beta r^* E_t r_{t+1} && (32)\\
y_t &= \omega(\alpha K_{t-1} + \alpha \psi r_t + (1-\alpha) l_t + e_t^x) && (33)\\
k_t &= (1-\delta) k_{t-1} + \delta i_t && (34)\\
\pi_t &= \frac{\beta}{1+\beta\mu_p} E_t \pi_{t+1} + \frac{\mu_p}{1+\beta\mu_p} \pi_{t-1} + \kappa_p m c_t && (35)\\
w_t &= \frac{\beta}{1+\beta} E_t w_{t+1} + \frac{1}{1+\beta} w_{t-1} + \frac{\beta}{1+\beta} E_t \pi_{t+1} - \frac{1+\beta\mu_w}{1+\beta} \pi_t + \frac{\mu_w}{1+\beta} \pi_{t-1} - \kappa_w \mu_t^W && (36)\\
l_t &= -w_t + (1+\psi) r_t + k_{t-1} && (37)\\
R_t &= \rho_R R_{t-1} + (1-\rho_R)(\gamma_\pi \pi_t + \gamma_y y_t) + e_t^R && (38)
\end{aligned}
$$

Support for the parameters

	Parameter	Support
σ_c	risk aversion coefficient	[1,6]
h	consumption habit	[0.0,0.8]
σ_l	inverse labor supply elasticity	[0.5,4.0]
ω	fixed cost	[1.0,1.80]
$1/\phi$	adjustment cost parameter	[0.0001,0.002]
δ	capital depreciation rate	[0.015,0.03]
α	capital share	[0.15,0.35]
$1/\psi$	capacity utilization elasticity	[0.1,0.6]
g_y	share of government consumption	[0.10,0.25]
ζ_p	degree of price stickiness	[0.4,0.9]
μ_p	price indexation	[0.2,0.8]
ζ_w	degree of wage stickiness	[0.4,0.9]
μ_w	wage indexation	[0.2,0.8]
ε^w	steady state markup in labor market	[0.1,0.7]
γ_R	lagged interest rate coefficient	[0.2,0.95]
γ_π	inflation coefficient on interest rate rule	[1.1,3.0]
ρ_y	output coefficient on interest rate rule	[0.0,1.0]
$\boldsymbol{\varrho}_i$	persistence of shocks $i = 1, \ldots, 7$	[0,0.9]

Question of interest: What is the relationship between hours and technology shocks? Do hours robustly fall or robustly increase?

Sign of the impact responses to shocks

	TFP	Monetary	Taste	Inv	Markup	L^s	G
Δy_t	+	+	+	+	+	+	+
π_t	-	+	+	-	-	-	+
Δc_t	+	+	+	-	+	+	-
Δgap_t	+	-	-	?	-	+	-
Δw_t	+	+	+	-	+	-	?

Identification restrictions for technology shocks

a) $\pi \downarrow, \Delta y \uparrow$.

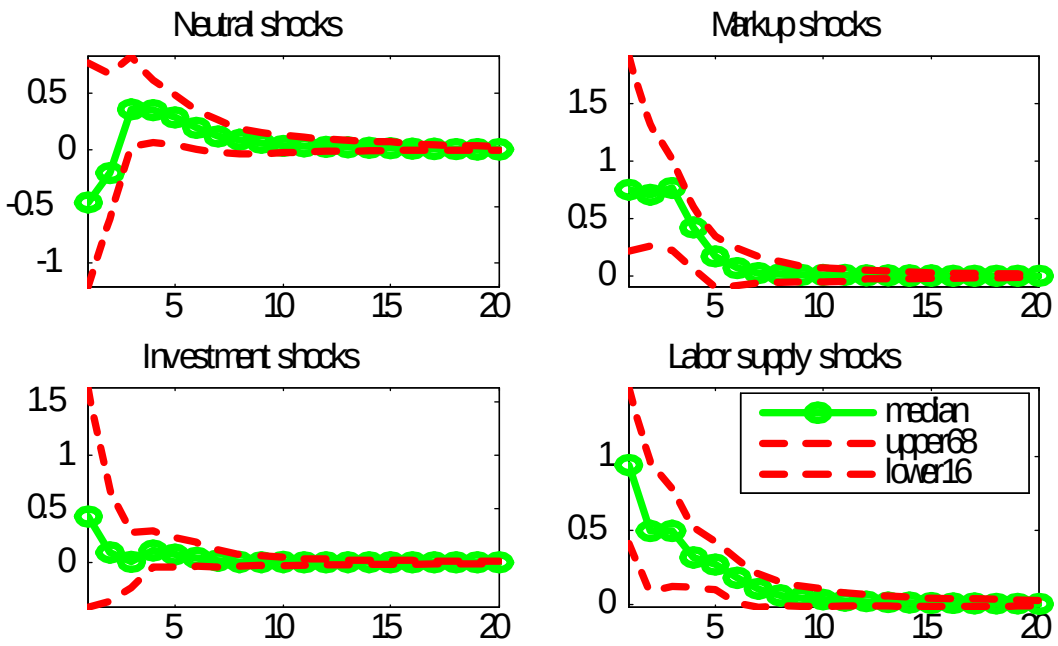
b) $\Delta c \downarrow$ with Investment shock, $\uparrow$ with others.

c) $\Delta gap \uparrow$ with TFP shocks, $\downarrow$ with markup

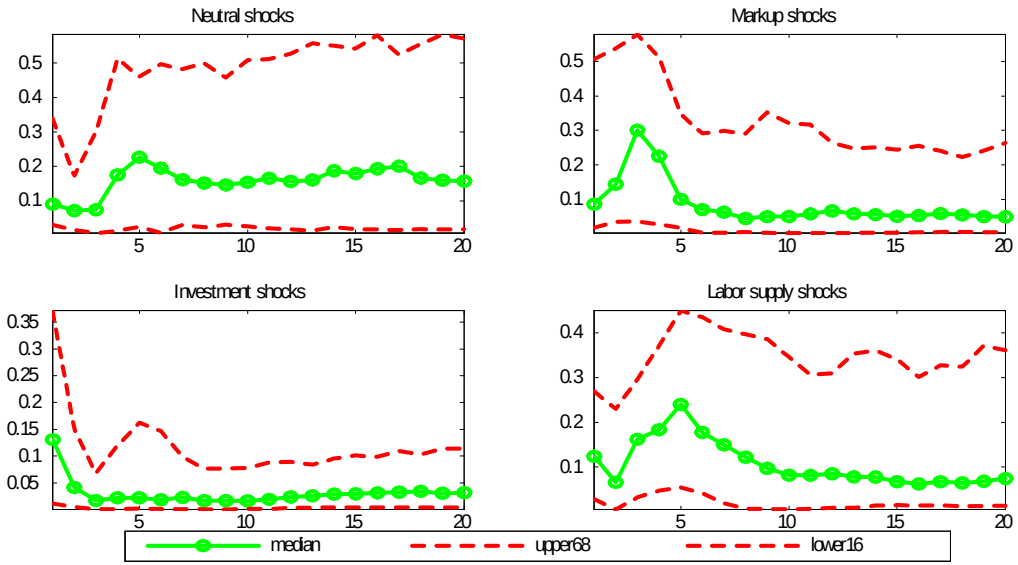
d) $\Delta w \downarrow$ with Labor supply and investment shocks, $\uparrow$ markup and TPF shocks.

- These restrictions produce mutually exclusive shocks.

- These restrictions do not involve hours. Once shocks are identified measure the response of hours and the contribution of various technology shocks to their variability.



Responses of hours to technology shocks



Share of hours volatility explained by technology shocks

Conclusions: a model based identification approach tells us:

- Response of hours depend on the source of technological disturbance

- With TFP shocks hours response insignificant contemporaneously, weakly positive after a while (i.e. it is neither NK not RBC).

- With the other shocks hours typically increase.

- Proportion of the variance of hours explained by TPF shocks large but very imprecisely estimated (can't really say if TPF shocks matter or not).

Example 7 *(use model based restrictions test RBC vs. NK transmission) (Pappa, 2009). RBC and NK models have different implications for the tranmisssion of goverment expenditure shocks to labor markets.*

- In RBC a g shocks make hours and wage increase.

- In a NK a g shock make hours and wage move in the opposite direction.

Which mechanism is more consistent with the facts?

- Take a general specification where you can nest a RBC model as a special case (take a NK model where the monopolistic distortions have been eliminated and consider either a sticky price or a flexible price version of the model).

- Find robust restrictions of the two class of models which do not involve either hours or real wages. There are many. Choose restrictions which are commonly satisfied across models.

Shock	Y	Deficit	C^g	I^g	N^g
C^g shock	> 0	> 0	> 0		
I^g shock	> 0	> 0		> 0	
N^g shock	> 0	> 0			> 0

- Contemporaneous effects only (the distinction at longer horizons is blurred).

- Sign consistent with a large range of parameter values.

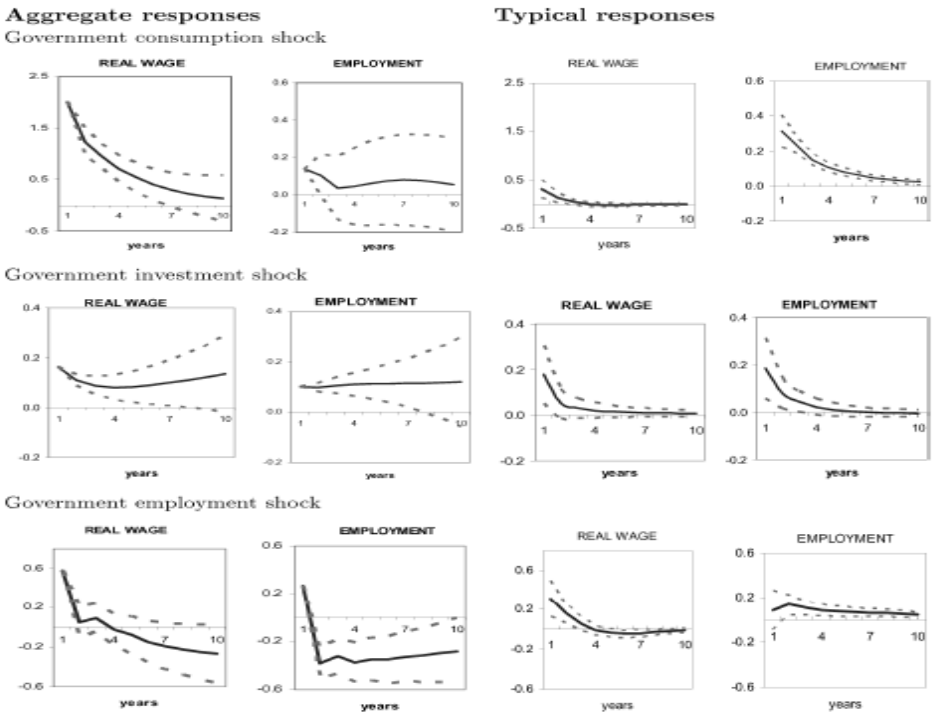


FIGURE 3

LABOR MARKET RESPONSES

Del Negro and Schorfheide (2004), Del Negro, et al. (2006): A VAR is a bridge between a DSGE model and the data.

Add to the actual data, simulated DSGE data organized in a VAR. Use a parameter to weight the two information. Model is the DGP of the data if parameter is ∞, model totally fails if parameter is 0.

Interpretation: DSGE is a VAR with cross equation restrictions. If restrictions are false better to relax them completely, i.e. use a VAR for the data.

Approach:

- Simulate data from model

- Append simulated data to VAR via a prior (use mean/variance of estimated parameters on simulated data to set up a Normal-Wishart prior).

- Choose proportion of simulated to actual data (to test model).

Let θ be DSGE parameters; α the VAR parameters. Prior is:

$g(\theta) = \prod_{i=1}^{k} g(\theta_k)$;

$g(\alpha) \sim N(\bar{\alpha}(\theta), \bar{\Sigma}_a(\theta))$;

$\Sigma_e \sim IW(T_1 \bar{\Sigma}(\theta), T_1 - k)$ where

$$\begin{aligned} \bar{\alpha}(\theta) &= (X^{s'} X^s)^{-1} (X^{s'} y^s) \\ \bar{\Sigma}_a(\theta) &= \Sigma_e(\theta) \otimes (T_1 X^{s'} X^s)^{-1} \\ \bar{\Sigma}(\theta) &= (y^{s'} y^s - (y^{s'} X^s) \bar{\alpha}(\theta)) \end{aligned} \tag{39}$$

y^s simulated data, X^s lags in the VAR of simulated data. $T_1 =$ length of simulated data. $\kappa = \frac{T_1}{T}$ measures the relative importance of two types of information. $\kappa \to 0$ ($\kappa \to \infty$) actual (simulated) data dominates.

Hierarchical structure: $f(\alpha, \Sigma_e|y)g(\alpha|\theta)g(\Sigma_e|\theta)g(\theta)$. Since the likelihood and the prior are conjugate:

$(\alpha|\theta, y, \Sigma_e) \sim N(\tilde{\alpha}(\theta), \tilde{\Sigma}(\theta))$;

$(\Sigma_e|\theta, y) \sim iW((\kappa + T)\tilde{\Sigma}(\theta), T + \kappa - k)$ where

$$\begin{aligned} \tilde{\alpha}(\theta) &= (T_1 X^{s'} X^s + X'X)^{-1}(T_1 X^{s'} y^s + X'y) \\ \tilde{\Sigma}_a(\theta) &= \Sigma_e(\theta) \otimes (T_1 X^{s'} X^s + X'X)^{-1} \\ \tilde{\Sigma}(\theta) &= \frac{1}{(1+\kappa)T}[(T_1 y^{s'} y^s + y'y) - (T_1 y^{s'} X^s + y'X)\tilde{\alpha}(\theta)] \quad (40) \end{aligned}$$

and $g(\theta|y) \propto g(\theta) \times |\Sigma_e|^{-0.5(T-M-1)}$
$\exp\{-0.5tr[\Sigma_e^{-1}(Y - X\alpha)'(Y - X\alpha)\} \times |\Sigma_e(\theta)|^{-0.5(T_1-M-1)}$
$\exp\{-0.5tr[\Sigma_e(\theta)^{-1}(Y^s - X^s\alpha(\theta))'(Y^s - X^s\alpha(\theta))\}$.

- Can estimate jointly θ and α but also possible to calibrate θ.

- All posterior moments in (40) conditional on κ. How do we select it?

- Use Rules of thumbs (e.g. $\kappa = 1$, T observation added).

- Maximize marginal likelihood.

Example 8 *In a basic sticky price-sticky wage economy, fix* $\eta = 0.66, \pi^{ss} = 1.005, N^{ss} = 0.33, \frac{c}{gdp} = 0.8, \beta = 0.99, \zeta_p = \zeta_w = 0.75, a_0 = 0, a_1 = 0.5, a_2 = -1.0, a_3 = 0.1$. *Run a VAR with output, interest rates, money and inflation using actual quarterly data from 1973:1 to 1993:4 and data simulated from the model conditional on these parameters. Overall, only a modest amount of simulated data (roughly, 20 data) should be used to set up a prior.*

Marginal Likelihood, Sticky price sticky wage model.

$\kappa = 0$	$\kappa = 0.1$	$\kappa = 0.25$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
-1228.08	-828.51	-693.49	-709.13	-913.51	-1424.61

5 Using loss functions to evaluate DSGE models

Schorfheide (2000): Compare two DSGE models both misspecified.

- PO ratio for misspecified models uninteresting. One model preferred but it may have very close to zero posterior probability.

Example 9 $PO = \frac{\pi_{1,T}}{\pi_{2,T}} = \frac{\pi_{1,0}}{\pi_{2,0}} \frac{ML(Y_T|\mathcal{M}_1)}{f(Y_T)} * \frac{f(Y_T)}{ML(Y_T|\mathcal{M}_2)}$. *If use 0-1 loss function the posterior risk is minimized by selecting* $\mathcal{M}_1$ *if* $PO > 1$.

Potential presence of a third (better specified) model does not affect PO *if the prior odds* $\frac{\pi_{1,0}}{\pi_{2,0}}$ *unchanged (*$\mathcal{M}_3$ *enters only in* $f(Y_T)$*, which cancels out).*

Problem if $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *have low posterior probability,* $\mathcal{M}_3$ *a large one.*

- Solution: Use loss functions.

Procedure

1. Compute the posterior distribution for the parameters of each model, using tractable priors and one of the available posterior simulators.

2. Obtain the marginal likelihood of the data, for each $\mathcal{M}_i$, that is, compute $f(y|\mathcal{M}_i) = \int f(y|\theta_i, \mathcal{M}_i) g(\theta_i|\mathcal{M}_i) d\theta_i$.

3. Compute posterior probabilities $\tilde{P}_i = \frac{\bar{P}_i f(y|\mathcal{M}_i)}{\sum_i \bar{P}_i f(y|\mathcal{M}_i)}$, where $\bar{P}_i$ is the prior probability of model i. Note that if the distribution of y is degenerated under $\mathcal{M}_i$ (e.g. if number of shocks is smaller than the number of endogenous variables), $\tilde{P}_i = 0$.

4. Calculate the posterior distribution of $h(\theta)$ for each model and average using posterior probabilities i.e. obtain $g(h(\theta)|y, \mathcal{M}_i)$, and $g(h(\theta)|y) = \sum_i \tilde{P}_i g(h(\theta)|y, \mathcal{M}_i)$. If all but model i' produce degenerate distributions for θ, $g(h(\theta)|y) = g(h(\theta)|y, \mathcal{M}_{i'})$.

5. Setup a loss function $\mathfrak{L}(h_T, h_i(\theta))$ measuring the discrepancy between model's i predictions and data h_T. Since the optimal predictor in model $\mathcal{M}_i$ is $\hat{h}_i(\theta) = arg\min_{h_i(\theta)} \int \mathfrak{L}(h_T, h_i(\theta)) g(h_i(\theta)|y, \mathcal{M}_i) dh_T$, one can compare models using the risk of $\hat{h}_i(\theta)$ under the overall posterior distribution $g(h(\theta)|y)$, i.e. $\mathfrak{R}(\hat{h}_i(\theta)|y) = \int \mathfrak{L}(h_T, \hat{h}_i(\theta)) g(h(\theta)|y) dh_T$.

In step 5) $\mathfrak{R}(\hat{h}_i(\theta)|y)$ measures how well model $\mathcal{M}_i$ predicts h_T. Note that while model comparison is relative, $g(h(\theta)|y)$ takes into account information from all models.

Taking step 5) further: for each i, θ can be selected so as to minimize $\mathfrak{R}(\hat{h}_i(\theta)|y)$. Such an estimate provides a lower bound to the posterior risk obtained by the "best" candidate model.

• Possible loss functions:

(a) Quadratic loss: $L_2(h(\theta), \hat{h}(\theta)) = (h(\theta) - \hat{h}(\theta))' W (h(\theta) - \hat{h}(\theta))$; W is a weighting matrix.

(b) Penalized Loss: $L_p(h(\theta), \hat{h}(\theta)) = I[g(h(\theta)|Y_T) > g(\hat{h}(\theta)|Y_T)]$, where $I(x, z) = 1$ if $x > z$ and zero otherwise.

(c) χ^2 loss: $L_{\chi^2}(h(\theta), \hat{h}(\theta)) = I[C_{\chi^2}(h(\theta)|Y_T) > C_{\chi^2}(\hat{h}(\theta)|Y_T)]$ where $C_{\chi^2}(h(\theta)|Y_T) = (h(\theta) - E(h(\theta)|Y_T))' V_\theta^{-1} (h(\theta) - E(h(\theta)|Y_T))$ and V_θ is a posterior covariance matrix of $h(\theta)$.

(d) 0-1 loss: $L_{01} = 1$ if $\hat{h}(\theta) \neq h(\theta)$ and zero otherwise.

- Results:

1) If $g(h(\theta)|Y_T)$ is normal $L_2 = L_{\chi^2}$.

2) Optimal predictor under L_2 and L_{χ^2} is $E(h(\theta)|Y_T, \mathcal{M}_i)$.

3) Optimal predictor for L_p is the posterior mode of $g(h(\theta)|Y_t, \mathcal{M}_i)$.

4) If for any positive definite W, $\mathcal{M}_1 > \mathcal{M}_2$ with probability one as $T \to \infty$, L_q selection is consistent and identical to a PO ratio.

5) If the two models are so misspecified that their posterior probability goes to zero as $T \to \infty$, the ranking depends on the discrepancy between $E(h(\theta)|y, \mathcal{M}_3) \approx E(h(\theta)|y)$ and $\hat{h}_i(\theta), i = 1, 2$. If $\mathcal{M}_3$ is any empirical model, then using a $\mathfrak{L}_2$ loss is equivalent to compare sample and population moments obtained from different models informally.

Simplest calibration exercises is optimal Bayesian decision using $\mathfrak{L}_2$ loss function and the models are highly misspecified.

Example 10 *Ferroni (2011).*

Take a cyclical DSGE. Possibility that the data is generated by the model plus three alternative specification for the non-cyclical part.

M1= DSGE + linear trend, M2= DSGE + HP trend, M3= DSGE + RW trend

1) Which model has the higher posterior probability (starting from an equal prior probability)?

- Log Bayes factor of M2 relative to M1=-31.80.

- Log Bayes factor of M3 relative to M2 = 98.47.

2) How do you robustify inference about DSGE parameters to trend uncertainty? Construct

$$g(\theta|y, DSGE) = \frac{\sum_j p(y|M_j)}{\sum_k p(y|M_k)} \int g(\theta|y, M_j, \alpha^j) d\alpha^j \tag{41}$$

where α_j represents the parameters of the non-cyclical specification for model j. (Hint to do this you need to estimate cyclical and non-cyclical parts jointly).

6 Some additional thoughts

Different kind of DSGE models:

• Academic model (typically small): generated by the idea of having internal consistency and well defined structure than the need to fit the data.

• Central Bank model (similar to academic model but typically large)

1) Why should a Central Bank model be big??

Can you figure out if it has a unique equilibrium?

Can you figure out what drives dynamics?

Is reality complex or are we unable to understand it?? "Sophisticated simplicity" Jeffreys(1963), Zellner(1981),

2) What defines a Central bank model?

"A tool to help to focus policy discussion around some stories rather than degenerating into a discussion of many special events as it often happens with data driven models"

• Operational Central Bank model (adjusted CB model to fit existing evidence, e.g. adding extra sources of errors or dynamics, converting theoretical variables in measurable ones).

• Central Bank forecasting model (OCB model adjusted to incorporate policymakers beliefs about the future (e.g. incorporate survey data information, information from anticipatory variables, etc.).

• The above evaluation procedures can be applied to any type of CB model. Careful if you use them with last two since posterior information is often used to setup your model.

• How should one move down from theoretical to policy oriented models?? Or should we go the other way around?