Bayesian methods for structural VARs

Fabio Canova
Norwegian Business School, CAMP, and CEPR
April 2023

Outline

- Bayesian vs. frequentist. Bayes theorem. Likelihood and prior selection.
- Posterior simulators. MCMC methods.
- Classical VARs. Identification restrictions: SVAR.
- BVAR and likelihood function for a VAR(q).
- Priors for VARs (Diffuse, Conjugate, Hierarchical).
- Structural analyses and forecasting with BVARs.
- Large scale BVARs.

References

Antolin, J. and Rubio, J (2018), Narrative sign restrictions. *American Economic Review*. 108, 2802-2829.

Arias, J., J. F. Rubio-Ramirez, and D. F. Waggoner (2018): Inference Based on SVARs Identified with Sign and Zero Restrictions: Theory and Applications, Econometrica, 86, 685-720.

Arias, J., J. F. Rubio-Ramirez, and D. F. Waggoner (2021): Uniform priors for impulse responses.

Banbura, M., Giannone, D., and Reichlin, L., 2010. Large Bayesian Vector Autoregressions. *Journal of Applied Econometrics*, 25, 71-92.

Belmonte,M. Koop, G. and D. Korobilis, D. (2014). Hierarchical Shrinkage in time varying parameters, *Journal of Forecasting*, 33, 80-94 .

Berger, J. (1995). Statistical decision theory and bayesian analysis, Springer and Verlag.

Blanchard, O. and Quah, D. (1989), "The Dynamic Effect of Aggregate Demand and Supply Disturbances", *American Economic Review*, 79, 655-673.

Blanchard, O., LeHuillier, G. Lorenzoni (2013), "News, noise and fluctuations: An empirical exploration", American Economic Review, 103, 3045-3070.

Baumeister, C. and J, Hamilton (2015), "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information", Econometrica, 83, 1963-1999.

Bekaert, G., Engstrom, E. and Ermolov, A, (2021), "Identifying aggregate demand and supply shocks using sign restrictions and higher order moments", manuscript.

Canova, F. and De Nicolo, G (2002), " Money Matters for Business Cycle Fluctuations in the G7", *Journal of Monetary Economics*, 49, 1131-1159.

Canova, F. and Pina, J. (2005) "Monetary Policy Misspecification in VAR models", in C. Diebolt, and Krystou, C. (eds.) *New Trends In Macroeconomics*, Springer Verlag.

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

Canova, F. and Ferroni, F. (2022) "Mind the Gap! Stylized facts and structural models, *American Economic Journal Macroeconomics*, 14, 104-135.

Canova, F. (2007). Methods for Applied Macroeconomic Research. Princeton University Press.

Canova, F. and F. Perez Forero (2015). "Estimating Overidentified, Non-recursive, TVC structural VARs",*Quantitative Economics*, 6, 359-384.

Carlstrom, C. Fuerst, T. and M. Paustian (2009) "Monetary Policy Shocks, Cholesky identification and DNK models", *Journal of Monetary Economics*, 56, 1014-1021.

Carriero, A., Kapetanios, G. and Marcellino, M. (2014). A Shrinkage Instrumental Variable Estimator for Large data sets. *L' Actualite Economique.*

Carriero, A., Clark, T. and Marcellino, M. (2019). Large vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Ec onometrics*, 212, 137-154.

Chari, V., Kehoe, P. and McGrattan, E. (2008) "Are Structural VARs with Long run restrictions useful in developing Business cycle Theory", *Journal of Monetary Economics*, 55, 1137-1352.

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.

Erceg, C, Guerrieri, L. and Gust, C. (2005) Can long run restrictions identify technology shocks?, *Journal of the European Economic Association*, 3, 1237-1278.

Faust, J. (1998), " On the Robustness of Identified VAR Conclusions about Money" , *Carnegie-Rochester Conference Series on Public Policy*, 49, 207-244.

Faust, J. and Leeper, E. (1997), "Do Long Run Restrictions Really Identify Anything?", *Journal of Business and Economic Statistics*, 15, 345-353.

Fernandez Villaverde, J., Rubio Ramirez, J., Sargent, T. and M. Watson (2007) The ABC and (D's) to understand VARs, *American Economic Review*, 97, 1021-1026.

Fry, R. and Pagan, A. (2011), "Sign restrictions in structural vector autoregressions: A Critical Review", *Journal of Economic Literature*, 49, 938-960.

Giannone, D., Primiceri, G. and Lenza, M. (2015). Prior selection for vector autoregression. *Review of Economics and Statistics*, 97, 412-435.

Giannone, D., Primiceri, G. and Lenza, M. (2019). Priors for the long run. *Journal of the American Statistical Association*, 114, 565-580.

Goncalves, S. and L. Kilian (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, 123(1): 89â€“120.

Hamilton, J. and C. Baumeister (2015). Sign Restrictions, Structural VARs and Useful prior information, *Econometrica*, 83, 1963-1999.

Hansen, L. and Sargent, T., (1991), "Two Difficulties in Interpreting Vector Autoregressions", in Hansen, L. and Sargent, T., (eds.), *Rational Expectations Econometrics*, Westview Press: Boulder London.

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

Kilian, L. and Murphy, D.(2012), "Why agnostic sign restrictions are not enough? Understanding the dynamics of oil market VAR models", *Journal of the European Economic Association*, 10, 1166-1188.

Korobilis, D. and Bauwens, L. (2012). Bayesian Methods, Handbook of Research Methods and Applications on Empirical Macroeconomics, edited by Nigar Hashimzade and Michael Thornton, Edward Elgar Publishing.

Lanne, M and Lutkepohl, H. (2008), "Identifying monetary policy shocks via changes in volatility", *Journal of Money, Credit and Banking*, 40, 1131-1149.

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.

Mavroeidis S. (2021), Identification at the ZLB, Econometrica, 89, 2855-2885.

Montiel Olea, L., M. Plagborg Moller and E. Quian (2022) SVAR identification through higher moments: has the simultanous causality problem been solved?, *AER: Papers and Proceedings*, 112, 481-485.

Pagan, A. and Robinson, T. (2019), Implications of partial information for applied macroeconomic modelling, forthcoming, European Economic Review

Rigobon, R. (2003) "Identification through heteroskedasticity", *Review of Economics and Statistics*, 85, 777-792.

Rao, C. R. (1975). Simultaneous estimation of Parameters in Different Linear Models and applications to Biometric Problems. *Biometrics*, 31, 545-554.

Robertson, J. and Tallman, E. (1999). Vector Autoregressions: Forecasting and Reality. *Federal Reserve Bank of Atlanta, Economic Review*, First quarter, 4-18.

Rubio, J., Waggonner, D. and T. Zha (2010) Structural Vector Autoregressions Theory of identification and Algorithms for inference, *Review of Economic Studies*, 77, 665-696.

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

Sims, C. and Zha, T. (1999), Error Bands for Impulse Responses, *Econometrica*, 67, 1113-1155.

Smith, A.F.M. (1973). A General Bayesian Linear Model. *Journal of the Royal Statistical Society*, ser B, 35, 67-75.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics,* 22, 1701-1762.

Uhlig, H. (2005) What are the Effects of Monetary Policy? Results from an agnostic Identification procedure, *Journal of Monetary Economics*, 52, 381-419.

Watson, M. (2020). Comment to Debotoli, Gali and Gambetti, On the empirical irrelevance of the zero lower bound, NBER Macroeconomic Annual, 182-193.

Zellner, A., Hong, (1989) Forecasting International Growth rates using Bayesian Shrinkage and other procedures. *Journal of Econometrics*, 40, 183-202.

1 Preliminaries

Classical and Bayesian analysis differ on a number of issues.

Classical analysis:

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

• Parameters: fixed, unknown quantities.

• Seek *unbiased* estimators because average value of sample estimator converge to true value via some law of large numbers (LLN). *Efficient* estimators preferable because they yield values closer to true parameter.

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

Bayesian analysis:

- Probabilities: degree of (typically subjective) beliefs of an event.

- Parameters: random, with a probability distributions.

- Unbiasedness and efficiency meaningless without a true value. Estimators are chosen to minimize expected loss functions (expectations taken with respect to the posterior distribution), conditional on the data. Probabilities quantify uncertainty.

- 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.

In large samples if the world is classical (under regularity conditions):

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

• Posterior distribution $\xrightarrow{P} N(\alpha_0, (T \times I(\alpha_0))^{-1})$, where $I(\alpha_0)$ is Fisher's information matrix (Asymptotic normality).

Classical and Bayesian analyses differ in small samples and in dealing with unit roots.

Bayesian analysis requires:

- Initial information $\longrightarrow$ Prior distribution.
- A model to organize the data $\longrightarrow$ Likelihood function.
- Prior and Likelihood $\longrightarrow$ Bayes theorem $\longrightarrow$ Posterior distribution.
- Can proceed recursively as new data comes in (mimic economic learning).

2 Bayes Theorem

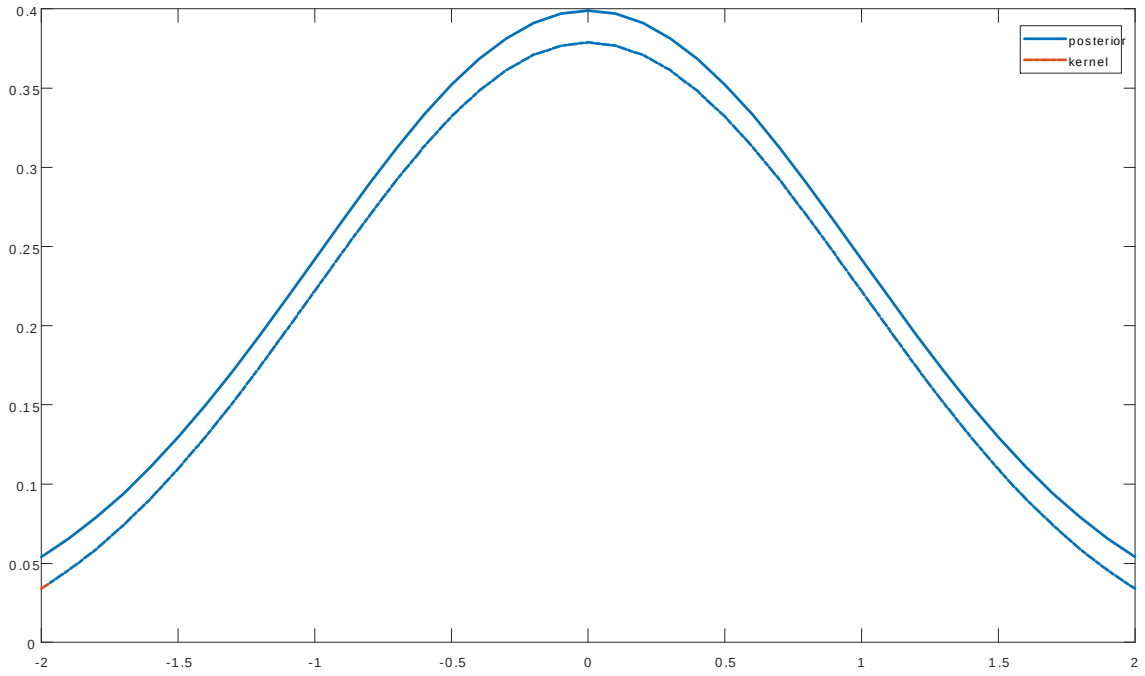
Parameters $\alpha \in A$, A compact subset of R^q. Prior information: $g(\alpha)$. Sample information: $f(y|\alpha) \equiv \mathcal{L}(\alpha|y)$.

• Bayes Theorem:

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

• $g(\alpha|y)$ is the posterior density (the posterior probability of α, after observing y); $\grave{g}(\alpha|y)$ is the posterior kernel, $g(\alpha|y) = \frac{\grave{g}(\alpha|y)}{\int \grave{g}(\alpha|y)d\alpha}$.

• $f(y) = \int f(y|\alpha)g(\alpha)d\alpha$ is the marginal likelihood, a constant (marginal data density). It is independent of α and is a measure of fit. It tells us how good the model is in predicting y, on average over all values of α which have a positive prior probability.



Posterior and posterior kernel are proportional. Kernel can be used to infer the location and spread of the posterior. It can't be used to construct $f(y)$ or draw from the posterior.

• Theorem uses: $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$ (an identity).

• Bayes theorem needs as inputs:

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

b) A model for the data, i.e. choose $f(y|\alpha)$.

• Theorem does not say what $g(\alpha)$ is, but how it should change when y is observed. That is $g(\alpha|y)$ is the updated belief about α once y is observed.

• Bayes Theorem with two (N) samples.

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

$$\breve{g} \equiv f(y_t|\alpha)g(\alpha) = f_2(y_{2t}|\alpha)f_1(y_{1t}|\alpha)g(\alpha) \propto f_2(y_{2t}|\alpha)g(\alpha|y_{1t}) \quad (1)$$

• Posterior for α can obtained in equivalent two ways. Find the posterior using y_t or finding the posterior of using y_{1t} and then, treating $g(\alpha|y_{1t})$ as a next stage prior, finding the posterior using y_{2t}.

• Sequential learning.

- y_{1t}, y_{2t} could be data from different regimes.

- y_{1t}, y_{2t} could be data from different countries.

• Independence is unnecessary. If y_{1t} and y_{2t} are not independent just use $f_2(y_{2t}|\alpha, y_{1t})$ in the first equality in (1).

2.1 Likelihood Selection

- Typically based on a economic or a time series model.

- It must represent well the data.

- Bayesian methods work also if the likelihood is poor (misspecified). It is the interpretation of the results that is affected (posterior estimates are meaningless).

2.2 Prior Selection

• Three basic methods for models which are linear in α.

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

- Location invariant prior: $g(\alpha) =$constant ($=1$ for convenience).

- Scale invariant prior $g(\sigma) = \sigma^{-1}$.

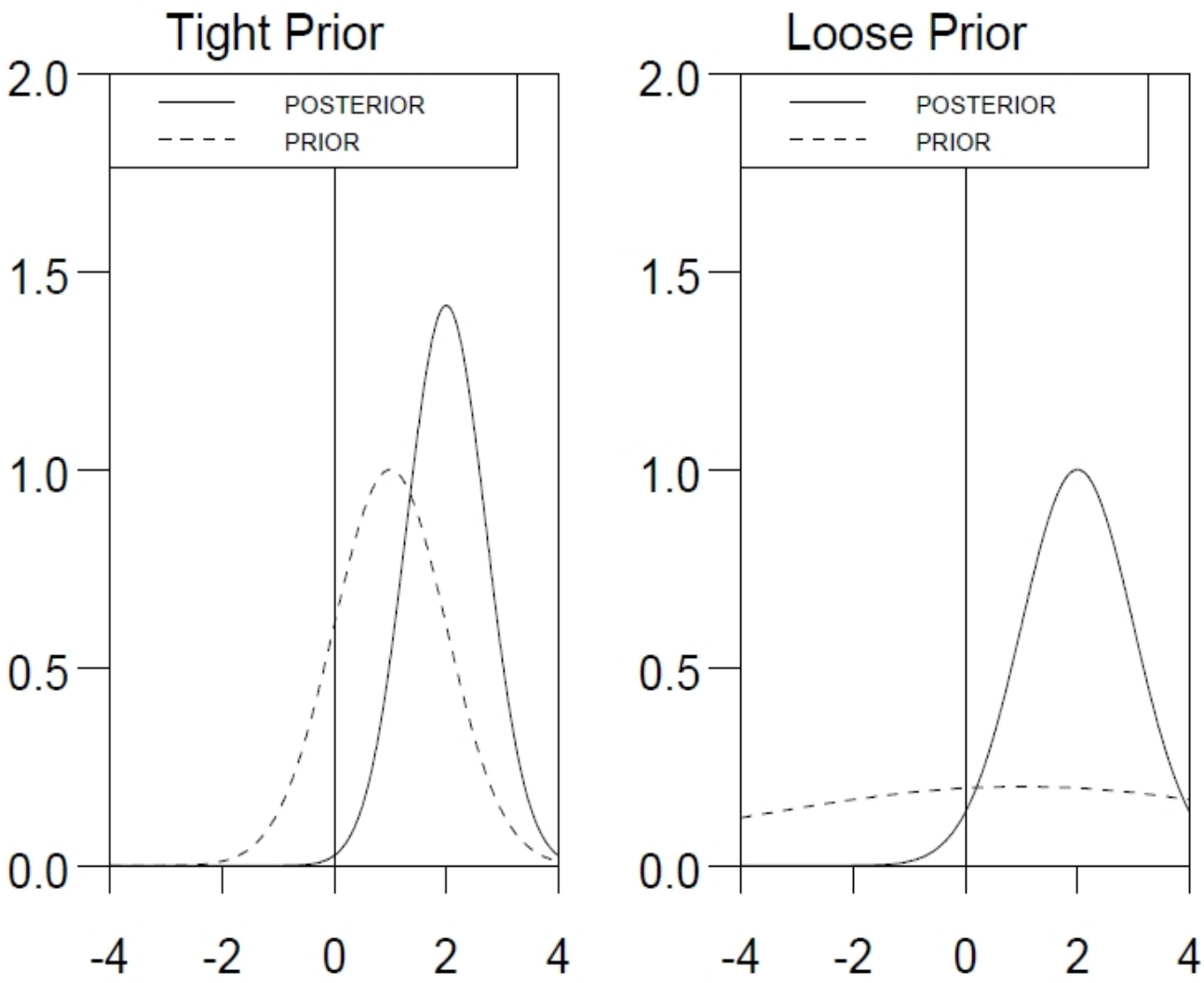
- Location-scale invariant prior : $g(\alpha, \sigma) = \sigma^{-1}$.

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

2) Conjugate Priors

• A prior is conjugate if the posterior has the same functional form as the prior. Thus, the format of the posterior will be analytically available. Only need to figure out its moments.

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


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

3) Objective priors (ML-II approach).

• Set $g(\alpha) = g(\alpha|\theta)$, where θ is a low dimension vector of hyperparameters (e.g. the mean and the variance of the prior of α)

• Marginal likelihood:

$$f(y) = \int \mathcal{L}(\alpha|y)g(\alpha|\theta)d\alpha \equiv \mathcal{L}(y|\theta) \tag{2}$$

Since $\mathcal{L}(\alpha|y)$ is fixed, $\mathcal{L}(y|\theta)$ reflects the plausibility of θ in the data.

• If θ_1 and θ_2 are two vectors and $\mathcal{L}(y|\theta_1) > \mathcal{L}(y|\theta_2)$, there is better support for θ_1. Hence, can estimate the "best" θ using $\mathcal{L}(y|\theta)$.

• 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.

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

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

- Prior here is data-based and not subjective (closer to frequentist ideas).

2.3 Posterior inference

• Objects of interest are typically functions of the posterior $h(\alpha|y)$, e.g.:

- Moments.

- Impulse responses/ variance/historical decompositions.

- Probability of an event.

- Estimate of a latent variable (potential output).

• Typically compute: $E(h(\alpha|y)) = \int h(\alpha)g(\alpha|y)d\alpha$.

• Calculating the expected value under the posterior is consistent with the investigator having a **quadratic** loss function. If the loss function is different, the optimal value of $h(\alpha|y)$ differ.

• With an **absolute** loss function, the optimal value is $h(\alpha_{0.5}|y)$.

• With a **zero-one** loss function the optimal value is $h(\alpha^*|y)$.

• No closed form expression exists for the optimal value of $h(\alpha|y)$ under different loss functions.

General problem: $E(h(\alpha|y))$, $h(\alpha_{0.5}|y)$, $h(\alpha^*|y)$ can not be generally evaluated, since $f(y)$ typically requires high dimensional integration.

• Occasionally, can evaluate the integral analytically. Most of the times, numerical methods are needed.

• If some $g^{AP}(\alpha|y)$ (a numerical approximation to the posterior) is available, compute $E(h(\alpha))$ using:

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

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

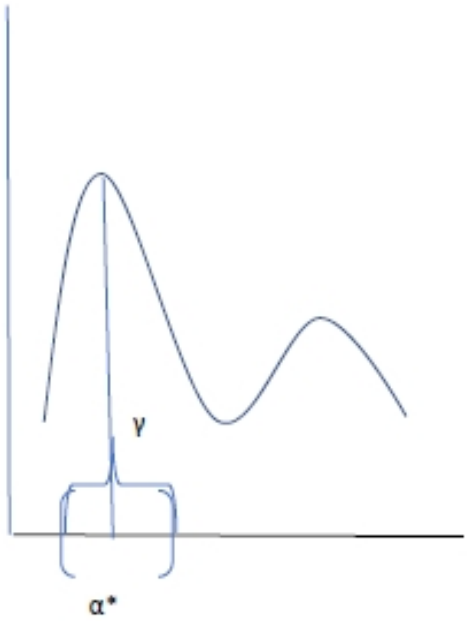
Example 1 *Need to compute* $Pr(\alpha > 0)$*. Draw* α^l *from* $g^{AP}(\alpha|y)$*. If* $\alpha^l > 0$*, set* $h(\alpha^l) = 1$*, else set* $h(\alpha^l) = 0$*. Draw L times and average* $h(\alpha^l)$ *over draws. The result is an estimate of* $Pr(\alpha > 0)$*.*

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

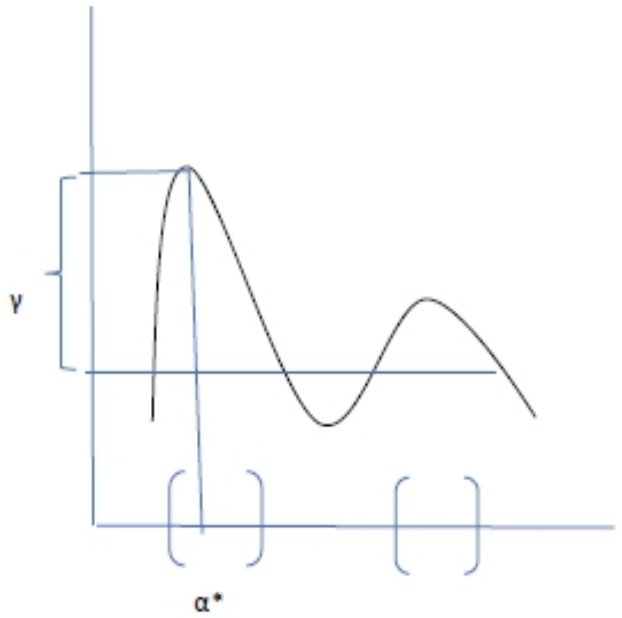
• By a central limit theorem (CLT) the difference between $\frac{1}{L}\sum_l h(\alpha^l)$ and E(h(α)) is normal with zero mean and fixed variance as L grows.

• Numerical standard errors ($\frac{1}{L}\sum_l (h(\alpha^l) - E(h(\alpha)))^2$), or numerical *credible sets* ($h_{0.05\gamma}(\alpha), h_{1-0.05\gamma}(\alpha)$) can be used to measure dispersion (uncertainty in the estimate).

• **Careful: credible sets are different from confidence intervals.**



Classical confidence interval



Bayesian credible set

• Can use the same approach for prediction.

• To compute $Ef(y^{T+\tau}|y^T) = \int f(y^{T+\tau}|y^T, \alpha)g(\alpha|y)d\alpha$ (the predictive density of future observations) or, e.g., $(f_{0.5\gamma}(y^{T+\tau}|y^T), f_{1-0.5\gamma}(y^{T+\tau}|y^T))$ (fan charts) use:

- Draw (iid) α^l from $g^{AP}(\alpha|y)$. Compute $f(y^{T+\tau}|y^T, \alpha^l)$

- Repeat draw L times. Average $f(y^{T+\tau}|y^T, \alpha^l)$ over draws or sort them increasingly and extract percentiles.

Summary

- Inputs: $g(\alpha),\ f(y|\alpha)$.

- Outputs: $g(\alpha|y) \propto f(y|\alpha)g(\alpha)$ (posterior distribution),

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

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

- Prior could be non-informative, conjugate, data based.

- In simple setups, $f(y)$, $g(\alpha|y)$, $f(y^{T+\tau}|y^T)$ are available.In general, need numerical approximations. Posterior statistics computed via Monte Carlo simulations, given analytical or numerical approximations.

3 Posterior simulators

- If $g(\alpha|y)$ is unavailable analytically, choose a $g^{AP}(\alpha|y)$ which is ''close" to ($g(\alpha|y)$ and easy to draw from.

- How do we choose $g^{AP}(\alpha|y)$?

- Normal Approximation

- Basic (non-normal) posterior simulators.

- Markov Chain Monte Carlo (MCMC) simulators.

- Sequential Monte Carlo (SMC) simulators.

3.1 Normal posterior simulators

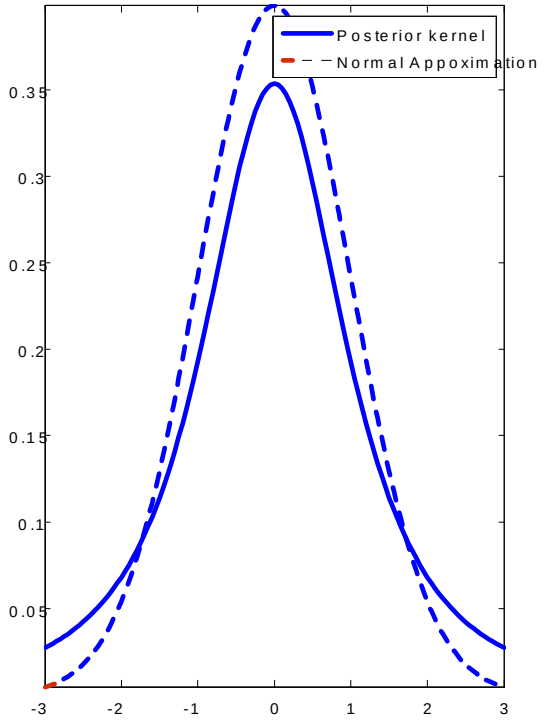
• If T is large, $g(\alpha|y) \approx f(\alpha|y)$, a normal density. If T is not large but $f(\alpha|y)$ is unimodal, roughly symmetric, and α^* is in the interior of A use:

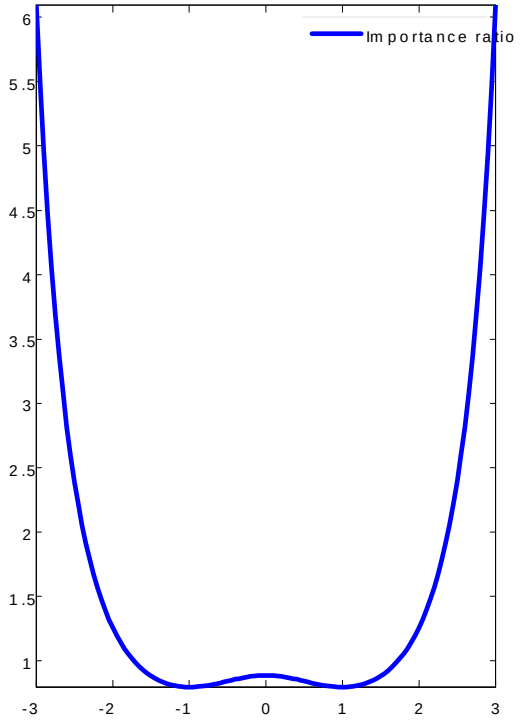
$$g(\alpha|y) \approx N(\alpha^*, \Sigma_{\alpha^*}) \tag{3}$$

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

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

• Need to check that the approximation is accurate. 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.


Posterior kernel
Normal Appoximation
0.35
0.3
0.25
0.2
0.15
0.1
0.05
-3
-2
-1
0
1
2
3

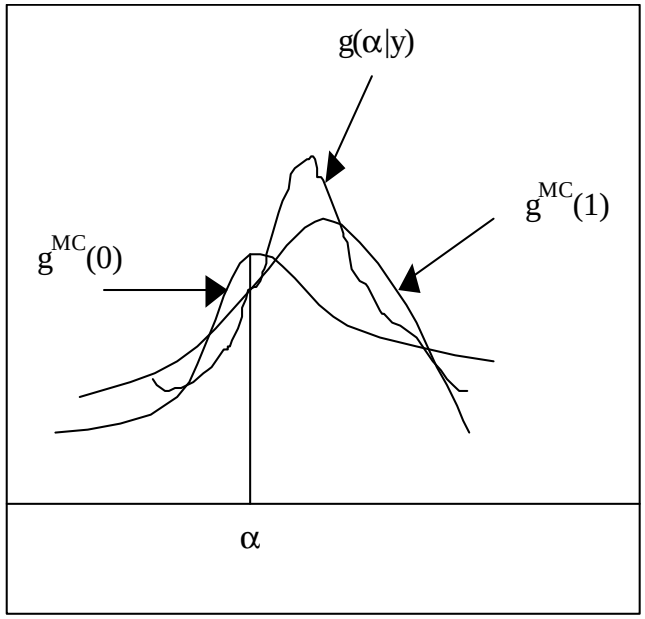

Importance ratio
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
-3
-2
-1
0
1
2
3

3.2 Markov Chain Monte Carlo simulators

• Problem with basic simulators: $g^{AP}(\alpha|y)$ is selected once and for all. If mistakes are made, they stay. With MCMC the location and the shape 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 $\mu \equiv g(\alpha|y)$, choose some initial $\mu(0)$ and a transition P. If conditions are right, iterate from $\mu(0)$ and the limiting distribution is $g(\alpha|y)$, the unknown posterior.



• The ergodicity of P insures consistency and asymptotic normality of estimates of any $h(\alpha)$ obtained with MCMC simulators.

Properties of P needed for MCMC to work:

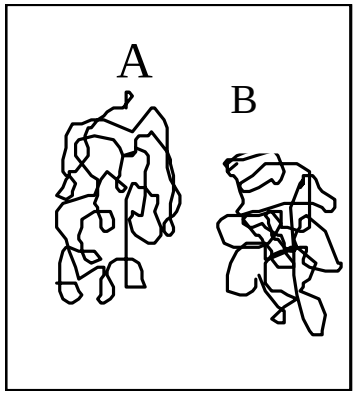
- P is irreducible, i.e. it has no absorbing state.

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

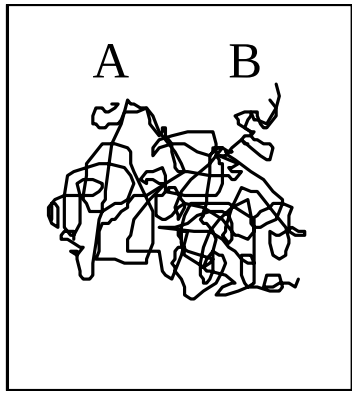
- P it is Harris recurrent, i.e. each cell is visited an infinite number of times with probability one.

- 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 $\alpha_0, A_1 \in A$.

Bad draws



Good draws



• If P has this properties, starting from any $\mu(0)$, the iterations will converge to the unique ergodic distribution.

• Can dispense with the finite number of state and the first order Markov assumption.

MCMC simulation strategy:

- Choose starting values α_0 and $\mu(0)$; choose a P with the right properties.

- Run MC simulations to obtain $g(\alpha|y)$.

- Check convergence.

- If convergence ok, compute E($h(\alpha)$) or percentiles of the distribution of α, etc with the draws of α after convergence is obtained.

Two main algorithms: **Gibbs sampler and Metropolis-Hastings**.

3.2.1 Gibbs sampler

1) Partition $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)$ to obtain $g(\alpha_k|\alpha_{k'}, y, k' \neq k)$ analytically.

2) Choose initial values $\alpha_1^{(o)}, \alpha_2^{(o)}, ..., \alpha_K^{(o)}$.

3) For $l = 1, 2, \ldots$, draw α_k^l as follows

i) $\alpha_1^{(l)}$ from $g(\alpha_1|\alpha_2^{(l-1)}, ..., \alpha_K^{(l-1)}, y)$.

ii) $\alpha_2^{(l)}$ from $g(\alpha_2|\alpha_1^{(l)}, ..., \alpha_K^{(l-1)}, y)$.

iii) $\vdots$

iv) $\alpha_K^{(l)}$ from $g(\alpha_k|\alpha_1^{(l)}, ..., \alpha_{K-1}^{(l)}, y)$.

4) Repeat step 3) $nL + \bar{L}$ times.

Drawing in this fashion produces a sequence which is the realization of a Markov chain with transition

$$P(\alpha^l, \alpha^{l-1}) = \prod_{k=1}^{K} g(\alpha_k^l | \alpha_{k'}^{l-1}(k' > k), \alpha_{k'}^l(k' < k), y) \tag{4}$$

- P in (4) satisfies the conditions for existence, uniqueness, convergence.

- If $\bar{L}$ is large, $\alpha_k^{L+j}, k = 1, \ldots K$ is a draw from $g(\alpha_k|y), j = 1, 2, \ldots$.

- With $\alpha^{L+j} = (\alpha_1^{L+j}, \ldots, \alpha_k^{l+j})$, compute $E(h(\alpha))$.

Intuition for Gibbs sampler come from integration by parts:

Choose α_2^0. Draw α_1^1 from $g(\alpha_1|\alpha_2^0, y)$, then draw α_2^1 from $g(\alpha_2|\alpha_1^1, y)$, α_1^2 from $g(\alpha_1|\alpha_2^1, y)$, etc. Each step is a draw by parts from $g(\alpha_1\alpha_2|y)$.

Example 2 *Suppose* $f(x,y) \propto \frac{n!}{x!(n-x)!} y^{x+\alpha_0-1}(1-y)^{n-x+\alpha_1-1}$, $x = 0,1,\ldots n$, $0 \leq y \leq 1$ *(binomial density for (x,y)) and the consider marginal of* $f(x)$*. Direct integration leads to*

$$f(x) \propto \frac{n!}{x!(n-x)!} \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \frac{\Gamma(x+\alpha_0)\Gamma(n-x+\alpha_1)}{\Gamma(\alpha_0+\alpha_1+n)}$$

which is the beta-binomial distribution. Hence, $f(x|y)$ *is binomial with parameters* (n,y)*, and* $f(y|x)$ *is Beta with parameters* $(x+\alpha_0, n-x+\alpha_1)$.

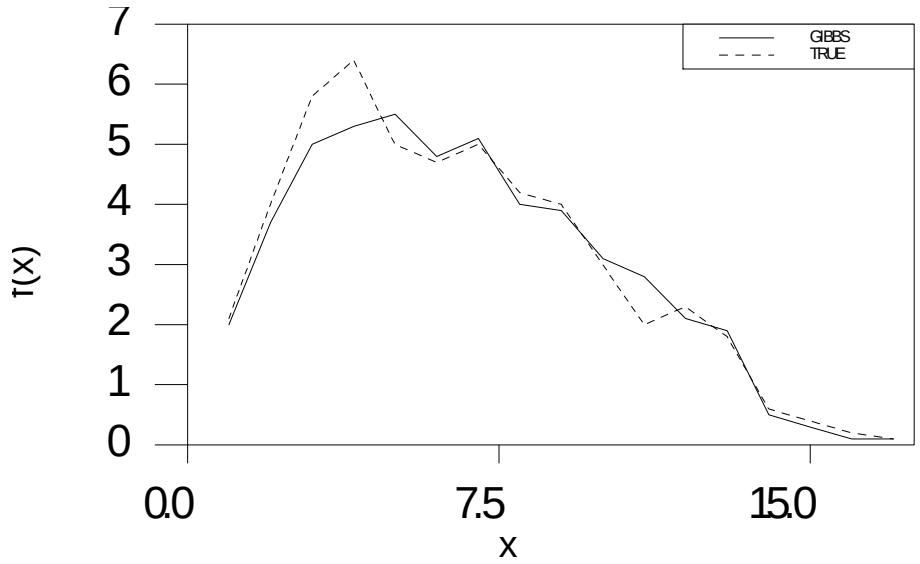


Figure: true/ Gibbs sampling marginal distribution for x with $L = 500,\ n = 100,\ \alpha_0 = 2,\ \alpha_1 = 4$ and $\bar{L} = 20$.

Example 3 *(seemingly unrelated regression) Let*

$$y_{it} = x_{it}'\alpha_i + e_{it}$$

$e_t = (e_{1t}, \ldots e_{mt})' \sim \mathbb{N}(0, \Sigma_e)$, $i = 1, \ldots, m$, $t = 1, \ldots, T$; α_i $k \times 1$ *vector. Stacking observations*

$$y_t = x_t\alpha + e_t$$

$y_t = (y_{it}, \ldots, y_{mt})', x_t = \mathsf{diag}(x_{it}', \ldots, x_{mt}')$, $\alpha = (\alpha_1', \ldots \alpha_m')'$ *is* $mk \times 1$ *vector.*

Suppose $g(\alpha, \Sigma^{-1}) = g(\alpha)g(\Sigma^{-1})$. *Posterior kernel is:*

$$\breve{g}(\alpha, \Sigma^{-1}|y) = g(\alpha)g(\Sigma^{-1})|\Sigma^{-1}|^{0.5T} \exp\{-0.5 \sum_t (y_t - x_t\alpha)'\Sigma^{-1}(y_t - x_t\alpha)\} \tag{5}$$

- *The target density is* $g(\alpha, \Sigma^{-1}|y) = \frac{\breve{g}(\alpha, \Sigma^{-1})}{\int \breve{g}(\alpha, \Sigma^{-1}) d\alpha d\Sigma}$.

- *if prior for α, Σ^{-1} is Normal-Wishart, conditional posteriors:*

$$(\alpha|Y, \Sigma^{-1}) \sim \mathbb{N}(\tilde{\alpha}, \tilde{\Sigma}_\alpha)$$

$$(\Sigma^{-1}|\alpha, Y) \sim W(T + v_0, \tilde{\Sigma})$$

$\tilde{\alpha} = \tilde{\Sigma}_\alpha^{-1}(\bar{\Sigma}_\alpha \bar{\alpha} + \sum_t x_t \Sigma_e^{-1} y_t)$; $\tilde{\Sigma}_\alpha = (\bar{\Sigma}_\alpha^{-1} + \sum_t x_t \Sigma^{-1} x_t)^{-1}$; $\tilde{\Sigma} = (\Sigma^{-1} + \sum_t (y_t - x_t \alpha_{ols})(y - x_t \alpha_{ols})')^{-1}$, $(\bar{\alpha}, \bar{\Sigma}_\alpha)$ *are the prior mean and variance,* Σ *is the prior scale matrix and* α_{ols} *is the OLS estimator of* α.

- *Use* α *and* Σ *as two Gibbs sampler blocks. When* L *is large obtain a sample such that* $\alpha^L \sim g(\alpha|y_1, \ldots y_t)$; $\Sigma^{-1(L)} \sim g(\Sigma^{-1}|y_1, \ldots, y_t)$

- *Can use this setup for VARs (a seemingly unrelated regression model).*

3.2.2 Implementation issues

A) Draw a sample with $nL + \bar{L}$ observations; throw away $\bar{L}$ of them. Keep elements $(L, 2L, \ldots, n * L)$ (to eliminate correlation in the draws: MC theorem is for iid draws).

B) How do you check convergence?

• Start from different α^0. Check if for the same $\bar{L}$ the remaining sample has same mean, variance, etc.

• Fix α_0, check for different $\bar{L}$ is mean, variance, etc. are the same ($\rightarrow$ CUMSUM statistic for mean, variance, etc.).

• For simple problems $\bar{L} \approx 50, L \approx 200$. For large models $\bar{L} \approx 100,000 - 200,000, L \approx 500,000$. If multiple modes are present, L could be larger.

C) Model comparisons.

- Compute marginal likelihood $f(y)$ of each model.

- Compute Log Bayes factors/ Log Posterior odds

$$BF = \frac{ML(m_1)}{ML(m_2)} \tag{6}$$

$$PO = BF * \frac{g(m_1)}{g(m_2)} \tag{7}$$

- If log $BF \geq 3(10)$, model 1 preferred (strongly preferred) .

4 What are VARs?

• VARs are multivariate linear time series models:

$$y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_q y_{t-q} + e_t \qquad e_t \sim (0, \Sigma_e) \quad (8)$$

where y_t is a $m \times 1$ vector; A_j are full rank, $m \times m$ matrices, $j = 1, \ldots q$; Σ_e full rank $m \times m$ matrix, A_0 deterministic components (constant, seasonals, etc.).

• e_t are the innovations (the news) in y_t, i.e. $e_t = y_t - E(y_t | I_{t-1})$, I_{t-1} information set at t.

• Can be derived from the Wold representation and some invertibility plus truncation conditions (see appendix)

• Small open economy version (VARX):

$$\begin{aligned} y_t &= A_0 + A_1 y_{t-1} + \ldots + A_q y_{t-q} + B_1 x_t + \ldots + B_p x_{t-p-1} + e_t \quad e_t \sim (0, \Sigma_e) \\ x_t &= G_0 + G_1 x_{t-1} + \ldots + G_q x_{t-q} + v_t \end{aligned} \tag{9}$$

where y_t endogenous, x_t is a vector of strictly exogenous (foreign) variables.

• Two country VAR:

$$\begin{aligned} y_t &= A_0 + A_1 y_{t-1} + \ldots + A_q y_{t-q} + B_1 x_t + \ldots + B_p x_{t-p-1} + e_t \quad e_t \sim (0, \Sigma_e) \\ x_t &= G_0 + F_1 y_{t-1} + \ldots + F_q y_{t-q} + G_1 x_{t-1} + \ldots + G_q x_{t-q} + v_t \end{aligned} \tag{10}$$

where y_t are endogenous variables of country 1 and x_t endogneous variables of country 2.

• Panel VAR:

$$w_t = M_0 + M_1 w_{t-1} + M_2 w_{t-2} + \ldots + M_q w_{t-q} + u_t \qquad u_t \sim (0, \Sigma_u) \tag{11}$$

$w_t = (y_{1t}, y_{2t}, \ldots y_{nt})$; M_j and Σ_u are $qn \times qn$ matrices.

• Advantages of VAR setup:

- Every variable is endogenous (no incredible exogeneity assumptions); they depend on all the others (no incredible exclusion restrictions).

- Simple to estimate.

• Disadvantages of VAR setup:

- No economic interpretation of the dynamics is possible Σ_e is not diagonal.

- Potentially difficult to relate VARs and theoretical models (these have a VARMA format).

Notation

- Lag operator, VAR and MAR (eliminating deterministic components)

$$\begin{aligned} y_t &= A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_q y_{t-q} + e_t \qquad e_t \sim (0, \Sigma_e) \quad (12) \\ &= A_1 \ell y_t + A_2 \ell^2 y_t + \ldots + A_q l^q y_t + e_t \quad (13) \end{aligned}$$

$$\begin{aligned} (I - A_1\ell - A_2\ell^2 - \ldots - A_q\ell^q) y_t &\equiv A(\ell) y_t = e_t \quad \text{VAR} \quad (14) \\ y_t = A(\ell)^{-1} e_t &\equiv D(\ell) e_t \quad \text{MAR} \quad (15) \end{aligned}$$

$\ell^k y_t \equiv y_{t-k}$.

- Alternative VAR representations:

$$y_t = A(\ell) y_{t-1} + e_t \quad (16)$$

$y_t,\ e_t$ $m \times 1$ vectors; A_j is $m \times m$,.

Companion format

- Transform a m-variable VAR(q) into a mq-variable VAR(1).

Example 4 *Consider a VAR(3). Let* $\mathbf{Y}_t = [y_t, y_{t-1}, y_{t-2}]'$; $\mathbf{E}_t = [e_t, 0, 0]'$;

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & A_3 \\ I_m & 0 & 0 \\ 0 & I_m & 0 \end{bmatrix} \quad \mathbf{\Sigma}_E = \begin{bmatrix} \mathbf{\Sigma}_e & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Then the VAR(3) can be rewritten as

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{E}_t \quad \mathbf{E}_t \sim \mathbf{N}(0, \mathbf{\Sigma}_E) \tag{17}$$

where $\mathbf{Y}_t, \mathbf{E}_t$ *are* $3m \times 1$ *vectors and* $\mathbf{A}$ *is* $3m \times 3m$.

Simultaneous equations format

1) Let $x_t = [y_{t-1}, y_{t-2}, \ldots]$; $\mathbf{X} = [x_1, \ldots, x_T]'$ (a $T \times mq$ matrix), $\mathbf{Y} = [y_1, \ldots, y_T]'$ (a $T \times m$ matrix); and if $\mathbf{A} = [A_1', \ldots A_q']'$ is a $mq \times m$ matrix

$$\mathbf{Y} \;=\; \mathbf{XA} + \mathbf{E} \tag{18}$$

2) Let i indicate the subscript for the $i-th$ column vector. The equation for variable i is $y_i = x\alpha_i + e_i$. Stacking the columns of y_i, e_i into where $mT \times 1$ vectors we have

$$y \;=\; (I_m \otimes x)\alpha + e \equiv X\alpha + e \tag{19}$$

How do you estimate (classical) VARs?

• Choose the lag length using some criteria (LR, AIC, BIC) and deterministic components (constant, polynomial trends) to be included.

• Check for structural breaks in the data.

• If variables are integrated may choose to use differenced data or leave it in level if cointegrated.

• Because the likelihood function of a VAR, conditional on the initial conditions is proportional to the sum of square errors: ML=OLS

• Because the regressors are the same in each equation, single equation OLS = system OLS (see appendix for details)

How do you estimate (Bayesian) VARs?

- Choose the generous lag length for the variables in level and use a constant (polynomial trends not usually included, see Giannone, et al., 2019)

- Specify a prior distribution for coefficients and covariance matrix and a likelihood function (choose the distribution of the error term)

- Compute posterior distribution of coefficients and covariance matrix analitically or via MCMC.

- Summarize the posteriors with a location and a spread measure.

- No pretesting of any sort ! (see later)

4.1 Structural VARs

VARs are reduced form models:

- Shocks e_t are linear combination of meaningful economic disturbances.
- Can't be used for certain policy analyses (Lucas critique).

• A SVAR is a linear dynamic structural model of the form:

$$\mathcal{A}_0 y_t = \mathcal{A}_1 y_{t-1} + \ldots + \mathcal{A}_q y_{t-q} + \varepsilon_t \qquad \varepsilon_t \sim (0, \Sigma_\varepsilon) \tag{20}$$

where Σ_ϵ is diagonal, ϵ_t structural shocks. Its reduced form (VAR) is:

$$y_t = A_1 y_{t-1} + \ldots + A_q y_{t-q} + e_t \qquad e_t \sim (0, \Sigma_e) \tag{21}$$

• (21) easy to estimate. We want to go from (21) to (20). Since $A_j = \mathcal{A}_0^{-1}\mathcal{A}_j,\ e_t = \mathcal{A}_0^{-1}\varepsilon_t$, we just need to find $\mathcal{A}_0$.

Note that (21) and (20) imply

$$\mathcal{A}_0^{-1}\Sigma_\epsilon \mathcal{A}_0^{\prime -1} = \Sigma_e \tag{22}$$

• SVAR problem: restrict and estimate $\mathcal{A}_0$, assuming $\Sigma_\epsilon = I$.

• To recover unknown elements in $\mathcal{A}_0$ from (22) we need at least as many equations as unknowns.

• Order condition: If the VAR has m variables, need $m(m-1)/2$ restrictions because there are m^2 free parameters on the left hand side of (22) and only $m(m+1)/2$ parameters in Σ_e ($m^2 = m(m+1)/2 + m(m-1)/2$).

• Exactly identified vs. overidentified (number of restrictions larger than $m(m-1)/2$), see Canova and Forero (2015) for overidentification.

• Rank condition: (see Hamilton, 1994, p.332-335) rank($\mathcal{A}_0^{-1}\Sigma_\epsilon\ \mathcal{A}_0^{'-1}$) = rank($\Sigma_e$).

• Rank and order conditions valid only for "local identification" (conditions need to be checked at one specific point).

• For global identification conditions see Rubio et al. (2010): work only under just identification.

Example 5 *i) Cholesky decomposition of* Σ_e *has exactly* $m(m-1)/2$ *zeros restrictions.* $\mathcal{A}_0^{-1}$ *is lower triangular and variable* i *does not affect variable* $i-1$ *simultaneously, but it affects variable* $i+1$*.*

ii) $y_t = [GDP_t, P_t, i_t, M_t]$*. Then we need at least 6 restrictions for local identification, e.g.*

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ \alpha_{01} & 1 & 0 & \alpha_{02} \\ 0 & 0 & 1 & \alpha_{03} \\ \alpha_{04} & \alpha_{05} & \alpha_{06} & 1 \end{bmatrix}.$$

How do you estimate a SVAR?

- Get (classical or Bayesian) estimates of $A(\ell)$ and Σ_e.

- Assume $\Sigma_\epsilon = I$ and estimate the free parameters of $\mathcal{A}_0$.

- Use $\mathcal{A}(\ell) = \mathcal{A}_0 A(\ell)$ to trace out structural dynamics in response to the structural shocks $\mathcal{A}_0 e_t = \epsilon_t$.

- Unless the system is in a Cholesky format, we need to estimate $\mathcal{A}_0$ by maximum likelihood.

4.2 Identification approaches

• Traditional: Cholesky, contemporaneous (Sims, 1980), long run (Blanchard and Quah, 1989),

• Newer: sign (Canova and De Nicolo' (2002), Faust (1998), Uhlig (2005)), quantity (DeSantis and Zimic, 2018), medium run forecast error variance (Barski and Sims, 2012).

• Latest: high frequency (Gertler and Karadi, 2015); IV (Martens and Ravn, 2013); narrative (Romer and Romer, 2004); sign and narrative (Antolin and Rubio, 2019, Been Zeev, 2018); heteroskedasticity (Lanne and Lutkepohl, 2008, Brunnemeier et al., 2021), higher moments (Lanne et al., 2017; Gourieroux et al. 2020), regime switching (Mavroedis, 2021).

Short and long run restrictions

• VAR/SVAR in MAR/SMAR format: (assuming unit roots and no cointegration)

$$\Delta y_t = D(\ell)e_t = D(1)e_t + D^*(\ell)\Delta e_t \quad (23)$$

$$\Delta y_t = \mathcal{D}(\ell)\mathcal{A}_0\epsilon_t = \mathcal{D}(\ell)(1)\mathcal{A}_0\epsilon_t + \mathcal{D}^*(\ell)\mathcal{A}_0\Delta\epsilon_t \quad (24)$$

where $D(\ell) = (I - A(\ell)\ell)^{-1}$, $\mathcal{D}(\ell) = (1 - \mathcal{A}(\ell)\ell)^{-1}$, $D^*(\ell) \equiv \frac{D(\ell)-D(1)}{1-\ell}$, $\mathcal{D}^*(\ell) \equiv \frac{\mathcal{D}(\ell)-\mathcal{D}(1)}{1-\ell}$. Matching coefficients: $\mathcal{D}(\ell)\mathcal{A}_0\epsilon_t = D(\ell)e_t$.

• Separating permanent and transitory components we have

$$\mathcal{D}(1)\mathcal{A}_0\epsilon_t = D(1)e_t \quad (25)$$

$$\mathcal{A}_0\Delta\epsilon_t = \Delta e_t \quad (26)$$

If y_t is stationary, $\mathcal{D}(1) = D(1) = 0$ and only (26) is available (in level).

• Two types of restrictions to estimate $\mathcal{A}_0$: short and long run.

Example 6 *In a bivariate VAR imposing (25) requires one restriction. Suppose that* $\mathcal{D}(1)^{12} = 0$ *(*ϵ_{2t} *has no long run effect on* y_{1t}*). If* $\Sigma_\epsilon = I$*, the three elements of* $\mathcal{D}(1)\mathcal{A}_0\Sigma_\epsilon\mathcal{A}_0'\mathcal{D}(1)'$ *can be obtained from the Cholesky factor of* $D(1)\Sigma_e D(1)'$*.*

• Blanchard-Quah (1999), use (25)-(26). If $y_t = [\Delta y_{1t}, y_{2t}]$, $(m \times 1)$; y_{1t} are I(1); y_{2t} are I(0) and $y_t = D_0 + D(\ell)\epsilon_t$, where $\epsilon_t \sim iid(0, \Sigma_\epsilon)$

$$\begin{pmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{pmatrix} = \begin{pmatrix} D_{01} \\ 0 \end{pmatrix} + \begin{pmatrix} D_1(1) \\ 0 \end{pmatrix} \epsilon_t + \begin{pmatrix} (1-\ell)D_1^\dagger(\ell) \\ (1-\ell)D_2^\dagger(\ell) \end{pmatrix} \epsilon_t \tag{27}$$

where $D_1(1) = [1, 0]$, $D^\dagger(\ell) = D(\ell) - D(1)$, y_{2t} is any set of stationary variables.

4.3 Problems with standard identification approaches

Example 7

$$p_t = a_{11}e_t^s \tag{28}$$
$$y_t = a_{21}e_t^s + a_{21}e_t^d \tag{29}$$

• *Price is set prior to knowing demand shocks. Equivalent to estimating* p *on lagged* p *and lagged* y *(this gives* e_t^s*) and then estimating* y *on lagged* y*, on current and lagged* p *(this gives* e_t^d*).*

$$y_t = a_{11}e_t^s \tag{30}$$
$$p_t = a_{21}e_t^s + a_{21}e_t^d \tag{31}$$

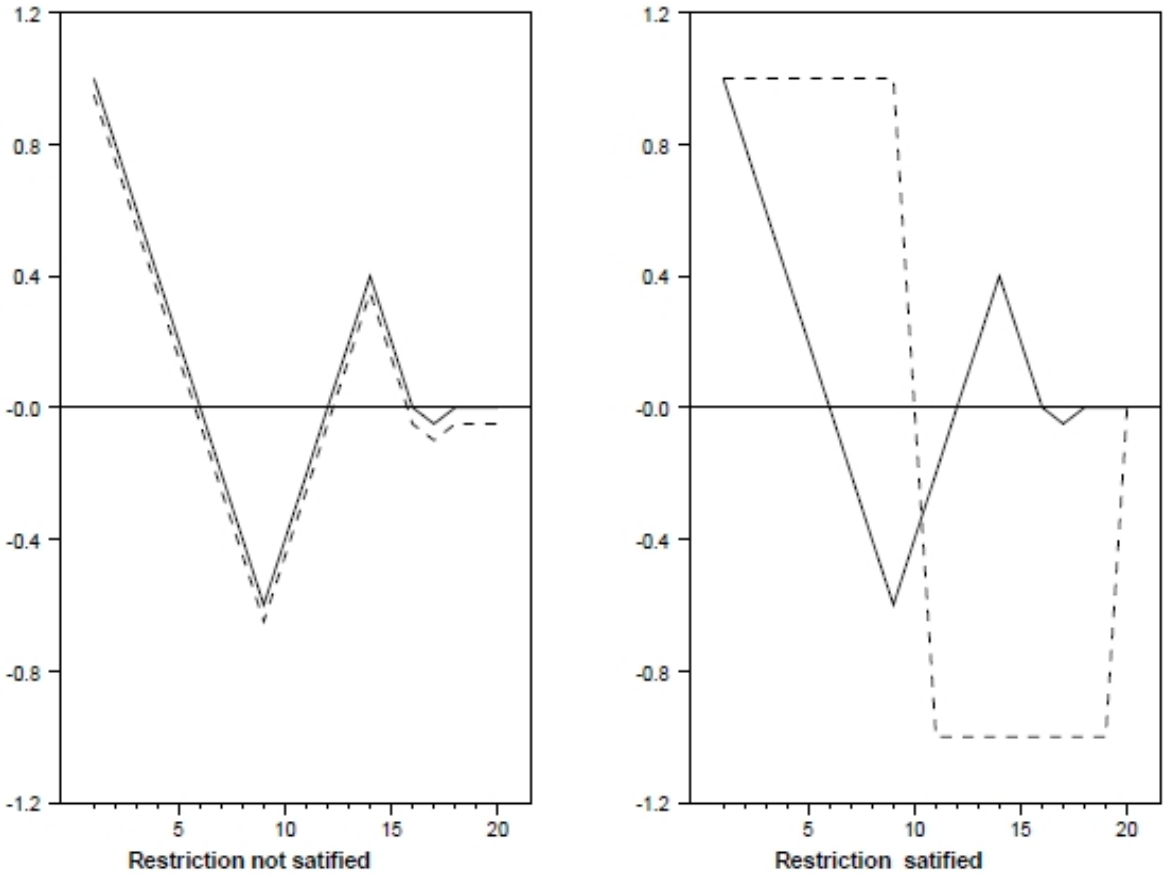
• *Quantity set prior to knowing demand shocks. Equivalent to estimating* y *on lagged* y *and lagged* p *(this gives* e_t^s*) and then estimating* p *on lagged* p*, on current and lagged* y *(this gives* e_t^d*).*

• Many Cholesky systems - which one to choose?

• **Without a structural model, it is difficult to choose between Cholesky systems.**

• Cooley-LeRoy (1985): unless strong restrictions are imposed (i.e. on the timing of information) dynamic RE models do not have a Cholesky structure.

- Problems with long run restrictions 1: Faust-Leeper (1997).

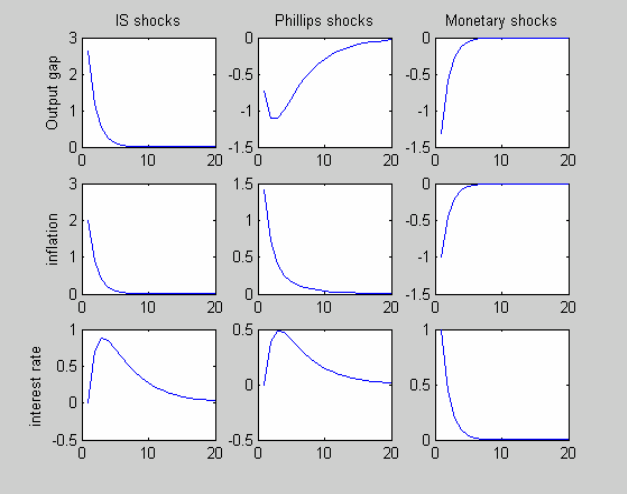


• Problems with long run restrictions 2: Erceg, et. al (2005): long run restrictions give poor identification in small samples.

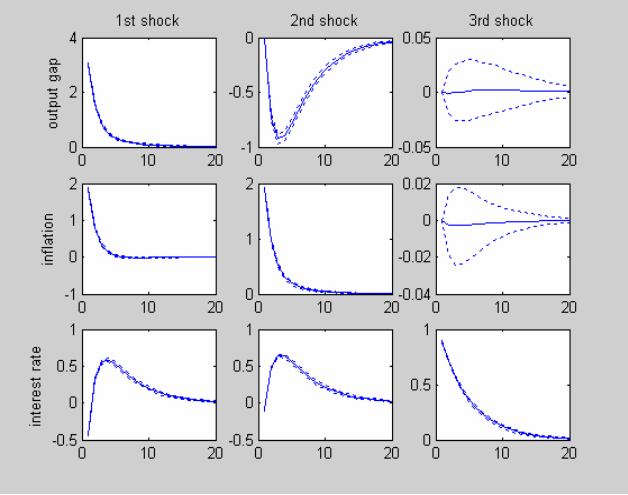
• Problems with long run restriction 3: Chari, et. al. (2006) potentially important truncation bias due to the estimation of a VAR(q), q finite.

• Problems with short run restrictions: Canova and Pina (2005), Fuerst, Carlstrom, Paustian (2009).

The DGP is a 3 equations New-Keynesian model



True responses



Inertial responses

Sign restrictions

Example 8 *i) Aggregate supply shocks:* $Y \uparrow$, $Inf \downarrow$*; aggregate demand shocks:* $Y \uparrow$, $Inf \uparrow \rightarrow$ *demand and supply shocks impose different sign restrictions on cov*(Y_t, INF_s)*. Restrictions shared by a large class of models with different foundations. Use these for identification.*

ii) Monetary Shocks: response of Y *is humped shaped, dies out in 3-4 quarters* $\rightarrow$ *shape restrictions on cov*(Y_t, i_s)*. Use these for identification.*

- Practical implementation of sign restrictions (Canova-De Nicolo',2002)

• Orthogonalize $\Sigma_e = \tilde{\mathcal{P}}\tilde{\mathcal{P}}'$ (e.g. Choleski or eigenvalue-eigenvector decomposition). Check if the shocks produce the required sign pattern for $y_{it}, i = 1, 2, \ldots$. If not:

• For any $\mathcal{H}$: $\mathcal{H}\mathcal{H}' = I$, $\Sigma_e = \tilde{\mathcal{P}}\mathcal{H}\mathcal{H}'\,\tilde{\mathcal{P}}' = \hat{\mathcal{P}}\hat{\mathcal{P}}'$. Check if any shock under new decomposition $\hat{\mathcal{P}}$ produces the required pattern for y_{it}. If not choose another $\mathcal{H}$, and repeat.

• Stop when you find a ε_{jt} with the right characteristics or compute the mean/ median (and s.e.) of the statistics of interest for all ε_{jt}^{l} satisfying the restrictions, where l is the number of shocks found with the right features.

• Many possible $\mathcal{H}$. Let $\mathcal{H} = \mathcal{H}(\omega)$, $\omega \in (0, 2\pi)$. $\mathcal{H}(\omega)$ are called rotation (Givens) matrices.

Example 9 *Let M=2. Then* $\mathcal{H}(\omega) = \begin{bmatrix} cos(\omega) & -sin(\omega) \\ sin(\omega) & cos(\omega) \end{bmatrix}$ *or* $\mathcal{H}(\omega) = \begin{bmatrix} cos(\omega) & sin(\omega) \\ sin(\omega) & -cos(\omega) \end{bmatrix}$. *Varying* $\omega \in (0, 2\pi)$, *we trace out all possible structural MA representations that could have generated the data.*

• Rotation matrices impractical in large scale systems (too many combinations of two variables to try).

4.4 Sign restrictions in large systems

• Use a QR decomposition (Rubio et al, 2010).

1. Start from some orthogonal representation $y_t = D(\ell)\epsilon_t$

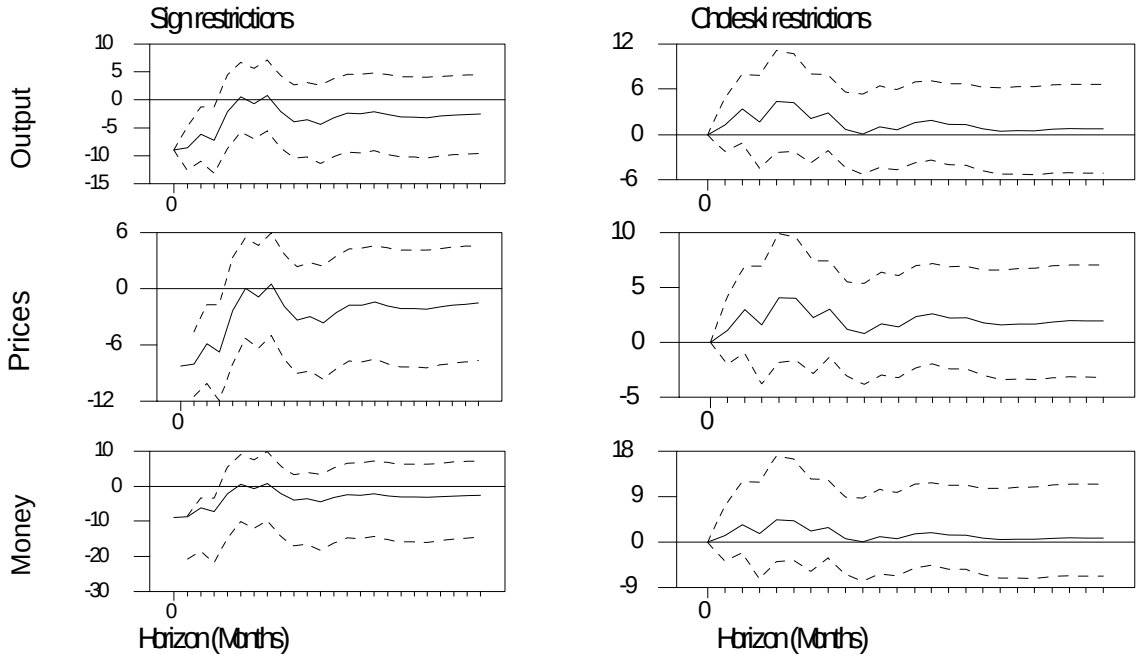
2. Draw an $m \times m$ matrix G from N(0,1). Find $G = QR$.

3. Compute responses as $D'(\ell) = D(\ell)Q$. Check if restrictions are satisfied (Note: shocks are now $R\epsilon_t$.

4. Repeat 2.-3. until L draws are found.

Fast even in large dimensional systems.

Example 10 *Comparing responses to US monetary shocks 1964-2001.*



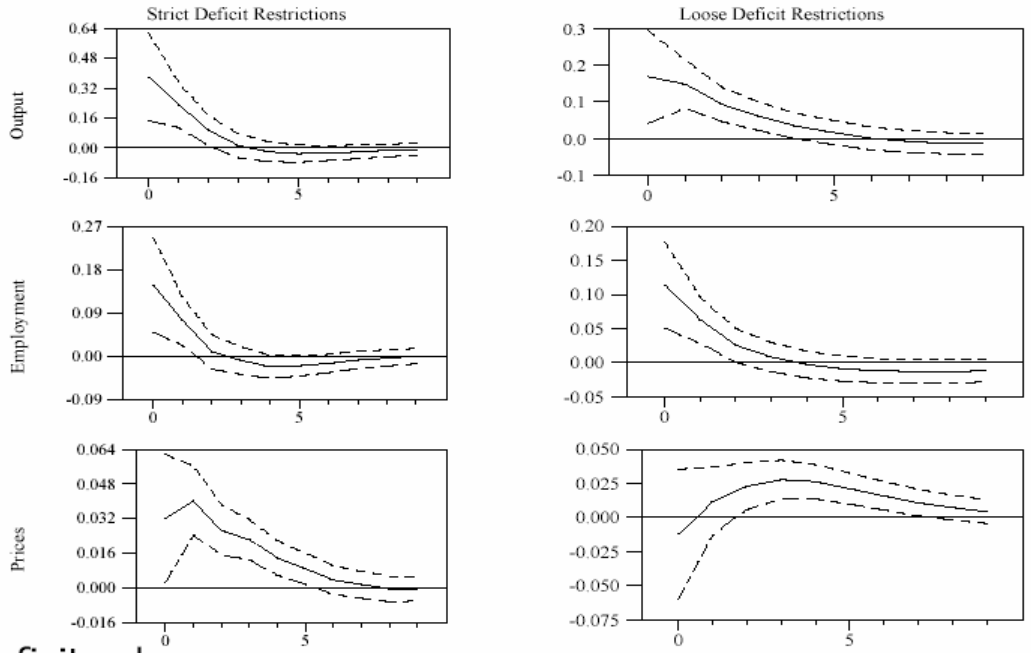
Example 11 *Studying the effects of fiscal shocks in US states: 1950-2005.*

	corr(G,Y)	corr(T,Y)	corr(G, DEF)	corr(T,DEF)	corr(G,T)
G shocks	> 0		> 0		> 0
BB shocks	< 0		$= 0$		$= 1$
Tax shocks		< 0		< 0	$= 0$

Table 1: Identification restrictions

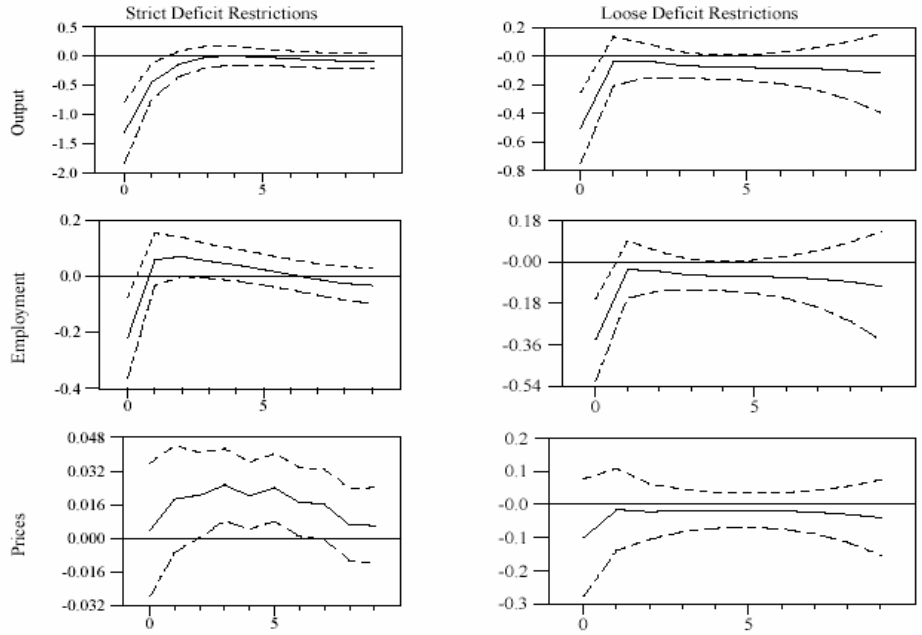
The results: Is the transmission of fiscal shocks different?

G shocks



The results: Is the transmission of fiscal shocks different?

BB shocks



Deficit rules

• Sign restrictions are weak: they identify a set not a point. Large uncertainty in the results compared with Cholesky SVARs.

• Sign restrictions may be poor if shock signal is weak (Canova and Paustian, 2011).

• Performance of sign restrictions improves if more variables are restricted and if more shocks are jointly identified (Canova and Paustian, 2011) even if they are not of interest.

• Use **robust** DSGE restrictions (see Dedola and Neri, 2007, Pappa, 2009, Persmann and Straub, 2009, Lippi and Nobili, 2011, etc.).

• Rank and order conditions do not apply here.

• Review of this literature: Fry and Pagan (2011), Kilian (2013).

Problems with Bayesian sign restrictions

Baumeister and Hamilton (2015): Bayesian inference in sign restricted SVARs may not be what you think it is.

• When sign restrictions are used, priors on contemporaneous parameters matter even when $T \to \infty$. This is not the case if contemporanous parameters are point identified.

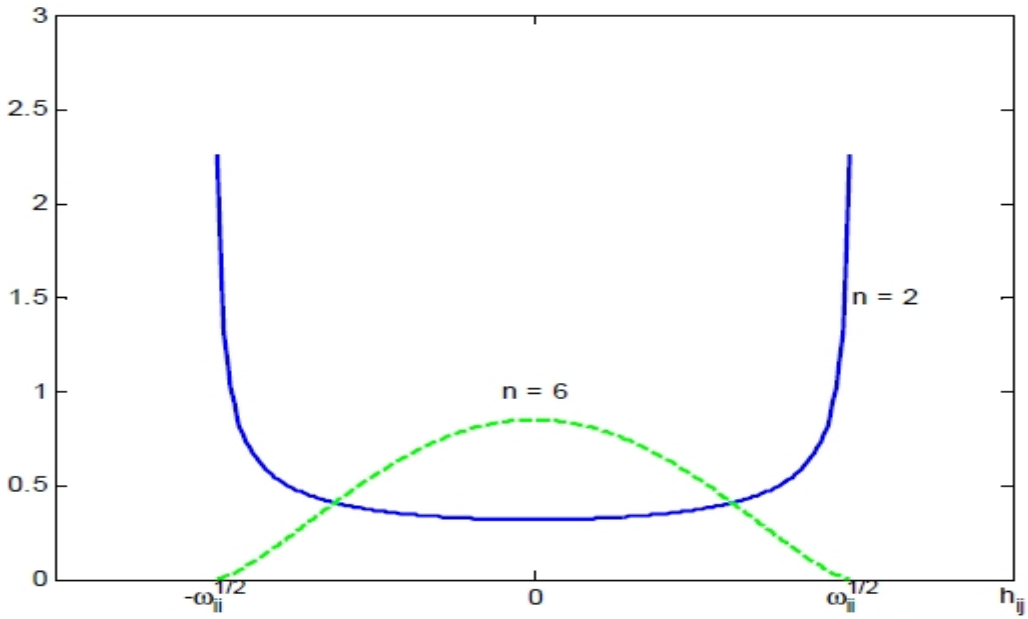
• A uniform (Haar) prior on the set of rotations used to generate candidate impulse responses implies very informative priors on elasticities or impulse responses (which are non-linear functions of the rotation parameters)

• Should use priors on elasticities or instantaneous responses directly.

• Why does a uniform prior for the rotations imply informative priors on economic objects?

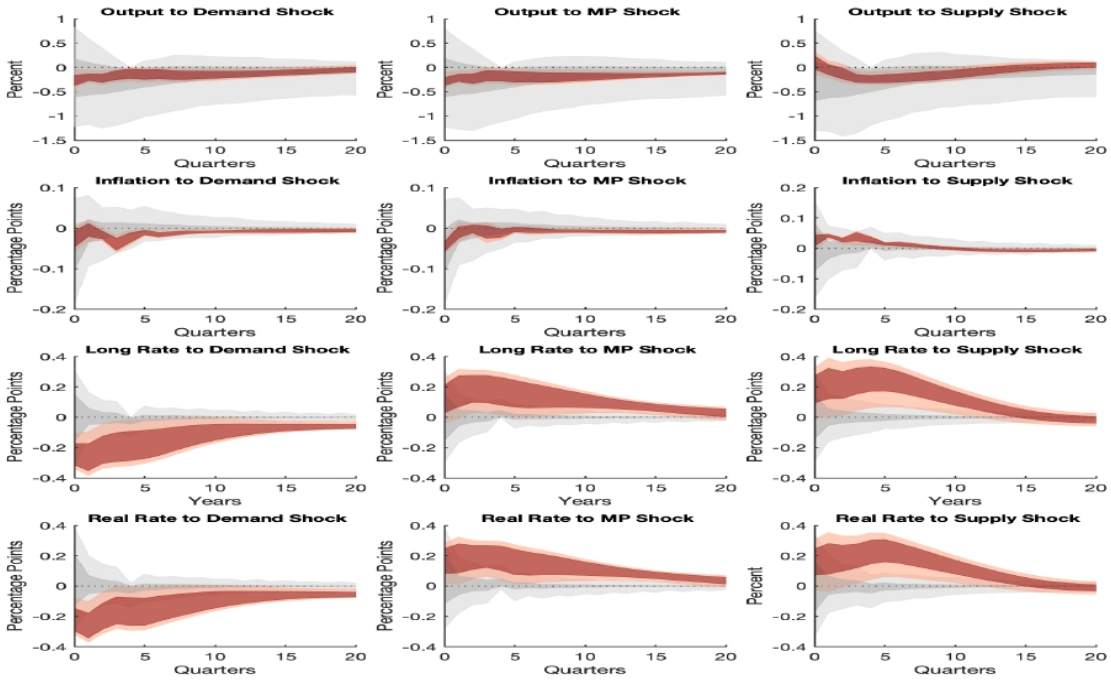
• Suppose $H = \begin{bmatrix} + & - \\ + & + \end{bmatrix}$ are the restrictions you want to impose. Using the QR decomposition one can show that $Q = \begin{bmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{bmatrix}$ with probability 0.5 and $Q = \begin{bmatrix} \cos(\omega) & \sin(\omega) \\ \sin(\omega) & -\cos(\omega) \end{bmatrix}$ with probability 0.5 and $\omega \sim U(-\pi, \pi)$.

• Because q_{ij} elements are nonlinear functions of ω, a uniform prior on ω implies a Beta distribution(1/2, (n-1)/2) for q_{ij}^2 where n is the number of VAR variables. When $n = 2$ the prior is like in the blue line. If $n > 2$ the shape of the prior changes. Why should it change?


3
2.5
2
1.5
1
0.5
0
n = 2
n = 6
$-\omega_{ii}^{1/2}$
0
$\omega_{ii}^{1/2}$
h_{ij}

• Arias et al. (2021): BH misleading. Result obtained conditioning on MLE estimates. If use parameter prior, prior on IRFs are uniform.

Effects of Prior



4.5 Additional identification restrictions

• **Maximal path restrictions**: an investment specific shock must generate e.g. the largest path for the structural shocks for the 1996-2000 period and impact only on investment at time zero (see Ben Zadev, 2013). Use the same technology as with sign restrictions: for each draw compute structural shock and check if restriction is satisfied.

• **Relative importance restrictions**: e.g. a TFP is the shock which is relatively more important for labor productivity (see Zimic, 2013). Similar technology as with sign restriction. Just add magnitude restrictions.

• **Variance decomposition restrictions**: at some horizon you maximize the variance of a variable explained one particular shock, see e.g. Barski and Sims (2012). Use principal components to find the rotation which gives you what you want.

• **Narrative sign restrictions**. Restrict sign of the responses and the sign or the contribution to historical decomposition of the identified shocks in certain periods (Antolin Diaz and Rubio, 2019).

• Can combine sign restriction with any of these (see Arias et al., 2018).

• **IV restrictions**. Typically used to identify one shock (say, the first one)

- Given a (vector of) instrument(s) z_t (from narrative, external, or high frequency information).

- First stage: $\hat{\epsilon}_{1t} = (z_t' z_t)^{-1}(z_t e_{1t}')z_t$, e_{1t} VAR residuals from equation 1.

- Second stage: compute instantaneous responses regressing y_t on $\hat{\epsilon}_{1t}$. Compute lagged responses using $A_i(\ell)$ and the relevant row of the estimated A_0 (see LP notes).

Implementation

VAR: $y_t = c + A(L)y_{t-1} + u_t$, $u_t \sim (0, \Sigma)$.

Companion form: $Y_t = C + AY_{t-1} + U_t$, $E_t \sim (0, \Sigma_U)$

Ortogonalization: $Y_t = C + AY_{t-1} + ME_t$, $E_t \sim (0, D)$

Projection: $Y_{t+h} = \sum_{i=0}^{h} A^i C + A^{h+1} Y_{t-1} + \sum_{i=0}^{h} A^i M E_{t+h-i}$

Rotations: $Y_{t+h} = \sum_{i=0}^{h} A^i C + A^{h+1} Y_{t-1} + \sum_{i=0}^{h} (A^i H)(H' M E_{t+h-i})$,
where $HH' = I$

Forecast error variance:
$\text{var}(Y_{t+h} - \sum_{i=0}^{h} A^i C - A^{h+1} Y_{t-1}) = \sum_{j=1}^{N} \sum_{i=0}^{h} ((A^i H)^2)_j (H' M E_{t+h-i})_j^2 \equiv$
$\sum_{j=1}^{N} \sum_{i=0}^{h} ((B^i)^2)_j var(v_{t+h-i,j})$.

• **Sign restrictions**. For some $\bar{\imath}, (A^{\bar{\imath}}H)_{lj}$ has a particular sign for variable l in response to shock j.

• **Magnitude restrictions**. For some $\bar{\imath}, (A^{\bar{\imath}}H)_{lj}$ has a particular sign for variable l in response to shock j and it is bounded above (below).

• **Variance decomposition restrictions**. $\frac{\sum_{i=0}^{\bar{h}}((B^i)^2)_{lj}var(e_{t+h-i,j})}{\sum_{j=1}^{N}\sum_{i=0}^{\bar{h}}((B^i)^2)_{j}var(e_{t+h-i,j})}$ is largest (smallest) for shock j,variable l at horizon $\bar{h}$.

• **Historical decomposition restrictions**. For some $t_1 < t < t_2$ and some l, $\sum_{i=0}^{h}(B^i)_{lj}e_{t+h-i,j}$ is largest (smallest) for shock j.

• **Narrative sign restrictions**: For some $\bar{\imath}, (A^{\bar{\imath}}H)_{lj}$ has a particular sign for variable l in response to shock j **and** $e_{t+h-i,j}$ has the right sign (from narrative) for $t_1 < t < t_2$.

Heteroskedasticity restrictions

• Assume the variance of the structural shocks changes over time: $var(e_t) = \Sigma_1, t = 1, \ldots, T_1, \quad var(e_t) = \Sigma_2, t = T_1 + 1, \ldots, T$, but the dynamics in response to the shocks do not.

• Lutkepohl (1996, Chapter 6.1.2): There exists a W and a diagonal Ω with typical element $\omega_i > 0, i = 1, 2, \ldots, m$ such that $\Sigma_1 = WW'$ and $\Sigma_2 = W\Omega W'$.

• W is a full matrix. It is unique up to sign changes if ω_i are distinct. Since W is the same in $\Sigma_1$and Σ_2, the impact effect of (structural) shocks is unchanged across regimes (and thus the dynamics unchanged).

• Ω incorporates volatility changes (if one $\omega_i \neq 1$ there is a change in volatility), so shocks are normalized to 1 in the first sample and to ω_i in the second.

• If $\mathcal{A}_0^{-1} = W$ all shocks of the system are identified.

• Heteroschedasticity restrictions typically sufficient to identify the shocks **without economic restrictions**. If economic restrictions exist, they become overidentifying and can be tested.

• If more than two regimes, variance changes may provide overidentification restrictions (see Rigobon, 2003).

• Lanne and Lutkepohl (2008): Markov switching structure in the variance of the shocks. Same idea applies.

Example 12

$$p_t = \beta y_t + \epsilon_{1t} \tag{32}$$
$$y_t = \alpha p_t + \epsilon_{2t} \tag{33}$$

where $E(\epsilon_{1t}\epsilon_{2t}) = 0$. *The covariance matrix of* $[p_t, y_t]'$ *satisfies*

$$V \equiv \begin{bmatrix} v_{11} & v_{12} \\ v_{12} & v_{22} \end{bmatrix} = \frac{1}{(1-\alpha\beta)^2} \begin{bmatrix} \beta^2\sigma_2^2 + \sigma_1^2 & \beta\sigma_2^2 + \alpha\sigma_1^2 \\ \beta\sigma_2^2 + \alpha\sigma_1^2 & \sigma_2^2 + \alpha^2\sigma_1^2 \end{bmatrix}$$

There are three free elements in V, $(v_{21} = v_{12})$ *and 4 structural parameters* $(\alpha, \beta, \sigma_1^2, \sigma_2^2)$. *The system is underidentified.*

- *Suppose* σ_1^2, σ_2^2 *depend on* $s = 1, 2$, *but* α, β *do not. Then*

$$V_1 = \frac{1}{(1-\alpha\beta)^2} \begin{bmatrix} \beta^2\sigma_{21}^2 + \sigma_{11}^2 & \beta\sigma_{21}^2 + \alpha\sigma_{11}^2 \\ \beta\sigma_{21}^2 + \alpha\sigma_{11}^2 & \sigma_{21}^2 + \alpha^2\sigma_{11}^2 \end{bmatrix}$$

$$V_2 = \frac{1}{(1-\alpha\beta)^2}\begin{bmatrix} \beta^2\sigma_{22}^2 + \sigma_{12}^2 & \beta\sigma_{22}^2 + \alpha\sigma_{12}^2 \\ \beta\sigma_{22}^2 + \alpha\sigma_{12}^2 & \sigma_{22}^2 + \alpha^2\sigma_{12}^2 \end{bmatrix}$$

There are six free elements in V_1, V_2 *and six structural parameters* $(\alpha, \beta, \sigma_{11}^2, \sigma_{12}^2\sigma_{21}^2, \sigma_{22}^2)$ *. System just identified by order condition!!*

• If we have three variance regimes, we have $(3 \times 3 = 9)$ reduced form parameters and 8 structural parameters (3×2 structural variances, α, β). System over-identified!

• Crucial restrictions:

i) α and β are unchanged across regimes.

ii) The variance of both shocks changes.

iii) Volatility changes must be independent across variables (no common factors), see Montiel Olea, et al. (2022).

iv) iii) requires structural shocks to be conditionally independent (strengthening from iid assumption).

• Could use the same approach if variance changes with season (seasonal heteroskedasticity).

• Careful: choice is not about homoskedastic or heteroskedastic shocks. OLS is quasi-MLE for iid, homosckedastic shocks, but consistency of OLS does not require these assumptions, see Goncalves and Kilian (2004).

Mechanics

• Assume $E(\epsilon_t|I_{t-1}) = 0,\ cov(\epsilon_{jt}, \epsilon_{it}|I_{t-1}) = 0, \forall j, i.$ Let $e_t = M\epsilon_t$, and $\sigma^2_{jt-1} = var(\epsilon_{jt}|I_{t-1}), \Sigma_{t-1} = var(e_t|I_{t-1})$.

• Then $\Sigma_{t-1} = Mdiag\{\sigma^2_{jt-1}\}M'$ and

$$\Sigma_t\Sigma_{t-1}^{-1} = Mdiag\{\frac{\sigma^2_{jt}}{\sigma^2_{jt-1}}\}M^{-1} \tag{34}$$

• Columns of M = eigenvectors of $\Sigma_t\Sigma_{t-1}^{-1}$: unique if $\frac{\sigma^2_{jt}}{\sigma^2_{jt-1}}$ distinct.

• Test conditional independence: examine if eigenvectors of $\Sigma_t\Sigma_{t-1}^{-1}$ are constant over time.

• Careful: need sufficient evidence of of heteroskedasticity. Otherwise, identification may become weak.

Structural Changes

• Heteroskedasticity identification exploits the fact that variance switches across s, but structural parameters do not.

• Can revert also assume that parameters change but variance do not. Can still get identification.

• Back to the simple model of example 11, but now let σ_i^2 be state independent and α, β change with the state. Then

Example 13

$$V_1 = \frac{1}{(1-\alpha_1\beta_1)^2}\begin{bmatrix} \beta_1^2\sigma_2^2+\sigma_1^2 & \beta_1\sigma_2^2+\alpha_1\sigma_1^2 \\ \beta_1\sigma_2^2+\alpha_1\sigma_1^2 & \sigma_2^2+\alpha_1^2\sigma_1^2 \end{bmatrix}$$

$$V_2 = \frac{1}{(1-\alpha_2\beta_2)^2}\begin{bmatrix} \beta_2^2\sigma_2^2+\sigma_1^2 & \beta_2\sigma_2^2+\alpha_2\sigma_1^2 \\ \beta_2\sigma_2^2+\alpha_2\sigma_1^2 & \sigma_2^2+\alpha_2^2\sigma_1^2 \end{bmatrix}$$

• *six free elements in* V_1, V_2 *and six structural parameters* $(\alpha_1, \beta_1, \alpha_2, \beta_2, \sigma_1^2, \sigma_2^2)$. *System is just identified.*

• As long as only structural parameters change, same idea applies. Structural breaks may help with identification.

Regime switches

• In some cases regime switches involve more than parameter (or variance) changes. It is the full solution that differs across regimes.

• Can exploit existing ideas to try to parameter identification

Example 14 *Simple model with Philips curve and Taylor rule*

$$\pi_t = c + \beta(r_t - r^n) + \psi b_t + \epsilon_{1t} \quad (35)$$

$$r_t = \max(r_t^*, 0) \quad (36)$$

$$r_t^* = r^n + \gamma\pi_t + \epsilon_{2t} \quad (37)$$

$$b_t = \min(\alpha r_t^*, 0) \quad (38)$$

- b_t *long term bonds (see Chen, Curdia and Ferrero, 2012)*
- ψ *UMP coefficient; if* $\psi = 0$*, MP ineffective at the ZLB.*
- $\alpha \geq 0$ *. If* $\alpha = 0$ *UMP not used.*
- r_t^* *shadow rate.*
- $(\epsilon_{1t}, \epsilon_{2t})$ *iid*$(0, diag(\sigma_1^2, \sigma_2^2))$*.*

- *(35) and (38) imply*

$$\pi_t = c + \beta(r_t - r^n) + \beta^* min(r_t^*, 0) + \epsilon_{1t} \tag{39}$$

where $\beta^* = \alpha\psi$.

- *If UMP removes the ZLB (MP unrestricted across regimes),* $\beta = \beta^*$ *and*

$$\pi_t = c + \beta(r_t^* - r^n) + \epsilon_{1t} \tag{40}$$

Philips curve has no kink.

- *In general (39) and (36) imply*

$$\pi_t = \tilde{c} + \tilde{\beta}(r_t^* - r^n) + v_{1t} \tag{41}$$

where $\tilde{\beta} = \frac{\beta - \beta^*}{1 - \gamma\beta^*}$, $\tilde{c} = \frac{c}{1 - \gamma\beta^*}$ $v_{1t} = \frac{\epsilon_{1t} + \beta^* \epsilon_{2t}}{1 - \gamma\beta^*}$

- *For existence and uniqueness of a solution need* $\gamma\tilde{\beta} < 1$ *or*

$$\frac{1-\gamma\beta}{1-\gamma\beta^*} > 0 \tag{42}$$

which is satisfied for $\beta, \beta^* < 0, \gamma > 0$ *(standard case).*

- *Under (42) the unique solution to the system is*

$$\pi_t = \mu_1 - \tilde{\beta} D(\mu_2 + u_{2t}) + u_{1t} \tag{43}$$

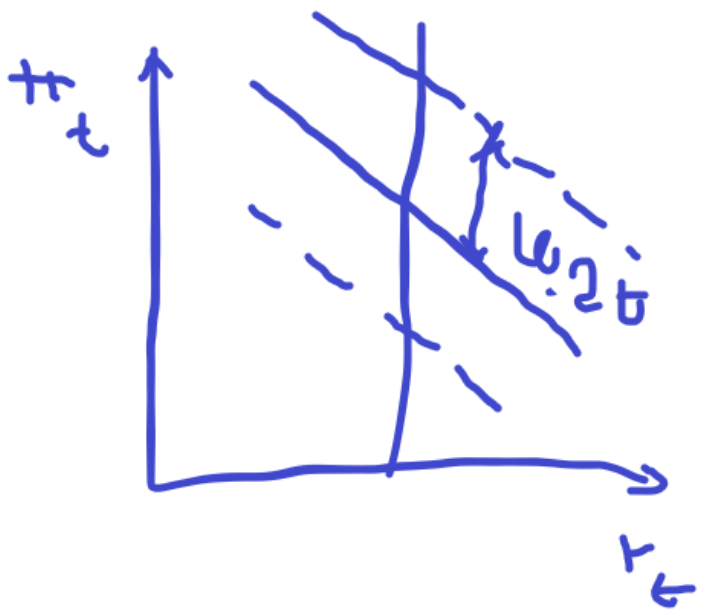
$$r_t = \max(\mu_2 + u_{2t}, 0) \tag{44}$$

where $D = 1$ *if* $r_t = 0$, $\mu_1 = \frac{c}{1-\gamma\beta}$, $\mu_2 = \frac{\gamma c}{1-\gamma\beta} + r^n$, $u_{1t} = \frac{\epsilon_{1t}+\beta\epsilon_{2t}}{1-\gamma\beta}$, $u_{2t} = \frac{\gamma\epsilon_{1t}+\epsilon_{2t}}{1-\gamma\beta}$.

- *Not only parameter change, but complete regime switching solution:*

$$\begin{aligned}
\pi_t &= \mu_1 + u_{1t} && (45)\\
r_t &= \mu_2 + u_{2t} \quad \text{if } r_t \neq 0 && (46)\\
\pi_t &= \mu_1 - \tilde{\beta} u_2 + u_{1t} - \tilde{\beta} u_{2t} && (47)\\
r_t &= 0 \quad \text{if } r_t = 0 && (48)\\
& && (49)
\end{aligned}$$

- *If $r_t = 0$ π_t responds to u_{1t}, u_{2t}; if $r_t \neq 0$ π_t responds to u_{1t} only, i.e. u_{2t} shifts (identifies) the Phillips curve.*



u(2t) shifts the Philips curve at the ZLB

- Problem: $u_t's$ are reduced form shocks. Interested in ϵ_{2t} (MP shock).

- Restrictions in different regimes

$$\begin{bmatrix} \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} \mu_{1t} \\ \mu_{2t} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \quad \begin{bmatrix} \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} \mu_{1t} - \tilde{\beta}\mu_{2t} \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & -\tilde{\beta} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{50}$$

- Restrictions in different regimes (in terms of structural shocks)

$$\begin{bmatrix} \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} \frac{c}{1-\gamma\beta} \\ -\frac{\gamma c}{1-\gamma\beta} \end{bmatrix} + \begin{bmatrix} \frac{1}{1-\gamma\beta} & \frac{\beta}{1-\gamma\beta} \\ \frac{\gamma}{1-\gamma\beta} & \frac{1}{1-\gamma\beta} \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \tag{51}$$

$$\begin{bmatrix} \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} \frac{(1-\gamma\tilde{\beta})c}{1-\gamma\beta} \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1-\gamma\tilde{\beta}}{1-\beta\gamma} & \frac{\beta-\tilde{\beta}}{1-\beta\gamma} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \tag{52}$$

- Mavroedis (Economerica, 2021) Let $\beta^* = 0$. If we identify β using tools from censored regressions models (Tobit), we can separate ϵ_{1t} and ϵ_{2t}. Procedure:

- Using observations outside the ZLB we have

$$E_t(\pi_t|r_t > 0) = c + \beta(r - R^n) + \rho(r_t - \mu_2 + \tau\frac{\phi(a)}{1 - \Phi(a)}) \tag{53}$$

where $a = -\frac{\mu_2}{\tau}, \tau = (var(u_{2t}))^{0.5}, \rho = \frac{cov(u_{1t}, u_{ut})}{\tau^2} - \beta = \frac{1-\gamma\beta}{\gamma^2\sigma_1^2+\sigma_2^2}$ and $\phi(.), \Phi(.)$ are the pdf and the cdf of a normal.

- $\rho =$ bias due to truncation

- μ_2, τ and thus $\phi(a), \Phi(a)$ can be recovered from

$$r_t = \mu_2 + u_{2t} \tag{54}$$

using observations outside the ZLB. Then ρ and β can obtained from (53) and, with an estimate of γ, we can separate ϵ_{1t} and ϵ_{2t}.

• Alternative 1 (magnitude restrictions across regimes): Since $\tilde{\beta} > 0$

- Responses of inflation at ZLB to $\epsilon_{1t}, \epsilon_{2t}$ **smaller** than outside the ZLB.

- Response of interest rates at ZLB to $\epsilon_{1t}, \epsilon_{2t}$ **smaller** than outside the ZLB.

• Draw rotations. Compute $\epsilon_{1t}, \epsilon_{2t}$ for each regimes. Check if impact effects are smaller at the ZLB.

- Alternative 2 (restrictions on the difference across regimes)
- A: $(\pi_t - E(\pi_t)|\epsilon_{1t}, r_t \neq 0) - (\pi_t - E(\pi)|\epsilon_{1t}, r_t = 0) = \gamma\tilde{\beta}/(1-\gamma\beta)$
- B: $(\pi_t - E(\pi_t)|\epsilon_{2t}, r_t \neq 0) - (\pi_t - E(\pi)|\epsilon_{2t}, r_t = 0) = \tilde{\beta}/(1-\gamma\beta)$
- C: $(r_t|\epsilon_{1t}) - (r_t = 0) = \gamma/(1-\gamma\beta)$
- D: $(r_t|\epsilon_{2t}) - (r_t = 0) = 1/(1-\gamma\beta)$

- Draw rotations, calculate $\epsilon_{1t}, \epsilon_{2t}$, keep only rotations that imply CB=A=DCB, A/B constant A/C constant ($\pm$ some tiny range, otherwise you get exact identification)

- For accepted rotations $A/B = \gamma$, $A/C = \tilde{\beta}$. Then we can trace out the differential effect of structural shocks on r_t across regimes. For a given γ, D identifies β and thus, given $\tilde{\beta}$ also β^*.

- We have all the ingredients to trace out the effect of shocks in two regimes $(\gamma, \beta, \tilde{\beta})$ and differentiate the impact of MP across regimes β, β^*.

4.6 Higher moments restrictions

• New literature using higher order moments for identification, see e.g. Lanne et al (2017), Gurieroux et al. (2020), Fiorentini and Sentana (2021), Bekaert et al. (2021)

• Identification via higher moments is possible only under non-normality. Skewnees and excess kurthosis may have information about parameters of the covariance matrix of the shocks.

• These moments help because they have different implications for the observables depending on which shock drive the fluctuations.

• Can use ML, quasi-ML or two step GMM approach to estimate the structural parameters.

Basic Idea

- Consider a bivariate VAR(1) with output growth and inflation

$$y_t = A_0 + A_1 y_{t-1} + e_t$$

- Structural model driven by supply and demand disturbances $\epsilon_t^s, \epsilon_t^d$.

- Mapping reduced-stuctural shocks (assuming $a_j^k > 0, \forall j, k$)

$$\begin{aligned} e_{\pi,t} &= -a_\pi^s \epsilon_t^s + a_\pi^d \epsilon_t^d \\ e_{y,t} &= a_y^s \epsilon_t^s + a_y^d \epsilon_t^d \end{aligned} \tag{55}$$

- Assuming var(ϵ_t^k)=1, the second moment (covariance) relationship is:

$$cov(e_t) = \begin{bmatrix} (a_\pi^s)^2 + (a_\pi^d)^2 & -a_\pi^s a_y^s + a_\pi^d a_y^d \\ -a_\pi^s a_y^s + a_\pi^d a_y^d & (a_y^s)^2 + (a_y^d)^2 \end{bmatrix} \tag{56}$$

- Four structural parameters, three reduced form covariances.

- Fourth moments (in excess from a normal distribution)

$$E(e_{\pi,t}^4) - 3 = \frac{(a_\pi^s)^4 * k_s + (a_\pi^d)^4 * k_d}{var^2(e_\pi)} \quad (57)$$

$$E(e_{y,t}^4) - 3 = \frac{(a_y^s)^4 * k_s + (a_y^d)^4 * k_d}{var^2(e_y)} \quad (58)$$

$$E(e_{\pi,t}^2 e_{y,t}^2) - 3 = \frac{(a_\pi^s)^2(a_y^s)^2 * k_s + (a_\pi^d)^2(a_y^s)^2 * k_d}{var(e_\pi)var(e_y)} \quad (59)$$

$$E(e_{\pi,t}^3 e_{y,t}) - 3 = \frac{-(a_\pi^s)^3 a_y^s * k_s + (a_\pi^d)^3 a_y^d * k_d}{var^{3/2}(e_\pi)var^{1/2}(e_y)} \quad (60)$$

$$E(e_{\pi,t} e_{y,t}^3) - 3 = \frac{-(a_\pi^s)(a_y^s)^3 * k_s + (a_\pi^d)(a_y^d)^3 * k_d}{var^{1/2}(e_\pi)var^{3/2}(e_y)} \quad (61)$$

k^s, k^d are the unconditional excess kurthosis of supply and demand shocks.

• (57)-(58) univariate fourth moments; (59) is a symmetric cross moment, (60)-(61) asymmetric cross moments.

• (57)-(58) add two moments and two parameters (k^s, k^d).

• (59)-(61) add three moments but no new parameter.

• Conclusion: If we use second and fourth moments (assuming that they are significantly different from 3), we have eight conditions and six parameters to estimate (those appearing in (56), k^s, k^d).

• Fourth moments source of overidentification.

Intuition

• (59) measures the covariance of the square of the two VAR residuals in excess of their square of the correlation. If inflation and output growth are volatile or quiescent at the same time, this moment is positive.

$$E(e_{\pi,t}^4) - 3 = \frac{cov(e_\pi^2, e_y^2)}{var(e_\pi) - var(e_y)} - \rho^2(e_\pi, e_y) \tag{62}$$

• To see how (60)-(61) may aid identification, suppose $k_s = k_d$. Then a larger a_π^s relative to a_π^d lowers the co-kurthosis moment with inflation to the third power much more then the co-kurthosis moment of output growth at the third power.

Estimation

• Let $H = [\sigma_\pi, \sigma_y, \rho_{\pi,y}, E(e_{\pi,t})^4 - 3, E(e_{y,t}^4) - 3, E(e_{\pi,t}^2 e_{y,t}^2) - 3, E(e_{\pi,t}^3 e_{y,t}) - 3, E(e_{\pi,t} e_{y,t}^3) - 3]$. These quantities can be estimated from the data conditional on the VAR parameters.

• Let $\theta = (a_\pi^d, a_\pi^s, a_y^d, a_y^s, k^s, k^d)$.

• Chose θ to min$(H - H(\theta))' W (H - H(\theta))$, where W is any weighting matrix and $H(\theta)$ the theoretical moments.

• Overidentified system: can test additional restrictions.

• Can check individual elements of H for large deviations.

• Use fourth moment since many financial (interest rate) variables display excess kurthosis.

• Could also use third moment if there is evidence of skewness.

• In VARs with more than 2 variables there are many cross-kurthosis moments. Need to be selective to maximize informativeness.

• Procedure strengthens the iid assumption to mutually independence for the shocks

• Excludes shocks with common volatility factor.

• Choice is not between Gaussian and non-Gaussian shocks (OLS is consistent even when shocks are non-Gaussian).

• Test independence assumption: $cov(\hat{e}_{jt}^2, \hat{e}_{it}^2) = 0$. Need strong evidence of independence; otherwise weak shock identification.

Summary

• Identification problem concerns the structural interpretation of the VAR residuals. Intepretation is as good as identification.

• It is independent of whether a classical or a Bayesian approach to estimation is used.

• It requires restrictions to be able to estimate the free parameters. Where are the restrictions coming from? Theory? Statistics?

5 Why do we want to use BVAR?

• VARs have lots of parameters to be estimated. If they are used for forecasting, their performance is poor.

• If sample is short, VAR estimate deviate considerable from large sample approximation.

• Hard to incorporate client prior views into classical VAR.

• BVARs are a flexible way to incorporate extraneous (client) information; help to reduce the dimensionality of the parameter space; and get more reasonable small sample estimates.

5.1 Likelihood function of a VAR(q)

Consider an M variable VAR with q lags (k=Mq coefficients each equation, Mk total coefficients in total), no constant (demeaned data).

$$y_t = A(L)y_{t-1} + e_t \quad e_t \sim N(0, \Sigma_e)$$

Letting $A = [A_1, \ldots A_q]$; $X_t = [y_{t-1}, \ldots y_{t-q}]$, $\beta = vec(A)$, the VAR is:

$$y = (I_M \otimes X)\beta + e \quad e \sim (0, \Sigma_e \otimes I_T) \tag{63}$$

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 $y_p = [y_{-1}, \ldots, y_{-q}]$:

$$\begin{aligned} L(\beta, \Sigma_e | y, y_p) &= \frac{1}{(2\pi)^{0.5MT}} |\Sigma_e \otimes I_T|^{-0.5} \\ &\times \exp\{-0.5(y - (I_M \otimes X)\beta)'(\Sigma_e^{-1} \otimes I_T)(y - (I_M \otimes X)\beta)\} \end{aligned}$$

After manipulations (see Canova,2007, Ch 10) the likelihood function is:

$$L(\beta, \Sigma_e | y, y_p) \propto N(\beta | \hat{\beta}, \Sigma_e, X, y, y_p) \times iW(\Sigma_e | \hat{\beta}, X, y, y_p, T - \nu) \quad (64)$$

where tr = trace of the matrix, $\hat{\beta} = (\Sigma_e^{-1} \otimes X'X)^{-1}(\Sigma_e^{-1} \otimes X)y$, and iW stands for inverted Wishart distribution

• 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 $(y - (x \otimes \Sigma_e)\hat{\beta})'(y - (x \otimes \Sigma_e)\hat{\beta})$ and $(T - \nu)$ degrees of freedom; $\nu = k + M + 1$.

• More info on Wishart distributions: see appendix and

https://en.wikipedia.org/wiki/Inverse-Wishart_distribution

• Bayesian inference: combine likelihood with a prior.

i) If the prior is conjugate and the parameters of the prior known (or estimable): closed form solution for the conditional and marginal of β and the marginal of Σ_e are available.

ii) If the parameters of the prior are random, need Gibbs sampler to get conditional and marginal distributions, even when the prior is conjugate.

iii) if the prior is not conjugate or hierarchical, always need MCMC simulation methods.

6 Priors for VARs

1. Diffuse prior for both β and Σ_e (conjugate).

2. Normal prior for β with Σ_e fixed (conjugate).

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: $g(\beta, \Sigma_e) \propto |\Sigma_e|^{-0.5(M+1)}$ - this is called Jeffrey's (flat) prior.

Joint posterior: $g(\beta, \Sigma_e|Y) = L(\beta, \Sigma_e|Y)g(\beta, \Sigma_e)$.

Posterior is similar to the likelihood: there is only an extra term in the normalizing constant. Thus

$$g(\beta|\Sigma_e, Y) = N(\beta|\hat{\beta}, \Sigma_e, X, y, y_p) \quad (65)$$

$$g(\Sigma_e|Y) = iW(\Sigma_e|\hat{\beta}, X, y, y_p, T-k) \quad (66)$$

where k number of parameters in each equation.

• If the prior is diffuse, the posterior mean is the OLS estimator. Classical analysis equivalent to Bayesian analysis with flat prior and a quadratic loss function (the posterior mean is the optimal point estimator)

• Posterior draws for β can be obtained in two steps:

1. Draw Σ_e from the posterior inverted Wishart.

2. Conditional on the value of Σ_e, draw β from a multivariate normal.

- Case 2: $\beta = \bar{\beta} + v,\ v \sim N(0, \Sigma_b)$, where $\bar{\beta}, \Sigma_b$ are known.

- Prior:

$$\begin{aligned} g(\beta) &\propto |\Sigma_b|^{-0.5} exp[-0.5(\beta - \bar{\beta})' \Sigma_b^{-1} (\beta - \bar{\beta})] \\ &= |\Sigma_b|^{-0.5} exp[-0.5(\Sigma_b^{-0.5}(\beta - \bar{\beta}))' \Sigma_b^{-0.5} (\beta - \bar{\beta})] \end{aligned} \quad (67)$$

- Posterior:

$$g(\beta|y) \propto exp[-0.5(\beta - \tilde{\beta})' \tilde{\Sigma}_b^{-1} (\beta - \tilde{\beta})] \quad (68)$$

$$\tilde{\beta} = [\Sigma_b^{-1} + (\Sigma_e^{-1} \otimes X'X)]^{-1} [\Sigma_b^{-1} \bar{\beta} + (\Sigma_e^{-1} \otimes X)' y] \quad (69)$$

$$\tilde{\Sigma}_b = [\Sigma_b^{-1} + (\Sigma_e^{-1} \otimes X'X)]^{-1} \quad (70)$$

- $g(\beta|y)$ is $N(\tilde{\beta}, \tilde{\Sigma}_b)$.

- If Σ_e is unknown, use $\hat{\Sigma}_e = \frac{1}{T-1} \hat{e}'\hat{e}$, where $\hat{e}_t = y_t - (I \otimes X)\beta_{ols}$.

- Alternatively

$$\begin{aligned} g(\beta|y) &\propto exp[-0.5(\beta-\tilde{\beta})'Z'Z(\beta-\tilde{\beta})] \\ \tilde{\beta} &= (Z'Z)^{-1}(Z'z) \\ Z &\equiv [\Sigma_b^{-0.5}, (\Sigma_e^{-0.5} \otimes X)]' \end{aligned} \tag{71}$$

- $\tilde{\beta}$ related to the classical least square under uncertain linear restrictions.

$$\begin{aligned} y_t &= x_t B + e_t \qquad e_t \sim (0, \sigma^2) \\ \bar{B} &= B - \epsilon \quad \epsilon \sim (0, \Sigma_b) \end{aligned} \tag{72}$$

where $B = [B_1, \ldots B_q]', x_t = [y_{t-1}, \ldots y_{t-q}]$.

- Set $z_t = [y_t, \bar{B}]', \ Z_t = [x_t, I]', \ E_t = [e_t, \epsilon]'$.

- Hence $z_t = Z_t B + E_t$ where $E_t \sim (0, \Sigma_E)$, $t = 1, \ldots, T$ and

$B_{GLS} = (Z'\Sigma_E^{-1}Z)^{-1}(Z'\Sigma_E^{-1}z) = \tilde{B}$ (Theil's mixed estimator).

• **Prior on VAR coefficients can be treated as a dummy observation added to the system of VAR equations.**

• **Prior can be thought as playing the role of an initial condition.** Can write it as

$$y_0 = x_0 B + e_0$$

where $y_0 = \sigma^2 W^{-1}\bar{B}, x_0 = \sigma^2 W^{-1}$, $e_0 = \sigma^2 W^{-1}\epsilon$, $WW' = \Sigma_b$.

Special case: Litterman (Minnesota) setup

• $\bar{\beta} = 0$ except $\bar{\beta}_{(j=i,\ell=1)} = 1$; $\Sigma_b = diag(\sigma(\phi))$ where:

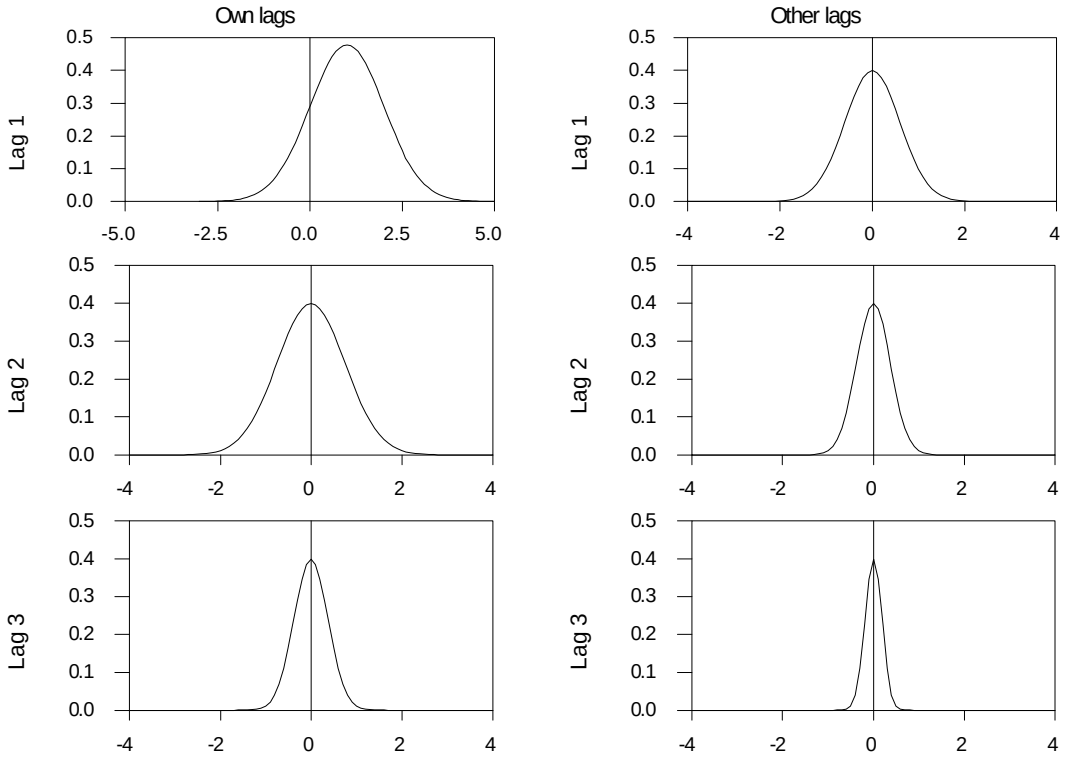
$$\sigma_{ij,\ell} = \frac{\phi_1}{h(\ell)} \text{ if } i = j \tag{73}$$

$$= \phi_1 \frac{\phi_2}{h(\ell)} * (\frac{\sigma_j}{\sigma_i})^2 \text{ otherwise} \tag{74}$$

$$= \phi_1 * \phi_4 * (\frac{\sigma_j}{\sigma_i})^2 \text{ for exogenous variables} \tag{75}$$

• $\phi_1 =$ general tightness; $\phi_2 =$ relative tightness on other variables; $h(\ell) =$ tightness on the variance of longer lags; $(\frac{\sigma_j}{\sigma_i})^2$ scaling factor; i=variable, j=equation, $\ell =$ lag.

• Useful structures for $h(\ell)$ (one decay parameter): harmonic decay $h(\ell) = \ell^{(2*\phi_3)}$; geometric decay $h(\ell) = \phi_3^{-\ell+1}$; linear decay $h(\ell) = \ell$.


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

• Prior mean so that VAR is M random walks (good for forecasting).

• Σ_b very big ($M \times M$). Decrease dimensionality by setting $\Sigma_b = \Sigma_b(\phi)$.

• Σ_b is a-priori diagonal (no expected relationship among equations and coefficients); ϕ_1 measure 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 $\phi_2 \leq 1$, relative importance of other variables.

• Variance of the exogenous variables is regulated by ϕ_4. If ϕ_4 is large, prior information on the exogenous variables diffuse.

Example 15 *Bivariate VAR(2) with $h(\ell) = \ell$; 4 coefficients each equation.*

$\bar{\beta} = [1, 0, 0, 0, 0, 0, 1, 0]$

$$\Sigma_b = \begin{bmatrix} \phi_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\phi_1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \phi_1\phi_2(\frac{\sigma_1}{\sigma_2})^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\phi_1}{2}\phi_2(\frac{\sigma_1}{\sigma_2})^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi_1\phi_2(\frac{\sigma_2}{\sigma_1})^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\phi_1}{2}\phi_2(\frac{\sigma_2}{\sigma_1})^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\phi_1}{2} \end{bmatrix}$$

• If Σ_b is diagonal and $\phi_2 = 1$ and the same variables belong to all equations. Then $\tilde{\beta}$=vec($\tilde{\beta}_i$), where $\tilde{\beta}_i$ computed equation by equation. If Σ_b is not diagonal and this result does not hold.

• Let $\alpha = (\beta, vech(\Sigma_b))$. Minnesota prior makes $\alpha = \psi(\phi)$, where ϕ is of small dimension. Hopefully, better estimates of ϕ than for α. Better forecasts.

• Minnesota prior imposes probability distributions on VAR coefficients (uncertain linear restrictions). It gives a reasonable account of the uncertainty faced by an investigator.

- $(\frac{\sigma_i}{\sigma_j})^2$ estimated from the data (univariate AR or VAR).

- How do we choose $\phi = (\phi_1,\ \phi_2, \ldots)$?

1) Rules of thumb. Default values: $\phi_1 = 0.2^2,\ \phi_2 = 0.5^2, \phi_4 = 10^5$, 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.

- If ϕ_1 is large Minnesota prior = diffuse prior (posterior=OLS)

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 - see later.

Example 16 *(ML-II approach)*

$$\begin{aligned} y_t &= Bx_t + u_t \quad u_t \sim N(0, \sigma_u^2) \\ B &= \bar{B} + \upsilon \quad \upsilon \sim N(0, \sigma_\upsilon^2 \end{aligned} \tag{76}$$

B *scalar,* σ_u^2 *known,* $\bar{B}$ *fixed and* $\sigma_\upsilon^2 = q(\phi)^2$, ϕ =*hyperparameters. Then:*

- $y_t = \bar{B}x_t + \epsilon_t$ *where* $\epsilon_t = e_t + \upsilon x_t$ *and posterior kernel is:*

$$\grave{g}(\beta, \phi|y) = \frac{1}{(2\pi)^{0.5}\sigma_u\sigma_\upsilon} \exp\{-0.5\frac{(y - Bx)^2}{\sigma_u^2} - 0.5\frac{(B - \bar{B})^2}{\sigma_\upsilon^2}\} \tag{77}$$

$y = [y_1, \ldots y_t]', x = [x_1, \ldots x_t]'$. *Integrating* B *out of (77), maximize:*

$$\tilde{g}(\phi|y) = \frac{1}{(2\pi q(\phi)^2 tr|X'X| + \sigma_u^2)^{0.5}} \exp\{-0.5\frac{(y - \bar{B}x)^2}{\sigma_u^2 + q(\phi)^2 tr|X'X|}\} \tag{78}$$

• Applications of ML-II approach

i) Giannone, Primiceri, Lenza (2015): 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 (2014): employ marginal likelihood to choose the informativeness of prior distribution for time variations in coefficients and in the variance, i.e. choose ζ in $\beta_t \sim N(\bar{\beta}, \Sigma \otimes \Omega\zeta)$

iii) Carriero, Kapetanios, Marcellino (2014): employ marginal likelihood to select the variance of the prior from a grid, i.e. choose ζ in $\beta \sim N(\bar{\beta}, \zeta I)$

Posterior simulations for case 2 prior

Easy. Since Σ_e is fixed.

- Draw β from the normal posterior, keeping Σ_e fixed.

Results for other cases (Kadiyala and Karlsson, 1997):

• Case 3) (semi-conjugate): $g(\beta, \Sigma_e)$ is Normal-diffuse, i.e. $g(\beta) \sim N(\bar{\beta}, \bar{\Sigma}_b)$; $\bar{\beta}$ and Σ_b known, and $g(\Sigma_e) \propto |\Sigma_e|^{-0.5(M+1)}$.

• The conditional posteriors as case 2) (moments are different) but the marginal posterior is unknown: $g(\beta|y) \propto \exp\{0.5(\beta - \tilde{\beta})'\bar{\Sigma}_b^{-1}(\beta - \tilde{\beta})\} \times |(y - X\hat{B})'(y - X\hat{B}) + (B - \hat{B})'(X'X)(B - \hat{B})|^{-0.5T}$.

• 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 $t(\tilde{\Omega}^{-1}, \tilde{\Sigma}_e, \tilde{B}, T + \bar{\nu})$.

• In cases 3)-4) there is posterior dependence among the equations (even with prior independence and $\phi_1 = 1$).

- Posterior simulations for case 3 require MCMC.

Posterior simulations for case 4 prior

1. Draw Σ_e from the posterior inverted Wishart.

2. Conditional on the draw for Σ_e, draw β from a multivariate normal.

• Logic of the prior as initial observation or as observation added to the data can be extended.

• **Any additional uncertain restrictions on the coefficients can be tagged on to the system as a set of additional observations.**

i) Sum of coefficient restriction:

$$\sum_i \beta_{ij} \sim N(\mu_{1j}, \sigma_{\mu_{1j}}) \; j = 1, 2, \ldots, m \tag{79}$$

ii) Cointegration restriction:

$$\sum_i \beta_{ij} - \sum_i \beta_{ih} \sim N(\mu_2, \sigma_{\mu_2}) \tag{80}$$

iii) Long run prior restrictions: Giannone et al. 2019.

Tips

- If ϕ are treated as fixed, need some sensitivity analysis.

- Rule-of-thumb parameters work well for forecasting. Careful in structural estimation.

- Set prior moments as you wish (subjective prior!!). For computational ease, σ_b should have a Kroneker product form.

- In short samples, prior dominate. Use large prior variances to avoid distortions.

- In persistent BVARs, standard statistics may blow up (probability of a root greater than one not small). Either throw away draws generating explosiveness or use version of trend restriction to keep draws stable.

7 Hierarchical priors

- If ϕ are random, computations become more difficult.

- No closed form solution for the posterior; no posterior moments.

- Need to use MCMC to draw sequences from the posterior.

- Are there gains from using random ϕ (relative to empirical based or rules of thumb choices)? Not much is known, see Carriero et al., (2014), Giannone et al., (2015).

Hierarchical BVARs

$$y_t = (I \otimes X_t)\beta + e \quad e \sim N(0, \Sigma) \tag{81}$$
$$\beta = M_0\theta + v \quad v \sim N(0, D_0) \tag{82}$$
$$\theta = M_1\mu + \zeta \quad \zeta \sim N(0, D_1) \tag{83}$$

• Priors: $p(\Sigma) \sim iW(\bar{S}, s)$; $p(D_0) \sim iW(\bar{D}_0, \rho)$; $p(\mu) \propto 1$, M_0, M_1, D_1 known.

• Conditional Posteriors:

1) $(\beta|\psi_{-\beta}, Y, \mathcal{X}) \sim N(\tilde{\beta}, \tilde{\Omega})$.

2) $(\Sigma|\psi_{-\Sigma}, Y, \mathcal{X}) \sim iW(\tilde{\Sigma}, s + T)$

3) $(\theta|\psi_{-\theta}, Y, \mathcal{X}) \sim N(\tilde{D}_1(D_1^{-1}M_1\mu + M_0'D_0^{-1}\beta), \tilde{D}_1)$

4) $(D_0|\psi_{-D_0}, Y, \mathcal{X}) \sim iW(\tilde{D}_0, \rho + 1)$
5)$(\mu|\psi_{-\mu}, Y, \mathcal{X}) \sim N(\hat{\mu}, \Sigma_\mu)$

where

$\tilde{\Omega} = (D_0^{-1} + \sum_t X_t' \Sigma^{-1} X_t)^{-1}$;
$\tilde{\beta} = \tilde{\Omega}(D_0^{-1} M_0 \theta + \sum_t X_t' \Sigma^{-1} y_t)$;
$\tilde{\Sigma}^{-1} = \bar{S} + \sum_t (Y_t - X_t \beta)(y_t - X_t \beta)'$;
$\tilde{D}_1 = (D_1^{-1} + M_0' D_0^{-1} M_0)^{-1}$;
$\tilde{D_0}^{-1} = D_0^{-1} + \sum_{g=1}^{M} (\beta_g - \theta)(\beta_g - \theta)'$
$\hat{\mu} = (M_1' M_1)^{-1}(M_1 \theta)$
$\Sigma_\mu = (\theta - \hat{\mu} M_1)'(\theta - \hat{\mu} M_1)$

Use these conditional posteriors in the Gibbs sampler.

- Another useful hierarchical VAR:

$$
\begin{aligned}
y_t &= (I \otimes X)\beta + e \quad e \sim N(0, \Sigma) && (84)\\
\beta &= \bar{\beta} + v \quad v \sim N(0, \Sigma \otimes \Omega * \zeta) && (85)\\
\zeta &= \bar{\zeta} + \epsilon \quad \epsilon \sim N(0, \eta) && (86)
\end{aligned}
$$

where $(\bar{\beta},\ \Omega, \bar{\zeta}, \eta)$ are known (or estimable).

- Want the joint posterior of (β, ζ, Σ).

- Interest is in $g(\zeta|y, X, y_p) = \int g(\zeta, \beta, \Sigma|y, X, y_p) d\beta d\Sigma$.

- One example where $g(\zeta|y, X, y_p)$ is analytically available is in Canova (2007, chapter 9). Otherwise, use Gibbs sampler.

8 Structural Analyses with BVARs

- Unusual to report estimates of coefficients, standard errors, and R^2.
- Most of VAR coefficients insignificant.
- R^2 always exceeds 0.99.
- How do we summarize VAR results?

$$y_t = A(\ell)y_{t-1} + e_t \quad e_t \sim N(0, \Sigma_e) \tag{87}$$

$$= D(\ell)e_t \tag{88}$$

where $D(\ell) = (1 - A(\ell)\ell)^{-1}$

8.1 Impulse responses (IR)

• What is the effect of a surprise cut in interest rates on inflation? What is the effect of foreign shocks on domestic employment? In a two variable VAR, with inflation being the first variable the contemporaneous effect is D_{012}, the effect at lag 1 is D_{112}, and the effect at lag q is D_{q12}.

• It traces out how y_{jt} is displayed from its steady state, given an orthogonal shock in ϵ_{it}.

• Similar to calculating the (micro) "causal effects of an intervention ".

$$IR^h(j,i) = E(y_{jt+h}|\epsilon_{it} = 1) - E(y_{jt+h}|\epsilon_{it} = 0) \tag{89}$$

- How are impulse responses computed?

- For each draw of VAR coefficients and covariance matrix Σ_e:

1) Transform the VAR into a companion form.

2) Solve backward the companion form.

3) Orthogonalize the shocks.

4) Store results.

Backward solution and orthogonalization

- Use the companion form:

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{E}_t \\ &= \mathbf{A}^t\mathbf{Y}_0 + \sum_{j=0}^{t-1} \mathbf{A}^j\mathbf{E}_{t-j} \qquad (90) \\ &= \mathbf{A}^t\mathbf{Y}_0 + \sum_{j=1}^{t-1} \tilde{\mathbf{A}}^j\tilde{\mathbf{E}}_{t-j} \qquad (91) \end{aligned}$$

where $\tilde{\mathbf{A}}^j = \mathbf{A}^j\mathbf{P}_E,\ \tilde{\mathbf{U}}_t = \mathbf{P}_E^{-1}\mathbf{E}_t,\ \mathbf{P}_E\mathbf{P}'_E = \mathbf{\Sigma}_E$;

- Draw parameters $l = 1, 2, \ldots, L$ times. Compute $\tilde{\mathbf{A}_1}^j$ for each horizon j. Store the results.

- Report the mean (median) of $\tilde{\mathbf{A}_1}^j$ and the percentiles, say, $[\tilde{\mathbf{A}_5}^j, \tilde{\mathbf{A}_{95}}^j]$

8.2 Variance decomposition: τ-steps ahead forecast error

- How much of output forecast error variance of due to supply shocks?

- Uses:

$$\hat{y}_t(\tau) \equiv y_{t+\tau} - y_t(\tau) = \sum_{j=0}^{\tau-1} \tilde{D}_j \tilde{e}_{t+\tau-j} \quad D_0 = I \tag{92}$$

$y_t(\tau)$ is the τ-steps ahead prediction of y_t based on the VAR.

- Computes share of the variance of $y_{i,t+\tau} - y_{i,t}(\tau)$ due to each $\tilde{e}_{i',t+\tau-j}$, $i, i' = 1, 2, \ldots, m$.

8.3 Historical decomposition

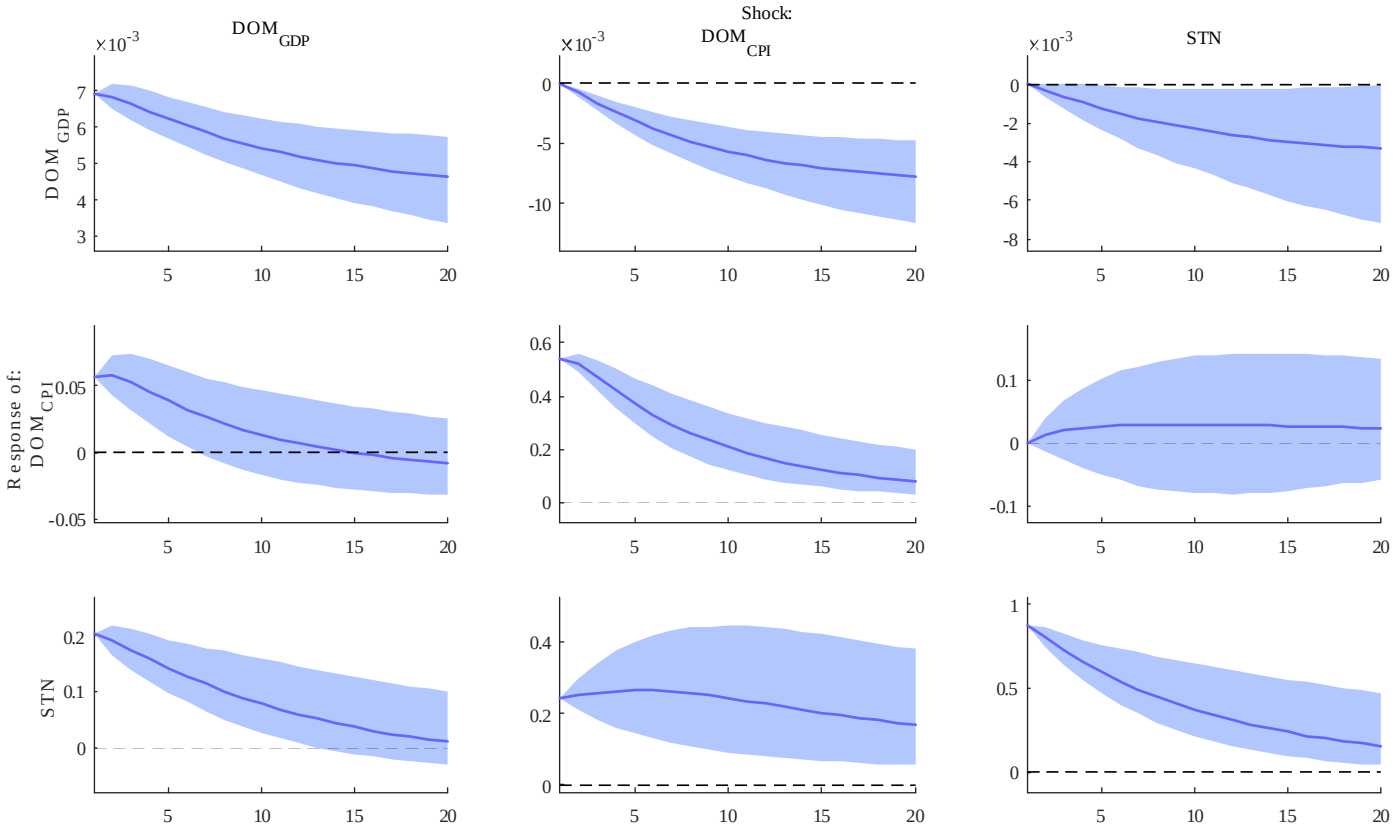
• What is the contribution of supply shocks to the productivity revival of the late 1990s?

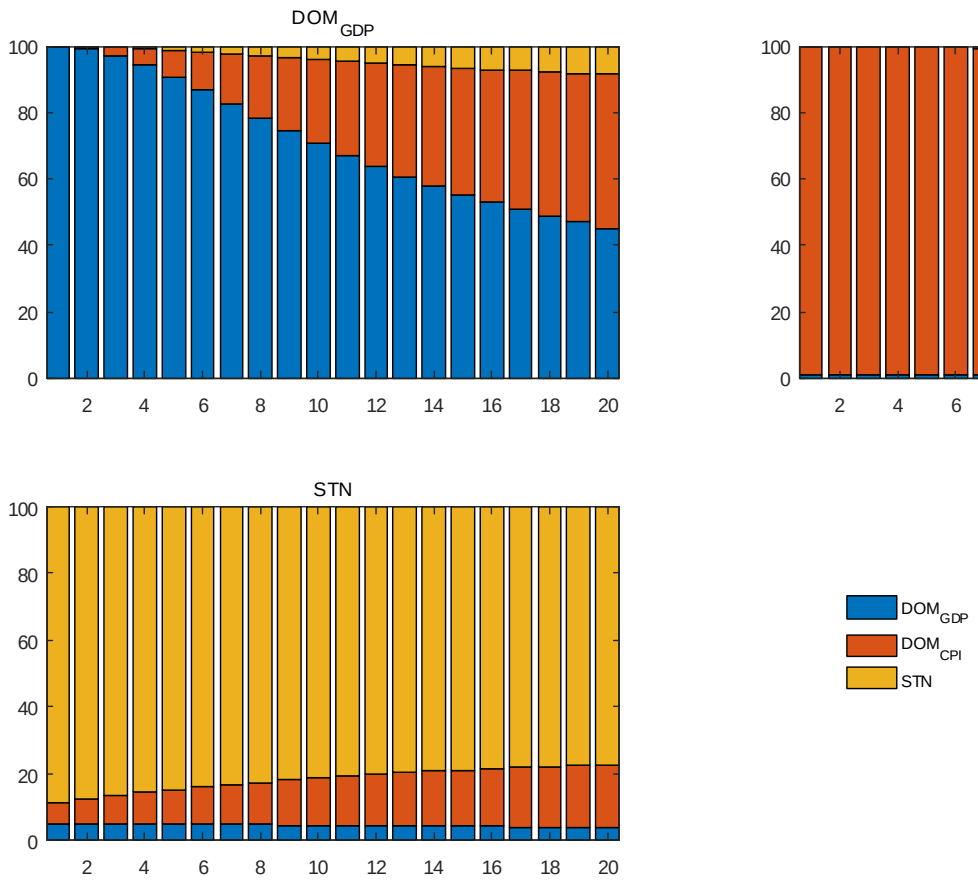
• Let $\hat{y}_{i,t}(\tau) = y_{i,t+\tau} - y_{i,t}(\tau)$ be the τ-steps ahead forecast error in the i-th variable of the VAR. Then:

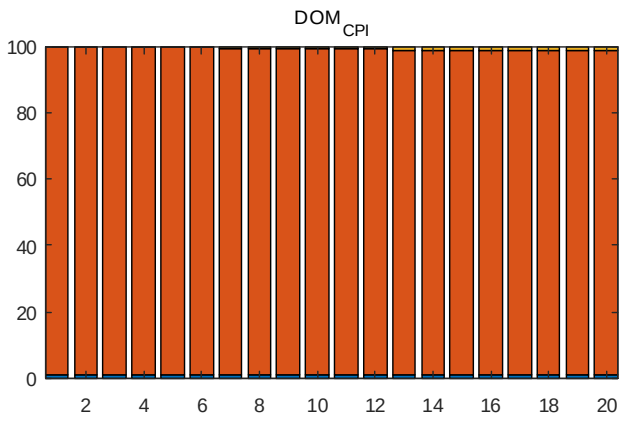
$$\hat{y}_{i,t}(\tau) = \sum_{i'=1}^{m} \tilde{D}^{i'}(\ell)\tilde{e}_{i't+\tau} \tag{93}$$

• Computes the path of $\hat{y}_{i,t}(\tau)$ due to each $\tilde{e}_{i'}$.

• Same ingredients appear in the computation of impulse responses, variance and historical decompositions. Different packaging!!

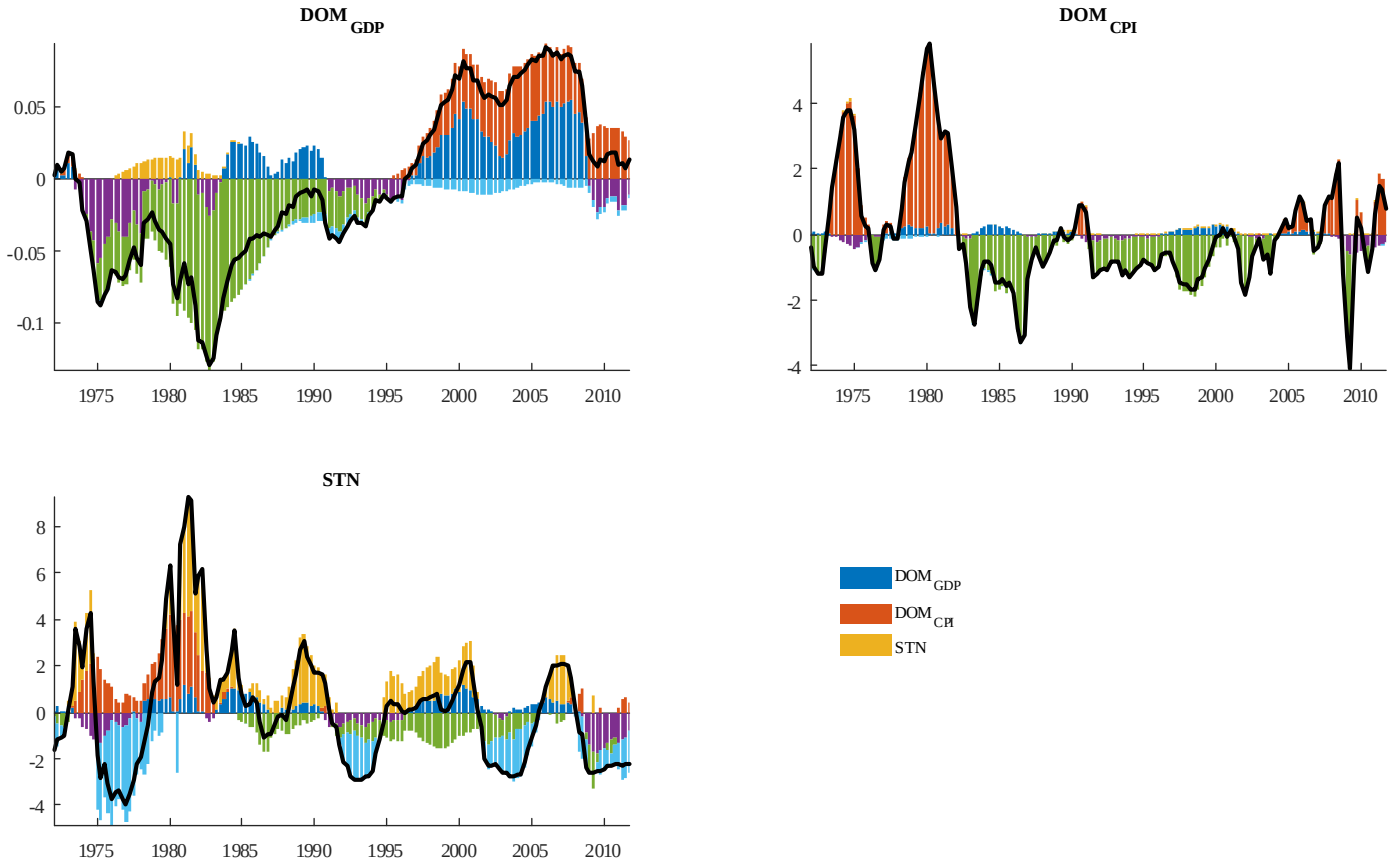

Shock:
DOM GDP
DOM CPI
STN
Response of:
DOM GDP
DOM CPI
STN
×10^-3


DOM GDP
100
80
60
40
20
0
2
4
6
8
10
12
14
16
18
20


DOM CPI
100
80
60
40
20
0
2
4
6
8
10
12
14
16
18
20

STN
100
80
60
40
20
0
2
4
6
8
10
12
14
16
18
20

DOM GDP
DOM CPI
STN


DOM GDP
0.05
0
-0.05
-0.1
1975
1980
1985
1990
1995
2000
2005
2010
DOM CPI
4
2
0
-2
-4
STN
8
6
4
2
0
-2
-4
DOM GDP
DOM CPI
STN

8.4 Impulse Responses in a VARX model

$$y_t = A(\ell)y_{t-1} + B(\ell)x_t + e_t \tag{94}$$

where x_t are exogenous variables, e.g. foreign variables for a domestic small open economy. Companion form:

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{R}x_t + \mathbf{E}_t \tag{95}$$

where

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \dots & A_p & B_1 & \dots & B_{q-1} & B_q \\ I_m & 0 & \dots & 0 & \dots & \dots & 0 & 0 \\ 0 & I_m & \dots & 0 & \dots & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & I_m & 0 \end{bmatrix}$$

$$\mathbf{E} = \begin{bmatrix} e_t \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} B_0 \\ 0 \\ \vdots \\ 0 \\ I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \mathbf{Y}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \\ x_t \\ x_{t-1} \\ \vdots \\ x_{t-q+1} \end{bmatrix}$$

• Let $J = [I_m, 0, 0 \ldots, 0]$. Repeatedly substituting into (95) we have

$$\mathbf{Y}_t = \sum_{i=0}^{\infty} J\mathbf{A}^i\mathbf{R}x_{t-i} + \sum_{i=0}^{\infty} J\mathbf{A}^i\mathbf{E}_{t-i} \tag{96}$$

• The responses of $\mathbf{Y}_t$ to a standardized shock in x_t are $D_i = J\mathbf{A}^i\mathbf{R}\, chol(x)$, where $chol(x)$ is the Cholesky decomposition of x_t.

9 Forecasting with BVAR (SBVAR) models

$$y_t = A(\ell)y_{t-1} + e_t \quad (97)$$

$$\mathcal{A}_0 y_t = \mathcal{A}(\ell)y_{t-1} + \epsilon_t \quad (98)$$

- Assume that we have a posterior distribution for $A(\ell), \mathcal{A}_0, \mathcal{A}(\ell)$.

- Transform (97) and(98) into a companion form.

$$Y_t = \mathbf{A}Y_{t-1} + E_t \quad (99)$$

$$A_0 Y_t = AY_{t-1} + \Upsilon_t \quad (100)$$

- Unconditional forecast: Set $E_{t+\tau} = 0 (\Upsilon_{t+\tau} = 0), \forall \tau > 0$.

- For fan charts (measuring forecast uncertainty):

  1. Draw $\mathbf{A}^l$ (A_0^l, A^l) from the available distribution, compute $Y_{t+\tau}^l$, $l = 1, 2, \ldots, L$, each horizon τ.

  2. Order $Y_{t+\tau}^l$ over l, each τ and extract median and posterior intervals (25-75, 16-84 or 2.5-97.5 percentiles).

  3. Or compute the mean and the standard deviation of $Y_{t+\tau}^l$ over l

• Conditional forecast 1: Manipulating shocks.

• This is the same as computing impulse responses (the impulse lasts longer than on period). Orthogonalize the disturbances if you have structural scenarios in mind.

• Choose $E_{jt+\tau} = \bar{E}_{jt+\tau}, (\Upsilon_{jt+\tau} = \bar{\Upsilon}_{jt+\tau}),\ \tau = 0, 1, 2, \ldots,$ some j.

• Given a draw $\mathbf{A}^l$ (A_0^l, A^l), find $Y_{t+\tau}^l = \mathbf{A}^l Y_{t+\tau-1} + E_{jt+\tau}$ $(A_0^l Y_{t+\tau} = A^l Y_{t+\tau-1} + \Upsilon_{t+\tau})$ and let the system run as in unconditional forecasts after the impulse has been exhausted.

• Use same algorithm employed for unconditional forecasts to quantify uncertainty

• Conditional Forecast 2: Manipulating endogenous variables.

• Separate $y_t = [y_t^A, y_t^B]$ and set $y_{t+\tau}^A = \bar{y}_{1t+\tau}^A,\ \tau = 0, 1, 2, ...$ Back out from the path of $E_{t+\tau}$ ($\Upsilon_{t+\tau}$) needed to produce $\bar{y}_{t+\tau}^A$. With this path compute the path for $y_{t+\tau}^B$ using (99)- (100).

• Potential identification problems. Many sources of shocks could produce the required path for $y_{t+\tau}^A$

Example 17 *Suppose that interest rates are (discretionarily) kept 50 basis point higher than the endogenous Taylor rule would imply. What is the effect on inflation? No identification problem: only a monetary shock enter the Taylor rule.*

Example 18 *Suppose that oil prices are expected to be 10 percent higher in the next two years. Problem: what has generated this increase? Is it demand? Is it supply? Is it a combination of the two?*

Example 19

$$x_{1t} = B_{11}(L)x_{1t-1} + B_{12}(L)x_{2t-1} + A_{11}u_{1t} + A_{12}u_{2t} \quad (101)$$
$$x_{2t} = B_{21}(L)x_{1t-1} + B_{22}(L)x_{2t-1} + A_{12}u_{1t} + A_{22}u_{2t} \quad (102)$$

where u_{1t} are real (domestic) and u_{2t} are nominal (international) shocks.

- *What is the effect of u_{2t} on x_{1t}? Problem x_{1t} is endogenously reacting to x_{2t-1}. Setting $u_{2t} = 0$ not enough.*

- *Needs to design shocks that make $x_{2t} = 0$*

1. Solve (102) for $x_{2t} = (I - B_{22}(L))^{-1}(B_{21}(L)x_{1t-1} + A_{21}u_{1t} + A_{22}u_{2t})$
2. Select $\hat{u}_{2t} = -A_{22}^{-1}(I - B_{22}(L))^{-1}(B_{21}(L)x_{1t-1} + A_{21}u_{1t})$ so that $x_{2t} = 0, \forall t$
3. Measure the effect of u_{1t} on x_{1t}, conditional on $\hat{u}_{2t}$.

10 Large Scale BVARs

• Can use same technology in large scale BVARs.

• Need to take care of a few details as computations in Gibbs sampler may become extraordinary costly when N is large.

• Standard assumptions (Sims and Zha, 1998, Banburra et al., 2010):

- Homoskedastic VAR errors: $e_t \sim (0, \Sigma)$

- Kroneker structure on VAR coefficients variance: $var(\beta) = \Sigma \otimes \Omega$.

- Minnesota prior on Ω, with $\psi_1 = h(N)$, where $h'(N) < 0$.

• With these assumptions, the conditional posterior for the VAR coefficients has precision matrix of the form $\Sigma \otimes (\Omega^{-1} + \sum_t y_{t-1} y_{t-1}^T)$ and the two terms in the Kroneker product can be manipulated separately.

• The system can be estimated equation by equation. Computation reduction from N^6 to N^3.

• Kroneker structure convenient but restrictive:

- It prevents asymmetries across equations a-priori.

- It implies that prior beliefs across equations are correlated (Σ full matrix).

• Alternative: Carriero et al. (2019): factorize likelihood and prior to estimate the model equation by equation which permits:

- Heteroskedastic VAR errors (stochastic volatility).

- General prior (besides conjugate Normal-Wishart can use independent Normal- inverted Wishart; normal-diffuse).

- Computational burden reduced from N^6 to N^4.

• Let $e_t = A^{-1}\Lambda_t^{-0.5}\epsilon_t$, where A is lower triangular and Λ_t is a diagonal matrix of stochastic volatility terms.

• Idea: Consider a VAR with two variables:

$$\begin{aligned} y_{1t} &= \beta_{10} + \sum_i \sum_l \beta_{1,i,l} y_{i,t-l} + h_{1t}^{-0.5} \epsilon_{1t} \\ y_{2t} &= \beta_{20} + \sum_i \sum_l \beta_{2,i,l} y_{i,t-l} + a_{21} h_{1t}^{-0.5} \epsilon_{1t} + h_{2t}^{-0.5} \epsilon_{2t} \end{aligned} \quad (103)$$

Given $\beta_{10}, \beta_{1,i,l}, A$, $h_{1t}^{-0.5}\epsilon_{1t}$ is known (from the first equation) when drawing $\beta_{20}, \beta_{2,i,l}$, so we can split the problem of drawing β's into two pieces.

• Same idea applies to the prior:$g(\beta) = g(\beta_1)g(\beta_2|\beta_1)$.

• Here we can draw the reduced form VAR parameters recursively equation by equation, i.e $g(\beta_1|A, \Lambda, y)g(\beta_2|\beta_1, A, \Lambda, y)$.

• Since this is applied to the reduced form system, the order in which we label the equations is irrelevant.

Appendix

1) Methods to sample from the posterior $g(\alpha|y)$

• Direct sampling (see example 1).

• Sampling by parts. If $g(\alpha|y)$ has a complicated structure could partition $\alpha = (\alpha_1, \alpha_2)$ and $g(\alpha|y) = g(\alpha_1|y, \alpha_2)g(\alpha_2|y)$ and sample separately from the two pieces.

Example 20 *We use sampling by parts when we construct the predictive distribution of forecasts. In fact* $f(y^{T+\tau}|y^T) = \int f(y^{T+\tau}|y^T, \alpha)g(\alpha|y)d\alpha$. *Hence sample* α^l *from the posterior, use the model to forecast* $y^{T+\tau}$ *given* α^l, *and average over draws.*

• Sampling by parts is typically used to obtain the marginal posterior of α in a linear regression model.

Example 21 *Suppose* $g(\alpha|y,\sigma^2)$ *is* $N(\bar{y},\sigma^2/T)$ *and that* $g(\sigma^2|y)$ *is* $IG(0.5(T-1),0.5(T-1)s^2)$ *where* $\bar{y}$ *and* s^2 *are the sample mean and variance of* y_t*. Since* $g(\alpha|y)=\int g(\alpha|y,\sigma^2)g(\sigma^2|y)d\sigma^2$*, draw* $(\sigma^2)^l$ *from* $g(\sigma^2|y)$*, and draw* α *from* $g(\alpha|y,(\sigma^2)^l)$*. As* L *goes to infinity we will have a sample from* $g(\alpha|y)$*.*

• Sampling by inversion. If $y=f(x)$ is $U(0,1)$ a draw for x can be obtained drawing from a uniform draw for y applying $x=f^{-1}(y)$.

2) Metropolis-Hastings algorithm

• MH is a general purpose MCMC algorithm that can be used when 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 $\alpha^{\ell} = \alpha^{\ell-1}$.

• Iterations define a mixture of continuous and discrete transitions:

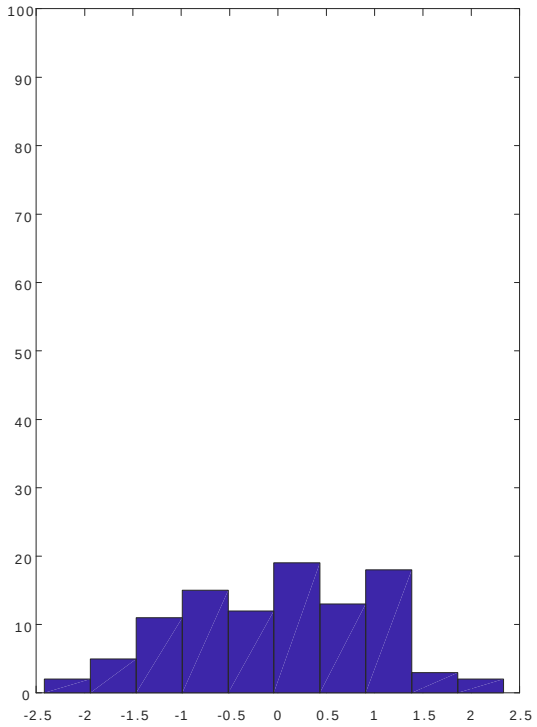
$$\begin{aligned} P(\alpha^{l-1},\ \alpha^{l}) &= q(\alpha^{l-1},\alpha^{l})\mathfrak{E}(\alpha^{l-1},\alpha^{l}) \qquad \text{if } \alpha^{l} \neq \alpha^{l-1} \\ &= 1 - \int_{A} q(\alpha^{l-1},\alpha)\mathfrak{E}(\alpha^{l-1},\alpha)d\alpha \quad \text{if } \alpha^{l} = \alpha^{l-1} \end{aligned} \tag{104}$$

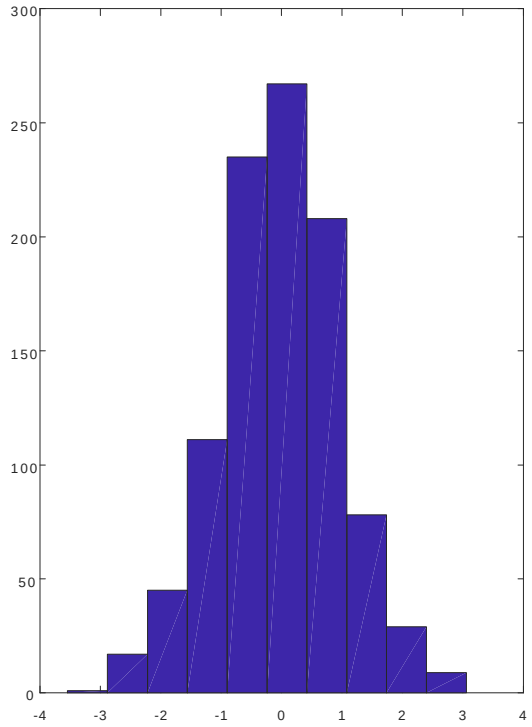
• $P(\alpha^{l-1},\ \alpha^{l})$ satisfies the conditions needed for existence, uniqueness and convergence.

• 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) $\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger}) = \frac{\breve{g}(\alpha^{l-1}|Y)}{\breve{g}(\alpha^{\dagger}|Y)}$. 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 $\mathfrak{E}(\alpha^{l-1}, \alpha^{\dagger})$.

• $q(\alpha^{l-1}, \alpha^{\dagger})$ 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.





- Left hand side 100 accepted draws; right hand side 1000 accepted draws.

• 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.

- 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 get reasonable acceptance rates. Standard choices for $\bar{q}$ are normal and t.

• General rule for selecting q. It 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 20-40%).

• Possible to nest Metropolis step within Gibbs sampler, if conditional distribution of some blocks does not have a closed form.

3) Matrix Algebra results

1) $A_{m\times n} \otimes B_{p\ \times q} = \begin{pmatrix} a_{11}B & a_{12}B & \dots & A_{1,n}B \\ \dots & \dots & \dots & \dots \\ a_{m1}B & a_{m2}B & \dots & A_{m,n}B \end{pmatrix}$

2) $vec(A)' = [a_{11}, a_{12}, \dots, a_{1n}, \dots, a_{m1}, a_{m2}, \dots, a_{mn}]$.

3) $vec(A')'vec(B) = tr(AB) = tr(BA) = vec(B')'vec(A)$

4) $vec(ABC) = (C' \otimes A)vec(B)$.

5)

$$
\begin{aligned}
tr(ABC) &= vec(A')'(C' \otimes I)vec(B) \\
&= vec(A')'(I \otimes B)vec(C) \\
&= vec(B')'(A' \otimes I)vec(C) \\
&= vec(B')'(I \otimes C)vec(A) \\
&= vec(C')'(B' \otimes I)vec(A) \\
&= vec(C')'(I \otimes A)vec(B)
\end{aligned}
$$

4) Some Distributions

1) Multivariate normal: $x_{(M\times 1)} \sim N(\mu, \Sigma)$

$$p(x) = (2\pi)^{-0.5M}|\Sigma|^{-0.5}exp\{0.5(x-\mu)'^{-1}(x-\mu)\} \tag{105}$$

2) Multivariate t: $x_{(M\times 1)} \sim t_v(\mu, \Sigma)$

$$p(x) = \frac{\Gamma(0.5(\nu+M))}{\Gamma(0.5\nu)(\nu\pi)^{0.5}\sigma}|\Sigma|^{-0.5}exp\{1+\frac{0.5}{\nu}(x-\mu)'^{-1}(x-\mu)\}^{-0.5(\nu+M)} \tag{106}$$

3) Inverse Wishart $A_{(M\times M)} \sim W(S^{-1}, \nu)$

$$\begin{aligned} p(A) &= (2^{0.5\nu M}\pi^{0.25M(M-1)}\prod_{i}^{M}\Gamma(0.5(\nu+1-i))^{-1}|S|^{0.5\nu} \\ &\quad |A|^{-0.5(\nu+M+1)}exp(-0.5tr(SA^{-1})) \end{aligned} \tag{107}$$

5) Wold theorem and the news

Wold Theorem: Under **linearity** and **stationarity**, any $m \times 1$ vector of time series $y_t^\dagger$ can be written as $y_t^\dagger = a y_{-\infty} + \sum_{j=0}^{\infty} D_j e_{t-j}$, where $y_{-\infty}$ is a $k \times 1$ vector which contains information about $y_t^\dagger$ known in the infinite past (i.e. constants, time trends), a is a $m \times k$ matrix of coefficients; e_{t-j} are the news at $t-j$, D_j are $m \times m$ matrices each j.

• The Wold theorem tells us that, apart from initial conditions, time series are the discounted accumulation of news.

• Let $y_t \equiv y_t^\dagger - a y_{-\infty}$, A news $e_t = 1$ has D_0 effect on y_t, D_1 effect on y_{t+1}, D_2 on y_{t+2},.... Thus, y_t is a moving average (MA) of the news, i.e. $y_t = D(\ell) e_t$.

• Without linearity, the Wold representation is $y_t^\dagger = f_1(y_{-\infty}) + \sum_j f_2(e_{t-j})$, where f_1 and f_2 are non-linear functions.

• Without stationarity, The Wold representation is $y_t \equiv y_t^\dagger - a_t y_{-\infty}, = D_t(\ell)e_t$ where a_t and D_t are time varying.

• What are the news? If $\mathcal{F}_{t-1}$ is the information set available at $t-1$, the news are:

$$e_t \equiv y_t - E[y_t|\mathcal{F}_{t-1}] \tag{108}$$

where $E[y_t|\mathcal{F}_{t-1}]$ is the mathematical conditional expectation of y_t.

Two issues

a) News are unpredictable, ($E(e_t|\mathcal{F}_{t-1}) = 0$), but contemporaneously correlated (Σ_e is not diagonal).

To give a *name* to the news, find a matrix $\tilde{\mathcal{P}}$ such that $\tilde{\mathcal{P}}\tilde{\mathcal{P}}' = \Sigma_e$. Then:

$$y_t = D(\ell)\tilde{\mathcal{P}}\tilde{\mathcal{P}}^{-1}e_t = \tilde{D}(\ell)\tilde{e}_t \quad \tilde{e}_t \sim (0, \tilde{\mathcal{P}}^{-1}\Sigma_e\tilde{\mathcal{P}}^{-1'} = I) \tag{109}$$

Examples of $\tilde{\mathcal{P}}$: Cholesky (lower triangular) factor; $\tilde{\mathcal{P}} = \mathcal{P}\Lambda^{0.5}$; where $\mathcal{P}$ is the eigenvector matrix, Λ the eigenvalue matrix, etc.

Example 22 *If* $\Sigma_e = \begin{bmatrix} 1 & 4 \\ 4 & 25 \end{bmatrix}$, $\tilde{\mathcal{P}} = \begin{bmatrix} 1 & 4 \\ 0 & 3 \end{bmatrix}$ *so that* $\tilde{\mathcal{P}}^{-1}e_t \sim (0, I)$.

• Orthrogonalization does not imply that the news have economic interpretation.

b) The news are not uniquely defined. For any matrix $\mathcal{H}$ such that $\mathcal{H}\mathcal{H}' = I$

$$y_t = D(\ell)e_t = D(\ell)\mathcal{H}\mathcal{H}'e_t = \tilde{D}(\ell)\tilde{e}_t \tag{110}$$

and $E(e_t e_t') = E(\tilde{e}_t \tilde{e}_t') = \mathbf{\Sigma}_e$.

• Standard to choose the **fundamental** representation: i.e. the one for which D_0 is the largest than all D_j coefficients, $j \geq 1$.
Formally: $det[D_0 E(e_t, e_t')D_0'] > det[D_j E(e_{t-j}e_{t-j}')D_j'],\ \forall j$.

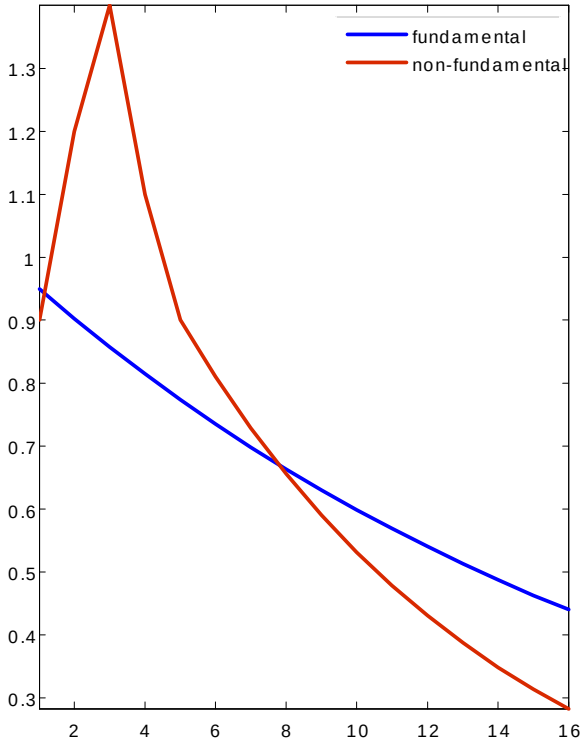
• Why are fundamental representations chosen?

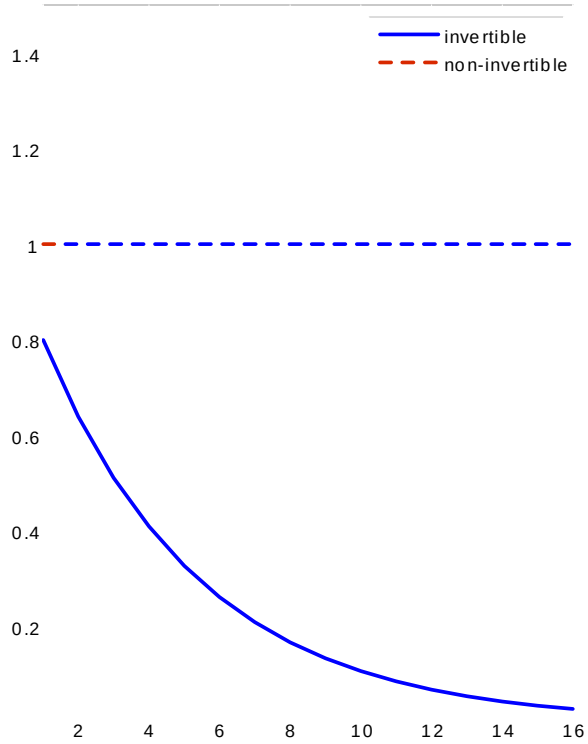
- In fundamental representations the information present in y_t's equals the information present in e_t's.

- In fundamental representations, the effect of news at t on y_t is larger than the effect on $y_{t+\tau}, \tau \neq 0$.

• If D_j decays with j, the importance of past news is smaller than importance of current news. Makes sense.

• Some economic models (e.g. models where news are anticipated - see later on) do not have this property and imply non-fundamental representations.





- Non-invertibility is a special case of non-fundamentalness.

From MAR to VARs

• If the D_j coefficients decay to zero "fast enough", $D(\ell)$ is *invertible* and

$$\begin{aligned} y_t &= D(\ell)e_t \\ D(\ell)^{-1}y_t &= e_t \\ y_t &= A(\ell)y_{t-1} + e_t \end{aligned} \tag{111}$$

where $I - A(\ell) \equiv D(\ell)^{-1}$. Note:

• $A(\ell)$ is of infinite length.

• With finite T, can't run a VAR(∞). A VAR(q), q fixed, approximates y_t well if D_j are close to zero for $j > q$.

• **Granger non-causality/Sims exogeneity**

Assume Σ_e diagonal and let $D(\ell) = \begin{bmatrix} D_{11}(\ell) & D_{12}(\ell) \\ D_{21}(\ell) & D_{22}(\ell) \end{bmatrix}$.

Granger Noncausality: y_{2t} fails to Granger Cause (GC) y_{1t} iff $D_{12}(\ell) = 0$.

Sims exogeneity: We can write $y_{2t} = Q(\ell)y_{1t} + e_{2t}$, with y_{1t} exogenous if and only if y_{2t} fails to GC y_{1t} and $D_{21}(\ell) \neq 0$.

• **Stability**: A VAR(1) stable if $det(I_m - Az) \neq 0, \forall |z| \leq 1$ (i.e. if all eigenvalues of A have modulus less than 1).

A VAR(q) stable if $det(I_{mq} - \mathbf{A}z) \neq 0, \forall |z| \leq 1$, where $\mathbf{A}$ is the companion matrix of the VAR (see below). This is equivalent to $det(I_m - A_1 z - \ldots - A_q z^q) \neq 0 \ \forall |z| \leq 1$.

Summary

- Any vector of time series has a linear VAR(∞) under certain assumptions.

- To give the news a name we need to orthogonalize them.

- With a finite sample of data need to carefully select the lag length of the VAR to make sure the news are unpredictable.

- If we want a constant coefficient VAR, we need stationarity of y_t. If non-stationarities are present a VAR representation exists, but with time varying coefficients.

6) Classical Specification

Selecting the lag length of a VAR

A) Likelihood ratio (LR) test

$$LR = 2[\ln \mathcal{L}(\alpha^{un}, \Sigma_e^{un}) - \ln \mathcal{L}(\alpha^{re}, \Sigma_e^{re})] \quad (112)$$

$$= T(\ln |\Sigma_e^{re}| - \ln |\Sigma_e^{un}|) \xrightarrow{D} \chi^2(\nu) \quad (113)$$

where $\mathcal{L}$ is the likelihood function, "UN"("RE") denotes the unrestricted (restricted) estimator, $\nu =$ number of restrictions of the form $R(\alpha) = 0$.

• LR test biased in small samples. If T small, it is better to use:

$$LR^c = (T - qm)(\ln |\Sigma^{re}| - \ln |\Sigma^{un}|)$$

where $q =$ number of lags, $m =$ number of variables.

• Sequential testing approach ("general-to-specific")

1) Choose an upper $\bar{q}$

2) Test a VAR($\bar{q}-1$) against a VAR($\bar{q}$), if do not reject

3) Test a VAR($\bar{q}-2$) against a VAR($\bar{q}-1$)

4) Continue until rejection.

LR is an in-sample criteria. What if we want a VAR for out-of-sample purposes?

Let $\Sigma_y(1) = \frac{T+mq}{T}\Sigma_e$ (q =number of lags, m = size of y_t, T sample size)

.

B) AIC criterion: $\min_q AIC(q) = \ln|\Sigma_y(1)|(q) + \frac{2qm^2}{T}$

• AIC is inconsistent. It overestimates true order q with positive probability.

C) HQC criterion: $\min_q HQC(q) = \ln|\Sigma_y(1)|(q) + (2qm^2)\frac{\ln\ln T}{T}$

• HQC is consistent (in probability).

D) SWC criterion: $\min_q SWC(q) = \ln|\Sigma_y(1)|(q) + (qm^2)\frac{\ln T}{T}$

• SWC is strongly consistent (a.s.).

• Criteria B)-D) trade-off the fit of the model (the size of Σ_e) with the number of parameters of the model ($m * q$), for a given sample size T. Hence, criteria B)-D) prefer smaller to larger scale models.

Criterion	T=40			T=80			T=120			T=200		
	q=2	q=4	q=6	q=2	q=4	q=6	q=2	q=4	q=6	q=2	q=4	q=6
AIC	1.6	3.2	4.8	0.8	1.6	2.4	0.53	1.06	1.6	0.32	0.64	0.96
HQC	0.52	4.17	6.26	1.18	2.36	3.54	0.83	1.67	2.50	0.53	1.06	1.6
SWC	2.95	5.9	8.85	1.75	3.5	5.25	1.27	2.55	3.83	0.84	1.69	2.52

Table 2: Penalties of AIC, HQC, SWC, m=4

- Penalties increase with q and fall with T. Penalty of SWC is the harshest.

- Ivanov and Kilian (2006): Quality of B)-D) depends on the frequency of data and on the DGP. Overall, HQC is the best.

- **Criteria A)-D) must be applied to the system not to single equations.**

Example 23 *VAR for the Euro area, 1980:1-2007:4; use output, prices, interest rates and M3, set* $\bar{q} = 7$.

Hypothesis	LR	LR^c	q	AIC	HQC	SWC
q=6 vs. q=7	2.9314e-5(∗)	0.1447	7	-7.556	-6.335	-4.482
q=5 vs. q=6	3.6400e-4	0.1171	6	-7.413	-6.394	-4.851
q=4 vs. q=5	0.0509	0.5833	5	-7.494	-6.675	-5.437
q=3 vs. q=4	0.0182	0.4374	4	-7.522	-6.905	-5.972
q=2 vs. q=3	0.0919	0.6770	3	-7.635(∗)	-7.219(∗)	-6.591
q=1 vs. q=2	3.0242e-7	6.8182e-3(∗)	2	-7.226	-7.012	-6.689(∗)

Table 3: Tests for the Lag length of a VAR

- *Different criteria choose different lag lenghts!.*

Checking Stationarity

All variable stationary/ all have unit roots $\longrightarrow$ easy.

Some cointegration. Transform VAR into VECM.

- Impose cointegration restrictions.
- Disregard cointegration restrictions.

Data is stationary. Can't see it because of small samples.

If Bayesian: stationarity/nonstationarity does not matter for inference.

Checking for Breaks

Wald test: $y_t = (A_1(\ell)\mathcal{I}_1)y_{t-1} + (A_2(\ell)\mathcal{I}_2)y_{t-1} + e_t$

$\mathcal{I}_1 = 0$ for $t \leq t_1$; $\mathcal{I}_1 = 1$ for $t > t_1$ and $\mathcal{I}_2 = 1 - \mathcal{I}_1$.

Use $S(t_1, T) = T(\ln |\Sigma_e^{re}| - \ln |\Sigma_e^{un}|) \xrightarrow{D} \chi^2(\nu)$; $\nu = dim(A_1(\ell))$ (Andrew and Ploberger (1994)).

If t_1 unknown, but belongs $[t^l, t^u]$ compute $S(t_1, T)$ for all the t_1 in the interval. Check for breaks using $max_{t_1} S(t_1, T)$.

Other useful specification tests

- Normality of residuals

Let $\hat{e}_t = y_t - \sum_j \hat{A}_j y_{t-j}$. Define $S_{1i} = \frac{1}{T}\sum_t \tilde{e}_{it}^3$; $S_{2i} = \frac{1}{T}\sum_t \tilde{e}_{it}^4$, $i = 1, \ldots, m$, $S_j = [S_{j1}, \ldots, S_{jm}]$, J $= 1, 2$. Then:

$$T^{0.5} \begin{bmatrix} S_1 \\ S_2 - 3 * I_m \end{bmatrix} \sim N(0, \begin{bmatrix} 6 * I_m & 0 \\ 0 & 24 * I_m \end{bmatrix})$$

- Non-linearities in the data

Regress $\hat{e}_t$ on y_{t-1}^2, $\log y_{t-1}$, etc. Check significance. Normality test is a kind of test for non-linearities.

Example 24 *Using a VAR(2) for the Euro area: test for normality; test for non-linearities using F-test of regression on squared lag residuals.*

	Skewness	Kurthosis	Nonlinear
e1	0.75	0.00	0.95
e2	0.99	0.00	0.99
e3	0.90	0.00	0.81
e4	0.96	0.00	0.97

Table 4: Normality and nonlinearity tests, VAR(2): significance level

Other specification issues

- Include constant? Constant and trend? Seasonal or break dummies? Check the news.

- Small scale vs. large scale VAR? Core VAR plus rotating variables?

- Maximize the sample size or look for stable periods?

- Standardize the data?

- Variables in level? Log-level? Growth rates? Real per-capita?

7) Classical parameters and covariance matrix estimation

Unrestricted VAR(q)

Assume that $y_{-q+1}, \ldots, y_0$ are known and $e_t \sim N(0, \Sigma_e)$ then

$$y_t|(y_{t-1}, \ldots, y_0, y_{-1}, y_{-q+1}) \sim N(A(\ell)y_{t-1}, \Sigma_e) \quad (114)$$

$$\sim N(\mathbf{A}_1'\mathbf{Y}_{t-1}, \Sigma_e) \quad (115)$$

where $\mathbf{A}_1'$ is the first row of $\mathbf{A}$ $(m \times mq)$. Let $\alpha = vec(\mathbf{A}_1)$.

Since $f(y_t|y_{t-1}, \ldots, y_{-q+1}) = \prod_j f(y_j|y_{j-1}, \ldots, y_{-q+1})$

$$\begin{aligned} \ln \mathcal{L}(\alpha, \Sigma_e) &= \sum_j \ln \mathcal{L}(y_j|y_{j-1}, \ldots, y_{-q+1}) \\ &= -\frac{Tm}{2}\ln(2\pi) + \frac{T}{2}\ln|\Sigma_e^{-1}| \\ &\quad -\frac{1}{2}\sum_t (y_t - \mathbf{A}_1'\mathbf{Y}_{t-1})'\Sigma_e^{-1}(y_t - \mathbf{A}_1'\mathbf{Y}_{j-1}) \quad (116) \end{aligned}$$

Setting $\frac{\partial \ln \mathcal{L}(\alpha,\Sigma_e)}{\partial \alpha} = 0$ we have

$$\mathbf{A}'_{1,ML} = [\sum_{t=1}^{T} \mathbf{Y}_{t-1}\mathbf{Y}'_{t-1}]^{-1}[\sum_{t=1}^{T} \mathbf{Y}_{t-1}y'_t] = \mathbf{A}'_{1,OLS} \tag{117}$$

and j-th column (a $1 \times mq$ vector) is

$$\mathbf{A}'_{1j,ML} = [\sum_{t} \mathbf{Y}_{t-1}\mathbf{Y}'_{t-1}]^{-1}[\sum_{t=1}^{T} \mathbf{Y}_{t-1}y_{jt}] = \mathbf{A}'_{1j,OLS} \tag{118}$$

• Why is OLS equivalent to maximum likelihood?

- If the initial conditions are known, maximizing the log-likelihood is equivalent to minimizing the sum of square errors!

• Why is it that single equation OLS is the same as full information ML?

- There are the same regressors in every equation! A VAR is a SUR system.

Plugging $\mathbf{A}_{1,ML}$ into $\ln \mathcal{L}(\alpha, \Sigma_e)$, we obtain the concentrated likelihood

$$\ln \mathcal{L}(\Sigma_e) = -\frac{T}{2}(m \ln(2\pi) + \ln |\Sigma_e^{-1}|) - \frac{1}{2}\sum_{t=1}^{T} e'_{t,ML}\Sigma_e^{-1}e_{t,ML} \quad (119)$$

where $e_{t,ML} = (y_t - \mathbf{A}_{1,ML}Y_{t-1})$. Using $\frac{\partial(b'\mathcal{Q}b)}{\partial\mathcal{Q}} = b'b$; $\frac{\partial \ln|\mathcal{Q}|}{\partial\mathcal{Q}} = (\mathcal{Q}')^{-1}$ we have $\frac{\partial \ln \mathcal{L}(\Sigma_e)}{\partial\Sigma} = \frac{T}{2}\Sigma'_e - \frac{1}{2}\sum_{t=1}^{T} e_{t,ML}e'_{t,ML} = 0$ or

$$\Sigma'_{ML} = \frac{1}{T}\sum_{t=1}^{T} e_{t,ML}e'_{t,ML} \quad (120)$$

and $\sigma_{i,i'} = \frac{1}{T}\sum_{t=1}^{T} e_{i't,ML}e'_{it,ML}$.

$\Sigma'_{ML} \neq \Sigma'_{OLS} = \frac{1}{T-1}\sum_{t=1}^{T} e_{t,ML}e'_{t,ML}$. Close for large T.

VAR(q) with restrictions

Assume restrictions are of the form $\alpha = R\theta + r$, where R is $mk \times k_1$ matrix of rank k_1; r is a $mk \times 1$ vector; θ a $k_1 \times 1$ vector.

Example 25 *i) Lag restrictions:* $A_q = 0$. *Here* $k_1 = m^2(q-1)$, $r = 0$, *and* $R = [I_{m_1}, 0]$.

ii) Block exogeneity of y_{2t} *in a bivariate VAR(2). Here* $R = blockdiag[R_1, R_2]$, *where* R_i, $i = 1, 2$ *is upper triangular.*

iii) Cointegration restrictions.

Plugging the restrictions in (19) we have

$$y = (I_m \otimes x)\alpha + e = (I_m \otimes x)(R\theta + r) + e$$

Let $y^\dagger \equiv y - (I \otimes x)r = (I \otimes x)R\theta + e$. Since $\frac{\partial \ln \mathcal{L}}{\partial \theta} = R\frac{\partial \ln \mathcal{L}}{\partial \alpha}$:

$$\theta_{ML} = [R'(\Sigma_e^{-1} \otimes x'x)R]^{-1}R[\Sigma_e^{-1} \otimes x]y^\dagger \quad (121)$$

$$\alpha_{ML} = R\,\theta_{ML} + r \quad (122)$$

$$\Sigma_e' = \frac{1}{T}\sum_t e_{ML}e'_{ML} \quad (123)$$

Alternative method

- Let the restrictions be of the form $R\alpha = r$, where R is a $mk \times mk$ matrix and r a $mk \times 1$ vector.

- Write the VAR as $\mathbf{Y}_t = \mathbf{A}_0 + \mathbf{A}_1\mathbf{Y}_{t-1} + \mathbf{E}_t,\ \mathbf{E}_t \sim \mathbf{N}(0, \mathbf{\Sigma}_E)$ where $\mathbf{Y}_t, \mathbf{E}_t$ are $mk \times 1$ vectors and $\mathbf{A}$ is $mk \times mk$.

- Minimize $\sum_t \mathbf{E}_t {\mathbf{\Sigma}_E}^{-1} \mathbf{E}_t$ subject to the restriction.

- Let $W = [1, \mathbf{Y}_{-1}], \alpha = vec([\mathbf{A}_0', \mathbf{A}_1']'), S = (\mathbf{\Sigma}_E \otimes W'W), \hat{\alpha} = I \otimes (W'W)^{-1}(W'\mathbf{Y})$.

- The objective function is

$$\min(\alpha - \hat{\alpha})' S (\alpha - \hat{\alpha}) + \lambda(R\alpha - r) \tag{124}$$

- FOC: $S(\alpha - \hat{\alpha}) = R\lambda$.

- Since $R\alpha = r = R\hat{\alpha} + RS^{-1}R'\lambda'$ the (restricted) estimator is

$$\alpha = \hat{\alpha} + S^{-1}R'(RS^{-1}R')^{-1}(r - R\alpha) \tag{125}$$

Summary

- For a VAR(q) without restrictions:

  - ML and OLS estimators of $\mathbf{A}_1$ coincide.

  - OLS estimation of $\mathbf{A}_1$, equation by equation, is consistent and efficient (if assumptions are correct).

  - OLS and ML estimators of Σ_e asymptotically coincide for large T.

• For a VAR(q) with restrictions:

- ML estimator of $\mathbf{A}_1$ is different from the OLS estimator.

- ML is consistent/efficient if restrictions are true. It is inconsistent if restrictions are false.

In general:

• OLS consistent if stationarity assumption is wrong (t-tests incorrect).

• OLS inconsistent if lag length wrong (regressors correlated with error term).