Bayesian methods for structural VARs

Fabio Canova
Norwegian Business School, CAMP, and CEPR
September 2022

Outline

- Bayesian vs. frequentist. Bayes theorem. Likelihood and prior selection.
- Posterior simulators. MCMC methods.
- 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.

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*.

Canova, F. and Ferroni, F. (2019) "Mind the Gap! Stylized facts and structural models, manuscript.

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. (2017). Large vector autoregressions with stochastic volatility and non-conjugate priors. Bocconi University manuscript.

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?, forthcoming, AER: Papers and Proceedings.

Pagan, A. and Robinson, T. (2019), Implications of partial information for applied macroeconomic modelling, manuscript.

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.

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 are 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 are 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 in the world is classical (and under appropriate regularity conditions):

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

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

Classical and Bayesian analyses differ in small samples and 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

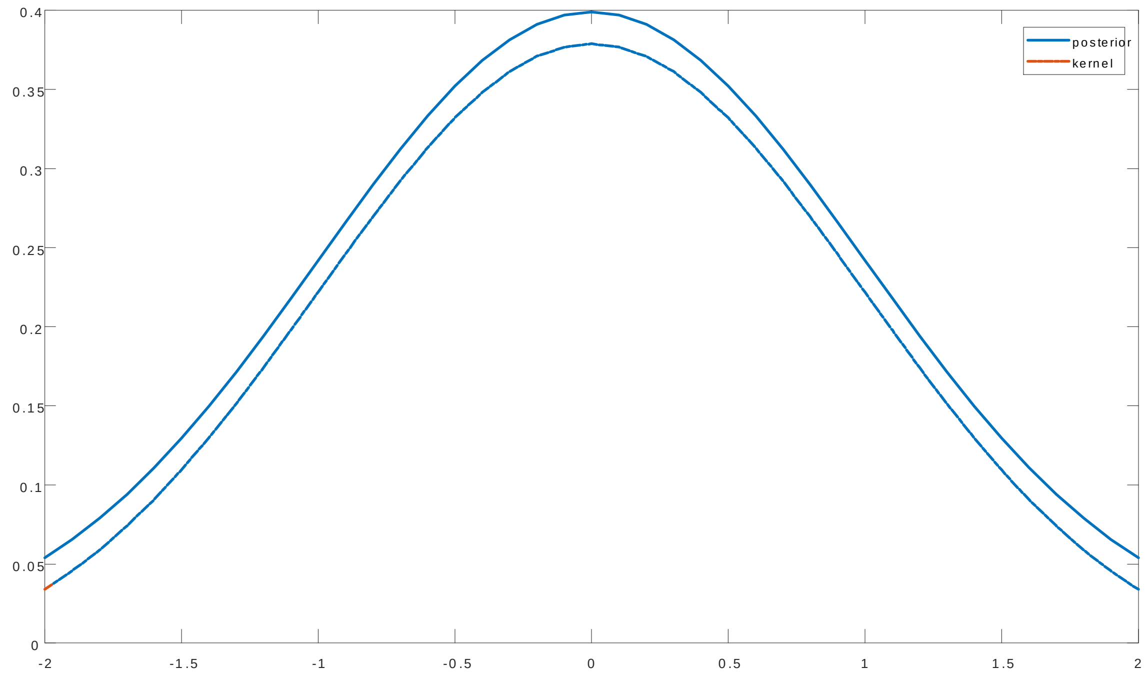
Parameters $\alpha \in A$, A compact subset of R^q. Prior information: $g(\alpha)$. Sample information: $f(y|\alpha) \equiv \mathcal{L}(\alpha|y)$.

• 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

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

• 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 uninterpretable).

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

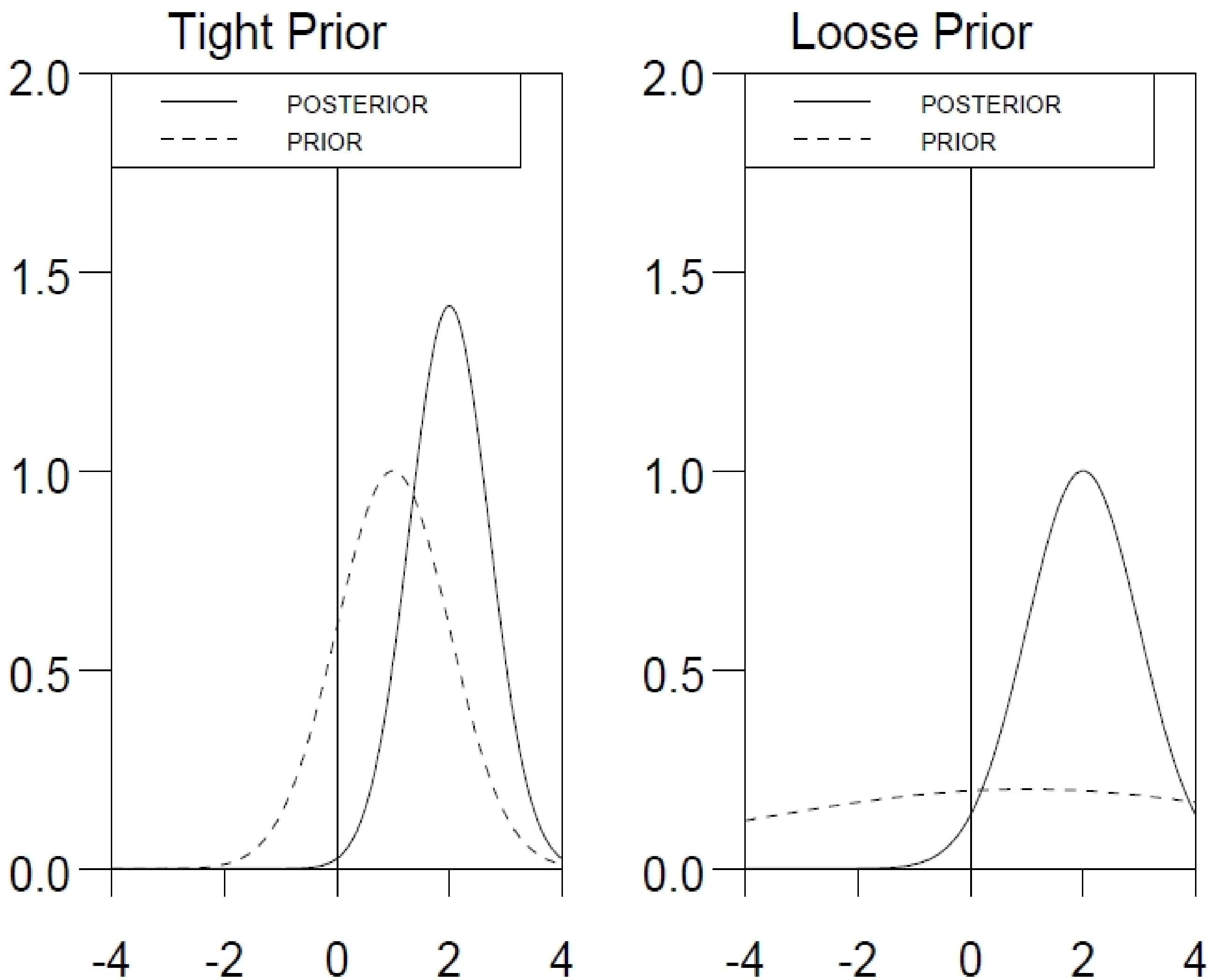
- Location-scale invariant prior : $g(\alpha, \sigma) = \sigma^{-1}$.

• 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
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:

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

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

- $y_1, \dots y_T$ 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).

- Bayesian computes: $E(h(\alpha|y)) = \int h(\alpha)g(\alpha|y)d\alpha$.

• 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 $h(\alpha_{.5}|y)$.

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

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

General problem: $E(h(\alpha|y))$, $h(\alpha_{.5}|y)$, $h(\alpha^*|y)$ can not be generally evaluated, since $f(y)$ typically requires integration in high dimensions

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

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

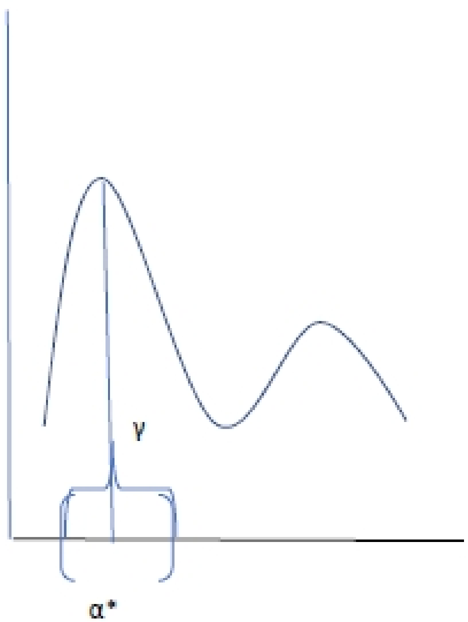
Example 1 *Need to compute* $Pr(\alpha > 0)$*. Draw* α^l *from* $g^{AP}(\alpha|y)$*. If* $\alpha^l > 0$*, set* $h(\alpha^l) = 1$*, else set* $h(\alpha^l) = 0$*. Draw L times and average* $h(\alpha^l)$ *over draws. The result is an estimate of* $Pr(\alpha > 0)$*.*

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

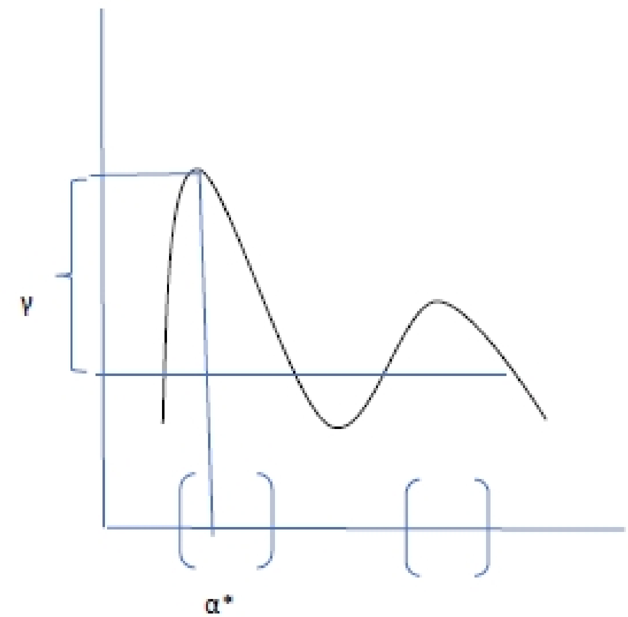
• By a central limit theorem (CLT) the difference between $\frac{1}{L}\sum_l h(\alpha^l)$ and E(h(α)) is normal with zero mean and fixed variance as L grows.

• Numerical standard errors $(\frac{1}{L}\sum_l (h(\alpha^l) - E(h(\alpha)))^2)$, or numerical *credible sets* $(h_{0.5\gamma}(\alpha), h_{1-0.5\gamma}(\alpha))$ 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

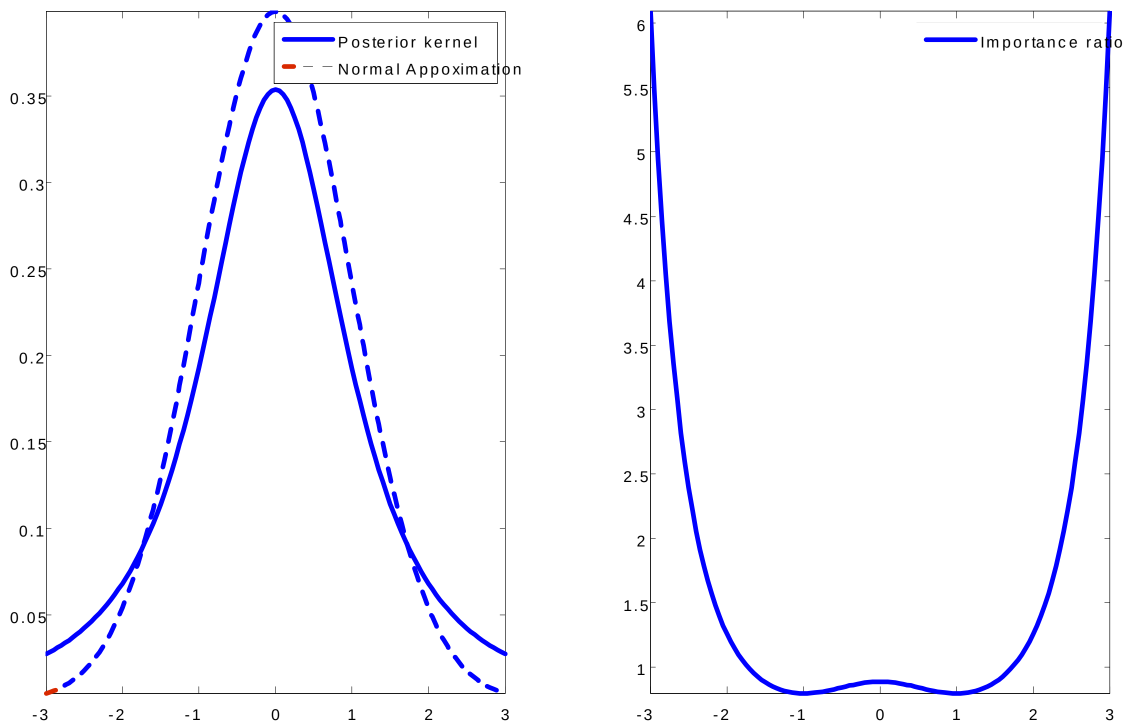
• If T is large then $g(\alpha|y) \approx f(\alpha|y)$, a normal density. If T is not large but $f(\alpha|y)$ is unimodal, roughly symmetric, and α^* is in the interior of A use:

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

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.



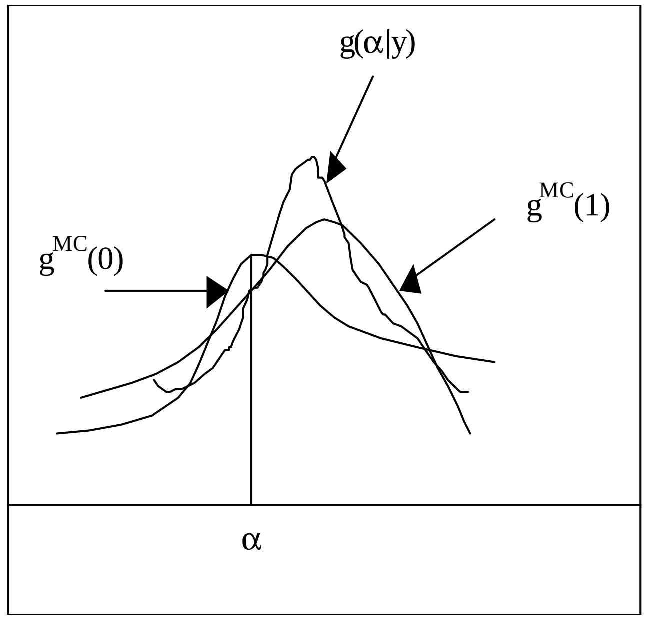
Posterior kernel
Normal Appoximation
0.35
0.3
0.25
0.2
0.15
0.1
0.05
-3
-2
-1
0
1
2
3
Importance ratio
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
-3
-2
-1
0
1
2
3

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 $(x_1, \dots x_n)$. Let $P(i,j) = Pr(x_{t+1} = x_j | x_t = x_i)$ and let $\mu(t) = (\mu_{1t}, \dots \mu_{nt})$ 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:

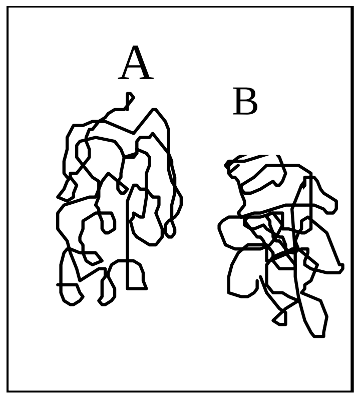
- P is irreducible, i.e. it has no absorbing state.

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

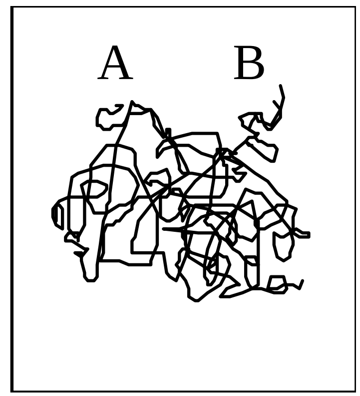
- P it is Harris recurrent, i.e. each cell is visited an infinite number of times with probability one.

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

- With $\alpha^{L+j} = (\alpha_1^{L+j}, \dots, \alpha_k^{l+j})$, 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* $f(x,y) \propto \frac{n!}{x!(n-x)!} y^{x+\alpha_0-1}(1-y)^{n-x+\alpha_1-1}, \quad x = 0,1,\ldots n,\ 0 \le y \le 1$ *(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* $(x+\alpha_0, n-x+\alpha_1)$*.*

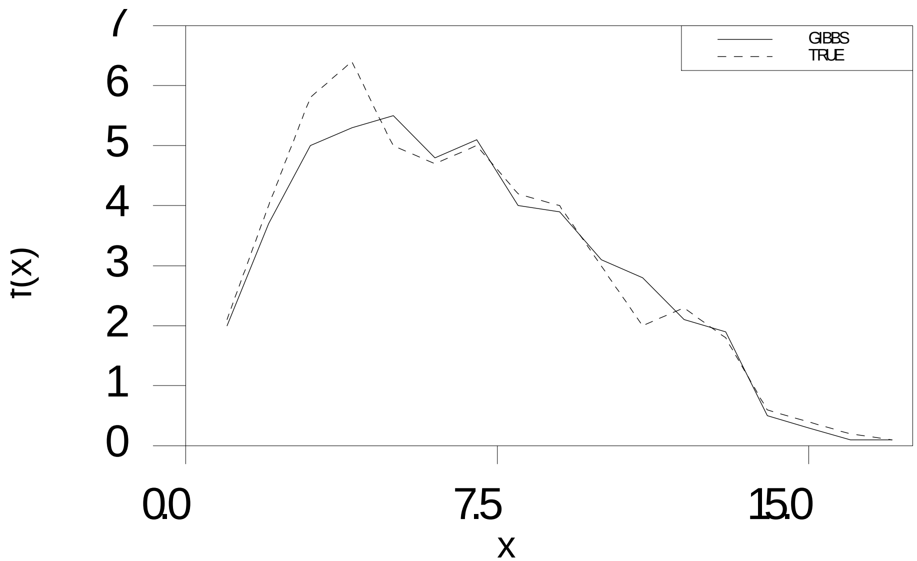


Figure: true/ Gibbs sampling marginal distribution for x with $L = 500,\ n = 100,\ \alpha_0 = 2,\ \alpha_1 = 4$ and $\bar{L} = 20$.

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*

$$y_t = x_t\alpha + e_t$$

$y_t = (y_{it}, \ldots, y_{mt})', x_t = \mathrm{diag}(x'_{it}, \ldots, x'_{mt})$, $\alpha = (\alpha'_1, \ldots \alpha'_m)'$ *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 of the form

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

where y_t is a $m \times 1$ vector; A_j are full rank, $m \times m$ matrices, $j = 1, \ldots q$; Σ_e is a full rank, $m \times m$ matrix.

all the correlations are taken care of the covariance matrix

- Small open economy version (VARX):

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

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

v_t is an imported shock, and uncorrelated with e_t

• Advantages of VARs setup:

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

- Simple to estimate.

• General disadvantages:

- 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).

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.

residuals are supposed to be serially uncorrelated -> test

• 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.

How do you estimate (Bayesian) VARs?

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

- Specify a prior distribution for coefficients and covariance matrix and a likelihood function.

Prior can be chosen to be anything. but different results from different priors have to be explained.

- Compute posterior distribution of coefficients and covariance matrix.

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

both depend on the loss function

4.1 Structural VARs

VARs are reduced form models:

i.e. structural shocks

- 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:

$$\mathcal{A}_0 y_t = \mathcal{A}_1 y_{t-1} + \ldots + \mathcal{A}_q y_{t-q} + \varepsilon_t \qquad \varepsilon_t \sim (0, \Sigma_\varepsilon) \tag{10}$$

where Σ_ϵ is diagonal. Its reduced form (VAR) is:

$$y_t = A_1 y_{t-1} + \ldots + A_q y_{t-q} + e_t \qquad e_t \sim (0, \Sigma_e) \tag{11}$$

• (11) easy to estimate. We want to go from (11) to (10). Since $A_j = \mathcal{A}_j \mathcal{A}_0^{-1}$, $e_t = \mathcal{A}_0^{-1}\varepsilon_t$, we just need to find $\mathcal{A}_0$.

Note that (11) and (10) imply

$$\mathcal{A}_0^{-1} \Sigma_\epsilon \mathcal{A}_0^{\prime -1} = \Sigma_e \tag{12}$$

linear combinations of structural shocks, with different weights defined by A_0^-1

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

• To recover unknown elements in $\mathcal{A}_0$ from (12) 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 (12) 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 methods in case of overidentification.

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

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

• For global identification conditions see Rubio et al. (2010).

Example 4 *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.*

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

No free columns which would be linear combinations of other columns. Zero can be replaced by -1 (perusteluineen)

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 and $\mathcal{A}_0^{-1}\epsilon_t = e_t$ to obtain structural shocks.

- 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); narrative and sign (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

- Write the VAR and SVAR in MA format:

growth-rate of y_t

$$\Delta y_t = D(\ell)e_t = D(1)e_t + D^*(\ell)\Delta e_t \quad (13)$$

$$\Delta y_t = \mathcal{D}(\ell)\mathcal{A}_0\epsilon_t = \mathcal{D}(\ell)(1)\mathcal{A}_0\epsilon_t + \mathcal{D}^*(\ell)\mathcal{A}_0\Delta\epsilon_t \quad (14)$$

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

Divide in two components

$$\mathcal{D}(1)\mathcal{A}_0\epsilon_t = D(1)e_t \quad (15)$$

Long-run shocks

$$\mathcal{A}_0\Delta\epsilon_t = \Delta e_t \quad (16)$$

If y_t is stationary, $\mathcal{D}(1) = D(1) = 0$ and only (16) is available.

If there is no cointegration, (15) is empty, as there is no long-run effect

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

Example 5 *In a bivariate VAR imposing (15) 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 (15)-(16). If $y_t = [\Delta y_{1t}, y_{2t}]$, $(m \times 1)$; y_{1t} are I(1); y_{2t} are I(0) and $y_t = \bar{y} + D(\ell)\epsilon_t$, where $\epsilon_t \sim iid(0, \Sigma_\epsilon)$

$$\begin{pmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ 0 \end{pmatrix} + \begin{pmatrix} D_1(1) \\ 0 \end{pmatrix} \epsilon_t + \begin{pmatrix} (1-\ell)D_1^\dagger(\ell) \\ (1-\ell)D_2^\dagger(\ell) \end{pmatrix} \epsilon_t \tag{17}$$

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

other is stationary, other is non-stationary

Problem: Provides Very weak identification, cholesky does not imply an unique system

4.3 Problems with standard identification approaches

- Many Cholesky systems - which one to choose?

Example 6

$$p_t = a_{11}e_t^s \tag{18}$$
$$y_t = a_{21}e_t^s + a_{21}e_t^d \tag{19}$$

- *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).*

$$y_t = a_{11}e_t^s \tag{20}$$
$$p_t = a_{21}e_t^s + a_{21}e_t^d \tag{21}$$

• *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*).*

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

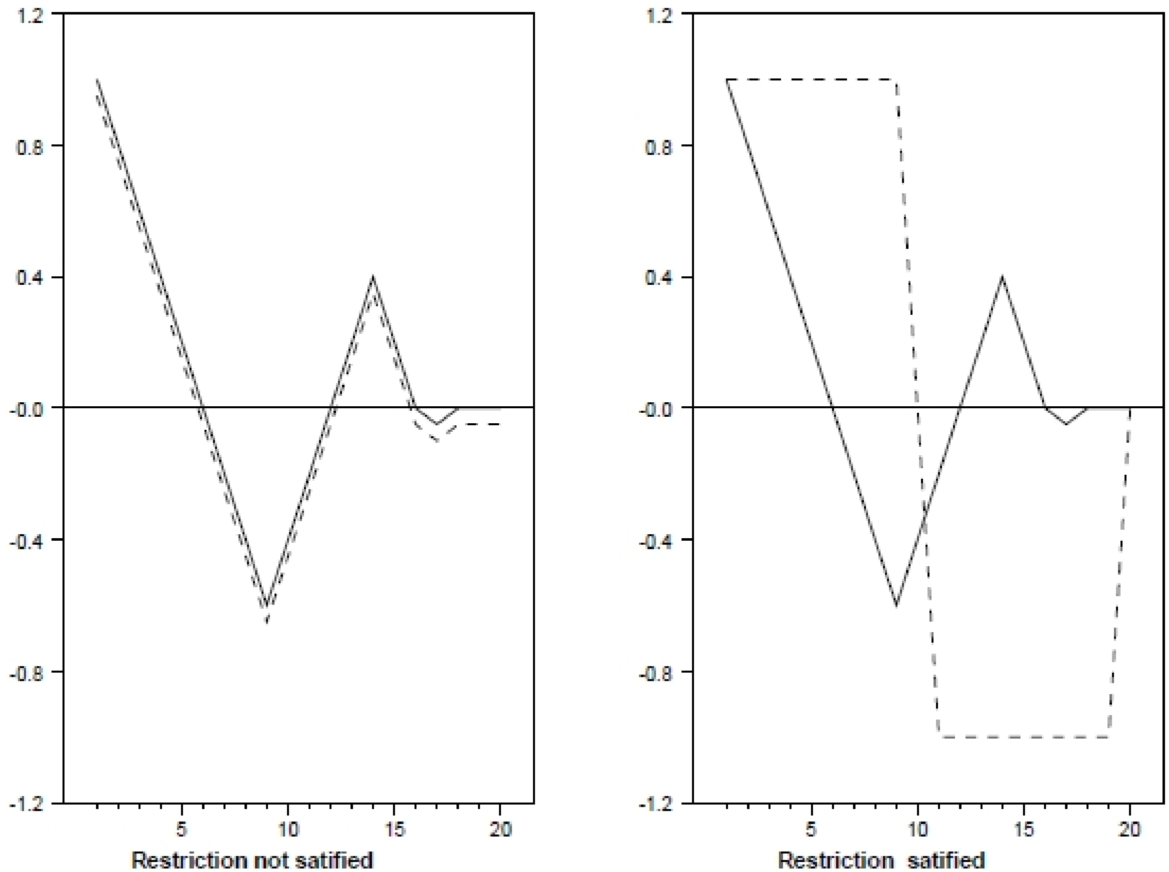
• Cooley-LeRoy (1985): unless strong restrictions are imposed (i.e. on the timing of information) dynamic models do not have a Cholesky structure.

Cholesky identification is bad -> make new methods -> just as bad -> go back to Cholesky (cycle!)

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

long-run identification restriction are restrictions on some coefficients, but do not tell anything about single coefficitent

Long-run identification requires a lot of data! 100 data points is not enough, shown by Monte-Carlo studies.



requires some smoothness condition that prevents something like pic on RHS happening

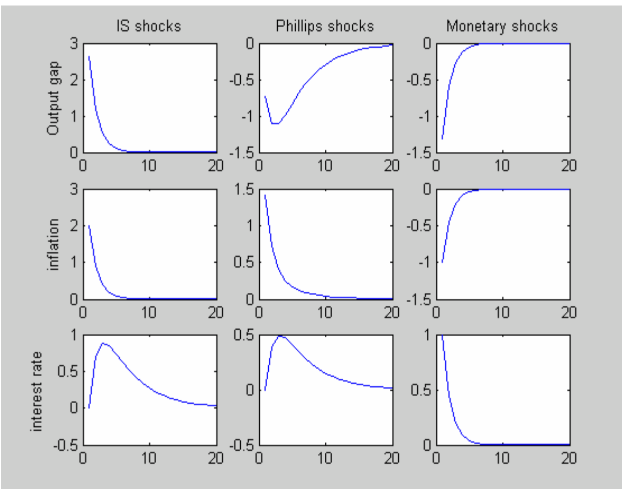
- Long run restrictions 2: Erceg, et. al (2005): long run restrictions give poor identification in small samples.

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

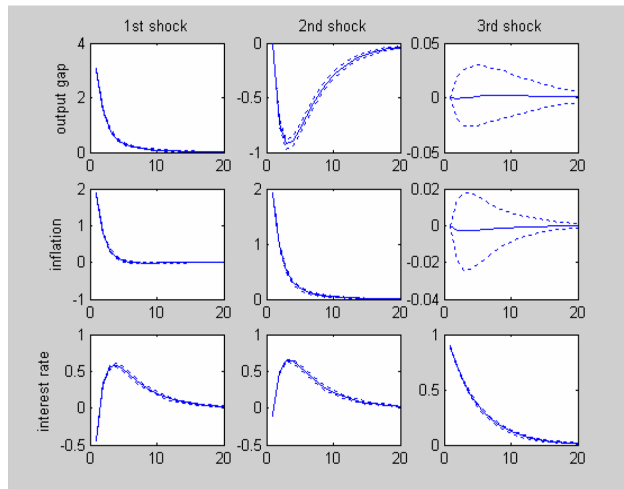
- Heteroschedasticity restrictions: is true that impact and dynamics unchanged and only variance changes? Why?

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

The DGP is a 3 equations New-Keynesian model



True responses



Inertial responses

some shocks pick up all the bad stuff, monetary policy shock in this case sign restrictions derived usually from theory

There is a set of sign restrictions that satisfy the requirement, often is seen what others have used.

Minimum and maximum path nowadays usually reported.

Sign restrictions

Example 7 *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 a $\mathcal{H}: \ \mathcal{H}\mathcal{H}' = I, \ \Sigma_e = \tilde{\mathcal{P}}\mathcal{H}\mathcal{H}' \ \tilde{\mathcal{P}}' = \hat{\mathcal{P}}\hat{\mathcal{P}}'$. 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^l_{jt} satisfying the restrictions, where l is the number of shocks found.

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

Example 8 *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* ω*, 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 $G = QR$.

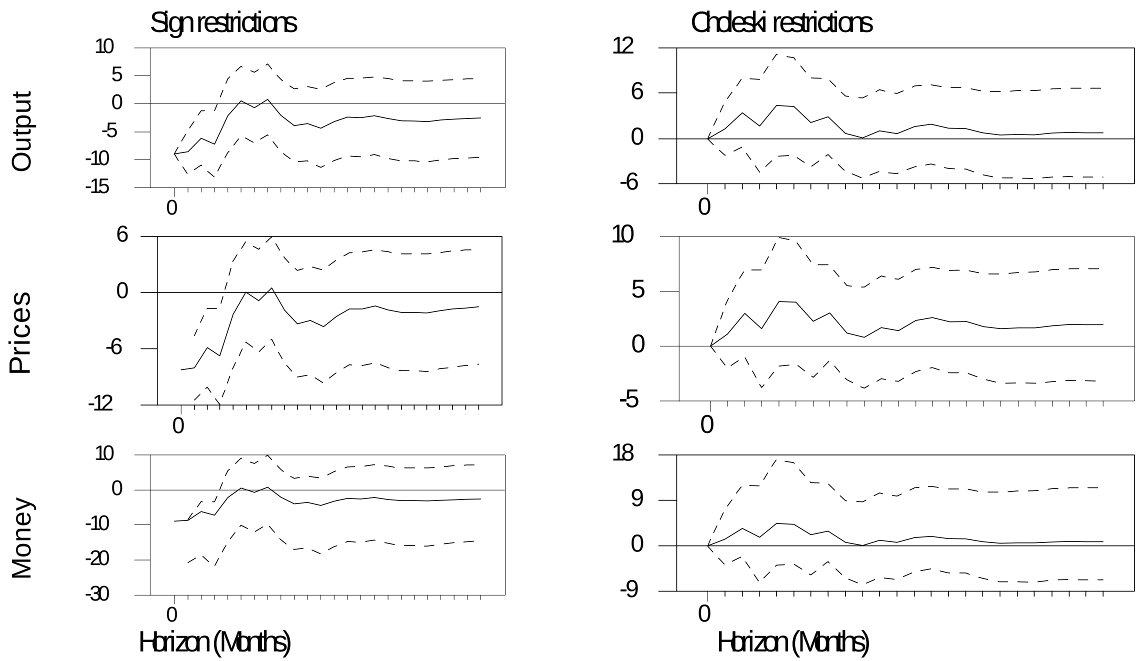
Q? is uniformly distributed.

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 9 *Comparing responses to US monetary shocks 1964-2001.*

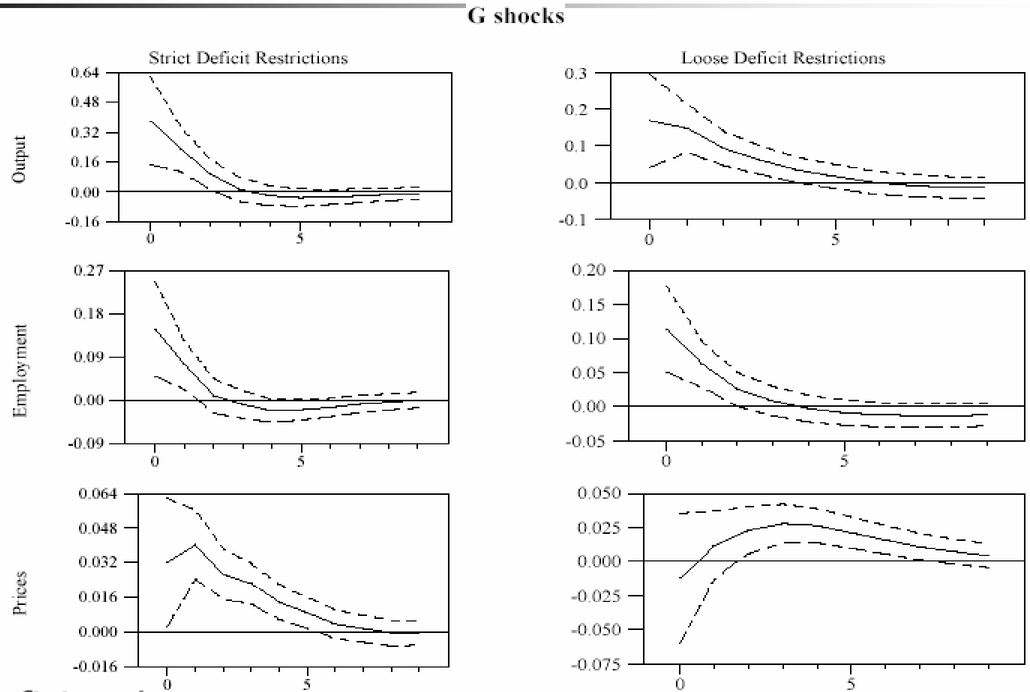


Example 10 *Studying the effects of fiscal shocks in US states: 1950-2005.*

	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	$= 0$

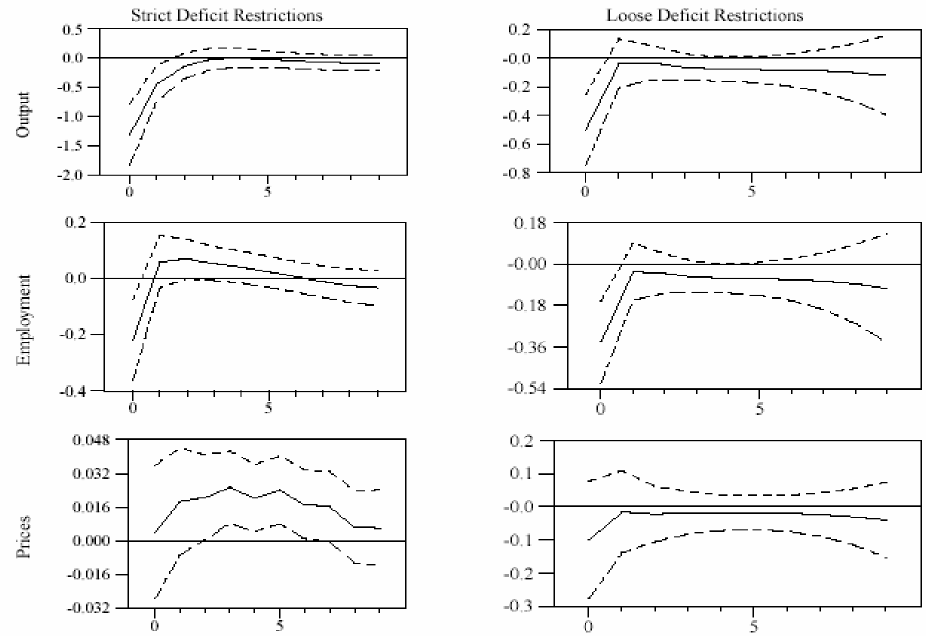
Table 1: Identification restrictions

The results: Is the transmission of fiscal shocks different?



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 too weak (Canova and Paustian, 2011).

signal-to-noise ratio will be very low, very high required (Monte-Carlo studies have shown)

- 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.).

you simulate for different values of parameters and see if they're robust

- Rank and order conditions do not apply here.

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

Kilian: inelasticity in the oil markets -> any results have to demonstrate this (kutakuinkin, katso paperi... ehkä soveltuu myös asuntomarkkinoille, jossa myös tarjonnan jäykkyyksiä

4.5 Additional identification restrictions

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

• **Maximal path restrictions**: e.g. an investment specific shock must generate the largest path for the structural shocks for the period 1996-2000 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.

SHocks should be i.i.d., saying for example the shock lasts for three months will make them hard to find

• **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)

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

usual IV requirements

- First stage: $\hat{\epsilon}_{1t} = (z_t' z_t)^{-1}(z_t \epsilon_{1t}')z_t$.

- Second stage: compute instantaneous responses regressing y_t on $\hat{\epsilon}_{1t}$. Compute lagged responses using $A_i(L)$ 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: $Y_t = C + AY_{t-1} + U_t$, $U_t \sim (0, \Sigma_U)$

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:
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.

No restrictions on the covariance matrix, except for the one equality

Example 11

$$p_t = \beta y_t + \epsilon_{1t} \quad (22)$$
$$y_t = \alpha p_t + \epsilon_{2t} \quad (23)$$

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

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

There are three free elements in $V, (v_{22} = v_{12})$ *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*

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

Volatility changes have to happen approximately at the same time!

$$V_2 = \frac{1}{(1-\alpha\beta)^2}\left[\begin{array}{cc} \beta^2\sigma_{22}^2 + \sigma_{12}^2 & \beta\sigma_{22}^2 + \alpha\sigma_{12}^2 \\ \beta\sigma_{22}^2 + \alpha\sigma_{12}^2 & \sigma_{22}^2 + \alpha^2\sigma_{12}^2 \end{array}\right]$$

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!!

3 in V_1 and 3 in V_2

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

• Crucial restrictions:

i) α and β are unchanged across regimes.

ii) The variance of both shocks changes.

iii) Volatility changes should 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 $E(e_t|I_{t-1}) = 0$, $cov(e_{jt}, e_{it}|I_{t-1}) = 0, \forall j, i$. Let $u_t = Me_t$, and $\sigma^2_{jt-1} = var(e_{jt}|I_{t-1}), \mathbf{\Sigma}_{t-1} = var(u_t|I_{t-1})$.

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

$$\mathbf{\Sigma}_t\mathbf{\Sigma}^{-1}_{t-1} = Mdiag\{\frac{\sigma^2_{jt}}{\sigma^2_{jt-1}}\}M^{-1} \qquad (24)$$

• Columns of M = eigenvectors of $\mathbf{\Sigma}_t\mathbf{\Sigma}^{-1}_{t-1}$. Unique if $\frac{\sigma^2_{jt}}{\sigma^2_{jt-1}}$ distinct.

• Test conditional independence: examine if eigenvectors of $\mathbf{\Sigma}_t\mathbf{\Sigma}^{-1}_{t-1}$ 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 12

$$V_1 = \frac{1}{(1-\alpha_1\beta_1)^2}\left[\begin{array}{cc} \beta_1^2\sigma_2^2+\sigma_1^2 & \beta_1\sigma_2^2+\alpha_1\sigma_1^2 \\ \beta_1\sigma_2^2+\alpha_1\sigma_1^2 & \sigma_2^2+\alpha_1^2\sigma_1^2 \end{array}\right]$$

$$V_2 = \frac{1}{(1-\alpha_2\beta_2)^2}\left[\begin{array}{cc} \beta_2^2\sigma_2^2+\sigma_1^2 & \beta_2\sigma_2^2+\alpha_2\sigma_1^2 \\ \beta_2\sigma_2^2+\alpha_2\sigma_1^2 & \sigma_2^2+\alpha_2^2\sigma_1^2 \end{array}\right]$$

• *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 13 *Simple model with Philips curve and Taylor rule*

is r_t the unobservable r_t^* or zero?

$$
\begin{aligned}
\pi_t &= c + \beta(r_t - r^n) + \psi b_t + \epsilon_{1t} && (25)\\
r_t &= \max(r_t^*, 0) && (26)\\
r_t^* &= r^n + \gamma\pi_t + \epsilon_{2t} && (27)\\
b_t &= \min(\alpha r_t^*, 0) && (28)
\end{aligned}
$$

- 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*$(0, diag(\sigma_1^2, \sigma_2^2))$.

- *(25) and (28) imply*

$$\pi_t = c + \beta(r_t - r^n) + \beta^* min(r_t^*, 0) + \epsilon_{1t} \tag{29}$$

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

zero lower bound

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

$$\pi_t = c + \beta(r_t^* - r^n) + \epsilon_{1t} \tag{30}$$

Philips curve has no kink.

- *In general (29) and (26) imply*

$$\pi_t = \tilde{c} + \tilde{\beta}(r_t^* - r^n) + v_{1t} \tag{31}$$

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*

$$\frac{1-\gamma\beta}{1-\gamma\beta^*} > 0 \tag{32}$$

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

- *Under (32) the unique solution to the system is*

enters the equation in certain regime state

$$\pi_t = \mu_1 - \tilde{\beta} D(\mu_2 + u_{2t}) + u_{1t} \tag{33}$$

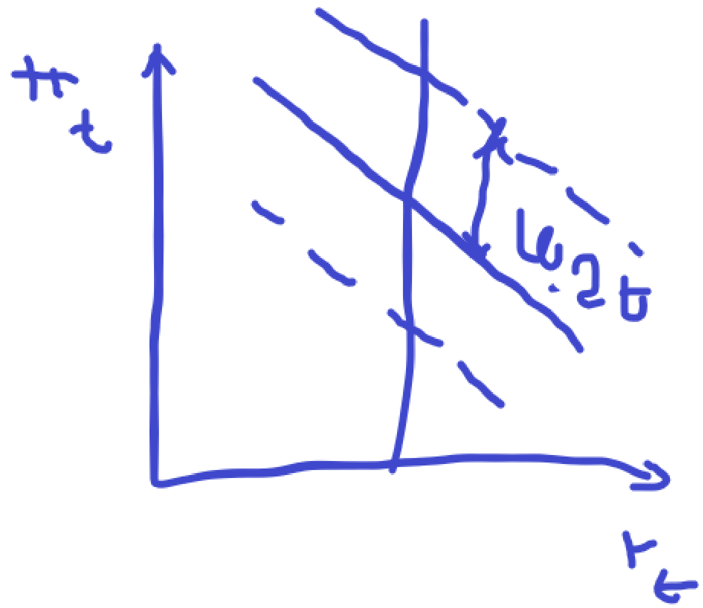
$$r_t = \max(\mu_2 + u_{2t}, 0) \tag{34}$$

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:*

$$\pi_t = \mu_1 + u_{1t} \tag{35}$$
$$r_t = \mu_2 + u_{2t} \quad \text{if } r_t \neq 0 \tag{36}$$
$$\pi_t = \mu_1 - \tilde{\beta} u_2 + u_{1t} - \tilde{\beta} u_{2t} \tag{37}$$
$$r_t = 0 \quad \text{if } r_t = 0 \tag{38}$$
$$\tag{39}$$

- *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

$$\begin{bmatrix} \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} \mu_{1t} \\ \mu_{2t} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \begin{bmatrix} \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} \mu_{1t} - \tilde{\beta}\mu_{2t} \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & -\tilde{\beta} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{40}$$

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

$$\begin{bmatrix} \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} \frac{c}{1-\gamma\beta} \\ -\frac{\gamma c}{1-\gamma\beta} \end{bmatrix} + \begin{bmatrix} \frac{1}{1-\gamma\beta} & \frac{\beta}{1-\gamma\beta} \\ \frac{\gamma}{1-\gamma\beta} & \frac{1}{1-\gamma\beta} \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \tag{41}$$

$$\begin{bmatrix} \pi_t \\ r_t \end{bmatrix} = \begin{bmatrix} \frac{(1-\gamma\tilde{\beta})c}{1-\gamma\beta} \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1-\gamma\tilde{\beta}}{1-\beta\gamma} & \frac{\beta-\tilde{\beta}}{1-\beta\gamma} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} \tag{42}$$

• Mavroedis (Econometrica, 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

$$E_t(\pi_t | r_t > 0) = c + \beta(r - R^n) + \rho(r_t - \mu_2 + \tau \frac{\phi(a)}{1 - \Phi(a)}) \tag{43}$$

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

$$r_t = \mu_2 + u_{2t} \tag{44}$$

using observations outside the ZLB. Then ρ and β can obtained from (43) 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.

effect is incremental, so it cannot be in interest rates

- 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^*.

Any financial variable is typically not normally distributed
- the fourth moment is not nornal

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

$$y_t = A + By_{t-1} + u_t$$

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

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

If you use fourth moment, you can identify all parameters in (45).

$$\begin{aligned} u_{\pi,t} &= -a_\pi^s \epsilon_t^s + a_\pi^d \epsilon_t^d \\ u_{y,t} &= a_y^s \epsilon_t^s + a_y^d \epsilon_t^d \end{aligned} \tag{45}$$

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

$$cov(u_t) = \left[\begin{array}{cc} (a_\pi^s)^2 + (a_\pi^d)^2 & -a_\pi^s a_y^s + a_\pi^d a_y^d \\ -a_\pi^s a_y^s + a_\pi^d a_y^d & (a_y^s)^2 + (a_y^d)^2 \end{array} \right] \tag{46}$$

- Four structural parameters, three reduced form covariances.

Deviation should be strong enough, otherwise identification is weak.

- Relationship between fourth moments (in excess from a normal distribution)

k=cumulants of supply and demand

using 47 48 49 and using fourth mome...

asymmetric cumulants

$$E(u_{\pi,t}^4) - 3 = \frac{(a_\pi^s)^4 * k_s + (a_\pi^d)^4 * k_d}{var^2(e_\pi)} \tag{47}$$

$$E(u_{y,t}^4) - 3 = \frac{(a_y^s)^4 * k_s + (a_y^d)^4 * k_d}{var^2(e_y)} \tag{48}$$

$$E(u_{\pi,t}^2 e_{y,t}^2) - 3 = \frac{(a_\pi^s)^2 (a_y^s)^2 * k_s + (a_\pi^d)^2 (a_y^s)^2 * k_d}{var(e_\pi) var(e_y)} \tag{49}$$

$$E(u_{\pi,t}^3 e_{y,t}) - 3 = \frac{-(a_\pi^s)^3 a_y^s * k_s + (a_\pi^d)^3 a_y^d * k_d}{var^{3/2}(e_\pi) var^{1/2}(e_y)} \tag{50}$$

$$E(u_{\pi,t} e_{y,t}^3) - 3 = \frac{-(a_\pi^s)(a_y^s)^3 * k_s + (a_\pi^d)(a_y^d)^3 * k_d}{var^{1/2}(e_\pi) var^{3/2}(e_y)} \tag{51}$$

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

• (47)-(48) univariate fourth moments; (49) is a symmetric cross moment, (50)-(51) asymmetric cross moments.

• (47)-(48) add two moments and two parameters (k^s, k^d).

• (49)-(51) 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 $cov(u_t)$ and k^s, k^d).

• Fourth moments source of overidentification.

• (49) 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.

$$E(u_{\pi,t}^4) - 3 = \frac{cov(u_\pi^2, u_y^2)}{var(u_\pi) - var(u_y)} - \rho^2(u_\pi, u_y) \tag{52}$$

• To see how (50)-(51) 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 $H = [\sigma_\pi, \sigma_y, \rho(\pi, y), E(u_{\pi,t})^4 - 3, E(u_{y,t}^4) - 3, E(u_{\pi,t}^2 u_{y,t}^2) - 3, E(u_{\pi,t}^3 u_{y,t}) - 3, E(u_{\pi,t} u_{y,t}^3) - 3]$. These quantities can be estimated from the data conditional on the VAR parameters.

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

- 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.

• Here 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 for the shocks to mutually independent.

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

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 prior views of the client into classical VAR.

• BVARs are a flexible way to incorporate extraneous (client) information; helps 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.

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

Letting $B = [B_1, \ldots B_q]; X_t = [y_{t-1}, \ldots y_{t-p}], \beta = vec(B)$, the VAR is:

$$y \;=\; (I_M \otimes X)\beta + e \quad e \sim (0, \Sigma_e \otimes I_T) \tag{53}$$

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 the likelihood function can be written as:

$$L(\beta, \mathbf{\Sigma}_e | y, y_p) \propto N(\beta | \hat{\beta}, \mathbf{\Sigma}_e, X, y, y_p) \times iW(\mathbf{\Sigma}_e | \hat{\beta}, X, y, y_p, T - \nu) \quad (54)$$

where tr = trace of the matrix, $\hat{\beta} = (\mathbf{\Sigma}_e^{-1} \otimes X'X)^{-1}(\mathbf{\Sigma}_e^{-1} \otimes X)y$, 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 $\mathbf{\Sigma}_e$, and an inverted Wishart distribution for $\mathbf{\Sigma}_e$, conditional on $\hat{\beta}$, with scale $(y - (x \otimes \mathbf{\Sigma}_e)\hat{\beta})'(y - (x \otimes \mathbf{\Sigma}_e)\hat{\beta})$ 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

$$
\begin{aligned}
g(\beta|\Sigma_e, Y) &= N(\beta|\hat{\beta}, \Sigma_e, X, y, y_p) && (55)\\
g(\Sigma_e|Y) &= iW(\Sigma_e|\hat{\beta}, X, y, y_p, T-k) && (56)
\end{aligned}
$$

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:

$$
\begin{aligned}
g(\beta) &\propto |\Sigma_b|^{-0.5} exp[-0.5(\beta - \bar{\beta})' \Sigma_b^{-1} (\beta - \bar{\beta})] \\
&= |\Sigma_b|^{-0.5} exp[-0.5(\Sigma_b^{-0.5}(\beta - \bar{\beta}))' \Sigma_b^{-0.5} (\beta - \bar{\beta})] \qquad (57)
\end{aligned}
$$

- Posterior:

$$
g(\beta|y) \propto exp[-0.5(\beta - \tilde{\beta})' \tilde{\Sigma}_b^{-1} (\beta - \tilde{\beta})] \qquad (58)
$$

$$
\begin{aligned}
\tilde{\beta} &= [\Sigma_b^{-1} + (\Sigma_e^{-1} \otimes X'X)]^{-1} [\Sigma_b^{-1} \bar{\beta} + (\Sigma_e^{-1} \otimes X)' y] \qquad (59) \\
\tilde{\Sigma}_b &= [\Sigma_b^{-1} + (\Sigma_e^{-1} \otimes X'X)]^{-1} \qquad (60)
\end{aligned}
$$

- $g(\beta|y)$ 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

$$
\begin{aligned}
g(\beta|y) &\propto exp[-0.5(\beta-\tilde{\beta})' Z' Z(\beta-\tilde{\beta})] \\
\tilde{\beta} &= (Z'Z)^{-1}(Z'z) \\
Z &\equiv [\mathbf{\Sigma}_b^{-0.5}, (\mathbf{\Sigma}_e^{-0.5} \otimes X)]'
\end{aligned} \tag{61}
$$

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

$$
\begin{aligned}
y_t &= x_t B + e_t \qquad e_t \sim (0, \sigma^2) \\
\bar{B} &= B - \epsilon \quad \epsilon \sim (0, \mathbf{\Sigma}_b)
\end{aligned} \tag{62}
$$

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 $E_t \sim (0, \mathbf{\Sigma}_E)$, $t = 1, \ldots, T$ 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 $y_0 = \sigma^2 W^{-1}\bar{B}, x_0 = \sigma^2 W^{-1}$, $e_0 = \sigma^2 W^{-1}\epsilon$, $WW' = \Sigma_b$.

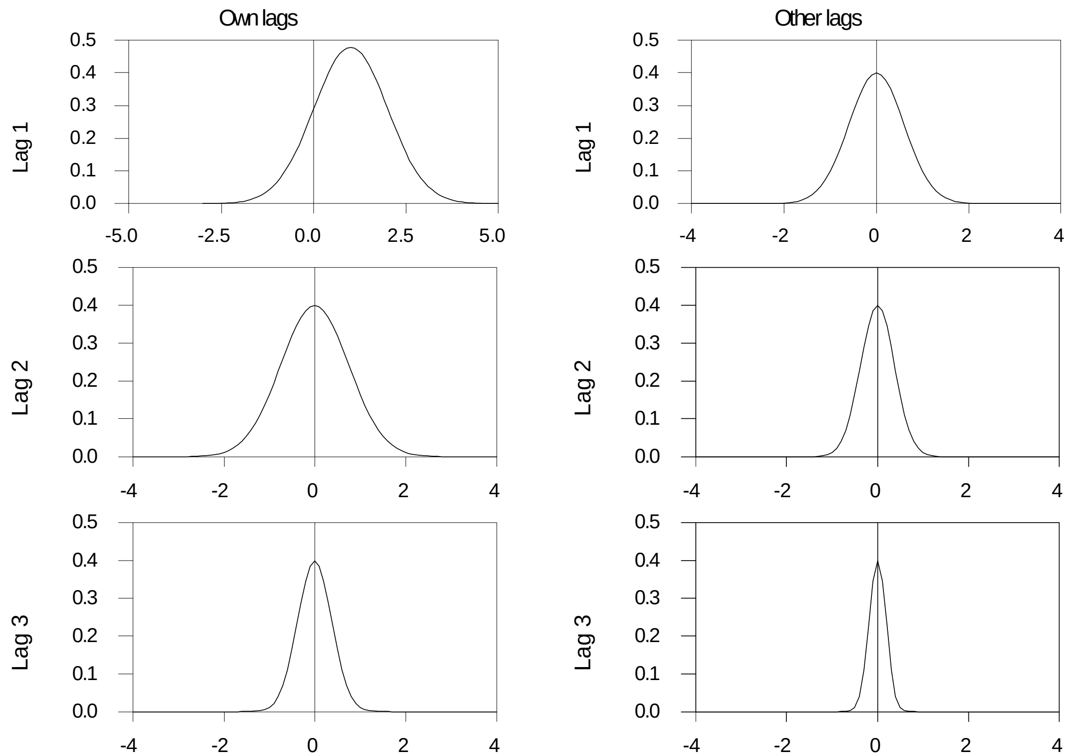
Special case: Litterman (Minnesota) setup

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

$$\begin{aligned}
\sigma_{ij,\ell} &= \frac{\phi_1}{h(\ell)} \text{ if } i = j && (63)\\
&= \phi_1 \frac{\phi_2}{h(\ell)} * (\frac{\sigma_j}{\sigma_i})^2 \text{ otherwise} && (64)\\
&= \phi_1 * \phi_4 * (\frac{\sigma_j}{\sigma_i})^2 \text{ for exogenous variables} && (65)
\end{aligned}$$

• $\phi_1 =$ 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
0.5
0.4
0.3
0.2
0.1
0.0
-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).

- $\mathbf{\Sigma}_b$ very big ($M \times M$). Decrease dimensionality by setting $\mathbf{\Sigma}_b = \mathbf{\Sigma}_b(\phi)$.

- $\mathbf{\Sigma}_b$ 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 14 *Bivariate VAR(2) with $h(\ell) = \ell$; 4 coefficients each equation.*

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

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

• 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 15 *(ML-II approach)*

$$\begin{aligned} y_t &= Bx_t + u_t \quad u_t \sim N(0, \sigma_u^2) \\ B &= \bar{B} + \upsilon \quad \upsilon \sim N(0, \sigma_\upsilon^2 \end{aligned} \tag{66}$$

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

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

$$\grave{g}(\beta, \phi|y) = \frac{1}{(2\pi)^{0.5}\sigma_u\sigma_\upsilon} \exp\{-0.5\frac{(y - Bx)^2}{\sigma_u^2} - 0.5\frac{(B - \bar{B})^2}{\sigma_\upsilon^2}\} \tag{67}$$

$y = [y_1, \ldots y_t]', x = [x_1, \ldots x_t]'$*. Integrating* B *out of (67), maximize:*

$$\tilde{g}(\phi|y) = \frac{1}{(2\pi q(\phi)^2 tr|X'X| + \sigma_u^2)^{0.5}} \exp\{-0.5\frac{(y - \bar{B}x)^2}{\sigma_u^2 + q(\phi)^2 tr|X'X|}\} \tag{68}$$

- Recent 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 is fixed.

- Draw β from the normal posterior, keeping Σ 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 ae 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:

$$\sum_i \beta_{ij} \sim N(\mu_{1j}, \sigma_{\mu_{1j}}) \; j = 1, 2, \ldots, m \tag{69}$$

ii) Cointegration restriction:

$$\sum_i \beta_{ij} - \sum_i \beta_{ih} \sim N(\mu_2, \sigma_{\mu_2}) \tag{70}$$

iii) Long run prior restrictions: Giannone et al. 2018.

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

$$
\begin{aligned}
y_t &= (I \otimes X_t)\beta + e \quad e \sim N(0, \Sigma) && (71)\\
\beta &= M_0\theta + v \quad v \sim N(0, D_0) && (72)\\
\theta &= M_1\mu + \zeta \quad \zeta \sim N(0, D_1) && (73)
\end{aligned}
$$

• 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)$(\mu|\psi_{-\mu}, Y, \mathcal{X}) \sim N(\hat{\mu}, \boldsymbol{\Sigma}_\mu)$

where

$\tilde{\Omega} = (D_0^{-1} + \sum_t X_t' \boldsymbol{\Sigma}^{-1} X_t)^{-1}$;
$\tilde{\beta} = \tilde{\Omega}(D_0^{-1} M_0 \theta + \sum_t X_t' \boldsymbol{\Sigma}^{-1} y_t)$;
$\tilde{\boldsymbol{\Sigma}}^{-1} = \bar{S} + \sum_t (Y_t - X_t \beta)(y_t - X_t \beta)'$;
$\tilde{D}_1 = (D_1^{-1} + M_0' D_0^{-1} M_0)^{-1}$;
$\tilde{D}_0^{-1} = D_0^{-1} + \sum_{g=1}^{M} (\beta_g - \theta)(\beta_g - \theta)'$
$\hat{\mu} = (M_1' M_1)^{-1} (M_1 \theta)$
$\boldsymbol{\Sigma}_\mu = (\theta - \hat{\mu} M_1)'(\theta - \hat{\mu} M_1)$

Use these conditional posteriors in the Gibbs sampler.

• Another useful hierarchical VAR:

$$y_t = (I \otimes X)\beta + e \quad e \sim N(0, \Sigma) \tag{74}$$
$$\beta = \bar{\beta} + v \quad v \sim N(0, \Sigma \otimes \Omega * \zeta) \tag{75}$$
$$\zeta = \bar{\zeta} + \epsilon \quad \epsilon \sim N(0, \eta) \tag{76}$$

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?

$$
\begin{aligned}
y_t &= A(L)y_{t-1} + e_t \quad e_t \sim N(0, \Sigma_e) && (77)\\
&= D(L)e_t && (78)
\end{aligned}
$$

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

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 what it is done in micro when calculating the "causal effects of an intervention ".

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

- How are impulse responses computed?
- For each draw of VAR coefficients and covariance matrix Σ_u:

1) Transform the VAR into a companion form.

2) Solve backward the companion form.

3) Orthogonalize the shocks.

4) Store results.

Companion form

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

Example 16 *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

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{E}_t \quad \mathbf{E}_t \sim \mathbf{N}(0, \mathbf{\Sigma}_E) \tag{79}$$

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

Backward solution and orthogonalization

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{E}_t \\ &= \mathbf{A}^t\mathbf{Y}_0 + \sum_{j=0}^{t-1} \mathbf{A}^j \mathbf{E}_{t-j} && (80) \\ &= \mathbf{A}^t\mathbf{Y}_0 + \sum_{j=1}^{t-1} \tilde{\mathbf{A}}^j \tilde{\mathbf{E}}_{t-j} && (81) \end{aligned}$$

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 at iteration $l = 1, 2, \ldots, L$. Compute $\tilde{\mathbf{A}_l}^j$ for each horizon j. Store the results.

- Report the mean (median) of $\tilde{\mathbf{A}_l}^j$ 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:

$$\hat{y}_t(\tau) \equiv y_{t+\tau} - y_t(\tau) = \sum_{j=0}^{\tau-1} \tilde{D}_j \tilde{e}_{t+\tau-j} \quad D_0 = I \tag{82}$$

$y_t(\tau)$ 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

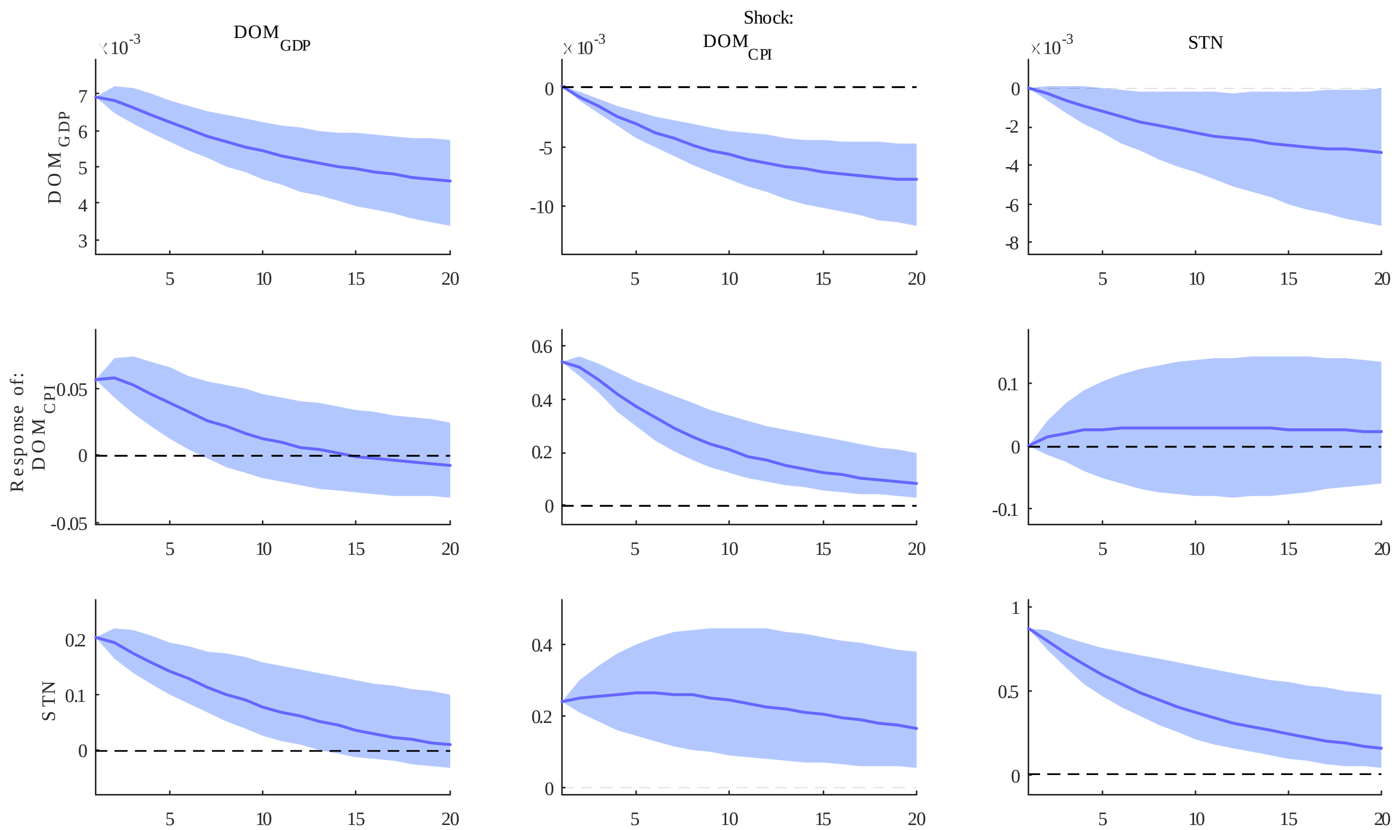
• What is the contribution of supply shocks to the productivity revival of the late 1990s?

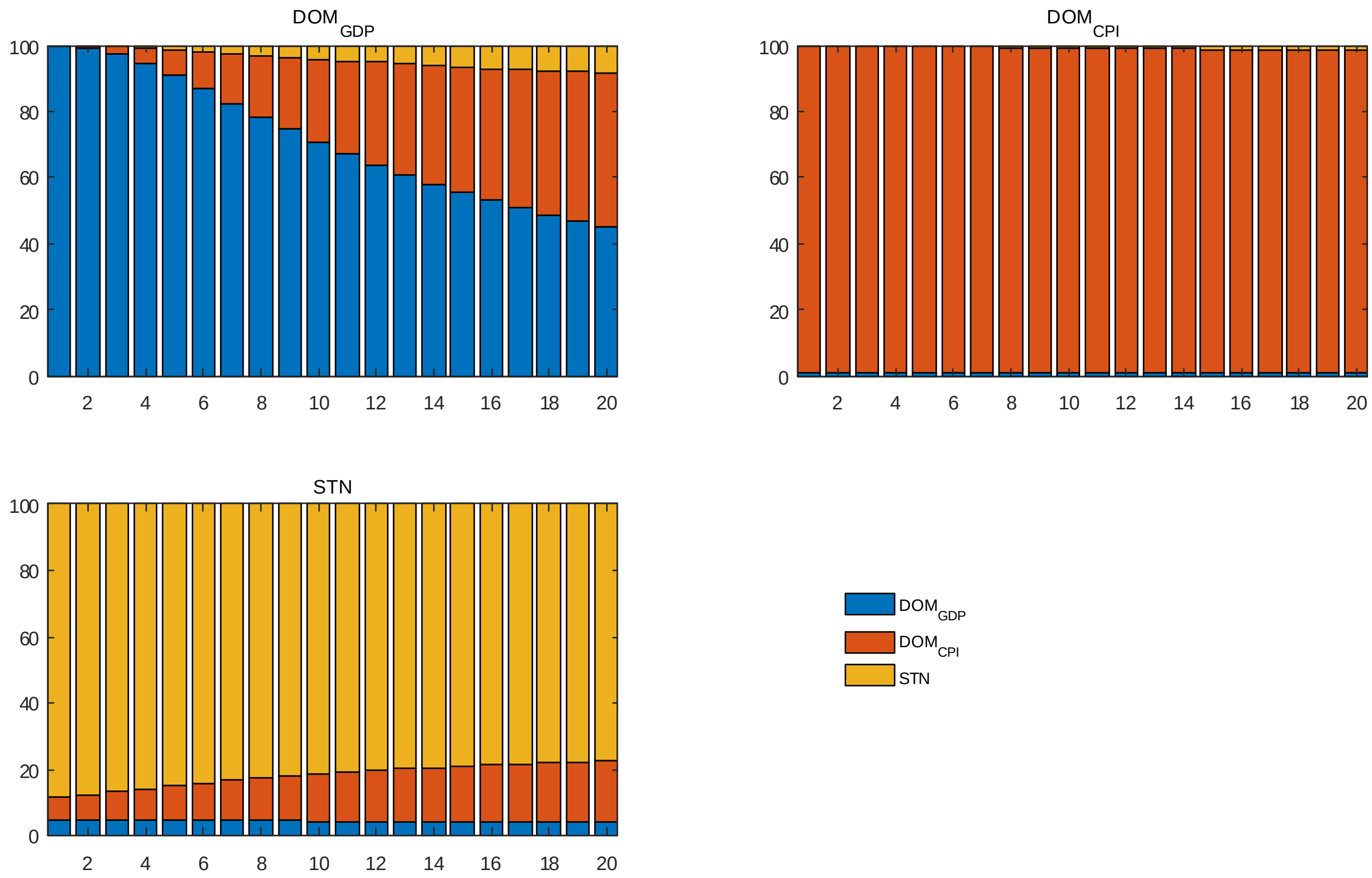
• Let $\hat{y}_{i,t}(\tau) = y_{i,t+\tau} - y_{i,t}(\tau)$ be the τ-steps ahead forecast error in the i-th variable of the VAR. Then:

$$\hat{y}_{i,t}(\tau) = \sum_{i'=1}^{m} \tilde{D}^{i'}(\ell)\tilde{e}_{i't+\tau} \tag{83}$$

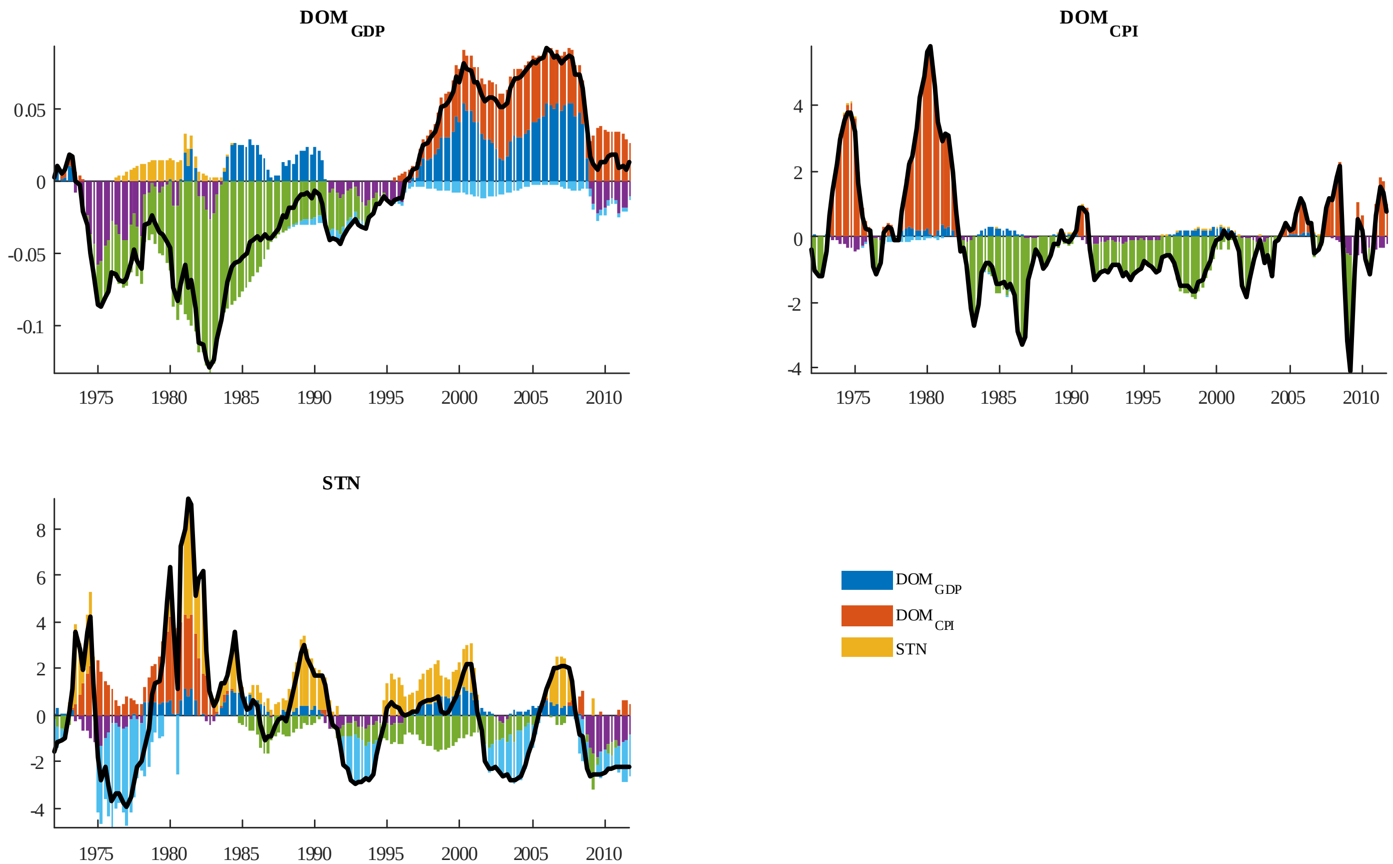
• Computes the path of $\hat{y}_{i,t}(\tau)$ due to each $\tilde{e}_{i'}$.

• 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
× 10^-3
× 10^-3



DOM_{GDP}
DOM_{CPI}
STN
0
20
40
60
80
100
2
4
6
8
10
12
14
16
18
20



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

8.4 Impulse Responses in a VARX model

$$y_t = A(\ell)y_{t-1} + B(\ell)x_t + e_t \tag{84}$$

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

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{R}x_t + \mathbf{E}_t \tag{85}$$

where

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

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

• Let $J = [I_m, 0, 0 \ldots, 0]$. Repeatedly substituting into (85) we have

$$\mathbf{Y}_t = \sum_{i=0}^{\infty} J\mathbf{A}^i\mathbf{R}x_{t-i} + \sum_{i=0}^{\infty} J\mathbf{A}^i\mathbf{E}_{t-i} \tag{86}$$

• 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.

Problems with Bayesian sign restrictions

Baumeister and Hamilton (2015) make two points about Bayesian inference in sign restricted SVARs:

- 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 prior on the set of rotations (Haar prior) 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 impulse responses directly.

9 Forecasting with BVAR (SBVAR) models

$$\begin{aligned} y_t &= A(\ell)y_{t-1} + e_t && (87) \\ \mathcal{A}_0 y_t &= \mathcal{A}(\ell)y_{t-1} + \epsilon_t && (88) \end{aligned}$$

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

- Transform (87) and(88) into a companion form.

$$\begin{aligned} Y_t &= \mathbf{A}Y_{t-1} + E_t && (89) \\ A_0 Y_t &= AY_{t-1} + \Upsilon_t && (90) \end{aligned}$$

- 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 $Y_{t+\tau}^l, \quad l = 1, 2, \ldots, L$, 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 (89)- (90).

• 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

$$x_{1t} = B_{11}(L)x_{1t-1} + B_{12}(L)x_{2t-1} + A_{11}u_{1t} + A_{12}u_{2t} \tag{91}$$
$$x_{2t} = B_{21}(L)x_{1t-1} + B_{22}(L)x_{2t-1} + A_{12}u_{1t} + A_{22}u_{2t} \tag{92}$$

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 (92) 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 BVAR

• 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 must be correlated (Σ full matrix).

• Alternative: Carriero et al. (2017): 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:

$$
\begin{aligned}
y_{1t} &= \beta_{10} + \sum_i \sum_l \beta_{1,i,l} y_{i,t-l} + h_{1t}^{-0.5} \epsilon_{1t} \\
y_{2t} &= \beta_{20} + \sum_i \sum_l \beta_{2,i,l} y_{i,t-l} + a_{21} h_{1t}^{-0.5} \epsilon_{1t} + h_{2t}^{-0.5} \epsilon_{2t} \qquad (93)
\end{aligned}
$$

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:

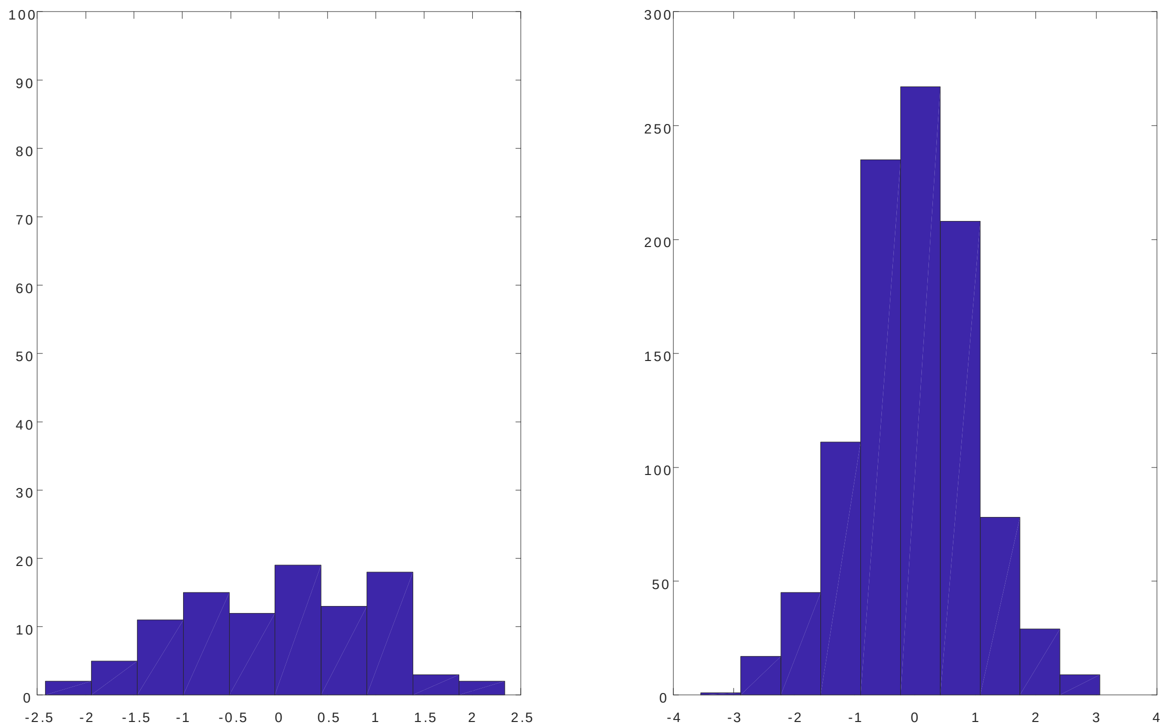
$$\begin{aligned} P(\alpha^{l-1},\ \alpha^{l}) &= q(\alpha^{l-1}, \alpha^{l})\mathfrak{E}(\alpha^{l-1}, \alpha^{l}) \qquad \text{if } \alpha^{l} \neq \alpha^{l-1} \\ &= 1 - \int_{A} q(\alpha^{l-1}, \alpha)\mathfrak{E}(\alpha^{l-1}, \alpha) d\alpha \ \text{ if } \alpha^{l} = \alpha^{l-1} \end{aligned} \tag{94}$$

• $P(\alpha^{l-1},\ \alpha^{l})$ 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) $A_{m\times n}\otimes B_{p\ \times q}=\left(\begin{array}{llll} a_{11}B & a_{12}B & \ldots & A_{1,n}B \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1}B & a_{m2}B & \ldots & A_{m,n}B \end{array}\right)$

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

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

$$p(x) = (2\pi)^{-0.5M}|\Sigma|^{-0.5}exp\{0.5(x-\mu)'^{-1}(x-\mu)\} \tag{95}$$

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

$$p(x) = \frac{\Gamma(0.5(\nu+M))}{\Gamma(0.5\nu)(\nu\pi)^{0.5}\sigma}|\Sigma|^{-0.5}exp\{1+\frac{0.5}{\nu}(x-\mu)'^{-1}(x-\mu)\}^{-0.5(\nu+M)} \tag{96}$$

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

$$\begin{aligned} p(A) \;\; = \;\; & (2^{0.5\nu M}\pi^{0.25M(M-1)}\prod_i^M \Gamma(0.5(\nu+1-i))^{-1}|S|^{0.5\nu} \\ & |A|^{-0.5(\nu+M+1)}exp(-0.5tr(SA^{-1})) \end{aligned} \tag{97}$$