Lecture 7: Bayesian Analysis: Intro and BVARs

Dr. Joao B. Duarte[1]

[1]Nova School of Business and Economics
University of Cambridge

Masters, Economics: Macroeconometrics
Lisbon
Spring 2017

Lecture Objectives:

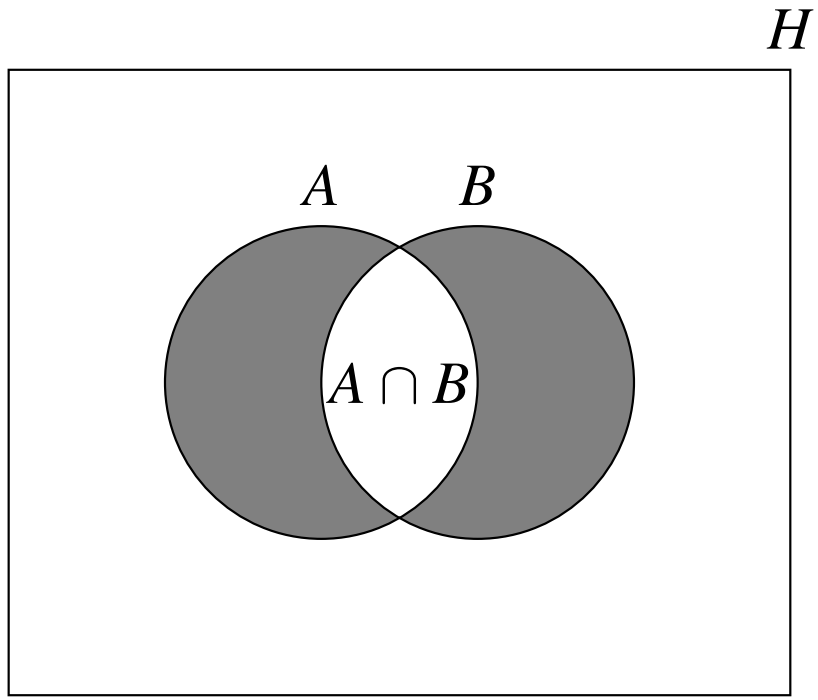
- Introduction to Bayesian analyzes.
- Bayesian vs. Frequentist perspective.
- Credible intervals vs. confidence intervals.
- Introduction to BVAR and its commonly used prior distributions.
- BVAR estimation and properties.
- Sign restrictions identification.

Secondary Readings:

- Chapter 9, Canova, Fabio
- Chapter 12, Time Series Analysis, Hamilton, James, first edition

Preliminaries

Let A and B be two events.



- **Conditional Probability**: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Preliminaries

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Proof:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$P(B \cap A) = P(B|A)P(A)$$

However, $P(B \cap A) = P(A \cap B)$. Hence, we have:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Preliminaries

- Frequentists agree with the Bayes theorem and use it.
- Where they differ from Bayesians is in the situations in which they use it.
- For Bayesians, one can treat A as model parameters and B as data.
- For Frequentists, that is unacceptable since there cannot be a probability statement about the model parameters because there exists a TRUE model.

Intro to Bayesian Analysis

- AR(1) example:

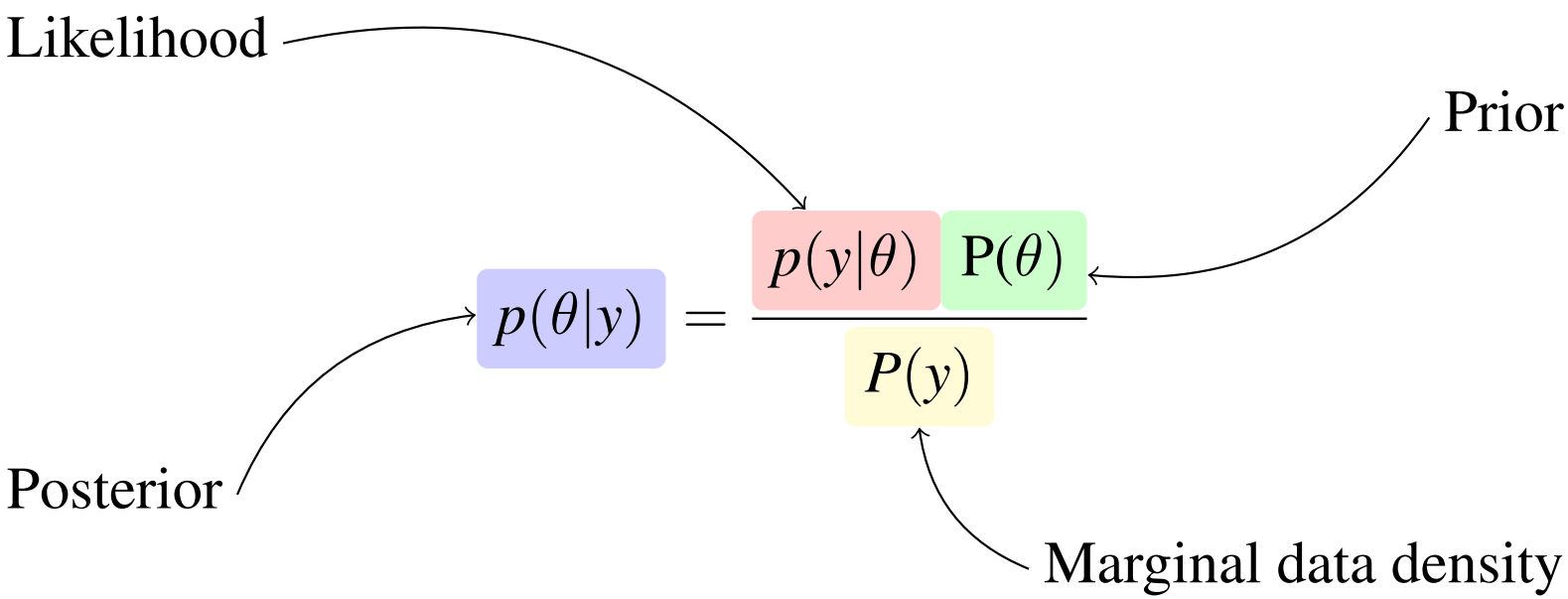
$$y_t = \rho y_{t-1} + \varepsilon_t$$

with $\varepsilon_t \sim N(0, \sigma^2)$

- If we assume that both the data y and the parameters ρ, σ are random, then we can treat:
  - $A = \theta = \{\rho, \sigma\}$
  - $B = y$

Intro to Bayesian Analysis

- Applying Bayes Theorem gives us:



Frequentists vs Bayesians

- Classical, or frequentist view: there is one model, and one tries to make inference about it, namely, deducing the probability that the model is true or false
- Bayesian view: there are many models, over which one forms a prior probability distribution, and uses the data to form a posterior

Frequentists vs Bayesians

- Likelihood:
  - Frequentist: likelihood is something to be maximised in knowledge that as data sample increases this maximum ($\hat{\theta}$) would approach the ONE TRUE THETA.
  - Bayesian: this is a distribution function, expressing the data's perspective on the probability mass on all possible thetas, with stance that there is NO SINGLE TRUE THETA.

Frequentists vs Bayesians

- Frequentist inference makes only pre-sample probability assertions.
  - A 95% confidence interval contains the true parameter value with probability .95 only before one has seen the data. After the data has been seen, the probability is zero or one.
  - Yet confidence intervals are universally interpreted in practice as guides to post-sample uncertainty.
  - They often are reasonable guides, but only because they often are close to posterior probability intervals that would emerge from a Bayesian analysis.
- People want guides to uncertainty as an aid to decision-making. They want to characterize uncertainty about parameter values, given the sample that has actually been observed. That it aims to help with this is the distinguishing characteristic of Bayesian inference.

Confidence Intervals vs Credible Sets

- Frequentists: There is a true model but the sample is random. If we repeat samples and construct confidence intervals, with what frequency would they include the true model? That is what the confidence interval gives us.
- Bayesians: Both data and model are random. Given the data we observe, with what probability would a model be true? This is the credible set, that can be easily computed from the posterior.

Bayesian Analysis of Different Models

- Bayesians naturally consider model uncertainty as there is no true model.
- We just need reinterpret θ as models.
- Example, consider an AR(1) that we will call M_1 and an AR(2), M_2:

$$\begin{aligned} \text{Model 1: } y_t &= \rho_1 y_{t-1} + \varepsilon_t \\ \text{Model 2: } y_t &= \rho_2 y_{t-1} + \gamma\rho y_{t-2} + \varepsilon_t \end{aligned}$$

- The parameters in each one are $\theta_1 = \{\rho_1, \sigma_1\}$ and $\theta_2 = \{\rho_2, \gamma, \sigma_2\}$

Bayesian Analysis of Different Models

- Now we need priors over the models, and also over the parameters given a specific model:

$$p(M) \text{ Prior over models}$$

$$p(\theta_1|M_1) \text{ Prior over } \theta_1 \text{ given } M_1$$

$$p(\theta_2|M_2) \text{ Prior over } \theta_2 \text{ given } M_2$$

$$\boxed{p(M|y) = \frac{p(y|M)p(M)}{p(y)}}$$

- This gives you the probability that a specific model is true given the data.

Linear Regression Example With Known Variance

- Take the linear regression example:

$$y = X\beta + \varepsilon$$

- Assuming normality we have that the OLS estimator:

$$\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$$

- That is given β, σ^2_{ols} we have that:

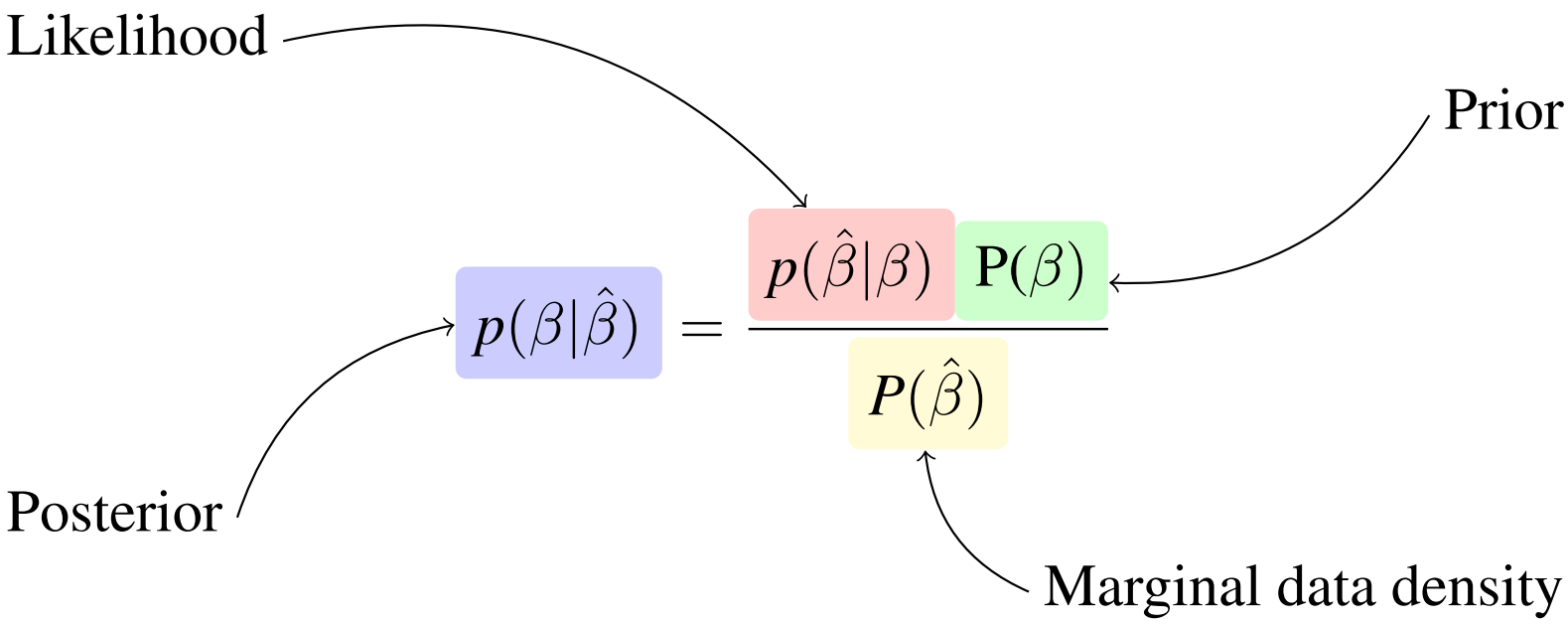
$$p(\hat{\beta}|\beta, \sigma^2_{ols}) = \frac{1}{\sqrt{2\sigma^2_{ols}\pi}} \exp\left(-\frac{(\hat{\beta} - \beta)^2}{2\sigma^2_{ols}}\right)$$

Linear Regression Example With Known Variance

- If we have a normal prior over $\beta \sim N(\beta_0, \sigma_0)$:

$$p(\beta) = \frac{1}{\sqrt{2\sigma_0^2\pi}} \exp\left(-\frac{(\beta - \beta_0)^2}{2\sigma_0^2}\right)$$

- We can find the posterior of β using Bayes theorem:



Linear Regression Example With Known Variance

- Note that the marginal data density is just a normalization factor to make sure the probability is well defined ($0 \leq p \leq 1$). Hence,

$$p(\beta|\hat{\beta}) \propto p(\hat{\beta}|\beta)p(\beta)$$

- After multiplying the likelihood by the prior we get that the posterior is given by:

$$p(\beta|\hat{\beta}) \propto \frac{1}{\sqrt{\frac{\sigma_{ols}^2\sigma_0^2}{\sigma_{ols}^2+\sigma_0^2}}} \exp\left(-\frac{(\beta-(w_1\hat{\beta}+w_2\beta_0))^2}{2\frac{\sigma_{ols}^2\sigma_0^2}{\sigma_{ols}^2+\sigma_0^2}}\right)$$

Where

$$\begin{aligned} w_1 &= \frac{\sigma_0^2}{\sigma_{ols}^2+\sigma_0^2} \\ w_2 &= \frac{\sigma_{ols}^2}{\sigma_{ols}^2+\sigma_0^2} \end{aligned}$$

Linear Regression Example With Known Variance

- Many important messages come from the posterior:

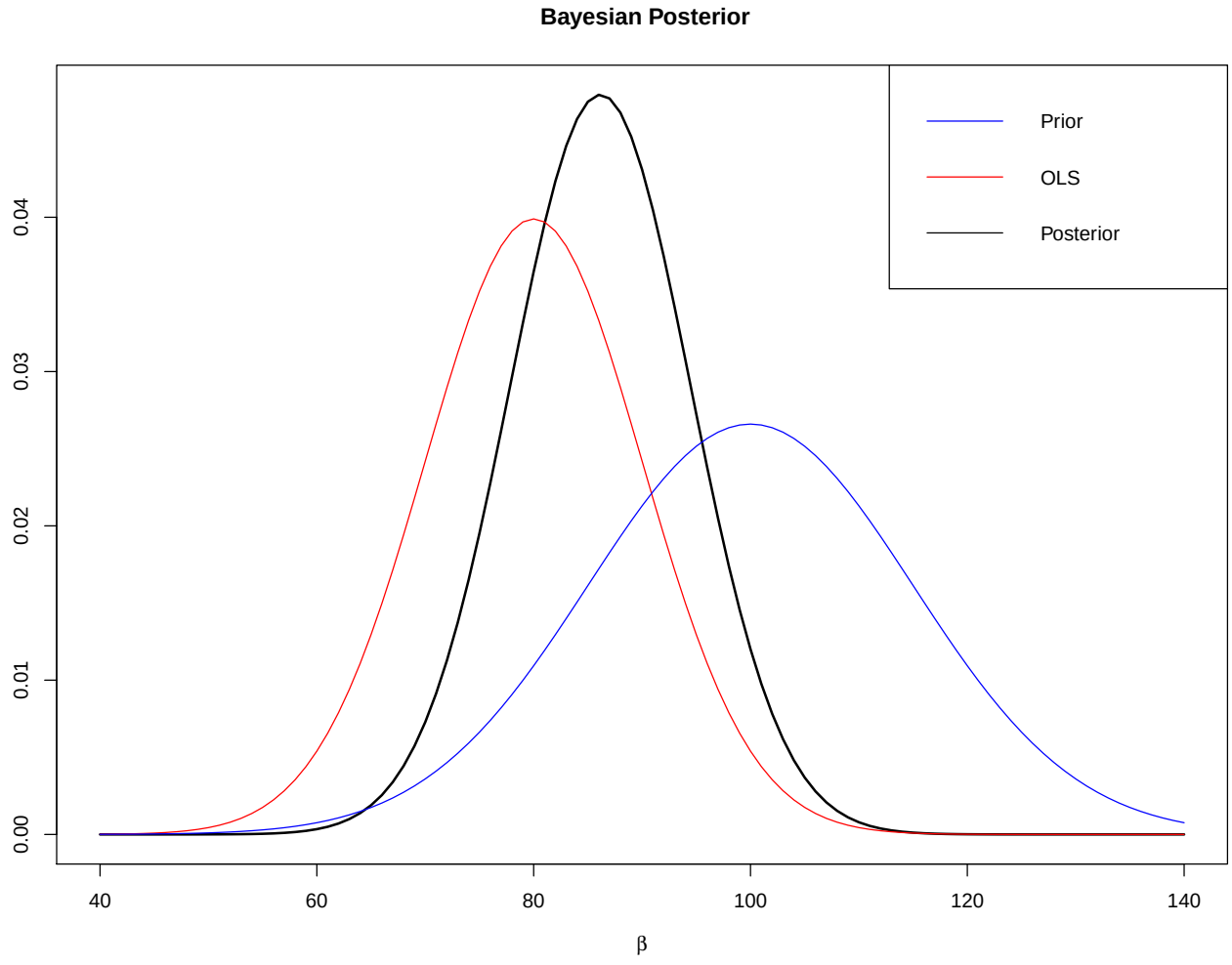
1. The posterior is itself normal! When the posterior has the same distribution of the prior we call them conjugate distributions. In this case, the prior is called a conjugate prior
2. The posterior follows $N(w_1\hat{\beta} + w_2\beta_0, \frac{\sigma^2_{ols}\sigma^2_0}{\sigma^2_{ols}+\sigma^2_0})$. The posterior mean is then:

$$\beta_{bayes} = w_1\hat{\beta} + w_2\beta_0$$

3. $w_1 + w_2 = 1$. Hence we can think of them as weights. The higher the variance of the OLS relative to the prior, the higher is going to be the weight given to the prior has we have more confidence in the prior.
4. As the sample size increases to infinity, the variance of the OLS estimate goes to zero $\Rightarrow w_1 \to 1$.

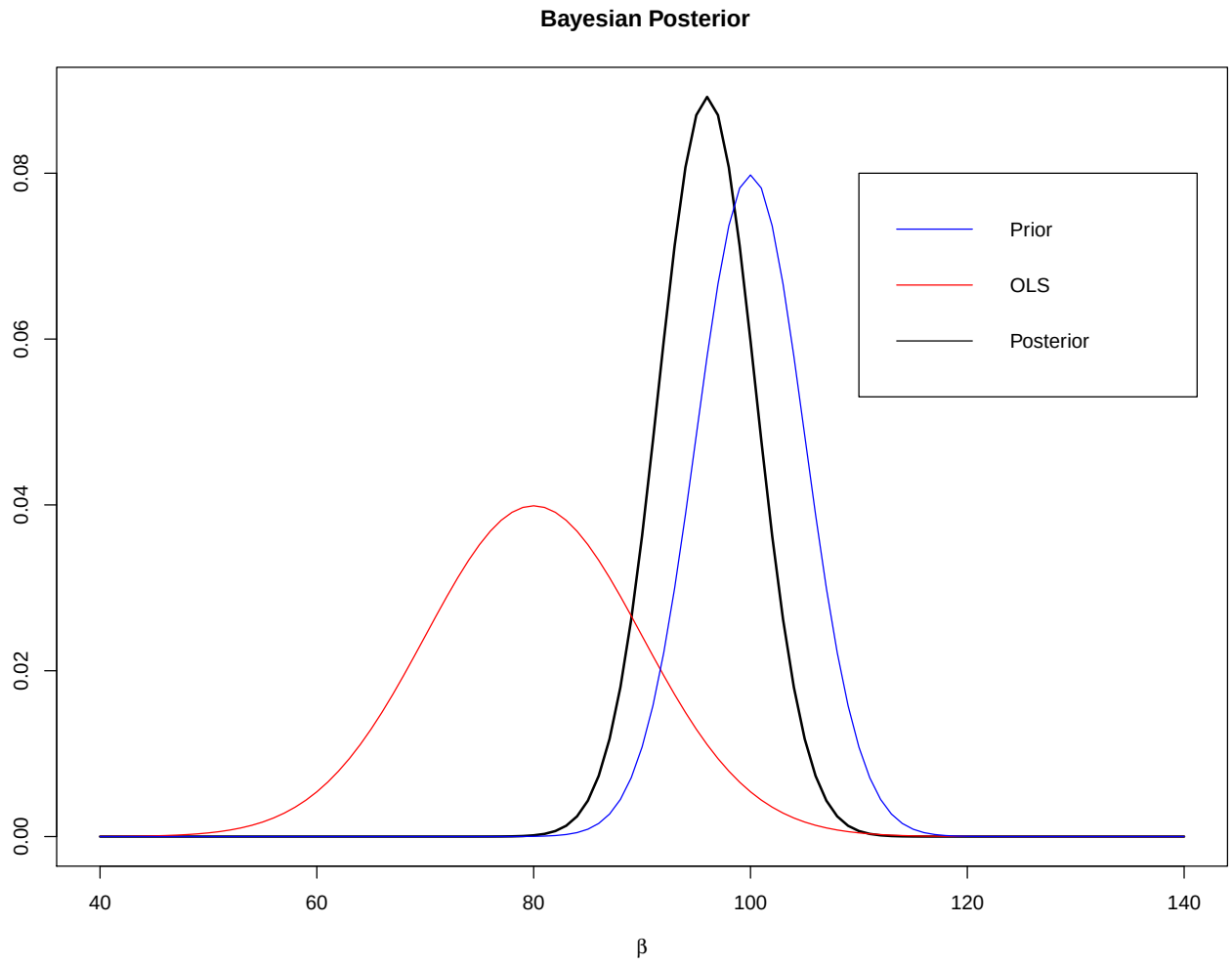
Linear Regression Example With Known Variance

- Examples: Here we can see that the variance of the estimate decreases and the posterior falls somewhere in between the prior and the OLS estimate.



Linear Regression Example With Known Variance

- Examples: If we have a more accurate prior notice how the OLS estimate is not taken into account:



Pragmatic motivation for estimating Bayesian VARs

- Large dimension to avoid omitted variables
- Curse of dimensionality: parameters increase with n^2. Quickly become ill-determined.
- Forecasting performance poor.
- Possibility of data driving you to misleading local maximum of the likelihood.
- Prior 'shrinkage' [shrinkage of probability mass around some mode, for example] alleviates these difficulties.

Bayesian VARs

- Bayesian methods can also be applied to VARs. The idea is again to impose a prior on the parameters.
- As you are probably by now more aware, the key issue is the choice of a prior.
- Also, a key assumption concerns the variance-covariance matrix. If it is known, we just need a prior for the reduced form estimates of the autoregressive components. If is not known, then we also need to impose a prior on the vcov matrix.

Estimation of Bayesian VARs Journey

1. Posterior for VAR parameters when we assume the vcov matrix is known. Use conjugate prior for easier computation of posterior.
2. Allow vcov itself to have a distribution. Notion of conjugacy in priors for this that guarantee known distribution for posterior that we can draw from.
3. Gibbs sampling when conjugacy is not possible or desirable.
4. Metropolis-Hastings, when you can't factor to leave distributions from which you can draw.
5. Particle filtering. When one cannot even evaluate the likelihood for a candidate parameter value.

Estimation of Bayesian VARs Journey

- In this lecture, we will only present the natural conjugate idea. Assuming, the prior on the VAR reduced form parameters is normal, the posterior will also be normal assuming a known vcov.
- That was the idea behind the first Bayesian VARs. The problem is that vcov is typically not know. The minnesota prior suggests ways of replacing the unknown vcov with an estimated one.

Minnesota Prior

- When Σ is replaced by an estimate, we only have to worry about a prior for the reduced-form parameters A_1. The Minnesota prior assumes:

$$A_1 \sim N(A_1^{Mn}, V^{Mn})$$

- The Minnesota prior can be thought of as a way of automatically choosing A_1^{Mn} and V^{Mn} n in a manner which is sensible in many empirical contexts.
- A big advantage of the Minnesota prior is that it leads to simple posterior inference involving only the Normal distribution.
- The disadvantage is the treatment of σ which is replaced by some estimate.

Minnesota Prior

- The prior is over the reduced-form parameters. After estimation, one can impose the identification assumptions of SVARs.
- It is possible to assume priors directly over the structural parameters, but we will not discuss it here.
- We will instead use the Bayesian knowledge to discuss the Uhlig (2005) sign restriction identification strategy.

Sign Restrictions Identification

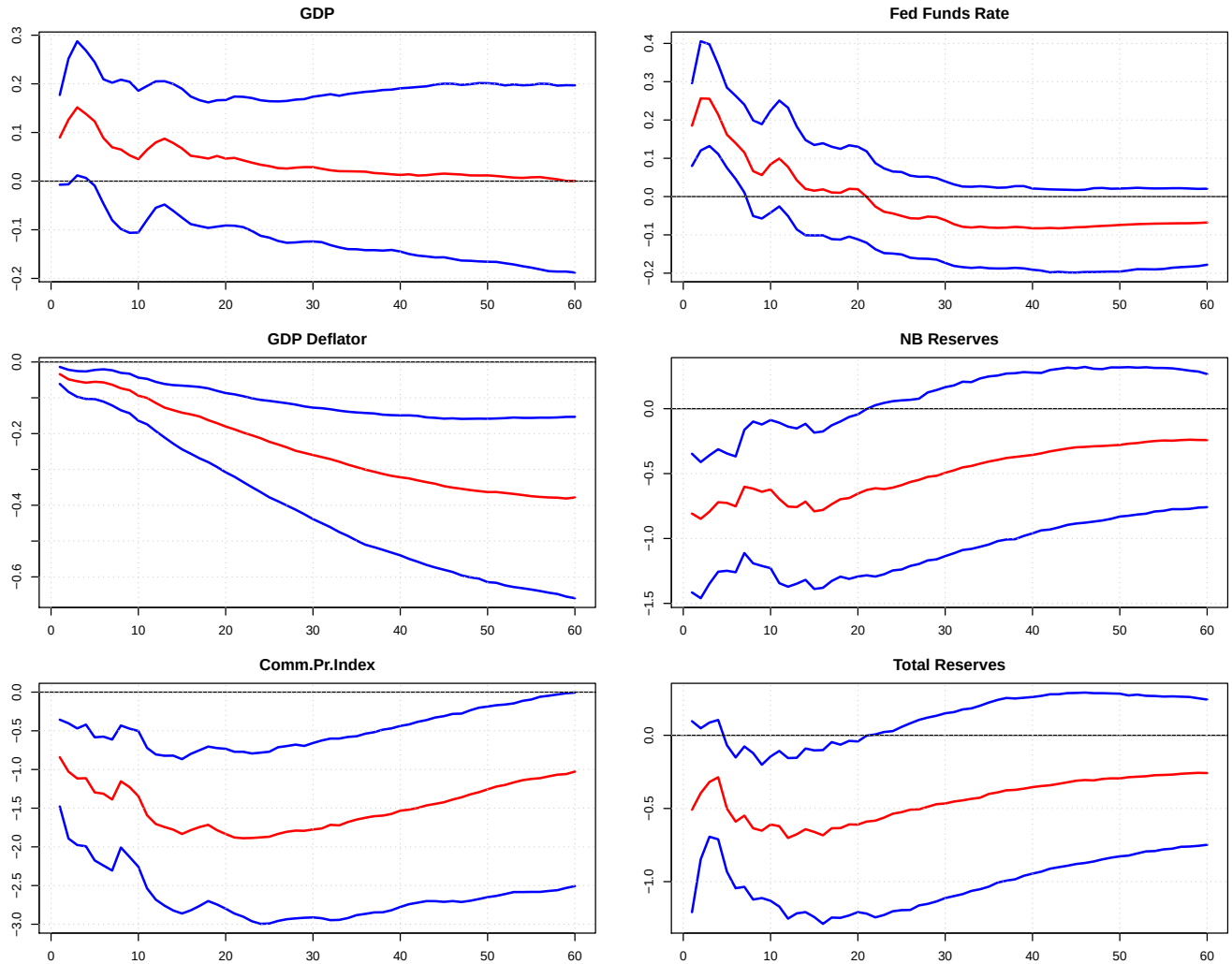
- The sign restrictions identification provides an alternative way of identifying structural shocks when the recursive identification is not plausible.
- The main idea is to restrict the sign of the impulse response functions for a number of periods k.
- For instance, a monetary policy shock is identified by restricting the impulse responses of prices, nonborrowed reserves and the federal funds rate.
- In particular it is assumed that prices and nonborrowed reserves both fall, and the federal funds rate rises.

Sign Restrictions Identification

- The idea is well defined in the Bayesian perspective.
- We have a prior that puts zero mass on some sets of impulse response functions. The posterior response functions are then the impulses responses functions that respect the restrictions.
- Here is how it works in practice:
  1. Do n_1 draws of the posterior of A_1 and Σ. The posterior will be a normal distribution if we use a conjugate prior Normal inverted-Wishart on (A_1, Σ)
  2. In each draw, extract the orthogonal innovations from the model using a Cholesky decomposition. The Cholesky decomposition here is just a way to orthogonalise shocks rather than an identification strategy.
  3. Do n_2 draws of impulse vectors and calculate the implied responses functions. If they respect restriction, keep them. Otherwise, discard them.

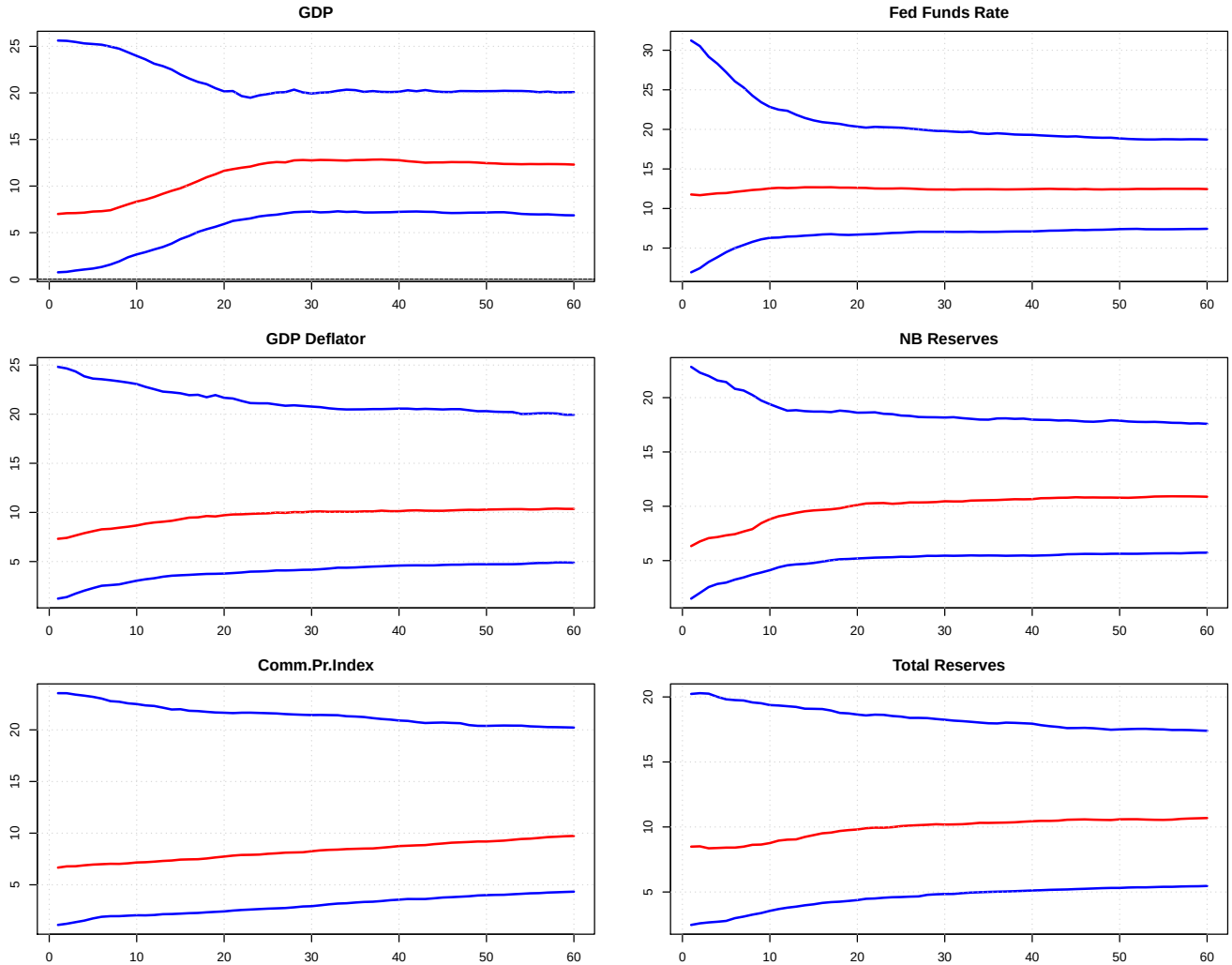
Sign-Restriction Uhlig (2005)

- Examples: Agnostic Perspective on the Effect of Monetary Policy on Output.



Sign-Restriction Uhlig (2005)

- Examples: Forecast Error Variance Decomposition



Questions to think about

- What is the difference between confidence and credible intervals?
- Why Bayesian analysis improve forecasts?
- What is a conjugate prior? And why are they useful?
- How does the sign-restriction identifies structural shocks?
- Why is it hard to formalize the sign-restriction approach in a frequentist perspective?