Bayesian Methods for DSGE models
Lecture 4
Bayesian Analysis

Kristoffer Nimark
Cornell University

July 5, 2018

Bayesian Analysis

The plan:

- Recap: Simulating the posterior distribution
- Convergence diagnostics for MCMCs
- Posterior densities of functions of θ
- Prior predictive analysis
- Model comparison and combination

Recap:

Last time we learned how to simulate the posterior density

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

for a simple DSGE model.

The probability density $p(\theta \mid y)$ describes what we know about θ given the data.

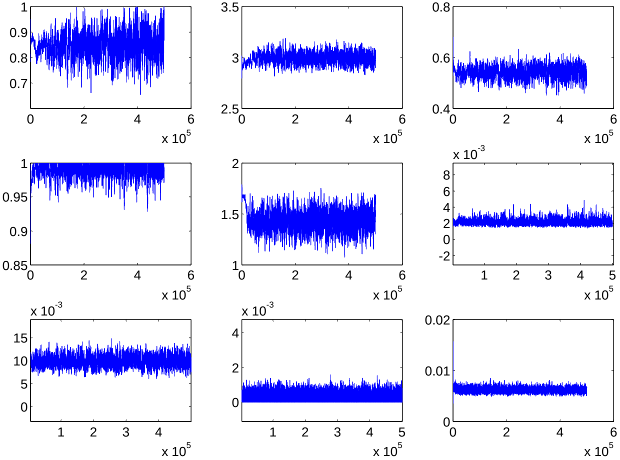
What does a simulated distribution look like?

It's a matrix:

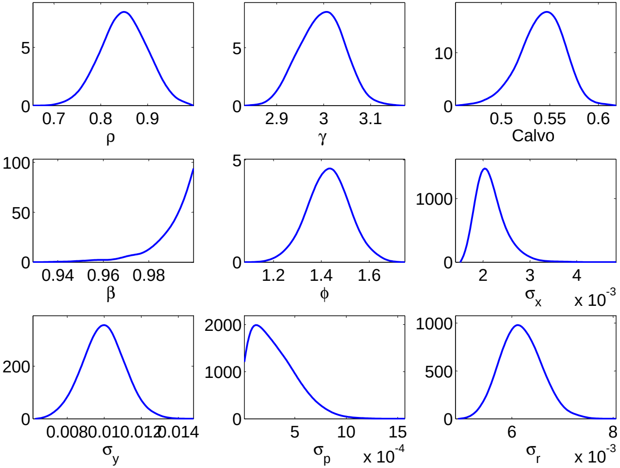
$$\begin{bmatrix} \theta_1^{(0)} & \theta_1^{(1)} & \cdots & \theta_1^{(s)} & \cdots & \theta_1^{(S)} \\ \vdots & \vdots & & \vdots & & \vdots \\ \theta_K^{(0)} & \theta_K^{(1)} & \cdots & \theta_K^{(s)} & \cdots & \theta_K^{(S)} \end{bmatrix}$$

K is the dimension of θ and S is the number of draws from the posterior.

Plotting the rows of the MCMC



Posterior distribution



Why does the MCMC converge to the target density (i.e. the posterior)?

It all depends on the rule that determines how we move from $\theta^{(s)}$ to $\theta^{(s+1)}$

$$\begin{bmatrix} \theta_1^{(0)} & \theta_1^{(1)} & \cdots & \theta_1^{(s)} & \cdots & \theta_1^{(S)} \\ \vdots & \vdots & & \vdots & & \vdots \\ \theta_K^{(0)} & \theta_K^{(1)} & \cdots & \theta_K^{(s)} & \cdots & \theta_K^{(S)} \end{bmatrix}$$

But what was the rule?

Metropolis-Hastings Algorithm

To simulate from the target density $p(\theta \mid y)$ by the Metropolis-Hastings Algorithm

1. Start with an arbitrary value $\theta^{(0)}$
2. Update from $\theta^{(s-1)}$ to $\theta^{(s)}$ (for $s = 1, 2, ...S$) by
   2.1 Generate a "candidate draw" $\theta^* \sim q(\theta^* \mid \theta^{(s-1)})$
   2.2 Define the acceptance probability

$$\alpha = \min\left(\frac{p(\theta^* \mid y)}{p\left(\theta^{(s-1)} \mid y\right)} \frac{q(\theta^{(s-1)} \mid \theta^*)}{q(\theta^* \mid \theta^{(s-1)})}, 1\right) \tag{1}$$

   2.3 Set $\theta^{(s)} = \theta^*$ if $U(0,1) \leq \alpha_s$ and $\theta^{(s)} = \theta^{(s-1)}$ otherwise.
3. Repeat Step 2 S times.
4. Output $\theta^{(1)}, \theta^{(2)}, \theta^{(3)} ..., \theta^{(S)}$

Metropolis-Hastings and the simulated posterior $p(\theta \mid y)$.

Inputs:

- Prior
- Data
- Likelihood function (e.g. the model)
    - DSGE model as a SSS
    - Kalman filter to compute the likelihood

The inputs all entered in the expression for the acceptance probability

$$\alpha = \min\left(\frac{p\left(\theta^{*} \mid y\right)}{p\left(\theta^{(s-1)} \mid y\right)} \frac{q(\theta^{(s-1)} \mid \theta^{*})}{q(\theta^{*} \mid \theta^{(s-1)})}, 1\right) \tag{2}$$

since

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$$

The DSGE model as a State Space System

The DSGE model can be viewed as a function $f(\theta) \to \{A, C, D, \Sigma_{vv}\}$ where A, C, D and Σ_{vv} are the matrices of a state space system

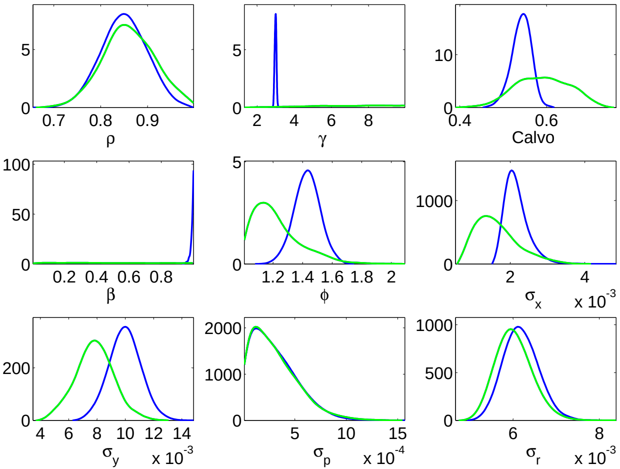
$$\begin{aligned} X_t &= AX_{t-1} + Cu_t \\ y_t &= DX_t + v_t \end{aligned}$$

We evaluated the log likelihood by computing

$$p(y \mid \theta) = -.5 \sum_{t=0}^{T} \left[p \ln(2\pi) + \ln |\Omega_t| + \widetilde{y}_t' \Omega_t^{-1} \widetilde{y}_t \right]$$

where $\widetilde{y}_t$ are the innovations from the Kalman filter

MLE $=$ posterior mode w/ uninformative priors



What can we do with the posterior?

Make probabilistic statements that allow us to quantify

- Posterior mean $E(\theta \mid y) = \int \theta p(\theta \mid y) d\theta$
- Posterior variance $var(\theta \mid y)$
- Posterior $prob(\theta_i > 0)$

These objects can all be written in the form

$$E(g(\theta) \mid y) = \int g(\theta) p(\theta \mid y) d\theta$$

where $g(\theta)$ is the function of interest.

Posterior simulation

There are only a few cases when the expected value of functions of interest can be derived analytically.

Instead, we rely on *posterior simulation* and *Monte Carlo integration.*

- Posterior simulation consists of constructing a sample from the posterior distribution $p(\theta \mid y)$
- Monte Carlo integration then uses that

$$\widehat{g}_S = \frac{1}{S} \sum_{s=1}^{S} g\left(\theta^{(s)}\right)$$

and that $\lim_{S \to \infty} \widehat{g}_S = E(g(\theta) \mid y)$ where $\theta^{(s)}$ is a draw from the posterior distribution.

Ergodicity in practice

A simulated posterior is a numerical approximation to the distribution $p(\theta \mid y)$

- We rely on ergodicity, i.e. that the moments of the constructed sample converges to the moments of the distribution $p(\theta \mid y)$ as S increases

But ergodicity is an asymptotic concept...how do we know that the chain has converged for a given S?

In other words, how can we decide on a stopping rule?

Convergence diagnostics for the MCMC

The most important diagnostic tool is .....

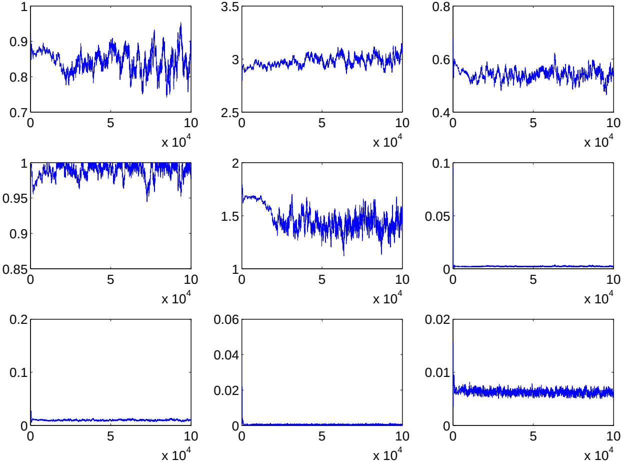
...YOUR EYES!

MCMC Diagnostics

There are several ways to check convergence, with varying degree of formality

- Ocular inspection of the raw MCMC is usually quite informative
- Plotting and inspecting recursive moments of the MCMC can also help

100 000 draws from the MCMC



Plotting the recursive means

A somewhat more formal way to check for convergence is to plot the recursive mean.

Remember: The chain is a matrix of the form

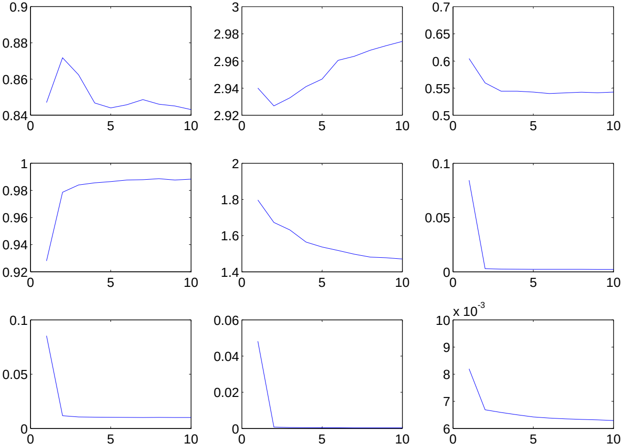
$$\begin{bmatrix} \theta_1^{(0)} & \theta_1^{(1)} & \cdots & \theta_1^{(s)} & \cdots & \theta_1^{(S)} \\ \vdots & \vdots & & \vdots & & \vdots \\ \theta_K^{(0)} & \theta_K^{(1)} & \cdots & \theta_K^{(s)} & \cdots & \theta_K^{(S)} \end{bmatrix}$$

For each $s = 0, 1, 2, ..., S$ compute μ_s^θ

$$\mu_s^\theta = \frac{1}{s} \sum_{\tau=0}^{s} \theta^{(\tau)}$$

and plot the results.

Recursive mean of MCMC 100 000 draws



Plotting recursive variance

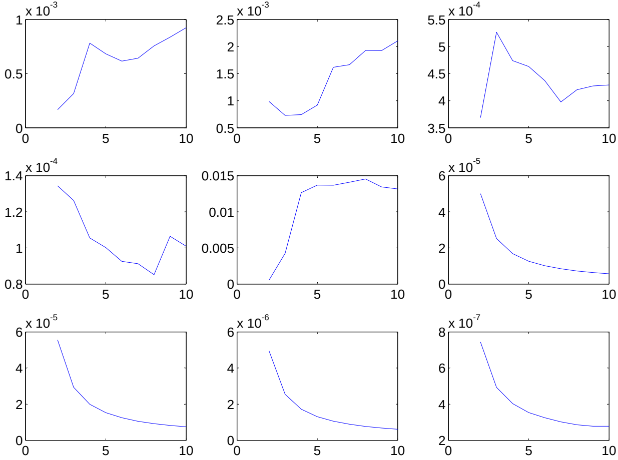
Often we care about convergence also of higher moments and in particular the second moment, i.e. the variance.

We can compute the recursive sample variance in a similar way:

$$\sigma^2_{\theta,s} = \frac{1}{s}\sum_{\tau=0}^{s}\left(\theta^{(\tau)} - \mu^{\theta}_{s}\right)^2$$

and plot the results

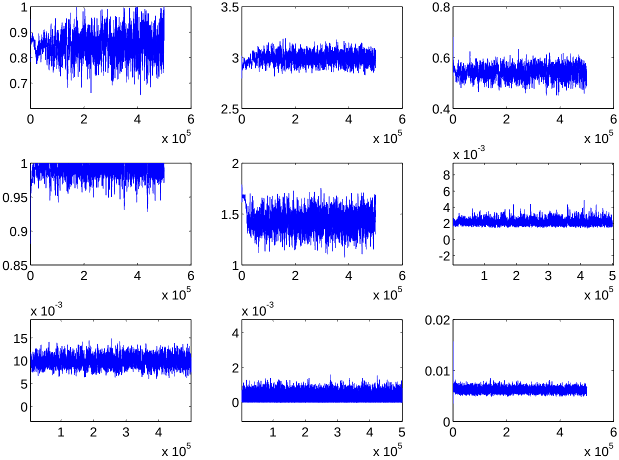
Recursive variance of MCMC, 100 000 draws



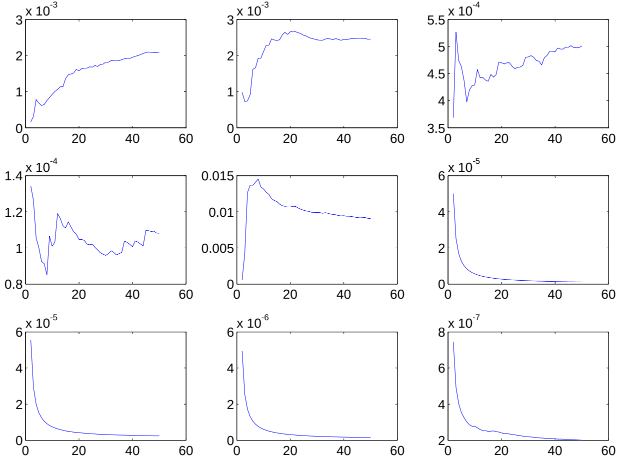
We need more draws

OK, so 100000 was not enough...how about 500 000?

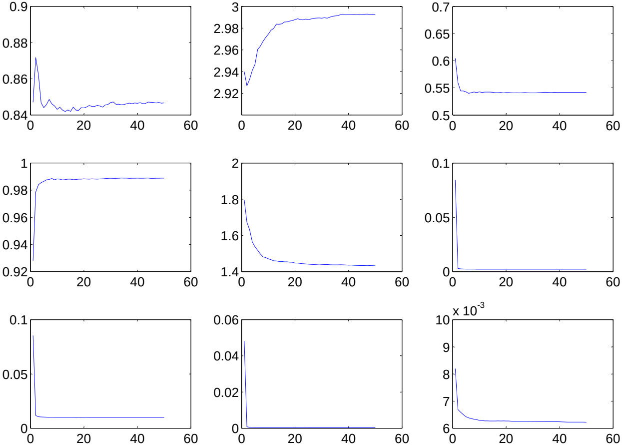
500 000 draws from the MCMC



Recursive variance of MCMC, 500 000 draws



Recursive mean of MCMC, 500 000 draws



Convergence diagnostics in Box:

`convcheck.m`

Formal MCMC convergence criteria

Koop proposes to use the CD statistic:

Divide the Markov chain into three parts, A, B and C and compute the function $g(\theta)$ for part A and C. The convergence diagnostic (CD) is then given by

$$CD = \frac{\widehat{g}_{SA} - \widehat{g}_{SC}}{\frac{\widehat{\sigma}_A}{\sqrt{S_A}} + \frac{\widehat{\sigma}_C}{\sqrt{S_C}}}$$

which should tend to a standard normal

$$CD \sim N(0,1)$$

where $\widehat{\sigma}$ is the numerical standard error of the relevant function $\widehat{g}$.

Formal convergence tests

The numerical standard error can by approximated by sample $\widehat{\sigma_g}$

$$\lim_{s\to\infty} \sqrt{s}\left\{\widehat{g}_s - E\left[g\left(\theta\right) \mid y\right]\right\} \sim N(0, \sigma_g^2)$$

where

$$\sigma_g^2 = varE\left[g\left(\theta\right) \mid y\right]$$

so that

$$\left\{\widehat{g}_s - E\left[g\left(\theta\right) \mid y\right]\right\} \sim N\left(0, \frac{\widehat{\sigma_g^2}}{\sqrt{s}}\right)$$

Burn-in sample

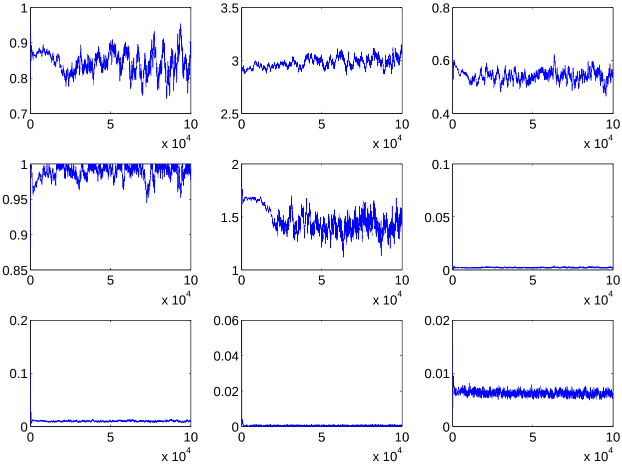
Burn-in sample

It is common practice to discard the first part of the chain

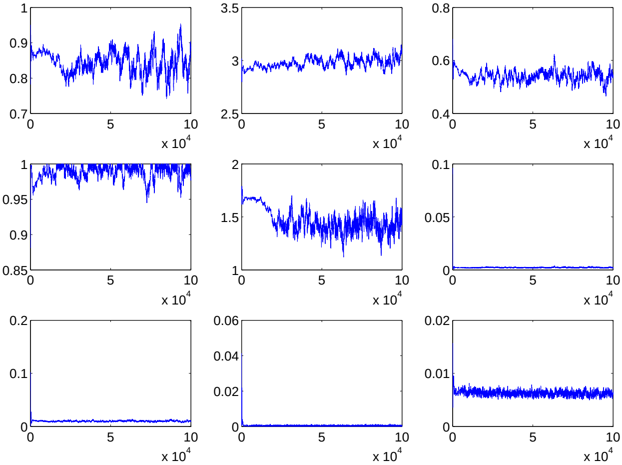
- Discard the part of chain that is not representative of invariant distribution
- The discarded part is called the *burn-in* sample
- Only the non-discarded part of the chain is then used for the analysis

There is no formal motivation for this, but it is nevertheless a good practice

Discarding a "burn-in sample"



Discarding a "burn-in sample"



Simulating posterior distributions of arbitrary functions of θ

Simulating posterior distributions of arbitrary functions of θ

We can use the MCMC to find the posterior distribution of any function $g(\theta)$

1. Draw an integer j on a uniform distribution between 1 and S
2. Compute $g(\theta)$ and save results
3. Repeat steps 1 and 2 J times ($J << S$ is ok)
4. Plot histogram of the $g(\theta)$ or find and plot percentiles

By the law of large numbers this converges to $p(g(\theta) \mid y)$ as J increases.

The MCMC

Remember: The chain is a matrix of the form

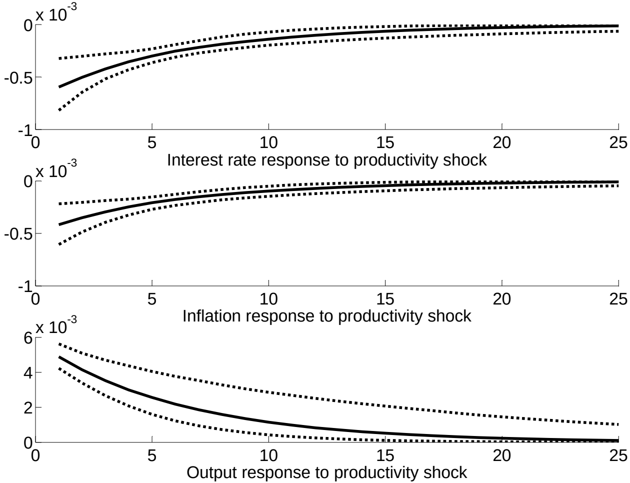
$$\begin{bmatrix} \theta_1^{(0)} & \theta_1^{(1)} & \cdots & \theta_1^{(s)} & \cdots & \theta_1^{(S)} \\ \vdots & \vdots & & \vdots & & \vdots \\ \theta_K^{(0)} & \theta_K^{(1)} & \cdots & \theta_K^{(s)} & \cdots & \theta_K^{(S)} \end{bmatrix}$$

The algorithm randomly picks elements from the chain and for each draw computes the function $g(\theta)$

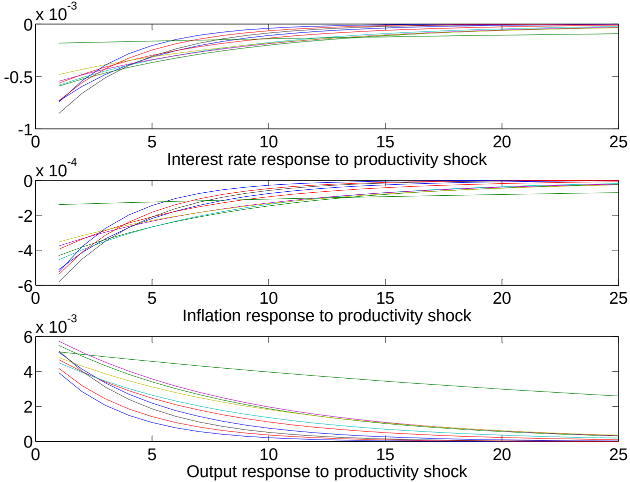
Example I: Probability intervals for impulse response function

1. Draw an integer j on a uniform distribution between 1 and S
2. Compute $DA^t C_i$ for $t = 0, 1, 2, \ldots$ using $\theta^{(j)}$ and save results.
3. Repeat steps 1 and 2 J times (usually $J < S$ is sufficient).
4. Find percentiles of the saved outputs from $DA^t C_i$ for each horizon t. These are the probability intervals of $A^t C_i$.
5. Plot.

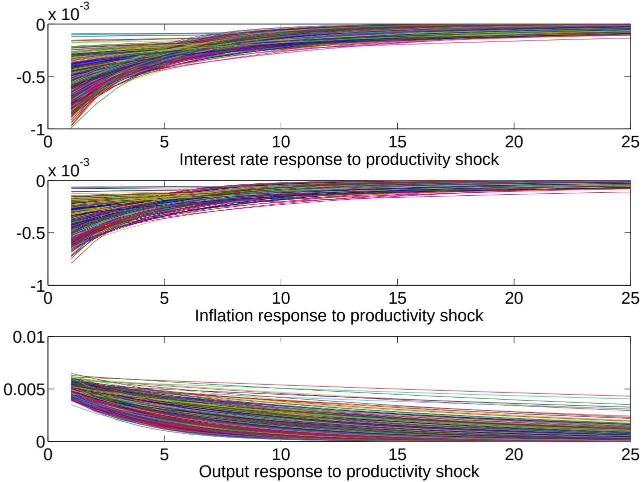
Posterior mean and 95% prob interval of IRF



Unsorted IRF S=10



Unsorted IRF S=500



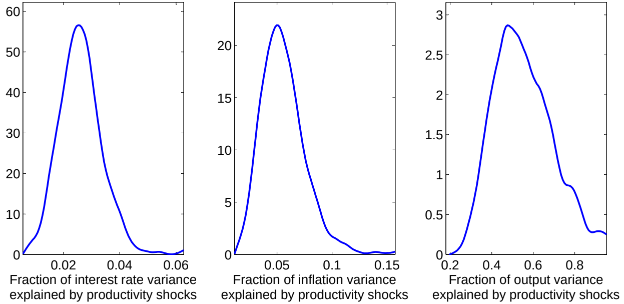
Example II: Probability intervals for variance decompositions

1. Draw an integer j on a uniform distribution between 1 and S
2. Compute variance decomposition using $\theta^{(j)}$
   - Unconditional variance Σ_Y:

     $$\Sigma_Y = D\Sigma_x D' + \Sigma_{VV}$$

   - Divide diagonal elements in Σ_{VV} by the corresponding diagonal elements of Σ_Y
   - Save results
3. Repeat steps 1 and 2 J times
4. Plot the posterior distributions.

Posterior of variance decompositions



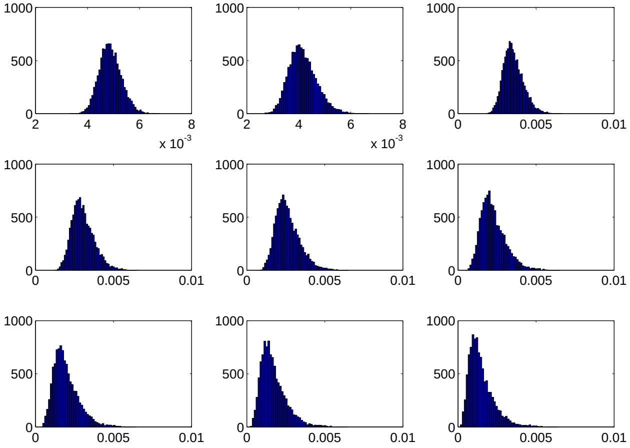
Example III: Posterior probabilities of logical statements

We can use the MCMC to find the posterior probability of logical statements such as $prob(\ f(\theta) > k \mid y)$

1. Set $c = 0$
2. Draw an integer j on a uniform distribution between 1 and S
3. Compute $f(\theta)$ and check if $f(\theta) > k$
    3.1 If statement true add c=c+1
4. Repeat steps 1 and 2 J times
5. $prob(\ f(\theta) > k \mid y) = c/J$

Example: Probability that a 1 s.d. productivity shock increases output by more than a 0.5 per cent = 0.36

Densities of output IRFs at different horizons



Prior predictive analysis

Prior predictive analysis

Prior predictive analysis is a tool to ask what is "possible" given

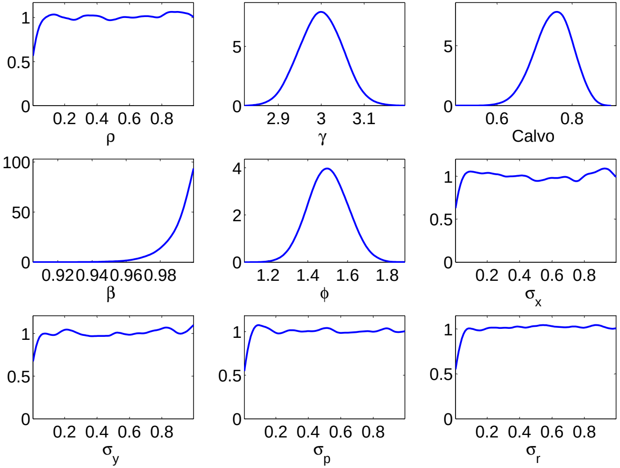
- Model
- Prior

How does it work?

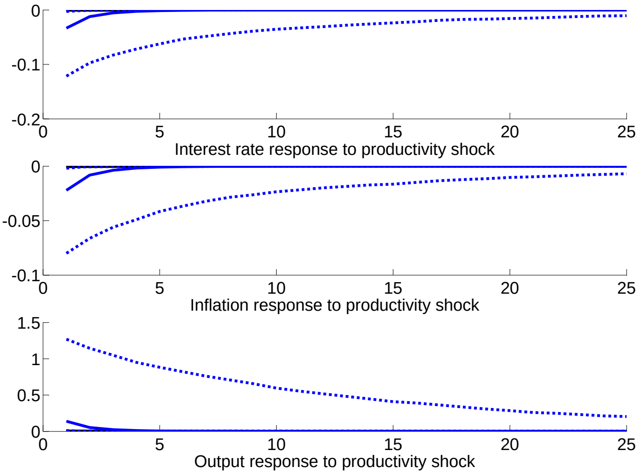
- Draw θ's from MCMC generated from prior distribution (or draw directly from prior distribution if possible)
- For each θ compute objects of interests

This is a good method to illustrate what components of the model outputs that are truly empirical results and what are implied by your choice of model and priors

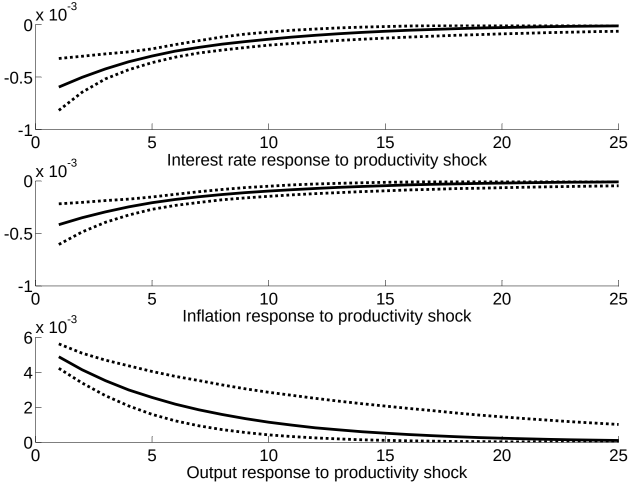
Prior densities



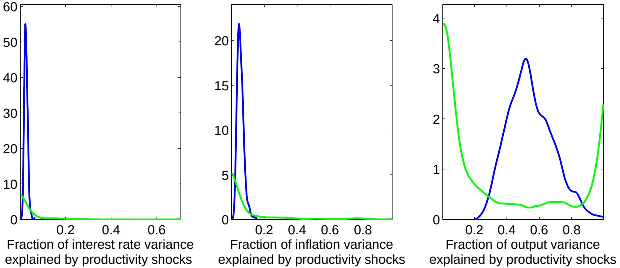
Prior predictive IRFs



Posterior mean and 95% prob interval of IRF



Prior predictive variance decomposition



Density forecasts

Density forecasts are a tool that allows us to express uncertainty around forecasts.

A Bayesian framework allows us to take into account:

- Parameter uncertainty
- State uncertainty
- Shock uncertainty

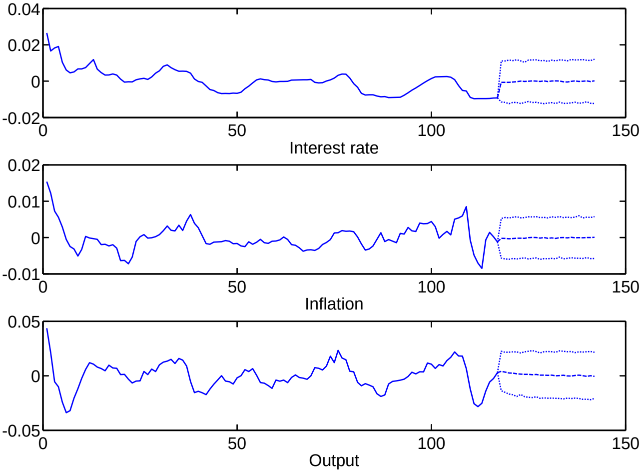
Density forecasts

1. Draw an integer j on a uniform distribution between 1 and S
2. Use

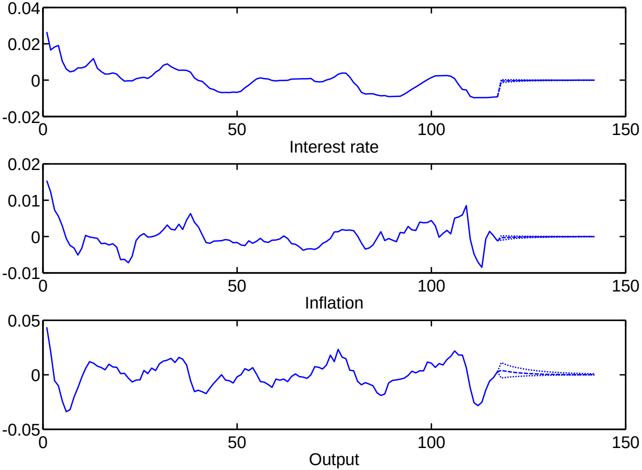
$$\begin{aligned} p(x_t \mid y, \theta) &\sim N(x_{t|t}, p_{t|t}) \\ y_{t+s} &= Dx_{t+s} + \mathbf{v}_{t+s} \\ x_{t+s} &= A^s x_t + \sum_{\tau=0}^{s} A^\tau u_{t+s-\tau} \end{aligned}$$

   to draw from the forecast distribution at each horizon and save results
3. Repeat steps 1 and 2 J times.
4. Plot the posterior distributions.

Density forecast



Parameter uncertainty and forecasts



Model comparison

Model comparison

A Bayesian approach to hypothesis testing

- We may have several models (or hypothesis) that may have generated the data
- What is the posterior probability that each theory/model is "correct"?

Examples:

- Is monetary policy better described as operating under discretion or commitment?
- Is Ricardian equivalence a good description of how households respond to changes in fiscal policy?

Bayes Factors described the relative strength of evidence for competing models/theories

Model comparison

We may consider several plausible models

Index different models by $i = 1, 2, ...m$

$$p(\theta \mid y, M_i) = \frac{p(y \mid \theta, M_i) p(\theta \mid M_i)}{p(y \mid M_i)}$$

Model comparison

The posterior model probability is given by

$$p(M_i \mid y) = \frac{p(y \mid M_i)p(M_i)}{p(y)}$$

where $p(y \mid M_i)$ is called the *marginal likelihood.* It can be computed from

$$\int p(\theta \mid y, M_i)d\theta = \int \frac{p(y \mid \theta, M_i)p(\theta, M_i)}{p(y \mid M_i)}d\theta$$

by using that $\int p(\theta \mid y, M_i)d\theta = 1$ so that

$$p(y \mid M_i) = \int p(y \mid \theta, M_i)p(\theta, M_i)d\theta$$

It is generally difficult to evaluate the marginal likelihood

The posterior odds ratio

The *posterior odds ratio* is the relative probabilities of two models conditional on the data

$$\frac{p(M_i \mid y)}{p(M_j \mid y)} = \frac{p(y \mid M_i)p(M_i)}{p(y \mid M_j)p(M_j)}$$

It is made up of

- The *prior odds* ratio $\frac{p(M_i)}{p(M_j)}$
- The *Bayes factor* $\frac{p(y|M_i)}{p(y|M_j)}$

The Bayes factor require computing the *marginal likelihood* $p(y \mid M_i)$ of each model

The marginal likelihood

Why is it difficult to compute?

- The marginal likelihood is not generally a function of the posterior distribution
- Simulation methods discussed earlier do not apply directly

What to do?

The Gelfand and Dey method to compute the marginal likelihood

Gelfand and Dey's method uses that we can rewrite Bayes Rule as

$$\frac{1}{p(Y)} = \frac{1}{p\left(Y \mid \theta\right) p\left(\theta\right)} p\left(\theta \mid Y\right)$$

Multiply both sides with $f\left(\theta\right)$ s.t. $\int f\left(\theta\right) d\theta = 1$ to get

$$\frac{1}{p(Y)} = \int \frac{f\left(\theta\right)}{p\left(Y \mid \theta\right) p\left(\theta\right)} p\left(\theta \mid Y\right) d\theta$$

$p(Y)$ can be approximated by

$$p(Y) \approx \left[\frac{1}{N} \sum_{i=1}^{N} \frac{f\left(\theta^i\right)}{p\left(Y \mid \theta^i\right) p\left(\theta^i\right)}\right]^{-1}$$

Works well only when $f(\theta) \simeq p(\theta \mid y, M_i)$

Geweke's harmonic mean estimate of the marginal likelihood

Geweke (1999) suggested to use the truncated normal

$$
\begin{aligned}
f(\theta) &= \tau^{-1}(2\pi)^{-d/2}|V_\theta|^{-1/2}\exp\left[0.5\left(\theta-\overline{\theta}\right)' V_\theta^{-1}\left(\theta-\overline{\theta}\right)\right] \\
&\quad \times I\left\{\left(\theta-\overline{\theta}\right)' V_\theta^{-1}\left(\theta-\overline{\theta}\right) \leq F^{-1}(\tau)\right\}
\end{aligned}
$$

in

$$p(Y) \approx \left[\frac{1}{N}\sum_{i=1}^{N}\frac{f\left(\theta^i\right)}{p\left(Y \mid \theta^i\right)p\left(\theta^i\right)}\right]^{-1}$$

Why truncate the tails?

- Avoids making the ratio infinite

Bayes Factors and priors

Priors must be proper densities

- Improper priors would imply that $p(\theta) = 0$

The role of priors for Bayes Factors tend to be larger than for posterior parameter densities

- An intentionally diffuse prior for parameters of a model will penalize that model's Bayes Factor

Bartlett's paradox: Bayes Factors may favor *strong-but-wrong* priors over uninformative priors.

Bayes Factors and priors

If models are non-nested, it can be difficult to ensure that priors do not penalize one model over the other

- Mapping between θ and $p(y \mid \theta)$ is often very indirect

What to do?

1. Estimate both models on a training sample with improper priors
2. Use posteriors from training sample as priors when estimating the models on the rest of the sample
3. Compute the implied Bayes Factors

Schwarz (1978) approximation

A simpler way that can be used to approximate the posterior odds ratio is to use the Schwarz approximation

$$PO \approx e^{\log L\left(\mathbf{y}^T|\widehat{\theta}_i\right) - \log L\left(\mathbf{y}^T|,\widehat{\theta}_j\right) - \frac{1}{2}\left(\dim\theta_i - \dim\theta_j\right)\ln T}$$

where $\widehat{\theta}_i$ is the posterior mode of the parameters of model i and $\dim\theta_i$ is the number of parameters in model i.

Penalty for large number of parameters is a fundamental aspect of marginal likelihoods

Posterior odds ratios thus have Occam's Razor built in

Words instead of numbers

While posterior odds ratios has clear probabilistic interpretations, Kass and Raftery (1995) suggest the following interpretation based on existing practice.

B_{10}	*Evidence against* M_0
1 - 3.2	Not worth more than a bare mention
3.2 - 10	Substantial
10 - 100	Strong
$>$100	Decisive

Model combination

Model combination

One common use of posterior odds ratios is to use it to combine models (so-called *Bayesian model averaging*)

- Useful when there is substantial model uncertainty
- Probabilistically coherent method for combining information from different models

How does it work?

Density forecasts using Bayesian model averaging

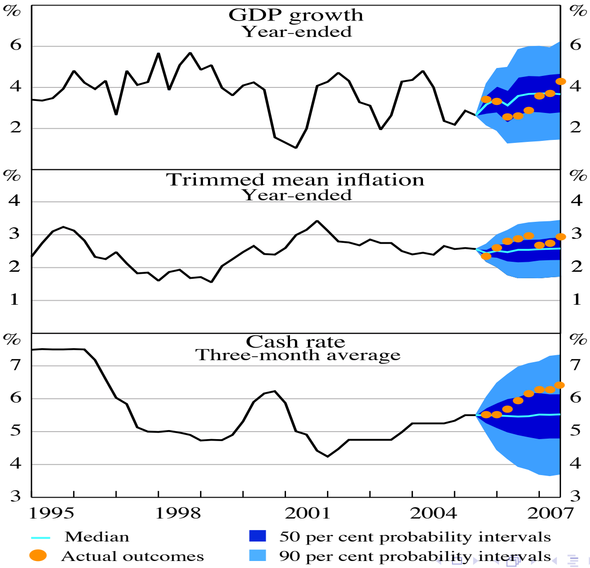
Consider two models with posterior odds ratio $0 < PO_{12} < 1$

1. Draw α from $U(0,1)$
2. If $\alpha < PO_{12}$ take a draw from the forecast distribution implied by model 1. Otherwise draw from forecast distribution implied by model 2. Save result.
3. Repeat steps 1 and 2 J times.
4. Plot the posterior distributions of the forecasts.

Conditional on a model, drawing from the forecast is done just like before.

Method works for any object of interest that can be written as a function of the model parameters.

Oz density forecast using DSGE, DFM and BVAR



Model comparison and combination in practice

Posterior odds ratios and Bayesian model averaging are logically consistent applications of probability theory

But:

- Posterior odds ratios often seem to provide "too strong" evidence in favour of one model over the other
- Equal model weights often outperform odds ratio based weights in out-of-sample forecasting

Summing up

Convergent checks on MCMC

- Non-convergence means simulated distribution different from posterior
- Non-formal convergence checks are often sufficient

Sampling from posterior MCMC allows us to construct probability intervals for any function$g(\theta)$

- Fewer draws needed relative to construction of MCMC since we can take independent samples

Model comparisons (posterior odds ratios) can be used to evaluate relative fit

- Conceptually straightforward, but somewhat involved in practice

Bayesian averaging allows for combination of several models

- No need to reject, or to choose models when several plausible models are available

That's it for today.