# Bayesian Estimation

## Evaluating the Posterior

Tools for Macroeconomists: The essentials

Petr Sedláček

# Evaluating the posterior

## Starting point

Aim is to be able to calculate something like

$$\mathbb{E}[g(\Psi)] = \frac{\int g(\Psi)P(\Psi|\mathcal{Y}^T)d\Psi}{\int P(\Psi|\mathcal{Y}^T)d\Psi}$$

- we know how to calculate $P(\Psi|\mathcal{Y}^T)$
- but we cannot draw from it
- the system is too large for numerical integration

# Principle of posterior evaluation

We cannot draw from the “target” distribution, but

- 1. can draw from a different, “stand-in”, distribution
- 2. can evaluate both stand-in and target distributions
- 3. comparing the two, we can re-weigh the draw “cleverly”

# Principle of posterior evaluation

- the above procedure is the idea of “importance sampling”
- MCMC methods effectively a version of importance sampling
    - traveling through the parameter space is more sophisticated
    - and or acceptance probability more sophisticated

# Evaluating the posterior

## A Few Examples

## A few simple examples

Problem:

- we want to simulate $x$
- $x$ comes from truncated normal with
  - mean $\mu$ and variance $\sigma^2$
  - and $a < x < b$

Solution:

- 1. draw $y$ from $N(\mu, \sigma^2)$
- 2a. if $y \in (a, b)$ then keep draw (accept) and go back to 1
- 2b. otherwise discard draw (reject) and go back to 1

## A few simple examples

Problem:

- want to draw $x$ from $F(x)$, but we cannot
- we can sample from $H(x)$ **and** $f(x) \leq ch(x)\ \forall x$

Solution:

- 1. sample $y$ from $H(y)$
- 2. accept draw with probability $\frac{f(y)}{ch(y)}$ and go back to 1

Note:

- acceptance rate higher for lower $c$
- optimal $c$ is $c = sup_x \frac{f(x)}{h(x)}$
- Metropolis-Hastings sampler (MCMC) is a generalization

# Evaluating the posterior

## Importance Sampling

# Importance sampling

Main idea very similar to the previous example:

- cannot draw from $P(\Psi|\mathcal{Y}^T)$
- but can draw from $H(\Psi)$
- be smart in reweighing (accepting) the draws

# Importance sampling

$$\mathbb{E}[g(\Psi)] = \frac{\int g(\Psi)P(\Psi|\mathcal{Y}^T)d\Psi}{\int P(\Psi|\mathcal{Y}^T)d\Psi}$$

$$= \frac{\int g(\Psi)P(\Psi|\mathcal{Y}^T)\frac{h(\Psi)}{h(\Psi)}d\Psi}{\int P(\Psi|\mathcal{Y}^T)\frac{h(\Psi)}{h(\Psi)}d\Psi}$$

$$= \frac{\int g(\Psi)\omega(\Psi)h(\Psi)d\Psi}{\int \omega(\Psi)h(\Psi)d\Psi}$$

Approximate the integral using MC integration:

$$\mathbb{E}[g(\Psi)] \approx \frac{\sum_{m=1}^{M} \omega(\Psi^{(m)}) g(\Psi^{(m)})}{\sum_{m=1}^{M} \omega(\Psi^{(m)})}$$

- $M$ is the number of draws from importance function $h(\Psi)$

How to best choose $h(.)$?

- we'd like $h(.)$ to have fatter tails compared to $f(.)$
- normal distribution has rather thin tails
- $\rightarrow$ often not a good importance function

# Evaluating the posterior

## Markov Chain Monte Carlo

## Some preliminaries for MCMC

Markov property:

- if for all $k \geq 1$ and all $t$ $P(x_{t+1}|x_t, x_{t-1}, ..., x_{t-k}) = P(x_{t+1}|x_t)$

Transition kernel:

- $\mathcal{K}(x, y) = P(x_{t+1} = y|x_t = x)$ for $x, y \in \mathcal{X}$
- $\mathcal{X}$ is the sample space

# Main idea behind MCMC methods

- as before, we'd like to sample from $P(\Psi|\mathcal{Y}^T)$, but we cannot
- MCMC methods provide a way to
  - create a **Markov chain transition kernel** ($\mathcal{K}$) for $\Psi$
    - that has an invariant density $P(\Psi|\mathcal{Y}^T)$
  - why is this useful?
    - starting with some initial values $P(\Psi_0)$
    - simulate the Markov chain $P' = \mathcal{K}P$
    - (eventually) distribution of Markov chain $\rightarrow P(\Psi|\mathcal{Y}^T)$

Ways of constructing such kernels

- Gibbs sampling
    - special case, more often in empirical work
- Metropolis (-Hastings) algorithm
    - we'll talk about this in detail

# Evaluating the posterior

## Gibbs Algorithm

# Gibbs algorithm

Special case: can sample from conditional distributions

- instead of draws of $\Psi$ from $P(\Psi|\mathcal{Y}^T)$
- portion $\Psi$ into $k$ blocks
- sample each from $P(\Psi^j|\mathcal{Y}^T, \Psi^{-j})$ for $j = 1, ..., k$
- iterate until convergence

# Gibbs sampling

Iterations ($k = 2$):

- initiate sample with $\Psi_0$
- then iterate according to:

$$\Psi^1_{i+1} \sim P(\Psi^1 | \mathcal{Y}^T, \Psi^2_i)$$

$$\Psi^2_{i+1} \sim P(\Psi^2 | \mathcal{Y}^T, \Psi^1_i)$$

- can prove that the above converges to $P(\Psi | \mathcal{Y}^T)$
- discard first $B$ number of draws to eliminate influence of $\Psi_0$

## Gibbs sampling

- once Markov chain has converged
- proceed as if we could sample directly:

$$\mathbb{E}[g(\Psi)] = \frac{1}{m}\sum_{i=1}^{m} g(\Psi_i)$$

- certain caveats, such as serial correlation of draws

# Evaluating the posterior

## Metropolis-Hastings algorithm

## Metropolis-Hastings algorithm

Main idea same as with importance sampling:

- 1. draw from a stand-in distribution $h(\Psi; \theta)$
  - $\theta$ explicitly shows parameters of stand-in distribution
  - e.g. mean ($\mu_h$ ) and variance ($\sigma_h^2$)
- 2. accept/reject based on probability $q(\Psi_{i+1}|\Psi_i)$
- 3. go back to 1
  - 3a. stand-in density does not change (indpendent MH)
  - 3b. mean of stand-in adjusts (random walk MH)
- can show convergence to target distribution

# Acceptance probability

"Metropolis"

$$q(\Psi_{i+1}|\Psi_i) = \min\left[1, \frac{P(\Psi^*_{i+1}|\mathcal{Y}^T)}{P(\Psi_i|\mathcal{Y}^T)}\right]$$

- $\Psi^*_{i+1}$ is the new candidate draw from stand-in distribution
- if $P(\Psi^*_{i+1}|\mathcal{Y}^T)$ high relative to $P(\Psi_i|\mathcal{Y}^T)$
- $\rightarrow$ accept candidate draw with certainty

## ACCEPTANCE PROBABILITY

"Metropolis-Hastings"

$$q(\Psi_{i+1}|\Psi_i) = \min\left[1, \frac{P(\Psi^*_{i+1}|\mathcal{Y}^T)}{P(\Psi_i|\mathcal{Y}^T)} \frac{h(\Psi_i;\theta)}{h(\Psi^*_{i+1};\theta)}\right]$$

- scale down by relative likelihood in stand-in density
    - a more "common" draw from the stand-in gets less "weight"
    - $\rightarrow q(\Psi_{i+1}|\Psi_i)$ is lowered
    - force algorithm to explore less likely areas of the state-space

# Updating the stand-in density

"Independence chain variant"

- stand-in distribution does not change
- it is independent across Monte Carlo replications
- this is also the case in importance-sampling

# Updating the stand-in density

"Random walk variant"

- candidate draws are obtained according to $\Psi^*_{i+1} = \Psi_i + \epsilon_{i+1}$
- $\epsilon_i$ from a symmetric density around 0 and variance $\sigma^2_h$
  - mean of the stand-in density adjusts with each accepted draw
  - in $\theta$, $\mu_h = \Psi_i$

## Summary of MCMC with MH algorithm

- 1. maximize posterior $P(\Psi|\mathcal{Y}^T)$
  - this yields the posterior mode $\widehat{\Psi}$
- 2. draw from a stand-in distribution $h(\Psi;\theta)$
  - should have fatter tails than posterior
- 3. accept/reject based on probability $q(\Psi_{i+1}|\Psi_i)$
  - Metropolis vs. Metropolis-Hastings specification
- 4a. go back to 2
  - random walk vs. independent chain variant
- 4b. stop
  - still need to discuss convergence criteria

# Summary of MCMC with MH algorithm

In the end, it doesn't seem so bad

- but where do you have to compute the likelihood?
- what does this entail?

So why not just be more clever and use conjugate priors?

# Evaluating the posterior

## Convergence

# Choice of stand-in density

- stand-in should have fatter tails
- variance parameter important for acceptance rate
- optimal acceptance rates:
    - around 0.44 for estimation of 1 parameter
    - around 0.23 for estimation of more than 5 parameters

## Choice of stand-in density

- often, stand-in is $N(\hat{\Psi}, c^2\Sigma_\Psi)$
    - $\hat{\Psi}$ is the posterior mode
    - $\Sigma_\Psi$ is the inverse (negative) Hessian at the mode
- tip: start with $c = 2.4/\sqrt{d}$
    - $d$ is number of estimated parameters
- increase (decrease) $c$ if acceptance rate is too high (low)

# Convergence statistics

- theory says that distribution will converge to target
- when does this happen?
- $\rightarrow$ diagnostic tests
    - sequence of draws should be from the invariant distribution
    - moments should not change within/between sequences

# Brooks and Gelman statistics

- $I$ draws and $J$ sequences

$$W = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{I-1} \sum_{i=1}^{I} \left(\Psi_{i,j} - \overline{\Psi}_j\right)^2$$

$$B = \frac{I}{J} \sum_{j=1}^{J} \left(\overline{\Psi}_j - \overline{\Psi}\right)^2$$

- $B/I$: estimate of the variance of the mean across sequences
- $W$: estimate of average variance within sequences

# Brooks and Gelman statistics

Combine the two measures of variance:

$$V = \frac{I-1}{I} W + \frac{B}{I}$$

- as the length of the simulation increases
- want these statistics to "settle down"

# Geweke statistic

- partition a sequence into 3 subsets $s = \{I, II, III\}$
- compute mean ($\overline{\Psi}^s$) and standard errors ($\sigma_{\Psi}^s$)
    - s.e.'s must be corrected for serial correlation
- then, under convergence CD is distributed $N(0,1)$

$$CD = \frac{\overline{\Psi}^I - \overline{\Psi}^{III}}{\sigma_{\Psi}^I + \sigma_{\Psi}^{III}}$$

# Evaluating the posterior

## Taking Stock

Evaluating the posterior distribution

- draw from a "stand-in" distribution
- evaluate draw under stand-in and posterior distribution
- use the relative probabilities to accept/reject the draw
- for each draw, need to re-solve the model


That's all Folks!