Introduction to Bayesian Estimation

Wouter J. Den Haan
London School of Economics

© 2011 by Wouter J. Den Haan

August 29, 2011

Overview

- A very useful tool: Kalman filter
- Maximum Likelihood
  - Singularity when #shocks $\leq$ number of observables
- Bayesian estimation
- Tools:
  - Metropolis Hastings
- Remaining issues

Rudolph E. Kalman

born in Budapest, Hungary, on May 19, 1930

Kalman filter

- Linear projection
- Linear projection with orthogonal regressors
- Kalman filter

The slides for the Kalman filter is based on Ljungqvist and Sargent's textbook

Linear projection

- y: $n_y \times 1$ vector of random variables
- x: $n_x \times 1$ vector of random variables

- First and second moments exist

$$
\begin{array}{lll}
\mathrm{E}y = \mu_y & \tilde{y} = y - \mu_y & \mathrm{E}\tilde{x}\tilde{x}' = \Sigma_{xx} \\
\mathrm{E}x = \mu_x & \tilde{x} = x - \mu_x & \mathrm{E}\tilde{y}\tilde{y}' = \Sigma_{yy} \\
 & & \mathrm{E}\tilde{y}\tilde{x}' = \Sigma_{yx}
\end{array}
$$

Definition of linear projection

The *linear projection* of y on x is the function

$$\widehat{\mathsf{E}}\left[y|x\right] = a + Bx,$$

a and B are chosen to minimize

$$\mathsf{E}\ \text{trace}\ \left\{(y - a + Bx)(y - a + Bx)'\right\}$$

Formula for linear projection

The *linear projection* of y on x is given by

$$\widehat{\mathsf{E}}\left[y|x\right] = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)$$

Difference with linear regression problem

- True model:

$$\begin{aligned} y &= \bar{B}x + \bar{D}z + \varepsilon, \\ \mathrm{E}x &= \mathrm{E}z = \mathrm{E}\varepsilon = 0, \mathrm{E}\left[\varepsilon | x, z\right] = 0,\ \mathrm{E}\left[z | x\right] \neq 0 \end{aligned}$$

$\bar{B}$: measures the effect of x on y *keeping all else—also* z *and* ε*—constant.*

- Particular regression model:

$$y = \bar{B}x + u$$

Difference with linear regression problem

Comments:

- Least-squares estimate $\neq \bar{B}$
- Projection:

$$\widehat{\mathsf{E}}\left[y|x\right] = Bx = \bar{B}x + \bar{D}\widehat{\mathsf{E}}\left[z|x\right]$$

- Projection well defined
  linear projection can include more than the direct effect:

Message:

- You can always define the linear projection
- you don't have to worry about the properties of the error term.

Linear Projection with orthogonal regressors

- $x = [x_1, x_2]$ and suppose that $\Sigma_{x_1 x_2} = 0$
- x_1 and x_2 could be vectors

$$\begin{aligned}
\widehat{\mathrm{E}}[y|x] &= \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x) \\
&= \mu_y + \left[\Sigma_{yx_1}\ \Sigma_{yx_2}\right]\begin{bmatrix} \Sigma_{x_1x_1}^{-1} & 0 \\ 0 & \Sigma_{x_2x_2}^{-1} \end{bmatrix}(x - \mu_x) \\
&= \mu_y + \Sigma_{yx_1}\Sigma_{x_1x_1}^{-1}(x_1 - \mu_{x_1}) + \Sigma_{yx_2}\Sigma_{x_2x_2}^{-1}(x_2 - \mu_{x_2})
\end{aligned}$$

Thus

$$\widehat{\mathrm{E}}[y|x] = \widehat{\mathrm{E}}[y|x_1] + \widehat{\mathrm{E}}[y|x_2] - \mu_y \qquad (1)$$

Time Series Model

$$x_{t+1} = Ax_t + Gw_{1,t+1}$$
$$y_t = Cx_t + w_{2,t}$$
$$Ew_{1,t} = Ew_{2,t} = 0$$
$$\mathsf{E}\begin{bmatrix} w_{1,t+1} \\ w_{2,t} \end{bmatrix}\begin{bmatrix} w_{1,t+1} \\ w_{2,t} \end{bmatrix}' = \begin{bmatrix} V_1 & V_3 \\ V_3' & V_2 \end{bmatrix}$$

Time Series Model

- y_t is observed, but x_t is not
- the coefficients are known (could even be time-varying)
- Initial condition:
  - x_1 is a random variable (mean μ_{x_1} & covariance matrix Σ_1)
- $w_{1,t+1}$ and $w_{2,t}$ are serially uncorrelated and orthogonal to x_1

Objective

The objective is to calculate

$$\begin{aligned}\widehat{\mathsf{E}}_t x_{t+1} &\equiv \widehat{\mathsf{E}}\left[x_{t+1}|y_t, y_{t-1}, \cdots, y_1, \tilde{x}_1\right] \\ &= \widehat{\mathsf{E}}\left[x_{t+1}|Y^t, \tilde{x}_1\right]\end{aligned}$$

where $\tilde{x}_1$ is an initial estimate of x_1 (Typically μ_{x_1})

Trick: get a recursive formulation

Orthogonalization of the information set

- Let
  - $\hat{y}_t = y_t - \widehat{\mathrm{E}}\left[y_t | \hat{y}_{t-1}, \hat{y}_{t-2}, \cdots, \hat{y}_1, \tilde{x}_1\right]$
  - $\hat{Y}^t = \{\hat{y}_t, \hat{y}_{t-1}, \cdots, \hat{y}_1\}$

- space spanned by $\{\tilde{x}_1, \hat{Y}^t\}$ = space spanned by $\{\tilde{x}_1, Y_t\}$

  - That is, anything that can be expressed as a linear combination with elements in $\{\tilde{x}_1, \hat{Y}^t\}$ can be expressed as a linear combination of elements in $\{\tilde{x}_1, Y_t\}$.

Orthogonalization of the information set

- Then

$$\widehat{\mathsf{E}}\left[y_{t+1}|Y^t,\tilde{x}_1\right] = \widehat{\mathsf{E}}\left[y_{t+1}|\hat{Y}^t,\tilde{x}_1\right] = C\widehat{\mathsf{E}}\left[x_{t+1}|\hat{Y}^t,\tilde{x}_1\right] \tag{2}$$

Derivation of the Kalman filter

From (1) we get

$$\widehat{\mathsf{E}}\left[x_{t+1}|\hat{Y}^t, \tilde{x}_1\right] = \widehat{\mathsf{E}}\left[x_{t+1}|\hat{y}_t\right] + \widehat{\mathsf{E}}\left[x_{t+1}|\hat{Y}^{t-1}, \tilde{x}_1\right] - \mathsf{E}x_{t+1} \qquad (3)$$

The first term in (3) is a standard linear projection:

$$\begin{aligned}\widehat{\mathsf{E}}\left[x_{t+1}|\hat{y}_t\right] &= \mathsf{E}x_{t+1} + \mathrm{cov}(x_{t+1}, \hat{y}_t)\left[\mathrm{cov}(\hat{y}_t, \hat{y}_t)\right]^{-1}(\hat{y}_t - \mathsf{E}\hat{y}_t) \\ &= \mathsf{E}x_{t+1} + \mathrm{cov}(x_{t+1}, \hat{y}_t)\left[\mathrm{cov}(\hat{y}_t, \hat{y}_t)\right]^{-1}\hat{y}_t\end{aligned}$$

Some algebra

- Similar to the definition of $\hat{y}_t$, let

$$\begin{aligned}\hat{x}_{t+1} &= x_{t+1} - \widehat{\mathrm{E}}\left[x_{t+1}|\hat{y}_t, \hat{y}_{t-1}, \cdots, \hat{y}_1, \tilde{x}_1\right] \\ &= x_{t+1} - \widehat{\mathrm{E}}_t x_{t+1}\end{aligned}$$

- Let $\Sigma_{\hat{x}_t} = \mathrm{E}\hat{x}_t\hat{x}_t'$

$$\mathrm{cov}(x_{t+1}, \hat{y}_t) = A\Sigma_{\hat{x}_t}C' + GV_3$$

$$\mathrm{cov}(\hat{y}_t, \hat{y}_t) = C\Sigma_{\hat{x}_t}C' + V_2$$

Using the derived expressions

$$\widehat{\mathsf{E}}\left[x_{t+1}|\hat{y}_t\right]$$

$$= \mathsf{E}x_{t+1} + \mathrm{cov}(x_{t+1}, \hat{y}_t)\left[\mathrm{cov}(\hat{y}_t, \hat{y}_t)\right]^{-1}\hat{y}_t$$

$$= \mathsf{E}x_{t+1} + \left(A\Sigma_{\hat{x}_t}C' + GV_3\right)\left(C\Sigma_{\hat{x}_t}C' + V_2\right)^{-1}\hat{y}_t \tag{4}$$

Derivation Kalman filter

- Now get an expression for the second term in (3).
- From $x_{t+1} = Ax_t + Gw_{1,t+1}$, we get

$$\widehat{\mathsf{E}}\left[x_{t+1}|\hat{Y}^{t-1},\tilde{x}_1\right] = A\widehat{\mathsf{E}}\left[x_t|\hat{Y}^{t-1},\tilde{x}_1\right] = A\widehat{\mathsf{E}}_{t-1}x_t \qquad (5)$$

Using (4) and (5) in (3) gives the *recursive* expression

$$\widehat{\mathsf{E}}_t x_{t+1} = A\widehat{\mathsf{E}}_{t-1} x_t + K_t \hat{y}_t$$

where

$$K_t = \left(A\Sigma_{\hat{x}_t} C' + GV_3\right)\left(C\Sigma_{\hat{x}_t} C' + V_2\right)^{-1}$$

Prediction for observable

From

$$y_{t+1} = Cx_{t+1} + w_{2,t+1}$$

we get

$$\widehat{\mathsf{E}}\left[y_{t+1} | Y_t, \tilde{x}_1\right] = C\widehat{\mathsf{E}}_t x_{t+1}$$

Thus

$$\hat{y}_{t+1} = y_{t+1} - C\widehat{\mathsf{E}}_t x_{t+1}$$

Updating the covariance matrix

- We still need an equation to update $\Sigma_{\hat{x}_t}$. This is actually not that hard. The result is

$$\Sigma_{\hat{x}_{t+1}} = A\Sigma_{\hat{x}_t}A' + GV_1G' - K_t(A\Sigma_{\hat{x}_t}C' + GV_3)'$$

- Expression is deterministic and does not depend particular realizations. That is, precision only depends on the coefficients of the time series model

Applications Kalman filter

- signal extraction problems
  - GPS, computer vision applications, missiles
- prediction
- simple alternative to calculating inverse policy functions
  - (see below)

Estimating DSGE models

- Forget the Kalman filter for now, we will not use it for a while
- What is next?
  - Specify the neoclassical model that will be used as an example
  - Specify the linearized version
  - Specify the estimation problem
  - Maximum Likelihood estimation
  - Explain why Kalman filter is useful
  - Bayesian estimation
  - MCMC, a necessary tool to do Bayesian estimation

Neoclassical growth model

First-order conditions

$$c_t^{-\nu} = \mathsf{E}_t \left[\beta c_{t+1}^{-\nu} \left(\alpha z_{t+1} k_t^{\alpha-1} + 1 - \delta \right) \right]$$

$$c_t + k_t = z_t k_{t-1}^{\alpha} + (1-\delta) k_{t-1}$$

$$z_t = (1-\rho) + \rho z_{t-1} + \varepsilon_t$$

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

Linearized solution

$$
\begin{aligned}
k_t &= \bar{k} + a_{k,k}(k_{t-1} - \bar{k}) + a_{k,z}(z_t - \bar{z}) \\
z_t &= (1-\rho) + \rho z_{t-1} + \varepsilon_t \\
&\quad \varepsilon_t \sim N(0, \sigma^2) \\
&\quad z_0 \sim N(1, \sigma^2/(1-\rho^2)) \\
&\quad k_0 \text{ is given}
\end{aligned}
$$

- $a_{k,k}$, $a_{k,z}$, and $\bar{k}$ are *known* functions of the structural parameters $\Longrightarrow$ better notation would be $a_{k,k}(\Psi)$, $a_{k,z}(\Psi)$, and $\bar{k}(\Psi)$
- Consumption has been substituted out
- Approximation error is ignored. Linearized model is treated as the true model with Ψ as the parameters

Estimation problem

Given data for capital, $\{k_t\}_0^T$, estimate the set of coefficients, Ψ

$$\Psi = [\alpha, \beta, \nu, \delta, \rho, \sigma, z_0]$$

- No data on productivity, z_t.
  - If you had data on $z_t \Longrightarrow$ Likelihood $= 0$ for sure
  - More on this below.

Formulation of the Likelihood

- Let Y^T be the complete sample

$$L(Y^T|\Psi) = p(z_0)\prod_{t=1}^{T} p(z_t|z_{t-1})$$

$p(z_t|z_{t-1})$ corresponds with probability of a particular value for ε_t

Formulation of the Likelihood

Basic idea:

- Given a value for Ψ and give the data set, Y^T, you can calculate the implied values for ε_t
- We know the distribution of $\varepsilon_t \Longrightarrow$
- We can calculate the probability (likelihood) of $\{\varepsilon_1, \cdots, \varepsilon_T\}$

Formulation of the Likelihood

$$k_t = \bar{k} + a_{k,k}(k_{t-1} - \bar{k}) + a_{k,z}(z_t - \bar{z})$$

$$\Longrightarrow$$

$$\begin{aligned}
z_t &= \frac{a_{k,z}\bar{z} - \bar{k} + a_{k,k}\bar{k}}{a_{k,z}} - \frac{a_{k,k}}{a_{k,z}}k_{t-1} + \frac{1}{a_{k,z}}k_t \\
z_t &= b_0 + b_1 k_{t-1} + b_2 k_t
\end{aligned}$$

$$\varepsilon_t = z_t - (1-\rho) - \rho z_{t-1}$$

Formulation of the Likelihood

- ε_t is obtained by **inverting** the policy function
- For larger systems, this inversion is not as easy to implement.
  - Below, we show an alternative

Formulation of the Likelihood

A bit more explicit

- Take a value for Ψ
- Given k_0 and k_1 you can calculate z_1
- Given z_0 you can calculate ε_1
- Continuing, you can calculate $\varepsilon_t \; \forall t$
- To make explicit the dependence of ε_t on Ψ, write $\varepsilon_t(\Psi)$
- The Likelihood can thus be written as

$$\prod_{t=1}^{T} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-\left(\varepsilon_t(\Psi)\right)^2}{2\sigma^2} \right\}$$

Too few unobservables & singularities

- Above we assumed that there was no data on z_t
- Suppose you had data on z_t

- There are two cases to consider
  - Data not exactly generated by this model (most likely case)
    $\Longrightarrow$ Likelihood $= 0$ for any value of Ψ
  - Data is exactly generated by this model
    $\Longrightarrow$ Likelihood $= 1$ for true value of Ψ *and*
    $\Longrightarrow$ Likelihood $= 0$ for any other value for Ψ

Too few unobservables & singularities

$$k_t = \bar{k} + a_{k,k}(k_{t-1} - \bar{k}) + a_{k,z}(z_t - \bar{z})$$

Using the values for 4 periods, you can pin down $\bar{k}$, $\bar{z}$, $a_{k,k}$, and $a_{k,z}$.

- What about values for additional periods?
  - Data generated by model (unlikely of course)
    $\Longrightarrow$ additional observations will fit this equation too
  - Data not generated by model
    $\Longrightarrow$ additional observations will not fit this equation
    $\Longrightarrow$ Likelihood = zero

Too few unobservables & singularities

- Can't I simply add an error term?

$$k_t = \bar{k} + a_{k,k}(k_{t-1} - \bar{k}) + a_{k,z}(z_t - \bar{z}) + u_t$$

- Answer: **NO** not in general
- Why not? It is ok in standard regression

Too few unobservables & singularities

Why is the answer NO in general?

1. u_t represents other shocks such as preference shocks
   $\Longrightarrow$ it's presence is likely to affect $\bar{k}$, $a_{k,k}$, and $a_{k,z}$
2. u_t represents measurement error
   $\Longrightarrow$ you are fine from an econometric stand point
   $\Longrightarrow$ but is residual only measurement error?

What if you also observe consumption?

Suppose you observe k_t, c_t, but not z_t?

$$\begin{aligned} k_t &= \bar{k} + a_{k,k}(k_{t-1} - \bar{k}) + a_{k,z}(z_t - \bar{z}) \\ c_t &= \bar{c} + a_{c,k}(k_{t-1} - \bar{k}) + a_{c,z}(z_t - \bar{z}) \end{aligned}$$

- Recall that the coefficients are functions of Ψ
- Given value of Ψ you can solve for z_t from top equation
- Given value of Ψ you can solve for z_t from bottom equation
- With real world data you will get inconsistent answers.

Unobservables and avoiding singularities

General rule:

- For every observable you need at least one unobservable shock
- Letting them be measurement errors is hard to defend
- The last statement does not mean that you cannot *also* add measurement errors

Using the Kalman filter

$$x_{t+1} = Ax_t + Gw_{1,t+1} \tag{6}$$

$$y_t = Cx_t + w_{2,t} \tag{7}$$

- (6) describes the equations of the model;
    - x_t consists of the "true" values of state variables like capital and productivity.
- (7) relates the observables, y_t, to the "true" values

Example

- consumption and capital are observed with error
  - $c_t^* = c_t + u_{c,t}$
  - $k_t^* = k_t + u_{k,t}$
- z_t is unobservable

- $x_t' = [k_{t-1} - \bar{k}, z_{t-1} - \bar{z}]$
- $w_{1,t+1} = \varepsilon_t$
- $y_t' = [k_{t-1}^* - \bar{k}, c_t^* - \bar{c}]$

Example

- (6) gives policy function for k_t and law of motion for z_t

$$\begin{bmatrix} k_t - \bar{k} \\ c_t - \bar{c} \\ z_{t+1} - \bar{z} \end{bmatrix} = \begin{bmatrix} a_{k,k} & a_{k,z} \\ a_{c,k} & a_{c,z} \\ 0 & \rho \end{bmatrix} \begin{bmatrix} k_{t-1} - \bar{k} \\ z_t - \bar{z} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \varepsilon_{t+1} \end{bmatrix}$$

- Equation (7) is equal to

$$\begin{bmatrix} k^*_{t-1} - \bar{k} \\ c^*_t - \bar{c} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ a_{c,k} & a_{c,z} \end{bmatrix} \begin{bmatrix} k_{t-1} - \bar{k} \\ z_t - \bar{z} \end{bmatrix} + \begin{bmatrix} u_{k,t} \\ u_{c,t} \end{bmatrix}$$

Back to the Likelihood

- y_t consists of k_t^* and c_t^* and the model is given by (6) and (7).
- From the Kalman filter we get $\hat{y}_t$ and $\Sigma_{\hat{y}_t}$

$$
\begin{aligned}
\widehat{\mathsf{E}}\left[x_t|Y^{t-1},\tilde{x}_1\right] &= A\widehat{\mathsf{E}}\left[x_{t-1}|Y^{t-2},\tilde{x}_1\right] + K_{t-1}\hat{y}_{t-1} \\
\widehat{\mathsf{E}}\left[y_t|Y^{t-1},\tilde{x}_1\right] &= C\widehat{\mathsf{E}}\left[x_t|Y^{t-1},\tilde{x}_1\right] \\
\hat{y}_t &= y_t - \widehat{\mathsf{E}}\left[y_t|Y^{t-1},\tilde{x}_1\right] \\
\Sigma_{\hat{x}_{t+1}} &= A\Sigma_{\hat{x}_t}A' + GV_1G' - K_t(A\Sigma_{\hat{x}_t}C + GV_3)' \\
\Sigma_{\hat{y}_t} &= C\Sigma_{\hat{x}_t}C' + V_2
\end{aligned}
$$

Back to the Likelihood

- $\hat{y}_{t+1}$ is normally distributed because
  - this is a linear model and underlying shocks are linear
- Kalman filter generates $\hat{y}_{t+1}$ and $\Sigma_{\hat{y}_t}$
  - (given Ψ and observables, Y^T)
- Given normality calculate likelihood of $\{\hat{y}_{t+1}\}$

Kalman Filter versus inversion

with measurement error

- have to use Kalman filter

withour measurement error

- could back out shocks using inverse of policy function
- but could also use Kalman filter
  - Dynare always uses the Kalman filter
  - hardest part of the Kalman filter is calculating the inverse of $C\Sigma_{\hat{x}_t}C' + V_2$ and this is typically not a difficult inversion.

Log-Likelihood

$$\begin{aligned}\ln(Y^T|\Psi) &= -\left(\frac{1}{2}\right)\left(n_x \ln(2\pi) + \ln(|\Sigma_{\widehat{x}_0}|) + \widehat{x}_0' \Sigma_{\widehat{x}_0}^{-1} \widehat{x}_0\right) \\ &\quad -\left(\frac{1}{2}\right)\left(T n_y \ln(2\pi) + \sum_{t=1}^{T}\left[\ln(|\Sigma_{\widehat{y}_t}|) + \widehat{y}_t' \Sigma_{\widehat{y}_t}^{-1} \widehat{y}_t\right]\right)\end{aligned}$$

n_y : dimension of $\hat{y}_t$

For the neo-classical growth model

- Start with $x_1 = [k_0, z_0]$, $y_1 = k_0^*$, and Σ_1
- Calculate

$$\begin{aligned}\hat{y}_1 &= y_1 - \widehat{\mathsf{E}}\left[y_1 | x_1\right] \\ &= y_1 - Cx_1\end{aligned}$$

- Calculate $\widehat{\mathsf{E}}\left[x_2 | y_1, x_1\right]$ using

$$\widehat{\mathsf{E}}_t x_{t+1} = A\widehat{\mathsf{E}}_{t-1} x_t + K_t \hat{y}_t$$

where

$$K_t = \left(A\Sigma_{\hat{x}_t} C' + GV_3\right)\left(C\Sigma_{\hat{x}_t} C' + V_2\right)^{-1}$$

For the neo-classical growth model

- Calculate

$$\begin{aligned}\hat{y}_2 &= y_2 - \widehat{\mathsf{E}}\left[y_2 | y_1, x_1\right] \\ &= y_2 - C\widehat{\mathsf{E}}\left[x_2 | y_1, x_1\right]\end{aligned}$$

- etc.

Bayesian Estimation

- Conceptually, things are not that different
- Bayesian econometrics combines
  - the likelihood, i.e., the data, with
  - the prior
- You can think of the prior as additional data

Posterior

The joint density of parameters and data is equal to

$$P(Y^T, \Psi) = L(Y^T|\Psi)P(\Psi) \quad \text{or}$$

$$P(Y^T, \Psi) = P(\Psi|Y^T)P(Y^T)$$

Posterior

From this we can get Bayes rule: $P(\Psi|Y^T) = \frac{L(Y^T|\Psi)P(\Psi)}{P(Y^T)}$

Reverend Thomas Bayes (1702-1761)

Posterior

- For the distribution of Ψ, $P(Y^T)$ is just a constant.
- Therefore we focus on

$$L(Y^T|\Psi)P(\Psi) \propto \frac{L(Y^T|\Psi)p(\Psi)}{P(Y^T)} = P(\Psi|Y^T)$$

- One can always make $L(Y^T|\Psi)P(\Psi)$ a proper density by scaling it so that it integrates to 1

Evaluating the posterior

- Calculating posterior for given value of Ψ not problematic.
- But we are interested in objects of the following form

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

- Examples
  - to calculate the mean of Ψ, let $g(\Psi) = 1$
  - to calculate the probability that $\Psi \in \Psi^*$,
    - let $g(\Psi) = 1$ if $\Psi \in \Psi^*$ and
    - let $g(\Psi) = 0$ otherwise
  - to calculate the posterior for j^{th} element of Ψ
    - $g(\Psi) = \Psi_j$

Evaluating the posterior

- Even *Likelihood* can typically only be evaluated numerically
- Numerical techniques also needed to evaluate the *posterior*

Evaluating the posterior

- Standard Monte Carlo integration techniques cannot be used
  - Reason: cannot *draw* random numbers directly from $P(\Psi|Y^T)$
  - being able to calculate $P(\Psi|Y^T)$ not enough to create a random number generator with that distribution
- Standard tool: Markov Chain Monte Carlo (MCMC)

Metropolis & Metropolis-Hasting

- Metropolis & Metropolis-Hasting are particular versions of the MCMC algorithm
- Idea:
  - travel through the state space of Ψ
  - weigh the outcomes appropriately

Metropolis & Metropolis-Hasting

- Start with an initial value, Ψ_0
  - discard the beginning of the sample, the burn-in phase, to ensure choice of Ψ_0 does not matter

Metropolis & Metropolis-Hasting

Subsequent values, Ψ_{i+1}, are obtained as follows

- Draw Ψ^* using the "stand in" density $f(\Psi^*|\Psi_i, \theta_f)$
  - θ_f contains the parameters of $f(\cdot)$
- Ψ^* is a *candidate* for Ψ_{i+1}
  - $\Psi_{i+1} = \Psi^*$ with probability $q(\Psi_{i+1}|\Psi_i)$
  - $\Psi_{i+1} = \Psi_i$ with probability $1 - q(\Psi_{i+1}|\Psi_i)$

Metropolis & Metropolis-Hasting

properties of $f(\cdot)$

- $f(\cdot)$ should have fat tails relative to the posterior
  - that is, $f(\cdot)$ should "cover" $P(\Psi|Y^T)$

Metropolis (used in Dynare)

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

- $P(\Psi^*|Y^T) \geq P(\Psi_i|Y^T) \Longrightarrow$
  - always include candidate as new element
- $P(\Psi^*|Y^T) < P(\Psi_i|Y^T) \Longrightarrow$
  - Ψ^* not always included; the lower $P(\Psi^*|Y^T)$ the lower the chance it is included

Metropolis-Hasting

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

- $P(\Psi_i|Y^T)/f(\Psi_i|\Psi_i,\theta_f)$ low $\Longrightarrow$
  - you should move away from this Ψ value $\Longrightarrow$ q should be high
- $P(\Psi^*|Y^T)/f(\Psi^*|\Psi_i,\theta_f)$ high:
  - probability of Ψ^* high & should be included with high prob.

Choices for f(.)

- Random walk MH:

$$\Psi^* = \Psi_i + \varepsilon \text{ with } \mathrm{E}\left[\varepsilon\right] = 0$$

  - and, for example,

$$\varepsilon \sim N(0, \theta_f^2)$$

- Independence sampler:

$$f(\Psi^* | \Psi_i, \theta_f) = f(\Psi^* | \theta_f)$$

Couple more points

- Is the singularity issue different with Bayesian statistics?
- Choosing prior
- Gibbs sampler

The singularity problem again

What happens in practice?

- lots of observations are available
- practioners don't want to exclude data $\Longrightarrow$
- add "structural" shocks

The singularity problem again

Problem with adding additional shocks

- measurement error shocks
    - not credible that this is reason for gap between model and data
- structural shocks
    - good reason, but wrong structural shocks $\Longrightarrow$ misspecified model

Possible solution to singularity problem?

Today's posterior is tomorrow's prior

Possible solution to singularity problem?

Suppose you want the following:

- use 2 observables and
- only 1 structural shock

Possible solution to singularity problem?

1. Start with first prior: $P_1(\Psi)$
2. Use first observable Y_1^T to form first posterior

$$F_1(\Psi) = L(Y_1^T|\Psi)P_1(\Psi)$$

3. Let second prior be first posterior: $P_2(\Psi) = F_1(\psi)$
4. Use second observable Y_2^T to form second posterior

$$F_2(\Psi) = L(Y_2^T|\Psi)P_2(\Psi)$$

Final answer:

$$\begin{aligned} F_2(\Psi) &= L(Y_2^T|\Psi)P_2(\Psi) \\ &= L(Y_2^T|\Psi)L(Y_1^T|\Psi)P_1(\Psi) \end{aligned}$$

Obviously:

$$\begin{aligned} F_2(\Psi) &= L(Y_2^T|\Psi)L(Y_1^T|\Psi)P_1(\Psi) \\ &= L(Y_1^T|\Psi)L(Y_2^T|\Psi)P_1(\Psi) \end{aligned}$$

Thus, it does not matter which variable you use first

Properties of final posterior

- Final posterior could very well have multiple modes
  - indicates where different variables prefer parameters to be
- This is only informative, not a disadvantage

Have we solved the singularity problem?

Problems of approach:

- Procedure avoids singularity problem by not considering *joint* implications of two observables
- Procdure misses some structural shock/misspecification

Key question:

- Is this worse than adding bogus shocks?

Have we solved the singularity problem?

Problems of approach:

- Procedure avoids singularity problem by not considering *joint* implications of two observables
- Procdure misses some structural shock/misspecification

Key question:

- Is this worse than adding bogus shocks?

How to choose prior

1. Without analyzing data, sit down and think
   problem in macro: we keep on using the same data
   so is this science or data mining?

2. Don't change prior depending on results

Uninformative prior

- $P(\Psi) = 1 \;\; \forall \Psi \in \mathbb{R} \Longrightarrow$ posterior = likelihood
- $P(\Psi) = 1/(b-a)$ if $\Psi \in [a,b]$ is not **un**informative
- Which one is the least informative prior?

$$P(\Psi) = 1/(b-a) \;\text{ if } \Psi \in [a,b]$$
$$P(\ln\Psi) = 1/(\ln b - \ln a) \;\text{ if } \Psi \in [\ln a, \ln b]$$

Uninformative prior

- $P(\Psi) = 1 \;\; \forall \Psi \in \mathbb{R} \Longrightarrow$ posterior = likelihood
- $P(\Psi) = 1/(b-a)$ if $\Psi \in [a, b]$ is not **un**informative
- Which one is the least informative prior?

$$P(\Psi) = 1/(b-a) \;\text{ if } \Psi \in [a, b]$$
$$P(\ln \Psi) = 1/(\ln b - \ln a) \;\text{ if } \Psi \in [\ln a, \ln b]$$

Uninformative prior

- $P(\Psi) = 1 \ \ \forall \Psi \in \mathbb{R} \Longrightarrow$ posterior $=$ likelihood
- $P(\Psi) = 1/(b-a)$ if $\Psi \in [a, b]$ is not **un**informative
- Which one is the least informative prior?

$$P(\Psi) = 1/(b-a) \ \text{ if } \Psi \in [a, b]$$
$$P(\ln \Psi) = 1/(\ln b - \ln a) \ \text{ if } \Psi \in [\ln a, \ln b]$$

The objective of Jeffrey's prior is to ensure that the prior is *invariant* to such reparameterizations

How to choose (not so) informative priors

Let the prior inherit invariance structure of the problem:

1. **location parameter:** If X is distributed as $f(x - \psi)$, then $Y = X + \phi$ have the same distribution but a different location. If the prior has to inherit this property, then it should be uniform.
2. **scale parameter:** If X is distributed as $(1/\sigma)f(x/\sigma)$, then $Y = \phi X$ has the same distribution as X except for a different scale parameter. If the prior has to inherit this property, then it should be of the form

$$P(\psi) = 1/\psi$$

Both are improper priors.
That is, they do not integrate to a finite number.

Not so informative priors

Let the prior be consistent with "total confusion"

3. **probability parameter:** If ψ is a probability $\in [0, 1]$, then the prior distribution

$$P(\psi) = 1/\left(\psi\left(1 - \psi\right)\right)$$

represents total confusion. The idea is that the elements of the prior correspond to different beliefs and everybody is given a new piece of info that the cross-section of beliefs would not change.
See notes by Smith

Gibbs sampler

Objective: Obtain T observations from $p(x_1, \cdots, x_J)$.
Procedure:

1. Start with initial observation $X^{(0)}$.
2. Draw period t observation, $X^{(t)}$, using the following iterative scheme:
    - draw $x_j^{(t)}$ from the conditional distribution:
    $p\left(x_j | x_1^{(t)}, \cdots, x_{j-1}^{(t)}, x_{j+1}^{(t-1)}, \cdots, x_J^{(t-1)}\right)$

Gibbs sampler versus MCMC

- Gibbs sampler does not require stand-in distribution
- Gibbs sampler still requires the ability to draw from conditional
  $\Longrightarrow$ not useful for estimation DSGE models

References

- Chib, S. and Greenberg, E., 1995, Understanding the Metropolis-Hastings Algorithm, The American Statistician.
  - describes the basics
- Ljungqvist, L. and T.J. Sargent, 2004, Recursive Macroeconomic Theory
  - source for the description of the Kalman filter
- Roberts, G.O., and J.S. Rosenthal, 2004, General state space Markov chains and MCMC algorithms, Probability Surveys.
  - more advanced articles describing formal properties

References

- Smith, G.P., Expressing Prior Ignorance of a Probability Parameter, notes, University of Missouri
  http://www.stats.org.uk/priors/noninformative/Smith.pdf
  on informative priors
- Syversveen, A.R, 1998, Noninformative Bayesian priors. Interpretation and problems with construction and applications
  http://www.stats.org.uk/priors/noninformative/Syversveen1998.pdf
  on informative priors