An Introduction to Bayesian econometrics

Bao H. Nguyen

School of Economics – UEH &
School of Business and Economics – University of Tasmania

January 8, 2020

Introduction

- Suppose we are interested in estimating the following regression model:

$$y_t = x_t\beta + \epsilon_t, \tag{1}$$

$$\epsilon \sim N(0, \sigma^2), \tag{2}$$

- where y_t is a $T \times 1$ matrix of dependent variable and x_t is a $T \times K$ matrix of the independent variables.
- We are concerned with estimating the $K \times 1$ vector of coefficients β and the variance of the error term σ^2

Introduction

- A classical econometrician (frequentist) obtains $\hat{\beta}$ and $\hat{\sigma}^2$ by maximising the likelihood function:

$$F(y_t|\beta,\sigma^2) = (2\pi\sigma^2)^{-T/2} exp\Big(-\frac{(y_t-\beta x_t)\prime(y_t-\beta x_t)}{2\sigma^2}\Big) \quad (3)$$

- In this case, we get the familiar OLS estimator for the coefficients and the error variance:

$$\hat{\beta}_{OLS} = (x\prime_t x_t)^{-1}(x\prime_t y_t), \quad (4)$$

$$\hat{\sigma}^2_{OLS} = \frac{\epsilon\prime_t \epsilon_t}{T} \quad (5)$$

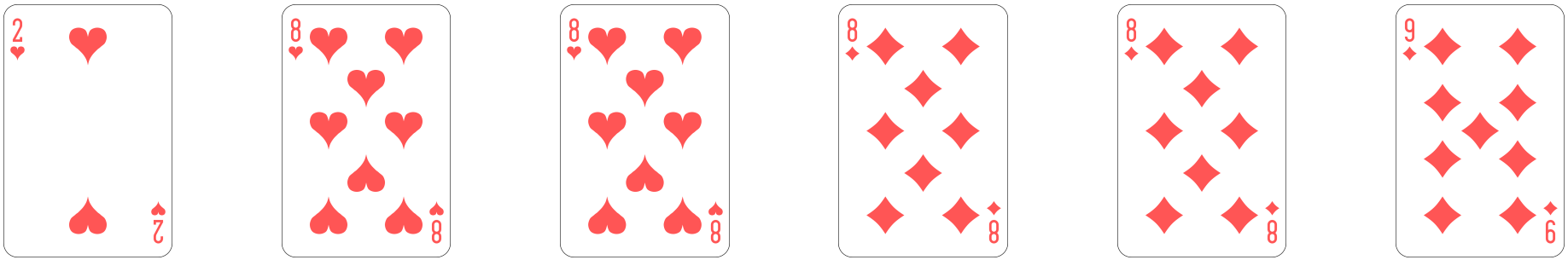
Classical vs. Bayesian analysis

- The estimates of the parameters of the model are solely based on information contained in data.
- β_{OLS} is a fixed, unknown quantity $\rightarrow$ "true value".
- The estimator $\hat{\beta}_{OLS}$ is a random variable an is evaluated via repeated sampling $\rightarrow$ the interval constructed will contain the true value in 95% of cases if we estimate for thousand different samples taken from a population.
- The estimator $\hat{\beta}_{OLS}$ is "best" in the sense of having the highest probability of being close to the true β_{OLS}.

Classical vs. Bayesian analysis

- Bayesian analysis allows the researcher to incorporate her prior beliefs about the parameters β and σ^2. In other words, she treats β and σ^2 as random variables and have a probability distribution
- The distribution summarizes our knowledge about the model parameter, reflecting two sources of information:
  - (1) Prior information (before seeing the data): subjective belief about how likely different parameter values are;
  - (2) Sample information: leads researcher to revise/update his prior beliefs
- How she does:
  - Step 1: Forms prior beliefs (PRIORS)
  - Step 2: Collect data and write down the LIKELIHOOD FUNCTION of the model
  - Step 3: Combines the prior distributions and the likelihood function to obtain the POSTERIORS

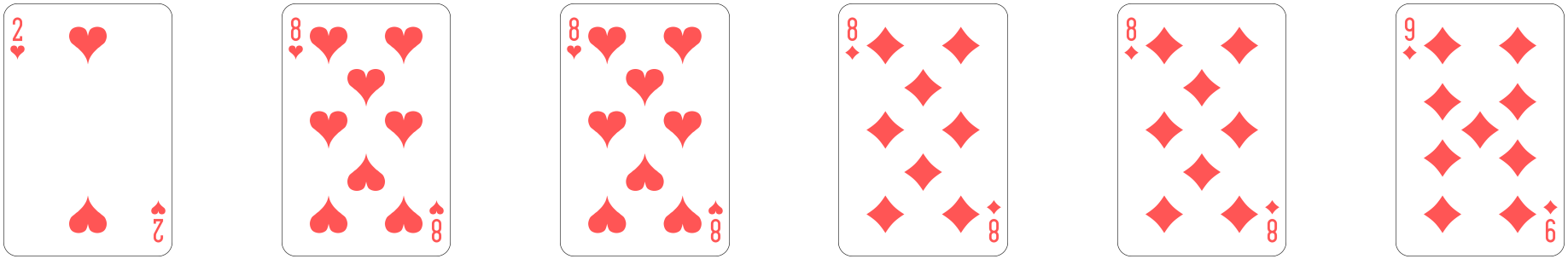
From Bayes' rule to Bayesian econometrics



$$P(2) = \frac{1}{6} \quad P(\heartsuit) = 1/2 \quad P(\heartsuit|2 = 1) \quad P(2|\heartsuit = \frac{1}{3})$$

$$P(2|\heartsuit) = P(2).\frac{P(\heartsuit|2)}{P(\heartsuit)} = \frac{1}{6}.\frac{1}{1/2} = \frac{1}{3}$$

From Bayes' rule to Bayesian econometrics



$$P(\diamondsuit|8) = P(\diamondsuit).\frac{P(8|\diamondsuit)}{P(8)}$$

$$\frac{1}{2} = \frac{1}{2}.\frac{2/3}{2/3}$$

$$P(9|\diamondsuit) = P(9).\frac{P(\diamondsuit|9)}{P(\diamondsuit)}$$

$$\frac{1}{3} = \frac{1}{6}.\frac{1}{1/2}$$

$$P(A|data) = P(A).\frac{P(data|A)}{P(data)}$$

From Bayes' rule to Bayesian econometrics

Let's map this rule into a simple regression model where we want to learn about a parameter $\boldsymbol{\theta} = \{B, \sigma\}$ given the data $\boldsymbol{y} = \{Y, X\}$:

$$P(\boldsymbol{\theta}|\boldsymbol{y}) = P(\boldsymbol{\theta}).\frac{P(\boldsymbol{y}|\boldsymbol{\theta})}{P(\boldsymbol{y})}$$

- the key object of interest: the posterior density $P(\boldsymbol{\theta}|\boldsymbol{y})$.
- prior density: $P(\boldsymbol{\theta})$. It does not depend on the data y; instead, it contains non-data information about $\boldsymbol{\theta}$
- likelihood function: $P(\boldsymbol{y}|\boldsymbol{\theta})$. It is the density of the data conditional on the parameters.
- marginal data density: $P(\boldsymbol{y})$. Since we are interested in learning about $\boldsymbol{\theta}$, we can ignore $P(\boldsymbol{y})$ since it does not involve $\boldsymbol{\theta}$.

From Bayes' rule to Bayesian econometrics

$$P(\boldsymbol{\theta}|\boldsymbol{y}) \propto P(\boldsymbol{\theta}).P(\boldsymbol{y}|\boldsymbol{\theta})$$

"The posterior is proportional to the likelihood times the prior."

- The posterior summarizes all we know about $\boldsymbol{\theta}$ after seeing the data. In other words, the posterior combines both data and non-data information.
- The equation can be viewed as an updating rule where the data allow us to update our prior views about $\boldsymbol{\theta}$.

Case 1: Normal model with known variance

Suppose the our goal is to obtain the posterior distribution $p(\mu|\boldsymbol{y})$ given the sample $\boldsymbol{y} = (y_1, \ldots, y_N)\prime$. To that end, we need:

- a likelihood function
- a prior for the parameter μ.

This is a normal model

$$(y_n|\mu) \sim N(\mu, \sigma^2), \quad n = 1, \ldots, N, \tag{6}$$

where the **variance σ^2 is assumed to be known**.

- A reasonable **prior** would be

$$\mu \sim N(\mu_0, \sigma_0^2), \tag{7}$$

  where both μ_0 and σ_0^2 are also known.
- the posterior distribution, which can be obtained by Bayes' theorem:

$$p(\mu|\boldsymbol{y}) = \frac{p(\mu)p(\boldsymbol{y}|\mu)}{p(\boldsymbol{y})}$$

Case 1: Normal model with known variance

It turns out that $p(\mu|\boldsymbol{y})$ is a Gaussian distribution. Recall that we say a random variable X follows a **normal** or **Gaussian distribution**, and we write $X \sim N(a, b^2)$, if its density is given by

$$\boxed{f(x, a, b^2) = (2\pi b^2)^{-\frac{1}{2}} e^{-\frac{1}{2b^2}(x-a)^2}} \tag{8}$$

It follows from (6), that the **likelihood function** is a product of N normal densities:

$$\begin{aligned} p(\mu|\boldsymbol{y}) &= \prod_{n=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_n-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\mu)^2} \end{aligned} \tag{9}$$

Similarly, the **prior** density $p(\mu)$ is given by

$$p(\mu) = (2\pi\sigma_0^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} \tag{10}$$

Case 1: Normal model with known variance

Next, we combine the likelihood (9) and the prior (10) to obtain the posterior distribution:

$$
\begin{aligned}
p(\mu|\boldsymbol{y}) &\propto p(\mu)p(\boldsymbol{y}|\mu) \\
&\propto e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\mu)^2} \\
&\propto exp\left[-\frac{1}{2}\left(\frac{\mu^2-2\mu\mu_0}{\sigma_0^2}\right)+\left(\frac{N\mu^2-2\mu\sum_{n=1}^{N}y_n}{\sigma^2}\right)\right] \\
&\propto exp\left[-\frac{1}{2}\left(\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)\right)\mu^2-2\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{N\bar{y}}{\sigma^2}\right)\right], \quad (11)
\end{aligned}
$$

where $\bar{y}=N^{-1}\sum_{n=1}^{N}y_n$ is the sample mean and ignore any constants that do not involve μ.

Case 1: Normal model with known variance

Since $p(\mu|\boldsymbol{y})$ is Gaussian, and it has the same family as the prior. In this case, the prior is called a **conjugate prior** for the likelihood function. We next determine the mean and variance of the distribution.

Now, suppose $(\mu|\boldsymbol{y}) \sim N(\hat{\mu}, D_\mu)$ and using the definition of the Gaussian density, we can rewrite the posterior distribution as

$$\begin{aligned} p(\mu|\boldsymbol{y}) &= (2\pi D_\mu)^{-\frac{1}{2}} e^{-\frac{1}{2D_\mu}(\mu-\hat{\mu})^2} \\ &\propto e^{-\frac{1}{2}(\frac{1}{D_\mu}\mu^2 - 2\mu\frac{\hat{\mu}}{D_\mu})} \end{aligned} \tag{12}$$

Now compare this expression with (11), they are identical for any $\mu \in R$, thus the coefficients must be the same, i.e.,

$$D_\mu = \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1} \qquad \frac{\hat{\mu}}{D_\mu} = \frac{\mu_0}{\sigma_0^2} + \frac{N\bar{y}}{\sigma^2}$$

Case 1: Normal model with known variance

Finally, we can write the posterior mean as

$$\begin{aligned}\hat{\mu} &= \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{N\bar{y}}{\sigma^2}\right) \\ &= \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}\mu_0 + \frac{\frac{N}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}\bar{y}\end{aligned}$$

The posterior mean is a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$, where the weights are, respectively, the inverse of the prior variance σ_0^2 and the inverse of the sample mean variance σ^2/N.

Numerical example 1

Suppose $\mu_0 = 10, \bar{y} = 20, \sigma_0^2 = \sigma^2 = 1$. We consider two cases: $N = 1$ an $N = 10$.

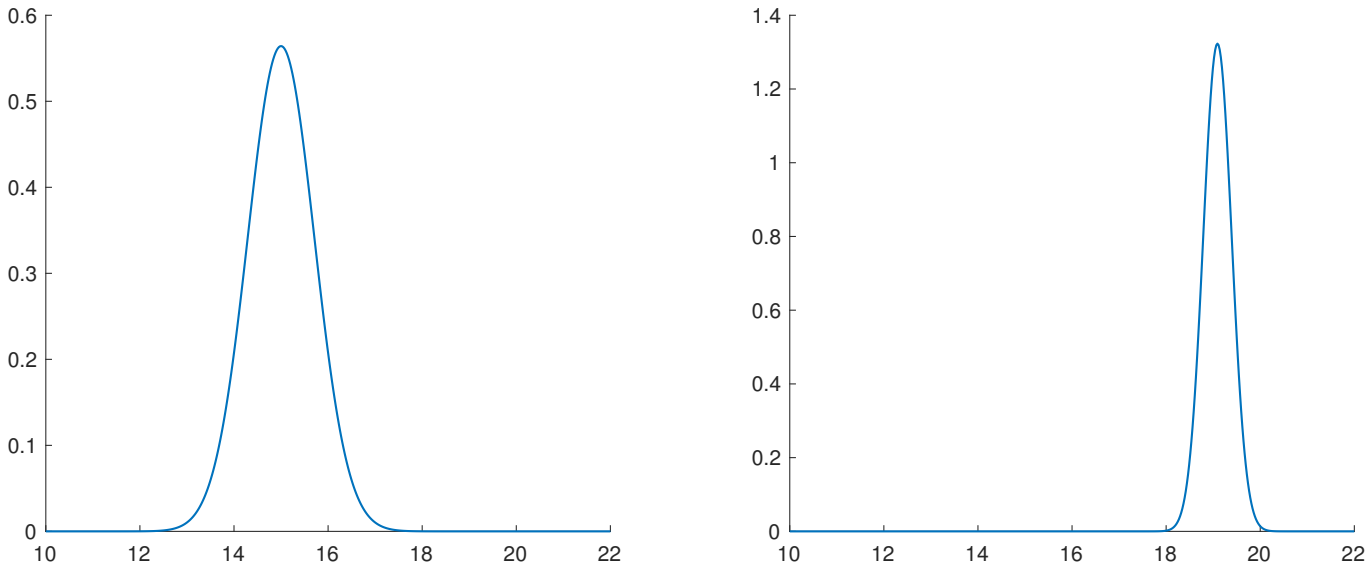


Figure: Posterior densities of μ for $N = 1$ (left panel) and $N = 10$ (right panel).

Case 2: Normal model with unknown variance

Now we extend the model to allow the variance σ^2 to be unknown. The model now is

$$(y_n|\mu, \sigma^2) \sim N(\mu, \sigma^2), \quad n = 1, \ldots, N, \tag{13}$$

where both μ and σ^2 are unknown. Assuming the same prior for μ, e.g., $\mu \sim N(\mu_0, \sigma_0^2)$. As for σ^2, which takes only positive values, a convenient prior is the inverse-gamma prior. With the prior and the corresponding likelihood, we can derive the joint posterior distribution $p(\mu, \sigma^2|\boldsymbol{y})$.

However, it is not obvious how we can compute analytically various quantities of interest, such as, $E(\mu|y)$, the posterior mean of μ or $Var(\sigma^2|y)$, the posterior variance of σ^2.

Solution: Markov Chain Monte Carlo (MCMC) simulation is used to approximate those quantities, called **Gibbs sampling**.

Case 2: Normal model with unknown variance

To construct a Gibbs sampler to draw from the posterior distribution $p(\mu, \sigma^2|\boldsymbol{y})$:

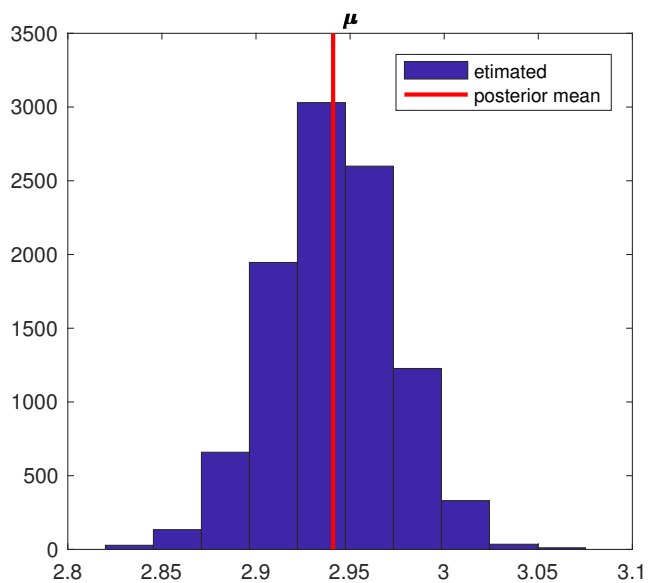
- derive two **conditional distributions**: $p(\mu|\boldsymbol{y}, \sigma^2)$ and $p(\sigma^2|\boldsymbol{y}, \mu)$. See appendix 1.
- Gibbs sampler
  - First step: We have to initialise the chain by setting both β an σ^2 to some initial values, e.g $\mu^{(0)}$ an $\sigma^{2(0)}$
  - Second step: We draw

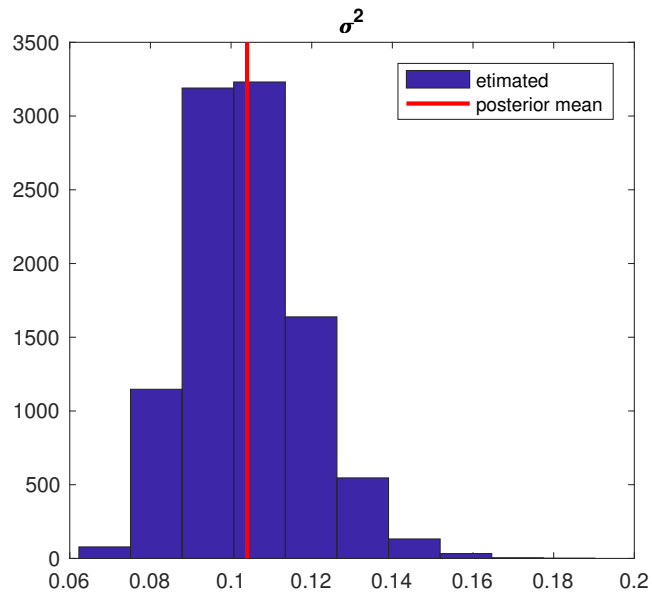
$$\mu^{(1)} \propto p(\mu|y, \sigma^{2(0)})$$
$$\sigma^{2(1)} \propto p(\mu|y, \mu^{(1)})$$
$$\vdots$$
$$\mu^{(M)} \propto p(\mu|y, \sigma^{2(M-1)})$$
$$\sigma^{2(M)} \propto p(\mu|y, \mu^{(M)})$$

- After discarding some first draws, K, as burn in, we use $\mu^{(M-K)}, \ldots, \mu^{(M)}$ and $\sigma^{2(M-K)}, \ldots, \sigma^{2(M)}$ to compute quantities of interest.

Numerical example 2

Suppose we have a dataset of $N = 100$ observations from the two-parameter normal model with $\mu = 3$ and $\sigma^2 = 0.1$. Priors for $\mu : \mu_0 = 0, \sigma_0^2 = 0$ and for $\sigma^2 : \nu_0 = 3, s_0 = 100$. Using $10,000$ posterior draws after a burn-in of $1,000$, the posterior means of μ and σ^2 are as follows





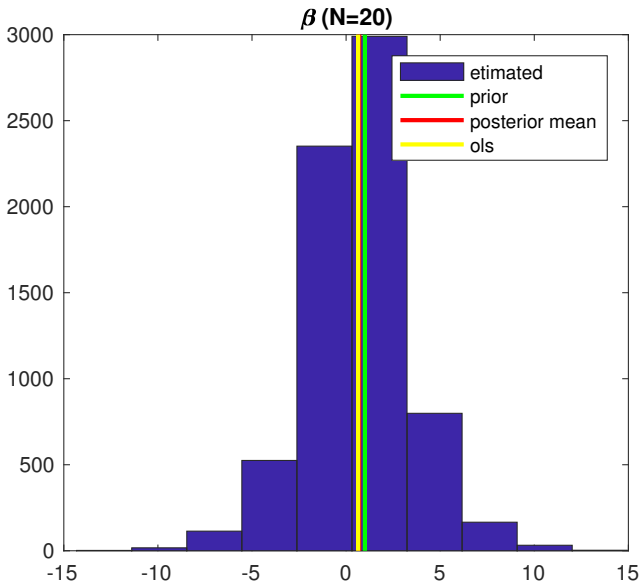
Numerical example 3

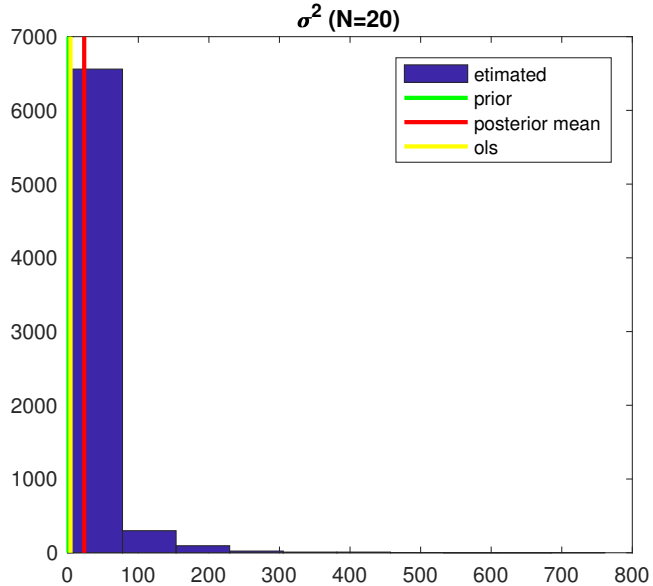
Consider equation

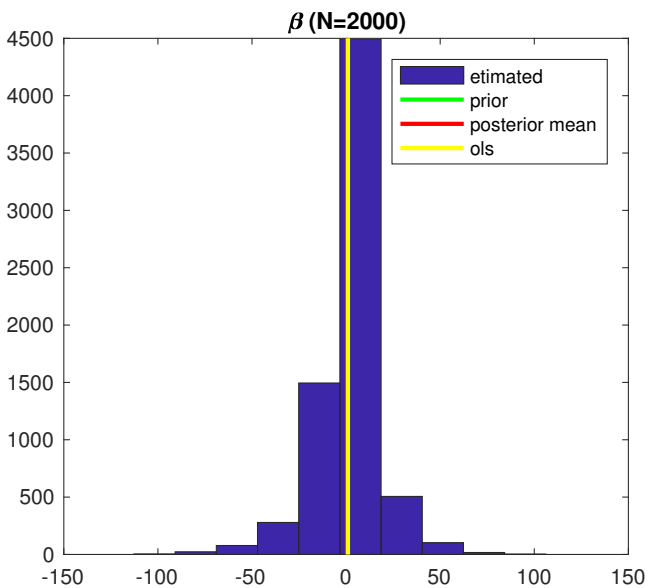
$$y_i = x_i\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2) \tag{14}$$

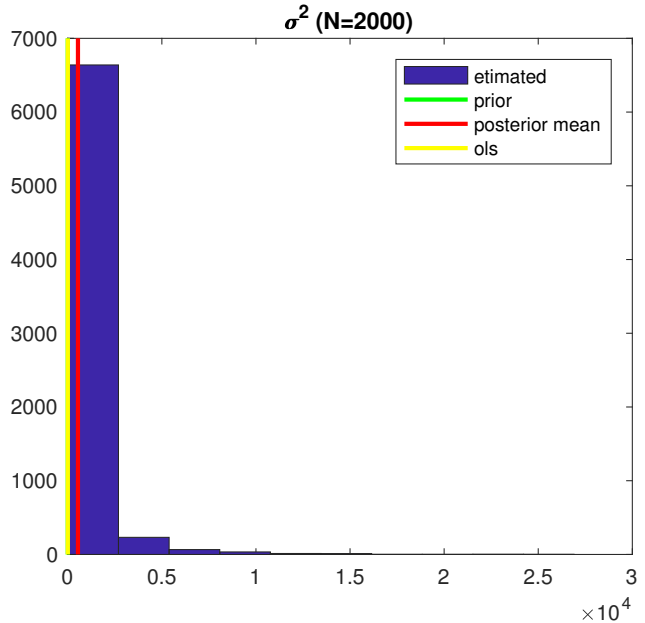
- data: $y \sim N(1,1)$ and $x \sim N(2,2)$. Case 1: N= 20; case 2 N=2000
- priors: $\beta \sim N(1,5)$ and for σ^2 are $\nu_0 = 1, s_0 = 1$
- posterior: 10000 draws, 3000 burn in

Numerical example 3



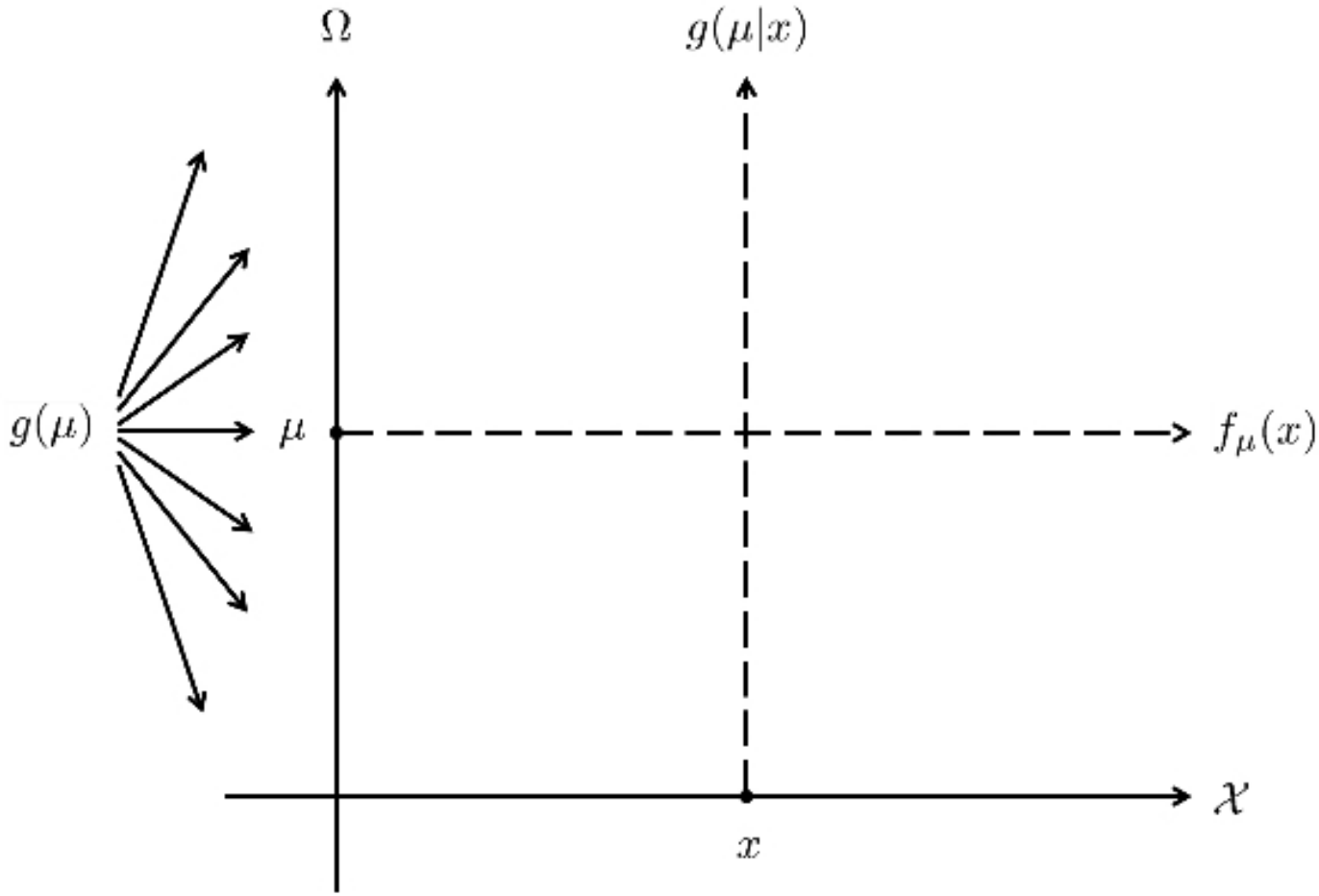






Conclusion

Bayesian inference proceeds vertically, given x; frequentist inference proceeds horizontally, given μ.



Why Bayesian?

- Why using prior? More information tends to be better than less!
- Prior sensitivity analysis
- Bayesian methods are crucial when you don't have much data.
- Bayesians are upfront and rigorous about including non-data information!
- Other advantages:
  - Unit root is relatively not important for applied Bayesian (frequentists care bout unit root because they use hypothesis testing methods, while Bayesians do not)
  - Frequentists also worry about unit roots due to spurious regression, but Bayesian improves the model by adding lags and "clean up" the strong correlation in the error.
  - Cointegration is also not a matter for Bayesian (State space models offer alternative way of modelling time series with common trends.

Want to learn about Bayesian econometrics?

- Joshua Chan
- Gary Koop
- Dimitris Korobilis
- Christiane Baumeister
- Haroon Mumtaz
- Michele Piffer

Appendix 1: derive conditional distribution

$p(\mu|\boldsymbol{y}, \sigma^2)$

Note that, given σ^2, this is the same normal model with known variance discussed previously. Thus,

$$D_\mu = \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1} \quad \hat{\mu} = D_\mu \left(\frac{\mu_0}{\sigma_0^2} + \frac{N\bar{y}}{\sigma^2}\right)$$

$p(\sigma^2|\boldsymbol{y}, \mu)$

Recall that, a random variable X is said to have an **inverse-gamma distribution** with shape parameter $\alpha > 0$ and scale parameter$\beta > 0$ if its density is given by

$$\boxed{f(x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x}} \tag{15}$$

$$\begin{aligned} p(\sigma^2|\boldsymbol{y}, \mu) &\propto p(\sigma^2) p(\boldsymbol{y}|\mu, \sigma^2) \\ &\propto (\sigma^2)^{-(\nu_0+1)} e^{-\frac{S_0}{\sigma^2}} (\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^N (y_n-\mu)^2} \\ &\propto (\sigma^2)^{-(\nu_0+N/2+1)} e^{-\frac{S_0+\sum_{n=1}^N (y_n-\mu)^2/2}{\sigma^2}} \end{aligned}$$

Thus,

$$p(\sigma^2|\boldsymbol{y}, \mu) \sim IG\left(\nu_0 + \frac{N}{2}, S_0 + \frac{1}{2}\sum_{n=1}^N (y_n - \mu)^2\right)$$

Appendix 2: Matlab code

Example 1 I

```
clc; clear; close all;
% compute posterior mean
mu0 = 10; sig02 = 1;
ybar = 20; sig2 = 1;
n_grid = 500;
mu_grid = linspace(10,22,n_grid)';
% case 1:
n = 1;
Dmu1 = 1/(1/sig02 + n/sig2);
mu_hat1 = Dmu1*(mu0/sig02 + n/sig2*ybar);
post1 = normpdf(mu_grid,mu_hat1,sqrt(Dmu1));
% case 2:
n = 10;
Dmu2 = 1/(1/sig02 + n/sig2);
mu_hat2 = Dmu2*(mu0/sig02 + n/sig2*ybar);
post2 = normpdf(mu_grid,mu_hat2,sqrt(Dmu2));

subplot(1,2,1);
```

Example 1 II

```
19 plot(mu_grid,post1,'linewidth',1); box off; ...
       xlim([10 22]);
20
21 subplot(1,2,2);
22 plot(mu_grid,post2,'linewidth',1); box off; ...
       xlim([10 22]);
23 set(gcf,'Position',[100 100 800 300])  %% 1 by 2 ...
       graphs
24
25 % Monte Carlo integration
26 R = 10000;
27 mu_hat = 19.09; Dmu = .09;
28 mu = mu_hat + sqrt(Dmu)*randn(R,1);
29 g_hat = mean(log(abs(mu)));
```

Example 2 I

```
clc; clear; close all;
% compute posterior mean
mu0 = 10; sig02 = 1;
ybar = 20; sig2 = 1;
n_grid = 500;
mu_grid = linspace(10,22,n_grid)';
% case 1:
n = 1;
Dmu1 = 1/(1/sig02 + n/sig2);
mu_hat1 = Dmu1*(mu0/sig02 + n/sig2*ybar);
post1 = normpdf(mu_grid,mu_hat1,sqrt(Dmu1));
% case 2:
n = 10;
Dmu2 = 1/(1/sig02 + n/sig2);
mu_hat2 = Dmu2*(mu0/sig02 + n/sig2*ybar);
post2 = normpdf(mu_grid,mu_hat2,sqrt(Dmu2));

subplot(1,2,1);
```

Example 2 II

```
19 plot(mu_grid,post1,'linewidth',1); box off; ...
       xlim([10 22]);
20
21 subplot(1,2,2);
22 plot(mu_grid,post2,'linewidth',1); box off; ...
       xlim([10 22]);
23 set(gcf,'Position',[100 100 800 300])  %% 1 by 2 ...
       graphs
24
25 % Monte Carlo integration
26 R = 10000;
27 mu_hat = 19.09; Dmu = .09;
28 mu = mu_hat + sqrt(Dmu)*randn(R,1);
29 g_hat = mean(log(abs(mu)));
```

Example 3 I

```
%% Bayesian estimate y=x\beta + \epsilon; epsilon ...
    ¬ N(0,sigma^2)

clear; clc; close all;
rng('shuffle')
nloop=10000; burnin=3000;
%% Generate the data
n=2000; x=zeros(n,1); y=zeros(n,1);
for i=1:n
    x(i,:)=1+sqrt(1)*randn;
    y(i,:)=2+sqrt(2)*randn;
end
%% Priors
beta0=1; betasig=5; %prior for beta
a=1; b=1; % priors for Sig2
%% Storage
store_beta=zeros(nloop-burnin,1);
store_Sig2=zeros(nloop-burnin,1);
```

Example 3 II

```
18 %% Initialized the Markow Chain
19 beta=1; Sig2=.5;
20 %% MCMC starts here randn('seed',sum(clock*100)); ...
       rand('seed',sum(clock*1000));
21 disp('Starting MCMC....  '); disp(' ' ); ...
       start_time = clock;
22 for loop=1:nloop
23     %% Sample beta
24     Vbeta=[sum(x.^2)/Sig2+(1/betasig)];
25     betahat=(1/Vbeta)*(sum(x.*y)/Sig2+beta0/betasig);
26     beta=betahat+sqrt(Vbeta)*randn;
27     %% Sample Sig2
28     err=y-x*beta;
29     lamda=sum(err.^2)/2+b;
30     alpha=n/2+a;
31     Sig2=1/gamrnd(alpha,1/lamda);
32     if loop > burnin
33     i = loop-burnin;
34     store_beta(i,:)=beta;
35     store_Sig2(i,:)=Sig2;
```

Example 3 III

```
36      end
37      if ( mod( loop, 2000 ) ==0 )
38  disp( [ num2str( loop ) ' loops...  '  ]  )
39      end
40  end
41  disp( ['MCMC takes ' num2str( etime( clock, ...
        start_time) ) ' seconds' ] ); disp(' ' );
42  posterior_beta=mean(store_beta); ...
        posterior_Sig2=mean(store_Sig2);
43  %% OLS
44  [ols,¬,r,¬,stats] = regress(y,x);
45  Sig2_ols=(r'*r)*(1/(n-1));
46  %% Figures
47  figure
48  subplot(1,2,1);
49  hist(store_beta); hold on
50  hold on;
51  line([beta0, beta0], ylim, 'LineWidth', 2, ...
        'Color', 'g');
52  hold on
```

Example 3 IV

```
53 line([posterior_beta, posterior_beta], ylim, ...
       'LineWidth', 2, 'Color', 'r'); hold on
54 line([ols, ols], ylim, 'LineWidth', 2, 'Color', ...
       'y'); hold off
55 legend('etimated', 'prior', 'posterior mean','ols')
56 title('\beta (N=2000)')
57
58 subplot(1,2,2);
59 hist(store_Sig2); hold on
60 line([a, a], ylim, 'LineWidth', 2, 'Color', 'g');
61 hold on
62 line([posterior_Sig2, posterior_Sig2], ylim, ...
       'LineWidth', 2, 'Color', 'r');hold on
63 line([Sig2_ols, Sig2_ols], ylim, 'LineWidth', 2, ...
       'Color', 'y');hold on
64 legend('etimated', 'prior', 'posterior mean','ols')
65 title('\sigma^2 (N=2000)')
66 set(gcf,'Position',[100 100 800 300])  %% 1 by 2 ...
       graphs
```