Bayesian Econometrics Primer

Stéphane Adjemian

stepan@adjemian.eu

January, 2019

Introduction

- In this chapter we present the Bayesian approach to econometrics.
- Basically, this approach allows to incorporate prior knowledge about the model and its parameters in the inference procedure.
- We will only deal with problems for which closed form solutions exist (linear models).
- In general DSGE models do not admit closed form solutions for the posterior distribution. We will deal with these models in the next chapter.

Outline

MV Estimation

- A model ($\mathcal{M}$) defines a joint probability distribution parameterized (by $\theta_{\mathcal{M}}$) function over a sample of variables (say $\mathcal{Y}_T$):

$$f(\mathcal{Y}_T|\theta_{\mathcal{M}}, \mathcal{M}) \tag{1}$$

- The parameters $\theta_{\mathcal{M}}$ can be estimated by confronting the model to the data through:
  - Some moments of the DGP.
  - The probability density function of the DGP (all the moments).
- The first approach is a method of moments, the second one corresponds to the Maximum Likelihood approach.
- Basically, a MV estimate for $\theta_{\mathcal{M}}$ is obtained by maximizing the density of the sample with respect to the parameters (we seek the value of $\theta_{\mathcal{M}}$ that maximizes the "probability of occurence" of the sample given by the Nature).
- In the sequel, we will denote $\mathcal{L}(\theta) = f(\mathcal{Y}_T|\theta)$ the likelihood function, omitting the indexation with respect to the model when not necessary.

MV Estimation

A simple static model

- As a first example, we consider the following model:

$$y_t = \mu_0 + \epsilon_t \tag{2-a}$$

where $\epsilon_t \underset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and μ_0 is an unknown finite real parameter.

- According to this model, y_t is normally distributed:

$$y_t|\mu_0 \sim \mathcal{N}(\mu_0, 1)$$

and $\mathbb{E}[y_t y_s] = 0$ for all $s \neq t$.

- Suppose that a sample $\mathcal{Y}_T = \{y_1, \ldots, y_T\}$ is available. The likelihood is defined by:

$$\mathcal{L}(\mu) = f(y_1, \ldots, y_T|\mu)$$

- Because the ys are iid, the joint conditional density is equal to a product of conditional densities:

$$\mathcal{L}(\mu) = \prod_{t=1}^{T} g(y_t|\mu)$$

MV Estimation

A simple static model

- Because the model is Gaussian:

$$\mathcal{L}(\mu) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_t-\mu)^2}{2}}$$

- Finally we have:

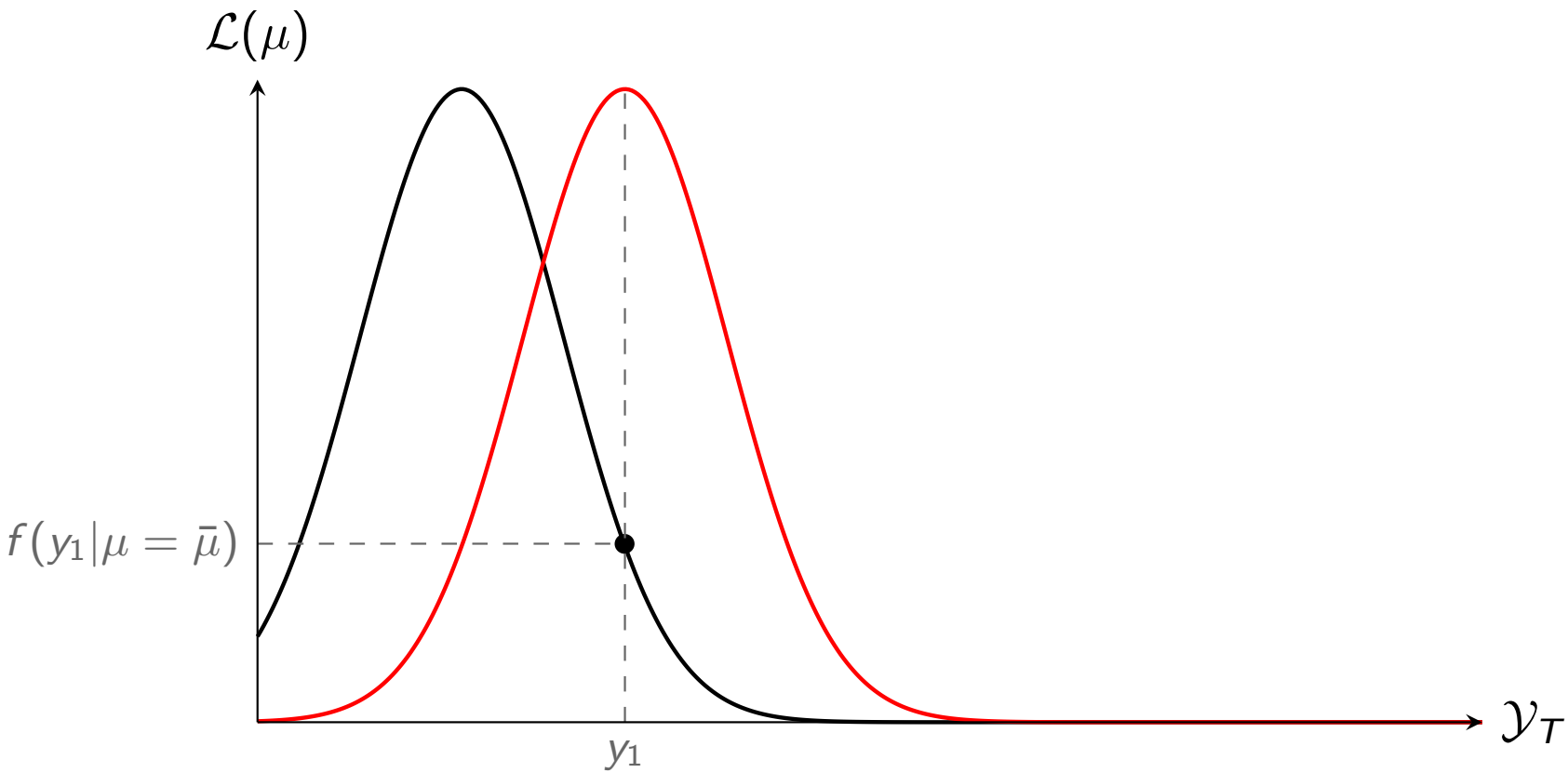
$$\mathcal{L}(\mu) = (2\pi)^{-\frac{T}{2}} e^{-\frac{1}{2}\sum_{t=1}^{T}(y_t-\mu)^2} \qquad \text{(2-b)}$$

- Note that the likelihood function depends on the data.

- Suppose that $T = 1$ (only one observation in the sample). We can graphically determine the ML estimator of μ in this case.

MV Estimation

A simple static model (cont'd)



Clearly, the value of the density of y_1 conditional on μ, *ie* the likelihood, is maximized for $\mu = y_1$: for any $\bar{\mu} \neq y_1$ we have $f(y_1|\mu = \bar{\mu}) < f(y_1|\mu = y_1)$

MV Estimation

A simple static model (cont'd)

$\Rightarrow$ If we have only one observation, y_1, the Maximum Likelihood estimator is the observation: $\widehat{\mu} = y_1$.

- This estimator is unbiased and its variance is 1.

- More generally, one can show that the maximum likelihood estimator is equal to the sample mean:

$$\widehat{\mu}_T = \frac{1}{T} \sum_{t=1}^{T} y_t \tag{2-c}$$

- This estimator is unbiased and its variance is given by:

$$\mathbb{V}\left[\widehat{\mu}_T\right] = \frac{1}{T} \tag{2-d}$$

- Because $\mathbb{V}\left[\widehat{\mu}\right]$ goes to zero as the sample size goes to infinity, we know that this estimator converges in probability to the true value μ_0 of the unknown parameter:

$$\widehat{\mu}_T \xrightarrow[T \to \infty]{\text{proba}} \mu_0$$

The ML estimator of μ must satisfy the following first order condition (considering the log of the likelihood):

$$\sum_{t=1}^{T} (y_t - \widehat{\mu}_T) = 0$$

$$\Leftrightarrow T\widehat{\mu}_T = \sum_{t=1}^{T} y_t$$

$$\Leftrightarrow \widehat{\mu}_T = \frac{1}{T} \sum_{t=1}^{T} y_t$$

We establish that this estimator is unbiased by showing that its unconditional expectation is equal to the true value of μ. We have:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_T\right] &= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} y_t\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mu_0 + \epsilon_t\right] \\
&= \frac{1}{T} T\mu_0 + 0 \\
&= \mu_0
\end{aligned}$$

where the second equality is obtained by linearity of the unconditional expectation and by substituting the DGP. Following the same steps,

we easily obtain the variance of the ML estimator:

$$\begin{aligned}
\mathbb{V}\left[\widehat{\mu}_T\right] &= \frac{1}{T^2}\mathbb{V}\left[\sum_{t=1}^{T} y_t\right] \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{V}\left[\mu_0 + \epsilon_t\right] \\
&= \frac{1}{T^2}T\mathbb{V}[\epsilon_t] + 0 \\
&= \frac{1}{T}
\end{aligned}$$

where the second equality is a consequence of the independence of the ys. If the variance of ϵ is not unitary we obtain $\mathbb{V}\left[\widehat{\mu}_T\right] = \sigma_\epsilon^2/T$ instead. The smaller is the size of the perturbation (or the greater is the sample), the more precise is the ML estimator of μ. This result is intuitive, the more noise we have in the sample (larger variance of ϵ) the more difficult is the extraction of the true value of μ.

MV Estimation

A simple dynamic model

- Suppose that the data are generated by an AR(1) model:

$$y_t = \varphi y_{t-1} + \epsilon_t$$

with $|\varphi| < 1$ and $\epsilon_t \underset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right)$.

- In this case, y_t depends (directly) on y_{t-1} and also on $y_{t-2}, y_{t-3}, \ldots$.
- It is no more legal to write the likelihood as the as a product of marginal densities of the observations.

Ex. 1

Show that the density of $\mathrm{y} \equiv (y_t, y_{t+1}, \ldots, y_{t+H-1})'$ is given by:

$$f(\mathrm{y}) = (2\pi)^{-\frac{H}{2}} |\Sigma_y|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathrm{y}'\Sigma_y^{-1}\mathrm{y}}$$

with

$$\Sigma_y = \frac{\sigma_\epsilon^2}{1-\varphi^2} \begin{pmatrix} 1 & \varphi & \varphi^2 & \ldots & \ldots & \varphi^{H-1} \\ \varphi & 1 & \varphi & \varphi^2 & \ldots & \varphi^{H-2} \\ \vdots & & & & & \\ \varphi^{H-1} & \varphi^{H-2} & \ldots & \ldots & \varphi & 1 \end{pmatrix}$$

under the assumption of stationarity.

MV Estimation

A simple dynamic model

Ex. 2

Let $\mathcal{Y}_T = \{y_1, y_2, \ldots, y_T\}$ be the sample. Write the likelihood function of the AR(1) model under the assumption of stationarity. Admitting that the inverse of the covariance matrix, Σ_y, can be factorized as $\Sigma_y^{-1} = \sigma_\epsilon^{-2} L'L$ with:

$$L = \begin{pmatrix} \sqrt{1-\varphi^2} & 0 & 0 & \ldots & 0 & 0 \\ -\varphi & 1 & 0 & \ldots & 0 & 0 \\ 0 & -\varphi & 1 & \ldots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & & & & -\varphi & 1 \end{pmatrix}$$

a $T \times T$ matrix, show that the likelihood function can be written as:

$$\mathcal{L}(\varphi, \sigma_\epsilon^2) = (2\pi)^{-\frac{T}{2}} \left(\frac{\sigma_\epsilon^2}{1-\varphi^2} \right)^{-\frac{1}{2}} \sigma_\epsilon^{T-1} e^{-\frac{1-\varphi^2}{2\sigma_\epsilon^2} y_1^2} e^{-\frac{1}{2\sigma_\epsilon^2} \sum_{t=2}^{T} \left(y_t - \varphi y_{t-1} \right)^2}$$

Bayes theorem

- Let A and B be two events.
- Let $\mathbb{P}(A)$ and $\mathbb{P}(B)$ be the marginal probabilities of these events.
- Let $\mathbb{P}(A \cap B)$ be the joint probability of events A and B.
- The Bayes theorem states that the probability of B conditional on A is given by:

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

- Or equivalently, that a joint probability can be expressed as the product of a conditional density and a marginal density:

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

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

- Same for continuous random variables.

Prior and posterior beliefs

- We assume that we are able to characterize our prior knowledge about a parameter with a probability density function.
- Let $p_0(\theta)$ be the prior density characterizing our beliefs about the vector of parameters θ.
- Our aim is to update our (prior) beliefs about θ with the sample information $(\mathcal{Y}_T)$ embodied in the likelihood function, $\mathcal{L}(\theta) = f(\mathcal{Y}_T|\theta)$.
- We define the posterior density, $p_1(\theta|\mathcal{Y}_T)$, which represents our updated beliefs.
- By the Bayes theorem we have:

$$p(\theta|\mathcal{Y}_T) = \frac{g(\theta, \mathcal{Y}_T)}{p(\mathcal{Y}_T)}$$

and

$$p(\theta|\mathcal{Y}_T) = \frac{f(\mathcal{Y}_T|\theta)p_0(\theta)}{p(\mathcal{Y}_T)}$$

where g is the joint density of the sample and the parameters.

Prior and posterior beliefs (cont'd)

- The posterior density is given by:

$$p(\theta|\mathcal{Y}_T) = \frac{\mathcal{L}(\theta)p_0(\theta)}{p(\mathcal{Y}_T)}$$

- Noting that the denominator does not depend on the parameters, we have that the posterior density is proportional (w.r.t θ) to the product of the likelihood and the prior density:

$$p(\theta|\mathcal{Y}_T) \propto \mathcal{L}(\theta)p_0(\theta)$$

- All the posterior inference about the parameters can be done with the posterior kernel: $\mathcal{L}(\theta)p_0(\theta)$.
- The denominator is the marginal density of the sample. Because a density has to sum up to one, we have:

$$p(\mathcal{Y}_T) = \int f(\mathcal{Y}_T|\theta)p_0(\theta)\mathrm{d}\theta$$

The marginal density is a weighted average of the likelihood function $\rightarrow$ will be used later for model comparison.

Prior and posterior beliefs

A simple static model (cont'd)

- For the sake of simplicity, we will see why later, we choose a Gaussian prior for the parameter μ, with prior expectation μ_0 and prior variance σ_μ^2:

$$p_0(\mu) = \frac{1}{\sigma_\mu \sqrt{2\pi}} e^{-\frac{1}{2\sigma_\mu^2}(\mu-\mu_0)^2}$$

- The smaller is the prior variance, σ_μ^2, the more informative is the prior.
- The posterior density is proportional to the product of the prior density and the likelihood:

$$p_1(\mu|\mathcal{Y}_T) \propto \frac{1}{\sigma_\mu \sqrt{2\pi}} e^{-\frac{1}{2\sigma_\mu^2}(\mu-\mu_0)^2} (2\pi)^{-\frac{T}{2}} e^{-\frac{1}{2}\sum_{t=1}^{T}(y_t-\mu)^2}$$

- One can show that the righthand side expression is proportional to a Gaussian density.

Prior and posterior beliefs

A simple static model (cont'd)

Ex. 3

Show that the likelihood can be equivalently written as:

$$\mathcal{L}(\mu) = (2\pi)^{-\frac{T}{2}} e^{-\frac{1}{2}\left(\nu s^2 + T(\mu - \widehat{\mu})^2\right)}$$

with $\nu = T - 1$ and

$$s^2 = \frac{1}{\nu} \sum_{t=1}^{T} (y_t - \widehat{\mu})^2$$

- s^2 and $\widehat{\mu}$ are sufficient statistics: they convey all the necessary sample information regarding the inference w.r.t μ.
- We use this alternative expression of the likelihood to show that the posterior density is Gaussian. We have:

$$p_1(\mu|\mathcal{Y}_T) \propto \frac{1}{\sigma_\mu \left(\sqrt{2\pi}\right)^{T+1}} e^{-\frac{1}{2\sigma_\mu^2}(\mu - \mu_0)^2 - \frac{\nu}{2} s^2 - \frac{T}{2}(\mu - \widehat{\mu})^2}$$

We have

$$\begin{aligned}
\sum_{t=1}^{T}(y_t - \mu)^2 &= \sum_{t=1}^{T}\left([y_t - \hat{\mu}] - [\mu - \hat{\mu}]\right)^2 \\
&= \sum_{t=1}^{T}(y_t - \hat{\mu})^2 + \sum_{t=1}^{T}(\mu - \hat{\mu})^2 - 2\sum_{t=1}^{T}(y_t - \hat{\mu})(\mu - \hat{\mu}) \\
&= \nu s^2 + T(\mu - \hat{\mu})^2 - 2\left(\sum_{t=1}^{T} y_t - T\hat{\mu}\right)(\mu - \hat{\mu}) \\
&= \nu s^2 + T(\mu - \hat{\mu})^2
\end{aligned}$$

the last term cancels out by definition of the sample mean.

Prior and posterior beliefs

A simple static model (cont'd)

- We can simplify the previous expression by omitting all the multiplicative terms not related to μ (this is legal because we are interested in a proportionality w.r.t μ):

$$p_1(\mu|\mathcal{Y}_T) \propto e^{-\frac{1}{2\sigma_\mu^2}(\mu-\mu_0)^2-\frac{T}{2}(\mu-\widehat{\mu})^2}$$

- We develop the quadratic forms and remove all the terms appearing additively (under the exponential function); we obtain:

$$p_1(\mu|\mathcal{Y}_T) \propto e^{-\frac{1}{2}\left(\sigma_\mu^{-2}+T\right)\left(\mu-\frac{T\widehat{\mu}+\mu_0\sigma_\mu^{-2}}{T+\sigma_\mu^{-2}}\right)^2}$$

- We recognize the expression of a Gaussian density (up to a scale parameter that does not depend on μ).

Let $A(\mu) = \frac{1}{\sigma_\mu^2}(\mu - \mu_0)^2 + T(\mu - \widehat{\mu})^2$. We establish the last expression of the posterior kernel by rewriting $A(\mu)$ as:

$$
\begin{aligned}
A(\mu) &= T(\mu - \widehat{\mu})^2 + \frac{1}{\sigma_\mu^2}(\mu - \mu_0)^2 \\
&= T\left(\mu^2 + \widehat{\mu}^2 - 2\mu\widehat{\mu}\right) + \frac{1}{\sigma_\mu^2}\left(\mu^2 + \mu_0^2 - 2\mu\mu_0\right) \\
&= \left(T + \frac{1}{\sigma_\mu^2}\right)\mu^2 - 2\mu\left(T\widehat{\mu} + \frac{1}{\sigma_\mu^2}\mu_0\right) + \left(T\widehat{\mu}^2 + \frac{1}{\sigma_\mu^2}\mu_0^2\right) \\
&= \left(T + \frac{1}{\sigma_\mu^2}\right)\left[\mu^2 - 2\mu\frac{T\widehat{\mu} + \frac{1}{\sigma_\mu^2}\mu_0}{T + \frac{1}{\sigma_\mu^2}}\right] + \left(T\widehat{\mu}^2 + \frac{1}{\sigma_\mu^2}\mu_0^2\right) \\
&= \left(T + \frac{1}{\sigma_\mu^2}\right)\left[\mu - \frac{T\widehat{\mu} + \frac{1}{\sigma_\mu^2}\mu_0}{T + \frac{1}{\sigma_\mu^2}}\right]^2 + \left(T\widehat{\mu}^2 + \frac{1}{\sigma_\mu^2}\mu_0^2\right) \\
&\qquad - \frac{\left(T\widehat{\mu} + \frac{1}{\sigma_\mu^2}\mu_0\right)^2}{T + \frac{1}{\sigma_\mu^2}}
\end{aligned}
$$

In the last equality, the two last additive terms do not depend on μ and can be therefore omitted.

Prior and posterior beliefs

A simple static model (cont'd)

- The posterior distribution is Gaussian with (posterior) expectation:

$$\mathbb{E}[\mu] = \frac{T\widehat{\mu} + \frac{1}{\sigma_\mu^2}\mu_0}{T + \frac{1}{\sigma_\mu^2}}$$

  and (posterior) variance:

$$\mathbb{V}[\mu] = \frac{1}{T + \frac{1}{\sigma_\mu^2}}$$

- As soon as the amount of prior information is positive ($\sigma_\mu^2 < \infty$) the posterior variance is less than the variance of the maximum likelihood estimator ($1/T$).

- The posterior expectation is a convex combination of the maximum likelihood estimator and the prior expectation.

Prior and posterior beliefs

A simple static model (cont'd)

- The Bayesian approach can be interpreted as a bridge between the calibration approach ($\sigma_\mu^2 = 0$, infinite amount of prior information) and the ML approach ($\sigma_\mu^2 = \infty$, no prior information):

$$\mathbb{E}[\mu] \xrightarrow[\sigma_\mu^2 \to 0]{} \mu_0$$

and

$$\mathbb{E}[\mu] \xrightarrow[\sigma_\mu^2 \to \infty]{} \widehat{\mu}$$

- The more important is the amount of information in the sample, the smaller will be the gap between the posterior expectation and the ML estimator.

Nuisance parameters

- One of the main advantages of the Bayesian approach is related to the treatment of the nuisance parameters.
- Suppose that the vector of estimated parameters is partitioned as $\theta' \equiv (\theta'_a, \theta'_b)$ and that we are only interested in θ_a (θ_b holds the nuisance parameters).
- The posterior density of θ_a is given by:

$$\begin{aligned} p_1(\theta_a|\mathcal{Y}_T) &= \int p_1(\theta_a, \theta_b|\mathcal{Y}_T)\mathrm{d}\theta_b \\ &= \int p_1(\theta_a|\theta_b, \mathcal{Y}_T)p_1(\theta_b|\mathcal{Y}_T)\mathrm{d}\theta_b \end{aligned}$$

- Nuisance parameters are eliminated by integrating them out!
- The marginal posterior density of θ_a is a weighted average of the conditional posterior density of θ_a knowing θ_b (the weights are given by the marginal posterior density of the nuisance parameters).

Nuisance parameters

A simple static model (cont'd)

- Suppose that the variance of ϵ_t is unknown, and has to be estimated jointly with μ.

- We need to choose a joint prior for μ and σ_ϵ^2, denoted $p_0(\mu, \sigma_\epsilon^2)$.
- This prior joint density can be factorized as:

$$p_0(\mu, \sigma_\epsilon^2) = p_0(\mu|\sigma_\epsilon^2) p_0(\sigma_\epsilon^2)$$

- We choose a Gaussian density for the prior conditional density of μ knowing σ_ϵ^2:

$$\mu|\sigma_\epsilon^2 \sim_0 \mathcal{N}\left(\mu_0, \sigma_\mu^2\right)$$

- We choose an inverse gamma density for the marginal prior density of σ_ϵ^2:

$$\sigma_\epsilon^2 \sim_0 \mathcal{IG}\left(\nu, s^2\right)$$

Ex. 4

Show that the posterior density of (μ, σ_ϵ^2) has the same shape than the prior density. Compute the marginal posterior density of μ.

Point estimate

- The outcome of the Bayesian approach is a (posterior) probability density function.
- But people generally expect much less information: a point estimate is often enough for most practical purposes (a single value for each parameter with a measure of uncertainty).

$\Rightarrow$ We need to reduce a function to a "representative" point.

- This is a well known in statistics and in microeconomics (choice under uncertainty).
- Let $L(\theta, \widehat{\theta})$ be the loss incurred if we choose $\widehat{\theta}$ while θ is the true value.
- The idea is to choose the value of θ that minimizes this loss... But the true value of θ is obviously unknown, so we minimize the (posterior) expected loss instead:

$$\theta^\star = \arg\min_{\widehat{\theta}} \mathbb{E}\left[L(\theta, \widehat{\theta})\right] = \arg\min_{\widehat{\theta}} \int L(\theta, \widehat{\theta}) p_1(\theta|\mathcal{Y}_T) \mathrm{d}\theta$$

- The choice of the loss function is purely arbitrary, for each loss we will obtain a different point estimate.

Point estimate

Quadratic loss function (L_2 norm)

- Suppose that the loss function is quadratic:

$$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})' \Omega (\theta - \widehat{\theta})$$

where Ω is a symmetric positive definite matrix. Note that this function returns a (real) scalar.

- The (posterior) expectation of the loss is:

$$\begin{aligned}
\mathbb{E}\left[L(\theta, \widehat{\theta})\right] &= \mathbb{E}\left[(\theta - \widehat{\theta})' \Omega (\theta - \widehat{\theta})\right] \\
&= \mathbb{E}\left[\left(\theta - \mathbb{E}\theta - \left(\widehat{\theta} - \mathbb{E}\theta\right)\right)' \Omega \left(\theta - \mathbb{E}\theta - \left(\widehat{\theta} - \mathbb{E}\theta\right)\right)\right] \\
&= \mathbb{E}\left[(\theta - \mathbb{E}\theta)' \Omega (\theta - \mathbb{E}\theta)\right] + (\widehat{\theta} - \mathbb{E}\theta)' \Omega (\widehat{\theta} - \mathbb{E}\theta)
\end{aligned}$$

- Noting that the first term does not depend on the choice variable, $\widehat{\theta}$, the expected loss is trivially minimized when $\widehat{\theta}$ is equal to the (posterior) expectation of θ:

$$\theta^\star = \mathbb{E}[\theta]$$

$\Rightarrow$ If the loss is quadratic the optimal point estimate is the posterior expectation.

Point estimate

Absolute value loss function (L_1 norm)

- Suppose that θ is a scalar defined over $[a, b]$ and that the loss function:

$$L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$$

- The (posterior) expectation of the loss is:

$$\begin{aligned}
\mathbb{E}\left[L(\theta, \widehat{\theta})\right] &= \int_a^b |\theta - \widehat{\theta}| p_1(\theta|\mathcal{Y}_T) \mathrm{d}\theta \\
&= \int_a^{\widehat{\theta}} (\widehat{\theta} - \theta) p_1(\theta|\mathcal{Y}_T) \mathrm{d}\theta + \int_{\widehat{\theta}}^b (\theta - \widehat{\theta}) p_1(\theta|\mathcal{Y}_T) \mathrm{d}\theta \\
&= \widehat{\theta} P(\widehat{\theta}|\mathcal{Y}_T) - \widehat{\theta}\left(P(1 - \widehat{\theta}|\mathcal{Y}_T)\right) + \int_{\widehat{\theta}}^b \theta p_1(\theta|\mathcal{Y}_T) \mathrm{d}\theta - \int_a^{\widehat{\theta}} \theta p_1(\theta|\mathcal{Y}_T) \mathrm{d}\theta
\end{aligned}$$

where $P(x|\mathcal{Y}_T)$ is the posterior cumulative distribution function.

Ex. 5

Show that the optimal point estimate is the median of the posterior distribution

Marginal density of the sample

- If we are only interested in inference about the parameters, the marginal density of the data, $p(\mathcal{Y}_T)$, can be omitted.
- We already saw that the marginal density of the data is:

$$p(\mathcal{Y}_T) = \int f(\mathcal{Y}_T|\theta)p_0(\theta)\mathrm{d}\theta$$

  - The marginal density of the sample acts as a constant of integration in the expression of the posterior density.
  - The marginal density of the sample is an average of the likelihood function (for different values of the estimated parameters) weighted by the prior density.

$\Rightarrow$ The marginal density of the sample is a measure of fit, which does not depend on the parameters (because we integrate them out).

- Note that, theoretically, it is possible to compute the marginal density of the sample (conditional on a model) without estimating the parameters.

Marginal density of the sample

A simple static model (cont'd)

- Suppose again that the sample size is $T = 1$. The likelihood is given by:

$$f(\mathcal{Y}_T|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1-\mu)^2}$$

- The marginal density is then given by:

$$\begin{aligned} p(\mathcal{Y}_T) &= \int_{-\infty}^{\infty} f(y_1|\mu) p_0(\mu) \mathrm{d}\mu \\ &= (2\pi\sigma_\mu)^{-1} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left((y_1-\mu)^2 + \frac{(\mu-\mu_0)^2}{\sigma_\mu^2}\right)} \mathrm{d}\mu \\ &= \frac{1}{\sqrt{2\pi(1+\sigma_\mu^2)}} e^{-\frac{(y_1-\mu_0)^2}{2(1+\sigma_\mu^2)}} \end{aligned}$$

- We can directly obtain the same result by noting that y_1 is the sum of two Gaussian random variables: $\mathcal{N}(0,1)$ and $\mathcal{N}\left(\mu_0, \sigma_\mu^2\right)$.

Marginal density of the sample

Model comparison

- Suppose we have two models $\mathcal{A}$ and $\mathcal{B}$ (with two associated vectors of deep parameters $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{B}}$) estimated using *the same sample* $\mathcal{Y}_T$.
- For each model $\mathcal{I} = \mathcal{A}, \mathcal{B}$ we can evaluate, at least theoretically, the marginal density of the data conditional on the model:

$$p(\mathcal{Y}_T|\mathcal{I}) = \int f(\mathcal{Y}_T|\theta_{\mathcal{I}}, \mathcal{I}) p_0(\theta_{\mathcal{I}}|\mathcal{I}) d\theta_{\mathcal{I}}$$

  by integrating out the deep parameters $\theta_{\mathcal{I}}$ from the posterior kernel.
- $p(\mathcal{Y}_T|\mathcal{I})$ measures the fit of model $\mathcal{I}$. If we have to choose between models $\mathcal{A}$ and $\mathcal{B}$ we will select the model with the highest marginal density of the sample.
- Note that models $\mathcal{A}$ and $\mathcal{B}$ need not to be nested (for instance, we do not require that $\theta_{\mathcal{A}}$ be a subset of $\theta_{\mathcal{B}}$) for the comparison to make sense, because the compared marginal densities do not depend on the parameters. The classical approach (comparisons of likelihoods) by requiring nested models is much less obvious.

Marginal density of the sample

Model comparison (cont'd)

- Suppose we have a prior distribution over models $\mathcal{A}$ and $\mathcal{B}$: $p(\mathcal{A})$ and $p(\mathcal{B})$.

- Again, using the Bayes theorem we can compute the posterior distribution over models:

$$p(\mathcal{I}|\mathcal{Y}_T) = \frac{p(\mathcal{I})p(\mathcal{Y}_T|\mathcal{I})}{\sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} p(\mathcal{I})p(\mathcal{Y}_T|\mathcal{I})}$$

- This formula may easily be generalized to a collection of N models.

- In the literature posterior odds ratio, defined as:

$$\frac{p(\mathcal{A}|\mathcal{Y}_T)}{p(\mathcal{B}|\mathcal{Y}_T)} = \frac{p(\mathcal{A})}{p(\mathcal{B})} \frac{p(\mathcal{Y}_T|\mathcal{A})}{p(\mathcal{Y}_T|\mathcal{B})}$$

  are often used to discriminate between different models. If the posterior odds ratio is large ($>$100) we can safely choose model $\mathcal{A}$.

Marginal density of the sample

Model comparison (cont'd)

- Note that we do not necessarily have to choose one model.
- Even if a model has a smaller posterior probability (or marginal density) it may provide useful informations in some directions (or frequencies), so we should not discard this information.
- An alternative is to mix the models.
- If these models are used for forecasting inflation, we can report an average of the forecasts weighted by the posterior probabilities, $p(\mathcal{I}|\mathcal{Y}_T)$, instead of the forecasts of the best model (in terms of marginal density) $\rightarrow$ Bayesian averaging.

Predictive density

- We often seek to use the estimated model to do inference about unobserved variables.
- The most obvious example is the forecasting exercise.
- In the Bayesian context the density of an unobserved variable (for instance the future growth of GDP) given the sample, is called a predictive density.
- Let $\tilde{y}$ be a vector of unobserved variables. The joint posterior density of $\tilde{y}$ and θ is:

$$p_1(\tilde{y}, \theta | \mathcal{Y}_T) = g(\tilde{y} | \theta, \mathcal{Y}_T) p_1(\theta | \mathcal{Y}_T)$$

- The predictive density is obtained by integrating out the parameters:

$$p(\tilde{y} | \mathcal{Y}_T) = \int g(\tilde{y} | \theta, \mathcal{Y}_T) p_1(\theta | \mathcal{Y}_T) \mathrm{d}\theta$$

The predictive density is the average of the density of $\tilde{y}$ knowing the parameters weighted by the posterior density of the parameters.

Predictive density

A simple static model (cont'd)

- Suppose that we want to do inference about the out of sample variable y_{T+1} (forecast).

- The density of y_{T+1} conditional on the sample and on the parameter is:

$$g(y_{T+1}|\mu, \mathcal{Y}_T) \propto e^{-\frac{1}{2}(y_{T+1}-\mu)^2}$$

  Note that this conditional density does not depend on $\mathcal{Y}_T$ because the model is static (for an autoregressive model, y_{T+1} and y_T would appear under the quadratic term).

- Remember that the posterior density of μ is:

$$p_1(\mu|\mathcal{Y}_T) \propto e^{-\frac{1}{2\mathbb{V}[\mu]}(\mu-\mathbb{E}[\mu])^2}$$

  where $\mathbb{E}[\mu]$ and $\mathbb{V}[\mu]$ are the posterior first and second order moments obtained earlier.

Predictive density

A simple static model (cont'd)

- The predictive density for y_{T+1} is given by:

$$p(y_{T+1}|\mathcal{Y}_T) = \int g(y_{T+1}|\mu, \mathcal{Y}_T) p_1(\mu|\mathcal{Y}_T) \mathrm{d}\mu$$
$$\propto \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y_{T+1}-\mu)^2 - \frac{1}{2\mathbb{V}[\mu]}(\mu-\mathbb{E}[\mu])^2} \mathrm{d}\mu$$

Ex. 6

Show that minus two times the terms under the exponential in the last expression can be rewritten as:

$$A(\mu) = \left[1 + \frac{1}{\mathbb{V}[\mu]}\right] \left(\mu - \frac{y_{T+1} + \mathbb{V}[\mu]^{-1}}{1 + \mathbb{V}[\mu]^{-1}}\right)^2 + \frac{\mathbb{V}[\mu]^{-1}}{1 + \mathbb{V}[\mu]^{-1}} (y_{T+1} - \mathbb{E}[\mu])^2$$

Predictive density

A simple static model (cont'd)

- Substituting $A(\mu)$ in the expression of the predictive density for y_{T+1} we obtain:

$$
\begin{aligned}
p(y_{T+1}|\mathcal{Y}_T) &\propto \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[1+\frac{1}{\mathbb{V}[\mu]}\right]\left(\mu-\frac{y_{T+1}+\mathbb{V}[\mu]^{-1}}{1+\mathbb{V}[\mu]^{-1}}\right)^2-\frac{1}{2}\frac{\mathbb{V}[\mu]^{-1}}{1+\mathbb{V}[\mu]^{-1}}(y_{T+1}-\mathbb{E}[\mu])^2}\,\mathrm{d}\mu \\
&\propto e^{-\frac{1}{2}\frac{\mathbb{V}[\mu]^{-1}}{1+\mathbb{V}[\mu]^{-1}}(y_{T+1}-\mathbb{E}[\mu])^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[1+\frac{1}{\mathbb{V}[\mu]}\right]\left(\mu-\frac{y_{T+1}+\mathbb{V}[\mu]^{-1}}{1+\mathbb{V}[\mu]^{-1}}\right)^2}\,\mathrm{d}\mu \\
&\propto e^{-\frac{1}{2}\frac{\mathbb{V}[\mu]^{-1}}{1+\mathbb{V}[\mu]^{-1}}(y_{T+1}-\mathbb{E}[\mu])^2}
\end{aligned}
$$

- Unsurprisingly, we recognize the Gaussian density:

$$y_{T+1}|\mathcal{Y}_T \sim \mathcal{N}\left(\mathbb{E}[\mu], 1+\mathbb{V}[\mu]\right)$$

- We would have obtained directly the same result by first noting that y_{T+1} is the sum of two Gaussian random variables: $\mathcal{N}\left(\mathbb{E}[\mu], \mathbb{V}[\mu]\right)$ (for the estimated parameter) and $\mathcal{N}\left(0, 1\right)$ (for the error term).

Predictive density

Point prediction

- For reporting our forecast, we may want to select one point in the predictive distribution.
- We proceed as for the point estimate by choosing an arbitrary loss function and minimizing the posterior expected loss.
- Usually the expectation of the predictive distribution is reported (rationalized with a quadratic loss function).

Asymptotic properties of the Bayesian approach

- The posterior density is proportional to the product of the likelihood and the prior density.
- As the sample gets larger the relative weight of likelihood increases (the prior does not depend on T).
- Asymptotically ($T \to \infty$) the Bayesian estimator inherits all the properties of the likelihood estimator.
- We know that under fairly general assumption, the likelihood is asymptotically Gaussian (even for nonlinear models).

$\Rightarrow$ Asymptotically, the posterior distribution is Gaussian.

- If the (finite sample) posterior distribution is untractable or does not possess a closed form expression, we can use an asymptotic approximation (with the Gaussian distribution).

Asymptotic properties of the Bayesian approach

- We know that:

$$\begin{aligned} p_1(\theta|\mathcal{Y}_T) &\propto p_0(\theta) f(\mathcal{Y}_T|\theta) \\ &\propto p_0(\theta) e^{\log f(\mathcal{Y}_T|\theta)} \end{aligned}$$

- Usually the log likelihood is $\mathcal{O}(T)$ while the prior is $\mathcal{O}(1)$

$\Rightarrow$ When T goes to infinity the density of the sample conditional on the parameters dominates the prior density (which can be neglected if T is large enough).

- For instance, in the simple static model we have:

$$\log f(\mathcal{Y}_T|\mu) = -\frac{T}{2}\log(2\pi) - \frac{T-1}{2}s^2 - T(\mu - \widehat{\mu})^2$$

which grows linearly with T.

Asymptotic properties of the Bayesian approach

Approximation

- Let $\hat{\theta}$ be the posterior mode obtained by maximizing the posterior kernel $\mathcal{K}(\theta) \equiv \mathcal{L}(\theta)p_0(\theta)$.

- With an order two Taylor expansion around $\hat{\theta}$, we have:

$$\log \mathcal{K}(\theta) = \log \mathcal{K}(\hat{\theta}) + (\theta - \hat{\theta}) \left. \frac{\partial \log \mathcal{K}(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta - \hat{\theta})' \left. \frac{\partial^2 \log \mathcal{K}(\theta)}{\partial \theta \partial \theta'} \right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \mathcal{O}\left(||\theta - \hat{\theta}||^3\right)$$

- Equivalently:

$$\log \mathcal{K}(\theta) = \log \mathcal{K}(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})'[\mathcal{H}(\hat{\theta})]^{-1}(\theta - \hat{\theta}) + \mathcal{O}(||\theta - \hat{\theta}||^3)$$

where $\mathcal{H}(\hat{\theta})$ is minus the inverse of the Hessian matrix evaluated at the posterior mode.

Asymptotic properties of the Bayesian approach

Approximation (cont'd)

- The posterior kernel can be approximated by:

$$\mathcal{K}(\theta) \doteq \mathcal{K}(\hat{\theta}) e^{-\frac{1}{2}(\theta-\hat{\theta})'[\mathcal{H}(\hat{\theta})]^{-1}(\theta-\hat{\theta})}$$

- Up to a constant

$$c = \mathcal{K}(\hat{\theta})(2\pi)^{\frac{k}{2}} |\mathcal{H}(\hat{\theta})|^{\frac{1}{2}}$$

  where k is the number of estimated parameters, we recognize the density of a multivariate Gaussian distribution.

- Completing for constant of integration we obtain an approximation of the posterior density:

$$p_1(\theta|\mathcal{Y}_T) \doteq (2\pi)^{-\frac{k}{2}} |\mathcal{H}(\hat{\theta})|^{-\frac{1}{2}} e^{-\frac{1}{2}(\theta-\hat{\theta})'[\mathcal{H}(\hat{\theta})]^{-1}(\theta-\hat{\theta})}$$

Asymptotic properties of the Bayesian approach

Approximation (cont'd)

- If the model is stationnary the hessian matrix is of order $\mathcal{O}(T)$, as T tends to infinity the posterior distribution concentrates around the posterior mode (which matches the ML estimator).

- This Gaussian approximation (namely the constant of integration c) is often used to approximate the marginal density of sample ($\rightarrow$ Laplace approximation).

- The asymptotic approximation is reliable iff the true (finite sample) posterior distribution is not too far from a Gaussian distribution.

Non informative priors

- Clearly, the inference will depend on the choice for the priors.
- The robustness of the results should be evaluated:
  - By checking that the results do not change too much if we increase the prior variance or consider more general prior shapes.
  - By checking that the results do not change too much if we change the parameterization of the model (which is equivalent to changing the shapes of the priors).
- Because the results depend crucially on the choice for the priors, we may want to do the inference with a non informative prior.
- Unfortunately there is no clear agreement in the literature about what should be a non informative prior.
- In the sequel we review two non informative priors proposed by Jeffrey.

Non informative priors

Jeffrey-I

- For a parameter that admits positive and negative values, we consider a uniform prior between $-\infty$ and ∞.
- If the parameter is constrained to be positive we consider a uniform prior between $-\infty$ and ∞ for the log of the parameter.

- For a real scalar parameter, Jeffrey's a priori density such that:

$$p_0(\theta)\mathrm{d}\theta \propto \mathrm{d}\theta$$

  For a vector of real parameters, take a product of such densities:

$$p_0(\theta)\mathrm{d}\theta \propto \mathrm{d}\theta_1\mathrm{d}\theta_2 \ldots \mathrm{d}\theta_n$$

- Obviously, this prior density is improper because the sum of the prior is not finite:

$$\int_{-\infty}^{\infty} \mathrm{d}\theta = \infty$$

- For Jeffrey, the improperness of this prior is precisely what we need to define a non informative prior.

Non informative priors

Jeffrey-I (cont'd)

- Because the prior is improper, we have:

$$\frac{\mathbb{P}_0(a < \theta < b)}{\mathbb{P}_0(c < \theta < d)} = \frac{0}{0}$$

  meaning that wee cannot compare the events $\theta \in (a, b)$ and $\theta \in (c, d)$

- For Jeffrey the improperness of the prior is the formalization of our ignorance.

- To understand this point, consider instead a bounded uniform prior distribution:

$$p_0(\theta)\mathrm{d}\theta = \frac{\mathrm{d}\theta}{2M} \quad \forall\ -M \leq \theta \leq M$$

- This proper uniform prior is informative because:

$$\frac{\mathbb{P}_0(a < \mu < b)}{\mathbb{P}_0(c < \mu < d)} = \frac{b-a}{d-c}$$

Non informative priors

Jeffrey-I (cont'd)

- If a parameter is constrained to be positive, Jeffrey suggest to put a uniform prior on the log of the parameter.
- A non informative prior on $\sigma > 0$ is defined as:

$$\theta = \log \sigma$$

$$p_0(\theta)\mathrm{d}\theta \propto \mathrm{d}\theta$$

- Because $\mathrm{d}\theta = \mathrm{d}\log\sigma = \frac{\mathrm{d}\sigma}{\sigma}$, we can equivalently write this prior as:

$$p_0(\sigma)\mathrm{d}\sigma = \frac{\mathrm{d}\sigma}{\sigma}$$

- This prior is improper because $\int_0^\infty \frac{\mathrm{d}\sigma}{\sigma}$ is not finite.
- We also have:

$$\int_0^a \frac{\mathrm{d}\sigma}{\sigma} = \infty \quad \text{and} \quad \int_b^\infty \frac{\mathrm{d}\sigma}{\sigma} = \infty$$

Non informative priors

Jeffrey-I (cont'd)

- Jeffrey's flat prior is invariant with respect to a power transformation.

- Suppose that $\phi = \sigma^n$.
- Then $\mathrm{d}\phi = n\sigma^{n-1}\mathrm{d}\sigma$ and consequently:

$$\frac{\mathrm{d}\phi}{\phi} \propto \frac{\mathrm{d}\sigma}{\sigma}$$

$\Rightarrow$ If we choose a flat prior for σ then we also have a flat prior on ϕ.

- This prior is not invariant with respect to other non linear transformations.

- The improperness of the prior is not an issue (w.r.t the inference about the parameters) as long as the posterior is proper (otherwise posterior inference would not be possible).

Non informative priors

A simple static model (cont'd)

- Suppose we change the Gaussian prior for μ by:

$$p_0(\mu)\mathrm{d}\mu \propto \mathrm{d}\mu$$

- The posterior density is then characterized by:

$$\begin{aligned} p_1(\mu|\mathcal{Y}_T) &= p_0(\mu) f(\mathcal{Y}_T|\mu) \\ &\propto e^{-\frac{1}{2}\left(\sum_{t=1}^{T}(y_t-\hat{\mu})^2+T(\mu-\hat{\mu})^2\right)} \\ &\propto e^{-\frac{T}{2}(\mu-\hat{\mu})^2} \end{aligned}$$

- We recognize the expression of a Gaussian density (up to a scaling term):

$$\mu|\mathcal{Y}_T \sim \mathcal{N}\left(\hat{\mu}, \frac{1}{T}\right)$$

- With a flat Jeffrey prior, the (Gaussian) posterior distribution is centered around the ML estimator. The posterior variance is equal to the variance of the ML estimator (no information in the prior).

Non informative priors

Jeffrey-II

- Years latter, Jeffrey came with another non informative prior generalizing the invariance property with respect to nonlinear transformations of the parameters.

- Basically the idea is to mimic the information in the data. Jeffrey proposes the following prior:

$$p_0(\theta) \propto |\mathrm{Inf}_\theta|^{\frac{1}{2}}$$

where Inf_θ is the Fisher information matrix:

$$\mathrm{Inf}_\theta = -\mathbb{E}_y \left[\frac{\partial^2 \log \mathcal{L}(\theta)}{\partial\theta\partial\theta'} \right]$$

where the expectation is an integral with respect to the density of the sample.

Non informative priors

Jeffrey-II (cont'd)

Invariance property

Suppose that we adopt the following priors for θ and $\eta = F(\theta)$ (where F is a differentiable function):

$$p_0(\theta) \propto |\mathrm{Inf}_\theta|^{\frac{1}{2}} \quad \text{and} \quad p_0(\eta) \propto |\mathrm{Inf}_\eta|^{\frac{1}{2}}$$

These two priors will lead to exactly the same posterior inference.

- This result states that the Jeffrey-II prior is invariant w.r.t. any non linear re-parameterization of the model.
- To establish this result we just have to note that:

$$|\mathrm{Inf}_\theta|^{\frac{1}{2}} = |J_F||\mathrm{Inf}_\eta|^{\frac{1}{2}}$$

and also

$$\mathrm{d}\theta = |J_F|^{-1}\mathrm{d}\eta$$

where J_F is the Jacobian matrix of F.

Non informative priors

A simple static model (cont'd)

Ex. 7

Show that the two Jeffrey's non informative priors are equivalent in the simple static model.

- Note that this result is not general. For instance in a dynamic model, these two priors lead to very different posterior inference. See the dispute about unit root testing in autoregressive models published in a special issue of the Journal of Applied Econometics (1991).

- The Jeffrey II prior is also improper in general.