Introduction to Bayesian estimation

Michel Juillard

CEPREMAP

A Bayesian approach

- Uncertainty and *a priori* knowledge about the model and its parameters are described by prior probabilities
- Confrontation to the data leads to a revision of these probabilities (posterior probabilities)
- Point estimates are obtained by minimizing a loss function (analogous to economic decision under uncertainty)
- Testing and model comparison is done by comparing posterior probabilities

Bayesian ingredients

- Choosing prior density
- Computing posterior mode
- Simulating posterior distribution
- Computing point estimates and confidence regions
- Computing posterior probabilities

Prior density

$$p(\boldsymbol{\theta}_A|A)$$

where A represents the model and $\boldsymbol{\theta}_A$, the parameters of that model.
The prior density describes *a priori* beliefs, before considering the data.

Likelihood function

- Conditional density

$$p(\boldsymbol{y}|\boldsymbol{\theta}_A, A)$$

- Conditional density for dynamic timeseries models

$$p(\mathbf{Y}_T|\boldsymbol{\theta}_A, A) = p(y_0|\boldsymbol{\theta}_A, A) \prod_{t=1}^{T} p(y_t|\mathbf{Y}_{T-1}, \boldsymbol{\theta}_A, A)$$

  where $\mathbf{Y}_T$ are the observations until period T

- Likelihood function

$$\mathcal{L}(\boldsymbol{\theta}_A|\mathbf{Y}_T, A) = p(\mathbf{Y}_T|\boldsymbol{\theta}_A, A)$$

Marginal density

$$
\begin{aligned}
p(\boldsymbol{y}|A) &= \int_{\Theta_A} p(\boldsymbol{y}, \boldsymbol{\theta}_A|A) d\boldsymbol{\theta}_A \\
&= \int_{\Theta_A} p(\boldsymbol{y}|\boldsymbol{\theta}_A, A) p(\boldsymbol{\theta}_A|A) d\boldsymbol{\theta}_A
\end{aligned}
$$

Posterior density

- Posterior density

$$p(\boldsymbol{\theta}_A|\mathbf{Y}_T, A) = \frac{p(\boldsymbol{\theta}_A|A)p(\mathbf{Y}_T|\boldsymbol{\theta}_A, A)}{p(\mathbf{Y}_T|A)}$$

- Unnormalized posterior density or posterior density kernel

$$p(\boldsymbol{\theta}_A|\mathbf{Y}_T, A) \propto p(\boldsymbol{\theta}_A|A)p(\mathbf{Y}_T|\boldsymbol{\theta}_A, A)$$

Posterior predictive density

$$
\begin{aligned}
p(\tilde{\mathbf{Y}}|\mathbf{Y}_T) &= \int_{\Theta_A} p(\tilde{\mathbf{Y}},\boldsymbol{\theta}_A|\mathbf{Y}_T,A)d\boldsymbol{\theta}_A \\
&= \int_{\Theta_A} p(\tilde{\mathbf{Y}}|\boldsymbol{\theta}_A,\mathbf{Y}_T,A)p(\boldsymbol{\theta}_A|\mathbf{Y}_T,A)d\boldsymbol{\theta}_A
\end{aligned}
$$

Bayes risk function

$$
\begin{aligned}
R(a) &= E\left[L(a,\boldsymbol{\theta})\right] \\
&= \int_{\Theta_A} L(a,\boldsymbol{\theta}_A)p(\boldsymbol{\theta}_A)d\boldsymbol{\theta}_A
\end{aligned}
$$

where $L(a,\boldsymbol{\theta})$ is the loss function associated with decision a when parameters take value $\boldsymbol{\theta}_A$.

Estimation

Action: deciding that the estimated value of $\boldsymbol{\theta}_A$ is $\widetilde{\boldsymbol{\theta}}_A$

- Point estimate:

$$\widehat{\boldsymbol{\theta}}_A = \arg\min_{\widetilde{\boldsymbol{\theta}}_A} \int_{\Theta_A} L(\widetilde{\boldsymbol{\theta}}_A, \boldsymbol{\theta}_A) p(\boldsymbol{\theta}_A | \mathbf{Y}_T, A) d\boldsymbol{\theta}_A$$

- Quadratic loss function:

$$\widehat{\boldsymbol{\theta}}_A = E(\boldsymbol{\theta}_A | \mathbf{Y}_T, A)$$

- Zero-one loss function: $\widehat{\boldsymbol{\theta}}_A$ = posterior mode

Credible sets

$$P(\theta \in C) = \int_C p(\theta) d\theta = 1 - \alpha$$

is a $100(1-\alpha)\%$ credible set for θ with respect to $p(\theta)$.

A $100(1-\alpha)\%$ highest probability density (HPD) credible set for θ with respect to $p(\theta)$ is a $100(1-\alpha)\%$ credible set with the property

$$p(\theta_1) \geq p(\theta_2) \quad \forall \theta_1 \in C \text{ and } \forall \theta_2 \in \bar{C}$$

Numerical integration

$$
\begin{aligned}
E(h(\boldsymbol{\theta}_A)) &= \int_{\Theta_A} h(\boldsymbol{\theta}_A) p(\boldsymbol{\theta}_A | \mathbf{Y}_T, A) d\boldsymbol{\theta}_A \\
&\approx \frac{1}{N} \sum_{k=1}^{N} h(\boldsymbol{\theta}_A^k)
\end{aligned}
$$

where $\boldsymbol{\theta}_A^k$ is drawn from $p(\boldsymbol{\theta}_A | \mathbf{Y}_T, A)$.

Metropolis algorithm

1. Draw a starting point $\boldsymbol{\theta}^{\circ}$ which $p(\boldsymbol{\theta}) > 0$ from a starting distribution $p^{\circ}(\boldsymbol{\theta})$.

Metropolis algorithm (continued)

2. For $t = 1, 2, \ldots$
    1. Draw a *proposal* $\boldsymbol{\theta}^*$ from a *jumping* distribution

$$J(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{t-1}) = N(\boldsymbol{\theta}^{t-1}, c\Sigma_{\text{mode}})$$

    2. Compute the acceptance ratio

$$r = \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{t-1})}$$

    3. Set

$$\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{t-1} & \text{otherwise.} \end{cases}$$

In practice . . .

- fix scale factor c so as to obtain a 25% average acceptance ratio
- discard first 50% of the draws

Potential Scale Reduction Factor

If we have simulated m independant sequences of n draws, a particular draw of scalar θ is noted θ_{ij} with $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

$$
\begin{aligned}
B &= \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\theta}_{.j} - \bar{\theta}_{..}\right)^2 \\
W &= \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{i=1}^{n} \left(\theta_{ij} - \theta_{.j}\right)^2 \\
\widehat{var}^{+}(\theta|\mathbf{Y}_T, A) &= \frac{n-1}{n} W + \frac{1/n}{B} \\
\hat{R} &= \sqrt{\frac{\widehat{var}^{+}(\theta|\mathbf{Y}_T, A)}{W}}
\end{aligned}
$$

ultivariate Potential Scale Reduction Fact

$$
\begin{aligned}
\hat{V} &= \frac{n-1}{n} W + \left(1 + \frac{1}{m}\right) B/n \\
W &= \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\boldsymbol{\theta}_{ij} - \overline{\boldsymbol{\theta}}_{\cdot j})(\boldsymbol{\theta}_{ij} - \overline{\boldsymbol{\theta}}_{\cdot j})' \\
B/n &= \frac{1}{m-1} \sum_{j=1}^{m} (\overline{\boldsymbol{\theta}}_{\cdot j} - \overline{\boldsymbol{\theta}_{\cdot\cdot}})(\overline{\boldsymbol{\theta}}_{\cdot j} - \overline{\boldsymbol{\theta}_{\cdot\cdot}})' \\
\hat{R}^p &= \frac{n-1}{n} + \frac{m+1}{m} \lambda_1
\end{aligned}
$$

λ_1 is the largest eigenvalue of $W^{-1}B/n$

Model comparison

The ratio of posterior probabilities of two models is

$$\frac{P(A_j|\mathbf{Y}_T)}{P(A_k|\mathbf{Y}_T)} = \frac{P(A_j)}{P(A_k)}\frac{p(\mathbf{Y}_T|A_j)}{p(\mathbf{Y}_T|A_k)}$$

In favor of the model A_j versus the model A_k:

- the **prior odds ratio** is $P(A_j)/P(A_k)$
- the **Bayes factor** is $p(\mathbf{Y}_T|A_j)/p(\mathbf{Y}_T|A_k)$
- the **posterior odds ratio** is $P(A_j|\mathbf{Y}_T)/P(A_k|\mathbf{Y}_T)$

Laplace approximation

$$\begin{aligned} p(\mathbf{Y}_T, A) &= \int_{\boldsymbol{\theta}_A} p(\boldsymbol{\theta}_A|\mathbf{Y}_T, A) p(\boldsymbol{\theta}_A|A) d\boldsymbol{\theta}_A \\ \hat{p}(\mathbf{Y}_T|A) &= (2\pi)^{\frac{k}{2}} |\Sigma_{\boldsymbol{\theta}^M}|^{-\frac{1}{2}} p(\boldsymbol{\theta}_A^M|\mathbf{Y}_T, A) p(\boldsymbol{\theta}_A^M|A) \end{aligned}$$

where $\boldsymbol{\theta}_A^M$ is the posterior mode.

Geweke (1999) modified harmonic mean

$$\begin{aligned}
p(\mathbf{Y}_T|A) &= \int_{\boldsymbol{\theta}_A} p(\boldsymbol{\theta}_A|\mathbf{Y}_T, A)p(\boldsymbol{\theta}_A|A)d\boldsymbol{\theta}_A \\
\hat{p}(\mathbf{Y}_T|A) &= \left[\frac{1}{n}\sum_{i=1}^{n}\frac{f(\boldsymbol{\theta}_A^{(i)})}{p(\boldsymbol{\theta}_A^{(i)}|\mathbf{Y}_T, A)p(\boldsymbol{\theta}_A^{(i)}|A)}\right]^{-1} \\
f(\boldsymbol{\theta}) &= p^{-1}(2\pi)^{\frac{k}{2}}|\Sigma_{\boldsymbol{\theta}}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})'\Sigma_{\boldsymbol{\theta}}{}^{-1}(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})\right\} \\
&\quad \times\left\{(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}})'\Sigma_{\boldsymbol{\theta}}{}^{-1}(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}) \leq F^{-1}_{\chi^2_k(p)}\right\}
\end{aligned}$$

with p an arbitrary probability and k, the number of estimated parameters.