I. Bayesian econometrics

A. Introduction
B. Bayesian inference in the univariate regression model
C. Statistical decision theory
   1. Example: portfolio allocation problem

a_j = quantity of asset j purchased

$j = 1, \ldots, J$

y = income

budget constraint:

$$\sum_{j=1}^{J} a_j = y$$

r_j = gross rate of return on asset j

c = future consumption

$$c = \sum_{j=1}^{J} r_j a_j$$

$$\max_{\{a_1,\ldots,a_J\}} EU\left(\sum_{j=1}^{J} r_j a_j\right)$$

$$\text{s.t. } \sum_{j=1}^{J} a_j = y$$

$\mathbf{r} = (r_1, \ldots, r_J)'$

$\mathbf{r}|\boldsymbol{\mu},\boldsymbol{\Omega} \sim N(\boldsymbol{\mu},\boldsymbol{\Omega})$

assume:

$\boldsymbol{\Omega}$ is known

$\boldsymbol{\mu}$ is unknown, must be estimated from **Y**

Classical econometrician:

Step 1: Solve optimization problem as if $\boldsymbol{\mu},\boldsymbol{\Omega}$ known with certainty

Step 2: Estimate $\hat{\boldsymbol{\mu}}$ from data $\mathbf{Y}$

Step 3: Plug results from Step 2 into Step 1

Example:

$$U(c) = -\exp(-\gamma c)$$
$$EU(c) = -E\exp(-\gamma \mathbf{a}'\mathbf{r})$$
$$= -\exp[-\gamma \mathbf{a}'\boldsymbol{\mu} + (\gamma^2/2)\mathbf{a}'\boldsymbol{\Omega}\mathbf{a}]$$

$$\max_{\{\mathbf{a}\}} \; -\exp[-\gamma \mathbf{a}'\boldsymbol{\mu} + (\gamma^2/2)\mathbf{a}'\boldsymbol{\Omega}\mathbf{a}]$$

s.t. $\mathbf{a}'\mathbf{1} = y$

$$\mathcal{L} = -\gamma \mathbf{a}'\boldsymbol{\mu} + (\gamma^2/2)\mathbf{a}'\boldsymbol{\Omega}\mathbf{a} + \lambda(y - \mathbf{a}'\mathbf{1})$$
$$-\gamma\boldsymbol{\mu} + \gamma^2\boldsymbol{\Omega}\mathbf{a} - \lambda\mathbf{1} = \mathbf{0}$$
$$\mathbf{a} = \gamma^{-2}[\boldsymbol{\Omega}^{-1}(\gamma\boldsymbol{\mu} + \lambda\mathbf{1})]$$
$$\mathbf{a}^* = \gamma^{-2}\left[\boldsymbol{\Omega}^{-1}(\gamma\hat{\boldsymbol{\mu}} + \hat{\lambda}\mathbf{1})\right]$$

Bayesian econometrician:

Solve optimization problem under uncertainty

posterior:

$$\boldsymbol{\mu}|\mathbf{Y} \sim N(\mathbf{m}^*, \mathbf{M}^*)$$
$$\mathbf{r} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Omega})$$
$$\mathbf{r}|\mathbf{Y} \sim N(\mathbf{m}^*, \boldsymbol{\Omega} + \mathbf{M}^*)$$

$$E[U(c)|\mathbf{Y}] = -E[\exp(-\gamma \mathbf{a}'\mathbf{r})|\mathbf{Y}]$$
$$= -\exp[-\gamma \mathbf{a}'\mathbf{m}^* + (\gamma^2/2)\mathbf{a}'(\boldsymbol{\Omega} + \mathbf{M}^*)\mathbf{a}]$$
$$\mathbf{a}^* = \gamma^{-2}[(\boldsymbol{\Omega} + \mathbf{M}^*)^{-1}(\gamma \mathbf{m}^* + \lambda^* \mathbf{1})]$$

uncertainty about $\boldsymbol{\mu}$ influences portfolio allocation decision (even if we have diffuse prior so that $\mathbf{m}^* = \hat{\boldsymbol{\mu}}$)

Bayesian considers the statistical inference problem to be: calculate the posterior distribution

How this distribution is used to come up with a "parameter estimate" requires specifying a loss function

I. Bayesian econometrics

C. Statistical decision theory

1. Example: portfolio allocation problem
2. General decision theory

$\boldsymbol{\theta}$ = unknown true value

$\hat{\boldsymbol{\theta}}$ = estimate

$\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ = loss function

= how much we are concerned if we announce an estimate of $\hat{\boldsymbol{\theta}}$ but the truth is $\boldsymbol{\theta}$

$\hat{\boldsymbol{\theta}}$ is solution to

$$\min_{\boldsymbol{\theta}} \int_{\aleph} \ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{Y}) d\boldsymbol{\theta}$$

where $\boldsymbol{\theta} \in \aleph$

Scalar examples:

(1) quadratic loss

$\ell(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$

Claim: optimal $\hat{\theta} = E(\theta|\mathbf{Y})$

Proof:

$$
\begin{aligned}
E_{\theta|\mathbf{Y}}\left[\theta - \hat{\theta}\right]^2
&= E_{\theta|\mathbf{Y}}\left[\theta - E(\theta|\mathbf{Y}) + E(\theta|\mathbf{Y}) - \hat{\theta}\right]^2 \\
&= E_{\theta|\mathbf{Y}}[\theta - E(\theta|\mathbf{Y})]^2 + \left[E(\theta|\mathbf{Y}) - \hat{\theta}\right]^2 \\
&\quad + 2E_{\theta|\mathbf{Y}}[\theta - E(\theta|\mathbf{Y})]\left[E(\theta|\mathbf{Y}) - \hat{\theta}\right] \\
&= E_{\theta|\mathbf{Y}}[\theta - E(\theta|\mathbf{Y})]^2 + \left[E(\theta|\mathbf{Y}) - \hat{\theta}\right]^2
\end{aligned}
$$

minimized at $\hat{\theta} = E(\theta|\mathbf{Y})$

Conclusion: for quadratic loss, optimal estimate is posterior mean

(2) absolute loss

$$\ell(\hat{\theta}, \theta) = |\theta - \hat{\theta}|$$

Claim: optimal $\hat{\theta} = \theta_{\text{med}}$

$$\int_{-\infty}^{\theta_{\text{med}}} p(\theta|\mathbf{Y})d\theta = 0.5$$

Proof:

$$\int_{-\infty}^{\infty} |\theta - \hat{\theta}| p(\theta|\mathbf{Y})d\theta$$

$$= \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) p(\theta|\mathbf{Y})d\theta$$

$$+ \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) p(\theta|\mathbf{Y})d\theta$$

differentiating with respect to $\hat{\theta}$ gives:

differentiating with respect to $\hat{\theta}$ gives:

$$(\hat{\theta} - \hat{\theta}) p(\hat{\theta}|\mathbf{Y}) + \int_{-\infty}^{\hat{\theta}} p(\hat{\theta}|\mathbf{Y})$$

$$-(\hat{\theta} - \hat{\theta}) p(\hat{\theta}|\mathbf{Y}) - \int_{\hat{\theta}}^{\infty} p(\hat{\theta}|\mathbf{Y})$$

minimized when

$$\int_{-\infty}^{\hat{\theta}} p(\hat{\theta}|\mathbf{Y}) = \int_{\hat{\theta}}^{\infty} p(\hat{\theta}|\mathbf{Y})$$

Conclusion: for absolute loss, optimal estimate is posterior median

(3) point loss (discrete case)

$\theta \in \{\theta_1, \ldots, \theta_J\}$

$\ell(\hat{\theta}, \theta) = 0$ if $\theta = \hat{\theta}$

$= 1$ if $\theta \neq \hat{\theta}$

$\hat{\theta} = \arg\min \sum_{j=1}^{J} \left[1 - \delta(\hat{\theta} = \theta_j)\right] P(\theta = \theta_j | \mathbf{Y})$

$\Rightarrow \hat{\theta} = \theta_j$ for which $P(\theta = \theta_j | \mathbf{Y})$ is highest

Conclusion: for point loss, optimal estimate is posterior mode

Returning to example from first lecture

$y_t | \mu \sim N(\mu, \sigma^2)$ (σ known)

$\mu \sim N(m, \tau^2)$ (prior)

$\mu | \mathbf{Y} \sim N(m^*, \tau^{*2})$ (posterior)

$$m^* = \left[\frac{(\sigma^2/T)}{(\sigma^2/T) + \tau^2}\right] m + \left[\frac{\tau^2}{(\sigma^2/T) + \tau^2}\right] \overline{y}$$

$$m^* = \left[\frac{(\sigma^2/T)}{(\sigma^2/T) + \tau^2}\right] m + \left[\frac{\tau^2}{(\sigma^2/T) + \tau^2}\right] \overline{y}$$

for any of these three loss functions (quadratic, absolute, point), the estimate would be m^*

diffuse prior: $\tau \to \infty$

$\Rightarrow \hat{\mu} = \overline{y}$

I. Bayesian econometrics

C. Statistical decision theory
   1. Example: portfolio allocation problem
   2. General decision theory
   3. Bayesian statistics and admissibility

More generally, we can consider some action a we plan to take, e.g.,

$a = \hat{\theta}$ means we announce that our estimate is $\hat{\theta}$

$a = 0$ if we accept H_0: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$a = 1$ if we reject H_0: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$\ell(a, \boldsymbol{\theta}) =$ loss if we take the action a when the true value of the parameter turns out to be $\boldsymbol{\theta}$

Definition: an action a is said to be inadmissible if there is an alternative action b such that $\ell(a, \boldsymbol{\theta}) \geq \ell(b, \boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ with strict inequality for some $\boldsymbol{\theta}$

The Bayes decision implied by
the probability distribution $p(\boldsymbol{\theta})$
is the action a for which $\int \ell(a,\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$
is minimized

Under certain regularity conditions:

(1) If action a is the Bayes decision implied by $p(\boldsymbol{\theta})$, then a is admissible

(2) If action a is admissible, then there exists a $p(\boldsymbol{\theta})$ for which a is the Bayes decision

Example: hypothesis testing

$a = 1$ (reject H_0: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$)

$a = 0$ (accept H_0)

$\ell(1,\boldsymbol{\theta}) = 0$ if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$

$= 1$ if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$\ell(0,\boldsymbol{\theta}) = 0$ if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$= c$ if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$

Bayes action: choose $a = 1$ if

$$E[\ell(1,\boldsymbol{\theta})|\mathbf{Y}] < E[\ell(0,\boldsymbol{\theta})|\mathbf{Y}]$$
$$P[\boldsymbol{\theta} \in \boldsymbol{\Theta}_0|\mathbf{Y}] < c\{1 - P[\boldsymbol{\theta} \in \boldsymbol{\Theta}_0|\mathbf{Y}]\}$$
$$P[\boldsymbol{\theta} \in \boldsymbol{\Theta}_0|\mathbf{Y}] < 1/(1 + c)$$

The hypothesis test

reject H_0 if $T(\mathbf{Y}) > t$

is said to be inadmissible if there exists an alternative test

reject H_0 if $S(\mathbf{Y}) > s$

such that:

(1) for every $\theta \in \boldsymbol{\Theta}_0$,

$$\int_{T(\mathbf{Y})>t} p(\mathbf{Y}|\boldsymbol{\theta})d\mathbf{Y} \geq \int_{S(\mathbf{Y})>s} p(\mathbf{Y}|\boldsymbol{\theta})d\mathbf{Y}$$

(2) for every $\theta \notin \boldsymbol{\Theta}_0$,

$$\int_{T(\mathbf{Y})>t} p(\mathbf{Y}|\boldsymbol{\theta})d\mathbf{Y} \leq \int_{S(\mathbf{Y})>s} p(\mathbf{Y}|\boldsymbol{\theta})d\mathbf{Y}$$

(3) there is some θ for which the the inequality in either (1) or (2) is strict

I. Bayesian econometrics

C. Statistical decision theory

D. Large sample results

Goal of this section:

A Bayesian is doing something with the data. How would a classical econometrician describe what that is?

I. Bayesian econometrics

C. Statistical decision theory

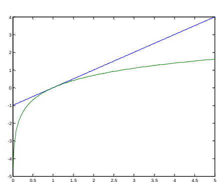
D. Large sample results

1. Background: The Kullback-Leibler information inequality

Claim:

$\log x \le x - 1$

equality only if $x = 1$



Implication:

$$E\log x \leq E(x) - 1$$

with equality only if $x = 1$ with probability 1

Application of claim to case of discrete parameter space and discrete random variables

$\theta \in \{\theta_1, \ldots, \theta_J\}$

$\theta^* =$ true value

$y_t \in \{1, \ldots, I\}$

Define

$$x(y_t, \theta_j) = \frac{P(Y = y_t | \theta = \theta_j)}{P(Y = y_t | \theta = \theta^*)}$$

This is a random variable (because y_t is random) that with probability $P(Y = i | \theta = \theta^*)$ takes on the value

$$\frac{P(Y = i | \theta = \theta_j)}{P(Y = i | \theta = \theta^*)}$$

$$E_{\theta^*}[x(y_t,\theta_j)]$$
$$= \sum_{i=1}^{I} \frac{P(Y = i|\theta = \theta_j)}{P(Y = i|\theta = \theta^*)} P(Y = i|\theta = \theta^*)$$
$$= \sum_{i=1}^{I} P(Y = i|\theta = \theta_j)$$
$$= 1$$

$$E_{\theta^*}[\log x(y_t,\theta_j)]$$
$$= \sum_{i=1}^{I} \log\left[\frac{P(Y = i|\theta = \theta_j)}{P(Y = i|\theta = \theta^*)} \right] P(Y = i|\theta = \theta^*)$$
$$= E_{\theta^*}\left\{ \log\left[\frac{p(y_t|\theta_j)}{p(y_t|\theta^*)} \right] \right\}$$

The claim

$$E \log x \leq E(x) - 1$$

implies for this case that

$$E_{\theta^*}\left\{ \log\left[\frac{p(y_t|\theta_j)}{p(y_t|\theta^*)} \right] \right\} \leq 1 - 1 = 0$$

with equality only if

$p(y_t|\theta_j) = p(y_t|\theta^*)\ \forall y_t$

Kullback-Leibler information inequality:

$$E_{\boldsymbol{\theta}^*}\left\{\log\left[\frac{p(\mathbf{y}_t|\boldsymbol{\theta})}{p(\mathbf{y}_t|\boldsymbol{\theta}^*)}\right]\right\} \leq 0$$

with equality only if $\boldsymbol{\theta} = \boldsymbol{\theta}^*$

I. Bayesian econometrics

C. Statistical decision theory

D. Large sample results

1. Background: The Kullback-Leibler information inequality
2. Implications of K-L for Bayesian posterior probabilities

will illustrate how data eventually overwhelm any prior

$$p(\boldsymbol{\theta}_s|\mathbf{Y}) = \frac{p(\boldsymbol{\theta}_s)p(\mathbf{Y}|\boldsymbol{\theta}_s)}{\sum_{j=1}^{J} p(\boldsymbol{\theta}_j)p(\mathbf{Y}|\boldsymbol{\theta}_j)}$$

$$= \frac{p(\boldsymbol{\theta}_s)\prod_{t=1}^{T} p(\mathbf{y}_t|\boldsymbol{\theta}_s)}{\sum_{j=1}^{J} p(\boldsymbol{\theta}_j)\prod_{t=1}^{T} p(\mathbf{y}_t|\boldsymbol{\theta}_j)}$$

$$= \frac{p(\boldsymbol{\theta}_s)\prod_{t=1}^{T}[p(y_t|\boldsymbol{\theta}_s)/p(\mathbf{y}_t|\boldsymbol{\theta}^*)]}{\sum_{j=1}^{J} p(\boldsymbol{\theta}_j)\prod_{t=1}^{T}[p(y_t|\boldsymbol{\theta}_s)/p(\mathbf{y}_t|\boldsymbol{\theta}^*)]}$$

$$= \frac{\exp\left\{\log p(\boldsymbol{\theta}_s) + \sum_{t=1}^{T}\log\left[\frac{p(\mathbf{y}_t|\boldsymbol{\theta}_s)}{p(\mathbf{y}_t|\boldsymbol{\theta}^*)}\right]\right\}}{\sum_{j=1}^{J}\exp\left\{\log p(\boldsymbol{\theta}_j) + \sum_{t=1}^{T}\log\left[\frac{p(\mathbf{y}_t|\boldsymbol{\theta}_j)}{p(\mathbf{y}_t|\boldsymbol{\theta}^*)}\right]\right\}}$$

LLN:

$$T^{-1}\sum_{t=1}^{T}\log\left[\frac{p(\mathbf{y}_t|\boldsymbol{\theta}_s)}{p(\mathbf{y}_t|\boldsymbol{\theta}^*)}\right]\underset{\boldsymbol{\theta}^*}{\to} E_{\boldsymbol{\theta}^*}\log\left[\frac{p(\mathbf{y}_t|\boldsymbol{\theta}_s)}{p(\mathbf{y}_t|\boldsymbol{\theta}^*)}\right]$$

which is < 0 if $\boldsymbol{\theta}_s \neq \boldsymbol{\theta}^*$

$= 0$ if $\boldsymbol{\theta}_s = \boldsymbol{\theta}^*$

$$p(\boldsymbol{\theta}_s|\mathbf{Y}) = \frac{\exp\left\{\log p(\boldsymbol{\theta}_s) + \sum_{t=1}^{T}\log\left[\frac{p(\mathbf{y}_t|\boldsymbol{\theta}_s)}{p(\mathbf{y}_t|\boldsymbol{\theta}^*)}\right]\right\}}{\sum_{j=1}^{J}\exp\left\{\log p(\boldsymbol{\theta}_j) + \sum_{t=1}^{T}\log\left[\frac{p(\mathbf{y}_t|\boldsymbol{\theta}_j)}{p(\mathbf{y}_t|\boldsymbol{\theta}^*)}\right]\right\}}$$

$$p(\boldsymbol{\theta}_s|\mathbf{Y}) \overset{p}{\to} \begin{cases} 0 & \text{if } \boldsymbol{\theta}_s \neq \boldsymbol{\theta}^* \\ 1 & \text{if } \boldsymbol{\theta}_s = \boldsymbol{\theta}^* \end{cases}$$

conclusion: Bayesian posterior distribution collapses to a spike at truth for i.i.d. discrete data

I. Bayesian econometrics

C. Statistical decision theory

D. Large sample results

1. Background: The Kullback-Leibler information inequality
2. Implications of K-L for Bayesian posterior probabilities
3. Bayesian posterior distribution as approximation to asymptotic distribution of MLE

$$\log p(\mathbf{Y}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \log p(\mathbf{y}_t|\boldsymbol{\theta})$$

define

$$\hat{\boldsymbol{\theta}}_T = \arg\max \log p(\mathbf{Y}|\boldsymbol{\theta})$$

$$\left.\frac{\partial \log p(\mathbf{Y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T} = \mathbf{0}$$

$$\begin{aligned}
&\log p(\mathbf{Y}|\boldsymbol{\theta}) \\
&= \log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}_T) + \left.\frac{\partial \log p(\mathbf{Y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T) \\
&\quad + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)' \left.\frac{\partial^2 \log p(\mathbf{Y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_T} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_T)
\end{aligned}$$

$$\tilde{\boldsymbol{\theta}}_T = \lambda_T \boldsymbol{\theta} + (1 - \lambda_T)\hat{\boldsymbol{\theta}}_T$$

$$\mathbf{H}_t(\boldsymbol{\theta}) \equiv -\frac{\partial^2 \log p(\mathbf{y}_t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}$$

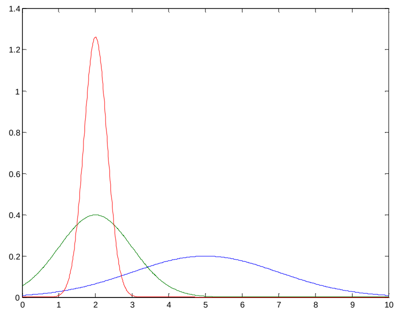
$$\mathbf{H}(\boldsymbol{\theta}) \equiv -E\frac{\partial^2 \log p(\mathbf{y}_t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}$$

$$\log p(\mathbf{Y}|\boldsymbol{\theta}) = \log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}_T) - \frac{1}{2}\sqrt{T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)' \times T^{-1}\sum_{t=1}^{T}\mathbf{H}_t(\tilde{\boldsymbol{\theta}}_T)\sqrt{T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)$$

$$\log p(\mathbf{Y}|\boldsymbol{\theta}) \simeq \log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}_T) - \frac{1}{2}\sqrt{T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)' \times \mathbf{H}(\boldsymbol{\theta}^*)\sqrt{T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)$$

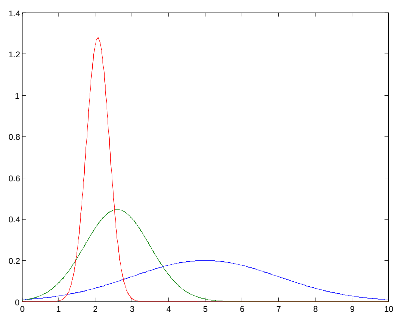
$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{Y}) &\simeq \tilde{k}_T p(\boldsymbol{\theta})\exp[-(1/2)\sqrt{T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)' \times \mathbf{H}(\boldsymbol{\theta}^*)\sqrt{T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)] \\ &= p(\boldsymbol{\theta})q_T(\boldsymbol{\theta}) \end{aligned}$$

$q_T(\boldsymbol{\theta})$ = kernel of $N(\mathbf{0}, \mathbf{H}(\boldsymbol{\theta}^*)^{-1})$ density for $\sqrt{T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_T)$



blue: $p(\theta)$ green: $q_{10}(\theta)$ red: $q_{100}(\theta)$

posterior distributions



blue: $T = 0$ green: $T = 10$ red: $T = 100$

Conclusions: the sequence of posterior distributions $p(\boldsymbol{\theta}|\mathbf{Y}_T)$ has the property

$$p(\boldsymbol{\theta}|\mathbf{Y}_T) \stackrel{p}{\to} 1 \text{ at } \boldsymbol{\theta} = \boldsymbol{\theta}^*$$
$$\stackrel{p}{\to} 0 \text{ at } \boldsymbol{\theta} \neq \boldsymbol{\theta}^*$$

Let $\boldsymbol{\theta}_T$ be sequence of random variables with distribution $p(\boldsymbol{\theta}|\mathbf{Y}_T)$. Then conditional on $\{\mathbf{Y}_T\}$ we have
$\sqrt{T}(\boldsymbol{\theta}_T - \hat{\boldsymbol{\theta}}_T) \xrightarrow{L} N(\mathbf{0}, \mathbf{H}(\boldsymbol{\theta}^*)^{-1})$
where distribution is across realizations of $\boldsymbol{\theta}_T$

Contrast with classical result:
$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{L} N(\mathbf{0}, \mathbf{H}(\boldsymbol{\theta}^*)^{-1})$
where distribution is across realizations of $\mathbf{Y}_T$

Implication: calculating the Bayesian posterior distribution is a way to find the asymptotic distribution of the MLE when regularity conditions hold

$y_t|\mu \sim N(\mu, \sigma^2)$ (σ known)

$\mu \sim N(m, \tau^2)$ (prior)

$\mu|\mathbf{Y} \sim N(m^*, \tau^{*2})$ (posterior)

$$m^* = \left[\frac{(\sigma^2/T)}{(\sigma^2/T) + \tau^2}\right] m + \left[\frac{\tau^2}{(\sigma^2/T) + \tau^2}\right] \overline{y}_T$$

$$\tau^{*2} = \frac{\tau^2 \sigma^2/T}{(\sigma^2/T) + \tau^2}$$

$$\tau^{*2} = \frac{\tau^2 \sigma^2/T}{(\sigma^2/T) + \tau^2}$$

Conditional on $\mathbf{Y}_T$, the variable $\mu|\mathbf{Y}_T$ has a distribution characterized by

$\tau^{*-1}(\mu_T - m_T^*) \sim N(0,1)$

$\frac{\sqrt{T}}{\sigma\tau}[(\sigma^2/T) + \tau^2]^{1/2}(\mu_T - m_T^*) \sim N(0,1)$

$$\frac{\sqrt{T}}{\sigma\tau}[(\sigma^2/T) + \tau^2]^{1/2}(\mu_T - m_T^*) \sim N(0,1)$$

As $T \to \infty$

$$\frac{\sqrt{T}}{\sigma}(\mu_T - \bar{y}_T) \sim N(0,1)$$

classical result:

$$\frac{\sqrt{T}}{\sigma}(\bar{y}_T - \mu^*) \sim N(0,1)$$

I. Bayesian econometrics

C. Statistical decision theory
D. Large sample results
E. Diffuse priors

Interpretations:

(1) Start with finite τ, calculate posterior, and consider limiting properties of sequence as $\tau \to \infty$

Interpretations:

(2) Start with $\tau = \infty$?

$$p(\mu) = \frac{1}{\sqrt{2\pi}\,\tau} \exp\left[-\frac{(\mu - m)^2}{2\tau^2} \right]$$

limit as $\tau \to \infty$ is not a density

(3) Just use kernels?

$$p(\mathbf{Y}|\mu) \propto \exp\left[-\frac{(\mu^2 - 2\mu\bar{y}]}{2(\sigma^2/T)}\right]$$

$p(\mu) \propto 1$?

(diffuse prior?)

implies

$$p(\mu|\mathbf{Y}) \propto \exp\left[-\frac{(\mu^2 - 2\mu\bar{y}]}{2(\sigma^2/T)}\right]$$

$\mu|\mathbf{Y} \sim N(\bar{y}, \sigma^2/T)$

gives the correct answer in this case

But $p(\mu) \propto 1$ is not a proper density for $\mu \in \Re^1$

$p(\mu) \propto 1$ is called an "improper" prior

In this case, it gave us the correct answer.

In other cases it can fail (with either analytical or numerical methods)

Another problem with the improper prior $p(\theta) \propto 1$ is that it is not invariant with respect to reparameterization.

Example: $T = 1$

$$p(y_1|\sigma;\mu) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(y_t-\mu)^2}{2\sigma^2}\right]$$

If parameter of interest is σ^{-1} and $p(\sigma^{-1}) \propto 1$ then

$$p(\sigma^{-1}|y_1;\mu) \propto \frac{1}{\sigma} \exp\left[-\frac{(y_t-\mu)^2}{2\sigma^2}\right]$$

The constant of proportionality needed to ensure $\int_0^\infty p(\sigma^{-1}|y_1;\mu)\ d\sigma^{-1} = 1$ is

$$p(\sigma^{-1}|y_1;\mu) = \frac{(y_1-\mu)^2}{\sigma} \exp\left[-\frac{(y_1-\mu)^2}{2\sigma^2}\right]$$

Suppose instead parameter of interest is taken to be σ^{-2} and prior is $p(\sigma^{-2}) \propto 1$

$$p(\sigma^{-2}|y_1;\mu) \propto \frac{1}{(\sigma^2)^{1/2}} \exp\left[-\frac{(y_t-\mu)^2}{2\sigma^2}\right]$$

$$p(\sigma^{-2}|y_1;\mu) = \frac{[(y_1-\mu)^2]^{3/2}}{\sqrt{2\pi}}(\sigma^{-2})^{1/2} \times \exp\left[-\frac{(y_1-\mu)^2}{2\sigma^2}\right]$$

(a $\Gamma(3/2,(y_1-\mu)^2/2)$ distribution)

Problem:

$$P[\sigma^{-1} > 1|y_1;\mu] = \int_1^\infty p(\sigma^{-1}|y_1;\mu)d\sigma^{-1}$$
$$\neq \int_1^\infty p(\sigma^{-2}|y_1;\mu)d\sigma^{-2}$$
$$= P[\sigma^{-2} > 1|y_1;\mu]$$

Issue: if $\theta \sim g(\theta)$ then

$w = \phi(\theta) \sim g[\phi^{-1}(w)]\left|\frac{d\phi^{-1}(w)}{dw}\right|$

Conclusion: the "improper priors" $p(\sigma^{-1}) \propto 1$ and $p(\sigma^{-2}) \propto 1$ represent different prior beliefs

Question: which (if either) should be called a "diffuse prior" corresponding to complete uncertainty?

Jeffreys prior:

$$p(\theta) \propto [h(\theta)]^{1/2}$$

$$h(\theta) = -\int_{\Re^T} \frac{\partial^2 \log p(\mathbf{y}|\theta)}{\partial \theta^2} p(\mathbf{y}|\theta) \ d\mathbf{y}$$

for $\mathbf{y} \in \Re^T$

Example: if $\theta = \sigma^{-1}$

$$\log p(\mathbf{y}|\theta) = -(T/2)\log 2\pi + T\log \sigma^{-1}$$

$$- (1/2)\textstyle\sum_{t=1}^{T}(y_t - \mu)^2(\sigma^{-1})^2$$

$$\partial \log p(\mathbf{y}|\theta)/\partial\theta = T/\sigma^{-1} - \textstyle\sum_{t=1}^{T}(y_t - \mu)^2\sigma^{-1}$$

$$\partial^2 \log p(\mathbf{y}|\theta)/\partial\theta^2 = -T/\sigma^{-2} - \textstyle\sum_{t=1}^{T}(y_t - \mu)^2$$

$$-E[\partial^2 \log p(\mathbf{y}|\theta)/\partial\theta^2] = T\sigma^2 + T\sigma^2 = 2T\sigma^2$$

$$p(\theta) \propto [h(\theta)]^{1/2} \Rightarrow p(\sigma^{-1}) \propto \sigma$$

If we instead take $\theta = \sigma^{-2}$:

$$\log p(\mathbf{y}|\theta) = -(T/2)\log 2\pi + (T/2)\log\sigma^{-2} - (1/2)\sum_{t=1}^{T}(y_t - \mu)^2\sigma^{-2}$$

$$\partial \log p(\mathbf{y}|\theta)/\partial\theta = -T/(2\sigma^{-2}) - (1/2)\sum_{t=1}^{T}(y_t - \mu)^2$$

$$\partial^2 \log p(\mathbf{y}|\theta)/\partial\theta^2 = -T/2\sigma^{-4}$$

$$-E[\partial^2 \log p(\mathbf{y}|\theta)/\partial\theta^2] = (T/2)\sigma^4$$

$$p(\theta) \propto [h(\theta)]^{1/2} \Rightarrow p(\sigma^{-2}) \propto \sigma^2$$

Advantage of Jeffreys prior:

Probabilities implied by $p(\sigma^{-1}|\mathbf{Y};\mu)$ derived from $p(\sigma^{-1}) \propto \sigma$ are identical to those implied by $p(\sigma^{-2}|\mathbf{Y};\mu)$ derived from $p(\sigma^{-2}) \propto \sigma^2$

Note: for the Normal-gamma prior

$$p(\sigma^{-2}) = \frac{(\lambda/2)^{(N/2)}}{\Gamma(N/2)}(\sigma^{-2})^{[(N/2)-1]} \times \exp\left[-\frac{\lambda\sigma^{-2}}{2}\right]$$

we characterized the diffuse prior as $N = 0, \lambda = 0$ or

$$p(\sigma^{-2}) \propto \sigma^2$$

Concerns about Jeffreys prior:
does not seem to represent "prior ignorance" in many examples

My recommendation:
Use improper prior $p(\theta) \propto 1$ or Jeffreys prior only for guidance, checking results, or in cases where operation is well understood.
Use mildly informative prior to avoid all problems.