Foundations of Statistical Inference

Julien Berestycki

Department of Statistics
University of Oxford

MT 2016

Lecture 7 : Prior distributions. Predictive Distributions. Summarizing inference.

Constructing priors

Subjective Priors: Write down a distribution representing prior knowledge about the parameter before the data is available. If possible, build a model for the parameter. If different scientists have different priors or it is unclear how to represent prior knowledge as a distribution, then consider several different priors. Repeat the analysis and check that conclusions are insensitive to priors representing 'different points of view'.

Constructing priors

Subjective Priors: Write down a distribution representing prior knowledge about the parameter before the data is available. If possible, build a model for the parameter. If different scientists have different priors or it is unclear how to represent prior knowledge as a distribution, then consider several different priors. Repeat the analysis and check that conclusions are insensitive to priors representing 'different points of view'.

Non-Subjective Priors: Several approaches offer the promise of an 'automatic' and even 'objective' prior. We list some suggestions below (Uniform, Jeffreys, MaxEnt). In practice, if one of these priors conflicted prior knowledge, we wouldnt use it. These approaches can be useful to complete the specification of a prior distribution, once subjective considerations have been taken into account.

Uniform priors

If we use a uniform prior ($\pi(\theta) \propto$ constant) then

Uniform priors

If we use a uniform prior ($\pi(\theta) \propto$ constant) then

$$\pi(\theta|x) = L(\theta; x) / \int_{\Theta} L(\theta; x) d\theta$$

and $\int_{\Theta} L(\theta; x) d\theta$ may not be finite. Such distributions are called improper and all inference is meaningless.

Uniform priors

If we use a uniform prior ($\pi(\theta) \propto$ constant) then

$$\pi(\theta|x) = L(\theta; x) / \int_\Theta L(\theta; x) d\theta$$

and $\int_\Theta L(\theta; x) d\theta$ may not be finite. Such distributions are called improper and all inference is meaningless.

Example $X \sim \text{Exp}(1/\mu)$ and $Y = \mathbb{I}(X < 1)$ yielding $Y = y$ with $y \in \{0, 1\}$. Suppose we observe $y = 0$. Now

$$L(\mu; y) = \exp(-1/\mu)$$

so if we take $\pi(\mu) \propto 1$ for $\mu > 0$ we have

$$\pi(\mu|y) \propto \exp(-1/\mu)$$

which is improper, as $\pi(\mu|y) \to 1$ as $\mu \to \infty$ so $\int_0^\infty \exp(-1/\mu) d\mu$ cannot exist.

Jeffreys' Priors

Jeffreys reasoned as follows. If we have a rule for constructing priors it should lead to the same distribution if we apply it to θ or some other parameterization ψ with $g(\psi) = \theta$.

Jeffreys' Priors

Jeffreys reasoned as follows. If we have a rule for constructing priors it should lead to the same distribution if we apply it to θ or some other parameterization ψ with $g(\psi) = \theta$.Jeffreys took

$$\pi(\theta) \propto \sqrt{I_\theta} \qquad \text{where} \qquad I_\theta = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] \text{ is the Fisher information.}$$

Now if $g(\psi) = \theta$ then

$$\pi_\Psi(\psi) \propto \pi(g(\psi))|g'(\psi)|,$$

so Jeffreys rule should yield $\pi_\Psi(\psi) \propto \sqrt{I_{g(\psi)}}|g'(\psi)|$.

Jeffreys' Priors

Jeffreys reasoned as follows. If we have a rule for constructing priors it should lead to the same distribution if we apply it to θ or some other parameterization ψ with $g(\psi) = \theta$.Jeffreys took

$$\pi(\theta) \propto \sqrt{I_\theta} \qquad \text{where} \qquad I_\theta = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] \text{ is the Fisher information.}$$

Now if $g(\psi) = \theta$ then

$$\pi_\Psi(\psi) \propto \pi(g(\psi))|g'(\psi)|,$$

so Jeffreys rule should yield $\pi_\Psi(\psi) \propto \sqrt{I_{g(\psi)}}|g'(\psi)|$. The rule gives $\pi_\Psi(\psi) \propto \sqrt{I_\psi}$. But $I_\psi = g'(\psi)^2 I_g(\psi)$, so

$$\sqrt{I_\psi} = \sqrt{I_{g(\psi)}}|g'(\psi)|$$

and the rule is consistent in this respect.

Jeffreys' Priors

Jeffreys reasoned as follows. If we have a rule for constructing priors it should lead to the same distribution if we apply it to θ or some other parameterization ψ with $g(\psi) = \theta$.Jeffreys took

$$\pi(\theta) \propto \sqrt{I_\theta} \qquad \text{where} \qquad I_\theta = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] \text{ is the Fisher information.}$$

Now if $g(\psi) = \theta$ then

$$\pi_\Psi(\psi) \propto \pi(g(\psi))|g'(\psi)|,$$

so Jeffreys rule should yield $\pi_\Psi(\psi) \propto \sqrt{I_{g(\psi)}}|g'(\psi)|$. The rule gives $\pi_\Psi(\psi) \propto \sqrt{I_\psi}$. But $I_\psi = g'(\psi)^2 I_g(\psi)$, so

$$\sqrt{I_\psi} = \sqrt{I_{g(\psi)}}|g'(\psi)|$$

and the rule is consistent in this respect.
It is sometimes desirable (on subjective grounds) to have a prior which is invariant under reparameterization.

Higher dimensions

If $\Theta \subset \mathbb{R}^k$, and $\ell(\theta; X) = \log(f(X;\theta))$, the Fisher information

$$[I_\theta]_{i,j} = -\mathbb{E}_\theta\left(\frac{\partial^2 \ell(\theta; X)}{\partial\theta_i \partial\theta_j}\right)$$

satisfies

$$-\mathbb{E}_\theta\left(\frac{\partial^2 \ell(\theta; X)}{\partial\theta_i \partial\theta_j}\right) = \mathbb{E}_\theta\left(\frac{\partial \ell(\theta; X)}{\partial\theta_i}\frac{\partial \ell(\theta; X)}{\partial\theta_j}\right)$$

subject to regularity conditions. A k-dimensional Jeffreys' prior

$$\pi(\theta) \propto |I_\theta|^{1/2}$$

($|A| \equiv \det(A)$) is invariant under 1-1 reparameterization.

Higher dimensions

If $\Theta \subset \mathbb{R}^k$, and $\ell(\theta; X) = \log(f(X;\theta))$, the Fisher information

$$[I_\theta]_{i,j} = -\mathbb{E}_\theta\left(\frac{\partial^2 \ell(\theta; X)}{\partial\theta_i \partial\theta_j}\right)$$

satisfies

$$-\mathbb{E}_\theta\left(\frac{\partial^2 \ell(\theta; X)}{\partial\theta_i \partial\theta_j}\right) = \mathbb{E}_\theta\left(\frac{\partial \ell(\theta; X)}{\partial\theta_i}\frac{\partial \ell(\theta; X)}{\partial\theta_j}\right)$$

subject to regularity conditions. A k-dimensional Jeffreys' prior

$$\pi(\theta) \propto |I_\theta|^{1/2}$$

($|A| \equiv \det(A)$) is invariant under 1-1 reparameterization.
Exercise Verify 1 to 1 $g(\psi) = \theta$ in R^k gives $\pi_\Psi(\psi) = \sqrt{I_{g(\psi)}}\left|\frac{\partial\theta^T}{\partial\psi}\right|$.

Maximum Entropy Priors

Choose a density $\pi(\theta)$ which maximizes the entropy

$$\phi[\pi] = -\int_{\Theta} \pi(\theta) \log \pi(\theta) d\theta$$

over *functions* $\pi(\theta)$ subject to constraints on π. This is a *Calculus of Variations problem*.

Maximum Entropy Priors

Choose a density $\pi(\theta)$ which maximizes the entropy

$$\phi[\pi] = -\int_\Theta \pi(\theta) \log \pi(\theta) d\theta$$

over *functions* $\pi(\theta)$ subject to constraints on π. This is a *Calculus of Variations problem*.

Example The distribution π maximizing $\phi[\pi]$ over all densities π on $\Theta = R$, subject to

$$\int_0^\infty \pi(\theta) d\theta = 1, \quad \int_0^\infty \theta\pi(\theta) d\theta = \mu, \quad \text{and} \quad \int_0^\infty (\theta-\mu)^2 \pi(\theta) d\theta = \sigma^2,$$

(normalized with $\mathbb{E}\vartheta = \mu$ and $\mathrm{Var}(\vartheta) = \sigma^2$) is the normal density

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\theta-\mu)^2/2\sigma^2}.$$

This is a special case of the following Theorem.

Theorem

The density $\pi(\theta)$ that maximizes $\phi(\pi)$, subject to

$$\mathbb{E}[t_j(\theta)] = t_j,\ j = 1, \dots, p$$

takes the p-parameter exponential family form

$$\pi(\theta) \propto \exp\left\{\sum_{i=1}^{p} \lambda_i t_i(\theta)\right\}$$

for all $\theta \in \Theta$, where $\lambda_1, \dots, \lambda_p$ are determined by the constraints.

(For the proof see Leonard and Hsu).

This is a special case of the following Theorem.

Theorem

The density $\pi(\theta)$ *that maximizes* $\phi(\pi)$*, subject to*

$$\mathbb{E}[t_j(\theta)] = t_j,\ j = 1, \ldots, p$$

takes the p-parameter exponential family form

$$\pi(\theta) \propto \exp\left\{\sum_{i=1}^{p} \lambda_i t_i(\theta)\right\}$$

for all $\theta \in \Theta$*, where* $\lambda_1, \ldots, \lambda_p$ *are determined by the constraints.*

(For the proof see Leonard and Hsu). Example In the normal case $t_1(\theta) = \theta$, $E(t_1) = \mu$, $t_2(\theta) = (\theta - \mu)^2$, $E(t_2) = \sigma^2$ gives $\pi(\theta) \propto \exp(\lambda_1\theta + \lambda_2(\theta - \mu)^2)$. Impose the constraints to get $\lambda_1 = 0$ and $\lambda_2 = -1/2\sigma^2$.

Example

Suppose prior probabilities are specified so that

$$P(a_{j-1} < \vartheta \leq a_j) = \phi_j, j = 1, \dots, p$$

with $\sum_j \phi_j = 1$ and

$$\vartheta \in (a_0, a_p),\ a_0 \leq a_1 \cdots \leq a_p \leq a_p.$$

We find the maximum entropy distribution subject to these conditions.

Example

Suppose prior probabilities are specified so that

$$P(a_{j-1} < \vartheta \leq a_j) = \phi_j, j = 1, \ldots, p$$

with $\sum_j \phi_j = 1$ and

$$\vartheta \in (a_0, a_p),\ a_0 \leq a_1 \cdots \leq a_p \leq a_p.$$

We find the maximum entropy distribution subject to these conditions. The conditions are equivalent to

$$\mathbb{E}[t_j(\vartheta)] = \phi_j, j = 1, \ldots, p$$

where $t_j(\vartheta) = \mathbb{I}[a_{j-1} < \vartheta \leq a_j]$. The posterior density of ϑ is

$$\pi(\theta) \propto \exp\left\{\sum_{j=1}^{p} \lambda_j \mathbb{I}[a_{j-1} < \theta \leq a_j]\right\},\ a_0 \leq \theta \leq a_p$$

where $\lambda_1, \ldots, \lambda_p$ are determined by the conditions. $\pi(\theta)$ is a histogram, with intervals $(a_0, a_1], (a_1, a_2], \ldots, (a_{p-1}, a_p]$.

Uninformative priors summary

- Uniform prior: $\pi(\theta) \propto \mathit{Cstt}$,
- Jeffrey's prior: $\pi(\theta) \propto \sqrt{I_\theta}$
- Entropy maximization: choose π to maximize $\Phi[\pi] = -\int_\Theta \pi(\theta) \log \pi(\theta) d\theta$ under constraints on π.

Priors for Exponential Families

Conjugate prior for an exponential family

$$f(x \mid \theta) = \exp\left\{ \sum_{j=1}^{k} A_j(\theta) \sum_{i=1}^{n} B_j(x_i) + \sum_{i=1}^{n} C(x_i) + nD(\theta) \right\}$$

Priors for Exponential Families

Conjugate prior for an exponential family

$$f(x \mid \theta) = \exp\left\{\sum_{j=1}^{k} A_j(\theta)\sum_{i=1}^{n} B_j(x_i) + \sum_{i=1}^{n} C(x_i) + nD(\theta)\right\}$$

The prior distribution based on sufficient statistics

$$\pi(\theta) \propto \exp\{\tau_0 D(\theta) + \sum_{j=1}^{k} A_j(\theta)\tau_j\}$$

(where $(\tau_0, \ldots, \tau_k)$ are constant prior parameters) is conjugate.

Priors for Exponential Families

The posterior density is proportional to

$$f(x \mid \theta)\pi(\theta \mid \tau_0, \ldots, \tau_k)$$

Priors for Exponential Families

The posterior density is proportional to

$$
\begin{aligned}
& f(x \mid \theta)\pi(\theta \mid \tau_0, \dots, \tau_k) \\
& \propto \exp\left\{ \sum_{j=1}^{k} A_j(\theta) \left[\sum_{i=1}^{n} B_j(x_i) + \tau_j \right] + (n + \tau_0) D(\theta) \right\}
\end{aligned}
$$

Priors for Exponential Families

The posterior density is proportional to

$$f(x \mid \theta)\pi(\theta \mid \tau_0, \ldots, \tau_k)$$
$$\propto \exp\left\{ \sum_{j=1}^{k} A_j(\theta) \left[\sum_{i=1}^{n} B_j(x_i) + \tau_j \right] + (n + \tau_0) D(\theta) \right\}$$

This is an updated form of the prior with

$$\begin{aligned} B_j'(x) &= \sum_{i=1}^{n} B_j(x_i) + \tau_j \\ n' &= n + \tau_0 \end{aligned}$$

Priors for Exponential Families: Example

$X_1, X_2, \ldots$ iid Poisson(θ).

$$p(y|\theta) \propto e^{-n\theta}\theta^{t(y)}, \qquad t(y) = \sum_{i=1}^{n} y_i.$$

Exponential with natural parameter $\phi(\theta) = \log\theta$.

Priors for Exponential Families: Example

$X_1, X_2, \ldots$ iid Poisson(θ).

$$p(y|\theta) \propto e^{-n\theta}\theta^{t(y)}, \qquad t(y) = \sum_{i=1}^{n} y_i.$$

Exponential with natural parameter $\phi(\theta) = \log\theta$. $D(\theta) = -n\theta$ so that the natural conjugate distribution

$$\pi(\theta) \propto e^{-\beta\theta + (\alpha-1)\log\theta}.$$

Gamma density with parameters (α, β).

Priors for Exponential Families: Example

$X_1, X_2, \ldots$ iid Poisson(θ).

$$p(y|\theta) \propto e^{-n\theta}\theta^{t(y)}, \qquad t(y) = \sum_{i=1}^{n} y_i.$$

Exponential with natural parameter $\phi(\theta) = \log\theta$. $D(\theta) = -n\theta$ so that the natural conjugate distribution

$$\pi(\theta) \propto e^{-\beta\theta + (\alpha-1)\log\theta}.$$

Gamma density with parameters (α, β).
Exercise: check that $p(\theta|y) \sim \text{Gamma}(\alpha + n\bar{y}, \beta + n)$.

Predictive distributions

$X_1, \ldots, X_n$ are observations from $f(x; \theta)$ and the predictive distribution of a further observation X_{n+1} is required.

Predictive distributions

$X_1, \ldots, X_n$ are observations from $f(x;\theta)$ and the predictive distribution of a further observation X_{n+1} is required.

Lemma

If $x = (x_1, \ldots, x_n)$ are iid from $f(x;\theta)$ then the posterior predictive distribution is

$$g(x_{n+1} \mid x) = \int f(x_{n+1};\theta)\pi(\theta \mid x)d\theta$$

Predictive distributions

$X_1, \ldots, X_n$ are observations from $f(x;\theta)$ and the predictive distribution of a further observation X_{n+1} is required.

Lemma

If $x = (x_1, \ldots, x_n)$ are iid from $f(x;\theta)$ then the posterior predictive distribution is

$$g(x_{n+1} \mid x) = \int f(x_{n+1};\theta)\pi(\theta \mid x)d\theta$$

Predictive distributions are useful for ...

Predictive distributions

$X_1, \ldots, X_n$ are observations from $f(x;\theta)$ and the predictive distribution of a further observation X_{n+1} is required.

Lemma

If $x = (x_1, \ldots, x_n)$ are iid from $f(x;\theta)$ then the posterior predictive distribution is

$$g(x_{n+1} \mid x) = \int f(x_{n+1};\theta)\pi(\theta \mid x)d\theta$$

Predictive distributions are useful for ...prediction.

Predictive distributions

$X_1, \ldots, X_n$ are observations from $f(x;\theta)$ and the predictive distribution of a further observation X_{n+1} is required.

Lemma

If $x = (x_1, \ldots, x_n)$ are iid from $f(x;\theta)$ then the posterior predictive distribution is

$$g(x_{n+1} \mid x) = \int f(x_{n+1};\theta)\pi(\theta \mid x)d\theta$$

Predictive distributions are useful for ...prediction.

They are used also for model checking. Divide the data in two groups, $Y = (X_1, ..., X_a)$ and $Z = (X_{a+1}, ..., X_n)$. If we fit using Y and check that the 'reserved data' Z overlap $g(x_{n+1} \mid x)$ in distribution.

Poisson example cont'd

The “prior predictive" distribution is just the marginal. Using

$$p(y) = \int p(y|\theta)\pi(\theta)d\theta = \frac{p(y|\theta)\pi(\theta)}{p(\theta|y)}$$

Poisson example cont'd

The “prior predictive" distribution is just the marginal. Using

$$p(y) = \frac{p(y|\theta)\pi(\theta)}{p(\theta|y)} \quad \text{we get } \; p(y) = \frac{\text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)}{\text{Gamma}(\theta|\alpha+y, 1+\beta)}$$

which reduces to

$$p(y) = \binom{\alpha+y-1}{y}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y}, \quad y \sim \text{ Neg-bin } (\alpha,\beta).$$

In other words

$$\text{Neg-bin } (y|\alpha,\beta) = \int \text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)d\theta.$$

Poisson example cont'd

The "prior predictive" distribution is just the marginal. Using

$$p(y) = \frac{p(y|\theta)\pi(\theta)}{p(\theta|y)} \quad \text{we get} \quad p(y) = \frac{\text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)}{\text{Gamma}(\theta|\alpha+y,1+\beta)}$$

which reduces to

$$p(y) = \binom{\alpha+y-1}{y}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y}, \quad y \sim \text{ Neg-bin } (\alpha,\beta).$$

In other words

$$\text{Neg-bin } (y|\alpha,\beta) = \int \text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)d\theta.$$

Therefore

$$p(y_{n+1}|y) = \int \text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha+n\bar{y},\beta+n)d\theta$$
$$\sim \text{ Neg-bin } (y|\alpha+n\bar{y},\beta+n).$$

Example : Normal with known variance

Data $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$ with σ^2 known and prior $\theta \sim N(\mu_0, \sigma_0^2)$.
Predict X_{n+1}.

Example : Normal with known variance

Data $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$ with σ^2 known and prior $\theta \sim N(\mu_0, \sigma_0^2)$. Predict X_{n+1}.

$$p(\theta|y) \propto \pi(\theta)p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2(y_i-\theta)^2}\right)$$

Example : Normal with known variance

Data $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$ with σ^2 known and prior $\theta \sim N(\mu_0, \sigma_0^2)$. Predict X_{n+1}.

$$p(\theta|y) \propto \pi(\theta)p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2(y_i-\theta)^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]\right)$$

Example : Normal with known variance

Data $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$ with σ^2 known and prior $\theta \sim N(\mu_0, \sigma_0^2)$. Predict X_{n+1}.

$$p(\theta|y) \propto \pi(\theta)p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2(y_i-\theta)^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta-\mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^n (y_i-\theta)^2\right]\right)$$

Example : Normal with known variance

Data $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$ with σ^2 known and prior $\theta \sim N(\mu_0, \sigma_0^2)$. Predict X_{n+1}.

$$p(\theta|y) \propto \pi(\theta)p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2(y_i-\theta)^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]\right)$$

Complete the squares to obtain

$$p(\theta|y) = p(\theta|\bar{y}) = N(\theta|\mu_n, \sigma_n^2)$$

where

$$\mu_n = \frac{\sigma_0^{-2}\mu_0 + n\sigma^{-2}\bar{y}}{\sigma_0^{-2} + n\sigma^{-2}} \text{ and } \sigma_n^{-2} = \sigma_0^{-2} + n\sigma^{-2}.$$

Example : Normal with known variance

Data $X_1, \dots, X_n$ are iid $N(\theta, \sigma^2)$ with σ^2 known and prior $\theta \sim N(\mu_0, \sigma_0^2)$. Predict X_{n+1}.

$$p(\theta|y) \propto \pi(\theta)p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2(y_i-\theta)^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta-\mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^n (y_i-\theta)^2\right]\right)$$

Complete the squares to obtain

$$p(\theta|y) = p(\theta|\bar{y}) = N(\theta|\mu_n, \sigma_n^2)$$

where

$$\mu_n = \frac{\sigma_0^{-2}\mu_0 + n\sigma^{-2}\bar{y}}{\sigma_0^{-2} + n\sigma^{-2}} \text{ and } \sigma_n^{-2} = \sigma_0^{-2} + n\sigma^{-2}.$$

Observe 1) that if $\sigma_0^2 = \sigma^2$ then the prior has same weight as one extra observation! 2) If n large then $p(\theta|y) \approx N(\theta|\bar{y}, \sigma^2/n)$.

Example : Normal with known variance

In order to calculate the posterior predictive density for X_{n+1} we need to evaluate

$$g(x_{n+1} \mid x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\theta-\mu_n)^2}{2\sigma_1^2}} d\theta$$

Example : Normal with known variance

In order to calculate the posterior predictive density for X_{n+1} we need to evaluate

$$g(x_{n+1} \mid x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\theta-\mu_n)^2}{2\sigma_1^2}} d\theta$$

We could complete the square to solve this.

Example : Normal with known variance

In order to calculate the posterior predictive density for X_{n+1} we need to evaluate

$$g(x_{n+1} \mid x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\theta-\mu_n)^2}{2\sigma_1^2}} d\theta$$

We could complete the square to solve this.Alternatively, think how X_{n+1} is built up.

Example : Normal with known variance

In order to calculate the posterior predictive density for X_{n+1} we need to evaluate

$$g(x_{n+1} \mid x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\theta-\mu_n)^2}{2\sigma_1^2}} d\theta$$

We could complete the square to solve this.Alternatively, think how X_{n+1} is built up.
We have $\theta|X \sim N(\mu_n, \sigma_n^2)$ and $X_{n+1} \sim \theta + N(0, \sigma^2)$.

Example : Normal with known variance

In order to calculate the posterior predictive density for X_{n+1} we need to evaluate

$$g(x_{n+1} \mid x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\theta-\mu_n)^2}{2\sigma_1^2}} d\theta$$

We could complete the square to solve this.Alternatively, think how X_{n+1} is built up.
We have $\theta|X \sim N(\mu_n, \sigma_n^2)$ and $X_{n+1} \sim \theta + N(0, \sigma^2)$.
If $Y, Z \sim N(0, 1)$ then

$$\theta = \mu_n + \sigma_n Z + \sigma Y.$$

Example : Normal with known variance

In order to calculate the posterior predictive density for X_{n+1} we need to evaluate

$$g(x_{n+1} \mid x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\theta-\mu_n)^2}{2\sigma_1^2}} d\theta$$

We could complete the square to solve this.Alternatively, think how X_{n+1} is built up.
We have $\theta|X \sim N(\mu_n, \sigma_n^2)$ and $X_{n+1} \sim \theta + N(0, \sigma^2)$.
If $Y, Z \sim N(0, 1)$ then

$$\theta = \mu_n + \sigma_n Z + \sigma Y.$$

It follows that $X_{n+1} \sim N(\mu_n, \sigma^2 + \sigma_n^2)$ is the posterior predictive density for $X_{n+1}|X_1, ..., X_n$.

Summarizing posterior inference

The posterior $p(\theta|y)$ contains all current information.

- Graphical display
- Contour and scatter plots in multidimensional cases

Summarizing posterior inference

The posterior $p(\theta|y)$ contains all current information.

- Graphical display
- Contour and scatter plots in multidimensional cases

Summary statistics

- mean, median, mode
- Standard deviation
- Central interval, highest posterior density interval (HPD).

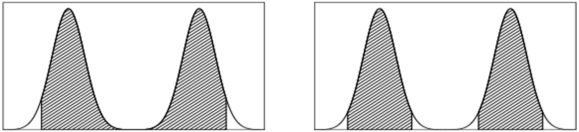
Summarizing posterior inference

The posterior $p(\theta|y)$ contains all current information.

- Graphical display
- Contour and scatter plots in multidimensional cases

Summary statistics

- mean, median, mode
- Standard deviation
- Central interval, highest posterior density interval (HPD).



Normal approximation

If $p(\theta|y)$ is unimodal and roughly symmetric, the a rough normal approximation can be convenient.

Normal approximation

If $p(\theta|y)$ is unimodal and roughly symmetric, the a rough normal approximation can be convenient. Taylor expansion around its modes $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right] (\theta - \hat{\theta}) + \dots$$

Normal approximation

If $p(\theta|y)$ is unimodal and roughly symmetric, the a rough normal approximation can be convenient. Taylor expansion around its modes $\hat{\theta}$

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right] (\theta - \hat{\theta}) + \dots$$

$$p(\theta|y) \approx N(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

Where $I(\theta)$ is the observed information

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

positive definite if $\hat{\theta}$ in the interior of Θ.

Normal approximation: example

$(y_1, \ldots, y_n)$ iid $N(\mu, \sigma^2)$. Prior is uniform density for $(\mu, \log \sigma)$.

Normal approximation: example

$(y_1, \ldots, y_n)$ iid $N(\mu, \sigma^2)$. Prior is uniform density for $(\mu, \log \sigma)$.
Observe that

$$\log p(\mu, \log \sigma | y) = -n \log \sigma - \frac{1}{2\sigma^2}((n-1)s^2 + n(\bar{y} - \mu)^2) + cstt$$

so that

$$\frac{d}{d\mu} \log p(\mu, \log \sigma | y) = \frac{n(\bar{y} - \mu)}{\sigma^2}$$

$$\frac{d}{d(\log \sigma)} \log p(\mu, \log \sigma | y) = -n + \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{\sigma^2}$$

so the posterior mode is

$$(\hat{\mu}, \log \hat{\sigma}) = \left(\bar{y}, \log \left(\sqrt{\frac{n-1}{n}} s \right) \right).$$

The second derivatives are

$$\frac{d^2}{d\mu^2}\log p(\mu,\log\sigma|y) = -\frac{n}{\sigma^2}$$

$$\frac{d^2}{d\mu d(\log\sigma)}\log p(\mu,\log\sigma|y) = -2n\frac{(\bar{y}-\mu)}{\sigma^2}$$

$$\frac{d^2}{d(\log\sigma)^2}\log p(\mu,\log\sigma|y) = -\frac{2}{\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2)$$

and thus

$$p(\mu,\log\sigma|y) \approx N\left(\left(\begin{array}{c}\mu\\ \log\sigma\end{array}\right)\Big|\left(\begin{array}{c}\bar{y}\\ \log(\hat{\sigma})\end{array}\right),\left(\begin{array}{cc}\hat{\sigma}^2/n & 0\\ 0 & 1/(2n)\end{array}\right)\right)$$

The second derivatives are

$$\frac{d^2}{d\mu^2}\log p(\mu,\log\sigma|y) = -\frac{n}{\sigma^2}$$

$$\frac{d^2}{d\mu d(\log\sigma)}\log p(\mu,\log\sigma|y) = -2n\frac{(\bar{y}-\mu)}{\sigma^2}$$

$$\frac{d^2}{d(\log\sigma)^2}\log p(\mu,\log\sigma|y) = -\frac{2}{\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2)$$

and thus

$$p(\mu,\log\sigma|y) \approx N\left(\left(\begin{array}{c}\mu\\ \log\sigma\end{array}\right)\middle|\left(\begin{array}{c}\bar{y}\\ \log(\hat{\sigma})\end{array}\right),\left(\begin{array}{cc}\hat{\sigma}^2/n & 0\\ 0 & 1/(2n)\end{array}\right)\right)$$

Exercise: Check that if we had conducted the analysis in terms of $p(\mu,\sigma^2)$ the second derivative matrix would be multiplied by the Jacobian of $\log\sigma \mapsto \sigma^2$ yielding a mode $\tilde{\sigma}^2 = \frac{n}{n+2}\hat{\sigma}^2$, the approx. post. dist would still have independent components with $p(\sigma^2|y) \approx N(\sigma^2|\tilde{\sigma}^2, 2\tilde{\sigma}^4/(n+2))$.

Large-sample theory

Suppose that the y_i are iid from some distribution $f(y)$ (the true distribution),

Large-sample theory

Suppose that the y_i are iid from some distribution $f(y)$ (the true distribution), that the data is modeled by a parametric family $p(y|\theta)$ with prior $\pi(\theta)$,

Large-sample theory

Suppose that the y_i are iid from some distribution $f(y)$ (the true distribution), that the data is modeled by a parametric family $p(y|\theta)$ with prior $\pi(\theta)$, and that $f(y) = p(y|\theta_0)$.

Theorem

Suppose that π is continuous and θ_0 is in the interior of the parameter space. For any neighborhood A of θ_0

$$\lim_{n\to\infty} \mathbb{P}(\theta \in A|y) = 1.$$

Large-sample theory

Suppose that the y_i are iid from some distribution $f(y)$ (the true distribution), that the data is modeled by a parametric family $p(y|\theta)$ with prior $\pi(\theta)$, and that $f(y) = p(y|\theta_0)$.

Theorem

Suppose that π is continuous and θ_0 is in the interior of the parameter space. For any neighborhood A of θ_0

$$\lim_{n \to \infty} \mathbb{P}(\theta \in A|y) = 1.$$

Furtthermore

$$p(\theta|y) \approx N(\theta_0, (nJ(\theta_0))^{-1})$$

where J is the Fisher information.