

Foundations of Statistical Inference

J. Berestycki & D. Sejdinovic

Department of Statistics
University of Oxford

MT 2019

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'.

Non-Subjective Priors: Several approaches offer the promise of an 'automatic' and even 'objective' prior. We list some suggestions below (Jeffreys, MaxEnt). They can be used in context of small or non reliable prior information or as references. These approaches can also be useful to complete the specification of a prior distribution, once subjective considerations have been taken into account.

Conjugate priors

Definition

Consider a sampling model $f(X; \theta), \theta \in \Theta$ for observables X . We say that a family of prior distributions $(\pi_\gamma, \gamma \in \Gamma)$ is conjugate if for all $\gamma \in \Gamma$ and all $x \in \mathcal{X}$, there exists $\gamma(x)$ such that

$$\pi_\gamma(\cdot|x) = \pi_{\gamma(x)}(\cdot)$$

Example : Normal distribution when the mean and variance are unknown.

$$X = (X_1, \dots, X_n), \quad X_i \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad \tau = 1/\sigma^2, \quad \theta = (\tau, \mu)$$

τ is called the precision.

Prior

$$[\mu|\sigma^2] \sim \mathcal{N}(\nu, \kappa\sigma^2), \quad \tau \sim \Gamma(\alpha, \beta), \quad \nu \in \mathbb{R}$$

The prior is

$$\pi(\tau, \mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \cdot (2\pi\kappa)^{-1/2} \tau^{1/2} \exp \left\{ -\frac{\tau}{2\kappa} (\mu - \nu)^2 \right\}$$

or

$$\pi(\tau, \mu) \propto \tau^{\alpha-1/2} \exp \left[-\tau \left\{ \beta + \frac{1}{2\kappa} (\mu - \nu)^2 \right\} \right]$$

Normal example - la suite

The likelihood is

$$f(x \mid \mu, \tau) = (2\pi)^{-n/2} \tau^{n/2} \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

Thus

$$\pi(\tau, \mu \mid x) \propto \tau^{\alpha+(n/2)-1/2} \exp \left[-\tau \left\{ \beta + \frac{1}{2\kappa} (\mu - \nu)^2 + \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \right]$$

Complete the square to see that

$$\begin{aligned} & (\mu - \nu)^2 / \kappa + \sum (x_i - \mu)^2 \\ &= (\kappa^{-1} + n) \left(\mu - \frac{\kappa^{-1} \nu + n \bar{x}}{\kappa^{-1} + n} \right)^2 + \frac{n}{n\kappa + 1} (\bar{x} - \nu)^2 + \sum (x_i - \bar{x})^2 \end{aligned}$$

Example Normal - the end

Thus the posterior is

$$\pi(\tau, \mu \mid x) \propto \tau^{\alpha'-1/2} \exp \left[-\tau \left\{ \beta' + \frac{1}{2\kappa'} (\nu' - \mu)^2 \right\} \right]$$

where

$$\begin{aligned} \alpha' &= \alpha + \frac{n}{2}, & \kappa' &= (n\kappa + 1)/n, & \nu' &= \frac{\kappa^{-1}\nu + n\bar{x}}{\kappa^{-1} + n} \\ \beta' &= \beta + \frac{1}{2} \cdot \frac{n}{n\kappa + 1} (\bar{x} - \nu)^2 + \frac{1}{2} \sum (x_i - \bar{x})^2 \end{aligned}$$

This is the same form as the prior, so the class is conjugate.

Marginalisation : posterior on μ

If we are interested in the posterior distribution of μ alone

$$\begin{aligned}\pi(\mu|x) &= \int \pi(\tau, \mu|x) d\tau \\ &\propto \int_0^\infty \tau^{\alpha'-1/2} \exp \left[-\tau \left\{ \beta' + \frac{1}{2\kappa'} (\nu' - \mu)^2 \right\} \right] d\tau\end{aligned}$$

We recognize a $\Gamma(\alpha' + 1/2, \beta' + (\nu' - \mu)^2/(2\kappa'))$ for τ so that

$$\pi(\mu|x) \propto (2\beta' + (\nu' - \mu)^2/\kappa')^{-(\alpha'+1/2)} \equiv \text{Student}(\alpha' + 1/2, \nu', (2\beta'\kappa')^{-1})$$

Conjugate priors for Exponential Families

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

The following family of priors is conjugate:

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

where $\tau = (\tau_0, \dots, \tau_k)$ are constant prior parameters

Priors for Exponential Families

The posterior density is proportional to

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

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, \dots iid $\text{Poisson}(\theta)$.

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

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 $p(\theta|y) \sim \text{Gamma}(\alpha + n\bar{y}, \beta + n)$.

About conjugate priors

- ▶ They are mathematically practical
- ▶ They are not justified on other grounds – but in some cases mathematical ease is important
- ▶ It is often easy to interpret the hyper parameters

Improper priors

In the Bayesian paradigm

$$[X|\theta] \sim f(x;\theta), \quad \theta \sim \pi$$

both are **probability densities**

We can generalize

Definition

We say that a prior distribution is **improper** if its mass is infinite

$$\int_{\Theta} \pi(\theta) d\theta = +\infty, \quad \pi(\theta) \geq 0$$

The posterior distribution is defined as soon as

$$\int_{\Theta} f(x;\theta) \pi(\theta) d\theta < +\infty, \quad \text{almost surely in } x$$

Examples

- ▶ Gaussian + Lebesgue measure **exercise**:

$$X \sim \mathcal{N}(\mu, 1), \quad \pi(\mu) = 1$$

- ▶ $X \sim \mathcal{B}(n, p)$ and $\pi(p) = [p(1-p)]^{-1}$: Haldane prior . Although it is used in some cases the posterior is not well defined.

$$\pi(p|x) \propto p^{x-1}(1-p)^{n-x-1} \quad \text{improper if } x = 0 \quad \text{or} \quad x = n$$

and for all $p \in]0, 1[$

$$P(X = 0; p) + P(X = n; p) > 0$$

Remark general case : if X is a discrete distribution then one cannot use an improper prior **Exercise : prove it**

Non-informative priors

When do we want to use noninformative priors ? Some will say never don't believe them !

There are cases where there is little or no prior information - at least on some aspects of the parameters.

How can we then choose a prior?

Useful to determine automatic procedures so that the choice becomes less arbitrary.

Uniform priors - a first (bad) choice

(Naive) representation of lack of information If little information one doesn't want to privilege some region of the space compared to others This leads to a *flat prior* :

$$\pi(\theta) = \textit{constant} = 1$$

i.e. Lebesgue measure on $\Theta \subset \mathbb{R}^d$. Then

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

exists if

$$\int_{\Theta} f(x; \theta) d\theta < +\infty, \quad \text{almost surely in } x$$

Example $X \sim \text{Exp}(\theta)$ $\pi(\theta) = 1$

$$\int_0^\infty e^{-\theta x} \theta d\theta < +\infty \Leftrightarrow x > 0$$

and for all $\theta > 0$ $P[X = 0; \theta] = 0$ The posterior is well defined But is it a *good* posterior ?

Not really because it is not invariant par parameterization. Let $\eta = \log \theta$

$$\tilde{\pi}(\eta) = \pi(\theta(\eta)) \frac{d\theta}{d\eta} = \frac{d\theta}{d\eta} = e^\eta \neq 1$$

As a prior in η $\tilde{\pi}$ is very *informative*

Remark This is also true for *weakly* informative priors like $\mathcal{N}(0, V)$ with V large.

Interesting only if θ is a discrete parameter

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

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

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.

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 \mathbb{R}^k gives $\pi_\Psi(\psi) = \sqrt{|I_{g(\psi)}|} \left| \frac{\partial \theta^T}{\partial \psi} \right|$.

Some comments on Jeffreys prior

Jeffreys prior is justified **in the case** $d = 1$ using various criteria of *non informativeness*

- ▶ Matching priors : Credible regions in the form

$$\{\theta \leq \theta_\alpha(x)\}, \quad \text{with} \quad \Pi(\theta \leq \theta_\alpha(x)|x) = 1 - \alpha$$

satisfy

$$P[\theta \leq \theta_\alpha(x); \theta] = 1 - \alpha + O(1/n) \quad \forall \theta \in \Theta$$

if and only if $\pi(\theta) = \sqrt{I_\theta}$ (Jeffreys)

- ▶ Reference prior : maximizes (asymptotically) the Kullback- Leibler distance (Shannon information) between the posterior and the prior
maximizes the relative weight of the likelihood (data) compared to the prior in terms of information brought to the inference

Not so good in the multivariate case: there exist extensions such as the reference prior approach

Partially informative priors : Maximum Entropy Priors

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

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

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

Example The distribution π maximizing $\text{Ent}[\pi]$ over all densities π on $\Theta = \mathbb{R}$, 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 $\text{Var}(\vartheta) = \sigma^2$) 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 $\text{Ent}(\pi)$, subject to

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

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).

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

Heuristics behind this approach

If $\Theta = \{1, \dots, N\}$

$$\sum_{i=1}^N -\log \pi_i \times \pi_i \leq \log N \quad \forall (\pi_1, \dots, \pi_N) \in \mathcal{S}_N \quad \textit{simplex}$$

with equality iff

$$\pi_1 = \dots = \pi_N = 1/N$$

with the idea that the uniform prior : most neutral prior (greatest disorder)

Hence Maximum entropy prior : looks for the least informative prior under some specific prior information constraints.

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

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

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

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

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

Comments on Max Ent priors

- ▶ The construction is independent of the model and the meaning of the parameter .
- ▶ It does not always exist. e.g. if the constraint is just $E(\theta) = \mu$

non informative priors summary

- ▶ Uniform prior: $\pi(\theta) \propto C_{stt}$: false good idea unless θ is discrete
- ▶ Jeffrey's prior: $\pi(\theta) \propto \sqrt{I_\theta}$
- ▶ partially informative prior : Entropy maximization: choose π to maximize $\text{Ent}[\pi] = - \int_{\Theta} \pi(\theta) \log \pi(\theta) d\theta$ under constraints on π .

Predictive distributions

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

Lemma

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

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

Predictive distributions are useful for ... prediction.

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

Poisson example cont'd

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

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

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

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

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

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

Therefore

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

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} .

$$\begin{aligned} p(\theta|y) &\propto \pi(\theta)p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right) \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta - \mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \theta)^2\right]\right) \end{aligned}$$

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

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} .

$$\begin{aligned} p(\theta|y) &\propto \pi(\theta)p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right) \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta - \mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \theta)^2\right]\right) \end{aligned}$$

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

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} .

$$\begin{aligned} p(\theta|y) &\propto \pi(\theta)p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right) \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\theta - \mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \theta)^2\right]\right) \end{aligned}$$

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

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

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 $X_{n+1}|X_1, \dots, X_n$.

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).

