A Very Brief Summary of Bayesian Inference, and Examples

Trinity Term 2009 Prof. Gesine Reinert

Our starting point are data $\mathbf{x} = x_1, x_2, \ldots, x_n$, which we view as realisations of random variables $X_1, X_2, \ldots, X_n$ with distribution (model) $f(x_1, x_2, \ldots, x_n|\theta)$. In the *Bayesian* framework, $\theta \in \Theta$ is random, and follows a prior distribution $\pi(\theta)$.

1. Priors, Posteriors, Likelihood, and Sufficiency

The posterior distribution of θ given x is

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

Abbreviating, we write: posterior $\propto$ prior $\times$ likelihood.

The *(prior) predictive distribution* of the data x on the basis π is

$$p(x) = \int f(x|\theta)\pi(\theta)d\theta.$$

Suppose data x_1 is available, and we want to predict additional data: the predictive distribution is

$$p(x_2|x_1) = \int f(x_2|\theta)\pi(\theta|x_1)d\theta.$$

Note that x_2 and x_1 are assumed conditionally independent given θ. They are **not**, in general, unconditionally independent.

Example. Suppose $y_1, y_2, \ldots, y_n$ are independent normally distributed random variables, each with variance 1 and with means $\beta x_1, \ldots, \beta x_n$, where β is an unknown real-valued parameter and $x_1, x_2, \ldots, x_n$ are known constants. Suppose as prior $\pi(\beta) = \mathcal{N}(\mu, \alpha^2)$. Then the likelihood is

$$L(\beta) = (2\pi)^{-\frac{n}{2}} \prod_{i=1}^{n} e^{-\frac{(y_i - \beta x_i)^2}{2}}$$

and

$$\pi(\beta) = \frac{1}{\sqrt{2\pi\alpha^2}} e^{-\frac{(\beta-\mu)^2}{2\alpha^2}} \propto \exp\left\{-\frac{\beta^2}{2\alpha^2} + \beta\frac{\mu}{\alpha^2}\right\}.$$

The posterior of β given $y_1, y_2, \ldots, y_n$ can be calculated as

$$\begin{aligned}\pi(\beta|\mathbf{y}) &\propto \exp\left\{-\frac{1}{2}\sum(y_i - \beta x_i)^2 - \frac{1}{2\alpha^2}(\beta-\mu)^2\right\} \\ &\propto \exp\left\{\beta\sum x_i y_i - \frac{1}{2}\beta^2\sum x_i^2 - \frac{\beta^2}{2\alpha^2} + \frac{\beta\mu}{\alpha^2}\right\}.\end{aligned}$$

Abbreviate $s_{xx} = \sum x_i^2$, $s_{xy} = \sum x_i y_i$, then

$$\begin{aligned}\pi(\beta|\mathbf{y}) &\propto \exp\left\{\beta s_{xy} - \frac{1}{2}\beta^2 s_{xx} - \frac{\beta^2}{2\alpha^2} + \frac{\beta\mu}{\alpha^2}\right\} \\ &= \exp\left\{-\frac{\beta^2}{2}\left(s_{xx} + \frac{1}{\alpha^2}\right) + \beta\left(s_{xy} + \frac{\mu}{\alpha^2}\right)\right\},\end{aligned}$$

which we recognize as

$$\mathcal{N}\left(\frac{s_{xy} + \frac{\mu}{\alpha^2}}{s_{xx} + \frac{1}{\alpha^2}}, \left(s_{xx} + \frac{1}{\alpha^2}\right)^{-1}\right).$$

To summarize information about θ we find a *minimal sufficient statistic* $t(\mathbf{x})$; $T = t(\mathbf{X})$ is sufficient for θ if and only if for all $\mathbf{x}$ and θ

$$\pi(\theta|\mathbf{x}) = \pi(\theta|t(\mathbf{x})).$$

As in the frequentist approach; posterior (inference) is based on sufficient statistics.

Choice of prior

There are many ways of choosing a prior distribution; using a coherent belief system, e.g.. Often it is convenient to restrict the class of priors to a particular family of distributions. When the posterior is in the same family of models as the prior, i.e. when one has a conjugate prior, then updating the distribution under new data is particularly convenient.

For the regular k-parameter exponential family,

$$f(x|\theta) = f(x)g(\theta)exp\{\sum_{i=1}^{k} c_i\phi_i(\theta)h_i(x)\}, \quad x \in \mathcal{X},$$

conjugate priors can be derived in a straightforward manner, using the sufficient statistics.

Non-informative priors favour no particular values of the parameter over others. If Θ is finite, choose we uniform prior. If Θ is infinite, there are several ways (some may be improper):

For a location density $f(x|\theta) = f(x-\theta)$, then the non-informative location-invariant prior is $\pi(\theta) = \pi(0)$ constant; we usually choose $\pi(\theta) = 1$ for all θ (improper prior).

For a scale density $f(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$ for $\sigma > 0$, the scale-invariant non-informative prior is $\pi(\sigma) \propto \frac{1}{\sigma}$; usually we choose $\pi(\sigma) = \frac{1}{\sigma}$ (improper).

The *Jeffreys prior* is $\pi(\theta) \propto I(\theta)^{\frac{1}{2}}$ if the information $I(\theta)$ exists; then is invariant under reparametrization; the Jeffreys prior may or may not be improper.

Example continued. Suppose $y_1, y_2, \ldots, y_n$ are independent, $y_i \sim \mathcal{N}(\beta x_i, 1)$, where β is an unknown parameter and $x_1, x_2, \ldots, x_n$ are known. Then

$$\begin{aligned} I(\beta) &= -E\left(\frac{\partial^2}{\partial\beta^2}\log L(\beta, \mathbf{y})\right) \\ &= -E\left(\frac{\partial}{\partial\beta}\sum(y_i - \beta x_i)x_i\right) \\ &= s_{xx} \end{aligned}$$

and so the Jeffreys prior is $\propto \sqrt{s_{xx}}$; constant and improper. With this Jeffreys prior, the calculation of the posterior distribution is equivalent to putting $\alpha^2 = \infty$ in the previous calculation, yielding

$$\pi(\beta|\mathbf{y}) \text{ is } \mathcal{N}\left(\frac{s_{xy}}{s_{xx}}, (s_{xx})^{-1}\right).$$

If Θ is discrete, and satisfies the constraints

$$E_\pi g_k(\theta) = \mu_k, \quad k = 1, \ldots, m$$

then we choose the distribution with the maximum entropy under these constraints; it is given by:

$$\tilde{\pi}(\theta_i) = \frac{exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right)}{\sum_i exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right)}$$

where the λ_i are determined by the constraints. There is a similar formula for the continuous case, maximizing the entropy relative to a particular reference distribution π_0 under constraints. For π_0 one would choose the "natural" invariant noninformative prior.

For inference, we check the influence of the choice of prior, for example by trying out different priors.

2. Point Estimation

Under suitable regularity conditions, and random sampling, when n is large, then the posterior is approximately $\mathcal{N}(\hat{\theta}, (nI_1(\hat{\theta}))^{-1})$, where $\hat{\theta}$ is the m.l.e.; provided that the prior is non-zero in a region surrounding $\hat{\theta}$. In particular, if θ_0 is the true parameter, then the posterior will become more and more concentrated around θ_0. Often reporting the posterior distribution is preferred to point estimation.

Bayes estimators are constructed to minimize the integrated risk. Recall from *Decision Theory:* Θ is the set of all possible states of nature (values of parameter), $\mathcal{D}$ is the set of all possible decisions (actions); a loss function is any function

$$L : \Theta \times \mathcal{D} \to [0, \infty).$$

For point estimation we choose $\mathcal{D} = \Theta$, and $L(\theta, d)$ is the loss in reporting d when θ is true. For a prior π and data $x \in \mathcal{X}$, the posterior expected loss of a decision is

$$\rho(\pi, d|x) = \int_\Theta L(\theta, d)\pi(\theta|x)d\theta.$$

For a prior π the integrated risk of a decision rule δ is

$$r(\pi, \delta) \;\; = \;\; \int_\Theta \int_\mathcal{X} L(\theta, \delta(x)) f(x|\theta) dx \pi(\theta) d\theta.$$

An estimator minimizing $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ that minimizes $\rho(\pi, \delta|x)$. A Bayes estimator associated

with prior π, loss L, is any estimator δ^π which minimizes $r(\pi, \delta)$. Then $r(\pi) = r(\pi, \delta^\pi)$ is the Bayes risk.

This is valid for proper priors, and for improper priors if $r(\pi) < \infty$. If $r(\pi) = \infty$, we define a generalized Bayes estimator as the minimizer, for every x, of $\rho(\pi, d|x)$.

For strictly convex loss functions, Bayes estimators are unique. For squared error loss $L(\theta, d) = (\theta - d)^2$, the Bayes estimator δ^π associated with prior π is the posterior mean. For absolute error loss $L(\theta, d) = |\theta - d|$, the posterior median is a Bayes estimator.

Example. Let x be a single observation from $\mathcal{N}(\mu, 1)$, and let μ have prior distribution

$$\pi(\mu) \propto e^\mu,$$

on the whole real line. What is the generalized Bayes estimator for μ under square error loss?

First we find the posterior distribution of μ given x.

$$\begin{aligned} \pi(\mu|x) &\propto exp\left(\mu - \frac{1}{2}(\mu - x))^2\right) \\ &\propto exp\left(-\frac{1}{2}(\mu - (x+1))^2\right) \end{aligned}$$

and so $\pi(\mu|x) = \mathcal{N}(x+1, 1)$.

The generalized Bayes estimator under square error loss is the posterior mean (from lectures), and so we have

$$\mu_B = 1 + x.$$

Note that, in this example, for μ_B, if we consider $X \sim \mathcal{N}(\mu, 1)$ and μ fixed, we have

$$MSE(\mu_B) = E(X + 1 - \mu)^2 = 1 + E(X - \mu)^2 = 2 > E(X - \mu)^2 = 1.$$

The m.l.e. for μ is X, so the mean-square error for the Bayes estimator is larger than the MSE for the maximum-likelihood estimator. The Bayes estimator is far from optimal here.

3. Hypothesis Testing

For testing $H_0 : \theta \in \Theta_0$, we set

$$\mathcal{D} = \{\text{accept } H_0, \text{ reject } H_0\} = \{1, 0\},$$

where 1 stands for acceptance; we choose as loss function

$$(1) \qquad L(\theta, \phi) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, \phi = 1 \\ a_0 & \text{if } \theta \in \Theta_0, \phi = 0 \\ 0 & \text{if } \theta \notin \Theta_0, \phi = 0 \\ a_1 & \text{if } \theta \notin \Theta_0, \phi = 1 \end{cases}.$$

Under this loss function, the Bayes decision rule associated with a prior distribution π is

$$\phi^\pi(x) = \begin{cases} 1 & \text{if } P^\pi(\theta \in \Theta_0 | x) > \frac{a_1}{a_0 + a_1} \\ 0 & \text{otherwise .} \end{cases}$$

The *Bayes factor* for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is

$$\begin{aligned} B^\pi(x) &= \frac{P^\pi(\theta \in \Theta_0 | x) / P^\pi(\theta \in \Theta_1 | x)}{P^\pi(\theta \in \Theta_0) / P^\pi(\theta \in \Theta_1)} \\ &= \frac{p(x | \theta \in \Theta_0)}{p(x | \theta \in \Theta_1)}; \end{aligned}$$

this is the ratio of how likely the data is under H_0 and how likely the data is under H_1. The Bayes factor measures the extent to which the data x will change the odds of Θ_0 relative to Θ_1.

Note that we have to make sure that our prior distribution puts mass on H_0 (and on H_1). If H_0 is simple, this is usually achieved by choosing a prior that has some point mass on H_0 and otherwise lives on H_1.

Example. Let X be binomially distributed with sample size n and success probability θ. Suppose that we would like to investigate $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, where θ_0 is specified. Our prior distribution is $\pi(H_0) = \beta$, $\pi(H_1) = 1 - \beta$, and given H_1, θ is Beta-distributed with parameters α and β, i.e.

$$\pi(\theta | H_1) = \frac{\theta^{\alpha - 1} (1 - \theta)^{\beta - 1}}{B(\alpha, \beta)},$$

where B is the Beta function,

$$B(\alpha,\beta)=\int_0^1\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

1. Find the Bayes factor for this test problem.
2. Find the Bayes decision rule for this test problem under the loss function (1).

For the Bayes factor we calculate $p(x|\theta\in\Theta_0)$ and $p(x|\theta\in\Theta_1)$. Firstly,

$$p(x|\theta\in\Theta_0) \;=\; p(x|\theta=\theta_0)=\binom{n}{x}\theta_0^x(1-\theta_0)^{n-x}.$$

For the alternative, we use the Beta prior,

$$\begin{aligned}p(x|\theta\in\Theta_1) &= \binom{n}{x}\frac{1}{B(\alpha,\beta)}\int_0^1\theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta\\ &= \binom{n}{x}\frac{1}{B(\alpha,\beta)}\int_0^1\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta\\ &= \binom{n}{x}\frac{B(x+\alpha,n-x+\beta)}{B(\alpha,\beta)}.\end{aligned}$$

Thus the Bayes factor is

$$\begin{aligned}B^\pi(x)\text{`} &= \frac{p(x|\theta\in\Theta_0)}{p(x|\theta\in\Theta_1)}\\ &= \frac{B(\alpha,\beta)}{B(\alpha+x,\beta+n-x)}\theta_0^x(1-\theta_0)^{n-x}.\end{aligned}$$

The Bayes decision rule is

$$\phi^\pi(x) \;=\; \begin{cases}1 & \text{if } P^\pi(\theta\in\Theta_0|x)>\frac{a_1}{a_0+a_1}\\ 0 & \text{otherwise}\end{cases}$$

and from lectures,

$$\begin{aligned}\phi^\pi(x)=1 &\iff B^\pi(x)>\frac{\frac{a_1}{a_0}}{\frac{\beta}{1-\beta}}\\ &\iff \frac{B(\alpha,\beta)}{B(\alpha+x,\beta+n-x)}\theta_0^x(1-\theta_0)^{n-x}>\frac{\frac{a_1}{a_0}}{\frac{\beta}{1-\beta}}\\ &\iff \frac{\beta B(\alpha,\beta)}{(1-\beta)B(\alpha+x,\beta+n-x)}\theta_0^x(1-\theta_0)^{n-x}>\frac{a_1}{a_0}.\end{aligned}$$

For robustness we consider least favourable Bayesian answers; suppose $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, and the prior probability on H_0 is $\rho_0 = 1/2$. For G a family of priors on H_1 we put

$$\underline{B}(x, G) = \inf_{g \in G} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta) g(\theta) d\theta}$$

and

$$\underline{P}(x, G) = \left(1 + \frac{1}{\underline{B}(x, G)}\right)^{-1}.$$

A Bayesian prior $g \in G$ on H_0 will then have posterior probability at least $\underline{P}(x, G)$ on H_0 (for $\rho_0 = 1/2$). If $\hat{\theta}$ is the m.l.e. of θ, and G_A the set of all prior distributions, then

$$\underline{B}(x, G_A) = \frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))}$$

and

$$\underline{P}(x, G_A) = \left(1 + \frac{f(x|\hat{\theta}(x))}{f(x|\theta_0)}\right)^{-1}.$$

4. Credible intervals

A $(1 - \alpha)$ *(posterior) credible interval (region)* is an interval (region) of $\theta-$values within which $1-\alpha$ of the posterior probability lies. Often we would like to find HPD (highest posterior density) region: a $(1 - \alpha)$ credible region that has minimal volume. When the posterior density is unimodal, this is often straightforward.

Example continued. Suppose $y_1, y_2, \ldots, y_n$ are independent normally distributed random variables, each with variance 1 and with means $\beta x_1, \ldots, \beta x_n$, where β is an unknown real-valued parameter and $x_1, x_2, \ldots, x_n$ are known constants. Under the Jeffreys prior, the posterior is $\mathcal{N}\left(\frac{s_{xy}}{s_{xx}}, (s_{xx})^{-1}\right)$. Hence a 95%-credible interval for β is

$$\frac{s_{xy}}{s_{xx}} \pm 1.96\sqrt{\frac{1}{s_{xx}}}.$$

The interpretation in Bayesian statistics is conditional on the observed $\mathbf{x}$; the randomness relates to the distribution of θ. In contrast, a frequentist confidence interval applies before $\mathbf{x}$ is observed; the randomness relates to the distribution of $\mathbf{x}$.

5. Not to forget about: *Nuisance parameters*

If $\theta = (\psi, \lambda)$, where λ nuisance parameter, and $\pi(\theta|x) = \pi((\psi, \lambda)|x)$, then we base our inference on the *marginal posterior* of ψ:

$$\pi(\psi|x) = \int \pi(\psi, \lambda|x) d\lambda.$$

That is, we just integrate out the nuisance parameter.