Markov chain Monte Carlo (MCMC)

TMA4300: Computer Intensive Statistical Methods
(Spring 2016)

Andrea Riebler

Lecture 7: Brief reminder – Bayesian model

- data: x
- likelihood model: $x|\theta \sim f(x|\theta)$
- prior distribution: $\theta \sim f(\theta)$
- posterior distribution:

$$\underbrace{f(\theta|x)}_{\text{Posterior}} \propto \underbrace{f(x|\theta)}_{\text{Likelihood}} \times \underbrace{f(\theta)}_{\text{Prior}}$$

Bayesian point estimates

Statistical inference about θ is based solely on the posterior distribution $f(\theta|x)$. Suitable point estimates are location parameters, such as:

- Posterior mean $\mathrm{E}(\theta|x)$:

$$\mathrm{E}(\theta|x) = \int \theta f(\theta|x) d\theta.$$

- Posterior mode $\mathrm{Mod}(\theta|x)$:

$$\mathrm{Mod}(\theta|x) = \arg\max_{\theta} f(\theta|x)$$

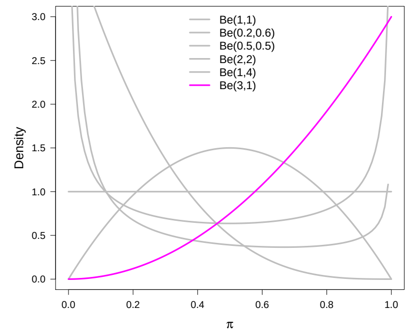
- Posterior median $\mathrm{Med}(\theta|x)$ is defined as the value a which satisfies

$$\int_{-\infty}^{a} f(\theta|x) d\theta = 0.5 \quad \text{and} \quad \int_{a}^{\infty} f(\theta|x) d\theta = 0.5$$

Binomial experiment

Let $X \sim \mathrm{Bin}(n, p)$ with n known and $p \in \Pi = (0, 1)$ unknown.

Since p is constrained to be within 0 and 1, a usual prior distribution is a beta distribution, so that $p \sim \mathrm{Be}(\alpha, \beta)$ with $\alpha, \beta > 0$ and $\mathcal{T} = (0, 1)$.



Binomial experiment (2)

$$X \sim \text{Bin}(n, p),\ x = 0, 1, \ldots, n, \qquad p \sim \text{Be}(\alpha, \beta),\ 0 < p < 1$$

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$\text{L}(p) = f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad f(p) = \frac{1}{\text{B}(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$\propto p^x(1-p)^{n-x} \qquad\qquad \propto p^{\alpha-1}(1-p)^{\beta-1}$$

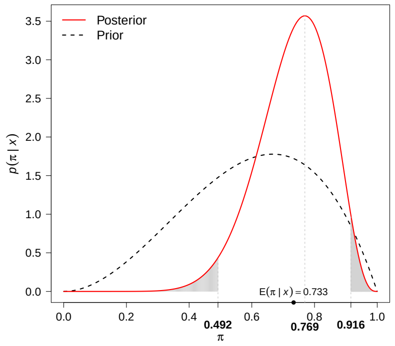
Thus, the posterior distribution results as:

$$\begin{aligned} f(p|x) &\propto f(x|p) \times f(p) \\ &= p^x(1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha+x-1}(1-p)^{\beta+n-x-1} \end{aligned}$$

This corresponds to the core of a beta distribution, so that

$$p|x \sim \text{Be}(\alpha + \underbrace{x}_{\text{successes}}, \beta + \underbrace{n-x}_{\text{failures}})$$

Binomial experiment: Simple example



Posterior density of $p|x$ for a Be(3, 2) prior and observation $x = 8$ in a binomial experiment with $n = 10$ trials. An equi-tailed 95% credible interval is also shown.

Using a Be(1,1) the posterior mode equals the Maximum Likelihood (ML) estimate.

Credible interval

For fixed $\alpha \in (0, 1)$, a $(1-\alpha)$ credible interval is defined through two real numbers t_l and t_u, so that

$$\int_{t_l}^{t_u} f(\theta|x) d\theta = 1 - \alpha.$$

The number $1-\alpha$ is called the credible level of the credible interval $[t_l, t_u]$.

There are infinitely many $(1-\alpha)$-credible intervals for fixed α. (At least if θ is continuous.)

Credible interval (II)

Equi-tailed credible interval

The same amount ($\alpha/2$) of probability mass is cut from the left and right tail of the posterior distribution, i.e. choose t_l as the $\alpha/2$-quantile and t_u as the $1-\alpha/2$-quantile.

Highest posterior density (HPD) intervals

Feature: The posterior density at any value of θ inside the credible interval must be larger than anywhere outside the credible interval.
HPD-interval have the smallest width among all $(1-\alpha)$ credible intervals.
For symmetric posterior distributions HPD intervals are also equi-tailed.

Properties of the beta-distribution

$\text{Be}(\alpha, \beta)$ can be interpreted as that which would have arisen if we had started with an "improper" $\text{Be}(0, 0)$ prior and then observed α successes in $\alpha + \beta$ trials. $\Rightarrow$ $n_0 = \alpha + \beta$ can be viewed as a prior sample size and $\alpha/(\alpha + \beta)$ as prior mean.
The posterior mean is given by:

$$E(p|x) = \frac{\alpha + x}{\alpha + \beta + n} = \underbrace{\frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta}}_{\text{Weighted prior mean}} + \underbrace{\frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}}_{\text{Weighted ML-estimate}}$$

The weights are proportional to the prior sample size and the data sample size.
$\Rightarrow$ Observing more data leads to a decreasing influence of the prior.

Bayesian learning

An important feature of Bayesian inference is the consistent processing of sequentially arising data.

- Suppose new independent data x_2 from a $\text{Bin}(n, p)$ arrive.
- The posterior distribution from the original observation (with x now called x_1) becomes the prior for x_2:

$$\begin{aligned} f(p|x_1, x_2) &\propto f(x_2|p, x_1) \times f(p|x_1) \\ &\propto f(x_2|p) \times f(p|x_1) \end{aligned}$$

Using $f(p|x_1) \propto f(x_1|p) \times f(p)$ an alternative formula is

$$\begin{aligned} f(p|x_1, x_2) &\propto f(x_2|p) \times f(x_1|p) \times f(p) \\ &= f(x_1, x_2|p) \times f(p) \end{aligned}$$

Thus, $f(p|x_1, x_2)$ is the same whether or not the data are processed sequentially.

Choice of prior distributions

- Under a uniform prior the posterior mode equals the MLE, as

$$f(\theta|x) \propto L_x(\theta)$$

- The prior distribution has to be chosen appropriately, which often causes concerns to practitioners.
- It should reflect the knowledge about the parameter of interest (e.g. a relative risk parameter in an epidemiological study).
- Ideally it should be elicited from experts.
- In the absence of expert opinions, simple informative prior distributions may still be a reasonable choice.

Choice of the prior distribution

Prior distributions incorporate prior beliefs in the Bayesian analysis. A pragmatic approach is to choose a prior distribution.

Conjugate prior distribution

Let $\text{L}_x(\theta) = \text{p}(x|\theta)$ denote a likelihood function based on the observation $X = x$. A class $\mathcal{G}$ of distributions is called conjugate with respect to $\text{L}_x(\theta)$ if the posterior distribution $\text{p}(\theta|x)$ is in $\mathcal{G}$ for all x whenever the prior distribution $\text{p}(\theta)$ is in $\mathcal{G}$.

Example

Binomial experiment Let $X|p \sim \text{Bin}(n, p)$. The family of beta distributions, $p \sim \text{Be}(\alpha, \beta)$, is conjugate with respect to $\text{L}_x(p)$, since the posterior distribution is again a beta distribution:
$p|x \sim \text{Be}(\alpha + x, \beta + n - x)$

List of conjugate prior distributions

Sequential processing:

- Sufficient to study conjugacy for one member of a random sample $X_1, \ldots, X_n$.
- The posterior after observing the first observation is of the same type as the prior and serves as new prior distribution for the next observation.
- Sequentially processing the data, only the parameters will change and not the type of prior.

List of conjugate prior distributions

Likelihood	Conjugate prior	Posterior distribution
$X\|p \sim \text{Bin}(n,p)$	$p \sim \text{Be}(\alpha,\beta)$	$p\|x \sim \text{Be}(\alpha+x, \beta+n-x)$
$X\|p \sim \text{Geom}(p)$	$p \sim \text{Be}(\alpha,\beta)$	$p\|x \sim \text{Be}(\alpha+1, \beta+x-1)$
$X\|\lambda \sim \text{Po}(e\cdot\lambda)$	$\lambda \sim \text{G}(\alpha,\beta)$	$\lambda\|x \sim \text{G}(\alpha+x, \beta+e)$
$X\|\lambda \sim \text{Exp}(\lambda)$	$\lambda \sim \text{G}(\alpha,\beta)$	$\lambda\|x \sim \text{G}(\alpha+1, \beta+x)$
$X\|\mu \sim \mathcal{N}(\mu,\sigma_\star^2)$	$\mu \sim \mathcal{N}(\nu,\tau^2)$	$\mu\|x \sim \mathcal{N}\left[(A)^{-1}\left(\frac{x}{\sigma^2}+\frac{\nu}{\tau^2}\right), (A)^{-1}\right]$
$X\|\sigma^2 \sim \mathcal{N}(\mu_\star,\sigma^2)$	$\sigma^2 \sim \text{IG}(\alpha,\beta)$	$\sigma^2\|x \sim \text{IG}(\alpha+\frac{1}{2}, \beta+\frac{1}{2}(x-\mu)^2)$

$_\star$: known.

$A = \frac{1}{\sigma^2} + \frac{1}{\tau^2}$

Improper prior distributions

Maybe you feel uncomfortable putting a prior on an unknown parameter. If you use a normal prior you can use a very large variance. In the limit this leads to an improper prior distribution.

Improper prior distribution

For example, let $\mu \sim \mathcal{N}(\mu, \infty)$, i.e. $f(\mu) \propto \text{const.} > 0$.

$$\int f(\mu) d\mu \approx \infty$$

Priors such as $f(\mu) = \text{const.}$, $f(\sigma) = 1/\sigma$ are improper, because they do not integrate to 1.

Improper prior distributions (II)

In most cases, improper priors can be used in Bayesian analyses without major problems. However, things to watch out for are:

- In a few models, the use of improper priors can result in improper posteriors.
- Use of improper priors makes model selection difficult.

Uninformative priors

Though conjugate priors are computationally nice, priors might be preferred which do not strongly influence the posterior distribution. Such a prior is called an uninformative prior.

- The historical approach, followed by Laplace and Bayes, was to assign flat priors.
- This prior seems reasonably uninformative. We do not know where the actual value lies in the parameter space, so we might as well consider all values equi-probable.
- However, this prior is not invariant to one-to-one transformations.

Harold Jeffreys' prior

Definition

Let X denote a random variable with likelihood function $p(x|\theta)$ where θ is an unknown scalar parameter. Jeffreys' prior or Jeffreys' rule is defined as

$$f(\theta) \propto \sqrt{J(\theta)},$$

where $J(\theta)$ is the expected Fisher information of θ.

Jeffreys' prior has certain desired properties, e.g. invariance property.

Jeffreys' prior for the geometric distribution

The geometric distribution models the number X of Bernoulli trials needed to get the first success. Let $X|\pi \sim \text{Geom}(\pi)$, i.e.

$$P(x|\pi) = \pi \cdot (1-\pi)^{x-1}.$$

Thus:

$$l_x(\pi) = \log(\pi) + (x-1)\log(1-\pi)$$

$$l'_x(\pi) = \frac{1}{\pi} - \frac{x-1}{1-\pi}$$

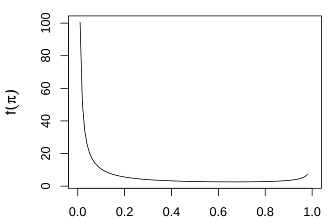
$$l''_x(\pi) = -\frac{1}{\pi^2} - \frac{x-1}{(1-\pi)^2}$$

$$J(\pi) = -\,\mathrm{E}\left(-\frac{1}{\pi^2} - \frac{x-1}{(1-\pi)^2}\right) = \frac{1}{\pi^2} + \frac{\frac{1}{\pi}-1}{(1-\pi)^2} = \frac{1}{\pi^2} + \frac{1-\pi}{\pi(1-\pi)^2} = \pi^{-2}(1-\pi)^{-1}$$

Jeffreys' prior results as:

$$f(\pi) \propto \sqrt{J(\pi)} = \pi^{-1}(1-\pi)^{-1/2}$$

(can be seen as "Be(0, 0.5)")



$\Rightarrow$ Small values are favoured.

New concept: Penalised complexity (PC) priors

There was recently a new concept developed here at NTNU to choose interpretable and meaningful prior distributions.
For more information see here

`http://arxiv.org/abs/1403.4630`

We may come back to this later, when we talk about Bayesian hierarchical models.