ST4234 Bayesian Statistics

Lingjie, May 2, 2022

Bayesian approach

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Ingredients for modelling

1. data
2. model
3. parameter
4. sampling distribution

Frequentist vs Bayesian approach

	Frequentist	Bayesian
tools	MLE, hypothesis testing, confidence interval	prior distribution, posterior distribution
true θ	fixed but unknown	not fixed but follows a distribution

Key terms

Techniques

term	meaning
Monte Carlo method	a method for approximating quantities via simulation of random variables
Markov chain Monte Carlo (MCMC)	generated variables form a Markov chain
Estimation vs prediction	Frequentist method estimates parameters by optimizing within a given dataset. Predictive performance on new dataset is not guaranteed, especially if there is a large number of parameters compared to sample size.
Shrinkage	Penalized likelihood to shrink the estimate towards a given value, such as 0 or the group mean. Shrinkage happen naturally in Bayesian (towards the prior or hyper-prior)
Hierarchical model	Models in which observations are clustered into groups. Parameters can be common within and across groups.

Bayesian

term	meaning
$p(\theta)$	prior density of θ
$p(\theta\|y)$	posterior density of θ given y
$p(y\|\theta)$	likelihood of θ at y
Odds	if an event occurs with probability p then odds of it occurring is $\frac{p}{1-p}$
Bayes factor	an integral likelihood ratio between two models in a Bayesian setting

Prior

term	meaning
Conjugate prior	family of prior distribution such that posterior distribution belong to the same family
Semi-Conjugate prior	prior for multiple parameters in which prior of each parameter conditioned on the other parameters is a conjugate prior
hyper-prior	Priors of hyper parameters
Within-group variability	variability between measurements of different units in the same group
Between-group variability	The variability between population means of different groups. (Across groups)
Improper prior	prior density which does not integrate to 1
non-informative prior	a prior which does not give the impression that you are favoring one parameter over another
Jeffrey's prior	prior invariant to a change of variable

Posterior distribution

term	meaning
Posterior mean	mean of posterior distribution
MAP	maximum a posterior probability, the mode of the posterior distribution
$\theta_{0.95}$	0.95− quantile of θ, threshold value of θ such that probability that the parameter is less than or equal to this value is 0.95

Credible set

term	meaning
Credible set (interval)	set of parameter values constructed from posterior distribution. It is an interval if parameter is 1-dimensional
95% credible set	credible set containing at least 95% of the parameters in the posterior distribution.
Exact 95% credible set	a credible set containing exactly 95% of the parameters in the posterior distribution
HPB region	Highest Posterior Density region. A credible interval with the shortest length or a credible set with the smallest area, volume etc at a particular level. The posterior density within HPD region is always uniformly larger than outside the region.

Confidence set (frequentist)

term	meaning
Confidence sets (intervals)	a set of parameter values constructed from data
95% confidence set	on average 95% of the data contains the true parameter
0.95−quantile of a test statistics	threshold value of a test statistic such that the probability that the test statistics is equal to or below this value is 0.95

Prediction

term	meaning
(Posterior) predictive distribution	distribution of future observation or test statistic based on the posterior distribution. Note: this is not posterior distribution, it is distribution of observation

Bayes Theorem

For events

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

For densities

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

Sensitivity analysis

Analysis of how estimate depends on the chosen prior

Credible interval

In general, $100(1-\alpha)\%$ credible set for θ is a set C s.t.

$$P(\theta \in C|y) = \int_C p(\theta|y)d\theta \geq 1-\alpha$$

There is no unique credible interval for $[\theta_z, \theta_{z+0.95}]$

common	$[\theta_{0.025}, \theta_{0.975}]$
highest posterior density (HPD)	density at θ_z is the same as $\theta_{0.95+z}$ (numerical solution)
∞	$[\theta_0, \theta_{0.95}]$

Predictive distribution

With the known posterior $P(\theta|Y)$, calculate the associated $P(Y=y)$ unconditional on θ.

Posterior density of the difference

$d := \theta_1 - \theta_2$ where θ_i are independent and $\theta_i \sim \Gamma \in [0, \infty]$
recall (assuming independent)

$$p(\theta_1 - \theta_2 \leq t) = \int_{D_2}\int_{D_1}^{\theta_2+t} f_{\theta_1}(\theta_1)f_{\theta_2}(\theta_2)d\theta_1 d\theta_2 = \int_{D_2} p(\theta_1 \leq \theta_2 + t|\theta_2)f_{\theta_2}(\theta_2)d\theta_2$$

Therefore,

$$p(d=t|Y,Z) = \int_{D_2} p(\theta_1 = \theta_2 + t|Y,\theta_2)p(\theta_2|Z)d\theta_2$$
$$p(d\leq t|Y,Z) = \int_{D_2} p(\theta_1 \leq \theta_2 + t|Y,\theta_2)p(\theta_2|Z)d\theta 2$$

Note:

- No close form, require numerical computation for each z
- Domain is $(0,\infty)$ due to *Gamma* distribution
- θ_1 is the prior distribution, $p(\theta_1|Y)$ is the posterior distribution with data Y

Reporting of posterior

Different way of reporting due to different target audience

- whole distribution
- credible interval
- summary statistics

Bayes factor

Informs which model the data favours

$$BF_{12} = \frac{m_1(y)}{m_2(y)} = \frac{\int p_1(\theta)p(y|\theta)d\theta}{\int p_2(\theta)p(y|\theta)d\theta}$$

Mixture of priors

Combining multiple priors as a mixture

$$p(\theta) = \alpha p_1(\theta) + (1-\alpha)p_2(\theta), \alpha \in [0,1]$$

α is probability model 1 is correct. Prior odds that first model is correct is $\frac{\alpha}{1-\alpha}$

Posterior $p(\theta|y)$

$$p(\theta|y) = \beta p_1(\theta|y) + (1-\beta)p_2(\theta|y)$$
$$\left(\frac{\beta}{1-\beta}\right) = BF_{12}\left(\frac{\alpha}{1-\alpha}\right)$$
$$\Rightarrow \beta = \frac{\alpha BF_{12}}{\alpha BF_{12} + (1-\alpha)} = \frac{\alpha m_1(y)}{\alpha m_1(y) + (1-\alpha)m_2(y)}$$

Bayes factor BF_{12} updates prior odds $\frac{\alpha}{1-\alpha}$ to posterior odds $\frac{\beta}{1-\beta}$, $BF_{12} > 1 \Rightarrow \beta > \alpha$

Three priors

$$p(\theta) = \alpha_1 p_1(\theta_1) + \alpha_2 p_2(\theta_2) + \alpha_3 p_3(\theta_3)$$
$$p(\theta|y) = \beta_1 p(\theta_1|y) + \beta_2 p(\theta_2|y) + \beta_3 p(\theta_3|y)$$

We have:

$$BF_{12} = \frac{m_1(y)}{m_2(y)} = \frac{m_1(y)/m_3(y)}{m_2(y)/m_3(y)} = \frac{BF_{13}}{BF_{23}}$$
$$\beta_i = \frac{\alpha_i m_i(y)}{\alpha_1 m_1(y) + \alpha_2 m_2(y) + \alpha_3 m_3(y)}$$

Hierarchical Modeling

Setup

- consider $\mathbf{y}_j$ be math scores for n_j students in School j, $j \in [1, m]$
- all test score: $\mathbf{y} = (\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_m})$
- individual score of ith student in School j: $Y_{ij} \sim N(\theta_j, \frac{1}{\lambda})$
- total number of students: $n = \sum_{j=1}^{m} n_j$

Prior

- $\theta_j \sim N(\mu, \frac{1}{\xi})$
- $\lambda \sim \Gamma(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$

Hyper-prior

- $\mu \sim N(\omega_0, \frac{1}{\gamma_0})$
- $\xi \sim \Gamma(\frac{\eta_0}{2}, \frac{\eta_0 \tau_0^2}{2})$

Interpretation of model

- Within group variation: $\frac{1}{\lambda}$ (different students in the same school)
- Across group variation: $\frac{1}{\xi}$ (differences between schools)
- σ_0^2 is prior belief of the variability of scores between students of the same school ($\because E(\lambda) = \frac{1}{\sigma_0^2}$)
- τ_0^2 is prior belief of how much θ_j varies between schools ($\because E(\xi) = \frac{1}{\tau_0^2}$)
- ω_0 is prior belief of typical math score

Bayes formula and likelihood

$\Theta = (\theta_1, \cdots, \theta_m, \mu, \xi, \lambda)$, $p(\Theta|\mathbf{y}) \propto p(\mathbf{y}|\Theta)p(\Theta)$
Likelihood $Y_{ij} \sim N(\theta_j, \frac{1}{\lambda})$

$$p(\mathbf{y}|\Theta) = \prod_{j=1}^{m} p(\mathbf{y}_j|\Theta) = \prod_{j=1}^{m} \lambda^{\frac{n_j}{2}} \exp\left[-\frac{\lambda}{2}\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2\right]$$

Prior
Assuming μ, λ, ξ are independent in prior $\Rightarrow p(\mu, \lambda, \xi) = p(\mu)p(\lambda)p(\xi)$

$$\begin{aligned}
\theta &= (\theta_1, \cdots, \theta_m) \\
p(\Theta) &= p(\mu, \lambda, \xi)p(\theta|\mu, \lambda, \xi) = p(\mu)p(\lambda)p(\xi)p(\theta|\mu, \lambda, \xi) \\
p(\theta|\mu, \lambda, \xi) &= \prod_{j=1}^{m} p(\theta_j|\mu, \lambda, \xi) \\
&\propto \xi^{\frac{m}{2}} \exp\left[-\frac{\xi}{2}\sum_{j=1}^{m}(\theta_j - \mu)^2\right]
\end{aligned}$$

Posterior

Priors are Semi-Conjugate
Let $\Theta^{(-\mu)}$:= all parameters except μ etc

$$\begin{aligned}
\theta_j|\mathbf{y}, \Theta^{(-\theta_j)} &\sim N\left(\mu_j, \frac{1}{\xi_j}\right) \\
\lambda|\mathbf{y}, \Theta^{(-\lambda)} &\sim \Gamma\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right) \\
\mu|\mathbf{y}, \Theta^{(-\mu)} &\sim N\left(\omega_n, \frac{1}{\gamma_n}\right) \\
\xi|\mathbf{y}, \Theta^{(-\xi)} &\sim \Gamma\left(\frac{\eta_n}{2}, \frac{\eta_n \tau_n^2}{2}\right)
\end{aligned}$$

$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$, sample average within group j
$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m} \theta_j$, average population means over the m groups
$\xi_j = \xi + n_j\lambda$
$\mu_j = \frac{\xi\mu + n_j\lambda\bar{y}_j}{\xi + n_j\lambda}$
$\nu_n = \nu_0 + n$
$\sigma_n^2 = \frac{\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{j=1}^{n_j}(y_{ij} - \theta_j)^2}{\nu_0 + n}$
$\gamma_n = \gamma_0 + m\xi$
$\omega_n = \frac{\gamma_0\omega_0 + m\xi\bar{\theta}}{\gamma_0 + m\xi}$
$\eta_n = \eta_0 + m$
$\tau_n^2 = \frac{\eta_0\tau_0^2 + \sum_{j=1}^{m}(\theta_j - \mu)^2}{\eta_0 + m}$

Mixture model

Setup

- consider n samples with unobserved group membership $X_i \in \{1, 2\}$, $p(X_i = 1) = p = 1 - p(X_i = 2)$
- $Y_i \sim N(\theta_1, \frac{1}{\lambda_1})$ when $X_i = 1$
- $Y_i \sim N(\theta_2, \frac{1}{\lambda_2})$ when $X_i = 2$
- n_1 := number of observations belong to group 1, n_2 is defined similarly

Prior

- $p \sim Beta(a, b)$
- $\theta_j \sim N(\mu_0, \frac{1}{\xi_0})$
- $\lambda_j \sim \Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$

$$\Theta = (p, \theta_1, \theta_2, \lambda_1, \lambda_2, X_1, \cdots, X_n)$$

$$p(p) \propto p^{a-1}(1-p)^{b-1} p^{n_1}(1-p)^{n_2} \propto p^{a+n_1-1}(1-p)^{b+n_2-1}$$

$$p(\theta_j) \propto \exp(-\frac{\xi_0}{2}(\theta_j - \mu_0)^2)$$

$$p(\lambda_j) \propto \lambda_j^{\frac{\nu_0}{2}-1} \exp\left(-\frac{\nu_0\sigma_0^2}{2}\lambda_j\right)$$

$$p(X_i|p) = \begin{cases} p, & X_i = 1 \\ 1-p, & X_i = 2 \end{cases}$$

Likelihood

$$\begin{aligned} p(\mathbf{y}|\Theta) &= \prod_{i:X_i=1} \lambda_1^{\frac{1}{2}} \exp\left(-\frac{\lambda_1}{2}[y_i - \theta_1]^2\right) \prod_{i:X_i=2} \lambda_1^{\frac{1}{2}} \exp\left(-\frac{\lambda_2}{2}[y_i - \theta_2]^2\right) \\ &= \lambda_1^{\frac{n_1}{2}} \exp\left(-\frac{\lambda_1}{2} \sum_{i:X_i=1} [y_i^2 + \theta_1^2 - 2y_i\theta_1]\right) \\ &\times \lambda_2^{\frac{n_2}{2}} \exp\left(-\frac{\lambda_2}{2} \sum_{i:X_i=2} [y_i^2 + \theta_2^2 - 2y_i\theta_2]\right) \end{aligned}$$

$$p(\mathbf{y}, p|\Theta^{(-p)}) \propto 1$$

$$p(\mathbf{y}, \theta_j|\Theta^{(-\theta_j)}) \propto \exp\left(-\frac{\lambda_j}{2}[n_j\theta_j^2 - 2n_j\bar{y}_j\theta_j]\right)$$

$$p(\mathbf{y}, \lambda_j|\Theta^{(-\lambda_j)}) \propto \lambda_j^{\frac{n_j}{2}} \exp\left(-\frac{\lambda_j}{2} \sum_{i:X_i=j} [y_i^2 + \theta_j^2 - 2y_i\theta_j]\right)$$

$$p(\mathbf{y}, X_i = j|\Theta^{(-X_i)}) = \sqrt{\frac{\lambda_j}{2\pi}} \exp\left(-\frac{\lambda_j}{2}[y_i^2 + \theta_j^2 - 2y_i\theta_j]\right)$$

Posterior

$$p(p|\Theta^{(-p)}, \mathbf{y}) = p(\mathbf{y}, p|\Theta^{(-p)})p(p|\Theta^{(-p)}) \propto p^{a+n_1-1}(1-p)^{b+n_2-1}$$

$$\begin{aligned} p(\theta_j|\Theta^{(-\theta_j)}, \mathbf{y}) &\propto \exp\left(-\frac{\xi_0 + \lambda_j n_j}{2}\left[\theta_j - \frac{\xi_0\mu_0 + n_j\bar{y}_j}{\xi_0 + \lambda_j n_j}\right]^2\right) \\ &\propto \exp\left(-\frac{\xi_n}{2}[\theta_j - \mu_n]^2\right) \end{aligned}$$

$$\begin{aligned} p(\lambda_j|\Theta^{(-\lambda_j)}, \mathbf{y}) &\propto \lambda_j^{\frac{\nu_0+n_j}{2}-1} \exp\left(-\frac{\lambda_j}{2}\left[\nu_0\sigma_0^2 + \sum_{i:X_i=j} [y_i^2 + \theta_j^2 - 2y_i\theta_j]\right]\right) \\ &\propto \lambda_j^{\frac{\nu_n}{2}-1} \exp\left(-\frac{\lambda_j}{2}\nu_n\sigma_n^2\right) \end{aligned}$$

$$p(X_i|\Theta^{(-X_i)}, \mathbf{y}) = \begin{cases} p\sqrt{\frac{\lambda_1}{2\pi}} \exp\left(-\frac{\lambda_1}{2}[y_i^2 + \theta_1^2 - 2y_i\theta_1]\right), & X_i = 1 \\ (1-p)\sqrt{\frac{\lambda_2}{2\pi}} \exp\left(-\frac{\lambda_2}{2}[y_i^2 + \theta_2^2 - 2y_i\theta_2]\right), & X_i = 2 \end{cases}$$

Sufficiency, exponential families and conjugate priors

Term	explanation
Sufficient statistic	summary data that provide as much information as the original data for parameter estimation or inference
Exponential families	family of distributions whose densities are of a given form. Goal here is to identify exponential families and their priors

Sufficiency

Summary test statistics $T(\mathbf{Y})$ is sufficient for θ if $\mathbf{Y}$ has a density $p(\mathbf{y}|\Theta)$ that can be factorized as

$$\begin{aligned} p(\mathbf{y}|\theta) &= h(\mathbf{y})g(\theta, T(\mathbf{y})) \\ \Rightarrow p(\theta|\mathbf{y}) &\propto p(\theta)g(\theta, T(\mathbf{y})) \end{aligned}$$

Exponential families

General form

$$p(y|\theta) = h(y)c_1(\theta)^{t_1(y)} \cdots c_J(\theta)^{t_J(y)}$$

Log transformed: $\eta_j = \log c_j(\theta)$

$$p(y|\eta) = h(y) \exp\left(\sum_{j=1}^{J} \eta_j t_j(y)\right)$$

Sufficient statistic for exponential families

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\ &= \left[\prod_{i=1}^{n} h(y_i)\right] [c_1(\theta)]^{T_1(\mathbf{y})} \cdots [c_J(\theta)]^{T_J(\mathbf{y})} \\ T_1(\mathbf{y}) &= \sum_{i=1}^{n} t_1(y_i), \cdots, T_J(\mathbf{y}) = \sum_{i=1}^{n} t_J(y_i) \end{aligned}$$

$(T_1(\mathbf{y}), \cdots, T_J(\mathbf{y}))$ is sufficient for $\mathbf{y}$

Conjugate priors for exponential families

Choose distribution with densities where $a_1, \cdots, a_J$ are hyperparameters

$$p(\theta) \propto [c_1(\theta)]^{a_1}, \cdots, [c_J(\theta)]^{a_J}$$

The posterior will have the "same" conjugate distribution but with hyperparameter updated to $(a_1 + T_1(\mathbf{y}), \cdots, a_J + T_J(\mathbf{y}))$

$$p(\theta|\mathbf{y}) \propto [c_1(\theta)]^{a_1+T_1(\mathbf{y})} \cdots [c_J(\theta)]^{a_J+T_J(\mathbf{y})}$$

Change of variables in priors. Improper, non-informative and Jeffrey's prior

Improper priors

prior $p(\theta) \propto f(\theta)$ is improper if $\int f(\theta)d\theta = \infty$
Therefore, it is impossible to simulate $\theta \sim p$.
However, improper priors are used because they often lead to proper posteriors.

Non-informative priors

A prior is non-informative if it does not give the impression that you are favouring one parameter over another.
Non-informative prior can be considered to avoid criticisms that prior favor a positive conclusion.
However, if there is outside information, then choosing non-informative prior leads to less accurate estimates.
Furthermore, shrinkage effect might be loss if non-informative prior is chosen.

Change in variable formula

Note: changing $p(\mathbf{y}|\theta)$ to $p(\mathbf{y}|\eta)$ only require replacing θ to $\eta(\theta)$ since θ is given.
For a change of variable θ to η, use Jacobian $\frac{d\theta}{d\eta}$

$$p(\eta) = p(\theta)\left|\frac{d\theta}{d\eta}\right|$$

Jeffrey's prior

Note: uniform prior is not invariant under a change of variables. Jeffrey's prior is an alternative non-informative prior that is in fact invariant. Note $I(\theta)$ is the fisher information.

$$\begin{aligned} p(\theta) &\propto \sqrt{I(\theta)} \\ I(\theta) &= \int \left[\frac{d}{d\theta}\log p(y|\theta)\right]^2 p(y|\theta)dy \\ p(\eta) &\propto \left|\frac{d\theta}{\eta}\right|\sqrt{I(\theta)} \propto \left|\frac{d\theta}{\eta}\right| p(\theta) \\ &\propto \sqrt{I(\eta)} \end{aligned}$$

Distributions (Prior and Predictive)

$$p(\theta)$$

Beta distribution

$a > 0, b > 0, B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$

$$\begin{aligned} p(\theta) &\propto \theta^{a-1}(1-\theta)^{b-1}, \theta \in (0,1) \\ &= \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

Properties

$$\begin{aligned} E(\theta) &= \frac{a}{a+b} \\ Var(\theta) &= \frac{ab}{(a+b)^2(a+b+1)} \\ Mode(\theta) &= \frac{a-1}{a+b-2}, \text{ if } a>1, b>1 \\ B(a,b) &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\ \Gamma(a) &= \int_0^\infty z^{a-1}e^{-z}dz \end{aligned}$$

Choice of a and b

Given the mean (or proportion), we set $E(\theta) = \frac{a}{a+b} = mean$
Choosing large a, b result in smaller variance (avoid)

Gamma distribution

$\theta \sim Gamma(a,b), a > 0, b > 0$

$$p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}, \theta > 0$$

$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx = (a-1)!$

Properties

- $E(\theta) = \frac{a}{b}$
- $Var(\theta) = \frac{a}{b^2}$
- $Mode(\theta) = \frac{a-1}{b}, a > 1$
- $cX \sim \Gamma(a, \frac{b}{c})$
- Chi-square: $\chi^2_\nu \sim \Gamma(\frac{\nu}{2}, \frac{1}{2})$

If $a = 1 \Rightarrow p(\theta) \sim exp(b)$ with mean $\frac{1}{b}$

Selection of a and b

With given mean, a large b will result in smaller variance ($Var(\theta) = \frac{E(\theta)}{b}$)

Negative binomial distribution

$Y \sim NB(r,p)$ if

$$P(Y=y) = \binom{r+y-1}{y} p^r(1-p)^y, y = 0,1,2,\cdots$$

This is the probability of r success and y number of failures (note we count the failures)

Dirichlet distribution

$Dir(a_1, \cdots, a_K)$ is extension of $Beta$ prior to $K \geq 2$

$$p(\theta) = \frac{1}{B(a_1, \cdots, a_K)} \theta_1^{a_1 - 1} \cdots \theta_K^{a_K - 1}$$
$$a_i \leq 1, \ \sum_{i \in K} \theta_i = 1$$

where $B(a_1, \cdots, a_K) = \frac{\Gamma(a_1) \times \cdots \Gamma(a_K)}{\Gamma(a_1 + \cdots + a_K)}$

Marginal Dirichlet is Beta

$$p(\theta_i | \theta^{(-i)}) \sim Beta\left(a_i, \sum_{j \neq i} a_j\right)$$

Normal distribution

$Y \sim (\theta, \sigma^2)$

$$p(y|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right)$$

Properties

- Symmetric about θ. Median and mode are both θ
- Affine transformation of Normal is Normal
- Standard normal $Z = \frac{Y - \theta}{\sigma}$
- Precision of sum is sum of precision $W = \sum_{i \in N} \lambda_i$
- $Q/(n-1) \sim \Gamma(\frac{n-1}{2}, \frac{n-1}{2})$

  $Y_1, \cdots, Y_n \sim N(\theta, \sigma^2)$ independently

  $Q := \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sim \chi^2_{n-1}$

Precision

Precise λ: measurements tend to be close to each other

$$\lambda = \frac{1}{\sigma^2} = \Sigma^{-1}$$

Highest precision (smallest variance)

Let $Y_1 \sim N(\theta, \sigma_1^2), Y_2 \sim N(\theta, \sigma_2^2)$

$$\begin{aligned}
Z &= wY_1 + (1-w)Y_2 \\
\Rightarrow E(Z) &= E(wY_1 + (1-w)Y_2) = \theta \\
\Rightarrow Var(Z) &= w^2 Var(Y_1) + (1-w)^2 Var(Y_2) \\
\text{solve } \min_w & \frac{w^2}{\lambda_1} + \frac{(1-w)^2}{\lambda_2} \\
\Rightarrow w^* &= \frac{\lambda_1}{\lambda 1 + \lambda 2} \\
\Rightarrow Var_{\min}(Z) &= \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^2 \frac{1}{\lambda_1} + \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^2 \frac{1}{\lambda_2} \\
&= \frac{1}{\lambda_1 + \lambda_2}
\end{aligned}$$

Generally the precision of Z is the sum of precision of Ys

$$\begin{aligned}
Z &= \left(\frac{\lambda_1}{\sum_{i \in n} \lambda_i}\right) Y_1 + \cdots + \left(\frac{\lambda_n}{\sum_{i \in n} \lambda_i}\right) Y_n \\
&= \frac{1}{\sum_{i \in n} \lambda_i} \\
\Rightarrow \lambda_Z &= \sum_{i \in n} \lambda_i
\end{aligned}$$

Student's t-distribution

$Z \sim N(0,1), Q \sim \chi^2_\nu$

$$T = \frac{z}{\sqrt{Q/\nu}} \sim t_\nu$$

Properties

- $T = \frac{\sqrt{n}(\bar{Y} - \theta)}{s_Y} \sim t_{n-1}$

  $Y_1, \cdots, Y_n \sim N(\theta, \sigma^2)$ independently

  $\bar{Y} = \frac{1}{n} \sum_i Y_i$

  $S_Y = \sqrt{\frac{1}{n-1} \sum_i (Y_i - \bar{Y})^2}$

Posterior distributions

$$p(\theta|y)$$

Binomial data distribution

$$p(y|\theta) \propto \theta^y (1-\theta)^{n-y}$$

Beta prior (conjugate prior)

$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \propto \theta^{a+y-1}(1-\theta)^{b+n-y}$

$$\Rightarrow p(\theta|y) \sim Beta(a+y, b+n-y)$$

Interpretation: a is the number of prior successes, b is the number of prior failures

Sensitivity analysis with beta prior

$w := \frac{a+b}{a+b+n}$, $\bar{y} = \frac{y}{n}$

$$\begin{aligned} E(\theta|y) &= \frac{a+y}{a+b+n} = w\left(\frac{a}{a+b}\right) + (1-w)\bar{y} \\ &= wE(\theta) + (1-w)\bar{y} \end{aligned}$$

Posterior mean is weighted average of prior mean and sample mean, weight determined by choice of a, b. However, as sample size n increase, influence of data increase.

Predictive distribution

3 coin tosses, $Y^* \sim Binom(2, \theta), \theta \sim p(\theta|Y)$

$$\begin{aligned} P(Y^* = 0|Y) &= E((1-\theta)^2|Y) \\ P(Y^* = 1|Y) &= E(2\theta(1-\theta)|Y) \\ P(Y^* = 2|Y) &= E(\theta^2|Y) \end{aligned}$$

Poisson data distribution

$Y \sim Poisson(\theta), \theta > 0$

$$\begin{aligned} p(y|\theta) = P(Y = y|\theta) &= \frac{\theta^y}{y!}e^{-\theta}, y \geq 0 \\ &\propto \theta^y e^{-\theta} \\ p(Y|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\ &\propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta} \end{aligned}$$

Properties

- $E(Y|\theta) = Var(Y|\theta) = \theta$
- If $Y_i \sim Poisson(\theta_i), i \in [1, n]$ are independent, then

$$S_n = \sum_{i=1}^{n} Y_i \sim Poisson\left(\sum_{i=1}^{n} \theta_i\right)$$

- If $\theta_1 = \cdots = \theta_n$, then $S_n \sim Poisson(n\theta)$

Relation with Poisson process

Consider a Poisson process on the positive real line with constant rate b

- the number of events Y observed on interval $[0, \theta]$ is distributed as $Poisson(b\theta)$
- waiting time till occurrence of the αth event $\sim Gamma(a, b)$

E.g. if on average $b = 10$ accidents per day, number of accident in 3 days $\sim Poisson(30)$, waiting time till 3rd accident is $Gamma(3, 10)$

Gamma prior (conjugate prior)

$p(\theta|Y) \propto p(\theta)p(Y|\theta) \propto \theta^{a+s-1}e^{-(b+n)\theta}$

$$\Rightarrow p(\theta|Y) \sim \Gamma(a+s, b+n)$$

$s = \sum y_i$

Sensitivity analysis with Gamma prior

$w := \frac{b}{b+n}, \bar{y} = \frac{y}{n}$

$$\begin{aligned} E(\theta|Y) &= \frac{a+s}{b+n} = w\left(\frac{a}{b}\right) + (1-w)\bar{y} \\ &= wE(\theta) + (1-w)\bar{y} \end{aligned}$$

b is the weightage (sample size) the prior

Predictive distribution: Negative binomial

If $Y \sim Poisson(\theta)$ with prior $p(\theta) \sim \Gamma(a, b)$, then unconditional on θ

$$\begin{aligned} P(Y = y) &= \int_0^\infty P(Y = y|\theta)p(\theta)d\theta \\ &= \binom{a+y-1}{y}\left(\frac{b}{b+1}\right)^a\left(\frac{1}{b+1}\right)^y \\ &\sim NB\left(r = a, p = \frac{b}{b+1}\right) \end{aligned}$$

Multinomial data distribution

$Y \sim$ Multinomial with $K, \theta_1, \theta_2, \cdots, \theta_K, \sum_{i \in K} \theta_i = 1$

$$\begin{aligned} p(y|\theta) &= P(Y = y|\theta) \\ &= \binom{n}{y_1, y_2, \cdots, y_k}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_K^{y_K} \\ &\propto \theta_1^{y_1}\theta_2^{y_2}\cdots\theta_K^{y_K} \end{aligned}$$

where $\binom{n}{y_1, y_2, \cdots, y_k} = \frac{n!}{y_1!\cdots y_K!}$

Dirichlet prior (conjugate prior)

$$\begin{aligned} p(\theta|Y) &\propto p(Y|\theta)p(\theta) \\ &\propto \theta_1^{y_1}\cdots\theta_K^{y_K}\theta_1^{a_1-1}\cdots\theta_K^{y_K-1} \\ &\sim Dir(a_1+y_1, \cdots, a_K+y_K) \end{aligned}$$

Normal data distribution (known variance)

$Y \sim N(\theta, \sigma^2) = N(\theta, \frac{1}{\lambda})$
$\theta :=$ mean, $\sigma^2 :=$ variance, $\lambda :=$ precision

$$\begin{aligned} p(y|\theta, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right) \\ &= \sqrt{\frac{\lambda}{2\pi}} exp\left(-\frac{\lambda}{2}(y-\theta)^2\right) \end{aligned}$$

Normal prior

Note:

- $\bar{y} = \frac{1}{n}\sum_{i \in N} y_i \Rightarrow p(\bar{y}|\theta) \sim N(\theta, \frac{1}{n\lambda})$
- $p(\theta) \sim N(\mu_0, \frac{1}{m_0\lambda})$
- We decide on $\sigma_0^2 = \frac{1}{m_0\lambda}$, then use population σ^2 to determine $\frac{1}{\lambda}$ and deduce m_0

$p(\theta|\bar{y}) \propto p(\bar{y}|\theta)p(\theta) \propto \exp(n\lambda(-\frac{\theta^2}{2} + \theta\bar{y}))\exp(m_0\lambda(-\frac{\theta^2}{2} + \theta\mu_0))$

$$p(\theta|\bar{y}) \sim N\left(\mu_n, \frac{1}{(m_0+n)\lambda}\right)$$

Sensitivity analysis with Normal prior

$w := \frac{m_0}{m_0+n}$

$$\begin{aligned} E(\theta|Y) = \mu_n &= \frac{m_0\mu_0 + n\bar{y}}{m_0+n} \\ &= w\mu_0 + (1-w)\bar{y} \end{aligned}$$

m_0 is the weightage (sample size) of prior

Predictive distribution

Let future observation be $Y = \theta + \epsilon$ (independent)
$\theta \sim N(\mu_n \frac{1}{(m_0+n)\lambda}), \epsilon \sim N(0, \frac{1}{\lambda})$

$$Y \sim N\left(\mu_n, \frac{1}{(m_0+n)\lambda + \frac{1}{\lambda}}\right)$$

Normal data distribution (unknown variance)

Same Normal data distribution with known variance case. However, now σ^2 is not known.

Normal-Gamma prior

- $\lambda(\nu_0, \sigma_0^2) \sim \Gamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$
- $E(\lambda) = \frac{1}{\sigma_0^2}$

  initial guess of λ is $\frac{1}{\sigma_0^2}$
- $Var(\lambda) = \frac{2}{\nu_0\sigma_0^4}$

  smaller $\nu_0 \Rightarrow$ higher confidence in prior λ
- $\theta|\lambda \sim N(\mu_0, \frac{1}{m_0\lambda})$

$$\begin{aligned} \theta|\lambda, \mathbf{y} &\sim N\left(\mu_n, \frac{1}{m_n\lambda}\right) \\ \lambda|\mathbf{y} &\sim \Gamma\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right) \end{aligned}$$

Where we have

$$\begin{aligned} \nu_n &= \nu_0 + n \\ m_n &= m_0 + n \\ m_n\mu_n &= m_0\mu_0 + n\bar{y} \\ \Rightarrow \mu_n &= \frac{m_0\mu_0 + n\bar{y}}{m_0+n} \\ \nu_n\sigma_n^2 + m_n\mu_n^2 &= \nu_0\sigma_0^2 + m_0\mu_0^2 + n\bar{y}^2 + (n-1)s^2 \\ \Rightarrow \sigma_n^2 &= \frac{\nu_0\sigma_0^2 + \frac{m_0 n}{m_0+n}(\bar{y}-\mu_0)^2 + (n-1)s^2}{\nu_0+n} \end{aligned}$$

Deriving Prior (θ, λ)

$$\begin{aligned} p(\theta, \lambda) &= p(\lambda)p(\theta|\lambda) \\ &\propto \lambda^{\frac{\nu_0-1}{2}} \exp\left(-\frac{\lambda}{2}\left[m_0\theta^2 - 2m_0\mu_0\theta + (\nu_0\sigma_0^2 + m_0\mu_0^2)\right]\right) \\ p(\mathbf{y}|\theta, \lambda) &= \prod_{i=1}^{n} p(y_i|\theta, \lambda) \\ &= \lambda^{\frac{n}{2}} \exp\left\{-\frac{\lambda}{2}\left[n\theta^2 - 2\theta n\bar{y} + n\bar{y}^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2\right]\right\} \\ p(\theta, \lambda|\mathbf{y}) &\propto p(\theta, \lambda)p(\mathbf{y}|\theta, \lambda) \end{aligned}$$

Inference on mean θ

Frequentist: 95% confidence interval

- If σ^2 is known

$$\left(\bar{y} \pm z_{0.975}\frac{\sigma}{\sqrt{n}}\right)$$

- If σ^2 is unknown

$$\left(\bar{y} \pm t_{0.975,n-1}\frac{s_Y}{\sqrt{n}}\right)$$

Bayesian: 95% credible interval

- If σ^2 is known

$$\left(\mu_n \pm z_{0.975}\frac{\sigma}{\sqrt{m_n}}\right)$$

$$Z = \frac{\theta-\mu_n}{\sigma/\sqrt{m_n}}$$

$$\theta = \mu_n + \frac{\sigma}{\sqrt{m_n}}Z$$

- If σ^2 is unknown

$$\left(\mu_n \pm t_{0.975,\nu_n}\frac{\sigma_n}{\sqrt{m_n}}\right)$$

$$T = \frac{Z}{\sqrt{\sigma_n^2\lambda}}, Z = \sqrt{m_n\lambda}(\theta-\mu_n)$$

$$\theta = \mu_n + \frac{\sigma_n}{\sqrt{m_n}}T$$

Difference in mean of two normal groups

Consider

- Model: $Y_i \sim N(\mu+\delta, \frac{1}{\lambda})$, $Z_i \sim N(\mu-\delta, \frac{1}{\lambda})$
- Prior: $\mu \sim N(\mu_0, \frac{1}{\lambda_0})$, $\delta \sim N(\delta_0, \frac{1}{\tau_0})$, $\lambda \sim \Gamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$
- Independence: $p(\mu,\delta,\lambda) = p(\mu)p(\delta)p(\lambda)$

Data likelihood

$$\begin{aligned} &p(\mathbf{y},\mathbf{z}|\mu,\delta,\lambda) \\ &\propto \lambda^{\frac{n+m}{2}} \exp\left\{-\frac{\lambda}{2}\left[\sum_{i=1}^{n}(y_i-\mu-\delta)^2 + \sum_{i=1}^{m}(z_i-\mu+\delta)^2\right]\right\} \end{aligned}$$

Posterior:

$$\begin{aligned} p(\mu|\mathbf{y},\mathbf{z},\delta,\lambda) &\propto p(\mathbf{y},\mathbf{z}|\mu,\delta,\lambda)p(\mu|\delta,\lambda) \\ &= N(\mu_n, \tfrac{1}{\lambda_n}) \\ p(\delta|\mathbf{y},\mathbf{z},\mu,\lambda) &= N(\delta_n, \tfrac{1}{\tau_n}) \\ p(\lambda|\mathbf{y},\mathbf{z},\mu,\delta) &= \Gamma(\tfrac{\nu_n}{2}, \tfrac{\nu_n\sigma_n^2}{2}) \end{aligned}$$

Updated parameters:

$$\begin{aligned} \lambda_n &= \lambda_0 + (n+m)\lambda \\ \mu_n &= \tfrac{\mu_0\lambda_0+[n(\bar{y}-\delta)+m(\bar{z}+\delta)]\lambda}{\lambda_0+(n+m)\lambda} \\ \tau_n &= \tau_n + (n+m)\lambda \\ \delta_n &= \tfrac{\delta_0\tau_0+[n(\bar{y}-\mu)+m(\mu-\bar{z})]\lambda}{\tau_0+(n+m)\lambda} \\ \nu_n &= \nu_0 + n + m \\ \sigma_n^2 &= \tfrac{\nu_0\sigma_0^2+\sum_{i=1}^{n}(y_i-\mu-\delta)^2+\sum_{i=1}^{m}(z_i-\mu+\delta)^2}{\nu_0+n+m} \end{aligned}$$

Techniques

MCMC techniques: Gibbs sampling, Metropolis-Hastings
MC techniques: rejection sampling, importance sampling, sampling importance resampling

MCMC and MCMC Diagnostics

Terms	Explanation
MCMC sample	time-series of the parameters generated from the posterior distribution using MCMC
MCMC diagnostics	statistical tools to judge the quality of MCMC samples, suggesting quality and area for improvements
coda	R package used for MCMC diagnostics
burn-in period	initial period of MCMC sample discarded during estimation. Suggested 20% burn-in
Autocorrelation	correlation between time-series and lags
Autocorrelation function (ACF)	function of ac with different lags
Naive SE	(wrong) standard error that assumes independence
Time-series SE	(right) standard error that takes into consideration time-series correlation
Effective sample size	relating time-series SE with independent Monte Carlo samples. For example, ESS = 100 means MCMC estimates has time-series ES equal to the SE of 100 independent samples
Trace plots	plots of MCMC samples as a function of time
Density plots	estimated density based on the MCMC samples
Summary stats	showing the average and time-series standard errors

Gibbs sampling (MCMC)

Key: generate samplings sequentially from conditional posterior distribution.
For large k, $\theta^{(k)}, \lambda^{(k)} \sim p(\theta,\lambda|\mathbf{y})$
Algorithm:

1. For $k \geq 1$:
   [1.1] Generate $\theta^{(k)} \sim p(\theta|\lambda^{(k-1)},\mathbf{y})$
   [1.2] Generate $\lambda^{(k)} \sim p(\lambda|\theta^{(k)},\mathbf{y})$

2. After K iterations return output $\{\theta^{(k)}, \mu^{(k)}\}_{k=1}^{K}$

Results such as credible interval can be obtained through the simulated samples. E.g. Estimating $E(f(\theta, \lambda)|\mathbf{y})$

$$\frac{1}{K}\sum_{k=1}^{K} f(\theta^{(k)}, \lambda^{(k)})$$

Metropolis-Hastings (MCMC)

Key: acceptance probability, probability that a proposed θ^* is accepted. If θ^* is not accepted, $\theta^{(k)} = \theta^{(k-1)}$. The desirable acceptance probability is large such that MCMC does not stuck at θ for long.

Generates Markovian sample $\theta = (\theta^{(1)}, \cdots, \theta^{(K)})$ with posterior $p(\theta|\mathbf{y})$ as the stationary distribution.

- Metropolis Algorithm: first version of the MH algorithm with symmetric proposal q
- Random-Walk MH: proposal q is a random walk (popular)
- Independence MH: proposal q is independent

Metropolis-Hastings is considered last resort in complicated Bayesian problems as it does not require for conditional posterior distribution and works for any distribution even without normalising constant. However, MH has high autocorrelation as a result of θ stucking in same θ value for a long time.

Understanding the MH algo

Consider two countries A, B with population $f(A), f(B)$. Probability of individual migrating from A to B is $q(B|A)$. Probability of individual migration application getting accepted is $\alpha(B|A)$.

Question: How to choose $\alpha(B|A), \alpha(A|B)$ such that population size are maintained?

Consider:

- Total number of people migrating: $f(A)q(B|A)$ and $f(B)q(A|B)$
- If $f(A)q(B|A) < f(B)q(A|B)$, net migration to A is more than B

  $\alpha(B|A) = 1$, B should accept everyone

  $\alpha(A|B) = \frac{f(A)q(B|A)}{f(B)q(A|B)}$, A should accept only a fraction of applicants such that total population unchanged i.e. $\alpha(A|B)f(B)q(A|B) = f(A)q(B|A)$
- Therefore, $\alpha(A|B) = \min\left\{1, \frac{f(A)q(B|A)}{f(B)q(A|B)}\right\}$, $\alpha(B|A) = \min\left\{1, \frac{f(B)q(A|B)}{f(A)q(B|A)}\right\}$

In the context of Metropolis-Hastings, the A, B are the transition from $\theta^{(k-1)}$ to $\theta^{(k)}$

Random walk

Generate $p(\theta|\mathbf{y}) \propto f(\theta)$ with $z \sim N(0,1)$

1. Generate $\theta^* = \theta^{(k-1)} + z$, $z \sim N(0,1) \Rightarrow \theta^* \sim N(\theta^{(k-1)}, 1)$

   $q(\theta^*|\theta^{(k-1)}) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(\theta^* - \theta^{(k-1)})^2)$
2. Accept θ^* with probability $\alpha(\theta^*|\theta^{(k-1)}) = \min(r, 1)$, $r = \frac{f(\theta^*)q(\theta^{(k-1)}|\theta^*)}{f(\theta^{(k-1)})q(\theta^*|\theta^{(k-1)})}$
3. If θ^* is accepted, $\theta^{(k)} = \theta^*$, else $\theta^{(k)} = \theta^{(k-1)}$

In general, we can use any g distribution such that $z \sim g$, $q(\theta^*|\theta^{(k-1)}) = g(\theta^* - \theta^{(k-1)})$

Symmetric proposal

If the choice of g is symmetric such that $g(z) = g(-z) \Rightarrow q(\theta^*|\theta^{(k-1)}) = q(\theta^{(k-1)}|\theta^*)$, then

$$r = \frac{f(\theta^*)}{f(\theta^{(k-1)})}$$

where the MH algorithm does not depend on the expression of q

Independence MH

Special case when $\theta^* \sim q$ that is independent of $\theta^{(k-1)}$ e.g. $\theta^* \sim N(0, \sigma^2)$, then $q(\theta^*|\theta^{(k-1)}) = q(\theta^*)$ and

$$r = \frac{f(\theta^*)q(\theta^{(k-1)})}{f(\theta^{(k-1)}q(\theta^*))} = \frac{w(\theta^*)}{w(\theta^{(k-1)})}$$

where $w(\theta) = \frac{f(\theta)}{q(\theta)}$ are the same weights in importance sampling and SIR. If the $\theta^{(k-1)}$ is under-represented in q then $w(\theta^{(k-1)})$ is large so probability of moving to a new value is low. Hence more samples are generated with $\theta^{(k-1)}$ to compensate for its under-representation in q.

Metropolis-within-Gibbs

MC

Terms	Explanation
Monte Carlo standard error (SE)	Errors due to randomness of Monte Carlo method.
Efficient	MC method is more efficient if it has a smaller Monte Carlo SE for the same computation time.
Target distribution	Distribution of interest, i.e. posterior $p(\theta\|\mathbf{y})$
Proposal distribution	distribution q which we simulate random parameters from. In direct MC $p = q$, in MCMC q are Markovian

Calculating the required sample to achieve the same efficiency

$$se = \frac{sd}{\sqrt{K}} \Rightarrow K = \frac{sd^2}{se^2}$$

Comparing the methods

- Rejection sampling:

  [key] some random samples from proposal is rejected

  [advantage] useful for e.g. predictive sampling since only useful samples are kept

  [disadvantage] probability of reject might be high

- Importance sampling:

  [key] all samples are kept, but re-sampled with over-represented values are given smaller importance weights.

  [advantage] useful for generating samples, since all samples are kept

  [advantage] No need to find M which is needed in rejection sampling

  [disadvantage] less useful for e.g. predictive samplings since even less important samples are kept, as weights have to be taken into consideration

- Sampling importance resampling:

  [key] samples from proposal density is resampled in proportion to the importance weight (improvement over important samplings for e.g. predictive samplings)

  [advantage] useful for e.g. predictive distribution since more important samples are sampled more often

Rejection sampling (MC)

Consider $p(\theta|\mathbf{y}) \propto f(\theta)$.
To generate $\theta^{(k)} \sim p(\theta|\mathbf{y})$, first find $M > 0$ s.t. $Mq(\theta) \geq f(\theta)$ for all θ

1. Generate $\theta \sim q$, where q is the proposal distribution
2. Generate $u \sim unif(0, 1)$

   [2.1] If $u \leq \frac{f(\theta)}{Mq(\theta)}$, accept θ and $\theta^{(k)} = \theta$

   [2.2] else repeat step 1

Importance sampling (MC)

Known constant

If $p(\theta|\mathbf{y})$ is known completely.
Generate $\theta^{(k)} \sim q$, where q is the proposal distribution.
Set weight as

$$w(\theta^{(k)}) = \frac{p(\theta^{(k)}|\mathbf{y})}{q(\theta^{(k)})}$$

and the estimated posterior mean is

$$\hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} w(\theta^{(k)}) \theta^{(k)}$$

Unknown constant

Same procedure as in the case of known constant. However, the weight $w(\theta^{(k)})$ is now normalised such that it sums up to 1

$$w^{(k)} = \frac{w(\theta^{(k)})}{\sum_{\ell=1}^{K} w(\theta^{(\ell)})}$$

and the estimated posterior mean is

$$\hat{\theta} = \sum_{k=1}^{K} w^{(k)} \theta^{(k)}$$

Sampling importance resampling (SIR) (MC)

Modification from importance sampling with unknown constant such that we do not need to deal with the weights.
Let $\theta_q = (\theta_q^{(1)}, \cdots, \theta_q^{(J)})$ denote i.i.d sample from q. Let $\mathbf{w} = (w^{(1)}, \cdots, w^{(J)})$ denote the normalized importance weights.
Create a resample $\theta_p = (\theta_p^{(1)}, \cdots, \theta_p^{(K)})$ representing a sample from the posterior $p(\theta|\mathbf{y})$ as followed:

1. Generate random index J_k such that for $1 \leq j \leq J$

$$P(J_k = j) = w^{(j)}$$

2. Let $\theta_p^{(k)} = \theta_q^{(J_k)}$

Linear Regression

Consider: $\mathbf{y} = \mathbf{X}\beta + \epsilon$

Simultaneous equation

If $\epsilon = \mathbf{0}$, $\beta = \mathbf{X}^{-1}\mathbf{y}$

Ordinary least squares (OLS)

$\hat{\beta}$ is the minimizer of $SSR(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2$
$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Regularized least squares (RLS)

Consider a Ridge regression with $L2$ norm.
$\hat{\beta}$ minimises $SSR^+(\beta) := SSR(\beta) + m_0 \sum_{j=1}^{d} \beta_j^2$
$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + m_0\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{y}$, $\mathbf{I}_d :=$ $(d \times d)$ identity matrix.

Bayesian regression

Ridge regression arrives at the same result with Bayesian models with normal errors and a Normal-Gamma prior. That is the β in Ridge is the posterior mean of β in Bayesian model.

Model:

$$Y_i = x_{i1}\beta_1 + \cdots x_{id}\beta_d + \epsilon_i$$
$$\epsilon_i \sim N\left(0, \frac{1}{\lambda}\right), i \in [1, n]$$

Parameters:

$$\mathbf{\Theta} = (\beta_1, \cdots, \beta_d, \lambda)$$

Likelihood:

$$\begin{aligned} p(\mathbf{y}|\mathbf{\Theta}) &= \prod_{i=1}^{n} p(\epsilon_i|\Theta) \\ &\propto \lambda^{\frac{n}{2}} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}\epsilon_i^2\right) \\ \because \epsilon &= \mathbf{y} - \mathbf{X}\beta \\ \therefore \sum_{i=1}^{n}\epsilon_i^2 &= ||\epsilon||^2 \\ &= ||\mathbf{y} - \mathbf{X}\beta||^2 \\ &= SSR(\beta) \\ \Rightarrow p(\mathbf{y}|\mathbf{\Theta}) &\propto \lambda^{\frac{n}{2}} \exp\left[-\frac{\lambda}{2}SSR(\beta)\right] \end{aligned}$$

Normal-Gamma prior

$$\lambda \sim \Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right), \nu_0 > 0, \sigma_0^2 > 0$$
$$\beta_j|\lambda \sim N\left(0, \frac{1}{m_0\lambda}\right), m_0 > 0, \beta_1, \cdots, \beta_d \text{ are independent}$$
$$\begin{aligned} p(\mathbf{\Theta}) &= p(\lambda)\prod_{j=1}^{d} p(\beta_j|\lambda) \\ &\propto \lambda^{\frac{\nu_0+d}{2}-1} \exp\left(-\frac{\lambda}{2}\left[m_0\sum_{j=1}^{d}\beta_j^2 + \nu_0\sigma_0^2\right]\right) \end{aligned}$$

Posterior

$$p(\mathbf{\Theta}|\mathbf{y}) \propto \lambda^{\frac{n+d+\nu_0}{2}-1} \exp\left(-\frac{\lambda}{2}\left[SSR^+(\beta) + \nu_0\sigma_0^2\right]\right)$$

Remark: It can be shown that $SSR^+(\beta) = SSR^+(\hat{\beta}) + \mathbf{z}^T(\mathbf{X}^T\mathbf{X} + m_0\mathbf{I}_d)\mathbf{z}$
Therefore, posterior can be expressed as

$$\begin{aligned} p(\Theta|\mathbf{y}) &\propto \lambda^{\frac{\nu_0+n}{2}-1} \exp\left(-\frac{\lambda}{2}\left[SSR^+(\hat{\beta}) + \nu_0\sigma_0^2\right]\right) \\ &\times \lambda^{\frac{d}{2}} \exp\left(-\frac{\lambda}{2}\left[\mathbf{z}^T(\mathbf{X}^T\mathbf{X} + m_0\mathbf{I}_d)\mathbf{z}\right]\right) \\ \Rightarrow \lambda|\mathbf{y} &\sim \Gamma\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right) \\ \nu_n &= \nu_0 + n \\ \sigma_n^2 &= \frac{\nu_0\sigma_0^2 + SSR^+(\hat{\beta})}{\nu_0 + n} \\ \Rightarrow \mathbf{z} &= \beta - \hat{\beta} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma^{-1}}) \\ \mathbf{\Sigma^{-1}} &= \frac{1}{\lambda}(\mathbf{X}^T\mathbf{X} + m_0\mathbf{I}_d)^{-1} \end{aligned}$$