Chapter 1. Introduction

1.1 Statistical Models for Different Data Types

- Categorical data
  - binary data
  - count data
  - ordinal data
  - nominal data
- Missing data
  - missing at random (MAR) data
  - non-ignorable missing data
- Hierarchical data
- Heterogenous data
- Longitudinal data
- Other non-normal data

1.1 Statistical Models for Different Data Types

- Categorical data

  Binary data

  - logistic regression model
  - probit regression model

  Count data

  - Poisson loglinear model
  - negative binomial loglinear model

  Ordinal data

  - cumulative logit model
  - probit model with latent variable

  Nominal data

  - multinomial logit model
  - multinomial probit model

1.1 Statistical Models for Different Data Types

- Non-ignorable missing data
  - patten mixture model
  - shared random effects and common factor model
  - independent binomial logit model
- Hierarchical data
  - multilevel model
  - generalized linear mixed effect model (GLMM)
- Heterogenous data
  - mixture model
  - semiparametric model
- Longitudinal data
  - GLMM
  - latent curve model
- Other non-normal data
  - semiparametric model
  - transformation model

Advantages of Bayesian approach

- It allows the use of genuine prior information to achieve better results.
- It does not rely on the large-sample asymptotic theory, thereby producing more reliable results even with small sample sizes.
- It is powerful in handling high-dimensional and complex data due to its sampling-based nature and the rapid development of modern computational techniques.

Notations

$\mathbf{Y}$ — observed data

$\mathbf{Z}$ — latent quantities

$\boldsymbol{\theta}$ — parameters

$p(\boldsymbol{\theta})$ — prior distribution

$p(\mathbf{Y})$ — marginal likelihood

$p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y})$ — joint posterior distribution

$p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$ — conditional posterior distribution of $\boldsymbol{\theta}$

$p(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta})$ — conditional posterior distributions of $\mathbf{Z}$

Bayes Theorem

Posterior inference for $\boldsymbol{\theta}$ is based on the following equality:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{Y}) &= p(\mathbf{Y},\boldsymbol{\theta})/p(\mathbf{Y}) \\ &= p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{Y}). \end{aligned}$$

Then,

$$\begin{gathered} p(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \quad \text{or} \\ \log p(\boldsymbol{\theta}|\mathbf{Y}) = \log p(\mathbf{Y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{constant}. \end{gathered} \tag{1.1}$$

Bayesian Inference

- Bayesian approach treats parameters as random variables and uses the data to update prior knowledge about parameters and latent quantities.
- The Bayesian sampling-based estimation techniques obtain samples from the joint posterior distribution of parameters and latent quantities.

Bayesian Inference

Important issues in Bayesian inference (Gilks et al., 1996; Carlin and Louis, 2006):

- How to choose prior distributions?
- The sensitivity or robustness of the Bayesian inference to the choice of priors.

Prior Specification

Conjugate prior

$p(\boldsymbol{\theta})$ is called a conjugate prior if the posterior $p(\boldsymbol{\theta}|\cdot)$ has the same form as the prior $p(\boldsymbol{\theta})$.

Commonly used conjugate priors (Congdon, 2006)

- Normal prior for regression coefficient
- Gamma prior for (inverse of) variance
- Inverse Wishart prior for covariance matrix
- Beta prior for binomial probability
- Dirichlet prior for multinomial probabilities

Prior Specification

Example 1: For a binomial model

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Consider the following prior density of θ:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{1.2}$$

which is a beta distribution with hyperparameters α and β.

Then,

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}, \end{aligned} \tag{1.3}$$

which is a beta distribution with parameters $y+\alpha$ and $n-y+\beta$.

Example 2: Let $y_1, \cdots, y_n$ are i.i.d. $\sim N[\mu, \sigma^2]$ with $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Let $\mathbf{Y} = (y_1, \cdots, y_n)$, the likelihood function is

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\sigma^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}.$$

Consider $p(\mathbf{Y}|\boldsymbol{\theta})$ as a function of μ (σ^2 is given), the likelihood is an exponential of a quadratic form in μ.

A conjugate prior distribution of μ can be parameterized as

$$p(\mu) \propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\},$$

that is, $\mu \stackrel{D}{=} N[\mu_0, \sigma_0^2]$, where μ_0 and σ_0^2 are hyperparameters.

Prior Specification

The conditional posterior density of $p(\mu|\mathbf{Y}, \sigma^2)$ is

$$\begin{aligned}
p(\mu|\mathbf{Y}, \sigma^2) &\propto p(\mu)p(\mathbf{Y}|\boldsymbol{\theta}) \\
&\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \\
&\propto \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma_0^2}(\mu-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}\right].
\end{aligned}$$

It can be shown that $[\mu|\mathbf{Y}, \sigma^2] \stackrel{D}{=} N[\tilde{\mu}, \tilde{\sigma}^2]$, where

$$\tilde{\mu} = \tilde{\sigma}^2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}y_i+\frac{\mu_0}{\sigma_0^2}\right), \quad \tilde{\sigma}^2 = \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}.$$

Prior Specification

If we consider $p(\mathbf{Y}|\boldsymbol{\theta})$ as a function of σ^2 (μ is given), then

$$p(\mathbf{Y}|\boldsymbol{\theta}) \quad \propto \quad (\sigma^2)^{-n/2} \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\Big\}.$$

A conjugate prior distribution of σ^2 can be parameterized as

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha_0+1)} \exp(-\beta_0/\sigma^2),$$

i.e., $\sigma^2 \overset{D}{=} IG(\alpha_0, \beta_0)$, where $IG(\alpha_0, \beta_0)$ is the inverse Gamma distribution with hyperparameters α_0 and β_0. Thus,

$$\begin{aligned} p(\sigma^2|\mathbf{Y},\mu) \quad &\propto \quad p(\sigma^2)p(\mathbf{Y}|\boldsymbol{\theta}) \\ &\propto \quad (\sigma^2)^{-(\frac{n}{2}+\alpha_0+1)} \exp\Big[-\frac{1}{2\sigma^2}\Big\{\sum_{i=1}^{n}(y_i-\mu)^2+\beta_0\Big\}\Big], \end{aligned}$$ or

$[\sigma^2|\mathbf{Y},\mu] \overset{D}{=} IG(\tilde{\alpha}, \tilde{\beta})$, with $\tilde{\alpha} = \frac{n}{2} + \alpha_0$, $\tilde{\beta} = \frac{1}{2}\Big\{\sum_{i=1}^{n}(y_i-\mu)^2+\beta_0\Big\}$.

Posterior Sampling

The Bayesian estimate of $\boldsymbol{\theta}$ is usually defined as the mean or the mode of $p(\boldsymbol{\theta}|\mathbf{Y})$. Theoretically, the mean of $p(\boldsymbol{\theta}|\mathbf{Y})$ can be obtained by integration, but it often doesn't have a closed form.

If we can simulate sufficient observations from $p(\boldsymbol{\theta}|\mathbf{Y})$, then the mean and other statistics of $[\boldsymbol{\theta}|\mathbf{Y}]$ can be approximated. However, directly sampling from $p(\boldsymbol{\theta}|\mathbf{Y})$ is difficult if $\boldsymbol{\theta}$ contains multiple components and/or latent quantities $\mathbf{Z}$ exist.

The strategy of data augmentation (Tanner and Wong, 1987) is to treat $\mathbf{Z}$ as hypothetical missing data and augment $\mathbf{Y}$ with $\mathbf{Z}$, so that $p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y})$ is relatively easy to handle. Specifically,

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y}) &= p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{Y})/p(\mathbf{Y}) \\ &= p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})p(\mathbf{Z}, \boldsymbol{\theta})/p(\mathbf{Y}) \\ &= p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})p(\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{Y}), \end{aligned} \tag{1.4}$$

or

$$p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta})p(\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1.5}$$

Gibbs sampler

At the jth iteration with current values of $\boldsymbol{\theta}^{(j)}$ and $\mathbf{Z}^{(j)}$,

a. generate $\mathbf{Z}^{(j+1)}$ from $p(\mathbf{Z}|\boldsymbol{\theta}^{(j)}, \mathbf{Y})$;

b. generate $\boldsymbol{\theta}^{(j+1)}$ from $p(\boldsymbol{\theta}|\mathbf{Z}^{(j+1)}, \mathbf{Y})$.

For sufficiently large j, the joint distribution of $(\boldsymbol{\theta}^{(j)}, \mathbf{Z}^{(j)})$ converges in distribution to the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y})$ (Geman and Geman, 1984; Tanner and Wong, 1987).

Check Convergence of MCMC Algorithm

a. The ‘estimated potential scale reduction (EPSR)’ values of the parameters. Convergence is claimed to be achieved if all EPSR values are less than 1.2 (Gelman, 1996).

b. Inspecting several parallel sequences of observations generated with different starting values (Gilks et al., 1996; among others).

Sample collection

a. The MCMC algorithm will continue to run for a sufficiently large number of iterations after convergence, so that the posterior distribution $p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y})$ can be approximated adequately by the empirical distribution of the simulated observations.

b. To reduce the serial correlation between consecutive observations, samples may be collected in cycles with indices $J_0 + s, J_0 + 2s, \cdots$, where J_0 is called burn-in. In many situations, a small s (e.g. $s = 1$) will suffice in estimation (Albert and Chib, 1993).

Bayesian estimation

Let $\{(\boldsymbol{\theta}^{(j)}, \mathbf{Z}^{(j)}) : j = 1, \cdots, J\}$ be a random sample generated from $p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y})$. The Bayesian estimates of $\boldsymbol{\theta}$ and $\mathbf{Z}$, and the standard error estimate of $\boldsymbol{\theta}$ can be obtained as follows:

$$\hat{\boldsymbol{\theta}} = \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\theta}^{(j)}, \quad \hat{\mathbf{Z}} = \frac{1}{J}\sum_{j=1}^{J}\mathbf{Z}^{(j)}, \tag{1.6}$$

$$\widehat{Var(\boldsymbol{\theta})} = \frac{1}{J-1}\sum_{j=1}^{J}(\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta})(\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta})^T. \tag{1.7}$$

Advantages of Maximum likelihood (ML) approach

- Prior specification and sensitivity analysis are not necessary.
- Asymptotic theories and nice properties of parameter estimators can be investigated.
- Relatively easy to apply computer package and computationally efficient.

Notations

$\mathbf{Y}$ — observed data

$\mathbf{Z}$ — latent quantities

$\boldsymbol{\theta}$ — parameters

$l(\boldsymbol{\theta})$ — observed-data log-likelihood function

$l_c(\boldsymbol{\theta})$ — complete-data log-likelihood function

In the presence of $\mathbf{Z}$, direct maximization of $l(\boldsymbol{\theta})$ is impossible because $l(\boldsymbol{\theta})$ has an intractable form

$$l(\boldsymbol{\theta}) = \log \int_{\mathbf{Z}} p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}.$$

EM Algorithm

EM algorithm (Dempster et al., 1977) regards $\mathbf{Z}$ as missing data and augments $\mathbf{Z}$ with the observed data $\mathbf{Y}$. At the tth iteration,

E-step: Compute Q-function:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E[l_c(\boldsymbol{\theta})|\mathbf{Y}, \boldsymbol{\theta}^{(t)}] = E[\log p(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta}^{(t)})], \tag{1.8}$$

where the expectation is taken with respect to $p(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}^{(t)})$.

M-step: Determine $\boldsymbol{\theta}^{(t+1)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ or equivalently by solving equation

$$\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} = E\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} l_c(\boldsymbol{\theta}) \middle| \mathbf{Y}, \boldsymbol{\theta}^{(t)} \right\} = 0. \tag{1.9}$$

Techniques in EM algorithm

1. In E-step, Monte Carlo integration and MCMC methods:

$$\widehat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \frac{1}{J}\sum_{j=1}^{J} \log p(\mathbf{Y}, \mathbf{Z}^{(j)}|\boldsymbol{\theta}^{(t)}),$$

   where $\{\mathbf{Z}^{(j)}, j = 1, \cdots, J\}$ are sampled from $p(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}^{(t)})$.

2. In M-step, conditional maximization, Newton Raphson algorithm, and other optimization methods may be used.

Check convergence of EM algorithm

1. Compute the observed-data likelihood and monitor its change via the ratio of two consecutive likelihood values. Let $R^{(t)} = l(\boldsymbol{\theta}^{(t+1)}) - l(\boldsymbol{\theta}^{(t)})$, then convergence is claimed to be achieved if the plot of $R^{(t)}$ against t shows a curve converging to zero (Meng and Schilling, 1996).

2. Check the absolute or relative error of the parameter estimates and monitor convergence via stopping rules (e.g., Shi and Copas, 2002; Lee and Song, 2004).

Akaike Information Criterion (AIC)

$$\text{AIC} = -2\{\log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) - d\} = -2\log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) + 2d, \tag{1.10}$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimate of $\boldsymbol{\theta}$, and d is the number of parameters involved. In the presence of $\mathbf{Z}$,

$$p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) = \int p(\mathbf{Y}|\mathbf{Z}, \hat{\boldsymbol{\theta}})p(\mathbf{Z}|\hat{\boldsymbol{\theta}})d\mathbf{Z}, \tag{1.11}$$

Bayesian Information Criterion (BIC)

$$\text{BIC} = -2\{\log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) - d\log n\} = -2\log p(\mathbf{Y}|\hat{\boldsymbol{\theta}}) + 2d\log n,$$

where n is the sample size.

Deviance Information Criterion (DIC)

DIC (Spiegelhalter et al., 2002) is an analog of AIC. It accounts for the goodness-of-fit and model complexity under a Bayesian framework. It is defined as

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D, \tag{1.12}$$

where $\overline{D(\boldsymbol{\theta})}$ measures the goodness-of-fit of the model, and

$$\overline{D(\boldsymbol{\theta})} = E_{\boldsymbol{\theta}}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta})|\mathbf{Y}\}. \tag{1.13}$$

p_D is the effective number of parameters, and is defined as

$$p_D = E_{\boldsymbol{\theta}}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta})|\mathbf{Y}\} + 2\log p(\mathbf{Y}|\widehat{\boldsymbol{\theta}}). \tag{1.14}$$

Computation of DIC

Let $\{\boldsymbol{\theta}_k^{(j)},\ j = 1, \cdots, J\}$ be a sample of observations simulated from the posterior distribution. The expectations in (1.13) and (1.14) can be estimated as follows:

$$E_{\boldsymbol{\theta}_k}\{-2\log p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)|\mathbf{Y}\} = -\frac{2}{J}\sum_{j=1}^{J}\log p(\mathbf{Y}|\boldsymbol{\theta}_k^{(j)}, M_k). \qquad (1.15)$$

The model with the smaller DIC value is selected. The cost of computing DIC is on simulating $\{\boldsymbol{\theta}_k^{(j)},\ j = 1, \cdots, J\}$ from the posterior distribution, and is lighter than that of Bayes factor.

Extension of DIC

Celeux et al. (2006) proposed an extension for incomplete data:

$$\text{DIC} = -4E_{\boldsymbol{\theta},\mathbf{Z}}\{\log p(\mathbf{Y},\mathbf{Z}|\boldsymbol{\theta})|\mathbf{Y}\} + 2E_{\mathbf{Z}}\{\log p(\mathbf{Y},\mathbf{Z}|E_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathbf{Y},\mathbf{Z}])|\mathbf{Y}\}, \tag{1.16}$$

where $\log p(\mathbf{Y},\mathbf{Z}|\boldsymbol{\theta})$ is the complete-data log-likelihood function. The first expectation of DIC is obtained by

$$E_{\boldsymbol{\theta},\mathbf{Z}}\{\log p(\mathbf{Y},\mathbf{Z}|\boldsymbol{\theta})|\mathbf{Y}\} \approx \frac{1}{J}\sum_{j=1}^{J}\log p(\mathbf{Y},\mathbf{Z}^{(j)}|\boldsymbol{\theta}^{(j)}), \tag{1.17}$$

where $\{(\mathbf{Z}^{(j)},\boldsymbol{\theta}^{(j)});j=1,\cdots,J\}$ are generated from $p(\mathbf{Z},\boldsymbol{\theta}|\mathbf{Y})$.

Extension of DIC

Let $\boldsymbol{\theta}^{(j,l)}, \; l = 1, \cdots, L$ be generated from $p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}^{(j)})$, we have

$$E_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}^{(j)}] \approx \bar{\boldsymbol{\theta}}^{(j)} = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{\theta}^{(j,l)}.$$

The second expectation of DIC is approximated by

$$E_{\mathbf{Z}}\{\log p(\mathbf{Y}, \mathbf{Z}|E_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}])|\mathbf{Y}\} \approx \frac{1}{J}\sum_{j=1}^{J} \log p(\mathbf{Y}, \mathbf{Z}^{(j)}|\bar{\boldsymbol{\theta}}^{(j)}). \quad (1.18)$$

Finally, we can obtain the approximation of the modified DIC:

$$\text{DIC} = -\frac{4}{J}\sum_{j=1}^{J} \log p(\mathbf{Y}, \mathbf{Z}^{(j)}|\boldsymbol{\theta}^{(j)}) + \frac{2}{J}\sum_{j=1}^{J} \log p(\mathbf{Y}, \mathbf{Z}^{(j)}|\bar{\boldsymbol{\theta}}^{(j)}). \quad (1.19)$$

Bayes factor

Let M_0 and M_1 be two competing models for the given data set $\mathbf{Y}$, $p(M_0)$ be the prior probability of M_0, $p(M_1) = 1 - p(M_0)$, and $p(M_k|\mathbf{Y})$ be the posterior probability for $k = 0,\ 1$. Then,

$$p(M_k|\mathbf{Y}) = \frac{p(\mathbf{Y}|M_k)p(M_k)}{p(\mathbf{Y}|M_1)p(M_1) + p(\mathbf{Y}|M_0)p(M_0)}, \quad k = 0,\ 1.$$

Hence,

$$\frac{p(M_1|\mathbf{Y})}{p(M_0|\mathbf{Y})} = \frac{p(\mathbf{Y}|M_1)p(M_1)}{p(\mathbf{Y}|M_0)p(M_0)}. \tag{1.20}$$

The Bayes factor for comparing M_1 and M_0 is defined as

$$B_{10} = \frac{p(\mathbf{Y}|M_1)}{p(\mathbf{Y}|M_0)}. \tag{1.21}$$

Bayes Factor

So, posterior odds = Bayes factor $\times$ prior odds. In the special case of $p(M_1) = p(M_0) = 0.5$, the Bayes factor is equal to the posterior odds. Bayes factor has the following features:

1. It may reject a null hypothesis associated with M_0, or may equally provide evidence in favor of M_0 or (alternative) M_1.
2. Bayes factor does not depend on the assumption that either model is 'true'.
3. The same data set used in the comparison. Thus, it does not favor the alternative hypothesis (M_1) in extremely large samples.
4. It can be applied to compare nonnested models.

Bayes Factor

The criterion for interpreting B_{10} (Kass and Raftery, 1995) :

B_{10}	$2\log B_{10}$	Evidence against $H_0(M_0)$
< 1	< 0	Negative (supports $H_0(M_0)$)
1 to 3	0 to 2	Not worth more than a bare mention
3 to 20	2 to 6	Positive (supports $H_1(M_1)$)
20 to 150	6 to 10	Strong
> 150	> 10	Decisive

Bayes Factor

Let $\boldsymbol{\theta}_k$ be the parameter vector associated with M_k. From

$$p(\boldsymbol{\theta}_k, \mathbf{Y}|M_k) = p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k),$$

we have

$$p(\mathbf{Y}|M_k) = \int p(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k, \tag{1.22}$$

where $p(\boldsymbol{\theta}_k|M_k)$ is the prior density of $\boldsymbol{\theta}_k$.

Computing B_{10} is difficult. Various numerical approximations have been proposed in the literature (Chib, 1995). We discuss the path sampling procedure (Gelman and Meng, 1998).

Bayes Factor

Let $\mathbf{Y}$ be the observed data and $\mathbf{\Omega}$ be the latent data. Note that

$$p(\mathbf{\Omega}, \boldsymbol{\theta}|\mathbf{Y}) = p(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta})/p(\mathbf{Y}).$$

Now, we consider the following class of densities, which are denoted by a continuous parameter t in $[0, 1]$:

$$p(\mathbf{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t) = \frac{1}{z(t)} p(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}|t), \tag{1.23}$$

where

$$z(t) = p(\mathbf{Y}|t) = \int p(\mathbf{Y}, \mathbf{\Omega}, \boldsymbol{\theta}|t) d\mathbf{\Omega} d\boldsymbol{\theta} = \int p(\mathbf{Y}, \mathbf{\Omega}|\boldsymbol{\theta}, t) p(\boldsymbol{\theta}) d\mathbf{\Omega} d\boldsymbol{\theta}. \tag{1.24}$$

Bayes Factor

Using $t \in [0, 1]$ to construct a path to link M_1 and M_0:

$$z(1) = p(\mathbf{Y}|1) = p(\mathbf{Y}|M_1), \quad z(0) = p(\mathbf{Y}|0) = p(\mathbf{Y}|M_0),$$

and $B_{10} = z(1)/z(0)$. Taking logarithm and then differentiating (1.24) with respect to t, and assuming the legitimacy of interchange of integration with differentiation, we have

$$\begin{aligned}
\frac{d \log z(t)}{dt} &= \int \frac{1}{z(t)} \frac{d}{dt} p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t) d\boldsymbol{\Omega} d\boldsymbol{\theta} = \int \frac{p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t)}{p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t)} \frac{d}{dt} p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t) \\
&= \int \frac{d}{dt} \log p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t) \cdot p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t) d\boldsymbol{\Omega} d\boldsymbol{\theta} \\
&= E_{\Omega,\theta} \left[\frac{d}{dt} \log p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}|t) \right],
\end{aligned}$$

where $E_{\Omega,\theta}$ denotes the expectation with respect to $p(\boldsymbol{\Omega}, \boldsymbol{\theta}|\mathbf{Y}, t)$.

Bayes Factor

Let $U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t) = \frac{d}{dt} \log p(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta} | t) = \frac{d}{dt} \log p(\mathbf{Y}, \boldsymbol{\Omega} | \boldsymbol{\theta}, t)$, then

$$\log B_{10} = \log \frac{z(1)}{z(0)} = \int_0^1 E_{\Omega,\theta}[U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t)]dt. \tag{1.25}$$

Let $0 = t_{(0)} < t_{(1)} < \cdots < t_{(S)} < t_{(S+1)} = 1$ be fixed grids. Then, the integral (1.25) can be obtained as follows:

$$\widehat{\log B_{10}} = \frac{1}{2} \sum_{s=0}^{S} (t_{(s+1)} - t_{(s)})(\bar{U}_{(s+1)} + \bar{U}_{(s)}), \tag{1.26}$$

where

$$\bar{U}_{(s)} = J^{-1} \sum_{j=1}^{J} U(\mathbf{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}), \tag{1.27}$$

and $\{(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}),\ j = 1, \cdots, J\}$ are drawn from $p(\boldsymbol{\Omega}, \boldsymbol{\theta} | \mathbf{Y}, t_{(s)})$.

Steps in implementing the path sampling procedure:

1. Define a link model M_t to link M_0 and M_1, such that when $t = 0$, $M_t = M_0$; and when $t = 1$, $M_t = M_1$.
2. Obtain $U(\mathbf{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t)$ by differentiating the logarithm of the complete-data likelihood function under M_t with respect to t
3. Estimate $\log B_{10}$ via (1.26) and (1.27). For most statistical models, $S = 20$ and $J = 1,000$ provide reliable results. Experiences indicate that $S = 10$ is also acceptable for simple models.

Bayesian Lasso

The least absolute shrinkage and selection operator (Lasso) was first introduced by Tibshirani (1996) in a linear model

$$\mathbf{y} = \mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n), \tag{1.28}$$

where $\mathbf{1}_n$ is a vector of all elements being 1, $\mathbf{X}$ is a standardized design matrix. The Lasso estimator of $\boldsymbol{\beta}$ can be viewed as the L_1-penalized least squares estimate obtained from

$$\arg\min_{\beta}\{(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \gamma \sum_{j=1}^{p} |\beta_j|\}, \tag{1.29}$$

where $\gamma \geq 0$, and $\tilde{\mathbf{y}} = \mathbf{y} - \mu \mathbf{1}_n$.

Bayesian Lasso

Park and Casella (2008) introduced Lasso to the Bayesian framework. The basic idea is to penalize $\boldsymbol{\beta}$ by imposing a conditional Laplace prior on $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\gamma}{2\sigma} e^{-\gamma|\beta_j|/\sigma}, \tag{1.30}$$

BLasso can be formulated by a hierarchical representation:

$$\begin{aligned}
&[\mathbf{y}|\mu, \mathbf{X}, \boldsymbol{\beta}, \sigma^2] \sim N_n(\mu\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \qquad (1.31)\\
&[\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2] \sim N_p(\mathbf{0}_p, \sigma^2\mathbf{D}_\tau), \ \mathbf{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2),\\
&\sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim \pi(\sigma^2)d\sigma^2 \prod_{j=1}^{p} \frac{\gamma^2}{2} e^{-\gamma^2\tau_j^2/2} d\tau_j^2, \ \sigma^2, \tau_1^2, \ldots, \tau_p^2 > 0.
\end{aligned}$$

(1.30) can be obtained by integrating out $\tau_1^2, \ldots, \tau_p^2$ from (1.31).

Bayesian Lasso

Specifically, the model can be reformulated in the following hierarchical representation:

$$
\begin{aligned}
&[\mathbf{y}|\mu, \mathbf{X}, \boldsymbol{\beta}, \sigma^2] \sim N_n(\mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \\
&[\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2] \sim N_p(\mathbf{0}_p, \sigma^2 \mathbf{D}_\tau), \quad \mathbf{D}_\tau = \text{diag}(\tau_1^2, \ldots, \tau_p^2), \\
&\tau_j^2 \sim \text{Gamma}(1, \frac{\gamma^2}{2}), \\
&\gamma^2 \sim \text{Gamma}(a_0, b_0), \\
&\sigma^2 \propto \frac{1}{\sigma^2} \ \text{(or an inverse-gamma prior for } \sigma^2\text{)}.
\end{aligned}
$$

Bayesian Lasso

Full conditional distributions:

$$[\boldsymbol{\beta}|\cdot] \sim N(\mathbf{A}^{-1}\mathbf{X}^T\tilde{\mathbf{y}},\ \sigma^2\mathbf{A}^{-1}),\ \ \mathbf{A} = \mathbf{X}^T\mathbf{X} + \mathbf{D}_\tau^{-1},$$

$$[\sigma^2|\cdot] \sim IG\Big(\frac{n-1}{2} + \frac{p}{2},\ \frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{D}_\tau^{-1}\boldsymbol{\beta}\Big),$$

$$[\frac{1}{\tau_j^2}|\cdot] \sim \text{IGaussian}\Big(\sqrt{\frac{\gamma^2\sigma^2}{\beta_j^2}},\ \gamma^2\Big),$$

$$[\gamma^2|\cdot] \sim \text{Gamm}(a_0 + p,\ b_0 + \sum_{i=1}^{p}\frac{\tau_j^2}{2}),$$

where $\mathbf{D}_\tau^{-1} = \text{diag}(1/\tau_1^2, \cdots, 1/\tau_p^2)$, and $a_0 = 1$ and $b_0 = 0.1$, making $p(\gamma^2)$ highly dispersed.

Bayesian Lasso

The above full conditionals form the basis for an efficient Gibbs sampler, with block updating of $\boldsymbol{\beta}$ and $(\tau_1^2, \cdots, \tau_p^2)$.

BLasso provides a posterior sample that can be used to summarize the entire distribution of $\boldsymbol{\beta}$. The posterior mean or mode of $\boldsymbol{\beta}$ can be regarded as its Lasso estimator.

Given that BLasso is a sampling-based method, it would not shrink the nonsignificant elements of $\boldsymbol{\beta}$ exactly to 0. A cutoff value must be set.

Bayesian Adaptive Lasso

A Bayesian version of adaptive Lasso can be obtained by assigning a conditional Laplace prior with coefficient-specific turning parameters as follows:

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\gamma_j}{2\sigma} e^{-\gamma_j|\beta_j|/\sigma}. \tag{1.32}$$

Bayesian adaptive Lasso introduces different penalties to various coefficients to enhance its capability of producing good estimation and model selection results.

Bayesian Adaptive Lasso

Full conditional distributions:

$$[\boldsymbol{\beta}|\cdot] \sim N(\mathbf{A}^{-1}\mathbf{X}^T\tilde{\mathbf{y}},\ \sigma^2\mathbf{A}^{-1}),\quad \mathbf{A} = \mathbf{X}^T\mathbf{X} + \mathbf{D}_\tau^{-1},$$

$$[\sigma^2|\cdot] \sim IG\Big(\frac{n-1}{2} + \frac{p}{2},\ \frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{D}_\tau^{-1}\boldsymbol{\beta}\Big),$$

$$[\frac{1}{\tau_j^2}|\cdot] \sim \text{IGaussian}\Big(\sqrt{\frac{\gamma_j^2\sigma^2}{\beta_j^2}},\ \gamma_j^2\Big),$$

$$[\gamma_j^2|\cdot] \sim \text{Gamm}(a_0 + 1,\ b_0 + \frac{\tau_j^2}{2})$$

where $\mathbf{D}_\tau^{-1} = \text{diag}(1/\tau_1^2, \cdots, 1/\tau_p^2)$, and $a_0 = 1$ and $b_0 = 0.1$, making $p(\gamma^2)$ highly dispersed.

1.5 Computer Software

WinBUGS

The freely available software WinBUGS (**W**indows version of **B**ayesian inference **U**sing **G**ibbs **S**ampling) is useful for producing Bayesian results for statistical models. WinBUGS can be downloaded from the website: http://www.mrc-bsu.cam.ac.uk/bugs/. The WinBUGS manual (Spiegelhalter *et al*., 2003) is available online.

R code

R can be used to conduct analysis for all the models introduced in this course.

R2WinBUGS

R package R2WinBUGS (Sturtz, Ligges and Gelman, 2005) provides tools to directly call WinBUGS after the manipulation in R. Then, it is possible to work on the results after importing them back into R.

The implementation of R2WinBUGS is mainly based on the R function ‘bugs($\cdots$)’, which takes data and initial values as input. It automatically writes a WinBUGS script, calls the model, and saves the simulation for easy access in R.

Stan

- A general program written in C++, Stan (Stan Development Team, 2017), with an R software interface for data inputs and for summarizing results.
- Stan implements gradient-based MCMC algorithms for Bayesian inference. It is an open-source, general purpose programming language for conducting Bayesian analysis with the Hamiltonian Monte Carlo (HMC) method.
- HMC directly analyzes the gradient of the log-posterior to avoid the sensitive random walk behavior of traditional MCMC methods. It provides efficient parameter space exploration even for correlated posteriors.
- The convergence of the HMC method is fast.
- The posterior samples obtained from the HMC algorithm provide summary measures, including posterior means and 95% credible intervals, of the parameters.

References

1. Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *JASA*, **88**, 669-679.
2. Carlin, B. P. and Louis, T. A. (2006). *Bayesian Methods for Data Analysis*. Third Edition. Chapman & Hall.
3. Celeux, et al. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651-674.
4. Chib, S. (1995) Marginal likelihood from the Gibbs output. *JASA*, **90**, 1313-1321.
5. Congdon, P. (2006). *Bayesian Statistical Modelling*. John Wiley & Sons.
6. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *JRSSB*, **39**, 1-38.
7. Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, et al. (Eds), *Markov Chain Monte Carlo in Practice*, 131-144. Chapman & Hall.
8. Gelman, A. and Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**, 163-185.
9. Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.
10. Gilks, W. R., et al. (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

References

11. Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *JASA*, **90**, 773-795.
12. Lee, S. Y. and Song, X. Y. (2004). Maximum likelihood analysis of a general latent variable model with hierarchically mixed data. *Biometrics*, **60**, 624-636.
13. Meng, X. L. and Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *JASA*, **91**, 1254-1267.
14. Park, T. and G. Casella (2008). The Bayesian Lasso. *JASA*, **103**, 681-686.
15. Shi, J. Q. and Copas, J. (2002). Publication bias and meta-analysis for 2×2 tables: an average Markov chain Monte Carlo EM algorithm. *JRSSB*, **64**, 221-236.
16. Spiegelhalter, D. J., et al. (2002). Bayesian measures of model complexity and fit (with discussion). *JRSSB*, **64**, 583-639.
17. Stan Development Team (2017). Stan Modeling Language User's Guide and Reference Manual.
18. Sturtz, S., Ligges, U. and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, **12**, 1-16.
19. Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *JASA*, **82**, 528-550.
20. Tibshirani, R. (1996). Regression Shrinkage and Selection via a Lasso. *JRSSB*, **58**, 267-288.