INDIAN STATISTICAL INSTITUTE

M.Stat. 2nd Year

BAYESIAN INFERENCE

ASSIGNMENT II

1. Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the vector of observations on the "dependent" variable, $\boldsymbol{X} = ((x_{ij}))_{n\times p}$ is of full rank, x_{ij} being the values of the nonstochastic regressor variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients and the components of $\boldsymbol{\epsilon}$ are independent, each following $N(0, \sigma^2)$. Consider the noninformative prior $\pi(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$, $\boldsymbol{\beta} \in \mathcal{R}^p$, $\sigma^2 > 0$.

(a) Find the $100(1-\alpha)\%$ HPD credible set for $\boldsymbol{\beta}$.

(b) Find the marginal posterior distribution of a particular β_j ($j = 1, \ldots, p$) and use it to find the $100(1-\alpha)\%$ HPD credible set for β_j.

2. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ variables where μ and σ^2 are both unknown. Consider the prior $\pi(\mu, \sigma^2) \propto 1/\sigma^2$. Show that the posterior predictive distribution of a future observation X_{n+1} is a t distribution with $n-1$ d.f., location $\bar{X}_n$ and scale $(1+1/n)^{1/2}s$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

3. Consider the setup of the theorem on asymptotic normality of posterior distribution (Theorem 4.2 of the book by Ghosh et al. (2006)), proved in the class for a proper prior.

(a) Suppose that the prior is improper but there is an n_0 such that the posterior distribution of θ given $x_1, \ldots, x_{n_0}$ is proper for *a.e.* $(x_1, \ldots, x_{n_0})$. Show that the theorem holds also in this case.

(b) In addition to the assumptions of Theorem 4.2, assume that the prior density $\pi(\theta)$ has a finite expectation. Proceeding as in the proof of Theorem 4.2 and using the assumption of finite expectation for π, show that

$$\int_{\mathcal{R}} |t|\, |\pi_n^*(t|X_1, \ldots, X_n) - \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)}|\, dt \to 0$$

with P_{θ_0}-probability one.

4. Suppose we have observations $X_1, \ldots, X_n$. Under model M_0, X_i are i.i.d. $N(0,1)$ and under model M_1, X_i are i.i.d. $N(\theta, 1)$, $\theta \in \mathcal{R}$. Consider the noninformative prior $g_1(\theta) \equiv 1$ for θ under M_1. Show that if we use training samples of size 2 and calculate the corresponding AIBF, the corresponding intrinsic prior will be $N(0,1)$.

Bayesian variable selection based on g-prior in normal linear regression models

Consider the regression problem with response variable y and a set of potential predictor variables $x_1, x_2, \ldots, x_p$. Let $\mathbf{y}_n = (y_1, y_2, \ldots, y_n)'$ be a vector of observations on the response variable and $\mathbf{X}_n = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ be an $n \times p$ design matrix. Here $\mathbf{x}_i$ is an $n \times 1$ vector of observations on the i^{th} regressor x_i and the j^{th} component of $\mathbf{x}_i$ is associated with y_j, $i = 1, \ldots, p$, $j = 1, \ldots, n$. We assume, without loss of generality, that the columns of $\mathbf{X}_n$ have been centered so that $\mathbf{1}_n' \mathbf{x}_i = 0$ for all i where $\mathbf{1}_n$ is a vector of 1's of length n. Let $\boldsymbol{\mu}_n$ denote $E(\mathbf{y}_n | \mathbf{X}_n)$ and assume

$$\mathbf{y}_n \sim N_n \left(\boldsymbol{\mu}_n, \sigma^2 I_n \right),$$

where σ^2 is unknown and I_n is the $n \times n$ identity matrix. We are interested in capturing the functional relationship, if any, between $\boldsymbol{\mu}_n$ and $\mathbf{X}_n$.

We restrict our search within the class of normal linear models under which $\boldsymbol{\mu}_n$ may be expressed as

$$\boldsymbol{\mu}_n = \mathbf{1}_n \beta_0 + \mathbf{X}_n \boldsymbol{\beta}, \tag{1}$$

where β_0 is an intercept and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of regression coefficients. Our problem is to select a subset of the potential predictor variables $x_1, x_2, \ldots, x_p$. Thus we have a model selection problem and our model space, denoted by $\boldsymbol{\mathcal{A}}$, may be indexed by α, where each α consists of a subset of size $p(\alpha)$ $(1 \leq p(\alpha) \leq p)$ of $\{1, 2, \ldots, p\}$, indicating which regressors are included in the model. The model M_α corresponding to $\alpha \in \boldsymbol{\mathcal{A}}$ may be expressed as a sub-model of (1),

$$M_\alpha \quad : \quad \boldsymbol{\mu}_n = \mathbf{1}_n\beta_0 + \mathbf{X}_{n\alpha}\boldsymbol{\beta}_\alpha, \tag{2}$$

where the intercept β_0 is common to all models, $\mathbf{X}_{n\alpha}$ is a sub-matrix of $\mathbf{X}_n$ consisting of the $p(\alpha)$ columns specified by α and $\boldsymbol{\beta}_\alpha$ is the $p(\alpha)$- dimensional vector of regression coefficients.

Bayesian model selection requires specification of prior distribution of the parameters $\boldsymbol{\theta}_\alpha = (\beta_0, \boldsymbol{\beta}_\alpha, \sigma^2) \in \Theta_\alpha$ under each model M_α and prior probabilities $p(M_\alpha)$ of the models. Let $p(\mathbf{y}_n|\boldsymbol{\theta}_\alpha, M_\alpha)$ denote the density of $\mathbf{y}_n$ given $\boldsymbol{\theta}_\alpha$ under M_α and $p(\boldsymbol{\theta}_\alpha|M_\alpha)$ denote the prior density of $\boldsymbol{\theta}_\alpha$ under M_α. Then the posterior probability of the model M_α, $\alpha \in \boldsymbol{\mathcal{A}}$, is given by

$$p(M_\alpha|\mathbf{y}_n) = \frac{p(M_\alpha)m_\alpha(\mathbf{y}_n)}{\sum_{\alpha\in\boldsymbol{\mathcal{A}}} p(M_\alpha)m_\alpha(\mathbf{y}_n)}, \tag{3}$$

$$\text{where} \quad m_\alpha(\mathbf{y}_n) = \int p(\mathbf{y}_n|\boldsymbol{\theta}_\alpha, M_\alpha)p(\boldsymbol{\theta}_\alpha|M_\alpha)d\boldsymbol{\theta}_\alpha \tag{4}$$

is the marginal density of $\mathbf{y}_n$ under M_α. In this paper, we consider the model selection procedure that selects the model with highest posterior probability.

A very popular conventional prior for the parameters $\boldsymbol{\beta}_\alpha$ is the conjugate g-prior due to Zellner (1986) given in (6). In the present scenario, β_0 and σ^2 may be regarded as parameters common to all the models and the suggested default priors are

$$p(\beta_0, \sigma^2|M_\alpha) = \frac{1}{\sigma^2} \tag{5}$$

$$\boldsymbol{\beta}_\alpha|\beta_0, \sigma^2, M_\alpha \sim N_{p(\alpha)}(\mathbf{0}, g\sigma^2(\mathbf{X}'_{n\alpha}\mathbf{X}_{n\alpha})^{-1}) \tag{6}$$

for some $g > 0$ (see, for example, Liang et al. (JASA, 2008), Section 2.1).

Given the priors (5) and (6), the marginal likelihood under the model M_α, $\alpha \in \boldsymbol{\mathcal{A}}$, is given by

$$\begin{aligned} m_\alpha(\mathbf{y}_n) &= \frac{\Gamma(n-1)/2}{\pi^{(n-1)/2}\sqrt{n}\ (1+g)^{p(\alpha)/2}} \\ &\quad \times \left[(1-a)\sum_{i=1}^{n}(y_i - \overline{y})^2 + a\mathbf{y}'_n(I_n - P_n(\alpha))\mathbf{y}_n\right]^{-(n-1)/2} \end{aligned} \tag{7}$$

where $a = g/(1+g)$ and $P_n(\alpha) = \mathbf{Z}_{n\alpha} \left[\mathbf{Z}'_{n\alpha}\mathbf{Z}_{n\alpha}\right]^{-1} \mathbf{Z}'_{n\alpha}$ is the projection matrix onto the span of $\mathbf{Z}_{n\alpha} = \left[\mathbf{1}_n, \mathbf{X}_{n\alpha}\right], \alpha \in \boldsymbol{\mathcal{A}}$. The model selection rule is to choose the model M_α with highest posterior probability, that is, we choose the model M_α for which $p(M_\alpha)m_\alpha(\mathbf{y}_n)$ is the largest among all $\alpha \in \boldsymbol{\mathcal{A}}$.

5. Show that the marginal likelihood $m_\alpha(\mathbf{y}_n)$ under the model M_α in the above variable selection problem is given by (7) above.

6. Consider the example of hierarchical Bayesian analysis of the usual one-way ANOVA (Example 7.13 of the book by Ghosh et al. 2006, page 227) discussed in the class.

Take $k = 10$ and $n_i = 25$ for all i. Generate 10 samples, each of size 25, from 10 normal populations. Choose 10 different values of the population means $\theta_1, \ldots, \theta_{10}$ and a common value of the population variance. Choose the values of the hyperparameters $a_1, a_2, b_1, b_2, \mu_0, \sigma_0^2$ so that the corresponding priors are not very informative. Use the Gibbs sampling procedure to estimate the poulation means. Do the same for five different choices of $(a_1, a_2, b_1, b_2, \mu_0, \sigma_0^2)$.

Do the above (a) using WinBUGS and also (b) using your own code written in your favourite programming language such as R or Python.