Indian Statistical Institute

M.Stat. Second year, Final Semester Examination 2025-26

Statistical Inference II

Maximum Marks: 100 Duration: 3 hours Date: 26.11.2025

Answer ALL question. Show your work to get full credit.

1. Consider multiple linear regression with p covariates and a single continuous response. Write the complete Hierarchical Bayes model of the Ridge regression for this setting. Derive the joint posterior distribution, all the full conditional distributions, and outline the MCMC algorithm that can be used to estimate the model parameters. [20]

2. (a) Consider the following model:
   $X_i|\lambda_i \overset{ind}{\sim}$ Poisson(λ_i), $i = 1, 2, \ldots, n$,
   $\lambda_i \overset{iid}{\sim}$ Gamma (α, β), α known, β unknown.

   Derive Empirical Bayes estimate of λ_i.

   (b) Let $X_1, X_2, \ldots, X_n$, $n > 3$, be independently distributed as follows:
   $X_i|\theta_i \overset{ind}{\sim} N(\theta_i, \sigma^2)$, where σ^2 is known.

   Define the James-Stein estimator of θ_i. Show that this estimator can indeed be viewed as an Empirical Bayes estimator. [10+10=20]

3. Suppose a psychometric study results in n binary outcomes $Y_1, Y_2, \ldots, Y_n$ with p continuous covariates measured for each subject under study. The goal is to develop a suitable predictive model with an appropriate subset of p covariates. Develop a Bayesian binary regression model following Albert and Chib (1993) considering an appropriate shrinkage prior for the regression coefficients. Specifically, (i) write down the complete Hierarchical Bayes model, (ii) derive the joint posterior distribution, and outline a suitable MCMC algorithm for estimating the model parameters. [10+10+5=25]

4. A medical study was conducted in the United States during 1980-1989, where 3071 counties were surveyed for the patients suffering from kidney cancer. Let Y_j denote the number of kidney cancer deaths from the j-th county in this period and we propose two competing models for analysing the dataset, and n_j is the population of the county.

   Model 1:

   $Y_j|\theta_j \overset{ind}{\sim}$ Poisson$(10n_j\theta_j)$, $\theta_j \overset{iid}{\sim} \sum_{h=1}^{K} \pi_h \delta_{\theta_h}$, $\theta_h \sim$ Gamma(α, β), with $\alpha = 20$, $\beta = 430,000$, $K = 25$, $\pi = (\pi_1, \pi_2, \ldots, \pi_k) \sim$ Dirichlet$(1/K, 1/K, \ldots, 1/K)$.

Model 2:

$Y_j|\theta_j \overset{ind}{\sim} \text{Poisson}(10n_j\theta_j)$, $\theta_j \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$, with $\alpha = 20$, $\beta = 430,000$.

(a) How does Model 1 differ from Model 2? Under what type of data set would you prefer Model 1?

(b) Suppose that you are not sure which model is a good fit for your data. You are advised to fit both models and compare them. How will you select the model most suitable for your dataset? Define the statistic you use and comment on its usefulness and limitation (if any). [8+7=15]

5. Suppose you have a mixture model with five components. For each data point you want to identify which of the five components it comes from. It would be best to use the full posterior distribution, but you need a point estimate. Which of the following would you prefer: the posterior mean, the posterior median, or the posterior mode? Assume each of these is done pointwise, that is, you are getting the posterior mean median and mode of the latent component for each data point, and not the joint mean, median, or mode for all the data points at once. [10]

6. Write short notes on any two of the following:

(i) Indian Buffet Process, (ii) Stick-Breaking formulation of Dirichlet Process, (iii) Horseshoe prior in variable selection. [5+5=10]