Stat 5444: Model Selection

Consider the ('semi') linear model of the form:

$$y_i = \sum_{j=1}^{p} f_j(x_{i,j})\beta_j + \epsilon_i, \qquad i = 1, \ldots, N, \tag{1}$$

where $\epsilon_i \sim Normal(0, \sigma^2)$. More specifically, we can write $Y = X\beta + \epsilon$, where $Y = [y_1, \ldots, y_n]^T$, $\beta = [\beta_1, \ldots, \beta_p]^T$, and $X_{i,j} = f_j(x_{i,j})$. Note: In general, figuring out the functions $f_j(\cdot)$ can be difficult, but as a cursory step it is common to look at polynomial evaluations of $x_{i,j}$ (which can potentially lead to the dimensionality of the problem (p) being very large).

Problem 1

Work out the Bayes factor for comparing 2 different models, where each model is of the form given in equation (1), but differ by the number of free parameters (i.e. the number of coefficients where $\beta_j \neq 0$).

Problem 2

Using the data provided in **Model1_5444.txt**, search for the 'true' model using the following selection procedures:

- **Least Absolute Shrinkage Selection Operator** (LASSO)

$$\hat{\beta}_{LASSO} = ||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

- **Bayesian Information Criterion (BIC)**: 'Deviance' $+ \log(N)\Delta_p$, where Δ_p denotes the difference in the number of parameters used in the compared models,

- **Akaike Information Criterion (AIC)**: 'Deviance' $+ 2\Delta_p$, where Δ_p denotes the difference in the number of parameters used in the compared models,

- **Stochastic Search Variable Selection** (SSVS):

$$\pi(\beta_j) = \pi_0\delta(\beta_j = 0) + (1 - \pi_0)N(\beta_j|0, \psi^2).$$

In your comparison between methods, describe in detail how you 'tuned' your method (if tuning is required), and report your selected model (including coefficient estimates).

Problem 3

Repeat the previous exercise using the data provided in **Model2_5444.txt**.

Problem 4

Repeat the previous exercise using the data provided in **Model3_5444.txt**