Results Needed for the Spline Fitting Theory

Result 1: Let $s(x)$ be the cubic interpolating spline for the points (x_i, y_i), $i = 0, 1, ..., n$. Suppose that $g(x)$ is twice continuously differentiable and also interpolates the given points (x_i, y_i). Show that

$$\int_{x_0}^{x_n} \left[g''(x) - s''(x)\right] s''(x)\, dx = 0.$$

Hint: break the integral into n sub-integrals and use integration by parts.

Result 2: Let $s(x)$ be the cubic interpolating spline for the points (x_i, y_i), $i = 0, 1, ..., n$. Show that

$$\int_{x_0}^{x_n} s''(x)^2\, dx = \gamma' R \gamma$$

where

$$\gamma = \begin{pmatrix} s''(x_1) \\ s''(x_2) \\ \vdots \\ s''(x_{n-1}) \end{pmatrix}$$

and R is the tri-diagonal matrix defined in the course notes.

Estimating A-B-O Alleles Frequencies Using Phenotype Data

Consider the ABO locus on the long arm of chromosome 9. This locus determines detectable antigens on the surface of red blood cells. There are three alleles, A, B and O, which correspond to an A antigen, a B antigen and the absence of either antigens, respectively. Phenotypes are recorded by reacting antibodies for A and B against a blood sample. The four observable phenotypes are A (antigen A alone detected), B (antigen B alone detected), AB (antigens A and B both detected), and O (neither antigen A nor B detected).

Phenotypes	Genotypes
A	A/A , A/O
B	B/B , B/O
AB	A/B
O	O/O

The phenotype of 600 individuals are reported below:

Phenotypes	Frequencies
A	210
B	50
AB	20
O	320

Assume that the population is at Hardy-Weinberg equilibrium, that is, the genotype probabilities are

Genotype	A/A	A/B	B/B	O/O	O/A	O/B
Probability	p_A^2	$2p_Ap_B$	p_B^2	p_O^2	$2p_Op_A$	$2p_Op_B$

where p_A, p_B and p_O are the frequencies of alleles A, B and O in the population. Use the EM algorithm to estimate p_A, p_B and p_O.

Mixture Distribution: The Faithful Data

The dataset "faithful" available in R gives the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. Suppose the data comes from a mixture of Gaussian distributions with unknown means and common covariance matrix. Use the EM algorithm to estimate the unknown model parameters and to assign the observations to the different populations.

Bayesian Updating

Suppose that given $\boldsymbol{\theta},$ the observations $Y_1, Y_2, ..., Y_n$ are i.i.d. with common density $f(y|\boldsymbol{\theta})$. Let $\pi(\boldsymbol{\theta})$ be the prior density.Set

$$
\begin{aligned}
\pi^{(1)}(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}|y_1) \\
\pi^{(2)}(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}|y_1, y_2) \\
&\vdots \\
\pi^{(k)}(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}|y_1, y_2, ..., y_k), \quad \text{etc.}
\end{aligned}
$$

Derive and interpret the following recursive formula:

$$\pi^{(k+1)}(\boldsymbol{\theta}) = \frac{f(y_{k+1}|\boldsymbol{\theta})\,\pi^{(k)}(\boldsymbol{\theta})}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(y_{k+1}|\boldsymbol{\theta})\,\pi^{(k)}(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

Bayesian Predictive Distribution

Show that, by the assumption of (conditional) independence

$$f(y_{n+1}, \boldsymbol{\theta} \mid y_1, y_2, ..., y_n) = f(\boldsymbol{\theta} \mid y_1, y_2, ..., y_n)\, f(y_{n+1} \mid \boldsymbol{\theta}, y_1, y_2, ..., y_n)$$

$$= \pi^{(n)}(\boldsymbol{\theta})\, f(y_{n+1} \mid \boldsymbol{\theta}) \qquad (1)$$

Threfore the predictive density for the next measurement, Y_{n+1} given $Y_1, Y_2, ..., Y_n$ is

$$h(y_{n+1}|y_1, y_2, ..., y_n) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(y_{n+1}|\boldsymbol{\theta})\,\pi^{(n)}(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

Explain how formula (1) suggests a procedure to sample from the predictive distribution. Use this procedure, along with the data from Problem 6, to give a 95% predictive interval for y_{n+1}.

Application of MCMC Algorithms

Suppose that $Y_i \sim B(n_i, p_i)$, $i = 1, 2, ..., n$ are independent binomial random variables with

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$$

for some fixed given x_i and n_i.

(a) Derive the MLE estimate $\left(\hat{\alpha}, \hat{\beta}\right)$ for (α, β). Notice that the estimates cannot be expressed in closed form. Propose an iterative algorithm to compute them.

(b) Derive the asymptotic distribution of

$$\sqrt{n}\left[\left(\hat{\alpha},\hat{\beta}\right)-(\alpha,\beta)\right].$$

Here you may proceed in a heuristic manner, without worrying too much about regularity assumptions.

(c) Construct approximate 95% confidence intervals for (α,β) and for $\alpha+3\beta$ using the dataset **"logistic.data.txt"** available online.