

# Statistics 581/582, Winter Quarter 2008

## Problem Set 14

**Reading:** Ferguson, Sections 16–20, 22–24.

**Problem 51 (estimation of a beta parameter, 4 points).** Suppose that  $X_1, \dots, X_n$  is a sample from a beta distribution with Lebesgue density

$$f(x, \theta) = \theta(\theta + 1)x^{\theta-1}(1 - x)$$

for  $x \in (0, 1)$ . Find a method of moments estimate  $\tilde{\theta}_n$  of the parameter  $\theta > 0$ . Show that the estimate is asymptotically normal, but not asymptotically efficient.

**Problem 52 (EM algorithm for a normal mixture model, 8 points).** Let the sample size  $n$  and the number of mixture components  $K$  be positive integers, and suppose that

$$f_{ik}, \quad i = 1, \dots, n, \quad k = 1, \dots, K,$$

are known real numbers. Let the observation vector  $Y = (Y_1, \dots, Y_n)$  have the density function

$$q(Y, \theta) = \prod_{i=1}^n \sum_{k=1}^K w_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - f_{ik})^2}{2\sigma^2}\right)$$

with respect to the Lebesgue measure on  $(\mathbb{R}^n, \mathcal{B}^n)$ , where  $\theta = (w_1, \dots, w_K, \sigma)$  with the constraints that  $w_1 > 0, \dots, w_K > 0$ ,  $w_1 + \dots + w_K = 1$  and  $\sigma > 0$ . This corresponds to a normal mixture model in which the means are known, while the weights  $w_1, \dots, w_K$  and the common variance  $\sigma^2$  are unknown. Closed form expressions for the maximum likelihood estimates do not exist.

A typical application is the following. Suppose that you are to predict a future quantity  $Y_i$ . You have access to  $K$  distinct sources of expertise, each of which provides a point forecast, resulting in a collection  $f_{i1}, \dots, f_{iK}$  of real-valued forecasts. You now aim to combine the discrete forecast ensemble into a continuous predictive probability density function, which takes the form of a normal mixture density.

- (a) Find the predictive mean  $\mu_i$  and the predictive variance  $\sigma_i^2$ , that is, the mean and the variance of a random variable  $Y_i$  with mixture density

$$q_i(y_i, \theta) = \sum_{k=1}^K w_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f_{ik})^2}{2\sigma^2}\right).$$

- (b) Find an EM algorithm for computing maximum likelihood estimates of the mixture weights  $w_1, \dots, w_k$  and the common variance  $\sigma^2$ .
- (c) I will shortly email a data matrix with  $n = 80$  rows and  $K + 2 = 10$  columns. Each row corresponds to a two days ahead ensemble forecast of the daytime high temperature at SeaTac Airport. The first column gives the initialization time, that is, the day on which the forecast was issued. The forecast horizon is two days, so the valid time is two days later. The initialization times range from 1 February 2007 to 27 April 2007. Days with missing values have been removed. Columns 2 through 9 show the  $K = 8$  ensemble member forecasts  $f_{i1}, \dots, f_{iK}$  from the University of Washington Mesoscale Ensemble (<http://www.atmos.washington.edu/~ens/uwme.cgi>). Column 10 shows the respective verifying observation. The unit used is degrees Kelvin. Feel free to convert to degrees Celsius or degrees Fahrenheit.

Implement the algorithm of part (b) and apply it to this dataset. Report the initial conditions, the convergence criterion, the number of iterations and the final estimate. Discuss the results and any difficulties with the implementation.

- (d) Check the goodness of fit of the normal mixture model. For example, you might apply the probability integral transform technique of Homework Problem 3(b).
- (e) The ensemble forecast initialized on April 29, and valid on May 1, was 289.89, 290.16, 289.49, 289.65, 290.02, 287.84, 290.19 and 290.40 degrees Kelvin, respectively. Find and plot the predictive probability density function for the daytime high temperature at SeaTac Airport on 1 May 2007. Compare to the verifying high temperature, which was 289.26 degrees Kelvin.

Tilmann Gneiting, February 22, 2008. Solutions are due Friday, February 29 at the beginning of the class session.