

## Homework III: Stat 246

Due, Tuesday, April 29

1. Let  $f(x; \theta)$  be a mixture of  $M$  univariate Gaussians, where

$$\theta = (\mu_1, \sigma_1^2, \dots, \mu_M, \sigma_M^2, \pi_1, \dots, \pi_M).$$

- (a) Describe this mixture model as a directed acyclic graphical model.  
(b) For a sample

$$(X_1, Y_1), \dots, (X_N, Y_N),$$

derive the maximum likelihood equations for the parameters  $\theta$ . Can you identify any problems with this estimate.

- (c) Now assume  $Y_n$ 's are not observed. For  $M = 2$ , assume you take the first observation  $X_1$  and define  $\mu_1 = X_1$ , with  $\pi_1 = \frac{1}{N}$ , and define  $\mu_2 = \frac{1}{N-1} \sum_{n=2}^N X_n$  with  $\pi_2 = \frac{N-1}{N}$ . Study the behavior of the log likelihood on the observed  $X_n$ 's as a function of the variances  $\sigma_1^2, \sigma_2^2$ . Show that for any fixed  $\sigma_2$  the likelihood goes to infinity as  $\sigma_1 \rightarrow 0$ . How do you interpret this result?  
(d) For  $M = 2$  assume  $\sigma_1 = \sigma_2$ . Does this resolve the problem of item (1c)?  
(e) For the setting in (1d), write out the MLE equations for the parameters when the  $Y_n$ 's are observed.  
(f) Derive the iterations of an EM algorithm for the situation in (1d) when the  $Y_n$ 's are not observed.  
(g) What other solutions can you propose for the degeneracy in (1c).

2. Derive the details of the EM algorithm for a mixture model with  $M$  components and each component a product of  $d$  independent Bernoulli variables:

$$f(x; \theta) = \sum_{m=1}^M \pi_m f_m(x; \theta_m), \quad f_m(x; \theta_m) = \prod_{\alpha=1}^d p_{\alpha, m}^{x_\alpha} (1 - p_{\alpha, m})^{(1-x_\alpha)},$$

where  $\theta = (\pi_1, \dots, \pi_M, \theta_1, \dots, \theta_M)$  and  $\theta_m = (p_{1, m}, \dots, p_{d, m})$ .

3. An experiment:

- (a) The file 'twos.asc' contains data for 99 handwritten 2's. Each image is  $30 \times 30$ , and at each pixel we set the value to 1 if there is some 'ink' over that pixel. If we order an image row by row this yields a binary vector  $X \in R^d, d = 900 = 30 \times 30$ . Each row in the file contains 900 zeros and ones corresponding to one image. Implement the EM algorithm on this data with  $M = 1, 3, 5$  components.

**Warnings:** 1. You want to avoid estimates of  $p_\alpha$  that are zero or one. What are possible solutions.  
2. Beware of multiplying lots of probabilities. You will quickly get below the rounding error of the computer. Work with sums of logs. When you have a ratio of terms involving probabilities work with the exponent of the differences of the logs.

- (b) After you estimate the different components show the probability vector you obtain as an image, i.e. reshape  $p_{\alpha,m}, \alpha = 1, \dots, 900$  as a  $30 \times 30$  array. What do you see?
- (c) **Extra credit:** Do you think the independence assumptions for each component are valid. Can you show any statistics that verify your claim.