Multivariate Data Analysis. Lab Assignment 1

Xijia Liu

There are two parts in this assignment. First, you will learn how to implement a Gaussian Mixture Model (GMM) in R, then apply it on a real data set. In the second part, you are required to solve some simple theoretical problems from lecture 1 and 2.

Part I: GMM and EM algorithm in R.

Task 1 GMM and estimation with EM algorithm has been implemented in a useful package 'mixtools' in R. Please install this package first.

Task 1.1 'mixtools' provides a function, 'rmvnorm' by which data can be simulated from multivariate normal distribution with arbitrary mean vector and covariance matrix. Please try to learn this function by typing '?rmvnorm', and generate 1000 realizations from a two dimension Gaussian distribution with mean vector $\boldsymbol{\mu} = (2, 3)'$, variance $\sigma_1 = 1$, $\sigma_2 = 4$, and correlation $\rho = 0.7$. Make a scatter plot for your random sample.

Task 1.2 Please simulate 1000 realizations from the following GMM:

- The latent (label) variable z_i belongs to Bernoulli distribution with parameter $p = 0.6$ for $i = 1, ..., 1000$
- The conditional distribution given the value of latent variable:

$$\mathbf{X}_i | z_i = 1 \sim \mathcal{N}_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \text{ and } \mathbf{X}_i | z_i = 0 \sim \mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

  where $\boldsymbol{\mu}_1 = (2, 3)'$ and $\boldsymbol{\mu}_2 = (3, 2)'$. For $\boldsymbol{\Sigma}_1$, the standard deviations are 0.2 and 0.6, the correlation is 0.5. For $\boldsymbol{\Sigma}_2$, the standard deviations are 0.4 and 0.3, the correlation is 0.5

Task 1.3 Please read the help document of function 'mvnormalmixEM', then fit a GMM on your simulated data and compare the estimation results of parameters with the true values.

Task 2: A puzzle. I picked up two species from the entire iris data, and only two variables 'petal length' and 'petal width' are saved in 'BlindingIris.txt'. Please find it from Cambro and read it into R. So your task is to answer me which are two species picked up from the original file? Please show me your answer, code and write down your idea and model. In table 1, you can find the mean vectors of 'petal length' and 'petal width' of each species which are calculated from the original iris data. (Note, find the answer by looking at a scatter plot and comparing with original data is not acceptable.) **Part II**: Theoretical problems

Solve the following theoretical exercises: 2.32, 3.18 and 4.16 in Applied Multivariate Statistical Analysis (AMSA). **Note**: there are (at least) two versions of AMSA, 6th edition, with the same ISBN-number. In the most recent one they have changed the numbering of the chapters: Ch 2 became Ch3 and Ch3 became Ch 2. The numbering of the exercises in this assignment refers to the older version. The easiest way to know what exercises to do is: In the chapter (2 or 3) with 42 exercises, do Number 32. In the chapter (2 or 3) with 20 exercises, do Number 18. The chosen exercise in Ch 4 has the same number in both versions.

Please submit your report no later than 25th September.

species	setosa	versicolor	virginica
petal_length	1.462	4.260	5.552
petal_width	0.246	1.326	2.026

Table 1: Table 1