



KTH Teknikvetenskap

## SF3961 Statistical Inference

2015/16

### Homework 7

Below you will find five projects/exercises on the EM-algorithm. You should hand in solutions to three of them.

#### Part I: Maximum likelihood estimation for mixtures

Let  $f_j(\cdot)$ ,  $j = 1, \dots, k$  be probability densities and suppose the random variable  $Y_j$  is a mixture with weights  $p_1, \dots, p_k$ . That is,  $Y_i$  has density

$$f_{Y_i|\Theta}(y_i | \theta) = \sum_{j=1}^k p_j f_j(y_i),$$

where  $\theta = (p_1, \dots, p_{k-1}) \in [0, 1]^{k-1}$  and  $p_k = 1 - (p_1 + \dots + p_{k-1}) \in [0, 1]$ . (You can visualize sampling from  $Y_i$  by first generating a variable  $W_i$  in  $\{1, \dots, k\}$  with weights  $p_j$ , and if  $W_i = j$  you sample  $Y_i$  from density  $f_j$ ).

Suppose we have observed a random sample of iid random variables,  $Y = (Y_1, \dots, Y_n)$  from the mixture. We want to estimate the parameters  $(p_1, \dots, p_{k-1})$ . You can view this as a missing data problem where the missing data is taken to be  $Z = (Z_1, \dots, Z_n)$  where  $Z_i = (Z_{i1}, \dots, Z_{ik})$  where  $Z_{ij}$  is an indicator being 1 if the  $i$ th sample was from mixture component  $j$ , i.e. from  $f_j(\cdot)$ . We denote the complete data by  $X = (Y, Z) = (Y_1, \dots, Y_n, Z_1, \dots, Z_n)$ .

- Write down the complete loglikelihood (for  $X$ ).
- E-step: Give an explicit expression for the function  $Q(\theta | \theta^{(m)})$  used in the E-step of the EM algorithm.
- M-step: Execute the M-step, i.e. maximize  $Q(\theta | \theta^{(m)})$  w.r.t.  $\theta$  and find an expression for the maximizer.
- Consider the data set `sp500.txt`. This data set consists of 2500 observations of daily returns in percent (actually log-returns) of the S&P 500 stock index, i.e.  $Y_i = 100 \times \log(S_i/S_{i-1})$  where  $S_i$  is the value of the index at closing time on day  $i$ . By plotting the data and making a histogram it may seem plausible that the data generating mechanism is a normal mixture with different variances. That is, suppose  $k = 2$  and  $f_1$  and  $f_2$  are normal densities with parameters  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ . For this exercise take  $\mu_1 = \mu_2 = 0$ ,  $\sigma_1 = 0.5$ , and  $\sigma_2 = 5/3$ . Implement the iterative EM-algorithm for  $k = 2$  to estimate the unknown proportions  $p_1$  and  $p_2$ . Report your chosen convergence criteria (you have to decide how to determine when the algorithm has converged) and the number of iterations needed for convergence.

---

|                              |   |
|------------------------------|---|
| Henrik Hult/Tatjana Pavlenko | E-mail: <a href="mailto:hult@kth.se">hult@kth.se</a> / <a href="mailto:pavlenko@kth.se">pavlenko@kth.se</a> |
| KTH                          | Website:  |
| Department of Mathematics    | Office hours:   |
| Lindstedtsv 25               | Phone: 8-790 6911/8466  |
| Stockholm 100 44             |   |

## Part II: Mixtures with unknown parameters

Suppose again that the densities  $f_i$  are normal ( $\mu_i, \sigma_i^2$ ) but the parameters are unknown.

Then the density for the data  $Y_i$  is given by

$$f_{Y_i|\Theta}(y_i | \theta) = \sum_{j=1}^k p_j f_j(y_i | \theta_j).$$

Suppose we have observed a random sample of iid random variables,  $Y = (Y_1, \dots, Y_n)$  from the mixture. The unknown parameters are then  $\theta = (\mu_1, \sigma_1^2, \dots, \mu_k, \sigma_k^2, p_1, \dots, p_{k-1})$  and your objective is to estimate these parameters. You can view this as a missing data problem where the missing data is taken to be  $Z = (Z_1, \dots, Z_n)$  where  $Z_i = (Z_{i1}, \dots, Z_{ik})$  where  $Z_{ij}$  is an indicator being 1 if the  $i$ th sample was from mixture component  $j$ , i.e. from  $f_{Y_j|\Theta}(\cdot | \theta_j)$ . We denote the complete data by  $X = (Y, Z) = (Y_1, \dots, Y_n, Z_1, \dots, Z_n)$ .

- (a) Write down the complete loglikelihood (for  $X$ ).
- (b) E-step: Give an explicit expression for the function  $Q(\theta | \theta^{(m)})$  used in the E-step of the EM algorithm.
- (c) M-step: Execute the M-step, i.e. maximize  $Q(\theta | \theta^{(m)})$  w.r.t.  $\theta$  and find an expression for the maximizer.
- (d) Consider the data set `sp500.txt`. This data set consists of 2500 observations of daily returns in percent (actually log-returns) of the S&P 500 stock index, i.e.  $Y_i = 100 \times \log(S_i/S_{i-1})$  where  $S_i$  is the value of the index at closing time on day  $i$ . By plotting the data and making a histogram it may seem plausible that the data generating mechanism is a normal mixture with different variances and possibly different means. That is, suppose  $k = 2$  and  $f_1$  and  $f_2$  are normal densities with parameters  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ . Implement the iterative EM-algorithm for  $k = 2$  to estimate the unknown proportions  $p_1$  and  $p_2$  as well as the unknown parameters  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ . Report your chosen convergence criteria (you have to decide how to determine when the algorithm has converged) and the number of iterations needed for convergence.

## Part III: EM algorithm for hidden Markov models

If you look closely at a serial plot of the data set (`sp500.txt`) you see that the series appear to have a serial dependence, there are periods of large values and periods of smaller values. One can try to capture this with a hidden Markov model. Here is how it works. Instead of assuming  $Y_1, \dots, Y_n$  are iid from the mixture distribution we assume that there is an unobserved Markov chain ( $Z_i$ ) that determines if the variable  $Y_i$  is generated from a normal with parameters  $\mu_1, \sigma_1^2$  or  $\mu_2, \sigma_2^2$ . Let the state space of the Markov chain ( $Z_i$ ) be  $\{1, 2\}$  and suppose the transition probabilities are  $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$ . That is,

$$P(Z_{i+1} = b | Z_i = a) = \pi_{ab}, \quad a, b \in \{1, 2\}.$$

Then,

$$P(Y_i = y | Z_i = a) = \frac{1}{\sqrt{2\pi\sigma_a^2}} \exp\left\{-\frac{1}{2\sigma_a^2}(y - \mu_a)^2\right\}.$$

---

|                              |   |
|------------------------------|---|
| Henrik Hult/Tatjana Pavlenko | E-mail: <a href="mailto:hult@kth.se">hult@kth.se</a> / <a href="mailto:pavlenko@kth.se">pavlenko@kth.se</a> |
| KTH                          | Website:  |
| Department of Mathematics    | Office hours:   |
| Lindstedtsv 25               | Phone: 8-790 6911/8466  |
| Stockholm 100 44             |   |

The parameters to be estimated are  $\Theta = (\pi_{11}, \pi_{22}, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ .

(a) Write down the complete loglikelihood (for  $X$ ).

(b) E-step: Give an explicit expression for the function  $Q(\theta | \theta^{(m)})$  used in the E-step of the EM algorithm.

(c) M-step: Execute the M-step, i.e. maximize  $Q(\theta | \theta^{(m)})$  w.r.t.  $\theta$  and find an expression for the maximizer.

(d) Consider the data set `sp500.txt`. Implement the iterative EM-algorithm for  $k = 2$  to estimate the unknown Markov transition probabilities  $\pi_{11}, \pi_{22}$  as well as the unknown parameters  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ .

#### Part IV: K-means algorithm

Read the text on pp. 284-290 in David Mackay's book "Information theory, Inference, and Learning Algorithms" available at <http://www.inference.phy.cam.ac.uk/itprnn/book.pdf>. Consider the soft K-means algorithm (Algorithm 20.7) on p. 289. Can this algorithm be viewed as a special case of the EM algorithm for fitting the parameters of a mixture model? If so, explain how!

Based on your experience in part I and II, suggest an improved soft K-means algorithm that can handle the "dissimilar" data set in Figure 20.5 (p. 288). If possible, suggest how you could construct an improved soft K-means algorithm to handle the data set in Figure 20.6 as well.

**Part V: Truncated Poisson**  $X$  is a discrete random variable with the probability mass function  $p_X(x)$ . Let us recall that the conditional probability of  $X$  given the event  $B$ , we write this as  $X | B$ , is denoted by  $p_{X|B}(x | B)$  and given by

$$p_{X|B}(x | B) = \begin{cases} \frac{p_X(x)}{p_X(B)} & \text{if } x \in B \\ 0 & \text{otherwise.} \end{cases}$$

If  $X \in \text{Po}(\lambda)$ , then we say that the random variable  $Y$  is *zero-truncated Poisson* random variable, or  $Y \in \text{ztrPo}(\lambda)$ , if

$$Y = X | X > 0.$$

We shall regard  $Y \in \text{ztrPo}(\lambda)$  as an incomplete observation of  $X \in \text{Po}(\lambda)$  in the sense of Sundberg (1974), since  $\text{Po}(\lambda)$  is an exponential family. The purpose of the exercise is to find the EM-algorithm for  $\lambda$  using zero-truncated samples.

1. a) Suppose that we have obtained  $n$  independent outcomes  $y_1, \dots, y_n$  of zero-truncated Poisson random variables  $Y_1, \dots, Y_n \in \text{ztrPo}(\lambda)$ , respectively. We set

$$\mathbf{y} = (y_1, \dots, y_n).$$

Show that the maximum likelihood estimate  $\hat{\lambda}_{\text{ml}}^{\text{tr}}$  of  $\lambda$  based on  $\mathbf{y}$  satisfies the equation

$$\hat{\lambda}_{\text{ml}}^{\text{tr}} = \bar{y}_{\text{obs}} \left( 1 - e^{-\hat{\lambda}_{\text{ml}}^{\text{tr}}} \right), \quad (1)$$

where

$$\bar{y}_{\text{obs}} = \frac{1}{n} \sum_{i=1}^n y_i. \quad (2)$$

- b) We regard  $Y_i \in \text{ztrPo}(\lambda)$ ,  $i = 1, 2, \dots, n$  as incomplete observations of the complete data  $X_i \in \text{Po}(\lambda)$ ,  $i = 1, \dots, n + m$ ,  $m \geq 0$ , in the sense that we do not know  $m$ , the number of zeros in the complete sample. Check that if  $x_1, \dots, x_{n+m}$  is a sample of independent Poisson random variables  $X_1, \dots, X_{n+m} \in \text{Po}(\lambda)$ , and  $\mathbf{y} = (y_1, \dots, y_n)$  is the corresponding truncated sample, then the (complete data) maximum likelihood estimate  $\hat{\lambda}_{\text{ml}}$  is given by

$$\hat{\lambda}_{\text{ml}} = \frac{n}{n + m} \bar{y}_{\text{obs}}. \quad (3)$$

The difficulty is thus that we only have obtained the truncated sample  $y_1, \dots, y_n$  and do not know  $m$ , the number of zeros in the sample  $x_1, \dots, x_{n+m}$ . We shall now develop an EM-algorithm to estimate  $\lambda$  in this situation.

2. a) **E-step**

Let  $\lambda_p$  be your current estimate. We need to evaluate

$$Q(\lambda | \lambda_p) = E[\ln f(X_1, \dots, X_{n+m} | \lambda) | \mathbf{y}, \lambda_p],$$

where  $f(X_1, \dots, X_{n+m} | \lambda)$  is the complete data joint density (w.r.t. the counting measure). A special issue arises in that we need to choose the conditional distribution  $P(m | \mathbf{y}, \lambda)$ . Give a negative binomial distribution (why?) suitable for this purpose and show that then  $Q(\lambda | \lambda_p)$  contains the factor

$$m_{p+1} = E[m | \mathbf{y}, \lambda_p] = \frac{e^{-\lambda_p}}{1 - e^{-\lambda_p}} \cdot n. \quad (4)$$

b) **M-step**

Show now that the M-step yields nothing else but a straightforward iterative solution to (1). Exploit and explain the connection of the M-step to (3).

c) Show that this EM-algorithm will for any initial value  $\lambda_0 > 0$  converge to the solution of (1).

The part 2.c) does not require any knowledge of statistical inference or of the lecture on EM, but is an exercise in proving computational convergence of an iteration for finding a fixed point of a transformation, see Section 10.2 in Luenberger (1969). (Optional: Find the speed of convergence.)

3. Let us quote from Blom et al. (2005, ch.13, p. 344):

I en klassisk datamängd undersöktes antalet ihjälsparkade soldater vid 14 tyska armékårer från 1875 till 1894 (20 år). De  $14 \cdot 20 = 280$  rapporterna fördelade sig som i tabellen.

| Antal döda | Antal rapporter | Andel  |
|------------|-----------------|--------|
| 0          | 144             | 0.5143 |
| 1          | 91              | 0.3250 |
| 2          | 32              | 0.1143 |
| 3          | 11              | 0.0393 |
| 4          | 2               | 0.0071 |
| $\geq 5$   | 0               | 0      |
| Summa      | 280             | 1      |

Zero-truncate this set of data, use  $\text{zerotrPo}(\lambda)$  as model, and estimate both  $\lambda$  and the number of zero observations using the EM-algorithm.

## References

Blom G., Englund G. et.al. *Sannolikhetsteori och statistikteori med tillämpningar*. Studentlitteratur, 2005.

Sundberg R.: Maximum Likelihood Theory for Incomplete Data from an Exponential Family *Scandinavian Journal of Statistics*, 1, pp. 49–58, 1974.

Luenberger D.G. : *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969.