

SKETCH of the SOLUTIONS

STK4051 - spring 2025

Exercise 1 EM Algorithm

- a As the widow can have 0 children both if it belongs to Group A and if belong to Group B and has 0 children,

$$Pr[Y_c = 0] = p + (1 - p)e^{-\mu}$$

where $e^{-\lambda} = Pr(X = 0)$ for $X \sim \text{Poisson}(\mu)$.

In all other cases, we just need to multiply the probabilities for the widow to be in Group B and to have c children,

$$Pr[Y_c = c] = (1 - p) \frac{\mu^c e^{-\mu}}{c!}.$$

We can recognise that Y_c is distributed as a Multinomial distribution, so that

$$L(p, \mu) = \frac{(\sum_{i=0}^6 y_i)!}{y_0! \times \dots \times y_6!} \times [p + (1 - p)e^{-\mu}]_0^{y_0} \prod_{c=1}^6 \left[(1 - p) \frac{\mu^c e^{-\mu}}{c!} \right]^{y_c}.$$

Taking the logarithm and calling the multinomial constant “const.”, we obtain

$$\ell(p, \mu) = \text{const} + y_0 \log[p + (1 - p)e^{-\mu}] + \sum_{c=1}^6 y_c [\log(1 - p) + c \log \mu - \mu - \log c!].$$

- b If we knew how many of the y_0 widows without children came from Group A (x_A) and from Group B (x_B), we would have the complete log-likelihood

$$\ell(p, \mu) = x_A \log p + x_B [\log(1 - p) - \mu] + \sum_{c=1}^6 y_c [\log(1 - p) + c \log \mu - \mu - \log c!]$$

- c $Q(p, \mu | p^{(k-1)}, \mu^{(k-1)})$ is the expectation of the joint log-likelihood based on the complete data, conditional on the observed data $Y_c = y_c$. So

$$\begin{aligned} Q(p, \mu | p^{(k-1)}, \mu^{(k-1)}) &= E[x_A \log p + x_B [\log(1 - p) - \mu] + \\ &\quad + \sum_{c=1}^6 y_c [\log(1 - p) + c \log \mu - \mu - \log c!]] \\ &= E[x_A] \log p + E[x_B] [\log(1 - p) - \mu] + \\ &\quad + \sum_{c=1}^6 y_c [\log(1 - p) + c \log \mu - \mu - \log c!] \end{aligned}$$

By noticing that the y_0 widows with 0 children can either be from group A or group B with probability proportional to p and $(1-p)e^{-\mu}$, namely $\frac{p}{p+(1-p)e^{-\mu}}$ and $\frac{(1-p)e^{-\mu}}{p+(1-p)e^{-\mu}}$, we can derive the terms

$$E[x_A|y_0, p, \mu] = y_0 \frac{p}{p + (1-p)e^{-\mu}}$$

$$E[x_B|y_0, p, \mu] = y_0 \frac{(1-p)e^{-\mu}}{p + (1-p)e^{-\mu}}$$

we get

$$\begin{aligned} Q(p, \mu | p^{(k-1)}, \mu^{(k-1)}) = & y_0 \frac{p^{(k-1)}}{p^{(k-1)} + (1-p^{(k-1)})e^{-\mu^{(k-1)}}} \log p + \\ & + y_0 \frac{(1-p^{(k-1)})e^{-\mu^{(k-1)}}}{p^{(k-1)} + (1-p^{(k-1)})e^{-\mu^{(k-1)}}} [\log(1-p) - \mu] + \\ & + \sum_{c=1}^6 y_c [\log(1-p) + c \log \mu - \mu - \log c!] \end{aligned}$$

d Call $\hat{x}_A^{(k-1)} = y_0 \frac{p^{(k-1)}}{p^{(k-1)} + (1-p^{(k-1)})e^{-\mu^{(k-1)}}}$, then

$$\begin{aligned} \frac{\partial Q(p, \mu | p^{(k-1)}, \mu^{(k-1)})}{\partial p} &= \frac{\hat{x}_A^{(k-1)}}{p} - \frac{y_0 - \hat{x}_A^{(k-1)}}{1-p} - \frac{\sum_{c=1}^6 y_c}{1-p} \\ &= \frac{\hat{x}_A^{(k-1)} - p\hat{x}_A^{(k-1)} - py_0 + p\hat{x}_A^{(k-1)} - p\sum_{c=1}^6 y_c}{p(1-p)} \\ &= \frac{\hat{x}_A^{(k-1)} - py_0 - p\sum_{c=1}^6 y_c}{p(1-p)}. \end{aligned}$$

Setting it equal to 0, we obtain

$$p^{(k)} = \frac{\hat{x}_A^{(k-1)}}{y_0 + \sum_{c=1}^6 y_c} = \frac{\hat{x}_A^{(k-1)}}{\sum_{c=0}^6 y_c}.$$

Similarly, defining $\hat{x}_B^{(k-1)} = y_0 \frac{(1-p^{(k-1)})e^{-\mu^{(k-1)}}}{p^{(k-1)} + (1-p^{(k-1)})e^{-\mu^{(k-1)}}}$, we obtain

$$\frac{\partial Q(p, \mu | p^{(k-1)}, \mu^{(k-1)})}{\partial \mu} = -\hat{x}_B^{(k-1)} + \frac{\sum_{c=1}^6 cy_c}{\mu} - \sum_{c=1}^6 y_c,$$

that leads to

$$\mu^{(k)} = \frac{\hat{x}_B^{(k-1)} + \sum_{c=1}^6 y_c}{\sum_{c=0}^6 cy_c}.$$

The first three iterations when starting from (0.75, 0.40) are

iter.	p	μ
0	0.75000	0.40000
1	0.61418	1.03548
2	0.61438	1.03601
3	0.61453	1.03643

e For each bullet point:

- The simplest way to construct a bootstrap confidence interval is to look at the percentiles of the histograms. In particular, if I want a 95% confidence interval, I need to look at the 0.025 and 0.975 quantiles. In numbers, it is something like $CI_{0.95}(p) = (0.585, 0.645)$ and $CI_{0.95}(\mu) = (0.95, 1.11)$;
- I could have applied a more precise approach like the accelerated bias-corrected percentile method (BC_a) or bootstrap t. For example, with the former I would have obtained different percentiles, in theory with a better coverage.
- An example of bootstrap sample:

c	0	1	2	3	4	5	6
y_c	3059	586	286	104	34	1	5

Exercise 2 Bootstrap

a The procedure is wrong because it assumes that the $Y_i|x_i$ are i.i.d., which is clearly false because they have different conditional means. Among the possible alternatives, in class we mentioned:

- bootstrap the residuals, i.e., apply a parametric bootstrap approach, which can be summarised as:
 - * fit the regression model;
 - * obtain the predictions $\hat{y}_i$ and the residuals $\hat{\epsilon}_i$;
 - * sample with replacement the residuals, obtaining $\hat{\epsilon}_i^*$;
 - * create the pseudo-responses $y_i^* = \hat{y}_i + \hat{\epsilon}_i^*$;
 - * regress Y^* on x and obtain $\hat{\beta}^*$;
 - * repeat many times and perform inference on β based on the empirical distribution of $\hat{\beta}^*$.

For all quantities, here $i = 1, \dots, n$.

- paired bootstrap, i.e., perform a bootstrap procedure on the data pairs $z_i = (x_i, y_i), i = 1, \dots, n$:

- * sample $z_1^*, \dots, z_n^*$ with replacement from the set of observed data pairs $z_1, \dots, z_n$;
- * fit the regression model on $z_1^*, \dots, z_n^*$ and obtain $\hat{\beta}^*$;
- * repeat many times and perform inference on β based on the empirical distribution of $\hat{\beta}^*$.

- b In this case bootstrap fails because the MLE is based on the highest value and therefore we cannot find a bootstrap sample around θ . We only get values smaller than it.

Alternatively, we can base our bootstrap procedure on the estimator based on the method of the moments. As it is based on a transformation of the sample mean, it allows computing values larger than θ and therefore we can compute a reasonable distribution of $\hat{\theta}$.

- c The three points:

- In the case of dependent data, the i.i.d. assumption is not met. While sampling with replacement, the dependence of an observation from its previous one is removed and therefore it is not present in the bootstrap sample;
- to preserve the dependency, the observation are sampled in blocks. In particular, in the case of moving block bootstrapping, the blocks can overlaps. The main advantage is the larger possible sample space, as more blocks are possible. Consider a 6-dimensional sample, and a block size of 3. With the moving block bootstrap we can sample from 4 blocks, while from the non-moving blocks only 2.
The main disadvantage is that we may have biased bootstrap estimator, for example that for the mean. The reason is that observations on the edges are under-represented (edge effect), as they are present in less samples than other observations.
- The performance of a block bootstrap technique depends on l , that must be chosen carefully: if its value is too small, the structure of interest, e.g. an autocorrelation structure, can be lost; if its value is too large, instead, there will be few blocks from which to sample and the estimator can be not efficient. Therefore, the optimal length should be chosen large enough to capture the structure of interest and small enough to have enough variability in the samples. It is clear that the optimal value depends on the sample size: approximatively, l should be of the order $n^{1/3}$.

Exercise 3 MCMC

a The three points:

- The plot should be wiggling enough, covering all support of the distribution, and move pretty fast from the starting point in the case this is far from the mass of the density;
- while the plot has many good characteristics, it does not provide a sample that can be used as representative of the distribution above. The reason is that it does not cover the whole support, in particular there are no points from the second Gaussian distribution. This example shows the importance of starting from different starting points of the support to detect possible anomalies. In this case, the problem would have been solved by increasing the variability of the proposal distribution.
- the value is satisfactory, as it should be (accordingly to the text book) between 0.25 and 0.4. On the one hand, we should avoid a procedure with too high acceptance rate to avoid very correlated samples or chains that do not fully explore the support of the distribution, especially in regions of lower density. On the other hand, we would like to also avoid too many rejection, otherwise the chain would have too many replications of the same point, maybe just because it is not able to reach a place with enough density.

b The two points:

- The idea behind the R statistic comes from an ANOVA-kind of argument: the between-chain and the within-chain variances should be comparable. If the former is much larger, the chain does not provide samples from the target distribution. The reason is that they do not have reached the stationary distribution or that the influence of the starting point is still too large after the burn-in removal. Therefore, it is used as a monitoring tool for checking the MCMC convergence: if its value is too large, more iterations and/or an increased burn-in period should be considered.
- If we were able to start the procedure from suitable points (i.e., also points from where the chain would have converged around 16) we would have get a large value, while if we would have started only from points not larger enough than 2, we may have even obtained a reasonable number, close to 1.