Jared Yu

Computational Statistics

Problem Set 3

NOTE: All code for each of the problems is in the Code Appendix at the end of the document.

1. Problem 2.5. Only do parts (a), (b), (c), and (e).
    a. Let $N_i \mid b_{i1}, b_{i2} \sim Poisson(\lambda_i);\ \lambda_i = \alpha_1 b_{i1} + \alpha_2 b_{i2}, i = 1, \cdots, n_T$

The density function for the above Poisson distribution is as follows:

$$f(n_i) = \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!} = \frac{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^{n_i} e^{-(\alpha_1 b_{i1} + \alpha_2 b_{i2})}}{n_i!}$$

The likelihood for the above p.d.f. is as follows:

$$L(\alpha_1, \alpha_2) = f\left(n_1, \cdots, n_{n_T}; \alpha_1, \alpha_2\right) = \prod_{i=1}^{n_T} \frac{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^{n_i}}{n_i!} e^{-\alpha_1 b_{i1} - \alpha_2 b_{i2}}$$

$$= \frac{\prod_{i=1}^{n_T} (\alpha_1 b_{i1} + \alpha_2 b_{i2})^{n_i}}{\prod_{i=1}^{n_T} n_i!} e^{\sum_{i=1}^{n_T} -\alpha_1 b_{i1} - \alpha_2 b_{i2}}$$

The log-likelihood of the p.d.f. is as follows:

$$l(\alpha_1, \alpha_2) = \ln L(\alpha_1, \alpha_2) = \sum_{i=1}^{n_T} n_i \ln(\alpha_1 b_{i1} + \alpha_2 b_{i2}) + \sum_{i=1}^{n_T} (-\alpha_1 b_{i1} - \alpha_2 b_{i2}) - \sum_{i=1}^{n_T} \ln n_i!$$

The following are the first and second derivatives of the log-likelihood:

Let $\theta = (\alpha_1, \alpha_2)^\top$

$$\frac{\partial l(\theta)}{\partial \alpha_1} = \sum_{i=1}^{n_T} \frac{n_i b_{i1}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} - \sum_{i=1}^{n_T} b_{i1}$$

$$\frac{\partial l(\theta)}{\partial \alpha_2} = \sum_{i=1}^{n_T} \frac{n_i b_{i2}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} - \sum_{i=1}^{n_T} b_{i2}$$

$$\frac{\partial l(\theta)}{\partial \theta} = \left(\frac{\partial l(\theta)}{\partial \alpha_1}, \frac{\partial l(\theta)}{\partial \alpha_2}\right)^\top = \left(\sum_{i=1}^{n_T} \frac{n_i b_{i1}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} - \sum_{i=1}^{n_T} b_{i1}, \sum_{i=1}^{n_T} \frac{n_i b_{i2}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} - \sum_{i=1}^{n_T} b_{i2}\right)^\top$$

$$\frac{\partial^2 l(\theta)}{\partial \alpha_1^2} = -\sum_{i=1}^{n_T} \frac{n_i b_{i1}^2}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2}$$

$$\frac{\partial^2 l(\theta)}{\partial \alpha_2^2} = -\sum_{i=1}^{n_T} \frac{n_i b_{i2}^2}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2}$$

$$\frac{\partial^2 l(\theta)}{\partial \alpha_1 \partial \alpha_2} = -\sum_{i=1}^{n_T} \frac{n_i b_{i1} b_{i2}}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2}$$

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} = \begin{bmatrix} -\sum_{i=1}^{n_T} \frac{n_i b_{i1}^2}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2} & -\sum_{i=1}^{n_T} \frac{n_i b_{i1} b_{i2}}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2} \\ -\sum_{i=1}^{n_T} \frac{n_i b_{i1} b_{i2}}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2} & -\sum_{i=1}^{n_T} \frac{n_i b_{i2}^2}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2} \end{bmatrix}$$

The Newton-Raphson update can be shown as follows:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{g}^{"}\left(\boldsymbol{\theta}^{(t)}\right)^{-1} \boldsymbol{g}'\left(\boldsymbol{\theta}^{(t)}\right)$$

where $\boldsymbol{\theta}^{(t)}$ is the $t'th$ iteration of the vector of parameters containing $\alpha_1^{(t)}$ and $\alpha_2^{(t)}$, $\boldsymbol{g}^{"}(\cdot)^{-1}$ is the inverse of the 2×2 Hessian matrix denoted above as $\frac{\partial^2 l(\theta)}{\partial \theta^2}$ and $\boldsymbol{g}'(\cdot)$ is the gradient of the log-likelihood denoted above as $\frac{\partial l(\theta)}{\partial \theta}$. Therefore, the above update for finding the MLE of this Poisson distribution can be expressed as:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{g}^{"}\left(\boldsymbol{\theta}^{(t)}\right)^{-1} \boldsymbol{g}'\left(\boldsymbol{\theta}^{(t)}\right)$$

$$= \boldsymbol{\theta}^{(t)} - \begin{bmatrix} -\sum_{i=1}^{n_T} \frac{n_i b_{i1}^2}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2} & -\sum_{i=1}^{n_T} \frac{n_i b_{i1} b_{i2}}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2} \\ -\sum_{i=1}^{n_T} \frac{n_i b_{i1} b_{i2}}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2} & -\sum_{i=1}^{n_T} \frac{n_i b_{i2}^2}{(\alpha_1 b_{i1} + \alpha_2 b_{i2})^2} \end{bmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n_T} \frac{n_i b_{i1}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} - \sum_{i=1}^{n_T} b_{i1} \\ \sum_{i=1}^{n_T} \frac{n_i b_{i2}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} - \sum_{i=1}^{n_T} b_{i2} \end{pmatrix}$$

b. The method for finding the Fisher scoring update will be shown below.

The following is the Fisher Information matrix that will be used in the Fisher scoring update. The Fisher Information matrix is denoted $I(\theta)$.

$$I(\theta) = -E\left[\frac{\partial^2 l(\theta)}{\partial \theta^2}\right] = \begin{bmatrix} \sum_{i=1}^{n_T} \frac{E(n_i) b_{i1}^2}{(\lambda_i)^2} & \sum_{i=1}^{n_T} \frac{E(n_i) b_{i1} b_{i2}}{(\lambda_i)^2} \\ \sum_{i=1}^{n_T} \frac{E(n_i) b_{i1} b_{i2}}{(\lambda_i)^2} & \sum_{i=1}^{n_T} \frac{E(n_i) b_{i2}^2}{(\lambda_i)^2} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n_T} \frac{b_{i1}^2}{\lambda_i} & \sum_{i=1}^{n_T} \frac{b_{i1}b_{i2}}{\lambda_i} \\ \sum_{i=1}^{n_T} \frac{b_{i1}b_{i2}}{\lambda_i} & \sum_{i=1}^{n_T} \frac{b_{i2}^2}{\lambda_i} \end{bmatrix}$$

Then the Fisher scoring update can be written as:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathbf{I}\left(\boldsymbol{\theta}^{(t)}\right)^{-1}\mathbf{l}'\left(\boldsymbol{\theta}^{(t)}\right),$$

where $\boldsymbol{\theta}^{(t)}$ is the $t'th$ iteration of the vector of parameters containing $\alpha_1^{(t)}$ and $\alpha_2^{(t)}$, $\mathbf{I}(\cdot)^{-1}$ is the inverse of the Fisher Information matrix written above as $I(\theta)$, and $\mathbf{l}'(\cdot)$ is the gradient of the log-likelihood function written above as $\frac{\partial l(\theta)}{\partial \theta}$. Therefore, the above update for finding the MLE through using the Fisher scoring update can be expressed as:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathbf{I}\left(\boldsymbol{\theta}^{(t)}\right)^{-1}\mathbf{l}'\left(\boldsymbol{\theta}^{(t)}\right)$$

$$= \boldsymbol{\theta}^{(t)} + \begin{bmatrix} \sum_{i=1}^{n_T} \frac{b_{i1}^2}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} & \sum_{i=1}^{n_T} \frac{b_{i1}b_{i2}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} \\ \sum_{i=1}^{n_T} \frac{b_{i1}b_{i2}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} & \sum_{i=1}^{n_T} \frac{b_{i2}^2}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} \end{bmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n_T} \frac{n_i b_{i1}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} - \sum_{i=1}^{n_T} b_{i1} \\ \sum_{i=1}^{n_T} \frac{n_i b_{i2}}{\alpha_1 b_{i1} + \alpha_2 b_{i2}} - \sum_{i=1}^{n_T} b_{i2} \end{pmatrix}$$

c. Implementing the two algorithms was not difficult since the Newton-Raphson algorithm was recently completed in the univariate case. The task begins with calculating the math by hand to first figure out the log-likelihood of the function along with its related gradient and Hessian matrix. After the math has been derived, the functions are programmed as separate pieces so that they can be looped through efficiently in RStudio.

   The process for both with and without Fisher scoring is similar in that they require some initialization points, a while-loop performs the steps of re-calculating the gradient and Hessian, and a flag that checks for relative convergence. The initialization points are all combinations of values from -1 to 3. These are based upon the visual analysis of the graph of the log-likelihood shown below (Figures 1 and 2):

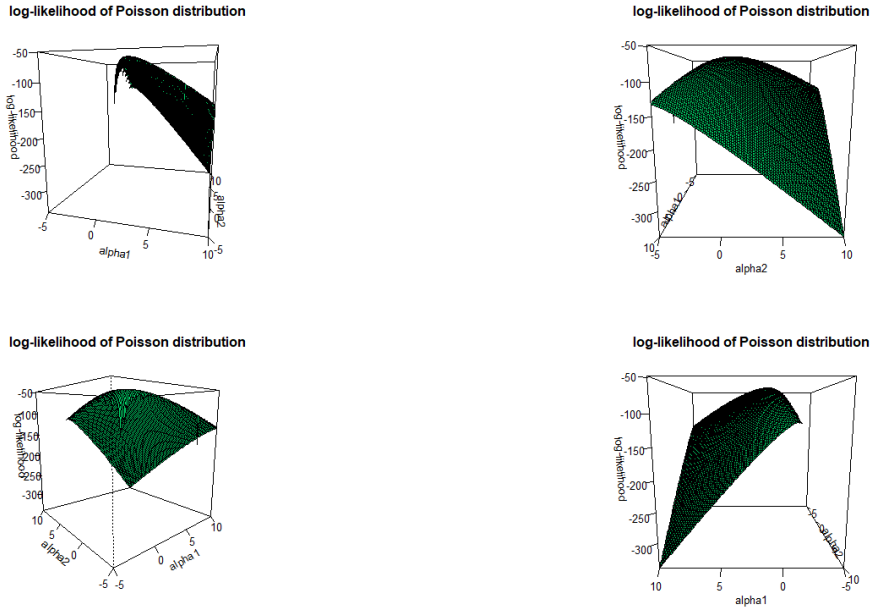

Figure 1 The above plot shows different angles of the log-likelihood of the Poisson distribution. By analyzing the hyperplane of the log-likelihood function, it can be estimated that the peak is roughly somewhere between -1 and 3.

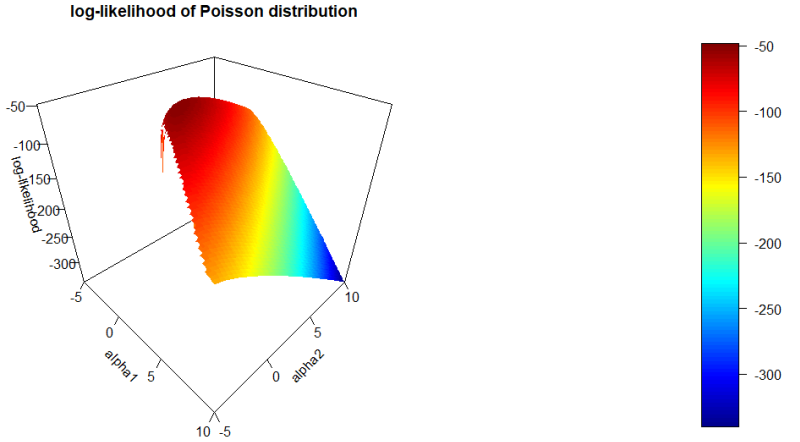

Figure 2 The above plot uses a heatmap of the data on the surface of the hyperplane to illustrate with color that the peak is in dark red. This indicates that the log-likelihood's max value is close to -50.

Both algorithms are run through all combinations of $-1, \cdots, 3$ and are represented in the Table 1 and 2 below. The columns on the far left, α_1 and α_2 represent the initial starting values of the algorithm in a certain trial. The columns $\hat{\alpha}_1$ and $\hat{\alpha}_2$ represent the approximated MLE's after utilizing the corresponding algorithm. The last column $Iterations$ represents the number of iterations required for an algorithm to converge.

Values of NA indicate that the algorithm failed to converge. One of the reasons is that there is a singularity error when trying to take the inverse of the matrix for the

update step. This was common for the Newton-Raphson algorithm. Another error is when dividing by 0 occurs somewhere leading to a failure to the algorithm to converge. The Fisher Scoring was more stable and less likely to output NA due to some calculation error. This is expected, since the Fisher Information matrix is a more efficient calculation compared to the Hessian matrix. Also, in terms of the stability, both methods appear to converge to the same value of $\hat{\alpha}_1 \approx 1.0972$ and $\hat{\alpha}_2 \approx 0.9376$.

There is a noticeable difference in speed when comparing the number of iterations required for convergence. The Newton-Raphson method would usually converge in under 10 iterations, while the Fisher Scoring method required at least 10 iterations to converge. This makes sense, since with optimization problems trade-offs are an expected part of any change in methodology. If the more stable Fisher Scoring method is used, then speed is the trade-off.

Table 1: Results for Newton-Raphson

α_1	α_2	$\hat{\alpha_1}$	$\hat{\alpha_2}$	Iterations
-1	-1	NA	NA	NA
0	-1	NA	NA	NA
1	-1	NA	NA	NA
2	-1	1.097152	0.9375546	8
3	-1	1.097152	0.9375546	8
-1	0	NA	NA	NA
0	0	NA	NA	NA
1	0	1.097152	0.9375546	6
2	0	1.097152	0.9375546	5
3	0	-5.112828	10.3490623	8
-1	1	NA	NA	NA
0	1	1.097152	0.9375546	8
1	1	1.097152	0.9375546	4
2	1	1.097152	0.9375546	7
3	1	NA	NA	NA
-1	2	-5.112828	10.3490623	8
0	2	1.097152	0.9375546	7
1	2	1.097152	0.9375546	6
2	2	3.004737	-1.9534767	10
3	2	NA	NA	NA
-1	3	-3.631235	8.1036416	7
0	3	1.097152	0.9375546	6
1	3	4.836828	-4.7300943	7
2	3	NA	NA	NA
3	3	NA	NA	NA

Table 2: Results for Newton-Raphson with Fisher Scoring

α_1	α_2	$\hat{\alpha_1}$	$\hat{\alpha_2}$	Iterations
-1	-1	1.097152	0.9375556	10
0	-1	1.097151	0.9375571	12
1	-1	NA	NA	NA
2	-1	1.097153	0.9375533	12
3	-1	1.097153	0.9375535	12
-1	0	1.097151	0.9375569	11
0	0	NA	NA	NA
1	0	1.097151	0.9375569	11
2	0	1.097151	0.9375569	11
3	0	1.097151	0.9375569	11
-1	1	NA	NA	NA
0	1	1.097151	0.9375571	12
1	1	1.097152	0.9375556	10
2	1	1.097154	0.9375519	10
3	1	1.097152	0.9375559	11
-1	2	1.097153	0.9375535	14
0	2	1.097151	0.9375571	12
1	2	1.097154	0.9375527	11
2	2	1.097152	0.9375556	10
3	2	1.097153	0.9375532	10
-1	3	1.097152	0.9375556	14
0	3	1.097151	0.9375571	12
1	3	1.097152	0.9375555	12
2	3	1.097153	0.9375534	11
3	3	1.097152	0.9375556	10

e. The steepest ascent method was also implemented for the same dataset. The technique to modify the learning rate α was to use step-halving, where if the approximated value of the log-likelihood function at the next iteration is not greater than or equal to the current iteration then $\alpha = 1$ will be halved until this happens. The reason that greater than or equal to is used is that otherwise the value of α will reduce to essentially 0 and the algorithm will fail to converge.

However, according to the lecture notes and the textbook, it seems to be implied that it should be strictly greater than. A possible explanation is that R is not capable of calculating and storing values to the tiny sizes required for the algorithm to function properly. Therefore, a greater than or equal to sign is required in the check for backtracking. This also occurs when Fisher scoring is used in place of the identity matrix in the update step.

To further clarify, the textbook mentions that in the Newton-Like methods such as steepest ascent, the Hessian matrix is replaced with another matrix that is simpler to calculate. A possibility is the identity matrix, and at the end of p.39 it mentions that, "backtracking with Fisher scoring would avoid stepping downhill." In an email conversation with the professor, it was made clear that the intention for this assignment is to utilize the identity matrix to replace the Hessian.

In the discussion problems for module 3, where the steepest ascent algorithm was used on the Normal distribution the results were ideal. The approximated MLE's in that case converged to the ideal values and the high number of iterations and backtracking steps coincides with what is understood to be the trade-off when implementing steepest ascent.

For this problem, steepest ascent was used in two different situations, the first will show the identity matrix as the substitute for the Hessian while the second will show the Fisher Information matrix as the substitute. Below are the two tables (Tables 3 and 4). From the tables it's evident that the identity matrix method fails to converge properly, while using Fisher scoring allows for convergence. The reason that convergence fails has been associated with the following calculation:

$$\log \lambda_i = \log \alpha_1 b_{i1} + \alpha_2 b_{i2}$$

The effect is that λ_i becomes negative, leading the log transformation evaluating to $-\infty$. However, the Fisher Information matrix prevents this from occurring as often.

Table 3: Results for Steepest Ascent with Identity Matrix

α_1	α_2	$\hat{\alpha_1}$	$\hat{\alpha_2}$	Iterations	Backtracks
-1	-1	NA	NA	NA	NA
0	-1	NA	NA	NA	NA
1	-1	NA	NA	NA	NA
2	-1	NA	NA	NA	NA
3	-1	NA	NA	NA	NA
-1	0	NA	NA	NA	NA
0	0	NA	NA	NA	NA
1	0	NA	NA	NA	NA
2	0	NA	NA	NA	NA
3	0	NA	NA	NA	NA
-1	1	NA	NA	NA	NA
0	1	NA	NA	NA	NA
1	1	1.097152	0.9375546	3416	13619
2	1	NA	NA	NA	NA
3	1	NA	NA	NA	NA
-1	2	NA	NA	NA	NA
0	2	NA	NA	NA	NA
1	2	NA	NA	NA	NA
2	2	NA	NA	NA	NA
3	2	NA	NA	NA	NA
-1	3	NA	NA	NA	NA
0	3	NA	NA	NA	NA
1	3	NA	NA	NA	NA
2	3	NA	NA	NA	NA
3	3	NA	NA	NA	NA

Table 4: Results for Steepest Ascent with Fisher Matrix

α_1	α_2	$\hat{\alpha_1}$	$\hat{\alpha_2}$	Iterations	Backtracks
-1	-1	NA	NA	NA	NA
0	-1	NA	NA	NA	NA
1	-1	NA	NA	NA	NA
2	-1	1.097152	0.9375546	38	23
3	-1	1.097152	0.9375546	36	15
-1	0	NA	NA	NA	NA
0	0	NA	NA	NA	NA
1	0	1.097152	0.9375546	35	11
2	0	1.097152	0.9375546	32	10
3	0	1.097152	0.9375546	32	10
-1	1	NA	NA	NA	NA
0	1	1.097152	0.9375546	37	22
1	1	1.097152	0.9375546	32	19
2	1	1.097152	0.9375546	33	14
3	1	1.097152	0.9375546	34	14
-1	2	NA	NA	NA	NA
0	2	1.097152	0.9375546	35	14
1	2	1.097152	0.9375546	38	21
2	2	1.097152	0.9375546	31	9
3	2	1.097152	0.9375546	35	22
-1	3	NA	NA	NA	NA
0	3	1.097152	0.9375546	34	10
1	3	1.097152	0.9375546	36	19
2	3	1.097152	0.9375546	34	14
3	3	1.097152	0.9375546	34	22