Finite Mixture Modeling of The number of Headaches In a Four Week Period.

Course: Topics In Advanced Modeling Techniques
Assignment: Finite Mixture Models
2019-2020

2nd year Master of Statistics
Hasselt University

Kubam Ivo (1850637)

Submission Date: 08/01/2020

Lecturers:
Prof. Dr. Geert Molenberghs
Prof. Dr. Geert Verbeke
Prof. Dr. Tomasz Burzykowski

1 Introduction

Migraine is a chronic neurovascular disorder characterized by frequent attacks of severe headache and autonomic and neurological symptoms. Acupuncture which is a complementary therapy is recommended by most health practitioners. It is widely used for chronic pain, including tension-type headache and migraine. (Price et al. 1985)

1.1 Scientific Question

This study was aimed at studying the distribution of the frequency of headache. Fit and explore the components of the finite mixture distribution then check if the components were related to the covariates in the model and if these covariates entirely explain the clusters seen in the response distribution.

2 Methodology

2.1 Data Description And Summary

Data set consist of five variables with 401 observations. The variable Frequency was the response which measures the number of headaches during a four week working period. Group(placebo an Acupuncture), gender(Female and male), Age and Migraine (yes or no) were considered the covariates.
The average number of headache was 9.3 with maximum of 28 and minimum 0. The variance was 71.564. Patient ages range from 18 to 65 years with a mean age of 45.5 years. 84% of the patients were female with just 19% male. Migraine yes group had 94% patients vs just 9% in the no group.51% against 49% for Acupuncture and placebo respectively.

2.2 Exploratory Analysis

Exploring the distribution of the observed response (frequency) as shown in figure 2 appendix, the different peaks (multi-modality) observed in the histogram suggests the presence of some underlying (latent) group structure.

2.3 Statistical Analysis

Due to multi-modality shown in figure 2, finite mixture model with a Poisson distributed components was considered.
The general expression for the finite mixture model (Böhning 1999) is given as

$$f(y) = \sum_{k}^{j=1} \pi_j(z, \alpha_j) p_j(y; x_j' \beta_j, \phi_j)$$

k repersents the number of components, π_j represents the mixture probabilities. These probabilities could depend on regressor variables z and α_j. P_j represents the component distributions which can also depend on regressor variables in x_j , regression parameters β_j, and ϕ_j. P_j are indexed by j because the distributions might belong to different families.
The distributions for the k components were assumed to poisson distributed with parameter λ.

$$Y|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim \begin{pmatrix} \lambda_1 & \lambda_2 & . & . & \lambda_k \\ \pi_1 & \pi_2 & . & . & \pi_k \end{pmatrix}$$

3 Results

3.1 Estimating The Number of Components(K)

In order to determine the number components (k),the Non parametric Maximum Likelihood Estimate (NPMLE) method was used to estimate the mixing distribution $\hat{G}$.
The Mixalg function in the CAMAN package was used to determine the number of support points in R. This led to six support points, but the first support point was dropped due to its extremely small weight. So the fitted model is

$$Y|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim G = \begin{pmatrix} 0.0214 & 3.515 & 7.430 & 13.005 & 24.083 \\ 0.251 & 0.056 & 0.279 & 0.248 & 0.158 \end{pmatrix}$$

The -2 log likelihood for the five component (2,388.2) was only slightly different from that with 6 component (2,387.7).To be sure the mixing distribution G was actually NPMLE, the gradient function was plotted as shown in figure 1. This is truely a NPMLE because

- $d(G, \lambda) \leq 1$ in the interval [0,28] where 0 and 28 are the minimum and maximum number of headaches.
- $d(G, \lambda) = 1$ at the points {0.0214, 3.515, 7.430, 13.005, 24.083}
- The obtain estimate is unique as the gradient function is not identically one.

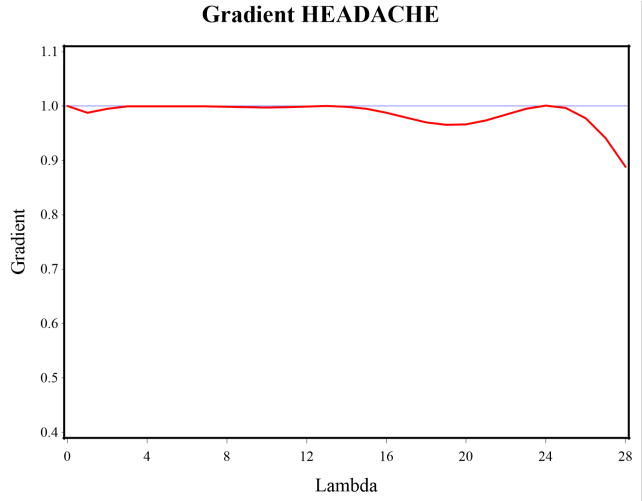


Figure 1: Gradient Function

3.2 Model One: Poisson Model

Based on the response variable (number of headache), a natural an obvious candidate model will be a poisson model.

$$Y_i \sim Poisson(\lambda_i), \quad log(\lambda_i) = \beta_0 + \beta_1 group_i + \beta_2 age_i + \beta_3 sex_i + \beta_4 migraine_i$$

where i=1,2,3...N and λ_i is the mean headache for each subject.
The results from fitting this model using the pearson scale showed strong evidence of over dispersion(scale=2.7593). Also the distribution of a single Poisson in figure 2 does not capture the distribution of the observed data as shown in figure 2. As a consequence of the over dispersion, an alternative is to fit a finite mixture model

3.3 Finite Mixture Models

A finite mixture model with K=5 better captures the distribution of data as seen in figure 2. Three different finite mixture models were considered in order to investigate if the five components determine using NPMLE were related to the covariates group, age, sex and migraine.

3.3.1 Model Two: Component Specific Regression Coefficients

$$Y_i \sim \pi_1 Poisson(\lambda_{1i}) + \pi_2 Poisson(\lambda_{2i}) + .. + (1 - \pi_1 - \pi_2 - \pi_3 - \pi_4) Poisson(\lambda_{5i})$$

$$log(\lambda_{ki}) = \beta_{k0} + \beta_1 group_{ki} + \beta_{k2} age_i + \beta_{k3} sex_i + \beta_{k4} migraine_i$$

where k=1,2,3,4,5
Model two had the following assumptions: The population consists of five sub populations. The proportion of each sub population does not depend on the covariates. The average number of headaches for each sub population depends on the covariates. The relation between the covariates and number of headaches is different amongst the sub populations.

Fitting this model resulted to a model with -2log likelihood of 2,388.2. A summary for all effects and mixing probabilities for the five components is seen in table 1

Effect	**Comp 1**	**comp 2**	**comp 3**	**Comp 4**	**Comp 5**
Intercept	3.758(0.315)	2.740(0.276)	0.076(1.151)	-31.324(655.25)	3.3704(0.296)
Age	0.005(0.005)	0.001(0.005)	0.031(0.020)	-0.064(0.077)	-0.010(0.006)
Migraine(Yes)	-0.757(0.193)*	-0.183(0.159)	-0.159(0.347)	-2.221(1.54)	-0.437(0.157)*
Group(Acupuncture)	-0.634(0.121)*	0.560(0.132)*	0.757(0.346)	21.277(0)	-0.434(0.123)*
Sex(Female)	-0.094(0.085)	-0.001(0.124)	-0.075(0.322)	11.274(655.24)	-0.236(0.114)*
Mixing Prob	0.1993	0.1341	0.1055	0.2564	0.3047

Table 1: Model 2:Parameter estimates and mixing prob for each component

Migraine, group and sex were significant in at least one of the components but age was found not to be significant in any of the components. For this reason, an extension of model 2 could be fixing age constant for all the components while letting the other covariates vary across the components.

3.3.2 Model Three (Extension of Model 2): Age Effect Kept Fixed For All Components

$$Y_i \sim \pi_1 Poisson(\lambda_{1i}) + \pi_2 Poisson(\lambda_{2i}) + .. + (1 - \pi_1 - \pi_2 - \pi_3 - \pi_4) Poisson(\lambda_{5i})$$

$$log(\lambda_{ki}) = \beta_{k0} + \beta_1 group_{ki} + \beta_2 age_i + \beta_{k3} sex_i + \beta_{k4} migraine_i$$

where k=1,2,3,4,5

Model three same assumptions as model two plus assuming age effect is the same for all components.
This model leads to a slight increase of -2log likelihood with a value of 2,398.1. The covariates group and sex both significant in at least one component while migraine insignificant in none of the groups.

3.3.3 Model Four: Mixture Weights Depend on Covariates

$$Y_i \sim \pi_{1i}Poisson(\lambda_{1i}) + \pi_{2i}Poisson(\lambda_{2i}) + .. + (1 - \pi_{1i} - \pi_{2i} - \pi_{3i} - \pi_{4i})Poisson(\lambda_{5i})$$

$$log(\lambda_{ki}) = \beta_{k0} + \beta_1 group_{ki} + \beta_2 age_i + \beta_{k3} sex_i + \beta_{k4} migraine_i$$

$$log(\pi_{ji}) = \alpha_0 + \alpha_{j1} group_i + \alpha_{j2} age_i + \alpha_{j3} sex_i + \alpha_{j4} migraine_i$$

where k=1,2,3,4,5. j=1,2,3,4.
This model assumptions that the proportion of each sub population depends on covariates. The average number of headaches for each sub population depends on the covariates and the relation between the covariates and number of headaches is different amongst the sub populations. A drop in -2log likelihood is recorded for this model, 2,374.5. Table 2 shows the mixing probabilities for each component by covariates in log scale.

Effect	**Comp 1**	**comp 2**	**comp 3**	**Comp 4**
Intercept	21.7124(1.1034)	8.4376(420.23)	23.0930(0.9648)	4.2698(0)
Age	0.05187(0.01701)*	0.04526(0.02484)	0.03675(0.01542)*	0.009989(0)
Migraine(Yes)	-1.4967(0.6228)*	-0.7126(1.5011)	-1.7199(0.5783)*	-1.9312(0)
Group(Acupuncture)	-1.5231(0.3386)*	-1.7548(0.4711)*	-1.1434 (0.3582)*	-1.6715(0)
Sex(Female)	-5.1017(0.4245)*	7.0983(420.23)	-6.0525(0.4231)*	-5.5496(0)

Table 2: Model 4: Mixing probabilities for each component (log scale)

The α for age are all positive in all four components but significant in component one and three. This implies an increase in age will lead to an increase in proportion of subjects in those components. Sex is significant in component one and three with negative estimates. This implies the proportion of subjects in component one will decrease for female subjects as compared to male subjects.

3.3.4 Comparing All Four Models

Model four has the lowest -2log likelihood but the difference is not too significant from that of model two. Also based on the fact that model four is an extension of model two with up to 45 parameters, model two seems to be the best model in the modelling the number of headaches in a four week period for subjects.

3.4 Final Model Interpretation

The final model formulation is given as follows:

$$Y_i \sim 0.199Poisson(\lambda_{1i}) + 0.134Poisson(\lambda_{2i}) + 0.106Poisson(\lambda_{3i}) + 0.256Poisson(\lambda_{4i}) + 0.305Poisson(\lambda_{5i})$$

$$Pop1 : log(\lambda_{1i}) = 3.7584 - 0.6339 group_i + 0.004765 age_i - 0.09392 sex_i - 0.7575 migraine_i$$

$$Pop2 : log(\lambda_{2i}) = 2.7401 + 0.5600 group_i + 0.001174 age_i - 0.00089 sex_i - 0.1830 migraine_i$$

$$Pop3 : log(\lambda_{3i}) = 0.07575 + 0.7568group_i + 0.03151age_i - 0.07484sex_i - 0.1591migraine_i$$

$$Pop4 : log(\lambda_{4i}) = -31.3244 + 21.2765group_i - 0.06422age_i - 0.07484sex_i - 2.2210migraine_i$$

$$Pop5 : log(\lambda_{5i}) = 3.3704 - 0.4344group_i - 0.01002age_i - 0.2364sex_i - 0.4374migraine_i$$

The population is sub divided into 5 populations with proportions 19.9%, 13.4%, 10.6%, 25.6% and 30.5% respectively. These proportions does not depend on the covariates. Subjects in the Acupuncture group in sub population one and five turn to have low average number of headaches as compared to subjects in the placebo group for these sub populations. The reverse is true for the other three sub populations. There is an increase in the average number of headaches with increase in age in sub population one, two and three. But a decrease in sub population four and five. Females subjects in all the sub populations have a low number of headaches on the average as compared to male subjects. Subjects with migraine also show on average a low number of headaches in all sub populations as compared to subjects not on migraine.
Using the posterior probabilities, the observations were classified into the five components as seen in table

Component	λ_k	π_k	**Freq**	**Prop**
One	0.0215	0.251	102	0.25
Two	3.515	0.056	23	0.06
Three	7.430	0.279	127	0.32
Four	13.005	0.248	89	0.22
Five	24.083	0.158	60	0.15
Total		1	401	1

Table 3: Classification

The prior probabilities (π_k) are closed to the component proportions in table 3.
In order to check if the five sub populations seen in this population could be describes by the covariates in the model, figure 3 in appendix shows that the number of headaches for each covariate is not well explained by the components as the boxes for observed and fitted don't perfectly overlap.

4 Discussion And Conclusion

Sometimes a single a distribution is not enough to describe the distribution of a response variable in a population. This was the case in this study were the number of headaches in a four week period was of interest. Fitting a single Poisson was not adequate as it could not describe the distribution of the response variable. Also it showed over dispersion. Finite mixture models offers natural models for unobserved population heterogeneity. This was seen for the 5 component mixture model of Poissons as it could describe the observed distribution much better than the single poisson 3 appendix. The four covariates in this study were not enough to explain the five sub populations of the mixture model. Based on the results of this study, a five component mixtures of poisson with components weight independent of the Covariates and the effects of Covariates different for each component is found to best fit.

References

Böhning, Dankmar (1999). *Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others*. Vol. 81. CRC press.

Price, Catherine J et al. (1985). "The developmental toxicity of ethylene glycol in rats and mice". In: *Toxicology and Applied Pharmacology* 81.1, pp. 113–127.

5 Appendix

5.0.1 Tables And Figures

Criteria	model 1	Model 2	Model 3	Model 4
Number of Parameters	5	29	29	45
-2log likelihood	4,709.4	2,388.2	2,398.1	2,374.5

Table 4: Models Comparison

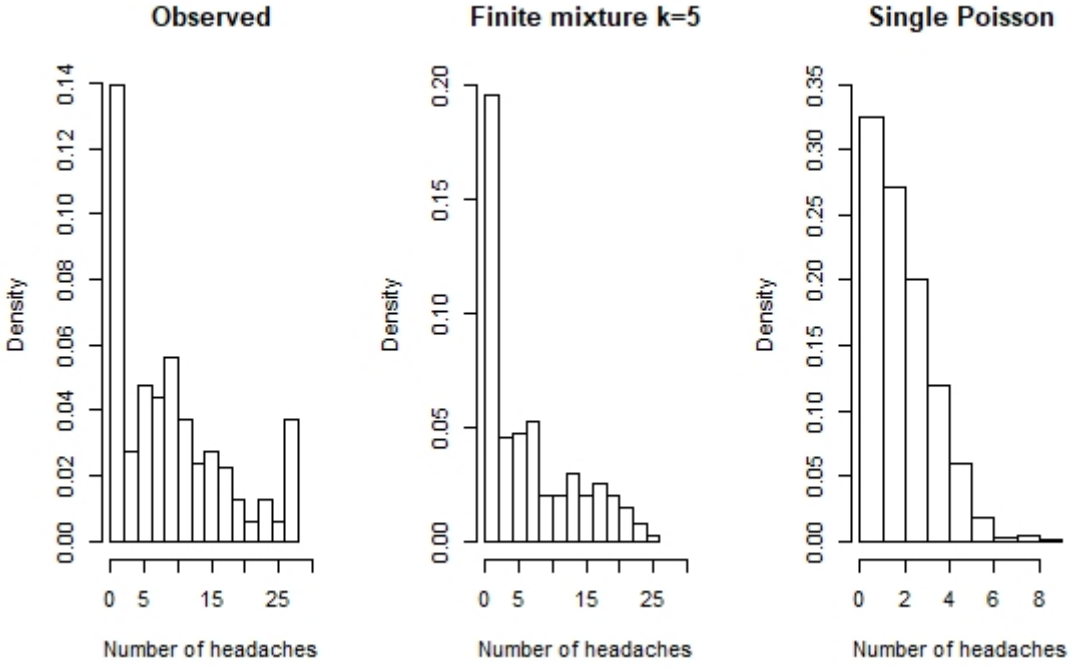
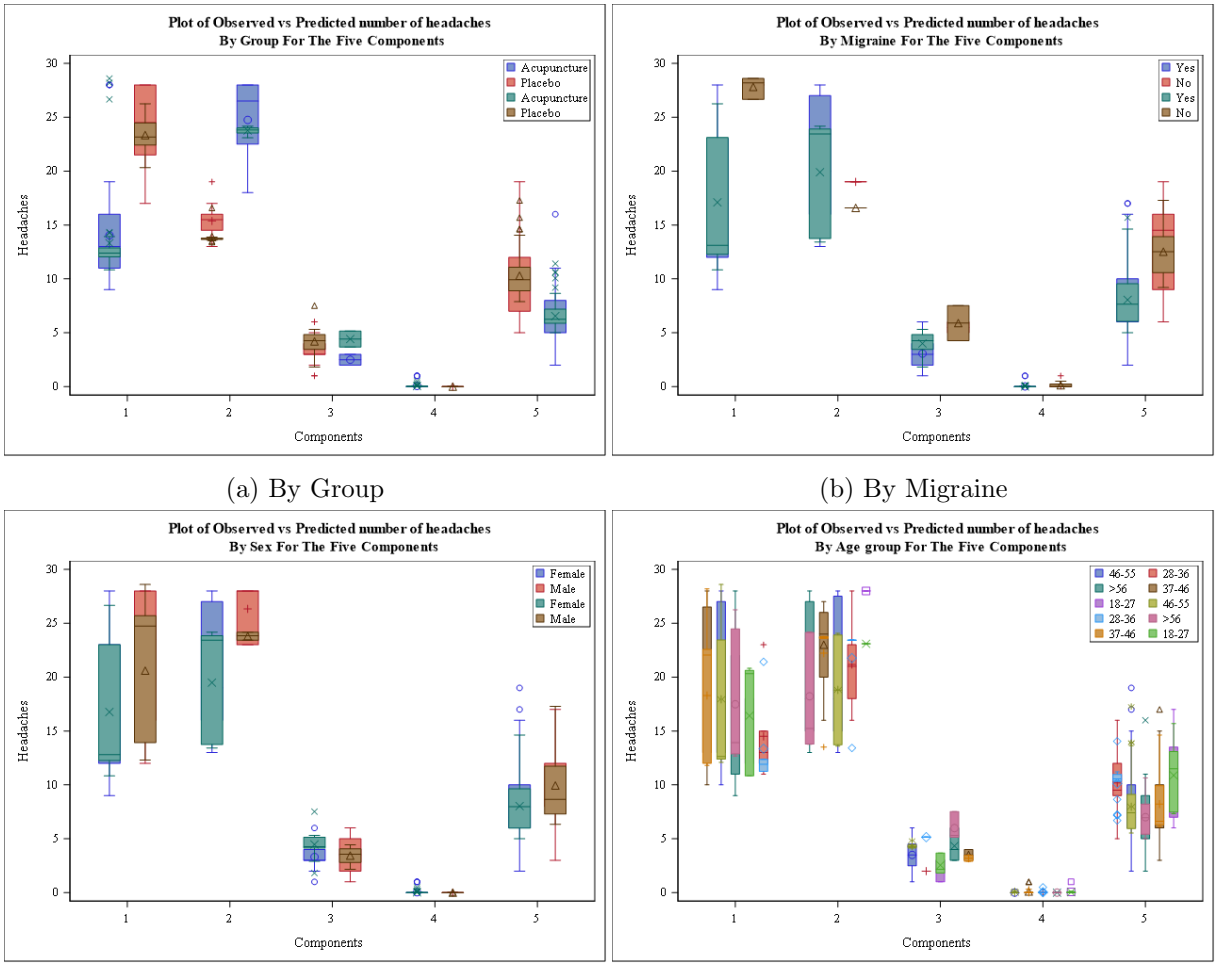


Figure 2: Distribution of Number of headaches



(a) By Group

(b) By Migraine

(c) By Sex

(d) By Age

Figure 3: Observed vs predicted number of headaches for each component by group, migraine and sex

5.1 SAS And R codes

```
R Codes
require("sas7bdat")
require("CAMAN")

#Reading in sas dataset;
headache<-read.sas7bdat("E:\\SEED\\OneDrive\\Msc. Biostatistics\\Level Two\\Advanced

#Estimating the number of support points using NPMLE
#Phase 1 and 2 combined
npml1<-mixalg(obs =headache2$Count,weights = headache2$freq,family = "poisson",data=h
npml1

SAS CODES
*Reading the data set;
data amt.headache;
set "E:\SEED\OneDrive\Msc. Biostatistics\Level Two\Advanced Modelling Techniques\Proj
```

```
format sex f_gender. migraine f_migraine. group f_group.;
frequency=ceil(frequency);
run;
/*Non parametric maximum likelihood estimate*/
data test;
set amt.headache;
run;
data test;
set amt.headache;
do x=0 to 28 by 1; output;
end;
run;
data test;set test;
gradient = pdf('POISSON',frequency,x)/(0.006711798*pdf('POISSON',frequency,0)+(0.2511
+0.279212202*pdf('POISSON',frequency, 7.42992944)+0.248212925 *pdf('POISSON',frequenc
run;
proc sort data = test;
by x;
proc means data=test;
var gradient;
by x;
output out=out;

data out;
set out;
if _STAT_ = 'MEAN';

title h=2.5 'Gradient HEADACHE';
proc gplot data = out;
plot gradient*x /nolegend haxis=axis1 vaxis=axis2 cvref=blue vref=1;
symbol c=red i=join w=5 l=1 mode=include;
axis1 label=(h=2 'Lambda') value=(h=1.5) order=(0 to 28 by 4) minor=none w=6;
axis2 label=(h=2 angle=90 'Gradient') value=(h=1.5) order=(0.4 to 1.1 by 0.1) minor=n
run;
quit;

/*MODEL EXTENSION WITH COVARIATES*/
/*Fitting a single poisson model*/
proc genmod data= amt.headache;
class migraine sex group;
model frequency = age migraine group sex /link=log dist=poisson scale=p;
run;

/*MODEL TWO K=6*/
ODS graphics on;
title "MODEL TWO K=6";
```

```
proc fmm data=amt.headache;
class migraine sex group;
model frequency = age migraine group sex /dist=poisson k=6;
run;
TITLE;

/*MODEL TWO*/
title "MODEL TWO K=5";
title2 "Covariates vary across components";
proc fmm data=amt.headache  order=freq componentinfo fitdetails ;
class migraine sex group;
model frequency = age migraine group sex /dist=poisson k=5 ;
ID frequency age migraine group sex  ;
output out=amt.out_pred class pred(components);
run;
TITLE;
/*MODEL TWO B*/

title "MODEL TWO K=5";
title2 "Age Fixed";
proc fmm data=amt.headache  order=freq componentinfo fitdetails ;
class migraine sex group;
model frequency = age migraine group sex /dist=poisson k=5 equate=effects(age);
ID frequency;
output out=amt.out_pred class pred(components);
run;
TITLE;

/*MODEL THREE*/
title "MODEL TWO K=5";
title2 "Equal Effect in the components";
proc fmm data=amt.headache order=freq componentinfo fitdetails ;
class migraine sex group;
model frequency = age migraine group sex /dist=poisson k=5 equate=effects(age migrain
output out=amt.out_pred class pred(component) maxprob;
run;
TITLE;

Title "MODEL Four";
title2 "Mixture Weights Depends on Covariates";
proc fmm data=amt.headache order=freq componentinfo noitprint;
class migraine sex group;
model frequency = age migraine group sex /dist=poisson k=5;
PROBMODEL age migraine group sex;
run;
```