Model based clustering for multivariate count data

Dimitris Karlis

Department of Statistics

Athens University of Economics

and

Loukia Meligkotsidou

Department of Mathematics and Statistics,

Lancaster University

Outline

- Motivation
- Model Based Clustering
- Multivariate Poisson models
- Multivariate Poisson mixtures
- Finite Multivariate Poisson mixtures
- Application to crime data
- Conclusions -Open problems

Motivation

Multivariate data are usually modelled via

- Multivariate Normal models
- Multinomial models (for categorical data)

What about multivariate count data?

- Small counts with a lot of zeros
- Normal approximation may not be adequate at all

Idea: Use multivariate Poisson models

Attractive idea but the models are computationally demanding.

Multivariate Count data

- Incidences of different diseases across time or space
- Different type of crimes in different areas
- Purchases of different products
- Accidents (different types or in different time periods)
- Football data
- Different types of faults in production systems
- Number of faults in parts of a large system etc

Clustering

Purpose: To find observations with similar characteristics

Methods used

- Hierarchical Clustering

  Large storage demands,dependence on the distance measure and the linkage method,not solid theoretical background

- K- Means

  Heuristic algorithm, computationally feasible, dependence on the initial solution, not solid theoretical background.

- Model Based Clustering

  Probabilistic method, strong theoretical background,inferential procedures available

Model Based Clustering (see, e.g. Banfield and Raftery, 1993)

- The population consists of k subpopulations
- For the q-dimensional observation $\boldsymbol{y}_i$ from the j -th subpopulation we have

$$\boldsymbol{y}_i \sim f_j(y_i \mid \boldsymbol{\theta}_j), \tag{1}$$

  ($\boldsymbol{\theta}_j$ unknown vector of parameters related to the j-th subpopulation)

- Unconditional density :

$$f(\boldsymbol{y}_i|\boldsymbol{\Theta}) = \sum_{j=1}^{k} p_j f_j(y_i \mid \boldsymbol{\theta}_j) \tag{2}$$

  Questions to be answered:
  - What is the value of k or how many clusters?
  - Estimate $\boldsymbol{\theta}_j$, i.e. estimate the characteristics of each cluster

Multivariate Poisson model

Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ and $X_i \sim Poisson(\theta_i)$, $i = 1, \ldots, m$. Then the general definition of multivariate Poisson models is made through the matrix $\mathbf{A}$ of dimensions $k \times m$, where the elements of the matrix are zero and ones and no duplicate columns exist.

Then the vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ defined as

$$\mathbf{Y} = \mathbf{AX}$$

follows a multivariate Poisson distribution.

Complete Specification

$$\mathbf{A} = [A_1 \quad A_2 \quad \dots \quad A_k]$$

where A_i is a matrix of dimensions $k \times \begin{pmatrix} k \\ i \end{pmatrix}$ where each column has exactly i ones and $k - i$ zeroes.

Example $k = 3$

$$A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad A_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad A_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

and then

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

This correspond to

$$\begin{aligned} X_1 &= Y_1 + Y_{12} + Y_{13} + Y_{123} \\ X_2 &= Y_2 + Y_{12} + Y_{23} + Y_{123} \\ X_3 &= Y_3 + Y_{13} + Y_{23} + Y_{123} \end{aligned} \tag{3}$$

where all Y_i's, are independently Poisson distributed random variables with parameter θ_i, $i \in (\{1\}, \{2\}, \{3\}, \{12\}, \{13\}, \{23\}, \{123\})$

Note: Parameters θ_{ij} are in fact covariance parameters between X_i and X_j. Similarly θ_{123} is a common 3-way covariance parameter.

Other cases

Independent Poisson variables

Corresponds to the case $A = A_1$.

Example for $k = 3$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

i.e. product of independent Poisson probability functions.

Full covariance structure

If we want to specify only up to 2-way covariances we take the form

$$\mathbf{A} = [A_1 \quad A_2]$$

Example $k = 3$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

This model is very interesting as it assumes different covariances between all the pairs and thus it resembles the multivariate normal model.

This correspond to

$$\begin{aligned} X_1 &= Y_1 + Y_{12} + Y_{13} \\ X_2 &= Y_2 + Y_{12} + Y_{23} \\ X_3 &= Y_3 + Y_{13} + Y_{23} \end{aligned}$$

where all Y_i's, are independently Poisson distributed random variables with parameter θ_i, $i \in (\{1\}, \{2\}, \{3\}, \{12\}, \{13\}, \{23\})$

the covariance matrix of (X_1, X_2, X_3) is now

$$Var(\mathbf{X}) = \begin{bmatrix} \theta_1 + \theta_{12} + \theta_{13} & \theta_{12} & \theta_{12} \\ \theta_{12} & \theta_2 + \theta_{12} + \theta_{23} & \theta_{23} \\ \theta_{13} & \theta_{23} & \theta_3 + \theta_{13} + \theta_{23} \end{bmatrix}$$

Properties

For the general model we have

$$E(X) = \mathbf{AM}$$

and

$$Var(X) = \mathbf{A\Sigma A^T}$$

where $\mathbf{M}$ and $\mathbf{\Sigma}$ are the mean vector and the variance covariance matrix for the variables $Y_0, Y_1, \ldots, Y_k$ respectively.

$\mathbf{\Sigma}$ is diagonal because of the independence of Y_i's and has the form

$$\mathbf{\Sigma} = diag(\theta_1, \theta_2, \ldots, \theta_m)$$

Similarly

$$\mathbf{M}^T = (\theta_1, \theta_2, \ldots, \theta_m)$$

Covariance Model

We concentrate on the 2-way full covariance model. A model with m variables has

$$m + \begin{pmatrix} m \\ 2 \end{pmatrix}$$

parameters.

Bad news: The joint probability function has at least m summations!

Good news: One may use recurrence relationships (clearly need to find efficient algorithms to do so)

Mixtures of multivariate Poisson distribution

Several different ways to define such mixtures:

- Assume

$$\begin{aligned}(X_1, \ldots, X_m) &\sim m - Poisson(\alpha\boldsymbol{\theta})) \\ \alpha &\sim G(\alpha)\end{aligned}$$

- Assume

$$\begin{aligned}(X_1, \ldots, X_m) &\sim m - Poisson(\boldsymbol{\theta})) \\ \boldsymbol{\theta} &\sim G(\boldsymbol{\theta})\end{aligned}$$

- Part of the vector $\boldsymbol{\theta}$ varies, while some of he parameters remain constant. For example (in 2 dimensions)

$$\begin{aligned}(X_1, X_2) &\sim Biv.Poisson(\theta_1, \theta_2, \theta_0) \\ \theta_1, \theta_2 &\sim G(\theta_1, \theta_2)\end{aligned}$$

Dependence Structure

Consider the case

$$\begin{aligned}(X_1, \ldots, X_m \mid \boldsymbol{\theta}) &\sim m - Poisson(\boldsymbol{\theta}))\\ \boldsymbol{\theta} &\sim G(\boldsymbol{\theta})\end{aligned}$$

The unconditional covariance matrix is given by

$$Var(X) = \mathbf{ADA}^T$$

where $\mathbf{A}$ is the matrix used to construct the conditional variates from the original independent Poisson ones and

$$\mathbf{D} = \begin{bmatrix} Var(\theta_1) + E(\theta_1) & Cov(\theta_1, \theta_2) & \dots & Cov\theta_1, \theta_m) \\ Cov(\theta_1, \theta_2) & Var(\theta_2) + E(\theta_2) & \dots & Cov(\theta_2, \theta_m) \\ & \dots & & \\ Cov(\theta_1, \theta_m) & \dots & & Var(\theta_m) + E(\theta_m) \end{bmatrix}$$

Important findings

Remark 1: The above formula imply that if the mixing distribution allows for any kind of covariance between the θ's then the resulting unconditional variables are correlated. Even in the case that one starts with independent Poisson variables the mixing operation can lead to correlated variables.

Remark 2: More importantly, if the covariance between the pairs (θ_i, θ_j) is negative the unconditional variables may exhibit negative correlation. It is well known that the multivariate Poisson distribution cannot have negative correlations, this is not true for its mixtures.

Remark 3: The covariance matrix of the unconditional random variables are simple expressions of the covariances of the mixing parameters and hence the moments of the mixing distribution. Having fitted a multivariate Poisson mixture model, one is able to estimate consistently the reproduced covariance structure of the data. This may serve as a goodness of fit index.

Finite Mixtures of multivariate Poisson distribution

If we assume that $\boldsymbol{\theta}$ can take only a finite number of different values finite multivariate Poisson mixture arise. The pf is given as

$$P(\boldsymbol{X}) = \sum_{j=1}^{k} p_j P(\boldsymbol{X} \mid \boldsymbol{\theta}_j)$$

where $P(\boldsymbol{X} \mid \boldsymbol{\theta}_j)$ denotes the pf of a multivariate Poisson distribution.

this model can be used for clustering multivariate count data Examples:

- Cluster customers of a shop according to their purchases in a series of different products
- Cluster areas according to the number of occurrences of different types of a disease etc

Inference

Standard inferential procedures as those for known model based clustering:

- Estimation through an EM algorithm using the latent structure of the mixture
- Choose the number of cluster via a standard method like AIC, BIC, NEC etc
- Allocate obsrvations to cluster using the posterior probability that an observation belong to a cluster (readily available through the EM)

Application

Multivariate Count data: number of 4 different type of crimes in Greece for the year 1997, for 50 perfectures.

Crime type:

rapes, arson , smuggling of antiquities and general smuggling.

The population of each prefecture is used as an offset.

The aim is to cluster the prefectures according to their profiles in those types of crimes.

Results (1)

- An EM type algorithm was used to fit the finite multivariate Poisson mixture model. The number k of components was considered as known for using the EM algorithm, but we fitted the model with increasing value of k in order to decide about the number of components.
- AIC criterion is used to find the optimal number of clusters For a model with k components there are $11k - 1$parameters to estimate. We selected a 4 clusters solution

Results (2)

Mixing proportions $\hat{p} = (0.5915, 0.2266, 0.0638, 0.1181)$.

Parameters in matrix form:

$$\Theta = \begin{bmatrix} \theta_1 & \theta_{12} & \theta_{13} & \theta_{14} \\ & \theta_2 & \theta_{23} & \theta_{24} \\ & & \theta_3 & \theta_{34} \\ & & & \theta_4 \end{bmatrix}, \tag{4}$$

Results (2)

$$\Theta_1 = \begin{bmatrix} 17.339 & 0 & 4.772 & 0 \\ & 2.530 & 0.198 & 1.977 \\ & & 34.112 & 2.398 \\ & & & 6.171 \end{bmatrix}, \Theta_2 = \begin{bmatrix} 0 & 0 & 3.675 & 0 \\ & 9.897 & 1.925 & 1.938 \\ & & 5.320 & 0 \\ & & & 2.172 \end{bmatrix},$$

$$\Theta_3 = \begin{bmatrix} 0 & 20.424 & 0 & 0 \\ & 55.868 & 24.323 & 0 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix}, \Theta_4 = \begin{bmatrix} 14.416 & 12.780 & 0 & 0 \\ & 3.048 & 0 & 0 \\ & & 8.934 & 0 \\ & & & 44.621 \end{bmatrix}$$

Interesting things

- Standard model-based clustering procedures can be applied. for example, estimation is feasible via EM algorithm, selection of the number of components can be used in a variety of criteria etc
- Since, mixing operation imposes structure is not a good idea to start with a model with a lot of covariance terms.
- Since we work with counts one may use the frequency table instead of the original observations. This speeds up the process and the computing time is not increased so much even if the sample size increases dramatically

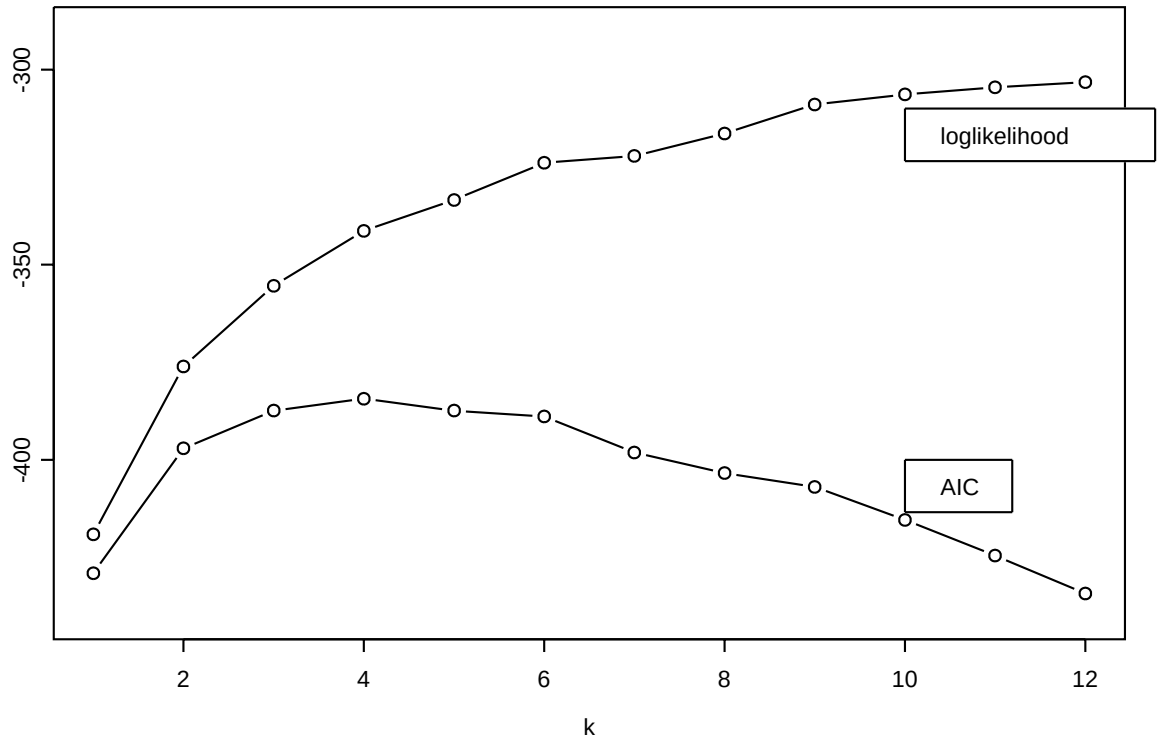
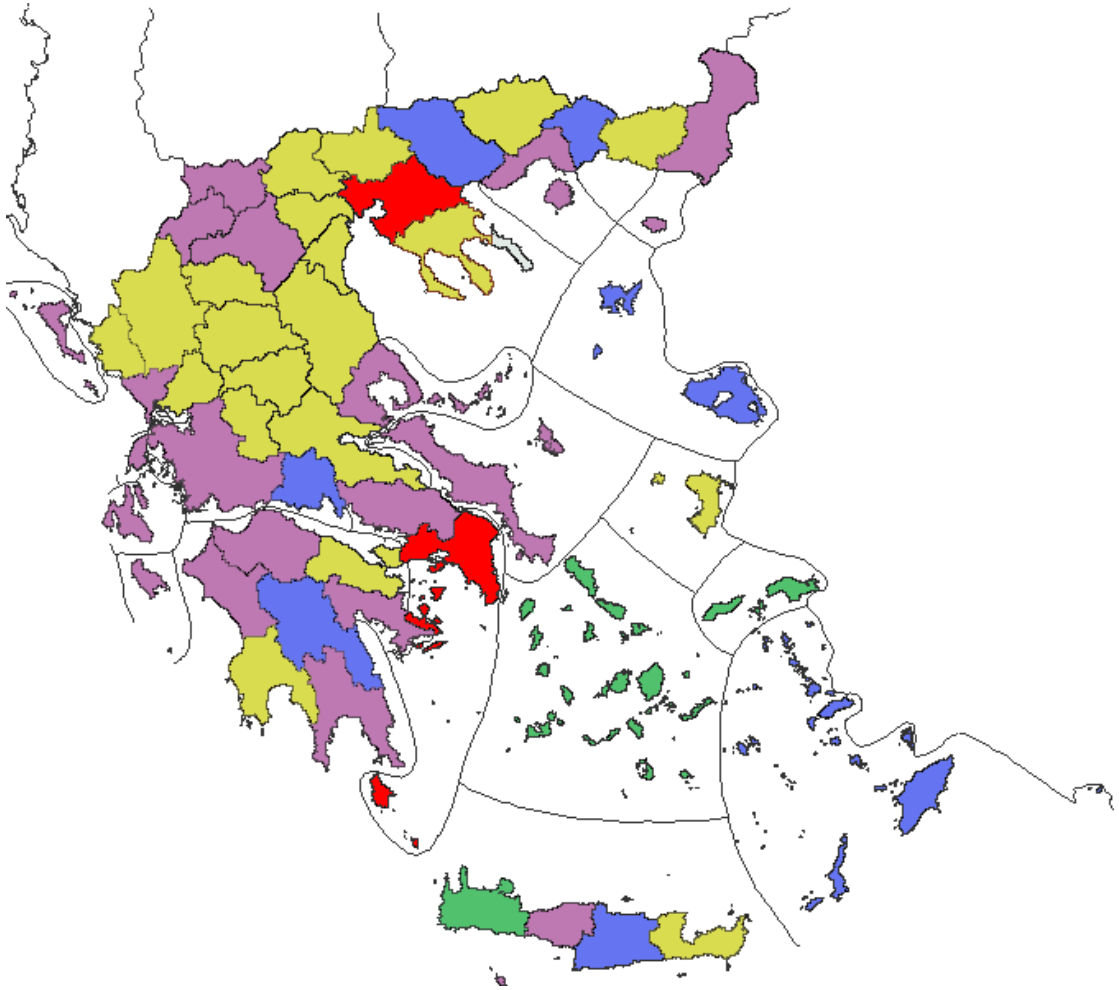


Figure 1: The loglikelihood and the AIC criterion (rescaled) for the crime data for different values of k



Summary

- Multivariate Poisson model similar in nature to multivariate normal were considered.
- The model can be generalized to have quite large (but unnecessary) structure.
- Using up to 2-way covariance term suffice to describe most data sets
- Estimation can be accomplished via EM algorithms (or MCMC schemes from the Bayesian perspective)

Open problem -Future and Ongoing research

- Need to speed up estimation, including quick calculation of the probabilities and improving the EM algorithm.
- Model selection procedure must be obtained that are suitable for the kind of data (e.g. selection of appropriate covariance terms)