MODEL BASED CLUSTERING FOR COUNT DATA

Dimitris Karlis

Department of Statistics

Athens University of Economics and Business,

Athens – April 2002

OUTLINE

- Clustering methods
- Model based clustering
    - The general model
    - Algorithmic issues
    - Inference
    - Multivariate Normal clustering
- Multivariate Poisson distributions
- Model Based clustering via the Multivariate Poisson distribution
    - The models
    - Estimation
- Application to marketing data
    - The data
    - The model
    - Results
- Generalizing the model
- Future research and open problems

CLUSTERING

Purpose: To find observations with similar characteristics

Other names: taxonomy, segmentation etc

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

The population consists of k subpopulations

For the q-dimensional observation $\mathbf{y_i}$ from the j-th subpopulation:

$y_i \sim f(y_i \mid \theta_j)$ (θ_j unknown vector of parameters)

Unconditional density :

$$f(y_i) = \sum_{j=1}^{k} p_j f(y_i \mid \theta_j)$$

Problem: Finding the values of the non-observable vector

$$\varphi = (\phi_1 , \phi_2 , \ldots, \phi_n)$$

$\phi_i = j$ if the i-th observation belongs to the j-th subpopulation.

Define the vector $$z_{ij} = \begin{cases} 1 & \phi_i = j \\ 0 & \phi_i \neq j \end{cases}$$

ESTIMATION

The **purpose** of model-based clustering is to estimate the parameters

$\Theta = (p_1, ..., p_{k-1}, \theta_1, ..., \theta_k)$.

Loglikelihood:

$$L(y; \theta, p) = \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{k} p_j \, f(y_i \mid \theta_j) \right)$$

Solution: *EM algorithm*

EM ALGORITHM FOR MODEL BASED CLUSTERING

Observed data y_i

Complete data (y_i, z_i)

- E-step: Estimate

$$w_{ij} = E(z_{ij} \mid y_i, \Theta)$$

- M-step: Update Θ by solving appropriate equations
  (usually weighted likelihood equations with weights w_{ij})

Variants: Clasiffication EM (CEM)

THE EM ALGORITHM

(Dempster *et al,* 1975, Meng and Van Dyk, 1997, McLachlan and Krishnan, 1997)

Complete Data $Z_i = (X_i, Y_i)$

X_i observed data

Y_i missing data (or they can be considered as missing)

E-step

Compute $Q(\varphi \mid \varphi^{(k)}) = E(\log p(Z \mid \varphi) \mid X, \varphi^{(k)})$

M-step

Update $\varphi^{(k+1)}$ by maximizing $Q(\varphi \mid \varphi^{(k)})$ with respect to φ

MODEL BASED CLUSTERING VIA MULTIVARIATE NORMAL MIXTURES

(Barnfield and Raftery, 1993)

Assume $y_i \sim f(x \mid \theta_j)$ is $MN(\mu_j, \Sigma_j)$

μ_j mean vector

Σ_j covariance matrix

Wide range of applications

(see, McLachlan and Basford, 1989, McLAchlan and Peel, 2001 etc)

INTERESTING FEATURES

$$\Sigma_k = \lambda_k D_k A D_k '$$

λ_k volume

A_k shape

D_k orientation

Model	**Covariance matrix of the k-th component** Σ_k	**Shape**	**Orientation**	**Volume**
1	λI	Spherical	None	Same
2	$\lambda_k I$	Spherical	None	Different
3	Σ	Same	Same	Same
4	$\lambda_k \Sigma$	Same	Same	Different
5	$\lambda D_k A D_k '$	Same	Different	Same
6	$\lambda_k D_k A D_k '$	Same	Different	Different
7	$\lambda_k DAD'$	Different	Same	Different
8	Σ_k	Different	Different	Different

NUMBER OF COMPONENTS

Criteria:

- Information criteria
  - AIC $AIC = -2L_k + 2\,d_k$
  - BIC $BIC = -2L_k + d_k \ln(n)$
  - AIC3, CAIC, JAAIC
  - Information Complexity Criteria
  - Minimum Information Ratio Criterion
- Cross-validated Criteria
- Classification Based Criteria
  - Normalized Entropy Criterion
- Bootstrap based criteria
  - Bootstrap LRT

INFERENTIAL PROBLEMS

General ML Theorem:

The number of components that can be fitted in any dataset is finite and it depends on the data.

Usually we are able to estimate a small number of components

Other Topics

- Goodness of fit
- Standard errors
- Classifying new observations

MULTIVARIATE COUNT DATA

Examples

- Number of different crimes in different areas/time
- Number of occurrence of different diseases in different areas/time
- Number of different transaction for a bank account
- Number of purchases of different products

Common elements

- correlation between the variables
- the data are counts, usually with large number of zero, so multivariate normal approach inappropriate

Approach Proposed:

Model based clustering based on the multivariate Poisson distribution

MULTIVARIATE POISSON DISTRIBUTION

$$X_i \sim Poisson(\lambda_i), \qquad i=1,\ldots,m$$

X_i's are independent

Multivariate Poisson distributions (MP) defined as

$$Y=\mathbf{A}X,$$

A is a q x m matrix of zeros and ones.

Example 1.

m=3, q=2

$$X = (X_1, X_2, X_3)$$

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

$$Y = (Y_1, Y_2)$$

$$Y_1 = X_1 + X_3$$
$$Y_2 = X_2 + X_3$$

(*Trivariate reduction technique*)

→ Important result: The marginal distributions are Poisson distributions

GENERAL MODEL

Suppose that the matrix **A** has the form

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

where $\mathbf{A_i}$ is a matrix of dimensions $q \times \binom{q}{i}$ where the columns of the matrix are all the combinations containing exactly *i* ones and *q-i* zeros.

$$E(Y) = \mathbf{A}\mathrm{M} \text{ and } Var(Y) = \mathbf{A\Sigma A^T}$$

where

$$\mathrm{M} = E(X) = (\lambda_1 , \lambda_2 , \ldots, \lambda_m)$$

and $\mathbf{\Sigma}$ is the variance/covariance matrix of X and is given as:

$$\mathbf{\Sigma} = Var(X) = diag(\lambda_1 , \lambda_2 , \ldots, \lambda_m)$$

Example

For instance, for q=3, we need

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

$$Y_1 = X_1 + X_{12} + X_{13} + X_{123}$$
$$Y_2 = X_2 + X_{12} + X_{23} + X_{123}$$
$$Y_3 = X_3 + X_{13} + X_{23} + X_{123}$$

where the form of the matrix **A** is now:

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

and $X = (X_1 , X_2 , X_3 , X_{12} , X_{13} , X_{23} , X_{123})$.

SOME INTERESTING THINGS

- Calculation of the probability function computationally demanding
- Estimation via EM algorithm
- Use of the entire covariance structure of little practical interest
- Very few applications, usually assuming just a common covariance term
- Even if we start with conditionally independent Poisson distribution the resulting mixture has covariance structure

COVARIANCE STRUCTURE

$X_i \mid \lambda_i \sim Poisson(\lambda_i)$, $i=1,\ldots,m$, X_i's are independent

Define $Y=\mathbf{A}X$, where $\mathbf{A}$ is a q x m matrix of zeros and ones.

Let $\theta = (\lambda_1, \lambda_2, ..., \lambda_m)$ the vector of parameters. Then,

$$Y \mid \theta \sim MultPoisson(\theta)$$
$$\theta \sim G(\theta),$$

The unconditional variance of Y is given as

$$Var(Y) = ADA'$$

where

$$D = \begin{bmatrix} Var(\lambda_1)+E(\lambda_1) & Cov(\lambda_1,\lambda_2) & ... & Cov(\lambda_1,\lambda_m) \\ Cov(\lambda_1,\lambda_2) & Var(\lambda_2)+E(\lambda_2) & ... & Cov(\lambda_1,\lambda_m) \\ ... & ... & ... & ... \\ Cov(\lambda_1,\lambda_m) & ... & ... & Var(\lambda_m)+E(\lambda_m) \end{bmatrix}$$

$$D = Var(\theta) + B,$$

$$B = \begin{bmatrix} E(\lambda_1) & 0 & ... & 0 \\ 0 & E(\lambda_2) & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & E(\lambda_m) \end{bmatrix}$$

Important things

- The unconditional covariance can be decomposed in two parts: the intrinsic covariance and the covariance due to mixing
- Even if the variables are conditionally uncorrelated, unconditionally there is covariance
- Resulting covariance can be negative as well (multivariate Poisson model has necessarily positive covariance)

APPLICATION

(Brijs, Karlis, et al. 2002)

The data

155 customers of a large super market chain in Belgium.

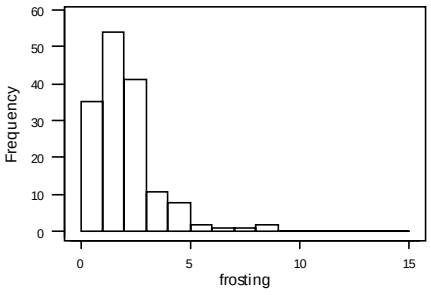
Purchase of 4 products were of interest

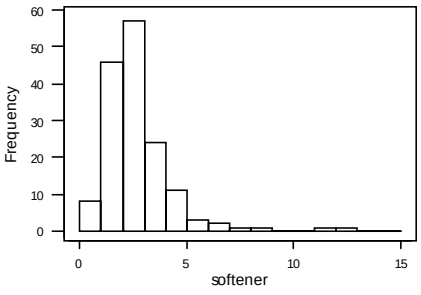
- Cakemix
- Frosting
- Softener
- Detergent

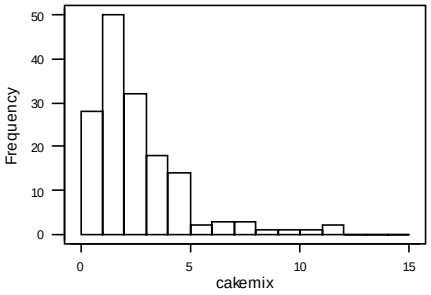
	Mean	Variance	Variance / mean
Cakemix	2.07742	4.46150	2.14762
Frosting	1.54839	2.18433	1.41071
Detergent	3.15484	6.52132	2.06709
Softener	2.20000	2.86234	1.30106

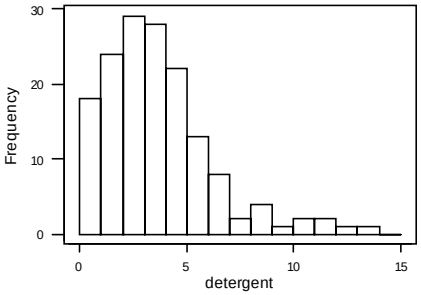
Univariate Clustering

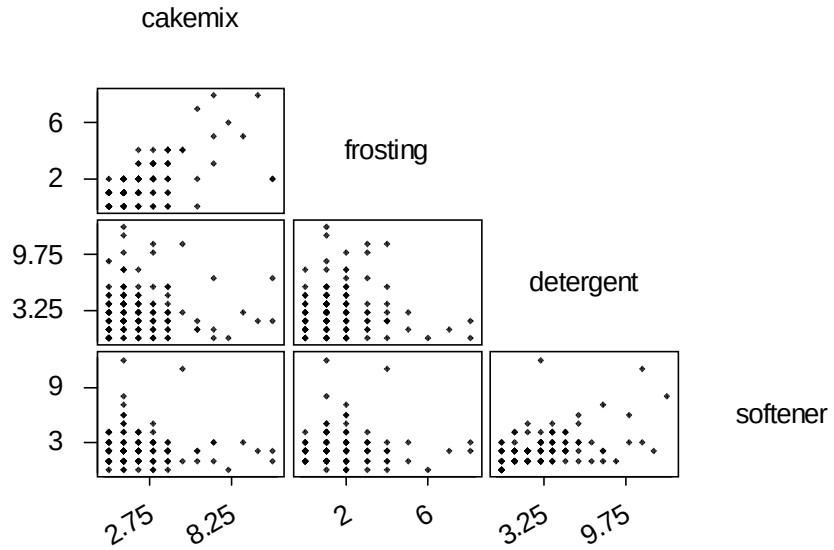
Product	Number of components
Softener	2
Detergent	3
Frosting	2
Cakemix	3


Frequency
frosting


Frequency
softener


Frequency
cakemix


Frequency
detergent


cakemix
frosting
detergent
softener
2.75
8.25
2
6
3.25
9.75

THE MODEL

- Not all the covariances were used
- Restricted covariance structure

Important covariances

Frostings and cakemix (r=0.66)

Detergent and softener (r=0.48)

The model

Latent variables

$X = (X_C, X_F, X_D, X_S, X_{CF}, X_{DS})$

The form of the matrix **A** is now:

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

the vector of parameters is now

$\theta = (\lambda_C, \lambda_F, \lambda_D, \lambda_S, \lambda_{CF}, \lambda_{DS})$.

Thus we have

$$Y_C = X_C + X_{CF}$$
$$Y_F = X_F + X_{CF}$$
$$Y_D = X_D + X_{DS}$$
$$Y_S = X_S + X_{DS}$$

Conditional probability function

$$P(y \mid \theta) = P(y_C, y_F, y_D, y_S \mid \theta)$$
$$= BP(y_C, y_F; \lambda_C, \lambda_F, \lambda_{CF}) BP(y_D, y_S; \lambda_D, \lambda_S, \lambda_{DS})$$

where

$$BP(y_1, y_2; \lambda_1, \lambda_2, \lambda_{12}) = \frac{e^{-\lambda_1} \lambda_C{}^{y_1}}{y_1!} \frac{e^{-\lambda_2} \lambda_2{}^{y_2}}{y_2!} \sum_{i=0}^{\min(y_1, y_2)} \binom{y_1}{i} \binom{y_2}{i} i! \left(\frac{\lambda_{12}}{\lambda_1 \lambda_2} \right)^i$$

with $y_1, y_2 = 0, 1, \ldots$

The unconditional probability mass function is given under a mixture with k-components model by

$$P(y) = \sum_{j=1}^{k} p_j P(y_C, y_F, y_D, y_S \mid \theta_j)$$

Estimation for the mixture model

- For a model with k components the number of parameters = $7k$-1.

- The likelihood function is quite complicated for direct maximization.

- EM type of algorithm is used.

THE EM ALGORITHM

Observed data

$Y_i = (Y_{Ci}, Y_{Fi}, Y_{Di}, Y_{Si})$

Unobserved data

$X_i = (X_{Ci}, X_{Fi}, X_{Di}, X_{Si}, X_{CFi}, X_{DSi})$ and

$Z_i = (Z_{1i}, Z_{2i}, \ldots, Z_{ki})$

Complete Data

the vectors (Y_i, X_i, Z_i).

Θ : the vector of parameters,

E-step: Using the current values of parameters calculate

$$w_{ij} = E(Z_{ji} \mid Y_i, \Theta) = \frac{p_j P(y_i \mid \theta_j)}{P(y_i)}, \quad i=1,\ldots n, j=1,\ldots k$$

$$x_{CFi} = E(X_{CF,i} \mid Y_i, \Theta) = \sum_{j=1}^{k} p_j BP(y_{Di}, y_{Si}; \lambda_{Dj}, \lambda_{Sj}, \lambda_{DSj}) \quad \times$$

$$\frac{\sum_{r=0}^{\min(y_C, y_F)} rPo(y_{Ci} - r \mid \lambda_{Cj}) Po(y_{Fi} - r \mid \lambda_{Fj}) Po(r \mid \lambda_{CFj})}{P(y_i)},$$

$$x_{DSi} = E(X_{DS,i} \mid Y_i, \Theta) = \sum_{j=1}^{k} p_j BP(y_{Ci}, y_{Fi}; \lambda_{Cj}, \lambda_{Fj}, \lambda_{CFj}) \quad \times$$

$$\frac{\sum_{r=0}^{\min(y_D, y_S)} rPo(y_{Di} - r \mid \lambda_{Dj}) Po(y_{Si} - r \mid \lambda_{Sj}) Po(r \mid \lambda_{DSj})}{P(y_i)},$$

$$x_{Ci} = y_{Ci} - x_{CFi}, \; x_{Fi} = y_{Fi} - x_{CFi},$$
$$x_{Di} = y_{Di} - x_{DSi}, \; x_{Si} = y_{Si} - x_{Dsi},$$

M-step: Update the parameters

$$p_j = \frac{\sum_{i=1}^{n} w_{ij}}{n}, \qquad j=1,\ldots k$$

$$\lambda_{ij} = \frac{\sum_{i=1}^{n} w_{ij} x_{ji}}{\sum_{i=1}^{n} w_{ij}}, \qquad j=1,\ldots k, \; i \in \{C, F, D, S, CF, DS\}$$

If some convergence criterions is satisfied, stop iterating, otherwise go back to the *E-step*.

Issues related to the EM

- Stopping criterion: $\left|\frac{L(k+1)-L(k)}{L(k)}\right| < 10^{-12}$

- Multiple maxima

Step 1: 10 sets of initial values are chosen randomly

Step 2: Each set run for 100 iterations and

Step 3: We keep iterating from the solution with the largest likelihood after the initial 100 iterations

The above procedure was run 20 times.

Check via the gradient function

Scalability:

- Working with frequency tables, not with original data speed up the process considerably.
- Appropriate for huge databases with million of customers
- Appropriate for Data mining
- Speed depends on
    - the number of components
    - the separation between components and
    - the relative size of the parameters

Number of components	Number of parameters	loglikelihood	AIC	BIC
1	6	. -1132.579565	1138.58	1139.15
2	13	-1099.44663	1112.447	1113.684
3	20	-1071.375534	1091.376	1093.279
4	27	-1061.069886	1088.07	1090.639
5	34	-1053.249673	1087.25	1090.485
6	41	-1045.666914	1086.667	1090.569
7	48	-1040.5518	1088.552	1093.12
8	55	-1035.695927	1090.696	1095.93
9	62	-1029.064983	1091.065	1096.965
10	69	-1024.21259	1093.213	1099.779
11	76	-1022.20919	1098.209	1105.442
12	83	-1020.641026	1103.641	1111.54
13	90	-1018.793612	1108.794	1117.359
14	97	-1018.185097	1115.185	1124.416
15	104	-1015.845674	1119.846	1129.743
16	111	-1015.295237	1126.295	1136.859

Other criteria used: NEC, MIRC, similar results

	λ_1	λ_2	λ_3	λ_4	λ_{CF}	λ_{DS}	p
Component	K=5						
1	0.207	0.295	1.507	8.431	4.639	0.000	0.088
2	0.427	0.279	1.093	1.347	0.031	1.955	0.575
3	0.908	0.441	0.555	0.000	1.030	0.977	0.216
4	2.000	0.792	4.292	0.000	0.524	1.187	0.062
5	4.668	0.000	1.223	3.166	1.161	0.702	0.059
	K=6						
1	0.205	0.171	1.523	6.116	0.000	2.061	0.066
2	0.356	0.000	2.063	8.698	10.331	0.000	0.019
3	0.424	0.311	1.061	1.275	0.026	2.083	0.578
4	0.897	0.425	0.587	0.000	1.047	0.972	0.215
5	1.975	0.776	4.287	0.000	0.521	1.192	0.062
6	4.684	0.000	1.219	3.040	1.085	0.782	0.059

Bivariate Poisson probability function with parameter vector 0,4,1

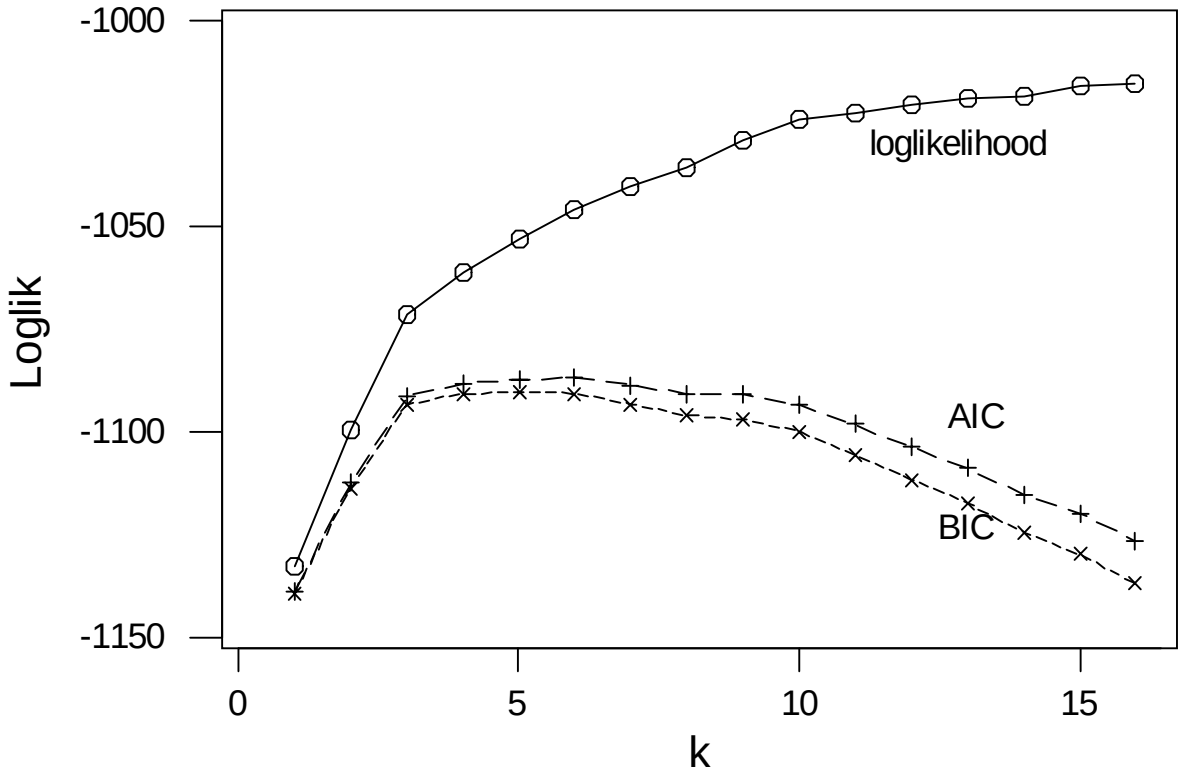
	0	1	2	3	4	5	6	7	8
0	0.00674	0.02695	0.05390	0.07187	0.07187	0.05750	0.03833	0.02190	0.01095
1	0.00000	0.00674	0.02695	0.05390	0.07187	0.07187	0.05750	0.03833	0.02190
2	0.00000	0.00000	0.00337	0.01348	0.02695	0.03594	0.03594	0.02875	0.01917
3	0.00000	0.00000	0.00000	0.00112	0.00449	0.00898	0.01198	0.01198	0.00958
4	0.00000	0.00000	0.00000	0.00000	0.00028	0.00112	0.00225	0.00299	0.00299
5	0.00000	0.00000	0.00000	0.00000	0.00000	0.00006	0.00022	0.00045	0.00060
6	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00004	0.00007
7	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001
8	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Probability function of a trivariate Poisson distribution
with full covariance structure

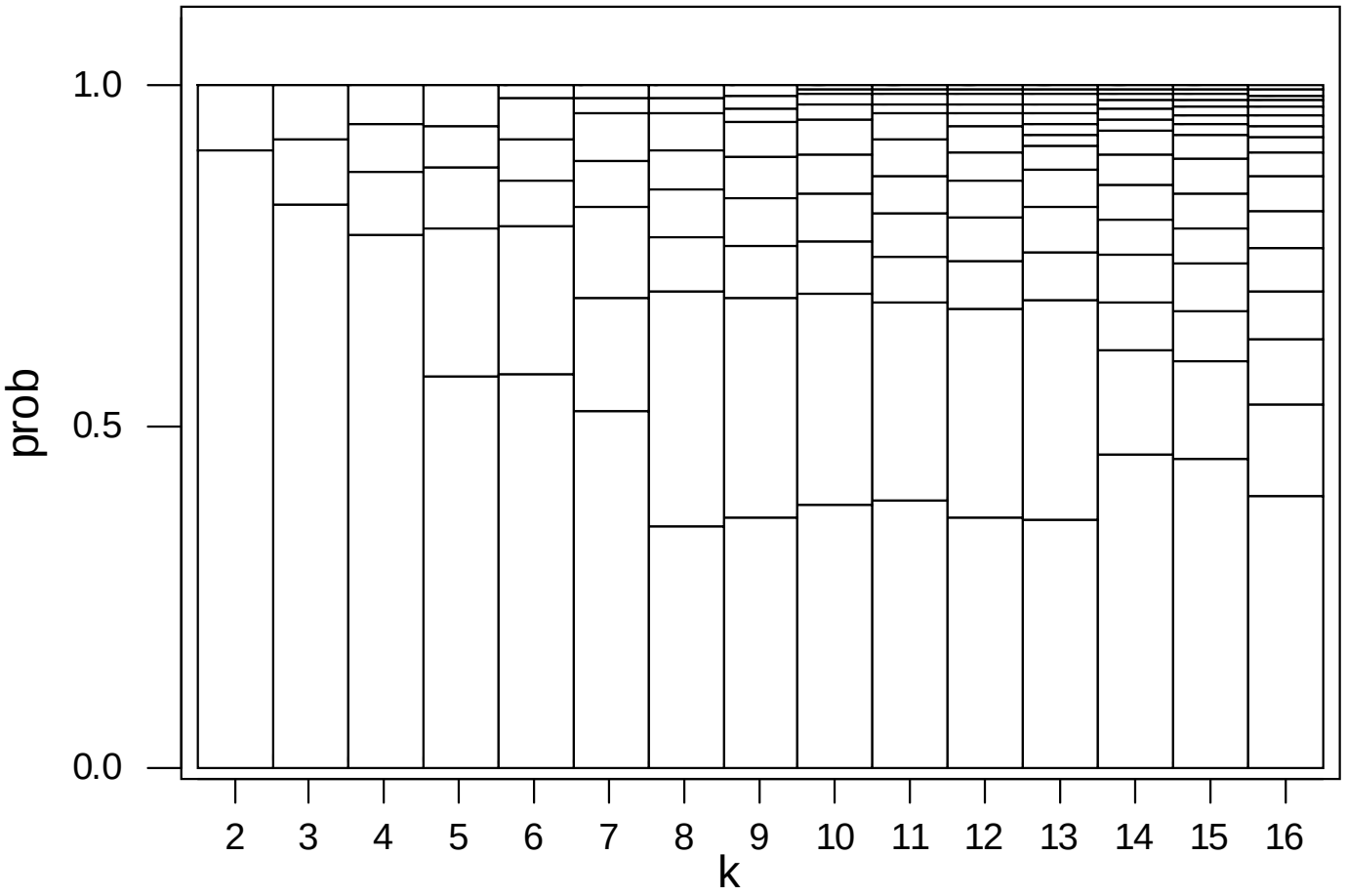
$$P(y_1, y_2, y_3) = \sum_{x_{12}=0}^{\min(y_1, y_2)} \sum_{x_{13}=0}^{\min(y_1 - x_{12}, y_3)} \sum_{x_{23}=0}^{\min(y_1 - x_{12}, y_3 - x_{13})} \exp\left(-\sum_{j \in A} \lambda_j \right) \times$$

$$\frac{\lambda_1^{\,y_1 - x_{12} - x_{13}}}{(y_1 - x_{12} - x_{13})!} \frac{\lambda_2^{\,y_2 - x_{12} - x_{23}}}{(y_2 - x_{12} - x_{23})!} \frac{\lambda_2^{\,y_3 - x_{13} - x_{23}}}{(y_3 - x_{13} - x_{23})!},$$

$$A = \{1, 2, 3, 12, 13, 14\}$$



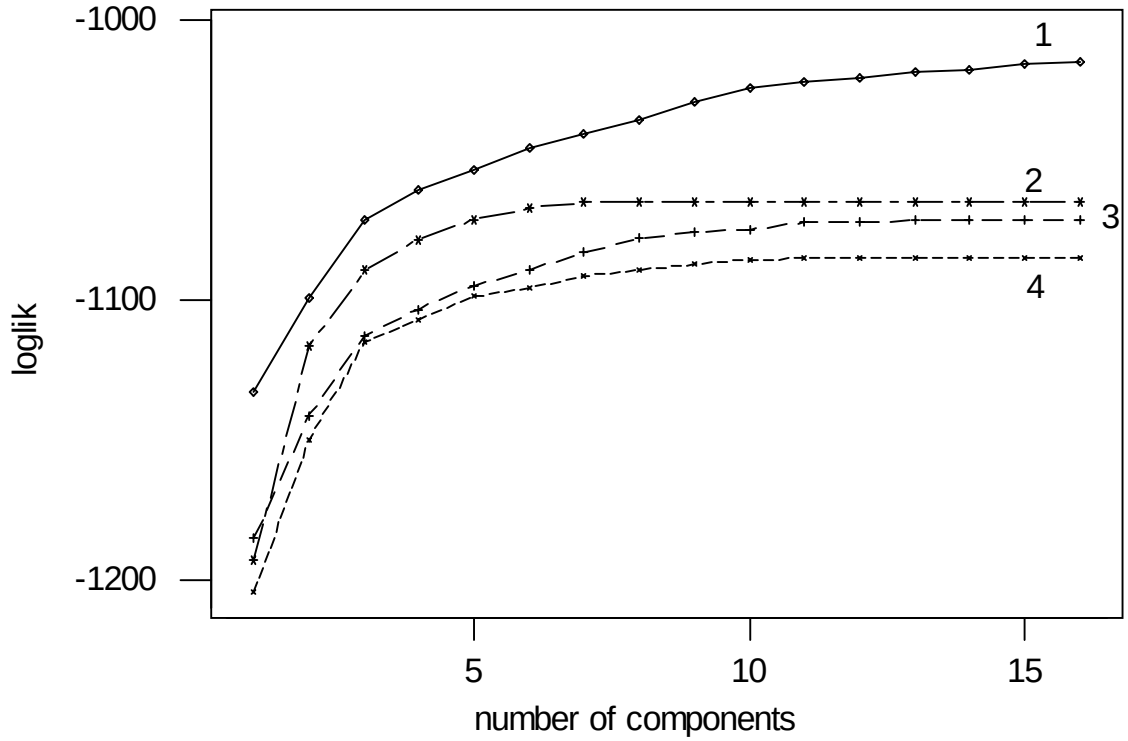
The Loglikelihood and the AIC and BIC criteria for our dataset



The mixing proportions for a wide range of models with varying number of clusters.

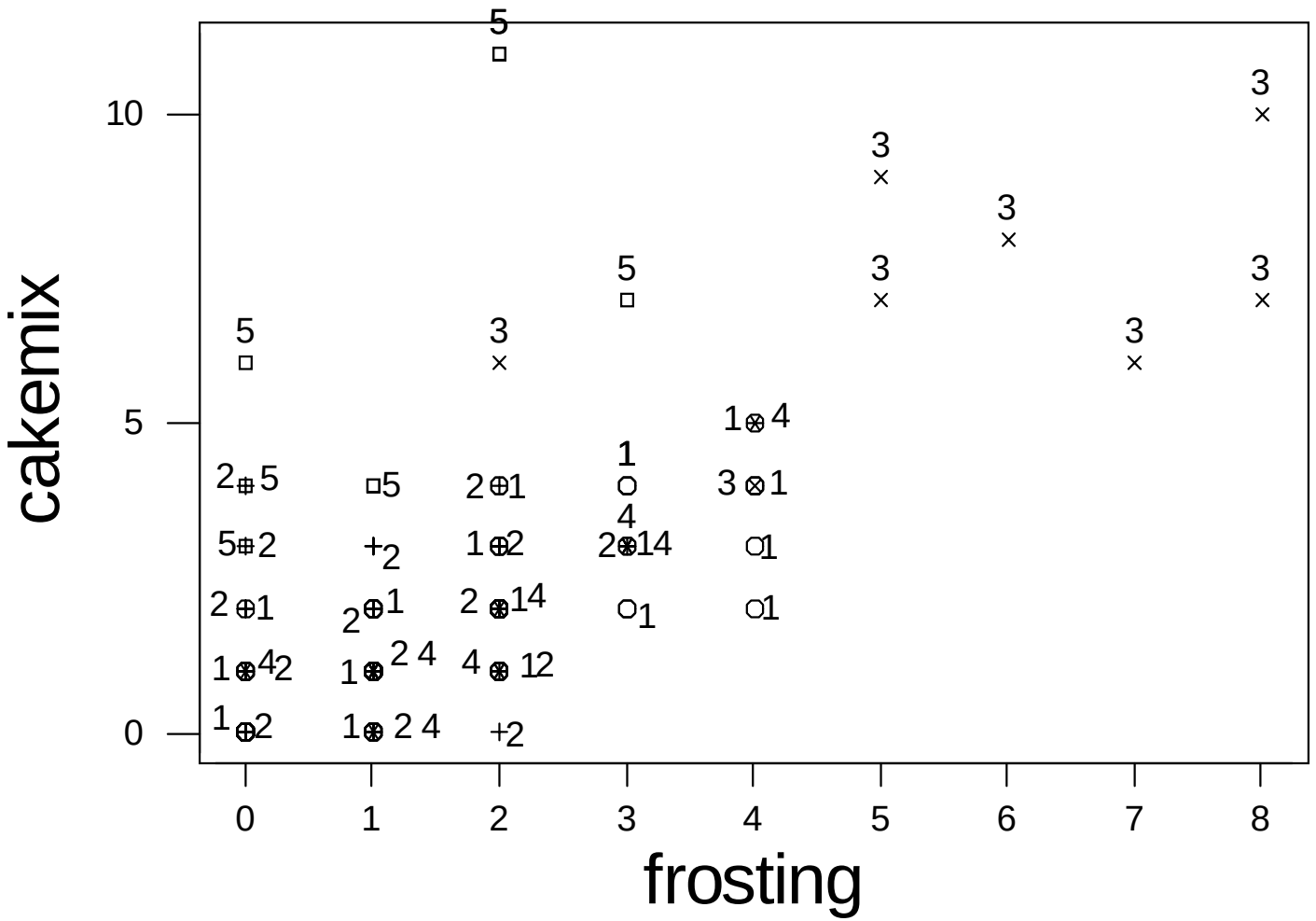
Cluster Centers

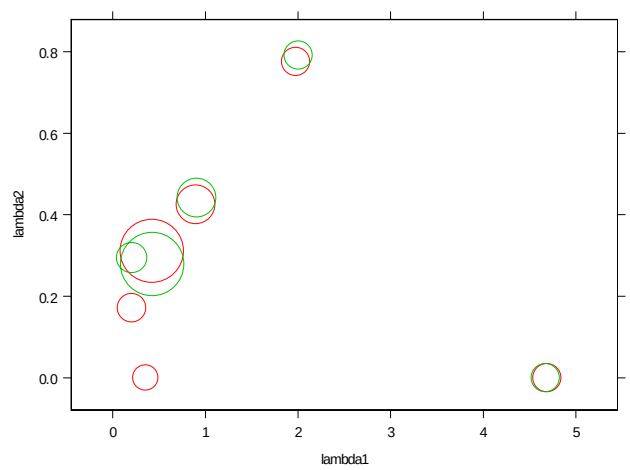
	Customers in the cluster	**cakemix**	**frosting**	**detergen**	**Softener**
1	93	1.5054	1.4194	3.2903	2.0215
2	34	1.6176	0.9706	0.9118	2.0000
3	8	7.1250	5.6250	1.0000	1.5000
4	12	1.6667	1.7500	9.2500	4.8333
5	8	6.2500	1.1250	4.1250	1.8750
	155	2.0774	1.5484	3.1548	2.2000
1	93	1.4516	1.3978	3.3548	2.1290
2	34	1.6176	0.9706	0.9118	2.0000
3	8	7.1250	5.6250	1.0000	1.5000
4	3	2.3333	2.0000	9.0000	10.3333
5	8	6.2500	1.1250	4.1250	1.8750
6	9	2.0000	1.8889	8.6667	1.8889
	155	2.0774	1.5484	3.1548	2.2000

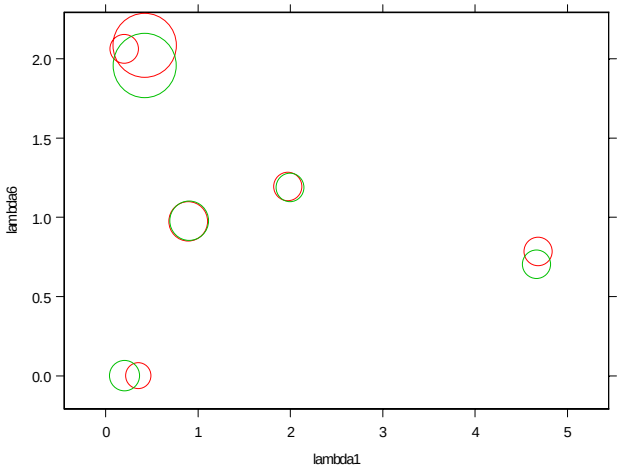


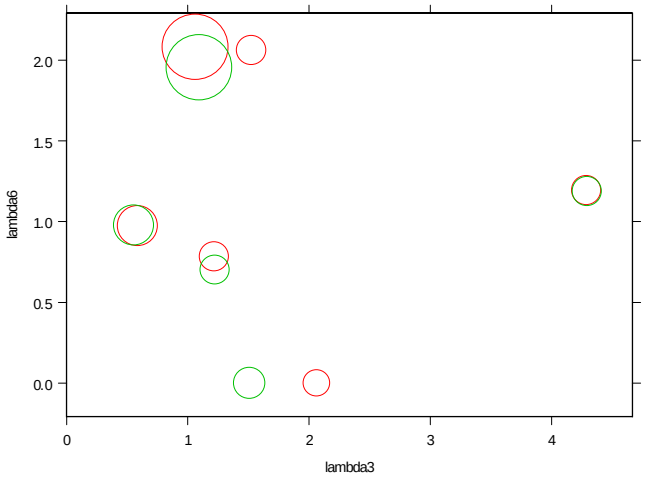
1- model described, 2-model with covariance between cakemix - detergent and softener -frosting

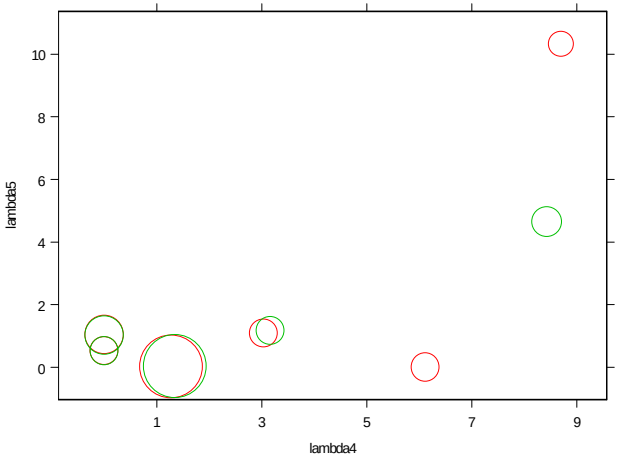
3- common covariance 4- no covariance model


cakemix
frosting
0
5
10
0
1
2
3
4
5
6
7
8


lambda2
lambda1
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5


lambda6
lambda1
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5


lambda6
lambda3
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4


lambda5
lambda4
0
2
4
6
8
10
1
3
5
7
9

QUALITY OF THE FITTED MODEL

Entropy criterion:

$$I(k) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\ln(w_{ij})}{n\ln(1/k)}$$

with the convention that $w_{ij}\ln(w_{ij}) = 0$ if $w_{ij}=0$.

Perfect classification → values near 1

Our solution **0.83**

Criterion can be used to select the number of components

AN INTERESTING CASE

Full covariance structure

$$A = [A_1, A_2] = \begin{bmatrix} 1 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & 1 & & 0 & 1 & 1 & & 0 \\ \vdots & & \ddots & & \vdots & 1 & \ddots & 1 \\ 0 & \ldots & & 1 & 0 & \ldots & & 1 \end{bmatrix}$$

Simple multivariate Poisson distribution (no mixture)

Parameters to be estimated

$$\begin{bmatrix} \lambda_1 & & & \\ \lambda_{12} & \lambda_2 & & \\ & & \ddots & \\ \lambda_{1m} & \lambda_{2m} & & \lambda_m \end{bmatrix}$$

- Off-diagonal elements are covariances

- Probability function too complicated
- Use of recursive relationships (problematic for 5 or more dimensions)

- Latent structure can be used to derive an EM
- EM does not need the calculation of the probability function

ONGOING RESEARCH

- EM algorithm for the general model
- Extension to the finite mixture model for clustering purposes
- Bayesian Approach: same data augmentation helps
- Generalization of the results in order to be applicable (and computationally feasible for large dimensions)