Tuning the Multivariate Poisson Mixture Model for Clustering Supermarket Shoppers

Tom Brijs[a] Dimitris Karlis[b] Gilbert Swinnen[a] Koen Vanhoof[a] Geert Wets[a]

[a]Department of Economics
Limburg University, Universitaire Campus
B-3590 Diepenbeek, BELGIUM
++3211 268757
tom.brijs@luc.ac.be

[b]Department of Statistics
Athens University of Economics
76 Patision, Str., 10434 Athens, GREECE
++3010 8203503
karlis@aueb.gr

ABSTRACT

This paper describes a multivariate Poisson mixture model for clustering supermarket shoppers based on their purchase frequency in a set of product categories. The multivariate nature of the model accounts for cross-selling effects that may exist between the purchases made in different product categories. However, because of computational difficulties, most multivariate approaches limit the covariance structure of the model by including just one common interaction term. Although this reduces the number of parameters significantly, it is often too simplistic in practice since typically multiple interactions exist on different levels. This paper proposes a theoretically correct variance/covariance structure of the multivariate Poisson model, based on domain knowledge or preliminary statistical analysis of significant purchase interaction effects in the data in order to produce the most parsimonious clustering model. As a result, the model does not contain more parameters than necessary whilst still accounting for the existing covariance in the data. From a practical point of view, the model can be used by retail category managers to devise customized merchandising strategies as illustrated in the text.

Categories and Subject Descriptors

[**Methods and Algorithms**]

Keywords

Mixture models, clustering, EM algorithm, multivariate Poisson.

1. INTRODUCTION

Today's competition forces consumer goods manufacturers and retailers to differentiate from their competitors by specializing and by offering goods/services that are tailored towards one or more subgroups or segments of the market. The retailer in the fast moving consumer goods (FMCG) sector is, however, highly limited in his ability to segment the market and to focus on the most promising segments, since the typical attraction area of the retail store is too small to afford neglecting a subgroup within the store's attraction area [7]. Nevertheless, given the huge amount of transactional and loyalty card data being collected by retail stores today, this leads to the intriguing question whether customer segmentation may provide a viable alternative to discover homogeneous customer segments for *in-store* segmentation, based on their shopping behavior. Indeed, scanner data reflect the frequency of purchases of products or product categories within the retail store and, as a result, they are extremely useful for modeling consumer purchase behavior. From a practical point of view, the discovery of different user groups will help the retail category manager to optimize his marketing mix towards different customer subgroups.

In this context, the use of mixture models (also called model based clustering) has recently gained increased attention as a statistically grounded approach to clustering [16, 18, 20]. More specifically, the *multivariate* Poisson mixture model will be introduced in this paper and it will be shown that the fully-parameterized model can be greatly simplified by preliminary statistical analysis of the existing purchase interactions in the transactional data, which will enable to remove free parameters from the variance/covariance matrix.

Cadez et al. [5] also used a mixture framework to model sales transaction data. However, their approach is different from ours in a number of respects. Firstly, they assume *individual* transactions to arise from a mixture of multinomials, whereas in our approach we model a customer's transaction history (i.e. the *aggregation* of the counts over the collection of transactions per individual) by means of a multivariate Poisson mixture model. The use of the Poisson distribution in this context is motivated in section 2. Secondly, the focus of their contribution is not really on finding groups of customers in the data, but rather on making predictions and profiling each individual customer's behavior. This is in contrast with our approach where the objective is to provide a methodology to discover groups of customers in the data having similar purchase rates in a number of product categories. Furthermore, our approach (being multivariate) explicitly accounts for interdependency effects between products which will lead to additional marketing insights as discussed in section 6 of this paper.

Another approach was taken by Ordonez et al. [19]. They used a mixture of Normal distributions to fit a sparse data set of binary vectors corresponding to the raw market baskets in a sales transaction database. Neither do they take correlations between product purchases into account by assuming diagonal covariance matrices.

Other examples include the use of the *univariate* Poisson mixture model [9] to find groups of customers with similar purchase rates of a particular candy product. The model identified light users (from 0 to 1.08 packs per week), heavy users (from 3.21 to 7.53 packs per week) and extremely heavy users (from 11.1 to 13.6 packs per week) of candy.

An example outside the marketing area includes the *bivariate* Poisson mixture model by Li et al. [15] for modeling outcomes of manufacturing processes producing numerous defect-free products. The mixture model was used to detect specific process equipment problems and to reduce multiple types of defects simultaneously.

To summarize, our model differs with existing approaches in basically two respects: 1) the use of the Poisson distribution instead of using Normal or multinomial distribution, 2) multivariate character of the model whereas most models treat the joint distribution as the product of the marginal univariate distributions.

This paper is structured as follows. Firstly, section 2 introduces the concept of model based clustering. Secondly, section 3 introduces the general formulation of the multivariate Poisson distribution. The reason is that the model proposed in this paper is a methodological enhancement to the more general multivariate Poisson model. Then, in section 4, model based clustering by using the multivariate Poisson mixture model is discussed together with its limitations. Subsequently, section 5 contains the core methodological contribution of this paper, i.e. the simplification of the general multivariate Poisson mixture model towards a more parsimonious formulation with restricted variance/covariance structure. The idea is implemented by using a concrete retail data example, which clearly illustrates the suggested simplification. Section 6 reports the results of the model and proposes a number of interesting strategic merchandising issues based on those results. Finally, section 7 is reserved for conclusions and topics for future research.

2. MODEL BASED CLUSTERING

Historically, cluster analysis has developed mainly through ad hoc methods based on empirical arguments. The last decade, however, there is an increased interest in model-based methodologies, which allow for clustering procedures based on statistical arguments and methodologies. The majority of such procedures are based on the multivariate normal distribution, see [3, 16] and others. The central idea of such models is the use of finite mixtures of multivariate normal distributions.

In general, in model-based clustering, the observed data are assumed to arise from a number of apriori unknown segments that are mixed in unknown proportions. The objective is then to ‘unmix’ the observations and to estimate the parameters of the underlying density distributions within each segment. The idea is that observations (in our case supermarket shoppers) belonging to the same class are similar with respect to the observed variables in the sense that their observed values are assumed to come from the same density distributions, whose parameters are unknown quantities to be estimated. The density distribution is used to estimate the probability of the observed values of the segmentation variable(s), conditional on knowing the mixture component from which those values were drawn.

The population of interest thus consists of k subpopulations and the density (or probability function) of the q-dimensional observation $\boldsymbol{y}$ from the j-th subpopulation is $f(\boldsymbol{y} \mid \theta_j)$ for some unknown vector of parameters θ_j. The interest lies on finding the values of the non-observable vector $\varphi = (\phi_1, \phi_2, \dots, \phi_n)$ which contains the cluster labels for each observation (1, …, n) and $\phi_i = j$ if the i-th observation belongs to the j-th subpopulation.

Since, we do not observe the cluster labels, the unconditional density of the vector $\boldsymbol{y}$ is a mixture density of the form

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

where $0 < p_j < 1$, and $\sum_{j=1}^{k} p_j = 1$ are the mixing proportions. Note that the mixing proportion is the probability that a randomly selected observation belongs to the j-th cluster.

This is the classical mixture model (see [4, 18]). The purpose of model-based clustering is to estimate the parameters $(p_1, \dots, p_{k-1}, \theta_1, \dots, \theta_k)$. Following the maximum likelihood (ML) estimation approach, this involves maximizing the loglikelihood

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

which is not easy since there is often not a closed-form solution for calculating these parameters. Fortunately, due to the finite mixture representation, an expectation-maximization (EM) algorithm is applicable.

The majority of model-based clustering is based on the multivariate normal distribution and hence it is based on the assumption of continuous data. If the data are not continuous, one can circumvent the problem by transforming the data to continuous data by using appropriate techniques, e.g. like correspondence analysis for categorical data. Such approaches, however, have serious limitations because useful information can be lost during transformation. Moreover, the normality assumption for each component may not be useful, especially for the case of count data with zeros, or when the normal approximation of the discrete underlined distribution is poor. For this reason, we will base our clustering on the multivariate Poisson distribution to allow for modeling the discrete nature of our data.

3. MULTIVARIATE POISSON DISTRIBUTION

Consider a vector $X = (X_1, X_2, \dots, X_m)$ where X_i‘s are independent and each follows a Poisson $Po(\lambda_i)$, $i = 1, \dots, m$ distribution. Usually, multivariate Poisson distributions (MP) are defined with Poisson marginals by multiplying the vector X with a matrix **A** of zeros and ones. Suppose that the matrix **A** has dimensions q x m, then the vector of random variables Y, defined as Y=**A**X, follows a multivariate Poisson distribution. The marginal distributions are simple Poisson distributions due to the properties of the Poisson distribution. In practice, the matrix **A** is structured so as to depict the covariances between the variables Y_i and Y_j. Such a structure assumes that **A** has the form

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \dots & A_m \end{bmatrix}$$

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. For instance, for q=3, we need

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

This construction of the matrix **A** has been used to define the multivariate Poisson distribution in its general form [11]. An implication of the above definition is that a multivariate Poisson distribution can be defined via a multivariate reduction technique. Suppose that q = 3 and slightly changing the notation to help the exposition, if one starts with independent Poisson random variables X_i , with means λ_i , $i \in$ S, with S = {1, 2, 3, 12, 13, 23, 123}, then the following representation is obtained for the Y_i's:

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

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

Interesting is the fact that X_{12} implies a covariance term between Y_1 and Y_2 whilst the term X_{123} implies a 3-fold covariance term. Furthermore, it is important to recognize that the λ_i 's have a similar interpretation. Indeed, it can be seen that the mean vector and the covariance matrix of the vector Y is given as:

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

where

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

and Σ is the variance/covariance matrix of X and is given as:

$$\Sigma = Var(X) = diag(\lambda_1 , \lambda_2 , \dots, \lambda_m)$$

This brings out the idea to create multivariate distributions with chosen covariances, i.e. not to include all the possible covariance terms but only to select covariance terms that are useful. Indeed, using all the m-fold covariance terms imposes too much structure while complicating the whole procedure without adding any further insight into the data. For this reason, after a preliminary examination, one may identify interesting covariance terms that may be included into the model and to exclude the others. This corresponds to fix the value of the Poisson parameter, i.e. the corresponding λ's.

The multivariate Poisson (MP) distribution in its general form, i.e. with all the m-fold covariance terms included, is computationally quite complicated as it involves multiple summations. Perhaps, this difficulty in the calculation of the probability mass function has been the major obstacle in the use of the multivariate Poisson distribution in its most general form. Kano and Kawamura [12] described recursive schemes to reduce the computational burden, but the calculation remains computationally demanding for large dimensions.

The next section introduces the construction, the estimation and the limitations of the general multivariate Poisson mixture model.

4. MODEL BASED CLUSTERING USING MULTIVARIATE POISSON

Let θ denote the vector of parameters of the general multivariate Poisson distribution. Then it holds that $\theta = (\lambda_1 , \lambda_2 , \dots, \lambda_m)$. We denote the multivariate Poisson distribution with parameter vector θ as $MP(\theta)$ and its probability mass function by $f(\boldsymbol{y} \mid \theta)$.

Consider the case where there are k clusters in the data and each cluster has parameter vector θ_j , $j = 1,\dots k$ and an observation Y_i conditional on the j-th cluster follows a $MP(\theta_j)$ distribution. Then unconditionally, each Y_i follows a k-component multivariate Poisson mixture. To proceed, one has to estimate using ML the parameters of the distribution. This task can be difficult, especially for the general case of a multivariate Poisson distribution with full covariances. In fact, estimation of the parameters can be carried out using the EM algorithm for finite mixtures. The main problem is that one has to maximize the likelihood of several multivariate Poisson distributions, which is extremely cumbersome. In a recent paper, Karlis [13] described an EM algorithm based on the multivariate reduction derivation of the multivariate Poisson distribution. An extended version of this EM algorithm will be used later in section 5.2.

Another important feature is the following. Even if we start with independent Poisson distributions, i.e. without assuming any covariance term, the finite mixture of such distributions will lead to non-zero covariances among the variables. The covariance is induced by the mixing distribution, i.e. the probability distribution with positive probability p_j at the simplex θ_j. This validates the above comment about the lack of need for imposing so much structure. If there is some covariance at the simple multivariate Poisson distribution, then the unconditional covariance can be decomposed in two parts, one due to the intrinsic covariance from the multivariate Poisson distribution and one from the mixing distribution.

5. RESTRICTED COVARIANCE MODEL

5.1 The model

The main contribution of this paper lies within the insight that the variance/covariance structure of the multivariate Poisson mixture model can be greatly simplified by examining the interaction effects between the variables of interest. Indeed, the statistical technique of log-linear analysis [2] is very well suited to assess the statistical significance of all *k*-fold associations between categorical variables in a multi-way contingency table [10]. The log-linear model is one of the specialized cases of generalized linear models for Poisson-distributed data.

More specifically, for the retailing example under study, this means that non-significant purchase interaction effects (i.e. cross-selling or substitution effects) between products or product categories can be used to cancel out free parameters in the variance/covariance structure of the multivariate Poisson model. In fact, the data used in this study contains the purchase history of 155 households over a period of 26 weeks in 4 product categories, i.e. cake mix (C), cake frosting (F), fabric detergent (D) and fabric softener (S). Log-linear analysis of the frequencies of co-occurrence showed that there was a strong purchase relationship between the 2-fold interactions cake mix and cake frosting, and between fabric detergent and fabric softener, but other *k*-fold combinations (e.g. 3-fold interaction between cake mix, cake frosting and fabric detergent) did not show to be statistically significant. The significance of the two-fold interactions is also illustrated by scatter matrix shown in figure 1.

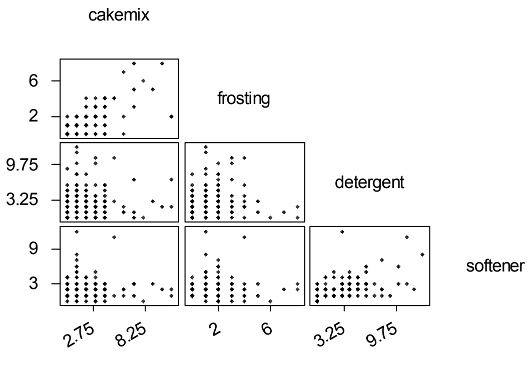


Figure 1: Scatter matrix for the 4 products

These scatter plots show that there exists a strong positive relation between cake mix and cake frosting, and fabric softener and fabric detergent, but not between other combinations of these products. This is further validated by calculating the sample Pearson correlation coefficients for the two-way product combinations. Only two correlations are significantly bigger than zero, i.e., *r*(mix, frosting) = 0.66, and *r*(detergent, softener) = 0.48, where *r*(A,B) denotes the Pearson correlation between the variables A and B.

Therefore, we make use of the latent variables $X = (X_C, X_F, X_D, X_S, X_{CF}, X_{DS})$ i.e. we use only two covariance terms and, for example, the term X_{DS} is the covariance term between detergent and softener. The interpretation of the parameters λ_{ij} are similar. The form of the matrix **A** is now:

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

and the vector of parameters is now $\theta = (\lambda_C, \lambda_F, \lambda_D, \lambda_S, \lambda_{CF}, \lambda_{DS})$. Thus we have

$$\begin{aligned} Y_C &= X_C + X_{CF} \\ Y_F &= X_F + X_{CF} \\ Y_D &= X_D + X_{DS} \\ Y_S &= X_S + X_{DS} \end{aligned}$$

Our definition of the model, in fact, assumes that the conditional probability function is the product of two bivariate Poisson distributions [14], one bivariate Poisson for cake mix and cake frosting, and another bivariate Poisson for fabric detergent and fabric softener. In general, we denote the probability mass function of the bivariate Poisson (BP) distribution as $BP(y_1, y_2; \lambda_1, \lambda_2, \lambda_{12})$, where λ_1, λ_2, λ_{12} are the parameters and the probability mass function is given as

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

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

Thus the conditional probability function of an observation $Y = (Y_C, Y_F, Y_D, Y_S)$ is given as

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

Thus the unconditional probability mass function is given under a mixture with *k*-components model by

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

As previously mentioned, the model assumes covariance between all the variables; the covariance is imposed by the mixing distribution. In addition, variables Y_C and Y_F and Y_D and Y_S have increased covariance due to their intrinsic covariance induced by our model.

The major issue of the above-defined model is how one can estimate the parameters of the model. For a model with k components the number of parameters equals $7k$-1. The likelihood function is quite complicated for direct maximization. Therefore, an EM type of algorithm is used. The algorithm is described in the next section.

5.2 Estimation: The EM Algorithm

The EM algorithm is a popular algorithm for ML estimation in statistics (see, e.g. Dempster *et al* [8] and McLachlan and Krishnan [17]). It is applicable to problems with missing values or problems that can be seen as containing missing values. Suppose that we observe data Y_{obs} and that there are unobservable/missing data Y_{mis}, that are perhaps missing values or even non-observable latent variables that we would like to be able to observe. The idea is to augment the observed and the unobserved data, taking the complete data $Y_{com} = (Y_{obs}, Y_{mis})$. The key idea is to iterate between two steps, the first step, the E-step, estimates the missing data using the observed data and the current values of the parameters, whilst the second step, the M-step, maximizes the complete data likelihood.

In our case, consider the multivariate reduction proposed above. The observed data are the q-dimensional vectors $Y_i = (Y_{Ci}, Y_{Fi}, Y_{Di}, Y_{Si})$ whereas the unobservable data are the vectors $X_i = (X_{Ci}, X_{Fi}, X_{Di}, X_{Si}, X_{CFi}, X_{DSi})$ that correspond to the original variables that led to the multivariate Poisson model plus the vectors $Z_i = (Z_{1i}, Z_{2i}, \ldots, Z_{ki})$ that correspond to the component memberships with $Z_{ji} = 1$ if the i-th observation belongs to the j-th component, and 0 otherwise. Thus, the complete data are the vectors (Y_i, X_i, Z_i). If we denote with φ the vector of parameters, then at the E-step of the algorithm, one has to calculate the conditional expectations $E(X_{ji} \mid Y_i, \phi)$ for $i=1,\ldots n$, $j \in \{C,F,D,S,CF,DS\}$ and $E(Z_{ji} \mid Y_i, \phi)$ for $i=1,\ldots n$, $j=1,\ldots k$. Note that, given the conditional expectations of the E-step, updating of the parameters is straightforward, as it is simply ML estimation of Poisson parameters. More formally, the procedure can be described as follows:

E-step*:* Using the current values of parameters calculate

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

$$x_{CFi} = \sum_{j=1}^{k} p_j BP(y_{Di}, y_{Si}; \lambda_{Dj}, \lambda_{Sj}, \lambda_{DSj}) \quad \times \frac{\sum_{r=0}^{\min(y_C, y_F)} r Po(y_{Ci} - r \mid \lambda_{Cj}) Po(y_{Fi} - r \mid \lambda_{Fj}) Po(r \mid \lambda_{CFj})}{P(y_i)},$$

$$x_{DSi} = \sum_{j=1}^{k} p_j BP(y_{Ci}, y_{Fi}; \lambda_{Cj}, \lambda_{Fj}, \lambda_{CFj}) \quad \times \frac{\sum_{r=0}^{\min(y_D, y_S)} r Po(y_{Di} - r \mid \lambda_{Dj}) Po(y_{Si} - r \mid \lambda_{Sj}) Po(r \mid \lambda_{DSj})}{P(y_i)},$$

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

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 criterion is satisfied, stop iterating, otherwise go back to the *E-step*.

The similarities with the standard EM algorithm for finite mixtures is obvious. The quantities w_{ij} at the termination of the algorithm are the posterior probabilities that the i-th observation belongs to the j-th cluster and thus they can be used to assign observations to the cluster with higher posterior probability.

Scalability: An important feature of our model is that we may use frequency tables to simplify the calculations. However, the description of the EM algorithm above is given without using frequencies. Indeed, the discrete nature of the data allows for using frequency tables instead of raw data. As a result, the sample size is not at all important for the computing time, since the original data are collapsed to frequency tables. Consequently, our clustering model is scalable to very large databases. In fact, even with a very large database, the clustering is done without any additional effort. This is of particular interest in real applications, where usually the amount of observations is large.

In general, in order to examine the scalability of our algorithm, two issues should be taken into account, the dimensions of the problem and the covariance structure being considered. In fact, it is well known that the speed of the EM algorithm depends on the 'missing' information. One could measure the missing information as the ratio of the observed information to the missing information. Our specification of the model has only two latent variables. Adding more latent variables lead to more 'missing' information and thus adds more computing time.

The same is true as far as the number of dimensions is concerned. More dimensions lead to more latent variables. If the structure is not more complicated, the algorithm will perform relatively the same, but if the structure is more complicated, then we expect more effort. This is a strong indication that the structure imposed before the fit of the model must remain in moderate levels. However, for a category manager, who is usually responsible for a limited number of product categories, this will pose no practical problems.

Finally, it is worth mentioning that if the model is expanded to a larger number of product categories, the discovery of significant purchase interactions to include in the variance-covariance structure involves the log-linear analysis of a larger collection of contingency tables. However, the cell frequencies in each contingency table can be calculated quite easily from the frequent itemsets in association rule mining [1]. Indeed, it can be shown that the cells in a multi-way contingency table, containing a particular combination of items, can be calculated from their corresponding frequent itemsets. Thus, the specification of the variance-covariance structure for larger model specifications is not really more complicated than for smaller models.

6. EMPIRICAL STUDY

The EM algorithm was implemented sequentially for 1 to 16 components (k = 1,...,16) because the loglikelihood stopped increasing after k>16 (see figure 2). A well-documented drawback of the EM is its dependence on the initial values. In order to overcome this problem, 10 different sets of starting values were chosen at random. In fact, the mixing proportions (p) were uniform random numbers and rescaled so as to sum at 1, while the λ's were generated from a uniform distribution on the range of the data points. For each set of starting values, the algorithm was run for 100 iterations without caring about any convergence criterion. Then, from the solution with the largest likelihood, EM iterations were continued until a rather strict convergence criterion was satisfied, i.e. until the relative change of the loglikelihood between two successive iterations was smaller than 10^{-12}. This procedure was repeated 10 times for each value of k. As expected, problems with multiple maxima occurred for large values of k, while for smaller values of k the algorithm usually terminated at the same solution with small perturbations.

There is a wealth of procedures for selecting the number of clusters (or equivalently the number of components) in a mixture [18]. The majority of them have been proposed for the case of clustering via multivariate normal mixtures. In this paper, we based our selection on the Akaike Information Criterion (AIC) given as:

$$AIC = -2L_k + 2\, d_k$$

where L_k is the value of the maximized loglikelihood for a model with k components and d_k is the number of parameters for this model. In our case $d_k = 7k - 1$. Other criteria could have also been used.

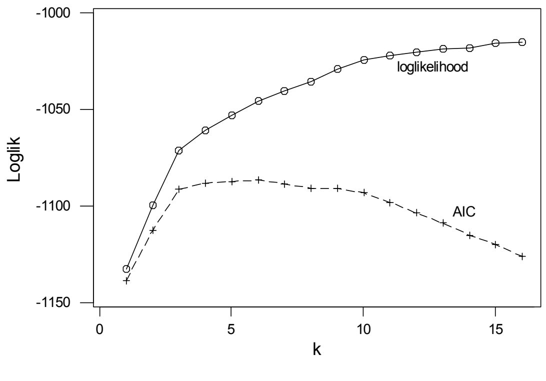


Figure 2: Loglikelihood and AIC (rescaled) against the number of components for the restricted MVP mixture model

Figure 2 shows that the AIC criterion selects 6 components as the optimal number of components. The depicted values are rescaled so as to be comparable to the loglikelihood. However, it is interesting to note that the AIC for the 5 component solution is extremely close to the 6 component solution and thus for reasons of parsimony, the 5 component solution could be selected. Furthermore, as we will describe in the sequel, the differences between the 5 and 6 components solutions are minor, hence, it seems plausible to choose the smaller model with 5 components.

Figure 3 shows the values of the mixing proportions for the entire range of models used (values of k ranging from 2 to 16). It is apparent from the graph that usually the additional component corresponds to a split of an existing component in two parts, perhaps with some minor modification for the rest components, especially if they have estimates close to the component split. This illustrates the stability of the model and the existence of two larger components, which together cover almost 80% of all observations (see also table 1 and 2).

It is also quite interesting to see that the solution with 5 and 6 components differ slightly. This is extremely interesting from the retailer point of view for which the existence of a limited number of clusters is very important. Indeed, if a large number of clusters would exist, it is impossible for the retailer to manage all segments separately, i.e. it would neither be cost-effective, nor practical to set-up different merchandising strategies for each (small) segment.

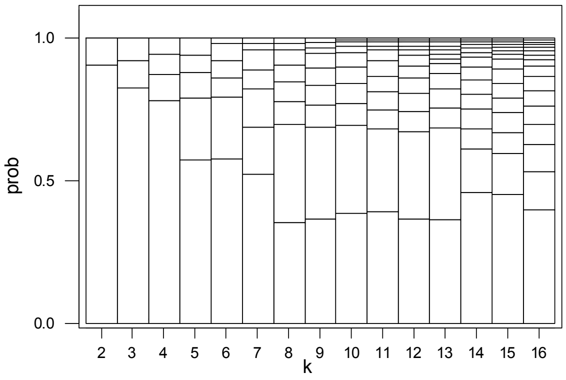
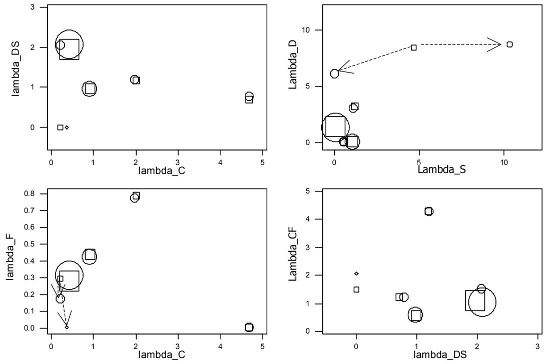


Figure 3: The mixing proportions for model solutions with k=2 to 16 components



□ = 5-component solution

○ = 6-component solution

Figure 4. Bubble plots for selected pairs of parameters.

Figure 4 on the previous page shows selected bubble-plots for pairs of the parameters. In fact, each graph depicts the joint mixing distribution for the selected pair. The plots depict both the 5 and the 6 cluster solution. The circles represent the 6-cluster solution and the squares the 5 cluster solution. The size of the circle/square reflects the mixing proportion, the larger the size the larger the mixing proportion. It is clear from the graph that the two solutions differ only slightly and that the 6 cluster solution just splits up one of the existing clusters as indicated by the arrows on figure 4.

Table 1 and table 2 contain the parameter estimates for the model with 5 components and 6 components respectively. One can see that all the components of the 5-cluster solution still exist in the 6-cluster solution, but an additional component appeared (number 2 in the 6-cluster solution) that takes observations from the old components 1 and 3 of the 5-cluster solution. In both solutions, there are 2 clusters of large size that are very similar, indicating the existence of two relatively stable clusters, which together account for almost 80% of all the customers.

	parameters						
cluster	λ_C	λ_F	λ_{CF}	λ_D	λ_S	λ_{DS}	p
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

Table 1: Estimated parameters for the 5-components model

	parameters						
cluster	λ_C	λ_F	λ_{CF}	λ_D	λ_S	λ_{DS}	p
1	0.205	0.171	1.523	6.116	0.000	2.061	0.066
2	0.356	0.000	2.063	8.698	10.33	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

Table 2: Estimated parameters for the 6-components model

The rule for allocating observation to clusters is the higher posterior probability rule. An observation is allocated to the cluster for which the posterior probability, as measured by w_{ij}, is larger. Note that w_{ij} are readily available after the termination of the EM algorithm, as they constitute the E-step of the algorithm. Careful examination of the results about the cluster that each customer belongs to reveals that there are only 10 customers that changed cluster between the two solutions, 8 of them switched to the new cluster.

In order to assess the quality of clustering we calculated the entropy criterion [18] as:

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

with the convention that $w_{ij}\ln(w_{ij}) = 0$ if w_{ij}=0. In the case of a perfect classification, for each i there is only one $w_{ij}=1$ and all the rest are 0. This implies a value for the criterion equal to 1.

On the contrary, for the worst case clustering the value of the criterion is 0. Thus, values near 1 show a good clustering. For our data, we found that *I(5)*=0.81 and *I(6)*=0.78, indicating a very good separation between the clusters.

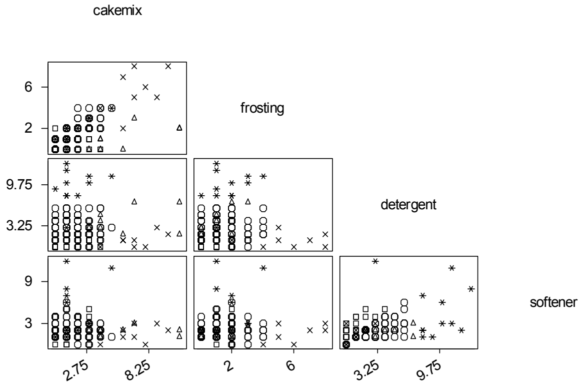


Figure 5: Pairwise clusters

Figure 5 depicts the clusters for the 5-clusters solution. All the pairs of variables are included and the different clusters are indicated by different symbols. Note that pair wise customers with the same values for some pair belong to different clusters due to their values for the rest variables. However some clusters can be identified by particular pairs, like the cluster indicated by '*' for the pair softener and detergent.

An interesting feature of the results in table 2 is the interpretation of the zero values. If the zero value corresponds to a covariance parameters (i.e. λ_{CF} , λ_{DS}) then this implies that the two variables are not correlated at all for this component, i.e. the purchase rate of a product is independent from the purchase rate of the other product. The interpretation of a zero value for the other lambdas is a little more complicated. A zero value leads to high correlation between the two variables, because the value for this product is usually very similar to the values for the other product.

cluster	*Cakemix*	*Frosting*	*Detergent*	*Softener*	*Obs.*
1	1.667	1.750	9.250	4.833	12
2	1.505	1.419	3.290	2.022	93
3	1.618	0.971	0.912	2.000	34
4	7.125	5.625	1.000	1.500	8
5	6.250	1.125	4.125	1.875	8
Overall mean	2.077	1.548	3.155	2.200	155

Table 3: Cluster centers for the 5-component mixture model

In order to interpret the cluster differences with regard to the original data, table 3 contains the cluster centers for the 5-components solution. The last row contains the sample centers.

Looking at the two major clusters (cluster 2 and 3) in table 1, it can be observed that they have a rather different profile. Especially with regard to fabric detergent and fabric softener, both clusters show indeed a rather different behavior.

Cluster 2 shows a very low average purchase rate of fabric softener (λ_S = 0.031) but a rather high covariance between fabric detergent and fabric softener (λ_{DS} = 1.955). This means that, for this cluster, the purchases of fabric softener are largely due to cross-selling with fabric detergent.

This is shown in table 3: people in cluster 2 have an average purchase rate of fabric softener of 2.022, which is largely due to the covariance with fabric detergent (λ_{DS} = 1.955). Consequently, sales of fabric softener in cluster 2 are almost non-existent and if they occur, they have occurred as a result of cross-selling with fabric detergent.

In contrast, cluster 3 shows a somewhat opposite profile. Cluster 3 shows no purchases of fabric detergent at all (λ_D = 0.000), but again a relatively strong covariance with fabric softener (λ_{DS} = 0.977). This means that, for this cluster, the purchases of fabric detergent are exclusively due to cross-selling with fabric softener. This is again shown in table 3: people in cluster 3 have an average purchase rate of fabric detergent of 0.912, which is rather low compared to the total sample average, but this purchase rate is exclusively due to the covariance with fabric softener (λ_{DS} = 0.977). Consequently, it can be concluded that sales of fabric detergent on its own in cluster 3 are non-existing and if they occur, they have occurred exclusively as a result of cross-selling with fabric softener.

These are important findings since they have interesting implications for marketing decision making, e.g. for targeted merchandising strategies. For instance, if it is known that in cluster 3 detergent is purchased only if one also purchases softener, then softener could be positioned as a loss-leader brand in this segment. It means that by reducing the price of a leader brand like of softener (even making a loss on softener) customers will not only buy more softener, but they will also buy more detergent, which is highly dependent on softener (as shown in our analysis). This way, an overall positive profit can be achieved because the high margin detergent purchase will compensate for the low margin softener purchase. As such, the softener acts as a decoy, i.e., the softener brand is positioned in an attractive way in order to encourage people to purchase detergent too.

Moreover, it is a common practice in retailing to allocate special display space (e.g. at the end-of-aisle) to some products in the store. This is done to stimulate the sales of the product on display. For instance, by putting softener on special display, more people will be encouraged to purchase softener. But, since people who buy softener also buy detergent, some of them in this segment will actually purchase detergent too. As a result, by putting softener on special display, sales of both products are boosted.

Similarly, coupon programs can be designed to encourage consumers to increase the number of categories considered on a shopping trip. In this respect, knowledge about correlated category usage patterns enables category managers to implement cross-category marketing strategies. For instance, Catalina Marketing [6] sells point-of-purchase electronic couponing systems that can be implemented to print coupons for a particular category, based on the purchases made in other categories. For example, cluster 2 consumers in table 1 could be given detergent coupons, not only to stimulate sales of detergent, but to stimulate sales of softener too, given that the sales of softener are dependent on detergent sales.

Another interesting merchandising strategy would be to put highly interdependent products closer together in the store to please the shopping behavior of so-called ‘run-shoppers’ who typically don’t want to waste time by looking for items in the store.

Finally, cluster 1 in the 5-component solution shows another interesting, yet different profile compared to the two bigger clusters discussed before. Customers in this cluster purchase large quantities of fabric detergent (λ_D = 8.431) and fabric softener (λ_S = 4.639), however, the purchases are not correlated at all (λ_{DS} = 0.000). This means that although customers in cluster 1 purchase high amounts of detergent and softener, their values are not the result of cross-selling between both products. Consequently, whatever promotional campaign (like price reduction or special display) on one of both products (say detergent), it will not influence the sales of the other product (softener). An explanation could be that customers in cluster 1 are deal-prone customers, and thus will only purchase a product when it is on promotion or when it is out of stock at home.

7. CONCLUSIONS AND FUTURE RESEARCH

7.1 Conclusions

In this paper, a multivariate Poisson mixture model was introduced to cluster supermarket shoppers based on their purchase rates in a number of predefined product categories. However, instead of using the classical definition of the multivariate Poisson distribution, i.e. with a fully-saturated variance/covariance matrix, we showed that the number of free parameters can be reduced significantly by preliminary examination of the interdependencies between the product category purchases. Knowledge about these interactions can be obtained in different ways. For instance, via a data mining approach where product interdependencies can be discovered by means of association rule analysis, or via statistical analysis of the multi-way contingency table containing the frequencies of product category purchase co-occurrences. The result is a much more parsimonious version of the multivariate Poisson mixture model that is easier and faster to estimate whilst it still accounts for the existing covariances in the underlying data. Furthermore, an EM algorithm was presented to estimate the parameters of the model.

The model was tested on a real supermarket dataset including the purchases of 155 households over a period of 26 weeks in 4 product categories. The results of the model indicated that two big clusters, accounting for almost 80% of the observations, could be found with a distinct purchasing profile in terms of the purchase rates and purchase interactions between the product categories considered. Moreover, a number of concrete merchandising strategies were proposed for the discovered clusters based on this purchasing profile.

7.2 Limitations

The model, however, also has some limitations.

First of all, neither pricing, nor promotional elasticity data were available to this study. As a result, the suggested merchandising strategies should be adopted with care. Indeed, the price and promotion sensitivity of consumers is not known so that results of merchandising strategies, based on knowledge about purchase rates, cannot be predicted with great certainty.

Secondly, the segment specific purchase rates are treated as static parameters in the model, whereas in practice, they can probably change over time. This could result in customers switching from one cluster to another, or entire clusters to change profile over time, i.e. moving from one position to another. This dynamic aspect has not been accounted for in the study.

Finally, although it was shown in this paper how to significantly reduce the complexity of the multivariate Poisson mixture model, and how the model scales towards more observations (e.g. more customers or a longer transaction history), the scalability towards including more product categories requires more empirical study.

7.3 Future Research

Future research topics include the evaluation of Markov Chain Monte Carlo (MCMC) methods to estimate the parameters for larger versions of the multivariate Poisson mixture model. Moreover, EM algorithms for the general multivariate Poisson mixture model discussed in section 4 are of interest. Lastly, issues related to our model will be addressed too. Such an example is the construction of efficient recursive schemes for the evaluation of the probability mass function of a multivariate Poisson model. Such schemes can speed up the estimation of our model considerably.

8. ACKNOWLEDGEMENT

The authors want to thank Prof. Puneet Manchanda from the Graduate School of Business at the University of Chicago for his data support.

9. REFERENCES

[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. (1996), Fast discovery of association rules, in: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining: AAAI Press, pp. 307-328.

[2] Agresti, A., An introduction to categorical data analysis, Wiley Series in Probability and Statistics, 1996.

[3] Banfield, J.D., and Raftery, A.E., Model-based Gaussian and non-Gaussian clustering, in: Biometrics, Vol. 49, pp. 803-821.

[4] Böhning, D., Computer assisted analysis of mixtures & applications in meta-analysis, disease mapping & others, CRC press, 1999.

[5] Cadez, I.V., Smyth, P., and Mannila, H. Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco (CA), USA, pp. 37-46, 2001.

[6] Catalina Marketing, Catalina CheckOut Coupon ®, *Catalina Marketing Website*, www.catalinamarketing.com.

[7] Corstjens, M.L., and Corstjens, J., Store Wars: The battle for mindspace and shelfspace. Wiley, 1995.

[8] Dempster, A.P., Laird, N.M., and Rubin, D., Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B 39, 1-38, 1977.

[9] Dillon, W.R., and Kumar, A., Latent structure and other mixture models in marketing: an integrative survey and overview, in: Advanced Methods in Marketing Research, Richard P. Bagozzi (ed.), Cambridge, MA: Blackwell, pp. 295-351, 1994.

[10] DuMouchel, W., Pregibon, D., Empirical bayes screening for multi-item associations, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco (CA), USA, pp. 67-76, 2001.

[11] Johnson, N., Kotz, S., and Balakrishnan, N., Discrete multivariate distributions. Wiley, 1997.

[12] Kano, K., and Kawamura, K., On recurrence relations for the probability function of multivariate generalized Poisson distribution, in: Communications in Statistics – Theory and Methods, Vol. 20, pp. 165-178, 1991.

[13] Karlis, D. An EM algorithm for multivariate Poisson distribution and related models. *(submitted)*, 2001

[14] Kocherlakota, S., and Kocherlakota, K., Bivariate discrete distributions, New York: Marcel Dekker, 1992.

[15] Li, C., Lu, J., Park, J., Kim, K., Brinkley P.A., and Peterson, J.P., Multivariate zero-inflated poisson models and their applications, in: Technometrics, Vol. 41(1), pp. 29-38, 1999.

[16] McLachlan, G.J., and Basford, K.E., Mixture models: inference and applications to clustering, New York: Marcel Dekker, 1988.

[17] McLachlan, G.J. and T. Krishnan, The EM algorithm and extensions. Wiley, New York, 1997.

[18] McLachlan, G., and Peel, D., Finite mixture models, Wiley series in probability and statistics, 2000.

[19] Ordonez, C., Omiecinski, E., and Ezquerra, N., A fast algorithm to cluster high dimensional basket data, in : Proceedings of the IEEE International Conference on Data Mining, San Jose (CA), USA, pp. 633-636, 2001.

[20] Wedel, M., and Kamakura, W.A., Market segmentation: conceptual and methodological foundations, Kluwer: Dordrecht, 1999.