Statistica Neerlandica (2004) Vol. 58, nr. 3, pp. 322–348

A multivariate Poisson mixture model for marketing applications

Tom Brijs*

Department of Economics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

Dimitris Karlis†

Department of Statistics, Athens University of Economics, 76 Patision, Str., 10434 Athens, Greece

Gilbert Swinnen, Koen Vanhoof, Geert Wets

Department of Economics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

Puneet Manchanda‡

Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637, USA

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 between the purchases made in different product categories. However, for computational reasons, most multivariate approaches limit the covariance structure by including just one common interaction term, or by not including any covariance at all. Although this reduces the number of parameters significantly, it is often too simplistic as typically multiple interactions exist on different levels. This paper proposes a theoretically more complete variance/covariance structure of the multivariate Poisson model, based on domain knowledge or preliminary statistical analysis of significant purchase interaction effects in the data. Consequently, the model does not contain more parameters than necessary, whilst still accounting for the existing covariance in the data. Practically, retail category managers can use the model to devise customized merchandising strategies.

Key Words and Phrases: mixture models, clustering, EM algorithm, multivariate Poisson, product purchasing.

*tom.brijs@luc.ac.be

†karlis@aueb.gr

‡puneet.manchanda@gsb.uchicago.edu

1 Introduction

Today's competition forces consumer goods manufacturers and retailers to differentiate themselves from their competitors by specializing and by offering goods/services that are tailored towards one or more subgroups or segments of the market. In this context, transactional market basket data (also known as scanner data) provide excellent opportunities for a retailer to segment the customer population into different groups based on differences in their purchasing 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. Furthermore, scanner data reflect the interdependencies that exist between purchases made in different product categories. In fact, previous research has shown that failing to consider these interdependencies may lead to marketing actions with disappointing results (see for instance, MANCHANDA *et al.*, 1999; AINSLIE and ROSSI, 1998; MULHERN and LEONE, 1991).

Among the existing clustering techniques, the use of mixture models (also called model based clustering) has recently gained increased attention as a statistically grounded approach to clustering (MCLACHLAN and BASFORD, 1988; MCLACHLAN and PEEL, 2000; WEDEL and KAMAKURA, 1999). In the retailing context, the kind of data that are often collected relating to the purchase behavior consists of counts which usually have a large number of zeroes. For example, HOOGENDOORN (1999) considered multivariate Poisson processes for purchase incidences, while MORRISON and SCHMITTLEIN (1988) discussed a multivariate negative binomial model for the same reason.

Given the kind of data available, the standard model based clustering procedures based on multivariate normal mixtures are not sound. Alternative models related to the discrete nature of the data are needed. In the present paper, we develop a multivariate Poisson model for this reason. It will be shown that the fully parameterized model can be greatly simplified by preliminary statistical analysis of the existing purchase interactions in the data, which will enable one to remove free parameters from the variance–covariance matrix.

CADEZ *et al.* (2001) also used a mixture of multinomials to model sales transaction data. However, they model binary purchases (yes/no) of a product, while we concentrate on the joint modeling of the number of purchases (counts) for a series of different product categories. Furthermore, our approach explicitly accounts for interdependency effects between product categories, which will lead to additional marketing insights as discussed in section 6 of this paper. Another approach was taken by ORDONEZ *et al.* (2001). 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. They do not take correlations between product purchases into account by assuming diagonal covariance matrices. Other examples of applications of model based clustering in marketing include the use of the univariate Poisson mixture model (DILLON and KUMAR, 1994) to find groups of customers with similar purchase rates of a particular candy product and

WEDEL *et al*. (1995) proposed a hazard model of brand switching using a finite mixture of Poisson distributions.

The content of this paper is structured as follows. First, section 2 introduces the concept of model based segmentation. 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 a restricted variance/covariance structure. Furthermore, our approach is also contrasted with two other well known approaches, i.e. the local independence model and the common covariance model. All three models are implemented by using a real retail data example, which clearly illustrates the suggested simplification. Section 6 reports the results of the models and proposes a number of interesting strategic merchandising issues based on these results. Finally, section 7 is reserved for conclusions and discussion.

2 Model based segmentation

Historically, cluster analysis has developed mainly through ad-hoc methods based on empirical arguments. In 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 for instance BANFIELD and RAFTERY, 1993; MCLACHLAN and BASFORD, 1988). 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 a priori 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 function within each segment. The idea is that observations (in our case supermarket shoppers) belonging to the same segment are similar with respect to the observed variables in the sense that their observed values are assumed to come from the same density functions, whose parameters are unknown quantities to be estimated. The density function 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 y from the j-th subpopulation is $f(y \mid \theta_j)$ for some unknown vector of parameters θ_j. Since we do not observe the cluster labels, the unconditional density of the vector 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 BÖHNING, 1999; MCLACHLAN and PEEL, 2000). The purpose of model based clustering is to estimate the parameters $(p_1,\ldots, p_{k-1}, \theta_1,\ldots, \theta_k)$. Following the maximum likelihood (ML) estimation approach, this involves maximizing the loglikelihood

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

which is not easy as there is often no 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, via appropriate techniques, e.g. 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 zeroes, or when the normal approximation of the discrete underlying 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 $X = (X_1, X_2,\ldots, X_m)$ where X_i's are independent and each follows a Poisson distribution with parameter λ_i, denoted hereafter as $Po(\lambda_i)$, $i = 1,\ldots, m$. Usually, multivariate Poisson distributions (MP) are defined with Poisson marginals by multiplying the vector X with a matrix $\mathbf{A}$ of zeroes and ones. Suppose that the matrix $\mathbf{A}$ has dimensions $q \times m$, then the vector of random variables Y, defined as $Y = \mathbf{A}X$, follows a multivariate Poisson distribution. The marginal distributions are simple Poisson distributions due to the properties of the Poisson distribution. In practice, the matrix $\mathbf{A}$ is structured so as to depict the covariances between the variables Y_i and Y_j. Such a structure assumes that $\mathbf{A}$ 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 $q - i$ zeroes. 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 (JOHNSON *et al.*, 1997). 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:

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

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

An interesting fact is that X_{12} implies a covariance term between Y_1 and Y_2 whilst the term X_{123} implies a three-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:

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

where

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

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

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

This brings out the idea of creating multivariate distributions with chosen covariances, i.e. not to include all the possible covariance terms but to select only covariance terms that are useful. Indeed, in some cases 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 of the data, one may identify interesting covariance terms that may be included in the model and to exclude the others. This corresponds to fixing the values of the Poisson parameters, i.e. the corresponding λ's equal to 0.

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 (1991) described recursive schemes to reduce the computational burden, but the calculation remains computationally demanding for large dimensions.

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

4 Model based clustering using multivariate Poisson

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

Consider the case where there are k clusters in the data and each cluster has parameter vector θ_j, $j = 1, \ldots, k$ 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 (2003) 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 5.1.2.

Another important feature is the following. Even if we start with independent Poisson distributions (see the local independence model in section 5.3), 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 into two parts: one due to the intrinsic covariance from the multivariate Poisson distribution and one from the mixing distribution (see, e.g. SUBRAHMANIAM, 1966).

For example, the model of AITCHINSON and HO (1989) starts from independent Poisson variates and the covariance is imposed by the multivariate lognormal mixing distribution. The novel idea in our model is that we start from variables that conditionally are not independent. Hence, the mixing distribution just adds to the existing dependence structure. In order to make the model parsimonious, we restrict the number of covariance terms (and thus the number of parameters to be estimated)

by setting some particular parameters (corresponding to covariances) equal to 0. This can be done either subjectively (by selecting some covariances based on prior knowledge or based on examination of the data, as in our case) or objectively by selecting terms using a model selection criterion.

5 Multivariate Poisson mixture models

In this section, different versions of the general multivariate Poisson mixture model are introduced depending on the structure of the variance–covariance matrix. We do not pursue the full model mainly because it imposes too much structure (i.e. too many covariance terms) to be practical. However, the ideas presented can be easily extended to cover more general cases, but this is not needed for the application presented in section 6.

5.1 *Restricted covariance model*

5.1.1 *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. For instance, log–linear analysis (AGRESTI, 1996) is very well suited to assess the statistical significance of all k -fold associations between categorical variables in a multi-way contingency table (DUMOUCHEL and PREGIBON, 2001). 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 are from a large metropolitan market area in the western United States and contain the purchase rates of 155 households over a period of 26 weeks in four product categories, i.e. cake mix (C), cake frosting (F), fabric detergent (D) and fabric softener (S). Figure 1 and Table 1 show the distribution of purchase rates, the mean, the variance and the variance to the mean ratio for each product category.

Log–linear analysis of the frequencies of co-occurrence showed that there was a strong purchase relationship between the two-fold interactions cake mix and cake frosting, and between fabric detergent and fabric softener, but other k-fold combinations (e.g. three-fold interaction between cake mix, cake frosting and fabric detergent) were not shown to be statistically significant. The significance of the two-fold interactions is also illustrated by the scatter matrix shown in Figure 2.

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

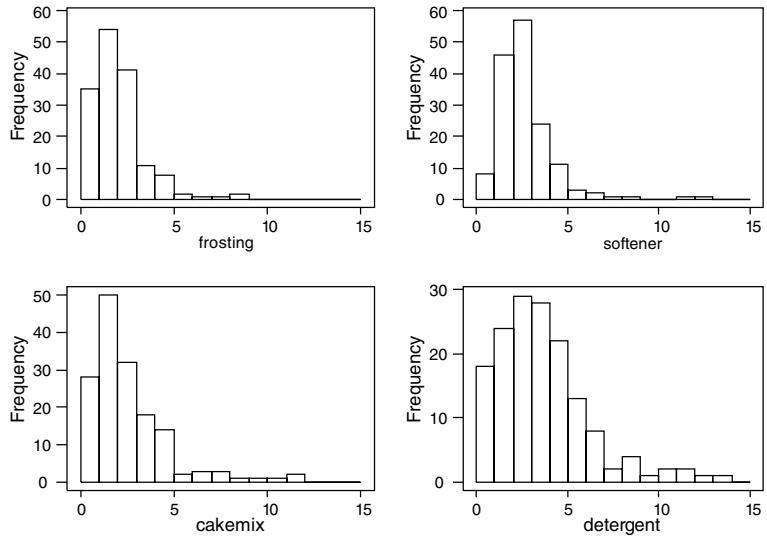

Fig. 1. Distribution of counts for each product category.

Table 1. Mean, variance and variance to the mean ratio for each product category.

	Mean	Variance	Variance to the mean ratio
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

sample Pearson correlation coefficients for the two-way product combinations. Only two correlations are significantly larger 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 λ's is 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 $\theta = (\lambda_C, \lambda_F, \lambda_D, \lambda_S, \lambda_{CF}, \lambda_{DS})$. Thus we have

$$\begin{aligned} Y_C &= X_C + X_{CF} \\ Y_F &= X_F + X_{CF} \\ Y_D &= X_D + X_{DS} \\ Y_S &= X_S + X_{DS} \end{aligned} \tag{1}$$

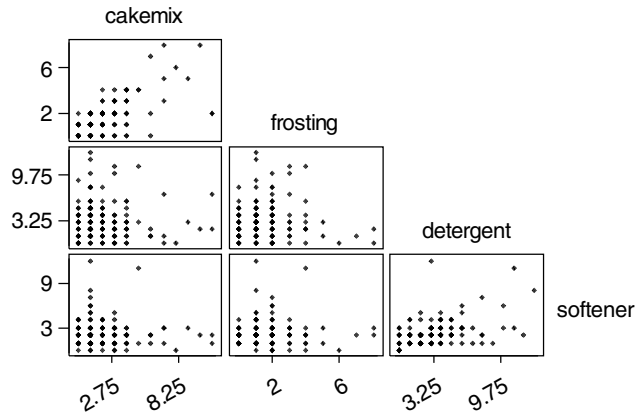

Fig. 2. Scatter matrix for the four products.

Our definition of the model, in fact, assumes that the conditional probability function is the product of two bivariate Poisson distributions (KOCHERLAKOTA and KOCHERLAKOTA, 1992), 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

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

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

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

Therefore, 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 total covariance produced by the model is two-fold. First, the model assumes covariance between all the variables as a result of 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 and four observed

variables 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.

5.1.2 *Estimation: the EM algorithm*

The EM algorithm is a popular algorithm for ML estimation in statistics (see for instance DEMPSTER *et al.*, 1977; MCLACHLAN and KRISHNAN, 1997). 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}, which 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, computes the conditional expectation of the complete data loglikelihood with respect to the missing data, 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 4-dimensional vectors $Y_i = (Y_{Ci}, Y_{Fi}, Y_{Di}, Y_{Si})$. We use the standard data augmentation used for finite mixture models by introducing as latent variables the vectors $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ik})$ that correspond to the component memberships with $Z_{ij} = 1$ if the i-th observation belongs to the j-th component, and 0 otherwise. Furthermore we introduce some more latent variables as follows:

We introduce component specific latent variables, i.e. for the j-th component we use the unobservable vectors $X_i^j = (X_{Ci}^j, X_{Fi}^j, X_{Di}^j, X_{Si}^j, X_{CFi}^j, X_{DSi}^j)$ where the superscript indicates the component, and the variables are the latent variables used to construct the model in (1). Thus, the complete data are the vectors (Y_i, X_i, Z_i). If we denote with φ the vector of the parameters, then the complete loglikelihood takes the following form

$$\begin{aligned} L(\varphi) &= \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\left(\log p_j + \log \prod_{t\in\Omega} f(X_{ti}^j \mid \lambda_{tj})\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\log p_j + \sum_{i=1}^{n}\sum_{j=1}^{k} Z_{ij}\sum_{t\in\Omega}\left(-\lambda_{tj} + X_{ti}^j \log \lambda_{tj} - \log X_{ti}^j!\right) \end{aligned}$$

where $\Omega = \{C, F, D, S, CF, DS\}$. One can see that concerning λ_{tj}, $t \in \Omega$, the relevant part of the complete loglikelihood is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(-Z_{ij}\lambda_{tj} + Z_{ij}X_{ti}^j \log \lambda_{tj}\right)$$

and, hence, one needs the expectations $E(Z_{ij})$ and $E(X_{ti}^j Z_{ij})$. But for the latter, since Z_{ij} is a binary random variable, we have that X_{ti}^j is 0 if the observation does not belong to the j-th component and takes the value X_{ti}^j if $Z_{ij} = 1$. Thus

$E(X_{ti}^{j}Z_{ij}) = p(Z_{ij})E(X_{ti}^{j} \mid Z_{ij} = 1)$. The last expectation is the expectation of the latent variable X_{ti}^{j} given that it belongs to the j-th component.

Thus, at the E-step one needs the expectations $E(Z_{ij} \mid Y_i, \phi)$ for i=1,...n, $j = 1,...k$ and $E(X_{ti}^{j}|Y_i, Z_{ij} = 1, \phi)$ for $i = 1,...n$, $j = 1,...,$ and $t \in \Omega$.

More formally, the procedure can be described as follows:
E-step: Using the current values of parameters calculate

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

$$\begin{aligned} E(X_{CFi}^{j} \mid Y_i, Z_{ij} = 1, \phi) &= b_{CFi}^{j} \\ &= \sum_{r=0}^{\min(y_{Ci}, y_{Fi})} r \Pr[X_{CFi}^{j} = r \mid Y_{Ci}, Y_{Fi}, Z_{ij} = 1, \phi] \\ &= \sum_{r=0}^{\min(y_{Ci}, y_{Fi})} r \frac{\Pr[X_{CFi}^{j} = r, y_{Ci}, y_{Fi} \mid Z_{ij} = 1, \phi]}{\Pr[y_{Ci}, y_{Fi} \mid Z_{ij} = 1, \phi]} \\ &= \frac{\sum_{r=0}^{\min(y_{Ci}, y_{Fi})} rPo(y_{Ci} - r \mid \lambda_{Cj}) Po(y_{Fi} - r \mid \lambda_{Fj}) Po(r \mid \lambda_{CFj})}{P(y_i \mid \theta_j)} \end{aligned}$$

The corresponding expression for $E(X_{DSi}^{j}|Y_i, Z_{ij} = 1, \phi) = b_{DSi}^{j}$ follows by analogy. Then

$$\begin{aligned} E(X_{Ci}^{j}|Y_i, Z_{ij} = 1, \phi) &= b_{Ci}^{j} = y_{Ci} - b_{CFi}^{j} \\ E(X_{Fi}^{j}|Y_i, Z_{ij} = 1, \phi) &= b_{Fi}^{j} = y_{Fi} - b_{CFi}^{j} \\ E(X_{Di}^{j}|Y_i, Z_{ij} = 1, \phi) &= b_{Di}^{j} = y_{Di} - b_{DSi}^{j} \\ E(X_{Si}^{j}|Y_i, Z_{ij} = 1, \phi) &= b_{Si}^{j} = y_{Si} - b_{DSi}^{j} \end{aligned}$$

M-step: Update the parameters

$$p_j = \frac{\sum_{i=1}^{n} w_{ij}}{n} \text{ and } \lambda_{tj} = \frac{\sum_{i=1}^{n} w_{ij} b_{ti}^{j}}{\sum_{i=1}^{n} w_{ij}}, \quad \text{for } j = 1, \ldots k, \ t \in \Omega$$

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 are straightforward. 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.

REMARK. 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, as the original data are collapsed to frequency tables. Consequently, our clustering model is scalable to databases with large amounts of records. 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 number 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 which is related to the number of latent variables introduced. Adding more latent variables leads 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 more computational effort is needed. This is a strong indication that the structure imposed before the fit of the model must remain at moderate levels. However, for a retail category manager, who is usually responsible for a limited number of product categories, this will pose no practical problems.

The above algorithm is fully described for the case of two-way interactions. A more extensive version of the algorithm in the case where the covariance structure is more complicated is provided in KARLIS and MELIGKOTSIDOU (2003).

5.2 *Common covariance model*

Alternative approaches to reduce the complexity of the variance–covariance structure and thus the number of free parameters in the multivariate Poisson mixture model were previously proposed in the literature, such as the common covariance model (see, e.g. JOHNSON *et al.*, 1997, TSIONAS, 2001). In this approach, the four-variate Poisson distribution $(Y_C,\ Y_F,\ Y_D,\ Y_S)$ with one common covariance term is defined as:

$$Y_C = X_C + X_{CFDS}$$

$$Y_F = X_F + X_{CFDS}$$

$$Y_D = X_D + X_{CFDS}$$

$$Y_S = X_S + X_{CFDS}$$

with all X's independent univariate Poisson distributions with respective parameters $(\lambda_C, \lambda_F, \lambda_D, \lambda_S, \lambda_{CFDS})$. In fact, the joint probability function of the four-variate Poisson distribution $P(Y_C = y_C, Y_F = y_F,\ Y_D = y_D,\ Y_S = y_S)$ can then be written as:

$$\sum_{x=0}^{\min(y_C,y_F,y_D,y_S)} e^{-\lambda} \frac{\lambda_C^{y_C-x}}{(y_C-x)!} \frac{\lambda_F^{y_F-x}}{(y_F-x)!} \frac{\lambda_D^{y_D-x}}{(y_D-x)!} \frac{\lambda_S^{y_S-x}}{(y_S-x)!} \frac{\lambda_{CFDS}^{x}}{x!}$$

where $\lambda = \lambda_C + \lambda_F + \lambda_D + \lambda_S + \lambda_{CFDS}$. This k-segment four-variate common covariance Poisson mixture model requires the estimation of $6k-1$ parameters.

The estimation by means of an EM-algorithm can be carried out in a similar way as presented in section 5.1.2. For example, at the E-step one has to derive the expectations of the latent variables $E(X^j_{CFDSi} \mid Y_i, \phi, Z_{ij} = 1)$ and to update the corresponding λ_{CFDSj} correspondingly.

5.3 *Local independence model*

Apart from the previous models that explicitly incorporate covariance into the model structure specification, the local independence model assumes that the purchase rates within each segment are mutually independent. The joint probability therefore reduces to the product of the product category specific densities. The four-variate model can then be defined as:

$$Y_C = X_C$$

$$Y_F = X_F$$

$$Y_D = X_D$$

$$Y_S = X_S$$

In that case, the joint probability function or the general k-segment mixture model for the four product categories takes a very simple form:

$$P(y_C, y_F, y_D, y_S) = \sum_{j=1}^{k} p_j \prod_{i \in \{C,F,D,S\}} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

The number of parameters for the k-segment four-variate Poisson mixture model then amounts to $5k - 1$.

It is important to emphasize that while the local independence model assumes that within each cluster the products are uncorrelated, the unconditional model assumes correlation between the products due to mixing, as previously stated. Estimation can be carried out simply by using the EM algorithm of section 5.1 without updating the parameters λ_{CF}, λ_{DS}, which have been set equal to 0. The remaining formulas remain unchanged.

5.4 *Identifiability*

Identifiability of parameters is a major issue in finite mixture modelling because it ensures that the estimation task is plausible. For the models described above we show that identifiability of parameters holds.

The key result concerning identifiability of the local independence model is due to TEICHER (1967). The local independence model assumes merely that the conditional distributions are the product of simple Poisson distributions and, hence, identifiability holds.

Identifiability for bivariate Poisson mixtures have been proved by AL-HUSSAINI and AHMAD (1981). The restricted covariance model uses as conditional distributions the product of bivariate Poisson distributions and, hence, combining the results of AL-HUSSAINI and AHMAD (1981) and TEICHER (1967) identifiability can be shown. Finally, identifiability of the common covariance model can be seen following the proof of AL-HUSSAINI and AHMAD (1981) for the bivariate case.

In all cases and in order to ensure that interchanging the order of components is not allowed, we assume that the components are in lexicographical order. This is a rather technical assumption without special importance in the practical level. However, despite the fact that this restriction has been used during estimation, for improving the exposition of the results we have not ordered the components. Finally empty components are not allowed, by setting all p_j's larger than 0.

6 Empirical Study

The models presented in section 5 were applied to the scanner dataset (see section 5.1.1). The covariance structure of the restricted covariance model was identified after preliminary examination of the data as previously illustrated. The other two models can be considered as alternatives to this model. In order to examine the appropriateness of the model, we also used other models with the restricted nature of the model in section 5.1.1 but using different pairs of non-zero covariances. However, the restricted covariance model was found to be superior and for this reason we examine it in depth in section 6.1. A comparison of different models is provided in section 6.2 and managerial implications of the model are discussed in section 6.3.

6.1 *The restricted covariance model*

The EM algorithm described in section 5.1.2 was implemented sequentially for 1 to 16 components ($k = 1,\ldots, 16$). The loglikelihood stopped increasing after $k > 16$ (see Figure 3). 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 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 less 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 (McLachlan and Peel, 2000). The majority of these 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) (Akaike, 1974) given as:

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

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 also have been used. However, it is beyond the scope of the present paper to argue for or against such criteria. In fact, Andrews and Currim (2003) showed that it is not easy to determine the optimal selection criterion as it depends on the model structure and the statistical distributions being used. In practice, we also seek solutions that are interpretable.

Figure 3 shows that the AIC criterion selects six components. The depicted values are rescaled so as to be comparable to the loglikelihood. It is interesting to note that the five-component solution is quite close and thus for reasons of parsimony it could

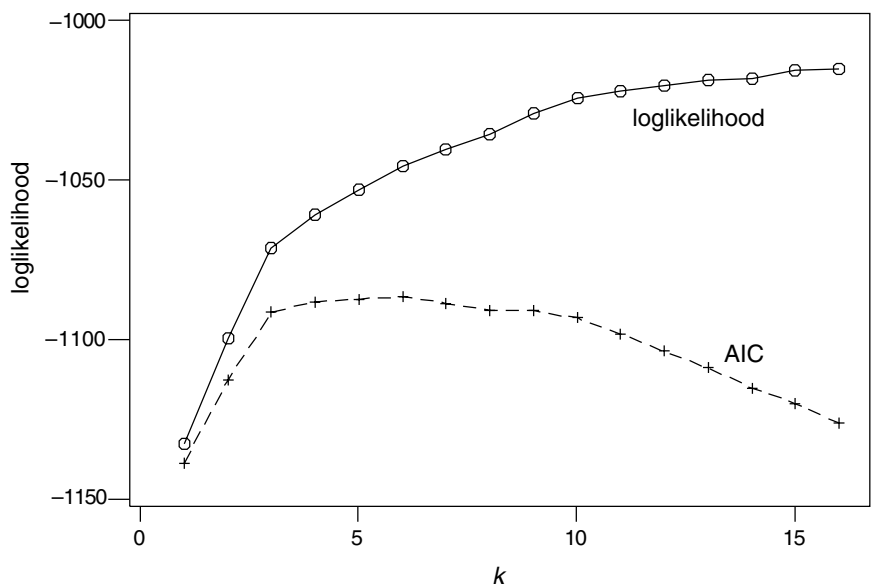

Fig. 3. LL and AIC (rescaled) against the number of components (k) for the restricted MVP mixture model.

be considered. As we will describe in the sequel, the differences between the five- and six-components solutions are minor, hence, it seems plausible to choose the smaller model with five components.

Figure 4 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 into two parts, perhaps with some minor modification for the rest of the 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 Tables 2 and 3). It is also quite interesting to see that the solutions with five and six components differ only slightly. This is extremely interesting from the retailer's point of view for whom the existence of a limited number of clusters is very important. Indeed, if a large number of clusters did exist, it would be 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. Note that in certain other disciplines small clusters may have interesting interpretations as, for example, in archaeometry or in outlier detection.

In Figure 5, selected bubble-plots are plotted for pairs of parameters. In fact, each graph depicts the joint mixing distribution for the selected pair. The plots depict both the five- and the six-cluster solution. The circles represent the six-cluster solution and the squares the five cluster solution. The size of the circle/square reflects the mixing proportions, the larger the size the larger the mixing proportion. It is clear from the

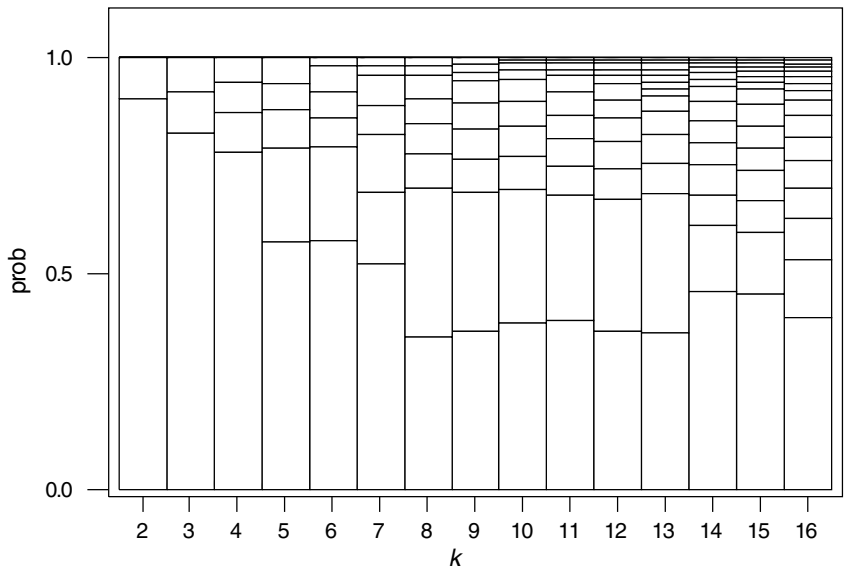

Fig. 4. The mixing proportions for model solutions with $k = 2$ to 16 components.

Table 2. Parameter estimates and standard errors for the five-component model.

Cluster		Parameters						
		λ_C	λ_F	λ_{CF}	λ_D	λ_S	λ_{DS}	p
1	ML estimate	0.207	0.295	1.507	8.431	4.639	0.000	0.088
	(standard error)	(0.492)	(0.504)	(1.252)	(1.682)	(2.272)	(0.000)	(0.056)
2	ML estimate	0.427	0.279	1.093	1.347	0.031	1.955	0.575
	(standard error)	(0.126)	(0.077)	(0.150)	(0.177)	(0.113)	(0.217)	(0.069)
3	ML estimate	0.908	0.441	0.555	0.000	1.030	0.977	0.216
	(standard error)	(0.322)	(0.231)	(0.293)	(0.039)	(0.306)	(0.290)	(0.048)
4	ML estimate	2.000	0.792	4.292	0.000	0.524	1.187	0.062
	(standard error)	(1.372)	(0.923)	(1.436)	(0.053)	(0.382)	(0.503)	(0.030)
5	ML estimate	4.668	0.000	1.223	3.166	1.161	0.702	0.059
	(standard error)	(1.043)	(0.188)	(0.518)	(1.040)	(0.868)	(0.813)	(0.028)

Table 3. Parameter estimates and standard errors for the six-component model.

Cluster		Parameters						
		λ_C	λ_F	λ_{CF}	λ_D	λ_S	λ_{DS}	p
1	ML estimate	0.205	0.171	1.523	6.116	0.000	2.061	0.066
	(standard error)	(0.554)	(0.419)	(1.008)	(1.138)	(0.063)	(0.697)	(0.082)
2	ML estimate	0.356	0.000	2.063	8.698	10.33	0.000	0.019
	(standard error)	(0.977)	(0.025)	(1.239)	(1.310)	(3.709)	(0.000)	(0.060)
3	ML estimate	0.424	0.311	1.061	1.275	0.026	2.083	0.578
	(standard error)	(0.185)	(0.117)	(0.196)	(0.241)	(0.138)	(0.258)	(0.099)
4	ML estimate	0.897	0.425	0.587	0.000	1.047	0.972	0.215
	(standard error)	(0.372)	(0.300)	(0.320)	(0.000)	(0.369)	(0.440)	(0.057)
5	ML estimate	1.975	0.776	4.287	0.000	0.521	1.192	0.062
	(standard error)	(1.414)	(0.650)	(1.127)	(0.060)	(0.739)	(1.005)	(0.041)
6	ML estimate	4.684	0.000	1.219	3.040	1.085	0.782	0.059
	(standard error)	(0.867)	(0.000)	(0.344)	(0.669)	(0.614)	(0.592)	(0.030)

graph that the two solutions differ only slightly and that the six-cluster solution just splits up one of the existing clusters.

Tables 2 and 3 contain the parameter estimates for the models with five components and six components respectively. One can see that all the components of the five-cluster solution still exist in the six-cluster solution, but an additional component appeared (number 2 in the six-cluster solution) that took observations from the old components 1 and 3 of the five-cluster solution. In both solutions, there are two 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.

In both Tables 2 and 3, one can also see bootstrap standard errors for the parameters. Note, however, that standard errors in a mixture model are not so easy to obtain. There are several reasons for this. First of all, asymptotic standard errors need a large sample size in order to be valid. In our case, the

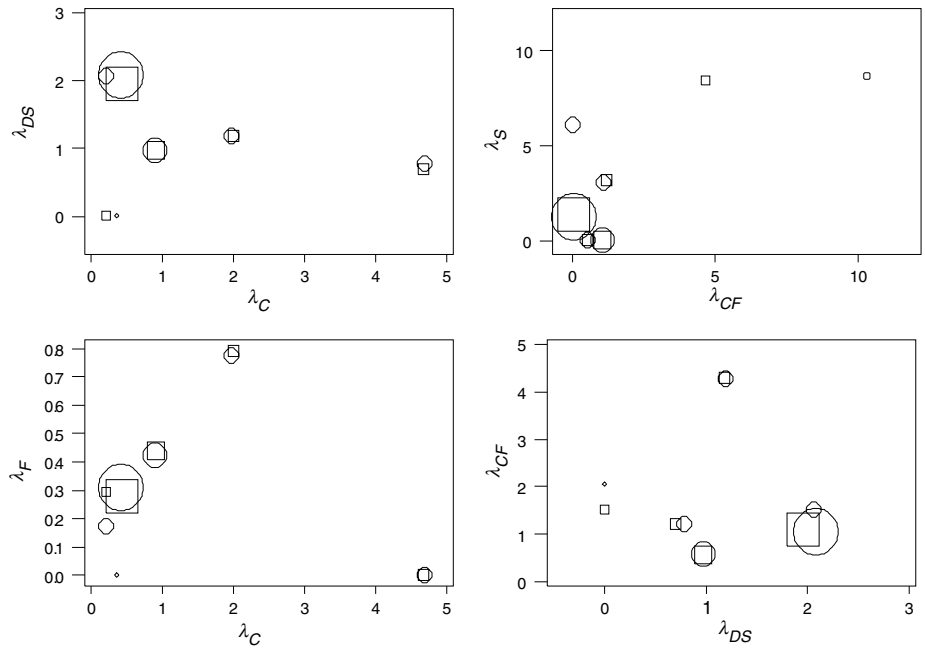

Fig. 5. Bubble plots for selected pairs of parameters (the squares indicate the five-component solution and the circles the six-component solution).

sample size is rather small and thus they are not appropriate at all (see the discussion in McLachlan and Peel, 2000). Alternatively, one may consider bootstrap standard errors but they suffer from the fact that resampling may lead to estimates with fewer components and hence the parameters are not identifiable. Sometimes, the order of the components may change during the execution of the algorithm and thus label switching problems occur. Tables 2 and 3 show bootstrap standard errors, mainly in order to have a rough feeling for the standard errors, however the errors should be interpreted with caution. Although special care was taken to avoid label switching, we believe that for small components with small mixing proportions the estimated standard errors may overestimate the true ones because of label switching. Parameters with zero estimated values and zero standard errors in fact can be interpreted as they do not differ significantly from zero.

The rule for allocating observation to clusters is the higher posterior probability rule. This means that an observation is allocated to the cluster for which the posterior probability, as measured by w_{ij}, is largest. 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 ten customers that changed cluster between the five and six-cluster solutions, eight of them switched to the new cluster.

In order to assess the quality of clustering we calculated the entropy criterion (McLachlan and Peel, 2000) as:

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

with the convention that $w_{ij}\log w_{ij} = 0$ 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. Conversely, 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 relatively good separation between the clusters.

An interesting feature of the results in Tables 2 and 3 is the interpretation of the zero values. If the zero value corresponds to a covariance parameter (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 of 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 there leads to high correlation between the two variables, so the value for this product is usually very similar to the values for the other product.

In order to interpret the cluster differences with regard to the original data, Table 4 contains the cluster centers for the five-components solution, i.e. the average purchase rate per product for all the households allocated to that cluster. The last row contains the sample centers for the entire data, corresponding to the data in Table 1, while the last column shows the number of observations allocated to that cluster based on a criterion that assigns each observation to the cluster with the higher posterior probability.

Looking at the two major clusters (clusters 2 and 3) in Table 2, it can be observed that they have rather different profiles. With regard to fabric detergent and fabric softener, in particular, both clusters indeed show rather different behaviors. Cluster 2 shows a very low average purchase rate of fabric softener ($\lambda_S = 0.031$) but a rather high covariance between fabric detergent and fabric softener ($\lambda_{DS} = 1.955$). 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 4: customers in cluster 2 have an

Table 4. Cluster centers for the five-component mixture model

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

average purchase rate of fabric softener of 2.022, which is largely due to the covariance with fabric detergent ($\lambda_{DS} = 1.955$). 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 because from (1) it follows that $Y_S = X_S + X_{DS}$ and thus, the observed sales of softener for that cluster can be roughly decomposed into its own latent effect ($\lambda_S = 0.031$) plus the latent effect due to cross-selling with detergent ($\lambda_{DS} = 1.955$).

By contrast, cluster 3 shows a somewhat different profile. Cluster 3 shows no purchases of fabric detergent at all ($\lambda_D = 0.000$), but again has a relatively strong covariance with fabric softener ($\lambda_{DS} = 0.977$). 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 4: people in cluster 3 have an average purchase rate of fabric detergent of 0.912, which is rather low compared with the total sample average, but this purchase rate is exclusively due to the covariance with fabric softener ($\lambda_{DS} = 0.977$). Consequently, it can be concluded that sales of fabric detergent on its own in cluster 3 are non-existent and, if they occur, they have occurred exclusively as a result of cross-selling with fabric softener.

These are important findings because they have interesting implications for marketing decision making, e.g. for targeted merchandising strategies. For instance, cluster 2 in the five-segment solution (see Table 2) represents an important customer segment because it contains almost 60% of all observations. However, more importantly, the purchase rate of softener in this cluster is rather low compared with the purchase rate of detergent, but the covariance between the purchase rates of softener and detergent is very high. Therefore, in order to take advantage of this interdependency effect, retail management could decide to make softener more salient, i.e. to bring softener more to the attention of those customers shopping for detergent, for instance by putting both products in the same display. Moreover, placing highly interdependent products closer together in the store may also have a positive effect on the shopping experience of time-pressured customers who typically do not want to waste too much time looking for items in the store.

Furthermore, knowledge about correlated category usage patterns enables category managers (and manufacturers) to implement cross-category marketing strategies. For instance, Catalina Marketing (www.catalinamarketing.com) in the US 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, consumers belonging to cluster 2 (Table 2) could be given detergent coupons, not only to stimulate sales of detergent, but to stimulate sales of softener too, based on this strong interdependency effect. It is important to note, however, that since we do not possess covariate information in the form of marketing mix variables, the sensitivity towards particular marketing actions remains uncertain. On the other hand, earlier research by RUSSEL and KAMAKURA (1997) found that consumers with high purchase rates for a particular product category respond more vigorously to the marketing activities in that category and that, due to strongly positively correlated

category preferences, promotions in one category may not only lead to higher sales in that category but also in the other positively correlated categories.

Another interesting merchandising strategy could be to put highly interdependent products closer together in the store to enhance the shopping experience of so-called 'run-shoppers' who typically don't want to waste time looking for items in the store.

Finally, cluster 1 in the five-component solution shows another interesting, yet different profile compared with the two bigger clusters discussed before. Customers in this cluster purchase large quantities of fabric detergent ($\lambda_D = 8.431$) and fabric softener ($\lambda_S = 4.639$), however, the purchases are not correlated at all ($\lambda_{DS} = 0.000$). This means that although customers in cluster 1 purchase large amounts of detergent and softener, their values are not the result of cross-selling between both products. Consequently, promotional campaigns (like price reduction or special displays) on one of the two products (say detergent) will probably not affect the sales of the other product (softener).

6.2 *Comparison with other models*

In order to examine the proposed model, we also fitted various other models to the data. For example, we considered other restricted covariance models with different pairs of correlated products within each cluster. The model used for Figure 6 is the one with covariances for the pairs cakemix with detergent and frosting with softener. This was the second best model with paired covariances. This model is referred to as RC2. We also considered the model with common covariances (CC for the figures) and the local independence model (LI).

Figure 6a depicts the loglikelihood for various numbers of clusters while Figure 6b depicts the AIC for the same models. Judging from the loglikelihood one can see that the proposed model is superior for all values of k, while model RC2 is the second best, keeping in mind that it has more parameters than both the CC and LI models. When taking into account the different number of parameters, one can see from Figure 6b that the common covariance model becomes better than RC2 model.

Perhaps more importantly the LI model is superior to the RC2 model when using the AIC. This implies that introducting non-significant covariances as we did in model RC2 does not help the model fitting at all. This clearly implies that introducing too much artificial structure to the model cannot help to improve the model at the cost of more computational burden and a non-parsimonious model.

It is interesting also to point out that all the models select five clusters apart from model RC1, which selects six clusters. Recall that we have discussed the minor differences of this model with five and six clusters. From the above it becomes obvious that the proposed model fits the data better.

6.3 *Managerial implications*

In order to assess the practical relevance of the model for marketing managers, WEDEL and KAMAKURA (1999) suggested six criteria for effective segmentation, i.e.

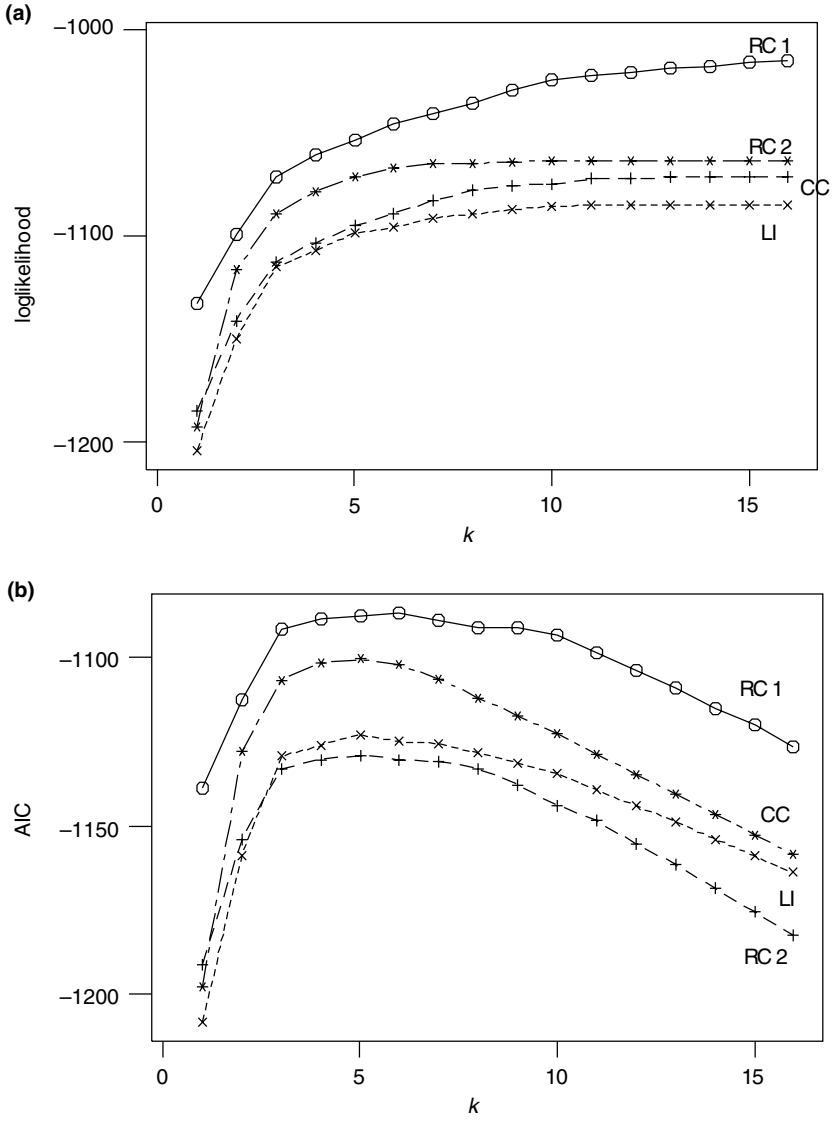

Fig. 6. Comparison of different models, using either the loglikelihood (a) or the AIC (b).

clusters should be identifiable, substantial, accessible, stable, responsive and actionable.

With regard to the *identifiability* of the multivariate Poisson mixture model with restricted covariance structure, it can be concluded from the entropy statistic (see section 6.1) that the segments are quite well separated and that the characteristics of the clusters are very different (see Figure 4 and Tables 2 and 3).

Furthermore, the cluster solution produces a limited set of clusters, with two of them (clusters 2 and 3, see Table 4) of a *substantial size*. In this respect, the model

with restricted covariance structure produces even better results than the local independence model, which for a larger number of components produces clusters of almost equal size. Indeed, in the restricted model two substantial clusters appear, containing almost 80% of all observations, which remain almost unchanged for larger component solutions (see Figure 4).

The clusters of the proposed model are *moderately accessible* because the basis for segmentation has been chosen as the category purchase rates, which can easily be tracked and stored in a database, and which offer opportunities at the checkout to differentiate between customers of different segments. Indeed, the basis for segmentation is observable and product specific and thus enables targeted marketing campaigns, e.g. for printing customized coupons at the checkout. On the other hand, no covariate information is included in the model (e.g. socio-demographic and/or lifestyle data) such that the differences in the structure of each cluster cannot be explained by means of covariate information (as in mixture regression models). This may limit the applicability of the model in a retailing context because the proposed model does not link the cluster solution to the loyalty card information that might be available from the customers.

The *structural stability* of the clusters is satisfactory. It is shown that usually the introduction of an additional component corresponds to a split of an existing component into two parts (see Figure 4), perhaps with some minor modification for the rest of the components, especially if they have estimates close to the component split. The *temporal stability* of the clusters cannot be examined as we do not possess purchase data for the same individuals at a later period in time. The existence of such data would enable one to compare the cluster solution and cluster membership of the observations in order to evaluate their stability over time. Furthermore, the lambda parameters in the multivariate Poisson are specified as stationary parameters, which are assumed not to change over time.

The model does not provide direct insights into the *responsiveness* of the discovered clusters. For instance, cluster 1 in the five component solution of the restricted covariance model (see Table 2) shows no interdependence between the purchases made in the categories fabric detergent and fabric softener. It could therefore be expected that promotions on fabric detergent would have no effect on the purchases of fabric softener in that particular segment. However, the model does not support such interpretations. The reason is that we do not possess information about the (cross-)promotional elasticity of the products, which would be necessary to assess the responsiveness of particular promotional campaigns. Yet, earlier work by RUSSELL and KAMAKURA (1997) found that consumers with high purchase rates for a particular product category respond more vigorously to the marketing activities in that category and that, due to strongly positively correlated category preferences, promotions in one category not only lead to higher sales in that category but also in the other positively correlated categories.

Finally, the *actionability* of the cluster solution depends on the strategic positioning of the retail firm. The extent to which the retailer wants to use the

results of the model to devise customized marketing campaigns will partially depend on the information technology available to the retailer and their willingness to experiment. Currently, many (European) retailers are not very keen on customizing promotions towards their customers because of the risk of wrong perception by the consumer. In fact, one supermarket retailer told us that he was afraid not to treat all consumers alike because of the risk that the consumers might feel manipulated by the retailer (Why does my neighbor get different promotions?) or may feel it to be an intrusion into their personal sphere of life (privacy).

7 Conclusions and discussion

In this paper, a multivariate Poisson mixture model was introduced and adopted to cluster supermarket shoppers based on their purchase rates in a number of predefined product categories. However, instead of using the multivariate Poisson distribution with a fully-saturated variance–covariance structure, we showed that the number of free parameters can be reduced either subjectively (by selecting some covariances based on prior knowledge or on preliminary examination of the data, as in our case) or objectively by selecting terms using a model selection criterion. In our application we used log–linear analysis as a preliminary tool to select particular covariances, while we also employed a model selection criterion to verify that the selected structure is preferable. 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 fitted on a real supermarket dataset that included the purchases of 155 households over a period of 26 weeks in four product categories. The results of the model indicated that two large clusters, accounting for almost 80% of the observations, could be found to have distinct purchasing profiles in terms of the purchase rates and purchase interactions between the product categories considered. Moreover, a number of real merchandising strategies were proposed for the discovered clusters based on this purchasing profile.

The model, however, also has some limitations. First, the model does not contain explanatory variables. For instance, marketing mix variables could be used as covariates in the model in order to identify segments of customers who react differently (in terms of their purchase behavior) towards particular marketing mix decisions, such as pricing or promotions. Knowledge of this kind could be used by the marketing manager to influence the purchase behavior of particular customer segments that are most sensitive towards particular changes in the marketing mix. Another type of covariate could be socio-demographic variables per customer (e.g. obtained by means of a loyalty card) in order to profile the segments. Knowledge of this kind could be used by the marketing manager to target particular socio-demographic segments, e.g. by electronic couponing systems based on loyalty card

information. Yet, the absence of explanatory variables in our model can also be viewed as a strength. In fact, the suggested model only requires information on category consumption to segment customers. Indeed, on some occasions, marketing mix variables are simply not available. Examples of such data sources include A.C. Nielsen wand panels in which consumers record purchases by scanning UPC product codes at home, or consumption data obtained from a retailer's 'loyalty card club' participants. In other words, the method presented in this paper can be used when other data models are impractical.

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 could another, or entire clusters could change profile over time, i.e. move from one position to another. However, since we do not expect such behavior to take place in such a short time, this dynamic aspect has not been accounted for in this study.

Thirdly, the multivariate Poisson mixture model assumes that, within each cluster, the correlations are positive. This is an important limitation. However, unfortunately there is a shortage of multivariate count models that allow for negative correlations. For example, the Poisson–lognormal model of AITCHINSON and HO (1989), used in CHIB and WINKELMANN (2001), assumes that conditionally the counts are uncorrelated and negative correlation is induced due to the mixing. In fact, our model allows negative correlation in the population through the same procedure and with respect to the above mentioned model, our model is more tractable. There is definitely a need to develop a simple model for multivariate counts that allows for negative correlation as well. From a marketing point of view, the inability of the model to account for negative correlations implies that we deal only with cross-selling (say complementarity) effects and not with substitution effects. However, following RUSSELL and KAMAKURA (1997), substitution effects are more dominant within product categories, and to a much lesser extent across product categories (where complementarity effects play a much more important role). Therefore, since in this paper we deal with category purchases across (and not within) different product categories, we believe that this limitation of the model poses no fundamental problems in this context.

Fourthly, it is not clear to what extent the cross-selling effects being found are the result of past marketing actions (promotions, pricing, etc.) during the period over which the data were collected. As a result of this, some dependencies between categories may be over-estimated. To solve this problem, one would need detailed information about marketing mix decisions taken by the retailer, and also by competing stores, during the period of data collection in order to estimate the impact of such decisions in a longitudinal study.

Fifthly, our model does not discuss joint purchases made at the same shopping event, such as in the multivariate probit model by MANCHANDA *et al.* (1999). The model discusses purchase rates for different product categories over a given period of time. Therefore, the model cannot be used to make predictions about particular shopping visits.

Finally, although it has been shown in this paper how to significantly reduce the complexity of the multivariate Poisson mixture model, the scalability towards including more product categories requires more empirical study.

Acknowledgement

Dimitris Karlis kindly appreciates the financial support from the Data Analysis and Modeling Group of Limburg University Center. Part of the paper was written during his visit to Limburg, Belgium.

References

AGRESTI, A. (1996), *An introduction to categorical data analysis*, Wiley Series in Probability and Statistics.

AINSLIE, A. and P. E. ROSSI (1998), Similarities in choice behavior across product categories, *Marketing Science* **17**, 91–106.

AITCHINSON, J. and C. H. HO (1989), The multivariate Poisson–log normal distribution, *Biometrika* **75**, 621–629.

AKAIKE, H. (1974), A new look at statistical model identification, *IEEE Transactions on Automatic Control* **AC-19**, 716–723.

AL-HUSSAINI, E. K. and K. E. D. AHMAD (1981), On the identifiability of finite mixtures of distributions, *IEEE Transactions on Information Theory* **27**, 664–668.

ANDREWS, R. L. and I. S. CURRIM (2003), Retention of latent segments in regression-based marketing models, *International Journal of Research in Marketing* **20**, 315–321.

BANFIELD, J. D. and A. E. RAFTERY (1993), Model based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

BÖHNING, D. (1999), *Computer assisted analysis of mixtures and applications in meta-analysis, disease mapping and others*, CRC Press.

CADEZ, I. V., P. SMYTH and H. MANNILA (2001), Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction, in: F. PROVOST and R. SRIKANT (eds), *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco (CA), USA, 37–46.

CHIB, S. and R. WINKELMANN (2001), Markov Chain Monte Carlo analysis of correlated count data, *Journal of Business and Economic Statistics* **19**, 428–435.

DEMPSTER, A. P., N. M. LAIRD and D. RUBIN (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* B **39**, 1–38.

DILLON, W. R. and A. KUMAR (1994), Latent structure and other mixture models in marketing: an integrative survey and overview, in: R. P. BAGOZZI (ed.), *Advanced Methods in Marketing Research*, Blackwell, Cambridge MA, 295–351.

DUMOUCHEL, W. and D. PREGIBON (2001), Empirical Bayes screening for multi-item associations, in: F. PROVOST and R. SRIKANT (eds), *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco (CA), USA, 67–76.

JOHNSON, N., S. KOTZ and N. BALAKRISHNAN (1997), *Discrete multivariate distributions*, Wiley, New York.

HOOGENDOORN, A. W. (1999), Description of purchase incidence by multivariate heterogeneous Poisson processes, *Statistica Neerlandica* **53**, 21–35.

KANO, K. and K. KAWAMURA (1991), On recurrence relations for the probability function of multivariate generalized Poisson distribution, *Communications in Statistics – Theory and Methods* **20**, 165–178.

KARLIS, D. (2003), An EM algorithm for multivariate Poisson distribution and related models, *Journal of Applied Statistics* **30**, 63–77.

KARLIS, D. and L. MELIGKOTSIDOU (2003), Finite mixtures of multivariate Poisson distributions with application, *Technical Report*, Department of Statistics, Athens University of Economics and Business, Greece.

KOCHERLAKOTA, S. and K. KOCHERLAKOTA (1992), *Bivariate discrete distributions*, Marcel Dekker, New York.

MANCHANDA, P., A. ANSARI and S. GUPTA (1999), A model for multi-category purchase incidence decisions, *Marketing Science* **18**, 95–114.

MCLACHLAN, G. J. and K. E. BASFORD (1988), *Mixture models: inference and applications to clustering*, Marcel Dekker, New York.

MCLACHLAN, G. J. and T. KRISHNAN (1997), *The EM algorithm and its extensions*, Wiley, New York.

MCLACHLAN, G. J. and D. PEEL (2000), *Finite mixture models*, Wiley, New York.

MORRISON, D. G. and D. C. SCHMITTLEIN (1988), Generalizing the NBD model for customer purchases: what are the implications and is it worth the effort?, *Journal of Business and Economic Statistics* **6**, 145–166.

MULHERN, F. J. and R. P. LEONE (1991), Implicit price bundling of retail products: a multiproduct approach to maximizing store profitability, *Journal of Marketing* **55**, 63–76.

ORDONEZ, C., E. OMIECINSKI and N. EZQUERRA (2001), A fast algorithm to cluster high dimensional basket data, in: N. CERCONE, T. LIN and X. WU (eds), *Proceedings of the IEEE International Conference on Data Mining*, San Jose (CA), USA, 633–636.

RUSSELL, G. J. and W. A. KAMAKURA (1997), Modeling multiple category brand preference with household basket data, *Journal of Retailing* **73**, 439–461.

SUBRAHMANIAM, K. (1966), A test for intrinsic correlation in the theory of accident proneness, *Journal of the Royal Statistical Society* B **28**, 180–189.

TEICHER, H. (1967), Identifiability of product measures, *Annals of Mathematical Statistics*, 1300–1302.

TSIONAS, E. G. (2001), Bayesian multivariate Poisson regression, *Communications in Statistics – Theory and Methods* **30**, 243–255.

WEDEL, M. and W. A. KAMAKURA (1999), *Market segmentation: conceptual and methodological foundations*, Kluwer, Dordrecht.

WEDEL, M., W. A. KAMAKURA, W. S. DESARBO and F. TER HOFSTEDE (1995), Implications for asymmetry, nonproportionality, and heterogeneity in brand switching from piece-wise exponential mixture hazard models, *Journal of Marketing Research* **32**, 457–462.

Received: February 2003. Revised: March 2004.