Boztuğ, Yasemin; Reutterer, Thomas

Working Paper

A combined approach for segment-specific analysis of market basket data

SFB 649 Discussion Paper, No. 2006,006

Provided in Cooperation with:
Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

Suggested Citation: Boztuğ, Yasemin; Reutterer, Thomas (2006) : A combined approach for segment-specific analysis of market basket data, SFB 649 Discussion Paper, No. 2006,006, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

This Version is available at:
https://hdl.handle.net/10419/25089

SFB 649 Discussion Paper 2006-006

A Combined Approach for Segment-Specific Analysis of Market Basket Data

Yasemin Boztuğ*
Thomas Reutterer**

* Institute of Marketing, Humboldt-Universität zu Berlin, Germany
** Institute of Retailing and Marketing, Vienna University of Economics and Business Administration, Austria

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

A Combined Approach for Segment-Specific Analysis of Market Basket Data *

Yasemin Boztuğ
Humboldt University Berlin,
Institute of Marketing,
Spandauer Str. 1, D-10178 Berlin
Germany,
boztug@wiwi.hu-berlin.de

Thomas Reutterer
Vienna University of Economics and Business Administration,
Institute of Retailing and Marketing,
Augasse 2–6, A-1090 Vienna
Austria,
thomas.reutterer@wu-wien.ac.at

January 24, 2006

*We are very grateful to the DFG for generously supporting this research through the project #BO1952/1 and through the SFB 649 "Economic Risk".

Abstract

There are two main research traditions for analyzing market basket data that exist more or less independently from each other, namely exploratory and explanatory model types. Exploratory approaches are restricted to the task of discovering cross-category interrelationships and provide marketing managers with only very limited recommendations regarding decision making. The latter type of models mainly focus on estimating the effects of category-level marketing mix variables on purchase incidences assuming cross-category dependencies. We propose a procedure that combines these two modeling approaches in a novel two-stage procedure for analyzing cross-category effects based on shopping basket data: In a data compression step we first derive a set of market basket prototypes and generate segments of households with internally more distinctive (complementary) cross-category interdependencies. Utilizing the information on categories that are most responsible for prototype construction, segment-specific multivariate logistic models are estimated in a second step. Based on the data-driven way of basket construction, we can show significant differences in cross-effects and related price elasticities both across segments and compared to the global (segment-unspecific) model.

Keywords: Marketing, Choice Models, Market Basket Analysis, Cross-Category Effects, Segmentation

JEL-classification: C31, C33, C35, C63, M31

1 Introduction

A market or shopping basket is representing the result of a specific consumer's decision making process on the choice or non-choice of product categories among the assortment offered by a retail outlet during one and the same shopping trip. Retail managers are interested in better understanding the interdependency structure among categories purchased jointly by their customers for several reasons. Traditionally, insights into cross-category dependencies and corresponding marketing mix effects are of particular interest for optimizing the overall profitability of retail category management (cf., Müller-Hagedorn 1978; Manchanda et al. 1999; Song and Chintagunta 2003; Chen et al. 2005). However, most of the attempts towards this direction are restricted to fairly small selections of (sub-) categories. Naturally, today's large retail assortments not

only make the consideration of complete category ranges prohibitive but also managerially inexpediently. Nevertheless, in most empirical applications both types and numbers of included categories seem to be rather guided by analytical viability than by sound managerial considerations. Hence, the question arises which categories to be included in models for predicting cross-category effects that adequately represent consumers multicategory decisions.

More recently, numerous retailers have equipped members of their loyalty programs with bar-coded plastic cards and provided various incentives (such as discounts or check cashing privileges) to encourage their regular customers to present their membership cards at each purchase occasion (cf., e.g., Passingham 1998). Combined with modern point-of-sale (POS) scanning technologies, those retailers are nowadays collecting tremendous amounts of personally identifiable POS transaction data. Among other things, the latter are dissembling valuable behavioral information on cross-category purchase patterns of their prime customers. Furthermore, the meaningful linkage of such household-level purchase transaction histories with relevant data on respective store characteristics and marketing activities can provide valuable managerial support for designing and targeting segment-specific (or even individually) customized cross- and up-selling initiatives within advanced customer relationship management (CRM) programs (Rossi et al. 1996).

As a consequence of these developments and corresponding managerial requirements, the analytical focus for studying cross-category dependencies and associated marketing-mix effects needs to be shifted to a more disaggregate (i.e., individual or customer segment) level. In particular, to satisfy decision support needs in the framework of an effective management of loyalty card programs, information on customer segment-specific rather than aggregate cross-category effects is called for. In the next section, it is briefly reviewed that conventional approaches to market basket analysis exhibit inherent limitations to efficiently accommodate such information.

Following, we propose a procedure that combines two different schools of thought prevalent in the literature on market basket analysis in a stepwise manner. Within this analytical framework both the issue of adequate (i.e., consumer centric) category selection and segment construction can be suitably resolved. The methodology's capability to contribute to the above sketched information needs is illustrated in an empirical application study.

2 Linking exploratory and explanatory approaches to market basket analysis

There are two main research traditions for analyzing market basket data that exist more or less independently from each other, namely exploratory and explanatory types of models (for an overview cf. Mild and Reutterer 2003; Boztuğ and Silberhorn 2006). Exploratory approaches are restricted to the task of discovering distinguished cross-category interrelationships based on observed co-occurrence frequency patterns of jointly purchased items or product categories. In the marketing literature, this is also referred to as 'affinity analysis' (Russell et al. 1999). The majority of attempts contributed to this research field so far, however, are examining cross-category purchase effects on the aggregate level of household demand only. This especially applies to methods aiming at a parsimonious representation of pair-wise symmetric association measures derived from cross-tabulations of joint purchases across multiple categories (e.g., Böcker 1978; Dickinson et al. 1992; Julander 1992; Lattin et al. 1996).

In marketing research practice, meaningful cross-correlational structures are merely 'determined' by visual inspection. Thus, the marketing analyst is usually aiming at a parsimonious representation of the cross-category associations in a compressed and meaningful fashion. Multidimensional scaling techniques or hierarchical clustering are typically employed to accomplish this task. Besides the drawback of an 'average' (or aggregate) market view, the practical relevance of such attempts obviously suffers from their limitations to relatively small number of categories that are allowed for symmetric pair-wise relationships only.

The latter constraints are successfully resolved by a huge amount of research on association rule discovery, which initially was presented in the data mining literature (see, e.g., Agrawal et al. 1995; Anand et al. 1998; Brin et al. 1998; Hahsler et al. 2006) and have seen recent applications in the marketing-related literature (Chen et al. 2005; Van den Poel et al. 2004). Following a probabilistic concept, such rule mining techniques derive asymmetric implications (rules) for disjoint subsets of items or categories based on co-occurrence frequencies (associations) aggregated across households' and shopping baskets. Mining association rules is capable of dealing with both very large number of categories (or even single items) and shopping baskets. Albeit considerable advances, the issue of an 'average' (or aggregate) market view also applies to the various rule mining techniques.

The idea of representing cross-category purchase effects at a more disaggregate level

was introduced into the marketing community only recently (Schnedlitz et al. 2001; Decker and Monien 2003; Decker 2005). The authors employ various neural network architectures with unsupervised learning rules as a data compression device that results in a mapping of binary-valued vectors of category incidences within retail transactions onto a set of so-called prototypes. In their empirical applications, they illustrate that each of these prototypes is (post-hoc) responsible for a specific class of market baskets with internally more pronounced (complementary) cross-category purchase interdependencies as compared to the aggregate case.

Despite their usefulness for discovering meaningful cross-category interrelationships patterns, the managerial value of exploratory approaches to market basket analysis is obviously limited. Since no a-priori assumptions regarding the distinction between 'response' and 'effect' category is made and, more specifically, no marketing variables are directly incorporated in the analytical framework they provide marketing managers with only very limited recommendations regarding decision making. In contrary, explanatory (or predictive) types of models for analyzing market basket data mainly focus on estimating the effects of marketing-mix variables on category purchase incidence by explicitly accounting for cross-category dependencies among the retail assortment. Most of such explanatory models for market basket analysis introduced so far are either conceptualized as logit- or probit-type specifications within the framework of random utility theory (excellent state-of-the field reviews are provided by Russell et al. 1997, 1999; Seetharaman et al. 2004; Boztuğ and Silberhorn 2006). Modeling approaches that contribute to the estimation of segment-specific or even individual level marketing-mix effect parameters as claimed in the introduction of this paper include works by Russell and Kamakura (1997), Andrews and Currim (2002), Ainslie and Rossi (1998), Manchanda et al. (1999), Seetharaman et al. (1999) or Chib et al. (2002).

One practical problem with explanatory models, however, is that the set of categories to be incorporated in the model and analyzed simultaneously for cross-category effects on the selected response category is rather limited (typically, up to the size of four to five categories). Indeed, significant improvements of powerful Markov chain Monte Carlo simulation methodologies can help to successfully alleviate estimation problems, multivariate logit or probit approaches are confronted with when the number of product categories to be analyzed is increasing. Nevertheless, real-world retail assortments are typically consisting of dozens of potentially relevant product categories which still entail severe computational problems or the necessity to impose constraints on excessively large covariance matrices. Yet another problem concerns the rather ad hoc selection of relevant categories for basket creation, which often needs to be guided

by managerial intuition or practical considerations within the respective problem context.

To overcome the limitations inherent to both exploratory attempts (lack of implications for managerial decision making) and explanatory models (issue of proper category selection and computational restrictions) for analyzing market basket data, we next introduce a procedure that combines the specific merits of both approaches. Since segment-level results are intended, approaches that avoid early data aggregation are preferred in an exploratory step for complexity reduction. Thus, the estimation of segment-specific marketing-mix and cross-category effects on category choices is preceded by a data-driven strategy for basket construction and segment generation.

3 Analyzing segment-specific cross-category effects

The proposed analytical framework proceeds in the following stepwise manner: In a first step, shopping baskets from a customer transaction database are compressed onto a set of basket prototypes using a similar methodology as employed by Schnedlitz et al. (2001) and Decker and Monien (2003). These basket prototypes are constituting a 'generic' (i.e., customer-unspecific) classification of the observed market baskets which is characterized by more distinguished complementary cross-category coincidences within each of the derived shopping basket classes. In the subsequent stage of segment construction, we account for heterogeneity across customers by designating best-fitting basket class assignments. Finally, segment-specific adapted explanatory cross-category effect models including marketing-mix variables are estimated based on a multivariate logistic model specification similar to the Russell and Petersen (2000) approach.

3.1 Compression of market baskets and segment construction

As a starting point, for each customer $n = 1, ..., N$ included in a retail transaction database a sequence of t_n purchase incidence decisions across a set of J categories is observed. Consistent with prior work, these multicategory choice decisions are considered as 'pick-any/J' data (Manchanda et al. 1999; Russell and Petersen 2000). Hence, each shopping basket is represented as an J-dimensional binary vector $x_h = \{0, 1\}^J$, with h being a pointer to the elongated arrangement $\{t_1, t_2, ..., t_N\}$ of 'stacked' transaction sequences. This data format implies that utilization of the customer-specific

provenance of shopping baskets (indicated by x_h^n for the t_n transactions realized by customer n) is postponed to a later stage of the analysis.

The task of finding a partition of the data into a fixed number of K 'generic' basket classes $C = \{c_1, c_2, \ldots, c_K\}$ with more distinguished complementary joint purchase incidences within the detected classes requires resolution of the following objective function ('minimum dispersion criterion'):

$$\sum_k \sum_{h \in c_k} d(x_h, p(x_h)) \rightarrow \min_{C,P} \tag{1}$$

where $P = (p_1, p_2, \ldots, p_K)$ denotes a set of prototypes or centroids with $p_k \in \Re^J \forall k$ and $d(\cdot)$ being a distance measure. Minimization of (1) is also known as the principal point or K-centroids problem in the clustering and classification literature (Jain and Dubes 1988; Bock 1999). For any optimum configuration (C^*, P^*) the condition $p^*(x_h) = arg\,min\{d(x_h, p_k), k = 1, \ldots, K\}$ holds and warrants that each basket x_h is mapped onto it's minimum distant or closest prototype. Furthermore, when using the Euclidean distance metric it can be shown that the prototypes p_k^* are equal to class-specific means for the corresponding partition as generated by the optimal prototypes under stationarity conditions (Bock 1999).

Since the purchase incidences are encoded as (typically extremely sparse) binary vectors and we aim at detecting complementary cross-effects, the well-known asymmetric Jaccard coefficient giving more weight to joint purchases than to common zeros (i.e., non-purchases) is preferred here as a distance measure. A simple extension of the Jaccard coefficient for measuring the distance between a binary market basket vector and a real-valued prototype is given as follows:

$$d(x_h, p_k) = 1 - \frac{(x_h, p_k)}{||x_h||^2 + ||p_k||^2 - (x_h, p_k)}, \tag{2}$$

where (x_h, p_k) denotes the scalar product of vectors x_h and p_k. Notice that the subtrahend in expression (2) is often referred to as the Tanimoto similarity coefficient (Anderberg 1973).

The probably most prominent approach for solving the principal point problem is the iterative K-means clustering algorithm. Based on a given initial partition the K-means method minimizes criterion (1) recursively with respect to $C(\tau) \rightarrow P(\tau) \rightarrow C(\tau+1) \rightarrow P(\tau+1) \ldots$ and converges after a finite number of iterations τ. Albeit any arbitrary distance measure can be embedded in the algorithm (cf., MacQueen 1967; Anderberg 1973) it is predominantly implemented using Euclidean distances — hence, the term K-means. Though convergence to the next local minimum is guaranteed, the

quality of the final cluster solution heavily depends on the starting partition. To cope with this 'algorithmic variability' (Gordon and Vichi 1998; Hornik 2005), generation of cluster ensembles with different random initializations and subsequent selection of the 'best fitting' partition or heuristics for obtaining 'proper starting values' are recommended. Such strategies entailing the evaluation of multiple partitions, however, makes K-means methods computationally expensive and impractical when the number of data points becomes very large and high-dimensional, which is typically the case for shopping basket data derived from several hundred thousands of retail transactions and large assortment sizes.

Fortunately, there are also other methods available to solve the principal point problem. Descending from the field of machine learning, numerous 'online' versions of K-means type clustering are available known as competitive learning or vector quantization (VQ) algorithms (cf. Ripley 1996; Hastie et al. 2001). In contrast to 'off-line' K-means clustering, the VQ approach minimizes (1) via stochastic approximation. This is accomplished by directly manipulating the prototype system in a sequential updating scheme. Since only one single data point (e.g., a shopping basket accruing at the electronic retail POS check-out systems) is required at each iteration, adaptive VQ-type partitioning techniques are suitable to process data sets of practically unlimited size. The algorithm adopted here for market basket quantization proceeds as follows:

1. Start with a random initialization of the set of prototypes P by drawing K 'seed points' from the input data set.

2. Compute the distances between a randomly chosen market basket vector x_h and each prototype p_k according to (2).

3. Determine the minimum distant ('winning' or 'best fitting') prototype $d(x_h, p_k^*) = \min\{d(x_h, p_k), k = 1, \ldots, K\}$ to x_h.

4. Update the 'winning' prototype

$$p_k^* := p_k^* + \alpha(\tau)(x_h - p_k^*)$$

   where $\alpha(\tau)$ is a 'learning rate' monotonically decreasing with iteration time τ; to fulfill the conditions for stochastic approximation this is conceived such that $\lim_{\tau \to \infty} \alpha(\tau) = 0$.

5. Repeat steps 2-4 until convergence (i.e., if prototype improvements are becoming very small) or the pre-specified maximum number of iterations is reached.

Notice that the above VQ procedure differs from more conventional implementations in some respect. Due to data sparsity we are advocating Jaccard distances for determination of the 'winner' p_k^* but perform an Euclidean-like updating following step 4. The important practical reason for doing so is that after convergence we obtain centroids that coincide with respective class means and therefore can be interpreted as empirical expectations of observing a value of unity (cf., Leisch 2006). Consequently, in the present context, each j-element of an optimal prototype vector p_k^* denotes the purchase incidence probability of the corresponding product category within the 'generic' class of shopping baskets c_k^*. Exceptionally (un-)marked combinations of these class-conditional probabilities are indicative for (weaker) stronger cross-category purchase complementarities at the basket class level and will serve as a basis for further investigation.

As the term 'generic' suggests, the prototypes generated after convergence still apply to the pooled data set and does not yet recognize the customer identities behind the realized shopping baskets. To account for customer heterogeneity, knowledge about the customers' tendencies to fluctuate across the partition of basket classes is utilized. Next, construction of customer segments is based on a simple majority 'voting' for best-fitting class assignments. For each customer n we therefore calculate the following average distance-weighted number of basket class k assignments:

$$v_k^n := \frac{1}{t_n} \sum_{h=1}^{t_n} (1 - d(x_h^n, p_k^*))_{\{x_h^n \in c_k\}} \qquad \forall k \tag{3}$$

Though the sum across all K classes is not necessarily unity, this voting-measure is conceptually similar to fuzzy class memberships. In fact, v_k^n represents the 'degree of belongingness' of customer n to basket class k. Taking respective maximum values provide the final segmentation of customers:

$$s_k = \{n \in N | v_k^n = \max_{l=1,\ldots,K} (v_l^n)\} \tag{4}$$

where s_k indicates those customers whose past multicategory choice decisions can be characterized most accurately by the prototypical pattern represented by basket class k and are therefore assigned to the corresponding segment.

3.2 Explanatory analysis using multivariate logistic models

Utilizing the now available information on most distinguished categories responsible for prototype and subsequent segment construction, segment-specific multivariate logistic

models (cf. Hruschka 1991; Hruschka et al. 1999; Russell and Petersen 2000) are estimated in the second (explanatory) step of our procedure. A suitable model for segment k members $n \in s_k$ utilizes shopping baskets comprising categories corresponding to the top elements of basket prototype p_k. To obtain a model close to standard approaches of describing choice decisions (with respect to random utility theory), a utility function including marketing-mix parameters and household specific variables is chosen. Using an extended version of a multivariate logistic model (Boztuğ and Hildebrandt 2006), the utility function U has the following form

$$\begin{aligned} U(i,n,t) &= \beta_i + \delta_{1i} \ln[\text{TIME}_{int} + 1] + \delta_{2i}\text{LOYAL}_{in} \\ &\quad + \gamma_i \ln(\text{PRICE}_{int}) + \xi_i \text{DISPLAY}_{int} + \sum_{i \neq j} \theta_{int} C(j,n,t) + \epsilon_{int} \\ &= V(i,n,t) + \epsilon_{int} \end{aligned} \tag{5}$$

with category i, consumer n and time t. β is a category dummy variable and θ the cross-category parameter. The stochastic error term ϵ_{int} is assumed extreme value distributed, as in a standard multinomial logit (MNL) model. The utility in (5) is close to a standard MNL model for a single category; where the cross-category-term is used to cope for cross-category dependence. $C(j,n,t)$ is a binary variable, which is one if consumer n purchases category j at time t and zero otherwise.

Household specific variables are the time and a measure of loyalty for each category, where TIME is the time in weeks since the last purchase for a consumer in the category. LOYAL is defined as $\text{LOYAL}_{in} = \ln \frac{m(i,n)+0.5}{m(n)+1}$. $m(n)$ accounts for the purchases of a consumer in the initial period, and $m(i,n)$ is the number of purchases in category i during the initial period. LOYAL is a measure for the loyalty for one specific category of a consumer.

The marketing-mix variables are price and display. PRICE is described by an index of prices of a category by calculating the mean of prices of all purchased products in a specific category during one week. DISPLAY is the mean number of available displays per category calculated for each week. The cross-category variable θ is decomposed by $\theta_{ijn} = \psi_{ij} + \eta\,\text{SIZE}_n$ with SIZE being the mean basket size for consumer n in the initial period. θ is assumed as symmetric, so ψ has to be constrained to be symmetric. $X(i,b)$ is a 0-1-coded dummy variable, which takes the value of one, if category i is included in basket b and zero otherwise. Here, only the choice of a product in a specific category is inspected, but not the inner-category choice.

The probability choosing one specific category, conditional on the choices in the

other categories, can be expressed as

$$P(C(i,n,t)=1|C(j,n,t) \text{ for } j \neq i) = \frac{1}{1+\exp(-V(i,k,t))}. \tag{6}$$

The market basket of a consumer n at time t is described by an q-tuple $B(n,t)$ with $B(n,t) = \{C(1,n,t),\ldots,C(q,n,t)\}$, $C(i,n,t)=1$ if consumer n purchases in category i at time t. This kind of choice representation induces 2^q different baskets. We exclude the Null basket (no choice in any category) in our analysis, so we end up with 2^q-1 possible baskets. Using Besag's Factorization theorem (Besag 1974; Cressie 1991), the utility function (6) and the binary description of a choice for a category, the probability of choosing a specific basket b is (Russell and Petersen 2000)

$$\begin{aligned} P(B(n,t)=b) &= \frac{\exp\{\mu(b,n,t)\}}{\sum_{b^*}\exp\{\mu(b^*,n,t)\}} \\ \mu(b,n,t) &= \sum_i \beta_i X(i,b) + \sum_i (\delta_{1i}\ln[\text{TIME}_{int}+1] + \delta_{2i}\text{LOYAL}_{in})X(i,b) \\ &\quad + \sum_i (\gamma_i \ln(\text{PRICE}_{int}) + \xi_i \text{DISPLAY}_{int})X(i,b) \\ &\quad + \sum_{i<j} \theta_{ijn} X(i,b)X(j,b) \end{aligned} \tag{7}$$

The model in (7) looks like a standard MNL approach with an additional cross-effects term described by θ_{ijn}. It should be kept in mind, that this model is not a result of an extension of a standard model, but is derived using methods from spatial statistics. To explain the different outcomes of $\mu(b,n,t)$ in (7), we present in table 1 exemplary a two-category case with only TIME and PRICE as explanatory variables. The θ parameter is only present if both categories are purchased simultaneously. It measures a bivariate relationship, which obtains more than one time in a basket containing at least three categories.

		Purchase in category 1	
		yes	no
Purchase in category 2	yes	$\beta_1 + \text{TIME}_{1nt} + \text{PRICE}_{1nt}$ $+\beta_2 + \text{TIME}_{2nt} + \text{PRICE}_{2nt}$ $+\theta_{12}$	$\beta_2 + \text{TIME}_{2nt} + \text{PRICE}_{2nt}$
	no	$\beta_1 + \text{TIME}_{1nt} + \text{PRICE}_{1nt}$	0

Table 1: Values for $\mu(b,n,t)$ in a two-category case

For managers, not only the parameter estimates are important, but especially cross-price elasticities. The price elasticities are defined relative to categories, but not to baskets. The sum over all baskets containing category i is named as $BC(i)_{nt}$, whereas $BC(i,j)_{nt}$ contains all baskets with category i and j. The summation over all possible baskets (including the null basket) is described as $BC(\text{all})_{nt}$. Therefore, the probability of choosing one basket, which includes category i is

$$\Lambda(i,j)_{nt} = \frac{BC(i)_{nt}}{BC(\text{all}_{nt})} \tag{8}$$

and for a basket containing category i and j

$$\Lambda(i,j)_{nt} = \frac{BC(i,j)_{nt}}{BC(\text{all}_{nt})} \tag{9}$$

The cross-price elasticities are defined as the percentage in change selecting category i with respect to a change in category j as

$$\text{E}(i,j)_{nt} = \frac{\partial(\log \Lambda(i)_{nt})}{\partial(\log \text{PRICE}_{jnt})} \tag{10}$$

This leads to the following expressions calculating the own and cross-price elasticities

$$\begin{aligned} \text{E}(i,i)_{nt} &= \gamma_i(1 - \Lambda(i)_{nt}) \\ \text{E}(i,j)_{nt} &= \gamma_j \Lambda(j)_{nt}(S(i,j)_{nt} - 1), \quad i \neq j \\ \text{with} \\ S(i,j)_{nt} &= \frac{\Lambda(i,j)_{nt}}{\Lambda(i)_{nt}\Lambda(j)_{nt}} \end{aligned} \tag{11}$$

In expression (11), γ_i and γ_j are expected to be negative (as usually is expected for price parameters). If they are not negative, they are set to a negative value. The own price elasticities are always negative, whereas the cross-price elasticities can be negative or positive as well. A negative elasticity implies a complementary relationship, a positive one a substitutional association between the inspected categories.

4 Empirical Application

Notice that from a data analytical standpoint the type of data illustrated in the introduction of this paper is equivalent to traditional household scanner data, with the

notable difference that they do not cover competitive information. For illustration purposes of our approach, we therefore use the well-known ZUMA data set set[1]. A total number of 470,825 retail transactions with pick-any choices among an assortment of $J = 65$ categories reported from 4,424 households over a period of one year were first subject to the data compression step and subsequent segment formation. The data contains information about the purchase date and which product was chosen by whom (and therefore also the chosen category). Additionally, we know how many items were purchased at which price and if the product was placed on a display or not. Almost all categories, which are present in a common supermarket, are reported, with the exemption of fresh products as meat and fruits. So it is possible to describe daily purchases containing all regular purchased items by a standard household. After an examination of the derived classification of shopping baskets and household segments, in the following we compare parameter estimates for segment-specific multivariate logistic model specifications with those resulting from an aggregate counterpart.

4.1 Construction of basket classes and household segments

In the clustering literature many authors expressed their doubts about the existence of 'quasi-natural' groupings in empirical data sets (cf., e.g., Dubes and Jain 1979; Aldenderfer and Blashfield 1984). Even though one is apt to accept this assumption, it is very unlikely that this 'natural' grouping is detectable with an efficiently manageable and managerially acceptable number of classes in light of the excessively large and high-dimensional data set of joint category purchase incidences at hand. In fact, finding a number of classes that balances adequate fit with the data (in terms of low within-class dispersion) and parsimony is not an easy task. Numerous heuristics exist to help the analyst in this respect (for a comparative overview see Milligan and Cooper 1985; Dimitriadou et al. 2002). Once combined, however, they often yield ambiguous or even contradictory recommendations. Nevertheless, in order to avoid obvious inferior solutions, the derived partition of shopping baskets can be required to be 'structurally stable' in a sense that replications of the same algorithm on different samples from the

[1]The data used for this analysis are part of a subsample of the 1995 GfK ConsumerScan Household panel data and were made accessible by ZUMA. The ZUMA data set includes all households having continuously reported product purchases during the entire year 1995. For a description of this data set cf. Papastefanou, G. 2001. The ZUMA data file version of the GfK ConsumerScan Household Panel. In: Papastefanou, G., Schmidt, P., Börsch-Supan, A., Lüdkte, H., Oltersdorf, U. (Eds.), Social and Economic Analyses of Consumer Panel Data, Zentrum für Umfragen, Meinungen und Analysen (ZUMA), Mannheim, pp. 206-212.

data set return similar partitions (Strehl and Ghosh 2002; Hornik 2005).

To cope with the size of the data set, we split it randomly into several smaller subsets and used those for successive clusterings similar to the CLARA (Clustering LARge Applications) procedure by Kaufman and Rousseeuw (1990). After each clustering, a classification of the entire data is accomplished by assigning each of the remaining shopping basket not belonging to the current sample to the class represented by the closest prototype. The k-medoid partitioning method employed within the standard CLARA procedure, however, was substituted by the above described VQ algorithm. Furthermore, each VQ replication was initialized with the 'optimal' prototypes for the previous sample as long as the partitioning quality of the entire data set is further enhanced. To measure the quality of the current classification the average Jaccard distance between each basket and it's 'best-fitting' prototype is computed. Hence, the prototype system is allowed to be continuously improved until the overall classification quality degrades (which is usually the case after a few iterations).

Given the number of classes K, 100 reiterations of this procedure yield a collection of individual solutions. For a sequence of increasing K these 'cluster ensembles' (Hornik 2005) can serve as a basis for further inspection of structural stability. As a measure of partition agreement the popular Rand index (Rand 1971; Hubert and Arabie 1985) was used to compare each possible pair of the K partitions. The box plots depicted in figure 1 nicely illustrate that the correspondence between partitions (and hence, the stability) is dramatically improved with increasing number of classes.

Representative for the various measures of internal cluster validity, we computed the statistic proposed by Davies and Bouldin (1979) to fortify the decision on a suitable number of classes. A traditional approach is to plot the index values by number of classes and to hope that an obvious 'elbow' or kink indicating the correct number of classes is observable. Though this is usually done by visual inspection, it can be formalized by looking at the most significant local peak of the index curve (Thorndike 1953). Using the procedure described by Dimitriadou et al. (2002) we determined the recommended class number based on this 'elbow-heuristic' for the complete set of cluster ensembles. The resulting distribution of recommendation frequencies is shown in figure 2. As expected, no clear recommendation in favor of a specific number of classes can be derived from this picture.

Bearing in mind that from a practitioner's view partitions with 20 or even more classes are becoming managerially prohibitive, priority is given to solutions with smaller class numbers but still structurally stable partitioning results. Putting the available

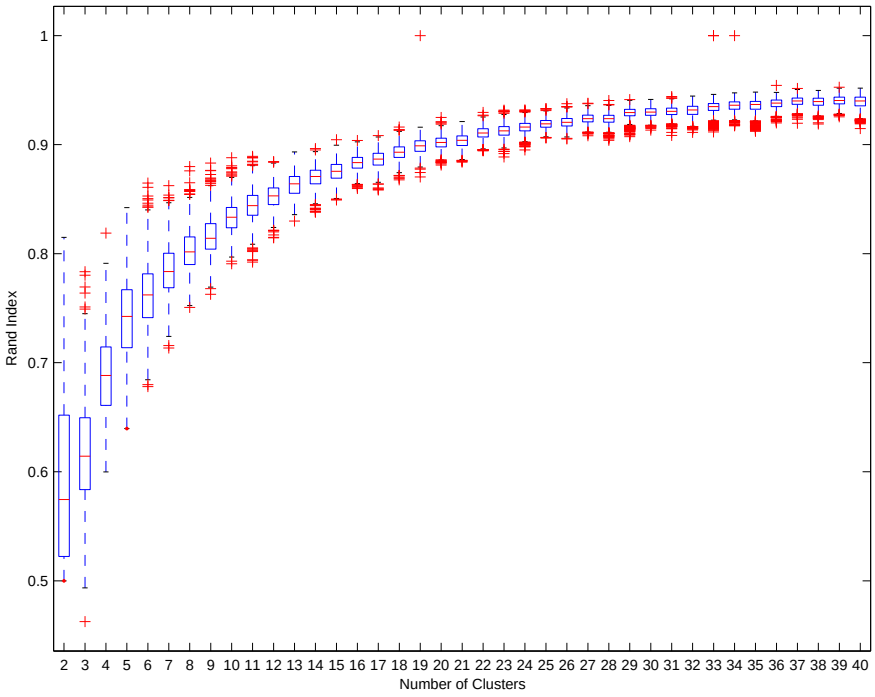

Figure 1: Distribution of the Rand index for increasing number of classes

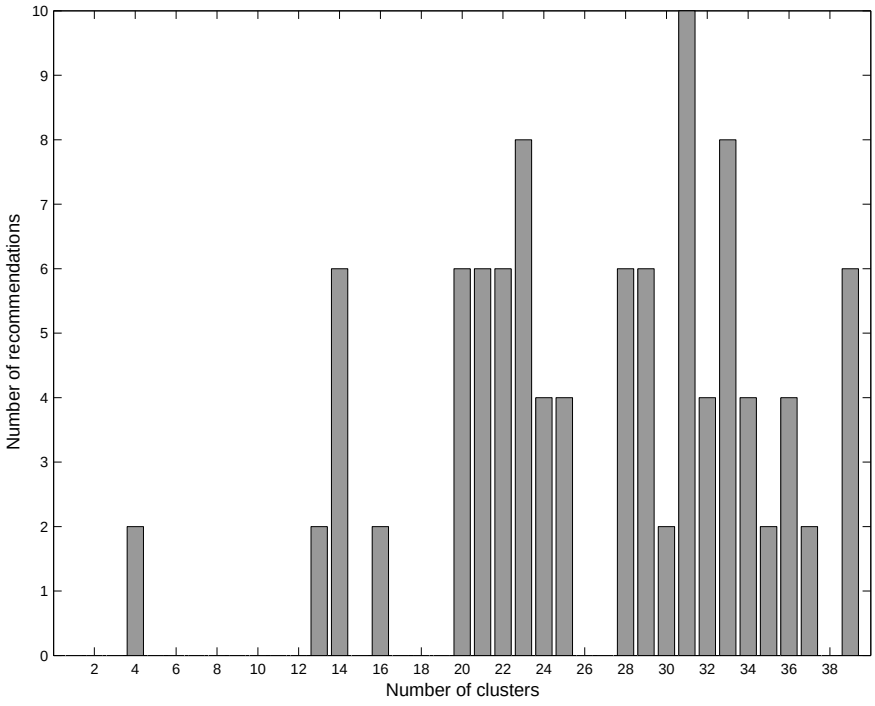

Figure 2: Number of classes recommendations based on the Davies-Bouldin statistic

pieces of information together, a number of $K = 14$ basket classes seems to provide a decent and adequate representation of the observed shopping baskets. Hence, we further elaborate on this solution for the data compression step of the proposed pro-

cedure.

Seg.	Most distinguished complementary product categories	Relative size (%)	
k	(top five j-elements of prototype p_k)	basket	household
1	**Milk**, soft cheese, curd, coffee, soft drinks	13.2	20.2
2	**Cream**, *milk*, *soft cheese*, *curd*, *yogurt*	13.3	15.2
3	**Yogurt**, *milk*, *curd*, *soft cheese*, soft drinks	11.9	14.4
4	**Hard cheese**, *soft cheese*, *milk*, *yogurt*, *curd*	11.6	11.5
5	*Soft cheese*, toilet paper, wine, cereals, instant coffee	5.0	0.4
6	**Curd**, soft cheese, pudding, cling films, coffee cream	3.5	1.2
7	*Coffee*, *coffee cream*, spirits, filter paper, soft cheese	6.2	5.0
8	**Pet food**, milk, milk products, coffee, soft cheese	3.7	5.5
9	Toothpaste, detergent, bath add., soap, dishwashing l.	5.1	0.1
10	**Water**, *beer*, milk, lemonade, coffee	9.9	16.0
11	**Soft drinks**, *water*, lemonade, soft cheese, milk	5.2	2.8
12	**Beer**, milk, soft drinks, lemonade, coffee	5.1	7.4
13	*Frozen veget.*, *ice*, froz. cookies, instant meal & fish	3.6	0.1
14	*Tea*, *cola drinks*, mayonnaise, lemonade, soft drinks	2.7	0.3

Bold: Class-conditional purch. prob. $p_{jk} > .75$; *Italic*: Class-conditional purch. prob. $p_{jk} > .25$

Table 2: Main characteristics of shopping basket classes and household segments

Table 2 provides a summary of the most important features of the derived shopping basket classes and corresponding household segments. As a result of the exploratory part of our procedure, each basket class can now be best characterized by it's generic profile of prototypical category purchase probabilities, with combinations of particularly outstanding values signalling stronger degrees of cross-category purchase complementarities. Hence, further examination of those categories exhibiting highest class-conditional purchase incidences in the subsequent (explanatory) step for estimating segment-specific cross-category effects models is recommended. In table 2 a selection of those five categories represented with the highest respective prototype values is highlighted for each of the basket classes. Quite obviously, they can be further organized into two different substructures: One is characterized by differential combinations of various dairy products (classes no. 1 to 4) and another is dominated by categories of beverages (classes no. 10 to 12). Most of the remaining classes are representing either some mixture types of the former or are marked by strongly discriminating product categories like pet food, etc. The last two columns of table 2 also provide information on the relative magnitudes of basket classes and corresponding segments. Although partly considerable differences can be observed (which is due to the specific assignment

rule adopted for segment construction), the two substructures can be clearly detected both on the generic basket class and segment level.

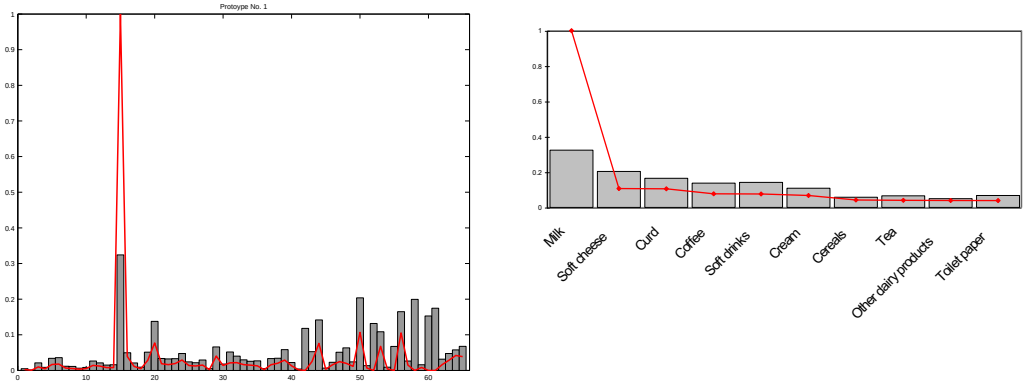

Figure 3: Category choice probabilities according to prototype no. 1

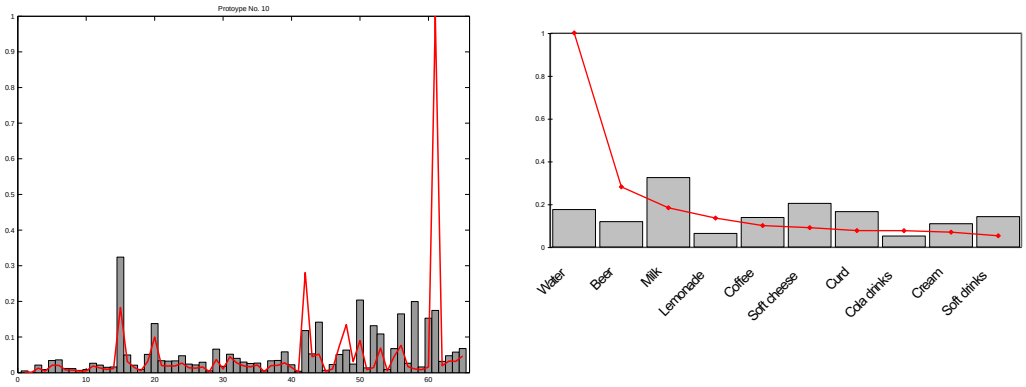

Figure 4: Category choice probabilities according to prototype no. 10

Let us therefore draw our further attention on two representative segments out of these diverse substructures, namely segment no. 1 and segment no. 10. Consider, for example, the pictorial representation of the before-mentioned prototypical profile of category choice probabilities representing household segment no. 1 according to the solid line in the right-hand side graph of figure 3. Instead, the grey bars are representing the unconditional purchase probabilities. From the right-hand side graph in the same figure (emphasizing the top ten categories in terms of class-conditional probabilities) it becomes obvious that the purchase behavior of this segment of households is clearly dominated by remarkably high purchase incidences of the milk category and only moderate class-conditional choice probabilities in the remaining dairy categories. Albeit different with regard to the dominance of only one single category, household segment no. 10 is marked by high purchase incidences in the equally dominating mineral water category, followed by the beer and lemonade categories (see figure 4). Milk,

in contrast, is expected to be chosen less frequently by the households assigned to this segment as compared to the aggregate (segment-unconditional) case.

Of course, other basket classes are characterized by their own prototypical basket compositions (i.e., cross-category purchase interdependencies) that are clearly distinctive from those further investigated here.

4.2 Investigating segment-specific versus aggregate cross-category effects

As already claimed, one of the primary purposes of the data compression step employed prior to the model-based estimation of cross-category purchase effects is the reduction of model complexity and to provide a data driven strategy for selection of categories that are meaningful and relevant to a specific segment of households. Consequently, estimation of segment-specific multivariate logistic models according to stage two of our procedure was restricted to the respective most distinguished product categories including associated marketing-mix variables.

For illustration purposes, we exemplary describe the estimation results of the multivariate logistic models for two household segments only, namely for the above described segments no. 1 and no. 10. All other results are available from the authors upon request. For segment no. 1, we inspected the five top categories according it's corresponding prototype (milk, soft cheese, curd, coffee, and soft drinks), while for segment no. 10 we used four categories (water, beer, milk, and lemonade). 893 households are members of segment no. 1 with a total number of 117,570 purchase occasions. Out of these at least one of the previously selected five categories was purchased on 89,340 occasions. Segment no. 10 comprises 709 households with a total number of 69,736 transactions and 38,912 purchase occasions containing at least one of the four categories of interest.

First, we present the parameter estimates for both prototypes for all consumers and for the segment-specific consumers in table 3.

	Prototype 1		Prototype 10	
	all	segment	all	segment
β_1	−3.60	12.48	−10.32	−7.40
	(2.01)	(6.32)	(1.90)	(5.78)
β_2	−0.71	−1.22	14.04	19.99
	(0.09)	(0.20)	(2.41)	(5.97)
β_3	−1.08	−1.61	−3.56	0.43
	(0.47)	(0.98)	(2.29)	(4.98)
β_4	−0.58	−1.26	1.62	3.73
	(0.19)	(0.41)	(1.12)	(2.92)
β_5	−2.55	−4.64		
	(0.55)	(1.18)		
δ_{11}	−0.79	−1.02	−0.29	−0.30
	(0.02)	(0.07)	(0.02)	(0.07)
δ_{12}	−0.30	−0.21	−0.36	−0.45
	(0.02)	(0.03)	(0.02)	(0.05)
δ_{13}	−0.37	−0.43	−0.84	−0.59
	(0.02)	(0.03)	(0.02)	(0.04)
δ_{14}	−0.04	0.04	−0.49	−0.79
	(0.02)	(0.03)	(0.02)	(0.06)
δ_{15}	−0.54	−0.47		
	(0.02)	(0.03)		
δ_{21}	0.67	0.73	0.48	0.17
	(0.01)	(0.04)	(0.01)	(0.03)
δ_{22}	0.80	0.81	0.51	0.57
	(0.01)	(0.02)	(0.01)	(0.02)
δ_{23}	0.66	0.57	0.70	0.51
	(0.01)	(0.02)	(0.01)	(0.017)
δ_{24}	0.87	0.84	0.59	0.63
	(0.01)	(0.02)	(0.01)	(0.02)
δ_{25}	0.46	0.43		
	(0.01)	(0.01)		

	Prototype 1		Prototype 10	
	all	segment	all	segment
γ_1	-1.90	4.52	-3.84	-3.42
	(0.88)	(2.76)	(0.76)	(2.32)
γ_2	3.76	4.79	9.05	12.72
	(0.58)	(1.23)	(1.46)	(3.60)
γ_3	-0.91	-1.33	-2.18	-0.02
	(0.51)	(1.06)	(1.00)	(2.18)
γ_4	-0.08	0.34	1.38	2.20
	(0.30)	(0.65)	(0.50)	(1.31)
γ_5	-0.88	-1.84		
	(0.30)	(0.64)		
ξ_1	0	0	1.07	0.26
	(-)	(-)	(0.77)	(2.37)
ξ_2	0	0	-0.84	-1.86
	(-)	(-)	(0.53)	(1.31)
ξ_3	0	0	0	0
	(-)	(-)	(-)	(-)
ξ_4	3.88	4.29	1.42	0.65
	(0.55)	(1.21)	(0.54)	(1.43)
ξ_5	2.07	1.46		
	(0.54)	(1.16)		
ψ_{12}	-0.27	-0.35	-0.13	-0.36
	(0.01)	(0.04)	(0.01)	(0.05)
ψ_{13}	-0.14	-0.14	-1.25	-1.37
	(0.01)	(0.04)	(0.02)	(0.04)
ψ_{14}	-0.46	-0.48	-0.05	-0.45
	(0.01)	(0.04)	(0.02)	(0.05)
ψ_{15}	-0.18	-0.25	-0.91	-0.49
	(0.01)	(0.04)	(0.02)	(0.04)
ψ_{23}	0.16	0.31	0.07	0.04
	(0.01)	(0.03)	(0.02)	(0.05)
ψ_{24}	-0.27	-0.06	-0.73	-0.43
	(0.01)	(0.03)	(0.02)	(0.04)

	Prototype 1		Prototype 10	
	all	segment	all	segment
ψ_{25}	-0.09	0.15		
	(0.01)	(0.03)		
ψ_{34}	-0.27	-0.08		
	(0.01)	(0.03)		
ψ_{35}	-0.13	0.14		
	(0.01)	(0.03)		
ψ_{45}	-0.22	0.06		
	(0.01)	(0.03)		
η	0.34	0.32	0.68	0.83
	(0.01)	(0.02)	(0.01)	(0.03)

Table 3: Parameter estimates for categories selected according to prototype 1 and prototype 10, for all and segment-specific consumers, the standard errors are given in parentheses

In the first two columns of table 3, the estimation results for the top-five categories (namely milk, soft cheese, curd, coffee, and soft drinks) of prototype 1 are shown. For all categories, a positive and significant loyalty parameter ($\delta_{2\cdot}$) is estimated. Comparing all consumers with the ones for the segment-specific consumers, the estimated loyalty parameters are smaller, except for the milk category. Both effects can be explained regarding the growing respectively falling percentage of purchasing the inspected categories. All estimated significant time parameters ($\delta_{1\cdot}$) are negative, but in the comparison, no clear relationship between all and the segment-specific consumers can be detected. The price parameters (γ) are larger measured in absolute values for the segment-specific consumers, this indicates a higher price sensibility for them. For the rarely significant display parameters (ξ), no clear direction of the values can be given. The cross-effects parameters (ψ and η) will be discussed in more detail regarding hypothesis 1 (which will be presented later on).

For the four most distinguished categories of prototype 10 (namely water, beer, milk, and lemonade), the estimation results are given in the last two columns of table 3. We find similar results as for prototype 1. Especially, the price sensitivity (γ) is again much higher (in absolute values) for the segment-specific consumers.

Turning now to a deeper inspection of the cross-category effects, we formulate the

following statement:

Hypothesis 1: Segment-specific cross-effects (θ) are higher than those for all consumers.

This hypothesis is based on the assumption that segment-specific households were responsible for prototype construction. So their cross-category purchase patterns are more revealed as for all consumers. Notice that cross-effects in table 4 and in table 5 are computed using the formula for cross-effects ($\theta_{ijn} = \psi_{ij} + \eta\,\mathrm{SIZE}_n$) for an average consumer (using the mean value for SIZE). We verify this hypothesis for all cross-category relationships, except for milk (see table 4).

	Milk	Soft cheese	Curd	Coffee	Soft drinks
Milk		0.501	0.636	0.313	0.593
Soft cheese	0.404		0.929	0.504	0.681
Curd	0.608	1.058		0.502	0.640
Coffee	0.268	0.694	0.669		0.557
Soft drinks	0.504	0.903	0.890	0.814	

Table 4: Cross-effects of prototype 1 for an average consumer, in the upper triangle values for all consumers, in the lower one for segment-specific ones

One possible reason could be that the class conditional purchase probability for milk is one and substantially smaller for all consumers. Therefore, the within-segment cross-effects for milk are lower regarding all consumers, because compared to the other categories, joint purchases with other categories become less. Not surprisingly, all cross-effects imply a complementary relationship, because the cross-category values are positive. This result is as expected due to the prior data compression step, which focuses on jointly purchased categories.

	Water	Beer	Milk	Lemonade
Water		0.975	−0.142	1.056
Beer	1.088		0.200	1.161
Milk	0.081	0.959		0.383
Lemonade	0.997	1.488	1.025	

Table 5: Cross-effects of prototype 10 for an average consumer, in the upper triangle values for all consumers, in the lower one for segment-specific ones

Hypothesis 1 is also inspected for prototype 10 in table 5. It is confirmed for the cross-effects of beer with all other categories, and for milk with all others. It could

only not be shown for the relationship between water and lemonade. Interestingly, a negative cross-effect value which induces a substitutional relationship between milk and water for all consumers changes to a positive one (a complementary relationship) for the segment-specific consumers. The change in the relationship between these two categories is quite interesting, because it shows that a segment-specific examination could lead to different results regarding an overall view. So managers assume by inspecting the all-consumer-results, that the consumers do decide between the purchase of water and milk, but within the target group of consumers (the segment specific) the opposite is true.

For merchandise managers, the cross-price elasticities (presented from table 6 to table 9) are even more interesting. The values represent the change in percent of the share of choice of the row category for a 1 % price increase in the column category. The elasticities account for consumer heterogeneity and can be interpreted as the average elasticities per week. Negative values imply a complementary relationship between the inspected categories (as was also reported in the cross-category values in table 4 and table 5).

Hypothesis 2: Segment-specific cross-price elasticities are higher than for all consumers.

This proposition is made because segment-specific consumers are by definition jointly purchasing the categories under investigation more often than the 'average' consumers and therefore being more affected by price changes. This hypothesis can be verified for all categories, except for milk for the prototype 1 (see table 6 for all consumers and table 7 for the segment-specific consumers[2]).

	Milk	Soft cheese	Curd	Coffee	Soft drinks
Milk	−0.333	−0.071	−0.021	−0.001	−0.019
Soft cheese	−0.040	−0.837	−0.038	−0.002	−0.029
Curd	−0.052	−0.169	−0.222	−0.002	−0.032
Coffee	−0.030	−0.109	−0.028	−0.022	−0.030
Soft drinks	−0.052	−0.144	−0.035	−0.003	−0.248

Table 6: Cross-price elasticities of prototype 1 for all consumers

For milk, the change in the purchase probability is higher for all consumers. One

[2]For curd and coffee, nonsignificant values were estimated, but we used them to calculate the elasticities. Also, some price parameters are positive, which leads to a wrong sign in the elasticities, so we changed them to negative in the table but not in the calculation.

	Milk	Soft cheese	Curd	Coffee	Soft drinks
Milk	−0.289	−0.028	−0.009	0.001	−0.011
Soft cheese	−0.043	−1.249	−0.079	0.016	−0.105
Curd	−0.063	−0.169	−0.463	0.022	−0.149
Coffee	−0.037	−0.274	−0.081	0.114	−0.125
Soft drinks	−0.054	−0.329	−0.097	0.022	−0.604

Table 7: Cross-price elasticities of prototype 1 for segment-specific consumers

reason could be that the segment-specific consumers do nearly always buy milk, independently from the other categories. So a price change in the other categories do effect them less in comparison to all consumers.

In prototype 10, the hypothesis 2 is fulfilled for most of the cross-price elasticities (see table 8 and table 9). Only in the beer or milk category price changes have a smaller impact on the choice share for the segment-specific consumers. This phenomenon can be explained again because both categories are purchased anyway.

	Water	Beer	Milk	Lemonade
Water	−1.332	−0.542	−0.046	−0.047
Beer	−0.264	−2.935	−0.001	−0.053
Milk	−0.004	−0.001	−0.042	−0.001
Lemonade	−0.376	−0.874	−0.005	−0.628

Table 8: Cross-price elasticities of prototype 10 for all consumers

	Water	Beer	Milk	Lemonade
Water	−1.008	−0.584	−0.000	−0.019
Beer	−0.216	−3.504	−0.001	−0.321
Milk	−0.047	−0.694	−0.007	−0.034
Lemonade	−0.240	−1.105	−0.001	−1.091

Table 9: Cross-price elasticities of prototype 10 for segment-specific consumers

5 Discussion and outlook

We proposed and empirically illustrated a two-stage procedure that combines features from exploratory with explanatory models for market basket analysis. It can be shown that the employed data compression step is capable to identify customer segments

with internally more distinctive and distinguished complementary cross-category interdependencies as compared to the aggregate case. Moreover, in the second stage of the proposed procedure, significantly different cross-effects and related cross-price elasticities both across previously determined segments and compared to the 'average' customer can be detected.

Both marketing analysts and retail marketing managers can directly benefit from the proposed methodology in at least two ways: First, a data-driven strategy for selecting product categories to be included in models for predicting cross-category effects is provided. The data compression task warrants that the selected categories adequately represent the meaningful (sub-)structures of consumers multicategory decision making processes. Secondly, information on segment-specific cross-category dependencies and associated marketing-mix effects becomes available. Retail marketing managers making use of this information can thus be assisted in designing targeted direct marketing actions within their loyalty programs.

As a useful side effect, the procedure could also be potentially useful as a framework for partitioning a retailer's overall (and typically considerably large) portfolio of product categories into smaller sub-portfolios as required in the category management process. This could be accomplished by collecting the most distinguished categories responsible for the formation of 'adjacent' (e.g., for meaningful substructures of) basket classes. These categories can be shown to be more 'homogeneous' in terms of independence with categories not included in a specific sub-portfolio and thus may be managed more easily. Furthermore, retailers would be enabled to customize their marketing decisions including pricing and promotional activities for each corresponding customer segment to optimize profits across these sub-portfolios (see Manchanda et al. 1999).

Regarding the construction of customer segments, the proposed approach is flexible enough to account for any (stronger or weaker) degree of cross-category complementarities simply by introducing user-defined threshold weights in the voting scheme adopted in the segment formation step. Finally, in order to expand the empirical performance and to fine-tune the proposed procedure to other settings, further application studies using different data sets including personalized retail transaction data for a variety of retail industries can be recommended.

References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I. 1995. Fast Discovery of Association Rules. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press and The MIT Press, Menlo Park, CA, pp. 307-328.

Ainslie, A., Rossi, P.E. 1998. Similarities in choice behavior across product categories, Marketing Science 17 91-106.

Aldenderfer, M.S., Blashfield, R.K. 1984. Cluster Analysis, Sage Publications, Beverly Hills, USA.

Andrews, R.L., Currim, I.S. 2002. Identifying segments with identical choice behaviors across product categories: An Intercategory Logit Mixture model, International Journal of Research in Marketing 19 65-79.

Anderberg, M.R. 1973. Cluster Analysis for Applications, Academic Press, New York.

Anand, S., Patrick, A.R., Hughes, J.G., Bell, D.A. 1998. A Data Mining Methodology for Cross-Sales, Knowledge Based Systems 10 449-461.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems, Journal of the Royal Statistical Society, Series B 36 192-236.

Bock, H.-H. 1999. Clustering and Neural Network Approaches. In: Gaul, W., Locarek-Junge, H. (Eds.), Classification in the Information Age, Springer, Heidelberg, pp. 42-57.

Böcker, F. 1978. Die Bestimmung der Kaufverbundenheit von Produkten, Duncker und Humblot, Berlin.

Boztuğ, Y., Hildebrandt, L. 2006. A Market Basket Analysis Conducted with a Multivariate Logit Model. In: Gaul, W., Kruse, R. (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, New York (forthcoming).

Boztuğ, Y., Silberhorn, N. 2006. Modellierungsansätze in der Warenkorbanalyse im Überblick, Working Paper, Humboldt-Universität zu Berlin.

Brin, S., Siverstein, C., Motwani, R. 1998. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules, Data Mining and Knowledge Discovery 2 39-68.

Chen, Y.-L., Tang, K., Hu, Y.-H. 2005. Market basket analysis in a multiple store environment, Decision Support Systems 40(2) 339-354.

Chib, S., Seetharaman, P.B., Strijnev, A. 2002. Analysis of multi-category purchase incidence decisions using IRI market basket data. In: Franses, P.H., Montgomery, A.L. (Eds.), Econometric Models in Marketing, Volume 16, Elsevier Science, Amsterdam, pp. 57-92.

Cressie, N.A.C. 1991. Statistics for spatial data, John Wiley Sons.

Davies, D.L., Bouldin, D.W. 1979. A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 224-227.

Decker, R., Monien, K. 2003. Market basket analysis with neural gas networks and self-organising maps, Journal of Targeting, Measurement and Analysis for Marketing 11 (4) 373-386.

Decker, R. 2005. Market Basket Analysis by Means of a Growing Neural Network, The International Review of Retail, Distribution and Consumer Research 15 (2) 151-169.

Dickinson, R., Harris, F., Sircar, S. 1992. Merchandise compatibility: An exploratory study of its measurement and effect on department store performance, International Review of Retail, Distribution and Consumer Research 2 (4) 351-379.

Dimitriadou, E., Dolnicar, S., Weingessel, A. 2002. An examination of indexes for determining the number of clusters in binary data sets, Psychometrika 67 137-160.

Dubes, R., Jain, A.K. 1979. Validity studies in clustering methodologies, Pattern Recognition 11 235-254.

Gordon, A.D., Vichi, M. 1998. Partitions of partitions, Journal of Classification 15 265-285.

Hahsler, M., Hornik, K., Reutterer, T. 2006. Implications of Probabilistic Data Modeling for Mining Association Rules. In: Gaul, W., Kruse, R. (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, New York (forthcoming).

Hastie, T., Tibshirani, R., Friedman, J. 2001. The Elements of Statistical Learning, Springer, New York.

Hornik, K. 2005. Cluster ensembles. In: Weihs, C., Gaul, W. (Eds.), Classification - The Ubiquitous Challenge, Springer, Heidelberg, pp. 65-72.

Hruschka, H. 1991. Bestimmung der Kaufverbundenheit mit Hilfe eines probabilistischen Meßmodells, Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung 43 418-434.

Hruschka, H., Lukanowicz, M., Buchta, Ch. 1999. Cross-category sales promotion effects, Journal of Retailing and Consumer Services 6 99-105.

Hubert, L., Arabie, P. 1985. Comparing Partitions, Journal of Classification 2 193-218.

Jain, A.K., Dubes, R.C. 1988. Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs.

Julander, C-R. 1992. Basket Analysis. A New Way of Analyzing Scanner Data, International Journal of Retail and Distribution Management 20 10-18.

Kaufman, L., Rousseeuw, P.J. 1990. Finding Groups in Data. An Introduction to Cluster Analysis, Wiley, New York.

Lattin, J.M., Gooley, C., Lal, R., Padmanabhan, V. 1996. Category Coincidence in Grocery Market Baskets, Working Paper, Graduate School of Business, Stanford University.

Leisch, F. 2006. A toolbox for k-centroids cluster analysis, Computational Statistics and Data Analysis (forthcoming).

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In: Le Cam, L.M., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, pp. 281-297.

Manchanda, P., Ansari, A., Gupta, S. 1999. The "shopping basket": A model for multi-category purchase incidence decisions, Marketing Science 18 95-114.

Mild, A., Reutterer, T. 2003. An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data, Journal of Retailing and Consumer Services 10 (3) 123-133.

Milligan, G.W., Cooper, M.C. 1985. An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 159-179.

Müller-Hagedorn, L. 1978. Das Problem des Nachfrageverbundes in erweiterter Sicht, Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung 3 181-193.

Papastefanou, G. 2001. The ZUMA data file version of the GfK ConsumerScan Household Panel. In: Papastefanou, G., Schmidt, P., Börsch-Supan, A., Lüdkte, H., Oltersdorf, U. (Eds.), Social and Economic Analyses of Consumer Panel Data, Zentrum für Umfragen, Meinungen und Analysen (ZUMA), Mannheim, pp. 206-212.

Passingham, J. 1998. Grocery retailing and the loyalty card, Journal of Market Research Society 40 (January) 55-63.

Rand, W.M. 1971. Objective Criteria for the Evaluation of Clustering Methods, Journal of the Amercian Statistical Association 66 846-850.

Ripley, B.D. 1996. Pattern recognition and neural networks, Cambridge University Press, Cambridge, UK.

Rossi, P.E., McCulloch, R.E., Allenby, G.M. 1996. The value of purchase history data in target marketing, Marketing Science 15 321-340.

Russell, G.J., Bell, D., Bodapati, A., Brown, C.L., Chiang, J., Gaeth, G., Gupta, S., Manchanda, P. 1997. Perspectives on Multiple Category Choice, Marketing Letters 8 (3) 297-305.

Russell, G.J., Kamakura, W.A. 1997. Modeling multiple category brand preference with household basket data, Journal of Retailing 73 439-461.

Russell, G.J., Ratneshwar, S., Shocker, A.D., Bell, D., Bodapati, A., Degeratu, A., Hildebrandt, L., Kim, N., Ramaswami, S., Shankar, V.H. 1999. Multiple-Category Decision-Making: Review and Synthesis, Marketing Letters 10 319-332.

Russell, G.J., Petersen, A. 2000. Analysis of Cross Category Dependence in Market Basket Selection, Journal of Retailing 76 (3) 367-392.

Schnedlitz, P., Reutterer, T., Joos, W. 2001. Data-Mining und Sortimentsverbundanalyse im Einzelhandel. In: Hippner, H., Küsters, U., Meyer., M., Wilde, K. (Eds.), Handbuch Data Mining im Marketing, Vieweg, Wiesbaden, pp. 951-970.

Seetharaman, P.B., Ainslie, A., Chintagunta, P.K. 1999. Investigating Household State Dependence Effects Across Categories, Journal of Marketing Research 36 488-500.

Seetharaman, P.B., Chib, S., Ainslie, A., Boatwright, P., Chan, T., Gupta, S., Mehta, N. Rao, V., Strijnev, A. 2004. Models of Multi-Category Choice Behavior, Working Paper, Rice University.

Song, I., Chintagunta, P.K. 2003. Measuring Cross-Category Price Effects with Aggregate Store Data, Working Paper, Hong Kong University of Science and Technology.

Strehl, A., Ghosh, J. 2002. Cluster ensembles — a knowledge reuse framework for combining multiple partitions, Journal on Machine Learning Research 3 583-617.

Thorndike, R.L. 1953. Who belongs in the family? Psychometrika 18 (4) 267-276.

Van den Poel, D., Schamphelaere, J.D., Wets, G. 2004. Direct and indirect effects of retail promotions on sales and profits in the do-it-yourself market, Expert Systems with Applications 27 (1) 53-62.

SFB 649 Discussion Paper Series 2006

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Calibration Risk for Exotic Options" by Kai Detlefsen and Wolfgang K. Härdle, January 2006.
002 "Calibration Design of Implied Volatility Surfaces" by Kai Detlefsen and Wolfgang K. Härdle, January 2006.
003 "On the Appropriateness of Inappropriate VaR Models" by Wolfgang Härdle, Zdeněk Hlávka and Gerhard Stahl, January 2006.
004 "Regional Labor Markets, Network Externalities and Migration: The Case of German Reunification" by Harald Uhlig, January/February 2006.
005 "British Interest Rate Convergence between the US and Europe: A Recursive Cointegration Analysis" by Enzo Weber, January 2006.
006 "A Combined Approach for Segment-Specific Analysis of Market Basket Data" by Yasemin Boztuğ and Thomas Reutterer, January 2006.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".