Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 187 (2008) 294–312
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

A combined approach for segment-specific market basket analysis

Yasemin Boztuğ [a,*], Thomas Reutterer [b]

[a] Humboldt University Berlin, Institute of Marketing, Spandauer Str. 1, D-10178 Berlin, Germany
[b] Vienna University of Economics and Business Administration, Institute of Retailing and Marketing, Augasse 2–6, A-1090 Vienna, Austria

Received 17 January 2006; accepted 2 March 2007
Available online 2 April 2007

Abstract

Market baskets arise from consumers' shopping trips and include items from multiple categories that are frequently chosen interdependently from each other. Explanatory models of multicategory choice behavior explicitly allow for such category purchase dependencies. They typically estimate own and across-category effects of marketing-mix variables on purchase incidences for a predefined set of product categories. Because of analytical restrictions, however, multicategory choice models can only handle a small number of categories. Hence, for large retail assortments, the issue emerges of how to determine the composition of shopping baskets with a meaningful selection of categories. Traditionally, this is resolved by managerial intuition. In this article, we combine multicategory choice models with a data-driven approach for basket selection. The proposed procedure also accounts for customer heterogeneity and thus can serve as a viable tool for designing target marketing programs. A data compression step first derives a set of basket prototypes which are representative for classes of market baskets with internally more distinctive (complementary) cross-category interdependencies and are responsible for the segmentation of households. In a second step, segment-specific cross-category effects are estimated for suitably selected categories using a multivariate logistic modeling framework. In an empirical illustration, significant differences in cross-effects and price elasticities can be shown both across segments and compared to the aggregate model.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Marketing; Choice models; Market basket analysis; Cross-category effects; Segmentation

1. Introduction

A market or shopping basket represents a set of items or product categories included in a retail assortment that a consumer purchases 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; Chen et al., 2005; Song and Chintagunta, 2006). As outlined in more detail in the literature review below, this is the

[*] Corresponding author. Tel.: +49 3020935707.
E-mail addresses: boztug@wiwi.hu-berlin.de (Y. Boztuğ), thomas.reutterer@wu-wien.ac.at (T. Reutterer).

0377-2217/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2007.03.001

domain of various explanatory models of multicategory choice behavior. So far, however, most attempts towards this end introduced in the marketing literature, are restricted to fairly small subsets of product categories. Naturally, today's large retail assortments not only make the consideration of complete category ranges prohibitive, but also managerially unsuitable. Nevertheless, in most empirical applications, both the number and the combination of the selected categories seem to be guided by analytical viability or pure managerial intuition. Hence, the question arises as to which categories should be included in models for predicting cross-category effects that adequately represent consumers' multicategory purchase behavior.

There is another challenge for cross-category effect models emerging from the increasing interest of retailers in targeted direct marketing actions. Numerous retailers equip members of their loyalty programs with bar-coded plastic cards and provide 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 point-of-sale (POS) scanning technologies, those retailers are 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. Consider, for example, a retail marketing manager who wishes to tailor the company's direct marketing efforts and promotional activities to specific customer segments derived from the company's transactions database. We argue that consideration of multicategory purchase behavior patterns for both the segmentation of the customer base and the prediction of marketing-mix effects among carefully selected product categories can assist the retailer in this respect.

As our brief literature review in the next section will show, conventional approaches to market basket analysis exhibit inherent limitations to efficiently accommodate such information. In the remainder of the paper, we present the building blocks of a procedure that combines the estimation of segment-specific marketing-mix and cross-category effects on category choices with a preceding data-driven strategy for adequate (i.e., consumer-centric) category selection and segment generation. The methodology's capability to contribute to the mentioned information needs is illustrated in an empirical application study. Finally, we discuss implications for retail managers and outline some future research agenda.

2. Literature review

There are two main research traditions for analyzing market basket data, 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 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, examine cross-category purchase effects on the aggregate level of demand only. This especially applies to methods aiming at a parsimonious representation of pairwise 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 usually aims for 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. The practical relevance of such attempts obviously suffers from their limitations to a relatively small number of categories with symmetric pairwise relationships. These constraints are successfully resolved by a huge amount of research on association rule discovery stemming from the data mining literature (see, e.g., Agrawal et al., 1995; Anand et al., 1998; Brin et al., 1998; Hahsler et al., 2006), which have seen recent applications in the marketing-related literature (Brijs et al., 2004; Van den Poel et al., 2004; Chen et al., 2005). Following a probabilistic concept, rule-mining techniques derive asymmetric implications (rules) for disjoint subsets of items or categories based on aggregated co-occurrence frequencies (associations). Rule-mining algorithms are capable of dealing with both very large numbers of categories (or even single items) and shopping baskets. However, the issue of an 'average' (or aggregate) market view remains.

The idea of representing cross-category purchase effects at a more disaggregate level is not new to the marketing community but was introduced only recently by Schnedlitz et al. (2001), Decker and Monien (2003) and Decker (2005). The authors utilize neural networks with unsupervised learning rules as a data compression device which results in a mapping of category purchase incidence vectors onto a set of so-called basket prototypes. In empirical applications, they illustrate that each of these prototypes is responsible for a specific class of market baskets with internally more pronounced (complementary) cross-category purchase interrelationships as compared to the aggregate case. More recently, Reutterer et al. (2006) extended this approach towards a customer segmentation tool with campaign design options for target marketing selection and report encouraging findings from a controlled field experiment.

Despite their usefulness for discovering meaningful cross-category interrelationship patterns, the managerial value of all these exploratory approaches to market basket analysis is limited. Since no a priori assumptions are made regarding the distinction between 'response' and 'effect' category (that is, between categories that are affected by purchases of other categories and categories that exert a purchase effect) 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.

By contrast, explanatory (or predictive) types of multicategory choice models mainly focus on estimating the effects of marketing-mix variables on category purchase incidences by explicitly accounting for cross-category dependencies among the retail assortment. Most of these 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. (2005) and Boztuğ and Silberhorn (2006). As an integral part of our proposed procedure, we will highlight the multivariate logit model in detail in the methodology section. 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 are included in the works by Russell and Kamakura (1997), Ainslie and Rossi (1998), Manchanda et al. (1999), Seetharaman et al. (1999), Andrews and Currim (2002) or Chib et al. (2002).

One practical problem with explanatory models is that the set of categories to be incorporated and simultaneously analyzed for cross-category effects on the selected response category is rather limited (typically, up to four of five categories). Indeed, for multivariate logit or probit approaches, significant improvements of powerful Markov chain Monte Carlo simulation methodologies can help to successfully alleviate estimation problems when the number of product categories to be analyzed increases. Nevertheless, real-world retail assortments typically consist of dozens or even hundreds of potentially relevant product categories, which cause severe computational problems unless constraints are placed 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 a priori knowledge within the respective problem context.

To summarize, both exploratory approaches to market basket analysis (lack of implications for managerial decision making) and explanatory multicategory choice models (issue of proper category selection because of computational restrictions) are limited in meeting the initially mentioned information requirements of modern retail marketing. On the other hand, each of these approaches undoubtedly have their specific merits which are combined in the procedure presented in the next section.

3. Methodology

The proposed analytical framework proceeds in a stepwise manner as depicted in Fig. 1. The first exploratory step of the procedure intends a reduction in complexity of the diverse category interdependencies hidden in the numerous shopping baskets collected in a retailer's customer transaction database. Because segment-level results are intended, approaches that avoid early data aggregation are preferred here. Using a similar methodology as employed by Decker and Monien (2003) and Reutterer et al. (2006), the individual shopping baskets are compressed into a set of so-called basket prototypes that constitute a 'generic' (i.e., customer-unspecific) classification of the available set of market baskets. Since each of these prototypes can be characterized by some outstanding or more distinguished complementary cross-category purchase incidences, this information is used to determine the composition of shopping baskets for the second stage of our analysis. In addition, a segmentation of the customer base is derived by assigning each household to the prototype that best represents its overall purchase history. In the second step of our approach, segment-specific adapted multicategory cross-effect models including marketing-mix variables for the previously recommended product categories are estimated based on a multivariate logistic (MVL) model specification similar to Russell and Petersen (2000). In the following two subsections, we provide more details on the technical aspects of analytical steps entailed by the proposed modeling framework.

3.1. Compression of market baskets and segment construction

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

To find a partition of the data into a fixed number of K 'generic' basket classes $C = \{c_1, \ldots, c_K\}$ with outstanding or more distinguished complementary cross-category purchase incidences within the detected classes, resolution of the following objective function is required:

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

where $P = (p_1, \ldots, p_K)$ is a set of prototypes or centroids with $p_k \in \mathbb{R}^J \; \forall k$ and $d(\cdot)$ denoting a distance measure. In the clustering and classification

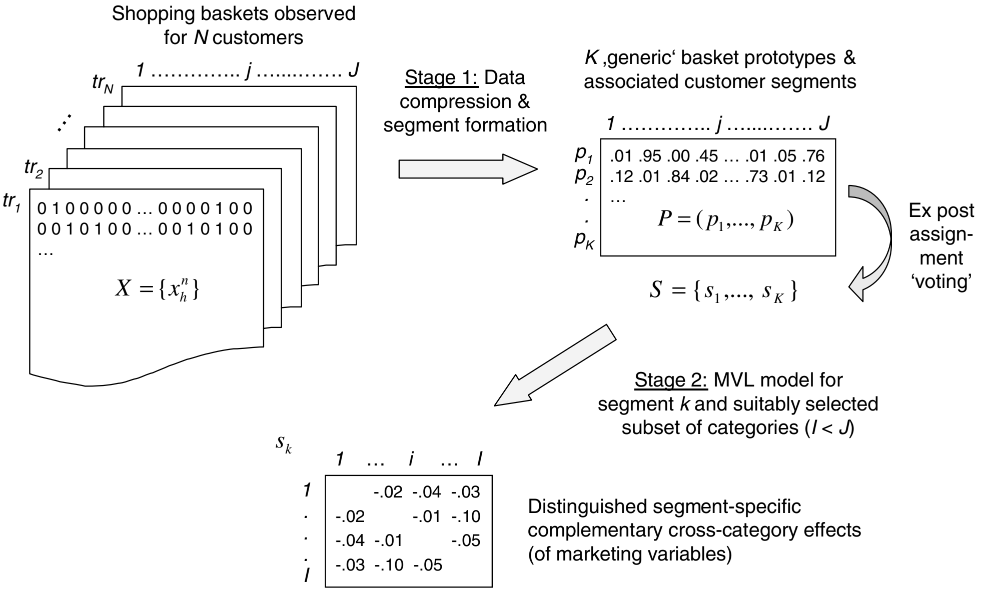
Shopping baskets observed for N customers
1 .... j .... J
tr_N
tr_2
tr_1
0 1 0 0 0 0 0 ... 0 0 0 0 1 0 0
0 0 1 0 1 0 0 ... 0 0 1 0 1 0 0
...
$X = \{x_h^n\}$
Stage 1: Data compression & segment formation
K ‚generic' basket prototypes & associated customer segments
1 .... j .... J
p_1 .01 .95 .00 .45 ... .01 .05 .76
p_2 .12 .01 .84 .02 ... .73 .01 .12
...
p_K
$P = (p_1,\ldots, p_K)$
$S = \{s_1,\ldots, s_K\}$
Ex post assignment 'voting'
Stage 2: MVL model for segment k and suitably selected subset of categories (I < J)
s_k
1 ... i ... I
1 -.02 -.04 -.03
-.02 -.01 -.10
-.04 -.01 -.05
I -.03 -.10 -.05
Distinguished segment-specific complementary cross-category effects (of marketing variables)

Fig. 1. Two-stage approach for analyzing market basket data.

literature, the 'minimum dispersion criterion' in (1) is also known as the principal point or K-centroids problem (Jain and Dubes, 1988; Bock, 1999). One important property of the successful resolution of (1) is that for any optimum configuration (C^*, P^*), the condition $p^*(x_h) = \text{argmin}\{d(x_h, p_k)\ \forall k\}$ holds, which warrants that each basket x_h is mapped onto its minimum distant or closest prototype. In addition, if $d(\cdot)$ is chosen as the Euclidean distance metric, it can be shown that the prototypes p_k^* are equal to their class specific means for the corresponding partition generated by the optimal prototypes under stationarity conditions (Bock, 1999).

Since the purchase incidences are encoded as (usually extremely sparse) binary vectors and we aim at detecting complementary cross-effects, an asymmetric distance measure giving more weight to joint purchases than to common zeros (i.e., non-purchases) is preferred. The well-known Jaccard coefficient has such properties and is used in the present application (cf. Sneath, 1957). An extension of the Jaccard coefficient for measuring the distance between a binary market basket vector and a real-valued prototype is given as follows:

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

with (x_h, p_k) denotes the scalar product of vectors x_h and p_k. Notice that $1 - d(x_h, p_k)$ is often referred to as the Tanimoto similarity coefficient (Anderberg, 1973).

The iterative K-means clustering algorithm is probably the most prominent approach for solving the principal point problem. Starting from any given initial partition, the K-means method recursively minimizes criterion (1) with respect to $C(\tau) \to P(\tau) \to C(\tau+1) \to P(\tau+1) \ldots$ and converges after a finite number of iterations τ to the next local minimum. Although any arbitrary distance measure can be embedded in the algorithm (cf. MacQueen, 1967; Anderberg, 1973), standard implementations use Euclidean distances—hence, the term K-means. One problem with K-means clustering is the 'algorithmic variability' of derived cluster solutions i.e., the quality of a final partition heavily depends on the starting values (Gordon and Vichi, 1998; Hornik, 2005). To cope with this issue, generation of cluster ensembles with different random initializations and selection of the 'best fitting' partition or heuristics for obtaining 'proper starting values' are recommended. Such strategies entail the evaluation of multiple partitions and make K-means type methods computationally expensive and impractical when the number of data points is very large and high-dimensional. The latter is typically the case for shopping basket data from hundreds of thousands of retail transactions and excessively large assortment sizes.

Fortunately, there are numerous 'online' versions of K-means type clustering methods available to solve the principal point problem. In the field of machine learning, they are also 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 achieved 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 virtually 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 distances between a randomly chosen market basket vector x_h and each prototype p_k according to (2).
3. Determine the minimum distant ('winning') prototype $d(x_h, p_k^*) = \min\{d(x_h, p_k)\ \forall k\}$ 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., changes of the prototypes become very small) or the prespecified maximum number of iterations is reached.

Notice that the procedure described above differs in some respects from more conventional VQ versions. Because of data sparsity, we advocate the use of Jaccard distances for identifying the 'winning' prototype p_k^*, but perform a Euclidean-like updating following step 4. We do this for the practical reason that after convergence, the resulting prototypes coincide with the mean values of respective basket classes and therefore can be interpreted as empirical expectations of observing a value of unity (cf. Leisch, 2006). Consequently, each j-element of an optimal prototype vector p_k^* denotes the corresponding product category's purchase incidence probability within the 'generic' shopping basket class c_k^*. Exceptionally (un-)marked combinations of these class-conditional probabilities are indicative of stronger (weaker) cross-category purchase complementarities at the basket class level and will serve as candidates for further investigation. In particular, those categories scoring highest in terms of class-conditional purchase probabilities provide a meaningful basis for basket selection in the second stage of our analysis.

As the term 'generic' suggests, the prototypes generated after convergence still apply to the pooled data set and do not yet recognize the customer identities behind the realized shopping baskets. Even though some households might be characterized by varying degrees of persistence in terms of their multicategory purchase behavior patterns, they are expected to fluctuate across the partition of basket classes from one purchase occasion to another. Typically, they will share some aspects of each of the multiple basket classes they belong to throughout their purchase history. In a mixture modeling terminology, which is well-introduced in the market segmentation literature (cf. Wedel and Kamakura, 2001), one would say that the households are assigned to the components of a mixture distribution according to their independent mixing proportions. Although the underlying statistical properties are different, the set of prototypes can also be regarded as a nonparametric equivalent to the probability density functions of a mixture distribution (in a similar context, see also the comments by Kohonen (1995, p. 78) or Bishop (1995, p. 60)). To determine the associated analogue of the 'mixing proportions', we now utilize our knowledge about the customers' fluctuations across the partition of basket classes. For each customer n, we therefore calculate the following average distance-weighted number of basket class k assignments:

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

Logical expression $1_{\{x_h^n \in c_k\}}$ equals one if shopping basket x_h^n of customer n is assigned to basket class k, otherwise it equals zero. $|t_n|$ is the number of transactions observed for customer n. v_k^n represents the 'degree of belongingness' of customer n to basket class k. Though the sum across all K classes is not necessarily unity for the above distance modified specification of the indicator function (it would be, however, for the raw values), this 'voting' measure of best-fitting class assignments is conceptually very similar to fuzzy class memberships. The latter in turn are very well-known to clustering-based segmentation methods in marketing (Hruschka, 1986; Wedel and Steenkamp, 1991). Notice that defuzzification of membership values allows for nonoverlapping and overlapping segments.[1] For example, by setting all membership values larger than a prespecified threshold value at one and all others at zero, a solution with overlapping segments is obtained (cf., e.g., Hruschka, 1986; Reutterer et al., 2006). On the other hand, setting the largest membership value of a customer across clusters at one and the other values at zero, results in a nonoverlapping segment solution. We opt here for the latter approach. Hence, the final segmentation of customers can be obtained by checking for the respective maximum values:

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

In the present context, segment s_k indicates the disjoint set of all those customers whose past multicategory purchase patterns can be characterized most accurately by prototype k and are therefore assigned to the corresponding segment.

3.2. Segment-specific multivariate logistic model

Utilizing the information now available on the most salient 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 members $n \in s_k$ of segment 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

[1] While one could argue that our approach refers to 'segments' more loosely in a sense of 'customer-types', we prefer to keep the term 'segments' to make it easier for the reader to link the remainder of the exhibited material to the proposed modeling framework. We thank an anonymous reviewer for pointing this out.

(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, forthcoming), the utility function U has the following form:

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

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 to be 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, whereas 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 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 $\mathrm{LOYAL}_{in} = \ln\frac{m(i,n)+0.5}{m(n)+1}$. $m(n)$ 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 $\theta_{ijn} = \psi_{ij} + \eta\mathrm{SIZE}_n$, 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 1 if category i is included in basket b, and 0 otherwise. Here, we inspect only the choice of any item from a specific category and do not model within category choices.

The probability of choosing one specific category, conditional on the choices in the other categories, can be expressed as

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

The market basket of a consumer n at time t is described by a 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, resulting in $2^q - 1$ 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)

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

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 an example

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

Purchase in category 2	Purchase in category 1	
	Yes	No
Yes	$\beta_1 + \mathrm{TIME}_{1nt} + \mathrm{PRICE}_{1nt} + \beta_2 + \mathrm{TIME}_{2nt} + \mathrm{PRICE}_{2nt} + \theta_{12}$	$\beta_2 + \mathrm{TIME}_{2nt} + \mathrm{PRICE}_{2nt}$
No	$\beta_1 + \mathrm{TIME}_{1nt} + \mathrm{PRICE}_{1nt}$	0

of 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 could be present more than once if a basket contains at least three categories.

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 categories 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

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

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 of selecting category i with respect to a change in category j as

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

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

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

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, and 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 panel 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.[2] A total number of 470,825 retail transactions with pick-any choices among an assortment of $J = 65$ categories reported from 4424 households over a 1-year period were first subject to the data compression step and subsequent segment formation. The data contains information about the purchase date and which item 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. With the exemption of fresh products such as meat and fruits, purchase behavior for all typical supermarket categories are recorded in the ZUMA household panel. So it is possible to describe daily shopping trips containing all regular purchased items by a standard household.

The presentation of our empirical findings is organized as follows: First, we examine the derived clustering of shopping baskets and corresponding household segments. Next, we present the parameter estimates for two different segment-specific multivariate logistic model specifications and compare them with those resulting from their aggregate counterpart.

4.1. Basket classes and household segments

In the clustering literature, many authors have 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 may accept this assumption, it is very unlikely that this 'natural' grouping is detectable with an efficiently manageable and managerially acceptable number of classes for the excessively large and high-dimensional data set 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

[2] 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 (2001).

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

To cope with the size of the data set, we randomly split it 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 baskets 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 its '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 Fig. 2 nicely illustrate that the correspondence between partitions (and hence 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

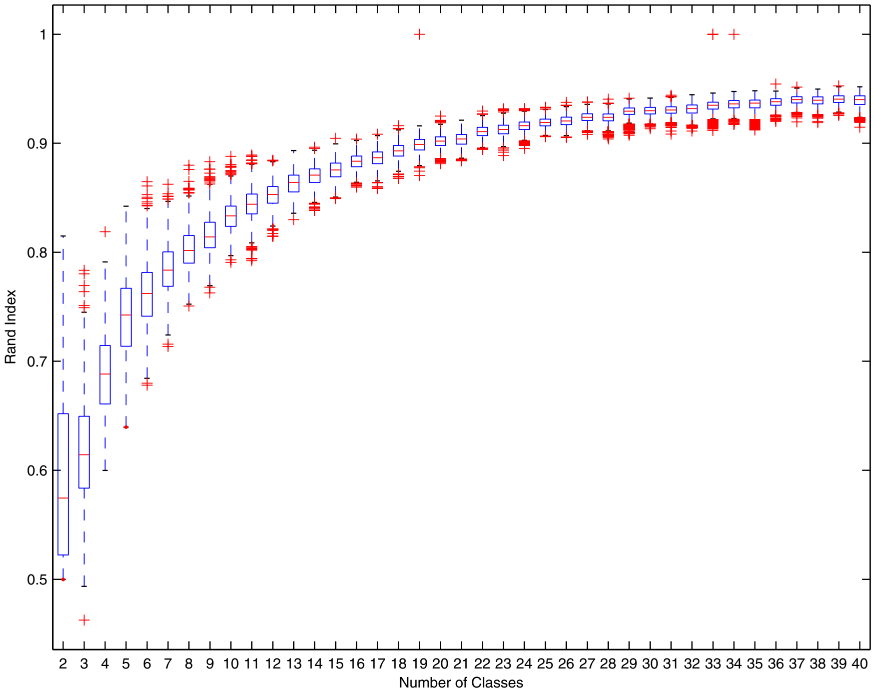

Fig. 2. Distribution of the Rand index for increasing number of classes.

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 derived a collection of class number recommendations based on this 'elbow-heuristic' for the complete set of cluster ensembles. The resulting distribution of recommendation frequencies is shown in Fig. 3. 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 become managerially prohibitive, priority is given to solutions with smaller class numbers but still structurally stable partitioning results. Putting the available 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 procedure.

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 first stage of our procedure, each basket class can now be characterized by its 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 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 represent 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 magnitude 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 at the level of the generic basket classes and the segments.

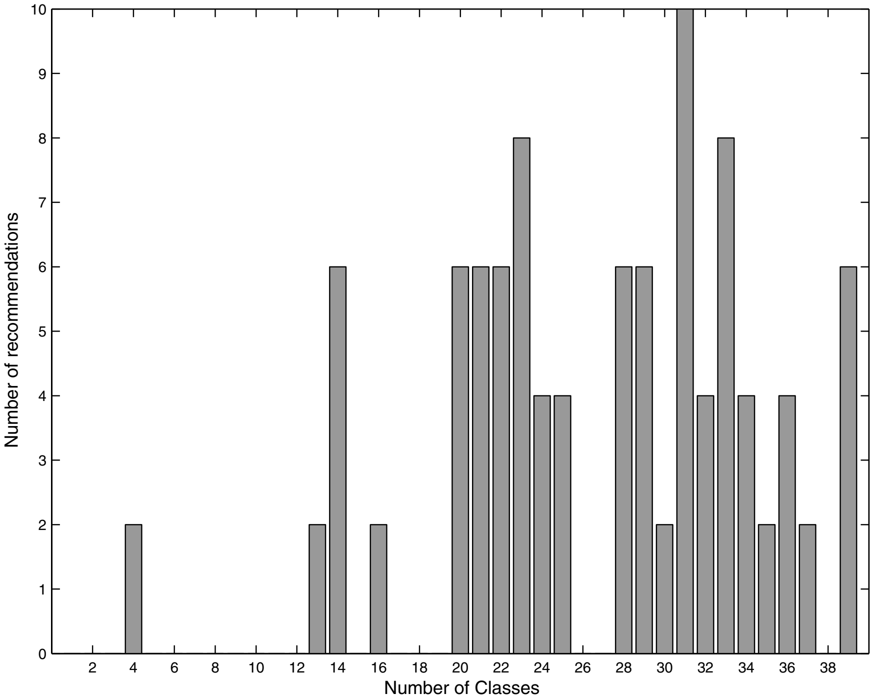

Fig. 3. Number of classes recommendations based on the Davies–Bouldin statistic.

Table 2
Main characteristics of shopping basket classes and household segments

Seg. k	Most distinguished complementary product categories (top five j-elements of prototype p_k)	Relative size (%)	
		Basket	Household
1	**Milk**, soft cheese, curds (e.g., cottage cheese, paneer), coffee, soft drinks	13.2	20.2
2	**Cream**, *milk*, *soft cheese*, *curds*, *yogurt*	13.3	15.2
3	**Yogurt**, *milk*, *curds*, *soft cheese*, soft drinks	11.9	14.4
4	**Hard cheese**, *soft cheese*, *milk*, *yogurt*, *curds*	11.6	11.5
5	*Soft cheese*, toilet paper, wine, cereals, instant coffee	5.0	0.4
6	**Curds**, soft cheese, pudding, cling films, cream	3.5	1.2
7	*Coffee*, *cream*, spirits, filter paper, soft cheese	6.2	5.0
8	**Pet food**, milk, other dairy products (e.g., buttermilk, kefir) coffee, soft cheese	3.7	5.5
9	Toothpaste, detergent, bath additives (e.g., bath salts), soap, dishwashing liquid	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 vegetables*, *ice*, frozen cookies, frozen meals & fish	3.6	0.1
14	*Tea*, *cola drinks*, mayonnaise, lemonade, soft drinks	2.7	0.3

Bold: Class-conditional purch. prob. $p_{jk} > 0.75$; *Italic*: Class-conditional purch. prob. $p_{jk} > 0.25$.

Let us concentrate on two representative segments out of these 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 associated with household segment no. 1 according to the solid line in the left-hand side graph of Fig. 4. Instead, the grey bars represent the unconditional purchase probabilities. From the right-hand 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. Although different with regard to the dominance of only one single category, household segment no. 10 is characterized by high purchase incidences in the equally dominating water category, followed by purchases in the beer, milk, and lemonade categories (see Fig. 5). Notice that milk 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. Segment-specific versus aggregate cross-category effects

As already claimed in the outline of the proposed methodology (see also Fig. 1), one of the primary goals of the data compression step employed prior to the estimation of cross-category purchase effects is the reduction of model complexity, which is

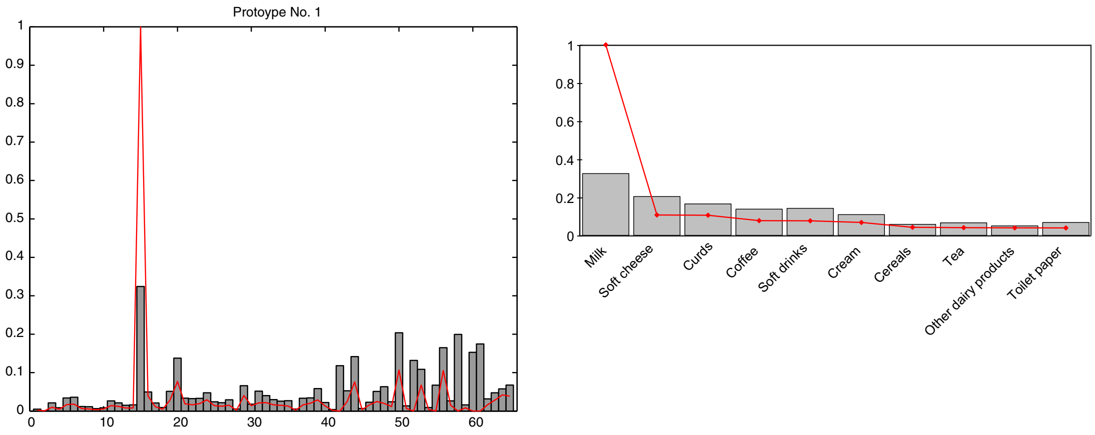

Fig. 4. Category choice probabilities according to prototype no. 1.

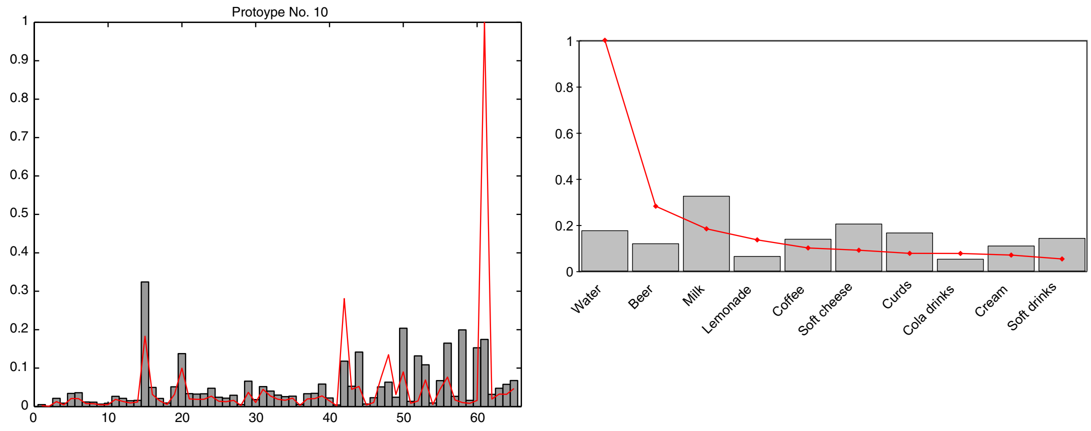

Fig. 5. Category choice probabilities according to prototype no. 10.

achieved by a data-driven strategy for category selection. As an additional condition, the selected categories are required to be meaningful and relevant to a specific household segment. As an obvious consequence of the first stage of our analysis, estimation of segment-specific multivariate logistic models was restricted to the respective most distinguished categories including associated marketing-mix variables. For illustration purposes, we focus on the parameter estimates for the two household segments already highlighted in the previous subsection, i.e., 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 their corresponding prototype values (milk, soft cheese, curds, coffee, and soft drinks), while for segment no. 10 we selected 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 out of the 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.

We first present the parameter estimates for models based on the baskets composed by the selected sets of categories. In doing so, aggregate results of model parameters estimated for the total sample of households are compared to the respective segment-level estimates. Tables 3 and 4 show a selection of parameters for models based on the basket compositions according to prototypes no. 1 and no. 10, respectively. For the sake of clarity, we concentrate our exposition on the price and cross-effect parameters, while the remaining coefficients are provided in Tables 11 and 12 in the Appendix. Overall, most of the statistically significant parameters have the expected sign. Interestingly, most of the segment-level price parameters are larger than their aggregate counterparts in absolute value, which implies a

Table 3
Price and cross-effects parameter estimates for categories selected according to prototype no. 1

	Price parameter estimates (γ_i)				
	Milk	Soft cheese	Curds	Coffee	Soft drinks
All households	−1.90	3.76	−0.91	−0.08	−0.88
	(0.88)	(0.58)	(0.51)	(0.30)	(0.30)
Segment no. 1 members	4.52	4.79	−1.33	0.34	−1.84
	(2.76)	(1.23)	(1.06)	(0.65)	(0.64)
	Cross–effect estimates (ψ_{ij})				
	Milk	Soft cheese	Curds	Coffee	Soft drinks
Milk	–	−0.27	−0.14	−0.46	−0.18
		(0.01)	(0.01)	(0.01)	(0.01)
Soft cheese	−0.35	–	0.16	−0.27	−0.09
	(0.04)		(0.01)	(0.01)	(0.01)
Curds	−0.14	0.31	–	−0.27	−0.13
	(0.04)	(0.03)		(0.01)	(0.01)
Coffee	−0.48	−0.06	−0.08	–	−0.22
	(0.04)	(0.03)	(0.03)		(0.01)
Soft drinks	−0.25	0.15	0.14	0.06	–
	(0.04)	(0.03)	(0.03)	(0.03)	
Basket size (η)	Aggregate: 0.34 (0.01)				
	Segment: 0.32 (0.02)				

Remark: The upper (lower) triangle of the cross-effects matrix shows aggregate (segment-level) estimates; standard errors are given in parentheses.

Table 4
Price and cross-effects parameter estimates for categories selected according to prototype no. 10

	Price parameter estimates (γ_i)			
	Water	Beer	Milk	Lemonade
All households	−3.84	9.05	−2.18	1.38
	(0.76)	(1.46)	(1.00)	(0.50)
Segment no. 10 members	−3.42	12.72	−0.02	2.20
	(2.32)	(3.60)	(2.18)	(1.31)
	Cross-effect estimates (ψ_{ij})			
	Water	Beer	Milk	Lemonade
Water	–	−0.13	−1.25	−0.05
		(0.01)	(0.02)	(0.02)
Beer	−0.36	–	−0.91	0.07
	(0.05)		(0.02)	(0.02)
Milk	−1.37	−0.49	–	−0.73
	(0.04)	(0.04)		(0.02)
Lemonade	−0.45	0.04	−0.43	–
	(0.05)	(0.05)	(0.04)	
Basket size (η)	Aggregate: 0.68 (0.01) Segment: 0.83 (0.03)			

Remark: The upper (lower) triangle of the cross-effects matrix shows aggregate (segment-level) estimates; standard errors are given in parentheses.

generally higher degree of price sensitivity within the two household segments under consideration.

The typically negative cross-effect parameters ψ_{ij} indicate the demand interdependencies among categories. However, they cannot be interpreted directly, because the (category unspecific) basket size loyalty varies across households. Notice that the coefficient η is slightly lower for segment no. 1 households and higher for segment no. 10 members as compared to the aggregate case. Although careful inspection of all these components gives no clear picture, one would expect complementary cross-category effects for the segment-specific models. More precisely, we propose the following statement:

Table 5
Cross-effects for prototype no. 1 categories

	Milk	Soft cheese	Curds	Coffee	Soft drinks
Milk		0.501	0.636	0.313	0.593
Soft cheese	0.404		0.929	0.504	0.681
Curds	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	

Remark: The upper (lower) triangle shows effects based on an average value of $\text{SIZE}_n = 2.24$ (2.36) for all (segment-specific) households.

Table 6
Cross-effects for prototype no. 10 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	

Remark: The upper (lower) triangle shows effects based on an average value of $\text{SIZE}_n = 1.63$ (1.75) for all (segment-specific) households.

Hypothesis 1. Segment-specific cross-category effects (θ_{ij}) are higher than those for all households.

This hypothesis is motivated by the way household segments were previously constructed and the data-driven strategy for basket selection. Since the retail transactions of segment member households were responsible for prototype construction, their characteristic cross-category purchase patterns are expected to be revealed more often than for the average household. The cross-effects depicted in Table 5 and 6 are computed using the formula $\theta_{ijn} = \psi_{ij} + \eta\text{SIZE}_n$ for a typical household by substituting SIZE_n with the respective average basket sizes across all households or segment members. However, it is important to note that the size of these cross-effects certainly differs across households.[3] Because the coefficient η is positive in both segment-specific and aggregate models, cross-effects tend to be higher (and positive) for households that purchase more categories and lower (or even negative) for households that exhibit smaller basket sizes. Since our model did not include empty baskets, SIZE_n is left-censored towards 1, and above-average cross-effects are more likely.

From an average household perspective, positive cross-effects among all categories included in Table 5 can be detected, which implies complementary demand interdependencies. Furthermore, the comparison of aggregate effects with segment-level effects confirms the above stated Hypothesis 1 for all cross-category relationships, except for milk. Most likely, this outlier role of milk is due to the fact that the unconditional choice probability for milk is substantially smaller than the probability

[3] The standard deviations of the basket size variable SIZE_n are 0.678 (0.654) for all (segment-specific) households using the basket composition according to prototype no. 1, and 0.484 (0.506) for all (segment-specific) households using baskets selected according to prototype no. 10.

of observing milk in transactions assigned to basket class no. 1, whereas the opposite is true for the remaining categories included in the segment-specific model. Therefore, the within-segment cross-effects are lower than those of the average consumer, because compared to the other categories, joint purchases with other categories become fewer.

Hypothesis 1 is also inspected for the categories selected according to prototype no. 10. According to the values depicted in Table 6, it can be confirmed for all cross-effects except the relationship between water and lemonade. Interestingly, a negative value of the segment-unspecific cross-effect between milk and water (indicating substitutability) changes to a positive or complementary relationship on the segment level of our analysis. This change of the type of a joint purchase relationship between categories demonstrates that a segment level examination can lead to completely different results compared to an aggregate view. As another important property of the derived cross-effects it should be kept in mind that these results also reflect cohesive consumption complementarity among the involved categories and/or the proximity of their presentations in the retail store's shelf space. In this regard, the beverage categories beer, lemonade, and water (see Table 6), but also the dairy categories soft cheese and curds (see Table 5), exhibit larger cross effects both on the aggregate and in particular on the segment level.

From a managerial standpoint, the cross-price elasticities displayed in Tables 7–10 are of primary interest. These values represent the percentage change in the share of choice of the row category with respect to a one percent price increase in the column category. While the cross-effects were constrained to be symmetric, this presentation implies that cross-price elasticities are asymmetric. Notice that the elasticities account for consumer heterogeneity and can be interpreted as the average elasticities per week. Negative cross-price elasticities indicate complementarity between the inspected categories, which generally would be consistent with the cross-effects reported in Tables 5 and 6.

Of the various price elasticity studies reported in the marketing literature, astonishingly little is known about the elasticity structure at the product category level of demand (Tellis, 1988; Bijmolt et al., 2005). According to previous findings by Russell and Petersen (2000) and Manchanda et al. (1999), cross-category elasticities tend to be small, whereas larger magnitudes could only be observed for obvious consumption complements. In general, however, category-level cross-price elasticities are expected to be negative but weak. Furthermore, the results from a study conducted by Narasimhan et al. (1996) on the brand level indicate that

Table 7
Aggregate cross-price elasticities for prototype no. 1 categories

	Milk	Soft cheese	Curds	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
Curds	−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 8
Segment-level cross-price elasticities for prototype no. 1 categories

	Milk	Soft cheese	Curds	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
Curds	−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 9
Aggregate cross-price elasticities for prototype no. 10 categories

	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 10
Segment-level cross-price elasticities for prototype no. 10 categories

	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

promotional elasticities increase with higher category penetration. The latter means that there is a larger pool of consumers that are interested in a specific category, are therefore more deal-prone and their consumption can be increased by promotion. A direct extension of this category penetration effect to specific combinations of categories at the segment level leads us to propose the following hypothesis:

Hypothesis 2. Segment-specific cross-price elasticities are higher than those for all households.

This proposition is justified by the simple fact that segment members by definition jointly purchase specific category combinations more frequently than the 'average' household. Therefore, they are also more affected by price changes in the respective categories or—in other words—their marginal propensity to increase consumption in response to promotions tends to be higher. For almost all combinations of categories selected according to prototype no. 1, this hypothesis can be supported (see Table 7 for all households and Table 8 for the households within the segment[4]). Consistent with our previous discussion on the cross-effects, milk is again the clear exception. Although the differences are only minor, changes in the purchase probability for the milk category in response to price changes of other categories are higher in the aggregate case. One reason could be that segment-specific households do almost always buy milk. Thus, their choice behavior within the milk category is less affected by price changes in other categories compared to the 'average' household.

Hypothesis 2 is also supported by most of the cross-price elasticities estimated for the combination of categories selected according to prototype no. 10 (see Tables 9 and 10). Only the water and milk category price changes have a weaker impact on choice shares for segment-specific households. This phenomenon can again be explained by dominant purchase frequencies and consistency with our findings for the respective cross-effect patterns.

As a result from careful inspection of Tables 7–10, the following aspects are notable: First, most of the own-category price elasticities depicted as diagonal elements in the tables are smaller than 1 in absolute value which implies an inelastic demand. As expected, this is in sharp contrast to the elastic demand patterns typically observed for brands competing within a single category. Especially the beer category is an exemption from this rule, but this is in conformity with the fact that beer is among the most intensively promoted product categories in the grocery retailing industry. Second, the cross-price elasticities are generally of a relatively small magnitude. This is consistent with the previous findings by Russell and Petersen (2000), who conclude that cross-category spillover effects due to price are only moderate in terms of category choice shares. While this seems to be true for the aggregate market level, the results of our analysis clearly indicate that this conclusion has to be modified if the focus is shifted towards a segment level of demand. In particular, our results demonstrate that higher cross-effects and elasticities can be obtained for suitably derived household segments.

Finally, some remarkable asymmetries can be detected in the cross elasticity matrices discussed above. It is important to note that this especially applies—albeit not exclusively—to the segment level elasticity structures. Consider, for example, the cross elasticities for the soft cheese category in Table 8. Clearly, price changes in this category affect the choice shares in other categories much more than the choice behavior within the soft cheese category is influenced by price changes in the remaining categories. The analogue applies to price promotions in the beer category as depicted in Table 10. These observations provide retail managers with valuable information regarding fine-tuning their promotional activities reflecting cross-category interrelationships. Again, the necessity of a segment-specific treatment becomes obvious.

5. Conclusions and implications

We propose and empirically illustrate a two-stage procedure that combines features from exploratory and model-based approaches of market basket analysis. It has been shown that the employed data compression step is capable of identifying customer

[4] Notice that for curds and coffee nonsignificant parameters were estimated. Nevertheless, we used them to calculate the elasticities. In addition, some positive price parameters result in a wrong sign of the elasticities. We changed them in the tables to the opposite sign as calculated, but we used the correct sign in the calculation of elasticities.

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 could 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. Second, information on segment-specific cross-category dependencies and associated marketing-mix effects become available. Retail marketing managers making use of this information can thus be assisted in designing targeted direct marketing actions within their loyalty programs.

From the above discussion of results obtained in our empirical demonstration study it should be clear that retail managers can enhance their direct marketing initiatives by more effectively targeting specific segments of households. In particular, this could be accomplished by exploiting the asymmetric elasticity structure of cross-category price effects. For example, a target marketing action plan designed for the above described segment no. 1 would be recommended to consider price reductions in the soft cheese category (which potentially could be combined with promotional activities in the soft drinks category). The retailer could select some attractive articles (e.g., based on item profitability or other managerial considerations) within this category and feature them using price promotions in a segment-specific adapted flyer directed to the segment members. Yet another example would be to promote the beer category in a similar way for segment no. 10 households. Because of the asymmetric structure of category level cross-price elasticities, such actions would have the potential to also boost demand in other categories that are not under promotion. According to our empirical results, an undifferentiated realization of the same marketing actions at the market level would have a much weaker impact as opposed to segment-specific targeting.

As a useful side effect, the procedure presented in this paper 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 independant of 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 through the simple introduction of user-defined threshold weights in the voting scheme adopted in the segment formation step (for an example, see Reutterer et al., 2006). In order to expand the empirical performance and to fine tune the proposed procedure to other retail settings, further application studies using different data sets including personalized retail transaction data for a variety of retail industries can be recommended. Finally, a comparison to the impacts of one-to-one targeting strategies and applications to non-retail industries would be helpful.

Acknowledgment

We are very grateful to the Deutsche Forschungsgemeinschaft (German Research Foundation) for generously supporting this research through project #BO1952/1.

Appendix A

In Tables 11 and 12 we present the remaining estimation results for segment no. 1 and no. 10 respectively, which we did not show due to place restrictions in Tables 3 and 4. TIME and LOYAL are household-specific variables, whereas DISPLAY represents one part of the marketing-mix variables.

Table 11
Parameter estimates for categories selected according to prototype no. 1 for all and segment-specific households

		Parameter estimates				
		Milk	Soft cheese	Curds	Coffee	Soft drinks
β_i	All	−3.60	−0.71	−1.08	−0.58	−2.55
		(2.01)	(0.09)	(0.47)	(0.19)	(0.55)
	Segment	12.48	−1.22	−1.61	−1.26	−4.64
		(6.32)	(0.20)	(0.98)	(0.41)	(1.18)
TIME (δ_{1i})	All	−0.79	−0.03	−0.37	−0.04	−0.54
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Segment	−1.02	−0.21	−0.43	0.04	−0.47
		(0.07)	(0.03)	(0.03)	(0.03)	(0.03)
LOYAL (δ_{2i})	All	0.67	0.80	0.66	0.87	0.46
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Segment	0.73	0.81	0.57	0.84	0.43
		(0.04)	(0.02)	(0.02)	(0.02)	(0.01)
DISPLAY (ξ_i)	All	0	0	0	3.88	2.07
		–	–	–	(0.55)	(0.54)
	Segment	0	0	0	4.29	1.46
		–	–	–	(1.21)	(1.16)

Remark: Standard errors are given in parentheses.

Table 12
Parameter estimates for categories selected according to prototype no. 10 for all and segment-specific households

		Parameter estimates			
		Water	Beer	Milk	Lemonade
β_i	All	−10.32	14.04	−3.56	1.62
		(1.90)	(2.41)	(2.29)	(1.12)
	Segment	−7.40	19.99	0.43	3.73
		(5.78)	(5.97)	(4.98)	(2.92)
TIME (δ_{1i})	All	−0.29	−0.36	−0.84	−0.49
		(0.02)	(0.02)	(0.02)	(0.02)
	Segment	−0.30	−0.45	−0.59	−0.79
		(0.07)	(0.05)	(0.04)	(0.06)
LOYAL	All	0.48	0.51	0.70	0.59
(δ_{2i})		(0.01)	(0.01)	(0.01)	(0.01)
	Segment	0.17	0.57	0.51	0.63
		(0.03)	(0.02)	(0.017)	(0.02)
DISPLAY	All	1.07	−0.84	0	1.42
(ξ_i)		(0.77)	(0.53)		(0.54)
	Segment	0.26	−1.86	0	0.65
		(2.37)	(1.31)		(1.43)

Remark: Standard errors are given in parentheses.

References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I., 1995. Fast discovery of association rules. In: Fayyad, G., 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.

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.

Bijmolt, T.H.A., Van Heerde, H.J., Pieters, R.G.M., 2005. New empirical generalizations on the determinants of price elasticity. Journal of Marketing Research 42, 141–156.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

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., forthcoming. Modeling joint purchases with a multivariate MNL approach. Schmalenbach Business Review.

Boztuğ, Y., Silberhorn, N., 2006. Modellierungsansätze in der Warenkorbanalyse im Überblick. Journal für Betriebswirtschaft 56 (2), 105–128.

Brijs, T., Swinnen, G., Vanhoof, K., Wets, G., 2004. Building an association rules framework to improve product assortment decisions. Knowledge Discovery and Data Mining 8, 7–23.

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, vol. 16. Elsevier Science, Amsterdam, pp. 57–92.

Cressie, N.A.C., 1991. Statistics for Spatial Data. Wiley, New York.

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: Spiliopoulou, M., Kruse, R., Borgelt, C., Nürnberger, A., Gaul, W. (Eds.), From Data and Information Analysis to Knowledge Engineering. Springer, Berlin, pp. 598–605.

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., 1986. Market definition and segmentation using fuzzy clustering. International Journal of Research in Marketing 3, 117–134.

Hruschka, H., 1991. Bestimmung der Kaufverbundenheit mit Hilfe eines probabilistischen Memodells. 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.

Kohonen, T., 1995. Self-Organizing Maps. Springer, Berlin.

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 51 (2), 526–544.

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.

Narasimhan, C., Neslin, S.A., Sen, S.K., 1996. Promotional elasticities and category characteristics. Journal of Marketing 60, 17–30.

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.

Reutterer, T., Mild, A., Natter, M., Taudes, A., 2006. A dynamic segmentation approach for targeting and customizing direct marketing campaigns. Journal of Interactive Marketing 20 (3/4), 43–57.

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, J.-B.E.M., Meyer, M., Wilde, K. (Eds.), 1991, 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, S., Gupta, S., Mehta, N., Rao, V., Strijnev, A., 2005. Models of multi-category choice behavior. Marketing Letters 16, 239–254.

Sneath, P.H., 1957. Some thoughts on bacterial classification. Journal of General Microbiology 17, 184–200.

Song, I., Chintagunta, P.K., 2006. Measuring cross-category price effects with aggregate store data. Management Science 52 (10), 1594–1609.

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

Tellis, G.J., 1988. The price elasticity of selective demand: A meta-analysis of econometric models of sales. Journal of Marketing Research 25, 331–341.

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.

Wedel, M., Kamakura, W.A., 2001. Market Segmentation: Conceptual and Methodological Foundations. Kluwer, Boston.

Wedel, M., Steenkamp, J.-B.E.M., 1991. A Clusterwise Regression Method for Simultaneous Fuzzy Market Structuring and Benefit Segmentation. Journal of Marketing Research 28 (4), 385–396.