Journal of Marketing Analytics
https://doi.org/10.1057/s41270-024-00331-0

ORIGINAL ARTICLE

Analyzing market basket data through sparse multivariate logit models

Harald Hruschka[1]

Revised: 19 February 2024 / Accepted: 17 May 2024

Abstract
Using multivariate logit models, we analyze purchases of product categories made by individual households. We introduce a sparse multivariate logit model that considers only a subset of all two-way interactions. A combined forward and backward selection procedure based on a cross-validated performance measure excludes about 74 % of the possible two-way interactions. We also specify random coefficient versions of both the non-sparse and the sparse model. The fact that the random coefficient models lead to better values of the Bayesian information criterion demonstrates the importance of latent heterogeneity. The random coefficients sparse model attains the best statistical performance if we consider model complexity and offers a better interpretability. We investigate the cross-purchase effects of household segments derived from this random coefficient model. As additional interpretation aid we cluster categories and category pairs by integer programming. We demonstrate what the best performing sparse model implies for cross-selling by product recommendations and store layout. The sparse model leads to managerial implications with respect to the effects of advertising in local newspapers and flyers that are as a rule close to those implied by its non-sparse counterpart.

Keywords Retailing · Multicategory choice · Market basket analysis · Multivariate logit model

Introduction

Multicategory choice models like the frequently applied multivariate logit (MVL) model analyze pick-any choices characterized by the fact that households may purchase multiple product categories on the same occasion (Hruschka et al. 1999; Russell and Petersen 2000; Boztuğ and Hildebrandt 2008; Boztuğ and Reutterer 2008; Dippold and Hruschka 2013; Aurier and Mejia 2014; Richards et al. 2018; Solnet et al. 2016; Hruschka 2024). The MVL model allows for two-way interactions between purchases of different product categories. A positive two-way interaction exists if the purchase of category j_1 increases the purchase probability of another category j_2. For example, the purchase of snacks could increase the purchase probability of beverages. In a negative two-way interaction, on the other hand, the purchase of category j_1 decreases the purchase probability of another category j_2 (e.g., the purchase of cold cereal could decrease the purchase probability of beer).

As a rule, the MVL model includes all two-way interactions between categories. For our data set with 31 categories, the number of all two-way interactions amounts to 465. Such a high number makes interpretation difficult. The use of sparse MVL (SVML) models in which purchases depend only on a subset of interactions improves interpretability. Despite this advantage of SMVL models, to our knowledge only few relevant publications exist. Hruschka et al. (1999) start from the MVL model with all two-way interactions applying a greedy stepwise backward elimination that stops if only significant coefficients remain. Dippold and Hruschka (2013) determine significant interactions by Bayesian variable selection techniques. The approach of Boztuğ and Reutterer (2008) consists of two steps. In the first step, these authors cluster market baskets by an online K-means algorithm. In the second step, they estimate one MVL model for the categories assigned to a cluster. This approach only allows two-way interactions between the categories of a cluster and sets interactions with categories of other clusters to zero.

✉ Harald Hruschka
harald.hruschka@ur.de

[1] Faculty of Business, Economics and Management Information Systems, University of Regensburg, Universitätsstrasse 31, 93053 Regensburg, Germany

Published online: 10 June 2024

The multivariate probit (MVP) model represents an alternative multicategory choice model that is quite often applied to market basket data (Chib et al. 2002; Duvvuri et al. 2007; Manchanda et al. 1999; Hruschka 2017; Aurier and Mejia 2014). In the MVL model the purchase probability of a category may be affected by current purchases of other categories. The MVP model does not include such current effects, it reproduces interdependences between categories by correlations of error terms. Error terms are assumed to follow a multivariate normal distribution whose parameters are constant across time. The fact that the MVP model puts more weight on joint non-purchases of category pairs, because they are much more frequent than joint purchases, is a related critical issue (Seetharaman et al. 2005). On the other hand, the MVL model takes only joint purchases into account. Which assumption on pairwise category interdependences leads to a better statistical performance, remains an empirical question.

For the MVL model selection of interactions comes down to straightforward selection of certain coefficients. The selection of pairwise correlations for the MVP model turns out to be more involved. Appropriate methods for the MVP model typically determine a spare inverse correlation matrix (Talhouk et al. 2012) whose elements are harder to interpret than elements of a correlation matrix.

In our MVL models we not only consider two-way interactions, we allow for three-way interactions as well. In a three-way interaction the joint purchase of two categories j_1 and j_2 increases (decreases) the probability of another category j_3. For the MVL and the SMVL models, expressions (4) and (7) show that the latent variable of any category linearly depends on purchases of all the other and of selected other categories or pairs of other categories, respectively.

The MVP model, however, does not include three-way interaction terms. Its latent variables have the structure of a linear seemingly unrelated regression (SUR) model (Zellner 1971). Each expected conditional latent variable of a category in the MVP model linearly depends on the product of the inverse error correlation matrix with the row and column indicating category j_1 eliminated and the vector of errors for the latent variables of the remaining categories (Albert and Chib 1993; Chib and Greenberg 1998).

In the following we discuss several machine learning methods that have recently been applied to market basket data, namely topic models (TMs), the restricted Boltzmann machine (RBM), the deep belief net (DBN), the skip-gram model (SGM) and a deep neural net with bottleneck layers (DNNBL). For these machine learning methods, the number of latent variables is usually lower than the number of categories. Whereas the MNL and MNP provide parameters (i.e., coefficients or correlations) directly measuring interactions, machine learning techniques require additional computations after estimation. Of course, this property of most machine prevents selection of interaction terms as part of the estimation process.

The discrete latent variables of TMs are called topics. TMs comprise two multinomial distributions, topic proportions of categories and topic proportions of baskets (Hruschka 2014b; Jacobs et al. 2016).

The restricted Boltzmann Machine (RBM) includes two-way interactions between latent variables and categories (Hruschka 2014a). Deep belief nets (DBN) stack several RBMs. Each higher level RBM processes the latent variables from the level immediately below (Hruschka 2014a).

The DBN also includes a further layer DBN connecting latent variables of the last layer to observed purchases of each category by a binary logistic function. For the TMs and the RBM two-way interaction measures between categories can be computed after estimation as dot product of two vectors each holding topic proportions and interaction coefficients with respect to latent variables, respectively.

Gabel et al. (2019) adapt the SGM, which was originally developed for natural language processing, to market basket analysis. The other models mentioned so far analyze either the probability of a whole market basket or the purchase probability of each category conditional on the remaining purchased categories. The SGM, on the other hand, considers cross-occurrences of (ordered) category pairs. Therefore, the number of equations to be estimated increases quadratically with the number of purchases categories contained in a basket. The probability of a cross-occurrence depends on a cross-occurrence score, specified as dot product of the estimated latent variables for the two categories. Gabel et al. (2019) demonstrate that cross-occurrence scores are strongly related to the error correlations of a MVP model.

The SGM represents an exploratory approach, which is especially appropriate if a large number of products should be analyzed. Gabel et al. (2019) suggest to use the SGM to decide which categories or products should be considered in a multicategory choice model like the MVL or MVP. Like the other machine learning methods discussed the estimation does not select interactions or cross-occurrences.

Gabel and Timoshenko (2022) develop a DNNBL for market basket analysis. These authors obtain summaries of purchase histories (i.e., market baskets of individual households across several periods) by applying several linear time series filters transformed by a neural activation function. Bottleneck layers capture cross-product relationships by compressing these summaries, average purchase frequencies and current coupons for all products. Outputs of bottleneck layers are projections of the compressed data back to the higher original dimension. Conditional purchase probabilities of each product result from plugging product-specific inputs and outputs of bottleneck layers into a binary logit function.

Our paper contributes over the extant literature as follows. Estimation of the SVML model consists of a forward selection stage and a backward elimination stage. The first stage selects in each step the predictor (e.g., a marketing variable, a two-way interaction, a three-way interaction) with the greatest performance improvement. The steps of the second stage look at reductions of the cross-validated performance measure. Cross-validation makes selection of predictors more robust as opposed to greedy backward elimination. To account for the heterogeneity of households, we extend the MVL and the SVML models to random coefficient models (abbreviated as RC-MVL and RC-SVML). To further improve interpretability, we cluster categories and pairs of categories being part of a three-way interaction based on estimation results in contrast to Boztuğ and Reutterer (2008) who form clusters beforehand.

In "Models" section we specify the MVL and SMVL as well as their random coefficient versions. We also deal with estimation and performance evaluation of these models. In addition we present finite mixture versions of the MVL and SMVL models. In "Cross-purchase effects of household segments" section we explain how we investigate the cross-purchase effects for household segments derived from the random coefficient version. "Clustering categories and category pairs" section introduces clustering of categories and category pairs being part of three-way interactions based on estimation results as interpretation aid for sparse models. In "Data" section we characterize the data set by means of descriptive statistics in "Data" section. "Estimation results" section presents estimation results. "Obtained cross-purchase effects of household segments" and "Obtained clusters of categories and category pairs" sections discuss obtained segment-specific cross-purchase effects and category clusters, respectively. In "Managerial implications" section we demonstrate what sparse models imply for cross-selling by product recommendations and store layout. We also investigate whether implications with respect to category-specific advertising by local newspapers and flyers differ between sparse and non-sparse models. In "Conclusion" section we summarize results and also discuss other applications and extensions of our approach in future research.

Models

J column vector y_{mt} denotes market basket t of household m and consists of binary purchase indicators (J symbolizes the number of product categories). If household m purchases category j on purchase occasion t, the respective element y_{jmt} equals one. Vector x_{mt} consists of regressors relevant for the market basket t of household m. In our study, these regressors consist of category loyalties and the category-specific marketing variable feature, i.e., advertising in local newspapers and flyers. Due to multicollinearity leading to many coefficients with implausible signs we decided not to add category-specific prices as regressors.

We compute the loyalty of household m for category j in market basket t in analogy to exponentially smoothed brand loyalties (Guadagni and Little 1983):

$$loy_{jmt} = \alpha\, y_{jmt-1} + (1-\alpha)\, loy_{jmt-1} \tag{1}$$

$0 \le \alpha \le 1$ denotes the smoothing constant. The binary purchase incidence y_{jmt-1} equals one, if household m purchases category j on the previous purchase occasion $t-1$. The current category loyalty depends on the previous purchase incidence y_{jmt-1} and the previous loyalty loy_{jmt-1}. In a manner similar to the brand loyalty of Guadagni and Little (1983) we set initial values loy_{jm0} equal to the relative purchase frequency of the respective category j across all households and shopping visits ($t = 1$ denotes the first shopping visit). The lower the smoothing constant α is, the less the loyalty variable reflects fluctuating purchases.

We use the Bayesian information (BIC) to evaluate models (Cameron and Trivedi 2007):

$$BIC = -2\, LPL + np\ \ln(N) \tag{2}$$

LPL denotes the total log pseudo-likelihood, np the number of parameters and N the number of observations. We explain the computation of the LPL for each model in "Multivariate logit model"–"Random coefficient and finite mixture models" sections. The BIC considers model complexity, i.e., it penalizes models with respect to the number of parameters. Models attaining low BIC values are to be preferred.

Multivariate logit model

Extending the expression for the MVL model without regressors (also known as auto-logistic model) given in Besag (1972) we define the probability of market basket y_{mt} conditional on regressors x_{mt} as follows:

$$\begin{aligned} &\exp(y'_{mt}a + x'_{mt}B\, y_{mt} + 1/2\, y'_{tm}Dy_{mt})/C \\ &\text{with} \quad C = \sum_{\psi \in \{0,1\}^J} \exp(\psi' a + x'_{mt}B\,\psi + 1/2\,\psi' D\psi) \end{aligned} \tag{3}$$

Expression (3) shows that computation of this probability requires division by the normalization constant C that is obtained by summing over all possible market baskets represented by different binary vectors ψ. Coefficients contained in the (J, J) matrix D measure two-way interactions between categories. As a two-way interaction of a category with itself does not make sense, all diagonal elements of D are zero. Off-diagonal elements are symmetric, i.e., $d_{j1,j2} = d_{j2,j1}$. Column vector a consists of J category constants. The (K, J) matrix B holds the effect of K regressors on purchase probabilities. The MVL model has been applied to market basket data by Russell and Petersen (2000) building upon earlier publications in statistics (Cox 1972; Besag 1974).

We can write the purchase probability of category j in market basket t of household m conditional on purchases of the other categories collected in vector y_{-jtm}, the category-specific loyalty loy_{jmt} and the category-specific marketing variable $mvar_{jt}$ as:

$$P(y_{jmt} = 1|y_{-jmt}, x_{mt}) = \varphi(Z_{jmt}) \\ Z_{jmt} = a_j + b_j loy_{jmt} + c_j mvar_{jt} + \sum_{l \neq j} d_{j,l}\, y_{lmt} \tag{4}$$

$\varphi(Z)$ denotes the binomial logistic function $1/(1 + exp(-Z))$. Z_{jmt} can be interpreted as latent variable referring to category j and market basket t of household m.

Maximum likelihood estimation of the MVL model requires in each iteration the computation of the normalization constant (see expression (3)). For the 31 categories in our study, we would have to sum over more than 2.14×10^9 possible market baskets. Maximum pseudo-likelihood (MPL) estimation (Bel et al. 2018) offers a viable alternative maximizing the log pseudo-likelihood LPL across households, market baskets and categories:

$$LPL = \sum_{m=1}^{M} \sum_{t=1}^{T_m} \sum_{j=1}^{J} \log(\tilde{P}_{jmt}) \tag{5}$$

T_m symbolizes the number of market baskets of household m, $\tilde{P}_{jmt}$ the pseudo-probability of a (non) purchase of category j in market basket t of household m. Summing logarithmic pseudo-probabilities across J product categories makes MPL estimation feasible as it replaces the summation across all possible baskets, which would be necessary in maximum likelihood estimation. The pseudo-probability $\tilde{P}_{jmt}$ can be written as:

$$\tilde{P}_{jmt} = P(y_{jmt} = 1|y_{-jmt}, x_{mt})^{y_{jmt}} \\ (1 - P(y_{jmt} = 1|y_{-jmt}, x_{mt}))^{1-y_{jmt}} \tag{6}$$

Expression (4) shows how to compute the conditional probability $P(y_{jmt} = 1|y_{-jmt}, x_{mt})$ for the MVL model. y_{jmt} denotes the binary purchase indicator, which is set to one if basket t of household m contains category j. One can see from Eq. (6) that its first part is relevant if category j is purchased and its second part if category j is not purchased. Briefly, LPL estimation looks at J different binomial logit models representing conditional probabilities.

Sparse multivariate logit model

Like the MVL we estimate the SMVL by maximizing the LPL using J binomial logit models. The SMVL differs from the MVL by the specification of the conditional probabilities that:

- may exclude the marketing variable or the category loyalty of the respective category.
- as a rule includes a subset of the purchases of other categories only.
- may include joint purchases of pairs of other categories.

Consequently, we write the conditional purchase probability of category j in market basket t of household m for the SMVL model as:

$$P(y_{jmt} = 1|y_{-jmt}, x_{mt}) = \varphi(Z_{jmt}) \tag{7}$$

$$Z_{jmt} = a_j + u_j^1 b_j loy_{jmt} + u_j^2 c_j mvar_{jt} \\ + \sum_{k \in \mathbb{J}_j^1} d_{j,k}\, y_{lmt} + \sum_{(l1,l2) \in \mathbb{J}_j^2} e_{j,l1,l2}\, y_{l1mt}\, y_{l2mt}$$

loy_{jmt} ($mvar_{jt}$) is exluded if the binary variable $u_j^1 j$ (u_j^2) equals 0. Two-way interactions of category j with a category k are included if k belongs to set $\mathbb{J}_j^1$ that must not contain category j. Sparsity mainly results from excluding purchases of many other categories. Three-way interactions of category j with categories $l1$ and $l2$ are included if the pair $(l1, l2)$ belongs to set $\mathbb{J}_j^2$ that must not contain pairs having category j as one of their two elements.

For each of J categories we determine optimal values of two hyperparameters based on a five-fold cross-validation. The first hyperparameter has 19 values, i.e., the integer number of coefficients $(2, 3, \dots, 20)$. The second hyperparameter has two values that indicate whether three-way interaction may be included or not. We perform a grid search to select the hyperparameter mix from $38 = 19 \times 2$ combinations. This grid search provides the two binary variables u_j^1, u_j^1 and the two sets $\mathbb{J}_j^1, \mathbb{J}_j^2$ of expression (7). Please note that each of these two sets may be empty meaning that two-way interactions and three-way interactions are excluded, respectively.

The Multivariate Adaptive Regression Splines (MARS) method (Friedman 1991; Kuhn and Johnson 2013) serves to search specifications with log loss ll_j of category j as performance measure:

$$ll_j = \sum_{m=1}^{M} \sum_{t=1}^{T_m} -\log(\tilde{P}_{jmt})/N \\ \tilde{P}_{jmt} = P(y_{jmt} = 1|y_{-jmt}, x_{mt})^{y_{jmt}} \\ (1 - P(y_{jmt} = 1|y_{-jmt}, x_{mt}))^{1-y_{jmt}} \tag{8} \\ N = \sum_{m=1}^{M} T_m$$

Log losses are related to the log pseudo-likelihood of expression (5) as follows:

$$LPL = -N \sum_{j=1}^{J} ll_j \tag{9}$$

MARS builds additive models based on the set of predictor variables (here: category loyalty, marketing variable, two-way and three-way interactions). In the first stage the algorithm operates recursively, incorporating at each step the predictor variables leading to the greatest performance improvement. The second stage consists of a backwards elimination routine that looks at reductions of the cross-validated performance measure. Optimal values of hyper-parameters are taken from the minimum mean out-of-fold performance measure. MARS then rebuilds the corresponding model using all the data.

Please be aware that we do not use splines, as all predictors are linearly related to the latent variable Z_{jmt} in expression (7). Most of the considered predictors are binary. MARS is known to perform well for binary predictors for which it was not originally developed (Ruczinski et al. 2003).

Random coefficient and finite mixture models

To take latent heterogeneity of households into account we extend the MVL and SMVL models to versions with random coefficients. Each random coefficient model consists of J different random coefficient binomial logit models, i.e., one for each product category. To simplify matters, we speak of a random coefficient multivariate logit model in the following to denote such a set of models.

We estimate random coefficient models by maximum simulated pseudo-likelihood using Halton draws and normal mixing distributions for category constants and random coefficients of predictors (Train 2003). Additional suffixes r symbolize the r-th draw of a category constants or a coefficient from the mixing distribution. For the random coefficient multivariate logit (RC-MVL) model we specify the r-th draw of latent variable Z_{jmtr} of category j in market basket t of household m as:

$$Z_{jmtr} = a_{jr} + b_{jr} loy_{jmt} + c_{jr} mvar_{jt} + \sum_{l \neq j} d_{j,l,r}\, y_{lmt} \tag{10}$$

The random coefficient sparse multivariate (RC-SMVL) model encompasses the predictors selected by the SVML model; it differs from the former by considering random draws of parameters. The r-th draw of latent variable Z_{jmtr} of category j in market basket t of household m is:

$$\begin{aligned} Z_{jmtr} =& a_{jr} + u_j^1 b_{jr} loy_{jmt} + u_j^2 c_{jr} mvar_{jt} \\ &+ \sum_{k \in \mathbb{J}_j^1} d_{j,k,r}\, y_{lmt} + \sum_{(l1,l2) \in \mathbb{J}_j^2} e_{j,l1,l2,r}\, y_{l1mt}\, y_{l2mt} \end{aligned} \tag{11}$$

For each random coefficient model the conditional purchase probability of category j in market basket t of household m based on the r-th draw is:

$$P_r(y_{jmt} = 1 | y_{-jmt}, x_{mt}) = \varphi(Z_{jmtr}) \tag{12}$$

The corresponding pseudo-probability $\tilde{P}_{jmtr}$ can be written as:

$$\begin{aligned} \tilde{P}_{jmtr} =& P_r(y_{jmt} = 1 | y_{-jmt}, x_{mt})^{y_{jmt}} \\ & (1 - P_r(y_{jmt} = 1 | y_{-jmt}, x_{mt}))^{1-y_{jmt}} \end{aligned} \tag{13}$$

We consider that as a rule we observe several market baskets for a household. Both for the RC-MVL and the RC-SMVL model we therefore compute the pseudo-likelihood PL_{jm} for category j and household m over T_m baskets as average across R samples:

$$PL_{jm} = (1/R) \sum_{r=1}^{R} \prod_{t=1}^{T_m} \tilde{P}_{jmtr} \tag{14}$$

Summing across households and categories we obtain the total log pseudo-likelihood of a random coefficient model:

$$LPL = \sum_{m=1}^{M} \sum_{j=1}^{J} \log PL_{jm} \tag{15}$$

We now present finite mixture extensions of these models for the readers' benefit though we do not investigate these models in more detail. Coefficients differ between S household segments with $s = 1, \ldots, S$. We specify segment-specific latent variables for the MVL and the SMVL models as:

$$Z_{jmts} = a_{js} + b_{js} loy_{jmt} + c_{js} mvar_{jt} + \sum_{l \neq j} d_{j,l,s}\, y_{lmt} \tag{16}$$

$$\begin{aligned} Z_{jmts} =& a_{js} + u_j^1 b_{js} loy_{jmt} + u_j^2 c_{js} mvar_{jt} \\ &+ \sum_{k \in \mathbb{J}_j^1} d_{j,k,s}\, y_{lmt} + \sum_{(l1,l2) \in \mathbb{J}_j^2} e_{j,l1,l2,s}\, y_{l1mt}\, y_{l2mt} \end{aligned} \tag{17}$$

The total log pseudo-likelihood for the finite mixture extensions results from expressions (6) and (5) with the following conditional purchase probability of category j in market basket t of household m:

$$P(y_{jmt} = 1 | y_{-jmt}, x_{mt}) = \sum_{s=1}^{S} f_{sm}\, \varphi(Z_{jmts}) \tag{18}$$

Fig. 1 Cluster-Specific Directed Graphs. Explanation: Directed edges indicate that a category or category pair affects another category. Categories and category pairs are identified by the numbers given in Table 15. Examples for cluster 4: frozen pizza (11) affects frozen dinners (10), cold cereal (17) affects milk (6) and vice versa, the category pair cold cereal (6) and milk (17) affects yoghurt (31)

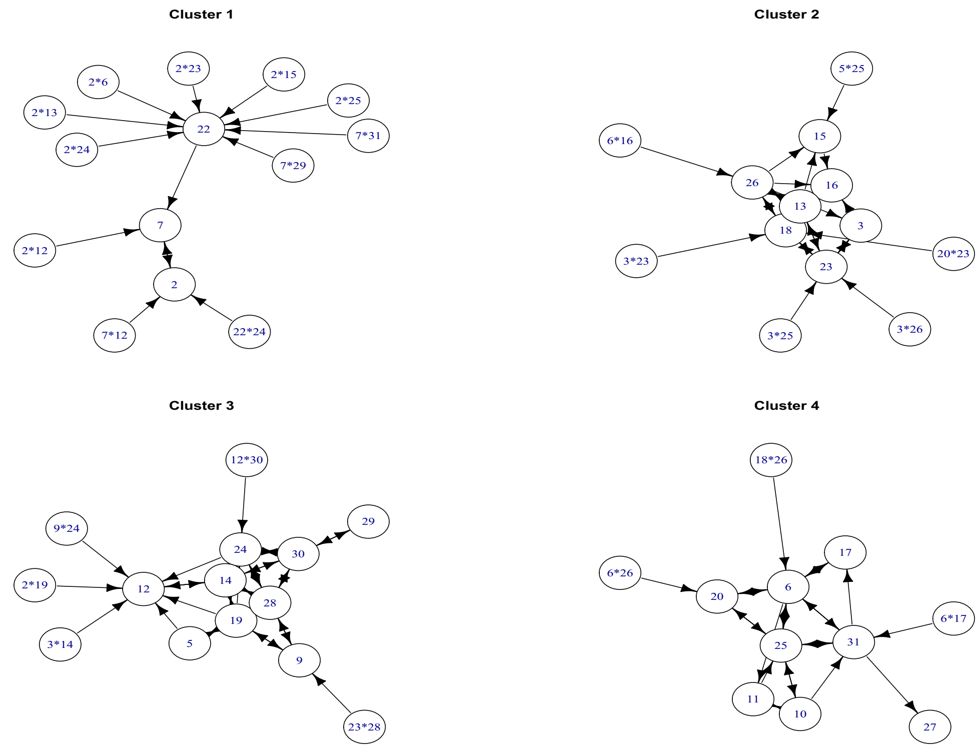

Table 1 Product categories and abbreviations

Beer & ale	beer	Blades	blades
Carbonated beverages	carbbev	Cigarettes	cigets
Coffee	coffee	Cold Cereal	coldcer
Deodorant	deod	Diapers	diapers
Facial tissue	factiss	Frozen dinners	fzdin
Frozen pizza	fzpizza	Household cleaners	hhclean
Frankfurters & hotdog	hotdog	Laundry detergent	laundet
Margarine & butter	margbutr	Mayonnaise	mayo
Milk	milk	Mustard & ketchup	mustketc
Paper towels	paptowl	Peanut butter	peanbutr
Photographic supplies	photo	Razors	razors
Salty snacks	saltsnck	Shampoo	shamp
Soup	soup	Spaghetti sauce	spagsauc
Sugar substitutes	sugarsub	Toilet tissue	toitisu
Tooth brush	toothbr	Toothpaste	toothpa
Yogurt	yogurt		

f_{sm} is the probability that household m belongs to segment s with $\sum_{s=1}^{S} f_{sm} = 1.0$.

Cross-purchase effects of household segments

We want to investigate how households differ in terms of their cross-purchase behavior implied by a random coefficient model. To this end we determine household-specific coefficients in the first step (Train 2003; Greene 2003).

For each random coefficient we draw R random samples from the normal distribution using means and standard deviations that were estimated before. From expression (13) we obtain the log pseudo-likelihood PL_{jmr} of draw r for category j and household m over T_m baskets as:

Table 2 Relative marginal frequencies

milk	0.476	carbbev	0.400	saltsnck	0.351	coldcer	0.280	yogurt	0.202
soup	0.197	spagsauc	0.184	toitisu	0.171	margbutr	0.158	paptowl	0.140
coffee	0.136	laundet	0.118	fzpizza	0.110	mayo	0.109	hotdog	0.103
mustketc	0.102	fzdin	0.090	factiss	0.084	peanbutr	0.080	beer	0.076
toothpa	0.059	shamp	0.053	deod	0.039	cigets	0.032	hhclean	0.030
diapers	0.020	blades	0.019	toothbr	0.014	sugarsub	0.011	photo	0.007
razors	0.002								

Table 3 Relative pairwise frequencies

carbbev	milk	0.199	carbbev	saltsnck	0.189	milk	saltsnck	0.176
coldcer	milk	0.154	coldcer	saltsnck	0.128	carbbev	coldcer	0.127
milk	yogurt	0.115	milk	soup	0.107	milk	spagsauc	0.094
carbbev	soup	0.092	milk	toitisu	0.089	carbbev	yogurt	0.089
carbbev	spagsauc	0.088	saltsnck	yogurt	0.088	saltsnck	soup	0.087
coldcer	yogurt	0.087	margbutr	milk	0.086	saltsnck	spagsauc	0.085
carbbev	toitisu	0.084	saltsnck	toitisu	0.080			

Shows the 20 highest relative pairwise frequencies

Table 4 Average category loyalties

milk	0.359	carbbev	0.307	saltsnck	0.274	coldcer	0.218
yogurt	0.161	soup	0.149	spagsauc	0.142	toitisu	0.133
margbutr	0.119	paptowl	0.109	coffee	0.103	laundet	0.092
fzpizza	0.084	mayo	0.084	hotdog	0.081	mustketc	0.081
fzdin	0.070	factiss	0.065	peanbutr	0.062	beer	0.058
toothpa	0.046	shamp	0.041	deod	0.032	cigets	0.026
hhclean	0.023	diapers	0.015	blades	0.014	toothbr	0.010
sugarsub	0.009	photo	0.004	razors	0.001		

Table 5 Average features

fzdin	0.187	yogurt	0.179	carbbev	0.175	fzpizza	0.174
diapers	0.171	spagsauc	0.169	saltsnck	0.154	coldcer	0.151
peanbutr	0.133	margbutr	0.130	milk	0.129	coffee	0.124
factiss	0.119	soup	0.112	laundet	0.106	mayo	0.100
toitisu	0.095	hotdog	0.094	shamp	0.094	razors	0.093
toothpa	0.089	deod	0.083	paptowl	0.067	beer	0.061
hhclean	0.041	mustketc	0.041	blades	0.040	photo	0.039
toothbr	0.017	sugarsub	0.008	cigets	0.000		

Table 6 Model evaluation results

Model	Pseudo log-likelihood	Number of parameters	BIC
Without loyalties and features			
MVL	– 224,868	496	454,740
SMVL	– 226,047	267	454,788
With loyalties and features			
MVL	– 181,863	558	369,356
SMVL	– 182,877	234	368,115
RC-MVL	– 177,151	589	360,244
RC-SMVL	– 178,644	265	359,962

24,074 observations; values rounded to the nearest integer

$$\log PL_{jmr} = \sum_{t=1}^{T_m} \log \tilde{P}_{jmtr} \tag{19}$$

Each household-specific coefficient corresponds to a weighted average of its draws. These weights $w_{mr} = \exp(\log PL_{jmr}) / \sum_{r=1}^{R} \exp(\log PL_{jmr})$ indicate the relative importance of random draw r for household m.

In the second step we apply K-means to the household-specific coefficients to obtain household segments. Each segment is characterized by its centroid, i.e., the coefficients averaged across all households allocated to this segment.

Now we are able to compute segment-specific cross-purchase effects for a segment s using its averaged coefficients. We measure the cross-purchase effects of category j on category j' of category by the difference of purchase probabilities of category j' conditional on a high and a low purchase probability of category j. Please note that cross-purchase effects are as a rule not symmetric, i.e., the effect of category j on category j' differs from the effect of category j' on category j.

As expressions (4) and (7) give conditional probabilities only we have to determine unconditional purchase

Table 7 RC-SMVL model: equations of latent variables Z_{jmt} (1)

beer:	− 3.8654 + 1.9205 * loyalty + 4.3206 * feature
blades:	− 5.0623 + 5.6272 * feature + 1.4164 * deod + 4.6913 * razors * shamp
carbbev:	− 1.6558 + 0.9112 * loyalty + 2.6125 * feature + 0.7931 * saltsnck + 0.2868 * paptowl + 0.2169 * spagsauc
cigets:	− 6.7845 + 1.1945 * loyalty
coffee:	− 3.8217 + 1.6336 * loyalty + 6.0958 * feature + 0.5102 * toitisu + 0.3660 * coldcer + 0.4069 * spagsauc
coldcer:	− 2.6006 + 0.5177 * loyalty + 3.9523 * feature + 0.5978 * yogurt + 0.4415 * spagsauc + 0.4450 * saltsnck + 0.4823 * peanbutr + 0.2989 * milk + 0.3100 * soup + 0.4752 * shamp + 0.3521 * margbutr + 0.6267 * mustketc * spagsauc
deod:	− 4.3890 + 1.6861 * loyalty + 4.0004 * feature + 1.3591 * blades + 0.9314 * toothpa + 0.9754 * shamp + 0.5181 * toitisu + 1.5004 * razors
diapers:	− 7.6713 + 3.0023 * feature
factiss:	− 4.1727 + 1.2551 * loyalty + 4.8714 * feature + 0.7260 * toitisu + 0.7676 * paptowl + 0.3586 * soup + 0.3712 * saltsnck * toitisu
fzdin:	− 3.9223 + 2.0720 * loyalty + 2.3516 * feature + 0.8686 * fzpizza + 0.4725 * soup
fzpizza:	− 4.4201 + 3.2291 * loyalty + 5.1048 * feature + 0.9589 * fzdin + 0.4258 * coldcer + 0.4627 * spagsauc + 0.4336 * saltsnck
hhclean:	− 4.6512 + 2.9284 * loyalty + 3.7339 * feature + 0.5583 * laundet + 0.5566 * paptowl + 0.4528 * coffee + 0.3913 * soup
hotdog:	− 3.8334 + 1.1528 * loyalty + 5.0071 * feature + 0.5428 * mustketc + 0.4939 * mayo + 0.3910 * coldcer + 0.4138 * spagsauc + 0.4359 * saltsnck
laundet:	− 3.7636 + 1.3764 * loyalty + 5.6510 * feature + 0.6265 * toitisu + 0.9698 * hhclean + 0.5804 * paptowl + 0.7011 * toothpa + 0.3917 * coldcer

Equations include fixed coefficients and mean category constants with a significance level less equal 0.001

Table 8 RC-SMVL model: equations of latent variables Z_{jmt} (2)

margbutr:	− 3.1128 + 3.5345 * feature + 0.3393 * soup + 0.4202 * coldcer + 0.3746 * spagsauc + 0.3902 * hotdog + 0.3661 * toitisu + 0.5042 * coffee * soup
mayo:	− 3.3732 + 4.2946 * feature + 1.0759 * mustketc + 0.3532 * spagsauc + 0.4624 * hotdog + 0.3359 * coldcer + 0.2799 * carbbev + 0.2814 * margbutr
milk:	− 1.0456 + 1.0823 * loyalty + 2.7768 * feature + 0.3527 * yogurt + 0.3188 * coldcer
mustketc:	− 3.2664 + 5.1842 * feature + 1.0724 * mayo + 0.5911 * hotdog + 0.4202 * spagsauc + 0.4015 * soup + 0.4672 * peanbutr * saltsnck + 0.3598 * carbbev * saltsnck
paptowl:	− 3.5425 + 1.6826 * loyalty + 4.4417 * feature + 1.2115 * toitisu + 0.6385 * factiss + 0.4990 * laundet + 0.3997 * coffee + 0.3346 * coldcer + 0.3484 * saltsnck
peanbutr:	− 4.0386 + 2.0645 * loyalty + 4.4338 * feature + 0.4771 * coldcer + 0.5018 * mustketc + 0.4079 * soup + 0.5088 * coldcer * spagsauc
photo:	− 5.9746 + 6.2445 * loyalty + 6.5125 * feature
razors:	− 6.8507 + 2.5543 * blades * shamp - 15.1689 * blades * soup + 3.2032 * blades * saltsnck + 2.4407 * deod * yogurt
saltsnck:	− 2.1784 + 0.9093 * loyalty + 3.1382 * feature + 0.6316 * carbbev + 0.4334 * coldcer + 0.3777 * mustketc + 0.3113 * toitisu + 0.3243 * fzpizza + 0.4111 * hotdog + 0.3592 * carbbev * spagsauc + 0.3191 * carbbev * soup
shamp:	− 4.5674 + 4.9523 * feature + 0.7080 * toothpa + 0.9777 * deod + 0.5191 * toitisu + 0.5049 * coldcer + 0.4715 * laundet + 0.4199 * spagsauc + 0.4055 * paptowl
soup:	− 3.0429 + 1.3980 * loyalty + 4.2701 * feature + 0.5044 * spagsauc + 0.3423 * coldcer + 0.3890 * margbutr + 0.3525 * yogurt + 0.3818 * fzdin + 0.3074 * paptowl + 0.3429 * peanbutr + 0.4546 * toothpa + 0.3707 * fzpizza + 0.3104 * mustketc

Equations include fixed coefficients and mean category constants with a significance level less equal 0.001

probabilities. We simulate purchases by iterated Gibbs-sampling from the appropriate conditional distributions (Besag 2004) using segment-specific coefficients. We estimate the unconditional purchase probability of a category by its marginal relative frequency across the simulated purchases.

Clustering categories and category pairs

As interpretation aid for of the SVML or RC-SVML models we determine clusters of categories and category pairs being part of three-way interactions. This clustering works on a graph whose nodes represent categories and category pairs (see expressions (7) and (11)). Nodes of the graph link each

Table 9 RC-SMVL model: equations of latent variables Z_{jmt} (3)

spagsauc:	− 3.2804 + 0.9140 * loyalty + 4.2215 * feature + 0.4383 * coldcer + 0.5388 * soup + 0.3846 * mustketc + 0.4171 * hotdog + 0.3817 * toitisu + 0.3895 * fzpizza + 0.3422 * coffee + 0.5004 * coldcer * mayo
sugarsub:	− 6.7917 + 0.7169 * yogurt
toitisu:	− 3.2136 + 0.6440 * loyalty + 4.2182 * feature + 1.1742 * paptowl + 0.8051 * factiss + 0.4987 * laundet + 0.3815 * spagsauc + 0.4904 * toothpa + 0.3724 * margbutr + 0.4174 * shamp + 0.2915 * saltsnck
toothbr:	− 5.1773 + 2.0573 * toothpa
toothpa:	− 4.1265 + 4.2189 * feature + 1.9137 * toothbr + 0.8941 * shamp + 0.9404 * deod + 0.6514 * laundet + 0.4581 * soup + 0.4918 * toitisu + 0.3569 * yogurt
yogurt:	− 3.2018 + 1.5958 * loyalty + 2.9473 * feature + 0.3524 * coldcer + 0.4222 * soup + 0.3189 * spagsauc + 0.4970 * fzdin + 0.3910 * coldcer * milk

Equations include fixed coefficients and mean category constants with a significance level less equal 0.001

Table 10 RC-SMVL model: standard deviations of category constants

beer	1.355	blades	1.075	carbbev	1.067	cigets	2.995
coffee	1.001	coldcer	0.882	deod	0.700	diapers	2.727
factiss	0.913	fzdin	1.211	fzpizza	0.678	hhclean	0.739
hotdog	0.875	laundet	0.743	margbutr	1.152	mayo	0.510
milk	0.925	mustketc	0.587	paptowl	0.843	peanbutr	0.710
photo	0.800	saltsnck	0.960	shamp	0.767	soup	0.791
spagsauc	0.824	sugarsub	2.300	toitisu	0.929	toothbr	1.053
toothpa	0.693	yogurt	1.258				

Shows standard deviations with a significance level less equal 0.001

category j to each of the other categories contained in $\mathbb{J}_j^1$ and to each three-way interaction contained in $\mathbb{J}_j^2$. The graph has no node for a category j if both $\mathbb{J}_j^1$ and $\mathbb{J}_j^2$ are empty and category j is contained neither in $\mathbb{J}_k^1$ nor in $\mathbb{J}_k^2$ for each of the other categories $k \neq j$. We say that such a category j is isolated.

We measure the quality of a clustering $\mathbb{C}$ with C clusters by its modularity (Brandes et al. 2008):

$$Q(\mathbb{C}) = \sum_{C \in \mathbb{C}} \left[\frac{m_c}{m} - \left(\frac{\sum_{v \in C} n(v)}{2m} \right)^2 \right] \tag{20}$$

m_C symbolizes the number of links that are assigned to cluster C, m the total number of links, $n(v)$ the number of links with node v.

Modularity reflects two conflicting objectives (Brandes et al. 2008). Its first term becomes higher for a low number of clusters, each with many links contained. On the other hand, its second term gets lower for a high number of clusters each with a small number of links. Modularity measures the difference between the total fraction of links that fall within clusters and the expected fraction if links were placed at random considering the number of links to nodes (Porter et al. 2009). To obtain clusters we maximize modularity by integer programming (Brandes et al. 2008). In contrast to heuristic search algorithms, integer programming is guaranteed to determine the global optimum and provides both the optimal number of clusters and the optimal assignment of links to these clusters.

Empirical study

Data

Our data refer to 24,047 shopping visits made by a random sample of 1500 households to one specific grocery store over a one-year period. For each shopping visit, we compose a market basket from the IRI data set Bronnenberg et al. (2008). We represent a market basket by a binary vector whose elements indicate whether a household purchases each of 31 product categories (see Table 1). The average number of shopping visits per household amounts to 16.031, its standard deviation to 13.464. The average basket size (i.e., number of purchased categories) is 3.852, its standard deviation 2.654.

Table 2 shows relative marginal purchase frequencies for the 31 categories. Milk (razors) is the category most (least) frequently purchased. Table 3 gives the highest 20 pairwise relative frequencies. Carbonated beverages and milk are the two categories most frequently purchased together, followed by carbonated beverages and salty snacks.

Table 4 contains the average loyalty across all market baskets for each category with a smoothing constant $\alpha = 0.1$, which puts most weight on the loyalty of the previous week. This value of the smoothing constant leads to the best performing MVL model with category loyalty according to a grid search over $[0.1, 0.2, 0.3, \dots, 0.9]$. Given such a value,

Table 11 RC-SVML model: number of interactions

Category	Interactions with other categories and category pairs	Interactions affecting other categories		
		Two-way	Three-way	Total
beer	0	0	0	0
blades	2	1	3	4
carbbev	3	2	3	5
cigets	0	0	0	0
coffee	3	3	1	4
coldcer	9	14	3	17
deod	5	3	1	4
diapers	0	0	0	0
factiss	4	2	0	2
fzdin	2	3	0	3
fzpizza	4	4	0	4
hhclean	4	1	0	1
hotdog	5	5	0	5
laundet	5	5	0	5
margbutr	6	4	0	4
mayo	6	2	1	3
milk	2	1	1	2
mustketc	6	6	1	7
paptowl	6	7	0	7
peanbutr	4	2	1	3
photo	0	0	0	0
razors	4	1	1	2
saltsnck	8	6	4	10
shamp	7	4	2	6
soup	10	10	3	13
spagsauc	8	12	3	15
sugarsub	1	0	0	0
toitisu	8	10	1	11
toothbr	1	1	0	1
toothpa	7	6	0	6
yogurt	5	5	1	6

Table 12 Segment-specific cross-purchase effects (1)

Segment 1					
saltsnck	carbbev	0.0128	carbbev	saltsnck	0.0108
paptowl	toitisu	0.0044	coldcer	saltsnck	0.0040
spagsauc	coldcer	0.0036	coldcer	milk	0.0034
toitisu	paptowl	0.0032			
Segment 2					
saltsnck	carbbev	0.0128	carbbev	saltsnck	0.0090
paptowl	toitisu	0.0064	coldcer	saltsnck	0.0058
toitisu	paptowl	0.0046	coldcer	milk	0.0042
coldcer	margbutr	0.0040	spagsauc	coldcer	0.0036
yogurt	coldcer	0.0036	coldcer	spagsauc	0.0030
Segment 3					
saltsnck	carbbev	0.0126	carbbev	saltsnck	0.0120
yogurt	coldcer	0.0104	yogurt	milk	0.0094
coldcer	milk	0.0056	coldcer	yogurt	0.0056
coldcer	saltsnck	0.0054	paptowl	toitisu	0.0048
spagsauc	coldcer	0.0042	toitisu	paptowl	0.0038
coldcer	soup	0.0038	soup	yogurt	0.0036
coldcer	margbutr	0.0034	mustketc	mayo	0.0030
Segment 4					
saltsnck	carbbev	0.0122	carbbev	saltsnck	0.0118
paptowl	toitisu	0.0050	spagsauc	coldcer	0.0048
coldcer	milk	0.0042	yogurt	coldcer	0.0040
coldcer	saltsnck	0.0040	toitisu	paptowl	0.0036
coldcer	spagsauc	0.0034	coldcer	margbutr	0.0032
mustketc	mayo	0.0032	spagsauc	carbbev	0.0030
yogurt	milk	0.0030			

Shows cross-purchase effects of the left category on the right category greater equal 0.0030

previous purchases are strongly smoothed. Milk attains the highest category loyalty, followed by carbonated beverages and salty snacks. Average category loyalties are remarkably similar to relative marginal purchase frequencies.

We measure features as weekly market share-weighted averages of UPC level variables in the respective category. Consequently, features take values between zero and one. Table 5 shows average values of features for each category. We obtain the highest (lowest) average feature value for frozen dinners (cigarettes).

Estimation results

Table 6 contains BIC values of the six investigated models. Adding loyalty and features as predictors improves BIC values both for the MVL and the SMVL model.

For both the RC-MVL and the RC-SMVL models standard deviations of predictors' random coefficients are as a rule not significantly different from zero. Therefore, we use fixed coefficients for loyalty, feature, 2-way and 3-way interactions, only category constants are random. These random coefficient models lead to clearly lower BIC values though they only have 31 parameters more. Overall, the RC-SMVL with predictors outperforms the other investigated models. From now on we only discuss the RC-MVL and RC- SMVL models as their performance is better than that of related models without random coefficients.

Tables 7, 8 and 9 contain the equations of the 31 category-specific latent variables Z_{jmt} for the RC-SMVL model. These equations only include mean category constants and fixed coefficients that are significant (to simplify terminology, we only write that a mean category constant or a fixed

Table 13 Segment-specific cross-purchase effects (2)

Segment 5					
saltsnck	carbbev	0.0134	carbbev	saltsnck	0.0084
paptowl	toitisu	0.0064	coldcer	milk	0.0050
toitisu	paptowl	0.0048	coldcer	saltsnck	0.0048
spagsauc	coldcer	0.0040	coldcer	margbutr	0.0040
coldcer	yogurt	0.0040	spagsauc	carbbev	0.0036
coldcer	spagsauc	0.0034			
Segment 6					
saltsnck	carbbev	0.0170	carbbev	saltsnck	0.0122
yogurt	milk	0.0052	coldcer	saltsnck	0.0052
paptowl	toitisu	0.0052	spagsauc	coldcer	0.0048
yogurt	coldcer	0.0046	toitisu	paptowl	0.0042
coldcer	milk	0.0040	coldcer	yogurt	0.0038
coldcer	spagsauc	0.0034	coldcer	margbutr	0.0039
coldcer	soup	0.0030			
Segment 7					
saltsnck	carbbev	0.0144	carbbev	saltsnck	0.0100
paptowl	toitisu	0.0052	coldcer	saltsnck	0.0050
saltsnck	carbbev	0.0144	carbbev	saltsnck	0.0100
paptowl	toitisu	0.0052	coldcer	saltsnck	0.0050
yogurt	coldcer	0.0044	spagsauc	coldcer	0.0042
coldcer	margbutr	0.0042	yogurt	milk	0.0042
coldcer	milk	0.0040	toitisu	paptowl	0.0040
coldcer	yogurt	0.0040	soup	spagsauc	0.0036
spagsauc	soup	0.0034	spagsauc	carbbev	0.0032
coldcer	soup	0.0032	coldcer	spagsauc	0.0030

Shows cross-purchase effects of the left category on the right category greater equal 0.0030

Table 14 Additional segment-specific cross-purchase effects

Segment 2 vs. Segment 1					
toitisu	paptowl	0.0046	coldcer	margbutr	0.0040
yogurt	coldcer	0.0036	coldcer	spagsauc	0.0030
Segment 3 vs. Segment 2					
yogurt	milk	0.0094	coldcer	yogurt	0.0056
coldcer	soup	0.0038	soup	yogurt	0.0036
mustketc	mayo	0.0030			
Segment 4 vs. Segment 2					
mustketc	mayo	0.0032	spagsauc	carbbev	0.0030
yogurt	milk	0.0030			
Segment 5 vs. Segment 2					
spagsauc	carbbev	0.0036			
Segment 6 sv. Segment 3					
coldcer	spagsauc	0.0034			
Segment 7 vs. Segment 6					
soup	spagsauc	0.0036	spagsauc	soup	0.0034
spagsauc	carbbev	0.0032			
coldcer	soup	0.0032			

Shows cross-purchase effects of the first segment additional to those of the second segment greater equal 0.0030

coefficients is significant if significant at a confidence level of 0.001).

The majority of equations include both loyalty and feature. Their coefficients are positive, i.e., higher values of these predictors increase the conditional purchase probability. Exceptions are the equations for blades, diapers, margarine & butter, mayonnaise, mustard & ketchup, shampoo and tooth paste, which include only feature and the equation for cigarettes which include only loyalty. Equations for razors, sugar substitutes and tooth brush exclude both loyalty and feature.

According to Table 10 standard deviations of constants are significant for all categories except blades. Both the size of standard deviations of category constants and the superior statistical performance of the RC-SMVL model (see Table 6) underline that taking the latent heterogeneity of households into account is important.

The total number of significant two-way interactions amounts to 120. These two-way interactions are all positive, indicating that the categories involved are purchase complements (i.e., purchase of category j_1 increases the conditional purchase probability of category j_2). Many food/drink categories are often affected only by other food/drink categories (frozen dinners, frozen pizza, frankfurters & hotdog, mayonnaise, mustard & ketchup, milk, peanut butter, salty snacks, sugar substitutes, yogurt). In an analogous way, several non-food categories are affected only by other non-food categories (blades, deodorant, tooth brush).

Three-way interactions are as a rule positive, i.e., a joint purchase of categories j_1 and j_2 (e.g., razors and shampoo) increases the conditional probability of the affected category j_3 (e.g., blades). We obtain only one negative three-way interaction, namely for razors with the two categories blades and soup. A joint purchase of blades and soup lowers the conditional purchase probability of razors.

Table 11 gives the number of interactions for each category with other categories and category pairs. This table also contains the number of categories affected by two-way and three-way interactions. Four categories are affected neither by purchases of other categories nor by purchases of category pairs (beer & ale, cigarettes, diapers, photographic supplies). For the remaining 27 categories, the number of such interactions varies between one and ten (e.g., 10 for soup, 9 for cold cereal, 8 for salty snacks, spaghetti sauce and toilet tissue).

Obtained cross-purchase effects of household segments

To determine segment-specific cross-purchase effects for the RC-SVML model we apply the procedure explained in "Cross-purchase effects of household segments" section. The elbow criterion suggests a solution with eight segments

Table 15 Clusters of categories and category pairs

Cluster 1
7 deod, 7 * 12 deod * hhclean, 22 * 24 razors * shamp, 2 blades, 22 razors, 2 * 12 blades * hhclean, 2 * 24 blades * shamp, 2 * 15 blades * margbutr, 2 * 13 blades * hotdog, 2 * 25 blades * soup, 2 * 23 blades * saltsnck, 2 * 6 blades * coldcer, 7 * 29 deod * toothbr, 7 * 31 deod * yogurt
Cluster 2
23 saltsnck, 26 spagsauc, 15 margbutr, 18 mustketc, 16 mayo, 13 hotdog, 5 * 25 coffee * soup, 3 carbbev, 20 * 23 peanbutr * saltsnck, 3 * 23 carbbev * saltsnck, 3 * 26 carbbev * spagsauc, 3 * 25 carbbev * soup, 6 *16 coldcer * mayo
Cluster 3
19 paptowl, 28 toitisu, 24 shamp, 30 toothpa, 23 * 28 saltsnck * toitisu, 14 laundet, 3 * 14 carbbev * laundet, 5 * coffee, 2 * 19 blades * paptowl, 9 * 24 factiss * shamp, 12 hhclean, 9 factiss, 12 * 30 hhclean * toothpa, 29 toothbr
Cluster 4
6 coldcer, 31 yogurt, 20 peanbutr, 17 milk, 25 soup, 18 * 26 mustketc * spagsauc, 11 fzpizza, 10 fzdin, 6 * 26 coldcer * spagsauc, 6 * 17 coldcer * milk, 27 sugarsub

*Indicates a category pair which is part of a three-way interaction

whose shares amount to 0.30, 0.19, 0.131, 0.105, 0.101, 0.071, 0.071, and 0.017, respectively. We ignore segment eight because of its low share in the following.

We obtain positive cross-purchase effects only. Therefore, we classify the involved categories as purchase complements in accordance with Betancourt and Gautschi (1990), who consider two categories as purchase complements if they are purchased jointly more frequently than expected under stochastic independence.

Cross-purchase effects greater equal 0.003 are listed in Tables 12 and 13. The number of these cross-purchase effects differs between segments (e.g., seven for segment 1 versus sixteen for segment 7).

The seven cross-purchase effects of segment 1 arise in the other segments as well. Nonetheless, categories involved in cross-purchase effects are heterogenous across segments. For selected segment pairs Table 14 shows cross-purchase effects turning out for a segment in addition to those relevant for the other segment. Interestingly, these additional cross-purchase effects are restricted to food categories and do not involve non-food categories. Moreover, several values of cross-purchase effects for the same two categories are clearly different in two segments (e.g., yogurt on milk: 0.0094 in segment 1 and 0.0030 in segment 4; carbonated beveragse on salty snacks: 0.0084 in segment 5 and 0.0122 in segment 6; toilet tissue on paper towels: 0.0032 in segment 1 and 0.0048 in segment 5).

Obtained clusters of categories and category pairs

As interpretation aid for the RC-SVML model we cluster an undirected graph whose nodes represent categories and category pairs in accordance with "Clustering categories and category pairs" section. Maximizing modularity yields four clusters, which are described in Table 15. Four categories (beer & ale, cigarettes, diapers, photographic supplies) are isolated and consequently do not belong to any cluster.

Cluster 1 contains non-food categories (deodorant, blades, razors). It also contains seven pairs and three pairs in which the categories blades and deodorants participate, respectively. Only food categories (salty snacks, spaghetti sauce, margarine & butter, mustard & ketchup, mayonnaise, frankfurters & hotdog, carbonated beverages) are assigned to cluster 2. Its category pairs involve food categories only. Cluster 3 consists of non-food categories (paper towels, toilet tissue, shampoo, tooth paste, laundry detergent, household cleaners, facial tissue, tooth brush) except for coffee. Cluster 4 contains food categories only (cold cereal, yoghurt, peanut butter, milk, soup, frozen pizza, frozen dinners, sugar substitutes). In a comparable manner its category pairs involve food categories. Overall, these descriptions show clear differences between clusters.

Categories with a high number of links to other categories and category pairs can be rated as central for a cluster. This centrality measure is known as degree of a node and frequently used to characterize graphs (Diestel 2005). Cluster-specific central categories are:

Table 16 Differences of feature effects between the RC-SMVL and RC-MVL models

Featured category	Affected category	RC-SMVL	RC-MVL	Difference
Negative differences ≤ -0.0010				
coldcer	coldcer	0.0158	0.0198	−0.0040
spagsauc	spagsauc	0.0138	0.0172	−0.0034
beer	beer	0.0014	0.0038	−0.0024
saltsnck	saltsnck	0.0168	0.0188	−0.0020
soup	soup	0.0100	0.0116	−0.0016
fzdin	coldcer	0.0000	0.0014	−0.0014
soup	coldcer	0.0000	0.0014	−0.0014
paptowl	paptowl	0.0032	0.0046	−0.0014
peanbutr	peanbutr	0.0038	0.0052	−0.0014
coffee	coldcer	0.0000	0.0012	−0.0012
margbutr	coldcer	0.0000	0.0012	−0.0012
factiss	factiss	0.0040	0.0052	−0.0012
carbbev	saltsnck	0.0020	0.0032	−0.0012
fzpizza	saltsnck	0.0000	0.0012	−0.0012
milk	coldcer	0.0000	0.0010	−0.0010
fzpizza	fzpizza	0.0098	0.0108	−0.0010
spagsauc	margbutr	0.0000	0.0010	−0.0010
laundet	paptowl	0.0000	0.0010	−0.0010
coffee	spagsauc	0.0000	0.0001	−0.0010
coldcer	spagsauc	0.0006	0.0016	−0.0010
spagsauc	yogurt	0.0000	0.0010	−0.0010
Positive differences ≥ 0.0010				
toitisu	toitisu	0.0084	0.0056	0.0028
diapers	diapers	0.0020	0.0006	0.0014
margbutr	margbutr	0.0096	0.0082	0.0014
mayo	mayo	0.0060	0.0046	0.0014
milk	milk	0.0170	0.0156	0.0014
fzpizza	laundet	0.0000	−0.0010	0.0010
toitisu	laundet	0.0010	0.0000	0.0010
yogurt	yogurt	0.0112	0.0102	0.0010

Only contains feature effects with differences at least 0.0010 in absolute size

- razors (22) for cluster 1.
- mustard & ketchup (18) and hotdog & frankfurters (13) for cluster 2.
- toilet tissue (28), paper towels (19) and laundry detergent (14) for cluster 3.
- cold cereal (6) and soup (25) for cluster 4.

Fig. 1 contains directed graphs for each of these four clusters. Directed links indicate that a category or category pair affects another category. One can also see from these graphs which categories are central for a cluster.

Managerial implications

We indicate what the RC-SMVL model implies with respect to cross selling. Cross-selling can be be supported by recommending categories not purchased, e.g. by printing at checkout or as part of a mobile phone message. Such recommendations can be based on high positive interactions (see Tables 7, 8 and 9). For example, the category shampoo (frozen dinner) can be recommended if the basket of a customer contains tooth paste and deodorants (frozen pizza and soup).

Segment-specific recommendations require that a household be allocated to the segment with the highest log pseudo-likelihood across observed purchase categories. Recommendations can rely on higher positive cross-purchase effects (see Tables 12, 13 and 14). Category yoghurt (soup) can be recommended to a household of segment 3 (7) purchasing cold cereal and soup (cold cereal and spaghetti sauce). Across all segments the category salty snacks can be recommended to a households purchasing carbonated beverages and cold cereal.

Management may enhance dross-selling also by appropriate positioning of categories in aisles and shelves of a store. Categories exert more effects on other categories belonging to the same cluster (see "Obtained clusters of categories and category pairs" section). Categories belonging to the same cluster are placed near to each other (e.g., carbonated beverages and salty snacks near to each other and far away from categories such as cold cereal and milk). Central categories can be positioned in the middle of the other categories belonging to the same cluster (e.g., mustard & ketchup and hotdog & frankfurters for cluster 2, cold cereal and soup for cluster 4). Isolated categories on the other hand at g (e.g., beer & ale, photographic supplies) can be positioned at greater distances to products belonging to a cluster.

Next, we assess whether managerial implications with respect to category-specific features vary between the RC-SMVL and RC-MVL models. To this end we compute differences of their respective feature effects. Feature effects are measured by the difference of purchase probabilities of category j' between high and low values of features for category j. For $j' = j$ we obtain an own effect, otherwise a cross effect. High (low) values of features for category j result from multiplying their average value by a multiplicative factor greater (less) than zero. We set this factor to 1.1 (0.9) and keep values of loyalties and features of other categories $j' \neq j$ at their average values.

Please note that the estimated coefficients presented in expressions (4), (7) and (12) refer to conditional probabilities and do not directly reflect effects on unconditional purchase probabilities. To determine unconditional purchase probabilities, we generate simulated purchases by iterated Gibbs-sampling from the appropriate conditional distribution (Besag 2004). In case of the RC-SMVL and RC-MVL

Table 17 Feature effects of the RC-SMVL model

Featured category	Affected category		Featured category	Affected category	
carbbev	carbbev	0.0208	milk	milk	0.0170
saltsnck	saltsnck	0.0168	coldcer	coldcer	0.0158
spagsauc	spagsauc	0.0138	yogurt	yogurt	0.0112
coffee	coffee	0.0100	soup	soup	0.0100
fzpizza	fzpizza	0.0098	margbutr	margbutr	0.0096
toitisu	toitisu	0.0084	laundet	laundet	0.0066
7 mayo	mayo	0.0060	hotdog	hotdog	0.0054
fzdin	fzdin	0.0046	factiss	factiss	0.0040
peanbutr	peanbutr	0.0038	saltsnck	carbbev	0.0034
toothpa	toothpa	0.0034	paptowl	paptowl	0.0032
shamp	shamp	0.0022	diapers	diapers	0.0020
carbbev	saltsnck	0.0020			

Only contains feature effects at least 0.0020 in absolute size

models, we generate simulated purchases for each of 500 sampled category constants and coefficients. We estimate the unconditional purchase probability of a category by its marginal relative frequency across the simulated purchases.

Table 16 shows that only five of the total 930 (= 31×30) feature effects differ between the RC-SMVL and RC-MVL models by at least 0.0020 in absolute size. All these five are own effects (cold cereal, spaghetti sauce, beer & ale, salty snacks, toilet tissue). Especially for zero feature effects that the RC-SVML model implies in contrast to the RC-MVL model, differences are not greater than 0.0010 in absolute size, many of these are much smaller. We therefore conclude that managerial implications with respect to feature effects are as a rule very similar for the two models.

Feature effects implied by the RC-SMVL of at least 0.0020 in absolute size are given in Table 17. These effects are all positive, i.e., higher features increase the unconditional purchase probability of the affected category. Table 17 contains own effects with the exception of one cross effect (features for salty snacks affect purchases of carbonated beverages).

Conclusion

We introduce a sparse multivariate logit (SVML) model to analyze market basket data. In contrast to the conventional multivariate logit (MVL) model the SVML model that does not include all two-way interactions between product categories. A combined forward and backward selection procedure based on a cross-validated performance measure excludes most two-way interactions. Because of its lower complexity the SVML clearly outperforms the MVL model in terms of the Bayesian information criterion (BIC). Random coefficient versions of these models (RC-MVL and RC-SVML) further improve performance demonstrating that latent heterogeneity of households is important. Once again, the sparse variant RC-SVML beats the non-sparse variant RC-MVL in terms of BIC.

For our data set only about 26 % of all possible two-way interactions remain in the RC-SVML. This result facilitates interpretation of the RC-SVML in comparison to the RC-MVL model. As further interpretation aid we determine four clusters of categories and category pairs by maximizing modularity using the interaction coefficients detected by the RC-SVML model. As a rule, these four clusters contain either food or non-food categories. Four categories are isolated, i.e., they are not part of an interaction effect.

We show what the estimated RC-SVML model implies for cross-selling based on product recommendations and store layout. Moreover, we measure effects of increasing features (advertising in local newspapers and flyers) on the purchase probability of categories. Most feature effects turn out to be equal for the two models. Only 0.5 % of all possible effects differ by at least 0.002 in absolute size between the SVML and the MVL models. To sum up, the sparse model attains a better statistical performance if model complexity is taking into account and offer better interpretability. In addition to these advantages, the sparse model leads to managerial implications with respect to feature effects that are as a rule close to those implied by its non-sparse counterpart.

The data that we use in this paper originate from a food retailing context. The presented approach could be readily applied to purchases of non-food product categories or browsing behavior across pages of a website or across multiple websites. Moreover, future research efforts might extend our approach in several ways. One possibility consists in analyzing purchases at a more detailed level (e.g., the brand level). Moreover, binary purchase incidences could be replaced or supplemented by response variables like purchase amount and purchase quantity.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The dataset used in the current study is not publicly available as it contains proprietary information that the authors acquired through a license. Information on how to obtain it and reproduce the analysis is available from the corresponding author on request.

Declarations

Conflict of interest The author states that there is no conflict of interest.

References

Albert, J., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of American Statistical Association* 88: 669–679.

Aurier, P., and V. Mejia. 2014. Multivariate logit and probit models for simultaneous purchases: Presentation, uses, appeal and limitations. *Recherche et Applications en Marketing* 29: 79–98.

Bel, K., D. Fok, and R. Paap. 2018. Parameter estimation in multivariate logit models with many binary choices. *Economic Review* 37: 534–550.

Besag, J. 1972. Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society: Series B* 34: 75–83.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B* 35: 192–236.

Besag, J. 2004. An introduction to Markov chain Monte Carlo methods. In *Mathematical foundations of speech and language processing*, ed. M.E. Johnson, S.P. Khudanpur, M. Ostendorf, et al., 247–270. New York: Springer.

Betancourt, R., and D. Gautschi. 1990. Demand complementarities, household production, and retail assortments. *Marketing Science* 9: 146–161.

Boztuğ, Y., and L. Hildebrandt. 2008. Modeling joint purchases with a multivariate MNL approach. *Schmalenbach Business Review* 60: 400–422.

Boztuğ, Y., and T. Reutterer. 2008. A combined approach for segment-specific market basket analysis. *European Journal of Operational Research* 187: 294–312.

Brandes, U., D. Delling, M. Gaertler, et al. 2008. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* 20: 172–188.

Bronnenberg, B.J., M.W. Kruger, and C.F. Mela. 2008. Database paper: The IRI marketing data set. *Marketing Science* 27: 745–748.

Cameron, A.C., and P.K. Trivedi. 2007. *Microeconometrics*. Cambridge: Cambridge University Press.

Chib, S., and E. Greenberg. 1998. Bayesian analysis of multivariate probit models. *Biometrika* 85: 347–361.

Chib, S., P.B. Seetharaman, and A. Strijnev. 2002. Analysis of multi-category purchase incidence decisions using IRI market basket data. In *Econometric Models in Marketing*, ed. P.H. Franses and A.L. Montgomery, 57–92. Amsterdam: JAI.

Cox, D.R. 1972. The analysis of multivariate binary data. *Journal of the Royal Statistical Society C* 21: 113–120.

Diestel, R. 2005. *Graph theory*, 3rd ed. Berlin: Springer.

Dippold, K., and H. Hruschka. 2013. Variable selection for market basket analysis. *Computational Statistics* 28: 519–539.

Duvvuri, S.D., V. Ansari, and S. Gupta. 2007. Consumers- price sensitivities across complementary categories. *Management Science* 53: 1933–1945.

Friedman, J.H. 1991. Multivariate adaptive regression splines. *The Annals of Statistics* 19: 1–67.

Gabel, S., and A. Timoshenko. 2022. Product choice with large assortments: A scalable deep-learning model. *Management Science* 68: 1808–1827.

Gabel, S., D. Guhl, and D. Klapper. 2019. P2V-MAP: Mapping market structures for large retail assortments. *Journal of Marketing Research* 56: 557–580.

Greene, W.H. 2003. *Econometric analysis*, 5th ed. Upper Saddle River: Pearson Education.

Guadagni, P.M., and J.D.C. Little. 1983. A logit model of brand choice calibrated on scanner data. *Marketing Science* 2: 203–238.

Hruschka, H. 2014. Analyzing market baskets by restricted Boltzmann machines. *OR Spectrum* 36: 209–228.

Hruschka, H. 2014. Linking multi-category purchases to latent activities of shoppers: Analysing market baskets by topic models. *Marketing ZFP* 36: 268–274.

Hruschka, H. 2017. Analyzing the dependences of multicategory purchases on interactions of marketing variables. *Journal of Business Economics* 87: 295–313.

Hruschka, H. 2024. Relevance of dynamic variables in multicategory choice models. *OR Spectrum* 46: 109–133.

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

Jacobs, B., B. Donkers, and D. Fok. 2016. Model-based purchase predictions for large assortments. *Marketing Science* 35: 389–404.

Kuhn, M., and K. Johnson. 2013. *Applied predictive modeling*. New York: Springer.

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

Porter, M.A., J.P. Onnela, and P.J. Mucha. 2009. Communities in networks. *Notices of the AMS* 56: 1082–1097.

Richards, T.J., S.F. Hamilton, and K. Yonezkawa. 2018. Retail market power in a shopping basket model of supermarket competition. *Journal of Retailing* 94: 328–342.

Ruczinski, I., C. Kooperberg, and M. Leblanc. 2003. Logic regression. *Journal of Computational and Graphical Statistics* 12: 475–511.

Russell, G.J., and A. Petersen. 2000. Analysis of cross category dependence in market basket selection. *Journal of Retailing* 76: 69–392.

Seetharaman, P.B., S. Chib, A. Anslie, et al. 2005. Models of multi-category choice behavior. *Marketing Letters* 16: 239–254.

Solnet, D., Y. Boztuğ, and S. Dolnicar. 2016. An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue. *The International Journal of Hospitality Management* 56: 119–125.

Talhouk, A., A. Doucet, and K. Murphy. 2012. Inference for multivariate probit models with sparse inverse correlation matrices. *The Journal of Computational and Graphical Statistics* 21: 739–757.

Train, K.E. 2003. *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.

Zellner, A. 1971. *An introduction to Bayesian inference in econometrics*. New York: Wiley.

Harald Hruschka is Professor Emeritus at the University of Regensburg, Germany, where he previously held a marketing char. His main research interests refer to choice and sales response modeling as well as to direct marketing. He has published in journals such as Journal of Forecasting, European Journal of Operational Research, OR Spectrum, Journal of Interactive Marketing, Marketing Letters, Journal of Retailing and Consumer Services, International Journal of Research in Marketing. Before joining the University of Regensburg,he was associate professor at the Vienna University of Business Administration and Economics, where he also obtained his Ph.D.